SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di Statistica e Metodi Quantitativi

Dottorato di Ricerca in Statistica e Matematica per la Finanza - Ciclo XXX

Curriculum: Statistica

Learning Markov Equivalence Classes of Gaussian DAGs via Observational and Interventional Data: an Objective Bayes Approach

Candidato: Federico Castelletti

Matricola: 798523

Tutor: Prof. Guido Consonni

Coordinatore: Prof. Giorgio Vittadini

Anno Accademico 2016/2017

Learning Markov Equivalence Classes of Gaussian DAGs via Observational and Interventional Data: an Objective Bayes Approach

Federico Castelletti

Contents

1 Introduction 1

1.1 Graphical models 1

1.2 Objective Bayes model comparison 4

1.3 Structure and main results of the thesis 9

2 Gaussian Graphical Models 11

2.1 Graph notation and background 11

2.2 Marginal data distribution 14

2.3 Fractional Bayes factor 17

2.3.1 General setting 17

2.3.2 FBF for Gaussian data 18

2.4 Priors for graphical model comparison 20

2.4.1 Comparison of DAG models 20

2.4.2 Comparison of decomposable UG models 22

2.4.3 Objective comparison of decomposable UG models 24

3 Model Comparison of Gaussian Essential Graphs 27

3.1 Markov equivalence and essential graphs 28

3.2 Gaussian Essential Graphs 34

3.2.1 Likelihood and prior factorization 34

3.2.2 Marginal likelihood 35

4 Model Comparison with Observational and Interventional Data 41

4.1 Interventional essential graphs 42
4.1.1 Interventions on DAGs 42
4.1.2 Interventions and Markov equivalence 43
4.1.3 Interventions and essential graphs 45
4.2 Gaussian interventional essential graphs 47
4.2.1 Likelihood and prior factorization 48
4.2.2 Marginal likelihood 50

5 MCMC Methods **55**
5.1 Markov chains on equivalence classes of DAGs 55
5.1.1 Operators on graphs 56
5.1.2 Markov chains and related properties 60
5.1.3 Exploring the EG space 62
5.1.4 Extension to the I-EG space 64
5.1.5 Algorithms and further details 70
5.2 MCMC for structural learning 75
5.2.1 Markov chains on I-EGs for model selection 75
5.2.2 MCMC scheme 76

6 Experiments **79**
6.1 Preliminaries 79
6.2 Simulations 81
6.2.1 Simulation setting 82
6.2.2 Data generation 82
6.2.3 MCMC parameters 83
6.2.4 Results 84
6.2.5 Comparison with GIES 89
6.3 Protein-signaling data 93
6.3.1 Data set 93
6.3.2 Model searching and results 95

7 Conclusions and Further Work **101**

A Some multivariate distributions **103**

A.1 Multivariate Normal distribution 103

A.2 Matrix Normal distribution 104

A.3 Wishart distribution 105

B Marginal distribution of Gaussian data: conjugate analysis **107**

B.1 Gaussian data with zero mean 107

B.1.1 Posterior 108

B.1.2 Marginal likelihood 109

B.2 Gaussian data with non-zero mean 109

B.2.1 Posterior 110

B.2.2 Marginal likelihood 111

B.3 Multivariate linear regression 112

B.3.1 Prior 113

B.3.2 Posterior 113

B.3.3 Marginal likelihood 114

C Marginal likelihood of Gaussian models: FBF setting **117**

C.1 Gaussian data with zero mean 117

C.1.1 Fractional prior 118

C.1.2 Posterior 118

C.1.3 Marginal likelihood 119

C.2 Gaussian data with non-zero mean 120

C.2.1 Fractional prior 120

C.2.2 Posterior 121

C.2.3 Marginal likelihood 122

C.3 Multivariate linear regression 123

C.3.1 Fractional prior 123

C.3.2 Posterior 124

C.3.3 Marginal likelihood 125

Abstract

Graphical models based on Directed Acyclic Graphs (DAGs) are a very common tool in many scientific areas for the investigation of dependencies among variables. Typically, the objective is to infer models from the data or measuring dependence relationships between variables. The set of all (marginal and) conditional independencies encoded by a DAG determines its *Markov property*. However, it is well known that we cannot distinguish between DAGs encoding the same set of conditional independencies (Markov equivalent DAGs) using observational data. Markov equivalent DAGs are then collected in equivalence classes each one represented by an Essential Graph (EG), also called Completed Partially Directed Graph (CPDAG). When the interest is in model selection it is then convenient to explore the EG space, rather than the whole DAG space, even if the number of EGs increases super-exponentially with the number of vertices. An exhaustive enumeration of all EGs is not feasible and so structural learning in the EG space has been confined to small dimensional problems. However, to avoid such limit, several methods based on Markov chains have been proposed in recent years.

In many applications (such as genomics) we have both observational and interventional data produced after an exogenous perturbation of some variables or from randomized intervention experiments. The concept of intervention is strictly related to the causal interpretation of a DAG. Interventions destroy the original *causal dependency* on the intervened variables and modify the Markov property of a DAG. This results in a finer partition of DAGs into equivalence classes, each one represented by an Interventional Essential Graph (I-EG). Hence, model se-

lection of DAGs in the presence of observational and interventional data can be performed over the I-EG space, thus improving the identifiability of the true data generating model.

In this work we deal with the problem of Gaussian DAG model selection from a Bayesian perspective. In particular, we adopt an objective Bayes approach based on the notion of fractional Bayes factor. We then obtain a closed formula to compute the marginal likelihood of an I-EG given a collection of observational and interventional data. Next, we construct a Markov chain to explore the I-EG space possibly accounting for *sparsity* constraints. Hence, we propose an MCMC algorithm to approximate the posterior distribution of I-EGs and provide a quantification of inferential uncertainty by measuring some features of interest, such as probabilities of edge inclusion. We apply our methodology, that we name Objective Bayesian Interventional Essential graph Search (OBIES) to simulation settings and to the analysis of protein-signaling data, where a collection of observations measured under different experimental conditions can be interpreted as interventional data.

List of Symbols

$\mathcal{G}$	graph
V	set of vertices of $\mathcal{G}$
E	set of edges of $\mathcal{G}$
q	number of nodes in a graph
$\vert\mathcal{G}\vert$	number of edges in $\mathcal{G}$
$\mathcal{D}$	DAG
$\mathrm{pa}_{\mathcal{G}}(u) = \Pi_u$	set of parents of node u in $\mathcal{G}$
$\mathrm{ne}_{\mathcal{G}}(u) = N_u$	set on neighbors of node u in $\mathcal{G}$
$\mathrm{fa}_{\mathcal{G}}(u)$	family of node u in $\mathcal{G}$
$\Omega_{u,v}$	$\Pi_u \cap N_v$
$\mathcal{G}_A$	subgraph of $\mathcal{G}$ induced by $A \in V$
$\mathcal{G}^{(u)}$	skeleton of $\mathcal{G}$
$\mathcal{G}^{<}$	perfect directed version of the decomposable graph $\mathcal{G}$
$\mathcal{T}$	set of chain components of a CG
$[\mathcal{D}]$	Markov equivalence class of $\mathcal{D}$
$\mathcal{G}(\mathcal{D})$	essential graph of $\mathcal{D}$
$f(\cdot)$	probability density function
$f_{\mathcal{G}}(\cdot)$	probability density function constrained by $\mathcal{G}$
$\boldsymbol{Y}$	data matrix
n	total sample size

I_k	intervention target
$Y_{I_k} \leftarrow U_{I_k}$	intervention on Y_{I_k}
$\mathcal{I}$	family of intervention targets
$\mathcal{D}^{I_k}$	intervention DAG
$[\mathcal{D}]_{\mathcal{I}}$	interventional Markov equivalence class of $\mathcal{D}$
${}_{\mathcal{I}}\mathcal{G}(\mathcal{D})$	$\mathcal{I}$-EG of $\mathcal{D}$
$\boldsymbol{Y}^k$	data matrix with interventional data associated to I_k
$n^{(k)}$	number of interventional data associated to I_k
$\boldsymbol{\Sigma}, \boldsymbol{\Omega}$	unconstrained covariance and precision matrices
$\boldsymbol{\Sigma}_{\mathcal{G}}, \boldsymbol{\Omega}_{\mathcal{G}}$	covariance and precision matrices constrained by $\mathcal{G}$
$m_{\mathcal{G}}(\cdot)$	marginal likelihood of $\mathcal{G}$
$\mathcal{S}_q$	set of all EGs on q nodes
r	sparsity constraint on the maximum number of nodes
$\mathcal{S}_q^r$	set of all EGs on q nodes satisfying $\lvert\mathcal{G}\rvert \leq rq$
$\mathcal{O}_{\mathcal{G}}$	perfect set of operators on $\mathcal{G}$
$\{\mathcal{G}^{(t)}\}$	Markov chain on the ($\mathcal{I}$)-EG space
$p_{\mathcal{G},\mathcal{G}'}$	transition probability from $\mathcal{G}$ to $\mathcal{G}'$
$q(\mathcal{G}\,\vert\,\cdot)$	proposal distribution for $\mathcal{G}$

UG	Undirected Graph
CG	Chain Graph
EG	Essential Graph
$\mathcal{I}$-EG	Interventional Essential Graph (or simply I-EG)
BF	Bayes Factor
FBF	Fractional Bayes Factor
OBIES	Objective Bayes Interventional Essential graph Search
GIES	Greedy Interventional Equivalence Search

Chapter 1

Introduction

In recent years, Directed Acyclic Graphs (DAGs) have become a very popular and attractive tool for the investigation of dependencies among variables. Applications of DAG models in various scientific areas abound, especially in genomics and biology; see for instance Sachs et al. (2005), Nagarajan et al. (2013). Typically, the true DAG generating model is unknown and then the objective is to perform model selection by scoring graphs belonging to a suitable model space.

In this Chapter we introduce two topics which are at the basis of the current work: graphical models and objective Bayes model comparison. In particular, in Section 1.1 we give an overview of graphical models both from a *non-causal* (observational) and *interventional* perspective. In Section 1.2 we then introduce objective Bayes methods for model comparison with reference to the most common techniques in literature. Finally, we present the structure and the main results of the thesis in Section 1.3.

1.1 Graphical models

Graphical models are widely used to represent dependency relationships among potentially many variables. One of the main objectives in graphical modelling is to discover connections between variables, or "learning a model from the data", where such model can be used as an approximation to our past experience (Friedman, 2004). This process is also called *structural learning*; see for instance Chickering (2002). From a probabilistic perspective, we can use a graph-based representation to encode a complex distribution over a high dimensional space (Koller & Friedman, 2009). A graph is made up of two components: a set of nodes (or vertices) and a set of edges. Nodes are associated to variables, while edges correspond to direct probabilistic interactions among them. Different typologies of graphical models are present in literature. Graphical models based on undirected graphs, also called *Markov random fields* are particularly used in spatial statistics (Besag, 1974). In the following, we focus on Directed Acyclic Graphs (DAGs). These are commonly used in many scientific areas, principally in biology and genomics; see for instance Friedman (2004), Sachs et al. (2005), Shojaie & Michailidis (2009).

A DAG encodes a set of independencies between variables which are of the form "Y is independent of X given Z". As an example, consider a set of variables $Y = \{Y_1, Y_2, Y_3, Y_4\}$ measuring four distinct gene expressions. One can be interested in discovering dependence relationships among them. Let $f(\cdot)$ be a joint probability distribution over Y and assume the representation through DAG $\mathcal{D}$ in Figure 1.1.

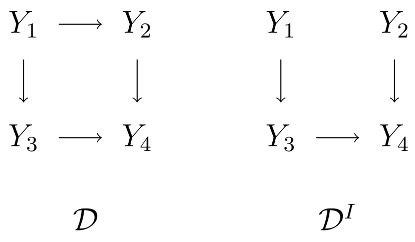


Figure 1.1: A DAG $\mathcal{D}$ representing dependence relationships among four variables and the intervention DAG $\mathcal{D}^I$ obtained after an intervention on Y_2.

As we associate each variable to a node in the DAG $\mathcal{D}$, we constrain the probability distribution of Y by the edges in $\mathcal{D}$. Hence, we can write

$$f_{\mathcal{D}}(y_1 \ldots, y_4) = f(y_1)f(y_2 \,|\, y_1)f(y_3 \,|\, y_1)f(y_4 \,|\, y_2, y_3). \tag{1.1}$$

All the (marginal and) conditional independencies between variables can be deduced from the DAG using the notion of *d-separation* (Pearl, 2000) or the *moral graph* representation (Lauritzen, 1996). For instance, in $\mathcal{D}$, Y_2 is independent of Y_3 given Y_1. Equivalently, we write $Y_2 \perp\!\!\!\perp Y_3 \,|\, Y_1$; see also Section 3.1 for details.

The factorization in Equation (1.1) determines the *Markov property* of DAG $\mathcal{D}$. It is well known that different DAGs can encode the same set of conditional independencies, and thus one cannot distinguish between *Markov equivalent* DAGs using observational data; see Chickering (2002). All DAGs encoding the same conditional independencies form a Markov equivalence class, which can be represented by a Completed Partially Directed Acyclic Graph (CPDAG) (Chickering, 2002), also called Essential Graph (EG) by Andersson et al. (1997). An EG is a particular chain graph whose chain components are decomposable undirected graphs linked by arrowheads; see Lauritzen (1996). Hence, an EG may contain both directed and undirected edges. In general, the number of undirected edges can be used as a measure of complexity of "causal learning" (He & Geng, 2008).

DAGs have also a natural causal interpretation; see for instance Pearl (2000). In addition, causal inference is strictly related to the concept of *intervention*. The causal structure of a DAG can be discovered by means of interventions on variables from randomized experiments or from exogenous perturbations of the true data generating model (Eberhardt & Scheines, 2007). In general, an intervention can be realized by "forcing the value of one or several random variables of the system to chosen values" (Hauser & Bühlmann, 2012). In doing so, we destroy the original *causal dependency* on the intervened variables. Assume for instance an intervention on Y_2 in DAG $\mathcal{D}$ of Figure 1.1. This results in the *intervention DAG* $\mathcal{D}^I$. Hence, the *post-intervention* distribution of Y can be obtained using the *do-operator* (Pearl, 1995) as

$$f_{\mathcal{D}^I}(y_1 \dots, y_4 \,|\, do(Y_2 = \tilde{y}_2)) = f(y_1)\tilde{f}(y_2)f(y_3 \,|\, y_1)f(y_4 \,|\, y_2, y_3). \tag{1.2}$$

A natural extension of the Markov equivalence property of DAGs, called *interventional Markov equivalence* is formalized in Hauser & Bühlmann (2012). They show that the interventional Markov equivalence property defines a finer partition of DAGs into equivalence classes, each one represented by an *interventional essential graph* which is still a chain graph with decomposable chain components. As discussed in Chapter 4, through interventions on variables it is then possible to *identify* the direction of undirected edges and then *distinguish* between (observationally) Markov equivalent DAGs.

In this work we generally assume that the structure of the DAG governing the joint distribution of Y is unknown. Hence, when the objective is to infer the true data generating (DAG) model, modelling jointly observational and interventional data can greatly improve the identifiability of the true underlying DAG. In this perspective, we perform graphical model selection by learning interventional Markov equivalence classes of DAGs. In particular, we tackle the problem through the Bayes factor (Kass & Raftery, 1995) by adopting an *objective Bayes* approach. In the following section we then give an overview of objective Bayes model comparison principles.

1.2 Objective Bayes model comparison

Let $Y_1, \ldots, Y_q$ be a collection of real valued random variables from which we observe n i.i.d. q-dimensional observations $\boldsymbol{y}_i$ $(i = 1, \ldots, n)$ collected in the data matrix $\boldsymbol{Y}$. A statistical model $\mathcal{M}$ consists in a probability density function $f_{\mathcal{M}}(\cdot)$ assigned to $Y_1, \ldots, Y_q$. We assume $f_{\mathcal{M}}(\cdot)$ belonging to some parametric family and then write $f_{\mathcal{M}}(y_1, \ldots, y_q \,|\, \boldsymbol{\theta}_{\mathcal{M}})$ where $\boldsymbol{\theta}_{\mathcal{M}}$ is a vector parameter taking values in the parametric space $\boldsymbol{\Theta}_{\mathcal{M}}$. Moreover, a Bayesian model can be expressed through the joint distribution

$$f(y_1, \ldots, y_q, \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}) = f_{\mathcal{M}}(y_1, \ldots, y_q \,|\, \boldsymbol{\theta}_{\mathcal{M}}) p(\boldsymbol{\theta}_{\mathcal{M}}) p(\mathcal{M})$$

which encodes assumptions on the sampling distribution of the data, $f_{\mathcal{M}}(y_1, \ldots, y_q \,|\, \boldsymbol{\theta}_{\mathcal{M}})$, together with a prior *belief* on $\boldsymbol{\theta}_{\mathcal{M}}$, $p(\boldsymbol{\theta}_{\mathcal{M}})$, and the model itself, $p(\mathcal{M})$. Such factorization is particularly relevant when the interest is in model selection. To this end, suppose to have a collection of K different models, $\mathcal{M}_1, \ldots, \mathcal{M}_K$.

Given the collection of observations $\boldsymbol{Y}$, we might be interested in computing the posterior model probability of $\mathcal{M}_k$ for each $k = 1, \ldots, K$,

$$p(\mathcal{M}_k \,|\, \boldsymbol{Y}) = \frac{m(\boldsymbol{Y} \,|\, \mathcal{M}_k)p(\mathcal{M}_k)}{\sum_k m(\boldsymbol{Y} \,|\, \mathcal{M}_k)p(\mathcal{M}_k)} \propto m(\boldsymbol{Y} \,|\, \mathcal{M}_k)p(\mathcal{M}_k),$$

which is proportional to the product of two terms. The first, $m(\boldsymbol{Y} \,|\, \mathcal{M}_k)$, is the marginal distribution of the data $\boldsymbol{Y}$ under model $\mathcal{M}_k$ (also called *marginal likelihood of* $\mathcal{M}_k$ when the emphasis is on model $\mathcal{M}_k$) which is defined as

$$m(\boldsymbol{Y} \,|\, \mathcal{M}_k) = \int\limits_{\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k} f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k,$$

where $f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k) = \prod_{i=1}^{n} f_k(\boldsymbol{Y}_i \,|\, \boldsymbol{\theta}_k)$ is the sampling distribution of the data. The marginal data distribution typically represents the most demanding element for the computation of posterior model probabilities. The second term is the prior probability of model $\mathcal{M}_k$ which can be set in absence of substantive information to $1/K$ but is often "modified" in other cases, such as regression variable selection or graphical modelling; see also Chapter 5. When the objective is the comparison of two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we can consider the ratio of posterior model probabilities

$$\frac{p(\mathcal{M}_1 \,|\, \boldsymbol{Y})}{p(\mathcal{M}_2 \,|\, \boldsymbol{Y})} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \cdot \frac{f_1(\boldsymbol{Y} \,|\, \mathcal{M}_1)}{f_2(\boldsymbol{Y} \,|\, \mathcal{M}_2)}.$$

This is made up by two terms; the second one, which is usually the dominant term especially when the sample size n is large, is the so-called *Bayes factor* of model $\mathcal{M}_1$ against $\mathcal{M}_2$, $\text{BF}_{1,2}(\boldsymbol{Y})$ for short; see also O'Hagan & Forster (2004).

A Bayesian approach for model selection requires the specification of prior distributions for each model parameter. If a *subjective* approach is adopted, these priors should be based on some prior information gathered from past experience or from experts' knowledge. Pericchi (2005) underlines some limits of pure subjective Bayesian analysis. The first one concerns prior elicitations; when the number of parameters is large, a huge number of prior elicitations can be problematic especially in lack of substantive prior information. This problem is not negligible in graphical modelling where the number of parameters typically grows *exponentially* in the number of variables. The second one is related to model comparison via Bayes factors which can be "very nonrobust with respect to seemingly innocent prior". At last, subjective approaches typically prevent connectivity (and then compatibility) of the conditional priors. We are now able to introduce the main principles of objective Bayes model comparison.

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a collection of K distinct models for the data $\boldsymbol{Y}$, $p_k^N(\boldsymbol{\theta}_k)$, $k = 1, \ldots, K$, some ordinary (objective) non informative priors. The corresponding marginal distribution of $\boldsymbol{Y}$ is then

$$m^N(\boldsymbol{Y} \,|\, \mathcal{M}_k) = \int\limits_{\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k} f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k) p^N(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k.$$

It is well known that objective priors are often improper and then defined only up to an arbitrary constants. Hence, they cannot be used naively to compute marginal likelihoods and Bayes factors; see Berger et al. (2001).

In the last decades several objective Bayes methods were proposed. Among these we mention the Intrinsic Bayes Factor (IBF), the Expected Posterior Prior (EPP) and the Fractional Bayes Factor (FBF) approaches. The latter is at the basis of the methodology developed in this work, while the others represent valid (although not trivial) alternative approaches. Most of the objective Bayes methods are based on the notion of *minimal training sample*. Let $\boldsymbol{Y}(l)$ be a subset of the sample $\boldsymbol{Y}$. $\boldsymbol{Y}(l)$ is proper if $0 < m^N(\boldsymbol{Y}(l) \,|\, \mathcal{M}_k) < \infty$ for all $\mathcal{M}_k$; moreover, $\boldsymbol{Y}(l)$ is minimal if it is proper and no subset is proper.

We start from the intrinsic Bayes factor approach. Let $\boldsymbol{Y}(l)$, $l = 1, \ldots, L$, be a collection of proper and minimal training samples of size n_0 "as small as possible", so that the updated posterior $m^N(\boldsymbol{Y}(l) \,|\, \mathcal{M}_k)$ under each model $\mathcal{M}_k$ becomes proper. Then, $\boldsymbol{Y}(l)$ are used to convert the improper priors $p_k^N(\boldsymbol{\theta}_k)$ to proper posteriors $p_k^N(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}(l))$. The Bayes factor of model $\mathcal{M}_k$ against $\mathcal{M}_{k'}$ for the rest of the data $\boldsymbol{Y}(-l)$ is then

$$\mathrm{BF}_{k,k'}(\boldsymbol{Y}, l) = \frac{\int f_k(\boldsymbol{Y}(-l) \,|\, \boldsymbol{\theta}_k, \boldsymbol{Y}(l)) p_k^N(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}(l)) d\boldsymbol{\theta}_k}{\int f_{k'}(\boldsymbol{Y}(-l) \,|\, \boldsymbol{\theta}_{k'}, \boldsymbol{Y}(l)) p_{k'}^N(\boldsymbol{\theta}_{k'} \,|\, \boldsymbol{Y}(l)) d\boldsymbol{\theta}_{k'}}.$$

$\mathrm{BF}_{k,k'}(\boldsymbol{Y},l)$ depends on the choice of l. Hence, $\mathrm{BF}_{k,k'}(\boldsymbol{Y},l)$ are "averaged" over all the possible training samples $\boldsymbol{Y}(l)$, $l=1,\ldots,L$, usually by taking the arithmetic mean or the median. In the first case we get

$$\mathrm{BF}^{I}_{k,k'}(\boldsymbol{Y}) = \frac{1}{L}\sum_{l=1}^{L}\mathrm{BF}_{k,k'}(\boldsymbol{Y},l). \tag{1.3}$$

The result is the so-called intrinsic Bayes factor; see Berger & Pericchi (1996) for further details.

We now consider the expected posterior prior approach. Let $\boldsymbol{Y}^*$ be an *imaginary* training sample generated by $f^*(\cdot)$, so that $f^*(\boldsymbol{Y}^*)$ is the density of $\boldsymbol{Y}^*$. We start assuming noninformative priors $p_k^N(\boldsymbol{\theta}_k)$, $k=1,\ldots,K$. The posterior distribution of $\boldsymbol{\theta}_k$ given $\boldsymbol{Y}^*$ is then

$$p_k^*(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}^*) = \frac{f_k(\boldsymbol{Y}^* \,|\, \boldsymbol{\theta}_k, \boldsymbol{Y}^*)p_k^N(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}^*)}{\int f_k(\boldsymbol{Y}^* \,|\, \boldsymbol{\theta}_k, \boldsymbol{Y}^*)p_k^N(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}^*)d\boldsymbol{\theta}_k}.$$

The *expected posterior prior* (EPP) for model $\mathcal{M}_k$ is obtained as an expectation with respect to $f^*(\boldsymbol{Y}^*)$ of the posteriors $p_k^*(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}^*)$,

$$p_k^*(\boldsymbol{\theta}_k) = \int p_k^*(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}^*)f^*(\boldsymbol{Y}^*)d\boldsymbol{Y}^*,$$

where $f^*(\boldsymbol{Y}^*)$ is common for all models. See Perez & Berger (2002) for details.

The intrinsic prior approach was introduced by Berger & Pericchi (1996). Let $\mathcal{M}_k$, $k=1,2$, be two nested models, each one consisting in a density $f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k)$ and a conventional, typically improper prior $p_k^N(\boldsymbol{\theta}_k)$. Suppose that $f_1(\cdot)$ is nested in $f_2(\cdot)$. We can start by observing that the intrinsic Bayes factor as defined in Equation (1.3) is not *properly* a Bayes factor. This means that the condition

$$\mathrm{BF}^{I}_{k,k'}(\boldsymbol{Y}) = \frac{1}{\mathrm{BF}^{I}_{k',k}(\boldsymbol{Y})}$$

is not in general satisfied. Hence, we call *intrinsic priors*, $p_k^I(\boldsymbol{\theta}_k)$, $k=1,2$, any two priors such that

$$\mathrm{BF}^{I}_{k,k'}(\boldsymbol{Y}) = \mathrm{BF}_{k,k'}(\boldsymbol{Y})$$

as $n \to \infty$, being

$$\mathrm{BF}_{k,k'}(\boldsymbol{Y}) = \frac{\int f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k, \boldsymbol{Y}) p_k^I(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y}) d\boldsymbol{\theta}_k}{\int f_{k'}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{k'}, \boldsymbol{Y}) p_{k'}^I(\boldsymbol{\theta}_{k'} \,|\, \boldsymbol{Y}) d\boldsymbol{\theta}_{k'}};$$

see also Casella & Moreno (2006) for details.

The fractional Bayes factor approach for model comparison was first introduced by O'Hagan (1995). Let $p_k^N(\boldsymbol{\theta}_k)$ be a default prior for model $\mathcal{M}_k$, $b \in (0,1)$. Basically, the FBF approach uses a fraction of the whole likelihood, $\{f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k)\}^b$ to make the improper prior $p_k^N(\boldsymbol{\theta}_k)$ proper. The *implied fractional prior* thus obtained is then combined with the discounted likelihood $\{f_k(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k)\}^{1-b}$ to compute the posterior $p_k(\boldsymbol{\theta}_k \,|\, \boldsymbol{Y})$ under each model $\mathcal{M}_k$. A common choice is $b = n_0/n$, $1 < n_0 < n$, where n_0 is the minimum training sample size such that the implied fractional prior is proper. The main difference between the first two approaches is that the FBF does not use training samples to compute Bayes factors but rather a *discounted* likelihood, which requires only the choice of n_0, which is typically equal for each model $\mathcal{M}_k$. A more detailed exposition of the FBF approach is given in Section 2.3 of the current work.

An alternative approach, although not "purely objective" but still based on discounted likelihoods, concerns the so-called *power priors*, which were originally proposed by Ibrahim & Chen (2000) in a regression context. Suppose we have some historical data $\boldsymbol{Y}_0$ of size n_0 from some previous study. The power prior is defined as the normalized likelihood function based on the historical data raised to a power a_0, where $0 < a_0 < 1$. Typically, as in the FBF approach, a_0 is set as small as possible so that the dependence on historical data is *weak*. The implied power prior is then combined with the "complete" likelihood based on $\boldsymbol{Y}$ to compute posterior distributions. A recent approach which combines power priors and EPP, named *power-expected-posterior priors*, is presented in Fouskakis et al. (2015) within the context of variable selection for Gaussian linear models.

1.3 Structure and main results of the thesis

In this thesis we approach the problem of model selection of Gaussian graphical models from an objective Bayes perspective. We consider a multivariate setting where we collect both observational and interventional data, the latter produced after an exogenous perturbation of the data generating model. To this end, we directly score Markov equivalence classes of DAGs by deriving a closed formula for the marginal likelihood of an interventional essential graph. Our approach relies on the methodology for model comparison of decomposable undirected graphs developed by Consonni & La Rocca (2012) and Consonni et al. (2017). This is extended to model comparison of essential graphs in the presence of observational and interventional data. Then, we propose an MCMC strategy to explore the interventional essential graph space and perform structural learning of Markov equivalence classes of DAGs. The rest of the thesis is organized as follows.

In Chapter 2 we first summarize some notation and background on graphical models. Then, we focus on Gaussian multivariate models. We give closed formulas for the computation of the marginal distribution of Gaussian data using proper (subjective) priors under different sampling assumptions. Next, we consider the same problem from an objective perspective using the fractional Bayes factor approach. We pay particular attention to the choice of priors for model comparison of Gaussian DAG models and show how the previous results can be used to compute the marginal likelihood of any DAG model.

In Chapter 3 we introduce the Markov equivalence property of DAGs and the notion of essential graph. We use the results of Chapter 2 to compute the marginal likelihood of an essential graph given a collection of observational data. In Chapter 4 we discuss interventions on DAGs together with interventional Markov equivalence, as introduced by Hauser & Bühlmann (2012). We extend the methodology of Chapter 3 for the computation of the marginal likelihood of an interventional essential graph in the presence of both observational and interventional data.

In Chapter 5 we introduce Markov chains on Markov equivalence classes of

DAGs based on the proposal of He et al. (2013). These are used for the exploration of the interventional essential graph space and for the construction of an MCMC algorithm to perform structural learning of Markov equivalence classes of DAGs.

In Chapter 6 we apply the proposed methodology to simulation settings and to the analysis of the protein-signaling data of Sachs et al. (2005). Finally, we present in Chapter 7 some conclusions and possible future developments. In Appendix we also collect some useful definitions and theoretical results.

Chapter 2

Gaussian Graphical Models

In this chapter we introduce Gaussian graphical models. To this end, we first resume in Section 2.1 some notation and theory about graphs which must be intended as supporting material for the topics presented in this work; for an exhaustive introduction to graphical models see for instance Pearl (2000). Then, we move to a Gaussian framework (Section 2.2). Here we assume that the normally distributed random variable Y is the responsible of the data generating process of a collection of i.i.d. observations. Starting from proper (informative) priors for the *unconstrained* model parameters (mean vector and precision matrix), we compute the marginal distribution of such Gaussian data. Next, we deal with the same problem from an *objective Bayes* perspective, relying on the notion of *fractional Bayes factor* (Section 2.3). By exploiting the methodology for prior construction resumed in Section 2.4, the previous results are then used to compute the marginal likelihood of Gaussian DAG models, that is when the sampling distribution (and so the precision matrix of Y) is *constrained* by a DAG.

2.1 Graph notation and background

A graph $\mathcal{G}$ is a pair (V, E) where $V = \{1, \dots, q\}$ is a set of vertices (or nodes) and $E \subseteq V \times V$ a set of edges. Let $u, v \in V$, $u \neq v$. If $(u, v) \in E$ and $(v, u) \notin E$ we say that $\mathcal{G}$ contains the directed edge $u \to v$. If instead $(u, v) \in E$ and $(v, u) \in E$

we say that $\mathcal{G}$ contains the undirected edge $u - v$. Two vertices u, v are adjacent if they are connected by an edge (directed or undirected). Moreover, if $u - v$ is in $\mathcal{G}$ we say that u is a *neighbor* of v in $\mathcal{G}$. The neighbor set of v is denoted by $\mathrm{ne}_{\mathcal{G}}(v)$; the common neighbor set of u and v is then $\mathrm{ne}_{\mathcal{G}}(u, v) = \mathrm{ne}_{\mathcal{G}}(u) \cap \mathrm{ne}_{\mathcal{G}}(v)$. For any pair of distinct nodes $u, v \in V$ we say that u is a *parent* of v if $u \rightarrow v$. Conversely, we say that v is a son of u. The set of all parents of u in $\mathcal{G}$ is denoted by $\mathrm{pa}_{\mathcal{G}}(u)$.

A graph is called *directed* (*undirected*, UG) if it contains only directed (undirected) edges. A sequence of distinct vertices $\{v_0, v_1, \ldots, v_k\}$ in $\mathcal{G}$ is a *path* from v_0 to v_k if $\mathcal{G}$ contains $v_{j-1} - v_j$ or $v_{j-1} \rightarrow v_j$ for all $j = 1, \ldots, k$. A path is directed (undirected) if all edges are directed (undirected). Moreover, we say that a path is *partially directed* if it contains at least one directed edge. If there exists a path from v_0 to v_k we also say that v_k is a *descendant* of v_0. A sequence of nodes $\{v_0, v_1, \ldots, v_k\}$ with $v_0 = v_k$ such that $v_{j-1} - v_j$ or $v_{j-1} \rightarrow v_j$ for all $j = 1, \ldots, k$ is called a *cycle*. A cycle is then directed (undirected) if it contains only directed (undirected) edges. Let $A \subseteq V$. We denote with $\mathcal{G}_A = (A, E_A)$ the *subgraph* of $\mathcal{G} = (V, E)$ induced by A, whose edge set $E_A = \{(u, v) \in V \mid u \in A, v \in A\}$. An undirected (sub)graph is complete if its vertices are all adjacent.

A particular class of undirected graphs is represented by *decomposable* graphs, also called *chordal* or *triangulated*; see for instance Lauritzen (1996). An undirected graph is decomposable if every cycle of length $l \geq 4$ has a *chord*, that is two nonconsecutive adjacent vertices. For a decomposable graph $\mathcal{G}$ on the set of vertices V, a complete subset that is maximal with respect to inclusion is called a *clique*; see for instance graph $\mathcal{G}$ in Figure 2.1. Let $\mathcal{C} = \{C_1, \ldots, C_K\}$ be a *perfect* sequence of cliques of the decomposable graph $\mathcal{G}$ (Lauritzen, 1996, p.18). We introduce for $k = 2, \ldots, K$ the three types of sets

$$\begin{aligned} H_k &= C_1 \cup \cdots \cup C_k, \\ S_k &= C_k \cap H_{k-1}, \\ R_k &= C_k \setminus H_{k-1}, \end{aligned}$$

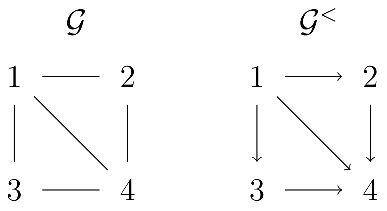


Figure 2.1: A decomposable graph $\mathcal{G}$ on the set of vertices $V = \{1, 2, 3, 4\}$; the cycle $\{1, 2, 4, 3\}$ of length $l = 4$ contains the chord $1 - 4$. $\mathcal{G}$ has the perfect sequence of cliques $\{C_1, C_2\}$, with $C_1 = \{1, 2, 4\}, C_2 = \{1, 3, 4\}$, and then $H_2 = V, S_2 = \{1, 4\}, R_2 = \{3\}$. $\mathcal{G}^<$ is the perfect directed version of $\mathcal{G}$.

which are called *history*, *separators* and *residuals* respectively, and set $R_1 = H_1 = C_1, S_1 = \emptyset$. Note that $C_1 \cup R_2 \cup \cdots \cup R_K = V$ and also $R_k \cap R_{k'} = \emptyset$. Moreover, it is possible to number the vertices of a decomposable graph starting from those in C_1, then those in R_1, R_2 and so on. In doing so we obtain a *perfect numbering of vertices*; see again Lauritzen (1996). Given a perfect numbering of the vertices in $\mathcal{G}$ we can construct its *perfect directed version* $\mathcal{G}^<$ by directing its edges from lower to higher numbered vertices; see also Figure 2.1.

A graph with only directed edges is called a Directed Acyclic Graph (DAG for short, denoted by $\mathcal{D}$) if it does not contain cycles. Let u, v be two distinct vertices of a DAG. If there exists a (directed) path from u to v but no paths from v to u, we say that u is an *ancestor* of v; conversely, v is a *descendant* of u. We then denote with $\text{an}(v)$ and $\text{de}(v)$ the set of all ancestors and descendants of v respectively.

A graph with no semi-directed cycles that may contain both directed and undirected edges is called a *chain graph* (CG) or simply *partially directed acyclic graph* (PDAG). For a chain graph $\mathcal{G}$ we call *chain component* $\tau \subseteq V$ a set of nodes that are joined by an undirected path. The set of chain components of a CG is denoted by $\mathcal{T}$. See for instance Figure 2.2. A subgraph of the form $u \rightarrow z \leftarrow v$, where there are no edges between u and v, is called a *v-structure* (or *immorality*). See for example $5 \rightarrow 6 \leftarrow 7$ in Figure 2.2. The *skeleton* of a graph $\mathcal{G}$ is the

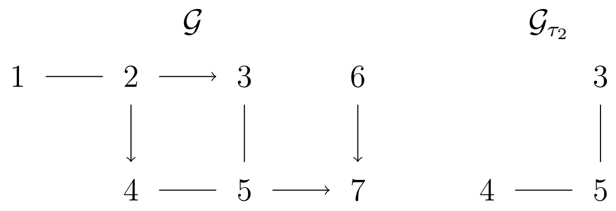


Figure 2.2: A chain graph $\mathcal{G}$ with set of chain components $\mathcal{T} = \{\tau_1, \tau_2, \tau_3, \tau_4\}$, $\tau_1 = \{1, 2\}, \tau_2 = \{3, 4, 5\}, \tau_3 = \{6\}, \tau_4 = \{7\}$; $\mathcal{G}_{\tau_2}$ is the subgraph induced by τ_2.

undirected graph on the same set of vertices obtained by removing the orientation of all its edges.

2.2 Marginal data distribution

In this section we focus on Gaussian models. We consider a multivariate framework where $Y = (Y_1, \dots, Y_q)^\top$ is a *q-dimensional* random variable and $\boldsymbol{Y}$ a $n \times q$ data matrix of i.i.d. observations from Y. $\boldsymbol{y}_i$ denotes the *i-th* row in $\boldsymbol{Y}$, $i = 1, \dots, n$. We analyse three cases. In the first one we consider Gaussian data with zero expectation, where the only parameter on interest is the covariance (or precision) matrix of Y. In the second the approach is extended to Gaussian data with non-zero mean, while in the last we consider a multivariate linear regression model and then allows for the presence of explanatory variables (covariates) too. In all the cases we assume that the precision matrix governing the distribution of $Y \sim \mathcal{N}(\cdot)$ is unconstrained, that is it has no zero entries; equivalently, there are no conditional independencies between the Y_js. This corresponds to the case in which $\boldsymbol{\Omega}$ is Markov with respect to a complete DAG; see also Chapter 3.

The objective is the computation of $m(\boldsymbol{Y}_A)$, the marginal data distribution of the $n \times |A|$ matrix $\boldsymbol{Y}_A$ containing columns indexed by $A \subseteq \{1, \dots, q\}$ in $\boldsymbol{Y}$, which is of particular interest for the current work. All the results are obtained using conjugate priors for model parameters based on Wishart and Normal-Wishart distributions. For references see Geisser & Cornfield (1963) or Gelman et al.

(2004). General definitions and detailed proofs are also reported in Appendix..

Gaussian data with zero expectation

Assume first that

$$(Y_1, \dots, Y_q) \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Omega}^{-1}),$$

where $\boldsymbol{\Omega}$ is a symmetric and positive definite (s.p.d.) but otherwise unconstrained $q \times q$ precision matrix. Assume

$$\boldsymbol{\Omega} \sim \mathcal{W}_q\big(a, \boldsymbol{R}\big),$$

with $a \in \mathbb{R}$, $a > q - 1$, and $\boldsymbol{R}$ a $q \times q$ s.p.d. matrix. Let $A \subseteq \{1, \dots q\}$ and $\bar{A}$ its complement. Then

$$m(\boldsymbol{Y}_A) \;=\; (\pi)^{-\frac{n|A|}{2}} \cdot \frac{\Gamma_{|A|}\big(\frac{a-|\bar{A}|+n}{2}\big)}{\Gamma_{|A|}\big(\frac{a-|\bar{A}|}{2}\big)} \cdot \frac{|\boldsymbol{R}_{AA}|^{\frac{a-|\bar{A}|}{2}}}{|\boldsymbol{R}_{AA} + \boldsymbol{S}_{AA}|^{\frac{a-|\bar{A}|+n}{2}}}, \tag{2.1}$$

where $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^\top$ and $\boldsymbol{S}_{AA}$ denotes the $|A| \times |A|$ matrix containing rows and columns indexed by A in $\boldsymbol{S}$; similarly for $\boldsymbol{R}_{AA}$. See also Appendix B.1.

Gaussian data with non-zero expectation

Assume now that

$$(Y_1, \dots, Y_q) \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}),$$

where $\boldsymbol{\mu}$ is the $q \times 1$ mean vector and $\boldsymbol{\Omega}$ the $q \times q$ unconstrained precision matrix. Assume

$$\begin{aligned} \boldsymbol{\mu} \,|\, \boldsymbol{\Omega} \;&\sim\; \mathcal{N}_q\big(\boldsymbol{m}_0, (a_\mu \boldsymbol{\Omega})^{-1}\big), \\ \boldsymbol{\Omega} \;&\sim\; \mathcal{W}_q\big(a_\Omega, \boldsymbol{R}\big), \end{aligned}$$

with $\boldsymbol{m}_0$ a $q \times 1$ vector, $a_\mu, a_\Omega \in \mathbb{R}$ $(a_\Omega > q - 1)$ and $\boldsymbol{R}$ a $q \times q$ s.p.d. matrix. Given $A \subseteq \{1, \dots q\}$ and $\bar{A}$ its complement, we have

$$\begin{aligned} m(\boldsymbol{Y}_A) \;=\; & (\pi)^{-\frac{n|A|}{2}} \cdot \left(\frac{a_\mu}{a_\mu + n}\right)^{\frac{|A|}{2}} \cdot \frac{\Gamma_{|A|}\big(\frac{a_\Omega-|\bar{A}|+n}{2}\big)}{\Gamma_{|A|}\big(\frac{a_\Omega-|\bar{A}|}{2}\big)} \\ & \cdot \; \frac{|\boldsymbol{R}_{AA}|^{\frac{a_\Omega-|\bar{A}|}{2}}}{\big|\boldsymbol{R}_{AA} + \boldsymbol{S}_{AA} + \frac{a_\mu n}{a_\mu+n}\boldsymbol{S}_{0,AA}\big|^{\frac{a_\Omega-|\bar{A}|+n}{2}}}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\top, \boldsymbol{S}_0 = (\bar{\boldsymbol{y}} - \boldsymbol{m}_0)(\bar{\boldsymbol{y}} - \boldsymbol{m}_0)^\top$ and $\bar{\boldsymbol{y}}$ the $q \times 1$ vector of sample means. $\boldsymbol{S}_{0,AA}$ is then the $|A| \times |A|$ matrix with rows and columns indexed by A in $\boldsymbol{S}_0$; similarly for $\boldsymbol{R}_{AA}$ and $\boldsymbol{S}_{AA}$. See also Appendix B.2.

Gaussian multivariate linear regression

Let $\boldsymbol{Y}$ be a $n \times q$ matrix of responses from $Y_1, \dots, Y_q$, $\boldsymbol{X}$ a $n \times (p+1)$ matrix of observations from a set of p explanatory variables (including the unit vector for the intercept) and $\boldsymbol{B}$ a $(p+1) \times q$ matrix of coefficients describing the effect of the explanatory variables on the responses. A Gaussian multivariate linear regression model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

where $\boldsymbol{E}$ is a $n \times q$ matrix of error terms, $\boldsymbol{E} \sim \mathcal{N}_{n,q}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1})$, $\boldsymbol{I}_n$ the $n \times n$ identity matrix, $\boldsymbol{\Omega}$ the unconstrained column precision matrix and $\boldsymbol{0}$ is the $n \times q$ null mean matrix. Equivalently, we can write using the matrix Normal notation (Appendix A.2)

$$\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}).$$

Assume now

$$\begin{aligned} \boldsymbol{B} \,|\, \boldsymbol{\Omega} \;&\sim\; \mathcal{N}_{p+1,q}\big(\underline{\boldsymbol{B}}, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}^{-1}\big), \\ \boldsymbol{\Omega} \;&\sim\; \mathcal{W}_q\big(a, \boldsymbol{R}\big), \end{aligned}$$

with $\underline{\boldsymbol{B}}$ a $(p+1) \times q$ matrix, $\boldsymbol{C}$ $(p+1) \times (p+1)$, $a \in \mathbb{R}$ $(a > q-1)$ and $\boldsymbol{R}$ a $q \times q$ s.p.d. matrix. Let $A \subseteq \{1, \dots q\}$, $\bar{A}$ its complement. Then

$$m(\boldsymbol{Y}_A \,|\, \boldsymbol{X}) = \frac{|\boldsymbol{C}|^{\frac{|A|}{2}} |\boldsymbol{R}_{AA}|^{\frac{a-|\bar{A}|}{2}} \Gamma_q\big(\frac{a-|\bar{A}|+n}{2}\big) 2^{\frac{|A|n}{2}}}{(2\pi)^{\frac{n|A|}{2}} |\boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}|^{\frac{|A|}{2}} |\boldsymbol{R}_{AA} + \hat{\boldsymbol{E}}_A^\top \hat{\boldsymbol{E}}_A + \boldsymbol{D}_{AA}|^{\frac{a-|\bar{A}|+n}{2}} \Gamma_{|A|}\big(\frac{a-|\bar{A}|}{2}\big)}, \tag{2.3}$$

where

$$\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}, \quad \hat{\boldsymbol{B}} = \big(\boldsymbol{X}^\top \boldsymbol{X}\big)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

and $\hat{\boldsymbol{E}}_A = \boldsymbol{Y}_A - \boldsymbol{X}\hat{\boldsymbol{B}}_A$, being $\hat{\boldsymbol{B}}_A$ the $n \times |A|$ submatrix of $\hat{\boldsymbol{B}}$ with columns indexed by A. See Appendix B.3 for details.

2.3 Fractional Bayes factor

In this section we briefly resume the Fractional Bayes Factor (FBF) as originally introduced by O'Hagan (1995); see also O'Hagan & Forster (2004). We then adopt the FBF approach to compute the marginal likelihood of the three Gaussian models described in Section 2.2, starting from objective (default) priors.

2.3.1 General setting

Let $\mathcal{M}_1, \mathcal{M}_2$ be two distinct models for the data $\boldsymbol{Y}$. In a Bayesian model comparison framework (Section 1.2), we might be interested in computing the Bayes factor of $\mathcal{M}_1$ over $\mathcal{M}_2$,

$$\mathrm{BF}(\mathcal{M}_1, \mathcal{M}_2) = \frac{f(\boldsymbol{Y} \mid \mathcal{M}_1)}{f(\boldsymbol{Y} \mid \mathcal{M}_2)} = \frac{m_1(\boldsymbol{Y})}{m_2(\boldsymbol{Y})},$$

where

$$m_k(\boldsymbol{Y}) = \int f(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k$$

is the marginal likelihood of model $\mathcal{M}_k$ $(k = 1, 2)$ given the data $\boldsymbol{Y}$. To this end we focus on the computation of $m_k(\boldsymbol{Y})$. In lack of substantive prior information, we would like to set $p(\boldsymbol{\theta}_k) = p^D(\boldsymbol{\theta}_k)$, where the latter is some objective default (non-informative) parameter prior. As mentioned in Section 1.2, objective priors are often improper and cannot be naively used to compute marginal likelihoods. Let $b = b(n)$, $0 < b < 1$, be a fraction of the number of observations n. The *fractional marginal likelihood* of model $\mathcal{M}_k$ is defined as

$$m_k(\boldsymbol{Y}; b) = \frac{\int f_k(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^D(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}{\int f_k^b(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^D(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k}, \tag{2.4}$$

where $f_k^b(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) = \{f_k(\boldsymbol{Y} \mid \boldsymbol{\theta}_k)\}^b$ is the sampling density under model $\mathcal{M}_k$ raised to the b-th power (fractional likelihood) and the two integrals are assumed to be finite and non-zero. Equation (2.4) can be rewritten as

$$m_k(\boldsymbol{Y}; b) = \int f_k^{1-b}(\boldsymbol{Y} \mid \boldsymbol{\theta}_k) p^F(\boldsymbol{\theta}_k \mid b, \boldsymbol{Y}) d\boldsymbol{\theta}_k,$$

where $p^F(\boldsymbol{\theta}_k \,|\, b, \boldsymbol{Y}) \propto f_k^b(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_k)p^D(\boldsymbol{\theta}_k)$ is the implied *fractional prior* (actually a "posterior" based on the fractional likelihood and the default prior). In other terms, we first consider a fraction b of the data $\boldsymbol{Y}$ which is used to make the default prior proper. The latter is then combined with the residual $(1-b)$ fraction of the likelihood to compute the posterior. The fractional prior depends on the choice of b, that we would like to take as small as possible so that the dependence of the prior on the data will be weak. O'Hagan suggests the default choice $b = n_0/n$ where n_0 is the minimal training sample size such that the fractional marginal likelihood is well defined. Other choices are possible, but Moreno (1997) argues in favor of the default choice. For general properties of the fractional Bayes factor see also O'Hagan (1995).

2.3.2 FBF for Gaussian data

Of particular interest for the current work is the use of the FBF to compute the marginal distribution of Gaussian data. To this end, we analyse the three cases presented in Section 2.2 from an objective Bayes perspective where parameter priors are obtained using the fractional Bayes factor. Default priors $p^D(\cdot)$ herein adopted assure the implied fractional priors to be conjugate to the Gaussian models. The objective is again the computation of $m(\boldsymbol{Y}_A)$, the marginal data distribution of the $n \times |A|$ matrix $\boldsymbol{Y}_A$ containing columns indexed by $A \subseteq \{1, \dots, q\}$ in $\boldsymbol{Y}$. For a detailed discussion about the use of FBFs in the more general framework of exponential families see Consonni & La Rocca (2012). Detailed proofs reported in Appendix C can also be found in Consonni & La Rocca (2012) and Consonni et al. (2017) within the context of Gaussian data with non-zero mean and multivariate regression respectively.

Gaussian data with zero expectation

Consider the case in which

$$(Y_1, \dots, Y_q) \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}),$$

where $\boldsymbol{\Omega}$ is the $q \times q$ unconstrained precision matrix. We start assuming the default prior for $\boldsymbol{\Omega}$,

$$p^D(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}},$$

(Geisser & Cornfield, 1963) and setting the fraction b equal to n_0/n, $n_0 < n$. Given $A \subseteq \{1, \dots, q\}$ and $\bar{A}$ its complement, we obtain

$$m^F(\boldsymbol{Y}_A) = (\pi)^{-\frac{(n-n_0)|A|}{2}} \cdot \frac{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n}{2}\right)}{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n_0}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{|A|(a_D-|\bar{A}|+n_0)}{2}} \cdot |\boldsymbol{S}_{AA}|^{-\frac{n-n_0}{2}}, \quad (2.5)$$

where $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{y}_i \boldsymbol{y}_i^\top$ and $\boldsymbol{S}_{AA}$ is corresponding $|A| \times |A|$ submatrix; see also Appendix C.1.

Gaussian data with non-zero expectation

Assume now

$$(Y_1, \dots, Y_q) \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}),$$

where $\boldsymbol{\mu}$ is the $q \times 1$ mean vector and $\boldsymbol{\Omega}$ the $q \times q$ unconstrained precision matrix. Assume the default prior for $(\boldsymbol{\mu}, \boldsymbol{\Omega})$,

$$p^D(\boldsymbol{\mu}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}},$$

and set $b = n_0/n$, $n_0 < n$. Given $A \subseteq \{1, \dots, q\}$, we obtain

$$m^F(\boldsymbol{Y}_A) = (\pi)^{-\frac{(n-n_0)|A|}{2}} \cdot \frac{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n-1}{2}\right)}{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n_0-1}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{|A|(a_D-|\bar{A}|+n_0)}{2}} \cdot |\boldsymbol{R}_{AA}|^{-\frac{n-n_0}{2}} \quad (2.6)$$

where $\boldsymbol{R} = \sum_{i=1}^n \boldsymbol{e}_i \boldsymbol{e}_i^\top, \boldsymbol{e}_i = \boldsymbol{y}_i - \bar{\boldsymbol{y}}$, and $\boldsymbol{R}_{AA}$ is the corresponding $|A| \times |A|$ submatrix; see Appendix C.2 for a detailed proof.

Gaussian multivariate linear regression

Consider now the multivariate linear regression model described in Section 2.2,

$$\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}),$$

where $\boldsymbol{B}$ is a $(p+1)\times q$ matrix of coefficients and $\boldsymbol{E}$ a $n \times q$ matrix of error terms, $\boldsymbol{E} \sim \mathcal{N}_{n,q}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1})$. Assume the default prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$,

$$p^D(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}}.$$

Setting $b = n_0/n$, $n_0 < n$, the marginal likelihood of $\boldsymbol{Y}_A$, $A \subseteq \{1, \dots, q\}$, is given by

$$m^F(\boldsymbol{Y}_A \,|\, \boldsymbol{X}) = (\pi)^{-\frac{(n-n_0)|A|}{2}} \frac{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n-p-1}{2}\right)}{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n_0-p-1}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{|A|(a_D-|\bar{A}|+n_0)}{2}} |\hat{\boldsymbol{E}}_A^\top \hat{\boldsymbol{E}}_A|^{-\frac{n-n_0}{2}}. \tag{2.7}$$

with $\hat{\boldsymbol{E}}$ as in Equation (2.3). See also Appendix C.3 for detailed results.

2.4 Priors for graphical model comparison

Geiger & Heckerman (2002) propose a method to construct parameter priors for model choice among DAG models. To this end, they introduce a set of assumptions that permits the construction of parameter priors for every DAG model starting from a small number of assessments. Such assumptions are "naturally" satisfied by discrete DAG models with conjugate priors (Multinomial-Dirichlet models) and Gaussian DAG models with Normal-Wishart priors. They then derive a formula to compute the marginal likelihood of any DAG model given a set of i.i.d. observations. The main consequence of their approach is that Markov equivalent DAGs have the same marginal likelihood. In this section we show how the general results of Section 2.2 can be used to compute the marginal likelihood of any DAG model by taking advantage of these assumptions. We then extend such approach to the case of decomposable graphs.

2.4.1 Comparison of DAG models

The approach of Geiger & Heckerman relies on five assumptions. The first three (*complete model equivalence, regularity, likelihood modularity*) concern the sampling distribution of the data. From these it follows that, in the presence of observational data, we cannot distinguish between two complete DAG models since they have the same marginal likelihood. Assumptions 4 and 5 concern instead the parameter distribution. Let $\boldsymbol{\theta}_j$ be the model parameter indexing vertex j in any DAG model; see also Chapter 3. Assumption 4 (*prior modularity*) says that, given two distinct DAG models with the same set of parents for vertex j, the parameter prior for $\boldsymbol{\theta}_j$ should be the same in both. Equivalently,

$$p(\boldsymbol{\theta}_j \,|\, \mathcal{D}_h) = p(\boldsymbol{\theta}_j \,|\, \mathcal{D}_k)$$

for any pair of distinct DAGs $\mathcal{D}_h$ and $\mathcal{D}_k$ such that $\mathrm{pa}_{\mathcal{D}_h}(j) = \mathrm{pa}_{\mathcal{D}_k}(j)$. Finally, Assumption 5 (*global parameter independence*) states that for every DAG model $\mathcal{D}$, the parameters $\boldsymbol{\theta}_j$ should be *a priori* independent, that is

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \prod_{j=1}^{q} p(\boldsymbol{\theta}_j \,|\, \mathcal{D}).$$

As a consequence of such conditions, it is then sufficient to specify a prior for the parameter of a complete DAG model; the parameter prior for each other (not complete) DAG model can be derived automatically. For instance, with reference to Gaussian DAG models with non-zero mean we only need to set the hyperparameters of a complete DAG model, that is $\boldsymbol{\mu}_0, a_\mu, a_\Omega, \boldsymbol{R}$ in the priors for model $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$ (Section 2.2). This result is summarized by the following theorem.

Theorem 2.4.1 (Geiger & Heckerman (2002))**.** *Let $\mathcal{D}^C$ be any complete DAG model. The marginal likelihood of any DAG $\mathcal{D}$ given the data $\boldsymbol{Y}$ is given under assumptions 1-5 of Geiger & Heckerman (2002) by*

$$m_{\mathcal{D}}(\boldsymbol{Y}) = \prod_{j=1}^{q} \frac{m_{\mathcal{D}^C}(\boldsymbol{Y}_{fa_{\mathcal{D}}(j)})}{m_{\mathcal{D}^C}(\boldsymbol{Y}_{pa_{\mathcal{D}}(j)})},$$

where $fa_{\mathcal{D}}(j) = pa_{\mathcal{D}}(j) \cup j$ is the family of node j in $\mathcal{D}$.

Proof. See Theorem 2 in Geiger & Heckerman (2002). □

As an important consequence of Theorem 2.4.1, Markov equivalent DAGs have the same marginal likelihood and so they are *score equivalent*. It follows that model selection of DAG models satisfying such assumptions can be performed in the space of *representative models* rather than in the whole space of DAGs; see for instance Madigan et al. (1996). This is at the basis of the methodology presented in Chapter 3, where model selection is performed by searching over the space of Markov equivalence classes of DAGs. Therefore, we can use Formulas (2.1), (2.2) and (2.3) of Section 2.2, which hold (only) for complete DAG models (equivalently, when $\boldsymbol{\Omega}$ is unconstrained), to compute the marginal likelihood of *any* Gaussian model, that is when the sampling distribution is constrained by a DAG. The assumptions of Geiger & Heckerman are in fact satisfied by the Gaussian models with Normal-Wishart priors presented in Section 2.2.

2.4.2 Comparison of decomposable UG models

We are now interested in evaluating the marginal distribution of the data when the sampling distribution is Markov with respect to a decomposable UG; see Section 2.1. Recall (Lauritzen, 1996) that the density under a decomposable UG $\mathcal{G}$ factorizes as

$$f_{\mathcal{G}}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{\mathcal{G}}) = \frac{\prod_{C \in \mathcal{C}_{\mathcal{G}}} f(\boldsymbol{Y}_C \,|\, \boldsymbol{\theta}_{\mathcal{G}})}{\prod_{S \in \mathcal{S}_{\mathcal{G}}} f(\boldsymbol{Y}_S \,|\, \boldsymbol{\theta}_{\mathcal{G}})}, \tag{2.8}$$

where $\mathcal{C}_{\mathcal{G}}$ is the set of cliques and $\mathcal{S}_{\mathcal{G}}$ the set of separators of $\mathcal{G}$. Moreover, it is well known that any decomposable UG $\mathcal{G}$ is Markov equivalent to some DAG $\mathcal{D}$.

Lemma 2.4.1. *Let $\mathcal{G} = (V, E)$ be a decomposable undirected graph, $\mathcal{C}_{\mathcal{G}}$ and $\mathcal{S}_{\mathcal{G}}$ its sets of cliques and separators; $\mathcal{G}^{<}$ a perfect directed version of $\mathcal{G}$ and $\boldsymbol{\theta}_{\mathcal{G}^{<}}$ its parameter. Let $m(\boldsymbol{Y})$ be the marginal data distribution of $\boldsymbol{Y}$ under any complete DAG. Then*

$$m_{\mathcal{G}^{<}}(\boldsymbol{Y}) = \frac{\prod_{C \in \mathcal{C}_{\mathcal{G}}} m(\boldsymbol{Y}_C)}{\prod_{S \in \mathcal{S}_{\mathcal{G}}} m(\boldsymbol{Y}_S)}, \tag{2.9}$$

where $\boldsymbol{Y}_A$ denotes the submatrix of $\boldsymbol{Y}$ with columns indexed by $A \subseteq V$.

Proof. For a decomposable undirected graph $\mathcal{G}$, let $\{C_1, \dots, C_K\}$ be the corresponding perfect sequence of cliques, $R_1, \dots, R_K$ the residuals and $S_1, \dots, S_K$ the separators, such that $R_1 = C_1, S_1 = \emptyset$ and $C_2 = R_2 \cup S_2, \dots, C_K = R_k \cup S_k$. Then

$$
\begin{aligned}
f_{\mathcal{G}}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{\mathcal{G}}) &= \prod_{k=1}^{K} \frac{f_{\mathcal{G}}(\mathbf{Y}_{C_k} \,|\, \boldsymbol{\theta}_{\mathcal{G}})}{f_{\mathcal{G}}(\mathbf{Y}_{S_k} \,|\, \boldsymbol{\theta}_{\mathcal{G}})} \\
&= \prod_{k=1}^{K} \frac{f_{\mathcal{G}}(\mathbf{Y}_{R_k \cup S_k} \,|\, \boldsymbol{\theta}_{\mathcal{G}})}{f_{\mathcal{G}}(\mathbf{Y}_{S_k} \,|\, \boldsymbol{\theta}_{\mathcal{G}})} \\
&= \prod_{k=1}^{K} f_{\mathcal{G}}(\mathbf{Y}_{R_k} \,|\, \mathbf{Y}_{S_k}; \boldsymbol{\theta}_{\mathcal{G}}).
\end{aligned}
$$

Consider now $\mathcal{G}^{<}$, a perfect directed version of $\mathcal{G}$. Then

$$
f_{\mathcal{G}^{<}}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_{\mathcal{G}^{<}}) = \prod_{k=1}^{K} \prod_{j \in R_k} f_{\mathcal{G}^{<}}\big(\mathbf{Y}_{R_{k,j}} \,|\, \mathbf{Y}_{S_k}, \mathbf{Y}_{R_{k,1}}, \dots, \mathbf{Y}_{R_{k,j-1}}; \boldsymbol{\theta}_{\mathcal{G}^{<}}\big),
$$

where $R_{k,j}$ is the j-th element in R_k. Then, under Assumptions 1-5 of Geiger & Heckerman (2002) we obtain

$$
\begin{aligned}
m_{\mathcal{G}^{<}}(\boldsymbol{Y}) &= \prod_{k=1}^{K} \prod_{j \in R_k} \frac{m(\mathbf{Y}_{S_k \cup R_{k,1} \cup \cdots \cup R_{k,j}})}{m(\mathbf{Y}_{S_k \cup R_{k,1} \cup \cdots \cup R_{k,j-1}})} \\
&= \prod_{k=1}^{K} \frac{m(\mathbf{Y}_{S_k \cup R_k})}{m(\mathbf{Y}_{S_k})} \\
&= \frac{\prod_{C \in \mathcal{C}_{\mathcal{G}}} m(\mathbf{Y}_C)}{\prod_{S \in \mathcal{S}_{\mathcal{G}}} m(\mathbf{Y}_S)}.
\end{aligned}
$$

□

Lemma 2.4.1 gives a formula to compute the marginal likelihood of a decomposable UG model, when the assumptions of Geiger & Heckerman hold, starting from the marginal data distribution obtained under a complete DAG; see Consonni & La Rocca (2012) and Consonni et al. (2017). Starting from proper (informative) priors for the parameter of a complete DAG model, we then obtain the marginal data distribution of any subset $\boldsymbol{Y}_A$ as in the more general setting of Section 2.3.2. Then, since the marginal likelihood of a decomposable UG is equal to the marginal likelihood of any of its perfect directed versions $\mathcal{G}^{<}$, we can use Formula (2.3) to compute (2.8). This is summarized in Proposition 2.4.1.

Proposition 2.4.1. *Let $\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1})$ with $\boldsymbol{\Omega}$ Markov with respect to a decomposable undirected graph $\mathcal{G}$. Then, under the assumptions 1-5 of Geiger & Heckerman (2002),*

$$m_{\mathcal{G}}(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{\prod_{C \in \mathcal{C}_{\mathcal{G}}} m(\boldsymbol{Y}_C \mid \boldsymbol{X})}{\prod_{S \in \mathcal{S}_{\mathcal{G}}} m(\boldsymbol{Y}_S \mid \boldsymbol{X})}$$

with $m(\boldsymbol{Y}_C \mid \boldsymbol{X}), m(\boldsymbol{Y}_S \mid \boldsymbol{X})$ as in Formula (2.3).

2.4.3 Objective comparison of decomposable UG models

We now use the FBF approach to compute the marginal likelihood of a decomposable UG model. Under a complete DAG, consider the multivariate linear regression model of Section 2.2,

$$\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}), \tag{2.10}$$

with $\boldsymbol{\Omega}$ unconstrained. Starting from the default prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$,

$$p^D(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D - q - 1}{2}},$$

one obtains the implied fractional prior as in the FBF setting of Section 2.3,

$$p^F(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D + n_0 - p - q - 2}{2}} \cdot \exp\left\{-\frac{n_0}{2}\mathrm{tr}\big(\boldsymbol{\Omega}\big\{(\boldsymbol{B} - \hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}\,(\boldsymbol{B} - \hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\big\}\big)\right\}; \tag{2.11}$$

see also Appendix C.3. The distribution (2.11) is a matrix normal Wishart, which is conjugate to model (2.10). It follows that to evaluate the marginal likelihood of the decomposable UG $\mathcal{G}$ we can adopt the same approach of Section 2.4.2. Specifically, we start computing the fractional marginal likelihood under any complete DAG model,

$$m^F(\boldsymbol{Y} \mid \boldsymbol{X}) = \int\int \left[f(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{B}, \boldsymbol{\Omega})\right]^{\frac{n - n_0}{n}} p^F(\boldsymbol{B}, \boldsymbol{\Omega}) d\boldsymbol{B} d\boldsymbol{\Omega}. \tag{2.12}$$

Then, we use (2.12) to compute the fractional marginal likelihood of the decomposable UG model $\mathcal{G}$,

$$m^F_{\mathcal{G}}(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{\prod_{C \in \mathcal{C}_{\mathcal{G}}} m^F(\boldsymbol{Y}_C \mid \boldsymbol{X})}{\prod_{S \in \mathcal{S}_{\mathcal{G}}} m^F(\boldsymbol{Y}_S \mid \boldsymbol{X})},$$

with $m^F(\boldsymbol{Y}_C \,|\, \boldsymbol{X}), m^F(\boldsymbol{Y}_S \,|\, \boldsymbol{X})$ as in Formula (2.7).

An alternative objective Bayes method for model selection of Gaussian decomposable UGs is presented in Carvalho & Scott (2009). They consider a zero-mean normally distributed random vector, $Y \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{G}})$ where $\boldsymbol{\Sigma}_{\mathcal{G}}$ is the covariance matrix constrained by the decomposable UG $\mathcal{G}$. Their approach is based on an Hyper-Inverse Wishart (HIW) distribution for $\boldsymbol{\Sigma}$ (Dawid & Lauritzen, 1993), $\boldsymbol{\Sigma} \,|\, \mathcal{G} \sim HIW_{\mathcal{G}}(b, \boldsymbol{D})$, which naturally resembles the factorization over cliques and separators in (2.8),

$$p(\boldsymbol{\Sigma} \,|\, \mathcal{G}) = \frac{\prod_{C \in \mathcal{C}_{\mathcal{G}}} p(\boldsymbol{\Sigma}_C \,|\, \mathcal{G}, d, \boldsymbol{D}_C)}{\prod_{S \in \mathcal{S}_{\mathcal{G}}} p(\boldsymbol{\Sigma}_S \,|\, \mathcal{G}, d, \boldsymbol{D}_S)}.$$

Then, they develop a default version of the HIW prior for constrained covariance matrices, called *hyper-inverse Wishart g-prior*, which corresponds to the implied fractional prior obtained using our FBF approach; see Consonni & La Rocca (2012) for a detailed discussion.

Chapter 3

Model Comparison of Gaussian Essential Graphs

A Directed Acyclic Graph (DAG) is a graphical model encoding a set of conditional independencies between q random variables (Pearl, 2000). Given a set of multivariate observations, we assume that the data generating model is *faithful* to a given DAG. DAGs encoding the same set of conditional independencies are called Markov equivalent and are uniquely represented by an Essential Graph (EG), also called Completed Partially Directed Graph (CPDAG). When the objective is model selection of graphical models it is then convenient to explore the EG space rather than the space of DAGs (Andersson et al., 1997), although the number of EGs still increases super-exponentially with the number of vertices (Gillispie & Perlman, 2002).

In this Chapter we introduce the Markov equivalence property of DAGs. We then focus on EGs which represent one of the model space of interest for the model selection problem described in the current work. Hence, we consider a Gaussian framework and compute the marginal likelihood of an EG adopting an objective Bayes approach based on the notion of fractional Bayes factor. For a concise background on graphical models please refer to Section 2.1.

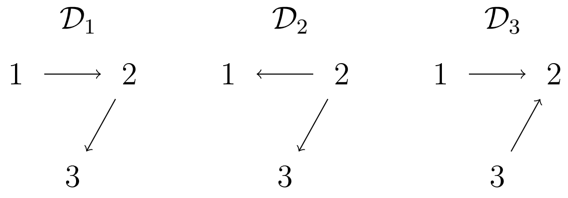
3.1 Markov equivalence and essential graphs

Consider q random variables $Y_1, \dots, Y_q$ with joint probability density function $f(\boldsymbol{y}) = f(y_1, \dots, y_q)$ and a graph $\mathcal{G} = (V, E)$. As we associate each variable Y_j to a vertex in $V \in \mathcal{G}$, we constrain $f(\boldsymbol{y})$ by the edges in $E \in \mathcal{G}$ and then write $f_{\mathcal{G}}(\boldsymbol{y})$. Consider now the case in which $\mathcal{G}$ is a DAG $\mathcal{D}$ on the set of vertices V. Then,

$$f_{\mathcal{D}}(\boldsymbol{y}) = \prod_{j \in V} f(y_j \,|\, \boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}). \tag{3.1}$$

$f_{\mathcal{D}}(\boldsymbol{y})$ encodes a set of (marginal and) conditional independences among $Y_1, \dots ,Y_q$ that can be read off from the covariance and (conditional) precision matrices of the random vector $(Y_1, \dots, Y_q)$ or directly from the DAG using the notion of *d-separation*; see Pearl (2000) for details. Le $\boldsymbol{\Sigma}_{\mathcal{D}}$ be the covariance matrix of $Y_1, \dots, Y_q$, with (k, j) element equal to $Cov(Y_k, Y_j)$ and $\boldsymbol{\Omega}_{\mathcal{D}} = \boldsymbol{\Sigma}_{\mathcal{D}}^{-1}$ the corresponding precision matrix. We remark that a zero entry at (k, j) in $\boldsymbol{\Sigma}_{\mathcal{D}}$ implies a marginal independence between Y_k and Y_j, while in $\boldsymbol{\Omega}_{\mathcal{D}}$ corresponds to a conditional independence between Y_k and Y_j given all the remaining variables. For any two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$, we then say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are *Markov equivalent* if and only if they encode the same (marginal and) conditional independencies.

Example 3.1.1. *Consider the following DAGs $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ on $q = 3$ nodes and three random variables Y_1, Y_2, Y_3 with joint density $f(\boldsymbol{y})$.*



The corresponding implied factorizations of $f(\boldsymbol{y})$ under each DAG are then

$$\begin{aligned}
f_{\mathcal{D}_1}(\boldsymbol{y}) &= f(y_1)f(y_2 \,|\, y_1)f(y_3 \,|\, y_2), \\
f_{\mathcal{D}_2}(\boldsymbol{y}) &= f(y_2)f(y_1 \,|\, y_2)f(y_3 \,|\, y_2), \\
f_{\mathcal{D}_3}(\boldsymbol{y}) &= f(y_1)f(y_3)f(y_2 \,|\, y_1, y_3).
\end{aligned}$$

Assume now for each DAG $\mathcal{D}$ the following set of linear equations

$$Y_j \,|\, \mathcal{D} = \sum_{k \in \mathrm{pa}_{\mathcal{D}}(j)} \beta_{k,j} Y_j + \varepsilon_j, \quad j = 1, 2, 3,$$

with $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ and $Cov(\varepsilon_k, \varepsilon_j) = 0$ for $k \neq j$. Equivalently, we assume $\beta_{k,j} \neq 0$ if and only if $k \to j$. For each DAG we can derive the corresponding covariance and precision matrices (we only highlight the non-zero entries with $*$):

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathcal{D}_1} &= \begin{bmatrix} 0 & * & * \\ * & 0 & * \\ * & * & 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathcal{D}_2} = \begin{bmatrix} 0 & * & * \\ * & 0 & * \\ * & * & 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathcal{D}_3} = \begin{bmatrix} 0 & * & 0 \\ * & 0 & * \\ 0 & * & 0 \end{bmatrix}; \\ \boldsymbol{\Omega}_{\mathcal{D}_1} &= \begin{bmatrix} 0 & * & 0 \\ * & 0 & * \\ 0 & * & 0 \end{bmatrix}, \quad \boldsymbol{\Omega}_{\mathcal{D}_2} = \begin{bmatrix} 0 & * & 0 \\ * & 0 & * \\ 0 & * & 0 \end{bmatrix}, \quad \boldsymbol{\Omega}_{\mathcal{D}_3} = \begin{bmatrix} 0 & * & * \\ * & 0 & * \\ * & * & 0 \end{bmatrix}.\end{aligned}$$

In the simple case $q = 3$, the only conditional independencies we can derive are "among two variables given the third" and so $\boldsymbol{\Sigma}_{\mathcal{D}}$ and $\boldsymbol{\Omega}_{\mathcal{D}}$ contain all the possible independence relations among pairs of variables. We observe that $\mathcal{D}_1$ and $\mathcal{D}_2$ encode the same marginal and conditional independencies, while $\mathcal{D}_3$ implies the additional marginal independence between Y_1 and Y_2, $Y_1 \perp\!\!\!\perp Y_2$, while Y_1 and Y_3 are not conditionally independent given Y_2, $Y_1 \not\!\perp\!\!\!\perp Y_3 \,|\, Y_2$. Hence, $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent, while $\mathcal{D}_3$ is not.

All the (marginal and) conditional independencies between variables can be also read from the DAG itself using the notion of *d-separation* (Pearl, 2000); an alternative approach is proposed by Lauritzen (1996) and based on the concept of *moral graph*. Let $\mathcal{D} = (V, E)$ be a DAG. The moral graph of $\mathcal{D}$, $\mathcal{D}^m$, is the undirected graph with the same vertex set of $\mathcal{D}$ and $u - v$ if and only if either $u \to v$ or $u \leftarrow v$ are in $\mathcal{D}$ or u and v are involved in a v-structure $u \to z \leftarrow v$.

Lemma 3.1.1 (Lauritzen (1996)). *Let $\mathcal{D} = (V, E)$ be a DAG. Then $A \perp\!\!\!\perp B \,|\, S$ whenever A and B are separated by S in $(\mathcal{D}_{\mathrm{an}(A \cup B \cup S)})^m$, the moral graph of the smallest ancestral set containing $A \cup B \cup S$.*

Consider for instance Figure 1.1 in Section 1.1 We observe that Y_2 and Y_3 are separated by Y_1 in the moral graph of $Y_2 \leftarrow Y_1 \rightarrow Y_3$ and so $Y_2 \perp\!\!\!\perp Y_3 \,|\, Y_1$. On the contrary, Y_2 and Y_3 are not separated in the moral graph of $Y_2 \rightarrow Y_4 \leftarrow Y_3$. Hence, $Y_2 \not\!\perp\!\!\!\perp Y_3 \,|\, Y_4$. Alternative formulations of Markov properties are also present in literature; see for instance Drton (2009) and Roverato (2005) for a detailed discussion. The following Theorem by Verma & Pearl (1991) provides a graphical criterion to establish whether two DAGs are Markov equivalent.

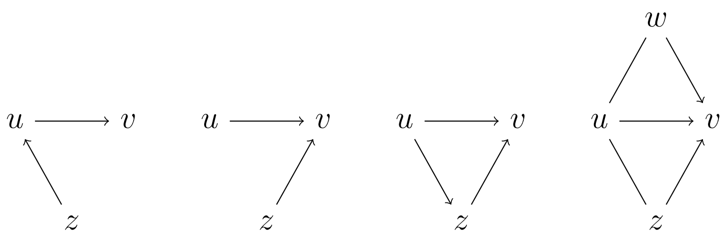
Theorem 3.1.1 (Verma & Pearl (1991))**.** *Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.*

For example, in Figure 3.1 we have three Markov equivalent DAGs, $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$. They have in fact the same skeleton and v-structures $(2 \rightarrow 4 \leftarrow 3)$.

Let $[\mathcal{D}]$ be the *Markov equivalence class* of $\mathcal{D}$, that is the set of all DAGs that are Markov equivalent to $\mathcal{D}$. Starting from a DAG $\mathcal{D}$, the objective is now to construct $[\mathcal{D}]$. We know from Theorem 3.1.1 that all DAGs in the equivalence class $[\mathcal{D}]$ have the same skeleton and the same v-structures of $\mathcal{D}$. However, other directed edges might occur in each member of $[\mathcal{D}]$. As an example, starting from DAG $\mathcal{D}_1$ in Figure 3.1, we construct its Markov equivalence class $[\mathcal{D}_1]$. Since $2 \rightarrow 4 \leftarrow 3$ is a v-structure, we have that $2 \rightarrow 4$ and $4 \leftarrow 3$ must occur in each DAG in $[\mathcal{D}_1]$. Next, we can reverse $1 \rightarrow 3$ in $1 \leftarrow 3$ and obtain $\mathcal{D}_2 \in [\mathcal{D}_1]$. Then, changing $1 \rightarrow 2$ in $1 \leftarrow 2$ is only possible in $\mathcal{D}_1$ (giving $\mathcal{D}_3 \in [\mathcal{D}_1]$) since in $\mathcal{D}_2$ would create an additional v-structure $2 \rightarrow 1 \leftarrow 3$. Moreover, reversing $1 \rightarrow 4$ would create a cycle in $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_3$. Therefore, there are no other possible orientations for the edges in $\mathcal{D}_1$. The Markov equivalence class of $\mathcal{D}_1$ is then $[\mathcal{D}_1] = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$. This kind of reasoning leads to the following definitions.

Definition 3.1.1. *Let $\mathcal{D}$ be a DAG. An edge $u \rightarrow v$ is essential in $\mathcal{D}$ if it occurs in all $\mathcal{D}^* \in [\mathcal{D}]$.*

Definition 3.1.2. *Let $\mathcal{G}$ be a graph. An arrow $u \rightarrow v$ is strongly protected in $\mathcal{G}$ if it occurs in at least one of the following four configurations as an induced* *subgraph of $\mathcal{G}$:*



(i) (ii) (iii) (iv)

Consider for instance Figure 3.1; it is easy to see that $2 \to 4$, $3 \to 4$ and $1 \to 4$ are strongly protected in $\mathcal{D}_1$ since they occur in configurations $(ii), (ii)$ and (iii) of Definition 3.1.2 respectively. Moreover, $1 \to 2$ $(1 \to 3)$ is also strongly protected in $\mathcal{D}_2$ $(\mathcal{D}_3)$ for configuration (i). We can now introduce the notion of *essential graph* (Andersson et al., 1997).

Definition 3.1.3. *Let $\mathcal{D}$ be a DAG and $[\mathcal{D}]$ its Markov equivalence class. The essential graph of $\mathcal{D}$ is defined as $\mathcal{G}(\mathcal{D}) := \bigcup_{\mathcal{D}^* \in [\mathcal{D}]} \mathcal{D}^*$.*

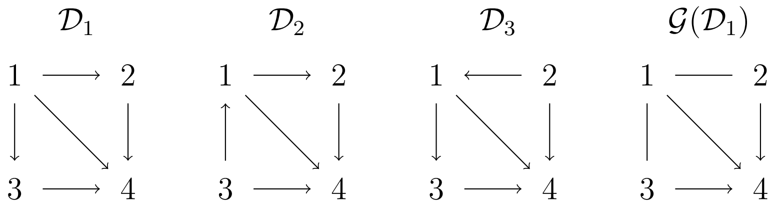


Figure 3.1: An equivalence class with three Markov equivalent DAGs $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ and the representative EG $\mathcal{G}(\mathcal{D}_1)$.

For a given Markov equivalence class $[\mathcal{D}]$, the corresponding EG $\mathcal{G}(\mathcal{D})$ is defined as the union of all DAGs $\mathcal{D}^* \in [\mathcal{D}]$, where such union is to be intended with respect to the edge sets of each $\mathcal{D}^* \in [\mathcal{D}]$. An EG then might contain both directed as well as undirected edges. From Definition 3.1.1 all the essential edges of $\mathcal{D}$ are directed in $\mathcal{G}(\mathcal{D})$ while the others undirected, since their orientation can vary from a DAG to another within the same equivalence class $[\mathcal{D}]$. In Figure 3.1, $\mathcal{G}(\mathcal{D}_1)$

is the EG associated to the equivalence class $[\mathcal{D}_1] = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$. The following Theorem by Andersson et al. (1997) gives necessary and sufficient conditions for a graph $\mathcal{G} = (V, E)$ to be the EG of some DAG $\mathcal{D}$.

Theorem 3.1.2 (Andersson et al. (1997))**.** *A graph $\mathcal{G}$ is the essential graph of a DAG $\mathcal{D}$ if and only if*

1. *$\mathcal{G}$ is a chain graph;*
2. *for each chain component $\tau \in \mathcal{T}$, $\mathcal{G}_\tau$ is chordal;*
3. *$\mathcal{G}$ has no induced subgraphs of the form $u \to v - z$ (flags);*
4. *every arrow $u \to v$ is strongly protected.*

For a chain graph $\mathcal{G} = (V, E)$ with set of chain components $\mathcal{T}$ (see Section 2.1), the joint p.d.f. of $Y_1, \ldots, Y_q$, $f(\boldsymbol{y})$, factorizes as

$$f_{\mathcal{G}}(\boldsymbol{y}) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{y}_\tau \,|\, \boldsymbol{y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}), \tag{3.2}$$

where $\boldsymbol{y}_\tau$ denotes the subset of $\boldsymbol{y} = (y_1, \ldots, y_q)$ indexed by $\tau \subseteq V$; see for example Andersson et al. (2001). Such factorization will be useful for the computation of the marginal likelihood of an EG (Section 3.2 and 4.2).

In this work, the set of all EGs with a given number of nodes q is generally taken as the model space of interest. For instance, in Figure 3.2 we have the set of all EGs with $q = 3$ nodes. When the objective is model selection of graphical models, it is well known that searching based on equivalence classes of DAGs can be more efficient than exploring the whole space on DAGs; see for instance Chickering (2002). However, an exhaustive enumeration of all possible Markov equivalent DAGs is computationally infeasible since the number of possible orientations of all edges that do not participate in the v-structures of a DAG grows exponentially in the number of such edges and superexponentially in the number of vertices (Andersson et al., 1997). The exact enumeration of all EGs with a given number of nodes is infeasible as well. Indeed, only EGs with a small number of nodes

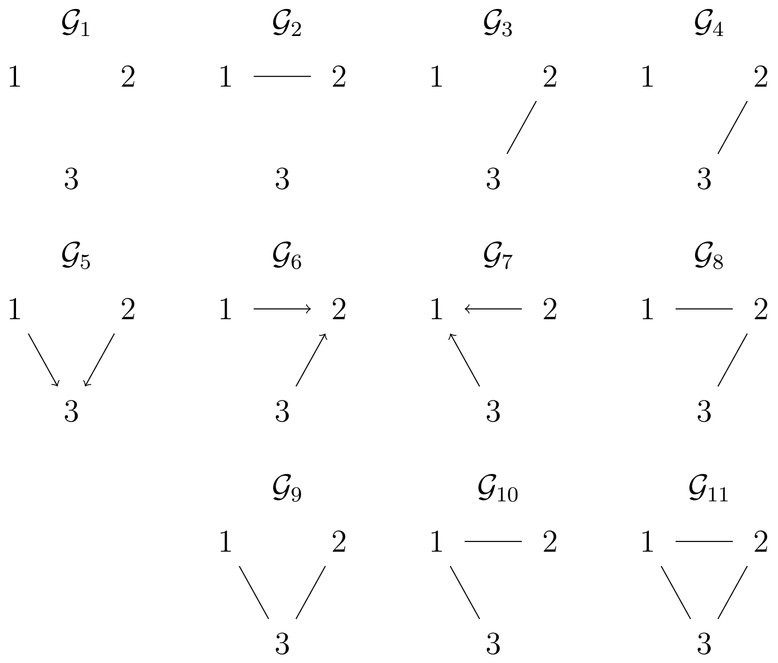


Figure 3.2: The set of all EGs on $q = 3$ nodes.

have been studied in detail in the literature; see for instance Gillispie & Perlman (2002). Nevertheless, in order to study properties of larger sets of EGs, Markov chain based algorithms were developed in recent years; see for instance He et al. (2013). We will discuss Markov chains on EGs in Section 5.1.

3.2 Gaussian Essential Graphs

In this section we consider Gaussian graphical models. The objective is to perform model selection of DAGs by scoring EGs using a Bayesian approach. To this end, we introduce objective priors for model selection based on the notion of fractional Bayes factor (Section 2.3) and compute the marginal likelihood of an EG given a set of observational data. The main ingredient of the current methodology is the result for Bayesian comparison of Gaussian multivariate regression models (Section 2.3.2) applied to decomposable UG models (Section 2.4.3); see also Consonni et al. (2017). The closed formula for the marginal likelihood thus obtained can be used to perform structural learning of EGs as described in Section 5.2. In Section 4.2 we extend the current approach to an *interventional* setting.

3.2.1 Likelihood and prior factorization

Let $\mathcal{G} = (V, E)$ be an EG. We consider a multivariate setting comprising q variables $Y_1, \dots, Y_q$ from which we collect n multivariate observations $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$, where $\boldsymbol{y}_i = (y_{i,1}, \dots, y_{i,q})^\top$, $i = 1, \dots, n$. We assume that these q-dimensional observations are i.i.d. from a parametric family of sampling distributions. From the theory presented in Andersson et al. (2001) and Drton & Eichler (2006), the joint density of $\boldsymbol{y}_i$ relative to the chain graph $\mathcal{G}$ factorizes as

$$f_{\mathcal{G}}(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_{\mathcal{G}}) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}), \tag{3.3}$$

where $\mathcal{G}_\tau$ is the (undirected) subgraph induced by $\mathcal{G}$ on τ (Section 2.1), while $\boldsymbol{y}_{i,\tau} = (y_{ij},\, j \in \tau)^\top$ denotes the (column) subvector of $\boldsymbol{y}_i$ whose components are indexed by the vertices in the chain component $\tau \subseteq V$; similarly for $\boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$, where $\mathrm{pa}_{\mathcal{G}}(\tau)$ denotes the *parents* in $\mathcal{G}$ of nodes contained in τ. Moreover, $\boldsymbol{\theta}_{\mathcal{G}}$ is the global parameter indexing the graphical model $\mathcal{G}$, while $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ is a local parameter for chain component τ indexing the conditional sampling distribution of $\boldsymbol{y}_{i,\tau}$ given $\boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$. If we let $\boldsymbol{\theta}_{\mathcal{G}} \in \boldsymbol{\Theta}_{\mathcal{G}}$ and $\boldsymbol{\theta}_\tau \in \boldsymbol{\Theta}_\tau$, we find $\boldsymbol{\Theta}_{\mathcal{G}} = \times_{\tau \in \mathcal{T}} \boldsymbol{\Theta}_\tau$, i.e., the components $\boldsymbol{\theta}_{\mathcal{G}_\tau}$s of $\boldsymbol{\theta}_{\mathcal{G}}$ are variation independent (Drton & Eichler, 2006). We now collect all n

observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ into the (n, q) data matrix $\boldsymbol{Y}$,

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1^\top \\ \vdots \\ \boldsymbol{y}_n^\top \end{pmatrix} \tag{3.4}$$

and then denote with $\boldsymbol{Y}_\tau$ the $n \times |\tau|$ matrix containing selected columns indexed by τ in $\boldsymbol{Y}$; similarly for the $n \times |\mathrm{pa}_\mathcal{G}(\tau)|$ matrix $\mathbf{Y}_{\mathrm{pa}_\mathcal{G}(\tau)}$. Recall that the observations, conditionally on $\mathcal{G}$, are i.i.d.; hence we obtain

$$\begin{aligned} f_\mathcal{G}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_\mathcal{G}) &= \prod_{i=1}^n \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) \\ &= \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}(\mathbf{Y}_\tau \,|\, \mathbf{Y}_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}). \end{aligned} \tag{3.5}$$

Since the $\boldsymbol{\theta}_{\mathcal{G}_\tau}$s are variation independent, we can further assume that the prior on $\boldsymbol{\theta}_\mathcal{G}$ factorizes as

$$p(\boldsymbol{\theta}_\mathcal{G}) = \prod_{\tau \in \mathcal{T}} p(\boldsymbol{\theta}_{\mathcal{G}_\tau}); \tag{3.6}$$

see also Castelo & Perlman (2004). Condition (3.6) extends the assumption of global (parameter) independence, which is typical for DAG models (Cowell et al., 1999, p. 193), to CG models. In this way we obtain

$$\begin{aligned} m_\mathcal{G}(\boldsymbol{Y}) &= \int_{\boldsymbol{\Theta}_\mathcal{G}} f_\mathcal{G}(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_\mathcal{G}) p(\boldsymbol{\theta}_\mathcal{G}) d\boldsymbol{\theta}_\mathcal{G} \\ &= \prod_{\tau \in \mathcal{T}} \int_{\boldsymbol{\Theta}_{\mathcal{G}_\tau}} f_{\mathcal{G}_\tau}(\mathbf{Y}_\tau \,|\, \mathbf{Y}_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) p(\boldsymbol{\theta}_{\mathcal{G}_\tau}) d\boldsymbol{\theta}_{\mathcal{G}_\tau} \\ &= \prod_{\tau \in \mathcal{T}} m_{\mathcal{G}_\tau}(\mathbf{Y}_\tau \,|\, \mathbf{Y}_{\mathrm{pa}_\mathcal{G}(\tau)}). \end{aligned} \tag{3.7}$$

From (3.7) it appears that the marginal distribution for the data matrix $\boldsymbol{Y}$ admits the same chain graph factorization that holds under the sampling distribution (3.5).

3.2.2 Marginal likelihood

Consider now a set of observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ which, conditionally on their mean vector $\boldsymbol{\mu}$ and their precision matrix $\boldsymbol{\Omega}_\mathcal{D}$ (inverse of the covariance matrix $\boldsymbol{\Sigma}_\mathcal{D}$), are i.i.d. $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}_{\mathcal{D}}^{-1})$, with $\boldsymbol{\Omega}_{\mathcal{D}}$ Markov with respect to a DAG $\mathcal{D}$. Let $\mathcal{G}$ be the EG for the equivalence class of $\mathcal{D}$. Then, we can write the factorization in the first display of (3.5) as

$$f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{\mu}_\tau + \boldsymbol{\Gamma}_\tau(\boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} - \boldsymbol{\mu}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}), \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{3.8}$$

or equivalently, letting $\boldsymbol{\alpha}_\tau = \boldsymbol{\mu}_\tau - \boldsymbol{\Gamma}_\tau \boldsymbol{\mu}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$,

$$f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{3.9}$$

$i = 1, \dots, n$, being $\boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} = \mathbb{E}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\alpha}_\tau, \boldsymbol{\Omega}_{\mathcal{G}_\tau})$, $\boldsymbol{\Gamma}_\tau$ the matrix of regression parameters and $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ the conditional precision matrix, $\boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1} = \mathbb{V}\mathrm{ar}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\alpha}_\tau, \boldsymbol{\Omega}_{\mathcal{G}_\tau})$; see also Appendix A.2. Recall from Theorem 3.1.2 that $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ is Markov with respect to a decomposable (chordal) graph $\mathcal{G}_\tau$. Collecting terms we can write

$$f_{\mathcal{G}_\tau}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\boldsymbol{y}_{i,\tau} \,|\, \boldsymbol{B}_\tau^\top \boldsymbol{x}_{i,\tau}, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{3.10}$$

where

$$\boldsymbol{x}_{i,\tau} = \begin{bmatrix} 1 \\ \boldsymbol{y}_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} \end{bmatrix}; \quad \boldsymbol{B}_\tau = \begin{bmatrix} \boldsymbol{\alpha}_\tau^\top \\ \boldsymbol{\Gamma}_\tau^\top \end{bmatrix}. \tag{3.11}$$

Notice that the matrix $\boldsymbol{B}_\tau$ consists of unconstrained components; this follows from Theorem 3.1.2 since $\mathcal{G}_\tau$ has no flags and then all nodes within the same chain component have the same parents. Letting

$$\boldsymbol{X}_\tau = \begin{pmatrix} \boldsymbol{x}_{1,\tau}^\top \\ \vdots \\ \boldsymbol{x}_{n,\tau}^\top \end{pmatrix}, \tag{3.12}$$

we can write

$$\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau, \boldsymbol{B}_\tau, \boldsymbol{\Omega}_{\mathcal{G}_\tau} \sim \mathcal{N}_{n,|\tau|}(\boldsymbol{X}_\tau \boldsymbol{B}_\tau, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{3.13}$$

so that

$$f_{\mathcal{G}}(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Omega}) = \prod_{\tau \in \mathcal{T}} \mathcal{N}_{n,|\tau|}(\boldsymbol{Y}_\tau \,|\, \boldsymbol{X}_\tau \boldsymbol{B}_\tau, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}), \tag{3.14}$$

where $\boldsymbol{X}$ is the collection (column binding) of the $\boldsymbol{X}_\tau$s.

Because of global parameter independence (3.6), we only need to specify priors separately for each chain component τ. Let $\boldsymbol{\Omega}_\tau$ denote the precision matrix of the variables in τ under a complete graph. A default prior on $(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau)$, with $\boldsymbol{\Omega}_\tau$ s.p.d., is

$$p^D(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \propto |\boldsymbol{\Omega}_\tau|^{\frac{a_D - |\tau| - 1}{2}}; \tag{3.15}$$

see also 2.3.2. Using the prior (3.15) and setting the fraction b equal to n_0/n, $n_0 < n$, the fractional prior for model (3.13) is given by

$$\begin{aligned} p^F(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \quad &\propto \quad |\boldsymbol{\Omega}_\tau|^{\frac{a_D + n_0 - |\mathrm{pa}_\mathcal{G}(\tau)| - |\tau| - 2}{2}} \\ &\cdot \quad \exp\left\{ -\frac{n_0}{2} \mathrm{tr}\Big(\boldsymbol{\Omega}_\tau \big\{ (\boldsymbol{B}_\tau - \hat{\boldsymbol{B}}_\tau)^\top \tilde{\boldsymbol{C}}_\tau \, (\boldsymbol{B}_\tau - \hat{\boldsymbol{B}}_\tau) + \tilde{\boldsymbol{R}}_\tau \big\} \Big) \right\}; \end{aligned} \tag{3.16}$$

where $\hat{\boldsymbol{B}}_\tau = (\boldsymbol{X}_\tau^\top \boldsymbol{X}_\tau)^{-1} \boldsymbol{X}_\tau^\top \boldsymbol{Y}_\tau$, $\hat{\boldsymbol{E}}_\tau = (\boldsymbol{Y}_\tau - \boldsymbol{X}_\tau \hat{\boldsymbol{B}}_\tau)$, $\tilde{\boldsymbol{C}}_\tau = n^{-1} \boldsymbol{X}_\tau^\top \boldsymbol{X}_\tau$, and $\tilde{\boldsymbol{R}}_\tau = n^{-1} \hat{\boldsymbol{E}}_\tau^\top \hat{\boldsymbol{E}}_\tau$. The distribution (3.16) is a matrix normal Wishart, which is conjugate to the sampling model (3.13); see also Appendix C.3. Distribution (3.16) is proper under two conditions: i) $a_D + n_0 - |\mathrm{pa}_\mathcal{G}(\tau)| > |\tau|$; ii) $n > |\tau| + |\mathrm{pa}_\mathcal{G}(\tau)|$; Condition ii) is a sparsity condition on the graph structure. Condition i) becomes $n_0 > |\mathrm{pa}_\mathcal{G}(\tau)| + 1$, upon setting $a_D = |\tau| - 1$.

To evaluate the EG $\mathcal{G}$ we need to compute the marginal likelihood for each $\mathcal{G}_\tau$, $\tau \in \mathcal{T}$, $m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \mid \boldsymbol{Y}_{\mathrm{pa}_\mathcal{G}(\tau)})$ in Formula (3.7). For the decomposable graph $\mathcal{G}_\tau$, let $\mathcal{C}_\tau$ be its set of (maximal) cliques and $\mathcal{S}_\tau$ be the corresponding set of separators. Then

$$m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \mid \boldsymbol{X}_\tau) = \frac{\prod_{C \in \mathcal{C}_\tau} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{C,\tau} \mid \boldsymbol{X}_\tau)}{\prod_{S \in \mathcal{S}_\tau} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{S,\tau} \mid \boldsymbol{X}_\tau)}; \tag{3.17}$$

see Dawid & Lauritzen (1993). Hence, from the theory resumed in Section 2.4.3, to compute the marginal likelihood $m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \mid \boldsymbol{X}_\tau)$ we need a formula for the density of $\boldsymbol{Y}_{J,\tau}$, the submatrix of $\boldsymbol{Y}_\tau$ containing the columns indexed by J. The result is obtained from the marginal likelihood of a Gaussian multivariate regression model

(Consonni et al., 2017); see also Section 2.3.2. Hence

$$\begin{aligned} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{J,\tau} \mid \boldsymbol{X}_\tau) &= \pi^{-\frac{(n-n_0)|J|}{2}} \frac{\Gamma_{|J|}\left(\frac{a_D+n-|\mathrm{pa}_{\mathcal{G}}(\tau)|-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n_0-|\mathrm{pa}_{\mathcal{G}}(\tau)|-1-|\bar{J}|}{2}\right)} \\ &\cdot \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{\boldsymbol{E}}_{J,\tau}^\top \hat{\boldsymbol{E}}_{J,\tau}|^{-\frac{n-n_0}{2}}, \end{aligned} \tag{3.18}$$

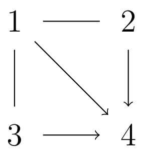
where $\bar{J} = \tau \setminus J$, so that $|\bar{J}| = |\tau| - |J|$, and $\hat{\boldsymbol{E}}_{J,\tau} = (\boldsymbol{Y}_{J,\tau} - \boldsymbol{X}_\tau \hat{\boldsymbol{B}}_{J,\tau})$, with $\hat{\boldsymbol{B}}_{J,\tau} = (\boldsymbol{X}_\tau^\top \boldsymbol{X}_\tau)^{-1} \boldsymbol{X}_\tau^\top \boldsymbol{Y}_{J,\tau}$.

We compute $m(\boldsymbol{Y}_{C,\tau} \mid \boldsymbol{X}_\tau)$ and $m(\boldsymbol{Y}_{S,\tau} \mid \boldsymbol{X}_\tau)$ in (4.10) by setting $J = C$ and $J = S$, respectively, in (3.18). Finally, using (3.7), we can recover the overall marginal likelihood of $\mathcal{G}$ given $\boldsymbol{Y}$, by multiplying each element in (4.10),

$$m_{\mathcal{G}}(\boldsymbol{Y}) = \prod_{\tau \in \mathcal{T}} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_\tau \mid \boldsymbol{X}_\tau). \tag{3.19}$$

In the following we report a simple four-node example to emphasize the likelihood factorization with respect to an EG. The same example is recalled in Section 4.2 to show how a similar factorization can be obtained in an interventional setting.

Example 3.2.1. *Consider the following EG $\mathcal{G}$ with $q = 4$ nodes and set of chain components $\mathcal{T} = \{\tau_1, \tau_2\}$, $\tau_1 = \{1, 2, 3\}, \tau_2 = \{4\}$.*



Observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from $(Y_1, \ldots, Y_4)$ are then collected in the $n \times 4$ data matrix $\boldsymbol{Y}$. We are interested in evaluating model $\mathcal{G}$ by scoring it. From equation (3.5) we obtain the factorization with respect to graph $\mathcal{G}$,

$$f_{\mathcal{G}}(\boldsymbol{Y} \mid \boldsymbol{\theta}_{\mathcal{G}}) \;=\; f_{\mathcal{G}_{\tau_1}}\big(\boldsymbol{Y}_{\tau_1} \mid \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\theta}_{\tau_1}\big) \cdot f_{\mathcal{G}_{\tau_2}}\big(\boldsymbol{Y}_{\tau_2} \mid \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_2)}, \boldsymbol{\theta}_{\tau_2}\big).$$

Then, assuming $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ *i.i.d.* $\mathcal{N}_4(\boldsymbol{\mu}, \boldsymbol{\Omega}_{\mathcal{G}}^{-1})$*, with* $\boldsymbol{\Omega}_{\mathcal{G}}$ *Markov with respect to the EG* $\mathcal{G}$ *we can write*

$$\begin{aligned} f_{\mathcal{G}_{\tau_1}}\big(\boldsymbol{Y}_{\tau_1} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\theta}_{\tau_1}\big) &= \mathcal{N}_{n,|\tau_1|}\big(\boldsymbol{Y}_{\tau_1} \,|\, \boldsymbol{X}_{\tau_1}\boldsymbol{B}_{\tau_1}, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}_{\tau_1}}^{-1}\big), \\ f_{\mathcal{G}_{\tau_2}}\big(\boldsymbol{Y}_{\tau_2} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_2)}, \boldsymbol{\theta}_{\tau_2}\big) &= \mathcal{N}_{n,|\tau_2|}\big(\boldsymbol{Y}_{\tau_2} \,|\, \boldsymbol{X}_{\tau_2}\boldsymbol{B}_{\tau_2}, \boldsymbol{I}_n, \boldsymbol{\Omega}_{\mathcal{G}_{\tau_2}}^{-1}\big), \end{aligned}$$

with

$$\boldsymbol{X}_{\tau_1} = \begin{bmatrix} \boldsymbol{1} \end{bmatrix}, \quad \boldsymbol{B}_{\tau_1} = \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 \end{bmatrix},$$

$$\boldsymbol{X}_{\tau_2} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y}_3 \end{bmatrix}, \quad \boldsymbol{B}_{\tau_2} = \begin{bmatrix} \mu_4 \\ \beta_{1,4} \\ \beta_{2,4} \\ \beta_{3,4} \end{bmatrix},$$

$$\boldsymbol{\Omega}_{\mathcal{G}_{\tau_1}} = \begin{bmatrix} \omega_1 & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_2 & 0 \\ \omega_{13} & 0 & \omega_3 \end{bmatrix}, \quad \boldsymbol{\Omega}_{\mathcal{G}_{\tau_2}} = \begin{bmatrix} \omega_4 \end{bmatrix}.$$

Chapter 4

Model Comparison with Observational and Interventional Data

In many areas, such as biology, we may have both observational and interventional data, the latter produced after a perturbation of the unknown data generating process. Interventions can be realized by forcing one or several variables of the system to "chosen values". In doing so, we destroy the original *causal dependency* on the intervened variables and modify the Markov property of the DAG model. This results in a *finer* partition of DAGs into equivalence classes, each one represented by an *interventional essential graph*. Hence, modelling jointly observational and interventional data can greatly improve the identifiability of the true data generating model.

In this chapter we extend the methodology for model comparison of Gaussian essential graphs (Chapter 3) to an interventional setting. After a short background on interventions on DAGs and interventional Markov equivalence, we present our approach to compute the marginal likelihood of an interventional essential graph given a collection of observational and interventional data.

4.1 Interventional essential graphs

The concept of intervention on DAGs goes back to Pearl (1995) and it is strictly related to the causal interpretation of a DAG model. A natural extension of the Markov equivalence property (interventional Markov equivalence) is then formalized in Hauser & Bühlmann (2012). They show that the interventional Markov equivalence property defines a finer partition of DAGs into equivalence classes, each one represented by a chain graph called Interventional Essential Graph (I-EG). In this section we introduce interventions on DAGs and resume the main results concerning I-EGs developed by Hauser & Bühlmann (2012). Of particular interest for the current work is the characterization theorem of an I-EG, which represents an extension of Theorem 4.1 in Andersson et al. (1997) for essential graphs. Such result has important consequences both for the computation of the marginal likelihood of an I-EG in the Gaussian setting of Section 4.2 and for the construction of a Markov chain on the I-EG space (Section 5.1), the two main components of the model selection problem we deal with in this work.

4.1.1 Interventions on DAGs

Let $Y_1, \dots, Y_q$ be a collection of random variables whose sampling distribution is constrained by a DAG $\mathcal{D} = (V, E)$. This implies the factorization (see also Section 3.1)

$$f_{\mathcal{D}}(\boldsymbol{y}) = \prod_{j \in V} f(y_j \,|\, \boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}). \tag{4.1}$$

where $\mathrm{pa}_{\mathcal{D}}(j)$ is the set of parents of node j in $\mathcal{D}$. Let now $I \subseteq V$ and $Y_I = \{Y_j, j \in I\}$ the corresponding subset of random variables. Following Pearl (2000), we define an intervention on I as the action of setting or forcing Y_I to the value of a random variable U_I with density $\tilde{f}(\cdot)$ such that U_I is independent of Y_j, for each $j \in \mathrm{pa}_{\mathcal{D}}(I)$; see also Hauser & Bühlmann (2012). We call I an *intervention target*. For a given DAG $\mathcal{D}$, an intervention on I destroys the original causal dependence between Y_I and its parents in $\mathcal{D}$ and leads to the following definition

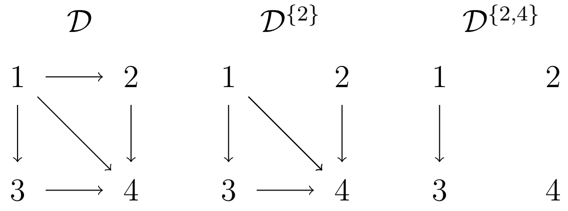


Figure 4.1: A DAG $\mathcal{D}$ and two intervention DAGs for the targets $\{2\}$ and $\{2, 4\}$.

of *intervention DAG*.

Definition 4.1.1. *Let $\mathcal{D} = (V, E)$ be a DAG, $I \subseteq V$ an intervention target. We call intervention graph of $\mathcal{D}$ the DAG $\mathcal{D}^I = (V, E^I)$, with $E^I := \{(u, v) : (u, v) \in E, v \notin I\}$.*

Consider for instance Figure 4.1. Starting from DAG $\mathcal{D}$ we assume an intervention on $I = \{2\}$ and then obtain the corresponding intervention DAG $\mathcal{D}^{\{2\}}$ by removing all edges $u \to 2$. Similarly, assuming $I = \{2, 4\}$ we obtain $\mathcal{D}^{\{2,4\}}$ by removing edges $u \to 2$ and $v \to 4$. The *post-intervention joint distribution* of $Y_1, \dots, Y_q$ is then obtained using the truncated factorization

$$f_{\mathcal{D}^I}(\boldsymbol{y} \,|\, Y_I \leftarrow U_I) = \prod_{j \notin I} f(y_j \,|\, \boldsymbol{y}_{\mathrm{pa}_{\mathcal{D}}(j)}) \prod_{j \in I} \tilde{f}(y_j). \tag{4.2}$$

Equivalently, we say that the sampling distribution of $Y_1, \dots, Y_q$ is *constrained* by DAG $\mathcal{D}^I$. With $I = \emptyset$ (no interventions are performed) and using the convention $f_{\mathcal{D}^\emptyset}(\boldsymbol{y} \,|\, Y_\emptyset \leftarrow U_\emptyset) = f_{\mathcal{D}}(\boldsymbol{y})$, Equation (4.2) reduces to Equation (4.1), which holds in the observational setting.

4.1.2 Interventions and Markov equivalence

Consider now a set of intervention targets $\mathcal{I} = \{I_k, k = 1, \dots, K\}$, also called a *family of targets*. Each target is associated to a random variable U_{I_k} with density $\tilde{f}_k(\cdot)$ assigned to the corresponding set of random variables Y_{I_k}. We also assume U_{I_k} independent of U_{I_h}, for each $k \neq h$. From the previous section we have that, for a given DAG $\mathcal{D}$, an intervention on the target I destroys the original dependencies on Y_I encoded by $\mathcal{D}$. Therefore, if we want to infer the *complete* structure (skeleton) of $\mathcal{D}$ we must require each edge in $\mathcal{D}$ to appear at least in one of the intervention DAGs $\mathcal{D}^I$. This property is formalized in the following definition.

Definition 4.1.2. *Let $\mathcal{I}$ be a family of targets. We say that $\mathcal{I}$ is conservative if for each $j \in V$ there is at least one $I \in \mathcal{I}$ such that $j \notin I$.*

We only remark that if $\emptyset \in \mathcal{I}$, that is we have observational data as well, then $\mathcal{I}$ is conservative. The Markov equivalence property of DAGs is now extended to the interventional setting. We know that two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are (observationally) Markov equivalent if $f_{\mathcal{D}_1}(\cdot)$ and $f_{\mathcal{D}_2}(\cdot)$ encode the same conditional independencies (Section 3.1). Markov equivalence with respect to a family of intervention targets $\mathcal{I}$ states that $\mathcal{D}_1$ and $\mathcal{D}_2$ are *interventionally Markov equivalent* if $f_{\mathcal{D}_1^I}(\cdot)$ and $f_{\mathcal{D}_2^I}(\cdot)$ encode the same conditional independenicies for each $I \in \mathcal{I}$. Theorem 4.1.1 provides a graphical criterion to establish if $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent under the conservative family of targets $\mathcal{I}$.

Theorem 4.1.1 (Hauser & Bühlmann (2012))**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs and $\mathcal{I}$ a conservative family of targets. We say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are $\mathcal{I}$-Markov equivalent ($\mathcal{D}_1 \sim_{\mathcal{I}} \mathcal{D}_2$) if for each $I \in \mathcal{I}$, $\mathcal{D}_1^I$ and $\mathcal{D}_2^I$ have the same skeleton and v-structures.*

Theorem 4.1.1 is a generalization to the interventional case of Theorem 1 of Verma & Pearl (1991) (see also Section 3.1); when $\mathcal{I} = \{\emptyset\}$ the two results coincide and we also say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent *in the observational sense*. As a consequence, $\mathcal{D}_1$ and $\mathcal{D}_2$ are $\mathcal{I}$-Markov equivalent if the corresponding intervention DAGs $\mathcal{D}_1^I$ and $\mathcal{D}_2^I$ are Markov equivalent in the observational sense for each $I \in \mathcal{I}$. From Theorem 4.1.1 it also follows that $\mathcal{D}_1$ and $\mathcal{D}_2$ are $\mathcal{I}$-Markov equivalent if they have the same skeleton and the same v-structures (they are Markov equivalent in the observational sense) and $\mathcal{D}_1^I$ and $\mathcal{D}_2^I$ have the same skeleton for each $I \in \mathcal{I}$. See also Theorem 10 in Hauser & Bühlmann (2012).

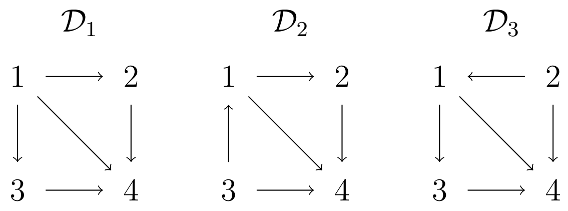


Figure 4.2: An equivalence class with three Markov equivalent DAGs $\mathcal{D}_1$, $\mathcal{D}_2$, D_3. Assuming $\mathcal{I} = \{\emptyset, \{2\}\}$, $\mathcal{D}_3$ is no longer $\mathcal{I}$-Markov equivalent to $\mathcal{D}_1$ and $\mathcal{D}_2$.

Let now $[\mathcal{D}]_\mathcal{I}$ be the $\mathcal{I}$-Markov equivalence class of $\mathcal{D}$, that is the set of all DAGs that are $\mathcal{I}$-Markov equivalent to $\mathcal{D}$. As a consequence of Theorem 4.1.1 we have that interventions based on a conservative family of targets define a finer partition of DAGs into equivalence classes. This is not true in general when $\mathcal{I}$ is not conservative as discussed in Hauser & Bühlmann (2012).

4.1.3 Interventions and essential graphs

Each interventional Markov equivalence class can be uniquely represented by a special chain graph called *interventional essential graph.* In the following we introduce the notion of *interventional essential graph* ($\mathcal{I}$-EG) together with the corresponding characterization theorem.

Definition 4.1.3. *Let $\mathcal{D}$ be a DAG and $\mathcal{I}$ a conservative family of targets. The $\mathcal{I}$-essential graph of $\mathcal{D}$ is defined as ${}_\mathcal{I}\mathcal{G}(\mathcal{D}) = \bigcup_{\mathcal{D}^* \in [\mathcal{D}]_\mathcal{I}} \mathcal{D}^*$.*

Clearly, ${}_\mathcal{I}\mathcal{G}(\mathcal{D}) =_\mathcal{I} \mathcal{G}(\mathcal{D}^*)$ for each $\mathcal{D}^* \in [\mathcal{D}]_\mathcal{I}$. In the sequel we often use ${}_\mathcal{I}\mathcal{G}$ to identify an $\mathcal{I}$-EG without making explicit its originating DAG.
In Figure 4.3, we have two $\mathcal{I}$-EGs for the family of targets $\mathcal{I} = \{\emptyset, \{2\}\}$, defining a partition of the equivalence class $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$ of Figure 4.2 which was represented by $\mathcal{G}$.

Definition 4.1.4. *Let $\mathcal{D}$ be a DAG. An edge $u \to v \in \mathcal{D}$ is $\mathcal{I}$-essential in $\mathcal{D}$ if it occurs in all $\mathcal{D}^* \in [\mathcal{D}]_\mathcal{I}$.*

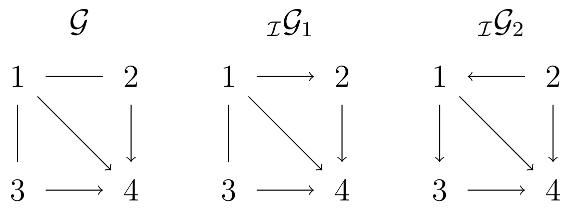
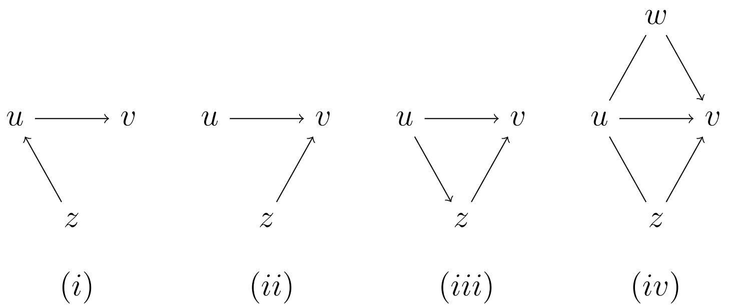


Figure 4.3: An EG $\mathcal{G}$ and two $\mathcal{I}$-EGs $_{\mathcal{I}}\mathcal{G}_1$ and $_{\mathcal{I}}\mathcal{G}_2$ for $\mathcal{I} = \{\emptyset, \{2\}\}$. $\mathcal{G}$ is the representative of the Markov equivalence class $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$ in Figure 4.2 which is partitioned into two $\mathcal{I}$-Markov equivalence classes: $\{\mathcal{D}_1, \mathcal{D}_2\}$ represented by $_{\mathcal{I}}\mathcal{G}_1$ and $\{\mathcal{D}_3\}$ represented by $_{\mathcal{I}}\mathcal{G}_1 \equiv \mathcal{D}_3$.

It is easy to show that for a given edge $u \to v$, if there exists an intervention target $I \in \mathcal{I}$ such that $|\{u, v\} \cap I| = 1$ then $u \to v$ is $\mathcal{I}$-essential. This means that if in some EG we have an undirected edge $u - v$, with an intervention on a target I such that $|\{u, v\} \cap I| = 1$ (for example $I = \{u\}$) we can *identify* the orientation of $u - v$ and then *distinguish* between $u \to v$ and $u \leftarrow v$.

Definition 4.1.5. *Let $\mathcal{G}$ be a graph. An arrow $u \to v \in \mathcal{G}$ is strongly $\mathcal{I}$-protected in $\mathcal{G}$ if there is some $I \in \mathcal{I}$ such that $|(u, v) \cap I| = 1$ or the arrow $u \to v$ occurs in at least one of the following four configurations as an induced subgraph of $\mathcal{G}$:*



Consider for instance DAG $\mathcal{D}_1$ in Figure 4.2. It is easy to see that $2 \to 4$, $3 \to 4$, $1 \to 4$ are strongly $\mathcal{I}$-protected since they occur in configurations (ii), (ii) and (iii) respectively. Moreover, assuming $\mathcal{I} = \{\emptyset, \{2\}\}$, $1 \to 2$ is strongly $\mathcal{I}$-protected as well and then results in the $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}_1$ of $\mathcal{D}_1$ in Figure 4.3. For simplicity of notation, in the following we will denote an $\mathcal{I}$-EG ($_{\mathcal{I}}\mathcal{G}$) by $\mathcal{G}$ when the family $\mathcal{I}$

is clear from the context. We can now give the characterization theorem of an $\mathcal{I}$-EG.

Theorem 4.1.2 (Hauser & Bühlmann (2012)). *Let $\mathcal{D}$ be a DAG on the set of vertices V and $\mathcal{I}$ a conservative family of targets. A graph $\mathcal{G}$ is the $\mathcal{I}$-essential graph of $\mathcal{D}$ if and only if*

1. *$\mathcal{G}$ is a chain graph;*
2. *for each chain component $\tau \in \mathcal{T}$, $\mathcal{G}_\tau$ is chordal;*
3. *$\mathcal{G}$ has no induced subgraphs of the form $u \rightarrow v - z$ (flags);*
4. *every arrow $u \rightarrow v$ is strongly $\mathcal{I}$-protected;*
5. *$\mathcal{G}$ has no line $u - v$ for which there exists some $I \in \mathcal{I}$ such that $|I \cap \{u, v\}| = 1$.*

While Conditions $1, 2, 3$ are the same as in Theorem 3.1.2 of Section 3.1, Condition 4 is a natural extension of the corresponding one. Condition 5 is instead specific to the interventional setting and of particular interest for this work. It says that, given a family of targets $\mathcal{I}$, each $\mathcal{I}$-EG ${}_{\mathcal{I}}\mathcal{G}$ is such that it has no chain components containing nodes on which at least one intervention was performed together with nodes on which no interventions were done. As mentioned, this result has important consequences both for the derivation of the marginal likelihood of a chain graph in Section 4.2 and for the construction of a Markov chain on the $\mathcal{I}$-EG space (Section 5.1).

4.2 Gaussian interventional essential graphs

In this section we extend the methodology for the computation of the marginal likelihood of an essential graph (Section 3.2) to the interventional setting described in Section 4.1. In doing so, we derive a method to perform model selection of Gaussian graphical models in the presence of interventional data. To this end we first introduce the data structure together with the likelihood factorization for a chain graph. As for the observational case, all the results are based on the marginal likelihood of multivariate regression graphical models. We assume again objective priors and then obtain conjugate priors based on the fractional Bayes factor for both *standard* and *interventional* model parameters that we assume *a priori* independent.

4.2.1 Likelihood and prior factorization

In the interventional setting of Section 4.1 a dataset consists in a collection of multivariate observations each one associated to an intervention target. In particular, we assume to have, for each target $I_k \in \mathcal{I}$ $(k = 1, \dots, K)$ $n^{I_k} \equiv n^{(k)}$ i.i.d. q-variate observations from the sampling distribution of $(Y_1, \dots, Y_q) \,|\, Y_{I_k} \leftarrow U_{I_k}$ collected in the $(n^{(k)}, q)$ matrix $\boldsymbol{Y}^{I_k} \equiv \boldsymbol{Y}^k$. By binding the matrices $\boldsymbol{Y}^k$, $I_k \in \mathcal{I}$, we then obtain the (n, q) data matrix $\boldsymbol{Y}$, where $n = \sum_k n^{(k)}$. Recall from Theorem 4.1.2 that an $\mathcal{I}$-essential graph $\mathcal{G} = (V, E)$ is a chain graph with set of chain components $\mathcal{T}$. Therefore we can write the factorization

$$f_{\mathcal{G}}\left(\boldsymbol{Y} \,|\, \boldsymbol{\Theta}_{\mathcal{G}}\right) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_\tau}\left(\boldsymbol{Y}_\tau \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}\right), \tag{4.3}$$

where $\boldsymbol{Y}_\tau$ denotes selected columns of the data matrix $\boldsymbol{Y}$ corresponding to the subset $\tau \subseteq V$. $\boldsymbol{\Theta}_{\mathcal{G}}$ is instead a global parameter indexing the graphical model $\mathcal{G}$ and $\boldsymbol{\Theta}_{\mathcal{G}_\tau}$ a local parameter for chain component τ.

For simplicity, consider first the case in which $|I_k| = 1$ for all $I_k \in \mathcal{I}$, that is assume *single-node* interventions. Let ${}_{\mathcal{I}}\mathcal{S}_q$ be the set of all $\mathcal{I}$-EGs on q nodes. From Condition 5 of Theorem 4.1.2 we have that each $\mathcal{I}$-EG $\mathcal{G} \in {}_{\mathcal{I}}\mathcal{S}_q$ contains a chain component $\tau = I_k$ for each $I_k \in \mathcal{I}$. Recall that $\tau \subseteq V$ indexes columns of the data matrix $\boldsymbol{Y}$, while I_k refers to subsets of rows. Then, we can write for each $I_k \in \mathcal{I}$ and $\tau \in \mathcal{T}$

$$f_{\mathcal{G}_\tau}\left(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}^k, \boldsymbol{\Theta}_{\mathcal{G}_\tau}\right) = \begin{cases} f_{\mathcal{G}_\tau}\left(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}^k, \boldsymbol{\theta}_{\mathcal{G}_\tau}\right) & \text{if } I_k \neq \tau, \\ \tilde{f}_{\mathcal{G}_\tau}\left(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{\psi}_{\mathcal{G}_\tau}^k\right) & \text{if } I_k = \tau, \end{cases} \tag{4.4}$$

where $\boldsymbol{Y}_\tau^k$ denotes columns indexed by τ of $\boldsymbol{Y}^k$. The first case of Equation (4.4) is the usual factorization for subgraph $\mathcal{G}_\tau$ which holds for all $\boldsymbol{Y}_\tau^k$ such that $I_k \neq \tau$, that is when no interventions are performed on chain component $I_k = \tau$. The second case corresponds instead to the (intervention) density $\tilde{f}_{\mathcal{G}_\tau}(\cdot)$, where the intervention on I_k destroys the original dependence between node $I_k = \tau$ and its parents. Moreover, $\boldsymbol{\psi}_{\mathcal{G}_\tau}^k$ is a parameter modelling the effect of the intervention on chain component τ, while $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ is the chain component parameter of the conditional distribution of those observations not arising from an intervention on τ. Implicitly we assume that $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ does not depend on I_k. Assuming now $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ independent we can write

$$
\begin{aligned}
f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau \,|\, \boldsymbol{Y}_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}\big) &= \prod_{I_k \neq \tau} f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{Y}^k_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\big) \cdot \prod_{I_k = \tau} \tilde{f}_{\mathcal{G}_\tau}\,\big(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{\psi}^k_{\mathcal{G}_\tau}\big) \\
&= f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau^* \,|\, \boldsymbol{Y}^*_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\big) \cdot \prod_{I_k = \tau} \tilde{f}_{\mathcal{G}_\tau}\,\big(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{\psi}^k_{\mathcal{G}_\tau}\big)\,, \qquad (4.5)
\end{aligned}
$$

being $\boldsymbol{Y}_\tau^*$ a $(n_\tau^*, |\tau|)$ matrix collecting all the observations $\boldsymbol{Y}_\tau^k$ such that $I_k \neq \tau$. See also the simple example with $q = 4$ nodes at the end of the current section.

We now extend the likelihood factorization in Equation (4.5) to the more general setting $|I_k| \geq 1$, $I_k \in \mathcal{I}$. Let $\mathcal{I}_\tau = \big\{I_k \in \mathcal{I} : |I_k \cap \tau| > 0\big\}$ be the set of all the intervention targets that *act* on chain component τ. We can write for $\boldsymbol{Y}_\tau$

$$
\begin{aligned}
f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau \,|\, \boldsymbol{Y}_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}\big) &= \prod_{I_k \notin \mathcal{I}_\tau} f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{Y}^k_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\big) \cdot \prod_{I_k \in \mathcal{I}_\tau} \tilde{f}_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{\psi}^k_{\mathcal{G}_\tau}\big) \\
&= f_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau^* \,|\, \boldsymbol{Y}^*_{\mathrm{pa}_\mathcal{G}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\big) \cdot \prod_{I_k \in \mathcal{I}_\tau} \tilde{f}_{\mathcal{G}_\tau}\big(\boldsymbol{Y}_\tau^k \,|\, \boldsymbol{\psi}^k_{\mathcal{G}_\tau}\big), \qquad (4.6)
\end{aligned}
$$

where again $\boldsymbol{Y}_\tau^*$ is a $(n_\tau^*, |\tau|)$ matrix collecting all the observations $\boldsymbol{Y}_\tau^k$ such that $I_k \notin \mathcal{I}_\tau$. As for the observational setting we assume that the prior on $\boldsymbol{\Theta}_\mathcal{G}$ factorizes as

$$
p(\boldsymbol{\Theta}_\mathcal{G}) = \prod_{\tau \in \mathcal{T}} p(\boldsymbol{\Theta}_{\mathcal{G}_\tau}), \qquad (4.7)
$$

while for each chain component parameter $\boldsymbol{\Theta}_{\mathcal{G}_\tau}$ we assume $\boldsymbol{\theta}_{\mathcal{G}_\tau}$ a priori independent of $\boldsymbol{\psi}^k_{\mathcal{G}_\tau}$, for all $I_k \in \mathcal{I}_\tau$. Hence, we obtain

$$
p(\boldsymbol{\Theta}_{\mathcal{G}_\tau}) = p(\boldsymbol{\theta}_{\mathcal{G}_\tau}) \cdot \prod_{I_k \in \mathcal{I}_\tau} p\big(\boldsymbol{\psi}^k_{\mathcal{G}_\tau}\big), \qquad (4.8)
$$

which extends global parameter independence (Equation (3.6) in Section 3.2) to the interventional setting. The marginal likelihood for the $\mathcal{I}$-EG $\mathcal{G}$ given the data $\boldsymbol{Y}$ is then

$$\begin{aligned} m_{\mathcal{G}}(\boldsymbol{Y}) &= \prod_{\tau\in\mathcal{T}} \int f_{\mathcal{G}_\tau}(\mathbf{Y}_\tau \,|\, \mathbf{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\mathcal{G}_\tau}) p(\boldsymbol{\Theta}_{\mathcal{G}_\tau}) d\boldsymbol{\Theta}_{\mathcal{G}_\tau} \\ &= \prod_{\tau\in\mathcal{T}} m_{\mathcal{G}_\tau}(\mathbf{Y}_\tau \,|\, \mathbf{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}). \end{aligned} \tag{4.9}$$

From Theorem 4.1.2 in Section 4.1 recall that $\mathcal{G}_\tau$ is a decomposable (chordal) graph. Let $\mathcal{C}_\tau$ be its set of (maximal) cliques and $\mathcal{S}_\tau$ the corresponding set of separators. Then

$$m_{\mathcal{G}_\tau}(\mathbf{Y}_\tau \,|\, \boldsymbol{X}_\tau) = \frac{\prod_{C\in\mathcal{C}_\tau} m_{\mathcal{G}_\tau}(\mathbf{Y}_{C,\tau} \,|\, \boldsymbol{X}_\tau)}{\prod_{S\in\mathcal{S}_\tau} m_{\mathcal{G}_\tau}(\mathbf{Y}_{S,\tau} \,|\, \boldsymbol{X}_\tau)}, \tag{4.10}$$

where

$$m_{\mathcal{G}_\tau}\left(\mathbf{Y}_{J,\tau} \,|\, \boldsymbol{X}_\tau\right) = m_{\mathcal{G}_\tau}\left(\mathbf{Y}^*_{J,\tau} \,|\, \boldsymbol{X}^*_\tau\right) \cdot \prod_{I_k\in\mathcal{I}_\tau} m_{\mathcal{G}_\tau}\left(\mathbf{Y}^k_\tau\right)$$

and $J \subseteq \tau$ refers to a generical clique or separator of $\mathcal{G}_\tau$.

4.2.2 Marginal likelihood

Let $\mathcal{I}$ be a conservative family of targets, $\mathcal{G}$ an $\mathcal{I}$-EG. Recall that for each τ, all the observations $\mathbf{Y}^k_\tau$ such that $I_k \notin \mathcal{I}_\tau$ are collected in the $(n^*_\tau, |\tau|)$ matrix $\mathbf{Y}^*_\tau$. For each observation $\boldsymbol{y}^*_{i,\tau}$ in $\boldsymbol{Y}^*_\tau$ assume

$$f_{\mathcal{G}_\tau}\left(\boldsymbol{y}^*_{i,\tau} \,|\, \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\right) = \mathcal{N}_{|\tau|}\left(\boldsymbol{y}^*_{i,\tau} \,|\, \boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Omega}^{-1}_{\mathcal{G}_\tau}\right), \tag{4.11}$$

$i = 1, \dots, n$, where $\boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}$ denotes the conditional mean, $\boldsymbol{\alpha}_\tau + \boldsymbol{\Gamma}_\tau \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} = \mathbb{E}(\boldsymbol{y}^*_{i,\tau} \,|\, \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\alpha}_\tau, \boldsymbol{\Omega}_{\mathcal{G}_\tau})$, $\boldsymbol{\Gamma}_\tau$ is the matrix of regression parameters and $\boldsymbol{\Omega}_{\mathcal{G}_\tau}$ the conditional precision matrix, $\boldsymbol{\Omega}^{-1}_{\mathcal{G}_\tau} = \mathbb{V}\mathrm{ar}(\boldsymbol{y}^*_{i,\tau} \,|\, \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\alpha}_\tau, \boldsymbol{\Omega}_{\mathcal{G}_\tau})$; see also Section 3.2.2. Now, letting

$$\boldsymbol{x}^*_{i,\tau} = \begin{bmatrix} 1 \\ \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)} \end{bmatrix}, \quad \boldsymbol{B}_\tau = \begin{bmatrix} \boldsymbol{\alpha}^\top_\tau \\ \boldsymbol{\Gamma}^\top_\tau \end{bmatrix}, \tag{4.12}$$

we can write

$$f_{\mathcal{G}_\tau}\left(\boldsymbol{y}^*_{i,\tau} \,|\, \boldsymbol{y}^*_{i,\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_{\mathcal{G}_\tau}\right) = \mathcal{N}_{|\tau|}\left(\boldsymbol{y}^*_{i,\tau} \,|\, \boldsymbol{B}_\tau^\top \boldsymbol{x}^*_{i,\tau}, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}\right). \tag{4.13}$$

As for the observational case, the matrix $\boldsymbol{B}_\tau$ consists of unconstrained component since the $\mathcal{I}$-EG $\mathcal{G}$ has no flags (Condition 3 of Theorem 4.1.2). In matrix notation we can write

$$f_{\mathcal{G}_\tau}\left(\boldsymbol{Y}^*_\tau \,|\, \boldsymbol{Y}^*_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\theta}_\tau\right) = \mathcal{N}_{n^*_\tau,|\tau|}\left(\boldsymbol{Y}^*_\tau \,|\, \boldsymbol{X}^*_\tau \boldsymbol{B}_\tau, \boldsymbol{I}_{n^*}, \boldsymbol{\Omega}_{\mathcal{G}_\tau}^{-1}\right), \tag{4.14}$$

being

$$\boldsymbol{X}^*_\tau = \begin{pmatrix} \boldsymbol{x}^{*\top}_{1,\tau} \\ \vdots \\ \boldsymbol{x}^{*\top}_{n^*,\tau} \end{pmatrix}.$$

Because of global parameter independence (4.8), we can specify priors separately for each chain component τ. Let $\boldsymbol{\Omega}_\tau$ denote the precision matrix of the variables in τ under a complete graph. Under the fractional Bayes setting described in Section 2.3 we start assuming the default prior on $(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau)$,

$$p(\boldsymbol{B}_\tau, \boldsymbol{\Omega}_\tau) \propto |\boldsymbol{\Omega}_\tau|^{\frac{a_D-|\tau|-1}{2}}. \tag{4.15}$$

The implied fractional prior is then obtained as in the Gaussian multivariate regression model resumed in Section 2.2 by setting $b = b^* = n^*_0/n^*$. Then, the marginal data distribution of $\boldsymbol{Y}^*_\tau$ can be computed accordingly. In particular, for any subset $J \subseteq \tau$ (selected columns indexed by J of $\boldsymbol{Y}^*_\tau$) we obtain from Formula (2.7) in Section 2.3.2,

$$\begin{aligned} m_{\mathcal{G}_\tau}\left(\boldsymbol{Y}^*_{J,\tau} \,|\, \boldsymbol{X}^*_\tau\right) \quad &= \quad (\pi)^{-\frac{(n^*_\tau-n^*_0)|J|}{2}} \cdot \frac{\Gamma_{|J|}\left(\frac{a_D+n^*_\tau-|\mathrm{pa}_{\mathcal{G}}(\tau)|-1-|\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D+n^*_0-|\mathrm{pa}_{\mathcal{G}}(\tau)|-1-|\bar{J}|}{2}\right)} \\ &\cdot \quad \left(\frac{n^*_0}{n^*}\right)^{\frac{|J|(a_D+n^*_0-|\bar{J}|)}{2}} \quad |\hat{\boldsymbol{E}}^{*\top}_{J,\tau}\hat{\boldsymbol{E}}^*_{J,\tau}|^{-\frac{n^*-n^*_0}{2}}. \end{aligned} \tag{4.16}$$

We remark that Formula (4.16) holds provided that $a_D + n^*_0 - |\mathrm{pa}_{\mathcal{G}}(\tau)| > |\tau|$, $n^*_\tau > |\tau| + |\mathrm{pa}_{\mathcal{G}}(\tau)|$. Focus now on the densities $\tilde{f}_{\mathcal{G}_\tau}(\cdot)$ of Formula (4.6). Assuming for each $\boldsymbol{y}_{i,\tau}^{k}$, $I_k \in \mathcal{I}_\tau$,

$$\tilde{f}_{\mathcal{G}_\tau}\left(\boldsymbol{y}_{i,\tau}^{k} \,|\, \boldsymbol{\psi}_{\mathcal{G}_\tau}^{k}\right) = \mathcal{N}_{|\tau|}\left(\boldsymbol{y}_{i,\tau}^{k} \,|\, \boldsymbol{\delta}_{\tau}^{k}, (\boldsymbol{\Phi}_{\tau}^{k})^{-1}\right), \tag{4.17}$$

we obtain

$$\tilde{f}_{\mathcal{G}_\tau}\left(\boldsymbol{Y}_{\tau}^{k} \,|\, \boldsymbol{\psi}_{\tau}^{k}\right) = \mathcal{N}_{n^{(k)},|\tau|}\left(\boldsymbol{Y}_{\tau}^{k} \,|\, \boldsymbol{\Delta}_{\tau}^{k}, \boldsymbol{I}_{n^{(k)}}, (\boldsymbol{\Phi}_{\tau}^{k})^{-1}\right), \tag{4.18}$$

where $\boldsymbol{\psi}_{\tau}^{k} = \left(\boldsymbol{\Delta}_{\tau}^{k}, \boldsymbol{\Phi}_{\tau}^{k}\right)$ is the chain component parameter modelling the effect of the intervention with target I_k on the chain component τ and

$$\boldsymbol{\Delta}_{\tau}^{k} = \mathbf{1}_{n^{(k)}} \boldsymbol{\delta}_{\tau}^{k\top}, \tag{4.19}$$

being $\mathbf{1}_{n^{(k)}}$ the unit vector of length $n^{(k)}$. A default prior for $\left(\boldsymbol{\Delta}_{\tau}^{k}, \boldsymbol{\Phi}_{\tau}^{k}\right)$ is then

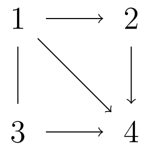
$$p\left(\boldsymbol{\Delta}_{\tau}^{k}, \boldsymbol{\Phi}_{\tau}^{k}\right) \propto |\boldsymbol{\Phi}_{\tau}^{k}|^{\frac{a_D^k - |\tau| - 1}{2}}.$$

Setting $b = b^{(k)} = n_0^{(k)}/n_\tau^{(k)}$, we obtain from Formula (2.6) the marginal data distribution of $\boldsymbol{Y}_{J,\tau}^{k}$, for $I_k \in \mathcal{I}_\tau$,

$$\begin{aligned} m_{\mathcal{G}_\tau}(\boldsymbol{Y}_{J,\tau}^{k}) \;&=\; (\pi)^{-\frac{\left(n_\tau^{(k)} - n_0^{(k)}\right)|J|}{2}} \cdot \frac{\Gamma_{|J|}\left(\frac{a_D^{(k)} + n_\tau^{(k)} - 1 - |\bar{J}|}{2}\right)}{\Gamma_{|J|}\left(\frac{a_D^{(k)} + n_0^{(k)} - 1 - |\bar{J}|}{2}\right)} \\ &\cdot \left(\frac{n_0^{(k)}}{n_\tau^{(k)}}\right)^{\frac{|J|\left(a_D^{(k)} + n_0^{(k)} - |\bar{J}|\right)}{2}} |\hat{\boldsymbol{Y}}_{J,\tau}^{k\top} \hat{\boldsymbol{Y}}_{J,\tau}^{k}|^{-\frac{n_\tau^{(k)} - n_0^{(k)}}{2}}, \end{aligned} \tag{4.20}$$

where $\hat{\boldsymbol{Y}}_{J,\tau}^{k} = \boldsymbol{Y}_{J,\tau}^{k} - \mathbf{1}_{n^{(k)}} \bar{\boldsymbol{Y}}_{J,\tau}^{k}$ and $\bar{\boldsymbol{Y}}_{J,\tau}^{k}$ is the $1 \times |J|$ mean vector of $\boldsymbol{Y}_{J,\tau}^{k}$. Observe that in such case we don't have a conditional structure any longer due to effect of the intervention on τ. Equivalently, Formula (4.20) can be derived from Equation (4.16) by setting $\mathrm{pa}_{\mathcal{G}}(\tau) = \emptyset$. We remark that Formula (4.20) holds provided that $a_D^{(k)} + n_0^{(k)} > |\tau|$, $n_\tau^{(k)} > |\tau|$.

Example 4.2.1. *Under the family of targets $\mathcal{I} = \{\emptyset, \{2\}\}$, consider the following $\mathcal{I}$-EG $\mathcal{G}$. The set of chain components of $\mathcal{G}$ is $\mathcal{T} = \{\tau_1, \tau_2, \tau_3\}$ with $\tau_1 = \{1, 3\}, \tau_2 = \{2\}, \tau_3 = \{4\}$.*



Given the family of targets $\mathcal{I}$, a dataset consists in a pair $(\boldsymbol{Y}^{\varnothing}, \boldsymbol{Y}^{\{2\}})^{\top}$, where we denote with $\boldsymbol{Y}^{\varnothing}$ the $n^{\varnothing} \times 4$ matrix of observational data, with $\boldsymbol{Y}^{\{2\}}$ the $n^{\{2\}} \times 4$ matrix of interventional data with target $\{2\}$ and $n = n^{\varnothing} + n^{\{2\}}$. We are interested in evaluating model $\mathcal{G}$ by scoring it. Given the likelihood factorization

$$f_{\mathcal{G}}\left(\boldsymbol{Y} \,|\, \boldsymbol{\Theta}_{\mathcal{G}}\right) = \prod_{\tau \in \mathcal{T}} f_{\mathcal{G}_{\tau}}\left(\boldsymbol{Y}_{\tau} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau)}, \boldsymbol{\Theta}_{\tau}\right),$$

we obtain from Equation (4.5)

$$\begin{aligned}
f_{\mathcal{G}_{\tau_1}}\left(\boldsymbol{Y}_{\tau_1} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\Theta}_{\tau_1}\right) &= f_{\mathcal{G}_{\tau_1}}\left(\boldsymbol{Y}^{\varnothing}_{\tau_1} \,|\, \boldsymbol{Y}^{\varnothing}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\theta}_{\tau_1}\right) \cdot f_{\mathcal{G}_{\tau_1}}\left(\boldsymbol{Y}^{\{2\}}_{\tau_1} \,|\, \boldsymbol{Y}^{\{2\}}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\theta}_{\tau_1}\right) \\
&= f_{\mathcal{G}_{\tau_1}}\left(\boldsymbol{Y}_{\tau_1} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\theta}_{\tau_1}\right), \\
f_{\mathcal{G}_{\tau_2}}\left(\boldsymbol{Y}_{\tau_2} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_2)}, \boldsymbol{\Theta}_{\tau_2}\right) &= f_{\mathcal{G}_{\tau_2}}\left(\boldsymbol{Y}^{\varnothing}_{\tau_2} \,|\, \boldsymbol{Y}^{\varnothing}_{\mathrm{pa}_{\mathcal{G}}(\tau_2)}, \boldsymbol{\theta}_{\tau_2}\right) \cdot \tilde{f}_{\mathcal{G}_{\tau_2}}\left(\boldsymbol{Y}^{\{2\}}_{\tau_2} \,|\, \boldsymbol{\psi}^{\{2\}}_{\tau_2}\right), \\
f_{\mathcal{G}_{\tau_3}}\left(\boldsymbol{Y}_{\tau_3} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_3)}, \boldsymbol{\Theta}_{\tau_3}\right) &= f_{\mathcal{G}_{\tau_3}}\left(\boldsymbol{Y}^{\varnothing}_{\tau_3} \,|\, \boldsymbol{Y}^{\varnothing}_{\mathrm{pa}_{\mathcal{G}}(\tau_3)}, \boldsymbol{\theta}_{\tau_3}\right) \cdot f_{\mathcal{G}_{\tau_3}}\left(\boldsymbol{Y}^{\{2\}}_{\tau_3} \,|\, \boldsymbol{Y}^{\{2\}}_{\mathrm{pa}_{\mathcal{G}}(\tau_3)}, \boldsymbol{\theta}_{\tau_3}\right) \\
&= f_{\mathcal{G}_{\tau_3}}\left(\boldsymbol{Y}_{\tau_3} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_3)}, \boldsymbol{\theta}_{\tau_3}\right).
\end{aligned}$$

According to the notation used in Formula (4.5), we have $\boldsymbol{Y}^{*}_{\tau_1} \equiv \boldsymbol{Y}_{\tau_1}, \boldsymbol{Y}^{*}_{\tau_2} \equiv \boldsymbol{Y}^{\varnothing}_{\tau_2}$ and $\boldsymbol{Y}^{*}_{\tau_3} \equiv \boldsymbol{Y}_{\tau_3}$. From Equation (4.14) we obtain for $f_{\mathcal{G}_{\tau_1}}(\cdot)$

$$f_{\mathcal{G}_{\tau_1}}\left(\boldsymbol{Y}_{\tau_1} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_1)}, \boldsymbol{\Theta}_{\tau_1}\right) = \mathcal{N}_{n,|\tau_1|}(\boldsymbol{Y}_{\tau_1} \,|\, \boldsymbol{X}_{\tau_1}\boldsymbol{B}_{\tau_1}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}_{\tau_1}),$$

where

$$\boldsymbol{X}_{\tau_1} = \left[\ \boldsymbol{1}_n\ \right]; \quad \boldsymbol{B}_{\tau_1} = \left[\ \mu_1 \quad \mu_3\ \right].$$

Then, for $f_{\mathcal{G}_{\tau_2}}(\cdot)$ we can write according to Equations (4.14) and (4.18),

$$\begin{aligned}
f_{\mathcal{G}_{\tau_2}}\left(\boldsymbol{Y}_{\tau_2} \,|\, \boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_2)}, \boldsymbol{\Theta}_{\tau_2}\right) &= \mathcal{N}_{n^{\varnothing},|\tau_2|}(\boldsymbol{Y}^{\varnothing}_{\tau_2} \,|\, \boldsymbol{X}^{\varnothing}_{\tau_2}\boldsymbol{B}_{\tau_2}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}_{\tau_2}) \\
&\cdot\ \mathcal{N}_{n^{\{2\}},|\tau_2|}(\boldsymbol{Y}^{\{2\}}_{\tau_2} \,|\, \boldsymbol{\Delta}^{\{2\}}_{\tau_2}, \boldsymbol{I}_{n^{\{2\}}}, (\boldsymbol{\Phi}^{\{2\}}_{\tau})^{-1})
\end{aligned}$$

where

$$\boldsymbol{X}^{\varnothing}_{\tau_2} = \begin{bmatrix} \mathbf{1}_{n^{\varnothing}} & \boldsymbol{y}^{\varnothing}_1 \end{bmatrix}; \quad \boldsymbol{B}_{\tau_2} = \begin{bmatrix} \mu_2 \\ \beta_{1,2} \end{bmatrix}.$$

Finally, for $f_{\mathcal{G}_{\tau_3}}(\cdot)$ *we have*

$$f_{\mathcal{G}_{\tau_3}}\big(\boldsymbol{Y}_{\tau_3}\,|\,\boldsymbol{Y}_{\mathrm{pa}_{\mathcal{G}}(\tau_3)}, \boldsymbol{\Theta}_{\tau_3}\big) \;=\; \mathcal{N}_{n,|\tau_3|}(\boldsymbol{Y}_{\tau_3}\,|\,\boldsymbol{X}_{\tau_3}\boldsymbol{B}_{\tau_3}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}_{\tau_3}),$$

with

$$\boldsymbol{X}_{\tau_3} = \begin{bmatrix} \mathbf{1}_n & \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y}_3 \end{bmatrix}; \quad \boldsymbol{B}_{\tau_3} = \begin{bmatrix} \mu_4 \\ \beta_{1,4} \\ \beta_{2,4} \\ \beta_{3,4} \end{bmatrix}.$$

Chapter 5

MCMC Methods

In this chapter we introduce the Markov Chain Monte Carlo (MCMC) methods that we adopt to perform structural learning of (interventional) essential graphs (I-EGs). To this end, we first construct a Markov chain on (interventional) Markov equivalence classes of DAGs that we use to explore the I-EG space and study related properties of interest (Section 5.1). The same is at the basis of the MCMC algorithm presented in Section 5.2 which allows for posterior inference on the model space. All the algorithms herein presented are implemented in R (**?**).

5.1 Markov chains on equivalence classes of DAGs

Markov chains represents a fundamental tool in model selection problems both from a Bayesian and a frequentist perspective. Once we collect a set of observations that we assume to be faithful to *some* statistical model (e.g. a DAG), we might be interested in evaluating *all* possible models belonging to a given space in order to "identify" the true data generating model or approximate the posterior distribution across models; see also Chapter 1.2. However, this is infeasible in many cases since the enumeration of such models is not possible or computationally demanding. Furthermore, in a Bayesian setting, the construction of a Markov chain on the model space is even fundamental when the objective is the approximation of posterior model probabilities (or related features) via MCMC procedures. In this section we resume Markov chains on EGs as introduced by Chickering (2002) and He et al. (2013) and then provide a coherent extension to the interventional setting described in Section 4.1.

5.1.1 Operators on graphs

From our perspective, a set of EGs can be taken in general as the model space. We can consider for instance the set of all EGs on q nodes (denoted by $\mathcal{S}_q$) or a proper subset. In the following we will often refer to the concept of sparsity. Let $|\mathcal{G}|$ be the number of edges (directed or undirected) in the EG $\mathcal{G}$. Given a threshold M we say that $\mathcal{G}$ is *sparse* if $|\mathcal{G}| \leq M$. M is usually a small multiple of the number of nodes q, for instance $M = rq$ with $r \in [1, 3]$ (He et al., 2013). We then denote with $\mathcal{S}_q^r$ the set of all EGs on q nodes satisfying the sparsity constraint $|\mathcal{G}| \leq rq$. Observe that if $rq \geq q(q-1)/2$, $\mathcal{S}_q^r$ and $\mathcal{S}_q$ coincide; when not specified $\mathcal{S}_q^r$ will include the case $\mathcal{S}_q$ as well. Our definition of sparsity is the same of He et al. (2013); other definitions of sparsity concern the maximum number of adjacent nodes in the graph, that is the maximum neighbourhood size.

To construct a Markov chain on EGs we first need to define the transitions among them. Chickering (2002) introduces a set of operators that can modify *locally* an EG. In other words, each operator involves a pair (or a triple) of nodes only. We consider six types of operators: inserting an undirected edge (InsertU for short), deleting an undirected edge (DeleteU), inserting a directed edge (InsertD), deleting a directed edge (DeleteD), making a v-structure (MakeV) and removing a v-structure (RemoveV). Each operator is then determined by two parts: the type and the modified edges. Therefore, the *modified graph* of an operator on $\mathcal{G}$ is the same as $\mathcal{G}$ except for the modified edges.

The modified graph of an operator on $\mathcal{G}$ is not in general an EG. For example in Figure 5.1, starting from the EG $\mathcal{G}$, the modified graphs $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_4$ are not EGs. Anyway, such operator can be *valid* as well in the sense that might *result* in a transition to an EG (He et al., 2013); see Definition 5.1.1. This is substantially

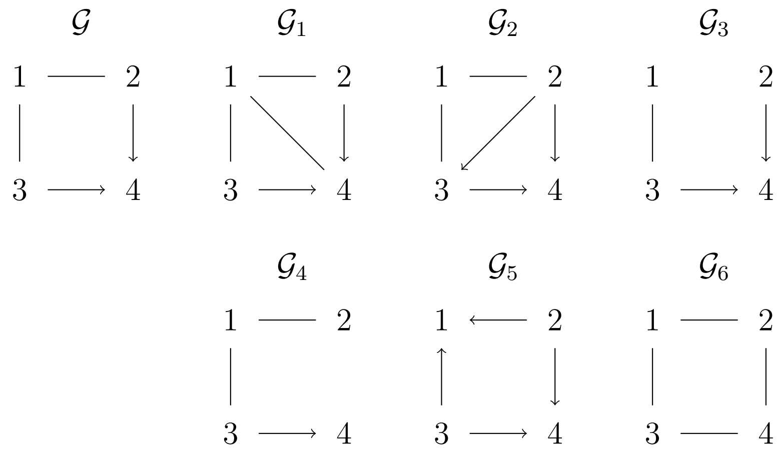


Figure 5.1: An EG $\mathcal{G}$ and the six corresponding modified graphs of the operators InsertU $1 - 4$ ($\mathcal{G}_1$), InsertD $2 \to 3$ ($\mathcal{G}_2$), DeleteU $1 - 2$ ($\mathcal{G}_3$), DeleteD $2 \to 4$ ($\mathcal{G}_4$), MakeV $2 \to 1 \leftarrow 3$ ($\mathcal{G}_5$), RemoveV $2 \to 4 \leftarrow 3$ ($\mathcal{G}_6$).

different from other authors, for instance Madigan et al. (1996), that admit only operators whose modified graph *is* an EG, which significantly reduces the number of possible transitions from each state (EG) of the chain. A valid operator is then defined as follows.

Definition 5.1.1. *Let $\mathcal{G}$ be an EG. An operator on $\mathcal{G}$ is valid if (1) the modified graph of the operator is a PDAG and has a consistent extension and (2) all modified edges in the modified graph occur in the resulting EG.*

Recall that a consistent extension of a Partially Directed Acyclic Graph (PDAG) $\mathcal{G}$ is a DAG on the same underlying set of edges, with the same orientations on the directed edges of $\mathcal{G}$ and the same set of v-structures (Dor & Tarsi, 1992). The two conditions of Definition 5.1.1 guarantee that a valid operator on the EG $\mathcal{G}$ brings about another EG (1) and the operator is *effective* (2), that is the modified edges of the operator appear in the resulting EG. In Figure 5.2, starting from the EG $\mathcal{G}^{(0)}$ we obtain the modified graph $\mathcal{G}^*$ of the operator RemoveV $2 \to 4 \leftarrow 3$. $\mathcal{G}^*$ is not an EG but admits a consistent extension ($\mathcal{G}^{**}$) and then condition (1) is satisfied. Moreover, since the modified edges $2 - 4 - 3$ occur the resulting EG

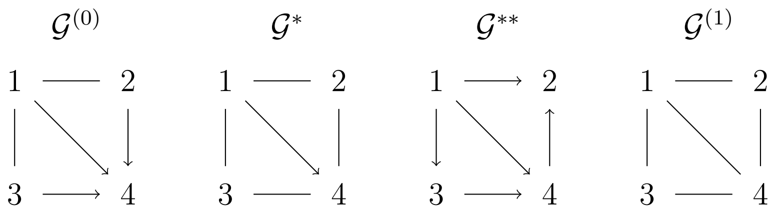


Figure 5.2: A graph $\mathcal{G}^{(0)}$, the corresponding modified graph of the operator RemoveV $2 \rightarrow 4 \leftarrow 3$ ($\mathcal{G}^{*}$), a consistent extension of $\mathcal{G}^{*}$ ($\mathcal{G}^{**}$) and the resulting EG $\mathcal{G}^{(1)}$.

Operator	Sparsity conditions	
InsertU $x - y$	(iu_1)	$\lvert\mathcal{G}\rvert < M$
InsertD $x \rightarrow y$	(id_1)	$\lvert\mathcal{G}\rvert < M$

Table 5.1: Sparsity conditions for operators InsertU, InsertD.

$\mathcal{G}^{(1)}$, the operator is valid.

Chickering (2002) and He et al. (2013) introduce a set of conditions that guarantee the validity of each operator. Such conditions are based on graphical features of the EG and the modified edges of each operator. We summarize them in Table 5.2, where $\Pi_x = \mathrm{pa}_{\mathcal{G}}(x)$, $\Omega_{xy} = \Pi_x \cap N_y$ and $N_x = \mathrm{ne}_{\mathcal{G}}(x)$; see also Section 2.1. Observe that in order to guarantee that the resulting EG of an operator is in $\mathcal{S}_q^r$, which generally depends on a sparsity constraint, the condition $|\mathcal{G}| < M$ where $M = rq$ must be satisfied as well for the operators InsertU and InsertD on the EG $\mathcal{G}$. In other terms, if the number of edges in $\mathcal{G}$ is equal to the upper bound M an edge addition is no longer possible. Such conditions are denoted by iu_1 and id_1 and reported in Table 5.1.

For each EG $\mathcal{G} \in \mathcal{S}_q^r$, let now $\mathcal{O}_{\mathcal{G}}$ be the corresponding set of valid operators. The set of all valid operators on $\mathcal{S}_q^r$ is then

$$\mathcal{O} = \bigcup_{\mathcal{G} \in \mathcal{S}_q^r} \mathcal{O}_{\mathcal{G}}. \tag{5.1}$$

Hence, we can define a Markov chain $\{\mathcal{G}^{(t)}\}$ on $\mathcal{S}_q^r$ as follows.

Operator	Validity conditions	
InsertU $x - y$	(iu_{21})	$\Pi_x = \Pi_y$
	(iu_{22})	every undirected path from x to y contains a node in $N_{x,y}$
DeleteU $x - y$	(du_1)	$N_{x,y}$ is a clique
InsertD $x \to y$	(id_{21})	$\Pi_x \neq \Pi_y$
	(id_{22})	$\Omega_{x,y}$ is a clique
	(id_{23})	every partially directed path from y to x contains a node in $\Omega_{x,y}$
DeleteD $x \to y$	(dd_1)	N_y is a clique
MaveV $x - z - y$	(mv_1)	every undirected path between x and y contains a node in $N_{x,y}$
RemoveV $x \to z \leftarrow y$	(rv_{11})	$\Pi_x = \Pi_y$
	(rv_{12})	$\Pi_x \cap N_{x,y} = \Pi_z \setminus \{x, y\}$
	(rv_{13})	every undirected path between x and y contains a vertex in $N_{x,y}$

Table 5.2: Six types of operators with the corresponding validity conditions.

Definition 5.1.2. *The Markov chain $\{\mathcal{G}^{(t)}\}$ determined by a set of valid operators $\mathcal{O}$ is generated as follows: start at an arbitrary EG $\mathcal{G}^{(0)} \in \mathcal{S}_q^r$ and repeat the following steps for $t = 0, 1, \dots$:*

1. *at the t-th step we are at an EG $\mathcal{G}^{(t)}$;*

2. *we choose an operator $o_{\mathcal{G}^{(t)}}$ uniformly from $\mathcal{O}_{\mathcal{G}^{(t)}}$; if the resulting EG $\mathcal{G}^*$ of $o_{\mathcal{G}^{(t)}}$ is in $\mathcal{S}_q^r$, we move to $\mathcal{G}^*$ and set $\mathcal{G}^{(t+1)} = \mathcal{G}^*$; otherwise we stay at $\mathcal{G}^{(t)}$ and set $\mathcal{G}^{(t+1)} = \mathcal{G}^{(t)}$.*

5.1.2 Markov chains and related properties

In this paragraph we resume some properties of finite discrete-time Markov chains. Let $p_{\mathcal{G}\mathcal{G}'}$ be the one-step transition probability from $\mathcal{G}$ to $\mathcal{G}'$ for any EG $\mathcal{G}$ and $\mathcal{G}'$ in $\mathcal{S}_q^r$. A Markov chain $\{\mathcal{G}^{(t)}\}$ on $\mathcal{S}_q^r$ is *irreducible* if starting at any state it can reach any EG in $\mathcal{S}_q^r$. If $\{\mathcal{G}^{(t)}\}$ is irreducible, there exists a unique distribution $\boldsymbol{\pi} = (\pi_{\mathcal{G}}, \mathcal{G} \in \mathcal{S}_q^r)$ satisfying balance equations

$$\pi_{\mathcal{G}} = \sum_{\mathcal{G} \in \mathcal{S}_q^r} \pi_{\mathcal{G}'} p_{\mathcal{G}'\mathcal{G}}; \tag{5.2}$$

see for instance Norris (1997), Theorems 1.7.7 and 1.5.6. Then, an irreducible Markov chain $\{\mathcal{G}^{(t)}\}$ is *reversible* if there exists a probability distribution $\boldsymbol{\pi}$ such that

$$\pi_{\mathcal{G}} p_{\mathcal{G}\mathcal{G}'} = \pi_{\mathcal{G}'} p_{\mathcal{G}'\mathcal{G}}. \tag{5.3}$$

In addition, $\boldsymbol{\pi}$ is the unique stationary distribution of the discrete-time Markov chain $\{\mathcal{G}^{(t)}\}$ if it is finite, reversible, and irreducible. In our case, a Markov chain on a set of EGs is reversible if the operators that determine the transitions among its states are reversible. He et al. (2013) introduce further properties of $\mathcal{O}$ with related conditions that guarantee that the Markov chain $\{\mathcal{G}^{(t)}\}$ is irreducible and reversible. These properties are *distinguishability*, *irreducibility* and *reversibility*. According to distinguishability, for each $\mathcal{G} \in \mathcal{S}_q^r$ and each of its direct successors

Operator	Perfectness condition	
InsertU $x - y$	(iu_3)	for any u that is a common child of x, y in $\mathcal{G}$, both $x \to u$ and $y \to u$ occur in the resulting EG of InsertU $x - y$
InsertD $x \to y$	(id_3)	for any u that is a common child of x, y in $\mathcal{G}$, $y \to u$ occurs in the resulting EG of InsertD $x \to y$
DeleteD $x \to y$	(dd_2)	for any v that is a parent of y but not a parent of x, directed edge $v \to y$ in $\mathcal{G}$ occurs in the resulting EG of DeleteD $x \to y$

Table 5.3: Perfectness conditions for operators InsertU, InsertD and DeleteD.

$\mathcal{G}'$, there is a unique operator that transforms $\mathcal{G}$ in $\mathcal{G}'$. It follows that, for each $\mathcal{G} \in \mathcal{S}_q^r$, different operators in $\mathcal{O}_\mathcal{G}$ result in different EGs. Equivalently, there is a one-to-one correspondence between operators in $\mathcal{O}_\mathcal{G}$ and resulting EGs (the *direct successors* of $\mathcal{G}$). Irreducibility states that, starting from $\mathcal{G} \in \mathcal{S}_q^r$, there is a positive probability to reach any other EG in $\mathcal{S}_q^r$ via a sequence of operators. Finally, according to reversibility, if $\mathcal{G}'$ is a direct successor of $\mathcal{G}$, then $\mathcal{G}$ is also a direct successor of $\mathcal{G}'$. A collection of valid operators satisfying these three properties is said to be *perfect*. In Table 5.3 we report three conditions for operators InsertU, InsertD, DeleteD which guarantee that $\mathcal{O}_\mathcal{G}$ is perfect. See also the original paper for further details about these properties.

To prove such conditions in an efficient way, three algorithms based on the notion of strong protection (see Definition 3.1.2) are also provided by the authors (He et al. 2013, Supplement).

One of the main advantages of the above mentioned properties is that the stationary probabilities $\pi_\mathcal{G}$ can be calculated efficiently if the Markov chain is reversible. Let $|\mathcal{O}_\mathcal{G}|$ be the number of operators in $\mathcal{O}_\mathcal{G}$; the transition probability

$p_{\mathcal{G}\mathcal{G}'}$ is given by

$$p_{\mathcal{G}\mathcal{G}'} = \begin{cases} 1/|\mathcal{O}_{\mathcal{G}}|, & \text{if } \mathcal{G}' \text{ is a direct successor of } \mathcal{G}; \\ 0, & \text{otherwise.} \end{cases} \tag{5.4}$$

An important result is summarized in Proposition 1 of He et al. (2013) which characterizes the stationary distribution of $\{\mathcal{G}^{(t)}\}$.

Proposition 5.1.1. *For the Markov chain $\{\mathcal{G}^{(t)}\}$ on $\mathcal{S}_q^r$ generated according to Definition 5.1.2, if $\mathcal{O}$ is perfect then:*

1. *the Markov chain $\{\mathcal{G}^{(t)}\}$ is irreducible and reversible;*
2. *there exists a unique stationary distribution $\boldsymbol{\pi}$ of $\{\mathcal{G}^{(t)}\}$ and $\pi_{\mathcal{G}} \propto |\mathcal{O}_{\mathcal{G}}|$.*

In Figure 5.3 we have the one-step transition matrix $\boldsymbol{P}$ for a Markov chain on $\mathcal{S}_3$, the set of all EGs on $q = 3$ nodes (see Figure 3.2 in Section 3.1). It is easy to prove that the corresponding Markov chain is irreducible and reversible. Its stationary distribution is then

$$\boldsymbol{\pi} \propto (3, 3, 3, 3, 4, 4, 4, 1, 1, 1, 3)^{\top}.$$

5.1.3 Exploring the EG space

The Markov chain summarized in the previous section can be used to describe properties of a set of EGs. This is the main objective of He et al. (2013). A sample of EGs is obtained as the result of a Markov chain and a feature of interest (for instance the number of undirected edges) is calculated for each EG. The aim is then to approximate the distribution of such feature. In this case the stationary distribution of the chain, $\boldsymbol{\pi}$, plays an essential role in the sense that samples thus obtained if re-weighted *according to* $\boldsymbol{\pi}$ can be considered as uniformly generated from the EG space under consideration. Without going into the details we simply obtain some results that are particularly relevant when model selection of EGs

$$\boldsymbol{P} = \left[\begin{array}{ccccccccccc}
0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\
0 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 0 & 0 & 1/4 & 1/4 \\
0 & 0 & 1/4 & 1/4 & 0 & 0 & 0 & 0 & 1/4 & 0 & 1/4 \\
0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 & 1/4 & 0 & 0 & 1/4 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0
\end{array}\right]
\begin{array}{c}
\mathcal{G}_1 \\ \mathcal{G}_2 \\ \mathcal{G}_3 \\ \mathcal{G}_4 \\ \mathcal{G}_5 \\ \mathcal{G}_6 \\ \mathcal{G}_7 \\ \mathcal{G}_8 \\ \mathcal{G}_9 \\ \mathcal{G}_{10} \\ \mathcal{G}_{11}
\end{array}$$

Figure 5.3: One-step transition matrix for a Markov chain on $\mathcal{S}_3$, the set of all EGs on $q = 3$ nodes (labelled as in Figure 3.2 of Section 3.1). Each probability $p_{\mathcal{G}_i \mathcal{G}_j}$ is obtained from Equation (5.4) according to conditions of Tables 5.2 and 5.3.

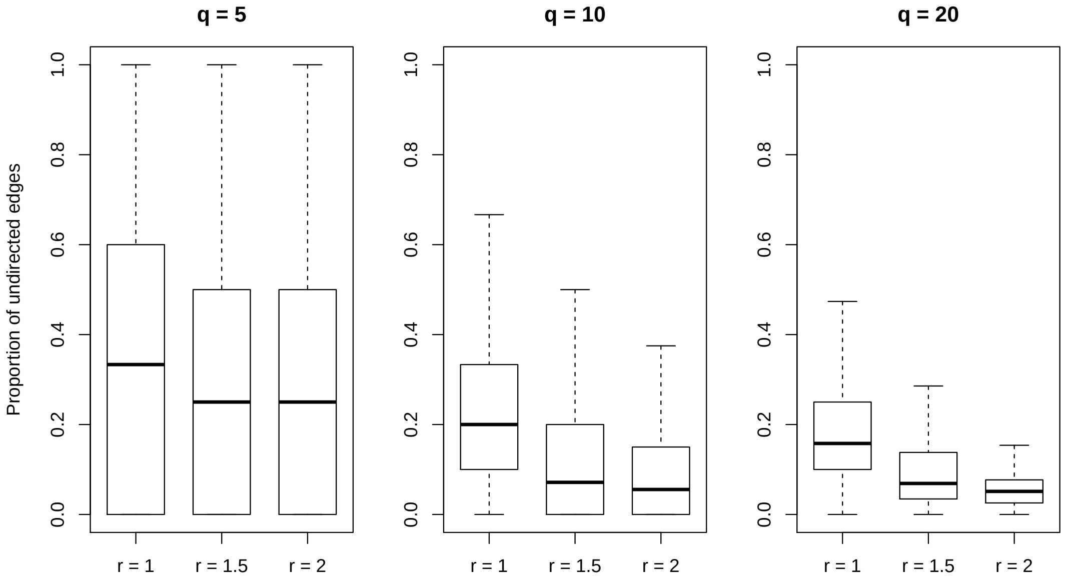


Figure 5.4: Box-plots of the proportion of undirected edges in the EG space with $q = \{5, 10, 20\}$ nodes and sparsity parameter $r = \{1, 1.5, 2\}$.

is performed in an interventional setting (Chapter 4). In Figure 5.4 we have the distribution of the proportion of undirected edges in the EG space $\mathcal{S}_q^r$, for $q \in \{5, 10, 20\}$ and $r \in \{1, 1.5, 2\}$. The number of undirected edges in an EG can be used in general as a measure of complexity of *causal learning* (He & Geng, 2008). Intuitively, if most EGs in the space have a *large* number of undirected edges, many interventions may be needed to direct them and then improve the identifiability of the true data generating DAG. From Figure 5.4 it appears that the proportion of undirected edges is relatively small and decreasing in q. Hence, it seems to be possible to infer the directions of the undirected edges via a “small” number of interventions.

5.1.4 Extension to the I-EG space

We now discuss how to extend the Markov chain above resumed in the interventional setting of Section 4.1. In such a case we are interested in exploring a set of interventional essential graphs ($\mathcal{I}$-EGs). Please recall that the $\mathcal{I}$-EG space depends on the family of intervention targets $\mathcal{I}$ (see Section 4.1). For a given $\mathcal{I}$ we then denote with ${}_{\mathcal{I}}\mathcal{S}_q$ the set of all $\mathcal{I}$-EGs ${}_{\mathcal{I}}\mathcal{G}$ on q nodes, with ${}_{\mathcal{I}}\mathcal{S}_q^r$ the set of all $\mathcal{I}$-EGs satisfying the sparsity constraint on $|{}_{\mathcal{I}}\mathcal{G}|$ (see Section 5.1.1). We introduce a novel definition of validity, called $\mathcal{I}$-validity, which is a natural extension of Definition 5.1.1 to the interventional setting. Related conditions that guarantee $\mathcal{I}$-validity ($iu_2^{\mathcal{I}}, id_2^{\mathcal{I}}, rv_2^{\mathcal{I}}$) are then provided.

Definition 5.1.3. *Let $\mathcal{I}$ be a conservative family of targets, ${}_{\mathcal{I}}\mathcal{G}$ an $\mathcal{I}$-EG. An operator on ${}_{\mathcal{I}}\mathcal{G}$ is $\mathcal{I}$-valid if (1) the modified graph of the operator is a PDAG and has a consistent extension and (2) all modified edges in the modified graph occur in the resulting $\mathcal{I}$-EG.*

Similarly to Definition 5.1.1, the two conditions guarantee that an $\mathcal{I}$-valid operator on ${}_{\mathcal{I}}\mathcal{G}$ brings about another $\mathcal{I}$-EG and the operator is effective. Recall now the following proposition from Theorem 4.1.2 in Section 4.1.

Proposition 5.1.2. *Let $\mathcal{I}$ be a conservative family of targets, ${}_{\mathcal{I}}\mathcal{G}$ an $\mathcal{I}$-EG. Then ${}_{\mathcal{I}}\mathcal{G}$ has no line $x - y$ for which there exists some $I \in \mathcal{I}$ such that $|\{x, y\} \cap I| = 1$.*

We first focus on the operator InsertU.

Let x, y be two non adjacent vertices and suppose to have $I \equiv x$ for some $I \in \mathcal{I}$, so that $|\{x, y\} \cap I| = 1$. Suppose then to apply the operator InsertU $x - y$. We know from Proposition 5.1.2 that $x - y$ does not occur in any ${}_{\mathcal{I}}\mathcal{G} \in {}_{\mathcal{I}}\mathcal{S}_q$. Hence, the modified edge $x - y$ cannot occur in the modified graph of the operator and then InsertU $x - y$ is not $\mathcal{I}$-valid.

It follows that an additional necessary condition to insert the undirected edge $x - y$ is that $|\{x, y\} \cap I| \neq 1$ for all $I \in \mathcal{I}$ (iu_{2I}), which means that x and y are not *singularly* involved in any intervention. The $\mathcal{I}$-validity condition for InsertU is then the following.

Definition 5.1.4. *Given the family of targets $\mathcal{I}$, the operator InsertU $x - y$ on the $\mathcal{I}$-EG ${}_{\mathcal{I}}\mathcal{G}$ is $\mathcal{I}$-valid ($iu_2^{\mathcal{I}}$) if and only if (iu_{2I}) $|\{x, y\} \cap I| \neq 1$ for all $I \in \mathcal{I}$, (iu_{21}) $\Pi_x = \Pi_y$, (iu_{22}) every undirected path from x to y contains a node in $N_{x,y}$.*

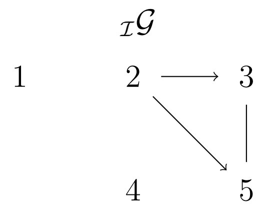


Figure 5.5: An $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}$ for $\mathcal{I} = \{\emptyset, \{1,2\}, \{4\}\}$. Operator InsertU $1 - 2$ is $\mathcal{I}$-valid according to Definition 5.1.4, while InsertU $2 - 4$ is not since (iu_{2I}) does not hold.

For instance, in Figure 5.5, given the family of targets $\mathcal{I} = \{\emptyset, \{1,2\}, \{4\}\}$, the operator InsertU $1 \to 2$ is $\mathcal{I}$-valid according to 5.1.4. However, InsertU $2 \to 4$ is not since condition iu_{2I} is not satisfied.

Consider now the operator InsertD. As a consequence of Proposition 5.1.2, we now have "more" directed edges (less undirected) in addition to the ones that we had in the (observational) EG space. The insertion of a directed edge $x \to y$ is also *possible* when $|\{x,y\} \cap I| = 1$ for some $I \in \mathcal{I}$ (id_{2I}) provided that such an insertion results in an $\mathcal{I}$-EG. More specifically, for two non adjacent vertices x and y, suppose that $|\{x,y\} \cap I| = 1$ for some $I \in \mathcal{I}$. In such case condition id_{21} (Table 5.2) is not necessary anymore for InsertD $x \to y$. See also Figure 5.6. We only need to assure that the insertion of $x \to y$ results in a chain graph with the features of an $\mathcal{I}$-EG which is guaranteed by conditions id_{22} and id_{23}; see He et al. (2013) for details. Hence, the $\mathcal{I}$-validity condition for the operator InsertD $x \to y$ can be stated as follows.

Definition 5.1.5. *Given the family of targets $\mathcal{I}$, the operator InsertD $x \to y$ on the $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}$ is $\mathcal{I}$-valid $(id_2^{\mathcal{I}})$ if and only if one of the following statements holds:*

1. *InsertD $x \to y$ is valid according to Definition 5.1.1 (or equivalently id_2 holds);*

2. *(id_{2I}) $|\{x,y\} \cap I| = 1$ for some $I \in \mathcal{I}$, (id_{22}) $\Omega_{x,y}$ is a clique, (id_{23}) every partially directed path from y to x contains a node in $\Omega_{x,y}$.*

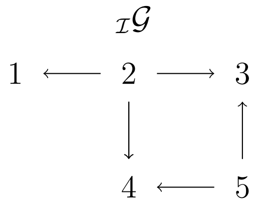


Figure 5.6: An $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}$ for $\mathcal{I} = \{\emptyset, \{1,2\}, \{5\}\}$. Operator InsertD $1 \to 4$ is $\mathcal{I}$-valid according to Condition 2 of Definition 5.1.5, while InsertD $3 \to 4$ is not.

Consider for instance Figure 5.6 and the family of targets $\mathcal{I} = \{\emptyset, \{1,2\}, \{5\}\}$. The operator InsertD $1 \to 4$ is $\mathcal{I}$-valid, since Condition 1 of Definition 5.1.5 holds for the target $I = \{1,2\}$, while InsertD $3 \to 4$ is not.

At last, consider the operator RemoveV $x \to z \leftarrow y$. Such operator basically converts two directed edges $x \to y \leftarrow z$ into the undirected v-structure $x - y - z$. Suppose first that $|\{x,z,y\} \cap I| = 1$ for some $I \in \mathcal{I}$. Then, the operator RemoveV $x \to y \leftarrow z$ cannot be $\mathcal{I}$-valid because it would result in $x - y - z$ which does not occur in any $_{\mathcal{I}}\mathcal{G} \in {_{\mathcal{I}}\mathcal{S}_q}$ according to Proposition 5.1.2. Hence, condition $|\{x,z,y\} \cap I| \neq 1$ is necessary. Moreover, suppose that $|\{x,z,y\} \cap I| = 2$ for some $I \in \mathcal{I}$. As before, the operator cannot result in a valid transition to any $_{\mathcal{I}}\mathcal{G}$ for Proposition 5.1.2. Then, condition $|\{x,z,y\} \cap I| \neq 2$ is also necessary. Suppose now that $|\{x,z,y\} \cap I| = 0$ for all $I \in \mathcal{I}$, which means that no interventions are made on the three nodes. Since Proposition 5.1.2 is satisfied, we obtain that the insertion of $x - z - y$ is allowed (provided that rv_1 holds as well). Similarly, if $|\{x,z,y\} \cap I| = 3$ for all $I \in \mathcal{I}$, then Proposition 5.1.2 is satisfied for both (x,z) and (z,y). Therefore, a necessary condition for the $\mathcal{I}$-validity of RemoveV $x \to z \leftarrow y$ is that $|\{x,z,y\} \cap I| \in \{0,3\}$ for all $I \in \mathcal{I}$.

Definition 5.1.6. *Given the family of targets $\mathcal{I}$, the operator RemoveV $x \to z \leftarrow y$ on the $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}$ is $\mathcal{I}$-valid ($rv_1^{\mathcal{I}}$) if and only if (rv_{1I}) $|\{x,z,y\} \cap I| \in \{0,3\}$ for all $I \in \mathcal{I}$, (rv_{11}) $\Pi_x = \Pi_y$, (rv_{12}) $\Pi_x \cap N_{x,y} = \Pi_z \setminus \{x,y\}$, ($rv_{13}$) every undirected path from x to y contains a node in $N_{x,y}$.*

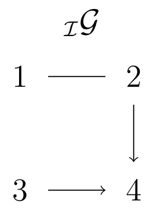


Figure 5.7: An $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}$ for $\mathcal{I} = \{\emptyset, \{1, 2\}\}$. Operator RemoveV $2 \to 4 \leftarrow 3$ is not $\mathcal{I}$-valid according to Definition 5.1.6.

For the remaining operators, DeleteU, DeleteD and MakeV, validity conditions are exactly the same as in the observational setting. In fact, the first two operators do not involve any intervention target; they change the skeleton of the $\mathcal{I}$-EG they are applied to, which always results in a different $\mathcal{I}$-EG, provided that validity holds. Instead, the operator MakeV introduces a v-structure, which is a distinguishing feature of $\mathcal{I}$-EGs and then brings about a different $\mathcal{I}$-EG. In the following six definitions we also extend the reversibility conditions of He et al. (2013) (Table 5.3) to the interventional setting. Recall that an operator is *perfect* if and only if it satisfies four conditions: validity, distinguishability, irreducibility and reversibility. For a given $\mathcal{I}$-EG $\mathcal{G}$ (we skip for simplicity the $\mathcal{I}$ subscript from $_{\mathcal{I}}\mathcal{G}$) on which the six types of operators are defined, let $IU_{\mathcal{G}}, DU_{\mathcal{G}}, ID_{\mathcal{G}}, DD_{\mathcal{G}}, MV_{\mathcal{G}}, RV_{\mathcal{G}}$ be the corresponding sets of perfect operators. These are constructed according to the following definitions.

Definition 5.1.7 (Perfect set $IU_{\mathcal{G}}$)**.** *For any two vertices x, y that are not adjacent in $\mathcal{G}$, the operator InsertU $x - y$ on $\mathcal{G}$ is in $IU_{\mathcal{G}}$ if and only if*

- *$|\mathcal{G}| < M$ (iu_1),*
- *InsertU $x - y$ is $\mathcal{I}$-valid ($iu_2^{\mathcal{I}}$),*
- *for any u that is a common child of x, y in $\mathcal{G}$, both $x \to u$ and $y \to u$ occur in the resulting $\mathcal{I}$-EG of InsertU $x - y$ (iu_3).*

Definition 5.1.8 (Perfect set $DU_{\mathcal{G}}$)**.** *For any undirected edge $x - y$ in $\mathcal{G}$, the operator DeleteU $x - y$ on $\mathcal{G}$ is in $DU_{\mathcal{G}}$ if and only if DeleteU $x - y$ is valid* (du_1)*.*

Definition 5.1.9 (Perfect set $ID_{\mathcal{G}}$)**.** *For any two vertices x, y that are not adjacent in $\mathcal{G}$, the operator InsertD $x \rightarrow y$ on $\mathcal{G}$ is in $ID_{\mathcal{G}}$ if and only if*

- *$|\mathcal{G}| < M$ (id_1),*
- *InsertD $x \rightarrow y$ is $\mathcal{I}$-valid ($id_2^{\mathcal{I}}$),*
- *for any u that is a common child of x, y in $\mathcal{G}$, $y \rightarrow u$ occurs in the resulting $\mathcal{I}$-EG of InsertD $x - y$ (id_3).*

Definition 5.1.10 (Perfect set $DD_{\mathcal{G}}$)**.** *For any directed edge $x \rightarrow y$ in $\mathcal{G}$, the operator DeleteD $x \rightarrow y$ on $\mathcal{G}$ is in $DD_{\mathcal{G}}$ if and only if*

- *DeleteD $x \rightarrow y$ is valid (dd_1),*
- *for any v that is a parent of y but not a parent of x, directed edge $v \rightarrow y$ in $\mathcal{G}$ occurs in the resulting $\mathcal{I}$-EG of DeleteD $x \rightarrow y$ (dd_2).*

Definition 5.1.11 (Perfect set $MV_{\mathcal{G}}$)**.** *For any subgraph $x - z - y$ in $\mathcal{G}$, the operator MakeV $x \rightarrow z \leftarrow y$ on $\mathcal{G}$ is in $MV_{\mathcal{G}}$ if and only if MakeV $x \rightarrow z \leftarrow y$ is valid (mv_1).*

Definition 5.1.12 (Perfect set $RV_{\mathcal{G}}$)**.** *For any v-structure $x \rightarrow z \leftarrow y$ in $\mathcal{G}$, the operator RemoveV $x \rightarrow z \leftarrow y$ on $\mathcal{G}$ is in $RV_{\mathcal{G}}$ if and only if RemoveV $x \rightarrow z \leftarrow y$ is $\mathcal{I}$-valid ($rv_1^{\mathcal{I}}$).*

Consider now the distribution of the proportion of undirected edges in the $\mathcal{I}$-EG space ${}_{\mathcal{I}}\mathcal{S}_q^r$. As mentioned, this feature is particularly interesting in an interventional setting because in principle we can improve the identifiability of the true data generating DAG by performing interventions on nodes. As a consequence, we “reduce” the number of undirected edges in the model space. Consider for instance the $\mathcal{I}$-EG space with sparsity parameter $r = 1$, ${}_{\mathcal{I}}\mathcal{S}_q^1$, which corresponds to

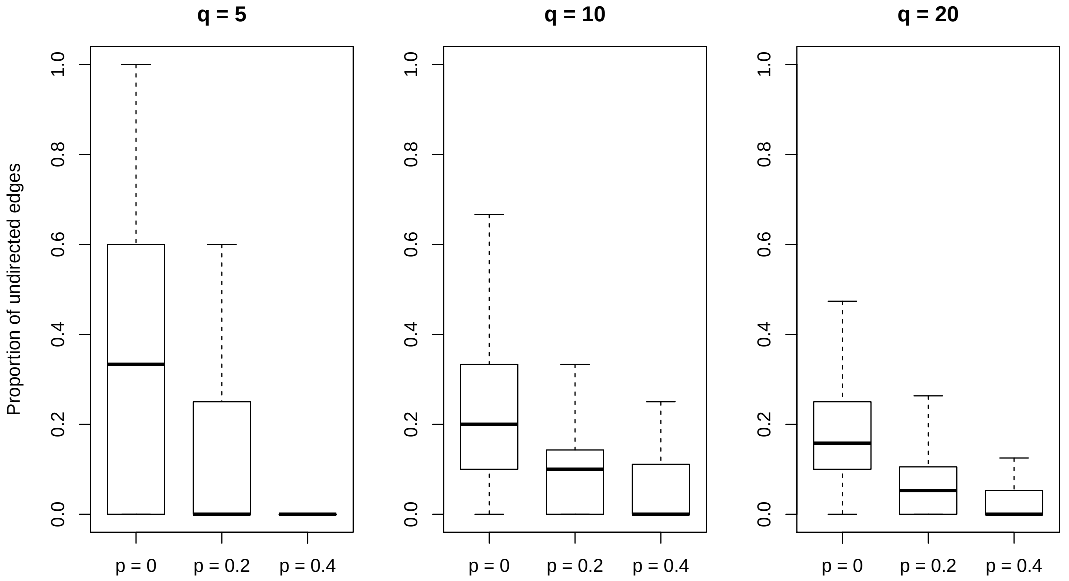


Figure 5.8: Box-plots of the proportion of undirected edges in the $\mathcal{I}$-EG space with $q \in \{5, 10, 20\}$ and sparsity parameter $r = 1$ under three scenarios: $p = 0$ (no interventions), $p = 0.2$ (interventions on 20% of nodes), $p = 0.4$ (interventions on 40% of nodes).

the setting in which interventions appear to be more effective; see Figure 5.4. Let $p \in [0, 1]$ be the proportion of intervened nodes. For each $q \in \{5, 10, 20\}$ we then consider three scenarios. In the first no interventions are performed, $p = 0$, in the second (third) we randomly assign to distinct single-node interventions (that is $|I| = 1$ for each $I \in \mathcal{I}$) the 20% (40%) of the nodes, $p = 0.2$ ($p = 0.4$). Figure 5.8 shows that the proportion of undirected edges rapidly decreases as we increase the number of intervened nodes. Moreover, since the proportion of undirected edges is decreasing in q, interventions are particularly effective for a moderate number of nodes.

5.1.5 Algorithms and further details

In the following we describe the implementation of the Markov chain on $\mathcal{I}$-EGs introduced in the current section together with some illustrative examples. For a given $\mathcal{I}$-EG $\mathcal{G}$, we first need to construct the corresponding perfect set of operators

$\mathcal{O}_{\mathcal{G}}$. To this end we introduce five sets. Let $I_{\mathcal{G}}$ be the set of all non adjacent pairs of vertices in $\mathcal{G}$, $U_{\mathcal{G}}$, $D_{\mathcal{G}}$, $V_{\mathcal{G}}^{D}$, $V_{\mathcal{G}}^{U}$ the set of undirected edges, directed edges, directed v-structure and undirected v-structures in $\mathcal{G}$ respectively. We use them to construct the set of all possible operators we are are going to test; see also the example in Table 5.4. Then, for a given $\mathcal{I}$-EG $\mathcal{G}$, the corresponding perfect set of operators is constructed according to Algorithm 2. Algorithm 1 returns a Markov chain of size T on the set of all $\mathcal{I}$-EGs on q vertices satisfying the sparsity constraint $|\mathcal{G}| < M$ $(M = rq)$.

Algorithm 1: Construction of a Markov chain on the $\mathcal{I}$-EG space

```
Data: $q$ (number of vertices), $M$ (maximum number of edges),
$T$ (length of the Markov chain), $\mathcal{I}$ (family of intervention targets)
Result: A Markov chain $\{\mathcal{G}^{(t)}, |\mathcal{O}_{\mathcal{G}^{(t)}}|\}$
Start from $\mathcal{G}^{(0)}$ (e.g. the empty graph);
for $t = 1, \dots, T$ do
    Construct the set of operators $\mathcal{O}_{\mathcal{G}^{(t)}}$ according to Algorithm 2 and
      compute $|\mathcal{O}_{\mathcal{G}^{(t)}}|$;
    Randomly sample an operator $o_{\mathcal{G}^{(t)}}$ in $\mathcal{O}_{\mathcal{G}^{(t)}}$;
    Apply $o_{\mathcal{G}^{(t)}}$ to $\mathcal{G}^{(t)}$ and obtain $\mathcal{G}^{(t+1)}$ as the resulting $\mathcal{I}$-EG of the
      operator
end
```

Consider for instance the EG $\mathcal{G}^{(0)}$ in Figure 5.9. We first construct the five sets $I_{\mathcal{G}^{(0)}}, U_{\mathcal{G}^{(0)}}, D_{\mathcal{G}^{(0)}}, V_{\mathcal{G}^{(0)}}^{D}, V_{\mathcal{G}^{(0)}}^{U}$ (first column of Table 5.4) and then check the corresponding operators to construct the perfect set $\mathcal{O}_{\mathcal{G}^{(0)}}$. We obtain

$$\begin{aligned}\mathcal{O}_{\mathcal{G}^{(0)}} = \{&\text{InsertD } 1 \rightarrow 4,\ \text{DeleteU } 1 - 2,\ \text{DeleteU } 1 - 3,\ \text{DeleteD } 2 \rightarrow 4,\\ &\text{DeleteD } 3 \rightarrow 4,\ \text{RemoveV } 2 \rightarrow 4 \leftarrow 3,\ \text{MakeV } 2 \rightarrow 1 \leftarrow 3\}\end{aligned}$$

and then $|\mathcal{O}_{\mathcal{G}^{(0)}}| = 7$. Hence, the transition probability from $\mathcal{G}^{(0)}$ to any of its direct successors $\mathcal{G}'$ is $p_{\mathcal{G}^{(0)},\mathcal{G}'} = 1/7$. Next, we randomly sample an operator

Algorithm 2: Construction of the perfect set of operators $\mathcal{O}_{\mathcal{G}}$

```
Data: An $\mathcal{I}$-EG $\mathcal{G}$, a family of intervention targets $\mathcal{I}$
Result: The corresponding perfect set of operators $\mathcal{O}_{\mathcal{G}}$
Set $\mathcal{O}_{\mathcal{G}}$ as empty set;
For each undirected edge $x - y$ in $U_{\mathcal{G}}$, consider the operator DeleteU
  $x - y$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $du_1$ holds;
For each directed edge $x \rightarrow y$ in $D_{\mathcal{G}}$, consider the operator DeleteD $x \rightarrow y$
  and add it to $\mathcal{O}_{\mathcal{G}}$ if $dd_1, dd_2$ hold;
For each directed $v$-structure $x \rightarrow y \leftarrow z$ in $V_{\mathcal{G}}^{D}$, consider the operator
  RemoveV $x \rightarrow y \leftarrow z$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $rv_1^{\mathcal{I}}$ holds;
For each undirected $v$-structure $x - y - z$ in $V_{\mathcal{G}}^{U}$, consider the operator
  MakeV $x \rightarrow y \leftarrow z$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $mv_1$;
if $|\mathcal{G}| < M$ $(id_1, iu_1)$ then
    for each pair of vertices $(x, y)$ in $I_{\mathcal{G}}$;
    consider InsertU $x - y$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $iu_2^{\mathcal{I}}, iu_3$ hold;
    consider InsertD $x \rightarrow y$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $id_2^{\mathcal{I}}, id_3$ hold;
    consider InsertD $x \leftarrow y$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $id_2^{\mathcal{I}}, id_3$ hold.
end
```

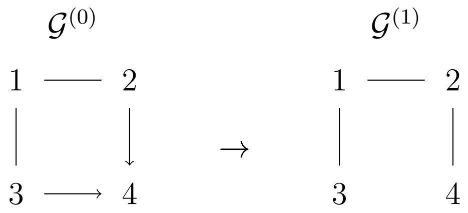


Figure 5.9: A step of a Markov chain on the EG space $\mathcal{S}_4$. We start from the EG $\mathcal{G}^{(0)}$; from the perfect set of operators $\mathcal{O}_{\mathcal{G}^{(0)}}$ we randomly sample $o_{\mathcal{G}^{(0)}} =$ DeleteD $3 \to 4$ and obtain the resulting EG $\mathcal{G}^{(1)}$ of the operator.

Set	$\mathcal{G}^{(0)}$	${}_{\mathcal{I}}\mathcal{G}^{(0)}$
$I_{\mathcal{G}}$	$\{(1,4),(2,3)\}$	$\{(1,4),(2,3)\}$
$U_{\mathcal{G}}$	$\{(1,2),(1,3)\}$	$\{(1,3)\}$
$D_{\mathcal{G}}$	$\{(2,4),(3,4)\}$	$\{(1,2),(2,4),(3,4)\}$
$V_{\mathcal{G}}^{D}$	$\{(2,4,3)\}$	$\{(2,4,3)\}$
$V_{\mathcal{G}}^{U}$	$\{(2,1,3)\}$	$\emptyset$

Table 5.4: Five sets of nodes involved in the construction of the sets of possible operators on $\mathcal{G}^{(0)}$ and ${}_{\mathcal{I}}\mathcal{G}^{(0)}$ to be tested.

$o_{\mathcal{G}^{(0)}} \in \mathcal{O}_{\mathcal{G}^{(0)}}$, for instance $o_{\mathcal{G}^{(0)}} =$ DeleteD $3 \to 4$, and obtain $\mathcal{G}^{(1)}$ as the resulting EG.

Given $\mathcal{I} = \{\emptyset, \{2\}\}$, consider now the $\mathcal{I}$-EG ${}_{\mathcal{I}}\mathcal{G}^{(0)}$ in Figure 5.10. From the five sets listed in the second column of Table 5.4 we obtain

$$\begin{aligned}\mathcal{O}_{{}_{\mathcal{I}}\mathcal{G}^{(0)}} = \{&\text{InsertD } 1 \to 4,\ \text{DeleteU } 1 - 3,\ \text{DeleteD } 1 \to 2,\\ &\text{DeleteD } 3 \to 4,\ \text{InsertD } 2 \to 3,\ \text{InsertD } 3 \to 2\}\end{aligned}$$

by checking the corresponding conditions. We randomly sample operator DeleteD $3 \to 4$ and obtain ${}_{\mathcal{I}}\mathcal{G}^{(1)}$ as the resulting $\mathcal{I}$-EG.

Algorithm 1 can be adapted for an accelerated version. For a given $\mathcal{I}$-EG $\mathcal{G}$ let $\mathcal{O}^{*}_{\mathcal{G}}$

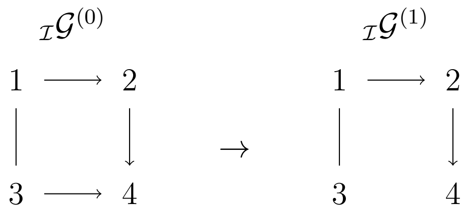

$_{\mathcal{I}}\mathcal{G}^{(0)}$
1
2
3
4
→
$_{\mathcal{I}}\mathcal{G}^{(1)}$
1
2
3
4

Figure 5.10: Two steps of a Markov chain on the $\mathcal{I}$-EG space $_{\mathcal{I}}\mathcal{S}_4$. We start from the $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}^{(0)}$; from the perfect set of operators $\mathcal{O}_{_{\mathcal{I}}\mathcal{G}^{(0)}}$ we randomly sample operator DeleteD 3 → 4 and obtain the resulting $\mathcal{I}$-EG $_{\mathcal{I}}\mathcal{G}^{(1)}$ of the operator.

be the set of all possible operators obtained from $I_{\mathcal{G}}$, $U_{\mathcal{G}}$, $D_{\mathcal{G}}$, $V_{\mathcal{G}}^D$, $V_{\mathcal{G}}^U$ and $|\mathcal{O}_{\mathcal{G}}^*|$ its cardinality. We are interested in computing $|\mathcal{O}_{\mathcal{G}}|$. Instead of testing all the operators in $\mathcal{O}_{\mathcal{G}}^*$ we consider a sample without replacement $\mathcal{O}_{\mathcal{G}}^{check}$ of size $[\alpha|\mathcal{O}_{\mathcal{G}}^*|]$, $\alpha \in (0,1)$, where $[x]$ denotes the closest integer value to x. We then check all the operators in $\mathcal{O}_{\mathcal{G}}^{check}$ and construct the perfect set $\tilde{\mathcal{O}}_{\mathcal{G}}$. Let $|\tilde{\mathcal{O}}_{\mathcal{G}}|$ be the number of perfect operators in $\mathcal{O}_{\mathcal{G}}^{check}$; an appropriate estimate of $|\mathcal{O}_{\mathcal{G}}|$ is then

$$|\mathcal{O}_{\mathcal{G}}|^{est} = \frac{|\tilde{\mathcal{O}}_{\mathcal{G}}|}{|\mathcal{O}_{\mathcal{G}}^{check}|} \cdot |\mathcal{O}_{\mathcal{G}}^*|.$$

The accelerated version of Algorithm 1 provides an approximation of the quantity $|\mathcal{O}_{\mathcal{G}}|$ which depends on the value of α. Details for the choice of the acceleration parameter α and the implementation of such algorithm in the EG space can be found in He et al. (2013).

This is easily adapted to the $\mathcal{I}$-EG space and substitutes step 3 in Algorithm 1, where instead of computing $|\mathcal{O}_{\mathcal{G}^{(t)}}|$ we estimate it through $|\mathcal{O}_{\mathcal{G}^{(t)}}|^{est}$.

5.2 MCMC for structural learning

In this section we illustrate the MCMC algorithm that we adopt to perform structural learning of Gaussian graphical models. In doing so, we focus on the Interventional Essential Graph ($\mathcal{I}$-EG) space described in Section 4.1 and then perform model selection and posterior inference of graphical models by directly scoring $\mathcal{I}$-EGs; see Section 4.2. To this end we first define a proposal distribution based on the Markov chain on the $\mathcal{I}$-EG space of Section 5.1. Next, we specify a prior on the $\mathcal{I}$-EG space through a beta-binomial distribution on the number of edges in the graph. These two quantities are then combined with the marginal likelihood of the $\mathcal{I}$-EGs to compute the acceptance rate of a standard Metropolis Hastings algorithm which approximates the posterior distribution over the $\mathcal{I}$-EG space.

5.2.1 Markov chains on I-EGs for model selection

In the previous section we introduced Markov chains on Markov equivalence classes of DAGs together with an extension to the interventional setting based on the proposal of He et al. (2013). The objective was to explore the $\mathcal{I}$-EG space under sparsity constraints on the maximum number of edges and describe some features of interest (for instance the proportion of undirected edges).

The set of operators presented in Section 5.1.1 satisfies some optimality conditions. Among these, it guarantees that starting at any graph $\mathcal{G}^{(0)}$ there is a non zero probability to reach any other graph in the space (reversibility). However, when the objective is "model selection using scoring criteria to identify models" (Chickering, 2002), such set of operators can not be sufficient to guarantee an efficient exploration of the model space.

Let $\mathcal{G}^{(t)}$ represent the graph visited at time t. In an MCMC setting, a new graph $\mathcal{G}'$ is then proposed and accepted with *some* probability α (Section 5.2.2) which "strongly" depends on the score (marginal likelihood) assigned to graph $\mathcal{G}'$. It follows that even if $p_{\mathcal{G}^{(t)},\mathcal{G}'}$ is not negligible, the transition from $\mathcal{G}^{(t)}$ to $\mathcal{G}'$ can be accepted with a *very small* probability if there is a little evidence from the data in favour of model $\mathcal{G}'$. Anyway, this can be very problematic if the transition to a "high score" model $\mathcal{G}''$ passes through $\mathcal{G}'$. In addition, the convergence of the MCMC can be very slow and even difficult to be reached. To avoid this kind of problem, we "reinstate" from Chickering (2002) the operator ReverseD $x \rightarrow y$ which is not needed for the Markov chain to be irreducible and reversible but it adds "extra connectivity" to the model space. We also highlight that such operator does not affect the properties of distinguishability, irreducibility and reversibility of the resulting Markov chain (Section 5.1.2). Validity conditions for the operator ReverseD $x \rightarrow y$ are given in Table 5.5; see also Chickering (2002).

Operator	Validity conditions	
ReverseD $x \rightarrow y$	(rd_1)	$\Omega_{y,x}$ is a clique
	(rd_2)	every partially directed path from x to y that does not include the edge $x \rightarrow y$ contains a node in $\Omega_{y,x} \cup N_y$

Table 5.5: Validity conditions for the operator ReverseD $x \rightarrow y$.

Moreover, when the operator is adopted in an interventional setting under the family of targets $\mathcal{I}$, no additional validity conditions are required. This happens because the operator ReverseD $x \rightarrow y$ applies to a directed edge whose presence, conditionally on the information about the intervened nodes enclosed in $\mathcal{I}$, has already been tested in InsertD $x \rightarrow y$.

5.2.2 MCMC scheme

Let $\mathcal{I}$ be a family of intervention targets, ${}_{\mathcal{I}}\mathcal{S}_q$ the set of all $\mathcal{I}$-EGs on q nodes and $\mathcal{G} \in {}_{\mathcal{I}}\mathcal{S}_q$ any $\mathcal{I}$-EG belonging to the space ${}_{\mathcal{I}}\mathcal{S}_q$. ${}_{\mathcal{I}}\mathcal{S}_q^r$ is then the set of all $\mathcal{I}$-EGs on q nodes satisfying the sparsity constraint $|\mathcal{G}| \leq rq$ (see Section 5.1.1). For a given $\mathcal{I}$-EG $\mathcal{G}$ we first need to construct the corresponding perfect set of operators $\mathcal{O}_{\mathcal{G}}$ which is used to compute the transition probability $p_{\mathcal{G},\mathcal{G}'} = 1/|\mathcal{O}_{\mathcal{G}}|$. In a standard

MCMC scheme we then randomly sample an operator $o_{\mathcal{G}} \in \mathcal{O}_{\mathcal{G}}$ and obtain $\mathcal{G}'$ by applying $o_{\mathcal{G}}$ to $\mathcal{G}$. This corresponds to sampling from a proposal distribution for $\mathcal{G}' \,|\, \mathcal{G}$. We assign coherently

$$q(\mathcal{G}' \,|\, \mathcal{G}) = 1/|\mathcal{O}_{\mathcal{G}}|.$$

Let $D_{\mathcal{G}}$ be the set of all directed edges of $\mathcal{G}$. Then, a sample from the proposal distribution of $\mathcal{G}'$ given $\mathcal{G}$ is obtained according to Algorithm 3.

```
Algorithm 3: Sampling from the proposal distribution of $\mathcal{G}' \,|\, \mathcal{G}$
  Data: An $\mathcal{I}$-EG $\mathcal{G}$, a family of intervention targets $\mathcal{I}$
  Result: A direct successor of $\mathcal{G}$, $\mathcal{G}'$, and the transition probability
          $q(\mathcal{G}' \,|\, \mathcal{G}')$
  Set $\mathcal{O}_{\mathcal{G}}$ as empty set;
  Construct the perfect set of operators $\mathcal{O}_{\mathcal{G}}$ as in Algorithm 2;
  For each directed edge $x \to y$ in $D_{\mathcal{G}}$, consider the operator ReverseD
    $x \to y$ and add it to $\mathcal{O}_{\mathcal{G}}$ if $rd_1, rd_2$ hold;
  Randomly sample an operator $o_{\mathcal{G}} \in \mathcal{O}_{\mathcal{G}}$;
  Apply $o_{\mathcal{G}}$ to $\mathcal{G}$ and obtain $\mathcal{G}'$ as the resulting $\mathcal{I}$-EG of the operator;
  Compute $q(\mathcal{G}' \,|\, \mathcal{G}) = 1/|\mathcal{O}_{\mathcal{G}}|$.
```

Let $m_{\mathcal{G}}(\boldsymbol{Y})$ be the marginal likelihood of graph $\mathcal{G}$ given the data $\boldsymbol{Y}$, $p(\mathcal{G})$ a prior on $\mathcal{G}$ and $q(\mathcal{G} \,|\, \cdot)$ a proposal distribution. In order to have an appropriate posterior sample of $\mathcal{I}$-EGs, the transition from $\mathcal{G}$ to $\mathcal{G}'$ proposed by Algorithm 3 is accepted with probability

$$\alpha = \min\left\{1; \frac{m_{\mathcal{G}'}(\boldsymbol{Y})}{m_{\mathcal{G}}(\boldsymbol{Y})} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} \,|\, \mathcal{G}')}{q(\mathcal{G}' \,|\, \mathcal{G})}\right\}. \tag{5.5}$$

As the prior on $\mathcal{G}$, $p(\mathcal{G})$, a common assumption is a beta-binomial on the adiacency matrix of $\mathcal{G}^u$, where $\mathcal{G}^u$ is the skeleton of $\mathcal{G}$ (same arcs as in $\mathcal{G}$, but with no orientation),

$$\begin{aligned} \mathcal{G}^u_{(j)} \,|\, \pi &\overset{i.i.d}{\sim} \mathrm{Ber}(\pi), \quad j = 1, \dots, q(q-1)/2, \\ \pi &\sim \mathrm{Beta}(a, b), \end{aligned} \tag{5.6}$$

being $\mathcal{G}^u_{(j)}$ the j-th element of the vectorized lower triangular part of the adjacency matrix of $\mathcal{G}^u$ and $q(q-1)/2$ the maximum number of edges in an $\mathcal{I}$-EG on q nodes; see also Bhadra & Mallick (2013). Notice that the prior on $\mathcal{G}$ only depends on the skeleton of the graph: two $\mathcal{I}$-EGs with the same number of edges (directed or undirected) will be assigned the same prior probability. More elaborate priors, specifically targeted to EGs, to our knowledge are not available in the literature, and are beyond the scope of the present work. The ratio between the priors on $\mathcal{G}'$ and $\mathcal{G}$ is then

$$\frac{p(\mathcal{G}')}{p(\mathcal{G})} = \frac{\Gamma(|\mathcal{G}'|+a)}{\Gamma(|\mathcal{G}|+a)} \cdot \frac{\Gamma\left(\frac{q(q-1)}{2} - |\mathcal{G}'| + b\right)}{\Gamma\left(\frac{q(q-1)}{2} - |\mathcal{G}| + b\right)}, \tag{5.7}$$

where $|\mathcal{G}|$ denotes the number of edges in $\mathcal{G}$. A common choice for a and b is $a = b = 1$ so that $\pi \sim \text{Unif}(0,1)$. However, to favor sparsity, we can set $a < b$ which implies $\mathbb{E}(\pi) < 0.5$. Further details for the choice of a and b are given in the simulation setting of Chapter 6. Finally, an MCMC on ${}_{\mathcal{I}}\mathcal{S}^r_q$ can be constructed as in Algorithm 4.

Algorithm 4: An MCMC on the $\mathcal{I}$-EG space ${}_{\mathcal{I}}\mathcal{S}^r_q$

```
Data: $\mathcal{G}^{(0)}$ (an arbitrary initial graph), $T$ (length of the chain)
Result: A sample $\{\mathcal{G}^{(t)}\}$ from $p(\mathcal{G} \mid \boldsymbol{Y}), \mathcal{G} \in {}_{\mathcal{I}}\mathcal{S}^r_q$
Start from $\mathcal{G}^{(0)}$ (e.g. the empty graph);
for $t = 1, \dots, T$ do
    set $\mathcal{G} = \mathcal{G}^{(t-1)}$;
    sample $\mathcal{G}'$ and compute $q(\mathcal{G}' \mid \mathcal{G})$ as in Algorithm 3;
    compute the probability of acceptance $\alpha$ in Formula (5.5);
    update $\mathcal{G}^{(t)} = \mathcal{G}'$ with probability $\alpha$, $\mathcal{G}^{(t)} = \mathcal{G}^{(t-1)}$ with probability
      $(1-\alpha)$;
end
```

Chapter 6

Experiments

In this chapter we apply our methodology, that we name Objective Bayes Interventional Essential graph Search (OBIES), to simulation settings and to the analysis of protein-signaling data (Sachs et al., 2005). In addition, we compare OBIES with the Greedy Interventional Equivalence Search (GIES) method introduced by Hauser & Bühlmann (2012) and implemented in the `R` package `pcalg` (**?**); see also Hauser & Bhlmann (2015). To this end, we start with some considerations about the need to choose a single model estimate (even in a Bayesian setting) for comparison purposes.

6.1 Preliminaries

For a collection of distinct models, $\{\mathcal{M}_k, k = 1, \ldots, K\}$, and given the data $\boldsymbol{Y}$, assume to have computed the posterior probabilities $p(\mathcal{M}_k \,|\, \boldsymbol{Y})$; see also Section 1.2. We accept the general thinking that "in principle, all inference and prediction should be based on the overall joint posterior distribution", like in Bayesian model averaging techniques (Pericchi, 2005). However, in many cases a single model estimate is required. Barbieri & Berger (2004) highlights that the *highest probability model*, that is the model with the highest posterior probability associated, is not in general an optimal choice from a predictive viewpoint. Instead, in a Gaussian linear regression setting, it was shown that the *median probability* *model* is predictively optimal. In such context, the median probability model is obtained by including all variables whose posterior inclusion probability is greater than 0.5.

In an MCMC framework, once we observe $\{\mathcal{M}^{(0)}, \dots, \mathcal{M}^{(T)}\}$ as the result of an MCMC algorithm on a specific model space, the objective is typically to *approximate* the posterior distribution across models or equivalently to *estimate* posterior model probabilities; see again Section 1.2. The number of visits of each model over the total number of iterations T is generally used as such approximation.

In model selection of graphical models, the median probability model can be naively constructed by including all edges whose probability of inclusion is greater than 0.5. More specifically, we start defining the *marginal posterior probability of inclusion* of the edge $u \to v$ given the data $\boldsymbol{Y}$ as

$$p_{u \to v}(\boldsymbol{Y}) = \sum_{\mathcal{G} \in \mathcal{S}_{u \to v}} p(\mathcal{G} \,|\, \boldsymbol{Y}), \tag{6.1}$$

where $\mathcal{S}_{u \to v}$ is the set of visited $\mathcal{I}$-EGs containing the directed edge $u \to v$ (recall that an undirected edge $u - v$ is equivalent to $u \to v$ and $v \to u$). The *median probability (graph) model* is then defined as the graph containing only those directed edges $u \to v$ such that $p_{u \to v}(\boldsymbol{Y}) > 0.5$. In general, the median probability model thus obtained is not guaranteed to be an $\mathcal{I}$-EG, but it is a partially directed acyclic graph (PDAG). From the median probability model $\mathcal{G}$, which may contain both directed and undirected edges, we first obtain a directed version (DAG) as follows. We start taking all the directed edges in $\mathcal{G}$. Then, for each undirected edge $u - v$ in $\mathcal{G}$ we take $u \to v$ if and only if $p_{u \to v}(\boldsymbol{Y}) > p_{v \to u}(\boldsymbol{Y})$. Finally, from the DAG $\mathcal{D}$ thus obtained, we construct the corresponding $\mathcal{I}$-EG, that is the representative of the interventional equivalence class of $\mathcal{D}$. This is done through the function `dag2essgraph` in the `R` package `pcalg` (**?**). We call such result *projected median probability (graph) model.* We then use the projected median probability model for comparison purposes, typically when a single graph estimate is required to compute distance-based indexes between models (e.g the structural Hamming distance with respect to the *true DAG*).

The median probability model specifies a threshold for edge inclusion $k = 0.5$. Alternatively, it is possible to choose k by looking at the *expected false discovery rate* (FDR) (Benjamini & Hochberg, 1995).

Set for simplicity $p_{u\to v}(\boldsymbol{Y}) \equiv p_{u\to v}$; see Equation (6.1). Then, we can define the expected FDR for a given threshold $k \in [0,1]$ as

$$\mathrm{FDR}(k) = \frac{\sum_{u=1}^{q}\sum_{u\neq v}(1-p_{u\to v})\mathbb{I}(p_{u\to v}\geq k)}{\sum_{u=1}^{q}\sum_{u\neq v}\mathbb{I}(p_{u\to v}\geq k)}, \tag{6.2}$$

being $\mathbb{I}$ the indicator function. More specifically, we can start constructing for a grid of thresholds a collection of *projected quantile probability models*. Each one, being associated to a value of $k \in [0,1]$, is obtained by including all edges such that $p_{u\to v} > k$ and then constructing the corresponding projection over the $\mathcal{I}$-EG space (as for the median probability model). Assuming that the projected quantile probability model of order k is the true model, FDR(k) provides a measure of evidence against such assumption. The denominator of (6.2) corresponds to the number of edges in the projected quantile probability model of order k. The numerator is instead the sum of the "probabilities of non inclusion" $(1-p_{u\to v})$ for those edges that are included in the estimated graph, thus representing a measure of *false positiveness*. One can show that FDR(k) is a non decreasing function of k. Hence, one can select k so that the expected FDR is below a desired level, typically 0.05. See also Peterson et al. (2015) for the adoption of the FDR in multiple Gaussian graphical model selection.

6.2 Simulations

In this section we apply OBIES to simulated data sets. We first describe the data generating process of a collection of observational and interventional data under a given DAG $\mathcal{D}$ and a family of intervention targets $\mathcal{I}$. We then set the parameters of the MCMC algorithm described in Section 5.2 and present our results on a variety of simulation scenarios. At last, we compare OBIES with the Greedy Interventional Equivalence Search (GIES) method.

6.2.1 Simulation setting

A simulation framework is characterized by the triple $(q, n^{\emptyset}, p)$, where $q \in \{5, 10, 20, 40\}$ is the number of nodes, $n^{\emptyset} \in \{100, 200, 500, 1000\}$ the number of observational data and $p \in \{0, 0.2, 0.4, 0.8\}$ the proportion of intervened nodes. For simplicity we consider multiple but *single-node* interventions, that is $|I_k| = 1$ for each $I_k \in \mathcal{I}$. As an example, for $q = 5$ we will consider scenarios with no interventions (observational case, $p = 0$), interventions on one node randomly chosen, $\mathcal{I} = \{\emptyset, \{u\}\}$, $u \in \{1, \dots, 5\}$ $(p = 0.2)$, interventions on two nodes $\mathcal{I} = \{\emptyset, \{u\}, \{v\}\}$, $u, v \in \{1, \dots, 5\}$, $u \neq v$ $(p = 0.4)$ and similarly for $p = 0.8$. For each target of intervention I_k we then set the number of interventional data as $n^{(k)}(q, n^{\emptyset}) = n^{\emptyset}q/100$. Table 6.1 summarizes the simulation setting parameters.

	$n^{\emptyset} = 100$	$n^{\emptyset} = 200$	$n^{\emptyset} = 500$	$n^{\emptyset} = 1000$
$q = 5$	$n^{(k)} = 5$	$n^{(k)} = 10$	$n^{(k)} = 25$	$n^{(k)} = 50$
$q = 10$	$n^{(k)} = 10$	$n^{(k)} = 20$	$n^{(k)} = 50$	$n^{(k)} = 100$
$q = 20$	$n^{(k)} = 20$	$n^{(k)} = 40$	$n^{(k)} = 100$	$n^{(k)} = 200$
$q = 40$	$n^{(k)} = 40$	$n^{(k)} = 80$	$n^{(k)} = 200$	$n^{(k)} = 400$

Table 6.1: Simulation setting parameters; number of interventional data for each target I_k, $n^{(k)}$, as a function of the number of nodes q and the number of observational data $n^{\emptyset}$.

6.2.2 Data generation

For each scenario, 40 datasets, corresponding to 40 true DAGs, are generated. Each dataset, which contains both observational and interventional data, is obtained as follows. For a given q, we randomly generate a topologically ordered DAG $\mathcal{D}$ with probability of edge inclusion $p_{edge} = 3/(2q-2)$ (Peters & Bühlmann, 2014). The DAG thus obtained is the responsible of a data generating process and implies the set of equations

$$Y_{i,j} = \mu_j + \sum_{k \in \mathrm{pa}_D(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \tag{6.3}$$

for $i = 1, \ldots, n^{\emptyset}$, $j = 1, \ldots, q$, where $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ and $Y_{i,j}$ are independent with respect to i. For each j we fix $\mu_j = 0$ and $\sigma_j^2 = 1$, while regression coefficients $\beta_{k,j}$ are uniformly chosen in the interval $[-1, -0.1] \cup [0.1, 1]$, still following Peters & Bühlmann (2014). An observational dataset of size $n^{\emptyset}$ is then generated accordingly.

Next, for a given p (proportion of intervened nodes) we randomly sample without replacement $[pq]$ nodes in $\{1, \ldots, q\}$ which represents intervention targets of size one, $I_1, \ldots, I_{pq}$. For each I_k we first obtain the corresponding intervention DAG $\mathcal{D}^{I_k}$ (see Section 4.1) which implies the set of equations

$$Y_{i,j} = \begin{cases} \mu_j + \sum\limits_{k \in \mathrm{pa}_{\mathcal{D}}(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j} & \text{if } j \neq I_k, \\ \delta_j + \epsilon_{i,j} & \text{if } j = I_k, \end{cases} \tag{6.4}$$

for $i = 1, \ldots, n^{(k)}, j = 1, \ldots, q$, where again $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$, $\epsilon_{i,j} \sim \mathcal{N}(0, \phi_j^2)$ and $Y_{i,j}$ are independent with respect to i. For each j we fix $\delta_j = 0$ and $\phi_j^2 = 0.1$. $n^{(k)}$ interventional data are then generated accordingly.

6.2.3 MCMC parameters

For the MCMC algorithm of Section 5.2 we need to choose the hyperparameters a and b of the prior $p(\mathcal{G})$ which is involved in the acceptance probability α (Formula 5.5). We fix $a = 1, b = (2q - 2)/3 - 1$, so that $\mathbb{E}(\pi) = 3/(2q - 2)$. In this way the expected prior edge inclusion probability resembles the sparse simulation setting as defined in Section 6.2.2. We note however that all the results are generally insensitive to the choice of a and b, because the ratio of marginal likelihoods $m_{\mathcal{G}'}(\boldsymbol{Y})/m_{\mathcal{G}}(\boldsymbol{Y})$ is by far the leading factor in the acceptance probability of the proposed $\mathcal{I}$-EG.

As describe in Chapter 5.1 we can also constrain the $\mathcal{I}$-EG space ${}_{\mathcal{I}}\mathcal{S}_q$ to a subspace ${}_{\mathcal{I}}\mathcal{S}_q^r$ with no more than a given number of edges (rq), so that sparsity can be introduced to improve structural learning. Specifically, we fix $r = 2$, that is we require that the number of edges is not higher than 2 the number of nodes. This threshold is well above the number of edges expected in the true DAG in each simulation scenario (3.75, 7.5, 15 and 30 edges respectively for 5, 10, 20 and 40 nodes). For each scenario, an MCMC is then implemented in the following way: on the basis of pilot runs, we choose the total number of iterations T and the initial *burn-in* period to remove from the calculation of the final estimate of the graph. The pilot runs are also used as a diagnostic tool of appropriate MCMC convergence and mixing. We choose $T = \{5000, 25000, 50000, 100000\}$ for $q = \{5, 10, 20, 40\}$ respectively.

6.2.4 Results

In the following we present some results from the application of OBIES to the simulation scenarios described in the previous paragraphs. To understand how much interventions can improve the identifiability of the true data generating DAG, we measure the Structural Hamming Distance (SHD) between the projected median probability model and the true DAG. The SHD represents the number of edge insertions, deletions or flips needed to transform the estimated I-EG into the true I-EG. Please note that, under each setting defined by q, the benchmark of our comparison is represented by the same set of (40) true DAGs. In Figure 6.1 we report the boxplots of the SHD values over the 40 replicates under the simulation settings defined by $q, n^{\emptyset}$ and p. For each q and $n^{\emptyset}$ we observe that as the proportion of intervened nodes p increases, the SHDs between estimated graph and true DAG become smaller. Moreover, such reduction is all the more effective as $n^{\emptyset}$ (and so n^{I}, see Table 6.1) grows. As q increases, modelling jointly observational and interventional data produced under $p = 0.2$ results in a substantial reduction of the SHDs with respect to the true DAG. With reference to the $q = 20$ setting, we observe that interventions on the 40% of nodes *randomly chosen* are sufficient to strongly reduce the uncertainty around the true DAG estimate, especially for large sample sizes. This means that we have hope to discover the true DAG model via a *small* number of interventions selected according to an optimal design of experiments. In Table 6.2 we also report summary statistics (mean and standard deviation) of the SHDs represented in Figure 6.1. The behaviour observed in Figure 6.1 is confirmed by the average SHDs too.

As mentioned, our $\mathcal{I}$-EG estimate is the projected median probability model. The median probability model specifies a threshold for edge inclusion of 0.5. By varying this threshold one obtains distinct projected quantile probability models. Each such model can be used as an $\mathcal{I}$-EG estimate. Let TP denotes the number of true positive, that is the number of edges in the true DAG that are estimated correctly, and FP the number of edges in the estimated graph that are not present in the true DAG (false positive). Remember that an undirected edge $u - v$ is equivalent to $u \rightarrow v$ and $v \rightarrow u$. Figure 6.2 reports for the setting $q = 40$, $n^{\emptyset} = \{200, 500\}$ and $p \in \{0, 0.2, 0.4\}$, the Receiver Operating Characteristic (ROC) plots for the projected quantile probability models, obtained by "averaging" with respect to the 40 simulations.

$\boldsymbol{q = 5}$	$n^{\varnothing} = 100$	$n^{\varnothing} = 200$	$n^{\varnothing} = 500$	$n^{\varnothing} = 1000$
$p = 0$	4.95 (2.06)	4.45 (1.95)	4.28 (1.84)	3.79 (2.05)
$p = 0.2$	3.30 (1.45)	3.05 (1.43)	2.00 (1.50)	1.90 (1.35)
$p = 0.4$	2.88 (1.73)	2.25 (1.74)	1.91 (1.35)	1.80 (1.18)

$\boldsymbol{q = 10}$	$n^{\varnothing} = 100$	$n^{\varnothing} = 200$	$n^{\varnothing} = 500$	$n^{\varnothing} = 1000$
$p = 0$	6.78 (3.69)	5.78 (3.58)	5.20 (3.32)	5.08 (3.25)
$p = 0.2$	3.70 (2.03)	3.35 (1.90)	3.17 (2.31)	2.88 (2.21)
$p = 0.4$	3.10 (1.85)	2.17 (1.74)	2.13 (1.51)	1.50 (1.13)

$\boldsymbol{q = 20}$	$n^{\varnothing} = 100$	$n^{\varnothing} = 200$	$n^{\varnothing} = 500$	$n^{\varnothing} = 1000$
$p = 0$	11.07 (3.11)	9.57 (2.16)	8.40 (2.09)	8.00 (2.04)
$p = 0.2$	5.47 (2.72)	4.80 (2.43)	3.70 (1.91)	3.05 (1.52)
$p = 0.4$	4.25 (2.06)	3.25 (1.75)	1.32 (1.27)	1.15 (0.98)

$\boldsymbol{q = 40}$	$n^{\varnothing} = 100$	$n^{\varnothing} = 200$	$n^{\varnothing} = 500$	$n^{\varnothing} = 1000$
$p = 0$	21.75 (3.30)	18.75 (2.93)	16.45 (2.43)	15.35 (3.37)
$p = 0.2$	13.18 (3.74)	11.97 (3.48)	12.07 (3.27)	9.98 (3.30)
$p = 0.4$	8.40 (3.49)	8.00 (3.54)	7.12 (4.77)	6.38 (3.73)

Table 6.2: Mean (standard deviation) of the structural Hamming distances between OBIES estimate and true DAG over 40 data sets for number of nodes $q \in \{5, 10, 20, 40\}$, $n^{\varnothing} \in \{100, 200, 500, 1000\}$ and $p \in \{0, 0.2, 0.4\}$.

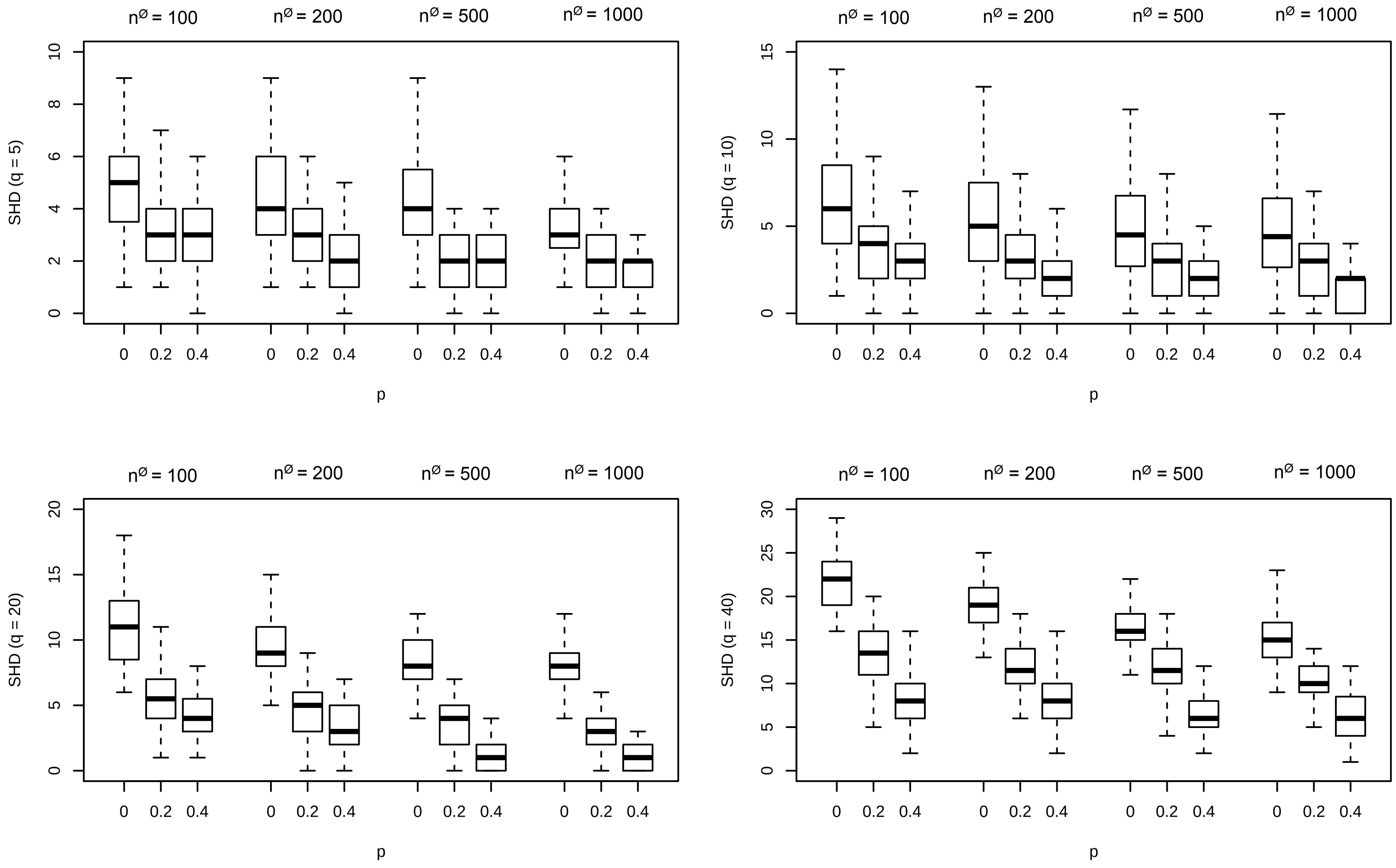
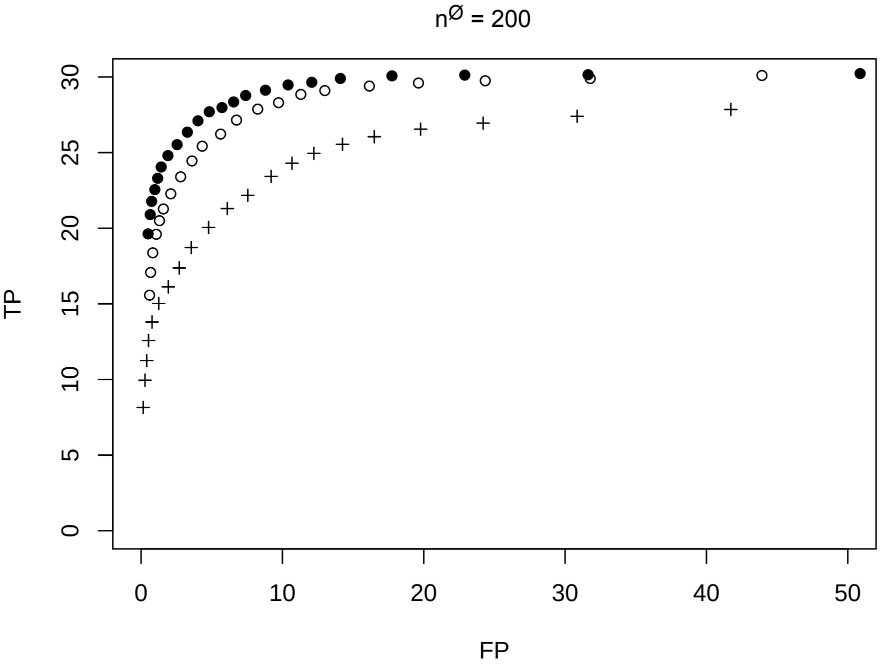


Figure 6.1: Box plots of the structural Hamming distances between OBIES estimate and true DAG over 40 data sets for number of nodes $q = \{5, 10, 20, 40\}$, number of observational data $n^{\emptyset} = \{100, 200, 500, 1000\}$ and proportion of intervened nodes $p \in \{0, 0.2, 0.4\}$.



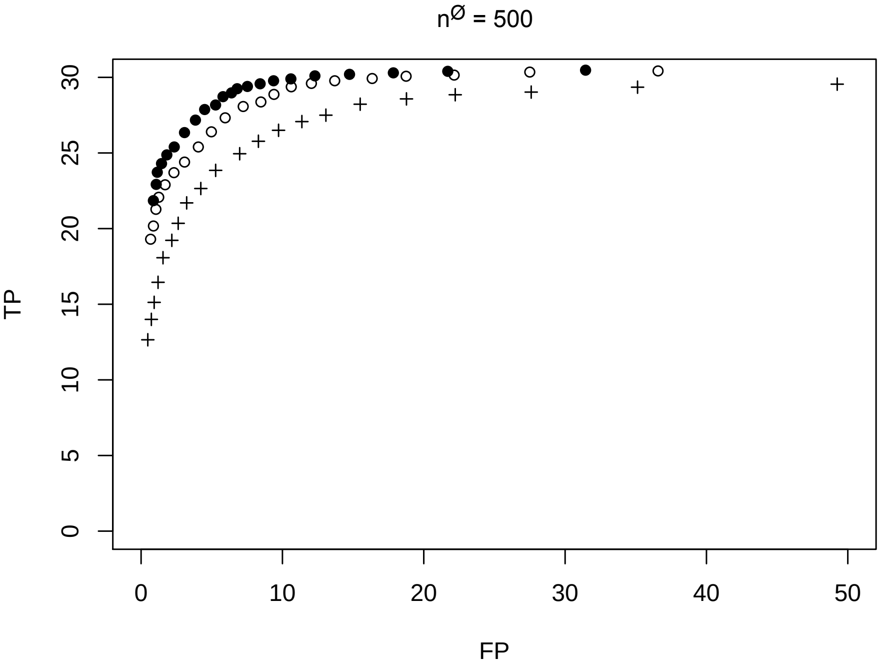


Figure 6.2: ROC plots (averaged over the 40 simulations) of the projected quantile probability model estimates for $q = 40$, $n^{\varnothing} = \{200, 500\}$ and proportion of intervened nodes $p = 0$ ($+$), $p = 0.2$ ($\circ$), $p = 0.4$ ($\bullet$).

6.2.5 Comparison with GIES

The Greedy Equivalence Search (GES) algorithm is a search-and-score method, based on maximum likelihood estimation, which provides an estimate of the true EG using the greedy equivalence search algorithm of Chickering (2002). Through additions and deletions of single edges, GES maximizes a score function in the space of the EGs. An extension of the GES algorithm, named Greedy Interventional Equivalence Search (GIES) algorithm, for structural learning of interventional essential graphs, was then introduced by Hauser & Bühlmann (2012). GES (and also GIES) can be implemented with different optimization criteria. Originally, it was proposed with the Bayesian Information Criterion (Schwarz, 1978) because of consistency. Anyway, any score equivalent and decomposable function can be adopted (Hauser & Bühlmann, 2012). In the following, we implement GIES for three different optimization criteria: the Bayesian Information Criterion and the Extended Bayesian Information Criterion with tuning coefficient $\gamma \in \{0.5, 1\}$ recommended in Foygel & Drton (2010); see also Chen & Chen (2008). We then refer to these three benchmarks as GIES 0, GIES 0.5 and GIES 1 respectively.

For each scenario and method we evaluate the performance in learning the graphical structure of the true DAG in terms of misspecification rate, specificity, sensitivity, precision and Matthews correlation coefficient, defined as

$$\text{MISR} = \tfrac{FN+FP}{q(q-1)}, \quad \text{SPE} = \tfrac{TN}{TN+FP}, \quad \text{SEN} = \tfrac{TP}{TP+FN},$$
$$\text{PRE} = \tfrac{TP}{TP+FP}, \quad \text{MCC} = \tfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives (respectively). The results in the simulation settings with number of nodes $q = 20$, $n^{\emptyset} \in \{100, 200, 500, 1000\}$ and $p \in \{0, 0.2, 0.4\}$ are summarized in Tables 6.3 and 6.4: with the exception of MISR, better performances correspond to higher indicators. We can observe that for all indicators and scenarios, OBIES is better than GIES 0.5, GIES 1 most of the time and it is almost uniformly better than GIES 0.

$n^{\varnothing} = 100$

	$p = 0$				$p = 0.2$				$p = 0.4$			
	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1
MISR	3.18	5.99	3.29	**2.99**	**2.36**	3.94	2.36	2.41	**1.89**	3.11	1.93	1.97
SPE	97.87	95.43	98.00	**98.42**	98.44	97.11	98.61	**98.78**	98.84	97.74	98.99	**99.14**
SEN	**72.79**	64.46	67.44	63.46	**79.36**	73.39	75.80	69.59	**81.10**	78.40	76.69	71.54
PRE	55.91	34.22	55.52	**60.34**	65.53	48.19	66.90	**68.33**	72.39	56.48	73.97	**75.57**
MCC	**63.10**	46.64	60.51	61.17	**71.35**	58.84	70.47	68.19	**76.04**	65.86	74.74	72.91

$n^{\varnothing} = 200$

	$p = 0$				$p = 0.2$				$p = 0.4$			
	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1
MISR	2.87	3.73	**2.65**	2.67	2.14	2.99	2.03	**1.93**	**1.57**	2.18	1.63	1.63
SPE	97.74	97.17	98.07	**98.28**	98.37	97.80	98.59	**98.76**	98.90	98.46	99.01	**99.07**
SEN	**84.67**	77.49	82.09	76.08	**86.49**	79.71	84.17	82.15	**88.07**	83.46	83.58	81.53
PRE	57.95	50.31	61.22	**62.50**	65.96	56.96	68.44	**70.78**	74.53	66.26	75.54	**76.27**
MCC	69.21	61.63	**70.10**	68.22	74.83	66.65	75.22	**75.59**	**80.38**	73.68	78.87	78.30

Table 6.3: Misspecification rate (MISR), specificity (SPE), sensitivity (SEN), precision (PRE) and Matthews correlation coefficient (MCC) for OBIES and GIES, for number of nodes $q = 20$, $n^{\varnothing} \in \{100, 200\}$ and $p \in \{0, 0.2, 0.4\}$.

	$n^{\varnothing} = 500$											
	$p = 0$				$p = 0.2$				$p = 0.4$			
	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1
MISR	2.57	2.75	**2.30**	2.30	**1.57**	2.20	1.69	1.73	**1.05**	1.77	1.15	1.17
SPE	97.74	97.70	98.03	**98.16**	98.66	98.27	98.70	**98.71**	99.12	98.56	99.15	**99.17**
SEN	**93.23**	88.87	92.49	89.13	**94.94**	88.63	91.00	89.54	**96.05**	91.91	92.63	91.74
PRE	60.45	58.84	63.39	**64.15**	**72.62**	65.83	72.11	72.01	80.86	70.79	80.72	**80.94**
MCC	74.12	71.42	**75.73**	74.83	**82.42**	75.68	80.39	79.67	**87.61**	79.93	85.97	85.66

	$n^{\varnothing} = 1000$											
	$p = 0$				$p = 0.2$				$p = 0.4$			
	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1	OBIES	GIES 0	GIES .5	GIES 1
MISR	2.31	3.02	**2.09**	2.14	1.61	1.97	**1.58**	1.60	**0.83**	1.20	0.97	0.97
SPE	97.86	97.46	98.13	**98.13**	98.57	98.37	98.69	**98.71**	**99.27**	99.05	99.23	99.25
SEN	**96.31**	87.69	95.31	93.78	**96.13**	91.57	93.52	92.61	**97.77**	93.82	95.21	95.02
PRE	62.45	56.05	**65.47**	65.14	71.59	67.82	72.55	**72.63**	**83.46**	78.98	82.76	82.89
MCC	76.68	69.21	**78.20**	77.37	**82.26**	78.11	81.76	81.39	**89.89**	85.60	88.33	88.32

Table 6.4: Misspecification rate (MISR), specificity (SPE), sensitivity (SEN), precision (PRE) and Matthews correlation coefficient (MCC) for OBIES and GIES, for number of nodes $q = 20$, $n^{\varnothing} \in \{500, 1000\}$ and $p \in \{0, 0.2, 0.4\}$.

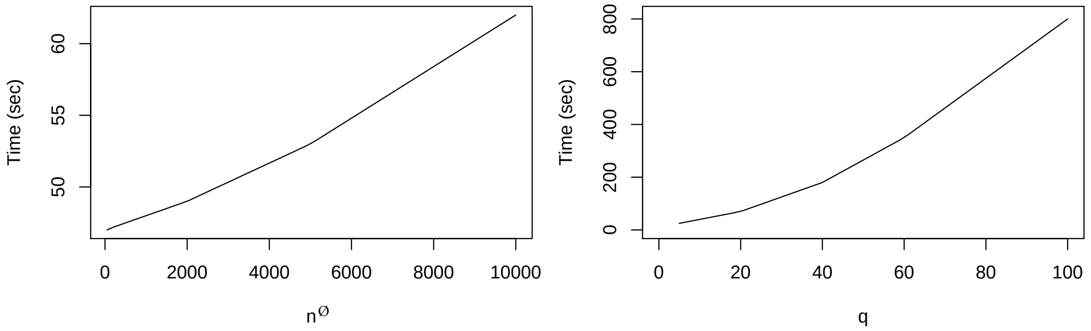


Figure 6.3: Computational time (in seconds) of 1000 MCMC iterations of OBIES, as a function of the sample size $n^{\emptyset}$ for fixed number of nodes $q = 10$ (left panel) and as a function of q for $n^{\emptyset} = 100$ (right panel), averaged over 40 simulated datasets corresponding to the $p = 0.2$ scenario.

We investigate the computational time of the proposed methodology as a function of the number of nodes q and of the sample size $n^{\emptyset}$: in the left panel of Figure 6.3 we report the time in seconds spent by our algorithm to perform 1000 MCMC iterations for $q = 20$, $p = 0.2$ and $n^{\emptyset}$ between 50 and 10000, whilst in the right panel we show the computational time for $n^{\emptyset} = 100$, $p = 0.2$ and q between 5 and 100 (all the results are averaged over the 40 simulated datasets). Algorithms were run on a Intel Core i7-8550U machine. The results suggest that the dependence of time on $n^{\emptyset}$ and q is, respectively, approximately linear and exponential.

6.3 Protein-signaling data

In this section we apply OBIES to the protein-signaling data set of Sachs et al. (2005). Data consist in a collection of observations measured under different experimental conditions and then can be considered as purely interventional. In the original work of Sachs et al. (2005) the objective was to infer a single DAG, whilst Friedman et al. (2008) used the same dataset to learn a single undirected graph. More recently, Peterson et al. (2015) analyzed the dataset from a *multiple graphs* perspective. In particular, they infer an undirected graph for each experimental condition, allowing for the possibility of shared structural features among graphs.

We adopt the methodology developed in the interventional setting of Section 4.2 to perform structural learning of interventional essential graphs. We then compare OBIES with the Greedy Interventional Equivalence Search method.

6.3.1 Data set

The data set, provided as a supplement to Sachs et al. (2005), is based on simultaneous measurements of multiple phosphorylated proteins and phospholipid components in individual primary human immune system cells. Observations are obtained from intracellular multicolor flow cytometry, which allows for simultaneous measurements in individual cells and then turns out in a large number of observations. Flow cytometry can also be used to measure protein modification states which from our perspective are interpreted as interventions on observed variables. Measurements of $q = 11$ phosphorylated proteins and phospholipids are collected after a series of stimulatory cues and inhibitory interventions obtained from the administration of reagents, each one being the responsible of the perturbation of a signaling node. This results in a collection of nine datasets, each containing observations measured under the same experimental condition. In Figure 6.4 we have a signaling network with different points of intervention. Signaling nodes in colour are measured directly, while in grey are not. See Sachs et al. (2005) for further details.

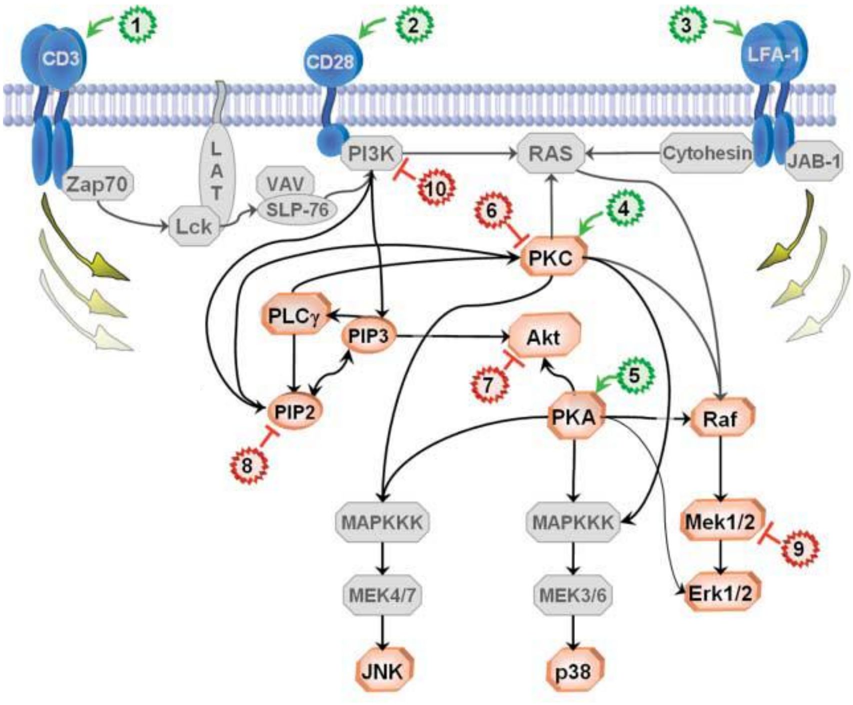


Figure 6.4: From Sachs et al. (2005). A classic signaling network with points of intervention. Signaling nodes in color are measured directly, while signaling nodes in grey are not. The interventions classified as activators are colored green and inhibitors are colored red.

$\boldsymbol{Y}^k$	$\boldsymbol{Y}^1$	$\boldsymbol{Y}^2$	$\boldsymbol{Y}^3$	$\boldsymbol{Y}^4$	$\boldsymbol{Y}^5$	$\boldsymbol{Y}^6$
I_k	Mek	PIP2	Akt	PKA	PKC	PKC
$n^{(k)}$	723	707	913	810	911	799

Table 6.5: Intervention targets and sample sizes for the six datasets included in the study.

Since our method cannot deal with interventions on latent variables, we consider the six datasets associated to interventions on observed variables, that is nodes representing variables which are measured directly. In Table 6.5 we report, for each dataset $\boldsymbol{Y}^k$ included in the study, the corresponding intervention target I_k and sample size. Please observe that we have two different datasets with the same target (PKC). For simplicity of notation variables included in the dataset are also numbered from 1 to 11 according to the original dataset of Sachs et al. (2005).

6.3.2 Model searching and results

We apply OBIES to learn the structure of an $\mathcal{I}$-EG from the dataset $\boldsymbol{Y} = \{\boldsymbol{Y}^1, \dots, \boldsymbol{Y}^6\}$; see Table 6.5. The corresponding family of intervention targets is

$$\mathcal{I} = \{\{2\}, \{4\}, \{7\}, \{8\}, \{9\}, \{9\}\}.$$

Please observe that $\mathcal{I}$ is conservative according to Definition 4.1.2. We explore the $\mathcal{I}$-EG space ${}_{\mathcal{I}}\mathcal{S}_{11}^{r}$ with sparsity parameter $r = 2$, which corresponds to a maximum number of edges of 22; see also Section 5.1.1. For the prior distribution $p(\mathcal{G})$ (Section 5.2.2) we set $a = 1, b = (2q - 2)/3 - 1$. We use an accelerated version of the Markov chain step described in Algorithm 1, by fixing the acceleration parameter $\alpha = 0.5$ (proportion of tested operators to compute $p(\mathcal{G}' \,|\, \mathcal{G})$). We then run $T = 100000$ iterations of Algorithm 4 with a burn-in period of 20000. This results in the MCMC output $\{\mathcal{G}^{(t)}\}$, $t = 20001, \dots, 100000$, that we use to make posterior inference on the $\mathcal{I}$-EG space.

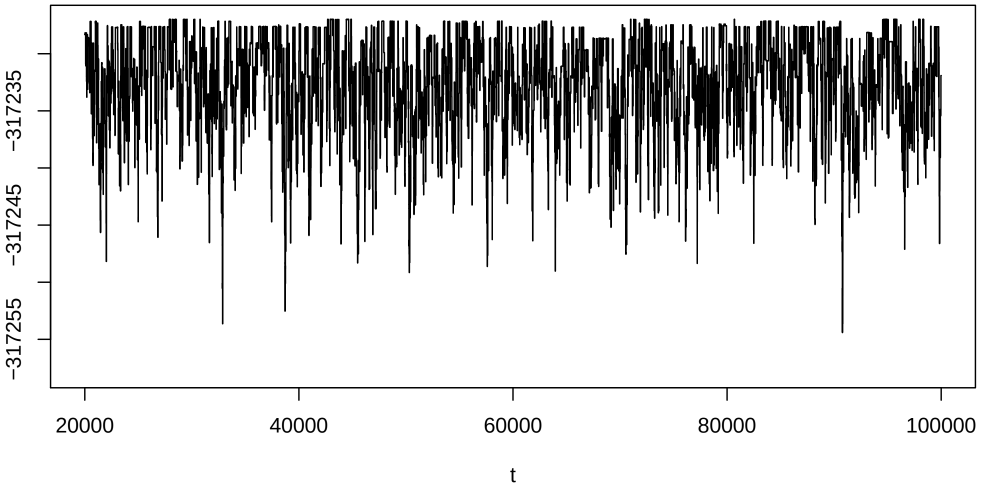


Figure 6.5: MCMC traceplot of visited $\mathcal{I}$-EGs log-scores.

In Figure 6.5 we have the traceplot of the visited $\mathcal{I}$-EGs log-score, which corresponds to the logarithm of $m_{\mathcal{G}^{(t)}}(\boldsymbol{Y})p(\mathcal{G}^{(t)})$ and it is proportional to the posterior probability of $\mathcal{G}^{(t)}$, $p(\mathcal{G}^{(t)} \,|\, \boldsymbol{Y})$. Figure 6.7 contains the heat map with the marginal posterior probabilities of edge inclusion computed according to Formula 6.1 in Section 6.1. In Table 6.6 we also report the marginal posterior probabilities of inclusion of the "top" 14 edges, $p_{u\to v}$, that is those edges that are most often present in the MCMC output. We can use such information to construct the median probability graph model, by including those edges $u \to v$ such that $p_{u\to v}(\boldsymbol{Y}) > 0.5$. Alternatively, we can compute for a grid of thresholds $k \in (0,1)$ the expected false discovery rate $\mathrm{FDR}(k)$ as defined in Equation (6.2) and then choose the maximum value of k such that $\mathrm{FDR}(k) < 0.05$. In doing so, we obtain $k^* = 0.480$. See Figure 6.6.

Hence, we obtain the corresponding $\mathcal{I}$-EG estimate by including all edges such that $p_{u\to v}(\boldsymbol{Y}) > 0.480$. We observe that the resulting graph of Figure 6.8 *is* an $\mathcal{I}$-EG and then no projection to the $\mathcal{I}$-EG space is required; see also Section 6.1.

Finally, we compare OBIES estimate of Figure 6.8 with the Greedy Interven-

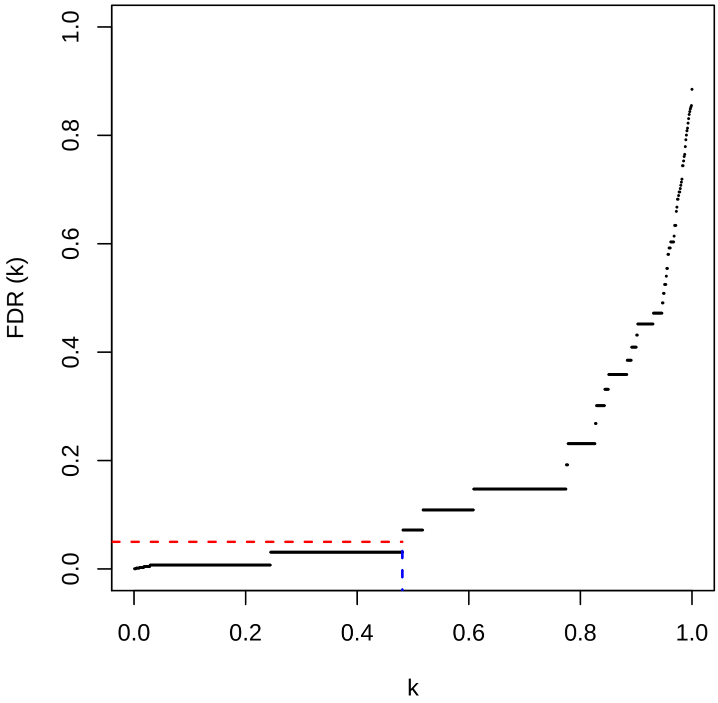


Figure 6.6: Expected false discovery rate FDR(k) as a function of the threshold k.

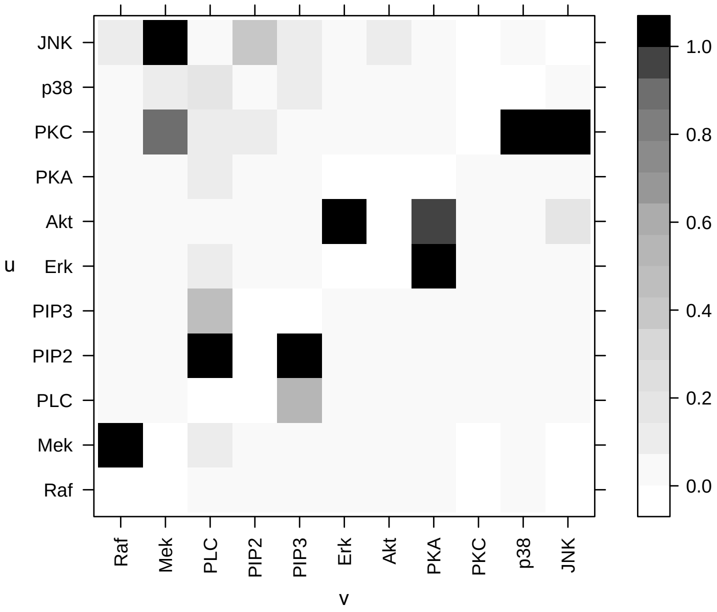


Figure 6.7: Heat map with marginal posterior probabilities of edge inclusion $p_{u \to v}$ under OBIES.

u	$\rightarrow$	v	$p_{u\rightarrow v}$	u	$\rightarrow$	v	$p_{u\rightarrow v}$
Mek	→	Raf	1.000	JNK	→	Mek	1.000
PIP2	→	PLC	1.000	Akt	→	PKA	0.993
PIP2	→	PIP3	1.000	PKC	→	Mek	0.906
Erk	→	PKA	1.000	PLC	→	PIP3	0.519
Akt	→	Erk	1.000	PIP3	→	PLC	0.481
PKC	→	p38	1.000	JNK	→	PIP2	0.376
PKC	→	JNK	1.000	Akt	→	JNK	0.188

Table 6.6: Marginal posterior probabilites of inclusion for the top 14 edges.

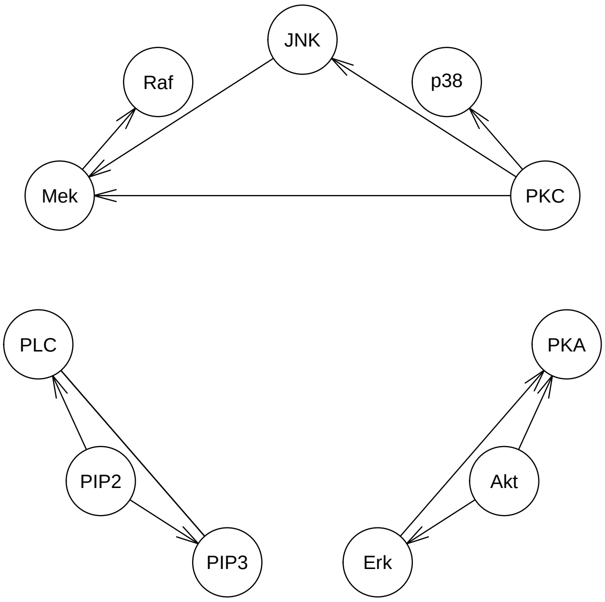


Figure 6.8: OBIES estimate obtained from the FDR criterion.

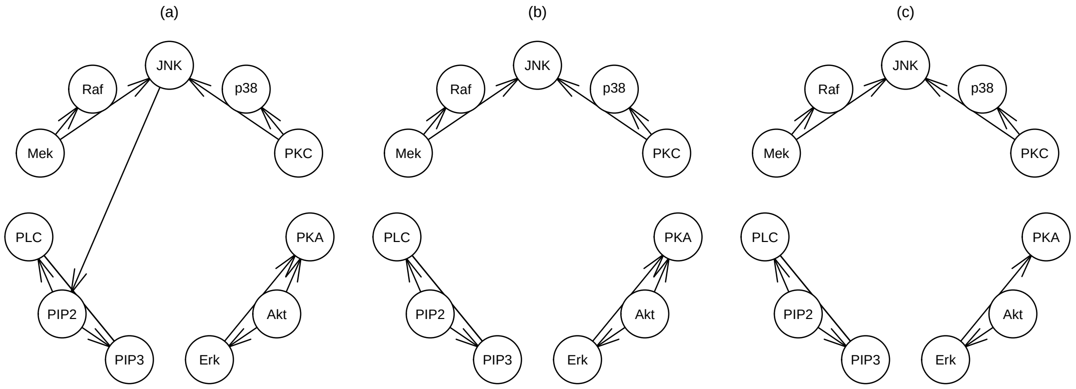


Figure 6.9: Estimated $\mathcal{I}$-EG under GIES 0 (a), GIES 0.5 (b) and GIES 1 (c).

tional Equivalence Search method (Hauser & Bühlmann, 2012). GIES is again computed for three different optimization criteria: the Bayesian Information Criterion (GIES 0) and the Extended Bayesian Information Criterion with tuning coefficient $\gamma \in \{0.5, 1\}$ (GIES 0.5 and GIES 1 respectively); see also Section 6.2.5. Results with $\mathcal{I}$-EG estimates are reported in Figure 6.9. As we can see, the tuning parameter can be used to intensify sparsity of the resulting graph. If compared with the GIES 0.5, our OBIES estimate of Figure 6.8 exhibits 9 edges in common. Instead, it appears that edge JNK → Mek is reversed in GIES 0.5, while PKC → Mek is not present.

Chapter 7

Conclusions and Further Work

Observational data cannot distinguish among Directed Acyclic Graphs (DAGs) encoding the same set of conditional independencies, that is among Markov equivalent DAGs. Each Markov equivalence class is represented by a special chain graph, known as Completed Partially Directed Graph (CPDAG) or Essential Graph (EG). Inteventional data from exogenous perturbations of variables or randomized intervention experiments lead to a finer partition of DAGs into equivalence classes, each one represented by an Interventional Essential Graph (I-EG), thus improving the identifiability of the true DAG generating model.

In this thesis we presented an objective Bayes approach based on the notion of fractional Bayes factor for model selection of Gaussian graphical models in the presence of both observational and interventional data. In addition, we proposed an MCMC sampler to explore the I-EG space under sparsity constraints and learning Markov equivalence classes of DAGs. We applied the proposed methodology, named OBIES, to simulation settings and to the analysis of protein-signaling data (Sachs et al., 2005), providing comparisons with the Greedy Interventional Equivalence Search (GIES) algorithm of Hauser & Bühlmann (2012).

We illustrated through simulations that OBIES is highly competitive with the Extended BIC GES, here considered for two different optimization criterion, and it outperforms BIC GES in producing a point estimate of the underlying I-EG. On the other hand, being fully Bayesian, our method yields a posterior distribution on the I-EG space. Accordingly, it can provide not only single estimate of the I-EG, but also an uncertainty evaluation of other features of interest, such as probabilities of edge inclusion. Finally, being objective, it is virtually free from prior specifications.

Randomized intervention experiments can be used to improve the identifiability of the true data generating model (He & Geng, 2008). By extending the family of intervention targets, (at least in principle) one can reduce each (observational) Markov equivalence class to a single DAG, because all edges that were undirected in the original essential graphs become directed. Ideally, one can identify the true underlying data generating model (rather than a *large* equivalence class) by means of a small number of interventions selected according to an optimal experimental plane. The analysis of such problem from a design of experiments perspective is currently under investigation.

The protein-signaling dataset was collected under distinct experimental conditions. From the theory presented in Sachs et al. (2005) each experimental condition can be interpreted as an intervention on *some* variables, both observed or latent. We analyse jointly all the datasets corresponding to interventions on one of the observed variables to infer a unique graphical structure (I-EG). Another possibility is to analyse all the datasets jointly in order to exploit potential shared features among single graphs (one for each dataset) from a *multiple-graphs* perspective. Joint structural learning for multiple Gaussian undirected graphical models is carried out in Peterson et al. (2015) through a suitable Markov random field prior which encourages common edges, as well as a spike-and-slab prior on the parameters that measure network relatedness. More recently, Tan et al. (2016) apply multiple Gaussian graphical models based on G-Wishart priors to metabolik association networks, using a logistic regression structure to link probability of edge inclusions among graphs.

Appendix A

Some multivariate distributions

In the following we briefly resume some theory about the multivariate random variables involved in this work. For further details see for instance Gelman et al. (2004), Lauritzen (1996) or Geisser & Cornfield (1963). As a matter of notation, we use bold characters for both vectors and matrices; however, just to avoid confusion, we use standard capital letters for random vectors while bold capital letters for the corresponding matrices of observations.

A.1 Multivariate Normal distribution

Let Y be a q-dimensional random vector. We say that $Y = (Y_1, \dots Y_q)^\top$ has a *multivariate Normal distribution* conditionally on the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$,

$$Y \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is a q-dimensional vector and $\boldsymbol{\Sigma}$ a $q \times q$ symmetric and positive definite matrix, if its probability density function is given by

$$f(\boldsymbol{y} \,|\, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right\}.$$

Equivalently, we can express $f(\boldsymbol{y} \,|\, \boldsymbol{\Sigma})$ in terms of the *precision matrix* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$:

$$Y \,|\, \boldsymbol{\Omega}, \boldsymbol{\mu} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}),$$

$$f(\boldsymbol{y}\,|\,\boldsymbol{\Omega}) = (2\pi)^{-\frac{q}{2}}|\boldsymbol{\Omega}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}(\boldsymbol{y}-\boldsymbol{\mu})\right\}.$$

Let now $A \subset \{1,\dots,q\}$, $\bar{A} = \{1,\dots,q\}\setminus A$ and $Y_A, Y_{\bar{A}}$ the corresponding subvectors of Y with components indexed by A and $\bar{A}$ respectively. Accordingly, we partition $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ as

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_A\\ \boldsymbol{\mu}_{\bar{A}}\end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{A\bar{A}}\\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{\bar{A}\bar{A}}\end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix}\boldsymbol{\Omega}_{AA} & \boldsymbol{\Omega}_{A\bar{A}}\\ \boldsymbol{\Omega}_{\bar{A}A} & \boldsymbol{\Omega}_{\bar{A}\bar{A}}\end{bmatrix},$$

where $\boldsymbol{\Sigma}_{A\bar{A}} = \boldsymbol{\Sigma}^\top_{\bar{A}A}$ and $\boldsymbol{\Omega}_{A\bar{A}} = \boldsymbol{\Omega}^\top_{\bar{A}A}$. For the marginal and conditional distributions of Y_A we have the two following results:

$$\begin{aligned} Y_A\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma} &\sim \mathcal{N}_{|A|}\big(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA}\big),\\ Y_A\,|\,Y_{\bar{A}};\boldsymbol{\mu},\boldsymbol{\Sigma} &\sim \mathcal{N}_{|A|}\big(\boldsymbol{\mu}_{A|\bar{A}}, \boldsymbol{\Sigma}_{A|\bar{A}}\big),\end{aligned}$$

where

$$\boldsymbol{\mu}_{A|\bar{A}} = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A\bar{A}}\boldsymbol{\Sigma}^{-1}_{\bar{A}\bar{A}}\big(\boldsymbol{Y}_{\bar{A}} - \boldsymbol{\mu}_{\bar{A}}\big), \quad \boldsymbol{\Sigma}_{A|\bar{A}} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{A\bar{A}}\boldsymbol{\Sigma}^{-1}_{\bar{A}\bar{A}}\boldsymbol{\Sigma}_{\bar{A}A};$$

see Lauritzen (1996, p. 254). Moreover, recall that

$$\begin{aligned}\boldsymbol{\Sigma}_{A|\bar{A}} = \boldsymbol{\Omega}^{-1}_{AA} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{A\bar{A}}\boldsymbol{\Sigma}^{-1}_{\bar{A}\bar{A}}\boldsymbol{\Sigma}_{\bar{A}A} = \boldsymbol{\Omega}^{-1}_{AA},\\ \boldsymbol{\Omega}_{A|\bar{A}} = \boldsymbol{\Sigma}^{-1}_{AA} = \boldsymbol{\Omega}_{AA} - \boldsymbol{\Omega}_{A\bar{A}}\boldsymbol{\Omega}^{-1}_{\bar{A}\bar{A}}\boldsymbol{\Omega}_{\bar{A}A} = \boldsymbol{\Sigma}^{-1}_{AA}.\end{aligned}$$

Therefore, $\boldsymbol{\Omega}_{AA}$ corresponds to the precision matrix of the conditional distribution $Y_A\,|\,Y_{\bar{A}}$. It follows that, for each pair of disjoint subsets $I, J \subset \{1,\dots,q\}$, $\boldsymbol{\Omega}_{IJ} = \boldsymbol{0}$ implies a *conditional* independence between Y_I and Y_J, given the remaining variables, while $\boldsymbol{\Sigma}_{IJ} = \boldsymbol{0}$ corresponds to a *marginal* independence between Y_I and Y_J.

A.2 Matrix Normal distribution

Let $\boldsymbol{Y}$ be a $n\times q$ random matrix. We say that $\boldsymbol{Y}$ has a *matrix Normal distribution* with mean matrix $\boldsymbol{M}$, row covariance matrix $\boldsymbol{\Phi}$ and column covariance matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{Y}\,|\,\boldsymbol{M},\boldsymbol{\Phi},\boldsymbol{\Sigma} \sim \mathcal{N}_{n,q}(\boldsymbol{M},\boldsymbol{\Phi},\boldsymbol{\Sigma}),$$

if its probability density function is given by

$$\begin{aligned} f(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}) &= (2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Phi}|^{-\frac{q}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \\ &\cdot \exp\left\{ -\frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^\top \boldsymbol{\Phi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\big)\right\}, \end{aligned}$$

where $\boldsymbol{M}$ is a $n \times q$ matrix, $\boldsymbol{\Phi}$ a $n \times n$ *symmetric and positive definite* (s.p.d.) matrix and $\boldsymbol{\Sigma}$ a $q \times q$ s.p.d. matrix. Please observe that in such case we do not make distinction between capital and lowercase letters to specify the random variable $\boldsymbol{Y}$ and its realization (both are of course matrices). Equivalently, in terms of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ (the *column precision matrix*) and $\boldsymbol{K} = \boldsymbol{\Phi}^{-1}$ (the *row precision matrix*) we can write

$$\begin{aligned} f(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{K}, \boldsymbol{\Omega}) &= (2\pi)^{-\frac{nq}{2}} |\boldsymbol{K}|^{\frac{q}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \\ &\cdot \exp\left\{ -\frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Omega}(\boldsymbol{Y}-\boldsymbol{M})^\top \boldsymbol{K}(\boldsymbol{Y}-\boldsymbol{M})\big)\right\}. \end{aligned}$$

Let now $A \subset \{1, \dots, q\}$ and $\boldsymbol{Y}_A$ the corresponding $n \times |A|$ submatrix of Y containing columns indexed by A in $\boldsymbol{Y}$. Similarly for $\bar{A} = \{1, \dots, q\} \setminus A$ and $\boldsymbol{Y}_{\bar{A}}$. If $\boldsymbol{M}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are partitioned accordingly,

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_A & \boldsymbol{M}_{\bar{A}} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{A\bar{A}} \\ \boldsymbol{\Sigma}_{\bar{A}A} & \boldsymbol{\Sigma}_{\bar{A}\bar{A}} \end{bmatrix}, \quad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{AA} & \boldsymbol{\Omega}_{A\bar{A}} \\ \boldsymbol{\Omega}_{\bar{A}A} & \boldsymbol{\Omega}_{\bar{A}\bar{A}} \end{bmatrix},$$

we can write for the marginal and conditional distributions of $\boldsymbol{Y}_A$

$$\begin{aligned} \boldsymbol{Y}_A \mid \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma} &\sim \mathcal{N}_{n,|A|}\big(\boldsymbol{M}_A, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{AA}\big), \\ \boldsymbol{Y}_A \mid \boldsymbol{Y}_{\bar{A}}; \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Sigma} &\sim \mathcal{N}_{n,|A|}\big(\boldsymbol{M}_{A|\bar{A}}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{A|\bar{A}}\big), \end{aligned}$$

where

$$\boldsymbol{M}_{A\,|\,\bar{A}} = \boldsymbol{M}_A - \big(\boldsymbol{Y}_{\bar{A}} - \boldsymbol{M}_{\bar{A}}\big)\boldsymbol{\Omega}_{\bar{A}A}\boldsymbol{\Omega}_{\bar{A}\bar{A}}^{-1}, \quad \boldsymbol{\Sigma}_{A\,|\,\bar{A}} = \boldsymbol{\Omega}_{AA}^{-1}.$$

A.3 Wishart distribution

Let $\boldsymbol{\Omega}$ be a $q \times q$ s.p.d. matrix. We say that $\boldsymbol{\Omega}$ has a *Wishart distribution* with parameters $a \in \mathbb{R}$ $(a > q - 1)$ and $\boldsymbol{R}$ a $q \times q$ s.p.d. matrix,

$$\boldsymbol{\Omega} \mid a, \boldsymbol{R} \sim \mathcal{W}_q(a, \boldsymbol{R}),$$

if its probability density function is given by

$$\begin{aligned} f(\boldsymbol{\Omega} \,|\, a, \boldsymbol{R}) &= 2^{-\frac{aq}{2}} \left[\Gamma_q\left(\frac{a}{2}\right)\right]^{-1} |\boldsymbol{R}|^{\frac{a}{2}} |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{R}\boldsymbol{\Omega}\big)\right\} \\ &\propto |\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{R}\boldsymbol{\Omega}\big)\right\}. \end{aligned}$$

A useful result is then contained in the following theorem.

Theorem A.3.1. *Let $\boldsymbol{\Omega} \,|\, a, \boldsymbol{R} \sim \mathcal{W}_q(a, \boldsymbol{R})$, with $\boldsymbol{R}$ a s.p.d. matrix and $a > q-1$. Given $A \subset \{1, \ldots, q\}, \bar{A} = \{1, \ldots, q\} \setminus A$ and the corresponding partitions of $\boldsymbol{\Omega}$ and $\boldsymbol{R}$,*

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{AA} & \boldsymbol{\Omega}_{A\bar{A}} \\ \boldsymbol{\Omega}_{\bar{A}A} & \boldsymbol{\Omega}_{\bar{A}\bar{A}} \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{AA} & \boldsymbol{R}_{A\bar{A}} \\ \boldsymbol{R}_{\bar{A}A} & \boldsymbol{R}_{\bar{A}\bar{A}} \end{bmatrix},$$

we have that

$$\boldsymbol{\Omega}_{A|\bar{A}} \sim \mathcal{W}_{|A|}(a - |\bar{A}|, \boldsymbol{R}_{AA}).$$

See also Lauritzen (1996, p. 261) for further properties of the Wishart distribution.

Appendix B

Marginal distribution of Gaussian data: conjugate analysis

In the following we report some results about the marginal distribution of Gaussian data under three different settings: Gaussian data with zero-expectation, Gaussian data with non-zero mean and the Gaussian multivariate linear regression model. In all cases we assume standard (informative) conjugate priors based on Wishart and Normal-Wishart distributions. Given the $n \times q$ matrix of observations $\boldsymbol{Y}$, we then obtain a formula to compute the marginal data distribution of $\boldsymbol{Y}_A$, the $n \times |A|$ submatrix of $\boldsymbol{Y}$ containing columns indexed by $A \subset \{1, \dots, q\}$. For details see for instance Gelman et al. (2004) and Geisser & Cornfield (1963).

B.1 Gaussian data with zero mean

Consider n q-variate observations $\boldsymbol{y}_i = (y_{i,1}, \dots, y_{i,q})^\top$, $i = 1, \dots, n$, from $(Y_1, \dots, Y_q) \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$ collected in the $n \times q$ matrix $\boldsymbol{Y}$. The likelihood function can be written as

$$\begin{aligned} f(\boldsymbol{Y}\,|\,\boldsymbol{\Omega}) &= \prod_{i=1}^{n} p(\boldsymbol{y}_i\,|\,\boldsymbol{\Omega}) \\ &= \prod_{i=1}^{n} (2\pi)^{-\frac{q}{2}}|\boldsymbol{\Omega}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{y}_i^{\top}\boldsymbol{\Omega}\boldsymbol{y}_i\right\} \\ &= (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Omega})\right\}, \end{aligned}$$

with $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{y}_i\boldsymbol{y}_i^{\top}$.

A standard conjugate prior for $\boldsymbol{\Omega}$ is $\boldsymbol{\Omega} \sim \mathcal{W}_q(a, \boldsymbol{R})$,

$$p(\boldsymbol{\Omega}) = c(a,q)\cdot|\boldsymbol{R}|^{\frac{a}{2}}\cdot|\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{R}\boldsymbol{\Omega}\big)\right\},$$

where

$$c(a,q) = 2^{-\frac{aq}{2}}\left[\Gamma_q\Big(\frac{a}{2}\Big)\right]^{-1}.$$

The prior normalizing constant is then

$$c(a,q)\cdot|\boldsymbol{R}|^{\frac{a}{2}}.$$

B.1.1 Posterior

The posterior distribution of $\boldsymbol{\Omega}$ is obtained as

$$\begin{aligned} p(\boldsymbol{\Omega}\,|\,\boldsymbol{Y}) &\propto p(\boldsymbol{Y}\,|\,\boldsymbol{\Omega})\cdot p(\boldsymbol{\Omega}) \\ &\propto |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Omega})\right\}\cdot|\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{R}\boldsymbol{\Omega}\big)\right\} \\ &= |\boldsymbol{\Omega}|^{\frac{a+n-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\Big[\big(\boldsymbol{R}+\boldsymbol{S}\big)\boldsymbol{\Omega}\Big]\right\}, \end{aligned}$$

from which we get

$$\boldsymbol{\Omega}\,|\,\boldsymbol{Y} \sim \mathcal{W}_q\big(a+n, \boldsymbol{R}+\boldsymbol{S}\big).$$

The posterior normalizing constant is then

$$c(a+n,q)\cdot|\boldsymbol{R}+\boldsymbol{S}|^{\frac{a+n}{2}},$$

$$c(a+n,q) = 2^{-\frac{(a+n)q}{2}}\left[\Gamma_q\Big(\frac{a+n}{2}\Big)\right]^{-1}.$$

B.1.2 Marginal likelihood

The marginal likelihood for the Gaussian model $\mathcal{N}_q(\mathbf{0}, \mathbf{\Omega}^{-1})$ given the data $\boldsymbol{Y}$ can be obtained, up to the constant term $(2\pi)^{-\frac{nq}{2}}$, as the ratio of prior and posterior normalizing constants,

$$\begin{aligned} m(\boldsymbol{Y}) &= (2\pi)^{-\frac{nq}{2}} \cdot \frac{c(a,q)}{c(a+n,q)} \cdot \frac{|\boldsymbol{R}|^{\frac{a}{2}}}{|\boldsymbol{R}+\boldsymbol{S}|^{\frac{a+n}{2}}} \\ &= (\pi)^{-\frac{nq}{2}} \cdot \frac{\Gamma_q\left(\frac{a+n}{2}\right)}{\Gamma_q\left(\frac{a}{2}\right)} \cdot \frac{|\boldsymbol{R}|^{\frac{a}{2}}}{|\boldsymbol{R}+\boldsymbol{S}|^{\frac{a+n}{2}}}. \end{aligned} \tag{B.1}$$

Given $A \subset \{1, \ldots, q\}$ and $\boldsymbol{Y}_A$ the corresponding $n \times |A|$ submatrix of $\boldsymbol{Y}$ we are also interested in computing $m(\boldsymbol{Y}_A)$, the marginal data distribution of $\boldsymbol{Y}_A$. From the theory resumed in Paragraph A.1 we know that

$$Y_A \,|\, \boldsymbol{\mu}, \mathbf{\Sigma} \sim \mathcal{N}_{|A|}\left(\boldsymbol{\mu}_A, \mathbf{\Omega}^{-1}_{A|\bar{A}}\right).$$

Moreover, from Theorem A.3.1, we have

$$\mathbf{\Omega}_{A|B} \sim \mathcal{W}_{|A|}(a - |B|, \boldsymbol{R}_{AA}),$$

where $\boldsymbol{R}_{AA}$ is the $|A| \times |A|$ submatrix of $\boldsymbol{R}$ with rows and columns indexed by A and $|\bar{A}| = q - |A|$. Hence,

$$\begin{aligned} m(\boldsymbol{Y}_A) &= (2\pi)^{-\frac{n|A|}{2}} \cdot \frac{c(a-|\bar{A}|, |A|)}{c(a-|\bar{A}|+n, |A|)} \cdot \frac{|\boldsymbol{R}_{AA}|^{\frac{a-|\bar{A}|}{2}}}{|\boldsymbol{R}_{AA}+\boldsymbol{S}_{AA}|^{\frac{a-|\bar{A}|+n}{2}}} \\ &= (\pi)^{-\frac{n|A|}{2}} \cdot \frac{\Gamma_{|A|}\left(\frac{a-|\bar{A}|+n}{2}\right)}{\Gamma_{|A|}\left(\frac{a-|\bar{A}|}{2}\right)} \cdot \frac{|\boldsymbol{R}_{AA}|^{\frac{a-|\bar{A}|}{2}}}{|\boldsymbol{R}_{AA}+\boldsymbol{S}_{AA}|^{\frac{a-|\bar{A}|+n}{2}}}, \end{aligned} \tag{B.2}$$

where $\boldsymbol{S}_{AA}$ is the submatrix of $\boldsymbol{S}$ with rows and columns indexed by A in $\boldsymbol{S}$.

B.2 Gaussian data with non-zero mean

Assuming $\boldsymbol{\mu}$ and $\mathbf{\Omega}$ unknown, we consider n q-variate observations $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,q})$, $i = 1, \ldots, n$, from $(Y_1, \ldots, Y_q)^\top \,|\, \boldsymbol{\mu}, \mathbf{\Omega} \sim \mathcal{N}_q(\boldsymbol{\mu}, \mathbf{\Omega}^{-1})$ collected in the $n \times q$ matrix

$\boldsymbol{Y}$. The likelihood function is then

$$\begin{aligned} f(\boldsymbol{Y}\,|\,\boldsymbol{\mu},\boldsymbol{\Omega}) &= \prod_{i=1}^{n} p(\boldsymbol{y}_i\,|\,\boldsymbol{\Omega}) \\ &\propto \prod_{i=1}^{n} |\boldsymbol{\Omega}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{\mu})^\top\boldsymbol{\Omega}(\boldsymbol{y}_i-\boldsymbol{\mu})\right\} \\ &= (2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{S}_\mu)\right\}, \end{aligned}$$

being

$$\boldsymbol{S}_\mu = \sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^\top.$$

Standard conjugate priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are

$$\begin{aligned} \boldsymbol{\mu}\,|\,\boldsymbol{\Omega} &\sim \mathcal{N}_q\big(\boldsymbol{m}_0,(a_\mu\boldsymbol{\Omega})^{-1}\big), \\ \boldsymbol{\Omega} &\sim \mathcal{W}_q\big(a_\Omega,\boldsymbol{R}\big) \end{aligned}$$

and then

$$\begin{aligned} p(\boldsymbol{\mu},\boldsymbol{\Omega}) &= c(a_\mu,a_\Omega,q)\cdot|\boldsymbol{\Omega}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}a_\mu(\boldsymbol{\mu}-\boldsymbol{m}_0)^\top\boldsymbol{\Omega}(\boldsymbol{\mu}-\boldsymbol{m}_0)\right\} \\ &\cdot\quad |\boldsymbol{R}|^{\frac{a_\Omega}{2}}|\boldsymbol{\Omega}|^{\frac{a-q-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{R}\boldsymbol{\Omega}\big)\right\}, \end{aligned}$$

where

$$c(a_\mu,a_\Omega,q) = (2\pi)^{-\frac{q}{2}}a_\mu^{\frac{q}{2}}\cdot 2^{-\frac{a_\Omega q}{2}}\left[\Gamma_q\Big(\frac{a_\Omega}{2}\Big)\right]^{-1}.$$

The prior normalizing constant is

$$c(a_\mu,a_\Omega,q)\cdot|\boldsymbol{R}|^{\frac{a_\Omega}{2}}.$$

B.2.1 Posterior

The posterior distribution of $(\boldsymbol{\mu},\boldsymbol{\Omega})$ is

$$\begin{aligned} \boldsymbol{\mu}\,|\,\boldsymbol{\Omega},\boldsymbol{Y} &\sim \mathcal{N}_q\big(\boldsymbol{\mu}_n,\big[(a_\mu+n)\boldsymbol{\Omega})\big]^{-1}\big), \\ \boldsymbol{\Omega}\,|\,\boldsymbol{Y} &\sim \mathcal{W}_q\Big(a_\Omega+n,\boldsymbol{R}+\boldsymbol{S}+\frac{a_\mu n}{a_\mu+n}\boldsymbol{S}_0\Big), \end{aligned}$$

where

$$\boldsymbol{\mu}_n = \frac{a_\mu}{a_\mu + n}\boldsymbol{m}_0 + \frac{n}{a_\mu + n}\bar{\boldsymbol{y}},$$

$$\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\top, \quad \boldsymbol{S}_0 = (\bar{\boldsymbol{y}} - \boldsymbol{m}_0)(\bar{\boldsymbol{y}} - \boldsymbol{m}_0)^\top$$

and $\bar{\boldsymbol{y}}$ the $q \times 1$ vector of sample means of $Y_1, \ldots, Y_q$; see for instance Gelman et al. (2004). The posterior normalizing constant is then

$$c(a_\mu + n, a_\Omega + n, q) \cdot \left|\boldsymbol{R} + \boldsymbol{S} + \frac{a_\mu n}{a_\mu + n}\boldsymbol{S}_0\right|^{\frac{a_\Omega + n}{2}},$$

$$c(a_\mu + n, a_\Omega + n, q) = (2\pi)^{-\frac{q}{2}}(a_\mu + n)^{\frac{q}{2}} \cdot 2^{-\frac{(a_\Omega + n)q}{2}}\left[\Gamma_q\left(\frac{a_\Omega + n}{2}\right)\right]^{-1}.$$

B.2.2 Marginal likelihood

The marginal likelihood for the Gaussian model $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$ given the data $\boldsymbol{Y}$ is obtained as

$$\begin{aligned} m(\boldsymbol{Y}) &= (2\pi)^{-\frac{nq}{2}} \cdot \frac{c(a_\mu, a_\Omega, q)}{c(a_\mu + n, a_\Omega + n, q)} \cdot \frac{|\boldsymbol{R}|^{\frac{a_\Omega}{2}}}{\left|\boldsymbol{R} + \boldsymbol{S} + \frac{a_\mu n}{a_\mu + n}\boldsymbol{S}_0\right|^{\frac{a_\Omega + n}{2}}} \\ &= (\pi)^{-\frac{nq}{2}} \cdot \left(\frac{a_\mu}{a_\mu + n}\right)^{\frac{q}{2}} \cdot \frac{\Gamma_q\left(\frac{a_\Omega + n}{2}\right)}{\Gamma_q\left(\frac{a_\Omega}{2}\right)} \cdot \frac{|\boldsymbol{R}|^{\frac{a_\Omega}{2}}}{\left|\boldsymbol{R} + \boldsymbol{S} + \frac{a_\mu n}{a_\mu + n}\boldsymbol{S}_0\right|^{\frac{a_\Omega + n}{2}}}. \end{aligned} \tag{B.3}$$

Let now $A \subset \{1, \ldots, q\}$ and $\boldsymbol{Y}_A$ the corresponding $n \times |A|$ submatrix of $\boldsymbol{Y}$. From Paragraph A.1 we know that

$$\begin{aligned} Y_A \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim \mathcal{N}_{|A|}\big(\boldsymbol{\mu}_A, \boldsymbol{\Omega}_{A|\bar{A}}^{-1}\big), \\ \boldsymbol{\mu}_A \,|\, \boldsymbol{\Omega} &\sim \mathcal{N}_{|A|}\big(\boldsymbol{m}_{0A}, (a_\mu \boldsymbol{\Omega}_{A|\bar{A}})^{-1}\big). \end{aligned}$$

Moreover, from Theorem A.3.1 we have

$$\boldsymbol{\Omega}_{A|B} \sim \mathcal{W}_{|A|}(a - |B|, \boldsymbol{R}_{AA}),$$

where $\boldsymbol{R}_{AA}$ is the $|A| \times |A|$ submatrix of $\boldsymbol{R}$ with rows and columns indexed by A and $|\bar{A}| = q - |A|$. Hence,

$$\begin{aligned} m(\boldsymbol{Y}_A) &= (2\pi)^{-\frac{n|A|}{2}} \cdot \frac{c(a_\mu, a_\Omega - |\bar{A}|, |A|)}{c(a_\mu, a_\Omega - |\bar{A}| + n, |A|)} \cdot \frac{|\boldsymbol{R}_{AA}|^{\frac{a_\Omega - |\bar{A}|}{2}}}{\left|\boldsymbol{R}_{AA} + \boldsymbol{S}_{AA} + \frac{a_\mu n}{a_\mu + n}\boldsymbol{S}_{0,AA}\right|^{\frac{a_\Omega - |\bar{A}| + n}{2}}} \\ &= (\pi)^{-\frac{n|A|}{2}} \cdot \left(\frac{a_\mu}{a_\mu + n}\right)^{\frac{|A|}{2}} \cdot \frac{\Gamma_{|A|}\left(\frac{a_\Omega - |\bar{A}| + n}{2}\right)}{\Gamma_{|A|}\left(\frac{a_\Omega - |\bar{A}|}{2}\right)} \cdot \frac{|\boldsymbol{R}_{AA}|^{\frac{a_\Omega - |\bar{A}|}{2}}}{\left|\boldsymbol{R}_{AA} + \boldsymbol{S}_{AA} + \frac{a_\mu n}{a_\mu + n}\boldsymbol{S}_{0,AA}\right|^{\frac{a_\Omega - |\bar{A}| + n}{2}}} \end{aligned} \tag{B.4}$$

where $\boldsymbol{S}_{AA}$ and $\boldsymbol{S}_{0,AA}$ are the submatrices of $\boldsymbol{S}$ and $\boldsymbol{S}_0$ containing rows and columns indexed by A in $\boldsymbol{S}$ and $\boldsymbol{S}_0$ respectively.

B.3 Multivariate linear regression

Let $\boldsymbol{Y}$ be a $n \times q$ matrix of responses from the q random variables $Y_1, \dots, Y_q$, $\boldsymbol{X}$ a $n \times (p+1)$ matrix of observations from a set of p explanatory variables (including the unit vector for the intercept) and $\boldsymbol{B}$ a $(p+1) \times q$ matrix of coefficients describing the effect of the explanatory variables on the responses. A Gaussian multivariate linear regression model can be written in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E},$$

where $\boldsymbol{E}$ is a $n \times q$ matrix of error terms, $\boldsymbol{E} \sim \mathcal{N}_{n,q}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1})$, with $\boldsymbol{I}_n$ the $n \times n$ identity matrix, $\boldsymbol{\Omega}$ is the s.p.d. (unconstrained) column precision matrix and $\boldsymbol{0}$ the $n \times q$ null matrix. Equivalently we write

$$\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}\big(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}\big)$$

and then

$$f(\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega}) = (2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\big]\right\}.$$

Letting $\hat{\boldsymbol{B}} = \big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$, we can write

$$\begin{aligned} f(\boldsymbol{Y} \,|\, \boldsymbol{B}, \boldsymbol{\Omega}) &= (2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \\ &\cdot \quad \exp\left\{ -\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}\{(\boldsymbol{B} - \hat{\boldsymbol{B}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{B} - \hat{\boldsymbol{B}}) + \hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}\}\big]\right\}, \end{aligned}$$

where $\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}$.

B.3.1 Prior

A conjugate prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$ is given by

$$\begin{aligned}\boldsymbol{B} \,|\, \boldsymbol{\Omega} &\sim \mathcal{N}_{p+1,q}\big(\underline{\boldsymbol{B}}, \boldsymbol{C}^{-1}, \boldsymbol{\Omega}^{-1}\big),\\ \boldsymbol{\Omega} &\sim \mathcal{W}_q\big(a, \boldsymbol{R}\big),\end{aligned}$$

whose probability densities are

$$\begin{aligned}p(\boldsymbol{B} \,|\, \boldsymbol{\Omega}) &= (2\pi)^{-\frac{1}{2}q(p+1)}|\boldsymbol{C}|^{\frac{q}{2}}|\boldsymbol{\Omega}|^{\frac{p+1}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}(\boldsymbol{B}-\underline{\boldsymbol{B}})^\top \boldsymbol{C}(\boldsymbol{B}-\underline{\boldsymbol{B}})\big]\right\}\\ p(\boldsymbol{\Omega}) &= \left[\Gamma_q\Big(\frac{a}{2}\Big)\right]^{-1} 2^{-\frac{aq}{2}}|\boldsymbol{R}|^{\frac{a}{2}}|\boldsymbol{\Omega}|^{\frac{a-q-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{R}\boldsymbol{\Omega}\big)\right\}\end{aligned}$$

The joint prior $p(\boldsymbol{B}, \boldsymbol{\Omega}) \propto p(\boldsymbol{B} \,|\, \boldsymbol{\Omega})p(\boldsymbol{\Omega})$ is then

$$p(\boldsymbol{B}, \boldsymbol{\Omega}) \propto \frac{|\boldsymbol{\Omega}|^{\frac{1}{2}[(p+1)+(a-q-1)]}}{K(\boldsymbol{C}, \boldsymbol{R}, a)} \exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}\big((\boldsymbol{B}-\underline{\boldsymbol{B}})^\top \boldsymbol{C}(\boldsymbol{B}-\underline{\boldsymbol{B}})+\boldsymbol{R}\big)\big]\right\},$$

where

$$K(\boldsymbol{C}, \boldsymbol{R}, a) = \frac{(2\pi)^{\frac{q(p+1)}{2}}\Gamma_q\big(\frac{a}{2}\big)2^{\frac{aq}{2}}}{|\boldsymbol{C}|^{\frac{q}{2}}|\boldsymbol{R}|^{\frac{a}{2}}}$$

is the prior normalizing constant.

B.3.2 Posterior

The posterior distribution of $(\boldsymbol{B}, \boldsymbol{\Omega})$ is

$$\begin{aligned}\boldsymbol{B} \,|\, \boldsymbol{\Omega}, \boldsymbol{Y} &\sim \mathcal{N}_{p+1,q}\big(\bar{\boldsymbol{B}}, (\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \boldsymbol{\Omega}^{-1}\big)\\ \boldsymbol{\Omega} \,|\, \boldsymbol{Y} &\sim \mathcal{W}_q\big(a+n, \boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}\big),\end{aligned}$$

where

$$\begin{aligned}\bar{\boldsymbol{B}} &= (\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y}+\boldsymbol{C}\underline{\boldsymbol{B}})\\ \boldsymbol{D} &= (\underline{\boldsymbol{B}}-\hat{\boldsymbol{B}})^\top\big(\boldsymbol{C}^{-1}+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big)^{-1}(\underline{\boldsymbol{B}}-\hat{\boldsymbol{B}});\end{aligned}$$

see also Geisser (1965). Hence,

$$\begin{aligned}p(\boldsymbol{B} \,|\, \boldsymbol{\Omega}, \boldsymbol{Y}) &= (2\pi)^{\frac{q(p+1)}{2}}|\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}}|\boldsymbol{\Omega}|^{\frac{p+1}{2}}\\ &\quad\cdot\ \exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}\big((\boldsymbol{B}-\bar{\boldsymbol{B}})^\top(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{B}-\bar{\boldsymbol{B}})\big)\big]\right\}\end{aligned}$$

$$p(\boldsymbol{\Omega}\,|\,\boldsymbol{Y}) = \frac{|\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}|^{\frac{a+n}{2}}}{2^{\frac{q(a+n)}{2}}\Gamma_q\big(\frac{a+n}{2}\big)}|\boldsymbol{\Omega}|^{\frac{a+n-q-1}{2}}\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\big(\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}\big)\Big]\Big\}$$

The joint density is then

$$\begin{aligned} p(\boldsymbol{B},\boldsymbol{\Omega}\,|\,\boldsymbol{Y}) &= K^{-1}(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X},\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D},a+n)\\ &\cdot\ |\boldsymbol{\Omega}|^{\frac{p+1}{2}}\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\big((\boldsymbol{B}-\bar{\boldsymbol{B}})^\top(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{B}-\bar{\boldsymbol{B}})\big)\Big]\Big\}\\ &\cdot\ |\boldsymbol{\Omega}|^{\frac{a+n-q-1}{2}}\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\big(\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}\big)\Big]\Big\} \end{aligned}$$

where

$$K(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X},\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D},a+n) = \frac{(2\pi)^{\frac{q(p+1)}{2}}\Gamma_q\big(\frac{a+n}{2}\big)2^{\frac{q(a+n)}{2}}}{|\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}}|\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}|^{\frac{a+n}{2}}}$$

is the posterior normalizing constant.

B.3.3 Marginal likelihood

The marginal data distribution is obtained (up to a constant term) as the ratio between prior and posterior normalizing constants,

$$\begin{aligned} m(\boldsymbol{Y}\,|\,\boldsymbol{X}) &= \frac{(2\pi)^{-\frac{nq}{2}}K^{-1}(\boldsymbol{C},\boldsymbol{R},a)}{K^{-1}(\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X},\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D},a+n)}\\ &= \frac{|\boldsymbol{C}|^{\frac{q}{2}}|\boldsymbol{R}|^{\frac{a}{2}}\Gamma_q\big(\frac{a+n}{2}\big)2^{\frac{qn}{2}}}{(2\pi)^{\frac{nq}{2}}|\boldsymbol{C}+\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}}|\boldsymbol{R}+\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}+\boldsymbol{D}|^{\frac{a+n}{2}}\Gamma_q\big(\frac{a}{2}\big)}. \end{aligned} \tag{B.5}$$

Let now $A\subset\{1,\dots,q\}$, $\boldsymbol{Y}_A$ the corresponding $n\times|A|$ submatrix of $\boldsymbol{Y}$ and $\boldsymbol{B}_A$ the $(p+1)\times|A|$ submatrix of $\boldsymbol{B}$, whose columns contain the regression coefficients for the selected responses in $\boldsymbol{Y}_A$. From the theory resumed in Paragraphs A.2 and A.3, we obtain for the regression model restricted to the subset A,

$$\boldsymbol{Y}_A\,|\,\boldsymbol{B},\boldsymbol{\Omega}\sim\mathcal{N}_{n,|A|}\big(\boldsymbol{X}\boldsymbol{B}_A,\boldsymbol{I}_n,\boldsymbol{\Omega}_{AA}^{-1}\big),$$

with the corresponding priors

$$\begin{aligned} \boldsymbol{B}_A\,|\,\boldsymbol{\Omega} &\sim \mathcal{N}_{p+1,|A|}\big(\underline{\boldsymbol{B}}_A,\boldsymbol{C}^{-1},\boldsymbol{\Omega}_{A|\bar{A}}^{-1}\big),\\ \boldsymbol{\Omega}_{A|\bar{A}} &\sim \mathcal{W}_{|A|}\big(a-|\bar{A}|,\boldsymbol{R}_{AA}\big), \end{aligned}$$

where $\underline{\boldsymbol{B}}_A$ is the submatrix of $\underline{\boldsymbol{B}}$ with columns indexed by $|A|$ and $|\bar{A}| = q - |A|$. The marginal likelihood of $\boldsymbol{Y}_A$ is then obtained by coherently updating the normalizing constants involved in $m(\boldsymbol{Y})$. Hence,

$$\begin{aligned} m(\boldsymbol{Y}_A \,|\, \boldsymbol{X}) &= \frac{(2\pi)^{-\frac{n|A|}{2}} K^{-1}(\boldsymbol{C}, \boldsymbol{R}_{AA}, a - |\bar{A}|)}{K^{-1}(\boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}, \boldsymbol{R}_{AA} + \hat{\boldsymbol{E}}_A^\top \hat{\boldsymbol{E}}_A + \boldsymbol{D}_{AA}, a - |\bar{A}| + n)} \\ &= \frac{|\boldsymbol{C}|^{\frac{|A|}{2}} |\boldsymbol{R}_{AA}|^{\frac{a-|\bar{A}|}{2}} \Gamma_q\big(\frac{a-|\bar{A}|+n}{2}\big) 2^{\frac{|A|n}{2}}}{(2\pi)^{\frac{n|A|}{2}} |\boldsymbol{C} + \boldsymbol{X}^\top \boldsymbol{X}|^{\frac{|A|}{2}} |\boldsymbol{R}_{AA} + \hat{\boldsymbol{E}}_A^\top \hat{\boldsymbol{E}}_A + \boldsymbol{D}_{AA}|^{\frac{a-|\bar{A}|+n}{2}} \Gamma_{|A|}\big(\frac{a-|\bar{A}|}{2}\big)}, \end{aligned} \tag{B.6}$$

where $\hat{\boldsymbol{E}}_A = \boldsymbol{Y}_A - \boldsymbol{X}\hat{\boldsymbol{B}}_A$, while $n, \boldsymbol{C}$ and $\boldsymbol{X}$ are unchanged.

Appendix C

Marginal likelihood of Gaussian models: FBF setting

In the following we adopt an objective Bayes approach based on the Fractional Bayes Factor (FBF) (Section 2.3) to evaluate the marginal likelihood of the Gaussian models presented in Section B. As before, we are interested in computing the marginal likelihood with respect to the $n \times |A|$ matrix $\boldsymbol{Y}_A$ which contains columns indexed by $A \subset \{1, \dots, q\}$ in $\boldsymbol{Y}$. The same results can be found in Consonni & La Rocca (2012) and Consonni et al. (2017) with reference to Gaussian data with zero expectation and multivariate Gaussian linear regression.

C.1 Gaussian data with zero mean

Consider the case in which $(Y_1, \dots, Y_q) \,|\, \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$ (Paragraph B.1). The likelihood function given the $n \times q$ data matrix $\boldsymbol{Y}$ can be written as

$$f(\boldsymbol{Y} \,|\, \boldsymbol{\Omega}) \quad = \quad (2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Omega}) \right\},$$

with $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^\top$.

C.1.1 Fractional prior

We start assuming the default prior for $\boldsymbol{\Omega}$,

$$p^D(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}}.$$

The implied fractional prior (see Section **??**) is then

$$\begin{aligned} p^F(\boldsymbol{\Omega}\,|\,b,\boldsymbol{Y}) &\propto f^b(\boldsymbol{Y}\,|\,\boldsymbol{\Omega})p^D(\boldsymbol{\Omega}) \\ &\propto \left[(2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Omega})\right\}\right]^b|\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}} \end{aligned}$$

Setting $b = n_0/n$ we obtain

$$p^F(\boldsymbol{\Omega}\,|\,b,\boldsymbol{Y}) \propto (2\pi)^{-\frac{n_0q}{2}}|\boldsymbol{\Omega}|^{\frac{a_D+n_0-q-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(n_0\boldsymbol{\Omega}\bar{\boldsymbol{S}})\right\}$$

where $\bar{\boldsymbol{S}} = \frac{1}{n}\boldsymbol{S}$ so that the fractional prior is

$$\boldsymbol{\Omega}\,|\,b,\boldsymbol{Y} \sim \mathcal{W}_q\big(a_D+n_0, n_0\bar{\boldsymbol{S}}\big).$$

The prior normalizing constant (see Section A.3) is

$$c(a_D+n_0,q)\cdot|n_0\bar{\boldsymbol{S}}|^{\frac{a_D+n_0}{2}},$$

$$c(a_D+n_0,q) = 2^{-\frac{(a_D+n_0)q}{2}}\left[\Gamma_q\left(\frac{a_D+n_0}{2}\right)\right]^{-1}.$$

C.1.2 Posterior

The posterior distribution of $\boldsymbol{\Omega}$ is obtained in the FBF setting of Section 2.3 as

$$\begin{aligned} p(\boldsymbol{\Omega}\,|\,b,\boldsymbol{Y}) &\propto f^{1-b}(\boldsymbol{Y}\,|\,\boldsymbol{\Omega})p^F(\boldsymbol{\Omega}\,|\,b,\boldsymbol{Y}) \\ &= \left[(2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Omega})\right\}\right]^{\frac{n-n_0}{n}}\cdot|\boldsymbol{\Omega}|^{\frac{a_D+n_0-q-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(n_0\boldsymbol{\Omega}\bar{\boldsymbol{S}})\right\} \\ &\propto |\boldsymbol{\Omega}|^{\frac{a_D+n-q-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}(\tilde{\boldsymbol{S}}+n_0\bar{\boldsymbol{S}})\big]\right\}, \end{aligned}$$

being

$$\tilde{\boldsymbol{S}} = \frac{n-n_0}{n}\boldsymbol{S} = (n-n_0)\bar{\boldsymbol{S}}.$$

Moreover, $\tilde{\boldsymbol{S}} + n_0\bar{\boldsymbol{S}} = n\bar{\boldsymbol{S}} = \boldsymbol{S}$ so that

$$p(\boldsymbol{\Omega} \,|\, n_0, \boldsymbol{Y}) \propto (2\pi)^{-\frac{(n-n_0)q}{2}} |\boldsymbol{\Omega}|^{\frac{a_D+n-q-1}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Omega}\boldsymbol{S}\big)\right\}.$$

Hence,

$$\boldsymbol{\Omega} \,|\, n_0, \boldsymbol{Y} \sim \mathcal{W}_q\big(a_D + n, \boldsymbol{S}\big).$$

The posterior normalizing constant is

$$c(a_D + n, q) \cdot |\boldsymbol{S}|^{\frac{a_D+n}{2}},$$

$$c(a_D + n, q) \;\;=\;\; 2^{-\frac{(a_D+n)q}{2}} \left[\Gamma_q\Big(\frac{a_D+n}{2}\Big)\right]^{-1}.$$

C.1.3 Marginal likelihood

The fractional marginal likelihood for the Gaussian model $\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$ is obtained in analogy with the standard conjugate case described in Section B.1 as

$$\begin{aligned}
m^F(\boldsymbol{Y}) &= (2\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{c(a_D+n_0,q)}{c(a_D+n,q)} \cdot \frac{|n_0\bar{\boldsymbol{S}}|^{\frac{a_D+n_0}{2}}}{|\boldsymbol{S}|^{\frac{a_D+n}{2}}} \\
&= (\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{\Gamma_q\left(\frac{a_D+n}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0}{2}\right)} \cdot \frac{|n_0\bar{\boldsymbol{S}}|^{\frac{a_D+n_0}{2}}}{|\boldsymbol{S}|^{\frac{a_D+n}{2}}} \\
&= (\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{\Gamma_q\left(\frac{a_D+n}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{q(a_D+n_0)}{2}} \cdot |\boldsymbol{S}|^{-\frac{n-n_0}{2}}. \qquad \text{(C.1)}
\end{aligned}$$

The fractional marginal likelihood with respect to $\boldsymbol{Y}_A$, the submatrix of $\boldsymbol{Y}$ with columns indexed by A is obtained accordingly as

$$\begin{aligned}
m^F(\boldsymbol{Y}_A) &= (2\pi)^{-\frac{(n-n_0)|A|}{2}} \cdot \frac{c(a_D-|\bar{A}|+n_0,|A|)}{c(a_D-|\bar{A}|+n,|A|)} \cdot \frac{|n_0\bar{\boldsymbol{S}}_{AA}|^{\frac{a_D-|\bar{A}|+n_0}{2}}}{|\boldsymbol{S}_{AA}|^{\frac{a_D-|\bar{A}|+n}{2}}} \qquad \text{(C.2)} \\
&= (\pi)^{-\frac{(n-n_0)|A|}{2}} \cdot \frac{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n}{2}\right)}{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n_0}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{|A|(a_D-|\bar{A}|+n_0)}{2}} \cdot |\boldsymbol{S}_{AA}|^{-\frac{n-n_0}{2}},
\end{aligned}$$

where $\boldsymbol{S}_{AA}$ is the submatrix of $\boldsymbol{S}$ with rows and columns indexed by A and $|\bar{A}| = q - |A|$.

C.2 Gaussian data with non-zero mean

Assume now $(Y_1, \dots, Y_q) \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$ as in Section B.2. Given the $n \times q$ data matrix $\boldsymbol{Y}$, the likelihood function can be written as

$$\begin{aligned} f(\boldsymbol{Y} \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega}) &= \prod_{i=1}^{n} p(\boldsymbol{y}_i \,|\, \boldsymbol{\Omega}) \\ &= (2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\text{tr}(\boldsymbol{\Omega}\boldsymbol{S}_\mu) \right\} \\ &= (2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\text{tr}\Big[\boldsymbol{\Omega}\Big\{ n(\boldsymbol{\mu} - \bar{\boldsymbol{y}})^\top (\boldsymbol{\mu} - \bar{\boldsymbol{y}}) + \sum_{i=1}^{n} \boldsymbol{e}_i \boldsymbol{e}_i^\top \Big\}\Big] \right\} \end{aligned}$$

where $\boldsymbol{e}_i = \boldsymbol{y}_i - \bar{\boldsymbol{y}}$; see also Gelman et al. (2004).

C.2.1 Fractional prior

We start assuming the default prior for $(\boldsymbol{\mu}, \boldsymbol{\Omega})$,

$$p^D(\boldsymbol{\mu}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D - q - 1}{2}}.$$

The implied fractional prior is then

$$p^F(\boldsymbol{\Omega} \,|\, b, \boldsymbol{Y}) \;\; \propto \;\; f^b(\boldsymbol{Y} \,|\, \boldsymbol{\mu}, \boldsymbol{\Omega}) p^D(\boldsymbol{\mu}, \boldsymbol{\Omega}).$$

Setting $b = n_0/n$ we obtain

$$\begin{aligned} p^F(\boldsymbol{\mu}, \boldsymbol{\Omega} \,|\, n_0, \boldsymbol{Y}) &\propto |\boldsymbol{\Omega}|^{\frac{n_0}{2}} \exp\left\{ -\frac{n_0}{2}\text{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{\mu} - \bar{\boldsymbol{y}})^\top (\boldsymbol{\mu} - \bar{\boldsymbol{y}}) + \bar{\boldsymbol{R}}\big\}\Big] \right\} \cdot |\boldsymbol{\Omega}|^{\frac{a_D - q - 1}{2}} \\ &= |\boldsymbol{\Omega}|^{\frac{n_0}{2}} \exp\left\{ -\frac{n_0}{2}\text{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{\mu} - \bar{\boldsymbol{y}})^\top (\boldsymbol{\mu} - \bar{\boldsymbol{y}})\big\}\Big] \right\} \\ &\cdot \;\; |\boldsymbol{\Omega}|^{\frac{a_D - q - 1}{2}} \exp\left\{ -\frac{n_0}{2}\text{tr}\big(\boldsymbol{\Omega}\bar{\boldsymbol{R}}\big) \right\}, \end{aligned}$$

where

$$\bar{\boldsymbol{R}} = \frac{1}{n}\boldsymbol{R} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{e}_i \boldsymbol{e}_i^\top.$$

Multiplying and dividing by $|n_0\boldsymbol{\Omega}|^{\frac{1}{2}}$ we obtain

$$\begin{aligned}\boldsymbol{\mu}\,|\,\boldsymbol{\Omega}, n_0, \boldsymbol{Y} &\sim \mathcal{N}_q\big(\bar{\boldsymbol{y}}, (n_0\boldsymbol{\Omega})^{-1}\big),\\ \boldsymbol{\Omega}\,|\,n_0, \boldsymbol{Y} &\sim \mathcal{W}_q\big(a_D + n_0 - 1, n_0\bar{\boldsymbol{R}}\big).\end{aligned}$$

The prior normalizing constant is then

$$c(a_D + n_0 - 1, q) \cdot |n_0\bar{\boldsymbol{R}}|^{\frac{a_D+n_0-1}{2}},$$

$$c(a_D + n_0 - 1, q) \;=\; (2\pi)^{-\frac{q}{2}} n_0^{\frac{q}{2}} \cdot 2^{-\frac{(a_D+n_0-1)q}{2}}\left[\Gamma_q\left(\frac{a_D + n_0 - 1}{2}\right)\right]^{-1}.$$

C.2.2 Posterior

The posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Omega})$ under the FBF setting is obtained as

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\Omega}\,|\,b, \boldsymbol{Y}) &\propto f^{1-b}(\boldsymbol{Y}\,|\,\boldsymbol{\mu}, \boldsymbol{\Omega}) p^F(\boldsymbol{\mu}, \boldsymbol{\Omega}\,|\,b, \boldsymbol{Y})\\ &\propto \left[(2\pi)^{-\frac{nq}{2}}|\boldsymbol{\Omega}|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\Big\{n(\boldsymbol{\mu}-\bar{\boldsymbol{y}})^\top(\boldsymbol{\mu}-\bar{\boldsymbol{y}}) + \boldsymbol{R}\Big\}\Big]\right\}\right]^{1-b}\\ &\cdot\quad |n_0\boldsymbol{\Omega}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\Big[n_0\boldsymbol{\Omega}\Big\{(\boldsymbol{\mu}-\bar{\boldsymbol{y}})^\top(\boldsymbol{\mu}-\bar{\boldsymbol{y}})\Big\}\Big]\right\}\\ &\cdot\quad |\boldsymbol{\Omega}|^{\frac{a_D+n_0-q-2}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\big[n_0\boldsymbol{\Omega}\bar{\boldsymbol{R}}\big]\right\}.\end{aligned}$$

Setting $b = n_0/n$ we obtain

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\Omega}\,|\,n_0, \boldsymbol{Y}) &\propto |\boldsymbol{\Omega}|^{\frac{n-n_0}{2}}\exp\left\{-\frac{n-n_0}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\Big\{(\boldsymbol{\mu}-\bar{\boldsymbol{y}})^\top(\boldsymbol{\mu}-\bar{\boldsymbol{y}}) + \bar{\boldsymbol{R}}\Big\}\Big]\right\}\\ &\cdot\quad |n_0\boldsymbol{\Omega}|^{\frac{1}{2}}\exp\left\{-\frac{n_0}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\Big\{(\boldsymbol{\mu}-\bar{\boldsymbol{y}})^\top(\boldsymbol{\mu}-\bar{\boldsymbol{y}})\Big\}\Big]\right\}\\ &\cdot\quad |\boldsymbol{\Omega}|^{\frac{a_D+n_0-q-2}{2}}\exp\left\{-\frac{n_0}{2}\mathrm{tr}\big[\boldsymbol{\Omega}(\bar{\boldsymbol{R}} + \bar{\boldsymbol{S}})\big]\right\}\end{aligned}$$

and so

$$\begin{aligned}p(\boldsymbol{\mu}, \boldsymbol{\Omega}\,|\,\boldsymbol{Y}, b) &\propto |\boldsymbol{\Omega}|^{\frac{1}{2}}\exp\left\{-\frac{n}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\Big\{(\boldsymbol{\mu}-\bar{\boldsymbol{y}})^\top(\boldsymbol{\mu}-\bar{\boldsymbol{y}})\Big\}\Big]\right\}\\ &\cdot\quad |\boldsymbol{\Omega}|^{\frac{a_D+n-q-2}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\big[\boldsymbol{\Omega}\boldsymbol{R}\big]\right\}.\end{aligned}$$

Therefore we get

$$\begin{aligned}\boldsymbol{\mu}\,|\,\boldsymbol{\Omega}, b, \boldsymbol{Y} &\sim \mathcal{N}_q\big(\bar{\boldsymbol{\mu}}, (n\boldsymbol{\Omega})^{-1}\big)\\ \boldsymbol{\Omega}\,|\,b, \boldsymbol{Y} &\sim \mathcal{W}_q\big(a_D + n - 1, \boldsymbol{R}\big).\end{aligned}$$

The posterior normalizing constant is then

$$c(a_D + n - 1, q) \cdot |\boldsymbol{R}|^{\frac{a_D+n-1}{2}},$$

$$c(a_D + n - 1, q) \;=\; (2\pi)^{-\frac{q}{2}} n^{\frac{q}{2}} \cdot 2^{-\frac{(a_D+n-1)q}{2}} \left[\Gamma_q\left(\frac{a_D + n - 1}{2}\right)\right]^{-1}.$$

C.2.3 Marginal likelihood

The fractional marginal likelihood for the Gaussian model $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$ given the data $\boldsymbol{Y}$ is obtained as

$$\begin{aligned} m^F(\boldsymbol{Y}) &= (2\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{c(a_D + n_0 - 1, q)}{c(a_D + n - 1, q)} \cdot \frac{|n_0\bar{\boldsymbol{R}}|^{\frac{a_D+n_0-1}{2}}}{|\boldsymbol{R}|^{\frac{a_D+n-1}{2}}}\\ &= (\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{\Gamma_q\left(\frac{a_D+n-1}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0-1}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{q}{2}} \cdot \frac{|n_0 n^{-1}\boldsymbol{R}|^{\frac{a_D+n_0-1}{2}}}{|\boldsymbol{R}|^{\frac{a_D+n-1}{2}}}\\ &= (\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{\Gamma_q\left(\frac{a_D+n-1}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0-1}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{q(a_D+n_0)}{2}} \cdot |\boldsymbol{R}|^{-\frac{n-n_0}{2}}, \end{aligned} \tag{C.3}$$

since $\boldsymbol{R}$ is a $q \times q$ matrix and so

$$|n_0 n^{-1}\boldsymbol{R}|^{\frac{a_D+n_0-1}{2}} = \left(\frac{n_0}{n}\right)^{\frac{q(a_D+n_0-1)}{2}} |\boldsymbol{R}|^{\frac{a_D+n_0-1}{2}}.$$

In analogy with the standard conjugate case described in Section B.2, we obtain the fractional marginal likelihood with respect to $\boldsymbol{Y}_A$ as

$$\begin{aligned} m^F(\boldsymbol{Y}_A) &= (2\pi)^{-\frac{(n-n_0)|A|}{2}} \cdot \frac{c(a_D - |\bar{A}| + n_0 - 1, |A|)}{c(a_D - |\bar{A}| + n - 1, |A|)} \cdot \frac{|n_0\bar{\boldsymbol{R}}_{AA}|^{\frac{a_D-|\bar{A}|+n_0-1}{2}}}{|\boldsymbol{R}_{AA}|^{\frac{a_D-|B|+n-1}{2}}}\\ &= (\pi)^{-\frac{(n-n_0)|A|}{2}} \cdot \frac{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n-1}{2}\right)}{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n_0-1}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{|A|(a_D-|\bar{A}|+n_0)}{2}} \cdot |\boldsymbol{R}_{AA}|^{-\frac{n-n_0}{2}} \end{aligned} \tag{C.4}$$

C.3 Multivariate linear regression

Consider now the Gaussian multivariate linear regression model of Section B.3,

$$\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega} \sim \mathcal{N}_{n,q}\big(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Omega}^{-1}\big)$$

and assume the default prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$,

$$p^D(\boldsymbol{B}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}}.$$

C.3.1 Fractional prior

The fractional prior is obtained as

$$p^F(\boldsymbol{B}, \boldsymbol{\Omega} \mid b, \boldsymbol{Y}) \propto f^b(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega}) p^D(\boldsymbol{B}, \boldsymbol{\Omega}).$$

Setting $b = n_0/n$ we obtain

$$\begin{aligned} f^b(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega}) &= \left[(2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}\big\}\Big]\right\}\right]^{\frac{n_0}{n}} \\ &= (2\pi)^{-\frac{n_0 q}{2}} |\boldsymbol{\Omega}|^{\frac{n_0}{2}} \exp\left\{-\frac{n_0}{2n}\mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}\big\}\Big]\right\} \\ &\propto |\boldsymbol{\Omega}|^{\frac{n_0}{2}} \exp\left\{-\frac{n_0}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\big\}\Big]\right\} \end{aligned}$$

where

$$\tilde{\boldsymbol{C}} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}, \quad \tilde{\boldsymbol{R}} = \frac{1}{n}\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}.$$

Therefore we can write the fractional prior for $(\boldsymbol{B}, \boldsymbol{\Omega})$ as

$$\begin{aligned} p^F(\boldsymbol{B}, \boldsymbol{\Omega} \mid n_0, \boldsymbol{Y}) &\propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}} |\boldsymbol{\Omega}|^{\frac{n_0}{2}} \exp\left\{-\frac{n_0}{2}\mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}(\boldsymbol{B}-\hat{\boldsymbol{B}})\big\}\Big]\right\} \\ &\cdot \quad \exp\left\{-\frac{n_0}{2}\mathrm{tr}\big[\boldsymbol{\Omega}\tilde{\boldsymbol{R}}\big]\right\}. \end{aligned}$$

Hence, by multiplying and dividing by $|\boldsymbol{\Omega}|^{p+1}$, we get

$$\begin{aligned} \boldsymbol{B} \mid \boldsymbol{\Omega}, b, \boldsymbol{Y} &\sim \mathcal{N}_{p+1,q}\big(\hat{\boldsymbol{B}}, (n_0\tilde{\boldsymbol{C}})^{-1}, \boldsymbol{\Omega}^{-1}\big), \\ \boldsymbol{\Omega} \mid b, \boldsymbol{Y} &\sim \mathcal{W}_q\big(a_D + n_0 - p - 1, n_0\tilde{\boldsymbol{R}}\big). \end{aligned}$$

The prior normalizing constant is

$$K(n_0\tilde{\boldsymbol{C}}, n_0\tilde{\boldsymbol{R}}, a_D + n_0 - p - 1) = \frac{(2\pi)^{\frac{q(p+1)}{2}}\Gamma_q\big(\frac{a_D+n_0-p-1}{2}\big)2^{\frac{q(a_D+n_0-p-1)}{2}}}{|n_0 n^{-1}\boldsymbol{X}^\top \boldsymbol{X}|^{\frac{q}{2}} |n_0 n^{-1}\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}|^{\frac{a_D+n_0-p-1}{2}}}$$

C.3.2 Posterior

The posterior distribution of $(\boldsymbol{B}, \boldsymbol{\Omega})$ is obtained from

$$p(\boldsymbol{B}, \boldsymbol{\Omega} \,|\, \boldsymbol{Y}) \propto f^{1-b}(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega}) p^F(\boldsymbol{B}, \boldsymbol{\Omega} \,|\, n_0, \boldsymbol{Y}),$$

where

$$\begin{aligned} f^{1-b}(\boldsymbol{Y} \mid \boldsymbol{B}, \boldsymbol{\Omega}) &= \left[(2\pi)^{-\frac{nq}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}\big\}\Big]\right\}\right]^{\frac{n-n_0}{n}} \\ &\propto |\boldsymbol{\Omega}|^{\frac{n-n_0}{2}} \exp\left\{ -\frac{n-n_0}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\big\}\Big]\right\}. \end{aligned}$$

Hence,

$$\begin{aligned} p(\boldsymbol{B}, \boldsymbol{\Omega} \,|\, \boldsymbol{Y}) &\propto |\boldsymbol{\Omega}|^{\frac{n-n_0}{2}} \exp\left\{ -\frac{n-n_0}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\big\}\Big]\right\} \\ &\cdot\ |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}} |\boldsymbol{\Omega}|^{\frac{n_0}{2}} \exp\left\{ -\frac{n_0}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\big\}\Big]\right\} \\ &\propto |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{n}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \tilde{\boldsymbol{C}}(\boldsymbol{B}-\hat{\boldsymbol{B}}) + \tilde{\boldsymbol{R}}\big\}\Big]\right\} \\ &= |\boldsymbol{\Omega}|^{\frac{a_D-q-1}{2}} |\boldsymbol{\Omega}|^{\frac{n}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{(\boldsymbol{B}-\hat{\boldsymbol{B}})^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{B}-\hat{\boldsymbol{B}})\big\}\Big]\right\} \\ &\cdot\ \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big[\boldsymbol{\Omega}\big\{\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}\big\}\Big]\right\}. \end{aligned}$$

Again, by multiplying and dividing by $|\boldsymbol{\Omega}|^{p+1}$, we get

$$\begin{aligned} \boldsymbol{B} \,|\, \boldsymbol{\Omega}, \boldsymbol{Y} &\sim \mathcal{N}_{p+1,q}\big(\hat{\boldsymbol{B}}, (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \boldsymbol{\Omega}^{-1}\big), \\ \boldsymbol{\Omega} \,|\, \boldsymbol{Y} &\sim \mathcal{W}_q\big(a_D + n - p - 1, \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}\big). \end{aligned}$$

The posterior normalizing constant is then

$$K(\boldsymbol{X}^\top \boldsymbol{X}, \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}, a_D + n - p - 1) = \frac{(2\pi)^{\frac{q(p+1)}{2}} \Gamma_q\big(\frac{a_D+n-p-1}{2}\big) 2^{\frac{q(a_D+n-p-1)}{2}}}{|\boldsymbol{X}^\top \boldsymbol{X}|^{\frac{q}{2}} |\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}|^{\frac{a_D+n-p-1}{2}}}.$$

C.3.3 Marginal likelihood

The fractional marginal likelihood is obtained as

$$\begin{aligned} m^F(\boldsymbol{Y} \,|\, \boldsymbol{X}) &= (2\pi)^{-\frac{(n-n_0)q}{2}} \frac{K^{-1}(n_0\tilde{\boldsymbol{C}}, n_0\tilde{\boldsymbol{R}}, a_D+n_0-p-1)}{K^{-1}(\boldsymbol{X}^\top\boldsymbol{X}, \hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}, a_D+n-p-1)} \\ &= (2\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{|n_0 n^{-1}\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}} |n_0 n^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}|^{\frac{a_D+n_0-p-1}{2}}}{|\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}} |\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}|^{\frac{a_D+n-p-1}{2}}} \\ &\cdot \frac{\Gamma_q\left(\frac{a_D+n-p-1}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0-p-1}{2}\right)} \cdot \frac{2^{\frac{q(a_D+n-p-1)}{2}}}{2^{\frac{q(a_D+n_0-p-1)}{2}}}. \end{aligned}$$

Moreover, since $\boldsymbol{X}^\top\boldsymbol{X}$ is a $(p+1)\times(p+1)$ matrix, we have

$$|n_0 n^{-1}\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}} = \left(\frac{n_0}{n}\right)^{\frac{q(p+1)}{2}} |\boldsymbol{X}^\top\boldsymbol{X}|^{\frac{q}{2}}$$

and similarly for the $q \times q$ matrix $\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}$. Hence,

$$\begin{aligned} m^F(\boldsymbol{Y} \,|\, \boldsymbol{X}) &= (2\pi)^{-\frac{(n-n_0)q}{2}} \cdot \left(\frac{n_0}{n}\right)^{\frac{q(p+1)}{2}} |\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}|^{-\frac{n-n_0}{2}} \cdot \frac{\Gamma_q\left(\frac{a_D+n-p-1}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0-p-1}{2}\right)} \\ &\cdot \left(\frac{n_0}{n}\right)^{\frac{q(a_D+n_0-p-1)}{2}} 2^{\frac{q(n-n_0)}{2}} \end{aligned}$$

and then

$$m^F(\boldsymbol{Y} \,|\, \boldsymbol{X}) = (\pi)^{-\frac{(n-n_0)q}{2}} \cdot \frac{\Gamma_q\left(\frac{a_D+n-p-1}{2}\right)}{\Gamma_q\left(\frac{a_D+n_0-p-1}{2}\right)} \cdot \left(\frac{n_0}{n}\right)^{\frac{q(a_D+n_0)}{2}} \cdot |\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}|^{-\frac{n-n_0}{2}}. \quad \text{(C.5)}$$

Let now $A \subset \{1,\dots,q\}$, $\boldsymbol{Y}_A$ the corresponding $n \times |A|$ submatrix of $\boldsymbol{Y}$ and $\boldsymbol{B}_A$ the $(p+1)\times|A|$ submatrix of $\boldsymbol{B}$, whose columns contain the regression coefficients for the selected responses in $\boldsymbol{Y}_A$. In analogy with the standard conjugate case described in Section B.3, we obtain the fractional marginal likelihood with respect to $\boldsymbol{Y}_A$ as

$$m^F(\boldsymbol{Y}_A \,|\, \boldsymbol{X}) = \frac{K(\boldsymbol{X}^\top\boldsymbol{X}, \hat{\boldsymbol{E}}_A^\top\hat{\boldsymbol{E}}_A, a_D-|\bar{A}|+n-p-1)}{(2\pi)^{\frac{(n-n_0)|A|}{2}} K(n_0\tilde{\boldsymbol{C}}, n_0\tilde{\boldsymbol{R}}_{AA}, a_D-|\bar{A}|+n_0-p-1)}$$

and then

$$m^F(\boldsymbol{Y}_A \,|\, \boldsymbol{X}) = (\pi)^{-\frac{(n-n_0)|A|}{2}} \frac{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n-p-1}{2}\right)}{\Gamma_{|A|}\left(\frac{a_D-|\bar{A}|+n_0-p-1}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{|A|(a_D-|\bar{A}|+n_0)}{2}} |\hat{\boldsymbol{E}}_A^\top\hat{\boldsymbol{E}}_A|^{-\frac{n-n_0}{2}}, \quad \text{(C.6)}$$

being $|\bar{A}| = q - |A|$ and $\hat{\boldsymbol{E}}_A = \boldsymbol{Y}_A - \boldsymbol{X}^\top\hat{\boldsymbol{B}}_A$.

Bibliography

ANDERSSON, S. A., MADIGAN, D. & PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* 25 505–541.

ANDERSSON, S. A., MADIGAN, D. & PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics* 28 33–85.

BARBIERI, M. M. & BERGER, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics* 32 870–897.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 289–300.

BERGER, J. O. & 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., PERICCHI, L. R., GHOSH, J. K., SAMANTA, T., SANTIS, F. D., BERGER, J. O. & PERICCHI, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series* 38 pp. 135–207.

BESAG, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society* 36 192–236.

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

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

Casella, G. & Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* 101 157–167.

Castelo, R. & Perlman, M. D. (2004). Learning essential graph Markov models from data. In *Advances in Bayesian networks*, vol. 146 of *Studies in Fuzziness and Soft Computing.* Springer, Berlin, 255–269.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95 759–771.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* 2 445–498.

Consonni, G. & La Rocca, L. (2012). Objective Bayes factors for Gaussian directed acyclic graphical models. *Scandinavian Journal of Statistics* 39 743–756.

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

Cowell, R. G., Dawid, P. A., Lauritzen, S. L. & Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* New York: Springer.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* 21 1272–1317.

Dor, D. & Tarsi, M. (1992). Simple algorithm to construct a consistent extension of a partially oriented graph. *Technical Report R-185, Cognitive Systems Laboratory, UCLA* .

Drton, M. (2009). Discrete chain graph models. *Bernoulli* 15 736–753.

Drton, M. & Eichler, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics* 33 247–257.

Eberhardt, F. & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science* 74 981–995.

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

Foygel, R. & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 23*. 2020–2028.

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

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* 303 799–805.

Geiger, D. & Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics* 30 1412–1440.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *The Annals of Mathematical Statistics* 36 150–159.

Geisser, S. & Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* 25 368–376.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2nd ed.

Gillispie, S. B. & Perlman, M. D. (2002). The size distribution for Markov equivalence classes of acyclic digraph models. *Artificial Intelligence* 141 137–155.

Hauser, A. & Bhlmann, P. (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society. Series B (Methodological)* 77 291–318.

Hauser, A. & Bühlmann, P. (2012). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* 13 2409–2464.

He, Y. & Geng, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research* 9 2523–2547.

He, Y., Jia, J. & Yu, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *The Annals of Statistics* 41 1742–1779.

Ibrahim, J. G. & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* 15 46–60.

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

KOLLER, D. & FRIEDMAN, N. (2009). *Probabilistic graphical models: principles and techniques.* MIT press.

LAURITZEN, S. L. (1996). *Graphical Models.* Oxford University Press.

MADIGAN, D., ANDERSSON, S., PERLMAN, M. & VOLINSKY, C. (1996). Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Communications in Statistics: Theory and Methods* 2493–2519.

MORENO, E. (1997). Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In Y. Dodge, ed., *L_1-Statistical Procedures and Related Topics.* Institute of Mathematical Statistics, 257–270.

NAGARAJAN, R., SCUTARI, M. & LBRE, S. (2013). *Bayesian Networks in R: With Applications in Systems Biology.* Springer Publishing Company, Incorporated.

NORRIS, J. R. (1997). *Markov Chains.* Cambridge University Press, Cambridge.

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

O'HAGAN, A. & FORSTER, J. J. (2004). *Bayesian Inference. Kendall's Advanced Theory of Statistics.* Arnold, 2nd ed.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* 82 669–688.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.

PEREZ, J. M. & BERGER, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* 89 491–511.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In D. Dey & C. R. Rao, eds., *Bayesian thinking: modeling and computation*, vol. 25 of *Handbook of Statistics*. Elsevier, 115–149.

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

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

Roverato, A. (2005). A unified approach to the characterization of equivalence classes of dags, chain graphs with no flags and chain graphs. *Scandinavian Journal of Statistics* 32 295–312.

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

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

Shojaie, A. & Michailidis, G. (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology* 16 407–26.

Tan, L. S. L., Jasra, A., De Iorio, M. & Ebbels, T. M. D. (2016). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *ArXiv e-prints* .

Verma, T. & Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI 90. New York, NY, USA: Elsevier Science Inc., 255–270.