Janne Leppä-aho
janne.leppa-aho@helsinki.fi
013612650

# Pseudo-Likelihood Learning of Gaussian Graphical Models

Master's thesis

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

| Tiedekunta/Osasto — Fakultet/Sektion — Faculty | Laitos — Institution — Department |
|---|---|
| Faculty of Science | Department of Mathematics and Statistics |

Tekijä — Författare — Author
Janne Leppä-aho

Työn nimi — Arbetets titel — Title

Pseudo-Likelihood Learning of Gaussian Graphical Models

Oppiaine — Läroämne — Subject
Mathematics

| Työn laji — Arbetets art — Level | Aika — Datum — Month and year | Sivumäärä — Sidoantal — Number of pages |
|---|---|---|
| Master's thesis | October 2014 | 57 p. |

Tiivistelmä — Referat — Abstract

Multivariate Gaussian distribution is an often encountered continuous distribution in applied mathematics and statistics due to its well known properties and wide applicability. In the graphical models framework, we make use of graphs to compactly represent the conditional independences between a set of random variables. Combining these two together leads to the class of Gaussian graphical models.

This thesis discusses learning of Gaussian graphical models from multivariate data. Given the data, our goal is to identify the graphical structure that specifies the conditional independence statements between the variables under consideration.

Following the footsteps of Pensar et al [10], we adopt a Bayesian, score-based approach for learning graphical models. Using *pseudo-likelihood* to approximate the true likelihood allows us to apply results of Consonni et al [4] to compute marginal likelihood integrals in closed form. This results in a method that can be used to make objective comparisons among Gaussian graphical models.

We test the method numerically and show that it can be readily applied in high-dimensional settings. According to our tests, the method presented here outperforms the widely used graphical LASSO method in accuracy.

The structure of this thesis is as follows. The chapters 2-4 discuss graphical models, multivariate Normal distribution and Bayesian model comparison in general. The fifth chapter goes through the results derived by Consonni, which are utilised in the next chapter to develop a scoring function and a learning algorithm for Gaussian graphical model selection. In the sixth chapter, we test the method in practice and present the obtained numerical results. The last appendix chapter is dedicated to the consistency proof, which gives the theoretical justification for the presented method.

Avainsanat — Nyckelord — Keywords
Gaussian graphical models, pseudo-likelihood, fractional Bayes factors

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpula Campus Library

Muita tietoja — Övriga uppgifter — Additional information

# Contents

# 1 Introduction

Multivariate Gaussian distribution is an often encountered continuous distribution in applied mathematics and statistics due to its well known properties and wide applicability. In the graphical models framework, we make use of graphs to compactly represent the conditional independences between a set of random variables. Combining these two together leads to the class of Gaussian graphical models.

This thesis discusses learning of Gaussian graphical models from multivariate data. Given the data, our goal is to identify the graphical structure that specifies the conditional independence statements between the variables under consideration.

Following the footsteps of Pensar et al [10], we adopt a Bayesian, score-based approach for learning graphical models. Using *pseudo-likelihood* to approximate the true likelihood allows us to apply results of Consonni et al [4] to compute marginal likelihood integrals in closed form. This results in a method that can be used to make objective comparisons among Gaussian graphical models.

We test the method numerically and show that it can be readily applied in high-dimensional settings. According to our tests, the method presented here outperforms the widely used graphical LASSO method in accuracy.

The structure of this thesis is as follows. The chapters 2-4 discuss graphical models, multivariate Normal distribution and Bayesian model comparison in general. The fifth chapter goes through the results derived by Consonni, which are utilised in the next chapter to develop a scoring function and a learning algorithm for Gaussian graphical model selection. In the seventh chapter, we test the method in practice and present the obtained numerical results. The last appendix chapter is dedicated to the consistency proof, which gives the theoretical justification for the presented method.

# 2 Probabilistic graphical models

In this section, we will discuss probabilistic graphical models. After a short general introduction, two important subclasses, directed acyclic graphs and undirected graphs, are considered. Main references used in writing this chapter are Whittaker's [13] and Koller's [8] books.

## 2.1 Graphical models framework

Probabilistic graphical models provide a convenient way to represent dependency structures of complex distributions in high dimensional spaces with the aid of mathematical objects called graphs. An ordinary graph is a fairly simple object specified by its nodes and edges connecting them. Each node is considered to correspond a random variable and the edges connected to it represent probabilistic interactions between it and some other variables. Absence of an edge between variables is a statement of a conditional independence.

More generally, the graph provides a complete representation of all the conditional independence assumptions that hold between the variables in consideration. Through specifying the conditional independence assumptions, the graph provides us a way to factorize a complex joint distribution into smaller components that are easier to handle and understand.

This compact and more tractable *representation* of complex objects is one of the advantages in the probabilistic graphical models framework mentioned by Koller and Friedman [8]. The other two important properties are related to *inference* and *learning*.

Knowing the underlying graphical model makes the *inference* easier. For example, consider a situation where we have hundreds of possibly interconnecting variables and we would be interested in the conditional distribution of only few of them, given the others. The conditional independences specified by the graph would provide us the information about the relevant variables, which are sufficient to take under consideration in order to do the inference.

The framework of probabilistic graphical models makes also efficient learning of models from data possible. Of course one could construct models by specifying them by hand, but usually it's more sensible to just provide a rough guideline, which should be fulfilled by the model and let other properties be determined automatically by a learning algorithm. This automatic approach might also reveal surprising connections between variables that might have

otherwise gone unnoticed.

The broad and flexible framework of probabilistic graphical models makes them also applicable in various real life problems. Applications can be found in numerous fields such as medical diagnosis, fault diagnosis, analysis of genetic and genomic data, communication and coding, analysis of marketing data, speech recognition and natural language understanding [8].

Two common and widely used subclasses of graphical models are *directed acyclic graphs* (DAG) and *undirected graphical models*. As is presumably clear from the name, DAGs are graphical models in which all the edges have a direction. This direction of probabilistic interaction allows one to also use DAGs for example to study causal relationships between variables. On the other hand, the undirected graphical models do not specify directions of connections. We will start the more formal treatment of the subject by introducing DAGs in detail.

## 2.2 Directed acyclic graphs

Let $X = (X_1, \ldots, X_p)^T$ denote a $p$-dimensional random vector. In this chapter and onwards, when referring to subvectors of $X$, we use similar notation as in [13]: By writing $X_a,\ a \subset \{1, 2, \ldots, p\}$, we are referring to a subvector of $X$ that consists of variables whose indices are included in the set $a$.

In the graphical models framework, each component $X_i,\ i = 1, \ldots p$ corresponds to a node in the graph. We will use $V = \{1, \ldots, p\}$ to denote the set of nodes. When considering DAGs, we assume that every edge is directed. An edge pointing from node $j$ to node $i$ is denoted by $(j, i)$. We use $E \subset V \times V$ to denote the set of edges. A graph $G$ is then defined as the pair $G = (V, E)$.

One motivation for using graphical models was their ability to provide us way a to factorize the joint probability distribution of several random variables. This is also the reason why we do not allow directed *cycles* in our graphs.

**Definition 2.2.1.** *A cycle is a finite sequence of nodes* $(v_1, \ldots, v_m)$*, where* $(v_j, v_{j+1}) \in E$ *for every* $j = 1, \ldots, m-1$ *and* $v_1 = v_m$*.*

Consider a simple example of a graph containing cycle. Let $V = \{1, 2, 3\}$ be the set of nodes and $E = \{(1, 2), (2, 3), (3, 1)\}$ the edge set. We would now hope that the joint density $f_{123}$ could be factorised into a product of

conditional densities according to

$$f_{123} = f_{2|1} \cdot f_{3|2} \cdot f_{1|3}.$$

However, the factorisation above defines a proper joint density in rare cases [13].

Directed graphs that do not posses any cycles can be equivalently characterized as graphs whose nodes can be completely ordered. Ordering means here that we can find a binary relation $\preccurlyeq$ for the elements of $V$ so that

1. $i \preccurlyeq j$ or $j \preccurlyeq i$ for every $i, j \in V$.
2. Relation $\preccurlyeq$ is irreflexive.
3. Relation $\preccurlyeq$ is transitive.

With the help of the relation, we can order the elements of $V$ as $1 \preccurlyeq 2 \preccurlyeq \cdots \preccurlyeq p$. This is summarized by the following theorem (Lemma 3.5.1 in [13])

**Theorem 2.2.2.** *In a directed graph $G$, the following statements are equivalent*

*(i) There is no directed cycle in $G$.*

*(ii) There exists a complete ordering of the nodes that is respected in the graph.*

Ordering of the nodes can be seen to provide us a way to define "past" and "present" for each of the variables. We can for example say rigorously if one node is a descendent of the other and speak of the *parents* of a node.

**Definition 2.2.3.** *Let $G = (V, E)$ be a DAG. The set of parents of node $j$ is denoted by $pa(j)$, and the set consists of nodes, which have an edge pointing to $j$. More formally*

$$pa(j) = \{i \in V |\ (i, j) \in E\}.$$

This allows us to specify the conditional independence assumptions implied by a DAG.

**Definition 2.2.4.** *Let $N(i)$ denote the set of variables that are not descendants of node $i$ in directed acyclic graph $G = (V, E)$. The DAG $G$ encodes the following set of conditional independence assumptions*

$$X_i \perp\!\!\!\perp N(i) \mid pa(i),\ i = 1, 2, \ldots p,$$

*that is, each variable $X_i$ is conditionally independent of its non-descendants given its parent nodes.*

**Definition 2.2.5.** *Let $p$ denote the joint probability distribution of $X = (X_1, \ldots, X_p)^T$. Assume $G = (V, E)$ is a directed acyclic graph, where the nodes correspond to components of $X$. We say that $p$ factorises according to $G$, if it can be expressed as*

$$p(X) = \prod_{i=1}^{p} p(X_i \mid pa(i)). \tag{1}$$

Now we are ready to define the DAG models.

**Definition 2.2.6.** *A DAG model $M$ is the pair $M = (G, \mathcal{F}_G)$, where $G$ is a directed acyclic graph and $\mathcal{F}_G$ is the set of allowable local distribution families. The local distributions refer to the components of (1). In other words, a DAG model is specified by the graph $G$ and the set of joint distributions that factorize according to $G$ into local distributions such that each distribution belongs to a family present in the set $\mathcal{F}_G$.*

Our definition of a DAG model is the same as used for example by Geiger and Heckerman in [5].

## 2.3 Undirected graphical models

Undirected graphical models, also called Markov networks or Markov random fields use undirected graphs to represent the conditional independence statements among a set of random variables. Statements that can be encoded with an undirected graph may differ from those discussed in the previous section.

As in the case of DAGs, we let $V = \{1, \ldots, p\}$ to denote a set of nodes, which correspond to random variables $X = (X_1, \ldots, X_p)^T$. Likewise, the set of edges is denoted by $E \subset V \times V$.

An undirected graph $G = (V, E)$ is a graph were each edge is undirected. If there is an undirected edge between nodes $i$ and $j$, then $(i, j)$ and $(j, i)$ are both in set the $E$. Absence of an edge in the graph is a statement of conditional independence between the corresponding random variables. More formally:

**Theorem 2.3.1.** *The undirected graph $G = (V, E)$ for a random vector $X$ encodes the following set of conditional independences:*

$$X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}} \text{ if, and only if } (i, j) \notin E \text{ and } (j, i) \notin E.$$

Somewhat similar concept as the set of parents in a DAG is the Markov blanket defined in undirected graphs.

**Definition 2.3.2.** *The Markov blanket of a node is the set of nodes directly connected to it. More formally*

$$mb(j) = \{i \in V \mid (i, j) \in E \text{ and } (j, i) \in E\}.$$

**Definition 2.3.3.** *We define a family of the node $j$ to be the set*

$$fa(j) = mb(j) \cup \{j\}.$$

We can find two equivalent formulations for the **Theorem 2.3.1**.

**Theorem 2.3.4.** *Assume the joint distribution of $X = (X_1, \ldots, X_p)^T$ is positive and let $G = (V, E)$ be an undirected graph. Now the following three statements are equivalent:*

1. $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ *if and only if* $(i, j) \notin E$ *and* $(j, i) \notin E$.
2. $X_i \perp\!\!\!\perp X_{V\setminus fa(i)} \mid mb(i)$ *for every* $i \in V$.
3. $X_a \perp\!\!\!\perp X_b \mid X_s$ *for every disjoint subsets* $a, b$ *and* $s$ *of* $V$, *such that* $s$ *separates* $a$ *from* $b$.

These statements are called pairwise, local and global Markov properties, respectively. "To separate" means here that we cannot follow the edges of $G$ to end up from a node in set the $a$ to a node in $b$ without passing a node of $s$ on our way.

In Markov networks the joint distribution of variables can be expressed as a product of individual factors corresponding to the *maximal cliques* in the underlying graph $G$.

**Definition 2.3.5.** *A clique $C \subset V$ is a set where every pair of nodes is connected, that is,*

$$\{i, j\} \in C, \text{ if } (i, j) \in E \text{ and } (j, i) \in E. \tag{2}$$

*A clique $C \subset V$ is maximal, if adding any node $i \in V$ to $C$ would contradict with (2).*

**Definition 2.3.6.** *Let $p$ be a positive distribution for a random vector $X$. Denote the set of maximal cliques related to the graph $G$ by $\mathcal{C}(G)$. We say that the $p$ factorizes according to $G$ if it can be expressed as*

$$p(X) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \Phi(X_C), \tag{3}$$

*where $Z = \int \prod_{C \in \mathcal{C}(G)} \Phi(X_C)$, is a normalizing constant, also called the partition function.*

An undirected graphical model is specified by the undirected graph $G$ and the family of multivariate probability distributions, that factorize according to $G$.

Exact inference and learning of undirected graphical models from data is hard due the global normalizing constant $Z$. In general, the constant $Z$ couples all the parameters of a model and prevents us from factorising the problem into a simpler sub-problems.

Inference gets easier if one assumes that the underlying graph is *chordal*, but this might be too restrictive [10]. Chordal graphs are undirected graphs, which do not have any cycle longer than three that would not contain a "short-cut", that is, an edge connecting a pair of nodes in the cycle.

# 3 Multivariate Normal distribution

In this chapter we present the multivariate Normal distribution and some of its basic properties. Main references used in writing this chapter are chapters 5 and 6 from the book of Whittaker [13] and all the results stated here can be found in the book.

**Definition 3.0.7.** *The $p$-dimensional random vector $X$ follows multivariate Normal distribution if and only if its density function can be written as*

$$f_X(x) = |\Sigma|^{-1/2}(2\pi)^{-p/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),\ x \in \mathbb{R}^p,$$

*or equivalently, using $\Omega = \Sigma^{-1}$,*

$$f_X(x) = |\Omega|^{1/2}(2\pi)^{-p/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Omega\ (x-\mu)\right),\ x \in \mathbb{R}^p,$$

*where $\Omega$ and $\Sigma$ are symmetric and positive definite $p \times p$ matrices, and $\mu$ is a fixed $p$-dimensional vector.*

To state that a random vector $X$ follows a $p$-variate Normal distribution with parameters $\mu$ and $\Sigma$, we write

$$X \sim N_p(\mu, \Sigma).$$

The meaning of these parameters is specified in the following theorem.

**Theorem 3.0.8.** *Assume that $X \sim N_p(\mu, \Sigma)$. Then the expected value and the variance of $X$ are*

$$\mathbb{E}(X) = \mu \text{ and } var(X) = \Sigma.$$

*Proof.* Can be found in [13]. □

## 3.1 Marginal and conditional distributions

The following theorems establish the fact that the class of multivariate Normal distributions is closed under marginalization and conditioning. In other words, every conditional or marginal density function obtained from multivariate Normal density is also multivariate Normal.

**Theorem 3.1.1.** *Assume that $X \sim N_p(\mu, \Sigma)$. Partition the vector $X$ as $X = (X_a, X_b)$. Partition the mean vector $\mu$ and covariance matrix $\Sigma$ accordingly as*

$$\mu = (\mu_a, \mu_b) \text{ and } \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}.$$

*Now,*

*(i) the marginal distribution of $X_a$ is Normal with mean $\mu_a$ and variance, $\Sigma_{aa}$*

*(ii) the conditional distribution of $X_b$ given $X_a = x_a$ is Normal with mean*

$$\mathbb{E}_{b|a}(X_b) = \mu_b + (\Sigma_{ba}\Sigma_{aa}^{-1})(x_a - \mu_a),$$

*and variance*

$$var_{\,b|a}(X_b) \equiv \Sigma_{bb|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}.$$

*Proof.* Can be found in [13]. □

Conditional mean and variance are connected to linear least squares prediction and partial variance under the Gaussian assumption. To see this, recall the definitions of the linear least squares predictor and the partial variance.

**Definition 3.1.2.** *Let $X$ and $Y$ be $p$- and $q$-dimensional random vectors, respectively. Assume that $\mathbb{E}[(X, Y)] = 0$ and $var[(X, Y)]$ is known. The linear least squares predictor of $Y$ from $X$ is then defined as*

$$\hat{Y}[X] = cov(X, Y)var(X)^{-1}X = BX,$$

*where $B = cov(X, Y)var(X)^{-1} \in \mathbb{R}^{q \times p}$ is referred as the matrix of the linear least squares prediction coefficients.*

Note that the zero mean assumption is not really necessary. We could get more general formula just by substituting $X = X - \mathbb{E}(X)$ and $Y = Y - \mathbb{E}(Y)$ to the formula above.

When $p = 1$, then $Y$ is a scalar random variable and $\hat{Y}$ is found by minimizing the expression $\mathbb{E}(Y - b^T X)$ with respect to vector $b$. It can be

shown that if and only if $b$ minimizes this expression, it solves *the normal equations*

$$\operatorname{cov}(Y - b^T X, X) = 0. \tag{4}$$

In case $p > 1$, linear least squares predictor is found by solving the normal equations (4), where instead of $b^T$, we have the matrix $B$.
Now we can define the partial covariance and variance.

**Definition 3.1.3.** *The partial covariance of $Y$ and $Z$ given $X$, is defined as*

$$cov\,(Y, Z|X) = cov\,(Y - \hat{Y}[X], Z - \hat{Z}[X]).$$

**Corollary 3.1.4.** *The partial covariance satisfies*

$$cov\,(Y, Z|X) = cov\,(Y, Z) - cov\,(Y, X) var\,(X)^{-1} cov\,(X, Z).$$

*Proof.* This can be seen easily by using the definition of partial covariance, bilinearity of covariance operator and the fact that predictor coefficient matrix satisfies the normal equations. □

From the **Corollary 3.1.4**, we obtain

**Corollary 3.1.5.** *The partial variance of $Y$ given $X$,*
*$var(Y|X) = cov(Y, Y|X)$ satisfies*

$$\begin{aligned} var\,(Y|X) &= var\,(Y) - var\,(\hat{Y}[X]) \\ &= var\,(Y) - cov\,(Y, X) var\,(X)^{-1} cov\,(X, Y). \end{aligned}$$

Using these, we can establish the following theorem

**Corollary 3.1.6.** *$X = (X_a, X_b) \sim N_p(\mu, \Sigma)$. Partition the covariance matrix respectively, as in Theorem 3.1.1. Now it holds that*

$$\mathbb{E}_{b|a}(X_b) = \hat{X}_b(x_a) \text{ and } var_{\,b|a}(X_b) = var(X_b|X_a).$$

*Proof.* Using the **Definition 3.1.2**, we can write the linear least squares predictor of $X_b$ from $X_a = x_a$

$$\begin{aligned} \hat{X}_b(x_a) &= \mathbb{E}(X_b) + \operatorname{cov}(X_b, X_a)\operatorname{var}(X_a)^{-1}(x_a - \mathbb{E}(X_a)) \\ &= \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a) \\ &= \mathbb{E}_{b|a}(X_b). \end{aligned}$$

For the second part of the statement, **Corollary 3.1.5** gives

$$\begin{aligned}\text{var}(X_b|X_a) &= \text{var}(X_b) - \text{cov}(X_b, X_a)\text{var}(X_a)^{-1}\text{cov}(X_a, X_b) \\ &= \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \\ &= \text{var}_{\ b|a}(X_b).\end{aligned}$$

□

In general, partial variance and conditional variance are two different things. The partial variance is evaluated in the joint distribution of $X_a$ and $X_b$, whereas the conditional variance is evaluated in the conditional distribution of $X_b$ given $X_a = x_a$. Under the assumption of Normality, these two are the same [13].

## 3.2 Independence and conditional independence

The following theorems consider independences and conditional independences of random Normal vectors and how these are reflected to the elements of covariance and precision matrices. These theorems are also the theoretical basis for defining the Gaussian graphical models.

**Theorem 3.2.1.** *Partition Normal random vector $X$ as $X = (X_a, X_b)$. $X_a$ and $X_b$ are independent, if and only if*

*(i)* $cov(X_a, X_b) = \Sigma_{ab} = 0$, *or*

*(ii)* $\Omega_{ab} = 0,\ \Omega = \Sigma^{-1}$.

**Theorem 3.2.2.** *Partition Normal random vector $X$ as $X = (X_a, X_b, X_c)$. $X_a$ and $X_b$ are independent conditional on $X_c$, $X_a \perp\!\!\!\perp X_b|X_c$, if and only if either*

*(i)* $cov(X_a, X_b|X_c) = 0$ *or*

*(ii)* $\Omega_{ab} = 0$.

*If we let $X_a$ and $X_b$ be one dimensional, then*

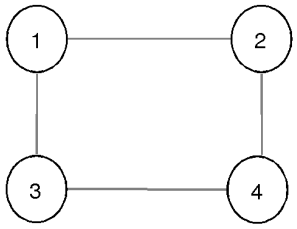
$$X_i \perp\!\!\!\perp X_j|\ X_{V\setminus\{i,j\}} \Leftrightarrow \Omega_{ij} = 0.$$

The message of these theorems is that the zeroes in the precision matrix of a Gaussian random vector are statements of conditional independence between the variables.

## 3.3 Gaussian graphical models

Let $X$ be a $p$-dimensional random vector with zero mean and precision matrix $\Omega$. Let $G$ be an undirected graph. We define the Gaussian graphical model (GGM) to be the family of multivariate Normal distributions for $X$ that satisfy the conditional independence statements implied by the graph. Due the **Theorem 3.2.2**, this means that we force some of the elements of the inverse covariance to be zero. The remaining elements can be chosen arbitrarily as long as the matrix is symmetric and positive definite.

A simple example of Gaussian graphical model is given in **Figure 1**.




$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} \quad \Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{13} & 0 \\ \omega_{21} & \omega_{22} & 0 & \omega_{24} \\ \omega_{31} & 0 & \omega_{33} & \omega_{34} \\ 0 & \omega_{42} & \omega_{43} & \omega_{44} \end{pmatrix}$$

Figure 1: An example network with four nodes and the corresponding covariance and precision matrices.

Absence of edges $(1, 4)$ and $(3, 2)$ in the graph implies that $\Omega_{14} = \Omega_{23} = 0$. Due the symmetry the corresponding elements below the diagonal are also zero. The Gaussian graphical model in this case would consist of all the multivariate Normal distributions, whose symmetric and positive definite precision matrices would have the same zero pattern as in the picture.

# 4 Bayes factors

This chapter discusses Bayes factors, which are tools for Bayesian model comparison. We will start with ordinary Bayes factors and their properties. In some cases, the use of Bayes factors is somewhat problematic. We study the problem where we would like to do objective model comparison by using improper priors for the model parameters. However, in this case the ordinary Bayes factors become unspecified.

Two solutions are proposed to overcome this: partial and fractional Bayes factors. This chapter is mainly based on O'Hagan's article [9], where the fractional Bayes factors were introduced for the first time. We will make use of fractional Bayes factors later when we consider the objective comparison of Gaussian DAGs.

## 4.1 Ordinary Bayes factors

Let $M_1$ and $M_2$ be two proposed models for data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Denote the sampling distribution of data under model $M_i$ by $f_i(\mathbf{x}|\boldsymbol{\theta}_i)$ and the corresponding prior distribution of parameters $\theta_i$ by $\pi_i(\theta_i)$. The marginal likelihood, or marginal data density of $\mathbf{X}$ under $M_i$ will be denoted by $p(\mathbf{X}|M_i)$.

**Definition 4.1.1.** *The marginal likelihood of $M_i$ given data $\boldsymbol{X}$ is*

$$p(\boldsymbol{X}|M_i) = \int \pi_i(\theta_i) f(\boldsymbol{X}|\theta_i) d\theta_i,$$

*where the integral is taken over all possible values of parameters $\theta_i$ of model $M_i$.*

The marginal likelihood can be seen to measure the fit of models to data, after the effect of parameters is marginalized out.

A natural way to compare models in Bayesian statistics is to compute their posterior probabilities given the observed data:

$$\frac{p(M_1|\mathbf{X})}{p(M_2|\mathbf{X})} = \frac{p(M_1)}{p(M_2)} \cdot \frac{p(\mathbf{X}|M_1)}{p(\mathbf{X}|M_2)} = \frac{p(M_1)}{p(M_2)} B_{12}(\mathbf{X}), \tag{5}$$

in which the quantity $B_{12}(\mathbf{X}) = p(\mathbf{X}|M_1)/p(\mathbf{X}|M_2)$ is defined to be the Bayes factor of model 1 against the model 2. Terms $p(M_1)$ and $p(M_2)$ are prior probabilities assigned to models themselves and they do not depend on the observed data.

We see that the Bayes factor is a ratio of the marginal-likelihoods of the models under comparison. If observed data provides more support to $M_1$ than $M_2$, we have $B_{12}(\mathbf{X}) > 1$ and vice versa.

Assuming identically distributed data $\mathbf{X}$ with sufficient regularity of the sampling distribution, one can show that Bayes factors are consistent, when comparing nested models. By saying that they are consistent, we mean that if the model 1 would be the right one, then

$$B_{12}(\mathbf{X}) \to \infty,$$

and the posterior probability $P(M_1|\mathbf{X}) \to 1$, as the sample size $n$ approaches infinity. Likewise, if the right model would have been model 2, then

$$B_{12}(\mathbf{x}) \to 0 \text{ and } P(M_2|\mathbf{X}) \to 1, \text{ when } n \to \infty.$$

For more thorough discussion on the consistency of ordinary Bayes factors we refer to O'Hagan [9] and further references therein.

## 4.2 Bayes factors with improper priors

Bayes factors are known to be sensitive to the choice of parameter prior $\pi_i(\theta_i)$. In order to do objective comparison between different models, we would like choose $\pi_i(\theta_i)$ to be as uninformative as possible. However, uninformative priors are often improper, which is problematic as shown in the following example.

Let $\pi_i(\theta_i)$ be an improper prior. This means that $\pi_i(\theta_i) \propto h_i(\theta_i)$, where $h_i(\theta_i)$ is a real function, but the integral

$$\int h_i(\theta_i) d\theta_i$$

diverges. Formally, we can write

$$\pi_i(\theta_i) = c_i \cdot h_i(\theta_i),$$

where the proportionality constant $c_i$ is unspecified and does not exist as a real number. Despite this, we can still write the posterior density for the

parameters $\theta_i$:

$$\begin{aligned}\pi_i(\theta_i|\mathbf{X}) &= \frac{\pi_i(\theta_i)f_i(\mathbf{X}|\theta_i)}{p(\mathbf{X}|M_i)} \\ &= \frac{c_i}{c_i} \cdot \frac{h_i(\theta_i)f_i(\mathbf{X}|\theta_i)}{\int h_i(t)f(\mathbf{X}|t)dt} \\ &= \frac{h_i(\theta_i)f_i(\mathbf{X}|\theta_i)}{\int h_i(t)f(\mathbf{X}|t)dt}, \end{aligned} \tag{6}$$

we see that unspecified constants cancel each other and the posterior density is proper if the integral in the denominator of (6) converges.

Consider now the situation of comparing two models, $M_1$ and $M_2$, for data $\mathbf{X}$. We assume improper prior for the parameters of model 1, so that $\pi_1(\theta_1) = c_1 \cdot h_1(\theta_1)$, with $c_1$ unspecified. Now the Bayes factor becomes

$$\begin{aligned} B_{12}(\mathbf{X}) &= \frac{p(\mathbf{X}|M_1)}{p(\mathbf{X}|M_2)} \\ &= c_1 \cdot \frac{\int h_1(\theta_1)f_1(\mathbf{X}|\theta_1)d\theta_1}{\int \pi_2(\theta_2)f_2(\mathbf{X}|\theta_2)d\theta_2}, \end{aligned} \tag{7}$$

we see that $c_1$ doesn't cancel and Bayes factor becomes unspecified. If we assume also $\pi_2$ to be improper, Bayes factor can be written as

$$\begin{aligned} B_{12}(\mathbf{X}) &= \frac{p(\mathbf{X}|M_1)}{p(\mathbf{X}|M_2)} \\ &= \frac{c_1}{c_2} \cdot \frac{\int h_1(\theta_1)f_1(\mathbf{X}|\theta_1)d\theta_1}{\int h_2(\theta_2)f_2(\mathbf{X}|\theta_2)d\theta_2}, \end{aligned} \tag{8}$$

which is also clearly unspecified due the dependency on the ratio $c_1/c_2$. One way to deal with this problem is based on the idea behind (6), where we noticed that with certain assumptions, the posterior density of parameters with improper prior is in fact proper. This notion leads us to partial Bayes factors.

## 4.3 Partial Bayes factors

Consider the same two model example as before and partition the data in two parts $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Our aim is to use $\mathbf{Y}$ as a training sample to get rid of the unspecified constants appearing in (7) and (8). The remaining data $\mathbf{Z}$ is then used to perform comparison between $M_1$ and $M_2$. We define partial Bayes factor as

$$B_{12}(\mathbf{Z}|\mathbf{Y}) = \frac{p(\mathbf{Z}|\mathbf{Y}, M_1)}{p(\mathbf{Z}|\mathbf{Y}, M_2)}, \tag{9}$$

where

$$p(\mathbf{Z}|\mathbf{Y}, M_i) = \int \pi_i(\theta_i|\mathbf{Y}) f(\mathbf{Z}|\mathbf{Y}, \theta_i) d\theta_i.$$

If we take $\pi_i(\theta_i)$ to be improper, we can use (6) to deduce that $\pi_i(\theta_i|\mathbf{Y})$ is a proper density given that the size of the training sample $\mathbf{Y}$ is sufficient. Using this approach, unspecified constants will not appear in (9) and the partial Bayes factor is well defined. Relationship between the full and the partial Bayes factor is also easily seen by

$$B_{12}(\mathbf{X}) = B_{12}(\mathbf{Y}) B_{12}(\mathbf{Z}|\mathbf{Y}).$$

Part of the problem is solved by using this approach, but partial Bayes factors are clearly dependent on the choice of the training sample $\mathbf{Y}$. One has to pick $\mathbf{Y}$ so that the integrals involved in updating the improper prior to a proper one converge, but otherwise the choice is arbitrary. If we denote the minimum sample size of $\mathbf{Y}$ by $m$, we have $\binom{n}{m}$ different possibilities to choose from.

One can show that the same asymptotic results mentioned in the treatment of the ordinary Bayes factors apply also here, if $m$ is held fixed. If we let $m$ to vary with $n$, then the ratio $n/m$ has to approach infinity as $n$ grows for consistency to be achieved.

To overcome the problems related to choosing $\mathbf{Y}$ one can compute partial Bayes factors for all of the possible data sets and average the results. This approach results in so called intrinsic Bayes factors, but it also comes with its own difficulties. Computing the partial Bayes factors for all possible training samples will quickly become inefficient if the $m$ is relatively large. Also the way one does the averaging has impact on the result. This motivation takes us to O'Hagan's solution, fractional Bayes factors.

## 4.4 Fractional Bayes factors

Denote $b = m/n$. We define the fractional Bayes factor as

$$B_{12}(\mathbf{X}; b) = \frac{p_b(\mathbf{X}|M_1)}{p_b(\mathbf{X}|M_2)}, \tag{10}$$

where the term $p_b(\mathbf{X}|M_i)$ denotes fractional marginal likelihood and can be written as

$$p_b(\mathbf{X}|M_i) = \frac{\int \pi_i(\theta_i) f_i(\mathbf{X}|\theta_i) d\theta_i}{\int \pi_i(\theta_i) f_i(\mathbf{X}|\theta_i)^b d\theta_i} \tag{11}$$

From (11) it is easy to see that any unspecified constants related to improper priors cancel, because they can be taken out of the integrals, which leaves fractional marginal likelihood well defined, provided that the integrals converge. The main idea behind the definition (11) is that if $m$ and $n$ and are large, then the likelihood based on the $m$ samples approximates the whole likelihood to the power $b$, that is

$$f_i(\mathbf{Y}|\theta_i) \approx f_i(\mathbf{X}|\theta_i)^b.$$

O'Hagan proposes the use of fractional Bayes factors also in cases when $m$ and $n$ are not large, even though the definition is justified asymptotically. Consistency of fractional Bayes factors can be proven assuming that $b \to 0$ as $n \to \infty$.

The fractional Bayes factor approach leaves the specification of the fraction $b$ open. This can be seen a far less critical problem than the arbitrary selection of the training fraction in partial Bayes factors. O'Hagan proposes three different choices for $b$. The most obvious choice is to pick $b = n_0/n$, where $n_0$ is the minimum sample size for everything to be well defined. Obvious choice means here that this selection of $b$ leaves as much data for model comparison as possible. If robustness is a great concern in sense of outliers, O'Hagan proposes taking $b = n^{-1}\max\{n_0, \sqrt{n}\}$. Last proposed value is to choose $b = n^{-1}\max\{n_0, \log n\}$, which represents an intermediate choice.

# 5 Objective comparison of Gaussian DAGs

The main goal of this chapter is to provide an analytical expression for the marginal likelihood of any Gaussian DAG model. We do this by following the work of Consonni and La Rocca (2012), who consider objective comparison of Gaussian Directed Acyclic Graphical models in their article [4]. Their approach to Gaussian DAG model comparison is based on using Bayes factors and uninformative, typically improper prior on the space of unconstrained covariance matrices. Concerns rising from using improper priors are dealt with utilizing the fractional marginal likelihood.

We first review a result concerning the computation of marginal likelihood in a more general setting, presented by Geiger and Heckerman [5]. They present 5 assumptions about the sampling distribution of data and the structure of the prior distribution for parameters, that allow one to construct parameter priors for every DAG model with given set of vertices just by specifying one parameter prior for any of the complete DAG models related to these given vertices.

This results in a convenient expression for the marginal likelihood, which is also used in Consonni's paper. Assumptions also guarantee that all Markov equivalent DAGs are scored equally, which is important property when DAGs are considered only as models of conditional independence instead of causality.

## 5.1 Marginal likelihood of a general DAG model

Let $X = (X_1, \ldots, X_p)^T$ denote a $p$-dimensional random vector and $M = (G, \mathcal{F}_s)$ be a DAG model defined over $X$. The DAG model $M$ is specified by the structure $G$, which is the set of conditional independences between the components of $X$ encoded in a graph. The notation $\mathcal{F}_G$ denotes the set of allowable local distribution families. $M_c$ denotes a complete DAG model, which is a model with no conditional independences. We use $\mathbf{x} = (x_1, \ldots, x_p)^T$ to denote a single observation of $X$ and boldface $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ to denote a complete random sample of size $n$. The following five assumptions are presented in [5]:

1. (Complete model equivalence). *Let $M_1 = (G_1, \mathcal{F}_{G_1})$ and $M_2 = (G_2, \mathcal{F}_{G_2})$ be complete DAG models for $X$. We assume, that $M_1$ and $M_2$ represent the same sets of joint probability distributions for $X$.*

2. (Regularity). *Let $M_1$ and $M_2$ be two complete DAG models for $X$. We assume that there exists a one-to-one mapping $f(\cdot)$ between the parameters $\theta_{m1}$ of $M_1$ and parameters $\theta_{m2}$ of $M_2$ such that the likelihoods satisfy $p(\boldsymbol{x}|\theta_{m1}, M_1) = p(\boldsymbol{x}|\theta_{m2}, M_2)$, where $\theta_{m2} = f(\theta_{m1})$. The Jacobian of the transformation is also assumed to exist and to be non-zero everywhere in the parameter space.*

3. (Likelihood modularity). *If variable $X_i$ has the same parents in models $M_1$ and $M_2$, then the local distributions for $x_i$ are the same in both models, that is, $p(x_i|pa(i), \theta_i, M_1) = p(x_i|pa(i), \theta_i, M_2)$.*

4. (Prior modularity). *If variable $X_i$ has same parents in models $M_1$ and $M_2$, then $p(\theta_i|M_1) = p(\theta_i|M_2)$.*

5. (Global parameter independence). *Let $M$ be arbitrary DAG model for $X$. Then $p(\theta|M) = \prod_{i=1}^{p} p(\theta_i|M)$ .*

Assumptions 1-3 consider the sampling the distribution of data. Geiger and Heckerman point out that one case where these three assumptions are satisfied, is if $X$ follows multivariate Normal distribution. One implication of assumption 1 is that any two complete DAG models have the same marginal likelihood for every dataset $\mathbf{X}$. This, in other words means that complete models cannot be distinguished based on the data. Geiger and Heckerman also state that in the multivariate Normal case with zero mean, both the assumptions 4 and 5 hold if and only if the prior on the precision matrix $\Omega$ is Wishart.

Using the assumptions above, the following propositions can be derived (Theorem 1 and 2 in [5])

**Theorem 5.1.1.** *Assume propositions 1-5. Now the parameter prior $p(\theta|M)$ for an arbitrary DAG model $M$ is determined by a parameter prior $p(\theta_c|M_c)$ for an arbitrary complete DAG model $M_c$.*

**Theorem 5.1.2.** *Let $M$ and $M_c$ be any DAG model and any complete DAG model for $X$, respectively. Let $\boldsymbol{X}$ denote a complete random sample. Assumptions 1-5 imply*

$$p(\boldsymbol{X}|M) = \prod_{i=1}^{p} \frac{p(\boldsymbol{X}_{fa(i)}|M_c)}{p(\boldsymbol{X}_{pa(i)}|M_c)}, \tag{12}$$

*where $X_{pa(i)}$ denotes the data belonging to $X_i$'s parents. Notation $fa(i) = pa(i) \cup i$ stands for the family of variable $X_i$.*

Consonni and Rocca make use of (12) to derive their formula for the marginal likelihood of the Gaussian DAGs. In order to apply (12), we need results concerning the marginal likelihoods of $X$'s subvectors when $X$ is multivariate Normal. The next sections follow quite strictly the treatment of the subject in Consonni's and Rocca's paper. We will start with preliminary results that are needed to go through the derivation.

## 5.2 Wishart distribution

The family of Wishart distributions is a conjugate prior family on the precision matrix of the multivariate Normal distribution. This allows us to get closed form solutions for the marginal likelihood integrals.

Let $X$ denote a $p$-dimensional random vector following multivariate normal distribution with zero mean and covariance matrix $\Omega^{-1} = \Sigma$,

$$X \sim N_p(\mathbf{0}, \Omega^{-1}).$$

We put no restrictions to elements of the precision matrix $\Omega$, other than requiring $\Omega$ to be symmetric and positive definite (s.p.d.).

Let $Y$ be a $p \times p$ symmetric and positive definite random matrix. We denote the set of all unconstrained s.p.d. matrices by $\mathcal{U}$. Write $Y \sim W_p(a, A)$ to say that $Y$ follows a Wishart distribution. The density of the Wishart distribution is given by

$$p^W(Y) = c(p,a)|A|^{\frac{a}{2}}|Y|^{\frac{a-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(YA)\right), \; Y \in \mathcal{U}, \tag{13}$$

and $p^W(Y) = 0$, when $Y \notin \mathcal{U}$. Here $A \in \mathcal{U}$ and $a \in \mathbb{R}$. For the density to be proper, $a > p - 1$ must hold. The term $c(p, a)$ denotes the normalising constant, and is defined as

$$\begin{aligned} c(p,a) &= \left(\int_{\mathcal{U}} |A|^{\frac{a}{2}}|Y|^{\frac{a-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(YA)\right) dY\right)^{-1} \\ &= \left(2^{ap/2}\pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{a+1-j}{2}\right)\right)^{-1}, \end{aligned} \tag{14}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

In order to obtain the marginal likelihood of any Gaussian DAG, we will need results concerning the marginal distributions of $X$'s subvectors when $X \sim N_p(\mathbf{0}, \Omega^{-1})$ and $\Omega \sim W_p(a, A)$.

Partition $X$ as $X = (X_v, X_w)$, where subvector dimensions are $p_v$ and $p_w$. Naturally, $p = p_v + p_w$. Partition the precision and the covariance matrix, $\Omega$ and $\Sigma$, respectively as

$$\Sigma = \begin{pmatrix} \Sigma_{vv} & \Sigma_{vw} \\ \Sigma_{wv} & \Sigma_{ww} \end{pmatrix}, \ \Omega = \begin{pmatrix} \Omega_{vv} & \Omega_{vw} \\ \Omega_{wv} & \Omega_{ww} \end{pmatrix}. \tag{15}$$

Using these, we have the following relationships between the blocks of matrices

$$\Sigma_{vv \cdot w} \equiv \text{var}(X_v | X_w) \equiv \Sigma_{vv} - \Sigma_{vw}\Sigma_{ww}^{-1}\Sigma_{wv} = (\Omega_{vv})^{-1}. \tag{16}$$

The quantity $\Sigma_{vv \cdot w}$ or $\text{var}(X_v|X_w)$ is called partial variance of $X_v$ given $X_w$. It is defined as a residual variance of $X_v$ after subtracting the variance based on the linear least squares predictor of $X_v$ from $X_w$. This equals also the conditional variance of $X_v$ given $X_w$ in our case.

Since $\Sigma = \Omega^{-1}$, we can switch the roles of $\Sigma$ and $\Omega$ in (16) to obtain

$$(\Sigma_{vv})^{-1} = \Omega_{vv} - \Omega_{vw}\Omega_{ww}^{-1}\Omega_{wv} \equiv \Omega_{vv \cdot w},$$

which can be further written as

$$(\Omega^{-1})_{vv} = (\Omega_{vv \cdot w})^{-1}, \tag{17}$$

since $\Sigma_{vv} = (\Omega^{-1})_{vv}$.

The following theorem (Theorem 2.1 in Consonni [4]) is of great use when we derive the marginal likelihood of the subvector $X_v$.

**Theorem 5.2.1.** *Assume $\Omega \sim W_p(a, A)$, where $A$ is s.p.d. matrix and $a > p - 1$. Use the partition in (15) for $\Omega$ and partition $A$ accordingly, then*

$$\Omega_{vv \cdot w} \sim W_{p_v}(a - p_w, A_{vv}) \tag{18}$$

*Proof.* We can use a theorem found in *Applied Multivariate Analysis* by Press [11] to write that, if $\Omega \sim \tilde{W}_p(a, A)$, then $\Omega_{vv \cdot w} \sim \tilde{W}_p(a - p_w, A_{vv \cdot w})$. We use tilde notation since Press defines Wishart distribution using $A^{-1}$ in place of $A$ in the formula for the density (13).

Now assume $\Omega \sim W_p(a, A)$. Using Press's parametrization, this is equivalent with $\Omega \sim \tilde{W}_p(a, (A)^{-1})$. Then the theorem referred above implies that $\Omega_{vv \cdot w} \sim \tilde{W}_p(a - p_w, (A^{-1})_{vv \cdot w})$, which in our notation means that $\Omega_{vv \cdot w} \sim W_p(a - p_w, ((A^{-1})_{vv \cdot w})^{-1})$. But using (17), one may conclude that $((A^{-1})_{vv \cdot w})^{-1} = A_{vv}$, which finishes the proof. □

## 5.3 Marginal likelihood with Wishart prior

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote an independent and identically distributed random sample obtained from $N_p(\mathbf{0}, \Omega^{-1})$. Denote $S = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$, which is the unscaled sample covariance matrix. Now the marginal likelihood can be computed

$$
\begin{aligned}
m(\mathbf{X}) &= \int_{\mathcal{U}} f(\mathbf{X}|\Omega) p^W(\Omega) d\Omega \\
&= \int_{\mathcal{U}} (2\pi)^{-\frac{np}{2}} |\Omega|^{\frac{n}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\Omega S)\right) c(p,a) |A|^{\frac{a}{2}} |\Omega|^{\frac{a-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\Omega A)\right) d\Omega,
\end{aligned}
$$

using the linearity of the trace operator, we get

$$
\begin{aligned}
m(\mathbf{X}) &= (2\pi)^{-\frac{np}{2}} c(p,a) |A|^{\frac{a}{2}} \int_{\mathcal{U}} |\Omega|^{\frac{n+a-p-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\Omega(S+A))\right) d\Omega \\
&= (2\pi)^{-\frac{np}{2}} \frac{c(p,a)}{c(p,a+n)} \frac{|A|^{\frac{a}{2}}}{|A+S|^{\frac{a+n}{2}}},
\end{aligned} \tag{19}
$$

where the last equality follows since the expression under the integral is proportional to a Wishart density with parameters $a + n$ and $S + A$, and thus integrates to $1/(c(p, a + n))|A + S|^{\frac{a+n}{2}}$. Note that $m(\mathbf{X})$ is a shorthand notation for $p(\mathbf{X}|M)$, where $M$ refers to the underlying model.

Recalling the definition (14), one can compute the ratio of normalising constants appearing in (19)

$$
\frac{c(p,a)}{c(p,a+n)} = 2^{\frac{np}{2}} \frac{\prod_{j=1}^p \Gamma\left(\frac{a+n+1-j}{2}\right)}{\prod_{j=1}^p \Gamma\left(\frac{a+1-j}{2}\right)}. \tag{20}
$$

Inserting this back into (19), gives

$$
m(\mathbf{X}) = (\pi)^{-\frac{np}{2}} \frac{\prod_{j=1}^p \Gamma\left(\frac{a+n+1-j}{2}\right)}{\prod_{j=1}^p \Gamma\left(\frac{a+1-j}{2}\right)} \frac{|A|^{\frac{a}{2}}}{|A+S|^{\frac{a+n}{2}}}. \tag{21}
$$

Consider next the marginal likelihood of a data related to a subvector of $X$, when $X \sim N_p(\mathbf{0}, \Omega^{-1})$ and $\Omega \sim W_p(a, A)$. We use the familiar partition $X = (X_v, X_w)$ and the corresponding partition for matrices defined in (15).

Now from the basic properties of multivariate Gaussians, it follows that $X_v \sim N_{p_v}(\mathbf{0}, (\Omega^{-1})_{vv})$. By using (17), we can also write that

$$X_v \sim N_{p_v}(\mathbf{0}, (\Omega_{vv \cdot w})^{-1}).$$

Then, **Theorem 5.2.1** tells us that $\Omega_{vv \cdot w} \sim W_{p_v}(a - p_w, A_{vv})$. But now the situation is essentially the same as in the deriving the marginal likelihood for the full vector. Thus, we get the marginal data density of $\mathbf{X}_v$ just by substituting

$$p \to p_v, \ a \to a - p_w, \ A \to A_{vv} \text{ and } S \to S_{vv}$$

into (21), which results in the following expression

$$m(\mathbf{X}_v) = (\pi)^{-\frac{np_v}{2}} \frac{\prod_{j=1}^{p_v} \Gamma\left(\frac{a-p_w+n+1-j}{2}\right)}{\prod_{j=1}^{p_v} \Gamma\left(\frac{a-p_w+1-j}{2}\right)} \frac{|A_{vv}|^{\frac{a-p_w}{2}}}{|A_{vv} + S_{vv}|^{\frac{a-p_w+n}{2}}}. \tag{22}$$

## 5.4 Exponential family setting

Consonni and La Rocca show that the expressions obtained for the marginal likelihoods (21) and (22), can be also derived in a more general context using exponential families paired with conjugate priors. We review this approach here, since it also gives us a simple way to compute the fractional marginal likelihood in closed form using improper priors. More thorough discussion on exponential families can be found in [8].

We say that the sampling distribution of data $y$ belongs to an *exponential family*, if the density can be written as

$$f(y|\theta) = h_n(y) \exp\{\langle \theta, s \rangle - nM(\theta)\}, \ y \in \mathcal{Y}, \tag{23}$$

where $n$ is sample size, $s = s(y)$ is the canonical statistic belonging to real vector space with inner product $\langle \cdot, \cdot \rangle$, $\theta$ denotes the canonical parameter and $\exp\{-nM(\theta)\}$ is the normalising constant for each given $\theta$. The leading factor is a product of base measures $h_n(y) = \prod_{i=1}^{n} h(y_i)$ not absorbed into the dominating measure. The dominating measure would be a product of Lebesgue measures in case of multivariate Gaussian data.

Suppose we have an i.i.d sample of multivariate Normal data, $\mathbf{x}_i \sim N_p(\mathbf{0}, \Omega^{-1}),\ i = 1, \ldots, n$. Now the sampling density can be written in the form of (23) by using the following notation

1. $s = -\dfrac{S}{2}$, where $S = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$.
2. The inner product $\langle A, B \rangle$ is the trace, $\mathrm{tr}(A^T B)$, where $A$ and $B$ are $p \times p$ real matrices.
3. Canonical parameter $\theta$ is the precision matrix $\Omega$ and $M(\theta) = -(1/2) \log |\Omega|$.
4. $h_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) = (2\pi)^{-\frac{np}{2}}$.

In the above setting, the family of conjugate prior densities on $\theta$ have the form

$$p^C(\theta) = K(n_\bullet, s_\bullet) \exp\{\langle \theta, s_\bullet \rangle - n_\bullet M(\theta)\} \tag{24}$$

with respect to Lebesgue measure. Here $s_\bullet$ is prior guess for $s$, $n_\bullet$ is the prior sample size and $K(n_\bullet, s_\bullet)$ is the normalizing constant, if it exists. The Wishart prior used in the last section is of the form (24), when we denote

1. $n_\bullet = a - p - 1$.
2. $s_\bullet = -\frac{A}{2}$.
3. The normalizing constant $K(n_\bullet, s_\bullet) = c(p, a)|A|^{a/2}$.

Pairing the sampling distribution (23) with the conjugate prior (24) leads to a posterior density of a form

$$p^C(\theta|y) = K(n_\bullet + n, s_\bullet + s) \exp\{\langle \theta, s_\bullet + s \rangle - (n_\bullet + n) M(\theta)\}. \tag{25}$$

Now since posterior density can be written as

$$\begin{aligned} p^C(\theta|y) &= \frac{p^C(\theta) f(y|\theta)}{\displaystyle\int_\Theta p^C(\theta) f(y|\theta) d\theta} \\ &= \frac{p^C(\theta) f(y|\theta)}{m^C(y)}, \end{aligned}$$

where $m^C(y)$ is the marginal likelihood of $y$, we can solve for $m^C(y)$ and obtain the following expression for the marginal likelihood

$$\begin{aligned} m^C(y) &= \frac{p^C(\theta) f(y|\theta)}{p^C(\theta|y)} \\ &= h_n(y) \frac{K(n_\bullet, s_\bullet)}{K(n_\bullet + n, s_\bullet + s)}, \end{aligned} \tag{26}$$

where terms involving exponential functions have cancelled. It is straightforward to check that the marginal likelihood for full data (21) derived in the last section can be obtained from (26). The general form (26) can be also used to derive the expression for the marginal likelihood of the subvector by using the **Theorem 5.2.1**.

## 5.5 Fractional marginal likelihood in general setting

Model comparison by using fractional Bayes factors (FBF) was more carefully discussed in the previous chapter. Essentially it boils down to computing a fractional marginal likelihood, since FBF is defined as a ratio of them. Let $b = n_0/n$ be the training fraction. We use notation $m(y, n_0)$ to denote the $b$-fractional marginal likelihood of $y$ under some model $M$, which is given according to

$$m(y, n_0) = \frac{\int f(y|\theta) p^D(\theta) d\theta}{\int f(y|\theta)^b p^D(\theta) d\theta}, \tag{27}$$

where $f(y|\theta)$ and $p^D(\theta)$ are the sampling density of data and the default parameter prior under $M$, respectively.

We can rewrite the expression (27) as a

$$m(y, n_0) = \int f(y|\theta)^{(1-b)} p^F(\theta) d\theta, \tag{28}$$

where $p^F(\theta)$ is the fractional prior, $p^F(\theta) \propto f(y|\theta)^b p^D(\theta)$. The fractional prior is a posterior obtained by updating the possibly improper prior $p^D(\theta)$ to a proper one by sacrificing a $b$-fraction of likelihood.

Our next goal is to provide an expression for (28) using the exponential family setting. To that end, assume that the sampling density belongs to an exponential family and the default prior has the form

$$p^D(\theta) \propto \exp\{\langle \theta, s_\bullet^D \rangle - n_\bullet^D M(\theta)\}, \tag{29}$$

where $s_\bullet^D$ and $n_\bullet^D$ are allowed to be chosen such that $p^D(\theta)$ is improper. The $b$-fractional likelihood can be written using (23) as

$$\begin{aligned} f(y|\theta)^b &= h_n(y)^{\frac{n_0}{n}} \exp\{\langle \theta, n_0 \bar{s} \rangle - n_0 M(\theta)\} \\ &= \bar{h}^{n_0} \exp\{\langle \theta, n_0 \bar{s} \rangle - n_0 M(\theta)\}, \end{aligned} \tag{30}$$

where we have used $\bar{h}$ to denote the geometric mean of data set $y$ and $\bar{s}$ to denote $s/n$. By comparing (30) to the general expression of the sampling density in exponential family setting (23), we see that the $b$-fractional likelihood can viewed as an ordinary likelihood based on $n_0$ observations with a leading factor $\bar{h}^{n_0}$ and a canonical statistic $n_0 \bar{s}$.

Similar reasoning applies to $f(y|\theta)^{(1-b)}$ in (28), since it can be written as

$$f(y|\theta)^{(1-b)} = \bar{h}^{n-n_0} \exp\{\langle \theta, (n-n_0)\bar{s} \rangle - (n-n_0) M(\theta)\}, \tag{31}$$

which implies that $f(y|\theta)^{(1-b)}$ can be seen as a likelihood based on $n - n_0$ observations, canonical statistic $(n - n_0)\bar{s}$ and a leading factor $\bar{h}^{n-n_0}$.

Now, since the fractional prior $p^F(\theta)$ is actually a posterior based on the likelihood (31) and the prior (29), we can use the formula (25) to write

$$p^F(\theta) \propto \exp\{\langle \theta, n_0 \bar{s} + s_\bullet^D \rangle - (n_0 + n_\bullet^D) M(\theta)\}, \tag{32}$$

where we assume that the training fraction $n_0$ is chosen such as (32) defines a proper density.

We have now seen that the fractional likelihood corresponds to a certain actual likelihood and the fractional prior is of the conjugate form. Thus, the fractional marginal likelihood

$$m(y, n_0) = \int f(y|\theta)^{(1-b)} p^F(\theta) d\theta,$$

can be computed utilizing the result (26), which leads to the expression

$$m(y, n_0) = \bar{h}^{n-n_0} \frac{K(n_0 + n_\bullet^D, n_0 \bar{s} + s_\bullet^D)}{K(n + n_\bullet^D, s_\bullet^D + s)} \tag{33}$$

All in all, we have shown that the fractional marginal likelihood corresponds to an ordinary conjugate marginal likelihood based on the reduced sample size $n - n_0$, the canonical statistic $\bar{s}$ computed from full data and the conjugate prior that depends on $n_0$ and $\bar{s}$. This means that the fractional marginal likelihood can be computed analytically in the Gaussian case using the results obtained at the beginning of this chapter.

## 5.6 Fractional marginal likelihood for Gaussian distributions

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be complete i.i.d sample from $N_p(\mathbf{0}, \Omega^{-1})$. Denote the unscaled empirical covariance matrix by $S = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ and $\bar{S} = S/n$. We select

$$p^D(\Omega) \propto |\Omega|^{\frac{a_\Omega - p - 1}{2}} \tag{34}$$

to be our default improper prior. Remember, that the default prior (29) used in the previous section had the form

$$p^D(\theta) \propto \exp\{\langle \theta, s_\bullet^D \rangle - n_\bullet^D M(\theta)\}.$$

By selecting $\theta = \Omega$, $s_\bullet^D = 0$, $n_\bullet^D = a_\Omega - p - 1$ and $M(\Omega) = -(1/2)\log|\Omega|$ in the formula above, we obtain (34). The updated fractional prior had the conjugate form

$$p^F(\theta) \propto \exp\{\langle \theta, n_0 \bar{s} + s_\bullet^D \rangle - (n_0 + n_\bullet^D) M(\theta)\}.$$

Specializing above to a multivariate Normal case with our default prior (34), leads to an expression

$$p^F(\Omega) \propto \exp\{\langle \Omega, -n_0 \frac{\bar{S}}{2} \rangle - (a_\Omega + n_0 - p - 1) M(\Omega)\} \tag{35}$$

which in this case shows, that the fractional prior for $\Omega$ is a Wishart distribution $W_p(a_\Omega + n_0, n_0 \bar{S})$. The Wishart density is proper if $a_\Omega + n_0 > p - 1$. If we select $a_\Omega = p - 1$, then the default prior has the improper form

$$p^D(\Omega) \propto |\Omega|^{-1} \tag{36}$$

and the corresponding minimum sample size to update this to a proper prior is $n_0 = 1$.

Consider now the previously used partition of $X = (X_v, X_w)$ and the fractional marginal likelihood corresponding to data related to variables in $X_v$. We have shown that the situation is now essentially the same as it was in deriving the expression (22). More specifically, we have a Gaussian likelihood based on $n - n_0$ observations and a conjugate prior $W_p(a_\Omega + n_0, n_0 \bar{S})$. Thus, by making the following substitutions

$$a \to a_\Omega + n_0, \quad A \to n_0 \bar{S}, \quad n \to (n - n_0), \quad S \to (n - n_0)\bar{S},$$

into (22), we obtain the fractional marginal likelihood related to data $\mathbf{X}_v$. This results in the following formula (the equation (22) of Consonni)

$$m(\mathbf{X}_v, n_0) = (\pi)^{-\frac{(n-n_0)p_v}{2}} \frac{\prod_{j=1}^{p_v} \Gamma\left(\frac{a_\Omega - p_w + n + 1 - j}{2}\right)}{\prod_{j=1}^{p_v} \Gamma\left(\frac{a_\Omega - p_w + n_0 + 1 - j}{2}\right)} \left(\frac{n_0}{n}\right)^{\frac{(a_\Omega - p_w + n_0)p_v}{2}} |S_{vv}|^{-\frac{n-n_0}{2}}. \tag{37}$$

The expression for the fractional marginal likelihood (37) is well defined if $S_{vv}$ is positive definite, which requires that $n \geq p_v$.

## 5.7 Marginal likelihood of any Gaussian DAG

Recall, that the general formula for the marginal likelihood of any DAG model $M$ defined over variables $X = (X_1, \ldots, X_p)^T$ was given according to (12) as

$$p(\mathbf{X}|M) = \prod_{j=1}^{p} \frac{p(\mathbf{X}_{fa(j)}|M_c)}{p(\mathbf{X}_{pa(j)}|M_c)},$$

where $\mathbf{X}_{pa(j)}$ and $\mathbf{X}_{fa(j)}$ and denote the data related to the parents and the family of node $j$, respectively.

We have now all the results needed to compute (12) in closed form. More in detail, take any complete Gaussian DAG $M_c$ and let the prior for parameters $\Omega$ be $W_p(a_\Omega + n_0, n_0\bar{S})$, which is the fractional prior obtained in the last section. Pair the fractional prior with ordinary Gaussian likelihood based on $n - n_0$ observations and a canonical statistic $\bar{S}$. Denote the size of the set $pa(j)$ by $p_j$. Then, the term $p(\mathbf{X}_{pa(j)}|M_c)$ is given by (37), when we do the following substitutions

$$v \to pa(j),\ p_v \to p_j,\ p_w = p - p_j.$$

This results in

$$\begin{aligned} p(\mathbf{X}_{pa(j)}|M_c) =& (\pi)^{-\frac{(n-n_0)p_j}{2}} \frac{\prod_{i=1}^{p_j} \Gamma\left(\frac{a_\Omega + n - p + p_j + 1 - i}{2}\right)}{\prod_{i=1}^{p_j} \Gamma\left(\frac{a_\Omega + n_0 - p + p_j + 1 - i}{2}\right)} \\ & \cdot \left(\frac{n_0}{n}\right)^{\frac{(a_\Omega + n_0 - p + p_j)p_j}{2}} |S_{pa(j)}|^{-\frac{n-n_0}{2}}, \end{aligned} \tag{38}$$

where we have used notation $S_{pa(j)}$ to mean $S_{pa(j)pa(j)}$, which refers to the unscaled sample covariance matrix of variables in $pa(j)$.

By choosing $a_\Omega = p - 1$, which corresponds to the improper prior of a type (36), the minimum sample size is $n_0 = 1$ and (38) becomes

$$p(\mathbf{X}_{pa(j)}|M_c) = \pi^{-\frac{(n-1)p_j}{2}} \frac{\prod_{i=1}^{p_j} \Gamma\left(\frac{n+p_j-i}{2}\right)}{\prod_{i=1}^{p_j} \Gamma\left(\frac{p_j+1-i}{2}\right)} n^{-\frac{p_j^2}{2}} |S_{pa(j)}|^{-\frac{n-1}{2}}. \tag{39}$$

The numerator $p(\mathbf{X}_{fa(j)}|M_c)$ of (12) can easily be written using (38). We need only to make substitutions $pa(j) \to fa(j)$ and $p_j \to p_j + 1$, since the set $fa(j)$ is always one element larger than the set $pa(j)$. This allows us to write

$$p(\mathbf{X}_{fa(j)}|M_c) = \pi^{-\frac{(n-1)(p_j+1)}{2}} \frac{\prod_{i=1}^{p_j+1} \Gamma\left(\frac{n+p_j+1-i}{2}\right)}{\prod_{i=1}^{p_j+1} \Gamma\left(\frac{p_j+2-i}{2}\right)} n^{-\frac{(p_j+1)^2}{2}} |S_{fa(j)}|^{-\frac{n-1}{2}}. \tag{40}$$

In order to obtain the final formula for the marginal likelihood of any Gaussian DAG, we need to compute the ratio $p(\mathbf{X}_{fa(j)}|M_c)/p(\mathbf{X}_{pa(j)}|M_c)$. This is not done explicitly in Consonni, but it is straightforward to see that Gamma functions cancel each other, except terms corresponding to $i = 1$ in the $p(\mathbf{X}_{fa(j)}|M_c)$. Thus, we'll get the following form for the local marginal likelihood

$$\frac{p(\mathbf{X}_{fa(j)}|M_c)}{p(\mathbf{X}_{pa(j)}|M_c)} = \pi^{-\frac{(n-1)}{2}} \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{p_j+1}{2}\right)} n^{-\frac{2p_j+1}{2}} \left(\frac{|S_{fa(j)}|}{|S_{pa(j)}|}\right)^{-\frac{n-1}{2}}. \tag{41}$$

And finally, for any Gaussian DAG $M$ over $X$, the marginal likelihood is given according to

$$p(\mathbf{X}|M) = \prod_{j=1}^{p} \pi^{-\frac{(n-1)}{2}} \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{p_j+1}{2}\right)} n^{-\frac{2p_j+1}{2}} \left(\frac{|S_{fa(j)}|}{|S_{pa(j)}|}\right)^{-\frac{n-1}{2}}, \tag{42}$$

where $S_{pa(j)}$ and $S_{fa(j)}$ have to be positive definite for every $j$.

# 6 Structure learning of Gaussian graphical models

This chapter discusses the problem of learning graphical models from multivariate Normal data. We adopt a score-based approach to learning, where each candidate graph can be given a score. Given enough data, our scoring function should assign the highest score to the true graph.

We present a fractional marginal pseudo-likelihood (FMPL), which will be used as our scoring function for learning Gaussian graphical models. Important property of this scoring function is the node-wise factorization, which makes it applicable also to a high-dimensional problems.

At the end of the chapter, we briefly present Graphical LASSO (GLASSO), a common method for estimating the inverse covariance matrices related to GGM's. GLASSO will be used in the next chapter, where we test the performance of FMPL numerically, and compare it to GLASSO.

## 6.1 Marginal likelihood

Consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and $E \subset V \times V$ is the set of edges. Let $X = (X_1, \ldots, X_p)^T$ be a $p$-dimensional Normal vector and by $\Omega$ denote the precision matrix of $X$. We assume that $\Omega$ is positive definite and for every $i \neq j$ we have, that $\Omega_{ij} = 0$, if and only if there is no edge between nodes $i$ and $j$. Assume zero mean for $X$ and denote $\Omega^{-1} = \Sigma$, so $X \sim N_p(\mathbf{0}, \Sigma)$.

Suppose we have a sample of independent and identically distributed multivariate Normal data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, coming from a distribution whose independence structure can be represented by a graph $G^*$. We would like to identify $G^*$ based on $\mathbf{X}$. This is done with a Bayesian approach by computing posterior probabilities of different graphs and picking the one maximizing it.

Posterior probability of a graph $G$ given data $\mathbf{X}$ is proportional to

$$p(G|\mathbf{X}) \propto p(G)p(\mathbf{X}|G), \tag{43}$$

where $p(G)$ is prior probability assigned to a specific graph and $p(\mathbf{X}|G)$ is the marginal likelihood. We leave the normalizing constant of posterior out, since it is the same for all the graphs and therefore cancels when doing comparison. At this stage, we are only interest in the marginal likelihood, since it is the

data dependent term in (43). Later on, we will make use of the prior $p(G)$ term in order to promote sparseness in graph structures.

We recall that the marginal likelihood measures the fit of a model to data after the effect of parameters has been taken out. By the definition, the marginal likelihood of $\mathbf{X}$ under $G$ is

$$p(\mathbf{X}|G) = \int_{\Theta_G} p(\theta|G)\ell(\theta|\mathbf{X}, G)d\theta, \tag{44}$$

where $p(\theta|G)$ denotes the parameter prior under $G$, the term $\ell(\theta|\mathbf{X}, G)$ is the likelihood function and the integral is taken over the set of all possible parameters under $G$.

However, computing the marginal likelihood for a general undirected graph is very difficult, due the global normalizing constant in the likelihood term. Closed form solution exists only for chordal graphs, which might be a too restrictive assumption in general [8].

## 6.2 Marginal pseudo-likelihood

We circumvent the problems involved in the true likelihood function by using *pseudo-likelihood*. Pseudo-likelihood was introduced for the first time in Besag's article (1972) [2]. The idea behind the pseudo-likelihood is to approximate the true likelihood by a product of conditional probabilities or densities, where in each factor variable is conditioned on all the rest. More formally, consider a likelihood of a single observation $\mathbf{x} = (x_1, \ldots, x_p)^T$. We use the chain rule of probability to write

$$p(\mathbf{x}|\theta) = \prod_{j=1}^{p} p(x_j|x_1, \ldots, x_{j-1}, \theta).$$

For every $j$, denote the remaining variables $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ by $\mathbf{x}_{-j}$. We use an approximation $p(x_j|x_1, \ldots, x_{j-1}) \approx p(x_j|\mathbf{x}_{-j})$, which lets us to write the pseudo-likelihood $\hat{p}(\mathbf{x}|\theta)$ as

$$\hat{p}(\mathbf{x}|\theta) = \prod_{j=1}^{p} p(x_j|\mathbf{x}_{-j}, \theta).$$

In general, pseudo-likelihood should not be considered as a numerically exact and accurate approximation of the likelihood but as an object that has

computationally more attractive form and which can be used to obtain consistent estimates of parameters. It can be shown that with certain regularity assumptions, the pseudo-likelihood estimate for model parameters coincides with the maximum likelihood estimator [8].

One advantage of using pseudo-likelihood instead of the true likelihood is that it allows us to replace the global normalization constant by $p$ local normalising constants related to conditional distributions of variables and thus makes the computations more tractable.

With help of pseudo-likelihood, our original problem (44) of computing the marginal likelihood can be written as

$$\begin{aligned} p(\mathbf{X}|G) &= \int_{\Theta_G} p(\theta|G)\ell(\theta|\mathbf{X}, G)d\theta \\ &\approx \int_{\Theta_G} p(\theta|G) \prod_{j=1}^{p} p(\mathbf{X}_j|\mathbf{X}_{-j}, \theta, G)d\theta \\ &= \hat{p}(\mathbf{X}|G) \end{aligned} \tag{45}$$

Term $\hat{p}(\mathbf{X}|G)$ is referred as *the marginal pseudo-likelihood* (MPL), introduced by Pensar et al in [10]. The global Markov property states that variable $X_j$ is conditionally independent of the rest given the variables in its Markov blanket $mb(j)$. More formally, we have that

$$p(X_j|X_{-j}, \theta) = p(X_j|X_{mb(j)}, \theta).$$

Thus, we get the following form for the marginal pseudo-likelihood

$$\begin{aligned} \hat{p}(\mathbf{X}|G) &= \int_{\Theta_G} p(\theta|G) \prod_{j=1}^{p} p(\mathbf{X}_j|\mathbf{X}_{-j}, \theta, G)d\theta \\ &= \int_{\Theta_G} p(\theta|G) \prod_{j=1}^{p} p(\mathbf{X}_j|\mathbf{X}_{mb(j)}, \theta)d\theta \end{aligned} \tag{46}$$

Assume global parameter independence, which means that the parameter prior factorizes according to

$$p(\theta|G) = \prod_{j=1}^{p} p(\theta_j).$$

This allows us to factor the MPL integral into integrals over individual parameter sets $\Theta_j$ related to conditional distributions $p(X_j|X_{mb(j)})$. The MPL integral (46) becomes

$$\hat{p}(\mathbf{X}|G) = \prod_{j=1}^{p} \int_{\Theta_j} p(\theta_j) p(\mathbf{X}_j|\mathbf{X}_{mb(j)}, \theta_j) d\theta_j. \tag{47}$$

## 6.3 Fractional marginal pseudo-likelihood

One can think the expression (47) for the MPL as a product of terms, where each term corresponds to a marginal likelihood of a DAG model. The idea is visualized in **Figure 2**. This approach offers an interesting way to compute MPL in closed form.

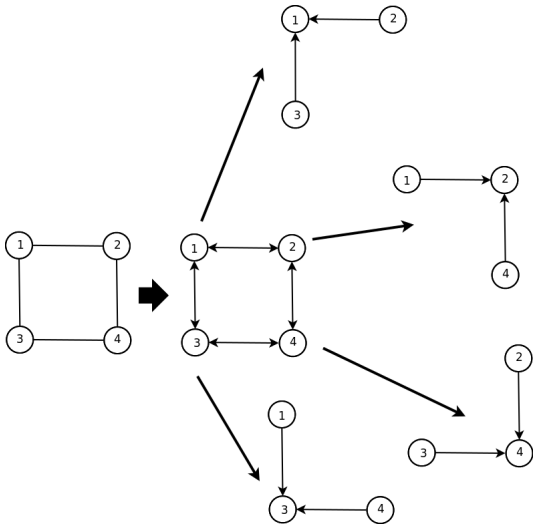



Figure 2: The factorization of a Markov network into simpler components with pseudo-likelihood.

Recall the general formula for a marginal likelihood of any DAG model $M$, used in the previous chapter:

$$p(\mathbf{Y}|M) = \prod_{j=1}^{p} \frac{p(\mathbf{Y}_{fa(j)}|M_c)}{p(\mathbf{Y}_{pa(j)}|M_c)},$$

It is useful to rewrite each factor in the expression above in a different form. Since $fa(j) = \{j\} \cup pa(j)$, we have

$$p(\mathbf{Y}_{fa(j)}|M_c) = p(\mathbf{Y}_j|\mathbf{Y}_{pa(j)}, M_c)p(\mathbf{Y}_{pa(j)}|M_c),$$

which allows us to write

$$p(\mathbf{Y}|M) = \prod_{j=1}^{p} p(\mathbf{Y}_j|\mathbf{Y}_{pa(j)}, M_c). \tag{48}$$

We can see the clear resemblance between the forms (48) and (47). In both of these, each factor corresponds to a marginal likelihood of a DAG model, where we have a node and its parent nodes. In the case of Markov networks, the set of node's parents is its Markov blanket, $mb(j)$.

Thus, we can use the closed form solution of (48) to compute the marginal pseudo-likelihood (47) we are after, just by changing $pa(j) \to mb(j)$ and defining $fa(j) = \{j\} \cup mb(j)$. Then the closed form solution (42) gives

$$\begin{aligned}\hat{p}(\mathbf{X}|G) &= \prod_{j=1}^{p} \pi^{-\frac{(n-1)}{2}} \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{p_j+1}{2}\right)} n^{-\frac{2p_j+1}{2}} \left(\frac{|S_{fa(j)}|}{|S_{mb(j)}|}\right)^{-\frac{n-1}{2}} \\ &\equiv \prod_{j=1}^{p} p(\mathbf{X}_j|\mathbf{X}_{mb(j)}),\end{aligned} \tag{49}$$

where $p_j = |mb(j)|$ and $S$ refers to the full $p \times p$ unscaled sample covariance matrix. As before, $S_{mb(j)}$ and $S_{fa(j)}$ refer to submatrices of $S$ restricted to variables in sets $mb(j)$ and $fa(j)$. From now on, $\hat{p}(\mathbf{X}|G)$ is referred as a *fractional marginal pseudo-likelihood* (FMPL), due the fractional Bayes factor approach used to derive the analytical form.

Derivation of (49) has been somewhat heuristic, but the next theorem provides justification for the steps taken and puts FMPL on a firmer ground. It also ultimately allows us to use FMPL efficiently as a scoring function for learning undirected graphical models from data.

**Theorem 6.3.1.** *Let $X \sim N_p(\boldsymbol{0}, (\Omega^*)^{-1})$ and $G^* = (V, E^*)$ denote the the undirected graph that completely determines the conditional independence statements between $X$'s components. Let $\{mb^*(1), \ldots, mb^*(p)\}$ denote the set of Markov blankets, which uniquely define $G^*$.*

*Suppose we have a complete random sample* $\boldsymbol{X}$ *of a size* $n$ *obtained from* $N_p(\boldsymbol{0}, (\Omega^*)^{-1})$. *Then the local FMPL estimator*

$$\widehat{mb}(j) = \underset{mb(j)\subset V\setminus\{j\}}{\arg\max}\ p(\boldsymbol{X}_j|\boldsymbol{X}_{mb(j)})$$

*is consistent, that is,* $\widehat{mb}(j) = mb^*(j)$ *with a probability tending to* 1, *as* $n \to \infty$.

*Proof.* The proof is presented in the appendix chapter. □

**Corollary 6.3.2.** *Let* $\mathcal{G}$ *denote the set of all undirected graphs with* $p$ *nodes. Now the global FMPL estimator*

$$\widehat{G} = \underset{G\in\mathcal{G}}{\arg\max}\ p(\boldsymbol{X}|G)$$

*is consistent, that is,* $\widehat{G} = G^*$ *with a probability tending to* 1, *as* $n \to \infty$.

*Proof.* **Theorem 6.3.1** guarantees, that each node's true Markov blanket is eventually found with probability tending to 1, as sample size increases. Since the structure of Markov network is uniquely determined by its Markov blankets, the result follows. □

## 6.4 Search algorithm for graph learning

Since FMPL is a consistent scoring function, we could in theory consider all the undirected graphs with $p$ nodes and score them. Given enough data, the true graph would eventually be identified. However, this approach is utterly doomed in practice if $p$ is even moderately large. To illustrate the inefficiency, assume that the number of nodes in the graph is $p$. Then there are $2^{\frac{p(p-1)}{2}}$ possible undirected graphs to consider.

We make use of the node-wise factorisation of FMPL-score and consistency of local estimators to obtain a more efficient way to learn the graph. We follow the approach used by Pensar et al [10], and the resulting algorithm is exactly the same as presented there.

Our problem of learning the graph from data $\mathbf{X}$, can be formulated as

$$\underset{G\in\mathcal{G}}{\arg\max}\ \hat{p}(\mathbf{X}|G)p(G), \tag{50}$$

where $\mathcal{G}$ is the set of undirected graphs with $p$ nodes. For now, we assume each graph equally probable *a priori*, so the prior term $p(G)$ can be ignored. Using the equation (49) and the consistency of local Markov blanket estimators, our problem can be factorised to independent sub-problems according to

$$\underset{mb(j)\subset V\setminus\{j\}}{\arg\max}\ p(\mathbf{X}_j|\mathbf{X}_{mb(j)}), \tag{51}$$

where $j = 1, \ldots, p$. We emphasize that the local consistency is the result that ultimately allows this approach. This factorisation to independent sub-problems is also the key that allows this method to be used when the number of variables grows large.

However, since consistency is an asymptotic result, this procedure might not produce consistent Markov blankets on small sample sizes. By consistent Markov blankets, we mean that if the node $k$ belongs to $l$'s Markov blanket, then $l$ has to be also part of $k$'s blanket. A method to overcome this will be presented soon, but we will first go through the search algorithm for Markov blanket discovery more in detail.

The search algorithm (Algorithm 1 in [10]) uses two operations, *add* and *delete*, to find the score optimal Markov blanket for each node in the graph. Given node $j$, we'll start with an empty set $\emptyset$ as a candidate Markov blanket $mb(j)$. The set of possible Markov blanket members for the given node $j$ is denoted by $C = V \setminus \{j\}$. At each step, the element of $C$ yielding the greatest improvement in local score $p(\mathbf{X}_j|\mathbf{X}_{mb(j)})$, is added to $mb(j)$. When the size of $mb(j)$ becomes larger than two, algorithm moves to a deletion phase.

In the deletion phase, we delete nodes from $mb(j)$ until there is no improvement in local score or $mb(j)$'s size is smaller than three. The algorithm terminates and returns the score optimal Markov blanket, when we cannot any more find a node, whose addition to $mb(j)$ would increase the local score. The described algorithm is presented more in detail using pseudo-code in **Algorithm 1**.

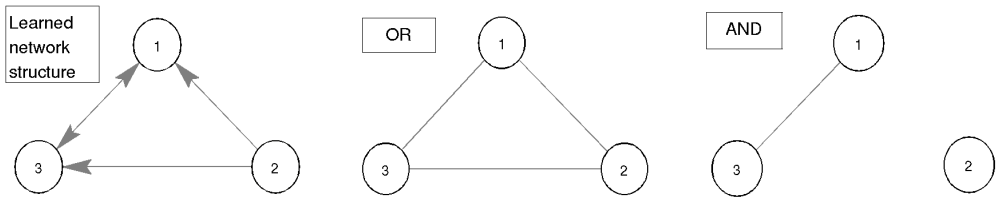
As mentioned earlier, the discovered Markov blankets might not be consistent. We present $AND$ and $OR$ methods that are able to cope with this problem and combine the found Markov blankets into an undirected graph. Consider a simple example presented in **Figure 3** to illustrate the methods.

The lefter most picture in **Figure 3** represents an example of learned network structure in case of a three node graph. We have found Markov blankets $mb(1) = \{2, 3\}$, $mb(2) = \emptyset$ and $mb(3) = \{1, 2\}$. These are combined

**Algorithm 1** Markov blanket discovery for node $j$

$mb(j) \leftarrow \emptyset$
$C \leftarrow V \setminus \{j\}$
$Add \leftarrow True$
**while** $Add = True$ **do**
    $mb^*(j) \leftarrow mb(j)$
    **for** $i \in C$ **do**
        **if** $p(\mathbf{X}_j|\mathbf{X}_{mb(j)\cup\{i\}}) > p(\mathbf{X}_j|\mathbf{X}_{mb^*(j)})$ **then**
            $mb^*(j) \leftarrow mb(j) \cup \{i\}$
        **end if**
    **end for**
    **if** $mb^*(j) \neq mb(j)$ **then**
        $mb(j) \leftarrow mb^*(j)$
        $C \leftarrow C \setminus mb(j)$
        $Add \leftarrow True$
        $Del \leftarrow True$
    **else**
        $Add \leftarrow False$
        $Del \leftarrow False$
    **end if**
    **while** $Del = True$ & $\text{size}(mb(j)) > 2$ **do**
        $mb^*(j) \leftarrow mb(j)$
        **for** $i \in mb(j)$ **do**
            **if** $p(\mathbf{X}_j|\mathbf{X}_{mb(j)\setminus\{i\}}) > p(\mathbf{X}_j|\mathbf{X}_{mb^*(j)})$ **then**
                $mb^*(j) \leftarrow mb(j) \setminus \{i\}$
            **end if**
        **end for**
        **if** $mb^*(j) \neq mb(j)$ **then**
            $mb(j) \leftarrow mb^*(j)$
            $Del \leftarrow True$
        **else**
            $Del \leftarrow False$
        **end if**
    **end while**
**end while**
**return** $mb(j)$

Figure 3: $AND$- and $OR$-method




to an edge set specifying an undirected graph by using $OR$-method

$$E_{OR} = \{(i,j) \subset E |\ i \in mb(j) \text{ or } j \in mb(i)\}$$

or AND-method

$$E_{AND} = \{(i,j) \subset E |\ i \in mb(j) \text{ and } j \in mb(i)\}.$$

In case of our example, $E_{OR} = \{(1,2),(2,1),(1,3),(3,1),(2,3),(3,2)\}$ and $E_{AND} = \{(1,3),(3,1)\}$. The corresponding graphs are presented in **Figure** 3.

## 6.5 Prior over local graphs

Until now we have assumed that every graph structure is *a priori* equally likely and thus the prior term $p(G)$ in (50) was ignored. Consider next a situation, where we would like to learn sparser graphs. This is done by using the prior term $p(G)$ to penalize nodes for having too many elements in their Markov blankets. We make an assumption, that the prior also factorizes node-wise, and thus the local score can be written as

$$p(G_j)p(\mathbf{X}_j|\mathbf{X}_{mb(j)}). \tag{52}$$

We use a similar approach as used in [3] to assign prior probabilities over local graph structures. In this approach, we imagine that the inclusion of an edge in a graph happens with some unknown probability $t$, which corresponds to a successful Bernoulli trial. A finite sequence of these inclusions is a repeated Bernoulli trial and thus binomially distributed. We obtain the following form for the local prior

$$p(G_j) \propto t^{p_j}(1-t)^{m-p_j}, \tag{53}$$

where $p_j$ is the proposed size of $j$'s Markov blanket, equivalently the number of edges connected to $j$ (number of successes in repeated Bernoulli trials). Here $m$ stands for the maximum number of edges, that could be present in the local graph, that has $p_j + 1$ nodes. So the $m$ corresponds to the number of trials.

Of course, true probability $t$ is unknown to us. It could be learned from data, but we adopt a Bayesian approach and put a prior on it. Choosing a conjugate prior $t \sim \text{Beta}(a, b)$ allows us to write (see [3])

$$p(G_j) \propto \int_0^1 p(G_j|t)p(t)dt \propto \frac{\beta(a + p_j, b + m - p_j)}{\beta(a, b)},$$

where $\beta(\cdot, \cdot)$ refers to the beta function. In the numerical tests, we use $a = b = 1/2$. Motivation for this choice is that $\text{Beta}(1/2, 1/2)$ is the Jeffreys prior for the probability parameter of the binomial distribution (see [6]).

## 6.6 Graphical LASSO

The graphical LASSO (GLASSO) method can be considered as the current state-of-the-art method for estimating the inverse covariance matrices related to GGM's. The method was introduced by Friedman et al [7] in 2008.

The idea of the method in general is to maximise $L_1$-penalized Gaussian log-likelihood with respect to the precision matrix $\Omega$. By imposing the $L_1$-penalty we force some amount of $\Omega$'s elements to be zero and thus obtain sparser matrices. And since the zero pattern in precision matrix identifies the underlying undirected graph, we obtain also sparser graphs.

To formulate the GLASSO method little more in detail, suppose we have $n$ samples from $N_p(\mu, (\Omega^*)^{-1})$. Denote the sample covariance matrix by $S$. Now, the aim of GLASSO is to find a symmetric positive definite matrix $\Omega$, that maximises the expression

$$\log |\Omega| + \text{tr}(S\Omega) - \alpha \|\Omega\|_1, \tag{54}$$

where $\text{tr}(\cdot)$ is the trace operator, $\alpha > 0$ is called the tuning or regularisation parameter. The $L_1$ norm $\|\Omega\|_1$ is defined as a sum of the absolute values of $\Omega$'s elements. The first two terms in (54) come from the Gaussian log-likelihood, which has been partly maximised with respect to mean the vector $\mu$.

Maximising the expression (54) is a convex optimisation problem. Friedman et al [7] propose a fast block coordinate descent algorithm to do the optimisation in practice. Selecting a sensible value for $\alpha$ is one of the main questions when applying GLASSO in practice. One way to determine $\alpha$ is to learn it from data using cross-validation, for instance. This is also the approach adopted in numerical examples presented in [7].

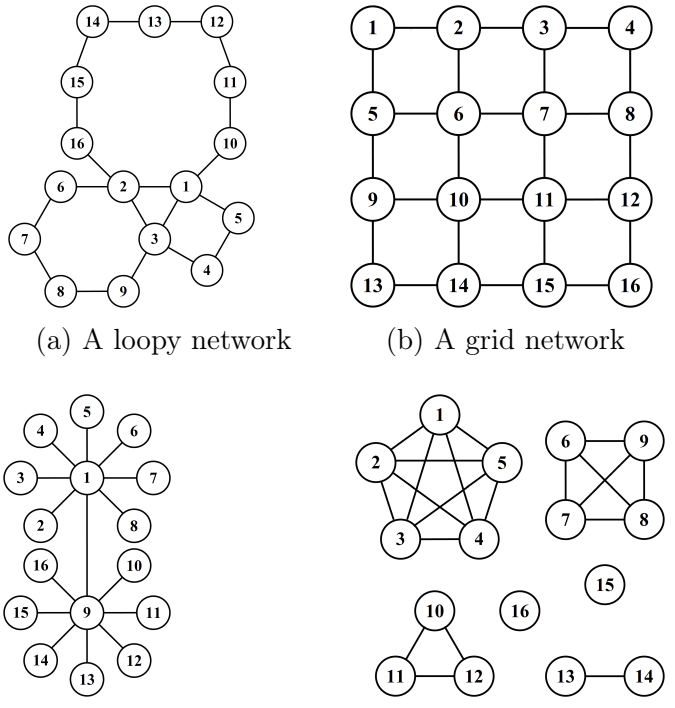
A GLASSO implementation written in R is available at Cran-repository [14].

# 7 Numerical results

In this section we study the performance of FMPL in learning the graphs from multivariate Normal data. We use Hamming distance, the true positive rate and the false positive rates to measure the accuracy of the FMPL method. The results are compared to ones obtained by using GLASSO method on same data sets.

## 7.1 Test setting

We create our data by first specifying the independence structure of the generating network by using 4 synthetic subgraphs, each of which containing 16 nodes. These smaller graphs are combined together to form a bigger 64 node graph. The subgraphs used in data generation are represented in **Figure 4**.



(a) A loopy network (b) A grid network

(c) A hub network (d) A clique network

Figure 4: Synthetic subgraphs. Pictures are taken from [10].

The bigger 64 node network, that is formed using synthetic the subgraphs as disconnected components is then replicated to obtain even bigger networks. We consider the cases of $64, 128, 256, 512$ and 1024 nodes.

After the independence structure is assigned, we construct the corresponding precision matrix by first setting the values implied by graph to zero. The absolute values of the remaining off-diagonal elements are chosen randomly between 0.1 and 0.9 so that the half of the elements are negative. The diagonal elements are also first chosen randomly from the same interval and then a suitable vector is added to ascertain the positive definiteness of the precision matrix.

Finally, the resulting matrix is inverted to obtain the covariance matrix, which is used to sample multivariate Normal data using Matlab's built-in functions.

We use three methods to learn graphs from simulated data: FMPL with uniform prior over graphs, FMPL with the sparsity promoting prior and GLASSO. AND-method is used to form the undirected graph from the Markov blankets discovered by FMPL. For every input data, we compute GLASSO using 10 different values for the tuning parameter $\alpha$. Then the parameter value leading to the least Hamming distance is chosen. The ten candidate parameter values are logarithmically spaced on the interval $[0.01, 1]$.

## 7.2 Measured quantities

We are interested in three quantities: Hamming distance, true positive rate and the false positive rate.

Since we know the true underlying graph structure that has been used to generate the data, we can measure how much a learned structure differs from it. For a learned graph $G$, the Hamming distance is defined as the number of edges, that has to be added and subtracted to obtain the real underlying graph.

True positive rate (TP) is defined as

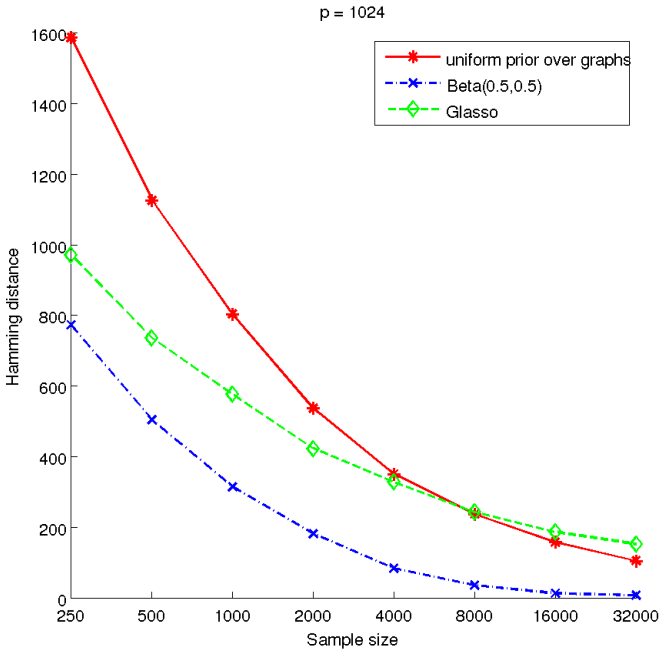
$$TP(G) = \frac{\text{The number of true edges found by a method}}{\text{Total number of edges in the true graph}}.$$

Accordingly, false positive rate is given by

$$FP(G) = \frac{\text{The number of false edges found by a method}}{\text{The number of "non-edges" in the true graph}}.$$

## 7.3 Results

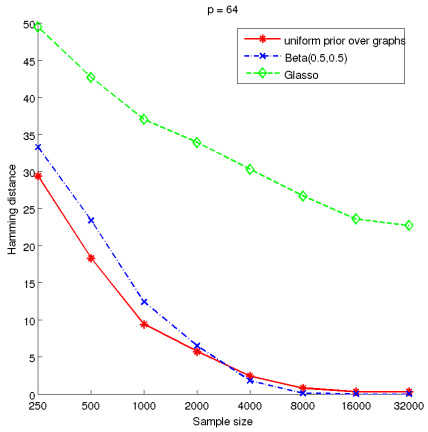
Figure 5: A figure presenting Hamming distances of three methods with different sample sizes, when the number of variables $p = 1024$.



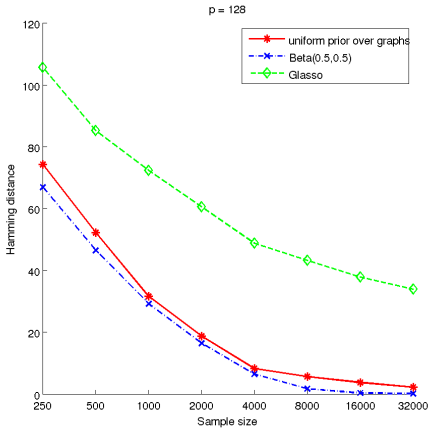

| Variables | Sample size | Uniform | | Beta(0.5,0.5) | | GLASSO | |
|---|---|---|---|---|---|---|---|
| | | TP | FP | TP | FP | TP | FP |
| 1024 | 250 | 0.5772 | 0.0020 | 0.5031 | 0.0003 | 0.2982 | 0.0002 |
| | 500 | 0.7417 | 0.0015 | 0.6690 | 0.0002 | 0.5174 | 0.0003 |
| | 1000 | 0.8522 | 0.0012 | 0.7953 | 0.0001 | 0.6619 | 0.0003 |
| | 2000 | 0.9244 | 0.0008 | 0.8847 | 7e-05 | 0.7547 | 0.0002 |
| | 4000 | 0.9728 | 0.0006 | 0.9508 | 4e-05 | 0.8554 | 0.0003 |
| | 8000 | 0.9949 | 0.0004 | 0.9856 | 3e-05 | 0.9002 | 0.0002 |
| | 16000 | 0.9998 | 0.0003 | 0.9993 | 2e-05 | 0.9482 | 0.0002 |
| | 32000 | 1.0000 | 0.0002 | 1.0000 | 1e-05 | 0.9718 | 0.0002 |

Figure 6: A table showing TP and FP rates for all the methods.

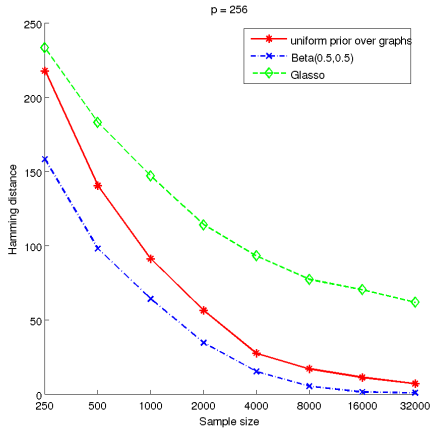
Figure 7: Hamming distance plots for dimensions $p = 64, 128, 256, 512$.



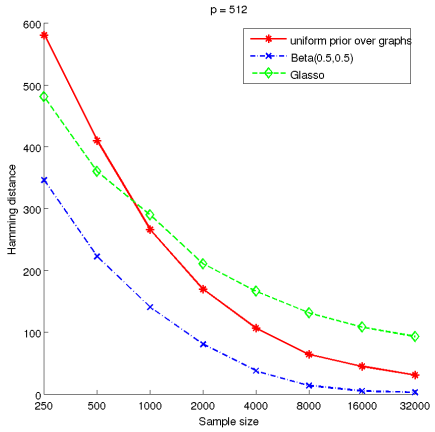

(a) $p = 64$




(b) $p = 128$




(c) $p = 256$




(d) $p = 512$

By examining the Hamming distance plots, it is clear that FMPL outperforms GLASSO in every setting under consideration. When the dimension grows, the generating network becomes sparser, which benefits GLASSO and the FMPL with sparsity promoting prior.

Figure 8: A table showing TP and FP rates in dimensions $p = 64, 128, 256, 512$.

| Variables | Sample size | Uniform | | Beta(0.5,0.5) | | Glasso | |
|---|---|---|---|---|---|---|---|
| | | TP | FP | TP | FP | TP | FP |
| 64 | 250 | 0.7128 | 0.0036 | 0.5846 | 0.0005 | 0.5910 | 0.0091 |
| | 500 | 0.8256 | 0.0024 | 0.7064 | 0.0003 | 0.6962 | 0.0098 |
| | 1000 | 0.9103 | 0.0012 | 0.8423 | 5e-05 | 0.7526 | 0.0091 |
| | 2000 | 0.9513 | 0.0010 | 0.9179 | 5e-05 | 0.8513 | 0.0115 |
| | 4000 | 0.9859 | 0.0007 | 0.9795 | 0.0001 | 0.8949 | 0.0114 |
| | 8000 | 1.0000 | 0.0004 | 1.0000 | 5e-05 | 0.9308 | 0.0110 |
| | 16000 | 1.0000 | 0.0002 | 1.0000 | 0 | 0.9538 | 0.0103 |
| | 32000 | 1.0000 | 0.0002 | 1.0000 | 0 | 0.9846 | 0.0111 |
| 128 | 250 | 0.6936 | 0.0033 | 0.5859 | 0.0003 | 0.4622 | 0.0027 |
| | 500 | 0.7917 | 0.0025 | 0.7103 | 0.0002 | 0.6212 | 0.0033 |
| | 1000 | 0.8782 | 0.0016 | 0.8179 | 0.0001 | 0.7115 | 0.0034 |
| | 2000 | 0.9429 | 0.0012 | 0.9006 | 0.0001 | 0.7929 | 0.0035 |
| | 4000 | 0.9808 | 0.0007 | 0.9615 | 5e-05 | 0.8776 | 0.0037 |
| | 8000 | 0.9955 | 0.0006 | 0.9923 | 6e-05 | 0.9103 | 0.0037 |
| | 16000 | 1.0000 | 0.0005 | 1.0000 | 4e-05 | 0.9423 | 0.0036 |
| | 32000 | 1.0000 | 0.0003 | 1.0000 | 1e-05 | 0.9827 | 0.0039 |
| 256 | 250 | 0.6494 | 0.0034 | 0.5372 | 0.0004 | 0.4042 | 0.0015 |
| | 500 | 0.7772 | 0.0022 | 0.7035 | 0.0002 | 0.5920 | 0.0017 |
| | 1000 | 0.8673 | 0.0015 | 0.8064 | 0.0001 | 0.6837 | 0.0015 |
| | 2000 | 0.9356 | 0.0011 | 0.8981 | 8e-05 | 0.8074 | 0.0017 |
| | 4000 | 0.9769 | 0.0006 | 0.9554 | 4e-05 | 0.8574 | 0.0015 |
| | 8000 | 0.9952 | 0.0005 | 0.9885 | 5e-05 | 0.9250 | 0.0016 |
| | 16000 | 0.9997 | 0.0003 | 0.9987 | 3e-05 | 0.9375 | 0.0016 |
| | 32000 | 1.0000 | 0.0002 | 1.0000 | 2e-05 | 0.9747 | 0.0017 |
| 512 | 250 | 0.6154 | 0.0026 | 0.5212 | 0.0004 | 0.3228 | 0.0005 |
| | 500 | 0.7593 | 0.0020 | 0.6838 | 0.0002 | 0.5604 | 0.0007 |
| | 1000 | 0.8587 | 0.0014 | 0.7992 | 0.0001 | 0.6883 | 0.0007 |
| | 2000 | 0.9253 | 0.0009 | 0.8880 | 8e-05 | 0.7620 | 0.0005 |
| | 4000 | 0.9713 | 0.0007 | 0.9503 | 5e-05 | 0.8505 | 0.0006 |
| | 8000 | 0.9946 | 0.0005 | 0.9872 | 4e-05 | 0.8897 | 0.0005 |
| | 16000 | 0.9995 | 0.0003 | 0.9989 | 3e-05 | 0.9412 | 0.0005 |
| | 32000 | 1.0000 | 0.0002 | 1.0000 | 2e-05 | 0.9644 | 0.0005 |

The tables presenting TP and FP rates show that adding a sparsity promoting prior to FMPL greatly reduces the false positive rate, which is vital in the high dimensional cases, where the generating network is sparse. These tests indicate that GLASSO can maintain fairly good FP rate in every setting used, but it needs quite a lot of data to be able the find all the true edges.

# 8 Conclusions

We have presented the fractional marginal pseudo-likelihood, a Bayesian method for learning graphical models from multivariate Normal data. A particular strength of the method is its objectivity: we do not need to assign any subjective prior beliefs on model parameters, which makes the method easily applicable by the user.

Often the interesting real world applications consider situations where the number of variables is much larger than the available sample size. The adopted Bayesian approach can cope with these situations naturally through the marginalization over the nuisance parameters in the model. By assigning a sparsity promoting prior over the graph structures, the method was also shown to estimate sparse graphs accurately from synthetic data.

The FMPL method has a sound theoretical basis. Given enough data, the true graph structure will be eventually identified. The result was proven by showing the consistency of the local Markov blanket estimators. This result also allows us to factorise our problem into independent sub-problems, which can be then solved parallel. The factorisation property is the key component, that makes the method applicable also in the high dimensional settings.

Future research would include further analysis of the method, its applications to real data and making the method more robust to outliers by relaxing the Gaussian assumption. The robust method could be compared against a method by Sun & Li [12], which was shown to perform better than GLASSO, when the data include outliers.

# References


[1] Abramowitz, M & Stegun, I. (editors). *A Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables.* Dover Publications, New York, 1972.

[2] Besag, J. E. *Nearest-Neighbour Systems and the Auto-Logistic Model for Binary Data.* Journal of the Royal Statistical Society. Series B (Methodological), 34 (1), 1972.

[3] Carvalho, C. & Scott, J. *Objective Bayesian model selection in Gaussian graphical models.* Biometrika, 96 (3), 2009.

[4] Consonni, G. & La Rocca, L. *Objective Bayes factors for Gaussian Directed Acyclic Graphical Models.* Scandinavian Journal of Statistics, 34 (4), 2012.

[5] Geiger, D. & Heckerman, D. *Parameter priors for directed acyclic graphical models and the characterization of several probability distributions.* The Annals of Statistics, 30 (5), 2002.

[6] Grünwald, P. *The minimum description length principle.* MIT Press, Cambridge MA, 2007.

[7] Friedman, J; Hastie, T and Tibshirani, R. *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics, 9 (3), 2007.

[8] Koller, D & Friedman, N. *Probabilistic Graphical Models: Principles and Tehniques.* MIT Press, Cambridge MA, 2009.

[9] O'Hagan, A. *Fractional Bayes Factors for Model Comparison.* Journal of the Royal Statistical Society, 57 (1), 1995.

[10] Pensar, J; Nyman, H; Corander, J. *Marginal Pseudo-Likelihood Inference for Markov Networks.* arXiv:1401.4988, 2014.

[11] Press, S. J. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference.* Robert E. Krieger Publishing Company, Malabar, 1982.

[12] Sun, L & Li, H. *Robust Gaussian graphical modeling via $l_1$ penalization.* Biometrics, 68 (4), 2012.

[13] Whittaker, J. *Graphical Models in Applied Multivariate Statistics.* Wiley, New York, 1990.

[14] R implementation of GLASSO (version 1.7). `http://cran.r-project.org/web/packages/glasso/`

# A Consistency proof

In this chapter we well provide the proof of the consistency of FMPL as a local Markov blanket estimator, as stated in the chapter 6. During the proof, we will make use of a couple well-known results, which are presented here. The proof itself has not appeared anywhere before and it is result of original work presented here. We will start by reviewing the notation and the setting used in the chapter 6.

## A.1 Statement of the theorem

Consider an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of nodes and $E \subset V \times V$ is the set of edges. Let $X = (X_1, \ldots, X_p)$ be a $p$-dimensional Gaussian random vector and $\Omega$ denote the precision matrix of $X$. We assume that $\Omega$ is positive definite and $\Omega_{ij} = 0$, if and only if there is no edge between nodes $i$ and $j$. Assume zero mean for $X$ and denote $\Omega^{-1} = \Sigma$, so $X \sim N_p(\mathbf{0}, \Sigma)$.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ denote a complete random sample obtained from this distribution. The local fractional marginal pseudo-likelihood (FMPL) for the node $X_j$ can be written as

$$p(\mathbf{X}_j | \mathbf{X}_{mb(j)}) = \pi^{-\frac{n-1}{2}} \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{1+p_j}{2}\right)} \left(\frac{1}{n}\right)^{\frac{1+2p_j}{2}} \left(\frac{|S_{fa(j)}|}{|S_{mb(j)}|}\right)^{-\frac{n-1}{2}}, \qquad (55)$$

in which $mb(j)$ denotes the Markov blanket of $X_j$ and $fa(j) = mb(j) \cup \{j\}$, $p_j$ is the cardinality of the set $mb(j)$ and $S$ is a matrix containing sums of squares and products of observation vectors' coordinates. $S_A$ is a submatrix of $S$ restricted to variables in the set $A$.

**Theorem A.1.1.** *Let $X \sim N_p(\boldsymbol{0}, (\Omega^*)^{-1})$ and $G = (V, E)$ denote the the undirected graph that completely determines the conditional independence statements between $X$'s components. Let $\{mb^*(1), \ldots, mb^*(p)\}$ denote the set of Markov blankets, which uniquely define $G$.*

*Suppose we have a complete random sample $\boldsymbol{X}$ of a size $n$ obtained from $N_p(\boldsymbol{0}, \Omega^*)$. Then the local FMPL estimator*

$$\widehat{mb}(j) = \underset{mb(j) \subset V \setminus \{j\}}{\arg\max}\ p(\boldsymbol{X}_j | \boldsymbol{X}_{mb(j)})$$

*is consistent, that is,* $\widehat{mb}(j) = mb^*(j)$ *with a probability tending to* 1*, as* $n \to \infty$.

## A.2 Preliminary results

The following propositions are based on the ones found in [13].

**Theorem A.2.1.** (6.7.1; *p.* 179) *Suppose the Normal random vector* $X$ *can be partitioned into three* $(X_A, X_B, X_C)$ *and all conditional independence constraints can be summarised by the single statement* $X_B \perp\!\!\!\perp X_C | X_A$*. If* $X_A, X_B$ *and* $X_C$ *are p-,q- and r-dimensional respectively, then the deviance*

$$dev(X_B \perp\!\!\!\perp X_C | X_A) = -n \log \frac{|S||S_A|}{|S_{A\cup B}||S_{A\cup C}|}$$

*has an asymptotic chi-squared distribution with* $qr$ *degrees of freedom.*

Here $S$ is defined as before, but in [13] $S$ is used to denote the sample covariance matrix. It is clear that this doesn't change the statement of the theorem in any manner. Note that theorem holds also if $A = \emptyset$, since complete independence can be considered a special case of the conditional independence. In this case, term $|S_A|$ in the expression of deviance just disappears.

**Theorem A.2.2.** (5.6.1; *p.* 138) *Consider the partitioned random vector* $X = (X_A, X_B, X_C)$*. Let* $\hat{X}_A[X_{B\cup C}]$ *denote the linear least squares predictor of* $X_A$ *from* $X_{B\cup C}$*. Now it holds that*

$$var(\hat{X}_A[X_{B\cup C}]) = var(\hat{X}_A[X_B]) + var(\hat{X}_A[X_C - \hat{X}_C[X_B]]),$$

*where* $var(\hat{X}_A[X_C - \hat{X}_C[X_B]])$ *can be expressed using partial covariance as*

$$cov(X_A, X_C|X_B) var(X_C|X_B)^{-1} cov(X_C, X_A|X_B).$$

The next theorem considering a determinant of a partitioned matrix can be found in [11].

**Theorem A.2.3.** (2.6.1; *p.* 26) *Let* $A$ *be an arbitrary* $(n+p)\times(n+p)$ *matrix. Partition* $A$ *as follows*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

*where* $A_{11}$ *is* $p \times p$ *matrix,* $A_{12}$ *is* $p \times n$*,* $A_{21}$ *is* $n \times p$ *and* $A_{22}$ *is* $n \times n$*. If* $|A_{22}| \neq 0$*, then*

$$|A| = |A_{22}| \cdot |A_{11} - A_{12}(A_{22})^{-1} A_{21}|.$$

In the proof, we will study the asymptotic behaviour of the local FMPL as sample size increases. The terms that are asymptotically constant with respect to the sample size $n$ will be ignored. We'll use notation $O(1)$ and $O_p(1)$ to denote this. More formally:

**Definition A.2.4.** *Let $(x_n)_{n=1}^{\infty}$ be a sequence of real numbers. We write*

$$x_n = O(1), \text{ as } n \to \infty,$$

*to say that the sequence $(x_n)$ is bounded. More formally, there is $M > 0$ such that*

$$|x_n| \leq M, \text{ for every } n \in \mathbb{N}.$$

**Definition A.2.5.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of real valued random variables. We write*

$$X_n = O_p(1),$$

*to say that the sequence $(X_n)$ is bounded in probability, that is, for any $\varepsilon > 0$ there is $M > 0$ such that*

$$P(|X_n| > M) < \varepsilon, \text{ for every } n \in \mathbb{N}.$$

The following asymptotic approximation for the logarithm of the Gamma function is based on the result found in [1].

**Theorem A.2.6.** (6.1.41; *p.* 257) *Let $x \to \infty$, $x > 0$. Now it holds, that*

$$\log \Gamma(x) \approx (x - \frac{1}{2}) \log x - x + O(1).$$

This theorem is referred in the proof as *Stirling's asymptotic formula.*

## A.3 The proof

The main idea of the proof is to examine the score given to node's true Markov blanket and compare it to a score given to some other Markov blanket candidate. As $n$ grows, we will show that the true Markov blanket will get the higher score, despite the choice of the set to be compared against.

The proof is divided in two parts. First, we show that the FMPL score does not overestimate. The true Markov blanket is preferred over the sets containing unnecessary nodes. The second part shows that the score does not underestimate. The set that does not contain all the members of the true Markov blanket will get strictly lower score. Combining these two results will prove our theorem.

## Overestimation

Let $mb^* \subset V$ and $fa^* \subset V$ denote the true Markov blanket and the true family of the node $X_j$, respectively. We denote the cardinality of $mb^*$ by $p_j$. Let $mb \subset V$ be a superset of the true Markov blanket $mb^*$. Denote $a = |mb| - p_j$. Since $mb^* \subset mb$, we have $a > 0$.

We want to show that

$$\log \frac{p(\mathbf{X}_j \mid \mathbf{X}_{mb^*})}{p(\mathbf{X}_j \mid \mathbf{X}_{mb})} \to \infty$$

in probability, as $n \to \infty$. Showing this will guarantee that FMPL prefers the true Markov blanket over its supersets as the sample size increases.

Consider next the log ratio of FMPLs. The term containing the power of $\pi$ appears in both of the FMPLs, and so it cancels. By noticing that

$$\frac{n^{-\left(\frac{1+2p_j}{2}\right)}}{n^{-\left(\frac{1+2(p_j+a)}{2}\right)}} = n^a,$$

we get the following form for the ratio

$$\begin{aligned}\log \frac{p(\mathbf{X}_j|\mathbf{X}_{mb^*})}{p(\mathbf{X}_j|\mathbf{X}_{mb})} &= \log \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{n+p_j+a}{2}\right)} + \log \frac{\Gamma\left(\frac{1+p_j+a}{2}\right)}{\Gamma\left(\frac{1+p_j}{2}\right)} \\ &\quad + a \log n - \left(\frac{n-1}{2}\right) \log \left(\frac{|S_{fa^*}||S_{mb}|}{|S_{mb^*}||S_{fa}|}\right). \qquad (56)\end{aligned}$$

The second term in (56) doesn't depend on $n$ so it can be omitted. Denote $m = (n + p_j)/2$. Clearly $m \to \infty$, as $n \to \infty$. Now we can write the first term in (56) as

$$\log \frac{\Gamma(m)}{\Gamma\left(m + \frac{a}{2}\right)} = \log \Gamma(m) - \log \Gamma\left(m + \frac{a}{2}\right). \qquad (57)$$

Now letting $n \to \infty$ and by using Stirling's asymptotic formula for each of the terms in (57), we get

$$\begin{aligned}\log \Gamma(m) - \log \Gamma\left(m + \frac{a}{2}\right) &= \left(m - \frac{1}{2}\right) \log m - m \\ &\quad - \left(\left(m + \frac{a}{2} - \frac{1}{2}\right) \log \left(m + \frac{a}{2}\right) - \left(m + \frac{a}{2}\right)\right) + O(1).\end{aligned}$$

We see that $m$ term cancels and the constant $a/2$ in the second term can be omitted. After rearranging terms, result can be written as

$$m \log \left( \frac{m}{m + \frac{a}{2}} \right) + \frac{1}{2} \log \left( \frac{m + \frac{a}{2}}{m} \right) - \frac{a}{2} \log \left( m + \frac{a}{2} \right) + O(1). \tag{58}$$

Asymptotically only the last log-term in (58) is relevant and the formula simplifies to

$$- \frac{a}{2} \log \left( m + \frac{a}{2} \right) + O(1). \tag{59}$$

This can be seen by noticing, that as $n \to \infty$, we have

$$m \log \left( \frac{m}{m + \frac{a}{2}} \right) = \frac{1}{2} \log \left( \frac{1}{1 + \frac{a}{2m}} \right)^{2m} \to \frac{1}{2} \log e^{-a} = -\frac{a}{2}$$

and

$$\frac{1}{2} \log \left( \frac{m + \frac{a}{2}}{m} \right) = \frac{1}{2} \log \left( 1 + \frac{a}{2m} \right) \to 0.$$

So far, we have shown that asymptotically

$$\log \frac{\Gamma(m)}{\Gamma\left(m + \frac{a}{2}\right)} = -\frac{a}{2} \log \left( m + \frac{a}{2} \right) + O(1),$$

or equivalently by using variable $n$

$$\log \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{n+p_j+a}{2}\right)} = -\frac{a}{2} \log \left( \frac{n + p_j + a}{2} \right) + O(1). \tag{60}$$

No we can simplify the original formula (56) by combining the first and the third term

$$\begin{aligned}
\log \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{n+p_j+a}{2}\right)} + a \log n &= -\frac{a}{2} \log \left( \frac{n + p_j + a}{2} \right) + \frac{a}{2} \log n^2 + O(1) \\
&= \frac{a}{2} \log \left( \frac{2n^2}{n + p_j + a} \right) + O(1) \\
&= \frac{a}{2} \log \left( \frac{2n}{1 + \frac{p_j}{n} + \frac{a}{n}} \right) + O(1) \\
&= \frac{a}{2} \log n + O(1).
\end{aligned} \tag{61}$$

Consider next the last term in (56)

$$-\left(\frac{n-1}{2}\right)\log\left(\frac{|S_{fa^*}||S_{mb}|}{|S_{mb^*}||S_{fa}|}\right). \tag{62}$$

Since $mb^* \subset mb$, we can write $mb = mb^* \cup R$, where $R$ denotes the set of unnecessary variables in $mb$. Recall the **Theorem A.2.1** and notice that by denoting

$$A = mb^*,\ B = \{j\} \text{ and } C = R,$$

it holds that $X_B \perp\!\!\!\perp X_C | X_A$, since $mb^*$ was $X_j$'s true Markov blanket. Note also that in this case $qr = 1 \cdot a = a$. Now the deviance can be written as

$$\text{dev}(X_j \perp\!\!\!\perp X_R | X_{mb^*}) = -n\log\left(\frac{|S_{fa}||S_{mb^*}|}{|S_{fa^*}||S_{mb}|}\right), \tag{63}$$

which is essentially just the determinant term (62) multiplied by a constant $-2$. Let us denote $D_n = \text{dev}(X_j \perp\!\!\!\perp X_R | X_{mb^*})$. The determinant term gets the following representation

$$\begin{aligned} -\left(\frac{n-1}{2}\right)\log\left(\frac{|S_{fa^*}||S_{mb}|}{|S_{mb^*}||S_{fa}|}\right) &= -\frac{n}{2}\log\left(\frac{|S_{fa^*}||S_{mb}|}{|S_{mb^*}||S_{fa}|}\right) + O_p(1) \\ &= -\frac{D_n}{2} + O_p(1). \end{aligned} \tag{64}$$

The $O_p(1)$ error on the first line comes from omitting the term

$$\frac{1}{2}\log\left(\frac{|S_{fa^*}||S_{mb}|}{|S_{mb^*}||S_{fa}|}\right).$$

**Theorem A.2.1** says that asymptotically, it holds that $D_n \sim \chi^2_a$. In other words

$$D_n \xrightarrow{d} D,\ D \sim \chi^2_a.$$

Convergence in distribution implies that the sequence $(D_n)$ is bounded in probability

$$D_n = O_p(1).$$

All in all, asymptotically

$$-\left(\frac{n-1}{2}\right)\log\left(\frac{|S_{fa^*}||S_{mb}|}{|S_{mb^*}||S_{fa}|}\right) = O_p(1).$$

Adding the results together, we have shown that, as $n \to \infty$

$$\log \frac{p(\mathbf{X}_j \mid \mathbf{X}_{mb^*})}{p(\mathbf{X}_j \mid \mathbf{X}_{mb})} = \frac{a}{2} \log n + O_p(1). \tag{65}$$

Now since $a > 0$, then

$$\log \frac{p(\mathbf{X}_j \mid \mathbf{X}_{mb^*})}{p(\mathbf{X}_j \mid \mathbf{X}_{mb})} \to \infty$$

in probability, as $n \to \infty$ . □

## Underestimation

Let $mb^*$ denote the true Markov blanket of node $X_j$ and $mb \subsetneq mb^*$. Let $A \subset V \backslash fa^*$. Remember that $fa^*$ was defined to be $mb^* \cup \{j\}$. Note that $A$ could also be an empty set. We want to show that

$$\log \frac{p(\mathbf{X}_j \mid \mathbf{X}_{mb^* \cup A})}{p(\mathbf{X}_j \mid \mathbf{X}_{mb \cup A})} \to \infty$$

in probability, as $n \to \infty$. Denote $|mb^* \cup A| = p_j$ and $a = |mb \cup A| - p_j$. Here $a < 0$, since $mb$ is a subset of the true Markov blanket. We can now proceed similarly as in the overestimation part, and write the log ratio as

$$\begin{aligned} \log \frac{p(\mathbf{X}_j|\mathbf{X}_{mb^* \cup A})}{p(\mathbf{X}_j|\mathbf{X}_{mb \cup A})} &= \log \frac{\Gamma\left(\frac{n+p_j}{2}\right)}{\Gamma\left(\frac{n+p_j+a}{2}\right)} + \log \frac{\Gamma\left(\frac{1+p_j+a}{2}\right)}{\Gamma\left(\frac{1+p_j}{2}\right)} \\ &\quad + a \log n - \left(\frac{n-1}{2}\right) \log \left(\frac{|S_{fa^* \cup A}||S_{mb \cup A}|}{|S_{mb^* \cup A}||S_{fa \cup A}|}\right). \end{aligned} \tag{66}$$

The first three terms are just the same ones appearing in (56), which allows us to write

$$\log \frac{p(\mathbf{X}_j|\mathbf{X}_{mb^* \cup A})}{p(\mathbf{X}_j|\mathbf{X}_{mb \cup A})} = \frac{a}{2} \log n - \left(\frac{n-1}{2}\right) \log \left(\frac{|S_{fa^* \cup A}||S_{mb \cup A}|}{|S_{mb^* \cup A}||S_{fa \cup A}|}\right) + O(1). \tag{67}$$

Consider next the determinant term

$$- \left(\frac{n-1}{2}\right) \log \left(\frac{|S_{fa^* \cup A}||S_{mb \cup A}|}{|S_{mb^* \cup A}||S_{fa \cup A}|}\right). \tag{68}$$

By the definition of $S$, it is clear that

$$\frac{S}{n} = \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}^i (\mathbf{X}^i)^T,$$

where $\hat{\Sigma}$ is the sample covariance matrix, or equivalently the maximum likelihood estimate of the true covariance matrix. As $n$ approaches infinity, the MLE converges in probability to the true covariance matrix $\Sigma$.

Letting $n \to \infty$, we can write the argument of logarithm in (68) as follows

$$\frac{\frac{|\Sigma_{fa^*\cup A}|}{|\Sigma_{mb^*\cup A}|}}{\frac{|\Sigma_{fa\cup A}|}{|\Sigma_{mb\cup A}|}} \tag{69}$$

We can simplify the numerator and denominator by noticing that,

$$\Sigma_{fa^*\cup A} = \begin{pmatrix} \operatorname{var}(X_j) & \operatorname{cov}(X_j, X_{mb^*\cup A}) \\ \operatorname{cov}(X_j, X_{mb^*\cup A})^T & \Sigma_{mb^*\cup A} \end{pmatrix},$$

where $\operatorname{var}(X_j)$ is the variance of variable $X_j$, $\operatorname{cov}(X_j, X_{mb^*\cup A})$ is a horizontal vector containing covariances between $X_j$ and each of the variables in set $mb^* \cup A$. Using the **Theorem A.2.3**, we have

$$\begin{aligned} |\Sigma_{fa^*\cup A}| &= |\Sigma_{mb^*\cup A}| \cdot \left(\operatorname{var}(X_j) - \operatorname{cov}(X_j,\ X_{mb^*\cup A}) \left(\Sigma_{mb^*\cup A}\right)^{-1} \operatorname{cov}(X_j,\ X_{mb^*\cup A})^T\right) \\ &= |\Sigma_{mb^*\cup A}| \cdot \left(\operatorname{var}(X_j) - \operatorname{var}(\hat{X}_j[X_{mb^*\cup A}])\right) \\ &= |\Sigma_{mb^*\cup A}| \cdot \operatorname{var}\left(X_j |\ X_{mb^*\cup A}\right). \end{aligned}$$

Last equality follows from the definition of partial variance, which is the residual variance of $X_j$ after subtracting the variance based on $\hat{X}_j[X_{mb^*\cup A}]$, the linear least squares predictor of $X_j$ from variables in $mb^* \cup A$. Using this, we get

$$\frac{|\Sigma_{fa^*\cup A}|}{|\Sigma_{mb^*\cup A}|} = \operatorname{var}\left(X_j \mid X_{mb^*\cup A}\right)$$

and

$$\frac{|\Sigma_{fa\cup A}|}{|\Sigma_{mb\cup A}|} = \operatorname{var}\left(X_j \mid X_{mb\cup A}\right).$$

Plugging these into (69) gives

$$\frac{\frac{|\Sigma_{fa^*\cup A}|}{|\Sigma_{mb^*\cup A}|}}{\frac{|\Sigma_{fa\cup A}|}{|\Sigma_{mb\cup A}|}} = \frac{\text{var}\left(X_j \mid X_{mb^*\cup A}\right)}{\text{var}\left(X_j \mid X_{mb\cup A}\right)} \tag{70}$$

The form (70) makes it easier to analyse the behaviour of the determinant term and we can write the log ratio in (66) as follows

$$\log \frac{p(\mathbf{X}_j|\mathbf{X}_{mb^*\cup A})}{p(\mathbf{X}_j|\mathbf{X}_{mb\cup A})} = \frac{a}{2}\log n - \frac{n}{2}\log \frac{\text{var}\left(X_j \mid X_{mb^*\cup A}\right)}{\text{var}\left(X_j \mid X_{mb\cup A}\right)} + O_p(1). \tag{71}$$

By looking at (71), it's clear that consistency is achieved if we can show that

$$\frac{\text{var}\left(X_j \mid X_{mb^*\cup A}\right)}{\text{var}\left(X_j \mid X_{mb\cup A}\right)} < 1. \tag{72}$$

The equation (72) is equivalent to

$$\begin{aligned} &\text{var}\left(X_j \mid X_{mb^*\cup A}\right) < \text{var}\left(X_j \mid X_{mb\cup A}\right) \\ \Leftrightarrow\ &\text{var}(X_j) - \text{var}(\hat{X}_j[X_{mb^*\cup A}]) < \text{var}(X_j) - \text{var}(\hat{X}_j[X_{mb\cup A}]) \\ \Leftrightarrow\ &\text{var}(\hat{X}_j[X_{mb^*\cup A}]) > \text{var}(\hat{X}_j[X_{mb\cup A}]). \end{aligned} \tag{73}$$

Now assume $mb \neq \emptyset$, and denote the missing true Markov blanket members by $R = mb^* \backslash mb$. Then with the help of **Theorem A.2.2**, we can write the left side of (73) as

$$\begin{aligned} \text{var}(\hat{X}_j[X_{mb^*\cup A}]) &= \text{var}(\hat{X}_j[X_{mb\cup A\cup R}]) \\ &= \text{var}(\hat{X}_j[X_{mb\cup A}]) + \text{var}(\hat{X}_j[X_R - \hat{X}_R[X_{mb\cup A}]]). \end{aligned}$$

The term $\text{var}(\hat{X}_j[X_R - \hat{X}_R[X_{mb\cup A}]]) > 0$, since elements of $R$ are in $X_j's$ Markov blanket. This shows that (72) holds.

If $mb = \emptyset$, the inequality (73) can be written as

$$\text{var}(\hat{X}_j[X_{mb^*\cup A}]) > \text{var}(\hat{X}_j[X_A]).$$

Using again the **Theorem A.2.2**, this becomes

$$\text{var}(\hat{X}_j[X_A]) + \text{var}(\hat{X}_j[X_{mb^*} - \hat{X}_{mb^*}[X_A]]) > \text{var}(\hat{X}_j[X_A]),$$

which clearly holds.
All in all, we have showed that

$$-\frac{n}{2}\log\frac{\operatorname{var}\left(X_j \mid X_{mb^*\cup A}\right)}{\operatorname{var}\left(X_j \mid X_{mb\cup A}\right)} \to \infty,$$

in probability, as $n \to \infty$. This implies that

$$\log\frac{p(\mathbf{X}_j \mid \mathbf{X}_{mb^*\cup A})}{p(\mathbf{X}_j \mid \mathbf{X}_{mb\cup A})} \to \infty$$

in probability, as $n \to \infty$, since $n$ increases faster than $(a/2)\log n$ decreases. □

To make it more transparent to see why these two proposition are sufficient to prove our theorem, consider a variable $X_j$ and its true Markov blanket $mb^*$. Take an arbitrary set $A \subset V \setminus \{j\},\ A \neq mb^*$. We will show that FMPL prefers $mb^*$ over $A$.

1. Assume $mb^* = \emptyset$. Now it holds that $mb^* \subset A$ and the overestimation part of the proof guarantees that $mb^*$ is preferred over $A$.

2. Assume $mb^* \neq \emptyset$. Now one of the following three cases has to hold.

   (i) We have $mb^* \cap A = \emptyset$. Now the underestimation part implies that $mb^* \cup A$ is preferred over $A$ and $mb^*$ is preferred over $mb^* \cup A$ (overestimation). Note that this covers also the case when $A = \emptyset$.

   (ii) We have $mb^* \cap A \subsetneq mb^*$. This case is similar to (i).

   (iii) We have $mb^* \cap A = mb^*$. This is equivalent to $mb^* \subset A$ and the overestimation statement implies that $mb^*$ is preferred over $A$.