



**HAL**  
open science

# Exact Bayesian Inference in Graphical Models : Tree-structured Network Inference and Segmentation

Loïc Schwaller

► **To cite this version:**

Loïc Schwaller. Exact Bayesian Inference in Graphical Models : Tree-structured Network Inference and Segmentation. Statistics [math.ST]. Université Paris Saclay (COMUE), 2016. English. NNT : 2016SACLS210 . tel-01401458

**HAL Id: tel-01401458**

**<https://theses.hal.science/tel-01401458v1>**

Submitted on 23 Nov 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2016SACLS210

THÈSE DE DOCTORAT  
DE L'UNIVERSITÉ PARIS-SACLAY  
PRÉPARÉE À L'UNIVERSITÉ PARIS-SUD

École doctorale n°574  
École doctorale de mathématiques Hadamard  
Spécialité de doctorat : Mathématiques appliquées

par

**M. LOÏC SCHWALLER**

Exact Bayesian Inference in Graphical Models: Tree-structured  
Network Inference and Segmentation

Thèse présentée et soutenue à Paris, le 9 septembre 2016.

Composition du Jury :

M.	ÉTIENNE BIRMELÉ	Université Paris Descartes	(Rapporteur)
M.	CHRISTOPHE GIRAUD	Université Paris-Sud	(Président du jury)
M.	STEFFEN LAURITZEN	University of Copenhagen	(Examinateur)
Mme	MARINA MEILÄ	University of Washington	(Rapporteur)
M.	STÉPHANE ROBIN	INRA	(Directeur de thèse)

**UMR MIA-Paris AgroParisTech/INRA**  
Université Paris-Saclay  
16, rue Claude Bernard  
75005 Paris



*“Nine worlds I knew, the nine in the tree  
With mighty roots beneath the mold.”*  
– THE POETIC EDDA



# Remerciements

*“For a moment, nothing happened. Then, after a second or so, nothing continued to happen.”*

— Douglas Adams, *The Hitchhiker’s Guide to the Galaxy*

*Voilà qui résume de manière assez fidèle les premiers instants du processus ayant abouti à ces remerciements. Ajoutez même quelques heures de néant à tout cela pour être au plus proche de la réalité. Rien de nouveau sous le soleil, mais l’exercice des remerciements est difficile. Difficile parce que personnel. Difficile parce que stylistique. Difficile parce que ce sont souvent les premières lignes sur lesquelles vont se poser les yeux du lecteur. Parfois même les seules. Ne mentez pas, ne rougissez pas, c’est quelque chose que tout le monde fait. À défaut d’être originaux, j’espère qu’ils transmettront le moins partiellement et erronément possible toute la reconnaissance que j’aurais souhaité qu’ils expriment.*

Stéphane, merci. Merci de m’avoir ouvert les portes de cette aventure que rien ne prédestinait à être graphique et arborescente, mais qui l’a pourtant été. Elle fut également passionnante, souriante et profondément formatrice. Il m’est littéralement impossible de concevoir un environnement doctoral plus épanouissant et plus agréable que celui que tu as pu m’offrir. Un savant mélange de liberté et de conseils avisés, dispensé avec générosité et disponibilité. Tes qualités sont si nombreuses qu’il m’est difficile de leur faire justice ici. La perspective embarrassante d’une tentative pédestre et maladroite m’invite à en rester là, mais que ce manque d’audace ne soit qu’une preuve supplémentaire de mon admiration.

Quelques mots en anglais, puisque c’est outre-Manche que l’histoire a commencé. Thank you Michael, for taking me in as a Master student and for having started all this, in a way. The Imperial College has been a special place for me.

Je remercie Marina Meilă et Étienne Birmelé d’avoir sacrifié une partie de leur été pour rapporter cette thèse. J’espère que la tâche ne fut pas trop pénible. Je remercie également Christophe Giraud et Steffen Lauritzen d’avoir accepté de faire partie de mon jury.

Plus que d’un laboratoire, c’est d’une communauté dont j’ai fait partie durant mes années de thèse. Une communauté éminemment scientifique et profondément humaine. Merci à chacun d’entre vous. Pour ces pauses café, que l’éclectisme des conversations rend si vivantes et imprévisibles. Pour ces moments passés, au Vieux Chêne et ailleurs, à travailler sur une définition claire du bobo, du hipster et du parisien, pour finalement toujours tomber dans

les pièges de l'induction. Pour ces quiz, où la bonne humeur vole la vedette à la mauvaise foi et à la culture. Que la tradition perdure. Ces instants cristallisent pour moi un esprit que je ne peux qu'espérer retrouver ailleurs, aussi loin que les vents me portent. Vous êtes vraiment de belles personnes.

Un grand merci au Bureau des Pauses, lieu de vie (et de travail) où l'entropie est étrangement hétérogène et où les idées les plus déjantées ont une certaine tendance à ne pas rester abstraites. Anna et Marie, je suis assez fier de ce que l'on a accompli, et je ne suis pas malheureux d'avoir partagé ce bureau avec vous. J'eus été bien démuni, sans votre soutien, devant ces pages web où il faut confirmer et confirmer encore. Petit conseil pour la fin : commencez dès maintenant à écrire vos remerciements. Merci à Pierre G. de nous avoir apporté cette saine émulation et de nous rappeler quotidiennement que l'humilité est la plus haute des qualités. Je remercie également nos voisins et chaperons pour leur bienveillance permanente.

À toutes les personnes ayant contribué à cette thèse de par la qualité des moments où je n'y travaillais pas, merci.  
Et elles sont nombreuses.  
Elles font avancer la recherche.  
Elles rient, elles dansent, elles chantent.  
Elles foulent les planches.  
Elles habitent loin. Ou pas.  
Elles rugissent comme tout le monde.  
Elles sont d'une candeur entière.  
Elles lézardent au soleil.  
Elles engueulent la pauvre Suzanne.  
Elles adorent (perdre) les paris.  
Elles mangent du pâté de campagne.  
Elles vont toujours bien.  
Elles poussent des cris (très) aigus.  
Elles sont au moins aussi folles que moi.  
Elles ont de nombreux petits-enfants.  
Elle n'aiment pas être bousculées.  
Elles ont des perches à selfie.  
Elles ne savent pas faire un duck face.  
Elles naissent et meurent pour la paix.  
Elles sont géniales.  
Encore une fois, merci.

J'adresse également des remerciements incommensurables à mes parents. Je ne vous dirai sûrement jamais assez à quel point vous êtes exceptionnels. Votre générosité sans bornes m'émerveille encore et toujours.

Enfin, les dernières lignes de ces remerciements vont à la personne qui a embrassé ma folie, mon indécision et mes lubies. Julien, merci d'être là, tout simplement.

# Contents

<b>0</b>	<b>Introduction</b>	<b>1</b>
<b>1</b>	<b>Background</b>	<b>7</b>
1.1	Graphical Models . . . . .	8
1.2	Algebra & Algorithms . . . . .	27
<b>2</b>	<b>Network Inference</b>	<b>37</b>
2.1	Introduction . . . . .	38
2.2	Background & Model . . . . .	39
2.3	Priors on Tree Structures & Distributions . . . . .	41
2.4	Inference in Tree Graphical Models . . . . .	44
2.5	Simulations . . . . .	52
2.6	Application to Cytometry Data . . . . .	55
<b>3</b>	<b>Segmentation</b>	<b>59</b>
3.1	Introduction . . . . .	60
3.2	Background . . . . .	62
3.3	Model & Properties . . . . .	65
3.4	Quantities of Interest . . . . .	68
3.5	Edge Status & Structure Comparisons . . . . .	70
3.6	Simulations . . . . .	72
3.7	Applications . . . . .	75
3.8	Discussion . . . . .	79
<b>4</b>	<b>Extensions</b>	<b>81</b>
4.1	Covariates . . . . .	83
4.2	Temporal Dependence . . . . .	90
4.3	Prior Distribution on Segmentations . . . . .	96
<b>5</b>	<b>En Bref</b>	<b>103</b>





# O

## Introduction

---

### Foreword .....

A network is a system of interconnected entities. These entities can be of various natures. They might for instance share their feelings on the Internet along with pictures of cats, in which case the resulting network is often qualified of *social*. But they might as well exchange information through copper wires and optical fibres, or via chemical compounds. The bottom line is that networks can be found at all scales, spanning the entire planet or the confined spaces of a cell. Once these networks have been observed, studying their properties is of crucial interest, as it can unravel deep knowledge about underlying mechanisms. Some might say it could even lead to the answer to life, the universe, and everything. But that is not our point.

In the problems we are interested in, the network is not observed. We just have access to snapshots giving the individual states of the involved entities. The idea is to retrieve the hidden network organizing the system, that we assume to exist, from these snapshots. Let us consider the following thought experiment as an example. You put ten imaginary people in a room. As the mastermind behind this *Gedankenexperiment*, you give these people instructions before leaving the imaginary room. They have to choose an order between them. Then, the first person tosses a coin. If it is a heads, this person sits otherwise they stand. The next person tosses a coin. If it is a heads, they behave like the person before. Otherwise, an other coin toss decides if they stand or sit, and so on. When you get back in the room, you can only see the posture of each person. Nonetheless, if this procedure is repeated many times with the same order, you might finally be able to make a good guess on the order your imaginary friends had chosen, based on the different configurations that you have seen. This is called *network inference*. Now, imagine that the procedure is repeated a hundred times, and that participants are allowed to change their order three times during the whole experiment. The changes can only be made between different runs of the procedure. Your goal is now to determine when the three changes occurred. This is called *change-point*



*detection*. These are basically the two problems that provided an occupation for yours truly during the passed few years.

## Graphical Models & Bayesian Inference

From a more formal point of view, networks find a natural representation in mathematical objects called *graphs*. A graph is made of a set of vertices, that are meant to represent the entities of interest, and a collection of edges linking these vertices. These edges can either be directed or not. In the example of our thought experiment, each vertex symbolises a participant, and directed edges are drawn between consecutive participants in the chosen order. Network inference supposes that, while being impossible to observe directly, the underlying graph is visible through the distribution of the quantities that we actually observe. Thus, the probabilistic models deployed in such circumstances have to take that into account. This leads to the development of the so-called *graphical models*. The rationale behind these models is to depict the conditional independence properties of a multivariate distribution by means of a graph. The graph acts as a more easy to handle proxy encoding abstract probabilistic properties in a visual manner. The model used in the example above belongs to one of the simplest classes of graphical models called *Markov chains*. Knowing the posture of a person makes the knowledge about all the previous participants in the order irrelevant to predict the posture of the following participants. The general framework of graphical models is meant to extend this reasoning to arbitrary dependence structures.

Performing network inference in graphical models requires to consider the graph itself as a parameter in a larger model, that includes the graph, the distribution of the observations and the observations themselves. In a Pirandellian twist, this model is itself a graphical model, whose graphical representation is given below.



Our choice was to consider this model from a Bayesian point of view. This decision was motivated by the fact that it would be inconvenient to deal with uncertainty on a graph in a frequentist setting because of their discrete nature. The posterior distributions provided by Bayesian inference are more practical than point estimates in this case. It was therefore necessary to specify prior distributions on graphs and distributions. Network inference then boils down to computing the posterior distribution on the space of graphs, conditionally to the observations. Said like this, it does not seem like much, but we are actually dealing with a distribution on a frighteningly large discrete set. For instance, considering all possible undirected graphs on  $p = 10$  vertices, we are faced with a set of size  $2^{p(p-1)/2} = 2^{45} \approx 3.5 \cdot 10^{13}$ , and usual problems might involve many more vertices. Performing exact inference under these circumstances would not be possible in a time amounting to less than a few geological eras.

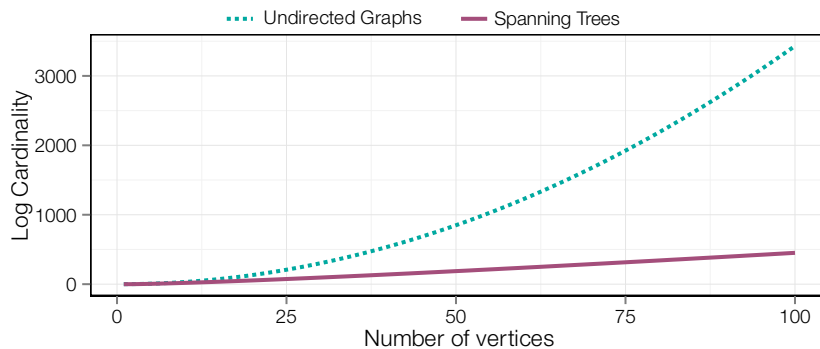
## Exact Inference through Algebra

Whenever exact inference is out of reach, sampling methods are one road to go down. As an example, let us imagine that we are given a probability distribution on the set of undirected graphs and that we want to evaluate the probability associated to the set of graphs with one connected component. We can try to compute this probability directly, but this might turn

out to be a quite tedious calculation if we consider more than a handful of vertices. Ruling out exact computation, an estimate of this probability can be obtained by sampling a lot of graphs from the distribution and looking at the proportion of these graphs that were in our set of interest, namely the graphs with one connected component. If we are sufficiently thorough in our exploration of graphs, we get a sketch of the distribution landscape that is accurate enough to produce a good estimate of the desired integral. This is called Monte Carlo integration. In cases where direct sampling is intractable, the roundabout way is to build a Markov chain whose stationary distribution is the targeted distribution. After a burnin period, samples of this chain can be used as surrogates for independent samples of the initial distribution. This is a brief description of the general principle behind Markov Chain Monte Carlo (MCMC) methods. Taking random walks in vast, jagged spaces, you might wander infinitely before getting anywhere, or find yourself caught in some local topographic feature. In spite of these potential pitfalls, MCMC algorithms have been extensively and successfully used in a variety of contexts, including graphical models.

We made the decision to follow a different path, shaded by trees, and taking short-cuts through algebraic loopholes. Our roadmap for graph-related inference was to focus on situations where exact calculations were tractable, with spanning trees as a common thread. We define such graphs as connected undirected graphs without any cycle. They can alternatively be seen as the sparsest connected graphs, or the most connected graphs without cycle. The basic principle of our approach on network inference was to replace a partial exploration of the full set of graphs by an exhaustive exploration of the subset made of spanning trees. As can be seen in Figure 1, the latter only represent a small fraction of undirected graphs. But in many practical cases, it is actually not necessary to allow all possible structures when trying to extract a network from the data. The Occam's razor principle states that assumptions introduced to explain a phenomenon must not be multiplied beyond necessity, and sparsity is a feature that is often sought-after in network inference. Restricting our attention to the bare spanning trees might therefore not be as ludicrous as it could seem at first sight.

As lovely as woodland paths can be, the main appeal of spanning trees, as far as we are concerned, lies in their specific algebraic properties. In the 19th century, Arthur Cayley and Gustav Kirchhoff both looked into the combinatorial analysis of spanning trees and showed



**Figure 1** – Log-cardinality of the set of undirected graphs and spanning trees as a function of the number of vertices.



that the number of trees embedded in a given graph could be computed in polynomial time, as the determinant of the Laplacian matrix associated to this graph. This result is either referred to as Cayley’s formula, Kirchhoff’s theorem or, in a less eponymous manner, as the Matrix-Tree theorem. In particular, it states that there are  $p^{p-2}$  spanning trees on  $p$  vertices. An extension of this theorem to weighted trees can be derived, under the assumption that the weight of a tree can be expressed as the product of the weights of its edges. Summation over spanning trees can therefore be performed in a reasonable amount of time and that mainly motivated our decision to use them for network inference.

Other situations involving summation over gigantic discrete sets are met by similar fortuitous algebraic blessing. Change-point detection is one of these situations. In this case, the discrete parameter is the partition of the observations into contiguous temporal segments, and summing over the  $\binom{N-1}{K-1}$  segmentations of time  $\llbracket 1; N \rrbracket$  into  $K$  segments can actually be performed in  $O(KN^2)$  time. Our goal is not to dress an extensive list of algebraic tricks. But it turns out that the aforementioned ones can be combined to perform efficient detection of change-points in the dependence structure of a multivariate time-series, which was our second objective.

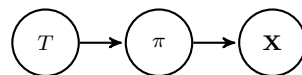
## Outline .....

### Chapter 1

This chapter is meant to supply the background underpinning Chapters 2, 3 and 4. The first section gives a detailed presentation of graphical models. It introduces Markov properties with respect to undirected graphs, as well as various extensions required for network inference. A few words are dedicated to directed graphical models for the sake of completeness, despite no further use hereabouts. A second section presents two algebraic results that are respectively used in Chapters 2 and 3 to perform summations over spanning trees and segmentations. This chapter can be read independently from the others. Some results are redundantly stated in Chapters 2 and 3 as they are based on submitted and accepted articles.

### Chapter 2

Chapter 2 is concerned with network inference. In such problems, the dependence structure of the multivariate distribution behind the observations has to be explicitly taken into account, and that is exactly what graphical models are meant to do. Following on from the works of Meilă & Jaakkola (2006), Kirshner (2007) and Lin et al. (2009), we make use of the hyper Markov property introduced by Dawid & Lauritzen (1993) to provide a full and formal framework for Bayesian inference in tree-structured graphical models. We consider a hierarchical model in which the top level is made of a spanning tree  $T$ , leading to a distribution  $\pi$  satisfying the Markov property with respect to this graph, and finally to observations  $\mathbf{X}$  drawn according to  $\pi$ .



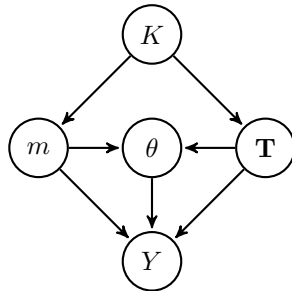
Advantage is taken of the Matrix-Tree theorem presented in Section 1.2.1 and its generalisation to forests to show that the inference of this model can basically be performed in

cubic time with respect to the number of vertices. In particular, the posterior probability for a given edge to appear in a random tree is derived in an exact manner, allowing for arbitrary prior edge appearance probability. We provide a new proof of a result formulated by Kirshner (2007) on the complexity of posterior edge probability computation. A simulation study addresses the influence of the tree-assumption on the accuracy of structure inference for non-tree-structured graphical models. The last section presents an application to flow cytometry data on a cellular signalling network.

This chapter is available as a preprint on *arXiv* (Schwaller et al., 2015). We also worked on a variation of the approach presented here in which  $\pi$  is integrated out through a BIC approximation, published in the french journal *Revue d'intelligence artificielle* (Schwaller & Robin, 2015).

### Chapter 3

We build on the network inference approach develop in Chapter 2 to perform change-point detection in the dependence structure of a multivariate time-series. The global model in which the inference is performed is depicted below.



For a number of segments fixed to  $K$ , a segmentation  $m$  of  $\llbracket 1; N \rrbracket$  and a series of  $K$  spanning trees  $\mathbf{T} = (T_k)_{k=1}^K$  are drawn. Then each segment of  $m$  is assigned a set of parameters  $\theta$  depending on the structure given by the corresponding tree. Finally, on each segment, observations  $Y$  are drawn independently according to a distribution governed by  $\theta$ . Using a result of Rigail et al. (2012) allowing efficient summation over the segmentations of  $\llbracket 1; N \rrbracket$  under some factorability assumption, we are able to compute quantities such as the posterior probability of a change-point occurring at a given time or posterior edge probabilities over time in  $O(N^2 p^3)$  time. We also provide a way to assess whether the status of an edge (or of the whole graph) remains identical throughout the time-series or not when segmentation  $m$  is given. We benchmarked our approach against a simpler one, in which dependence structure is not explicitly taken into account, on synthetic data. Finally, we tackled two datasets respectively originating from cellular biology and neuroscience.

This chapter has been published as an article in the journal *Statistics & Computing* (Schwaller & Robin, 2016).

### Chapter 4

Chapter 4 presents some extensions and perspectives to the work presented in Chapters 2 and 3.

The first section explains how covariates can be integrated in what has been previously presented. When observations are assumed to be normally distributed, adapting the model to take covariates into account is quite straightforward. We show conjugate priors for the



multivariate multiple linear regression induce hyperdistributions that can be used as a basis to build a compatible family of hyperdistributions. We also derive an expression for the marginal likelihood of the observations on any subset of vertices. Whenever observations cannot be modelled by a multivariate normal distribution, we suggested an approximate approach based on copulas, that we used on microbial ecology data (Jakuschkin et al., 2016).

In the second section, we explain how to introduce temporal dependence in the segmentation model described in Chapter 3. Our goal is to lift the independent process assumption that cannot reasonably be made in many practical cases. Our suggestion is to use Temporal Independence Models (TIMs) introduced by Siracusa (2009). Doing so forces us to leave the framework of undirected models to use their directed counter-parts. Under factorisation assumptions similar to the undirected independent case, we show that inference within TIM whose dependence structures are directed trees can be performed with cubic complexity with respect to the number of vertices.

Finally, we investigate prior specification on segmentations. We make out a non-exhaustive bestiary of priors fitting in our framework. We also describe a different kind of prior in which taking into account an initial guess on the segmentation or information coming from a previous study is straightforward.

# 1

## Background

---

<b>1.1 Graphical Models</b> .....	<b>8</b>
1.1.1 Graphs & Markov Properties	9
1.1.1.a Conditional Independence	9
1.1.1.b Undirected Graphs	10
1.1.1.c Markov Properties	12
1.1.1.d Factorisation	14
1.1.1.e Graphical Models	17
1.1.2 Network Inference in Graphical Models	17
1.1.2.a Hyper Markov Properties for Hyperdistributions	18
1.1.2.b Meta Markov Models	21
1.1.2.c Structural Markov Property for Graph Distributions	23
1.1.3 Directed Graphical Models & Markov Equivalence	24
<b>1.2 Algebra &amp; Algorithms</b> .....	<b>27</b>
1.2.1 Summing over Trees	28
1.2.1.a Undirected Trees	28
1.2.1.b Directed Trees	31
1.2.2 Summing over Segmentations	33

---

The aim of this chapter is to extract the main tools needed in the following chapters from the existing literature. In order to be as self-contained as possible, a reasonable amount of definition is included. This chapter will also serve as a reference concerning the notations used afterwards.

The first section is concerned with graphical models, which are at the center of the work presented here. As suggested by their explicit name, the idea behind these models is to make use of graphs to represent the conditional independence relationships satisfied by the distribution of a multivariate random variable. Each component of the random variable is





pictured by a vertex, and edges are supposed to represent dependences between the different components. It is however of crucial importance to carefully define what properties of the distribution can be derived from the graph, as there is actually more than one way to decode a graph into conditional independence statements. Their being equivalent in many practical cases should not hide the fact that they are not in general. We mainly focus on undirected graphical models, stating the different associated Markov properties and how they relate to each other. This section is mostly based on the work of Dawid & Lauritzen (1993), Lauritzen (1996) and Byrne (2011).

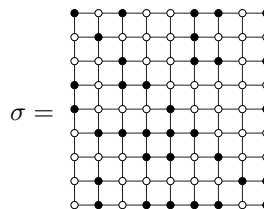
The second section details two algebraic results that motivated our committed stance on exact Bayesian inference in graphical models. Indeed, Bayesian inference typically relies on integrated quantities such as the marginal likelihood of the data, where parameters are integrated out. In graphical models, one of these parameters is the graph depicting the dependences. In this particular case, computing marginal likelihood requires to sum over a set of frightening cardinality. However, some subsets of graphs can be explored efficiently by making extensive use of algebra. Spanning trees form one of these sets, for which summation can be efficiently performed by relying on a result called the Matrix-Tree theorem (Chaiken, 1982). To be able to put this result to good use for Bayesian inference, the prior distributions on the other parameters have to be chosen carefully, and this is where the (hyper) Markov properties given in the previous section come into play. The second result that we present is related to segmentation problems, where an other obviously discrete parameter is involved. Integrating a function over the set of segmentations of  $\llbracket 1; N \rrbracket$  into  $K$  segments is tractable as soon as this function factorises over the segments. An algorithm based on dynamic programming principles can be used to perform the summation in  $O(KN^2)$  time (Rigaill et al., 2012).

## 1.1 Graphical Models

Informally speaking, a graphical model is a probabilistic model whose conditional (in)dependence structure between random variables is given by a graph. This framework has received a fair amount of attention recently, but the ideas can be traced back as far as the end of the 19th century with J. Willard Gibbs. Indeed, one of the scientific areas that popularised graphical models is statistical physics. As an example, let us consider a simple model, named after the physicist Ernst Ising, that can be used to describe a large system of magnetic dipoles (or *spins*). Spins can be in one of the two states  $\{\pm 1\}$ . They are spread on a graph (commonly a lattice) and can only interact with their neighbours. If  $\sigma = (\sigma_1, \sigma_2, \dots)$  is a state of the system giving assignments for all spins, the energy of  $\sigma$  and the associated Gibbsian distribution are respectively given by:

$$\mathcal{H}(\sigma) := -J \sum_{i \sim j} \sigma_i \sigma_j - H \sum_i \sigma_i,$$

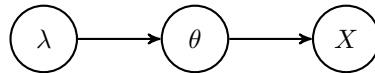
$$p(\sigma) := \frac{1}{Z} \exp(-\mathcal{H}(\sigma)),$$



where  $i \sim j$  means that vertices  $i$  and  $j$  are neighbours in the lattice. The graphical aspect of this model is rather obvious, since the definition of  $\mathcal{H}$  depends on the neighbourhood of each spin. This is an example of *undirected graphical model*. Such models are also called

*Markov random fields.*

Graphical models also naturally arise for instance when designing hierarchical models with sequentially drawn variables. Let us consider a classical Bayesian framework where observations  $X$  are drawn according to a distribution with parameters  $\theta$ , and where  $\theta$  is itself drawn from a distribution with (hyper)parameters  $\lambda$ . This model can be depicted by a directed graph.



Here we have an example of *directed graphical model*. The idea is that the graph indicates a way to factorise the joint probability distribution of all variables as a product of conditional probability distributions. For this factorisation to be possible, the graph cannot have any directed cycle. It has to be a *directed acyclic graph* (DAG). These models are often referred to as *Bayesian networks* (Jensen & Nielsen, 2007). Other classical examples include *Markov chains* and *hidden Markov models* (HMM).

Our main concern will be with undirected graphical models. Nonetheless, as the models that we are especially interested in, namely tree-structured graphical models, can be equally represented in both formalism, we give a brief description of directed graphical models at the end of this section.

### 1.1.1 Graphs & Markov Properties

#### 1.1.1.a Conditional Independence

In graphical models, the probabilistic notion that graphs are meant to represent is conditional (in)dependence. We give the most general definition of conditional independence for a triplet of random variables defined on their respective probability spaces, as stated by Lauritzen (1996).

**Definition 1.1** (Conditional independence). *Let  $X, Y, Z$  be random variables with a joint distribution  $P$ .  $X$  is said to be conditionally independent of  $Y$  given  $Z$ , written*

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

*if, for any measurable set  $A$  in the sample space of  $X$ , there exists a version of  $P(A|Y, Z)$  that is a function of  $Z$  alone.*

Whenever  $P$  admits a density with respect to some product measure, conditional independence can be more easily expressed.

**Proposition 1.1.** *Let  $X, Y, Z$  be random variables with a joint distribution  $P$  that admits a density  $p$  with respect to some product measure. Then*

$$X \perp\!\!\!\perp Y | Z \Leftrightarrow p(x, y | z) = p(x | z)p(y | z) \text{ a.s.}$$

Here are some properties satisfied by the ternary relation  $X \perp\!\!\!\perp Y | Z$ .

**Proposition 1.2.** *Let  $h$  be an arbitrary measurable function on the sample space of  $X$ . Then, it holds that*



- (C1) if  $X \perp\!\!\!\perp Y|Z$  then  $Y \perp\!\!\!\perp X|Z$ ;
- (C2) if  $X \perp\!\!\!\perp Y|Z$  and  $U = h(X)$ , then  $U \perp\!\!\!\perp Y|Z$ ;
- (C3) if  $X \perp\!\!\!\perp Y|Z$  and  $U = h(X)$ , then  $X \perp\!\!\!\perp Y|(Z, U)$ ;
- (C4) if  $X \perp\!\!\!\perp Y|Z$  and  $X \perp\!\!\!\perp W|(Y, Z)$ , then  $X \perp\!\!\!\perp (W, Y)|Z$ .

Whenever the joint distribution of  $X, Y, Z$  admits a continuous and positive density, it also holds that

- (C5) if  $X \perp\!\!\!\perp Y|Z$  and  $X \perp\!\!\!\perp Z|Y$  then  $X \perp\!\!\!\perp (Y, Z)$ .

### 1.1.1.b Undirected Graphs

Let  $V = \{1, \dots, p\}$ ,  $p \geq 2$ , and let  $\mathcal{P}_2(V)$  denote the subsets of  $V$  of size 2.

**Definition 1.2.** For  $E \subseteq \mathcal{P}_2(V)$ ,  $G = (V, E)$  is the undirected graph with vertices  $V$  and edges  $E$ . A graph  $G = (V, E)$  is said to be complete if  $E = \mathcal{P}_2(V)$ . For a subset  $A \subset V$ , the subgraph of  $G$  induced by  $A$  is defined as  $G_A := (A, E_A)$  with  $E_A := \{(i, j) \in E | i \in A, j \in A\}$ . Whenever  $G_A$  is complete for  $A \subseteq V$ ,  $A$  is said to be a clique of  $G$ .

For  $\alpha \in V$  and  $G = (V, E)$ , we define the following set of vertices:

$$\begin{aligned} \text{boundary of } \alpha : & \quad \text{bd}(\alpha) = \{\beta \in V \setminus \{\alpha\} | \{\alpha, \beta\} \in E\}, \\ \text{closure of } \alpha : & \quad \text{cl}(\alpha) = \text{bd}(\alpha) \cup \{\alpha\}. \end{aligned}$$

Vertices in  $\text{bd}(\alpha)$  are also called the *neighbours* of  $\alpha$ .

**Definition 1.3 (Path).** Let  $G = (V, E)$  be an undirected graph and  $\alpha, \beta$  be two distinct vertices in  $V$ . A path from  $\alpha$  to  $\beta$  is a sequence  $\alpha = \gamma_0, \dots, \gamma_n = \beta$ ,  $n \geq 1$ , of distinct vertices such that, for all  $1 \leq i \leq n$ ,  $\{\gamma_{i-1}, \gamma_i\} \in E$ .

A graph is *connected* if,  $\forall \{\alpha, \beta\} \in \mathcal{P}_2(V)$ , there is a path between  $\alpha$  and  $\beta$ .

**Definition 1.4.** A (*spanning*) tree is a connected graph on  $V$  with cliques of size at most 2.

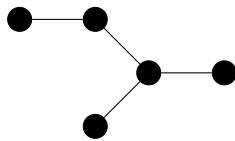


Figure 1.1 – An example of tree.

On undirected graphs, the core notion for graphical models is *separation*.

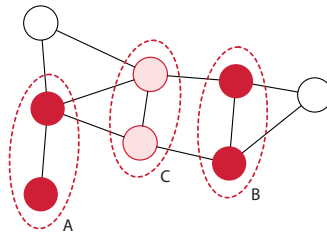
**Definition 1.5 (Separation).** Let  $A, B, C$  be subsets of  $V$ .  $C$  is said to separate  $A$  from  $B$  if any path from  $\alpha \in A$  to  $\beta \in B$  intersects  $C$ .

We let  $A \perp\!\!\!\perp_G B|C$  denote the statement “ $C$  separates  $A$  from  $B$  in  $G$ ”. The ternary relation  $\perp\!\!\!\perp_G$  satisfies Properties (C1) to (C4) given for conditional independence in Proposition 1.2, when  $h$  is replaced by set inclusion. Property (C5) holds whenever  $A, B, C$  are disjoint subsets of  $V$ . Let us illustrate this by rewriting (C2) for  $\perp\!\!\!\perp_G$ :

(C2') if  $A \perp\!\!\!\perp_G B|C$  and  $D \subseteq A$ , then  $D \perp\!\!\!\perp_G B|C$ .

Properties (C1) to (C5) can in fact be seen as pure formal statements related to the notion of “irrelevance”. Any ternary relation on the subsets of a finite set that satisfies properties (C1) to (C5), where  $h$  is replaced by set inclusion, is called a *graphoid*. If (C5) does not hold, it is only a *semi-graphoid*. The name *graphoid* goes back to Pearl & Paz (1987) who noticed that these properties were in fact perfectly captured by graphs. The rationale behind graphical models is to use a graphoid defined by a graph to represent the graphoid (or semi-graphoid) induced by a probability distribution.

*Example 1.1.*  $A$  and  $B$  are separated by  $C$ .



Before actually connecting the notions of conditional independence and separation, we describe a particular class of undirected graphs, which are of particular interest because they can be recursively broken down into their maximal cliques.

**Definition 1.6** (Decomposition). *A pair  $(A, B)$  of subsets of  $V$  is said to be a decomposition of  $G$  if  $V = A \cup B$ , the subgraph induced by  $G$  on  $A \cap B$  is complete and  $A \cap B$  separates  $A$  from  $B$ . If  $A$  and  $B$  are both proper subsets of  $V$ , the decomposition is said to be proper.*

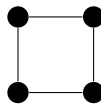
*A graph  $G$  is said to be decomposable if it is either complete or if there exists a proper decomposition of  $G$  into two decomposable subgraphs.*

Notice that the definition of a decomposable graph is recursive. From a conceptual point, the equivalent notion of chordality is easier to handle.

**Definition 1.7** (Chordal graphs). *A graph  $G$  is chordal if all cycles of four or more vertices have a chord, which is an edge that is not part of the cycle but connects two vertices of the cycle.*

**Proposition 1.3** (Lauritzen, 1996, Prop. 2.5).  *$(G \text{ is decomposable}) \Leftrightarrow (G \text{ is chordal})$*

*Example 1.2.* The smallest example of non-decomposable graph is the cycle on four vertices.



Whenever a graph is decomposable, one can build the set of its *minimal complete separators*. Let  $G$  be a decomposable graph. The set  $\mathcal{S}$  of minimal separators for  $G$  is given by the following algorithm. We begin with an empty set of separators  $\mathcal{S} = \emptyset$ . We consider a proper decomposition  $(A, B)$  of  $G$  such that  $A \cap B$  is of minimal cardinality. If there is no



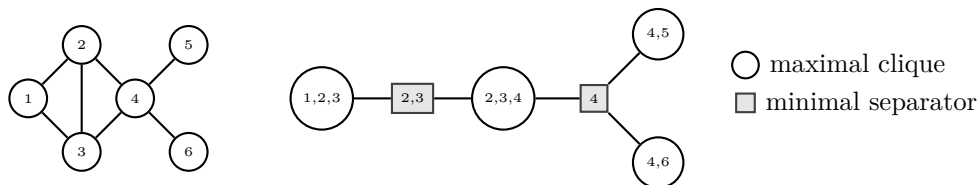
such decomposition, it means that  $G$  is complete and we stop. Otherwise, we add  $A \cap B$  to  $\mathcal{S}$  and apply the same procedure to the subgraphs  $G_A$  and  $G_B$ . For any subset  $S \in \mathcal{S}$ , we let  $\nu(S)$  denote the number of times it appeared in the procedure. An other product of this algorithm is the set  $\mathcal{C}$  containing the maximal cliques of  $G$ .

This is often referred to as the *junction tree algorithm*. Indeed, the result of this algorithm can be represented by a *factor graph*, *i.e.* a bipartite graph whose vertices are indexed by  $\mathcal{C}$  and  $\mathcal{S}$ . We put an edge between  $C \in \mathcal{C}$  and  $S \in \mathcal{S}$  if and only if  $C \cap S = S$ . The graph obtain this way is a tree, hence the name given to this algorithm (see Example 1.3). If  $d(S)$  denotes the degree of the node corresponding to  $S$  in the junction tree, we have that  $\nu(S) = d(S) - 1$ . We will see in Section 1.1.1.d that junction trees can be used to obtain an explicit form for the factorisation of a distribution.

This definition of the junction tree of a decomposable graph is different from the one given in (Lauritzen, 1996). It is sometimes referred to as an ‘Almond tree’, after Almond & Kong (1993).

---

*Example 1.3.* A decomposable graph and its junction tree.



### 1.1.1.c Markov Properties

Let  $X = (X_1, \dots, X_p)$  be a random vector taking values in a product space  $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$ . The set of probability distributions on  $\mathcal{X}$  is denoted by  $\Pi(\mathcal{X})$ . For any  $A \subset V$ , we let  $X_A$  denote  $(X_\alpha)_{\alpha \in A}$ . The same notation is used for any  $x \in \mathcal{X}$ . In the following, for any subsets  $A, B, C \subset V$ , we will use the short notation  $A \perp\!\!\!\perp B | C$  for  $X_A \perp\!\!\!\perp X_B | X_C$ .

**Definition 1.8** (Markov properties). *Let  $G = (V, E)$  be an undirected graph. A probability measure  $P$  on  $\mathcal{X}$  is said to obey*

(P) *the pairwise Markov property relative to  $G$ , if for any pair  $(\alpha, \beta) \in V^2$  such that  $\{\alpha, \beta\} \notin E$ ,*

$$\alpha \perp\!\!\!\perp \beta | V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property relative to  $G$ , if for any vertex  $\alpha \in V$ ,*

$$\alpha \perp\!\!\!\perp V \setminus cl(\alpha) | bd(\alpha);$$

(G) *the global Markov property relative to  $G$ , if for any triple  $(A, B, S)$  of disjoint subsets of  $V$  such that  $S$  separates  $A$  from  $B$  in  $G$ ,*

$$A \perp\!\!\!\perp B | S.$$

The Markov properties  $(P)$ ,  $(L)$  and  $(G)$  are related as described in the proposition below.

**Proposition 1.4** (Pearl & Paz, 1987). *For any undirected graph  $G$  and any probability distribution  $P$  on  $\mathcal{X}$ , it holds that*

$$(G) \Rightarrow (L) \Rightarrow (P).$$

*If  $P$  has a positive and continuous density with respect to a product measure, then*

$$(G) \Leftrightarrow (L) \Leftrightarrow (P).$$

*Proof.* We prove  $(G) \Rightarrow (L) \Rightarrow (P)$  in the general case. The proof for  $(P) \Rightarrow (G)$  in the case where  $P$  has a positive and continuous density with respect to a product measure is obtained through the Hammersley-Clifford theorem (Theorem 1.1 in Section 1.1.1.d).

$(G) \Rightarrow (L)$ : For  $\alpha \in V$ , any path from  $\alpha$  to a vertex in  $V \setminus \text{cl}(\alpha)$  has to go through a neighbour of  $\alpha$ . Therefore,  $\text{bd}(\alpha)$  separates  $\alpha$  from  $V \setminus \text{cl}(\alpha)$  and  $(G) \Rightarrow (L)$ .

$(L) \Rightarrow (P)$ : We suppose that  $(L)$  is true. Let  $\alpha$  and  $\beta$  be two distinct vertices of  $V$  such that  $\{\alpha, \beta\} \notin E$ . By  $(L)$ ,  $\alpha \perp\!\!\!\perp V \setminus \text{cl}(\alpha) \mid \text{bd}(\alpha)$ . As  $\{\alpha, \beta\} \notin E$ ,  $\beta \in V \setminus \text{cl}(\alpha)$ . By statement (C3), we obtain that

$$\alpha \perp\!\!\!\perp V \setminus \text{cl}(\alpha) \mid \underbrace{\text{bd}(\alpha) \cup (V \setminus \{\text{cl}(\alpha) \cup \beta\})}_{V \setminus \{\alpha, \beta\}}$$

and (C2) gives that  $\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\}$ .  $\square$

Whenever statement (C5) holds, all Markov properties are equivalent. But when it does not hold, it is possible to find distributions that satisfy one but not the other. We give an example of such a distribution below.

---

*Example 1.4.* Let  $X, Y, Z$  be three random variables with values in  $\{0, 1\}$ . Let  $X = Y = Z$  and  $P(X = 0) = P(X = 1) = 1/2$ . The probability distribution of  $(X, Y, Z)$  satisfies the pairwise Markov property but not the local Markov property with respect to the graph given below.



Indeed, we have that  $X \perp\!\!\!\perp Y \mid Z$  and  $X \perp\!\!\!\perp Z \mid Y$ , as  $X = Y = Z$  thus leading to trivial statements. But conditioning on the neighbours of  $X$ , *i.e.* not conditioning at all, does not make it independent from  $Y$  and  $Z$ .

---

For the complete graph on  $V$ , it is not possible to find disjoint subsets  $A, B, S$  such that  $S$  separates  $A$  from  $B$ . Thus, any distribution is globally Markov with respect to the complete graph. More generally, if  $P$  is globally Markov with respect to a graph  $G = (V, E)$  and if  $G' = (V, E')$  is such that  $E \subseteq E'$ , then  $P$  is globally Markov with respect to  $G'$ . Whenever a graph  $G$  perfectly describes the conditional independence properties of a distribution  $P$ , we are talking about *Markov faithfulness*.

**Definition 1.9** (Markov faithfulness). *A distribution  $P$  is said to be Markov faithful to a graph  $G$  if, for any triple  $(A, B, S)$  of disjoint subsets of  $V$ , it holds that*



$$A \perp\!\!\!\perp_G B | S \Leftrightarrow A \perp\!\!\!\perp B | S.$$

Markov faithfulness forbids the distribution to satisfy any conditional independence statement that is not encoded in the graph. If  $P$  is Markov faithful to a graph  $G$ ,  $G$  is said to be a *perfect map* (or *P-map*) for  $P$ . If only the implication

$$A \perp\!\!\!\perp_G B | S \Rightarrow A \perp\!\!\!\perp B | S.$$

holds, then  $G$  is said to be an *independence map* (or *I-map*). On the contrary, if we have that

$$A \perp\!\!\!\perp_G B | S \Leftarrow A \perp\!\!\!\perp B | S.$$

then  $G$  is called a dependence map (or, as you might have guessed by now, a *D-map*). A P-map is both a D-map and an I-map. Saying that  $P$  is Markov with respect to  $G$  is equivalent to saying that  $G$  is an I-map for  $P$ . There is just a change of focus in the terminology.

#### 1.1.1.d Factorisation

Closely related to conditional independence is the property of factorisation, as can be seen in Proposition 1.1.

**Definition 1.10** (Factorisation). *A probability measure  $P$  on  $\mathcal{X}$  is said to factorise over  $G$  (or to satisfy (F)) if it admits a density  $p$  with respect to some product measure on  $\mathcal{X}$  of the form*

$$p(x) = \prod_{A \in \mathcal{A}} \psi_A(x_A), \quad \forall x \in \mathcal{X}, \quad (F)$$

where  $\mathcal{A}$  are complete subsets of  $G$  or, equivalently, if

$$p(x) = \prod_{C \in \mathcal{C}} \tilde{\psi}_C(x_C), \quad \forall x \in \mathcal{X}, \quad (F)$$

where  $\mathcal{C}$  are the maximal cliques of  $G$ .

The relationship between property (F) and the different Markov properties is given by the following proposition.

**Proposition 1.5** (Lauritzen, 1996, Prop. 3.8). *For any undirected graph  $G$  and any probability distribution  $P$  on  $\mathcal{X}$ , it holds that*

$$(F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P).$$

If a distribution with density  $p$  is such that,  $\forall x \in \mathcal{X}$ ,  $p(x) > 0$ , we have seen that all Markov properties were equivalent. In this case, we also have that (P) implies (F).

**Theorem 1.1** (Hammersley & Clifford, 1971). *A probability distribution  $P$  with positive and continuous density  $p$  with respect to some product measure on  $\mathcal{X}$  satisfies the pairwise Markov property with respect to an undirected graph  $G$  if and only if it factorises according to  $G$ . Then*

$$(F) \Leftrightarrow (G) \Leftrightarrow (L) \Leftrightarrow (P).$$

*Proof.* The proof, drawn from (Lauritzen, 1996), is based on Möbius inversion lemma.

**Lemma** (Möbius inversion). *Let  $\phi$  be a function defined on the subsets of  $V$ , taking values in an Abelian group. We let  $\zeta$  be the function defined by*

$$\zeta(A) := \sum_{B \subseteq A} (-1)^{|A \setminus B|} \phi(B), \quad \forall A \subseteq V.$$

Then, it holds that

$$\phi(A) = \sum_{B \subseteq A} \zeta(B), \quad \forall A \subseteq V.$$

We assume that (P) is true. Let  $x^*$  be an arbitrary but fixed element of  $\mathcal{X}$ . For  $A \subseteq V$ , we define  $\phi_A$  and  $\zeta_A$  by

$$\begin{cases} \phi_A(x) &= \log p(x_A, x_{V \setminus A}^*), \\ \zeta_A(x) &= \sum_{B \subseteq A} (-1)^{|A \setminus B|} \phi_B(x), \end{cases} \quad \forall x \in \mathcal{X}.$$

Both  $\phi_A$  and  $\zeta_A$  are in fact functions of  $x_A$ . Möbius inversion lemma yields that

$$\phi_V(x) = \log p(x) = \sum_{A \subseteq V} \zeta_A(x), \quad \forall x \in \mathcal{X}.$$

It remains to show that  $\zeta_A \equiv 0$  as soon as  $A$  is not complete. Let  $A$  be a subset containing two distinct vertices  $\alpha$  and  $\beta$  such that  $\{\alpha, \beta\} \notin E$ . Subsets of  $A$  can be sorted according to their containing  $\alpha$  and  $\beta$ , so that, if  $C = A \setminus \{\alpha, \beta\}$ ,

$$\zeta_A = \sum_{B \subseteq C} (-1)^{|C \setminus B|} (\phi_B - \phi_{B \cup \{\alpha\}} - \phi_{B \cup \{\beta\}} + \phi_{B \cup \{\alpha, \beta\}}).$$

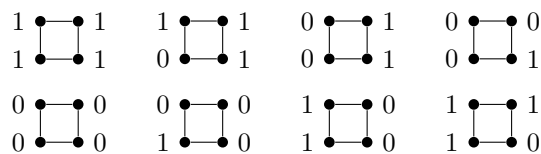
Let  $D = V \setminus \{\alpha, \beta\}$ . Then, using the pairwise Markov property,  $\forall x \in \mathcal{X}$ ,

$$\begin{aligned} \phi_{B \cup \{\alpha, \beta\}}(x) - \phi_{B \cup \{\alpha\}}(x) &= \log \frac{p(x_B, x_\alpha, x_\beta, x_{D \setminus B}^*)}{p(x_B, x_\alpha, x_\beta^*, x_{D \setminus B}^*)} = \log \frac{p(x_\alpha | x_B, x_{D \setminus B}^*) p(x_\beta, x_B, x_{D \setminus B}^*)}{p(x_\alpha | x_B, x_{D \setminus B}^*) p(x_\beta^*, x_B, x_{D \setminus B}^*)} \\ &= \log \frac{p(x_\alpha^* | x_B, x_{D \setminus B}^*) p(x_\beta, x_B, x_{D \setminus B}^*)}{p(x_\alpha^* | x_B, x_{D \setminus B}^*) p(x_\beta^*, x_B, x_{D \setminus B}^*)} = \log \frac{p(x_B, x_\alpha^*, x_\beta, x_{D \setminus B}^*)}{p(x_B, x_\alpha^*, x_\beta^*, x_{D \setminus B}^*)} \\ &= \phi_{B \cup \{\beta\}}(x) - \phi_B(x), \end{aligned}$$


therefore showing that  $\zeta_A \equiv 0$ . We have proved that (F) is true.  $\square$

In the general case, (F) is stronger than (G), has shown by the example below.

*Example 1.5.* (Moussouris, 1974) Here is an example of a distribution satisfying (G) but not (F). Consider the uniform distribution on the 8 (out of 16) configurations of  $\{0, 1\}^4$  displayed below.





To show that this distribution is globally Markov with respect to , one only has to show that two opposite vertices are independent conditionally on the other two. But when the values of two diagonally opposite vertices are fixed, so is the value of one of the other two vertices, the global Markov property is therefore satisfied.

We now show that the density does not factorise. Let us assume that it does factorise. Then

$$p(0, 0, 0, 0) = 1/8 = \psi_{12}(0, 0)\psi_{23}(0, 0)\psi_{34}(0, 0)\psi_{41}(0, 0),$$

so these factors are all strictly positive. Similarly reasoning on all 8 possible configurations yields that all factors  $\psi$  are strictly positive. This contradicts the fact that 8 configurations have probability 0.

For decomposable graphs, (F) and (G) are equivalent even if  $p$  is not positive, hence the fondness caused by decomposability.

**Proposition 1.6** (Lauritzen, 1996, Prop. 3.19). *Let  $G$  be a decomposable graph. Then it holds that*

$$(F) \Leftrightarrow (G).$$

Moreover, whenever  $G$  is decomposable, the junction tree algorithm described in Section 1.1.1.b can be used to obtain an explicit formula for the density of a distribution that factorises over  $G$  using marginal distributions over complete subsets. Indeed, if  $P$  is a distribution with density  $p$  that factorises over a decomposable graph  $G$ , and if  $\mathcal{S}$  and  $\mathcal{C}$  respectively denote the set of minimal separators and maximal cliques yielded by the junction tree algorithm, it can be shown that

$$p(x) \prod_{S \in \mathcal{S}} p_S(x_S)^{\nu(S)} = \prod_{C \in \mathcal{C}} p_C(x_C), \quad \forall x \in \mathcal{X},$$

where, for  $A \subseteq V$ ,  $p_A$  is the marginal density of  $p$  on  $X_A$ . If  $p$  is positive, then

$$p(x) = \frac{\prod_{C \in \mathcal{C}} p_C(x_C)}{\prod_{S \in \mathcal{S}} p_S(x_S)^{\nu(S)}}, \quad \forall x \in \mathcal{X}. \quad (1.1)$$

*Example 1.6.* On the graph given in Example 1.3, (1.1) would yield

$$p(x) = \frac{p(x_1, x_2, x_3)p(x_2, x_3, x_4)p(x_4, x_5)p(x_4, x_6)}{p(x_2, x_3)p(x_4)^2}.$$

Whenever  $G = (V, E)$  is a tree, the factorisation is directly given by

$$p(x) = \frac{\prod_{\{i,j\} \in E} p_{ij}(x_i, x_j)}{\prod_{i \in V} p_i(x_i)^{d(i)-1}} = \prod_{i \in V} p_i(x_i) \prod_{\{i,j\} \in E} \frac{p_{ij}(x_i, x_j)}{p_i(x_i)p_j(x_j)}, \quad \forall x \in \mathcal{X},$$

where  $d(i)$  stands for the degree of vertex  $i$  in  $G$ .

### 1.1.1.e Graphical Models

In the previous sections, we have seen how graphs could be used to represent the conditional independence properties of a distribution. This connection between graphs and distributions can be used to define the notion of graphical model.

For any undirected graph  $G$ , we let  $\mathcal{M}(G)$  denote the set of distributions on  $\mathcal{X}$  that are globally Markov with respect to  $G$ . What we call a *model* is a family of distributions on  $\mathcal{X}$ . Thus, a graphical model is a family of distributions sharing some Markov property with respect to a graph. The formal definition that we adopt here is the following.

**Definition 1.11.** *A graphical model is given by a graph  $G$  describing its structure and a family of distributions  $\mathcal{F} \subset \mathcal{M}(G)$  globally Markov with respect to  $G$ .*

If all distributions in  $\mathcal{F}$  admit a positive density, one might use any of the Markov properties, as they are all equivalent. Alternatively, one might require the distributions in  $\mathcal{F}$  to be Markov faithful to  $G$ . This definition is much more restrictive.

### 1.1.2 Network Inference in Graphical Models

Many scientific fields are concerned with retrieving an underlying network that is supposed to exist between entities of interest. This kind of problem can be given many names, but ours is network inference. In this situation, the graph and the distribution of the observations are themselves parameters of a hierarchical model that can be depicted as follows.

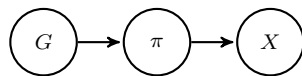


Figure 1.2 – Hierarchical model for network inference.

In this model,  $\pi$  is assumed to be globally Markov with respect to  $G$ , and  $X$  is distributed according to  $\pi$ . The distribution  $\pi$  is drawn from a family  $\mathcal{F}_G$  of distributions that satisfy the global Markov property with respect to  $G$ . Thus,  $(G, \mathcal{F}_G)$  is a graphical model, and network inference can be seen as a model selection problem between different graphical models.

There exists an abundant literature discussing network inference in a frequentist framework, maybe even vaster than its Bayesian equivalent. As we still decided to go the Bayesian way, this paragraph has no further claim than mentioning a very small fraction of this literature. Gaussian Graphical Models (GGMs) have been of particular interest because of the specific properties of normal distributions. Indeed, for such distributions, conditional independence properties can directly be read in the inverse covariance (or precision) matrix. Learning the dependence structure of the distribution therefore boils down to identifying the null entries in the precision matrix. This task can be seen as a set of variable selection problems, where the neighbours of each variable have to be selected individually. In a frequentist setting, sparse regularisation techniques such as the Lasso (Tibshirani, 2011) are well suited for this task. Various algorithms have been especially developed in this context, including ‘neighbourhood selection’ (Meinshausen & Bühlmann, 2006) and the ‘Graphical Lasso’ (Friedman et al., 2008). Score-based approaches have also been considered for network inference. In the case of DAGs for instance, Heckerman & Chickering (1995) introduced the BDe metric that was used in a greedy search algorithm called GES (Chickering, 2002).



When considering Bayesian inference, a prior distribution has to be specified for  $G$ . Without any restriction on the support of this distribution, exact inference can only be contemplated as long as there are no more than thirty or so variables of interest (Parviainen & Koivisto, 2009). In this context, sampling approaches such as MCMC have been extensively used to get access to the posterior graph distribution, whether it be for undirected graphs (Roverato, 2002; Atay-Kayis & Massam, 2005; Green & Thomas, 2013) or DAGs (Madigan et al., 1995; Friedman & Koller, 2003; Niinimäki et al., 2011). Some interest has also been dedicated to specific subsets of graphs, such as spanning trees (Chow & Liu, 1968; Meilä & Jordan, 2001; Meilä & Jaakkola, 2006; Kirshner, 2007; Lin et al., 2009). In this particular case, the inference can be performed with cubic complexity with respect to the number of variables, as we shall see in the following chapter.

Defining a full and formal framework for Bayesian inference in tree-structured graphical models, we will find it useful to extend Markov properties to the distributions of  $\pi$  and  $G$ . These extensions are respectively called hyper and structural Markov properties and are described in Sections 1.1.2.a and 1.1.2.c.

### 1.1.2.a Hyper Markov Properties for Hyperdistributions

Hyper Markov properties were introduced by Dawid & Lauritzen (1993) as an extension of Markov properties for distributions on sets of Markov distributions. In order to establish a clear distinction between the two levels of involved distributions, a distribution for  $\pi$  will be called a *hyperdistribution*. Whenever  $\pi$  is taken from a parametric family of distributions depending on a set of parameters  $\theta \in \Theta$ , a hyperdistribution  $\rho$  can be specified by giving a distribution on  $\Theta$ . This distribution can itself be taken in a parametric family, in which case these parameters are traditionally called hyperparameters in a Bayesian setting. Hence the term *hyperdistribution* that we have adopted to refer to  $\rho$ . In order to remain as general as possible, all results are nonetheless given in a non-parametric framework.

Let  $G = (V, E)$  be a decomposable graph and let  $m_G = (G, \mathcal{F}_G)$  be a graphical model with structure  $G$ . We consider a hyperdistribution  $\rho$  defined on  $\mathcal{F}_G$ . For  $\pi \in \mathcal{F}_G$  and  $A, B \subseteq V$ , we let  $\pi_A$  denote the marginal distribution obtained from  $\pi$  on the variables  $X_A$ , and  $\pi_{B|A}$  denote the collection of conditional distributions of  $X_B|X_A$  under  $\pi$ . We also let  $\rho_A$  and  $\rho_{B|A}$  respectively denote the marginal hyperdistribution induced by  $\rho$  on  $\pi_A$  and the collection of hyperdistributions induced by  $\rho$  on  $\pi_{B|A}$ .

**Definition 1.12.** A hyperdistribution  $\rho$  is said to be (weak) hyper Markov with respect to  $G$  if, for any decomposition  $(A, B)$  of  $G$ , it holds that, under  $\rho$ ,

$$\pi_A \perp\!\!\!\perp \pi_B | \pi_{A \cap B}. \quad (1.2)$$

As stated by Dawid & Lauritzen (1993), condition (1.2) is equivalent to either

$$\pi_{A|B} \perp\!\!\!\perp \pi_B | \pi_{A \cap B}, \quad \pi_{A|B} \perp\!\!\!\perp \pi_{B|A} | \pi_{A \cap B}.$$

The hyper Markov property can also be characterised through separation.

**Theorem 1.2** (Byrne, 2011, Th. 1.3.1). A hyperdistribution  $\rho$  is (weak) hyper Markov with respect to  $G$  if and only if, for any triple  $(A, B, S)$  of disjoint subsets of  $V$  such that  $S$  separates  $A$  from  $B$  in  $G$ , it holds that, under  $\rho$ ,

$$\pi_{A \cup S} \perp\!\!\!\perp \pi_{B \cup S} | \pi_S.$$

We will actually need a stronger version of the hyper Markov property, for reasons that we will expose shortly after.

**Definition 1.13.** A hyperdistribution  $\rho$  is said to be strong hyper Markov with respect to  $G$  if, for any decomposition  $(A, B)$  of  $G$ , it holds that, under  $\rho$ ,

$$\pi_A \perp\!\!\!\perp \pi_{B|A}.$$

If  $G = (V, E)$  and  $G' = (V, E')$  are such that  $E \subset E'$  and if  $\rho$  is weak hyper Markov with respect to  $G$ , then it is also hyper Markov with respect to  $G'$ . This is not true for the strong hyper Markov property. But as the strong hyper Markov property implies the weak version,  $\rho$  being strong hyper Markov with respect to  $G$  implies that  $\rho$  is weak hyper Markov with respect to  $G'$ .

**Proposition 1.7** (Dawid & Lauritzen, 1993, Prop. 5.1). *Let  $\rho$  be a hyperdistribution that is (weak) hyper Markov with respect to a graph  $G$ . Then, for any decomposition  $(A, B)$  of  $G$ , it holds that*

$$(X_A, \pi_A) \perp\!\!\!\perp (X_B, \pi_{B|A}) | (X_{A \cap B}, \pi_{A \cap B}).$$

If  $\rho$  is strong hyper Markov with respect to  $G$ , for any decomposition  $(A, B)$  of  $G$ , it further holds that

$$(X_A, \pi_A) \perp\!\!\!\perp (X_B, \pi_{B|A}) | X_{A \cap B}.$$

We remind that  $\rho$  is a prior distribution for the distribution  $\pi$  of  $X$ . If we abusively let  $\rho$  either denote the distribution itself or its density with respect to some product measure  $\eta$  on  $\Pi(\mathcal{X})$ , the density of the marginal distribution for  $X$  can be written as

$$p(x) := \int \pi(x) \rho(\pi) d\eta(\pi). \quad (1.3)$$

Whenever  $\rho$  is strong hyper Markov, for any decomposition  $(A, B)$  of  $G$ ,  $X_{A \cap B}$  is enough to isolate  $(X_A, \pi_A)$  from  $(X_B, \pi_{B|A})$ , so that it is possible to integrate over  $\pi$ , and the marginal distribution over  $X$  remains globally Markov with respect to  $G$ . This is not the case when  $\rho$  is only weak hyper Markov with respect to  $G$ .

**Proposition 1.8** (Dawid & Lauritzen, 1993, Prop. 5.6). *If  $\rho$  is strong hyper Markov with respect to  $G$ , the marginal distribution defined in (1.3) is globally Markov with respect to  $G$ .*

If  $G$  is chosen to be decomposable and if  $\rho$  is strong hyper Markov with respect to  $G$ , Propositions 1.6 and 1.8 show that the marginal likelihood factorises over  $G$ . The integral given in (1.3) can then be computed on the maximal cliques  $\mathcal{C}$  and minimal separators  $\mathcal{S}$  of  $G$  as

$$p_C(x_C) = \int \pi_C(x) \rho_C(\pi_C) d\eta(\pi_C), \quad \forall C \in \mathcal{C}, \quad (1.4)$$

$$p_S(x_S) = \int \pi_S(x) \rho_S(\pi_S) d\eta(\pi_S), \quad \forall S \in \mathcal{S}, \quad (1.5)$$

and the complete marginal likelihood is obtained through (1.1).



In the model described in Figure 1.2, we are actually interested in defining a collection of hyperdistributions  $\{\rho^G\}_{G \in \mathcal{G}}$ , where  $\mathcal{G}$  is the support of a graph distribution  $\xi$ . In this situation, it might be desirable for  $\rho^G$  and  $\rho^{G'}$  to be ‘similar’ whenever two graphs  $G$  and  $G'$  of  $\mathcal{G}$  are close. This can be achieved by building what is called a (*hyper*) *compatible family* of hyperdistributions. Let us consider a hyperdistribution on  $\Pi(\mathcal{X})$  such that, for any  $A \subseteq V$ , under  $\rho$ ,

$$\pi_A \perp\!\!\!\perp \pi_{V \setminus A} \mid \rho. \quad (1.6)$$

This means that  $\rho$  is strong hyper Markov with respect to the complete graph over  $V$ .

**Proposition 1.9.** (*Dawid & Lauritzen, 1993, §6.2*) *For any decomposable graph  $G$ , there exists a unique hyperdistribution  $\rho^G$  on  $\mathcal{M}(G)$  that is strong hyper Markov with respect to  $G$  and such that, for every clique  $C$  of  $G$ ,*

$$\rho_C^G = \rho_C.$$

If  $\mathcal{G}$  is a graph family,  $\{\rho^G\}_{G \in \mathcal{G}}$  is a (hyper) compatible family of strong hyper Markov hyperdistributions.

---

*Example 1.7.* We illustrate the hyper Markov property on the special case of normal distributions. We let  $\mathcal{N}_p$  denote the set of multivariate normal distributions on  $\mathcal{X} = \mathbf{R}^p$  with mean vector  $\mathbf{0}_p$ . We also let  $P_p$  denote the set of symmetric positive-definite real-valued matrices of size  $p \times p$ . There is a one-to-one correspondence between  $P_p$  and  $\mathcal{N}_p$  through the map

$$\Upsilon_p : \begin{cases} P_p & \longrightarrow \mathcal{N}_p \\ \Lambda & \longmapsto \mathcal{N}(\mathbf{0}_p, \Lambda^{-1}) \end{cases}$$

where  $\mathcal{N}(\mathbf{0}_p, \Lambda^{-1})$  stands for the multivariate normal distribution with mean vector  $\mathbf{0}_p$  and inverse covariance matrix (or precision matrix)  $\Lambda$ . For any undirected graph  $G$ , we also consider the set  $\mathcal{N}_p^G$  of normal distributions that are globally Markov with respect to  $G$ . As normal distributions are positive, all Markov properties are equivalent and we can actually drop the ‘globally’. The distributions in  $\mathcal{N}_p^G$  can be characterised by a condition on their precision matrices.

*Proposition 1.10.* *For a graph  $G = (V, E)$ , we define*

$$P_p^G = \{\Lambda \in P_p : \forall \{i, j\} \notin E, \Lambda_{ij} = 0\}.$$

*Then, it holds that*

$$\mathcal{N}_p^G = \Upsilon_p(P_p^G).$$

*Proof.* See for instance (Lauritzen, 1996, Prop. 5.2). □

The Wishart distribution  $W_p(\alpha, \Psi)$  with  $\alpha > p - 1$  degrees of freedom and scale matrix  $\Psi$  is defined on  $P_p$  and has density

$$h(\Lambda) = \frac{|\Psi|^{\frac{\alpha}{2}}}{2^{\frac{\alpha p}{2}} \Gamma_p(\frac{\alpha}{2})} |\Lambda|^{\frac{\alpha - p - 1}{2}} e^{-\frac{1}{2} \text{tr}(\Psi \Lambda)},$$

where the multivariate gamma function  $\Gamma_p$  is defined by

$$\Gamma_p\left(\frac{x}{2}\right) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^p \Gamma\left(\frac{x+1-j}{2}\right).$$

Therefore, any Wishart distribution induces a hyperdistribution  $\rho$  on  $\mathcal{N}_p$  through  $\Upsilon_p$ :

$$\rho(\pi) = h(\Upsilon_p^{-1}(\pi)), \quad \forall \pi \in \mathcal{N}_p.$$

In order to build a compatible family of strong hyper Markov hyperdistributions through Proposition 1.9, one has to prove that Condition (1.6) is satisfied for a hyperdistribution on  $\mathcal{N}_p$  induced by a Wishart distribution. For  $\pi \in \mathcal{N}_p$  with precision matrix  $\Lambda$ , classic results on normal distributions state that, for any partitioning  $(A, B)$  of  $V$ ,

$$\begin{aligned} \pi_A &= \mathcal{N}(\mathbf{0}_{|A|}, (\Lambda_{AA} - \Lambda_{AB}\Lambda_{BB}^{-1}\Lambda_{BA})^{-1}), \\ \pi_{B|X_A=x_A} &= \mathcal{N}(-\Lambda_{BB}^{-1}\Lambda_{BA}x_A, \Lambda_{BB}^{-1}). \end{aligned}$$

where  $[\Lambda_{AA}, \Lambda_{AB}, \Lambda_{BA}, \Lambda_{BB}]$  is a block partitioning of  $\Lambda$  according to  $(A, B)$ . Thus, showing that  $\pi_A \perp\!\!\!\perp \pi_{B|A}$  boils down to show that  $(\Lambda_{AA} - \Lambda_{AB}\Lambda_{BB}^{-1}\Lambda_{BA})$  is independent of  $\{\Lambda_{BB}, \Lambda_{AB}\}$ .

*Proposition 1.11.* *Let  $(A, B)$  be a partition of  $V$ , with  $|A| = a$  and let  $\Lambda \sim W_p(\alpha, \Psi)$ . We denote  $\Lambda_{AA \bullet B}$  the matrix  $\Lambda_{AA} - \Lambda_{AB}\Lambda_{BB}^{-1}\Lambda_{BA}$ . Then, it holds that*

$$\Lambda_{AA \bullet B} \perp\!\!\!\perp \{\Lambda_{BB}, \Lambda_{AB}\}.$$

Moreover,

$$\Lambda_{AA \bullet B} \sim W_a(\alpha - p + a, \Psi_{AA}), \quad (1.7)$$

where  $[\Psi_{AA}, \Psi_{AB}, \Psi_{BA}, \Psi_{BB}]$  is a block partitioning of  $\Psi$  according to  $(A, B)$ .

*Proof.* This is a particular case of Theorem 5 in (Geiger & Heckerman, 2002).  $\square$

Therefore, any Wishart distribution  $W_p(\alpha, \Psi)$  can be used to build a compatible family of strong hyper Markov hyperdistributions, and (1.7) states that the clique hyperdistributions of this family are given by

$$\rho_A(\pi_A) = h_A(\Upsilon_a^{-1}(\pi_A)), \quad \forall A \subseteq V,$$

where  $h_A$  stands for the density of  $W_a(\alpha - p + a, \Psi_{AA})$ . Exact formulas can be derived for the marginal likelihoods  $p_C$  and  $p_S$  given in (1.4) and (1.5) using classic results on conjugate priors.

### 1.1.2.b Meta Markov Models

Let  $\rho$  be a hyperdistribution on  $\Pi(\mathcal{X})$ . The support family of  $\rho$  is the subfamily of  $\Pi(\mathcal{X})$  defined by

$$\mathcal{F}_\rho := \{\pi \in \Pi(\mathcal{X}) : \rho(\pi) > 0\}.$$

The support family of a (strong) hyper Markov hyperdistribution has some intrinsic Markov property of its own. Expressing this property requires to define *variation independence*.



**Definition 1.14.** Let  $\varphi$ ,  $\varsigma$  and  $\varpi$  be functions on a common domain  $D$ .  $\varphi$  is said to be conditionally variation independent of  $\varsigma$  given  $\varpi$ , written

$$\varphi \dagger \varsigma | \varpi \quad [D],$$

if, for all  $v \in \varsigma(D)$ ,  $w \in \varpi(D)$ , it holds that

$$\varphi((\varsigma, \varpi)^{-1}(v, w)) = \varphi(\varpi^{-1}(w)).$$

If  $\varpi$  is trivial, we write  $\varphi \dagger \varsigma$ , meaning that,  $\forall v \in \varsigma(D)$ ,  $\varphi(\varsigma^{-1}(v)) = \varphi(D)$ .

Basically, the statement “ $\varphi \dagger \varsigma | \varpi \quad [D]$ ” means that, whenever the value taken by  $\varpi$  is fixed, the range of values that can be attained by  $\varphi$  does not change with the value taken by  $\varsigma$ . Conditional variation independence is a property related to a domain, just like conditional independence is related to a joint probability distribution. It can be seen as weaker than conditional independence, as the lack of variation independence proscribe the corresponding probabilistic independence statement.

**Proposition 1.12** (Dawid & Lauritzen, 1993, Lemma 4.2). *Properties (C1) to (C4) defined for conditional independence in Proposition 1.2 hold for conditional variation independence.*

For  $A, B \subseteq V$ , we consider the following functions on  $\Pi(\mathcal{X})$ :

$$\begin{aligned} \Xi_A &: \pi \longmapsto \pi_A, \\ \Xi_{B|A} &: \pi \longmapsto \pi_{B|A}. \end{aligned}$$

**Definition 1.15.** Let  $G$  be an undirected graph. A model  $\mathcal{F} \subseteq \Pi(\mathcal{X})$  is said to be (weak) meta Markov with respect to  $G$  if, for any decomposition  $(A, B)$  of  $G$  it holds that

$$\Xi_A \dagger \Xi_B | \Xi_{A \cap B} \quad [\mathcal{F}].$$

A model  $\mathcal{F} \subseteq \Pi(\mathcal{X})$  is said to be strong meta Markov with respect to  $G$  if, for any decomposition  $(A, B)$  of  $G$  it holds that

$$\Xi_A \dagger \Xi_{B|A} \quad [\mathcal{F}].$$

Within a strong meta Markov model, any value taken by  $\pi_A$  is logically compatible with any value taken  $\pi_{B|A}$ . The following propositions describe two cases involving meta Markov models.

**Proposition 1.13.** Let  $G$  be a graph. The family  $\mathcal{M}(G)$  made of all the distributions in  $\Pi(\mathcal{X})$  that are Markov with respect to  $G$  is a meta Markov model with respect to  $G$ .

**Proposition 1.14.** Let  $\rho$  be a hyperdistribution that is (strong) hyper Markov with respect to a graph  $G$ . Then the support family  $\mathcal{F}_\rho$  of  $\rho$  is a (strong) meta Markov model with respect to  $G$ .

Therefore, if  $\mathcal{F}$  is not (strong) meta Markov with respect to a graph  $G$ , it will not be possible to find a hyperdistribution with support  $\mathcal{F}$  that is (strong) hyper Markov with respect to  $G$ .

---

*Example 1.8.* The family of multivariate normal distributions on  $\mathbf{R}^p$  is a strong meta Markov model with respect to the complete graph on  $V$  (see Example 1.7).

The family of multinomial distributions on  $[[1; r]]^p$ ,  $r \in \mathbf{N}^*$ , is a strong meta Markov model with respect to the complete graph on  $V$ .

---

### 1.1.2.c Structural Markov Property for Graph Distributions

In the model depicted in Figure 1.2, the graph itself is a parameter, for which we therefore need a prior distribution. Pursuing an approach similar to what Dawid & Lauritzen (1993) had done with hyperdistributions, Byrne (2011) proposed an extension of Markov properties for graph distributions.

We let  $\mathfrak{U}$  denote the set of undirected decomposable graphs on  $V$ . A *covering pair* of  $V$  is a pair of sets  $(A, B)$  such that  $A \cup B = V$ . For any family of graphs  $\mathcal{G} \subseteq \mathfrak{U}$  and for any covering pair  $(A, B)$ , we define  $\mathcal{G}(A, B)$  to be the set of graphs  $G \in \mathcal{G}$  for which  $(A, B)$  is a decomposition. In particular,  $\mathfrak{U}(A, B)$  is the set of decomposable graphs for which  $(A, B)$  is a decomposition. We remind that, for  $G \in \mathfrak{U}(A, B)$ ,  $G_{A \cap B}$  is complete.

**Definition 1.16.** A graph distribution  $\xi$  is said to be *structurally Markov* if, for any covering pair  $(A, B)$  of  $V$ , it holds that

$$G_A \perp\!\!\!\perp G_B \mid \{G \in \mathfrak{U}(A, B)\}.$$

Structurally Markov graph distributions combine nicely with compatible families of strong hyper Markov hyperdistributions.

**Proposition 1.15** (Byrne, 2011, Th. 4.4.4). Let  $\xi$  be a structurally Markov graph distribution and let  $\{\rho^G\}_{G \in \mathfrak{U}}$  be a compatible family of strong hyper Markov distributions. Then, for any covering pair  $(A, B)$  of  $V$ , it holds that

$$(\pi_A, G_A) \perp\!\!\!\perp (\pi_{B|A}, G_B) \mid \{G \in \mathfrak{U}(A, B)\}.$$

Under these conditions, the posterior graph distribution is also structurally Markov.

---

*Example 1.9.* A graph distribution  $\xi$  on the set of trees, denoted  $\mathcal{T}$ , can easily be specified through an edge weight matrix. Indeed, if we are given a symmetric matrix  $\omega := (\omega_{i,j})_{1 \leq i,j \leq p}$  with non-negative entries, a distribution on  $\mathcal{T}$  can be obtained through

$$\xi(T) = \frac{1}{Z} \prod_{\{i,j\} \in E_T} \omega_{i,j}, \quad \forall T = (V, E_T) \in \mathcal{T}, \quad (1.8)$$

where  $Z$  is a normalising constant defined by  $Z := \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \omega_{i,j}$ . In this kind of distributions, the weight of a tree is just the product of the weights of its edges. Thus, an edge shared by two different trees contributes in the same way in both trees. It can easily be shown that the distributions on  $\mathcal{T}$  of the form given in (1.8) are structurally Markov. These distributions form what Byrne & Dawid (2015) call the *clique exponential family* on  $\mathcal{T}$ .

---

The meta Markov property that we defined for families of distributions (or models) can similarly be defined for graph families. For  $A \subset V$ , we consider the following function on  $\mathfrak{U}$ :

$$\mathfrak{G}_A : G \mapsto G_A.$$

**Definition 1.17.** A graph family  $\mathcal{G} \subseteq \mathfrak{U}$  is said to be *structurally meta Markov* if, for any covering pair  $(A, B)$  of  $V$ , it holds that

$$\mathfrak{G}_A \dagger \mathfrak{G}_B \quad [\mathcal{G}(A, B)].$$





For a covering pair  $(A, B)$  of  $V$ , if  $G, G' \in \mathfrak{U}(A, B)$ , we let  $G_A \otimes G'_B$  denote the one and only graph in  $\mathfrak{U}(A, B)$  such that  $(G_A \otimes G'_B)_A = G_A$  and  $(G_A \otimes G'_B)_B = G'_B$ . The structural meta Markov property can equivalently be defined as follows.

**Theorem 1.3** (Byrne, 2011, Th. 4.3.1). *A graph family  $\mathcal{G} \subseteq \mathfrak{U}$  is structurally meta Markov if and only if, for any covering pair  $(A, B)$  of  $V$  and for  $G, G' \in \mathcal{G}(A, B)$ , it holds that*

$$G_A \otimes G'_B \in \mathcal{G}(A, B).$$

Structural and structural meta Markov properties are closely related.

**Theorem 1.4** (Byrne, 2011, Th. 4.3.2). *The support of a structurally Markov graph distribution is a structurally meta Markov family.*

---

*Example 1.10.* The family of spanning trees  $\mathcal{T}$  is a structurally meta Markov family.

---

### 1.1.3 Directed Graphical Models & Markov Equivalence

Sections 1.1.1.b to 1.1.2.c described how graph theory on undirected graphs and conditional independence could be weaved together to produce undirected graphical models. The same results can be obtained for *directed acyclic graphs (DAGs)*, with all the Markov properties defined for distributions, hyperdistributions or graph distributions having directed counterparts. We will not roll out the whole theory of directed graphical models. We will just define the core notion of separation on DAGs, as it is less straightforward than in the undirected case.

**Definition 1.18** (Directed graph). *A directed graph is a pair  $D = (V, \mathcal{E})$ , where  $V$  is a set of vertices and the set of edges  $\mathcal{E}$  is a subset of  $V \times V$  of ordered pairs of distinct vertices. Notice that we do not allow edges from a vertex to itself.*

**Definition 1.19** (Directed path). *Let  $D = (V, \mathcal{E})$  be a directed graph and  $\alpha, \beta$  be two vertices in  $V$ . A (directed) path from  $\alpha$  to  $\beta$  is a sequence  $\alpha = \gamma_0, \dots, \gamma_n = \beta$ ,  $n \geq 1$ , of distinct vertices such that, for all  $1 \leq i \leq n$ ,  $(\gamma_{i-1}, \gamma_i) \in \mathcal{E}$ . If there is a path from  $\alpha$  to  $\beta$  in  $D$ , we say that  $\alpha$  leads to  $\beta$  and write  $\alpha \mapsto \beta$ .*

*A directed cycle is a directed path from a vertex  $\alpha$  to itself.*

**Definition 1.20** (DAG). *A directed acyclic graph or DAG is a directed graph with no directed cycles.*

The *skeleton* of a DAG  $D$  is the undirected graph obtained by forgetting the direction of the edges in  $D$ . For a DAG  $D = (V, \mathcal{E})$  and for  $\alpha \in V$ , we define the following set of vertices:

$$\begin{array}{ll} \text{parents of } \alpha: & \text{pa}(\alpha) = \{\beta \in V \mid (\beta, \alpha) \in \mathcal{E}\}, \\ \text{descendants of } \alpha: & \text{de}(\alpha) = \{\beta \in V \mid \alpha \mapsto \beta\}, \\ \text{non-descendants of } \alpha: & \text{nd}(\alpha) = V \setminus (\text{de}(\alpha) \cup \{\alpha\}). \end{array}$$

## Markov properties

$\pi$	<b>Markov property</b> with respect to $G = (V, E)$	<i>pairwise</i>	$\{\alpha, \beta\} \notin E \Rightarrow \alpha \perp\!\!\!\perp \beta   V \setminus \{\alpha, \beta\}$
		<i>local</i>	$\forall \alpha \in V, \alpha \perp\!\!\!\perp V \setminus \text{cl}(\alpha)   \text{bd}(\alpha)$
		<i>global</i>	$\forall A, B, S$ disjoint subsets of $V$ , $A \perp\!\!\!\perp_G B   S \Rightarrow A \perp\!\!\!\perp B   S$
$\rho$	<b>Hyper Markov property</b> with respect to decomposable graph $G = (V, E)$	<i>weak</i>	$\forall (A, B)$ decomposition of $G$ , $\pi_A \perp\!\!\!\perp \pi_B   \pi_{A \cap B}$
		<i>strong</i>	$\forall (A, B)$ decomposition of $G$ , $\pi_A \perp\!\!\!\perp \pi_B   A$
$\xi$	<b>Structural Markov property</b>		$\forall (A, B)$ covering pair of $G$ , $G_A \perp\!\!\!\perp G_B   \{G \in \mathfrak{U}(A, B)\}$

## Meta Markov properties

		$\Xi_A : \pi \mapsto \pi_A$	
		$\Xi_{A B} : \pi \mapsto \pi_{A B}$	
$\mathcal{F}$	<b>Meta Markov property</b> with respect to decomposable graph $G = (V, E)$	<i>weak</i>	$\forall (A, B)$ decomposition of $G$ , $\Xi_A \dagger \Xi_B   \Xi_{A \cap B} [\mathcal{F}]$
		<i>strong</i>	$\forall (A, B)$ decomposition of $G$ , $\Xi_A \dagger \Xi_B   A [\mathcal{F}]$
		$\mathfrak{G}_A : G \mapsto G_A$	
$\mathcal{G}$	<b>Structural Meta Markov property</b>		$\forall (A, B)$ covering pair of $G$ , $\mathfrak{G}_A \dagger \mathfrak{G}_B [\mathcal{G}(A, B)]$

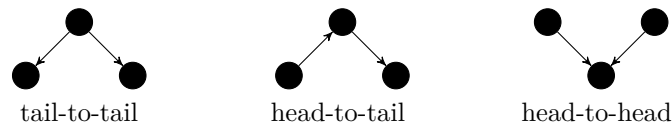
Table 1.1 – Handbook of Markov properties.



*Example 1.11.* A DAG and the sets associated with vertex  $\bullet$ .



Let us consider the possible directed graphs that one can obtain from an undirected graph linking three vertices in a chain. The result is from one of the three types described below.



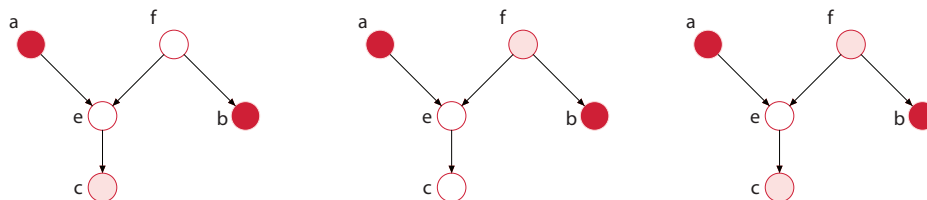
The head-to-head configuration is also called a *v-structure*. Whenever there is no edge between the parents of a v-structure, it is called an *immorality*. V-structures are of key importance in the definition of directed separation or *d-separation*.

**Definition 1.21** (d-separation). *Let  $A, B, S$  be subsets of  $V$ .  $S$  is said to d-separate  $A$  and  $B$  if for any  $\alpha \in A$ ,  $\beta \in B$  and any undirected path  $\gamma$  between  $\alpha$  and  $\beta$ , there exists a node  $\delta$  in  $\gamma$  such that either*

- $\delta \in S$ , and edges of  $\gamma$  do not meet head-to-head at  $\delta$ ;
- $\delta \notin S$  nor any of its descendants, and the arrows meet head-to-head at  $\delta$ .

$\delta$  is said to block the path  $\gamma$  from  $\alpha$  to  $\beta$ .

*Example 1.12.*



$c$  does not d-separate  $a$  and  $b$ .  $f$  d-separates  $a$  and  $b$ .  $\{c, f\}$  d-separates  $a$  and  $b$ .

For a DAG  $D$ , we let  $A \perp\!\!\!\perp_D B | C$  denote the statement “ $C$  d-separates  $A$  from  $B$  in  $D$ ”. Just like  $\perp\!\!\!\perp_G$  for undirected graphs  $G$ , the ternary relation  $\perp\!\!\!\perp_D$  is a graphoid. It can therefore be used to represent conditional independence.

**Definition 1.22** (Directed global Markov property). *Let  $D = (V, \mathcal{E})$  be a DAG. A probability distribution  $P$  on  $\mathcal{X}$  is said to obey the (directed) global Markov property relative to  $D$ , if for any triple  $(A, B, S)$  of disjoint subsets of  $V$  such that  $S$  d-separates  $A$  from  $B$  in  $D$ , it holds that  $A \perp\!\!\!\perp B | S$ .*

Similarly to the undirected case, weaker directed Markov properties can be defined on DAGs, but they are in fact equivalent to the global property without any further assumption on  $P$ . For more details, we refer the reader to (Lauritzen, 1996). Directed versions of the hyper Markov and structural Markov properties can also be easily derived (Dawid & Lauritzen, 1993; Byrne, 2011).

Two undirected graphs induce the same graphoid if and only if they are equal. With DAGs, the situation is more complicated, as several DAGs can lead to the same graphoid. It is also interesting to determine in which cases a DAG and an undirected graph can describe the same graphoids (see Example 1.13).

**Definition 1.23** (Markov equivalence). *Let  $K$  and  $K'$  be two graphs, each of them either an undirected graph or a DAG.  $K$  and  $K'$  are said to be Markov equivalent if, for any triple  $(A, B, S)$  of disjoint subsets of  $V$ , it holds that*

$$A \perp\!\!\!\perp_K B \mid S \Leftrightarrow A \perp\!\!\!\perp_{K'} B \mid S.$$

**Proposition 1.16** (Frydenberg, 1990). *Two DAGs are Markov equivalent if and only if they have the same skeleton and the same immoralities. A DAG  $D$  and an undirected graph  $G$  are Markov equivalent if and only if  $D$  has no immoralities and its skeleton is  $G$ .*

---

*Example 1.13.* All these graphs are Markov equivalent:



but not Markov equivalent to:



The first graph in Example 1.13 can be seen as the representative of the Markov equivalence class formed by itself and the three DAGs beside. More generally, *partially directed graphs* (PDG) were introduced to serve as representatives for Markov equivalence classes. As hinted by their names, these graphs are made of both directed and undirected edges. The PDG representing the Markov equivalence class of a DAG can be obtained by removing the directions of all edges that are not involved in an immorality. All the DAGs of a Markov equivalence class described by a PDG can be obtained by giving directions to undirected edges in a way that does not create additional immoralities. Consequently, the Markov equivalence class of an undirected spanning tree is made of all the DAGs that can be obtained by choosing a vertex as a root and directing all edges away from this root.

## 1.2 Algebra & Algorithms

The problems at hand in Chapters 2 and 3 are similar in the sense that they both require summations over large discrete sets. At first sight, these sets can seem quite hard to handle. In the first case, when we consider  $p$  vertices, we are dealing with  $p^{p-2}$  spanning trees. For  $p = 10$ , we are talking about a hundred million trees. When it comes to segmentations, the cardinality of the set  $\mathcal{M}_K$  made of all possible segmentations of  $\llbracket 1; N \rrbracket$  into  $K$  segments also grows fast, as it is equal to  $\binom{N-1}{K-1} \approx (N/K)^K$ . Nonetheless, in both cases, the summation can in fact be performed in polynomial time whenever the function to be summed up factorises in the right manner. The resulting algorithms heavily rely on algebraic results that we detail in the following sections.



## 1.2.1 Summing over Trees

### 1.2.1.a Undirected Trees

We remind that  $V = \{1, \dots, p\}$  and that  $\mathcal{T}$  denotes the set of spanning trees on  $V$ . Let  $f$  be a non-negative function defined on  $\mathcal{T}$ . Our goal is to compute

$$\Sigma(f) := \sum_{T \in \mathcal{T}} f(T).$$

Let us assume that  $f$  can be written as

$$f(T) = \prod_{\{i,j\} \in E_T} \omega_{i,j}, \quad \forall T = (V, E_T) \in \mathcal{T}, \quad (1.9)$$

where  $\omega = (\omega_{i,j})_{1 \leq i,j \leq p}$  is a symmetric matrix with non-negative entries giving the weight of any possible edge. The diagonal of  $\omega$  is set to zero. If the graph  $G_\omega = (V, E_\omega)$ , where

$$E_\omega = \{\{i,j\} \in \mathcal{P}_2(V) : \omega_{i,j} > 0\},$$

is not connected,  $f \equiv 0$ . If this factorisation assumption is true,  $\Sigma(f)$  is given by the following result.

**Theorem 1.5** (Matrix-Tree theorem). *Let  $\Delta$  be the Laplacian matrix associated to  $\omega$ , whose general term is given by*

$$\Delta_{i,j} = \begin{cases} -\omega_{i,j} & \text{if } i \neq j, \\ \sum_{k \in V} \omega_{k,j} & \text{if } i = j. \end{cases}$$

*Then all cofactors of  $\Delta$  are equal to  $\Sigma(f)$ .*

If all edges are given a weight equal to one, this theorem yields the number of trees in  $\mathcal{T}$ , which is  $p^{p-2}$ . This particular case is known as Cayley's formula. When the weight of an edge is set to either 0 or 1, it gives the number of spanning trees contained in the graph whose adjacency matrix is given by  $\omega$ . This result is often referred to as Kirchhoff's theorem.

We now prove a series of lemma that will be put together in the proof of Theorem 1.5. We consider a function  $f$  of the form given in (1.9), defined by a weight matrix  $\omega$ , and the associated Laplacian matrix  $\Delta$ . We assume that the graph  $G_\omega = (V, E_\omega)$  induced by the support of  $\omega$  is connected.

**Definition 1.24.** *The incidence matrix  $D^G$  of a graph  $G = (V, E)$  is a  $p \times |E|$  matrix, whose rows are indexed by  $V$  and columns are indexed by  $E$ . The  $l$ -th column of  $D^G$ , corresponding to  $\{i, j\} \in E$  with  $i < j$ , is defined by*

$$D_{k,l}^G = \begin{cases} 1 & \text{if } k = i, \\ -1 & \text{if } k = j, \\ 0 & \text{otherwise.} \end{cases}$$

For any matrix  $A$ , we let  $rk(A)$  denote the rank of  $A$ .

**Lemma 1.1.** *A graph  $G$  is connected if and only if  $rk(D^G) = p - 1$ .*

*Proof.* Suppose that  $G$  is connected. Each column having two non-zero entries, 1 and  $-1$ , summing all rows shows that  $rk(D^G) \leq p - 1$ . Let us consider a linear combination  $\alpha$  of  $d \leq p - 1$  rows of  $D^G$  whose coefficients are all non-zero and let us say that this combination is equal to zero. Since  $G$  is connected, there exists an edge  $e_j$  connecting a vertex corresponding to a chosen row to a vertex corresponding to a non-chosen one. There are only two non-zero entries in the  $j$ -th column of  $D^G$  and only one is in the  $d$  chosen rows: the  $j$ -th entry of  $\alpha$  cannot be equal to zero and that is absurd. So  $D^G$  has rank  $p - 1$ . Conversely, suppose that  $rk(D^G) = p - 1$  and that  $G$  is not connected, *i.e.* that  $G$  has a connected component of size  $d < p$ . By summing the rows corresponding to these  $d$  vertices, we would get a linear combination of size  $d$  equal to zero. That is not possible since  $rk(D^G) = p - 1$ .  $\square$

The Laplacian matrix associated to  $\omega$  can be expressed using the incidence matrix  $D$  of the complete graph on  $V$ . If  $m = |\mathcal{P}_2(V)| = p(p - 1)/2$ , we let  $\Omega$  denote the diagonal matrix of size  $m \times m$  giving the weights in  $\omega$  in the order defined by  $D$ .

**Lemma 1.2.**  $\Delta = D\Omega D^\top$ .

*Proof.*  $[D\Omega D^\top]_{i,j}$  is the inner product of the  $i$ -th row of  $D\Omega$  with the  $j$ -th row of  $D$ .  $D\Omega$  can be seen as a weighted incidence matrix for the complete graph. If  $i \neq j$ , the only non-zero entry in common between these two rows is in the column corresponding to the edge  $\{i, j\}$ . In this case, these two non-zero entries are either  $-\omega_{i,j}$  and 1 or  $\omega_{i,j}$  and  $-1$ , so that  $(D\Omega D^\top)_{i,j} = -\omega_{i,j}$ . Similarly, if  $i = j$ , we got  $[D\Omega D^\top]_{i,i} = \sum_{k \in V} \omega_{ik}$ .  $\square$

Note that this result remains true whenever one consider the incidence matrix of the graph  $G_\omega$  associated to the support of  $\omega$  instead of  $D$ . If  $\bar{\Omega}$  is the matrix obtained by removing the rows and columns corresponding to edges with weight zero from  $\Omega$ , it holds that

$$\Delta = (D^{G_\omega})\bar{\Omega}(D^{G_\omega})^\top. \quad (1.10)$$

We now give and prove two results on the incidence matrix that will be useful in the proof of the final result.

**Lemma 1.3.** Any square submatrix of  $D$  has determinant equal to 0 or 1 or  $-1$ .

*Proof.* The result can be proved recursively on the size of the submatrix. It is obviously true for submatrices of size 1. If  $1 < k \leq p$ , let us consider a square submatrix of  $D$  of size  $k$ . If all its entries are 0, its determinant is 0. Otherwise, two different cases can arise. All columns might have two non-zero entries, namely 1 and  $-1$ . Then we replace the first row by the sum of all rows and we get a null determinant. If not, there is one column with only one non-zero entry, be it 1 or  $-1$ , and we can develop our determinant according to this column, using our recursive hypothesis to conclude.  $\square$

**Lemma 1.4.** Let  $U$  be a subset of  $\mathcal{P}_2(V)$  with  $|U| = p - 1$ . Let  $D_U$  denote the  $(p - 1) \times (p - 1)$  submatrix of  $D$  consisting of the intersection of the  $p - 1$  columns corresponding to the edges in  $U$  and any set of  $p - 1$  rows of  $D$ . Then  $D_U$  is invertible if and only if the graph  $G_U = (V, U)$  is a spanning tree.

*Proof.* Suppose that  $G_U = (V, U)$  is a spanning tree. Then the submatrix  $D_U$  consists of  $p - 1$  rows of the incidence matrix  $D^{G_U}$  of  $G_U$  ( $D^{G_U}$  being a  $p \times (p - 1)$  submatrix of  $D$  itself). Since  $U$  is connected, the rank of  $D'$  is  $p - 1$  and  $D_U$  is invertible.



Conversely suppose that  $D_U$  is invertible. Then the incidence matrix  $D^{G_U}$  of  $G_U$  has an invertible  $(p-1) \times (p-1)$  submatrix and thus has rank  $p-1$ . It follows from Lemma 1.1 that  $G_U$  is a connected graph on  $V$ , with  $p-1$  edges. It is therefore a spanning tree.  $\square$

Finally, we give the following result on the determinant of a product of two rectangular matrices of transpose shape.

**Proposition 1.17** (Cauchy-Binet formula). *Let  $n, m$  be two integers such that  $n \leq m$ . Let  $A$  and  $B$  be two matrices of respective sizes  $n \times m$  and  $m \times n$ . For  $U \subseteq \llbracket 1; m \rrbracket$ , we let  $A_U$  and  $B_U$  denote the matrices respectively obtained by retaining the columns indicated by  $U$  in  $A$  and the rows indicated by  $U$  in  $B$ . Then, it holds that*

$$\det(AB) = \sum_{\substack{U \subseteq \llbracket 1; m \rrbracket, \\ |U|=n}} \det(A_U) \det(B_U).$$

*Proof.* See for instance (Broida & Williamson, 1989, §4.6).  $\square$

We now have all the elements necessary for the proof of Theorem 1.5.

*Proof of Theorem 1.5.* We first show that all cofactors of  $\Delta$  are equal. Let  $C_\Delta$  denote the cofactor matrix of  $\Delta$ . Then  $\Delta C_\Delta^\top = \det(\Delta)I = 0$  and every column of  $C_\Delta^\top$  is in the kernel of  $\Delta$ . As  $G_\omega$  is connected, by Lemma 1.1,  $D^{G_\omega}$  has rank  $p-1$  and by (1.10), so has  $\Delta$ . It follows that  $\Delta$  has a kernel of dimension 1. The vector  $(1, \dots, 1)^\top$  obviously belongs to that kernel. So each column of  $C_\Delta^\top$  has identical entries. But  $\Delta$  being symmetric, so is  $C_\Delta$  and all cofactors are equal.

Let  $\bar{D}$  denote the matrix obtained from  $D$  by removing the last row. Then  $\det(\bar{D}\Omega\bar{D}^\top)$  is a cofactor of  $\Delta$  by Lemma 1.2 and this determinant can be expanded thanks to the Binet-Cauchy formula:

$$\begin{aligned} \det(\bar{D}\Omega\bar{D}^\top) &= \sum_{\substack{U \subset \mathcal{P}_2(V), \\ |U|=p-1}} \det([\bar{D}\Omega]_U) \det(\bar{D}_U^\top) \\ &= \sum_{\substack{U \subset E, \\ |U|=p-1}} \det(\bar{D}_U)^2 \prod_{\{i,j\} \in U} \omega_{i,j} \end{aligned}$$

where  $\bar{D}_U$  and  $[\bar{D}\Omega]_U$  are respectively the square submatrices of  $\bar{D}$  and  $\bar{D}\Omega$  whose  $p-1$  columns correspond to the edges in  $U \subset \mathcal{P}_2(V)$ . Lemma 1.3 and 1.4 yield that  $\det(\bar{D}_U)^2$  is equal to 1 if  $G_U = (V, U)$  is a spanning tree on  $V$  and 0 otherwise, hence the result.  $\square$

Therefore, computing  $\Sigma(f)$  can be done in  $O(p^3)$  time as it only requires to compute the determinant of a  $(p-1) \times (p-1)$  matrix.

Other sums that  $\Sigma(f)$  might be of interest. For instance, in Chapter 2, we will actually need to sum  $f$  on the trees containing a particular edge  $\{k, l\}$ . The sum over the trees that do not contain edge  $\{k, l\}$  can be obtained by putting  $\omega_{k,l} = \omega_{l,k}$  to zero in  $\omega$  and by applying the Matrix-Tree theorem to this new weight matrix, and the difference with  $\Sigma(f)$  is the wanted sum.

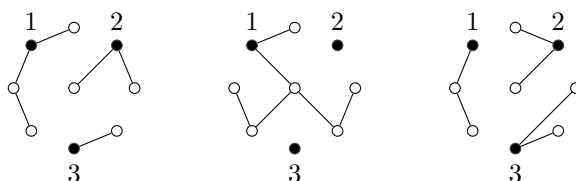
A more general version of the Matrix-Tree theorem can be given for graphs whose connected components are spanning trees on their respective sets of vertices. Such graphs include spanning trees, but also non-connected graphs. They are called *forests*, for a rather obvious reason.

**Theorem 1.6 (Chaiken, 1982).** *Let  $\Delta$  be the Laplacian of a weight matrix  $\omega$  and let  $U \subseteq V$ . We let  $\mathcal{F}_U$  denote the set of forests on  $V$  with  $|U|$  connected components such that, for any two vertices  $u_1, u_2 \in U$ ,  $u_1$  and  $u_2$  are not in the same connected component. We also let  $\Delta^U$  denote the matrix obtained from  $\Delta$  by removing the rows and columns corresponding to the vertices in  $U$ . Then it holds that*

$$|\Delta^U| = \sum_{F \in \mathcal{F}_U} \prod_{\{i,j\} \in E_F} \omega_{i,j}.$$

Briefly speaking,  $U$  can be seen as a set of “roots” (even though the models are not directed) for the trees of the forests in  $\mathcal{F}_U$ . If  $U$  is taken equal to a single vertex, then the forests in  $\mathcal{F}_U$  only have one connected component which is a tree and we get Theorem 1.5.

*Example 1.14.* Three examples of forests in  $\mathcal{F}_{\{1,2,3\}}$ .



A straightforward result can be deduced from Theorem 1.6 on the summation over the trees containing a given edge  $\{k, l\}$ .

**Corollary 1.6.1.** *Let  $\{k, l\}$  be an edge in  $\mathcal{P}_2(V)$ . Then it holds that*

$$\sum_{\substack{T \in \mathcal{T}, \\ \{k,l\} \in E_T}} \prod_{\{i,j\} \in E_T} \omega_{i,j} = \omega_{k,l} |\Delta^{\{k,l\}}|.$$

*Proof.* There is a one to one correspondence between the trees that contain edge  $\{k, l\}$  and the forests in  $\mathcal{F}_{\{k,l\}}$ . Each forest in  $\mathcal{F}_{\{k,l\}}$  can uniquely be transformed in a tree by adding edge  $\{k, l\}$ . Reciprocally, a tree containing edge  $\{k, l\}$  induces a forest in  $\mathcal{F}_{\{k,l\}}$  by removing  $\{k, l\}$ . Therefore, using Theorem 1.6, we have that

$$\sum_{\substack{T \in \mathcal{T}, \\ \{k,l\} \in E_T}} \prod_{\{i,j\} \in E_T} \omega_{i,j} = \omega_{k,l} \sum_{F \in \mathcal{F}_{\{k,l\}}} \prod_{\{i,j\} \in E_F} \omega_{i,j} = \omega_{k,l} |\Delta^{\{k,l\}}|.$$

□

### 1.2.1.b Directed Trees

Both versions of the Matrix-Tree theorem given above can actually be stated for directed trees and forests. A directed tree is a DAG  $D$  whose skeleton is an undirected spanning trees





and whose edges are oriented away from one root vertex. A directed forest is a *DAG* whose connected components are directed trees. We let  $\vec{\mathcal{T}}$  and  $\vec{\mathcal{F}}$  respectively denote the sets of directed trees and forests on  $V$ . The results that we are about to give will only be used in Section 4.2, when introducing temporal dependence in the modelling of multivariate time-series.

Let  $\omega$  be a  $p \times p$  matrix with entries in  $\mathbf{R}^+$ . The diagonal terms of  $\omega$  are assumed to be zeros. For  $(i, j) \in V^2$ ,  $\omega_{i,j}$  gives the weight of the edge going from  $i$  to  $j$ .  $\omega$  is not assumed to be symmetric. The Laplacian matrix  $\vec{\Delta}$  associated to  $\omega$  is defined as in the undirected case by

$$\vec{\Delta}_{i,j} = \begin{cases} -\omega_{i,j} & \text{if } i \neq j, \\ \sum_{k \in V} \omega_{kj} & \text{if } i = j. \end{cases}$$

The directed counter-part of Theorem 1.5 is given below.

**Theorem 1.7** (Directed Matrix-Tree theorem). *If we let  $C$  denote the cofactor matrix of  $\vec{\Delta}$  and  $\vec{\mathcal{T}}_r$ ,  $r \in V$ , denote the set of directed trees rooted at  $r$ , then it holds that*

$$\vec{Z}_r(\omega) := \sum_{D \in \vec{\mathcal{T}}_r} \prod_{(i,j) \in \mathcal{E}_D} \omega_{i,j} = C_{r,r}, \quad \forall r \in V.$$

If  $\omega$  is taken to be symmetric, we find that all  $\vec{Z}_r(\omega)$  are equal, as stated by the undirected version of the theorem.

Theorem 1.6 also admits an adaptation to directed forests. Stating this result, which is actually the one proved by Chaiken (1982), requires some definitions. The signature of  $A \subseteq V$  is defined by

$$\epsilon(A) := (-1)^{n(A)}, \quad n(A) := |\{\{i, j\} \in (V \setminus A) \times A : i < j\}|.$$

Let  $A, B$  be two subsets of  $V$  such that  $|A| = |B|$ . We let  $\vec{\mathcal{F}}_{A,B}$  denote the set of directed forests on  $V$  such that  $F \in \vec{\mathcal{F}}_{A,B}$  if and only if

- $F$  contains exactly  $|A| = |B|$  directed trees,
- each tree in  $F$  contains exactly one vertex in  $A$  and one vertex in  $B$ ,
- each edge in  $F$  is oriented away from the vertex in  $B$  of the tree containing that edge.

$\vec{\mathcal{F}}_{A,A}$  is the set of forests with skeleton in  $\mathcal{F}_A$ , the undirected forests rooted in  $A$ , and in which each edge is oriented away from the vertex in  $A$  of the tree containing that edge. Each forest  $F$  in  $\vec{\mathcal{F}}_{A,B}$  defines a bijective map  $\pi_F$  from  $A$  to  $B$  so  $\pi_F(j) = i$  if and only if  $i$  and  $j$  are in the same tree of  $F$ . A pair  $\{i, i'\} \in A$  is an inversion of  $\pi_F$  if  $i < i'$  and  $\pi_F(i) > \pi_F(i')$ . The number of inversions in  $\pi_F$  is denoted by  $n(\pi_F)$  and the signature of  $\pi_F$  is defined by

$$\epsilon(\pi_F) := (-1)^{n(\pi_F)}.$$

For every  $F \in \vec{\mathcal{F}}_{A,A}$ ,  $\pi_F$  is the identity and  $\epsilon(\pi_F) = 1$ .

**Theorem 1.8** (All Minors Matrix-Tree theorem, Chaiken, 1982). Let  $\vec{\Delta}$  be the directed Laplacian matrix associated to a directed edge weight matrix  $\beta$ . Let  $A, B$  be two subsets of  $V$  such that  $|A| = |B|$ . We let  $\vec{\Delta}^{A,B}$  denote the matrix obtained from  $\vec{\Delta}$  by removing the rows corresponding to the vertices in  $A$  and the columns corresponding to the vertices in  $B$ . For any DAG  $D = (V, \mathcal{E})$ , we define

$$\omega_D := \prod_{(i,j) \in \mathcal{E}} \omega_{i,j}.$$

Then, it holds that

$$|\vec{\Delta}^{A,B}| = \epsilon(A)\epsilon(B) \sum_{F \in \mathcal{F}_{A,B}} \epsilon(\pi_F) \omega_F.$$

Theorem 1.8 can be used to sum over the directed trees rooted in a vertex  $r$  and containing a given edge  $(i, j)$ .

**Corollary 1.8.1.** Let  $r$  be a vertex in  $V$  and  $(u, v)$  be a directed edge such  $v \neq r$ . Then, it holds that

$$\sum_{\substack{D \in \vec{T}_r, \\ (u,v) \in \mathcal{E}_D}} \omega_D = \omega_{u,v} \left| |\vec{\Delta}^{\{u,v\}, \{r,v\}}| \right|$$

where  $\left| |\vec{\Delta}^{\{u,v\}, \{r,v\}}| \right|$  is the absolute value of the determinant of  $\vec{\Delta}^{\{u,v\}, \{r,v\}}$ .

*Proof.* Let  $A = \{u, v\}$  and  $B = \{r, v\}$ . Note that, as  $v \neq r$  and  $u \neq v$ , neither set can be reduced to a single vertex. Let  $D$  be a tree rooted in  $r$  and containing  $(u, v)$ . Removing  $(u, v)$ , we are left with two directed trees. Vertex  $v$  cannot be in the connected component of  $r$ . Otherwise, the last edge in the directed path from  $r$  to  $v$  and  $(u, v)$  would form a  $v$ -structure at vertex  $v$ . Therefore, the unique directed path between  $r$  and  $v$  goes through  $(u, v)$  and  $v$  and  $r$  are in two distinct connected components when  $(u, v)$  is removed. This also implies that in the connected component containing  $v$ , all edges are oriented away from  $v$ . The forest obtained by removing  $(u, v)$  from  $D$  thus belongs to  $\vec{\mathcal{F}}_{A,B}$ . Reversely, each forest in  $\vec{\mathcal{F}}_{A,B}$  can be converted to a tree rooted in  $r$  and containing edge  $(u, v)$  by adding  $(u, v)$ . There is a one-to-one mapping between these two sets.

Now, for all  $F \in \vec{\mathcal{F}}_{A,B}$ , note that  $\pi_F$  is always such that  $\pi_F(v) = v$  and  $\pi_F(u) = r$ , so that  $\epsilon(\pi_F)$  has the same value no matter  $F$ , denoted  $\epsilon^*$ . Using Theorem 1.8, we obtain that

$$\sum_{\substack{D \in \vec{T}_r, \\ (u,v) \in \mathcal{E}_D}} \omega_D = \omega_{u,v} \sum_{D \in \vec{\mathcal{F}}_{A,B}} \omega_D = \epsilon^* \epsilon(A)\epsilon(B) \omega_{u,v} |\vec{\Delta}^{A,B}| = \omega_{u,v} \left| |\vec{\Delta}^{A,B}| \right|.$$

□

## 1.2.2 Summing over Segmentations

For  $0 < i < j$ , we let  $\llbracket i; j \rrbracket$  and  $\llbracket i; j \rrbracket$  respectively denote the sets  $\{i, \dots, j\}$  and  $\{i, \dots, j-1\}$ .



**Definition 1.25.** For  $0 < i < j$ , a segmentation  $m$  of  $\llbracket i; j \rrbracket$  is a partition of  $\{i, \dots, j-1\}$  into sets of consecutive elements, called segments. Thus, if  $m$  has  $K \geq 0$  segments, it can be written as

$$m = \{\llbracket \tau_k; \tau_{k+1} \rrbracket\}_{k=1}^K = \{r_k\}_{k=1}^K,$$

with  $i = \tau_0 < \tau_1 < \dots < \tau_{K-1} < \tau_K = j$ .  $\{\tau_1, \dots, \tau_{K-1}\}$  are called the change-points of  $m$ .

For  $K \geq 1$ , we let  $\mathcal{M}_K(\llbracket i; j \rrbracket)$  denote the set made of all the segmentations of  $\llbracket i; j \rrbracket$  into  $K$  segments. Let  $N \geq 0$ . We let  $\mathcal{M}_K$  be a shorthand for  $\mathcal{M}_K(\llbracket 1; N+1 \rrbracket)$ . Let  $f$  be a non-negative function defined on  $\mathcal{M}_K$ . Our goal is to compute

$$\Sigma(f) := \sum_{m \in \mathcal{M}_K} f(m).$$

Let us assume that  $f$  can be written as

$$f(m) = \prod_{\llbracket i; j \rrbracket \in m} A_{i,j}, \quad \forall m \in \mathcal{M}_K, \quad (1.11)$$

where  $A$  is a strictly upper triangular matrix of size  $(N+1) \times (N+1)$ . Matrix  $A$  basically gives the weight of all possible segments. Function  $f$  can therefore be extended to  $\mathcal{M}_k(\llbracket i; j \rrbracket)$  for all  $k \geq 0$  and any segment  $\llbracket i; j \rrbracket \subseteq \llbracket 1; N+1 \rrbracket$ .

**Theorem 1.9** (Rigaill et al., 2012). Let  $f$  be a function of the form given in (1.11) defined by a strictly upper triangular matrix  $A$ . Then, for  $k \geq 1$ , and  $\llbracket i; j \rrbracket \subseteq \llbracket 1; N+1 \rrbracket$ , it holds that

$$\sum_{m \in \mathcal{M}_k(\llbracket i; j \rrbracket)} f(m) = [A^k]_{i,j}.$$

The proof of Theorem 1.9 relies on the following lemma.

**Lemma 1.5.** Let  $A$  be a square matrix of size  $n \times n$ ,  $n \geq 1$ . For  $k \in \mathbf{N}^*$ , we define  $f_{A,k}$  as

$$f_{A,k}(i, j) = \sum_{(t_1, \dots, t_{k-1}) \in \llbracket 1; n \rrbracket^{k-1}} \prod_{s=0}^{k-1} A_{t_s, t_{s+1}}. \quad (1.12)$$

Then, it holds that

$$f_{A,k}(i, j) = [A^k]_{i,j}.$$

*Proof.*  $f_{A,k}(i, j) = [A^k]_{i,j}$  holds for  $k = 1$ . Suppose that  $f_{A,k}(i, j) = [A^k]_{i,j}$  holds for  $k \in \mathbf{N}^*$ . For  $k+1$ , we have that

$$\begin{aligned} f_{A,k+1}(i, j) &= \sum_{(t_1, \dots, t_k) \in \llbracket 1; n \rrbracket^k} \prod_{s=0}^k A_{t_s, t_{s+1}} \\ &= \sum_{t=1}^n \sum_{(t_1, \dots, t_{k-1}) \in \llbracket 1; n \rrbracket^{k-1}} \prod_{s=0}^{k-1} A_{t_s, t_{s+1}} \\ &= \sum_{t=1}^n f_{A,k}(i, t) A_{t,j}. \end{aligned}$$

Using our induction hypothesis and by definition of the matrix product, we obtain:

$$f_{A,k+1}(i, j) = \sum_{t=1}^n [A^k]_{i,t} A_{t,j} = [A^{k+1}]_{i,j}. \quad (1.13)$$

□

Proving Theorem 1.9 from Lemma 1.5 is straightforward.

*Proof of Theorem 1.9.* Whenever  $A$  is a strictly upper diagonal matrix, all the terms in  $f_{A,k}(i, j)$  that do not correspond to a segmentation of  $\llbracket i; j \rrbracket$ , *i.e.* for which we don't have  $i = t_0 < t_1 < \dots < t_{k-1} < t_k = j$ , are equal to zero. Indeed, if  $(t_0, \dots, t_k)$  is not a segmentation of  $\llbracket i; j \rrbracket$ , there exist  $s \in \llbracket 0; k-1 \rrbracket$  such that  $t_s \geq t_{s+1}$  and  $A_{t_s, t_{s+1}} = 0$  as  $A$  is strictly upper diagonal. Thus, we have that

$$f_{A,k}(i, j) = \sum_{m \in \llbracket i; j \rrbracket} \prod_{\llbracket t_s; t_{s+1} \rrbracket \in m} A_{t_s, t_{s+1}} = \sum_{m \in \mathcal{M}(\llbracket i; j \rrbracket)} f(m),$$

and Lemma 1.5 gives the sought-after result. □

Concerning complexity, (1.13) shows that, for  $i \in \llbracket 1; N+1 \rrbracket$  and  $K \leq N$ , quantities in

$$\{f_{A,k}(i, j)\}_{\substack{1 \leq i \leq N \\ 1 \leq k \leq K}}$$

can all be computed in  $O(KN^2)$  time as the  $i$ -th row of matrices  $A, A^2, \dots, A^K$ . The same quantities where  $j$  is fixed are obtained in  $O(KN^2)$  as the  $j$ -th column of matrices  $A, A^2, \dots, A^K$ .





# 2

## Network Inference

---

2.1	Introduction	38
2.2	Background & Model	39
2.2.1	Markov Properties & Graphical Models	39
2.2.2	Model for Bayesian Inference of Graphical Models Based on Trees	40
2.3	Priors on Tree Structures & Distributions	41
2.3.1	Prior on Tree Structures	41
2.3.2	Prior on Tree Distributions	41
2.3.3	Structural Markov Property and Structurally Meta Markov Families	43
2.4	Inference in Tree Graphical Models	44
2.4.1	Integration with Respect to $\pi$	45
2.4.2	Integration with Respect to $T$	48
2.4.3	Controlling prior edge probability	51
2.5	Simulations	52
2.5.1	Simulation Scheme	52
2.5.2	Results	53
2.6	Application to Cytometry Data	55
2.6.1	Data	56
2.6.2	Results	56

---

*Our angle on network inference is to assume that the structure of the underlying graphical model is a spanning tree. The chances are that the structure of interest do not belong to this particular class of graph. For this reason, instead of looking out for the most likely tree structure, we actually compute the posterior probability for any given edge to be borrowed by a random tree. This approach can somehow be justified by the fact that a graph can be close to a tree on a local scale, while having cycles or not being connected. Averaging over all spanning trees, we might therefore be able to derive meaningful information about local features, such as edges, while blurring out the strong global constraints imposed by individual*



trees. Having restrained our attention to spanning trees, the matrix containing the posterior probabilities of every possible edges can be computed in cubic time with respect to the number of vertices. This means that, from a complexity point of view, our approach could be used on problems involving up to a few hundreds variables in a reasonable amount of time. This chapter is available as a preprint on arXiv (Schwaller et al., 2015).

## 2.1 Introduction

Statistical models are getting more and more complex and can now involve very intricate dependency structures. Graphical models are both a natural and powerful way to depict such structures. Inferring a graphical model based on observed data is hence of great interest for many fields of applications. From a statistical point-of-view, considering the inference of a graphical model requires to consider the graphical model itself as a parameter. In a Bayesian context, it means that we have to define a full model and, more specifically, a prior distribution on graphical models, therefore on graphs themselves.

Regardless of whether we consider directed or undirected graphs, their sheer number make them difficult to deal with. Exact inference can only be contemplated as long as there are no more than thirty or so variables of interest (Parviainen & Koivisto, 2009). When exact inference is no longer tractable, sampling is used as a pragmatic alternative. Markov Chain Monte Carlo (MCMC) methods have for instance been used to sample from some sets of graphs, such as Directed Acyclic Graphs (DAGs) (Madigan et al., 1995; Friedman & Koller, 2003; Niinimäki et al., 2011) or decomposable graphs (Green & Thomas, 2013). The decomposability assumption for undirected graphical models, also called Markov random fields, is commonly made in the literature, although some interest has been devoted to the less easy to handle non-decomposable graphs (Roverato, 2002; Atay-Kayis & Massam, 2005). The sampling schemes developed in the aforementioned papers are often subject to standard issues related to MCMC sampling in high-dimensional spaces, namely slow mixing and difficulty to get to the stationary distribution.

One way to bypass these hurdles is to further restrict the exploratory space so as to make exact inference tractable. When a subset of graphs is considered, it sometimes becomes possible to get access to the full posterior distribution on graphs. The obvious drawback of this approach is that the “true” graph might not belong to this subset. In this case, computing a maximum a posteriori (MAP) estimate would for instance yield a systematically wrong answer. But usually, such methods are not intended to assess the global structure all at once but to separately assess a collection of local features of the graph (typically, edges). The idea is that the inference of such features is less affected by the restriction than the global structure. In that perspective, trees have been of particular interest as a subset of both decomposable graphs and DAGs (Chow & Liu, 1968; Meilä & Jordan, 2001; Meilä & Jaakkola, 2006; Kirshner, 2007; Lin et al., 2009; Burger & Van Nimwegen, 2010).

Our first contribution is to provide a well-defined fully Bayesian framework for graphical model inference based on trees. We use the work of Dawid & Lauritzen (1993) on hyper Markov laws to define priors on tree parameters and distributions that can easily be marginalized over. This framework spares us from requiring likelihood equivalence between Markov equivalent directed tree models, like Meilä & Jaakkola (2006) did building on the work of Heckerman & Chickering (1995). We also point out that it fits within the recent work of Byrne & Dawid (2015) on structurally Markov graph distributions. We then go through a series of typical models befitting this framework, namely tree-structured copulas

(Kirshner, 2007), multinomial distributions (Meilä & Jaakkola, 2006) and Gaussian distributions. Bayesian inference in this framework requires integration over the set of trees, that can be carried out exactly and efficiently thanks to an algebraic result called the Matrix-Tree theorem.

Our second contribution focuses on edge inference. When Meilä & Jaakkola (2006) and Kirshner (2007) were interested in the joint distribution of the observations, we are interested in the inference of the dependence structure. To this purpose, we are not concerned with the inference of the parameters but we need to account for the uncertainty of their estimates. The Bayesian construction we propose provides a natural framework to achieve this. We derive the exact posterior probability of any given edge in an exact manner, allowing for an arbitrary prior edge appearance probability.

Most works on tree-structured graphical model inference rely on the aforementioned Matrix-Tree theorem. As noticed by Kirshner (2007), the computation of posterior probabilities for all the edges in this setting can be achieved with cubic complexity with respect to the number of variables. We provide a new proof of this result relying on a generalization of the Matrix-Tree theorem to forests.

Our last contribution is a simulation study which addresses the influence of the tree assumption on the accuracy of structure inference for non-tree-structured graphical models. We demonstrate that, as long as edge inference is concerned, the computational efficiency following from this assumption can be obtained at a limited cost.

An R-language package **saturnin** implementing the approach presented here is available from the Comprehensive R Archive Network.

In Section 2.2, we provide some background on graphical models and Markov properties before writing down the full model in which the inference is performed. Priors for tree structures and distributions are defined in Section 2.3. Section 2.4 deals with the inference of the model. Integrations with respect to distributions and structures are respectively discussed in Sections 2.4.1 and 2.4.2. The simulation study and its results are described in Section 2.5. An application to flow cytometry data is presented in Section 2.6.

## 2.2 Background & Model

### 2.2.1 Markov Properties & Graphical Models

Let  $V = \{1, \dots, p\}$  and  $\mathbf{X} = (X_1, \dots, X_p)$  be a random vector taking values in a product space  $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$ . The set of distributions on  $\mathcal{X}$  is denoted by  $\mathcal{F}$ . For any subset  $A$  of  $V$ ,  $\mathbf{X}_A$  denotes the subvector of  $\mathbf{X}$  corresponding to the components in  $A$ . Let  $\mathcal{P}_2(V)$  denote the subsets of  $V$  of size 2. For  $E \subseteq \mathcal{P}_2(V)$ ,  $G = (V, E)$  is the undirected graph with vertices  $V$  and edges  $E$ . In the following, the notations of Dawid & Lauritzen (1993) will be used. We refer the reader to the appendix of their article for a quick introduction to graph terminology and graphical models, or to Lauritzen (1996) for a more detailed overview.

A pair  $(A, B)$  of subsets of  $V$  is said to be a decomposition of  $G$  if  $V = A \cup B$ , the subgraph induced by  $G$  on  $A \cap B$  is complete and  $A \cap B$  separates  $A$  from  $B$ . If  $A$  and  $B$  are both proper subsets of  $V$ , the decomposition is said to be proper. Here we restrain our attention to decomposable graphs, namely graphs that are either complete or for which there exists a proper decomposition into two decomposable subgraphs.





**Definition 2.1.** A distribution  $\pi \in \mathcal{F}$  is said to be Markov with respect to (w.r.t.) a decomposable graph  $G$  if, for any decomposition  $(A, B)$  of  $G$ , under  $\pi$ ,

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_{A \cap B}.$$

**Proposition 2.1.** (Hammersley & Clifford, 1971) Let  $\pi \in \mathcal{F}$ . If  $\pi$  is a positive distribution (for all  $\mathbf{x} \in \mathcal{X}$ ,  $\pi(\mathbf{x}) > 0$ ), being Markov w.r.t. a decomposable graph  $G$  is equivalent to the existence of a factorisation of  $\pi$  on the (maximal) cliques of  $G$ .

Here we will focus on distributions that are Markov w.r.t. to connected graphs without any cycles. Such graphs are called spanning trees and their maximal cliques are of size 2. Thus, a positive distribution that is Markov w.r.t. a tree  $T = (V, E_T)$  can be factorised on the edges of the tree, using the marginal distributions of order 1 and 2:

$$\forall \mathbf{x} \in \mathcal{X}, \pi(\mathbf{x}) = \prod_{i \in V} \pi_i(x_i) \prod_{\{i,j\} \in E_T} \frac{\pi_{ij}(x_i, x_j)}{\pi_i(x_i)\pi_j(x_j)}.$$

Such distributions will be called tree distributions in the following.

**Definition 2.2.** A graphical model  $m_G := (G, \mathcal{F}_G)$  is given by a decomposable graph  $G$  and a family of distributions  $\mathcal{F}_G \subseteq \mathcal{F}$  that are Markov w.r.t.  $G$ .

Let  $m_G = (G, \mathcal{F}_G)$  be a graphical model. To avoid any confusion, distributions on a set of distributions will be called hyperdistributions. If  $\pi \in \mathcal{F}_G$  and  $\rho$  is a hyperdistribution on  $\mathcal{F}_G$ , for any  $A, B \subseteq V$ ,  $\pi_A$  denotes the marginal distribution obtained from  $\pi$  on the variables  $\mathbf{X}_A$  and  $\pi_{B|A}$  the collection of conditional distributions of  $\mathbf{X}_B \mid \mathbf{X}_A$  under  $\pi$ . We also denote  $\rho_A$  the marginal hyperdistribution induced by  $\rho$  on  $\pi_A$  and  $\rho_{B|A}$  the collection of hyperdistributions induced by  $\rho$  on  $\pi_{B|A}$ .

**Definition 2.3.**  $\rho$  is said to be strong hyper Markov w.r.t.  $G$  if, for any decomposition  $(A, B)$  of  $G$ , under  $\rho$ ,

$$\pi_A \perp\!\!\!\perp \pi_{B|A}.$$

Such hyperdistributions will be useful to define priors on distribution spaces.

## 2.2.2 Model for Bayesian Inference of Graphical Models Based on Trees

Let  $\mathcal{T}$  denote the set of spanning trees on  $V$ . For any tree  $T \in \mathcal{T}$ , we consider a graphical model  $m_T = (T, \mathcal{F}_T)$  with a family of positive distributions  $\mathcal{F}_T \subseteq \mathcal{F}$  Markov w.r.t.  $T$ . Here we consider a Bayesian framework. We therefore need to define prior distributions for  $T$  and for  $\pi$  conditional on  $T$ . This is dealt with in Section 2.3. The full Bayesian model consists in first drawing a random tree  $T^*$ , then a distribution  $\pi$  in  $\mathcal{F}_T$  and finally  $\mathbf{X}$  according to  $\pi$  (Figure 2.1). Defining a prior on tree distributions could be especially troublesome since it needs to be defined for every graphical model  $m_T$ . The idea is to require these hyperdistributions to be strong hyper Markov w.r.t. to their trees, so that they can be built from local hyperdistributions defined on the edges and chosen once and for all trees.

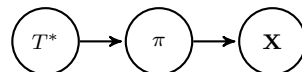


Figure 2.1 – Global hierarchical model.

This choice of prior and the fact that we only consider tree structures for the graphical models make the inference of the graph in our model tractable in an exact manner, thanks to the Matrix-Tree theorem.

## 2.3 Priors on Tree Structures & Distributions

Restraining the explored set of graphs to the spanning trees obviously helps to make the inference easier to perform, but it still leaves us with a super-exponential number,  $p^{p-2}$ , of graphs. Nonetheless, a suitable choice of priors on tree structures and parameters leads to a tractable situation. Meilă & Jaakkola (2006) defined what they call decomposable priors under which parameters can be dealt with at the edge level. The integration over the set of trees can then be performed exactly thanks to algebra. We will use strong hyper Markov hyperdistributions (Dawid & Lauritzen, 1993) to define our prior but the idea is basically the same. Let  $D = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$  be an independent sample of size  $n \geq 1$  drawn from  $\mathbf{X}$ . Our goal is to define a prior distribution on  $(T, \theta)$  such that the posterior distribution on trees  $\xi(\cdot|D)$  factorises over the edges, i.e.

$$\xi(T|D) = \frac{1}{Z} \prod_{\{i,j\} \in E_T} \omega_{i,j}, \quad \forall T \in \mathcal{T}, \quad (2.1)$$

where  $\omega = (\omega_{i,j})_{(i,j) \in V^2}$  is a symmetric matrix with non-negative values and

$$Z = \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \omega_{i,j} \quad (2.2)$$

is a normalising constant. Both  $\omega$  and  $Z$  obviously depend on the data  $D$  but we drop the dependence in the notations for sake of clarity.

### 2.3.1 Prior on Tree Structures

Let  $\beta = (\beta_{i,j})_{(i,j) \in V^2}$  be a symmetric matrix with non-negative values such that its support graph  $G_\beta = (V, E_\beta)$ , where  $E_\beta = \{\{i,j\} \in \mathcal{P}_2(V), \beta_{i,j} > 0\}$ , is connected. We consider a prior distribution  $\xi$  on  $T$  that factorises over the edges,

$$\xi(T) = \frac{1}{Z_0} \prod_{\{i,j\} \in E_T} \beta_{i,j}, \quad \forall T \in \mathcal{T}. \quad (2.3)$$

The assumption about  $\beta$  is here to serve as a guarantee that  $\beta$  induces a proper distribution on trees;  $\xi$  can typically be taken as a uniform on  $\mathcal{T}$ .

These distributions belong to the family of structurally Markov graph distributions described by Byrne & Dawid (2015) (see Section 2.3.3).

### 2.3.2 Prior on Tree Distributions

As Bayes' rule states that  $\xi(T|D) \propto \xi(T)p(D|T)$ , we are now interested in the marginal likelihood of the data under a tree model  $m_T$ ,

$$p(D|T) = \int_{\mathcal{F}_T} p(D|\pi)p(\pi|T)d\pi. \quad (2.4)$$



For every  $T \in \mathcal{T}$ , we have to define a prior distribution on  $\mathcal{F}_T$  such that the marginal likelihood  $p(D|T)$  can also be factorised on the edges.

Meilä & Jaakkola (2006) built their prior on multinomial tree distributions around three main assumptions, namely likelihood equivalence, parameter independence and parameter modularity. The first assumption requires that the prior treats all possible para-metrizations consistent with a given tree  $T$  (be it directed or undirected) as indistinguishable. As we only consider undirected parametrizations in our construction, we shall not need this assumption here. As for the parameter independence assumption, it can be broken down into local and global independences (Spiegelhalter & Lauritzen, 1990). Strong hyper Markov hyperdistributions satisfy global independence but not necessarily local independence. The latter is in fact not needed to get the desired factorisation property for the marginal likelihood. Finally, the parameter modularity assumption is ensured by the construction of a compatible family of strong hyper Markov hyperdistributions.

Let  $T$  be a tree and  $\rho^T$  a strong hyper Markov hyperdistribution on  $\mathcal{F}_T$ . Such hyperdistributions have an interesting property regarding the marginal likelihood  $p(D|T)$ .

**Proposition 2.2.** (Dawid & Lauritzen, 1993, Prop. 5.6) *If  $\rho^T$  is strong hyper Markov w.r.t.  $T$ , then the marginal likelihood  $p(D|T)$  is Markov w.r.t. to  $T$ .*

This means that the marginal likelihood can be factorised on the edges of  $T$ . For  $i \in V$ , let  $D_i = \{x_i^1, \dots, x_i^n\}$  be the observed data restricted to  $X_i$ . The integral given in (2.4) can then be rewritten as

$$\begin{aligned} p(D|T) &= \int p(D|\pi)p(\pi|T)d\pi = \int \pi(D)\rho^T(\pi)d\pi \\ &= \prod_{i \in V} p(D_i|T) \prod_{\{i,j\} \in E_T} \frac{p(D_i, D_j|T)}{p(D_i|T)p(D_j|T)} \end{aligned} \quad (2.5)$$

where, for all  $(i, j) \in V^2$ ,

$$p(D_i, D_j|T) = \int \pi_{ij}(D_i, D_j)\rho_{ij}^T(\pi_{ij})d\pi_{ij}; \quad (2.6)$$

$$p(D_i|T) = \int \pi_i(D_i)\rho_i^T(\pi_i)d\pi_i. \quad (2.7)$$

The calculation of these integrals will be addressed in Section 2.4.1.

We now explain how to choose  $\rho^T$  for all  $T$  so that the distributions of  $\{\pi_{ij}\}_{\{i,j\} \in \mathcal{P}_2(V)}$  do not depend on  $T$ . Let us consider a general hyperdistribution  $\rho$  on  $\mathcal{F}$  such that, for any  $A \subseteq V$  and under  $\rho$ ,

$$\pi_A \perp\!\!\!\perp \pi_{V \setminus A} | A. \quad (2.8)$$

This means that  $\rho$  is strong hyper Markov w.r.t. the complete graph over  $V$ .

**Proposition 2.3.** (Dawid & Lauritzen, 1993, §6.2) *For any tree  $T \in \mathcal{T}$ , there exists a unique hyperdistribution  $\rho^T$  on  $\mathcal{F}_T$  that is strong hyper Markov w.r.t.  $T$  and such that, for every edge  $\{i, j\} \in E_T$ ,*

$$\rho_{ij}^T = \rho_{ij}. \quad (2.9)$$

$\{\rho^T\}_{T \in \mathcal{T}}$  is said to be a (hyper) compatible family of strong hyper Markov hyperdistributions.

Proposition 2.3 guarantees that all  $\rho^T$  are strong hyper Markov w.r.t.  $T$ . By Proposition 2.2, for all  $T \in \mathcal{T}$ , the marginal likelihood under  $\rho^T$  is Markov w.r.t.  $T$ . Moreover, the compatibility of the family  $\{\rho^T\}_{T \in \mathcal{T}}$  makes the dependence on  $T$  in the local marginal distributions given in (2.6) and (2.7) irrelevant. They can be computed once and for all for every  $\{i, j\} \in \mathcal{P}_2(V)$ . This choice of priors for the distributions assures that (2.1) is satisfied with

$$\omega_{i,j} = \beta_{i,j} \frac{p(D_i, D_j)}{p(D_i)p(D_j)}, \quad \forall (i, j) \in V^2. \quad (2.10)$$

The model under which we are now working is fully described in Figure 2.2.

Proposition 2.3 shows that we do not need to have access to the full basis hyperdistribution to specify a compatible family of strong hyper Markov hyperdistributions. It is indeed enough to provide a consistent family of pairwise hyperdistributions  $\{\rho_{ij}\}_{\mathcal{P}_2(V)}$ , where the consistency property must be understood in the sense that two hyperdistributions involving a common vertex should induce the same marginal hyperdistribution on this vertex. This is automatically satisfied when  $\{\rho_{ij}\}_{\mathcal{P}_2(V)}$  is obtained from a fully specified hyperdistribution  $\rho$ . In order to obtain strong hyper Markov hyperdistributions when combining these pairwise hyperdistributions, we shall additionally require that, for all  $i, j \in V$ ,  $\pi_{i|j} \perp\!\!\!\perp \pi_j$  under  $\rho_{ij}$  (Dawid & Lauritzen, 1993, Prop. 3.16), meaning that  $\rho_{ij}$  is strong hyper Markov w.r.t. the graph on  $\{i, j\}$  where vertices  $i$  and  $j$  are connected.

### 2.3.3 Structural Markov Property and Structurally Meta Markov Families

The purpose of this section is to show how the model that we have described so far is related to the structural Markov property defined by Byrne & Dawid (2015). Indeed, trees have specific algebraic properties that will be taken advantage of in Section 2.4 for the inference of the model, but the model itself can be extended to other subsets of decomposable graphs.

Byrne & Dawid (2015) defined an extension of the (hyper) Markov properties described in Dawid & Lauritzen (1993) to undirected decomposable graphs (and to directed acyclic graphs, but this will not be discussed here) called the structural Markov property.

Let  $\mathcal{U}$  be the set of undirected decomposable graphs on  $V$ . A pair of subsets  $(A, B)$  of  $V$  is called a covering pair if  $A \cup B = V$ . For any family of graphs  $\mathcal{G} \subseteq \mathcal{U}$  and for any covering pair  $(A, B)$ , we define  $\mathcal{G}(A, B)$  to be the set of graphs  $G \in \mathcal{G}$  for which  $(A, B)$  is a decomposition.

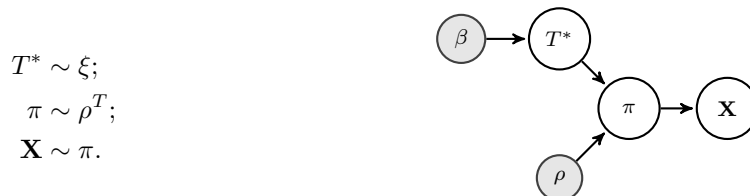


Figure 2.2 – Compatible strong hyper Markov tree model.



**Definition 2.4.** A distribution  $\xi$  for  $G \in \mathfrak{U}$  is said to be *structurally Markov* if for any covering pair  $(A, B)$  such that  $\xi(\mathfrak{U}(A, B)) > 0$ ,

$$G_A \perp\!\!\!\perp G_B | \{G \in \mathfrak{U}(A, B)\} \quad (2.11)$$

under  $\xi$ .

The support graph families of structurally Markov graph distributions have structural properties. They satisfy the so called structural meta Markov property.

**Definition 2.5.** Let  $\mathcal{G}$  be a family of undirected decomposable graphs on  $V$ . Then  $\mathcal{G}$  is *structurally meta Markov* if for any covering pair  $(A, B)$ , the set

$$\{G_A | G \in \mathcal{G}(A, B), G_B = J\}$$

is the same for all  $J \in \{G_B | G \in \mathcal{G}(A, B)\}$ .

The set of trees  $\mathcal{T}$  is an example of such a family (Byrne & Dawid, 2015, Ex. 3.1) and the distributions that we considered in Section 2.3.1 are structurally Markov.

These graph distributions naturally interact with (strong) hyper Markov hyperdistributions and Markov distributions when they are chosen carefully. Compatible hyperdistribution families as described in Proposition 2.3 conjugate nicely with graph distributions factorised on the edges, so that all hyperdistribution updates can be performed down to the edge level. But compatibility can be defined for any structurally meta Markov family  $\mathcal{G}$  (Byrne & Dawid, 2015, Definition 3.4). Then, the update can be performed locally on  $\mathcal{C}_G = \bigcup_{G \in \mathcal{G}} \mathcal{C}_G$  where  $\mathcal{C}_G$  denotes the cliques of graph  $G$ .

We finish this section by laying stress upon the fact that, among structurally meta Markov graph families, trees are of particular computational interest given their algebraic properties. One of the main difficulties in assessing graph distributions is to compute normalisation constants, but closed-form expressions can be derived for these constants in the case of trees (see Section 2.4.2).

## 2.4 Inference in Tree Graphical Models

Different inference tasks can be performed on graphical models. One might be interested in estimating the emission distribution of  $X$ . Chow & Liu (1968) gave an algorithm to get the tree distribution maximizing the likelihood of discrete multivariate data in the frequentist equivalent of the model described in the previous section. It can easily be adapted to MAP estimation in a full Bayesian framework (Meilă, 1999). It is also possible to look at the posterior predictive distribution (Meilă & Jaakkola, 2006),

$$p(\mathbf{x}|D) = \sum_{T \in \mathcal{T}} p(\mathbf{x}|T) \xi(T|D).$$

In some other situations, the structure of dependence between the variables, that is the graph  $G$ , might be the only object of interest. Lin et al. (2009) were for instance interested in the probability of an edge appearing in a tree. They looked out for the matrix  $\beta$  maximizing the likelihood of the data under a mixture of all possible tree models, where the probability of a tree model is defined just as in (2.3). The parameters of the models are estimated with

plug-in estimators. The distribution on trees cannot be called a prior in the traditional sense but the likeness to the model that we have described is obvious.

Here we are also interested in the probability for edges to appear in a tree, but in a full Bayesian framework. Formally, we would like to compute  $P(\{k, l\} \in E_{T^*} | D, \xi)$  for any edge  $\{k, l\}$ ,

$$P(\{k, l\} \in E_{T^*} | D, \xi) = \sum_{\substack{T \in \mathcal{T}, \\ E_T \ni \{k, l\}}} \xi(T | D). \quad (2.12)$$

The previous section shows that achieving this requires two things. First, we have to get access to  $\omega$  by computing local marginal likelihoods, which amounts to integrating w.r.t.  $\pi$  (Section 2.4.1). Then comes in the integration over the set of trees, that can be performed exactly thanks to an algebra result called the Matrix-Tree theorem (Section 2.4.2).

### 2.4.1 Integration with Respect to $\pi$

Thanks to the strong hyper Markov property required for the hyperdistributions, the integration on  $\pi$  can be performed locally and the compatibility ensures that these local integrated quantities can be passed from one model to another whenever they are needed. Thus, the integrations are always made on sets of bivariate distributions, with  $\frac{p(p+1)}{2}$  of them to be computed. The small dimension of each of the involved problems makes it possible to consider numerical or Monte Carlo integration. We begin by describing a framework based on tree-structured copulas where it might be needed, depending on the choice of local copulas. We then present two models where the local integrated likelihood terms can even be computed exactly thanks to conjugacy.

#### 2.4.1.a Tree-Structured Copulas

Let us assume that  $\mathcal{X} = [0, 1]^p$ . If we make the assumption that the marginal distribution of each variable is uniform, the joint distribution for  $\mathbf{X}$  is called a copula. Here we are interested in a subset of these distributions called the tree-structured copulas (Kirshner, 2007). We denote by  $\mathcal{U}$  the uniform distribution on  $[0, 1]$ . Let us assume that, for all  $i \in V$ ,  $X_i \sim \mathcal{U}$ . We are basically considering a copula model where the marginal data distributions have been dealt with in a relevant manner, independently from our model. For any  $i \in V$ , the marginal hyperdistribution  $\rho_i$  for  $\pi_i$  is then a Dirac distribution concentrated on  $\mathcal{U}$ , denoted  $\delta_{\mathcal{U}}$ . Defining a compatible family of hyperdistributions requires that we consider pairwise hyperdistributions whose marginals are concentrated on  $\mathcal{U}$ . Such hyperdistributions are in fact defined on bivariate copulas.

As an example, we consider the particular class of Archimedean copulas (Nelsen, 2006). Such copulas have simple expressions for their cdf. Let  $\psi : [0, 1] \rightarrow \mathbf{R}^+ \cup \{\infty\}$  be a continuous, strictly decreasing function such that  $\psi(1) = 0$ . Its pseudo-inverse  $\psi^{[-1]} : \mathbf{R}^+ \cup \{\infty\} \rightarrow [0, 1]$  is the continuous function defined by

$$\forall t \in \mathbf{R}^+ \cup \{\infty\}, \psi^{[-1]}(t) = \begin{cases} \psi^{-1}(t) & \text{if } 0 \leq t \leq \psi(0), \\ 0 & \text{otherwise.} \end{cases}$$

Let us remark that if  $\psi(0) = \infty$ ,  $\psi^{[-1]} = \psi^{-1}$ . The cdf of the Archimedean copula generated by  $\psi$  is given by

$$C_\psi(x_i, x_j) = \psi^{[-1]}(\psi(x_i) + \psi(x_j)).$$



$\psi$  is said to be a generator of the copula  $C_\psi$ . There is an extensive list of commonly used families of generators, many of them being governed by one or more parameters. Once again, we refer the reader to [Nelsen \(2006\)](#) for a detailed list of such generators. We can mention the well-known Gumbel copulas for instance, whose generator and inverse generator are given by

$$\begin{aligned}\psi_\theta(x) &= (-\log(x))^\theta & \forall x \in [0, 1], \\ \psi_\theta^{-1}(t) &= \exp(-t^{1/\theta}) & \forall t \in \mathbf{R}^+ \cup \{\infty\},\end{aligned}$$

with  $\theta \in [1, \infty)$  regulating the strength of the dependence.

Let  $\{i, j\}$  be a given edge. If we consider an identifiable parametric family of Archimedean copulas  $\{C_\theta\}_{\theta \in \Theta}$ ,  $\Theta \subseteq \mathbf{R}$ , defined by parametric generators  $\{\psi_\theta\}_{\theta \in \Theta}$ , there is a one-to-one mapping  $\Upsilon$  between  $\theta$  and the distributions  $\pi_{ij}$  on  $(X_i, X_j)$ . A pairwise hyperdistribution  $\rho_{ij}$  for  $\pi_{ij}$  is then easily defined by any distribution  $\kappa$  for  $\theta$  through the identity

$$\rho_{ij}(\pi_{ij}) = \kappa(\Upsilon^{-1}(\pi_{ij}))$$

and the integrated pairwise distribution  $p(x_i, x_j)$  is given by

$$p(x_i, x_j) = \int_{\Theta} \frac{\partial^2 C_\theta}{\partial x_i \partial x_j}(x_i, x_j) \kappa(\theta) d\theta, \quad \forall (x_i, x_j) \in [0, 1]^2. \quad (2.13)$$

Such a family of pairwise hyperdistributions is bound to be consistent since all marginals are equal to  $\delta_{\mathcal{U}}$ . Moreover, the global hyperdistributions that we obtain from this family are strong hyper Markov since it holds that, for  $i, j \in V$ ,  $\pi_{ij} \perp\!\!\!\perp \pi_j$  under  $\rho_{ij}$ .

The integrals given in (2.13) shall be computed exactly or through numerical integration depending on the choice of the copula family. This choice needs not be the same for all the edges. In the case of the Gumbel copula, a numerical or Monte Carlo integration is required. Obviously, bivariate Gaussian copulas would also be a valid choice. The pairwise hyperdistributions could then be specified through Wishart distributions for the precision matrices of the copulas, just like in the full Gaussian case described in Section 2.4.1.c.

#### 2.4.1.b Multinomial Distributions

We now consider the case where all  $X_i$  are discrete, taking their values in finite spaces  $\mathcal{X}_i$  of size  $r_i$  respectively. Let  $\mathcal{X}$  be the Cartesian product of spaces  $\mathcal{X}_i$ . A distribution for  $\mathbf{X}$  is given by a probability vector  $\theta$  in

$$\Theta = \left\{ \theta \in [0, 1]^{|\mathcal{X}|} \mid \sum_{\mathbf{x} \in \mathcal{X}} \theta(\mathbf{x}) = 1 \right\}. \quad (2.14)$$

$\Theta$  is the set of multinomial distributions on  $\mathcal{X}$ . It happens that the conjugate Dirichlet distribution is satisfying the condition given in (2.8) that is necessary to build a compatible family of strong hyper Markov hyperdistributions. Let  $\lambda = (\lambda(\mathbf{x}))_{\mathbf{x} \in \mathcal{X}}$  be a family of positive numbers indexed by  $\mathcal{X}$ . We denote  $\mathcal{D}(\lambda)$  the Dirichlet distribution for  $\theta \in \Theta$ , with density

$$f(\theta|\lambda) \propto \prod_{\mathbf{x} \in \mathcal{X}} \theta(\mathbf{x})^{\lambda(\mathbf{x})-1}.$$

**Proposition 2.4.** (*Dawid & Lauritzen, 1993, Lemma 7.2*) *Let  $\theta \sim \mathcal{D}(\lambda)$ . Then for all  $A \subseteq V$  and  $B = V \setminus A$ ,*

i.  $\theta_A \sim \mathcal{D}(\lambda_A)$ ;

ii.  $\theta_A \perp\!\!\!\perp \theta_{B|A}$ ;

with  $\lambda_A(\mathbf{x}_A) = \sum_{\mathbf{y}, \mathbf{y}_A = \mathbf{x}_A} \lambda(\mathbf{y})$  for all  $\mathbf{x}_A \in \mathcal{X}_A$ .

All these properties result from the fact that, if  $\{Y_k\}_{k=1}^K$  are independent random variables distributed as  $\Gamma(\lambda_k, \theta)$  respectively and  $V = \sum_{k=1}^K Y_k$ ,  $(Y_1/V, \dots, Y_K/V) \sim \mathcal{D}(\lambda)$ . (ii) assures that any  $\lambda$  gives rise to a hyperdistribution  $\rho$  on the multinomial family of distributions from which we can build a family of compatible strong hyper Markov hyperdistributions. (i) states that the marginal hyperdistributions are also Dirichlet distributed. The conjugacy can then be used locally to compute  $\omega$ . These hyperdistributions were referred to as hyper-Dirichlet laws in (Dawid & Lauritzen, 1993, §7.2.2).

As mentioned in Section 2.3.2, specifying a full set of hyperparameters  $\lambda$  is in fact not necessary to define the family of hyperdistributions  $\{\rho^T\}_{T \in \mathcal{T}}$ . We only need a consistent family of  $\{\lambda_{i,j}\}_{(i,j) \in V^2}$ , in the sense that, for  $(i, j, k) \in V^3$ ,  $\lambda_{i,j}$  and  $\lambda_{ik}$  should induce the same  $\lambda_i$ . A possibility is to set the prior hyperparameters on the edges thanks to an equivalent sample size  $N$ ,

$$\lambda_{i,j} := N/r_i r_j, \quad \lambda_i := N/r_i. \quad (2.15)$$

If all  $\mathcal{X}_i$  are of equal size  $r$ , a possibility is to set  $N = r^2/2$  so that all  $\lambda_{i,j}$  are equal to  $1/2$  to mimic Jeffreys priors for the bivariate distributions on the edges. However, this choice will not induce global Jeffreys priors, which are not hyper-Dirichlet hyperdistributions (York & Madigan, 1992). For an edge  $\{i, j\}$ , we denote  $\lambda'_{i,j}$  the updated hyperparameters for the edge  $\{i, j\}$ :

$$\lambda'_{i,j}(\ell, \ell') = \lambda_{i,j}(\ell, \ell') + \sum_{k=1}^n \delta_{x_i^k, \ell} \delta_{x_j^k, \ell'} \quad \forall (\ell, \ell') \in \mathcal{X}_i \times \mathcal{X}_j,$$

where  $\delta_{x, \ell} = 1$  if  $x = \ell$  and 0 otherwise.

The matrix  $\omega$  defined in (2.10) is then given by (Meilä & Jaakkola, 2006)

$$\omega_{i,j} = \beta_{i,j} \prod_{\ell \in \mathcal{X}_i} \frac{\Gamma(\lambda_i(\ell))}{\Gamma(\lambda'_i(\ell))} \prod_{\ell' \in \mathcal{X}_j} \frac{\Gamma(\lambda_j(\ell'))}{\Gamma(\lambda'_j(\ell'))} \prod_{(\ell, \ell') \in \mathcal{X}_i \times \mathcal{X}_j} \frac{\Gamma(\lambda'_{i,j}(\ell, \ell'))}{\Gamma(\lambda_{i,j}(\ell, \ell'))} \quad (2.16)$$

where  $\Gamma$  denotes the gamma function. If  $R = \max_{i \in V} r_i$ , computing  $\omega$  requires  $O(np^2 R^2)$  operations (Meilä & Jaakkola, 2006).

Let us finish this section by a remark on parameter independence. The following property of the Dirichlet distribution can be added to Proposition 2.4 even though it is not used here.

**Proposition 2.5.** (Dawid & Lauritzen, 1993, Lemma 7.2) *Let  $\theta \sim \mathcal{D}(\lambda)$ . Then for all  $A \subseteq V$  and  $B = V \setminus A$ ,  $\theta_{B|A}(\cdot | \mathbf{x}_A)$  are all independent and distributed as  $\mathcal{D}(\lambda_{B|A}(\cdot | \mathbf{x}_A))$  with  $\lambda_{B|A}(\mathbf{x}_B | \mathbf{x}_A) = \lambda(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{X}$  (up to a rearrangement of the components of  $\mathbf{x}$ ).*

Thus, although not required here, the local independence assumption made by Meilä & Jaakkola (2006) is in fact satisfied. In the multinomial case, Geiger & Heckerman (1997) even showed that, together with likelihood equivalence, global parameter independence and parameter modularity, the local parameter independence assumption constrains the prior to be locally Dirichlet distributed.





### 2.4.1.c Gaussian Distributions

Whenever  $\mathbf{X}$  is real-valued, one might work under the assumption that  $\mathbf{X}$  is Gaussian-distributed with mean  $\mu$  and inverse covariance matrix  $\Lambda$ . The conjugate normal-Wishart distribution is then a natural choice of prior for  $(\mu, \Lambda)$ . The normal-Wishart distribution is denoted by  $n\mathcal{W}(\nu, \lambda, \alpha, \Psi)$  and is hierarchically defined by

$$\begin{aligned}\Lambda &\sim \mathcal{W}(\alpha, \Psi), \\ \mu|\Lambda &\sim \mathcal{N}(\nu, (\lambda\Lambda)^{-1}),\end{aligned}$$

where  $\mathcal{W}(\alpha, \Psi)$  is the Wishart distribution with  $\alpha > p - 1$  degrees of freedom and positive-definite parametric matrix  $\Psi$ . Geiger & Heckerman (2002) showed that the normal-Wishart distribution satisfy the parameter independence property given in (2.8). They further proved that this property coerces the distribution to be normal-Wishart when  $p \geq 3$ . It can thus be used to build a compatible family of strong hyper Markov hyperdistributions. Moreover, for any partitioning  $(A, B)$  of  $V$ ,

$$\mathbf{X}_A \sim \mathcal{N}(\mu_A, (\Lambda_A - \Lambda_{AB}\Lambda_B^{-1}\Lambda_{AB}^T)^{-1})$$

and  $(\mu_A, \Lambda_A - \Lambda_{AB}\Lambda_B^{-1}\Lambda_{AB}^T)$  is also normal-Wishart-distributed with parameters

$$(\nu_A, \lambda, \alpha - p + l, \Psi_A - \Psi_{AB}\Psi_B^{-1}\Psi_{AB}^T)$$

where all indices are understood as partitioning of the corresponding vectors and matrices according to  $(A, B)$ .

The pairwise marginal likelihoods can then be computed by updating the hyperparameters of the basis hyperdistribution to  $(\nu', \lambda', \alpha', \Psi')$  thanks to classical Bayesian updating formulæ. The locally updated hyperparameters are then derived from the globally updated ones and

$$p(D_i, D_j) \propto \frac{|\Psi_{\{i,j\}}|^{\frac{\alpha-p+2}{2}}}{|\Psi'_{\{i,j\}}|^{\frac{\alpha'-p+2}{2}}}, \quad p(D_i) \propto \frac{|\Psi_i|^{\frac{\alpha-p+1}{2}}}{|\Psi'_i|^{\frac{\alpha'-p+1}{2}}}, \quad (2.17)$$

where, for a matrix  $M$  and  $i, j \in V$ ,  $M_{\{i,j\}}$  denotes the submatrix of size 2 corresponding to vertices  $i$  and  $j$ . This result is given in the work of Kuipers et al. (2014) as a correction to the erroneous result stated in Geiger & Heckerman (2002).

The compatible hyperdistributions built on  $(\mu, \Lambda)$  are called hyper-normal-Wishart distributions. One can notice that  $\Lambda^{-1}$  follows a hyper-inverse-Wishart distribution (Dawid & Lauritzen, 1993, §7.3.2).

### 2.4.2 Integration with Respect to $T$

We assume that we have knowledge of  $\omega$ . Consequently, we know  $\xi(\cdot|D)$  up to the normalising constant  $Z$ . For an edge  $\{k, l\}$ , gaining access to  $P(\{k, l\} \in E_{T^*}|D, \xi)$  means being able to sum the posterior tree distribution over the trees that possess the edge  $\{k, l\}$ . Because we are only considering trees, this is tractable thanks to the Matrix-Tree theorem.

Let  $\omega = (\omega_{i,j})_{(i,j) \in V^2}$  be a symmetric weight matrix such that, for all  $i \in V$ ,  $\omega_{i,i} = 0$ , the off-diagonal terms being non-negative. The weight of a graph  $G = (V, E_G)$  is defined as the product of the weights of its edges,

$$\omega_G := \prod_{\{i,j\} \in E_G} \omega_{i,j}.$$

The Laplacian  $\Delta = (\Delta_{i,j})_{(i,j) \in V^2}$  of  $\omega$  is given by

$$\Delta_{i,j} = \begin{cases} -\omega_{i,j} & \text{if } i \neq j, \\ \sum_j \omega_{i,j} & \text{if } i = j. \end{cases}$$

For  $U \subseteq V$ , we defined  $\Delta^U$  as the matrix obtained from  $\Delta$  by removing the rows and columns corresponding to  $U$ .

**Theorem 2.1** (Chaiken, 1982). *Let  $\Delta$  be the Laplacian of a weight matrix  $\omega$ . Then all minors  $|\Delta^{\{u\}}|$ ,  $u \in V$ , are equal and the following identity holds:*

$$|\Delta^{\{u\}}| = \sum_{T \in \mathcal{T}} \omega_T. \quad (2.18)$$

We directly get the normalising constant of  $\xi(T|D)$  from this result.

There is a more general version of this theorem concerning graphs whose connected components are spanning trees on their respective sets of vertices. Such graphs are called forests.

**Theorem 2.2** (All Minors Matrix-Tree theorem, Chaiken, 1982). *Let  $\Delta$  be the Laplacian of a weight matrix  $\omega$  and  $U \subseteq V$ . Let  $\mathcal{F}_U$  be the set of forests on  $V$  with  $|U|$  connected components such that, for any two vertices  $u_1, u_2 \in U$ ,  $u_1$  and  $u_2$  are not in the same connected component. Then*

$$|\Delta^U| = \sum_{F \in \mathcal{F}_U} \omega_F. \quad (2.19)$$

Briefly speaking,  $U$  can be seen as a set of ‘‘roots’’ (even though the models are not directed) for the trees of the forests in  $\mathcal{F}_U$ . If  $U$  is taken equal to a single vertex, then the forests in  $\mathcal{F}_U$  only have one connected component which is a tree and we get Theorem 2.1. This theorem will be used in the proof of the following result that was first stated by Kirshner (2007).

**Theorem 2.3** (Kirshner, 2007). *Let  $\omega$  be defined as in (2.10) and  $\Delta$  be the associated Laplacian. Let  $u$  be a vertex in  $V$ . We define  $Q = (\Delta^{\{u\}})^{-1}$ . Then, for all  $\{k, l\} \in \mathcal{P}_2(V)$ ,*

$$P(\{k, l\} \in E_{T^*} | D, \xi) = \begin{cases} \omega_{k,l} (Q_{k,k} + Q_{l,l} - 2Q_{k,l}) & \text{if } k \neq u, l \neq u, \\ \omega_{k,u} Q_{k,k} & \text{if } l = u, \\ \omega_{u,l} Q_{l,l} & \text{if } k = u. \end{cases} \quad (2.20)$$

A proof of this result is provided in the extended version of (Kirshner, 2007) available online. We provide a shorter version relying on the generalized version of the Matrix-Tree theorem given above (see Eq. 2.21).



*Proof.* Let  $\{k, l\}$  be an edge in  $\mathcal{P}_2(V)$ . Let  $Z$ ,  $Z_{kl}^+$  and  $Z_{kl}^-$  respectively denote the sums of  $\omega_T$  over the sets  $\mathcal{T}$ ,  $\{T \in \mathcal{T} : \{k, l\} \in E_T\}$  and  $\{T \in \mathcal{T} : \{k, l\} \notin E_T\}$ . It is immediate to see that  $Z = Z_{kl}^+ + Z_{kl}^-$ .

Lemma 3 of Meilä & Jaakkola (2006) states that

$$\frac{\partial Z}{\partial \omega_{k,l}} = M_{k,l} |\Delta^{\{u\}}| = M_{k,l} Z$$

where  $M$  is a symmetric matrix with 0 diagonal defined by

$$M_{i,j} = \begin{cases} (Q_{i,i} + Q_{j,j} - 2Q_{i,j}) & \text{if } i \neq u, j \neq u, \\ Q_{i,i} & \text{if } j = u, \\ Q_{j,j} & \text{if } i = u. \end{cases}$$

It is easy to see that  $Z_{kl}^-$  can be obtained by applying Theorem 2.1 to a weight matrix equal to  $\omega$  except for the terms  $\omega_{k,l}$  and  $\omega_{lk}$  that are set to 0. This means that  $Z_{kl}^-$  does not depend on  $\omega_{k,l}$  and

$$\frac{\partial Z}{\partial \omega_{k,l}} = \frac{\partial Z_{kl}^+}{\partial \omega_{k,l}}.$$

By Theorem 2.2,

$$Z_{kl}^+ = \omega_{k,l} \sum_{F \in \mathcal{F}_{\{k,l\}}} \omega_F = \omega_{k,l} |\Delta^{\{k,l\}}|. \quad (2.21)$$

$|\Delta^{\{k,l\}}|$  does not depend on  $\omega_{k,l}$  since the only terms of  $\Delta$  that depend on  $\omega_{k,l}$  are  $\Delta_{k,l}$ ,  $\Delta_{lk}$ ,  $\Delta_{k,k}$ ,  $\Delta_{l,l}$  and these terms are all withdrawn in  $\Delta^{\{k,l\}}$ . Therefore,

$$|\Delta^{\{k,l\}}| = \frac{\partial Z_{kl}^+}{\partial \omega_{k,l}} = \frac{\partial Z}{\partial \omega_{k,l}} = M_{k,l} Z. \quad (2.22)$$

Combining (2.21) and (2.22) with the fact that  $P(\{k, l\} \in E_T | D, \xi) = Z_{kl}^+ / Z$ , we get the claimed result.  $\square$

Theorem 2.3 shows that all posterior probabilities for the edges can be computed at once by inverting a matrix of size  $p - 1$ , amounting to a complexity of  $O(p^3)$ .

In a Bayesian framework, the posterior entropy gives insight about the concentration of the posterior distribution, which is for instance of particular interest when a MAP approach is considered. The computation of this quantity is not always straightforward, but here, it can be obtained at small cost once posterior probabilities for the edges have been computed.

**Proposition 2.6.** *The entropy of the posterior distribution on trees  $\xi(\cdot | D)$  can be computed with complexity  $O(p^3)$ .*

*Proof.* We show that the entropy has a simple expression depending on  $Z$  and  $(P(\{k, l\} \in E_{T^*} | D, \xi))_{\{k,l\} \in \mathcal{P}_2(V)}$  which can both be computed with complexity  $O(p^3)$  through Theorems

2.1 & 2.3. Indeed,

$$\begin{aligned}
H(\xi(\cdot|D)) &= - \sum_{T \in \mathcal{T}} \xi(T|D) \log(\xi(T|D)) \\
&= \sum_{T \in \mathcal{T}} \frac{1}{Z} \prod_{\{i,j\} \in E_T} \omega_{i,j} \left( \log(Z) - \sum_{\{k,l\} \in E_T} \log(\omega_{k,l}) \right) \\
&= \log(Z) - \sum_{\{k,l\} \in \mathcal{P}_2(V)} \frac{\log(\omega_{k,l})}{Z} \sum_{T \ni \{k,l\}} \prod_{\{i,j\} \in E_T} \omega_{i,j} \\
&= \log(Z) - \sum_{\{k,l\} \in \mathcal{P}_2(V)} \log(\omega_{k,l}) P(\{k,l\} \in E_{T^*} | D, \xi).
\end{aligned}$$

□

### 2.4.3 Controlling prior edge probability

If the distribution on trees is not strongly peaked, the prior probability for an edge to appear in a random tree can be quite small. For instance, the uniform distribution on  $\mathcal{T}$  leads to any edge appearing with probability  $2/p$ . Indeed, no edge is favoured and each tree borrows  $p - 1$  of the  $p(p - 1)/2$  possible edges. We consider an edge  $\{k, l\} \in \mathcal{P}_2(V)$  and the event  $\mathcal{E}_{kl} := \{T : \{k, l\} \in E_T\}$ . We let  $p_{kl}^0$  and  $p_{kl}$  respectively denote the prior and posterior probabilities of event  $\mathcal{E}_{kl}$ . These probabilities are obtained through Theorem 2.3.

In a decision perspective, it might be desirable to allow some control on the prior probability of  $\mathcal{E}_{kl}$ . To this aim, we use a binary random variable  $\epsilon_{kl} \sim \mathcal{B}(\lambda_{kl})$  explicitly controlling the status of edge  $\{k, l\}$  in the random tree:

$$p(T|\epsilon_{kl}, \xi) = \begin{cases} \xi(T|\mathcal{E}_{kl}) & \text{if } \epsilon_{kl} = 1 \\ \xi(T|\bar{\mathcal{E}}_{kl}) & \text{if } \epsilon_{kl} = 0 \end{cases}.$$

In particular, the choice  $\lambda_{kl} = 1/2$  takes us back to a non-informative prior configuration regarding  $\mathcal{E}_{kl}$ . We obtain the model represented in Figure 2.3 in which the fully marginal likelihood can be written as

$$p(D) = \lambda_{kl} p(D|\mathcal{E}_{kl}) + (1 - \lambda_{kl}) p(D|\bar{\mathcal{E}}_{kl}).$$

We are now interested in the posterior distribution of  $\epsilon_{kl}$ .

**Proposition 2.7.**

$$P(\epsilon_{kl} = 1|D) = \lambda_{kl} \frac{p_{kl}}{p_{kl}^0} \cdot \left[ \lambda_{kl} \frac{p_{kl}}{p_{kl}^0} + (1 - \lambda_{kl}) \frac{1 - p_{kl}}{1 - p_{kl}^0} \right]^{-1}$$

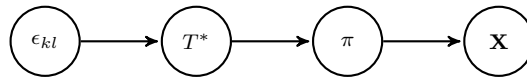


Figure 2.3 – Model with variable  $\epsilon_{kl}$  explicitly controlling the status of edge  $\{k, l\}$  in  $T^*$ .



*Proof.*

$$\begin{aligned}
P(\epsilon_{kl} = 1|D) &= \frac{p(D|\epsilon_{kl} = 1)P(\epsilon_{kl} = 1)}{p(D)} = \lambda_{kl} \frac{p(D|\mathcal{E}_{kl})}{p(D)} \\
&= \lambda_{kl} p(D|\mathcal{E}_{kl}) \cdot [\lambda_{kl} p(D|\mathcal{E}_{kl}) + (1 - \lambda_{kl}) p(D|\bar{\mathcal{E}}_{kl})]^{-1} \\
&= \lambda_{kl} \frac{p_{kl}}{p_{kl}^0} \cdot \left[ \lambda_{kl} \frac{p_{kl}}{p_{kl}^0} + (1 - \lambda_{kl}) \frac{1 - p_{kl}}{1 - p_{kl}^0} \right]^{-1}
\end{aligned}$$

□

The computation of  $P(\epsilon_{kl} = 1|D)$  for all edges can be achieved in  $O(p^2)$  time from the posterior edge probability matrix  $\{p_{kl}\}_{\{k,l\} \in \mathcal{P}_2(V)}$ . We can notice that  $P(\epsilon_{kl} = 1|D)$  is a strictly increasing function of  $p_{kl}$ . When the initial prior on trees  $\xi$  is uniform and all  $\lambda_{kl}$  are taken equal, the order induced on the edges by  $\{P(\epsilon_{kl} = 1|D)\}_{\{k,l\} \in \mathcal{P}_2(V)}$  is identical to the order induced by the posterior edge probability matrix. The ROC and PR curves that are commonly used to assess network inference accuracy therefore remain unchanged.

## 2.5 Simulations

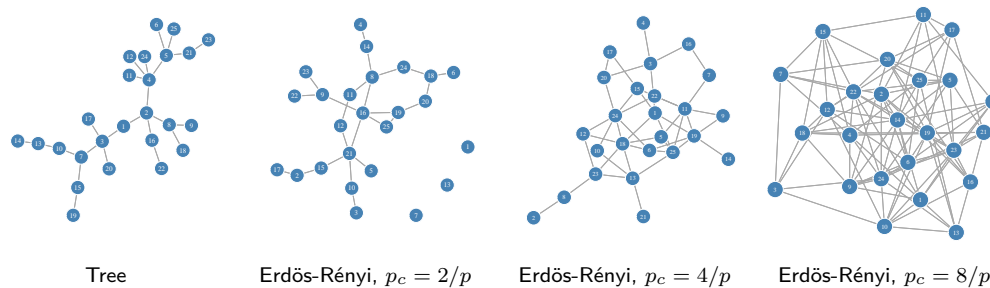
In this section, we use synthetic data to meet a twofold objective. On one hand, the aim of this study is to show that there is an advantage in averaging over trees rather than considering a single MAP estimate. On the other hand, we show that assuming a tree structure is not substantially more detrimental to the accuracy of the inference of non-tree-structured graphical models than assuming a DAG structure. To do so, we compare our method with another fully Bayesian inference method carried out on DAGs, described by Niinimäki et al. (2011) and implemented in the BEANDisco software. Computations for our approach were performed with the R package **saturnin**.

### 2.5.1 Simulation Scheme

We have chosen four typical networks with  $p = 25$  vertices, namely a tree and three Erdős-Rényi random graphs drawn with probabilities of connection  $p_c = 2/p, 4/p$  and  $8/p$ . These graphs are shown in Figure 2.4. Data sets are then simulated according to Gaussian graphical models. For all four adjacency matrices  $A$ , we used the Laplacian matrix of  $A$  augmented of 1 on the diagonal as precision matrix  $\Lambda_A$ . This construction ensures that  $\Lambda_A$  is non-singular. Independent samples were drawn according to  $\mathcal{N}(0, \Lambda_A^{-1})$  and discretized into  $r = 3$  bins. For  $n = 25, 50, 75, 100$  and  $200$ , we generated 100 data sets of size  $n$ .

We then considered the Multinomial/Dirichlet framework described in Section 2.4.1.b, setting the prior on trees  $\xi$  to the uniform and the equivalent prior sample size  $N$  to  $r^2/2 = 4.5$  (see Eq. (2.15)). For each data set, we computed

- the MAP tree structure in our model thanks to a Maximal Spanning tree algorithm applied to  $\omega$ ;
- the matrix of posterior edge probabilities  $P(\{k, l\} \in E_{T^*} | D)$  in our model. For all the edges, the prior appearance probability was brought back to  $q_0^{(kl)} = 1/2$  (see Section 2.4.3);



**Figure 2.4** – Gold standard networks in the simulation study.

- an estimation of the matrix of posterior edge appearance probabilities in a random DAG obtained by MCMC sampling (Niinimäki et al., 2011). We refer the reader to this paper for details on the prior distribution on DAGs. We ran the code provided by the authors with default parameters. The direction of the edges of the sampled DAGs was not taken into account to get empirical frequencies for all undirected edges.

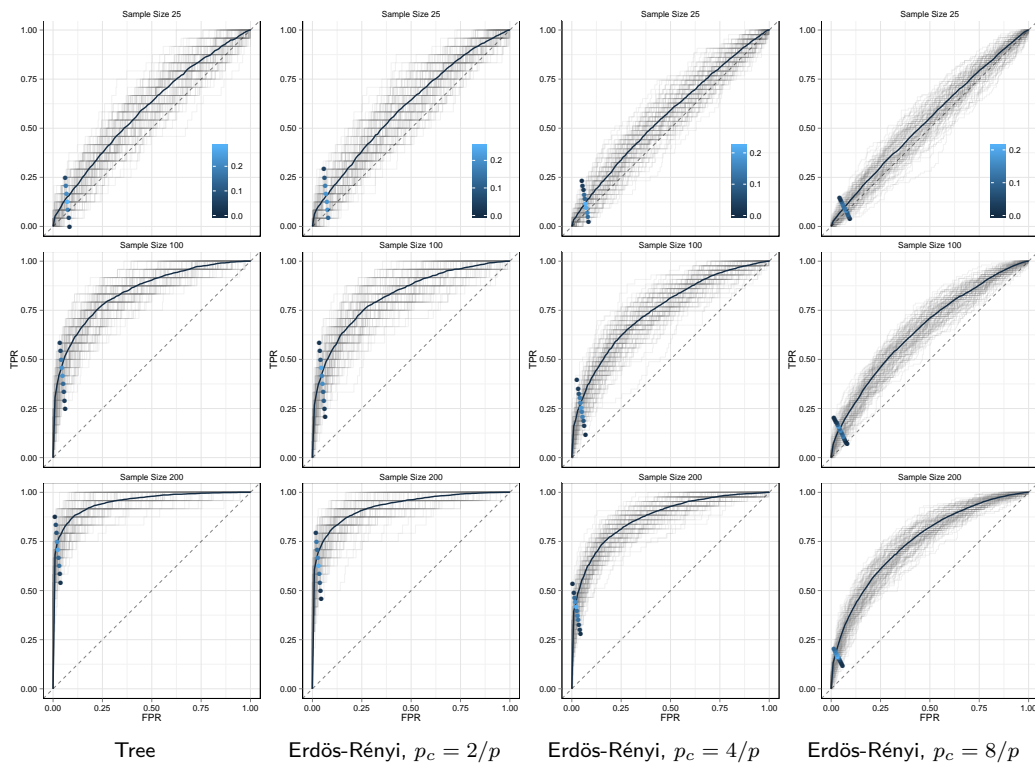
We also considered the normal-Wishart/Gaussian framework described in Section 2.4.1.c on the undiscretized data.

The accuracy of the inference was evaluated against the true adjacency matrix according to the yielded outputs. In the case of the MAP estimate, we calculated the True and False Positives Rates (TPR, FPR) between the best tree and the true graph. These rates are constrained by the fact that a spanning trees on  $p$  vertices has exactly  $p - 1$  edges. For the (estimated) posterior edge appearance probability matrices, ROC and PR curves against the true adjacency matrix are plotted and summarized by the area under the curves.

## 2.5.2 Results

**COMPARISON WITH MAP.** Figure 2.5 simultaneously represents the (TPR, FPR) scores and the ROC curves obtained for the MAP estimate and the tree posterior edge appearance probability matrix respectively. It makes sense to plot both results on the same graph since a ROC curve is just a succession of (TPR, FPR) points computed as more and more edges are selected, going from the most to the least likely. When  $p - 1$  edges are selected, both methods behave similarly. So, if there is external evidence that the true graph is in fact a tree, a MAP approach could be considered but using posterior edge probabilities would do as well. Nonetheless, when the true graph is not a tree, the MAP approach is penalized by its lack of flexibility. Computing posterior appearance probabilities for the edges allows to retain an arbitrary number of edges. The balance between selectivity and sensibility achieved by the MAP approach can obviously be improved by selecting more edges. An other argument in favor of considering the whole posterior distribution on trees instead of the MAP is presented in Figure 2.6. For all four simulation scenarios, posterior tree distributions are not really peaked around their modes, especially for small samples, with the second most probable tree always being very close to the MAP. Moreover, the entropy of the posterior distribution on trees behaves similarly across all scenarios. We could have



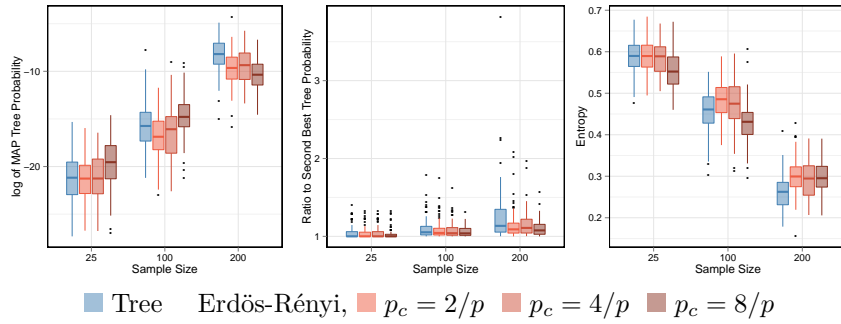


**Figure 2.5** – ROC curves for the posterior edge probabilities and (TPR, FPR) scores for the MAP estimate on data sets of size 25, 50 and 200 (from top to bottom). For the ROC curves, the mean curve is plotted in bold line. The color of a (TPR, FPR) point expresses its frequency within the 100 samples.

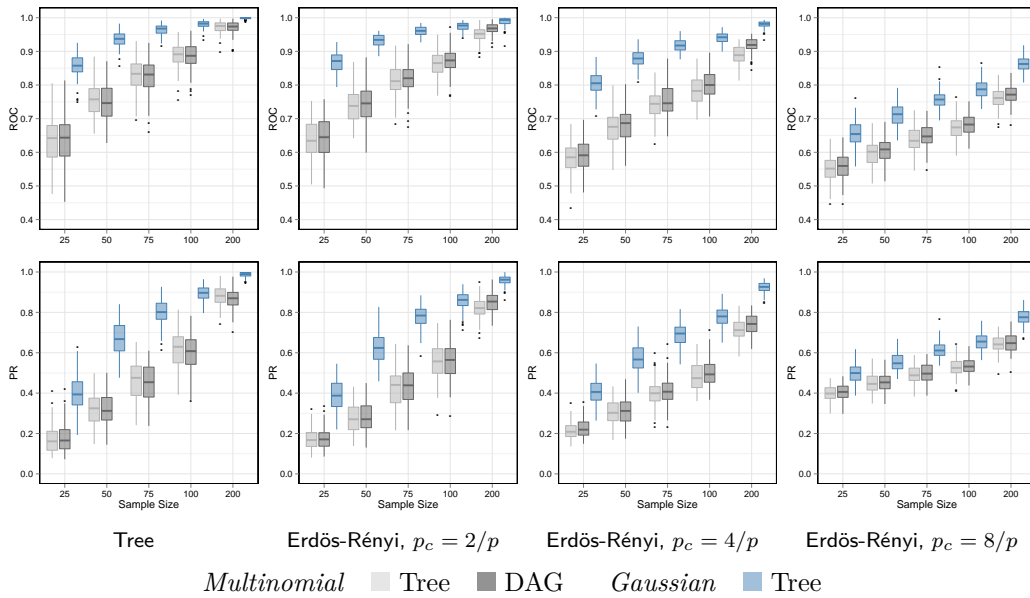
expected a singular behaviour in the tree scenario, but it is only slightly observed for  $n = 200$ .

**INFLUENCE OF THE TREE ASSUMPTION.** We now study the influence of the tree assumption on the accuracy of structure inference when the true graphical model is not tree-structured. With this end in view, we consider a similar model where DAGs are drawn instead of trees and use the posterior edge appearance probabilities yielded by this model as gold standard, as it achieves the same goal in terms of Bayesian inference within a larger class of graphs. Results are given in Figure 2.7. Both algorithms seem to perform equally well in all four situations. The accuracy of the inference expectedly increases with sample size. The results we get here indicate that the posterior probabilities for the edges to belong to a random tree can be relevant even when the true network is not a tree, with no clear evidence in favor of considering an inference within the broader class of DAG structures.

**RUNNING TIME.** We conclude this section on synthetic data by mentioning running times (Table 2.1). While retaining similar accuracy to the algorithm based on MCMC sampling in



**Figure 2.6** – Posterior probability of the MAP tree, ratio to the posterior probability of the second best tree and entropy of the posterior tree distribution (normalised by the entropy of the uniform distribution on  $\mathcal{T}$ , i.e.  $(p - 2) \log(p)$ ).



**Figure 2.7** – Area under the ROC (top) & PR (bottom) curves computed for the output of our approach and of the MCMC sampling algorithm in the DAGs on the multinomial samples and for the output of our approach in the Gaussian setting on the undiscretized data.

the space of DAGs that we used as a point of comparison, our algorithm runs significantly faster than the MCMC sampling ran with default parameters, especially for large networks. Of course, the accuracy of the MCMC sampling approach could be improved by augmenting the number of samples at the cost of even longer running times, but we have not observed any evidence going that way.

## 2.6 Application to Cytometry Data

This section presents an application of our approach to flow cytometry data. They have been collected by Sachs et al. (2005) and were used by Werhli et al. (2006) in a review of





Network Size	DAG MCMC	Tree
p=25	11 s	0.2 s
p=50	206 s	1 s
p=75	1393 s	2.2 s

**Table 2.1** – Average running times for different network sizes with our method (Tree) and the MCMC approach on DAGs (DAG MCMC) on data sets of size  $n = 100$ .

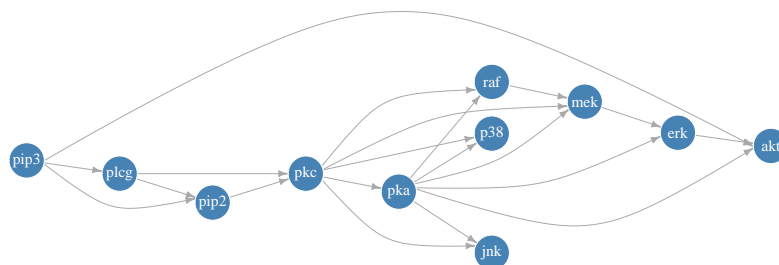
network inference techniques. They are related to the Raf cellular signalling network, which is involved in many different processes, including the regulation of cellular proliferation in human immune cells. The activation levels of the 11 proteins and phospholipids that are part of this pathway can be measured by flow cytometry. The generally accepted structure of the Raf pathway is given in Figure 2.8, but the true structure of this network, despite considerable experimental and theoretical efforts, may be more subtle. The undirected skeleton of this network will, however, be used as the gold standard network in our study.

### 2.6.1 Data

In flow cytometry experiments, cells are suspended in a stream of fluid and go through a laser beam one at a time. Different parameters are then measured on each cell by recovering the light that is reemitted by diffusion or fluorescence. We are interested in the activation levels (also called phosphorylation levels) of the involved proteins and phospholipids. Such experiments typically produce samples of several thousands observations. Since all biological network inference problems are not met by such a profusion of data, Werhli et al. (2006) sampled down 5 samples with 100 data points from the data provided by Sachs et al. (2005). We discretized each sample into  $r=3$  bins and performed the inference on each of them with our algorithm (Tree) and the MCMC sampling in DAGs algorithm (DAG), just like in the previous section. The accuracy of the inference was once again assessed by the area under the ROC and PR curves, averaged on all 5 samples.

### 2.6.2 Results

The results of the inference are reported in Table 2.2. The DAG approach performs globally better than our inference within trees. These results qualify those of the previous section. Nonetheless, we would like to make the following points. While not being as accurate, our approach still provides good results and might in fact be more adapted to bigger problems where MCMC sampling can hardly be contemplated. Moreover, unlike the simulation



**Figure 2.8** – Raf pathway.

study, the gold standard network against which the accuracy of the inference is assessed here, shown in Figure 2.8, is not perfectly known and may still differ quite considerably from the truth. The difference observed between the two approaches is small enough that it could be reversed if the status of an edge was to be changed in the gold standard network.

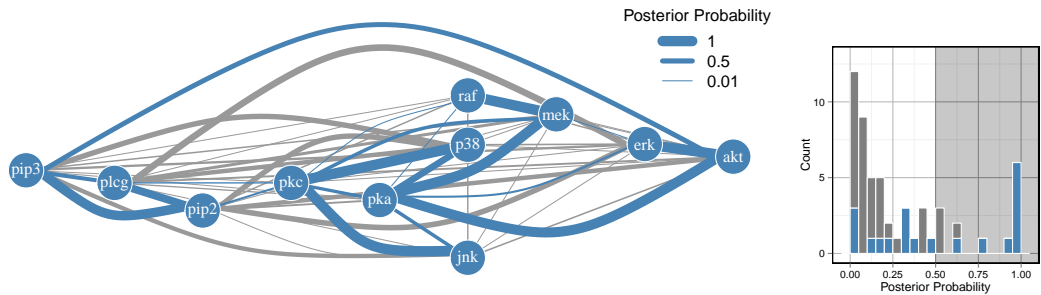
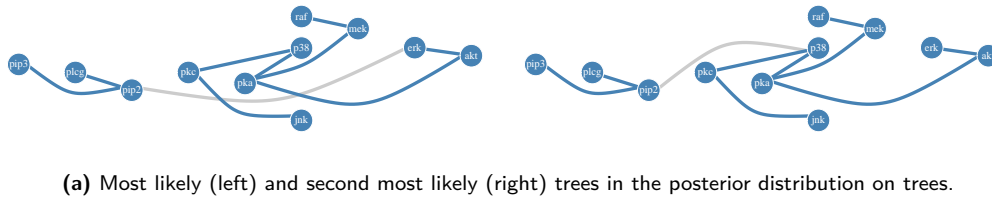
Figure 2.9 gives a graphical representation of the results obtained on one of the five data sets, offering a more detailed overview. We note that the gold standard network as defined here has 20 edges. The two likeliest trees in the posterior tree distribution are given in Figure 2.9a. Both trees have 9 true positives out of the  $p - 1 = 10$  edges they respectively selected. As expected, most of these edges also have strong posterior probabilities (Figure 2.9b). When the prior probabilities of all edges is brought back to  $1/2$ , we get 13 edges with posterior probabilities strictly greater than  $1/2$ , among which the same true positives as in the MAP estimate. More generally, one could consider using the histogram of posterior probabilities to empirically find a more appropriate cut-off.

We did not represent the empirical edge frequencies obtained for DAGs since prior appearance probabilities could not be easily accounted for in this case, thus making direct comparison with posterior edge probabilities in trees impossible.

As a conclusion, these results lead us to believe that it might be preferable to favour inference using DAGs for small problems. When that is no longer possible in a reasonable

	DAG	Tree
ROC	0.767 (0.068)	0.729 (0.047)
PR	0.725 (0.070)	0.690 (0.051)

**Table 2.2** – Inference results on flow cytometry data. Area under the ROC and PR curves for different discretization levels (standard deviation).



**Figure 2.9** – Graphical representation of the results obtained on one of the five data sets. The edges of the golden standard network are colored in blue.



amount of time, performing exact inference in a model based on trees is a computationally efficient alternative that can be used at a limited cost.

# 3

## Segmentation

---

<b>3.1</b>	<b>Introduction</b>	<b>60</b>
<b>3.2</b>	<b>Background</b>	<b>62</b>
3.2.1	Product Partition Models	62
3.2.2	Tree-structured Graphical Models	63
<b>3.3</b>	<b>Model &amp; Properties</b>	<b>65</b>
3.3.1	Model	65
3.3.2	Factorisation Properties	66
<b>3.4</b>	<b>Quantities of Interest</b>	<b>68</b>
3.4.1	Change-point Location	68
3.4.2	Number of Segments	68
3.4.3	Posterior Edge Probability	69
<b>3.5</b>	<b>Edge Status &amp; Structure Comparisons</b>	<b>70</b>
3.5.1	Edge Status Comparison	70
3.5.2	Structure Comparison	71
<b>3.6</b>	<b>Simulations</b>	<b>72</b>
3.6.1	Simulation Scheme	72
3.6.2	Results	73
<b>3.7</b>	<b>Applications</b>	<b>75</b>
3.7.1	Drosophila Life Cycle Microarray Data	75
3.7.2	Functional MRI Data	77
<b>3.8</b>	<b>Discussion</b>	<b>79</b>

---

*We consider the problem of change-point detection in multivariate time-series. The multivariate distribution of the observations is supposed to follow a graphical model, whose graph and parameters are affected by abrupt changes throughout time. Building on the approach develop in Chapter 2, we assume that all graphical models are tree-structured, and make the best use of this assumption to compute a variety of quantities of interest in polynomial time*



with respect to the number of vertices and the length of the series. This chapter has been published as an article in the journal *Statistics & Computing* (Schwaller & Robin, 2016).

### 3.1 Introduction

We are interested in time-series data where several variables are observed throughout time. An assumption often made in multivariate settings is that there exists an underlying network describing the dependences between the different variables. When modelling time-series data, one is faced with a choice: shall this network be considered stationary or not? Taking the example of genomic data, it might for instance be unrealistic to consider that the network describing how a pool of genes regulate each other remains identical throughout time. This network might slowly evolve, or undergo abrupt changes leading to the initialisation of new morphological development stages in the organism of interest. Here, we focus our interest on the second scenario.

The inference of the dependence structure ruling a multivariate time-series was first performed under the assumption that this structure was stationary (*e.g.* (Friedman et al., 1998; Murphy & Mian, 1999)). Non-stationarity has then been addressed in a variety of ways. Classical Dynamic Bayesian Networks (DBNs) can for instance be adapted to allow the directed graph (or Bayesian Network) describing the interactions between two consecutive time-points to change, leading to so-called switching DBNs (Robinson & Hartemink, 2010; Lèbre et al., 2010; Grzegorzczak & Husmeier, 2011). Some models alternatively suppose that the heterogeneity is the result of parameters changing smoothly with time (Zhou et al., 2010; Kolar et al., 2010). This is especially appropriate for Gaussian graphical models where the graph structure can directly be read in the non-zero terms of the precision (or inverse-covariance) matrix, therefore enabling smooth transitions within the otherwise discrete space of graphs. Hidden Markov Models (HMM) have also been used to account for heterogeneity in multivariate time-series (Fox et al., 2009; Barber & Cemgil, 2010). In the aforementioned models, the inference can rarely be performed exactly, and often relies on sampling techniques such as Markov Chain Monte Carlo (MCMC).

The model that we consider here belongs to the class of product partition models (PPM) (Barry & Hartigan, 1992). We assume that the observed data  $\{y^t\}_{t=1,\dots,N}$  are a realisation of a process  $\{Y^t\}_{t=1,\dots,N}$  where, for  $1 \leq t \leq T$ ,  $Y^t$  is a random vector with dimension  $p \geq 2$ . If  $m$  is a segmentation of  $\{1, \dots, T\}$  with change-points  $1 = \tau_0 < \tau_1 < \dots < \tau_{K-1} < \tau_K = N$ , the model has the general form

$$Y_t \sim \pi(G_r, \theta_r), \quad \text{if } t \in r \text{ and } r = \llbracket \tau_i, \tau_{i+1} \rrbracket,$$

where  $G_r$  and  $\theta_r$  respectively stand for the graph describing the dependence structure and

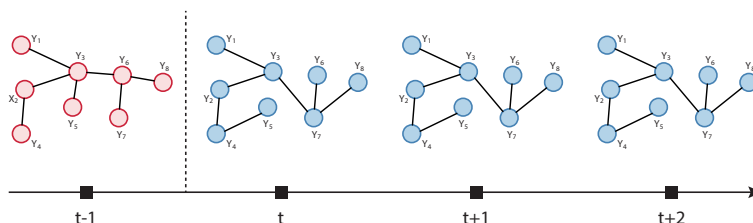


Figure 3.1 – Illustration of the change-point detection problem in the tree structure of a graphical model.

the distribution parameters on segment  $r$ . The parameters  $(G_r, \theta_r)$  are assumed to be independent between segments. This model is illustrated in Figure 3.1.

We are interested in retrieving all change-points at the same time, therefore performing off-line detection. It has been shown that both off-line (Fearnhead, 2006; Rigaiil et al., 2012) and on-line detection (Fearnhead & Liu, 2007; Caron et al., 2012) of change-points can be performed exactly and efficiently in this model thanks to dynamic programming. Xuan & Murphy (2007) explicitly consider this framework in a multivariate Gaussian setting. They estimate a set of possible structures for their model by performing regularized estimation of the precision matrix on arbitrary overlapping time segments. This graph family is then taken as a starting point in an iterative procedure where the segmentation and the graph family are sequentially updated to get the best segmentation and graph series.

**OUR CONTRIBUTION** From a Bayesian point of view, the problem at hand raises an interesting and quite typical problem as both continuous and discrete parameter are involved in the model. Indeed, the location and scale parameters or, more specifically, the means and (conditional) covariances associated with each segments are continuous but the location of the change-points and the structure of the graphical model within each segments are not. Denoting  $\theta$  the set of continuous parameters,  $Q$  the set of discrete parameters and  $y$  the observed data, Bayesian inference will typically rely on integrals such as the marginal likelihood of the data, that is

$$p(y) = \sum_{Q \in \mathcal{Q}} p(Q) \int_{\theta \in \Theta} p(y|\theta, Q) p(\theta|Q) d\theta.$$

In many situations, the use of conjugate priors allows us to compute the integral with respect to  $\theta$  in an exact manner. Still, the summation over all possible values for the discrete parameter  $Q$  is often intractable due to combinatorial complexity. One aim of this article is to remind that the algebraic properties of the space  $\mathcal{Q}$  can sometimes help to actually achieve this summation in an exact manner, so that a fully exact Bayesian inference can be carried out.

We show that exact and efficient Bayesian inference can be performed in a multivariate product partition model within the class of undirected graphs called spanning trees. These structures are connected graphs, with no cycles. When  $p$  nodes are considered, we are left with  $p^{p-2}$  possible spanning trees, but exact inference remains tractable by using algebraic properties pertaining to this set of graphs. On each independent temporal segment, we place ourselves in the framework developed by Schwaller et al. (2015), in which the likelihood of a segment  $\llbracket s; t \rrbracket$ , defined by

$$p(y^{\llbracket s; t \rrbracket}) := \sum_{T \in \mathcal{T}} \int p(y^{\llbracket s; t \rrbracket} | \theta, T) p(\theta | T) d\theta,$$

where  $\mathcal{T}$  stands for the set of spanning trees, can be computed efficiently. We provide explicit and exact formulas for quantities of interest such as the posterior distribution of change-points or posterior edge probabilities over time. We also provide a way to assess whether the status of an edge (or of the whole graph) remains identical throughout the time-series or not when the partition is given.

**OUTLINE** In Section 3.2, we provide some background on graphical models and product partition models. In particular, we give a more detailed presentation of the results of Rigaiil



et al. (2012) on dynamic programming used for change-point detection problems. We also introduce tree-structured graphical models. The model and its properties are presented in Section 3.3. Section 3.4 enumerates a list of quantities of interest that can be computed in this model, while Section 3.5 deals with edge and graph status comparison, when the segmentation is known. Sections 3.6 and 3.7 respectively present the simulation study and the applications to both biological and neuroscience data.

## 3.2 Background

In this section we introduce two models involving a discrete parameter, for which exact integration over this parameter is possible.

### 3.2.1 Product Partition Models

Let  $Y = \{Y^t\}_{t=1,\dots,N}$  be an independent random process and let  $y$  be a realisation of this process. For any time interval  $r$ , we let  $Y^r := \{Y^t\}_{t \in r}$  denote the observations for  $t \in r$ . PPMs as described in (Barry & Hartigan, 1992) work under the assumption that the observations can be divided in independent adjacent segments. Thus, if  $m$  is a partition of  $\llbracket 1; N \rrbracket$ , the likelihood of  $y$  conditioned on  $m$  can be written as

$$p(y|m) = \prod_{r \in m} p(y^r|r),$$

$$p(y^r|r) = \int \left( \prod_{t \in r} p(y^t|\theta_r) \right) p(\theta_r) d\theta_r,$$

where  $\theta_r$  is a set of parameters giving the distribution of  $Y^t$  for  $t \in r$ . For the sake of clarity, we let  $p(y^r)$  denote  $p(y^r|r)$  in the following.

For  $K \geq 1$ , we let  $\mathcal{M}_K$  denote the set made of the partitions of  $\llbracket 1; N \rrbracket$  into  $K$  segments. The cardinality of this set is  $\binom{N-1}{K-1}$ . More generally, we let  $\mathcal{M}_K(\llbracket s; t \rrbracket)$  denote the partitions of any interval  $\llbracket s; t \rrbracket$  into  $K$  segments. In order to get the marginal likelihood of  $y$  conditionally on  $K$ , one has to integrate out both  $m$  and  $\theta = \{\theta_r\}_{r \in m}$ :

$$p(y|K) = \sum_{m \in \mathcal{M}_K} p(m) \prod_{r \in m} p(y^r)$$

$$= \sum_{m \in \mathcal{M}_K} p(m) \prod_{r \in m} \int \left( \prod_{t \in r} p(y^t|\theta_r) \right) p(\theta_r) d\theta_r.$$

If the distribution of  $m$ , conditional on  $K$ , factorises over the segments with an expression of the form

$$p(m|K) = \frac{1}{C_K(a)} \prod_{r \in m} a_r, \quad (3.1)$$

where  $a_r$  are non-negative weights assigned to all segments and  $C_K(a) = \sum_{m \in \mathcal{M}_K} \prod_{r \in m} a_r$  is a normalising constant, these integrations can be performed separately. Rigaiill et al. (2012) introduced a matrix containing the weighted likelihood of all possible segments, whose general term is given by

$$A_{s,t} = \begin{cases} a_{\llbracket s;t \rrbracket} \cdot p(y^{\llbracket s;t \rrbracket}) & \text{if } 1 \leq s < t \leq N + 1, \\ 0 & \text{otherwise.} \end{cases} \quad (3.2)$$

This matrix can be used in an algorithm designed according to a dynamic programming principle to perform the integration on  $\mathcal{M}_K$  efficiently.

**Proposition 3.1** (Rigaill et al., 2012).

$$[A^K]_{s,t} = \sum_{m \in \mathcal{M}_K(\llbracket s;t \rrbracket)} \prod_{r \in m} a_r \cdot p(y^r)$$

where  $A^k$  denotes the  $k$ -th power of matrix  $A$  and  $[A^k]_{s,t}$  its general term. Moreover,

$$\mathcal{A}_K := \{[A^k]_{1,t}, [A^k]_{t,n+1}\}_{\substack{1 \leq k \leq K \\ 2 \leq t \leq N}}$$

can be computed in  $O(KN^2)$  time.

In particular,  $[A^K]_{1,n+1} = C_K(a) \cdot p(y|K)$ . Several quantities of interest share the same form: from  $\mathcal{A}_K$ , Rigaill et al. (2012) also derived exact formulas for the posterior probability of a change-point to occur at time  $t$  or for the posterior probability that a given segment  $r$  belongs to  $m$  (see Section 3.4.1). Classical Bayesian selection criteria for  $K$  are also given. One can notice that  $C_K(a)$  can be recovered by applying Proposition 3.1 not to matrix  $A$  but to a matrix defined similarly from  $a$ . For the uniform distribution on  $\mathcal{M}_K$ , *i.e.*  $a_r \equiv 1$ , we get  $C_K(a) = \binom{N-1}{K-1}$ .

Fearnhead (2006) worked under a slightly different model where  $m$  is not chosen conditionally on  $K$  but is instead drawn sequentially by specifying the probability mass function for the time between two successive change-points. They presented a filtering recursion to compute the marginal likelihood of the observations under their model where the integrations over parameters and segmentations are also uncoupled. Fearnhead & Liu (2007) showed that on-line and exact inference is also tractable in this model.

### 3.2.2 Tree-structured Graphical Models

In a multivariate setting, graphical models are used to describe complex dependence structures between the involved variables. A graphical model is given by a graph, either directed or not, and a family of distributions satisfying some Markov property with respect to this graph. We concentrate our attention on undirected graphical models, also called Markov random fields. We refer the reader to (Lauritzen, 1996) for a complete overview on the subject. Let  $V = \{1, \dots, p\}$  and  $Y = (Y_1, \dots, Y_p)$  be a random vector taking values in a product space  $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$ . We consider the set  $\mathcal{T}$  of connected undirected graphs with no cycles. These graphs are called spanning trees. For  $T \in \mathcal{T}$ , we let  $E_T$  denote the edges of  $T$ .

We consider a hierarchical model where one successively draws a tree  $T$  in  $\mathcal{T}$ , the parameters  $\theta$  of a distribution that factorises according to  $T$ , and finally a random vector  $Y$  according to this distribution. The marginal likelihood of the observations under this model, where both  $\theta$  and  $T$  are integrated out, is given by

$$p(y) = \sum_{T \in \mathcal{T}} p(T) \int p(y|T, \theta) p(\theta|T) d\theta.$$

These integrations can be performed exactly and efficiently by choosing the right priors on  $T$  and  $\theta$  (Meilä & Jaakkola, 2006; Schwaller et al., 2015). The distribution on trees is taken





to be factorised on the edges,

$$p(T) = \frac{1}{Z(b)} \prod_{\{i,j\} \in E_T} b_{i,j}, \quad (3.3)$$

where  $b_{i,j}$  are non-negative edge weights and

$$Z(b) := \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} b_{i,j} \quad (3.4)$$

is a normalizing constant. The prior on  $\theta$  has to be specified for all trees in  $\mathcal{T}$ . The idea is to require each of these priors to factorise on the edges and to specify a prior on  $\theta_{ij}$  once and for all trees,  $\theta_{ij}$  designating the parameters governing the marginal distribution of  $(Y_i, Y_j)$ . These priors must be chosen coherently, in the sense that, for all  $i, j, k \in V$ , the priors on  $\theta_{ik}$  and  $\theta_{jk}$  should induce the same prior on  $\theta_k$ . Some local Markov property is also needed. Schwaller et al. (2015) especially detailed three frameworks in which this can be achieved, namely multinomial distributions with Dirichlet priors, Gaussian distributions with normal-Wishart priors and copulas. We elaborate a little more on the particular case of Gaussian graphical models (GGMs). In a multivariate Gaussian setting,  $\theta = (\mu, \Lambda)$  where  $\mu$  and  $\Lambda$  respectively stand for the mean vector and precision matrix of the distribution. A classical result on GGMs states that if the  $(i, j)$ -th term of the precision matrix is equal to zero, there is no edge between nodes  $i$  and  $j$ . Thus, the support of  $p(\theta|T)$  is the set of sparse positive definite matrices whose non-zero terms are given by the adjacency matrix of  $T$ . The distribution of  $\theta|T$  can be defined for all trees at once by using a general normal-Wishart distribution defined on all positive-definite matrices (Schwaller et al., 2015, Sec. 4.1.3). Marginal distributions of this normal-Wishart distributions are used to build distributions for  $\{\theta|T\}_{T \in \mathcal{T}}$ .

When  $p(\theta|T)$  is carefully chosen, the integration on  $\theta$  can be performed independently from the integration on  $T$  and  $p(y|T)$  factorises on the edges of  $T$ :

$$p(y|T) = \prod_{i \in V} p(y_i) \prod_{\{i,j\} \in E_T} \frac{p(y_i, y_j)}{p(y_i)p(y_j)}$$

where

$$\begin{aligned} p(y_i, y_j) &= \int p(y_i, y_j | \theta_{ij}) d\theta_{ij}, \\ p(y_i) &= \int p(y_i | \theta_i) d\theta_i. \end{aligned} \quad (3.5)$$

Computing  $\{p(y|T)\}_{T \in \mathcal{T}}$  only requires  $p(p+1)/2$  computations of low-dimensional integrals, where  $p$  is the dimension of the model. As both  $p(T)$  and  $p(y|T)$  factorise on the edges, integrating the likelihood over  $T$  can be performed in  $O(p^3)$  time.

**Proposition 3.2.** *The marginal likelihood is given by*

$$p(y) = \frac{Z(\omega)}{Z(b)} \cdot \prod_{i \in V} p(y_i)$$

where  $Z(\cdot)$  is defined as in (3.4) and  $\omega$  is the posterior edge weight matrix whose general term is given by

$$\omega_{i,j} := b_{i,j} \frac{p(y_i, y_j)}{p(y_i)p(y_j)}. \quad (3.6)$$

Moreover,  $p(y)$  is obtained in  $O(p^3)$  time from  $b$  and  $\omega$ .

*Proof.*

$$\begin{aligned} p(y) &= \sum_{T \in \mathcal{T}} p(y|T)p(T) \\ &= \frac{1}{Z(b)} \left( \prod_{i \in V} p(y_i) \right) \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} b_{i,j} \frac{p(y_i, y_j)}{p(y_i)p(y_j)} \\ &= \frac{Z(\omega)}{Z(b)} \cdot \prod_{i \in V} p(y_i), \end{aligned}$$

with  $\omega$  as defined above. As  $Z(\cdot)$  can be computed in  $O(p^3)$  time using the Matrix-Tree theorem, we get the announced complexity.  $\square$

The posterior probability for an edge to belong to  $T$ ,  $P(\{i, j\} \in E_T | y)$ , can also be obtained for all edges at once in  $O(p^3)$  time (Schwaller et al., 2015, Th. 3).

### 3.3 Model & Properties

Sections 3.2.1 and 3.2.2 presented two models in which Bayesian inference requires us to integrate out a fundamentally discrete parameter (either the segmentation  $m$  or the spanning tree  $T$ ) and other (usually continuous) parameters  $\theta$ . In both situations, these integrations can be performed exactly and efficiently by uncoupling the problems. The integration over  $\theta$  is performed “locally” and the results are stored to be used in an algorithm that heavily relies on algebra to integrate over the discrete parameter. This is made possible by a careful choice of priors for both parameters. Our aim is to show that these algebraic tricks can be combined to perform exact Bayesian inference of multiple change-points in the dependence structure of multivariate time-series.

#### 3.3.1 Model

It is assumed that the observed data  $y = \{y^t\}_{t=1}^N$  are a realisation of a multivariate random process  $Y = \{Y^t\}_{t=1}^N$  of dimension  $p \geq 2$ . For  $1 \leq t \leq N$ ,  $Y^t = (Y_1^t, \dots, Y_p^t)$  is a multivariate random variable of dimension  $p$  taking values in a product space  $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$ . We model  $Y$  by a PPM where, at each time-point, observations  $Y^t$  are modelled by a tree-structured graphical model. If  $m$  is a segmentation with  $K$  segments, we let  $\mathbf{T} = \{T_k\}_{k=1}^K$  and  $\theta = \{\theta_r\}_{r \in m}$  respectively denote the tree structures and parameters for each segment. For  $r \in m$ , we also let  $\kappa(r|m)$  denote the position of  $r$  in  $m$ . We use index  $r$  rather than  $k$  as, in the following, all possible segments  $r$  will have to be considered *per se*, whatever the segmentation



they belong to. Our model can then be written as follows:

$$\begin{aligned}
 p(m|K) &= \frac{1}{C_K(a)} \prod_{r \in m} a_r, \\
 p(\mathbf{T}|K) &= \prod_{k=1}^K p(T_k) = \frac{1}{Z(b)^K} \prod_{k=1}^K \prod_{\{i,j\} \in E_{T_k}} b_{i,j}, \\
 p(\theta|m, \mathbf{T}) &= \prod_{r \in m} p(\theta_r | T_{\kappa(r|m)}), \\
 p(y|m, \theta, \mathbf{T}) &= \prod_{r \in m} \prod_{t \in r} p(y^t | T_{\kappa(r|m)}, \theta_r).
 \end{aligned}$$

For  $r \in m$ ,  $\{Y^t\}_{t \in r}$  are independent and identically distributed with structure  $T_{\kappa(r|m)}$  and parameters  $\theta_r$ . The priors on  $m$  and each of  $T_k$  are respectively taken of the form given in (3.1) and (3.3) through segment weights  $a$  and edge weights  $b$ . The distribution of  $\theta_r | \{T_{\kappa(r|m)} = T\}$  is assumed to factorise over the edges of  $T$ , coherently between all spanning trees  $T \in \mathcal{T}$ , as described in Section 3.2.2. A graphical representation of this model is given in Figure 3.2.

### 3.3.2 Factorisation Properties

In the model that we have described, the marginal likelihood of the observation, conditionally on  $K$ , is given by

$$p(y|K) = \sum_{m \in \mathcal{M}_K} \sum_{\mathbf{T} \in \mathcal{T}^K} \int p(y, m, \theta, \mathbf{T}|K) d\theta. \quad (3.7)$$

Integrating out the discrete parameters  $(m, \mathbf{T})$  requires to sum over a set of cardinality

$$|\mathcal{M}_K| \cdot |\mathcal{T}^K| = \binom{N-1}{K-1} \cdot p^{K(p-2)} \approx \left( \frac{Np^{p-2}}{K} \right)^K.$$

Nonetheless, the joint distribution of  $(m, \theta, \mathbf{T})$ , conditionally on  $K$ , factorises at different levels and integration can therefore be performed by combining the results given in Section 3.2.

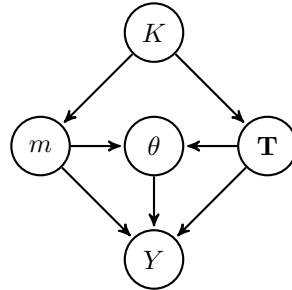


Figure 3.2 – Global graphical model.

**Proposition 3.3.** *The marginal likelihood  $p(y|K)$  can be computed in  $O(\max(K, p^3)N^2)$  time, where  $p$  and  $N$  respectively stand for the dimension of the model and the length of the series, from the posterior edge weight matrices computed on all possible segments  $r$ , whose general terms are given by*

$$\omega_{i,j}^{(r)} := b_{i,j} \frac{p(y_i^r, y_j^r)}{p(y_i^r)p(y_j^r)}. \quad (3.8)$$

$p(y_i^r, y_j^r)$  and  $p(y_i^r)$  are local integrals on  $\theta$  computed on edges and vertices as defined in (3.5).

*Proof.* For any segmentation  $m \in \mathcal{M}_K$  of  $\llbracket 1; N \rrbracket$  into  $K$  segments,  $\{(T_{\kappa(r|m)}, \theta_r)\}_{r \in m}$  are independent, so that  $p(y, m|K)$  can be written as

$$p(y, m|K) = \frac{1}{C_K(a)} \prod_{r \in m} a_r p(y^r),$$

where  $p(y^r)$  stands for the locally integrated likelihood of  $y^r$  on segment  $r$ ,

$$p(y^r) = \sum_{T \in \mathcal{T}} p(T) \int \left( \prod_{t \in r} p(y^t|T, \theta) \right) p(\theta|T) d\theta. \quad (3.9)$$

Thus,  $p(y, m|K)$  satisfies the factorability assumption required by Rigaiil et al. (2012) and once the weighted segment likelihood matrix  $A$ , defined by

$$A_{s,t} = \begin{cases} a_{\llbracket s;t \rrbracket} \cdot p(y^{\llbracket s;t \rrbracket}) & \text{if } 1 \leq s < t \leq n+1, \\ 0 & \text{otherwise,} \end{cases}$$

is computed, Proposition 3.1 can be used to gain access to  $p(y|K)$  in  $O(KN^2)$  time.

Computing matrix  $A$  requires to integrate the likelihood over tree structure  $T$  and parameters  $\theta$  for all possible segments  $r \subseteq \llbracket 1; N \rrbracket$ . On each segment, we fall back to the tree-structured model described in Section 3.2.2 and the integrated likelihood can be expressed using the local terms computed on vertices and edges that were defined in (3.5). Indeed, for  $r \subset \llbracket 1; N \rrbracket$ ,  $p(y^r)$  is obtained through Proposition 3.2 applied to  $\omega^{(r)}$  (defined in (3.8)):

$$p(y^r) = \frac{Z(\omega^{(r)})}{Z(b)} \cdot \prod_{i \in V} p(y_i^r).$$

where we remind that  $Z(\cdot)$  is the function giving the normalising constant of a tree distribution. As a consequence,  $A$  is computed in  $O(p^3 N^2)$  time from the posterior edge weight matrices  $\{\omega^{(r)}\}_{r \subseteq \llbracket 1; N \rrbracket}$ , hence the total complexity.  $\square$

The components of the matrices  $\omega^{(r)}$  result from the integration over  $\theta$ , which can be made separately and locally thanks to the assumptions made on its prior distribution in Section 3.3.1. This integration comes down to remove node  $\theta$  in the global graphical model displayed in Figure 3.2.

Marginal likelihood is only one of many quantities than might be of concern in this model. Yet, once matrix  $A$  has been calculated, other quantities of interest with respect to our model can be obtained at low cost. The next section provides a non-exhaustive list of such quantities.



## 3.4 Quantities of Interest

### 3.4.1 Change-point Location

For  $m \in \mathcal{M}_K$ , we let  $1 = \tau_0 < \tau_1 < \dots < \tau_K = N$  denote the change-points of  $m$  and, for  $1 \leq k \leq K$ , we let  $r_k = \llbracket \tau_{k-1}; \tau_k \rrbracket$  denote its  $k$ -th segment. In this section we are interested in computing the posterior probabilities of the following subsets of  $\mathcal{M}_K$ ,

$$\begin{aligned} \mathcal{B}_{K,k}(t) &:= \{m \in \mathcal{M}_K \mid \tau_k = t\}, \\ \mathcal{B}_K(t) &:= \bigcup_{k=1}^K \mathcal{B}_{K,k}(t), \\ \mathcal{S}_{K,k}(\llbracket s; t \rrbracket) &:= \{m \in \mathcal{M}_K \mid r_k = \llbracket s; t \rrbracket\} \\ \mathcal{S}_K(\llbracket s; t \rrbracket) &:= \bigcup_{k=1}^K \mathcal{S}_{K,k}(\llbracket s; t \rrbracket). \end{aligned}$$

Subsets  $\mathcal{B}_K(t)$  and  $\mathcal{S}_K(\llbracket s; t \rrbracket)$  are respectively the set of segmentations having a change-point at time  $t$  and the set of segmentations containing segment  $\llbracket s; t \rrbracket$ . We let  $B_{K,k}(t)$ ,  $B_K(t)$ ,  $S_{K,k}(\llbracket s; t \rrbracket)$  and  $S_K(\llbracket s; t \rrbracket)$  denote the respective posterior probabilities of these subsets.

Rigaiil et al. (2012) showed that, with the convention that  $[A^0]_{t_1, t_2} = 1$  for all  $1 \leq t_1 < t_2 \leq N + 1$ , these probabilities could be expressed as

$$\begin{aligned} B_{K,k}(t) &= \frac{[A^k]_{1,t} [A^{K-k}]_{t,N+1}}{[A^K]_{1,N+1}}, \\ B_K(t) &= \sum_{k=1}^{K-1} B_{K,k}(t), \\ S_{K,k}(\llbracket s; t \rrbracket) &= \frac{[A^{k-1}]_{1,s} A_{s,t} [A^{K-k}]_{t,N+1}}{[A^K]_{1,N+1}}, \\ S_K(\llbracket s; t \rrbracket) &= \sum_{k=1}^K S_{K,k}(\llbracket s; t \rrbracket). \end{aligned}$$

$\{B_{K,k}(t)\}_{t=1}^N$  provides the exact posterior distribution of the  $k$ -th change-point when  $m$  has  $K$  segments. Posterior segment probabilities  $\{S_K(\llbracket s; t \rrbracket)\}_{1 \leq s < t \leq N+1}$  will be useful in the following.

Once  $\{B_K(t)\}_{K \geq 2}$  is computed, the posterior probability  $B(t)$  of a change-point occurring at time  $t$  integrated on  $K$  is obtained as

$$B(t) = P(\cup_{K \geq 2} \mathcal{B}_K(t) \mid y) = \sum_{K \geq 2} p(K \mid y) B_K(t).$$

The computation of the posterior distribution on  $K$  is addressed below.

### 3.4.2 Number of Segments

The posterior distribution on  $K$  can also be derived from Proposition 3.1.

**Proposition 3.4.**

$$p(K|y) \propto \frac{p(K)[A^K]_{1,N+1}}{C_K(a)}.$$

*Proof.* Bayes' rule states that  $p(K|y) \propto p(K)p(y|K)$  and by Proposition 3.1,  $p(y|K) = [A^K]_{1,N+1}/C_K(a)$ .  $\square$

The best segmentation *a posteriori* can also be recovered efficiently by using matrix  $A$  in the Segment Neighbourhood Search algorithm (Auger & Lawrence, 1989). Thus, if one's interest lies in retrieving the number of segments  $K$ , two estimators can be considered

$$\begin{aligned} \hat{K}_1 &= \arg \max_K p(K|y), \\ \hat{K}_2 &= K(\arg \max_m p(m|y)). \end{aligned}$$

where  $K(m)$  stands for the number of segments in  $m$ .

### 3.4.3 Posterior Edge Probability

For any segment  $r \subseteq \llbracket 1; N \rrbracket$ , it is possible to compute the posterior edge probabilities corresponding to segment  $r$ :

$$P(\{i, j\} \in E_T | y^r), \quad \forall \{i, j\} \in \mathcal{P}_2(V),$$

where  $T$  is a random tree distributed as  $T_1, \dots, T_K$ . Whenever  $m$  is unknown, the segmentation can be integrated out to obtain instant posterior edge probabilities at any given time  $t$ . Conditionally on  $K$ , the instant posterior appearance probability of edge  $\{i, j\}$  at time  $t$  can be written as

$$\mathbf{p}_{ij}^K(t) := \sum_{m \in \mathcal{M}_K} p(m|y, K) P(\{i, j\} \in E_{T_{\kappa(t|m)}} | y, m),$$

where  $\kappa(t|m)$  gives the position of the segment containing  $t$  in  $m$ .

**Proposition 3.5.** *The instant posterior probability of edge  $\{i, j\}$  at time  $t$  is given by*

$$\mathbf{p}_{ij}^K(t) = \sum_{r \ni t} S_K(r) P(\{i, j\} \in E_T | y^r). \quad (3.10)$$

$\{\mathbf{p}_{ij}^K(t)\}_{\substack{1 \leq i, j \leq p \\ 1 \leq t \leq N}}$  can be computed in  $O(\max(K, p^3)N^2)$  time from  $A$  and  $\{\omega^{(r)}\}_{r \subseteq \llbracket 1; N \rrbracket}$ .

*Proof.* This formula is similar to the one giving the posterior mean of the signal in (Rigaille et al., 2012). If  $r \in m$  and  $t \in r$ , then  $P(\{i, j\} \in E_{T_{\kappa(t|m)}} | y, m) = P(\{i, j\} \in E_T | y^r)$ , hence the result.  $\{S_K(r)\}_{r \subseteq \llbracket 1; N \rrbracket}$  is obtained with complexity  $O(KN^2)$  and  $\{P(\{i, j\} \in E_T | y^r)\}_{r \subseteq \llbracket 1; N \rrbracket}$  with complexity  $O(p^3N^2)$ , and that gives an upper bound on total complexity.  $\square$

One could be interested in computing the posterior probability for an edge to keep the same status throughout time when  $m$  is integrated out, given  $K$ . Nonetheless, it would require to integrate on subsets of  $\mathcal{M}_K \otimes \mathcal{T}^K$  that are in direct contradiction with the factorability assumption, making the results that we have presented so far useless. Indeed, we would effectively be introducing dependency between segments, thus breaking up the factorability of  $p(y, m)$  with respect to  $r \in m$ . In this situation, Proposition 3.1 can no longer be used. A drastic workaround is to work under a fixed segmentation instead of integrating out  $m$ , and this is what we do in the following section.



## 3.5 Edge Status & Structure Comparisons

We now turn to the specific case where  $m$  is known and has  $K$  segments  $(r_1, \dots, r_K)$ . This situation is far less general than the framework we considered until now. Still, it corresponds to some practical situations where segment comparison is interesting and for which further exact inference can be carried out.

### 3.5.1 Edge Status Comparison

Let  $i, j$  be two distinct nodes in  $V$ . We are interested in computing the posterior probability of the subsets of  $\mathcal{T}^K$  defined by

$$\begin{aligned}\mathcal{E}_{ij}^+ &= \{\mathbf{T} = (T_1, \dots, T_K) \mid \forall k \in \llbracket 1, K \rrbracket, \{i, j\} \in E_{T_k}\}, \\ \mathcal{E}_{ij}^- &= \{\mathbf{T} = (T_1, \dots, T_K) \mid \forall k \in \llbracket 1, K \rrbracket, \{i, j\} \notin E_{T_k}\}, \\ \mathcal{E}_{ij} &= \mathcal{E}_{ij}^+ \cup \mathcal{E}_{ij}^-, \end{aligned}$$

that respectively correspond to the situations where edge  $\{i, j\}$  is always present, always absent, or has the same status in all trees. If  $\mathbf{T}$  belongs to  $\overline{\mathcal{E}}_{ij} = \mathcal{T}^K \setminus \mathcal{E}_{ij}$ , it means that there exists two segments in which  $\{i, j\}$  does not have the same status. We let  $(q_0^-, \bar{q}_0, q_0^+)$  respectively denote the prior probabilities of  $\mathcal{E}_{ij}^-, \overline{\mathcal{E}}_{ij}$  and  $\mathcal{E}_{ij}^+$ . These probabilities can be written as

$$\begin{aligned}q_0^- &= \prod_{k=1}^K P(\{i, j\} \notin E_{T_k}) = P(\{i, j\} \notin E_T)^K, \\ q_0^+ &= P(\{i, j\} \in E_T)^K, \quad \bar{q}_0 = 1 - q_0^- - q_0^+, \end{aligned}$$

where  $T$  is a tree distributed as  $T_1, \dots, T_K$ , and are obtained for all edges at once in  $O(p^3)$  time by computing the prior edge probability matrix  $(P(\{i, j\} \notin E_T))_{1 \leq i < j \leq p}$ .

Posterior probabilities  $(q^-, \bar{q}, q^+)$  for  $\mathcal{E}_{ij}^-, \overline{\mathcal{E}}_{ij}$  and  $\mathcal{E}_{ij}^+$  can be computed similarly but one posterior edge probability matrix has to be calculated per segment:

$$\begin{aligned}q^- &= \prod_{k=1}^K P(\{i, j\} \notin E_{T_k} \mid y^{r_k}), \\ q^+ &= \prod_{k=1}^K P(\{i, j\} \in E_{T_k} \mid y^{r_k}), \quad \bar{q} = 1 - q^- - q^+, \end{aligned}$$

However, if the prior distribution on trees is not strongly peaked, as events  $\mathcal{E}_{ij}^+$  and  $\mathcal{E}_{ij}^-$  only account for a relatively small number of tree series in  $\mathcal{T}^K$ ,  $q_0^-$  and  $q_0^+$  (as well as  $q^-$  and  $q^+$ ) will always be very small. To allow some control on the prior probabilities of these events, we use a random variable  $\epsilon_{ij}$  taking values  $\{-1; 0; 1\}$  with probabilities  $(\lambda^-, \bar{\lambda}, \lambda^+)$  and explicitly controlling the status of edge  $\{i, j\}$  in all trees:

$$p(\mathbf{T} \mid \epsilon_{ij}) = \begin{cases} p(\mathbf{T} \mid \mathcal{E}_{ij}^+) & \text{if } \epsilon_{ij} = 1, \\ p(\mathbf{T} \mid \overline{\mathcal{E}}_{ij}) & \text{if } \epsilon_{ij} = 0, \\ p(\mathbf{T} \mid \mathcal{E}_{ij}^-) & \text{if } \epsilon_{ij} = -1. \end{cases}$$

We obtain the model described in Figure 3.3, in which

$$p(y) = \lambda^+ p(y \mid \mathcal{E}_{ij}^+) + \lambda^- p(y \mid \mathcal{E}_{ij}^-) + \bar{\lambda} p(y \mid \overline{\mathcal{E}}_{ij}).$$

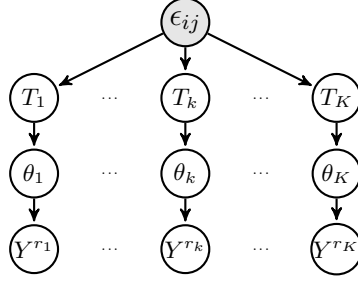


Figure 3.3 – Model for edge status comparison.

**Proposition 3.6.** *The vector of posterior probabilities for  $\epsilon_{ij}$  is proportional to  $(\lambda^- \frac{q^-}{q_0}, \bar{\lambda} \frac{\bar{q}}{q_0}, \lambda^+ \frac{q^+}{q_0})$ .*

*Proof.* We have that

$$\begin{aligned} p(\epsilon_{ij} = 1|y) &= \lambda^+ \frac{p(y|\mathcal{E}_{ij}^+)}{p(y)} \\ &= \frac{\lambda^+ p(y|\mathcal{E}_{ij}^+)}{\lambda^+ p(y|\mathcal{E}_{ij}^+) + \lambda^- p(y|\mathcal{E}_{ij}^-) + \bar{\lambda} p(y|\mathcal{E}_{ij}^-)} \\ &= \frac{\lambda^+ \frac{q^+}{q_0}}{\lambda^+ \frac{q^+}{q_0} + \lambda^- \frac{q^-}{q_0} + \bar{\lambda} \frac{\bar{q}}{q_0}}. \end{aligned}$$

We reason similarly with  $p(\epsilon_{ij} = -1|y)$  to get the result.  $\square$

### 3.5.2 Structure Comparison

The same reasoning can be applied for the global event

$$\mathcal{E} = \{\mathbf{T} = (T_1, \dots, T_K) | \exists T \in \mathcal{T}, \forall k \in \llbracket 1, K \rrbracket, T_k = T\},$$

which corresponds to a constant dependency structure across all segments (we remind that, in this section, the segments are known *a priori*), with possible changes for the parameters. The prior probability of  $\mathcal{E}$  is given by

$$q_0 := P(\mathcal{E}) = \frac{1}{Z(b)^K} \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} b_{i,j}^K = \frac{Z(b^{\odot K})}{Z(b)^K},$$

where  $b^{\odot K}$  stands for the element-wise  $K$ -th power of matrix  $b$ . On each segment  $r_k$ , the posterior distribution on trees factorises as

$$p(T_k | y^{r_k}) = \frac{1}{Z(\omega^{(k)})} \prod_{\{i,j\} \in T_k} \omega_{i,j}^{(k)},$$

and the posterior probability of  $\mathcal{E}$  is therefore given by

$$q := \sum_{T \in \mathcal{T}} \prod_{k=1}^K p(T | y^{r_k}) = \frac{Z(\bigodot_k \omega^{(k)})}{\prod_k Z(\omega^{(k)})}$$





where  $\odot$  denotes the element-wise matrix product.

Just as in the edge status comparison, we let a binary variable  $\epsilon \sim \mathcal{B}(\pi)$  control the prior probability of  $\mathcal{E}$ , with  $p(\mathbf{T}|\epsilon = 1) = p(\mathbf{T}|\mathcal{E})$ , and derive a similar formula for the posterior distribution of  $\epsilon$ .

**Proposition 3.7.**  $\epsilon|y \sim \mathcal{B}(\pi^*)$  with  $\pi^* := \frac{\pi \frac{q}{q_0}}{\pi \frac{q}{q_0} + (1-\pi) \frac{1-q}{1-q_0}}$ .

*Proof.* Similar to Proposition 3.6. □

## 3.6 Simulations

Our approach was especially concerned with explicitly modelling the structure of the graphical model within each segment, but a simpler model could be considered in which the structure remains implicit. In a Gaussian setting, that would mean that the precision matrix governing the distribution on a given segment would be drawn without any zero-term constraints. One goal of this simulation study is to show how both models (with and without structure constraints) comparatively behave when one is interested in retrieving the number of segments or the location of the change-points.

Another concern addressed by these simulations is the cost of the tree assumption when the true model is not tree-structured. How well can the number of segments, the change-points or even the structures be recovered when the true networks are not trees?

### 3.6.1 Simulation Scheme

For this study, we generated time-series of size  $N = 70, 140$  and  $210$ . We choose segmentations with four segments of lengths  $\frac{3}{7}N, \frac{1}{7}N, \frac{2}{7}N$  and  $\frac{1}{7}N$  such that the relative length of each segment is kept identical through all sample sizes. The number of variables was fixed to  $p = 10$ . To give an idea of the sizes of the discrete sets we are working with, for  $N = 210$ , the cardinalities of the segmentation and tree sets are respectively  $|\mathcal{M}_4| \approx 1.5 \cdot 10^6$  and  $|\mathcal{T}| = 10^8$ , so the size of the space to be explored is  $\approx 1.5 \cdot 10^{38}$ . We built three structure scenarios by sampling structures from the uniform distribution on spanning trees, or from an Erdős-Rényi random graph distribution with connection probability  $p_C = 2/p$  or  $4/p$ . Thus, for each scenario, we got a series  $\{\Delta_r\}_{r \in M_N}$  of adjacency matrices describing the structure of the graphical model on all segments. The observations on a segment  $r$  were then drawn according to a multivariate Gaussian distribution with mean vector zero and precision matrix  $\Lambda_r$  equal to the Laplacian matrix of  $\Delta_r$  augmented of 1 on the diagonal, rescaled so that each variable has unit variance. For each sample size and structure series, 100 datasets were generated.

As described in the introduction of this section, the inference was then performed in the two following models. The first one is the full precision matrix model, without any structure constraint, and is given by

$$\begin{aligned} \{\Lambda_r\}_{r \in m} & \text{ i.i.d.}, & \Lambda_r & \sim \mathcal{W}(\alpha, \Psi), \\ \{Y_t\}_{t=1}^N & \text{ independent}, & Y^t & \sim \mathcal{N}(\mathbf{0}_p, \Lambda_r^{-1}), \quad \forall t \in r. \end{aligned} \tag{3.11}$$

where  $\mathcal{W}(\alpha, \Psi)$  stands for the Wishart distribution with  $\alpha$  degrees of freedom and scale matrix  $\Psi$ . The second one is the corresponding model with tree-structure assumption, as

described in Section 3.3.1, and given by

$$\begin{aligned}
\{T_k\}_{k=1}^K &\text{ i.i.d.}, & T_k &\sim \mathcal{U}(\mathcal{T}), \\
\{\Lambda_r\}_{r \in m} &\text{ independent}, & \Lambda_r &\sim h\mathcal{W}(\alpha, \Psi, T_{\kappa(r|m)}), \\
\{Y_t\}_{t=1}^N &\text{ independent}, & Y^t &\sim \mathcal{N}(\mathbf{0}_p, \Lambda_r^{-1}), \quad \forall t \in r,
\end{aligned} \tag{3.12}$$

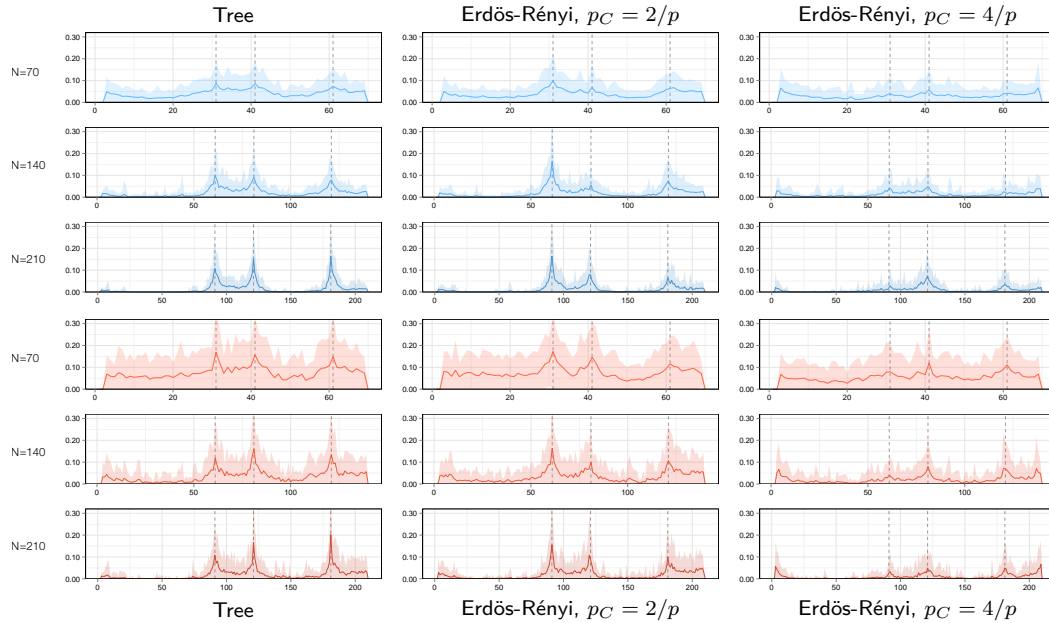
where we let  $h\mathcal{W}(\alpha, \Psi, T)$  denote the hyper-Wishart distribution based on  $\mathcal{W}(\alpha, \Psi)$  and with structure  $T$  (Schwaller et al., 2015). In both cases, we set  $\alpha = p + 10$  and  $\Psi = (\alpha - p - 1) \cdot \mathbf{I}_p$ , where  $\mathbf{I}_p$  stands for the identity matrix of size  $p$ . The distribution of  $m|K$  is set to the uniform on  $\mathcal{M}_K$  and  $K$  follows a Poisson distribution with parameter  $\gamma = 4$ , truncated to  $\llbracket 1; 10 \rrbracket$ .

We emphasize the fact that, when the tree-structured model is considered, the series of precision matrices  $\{\Lambda_r\}_{r \in m}$  used to generate the data only belongs to the support of the law in the first structure scenario. The graphs drawn from the Erdős-Rényi distributions are not trees and therefore cannot induce precision matrices in the support of a tree-structured hyper-Wishart distribution. On the contrary, the full model obviously allows such precision matrices.

Finally, for the sake of clarity, we limited our study to centered data and null mean models, but one could allow the mean to vary between segments by using a (hyper) normal-Wishart distribution for  $(\mu_r, \Lambda_r)$ , where  $\mu_r$  stands for the mean on segment  $r$ .

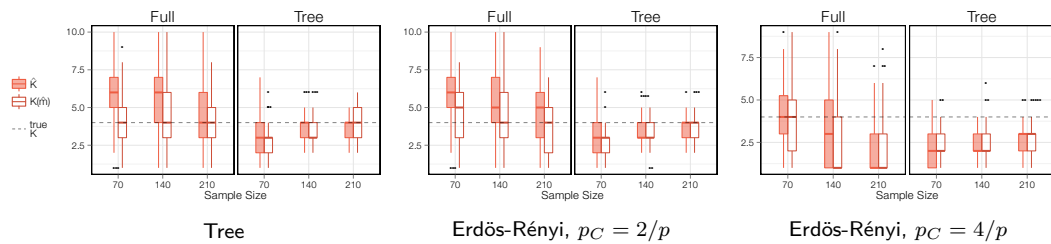
### 3.6.2 Results

We plotted the posterior probability of a change-point intervening at time  $t$ , integrated over  $K$ , as a function of  $t$  in the tree-structured and full models (Figure 3.4). In both cases,

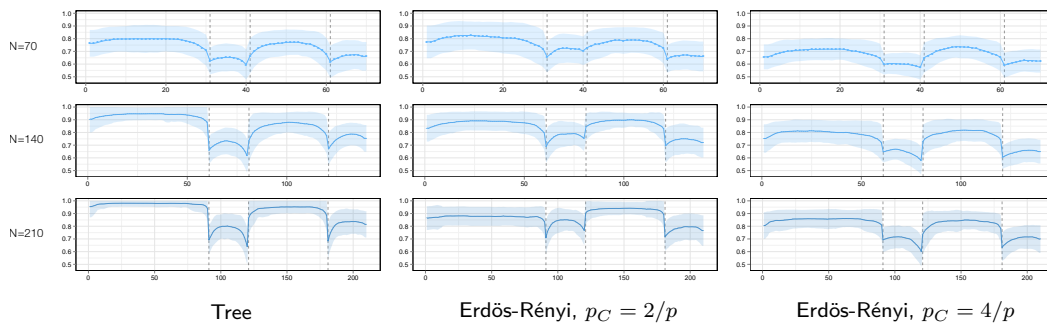


**Figure 3.4** – Posterior probability of observing a change-point for the tree-structured model (blue) and for the full model (red). The curve represents the mean value obtained from the 100 samples and the ribbon gives the standard deviation.

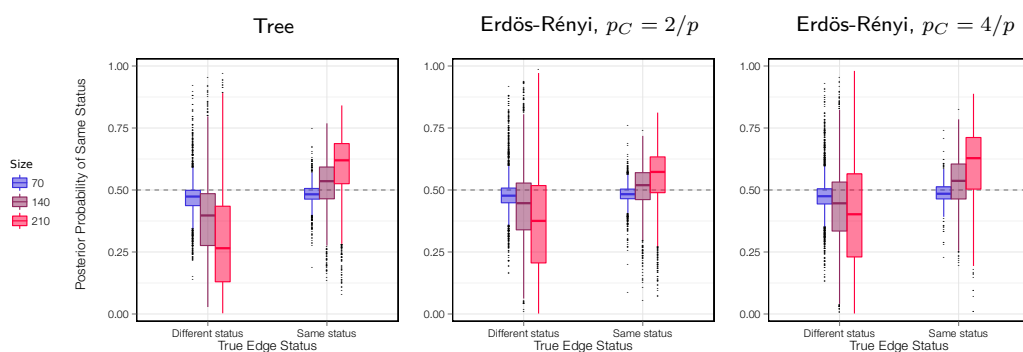




**Figure 3.5** – Boxplot of  $\hat{K} = \arg \max_K p(K|y)$  and  $K(\hat{m}) = K(\arg \max_m p(m|y))$  against sample size  $N$  for the full model (Full) and the tree-structured model (Tree).



**Figure 3.6** – Area under the ROC curve computed for the posterior edge probability matrix  $\left[ \mathbf{p}_{ij}^K(t) \right]_{i,j=1}^P$  with respect to the true adjacency matrix at time  $t$ . We set  $K$  to the true number of segments ( $K = 4$ ). The curve represents the mean value obtained from the 100 samples and the ribbon gives the standard deviation.



**Figure 3.7** – Boxplot of the posterior probability for an edge to have the same status throughout the time-series. Edges were separated according to their true status (either identical in all graphs or not). Each boxplot aggregates the results for all edges with a given status and all datasets.

change-points are hardly retrieved in the high-density Erdős-Rényi scenario, the inference performing better in the other two low-density scenarios. The standard deviations across samples are lower for the tree-structured model than for the full model. We can also observe a smoother behaviour with respect to time in the tree-structured model. Results on simulations with a greater number of segments (not shown) confirmed these observations. As expected, the shortest segments are hardly detected when the length of the series is small. These results seem to show that, when one is interested in retrieved change-point locations, the tree-structured model that we have presented can be considered in non-tree scenarios without any meaningful drop in performances.

**NUMBER OF SEGMENTS** For each sample, we computed  $\hat{K} = \arg \max_K p(K|y)$  and  $K(\hat{m}) = K(\arg \max_m p(m|y))$ . The results are given in Figure 3.5. In the full model, the number of segments selected by  $\hat{K}$  and  $K(\hat{m})$  varies a lot across samples and is usually higher than in the tree model. In the tree-structured model, both  $\hat{K}$  and  $K(\hat{m})$  tend to slightly underestimate the number of segments, especially in the highly-connected Erdős-Rényi scenario. They also display a more stable behaviour in the tree model. On small samples,  $K(\hat{m})$  seems to achieve better stability.

**POSTERIOR EDGE PROBABILITY** For  $t \in \llbracket 1; N \rrbracket$ , we computed the posterior edge probability matrix defined in (3.10) for  $K = 4$ . Figure 3.6 shows the area under the ROC curve of this matrix against the true adjacency matrix at time  $t$ . In all scenarios, the structure is better retrieved on long segments. A drop in the accuracy is systematically observed near true change-points. While presenting lower accuracy compared to the other two scenarios, the structure inference in the highly connected scenario still provides meaningful results.

**EDGE STATUS COMPARISON** The posterior probability for an edge to keep the same status throughout time was computed for all edges as explained in Section 3.5. We set the prior probability to change status at  $\bar{\lambda} = 0.5$  and the prior probabilities to be always present or absent to  $\lambda^+ = \lambda^- = 0.25$ . We expected edges changing status during the time-series to be given low posterior probabilities. For small samples and across all scenarios, the posterior probability to have the same status remains close to the prior probability 0.5 for all edges. When samples grow bigger, a small contrast sets up according to the edges effectively changing status or not. We nonetheless observe a large variability across samples and edges, that could be explained by the fact that some configurations are harder to detect than others. An edge only present on a small segment might for instance be considered absent through the whole series.

## 3.7 Applications

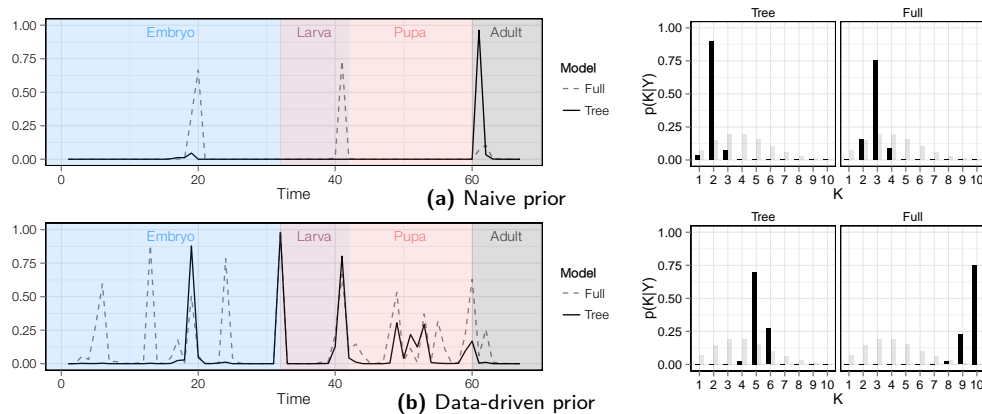
### 3.7.1 *Drosophila* Life Cycle Microarray Data

The life-cycle of *Drosophila melanogaster* is punctuated by four main stages of morphological development: embryo, larva, pupa and adult. The expression levels of 4028 genes of wild-type *Drosophila* were measured by Arbeitman et al. (2002) at 67 time-points throughout their life-cycle. We have here restricted our attention to eleven genes involved in wing muscle development and previously studied by Zhao et al. (2006) and Dondelinger et al. (2013). The expectation was that our approach would find change-points corresponding to the four different stages of development observed for *Drosophila melanogaster*.

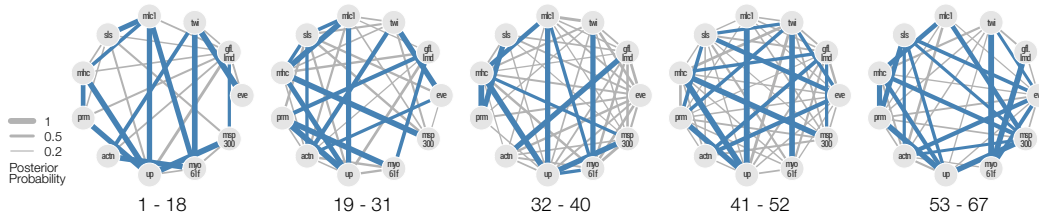


We used the normal-Wishart version of the model described in the simulation study. When using the naive prior parameters given in Section 3.6, we obtained poor results (Figure 3.8.a), probably because of the small number of time-points. We noticed that the results could be improved by using data-driven prior specification. We centered the data and set the prior scale matrix  $\phi$  of the normal-Wishart distribution with  $\alpha = p + 10$  degrees of freedom to  $\phi = (\alpha - p - 1) \cdot \Sigma_y$  where  $\Sigma_y$  stands for the sample covariance matrix. By doing this, the normal-Wishart distribution that we get has expectation  $(\mathbf{0}_p, \Sigma_y)$ . We then obtained the results given in Figure 3.8.b. For this prior, we looked closer to the results for  $\hat{K} = \arg \max_K p(K|y) = 5$  segments, *i.e.* one more than the number of development stages. The best segmentation  $\hat{m}_5$  with 5 segments has change-points at positions (19, 32, 41, 53). The posterior probability of observing a change-point at these locations is quite high (Figure 3.8.b). The larva stage is almost exactly recovered, with a shift of one position for the end of the segment. The embryo stage is divided into two segments and the separation between pupa and adult states is missed, the last segment including both adulthood and part of the pupa stage. These results are nonetheless encouraging. For each segment  $r$  of  $\hat{m}_5$ , we computed the posterior edge probability matrix given by  $(P(\{i, j\} \in E_T | y^r))_{1 \leq i, j \leq p}$ . On each segment, the prior probability for an edge to appear was set to 0.5 with an approach similar to what was done in Section 3.5. We give a graphical representation of the results in Figure 3.9. In the first segment, fewer edges have large posterior probabilities. However, this higher contrast in probabilities might just be a consequence of this segment being larger than the others.

Finally we compared our results with those obtained by [Dondelinger et al. \(2013\)](#) on the same dataset. As for the probability of change-point along time, the results we give in Figure 3.8.b are very similar to those displayed in Figure 12 of this reference. The comparison in terms of inferred networks is more complex as the networks they displayed correspond to the expected stages (embryo, larva, pupa and adult) and not to the one they actually inferred. We found good concordances between the network they inferred for the embryo stage and those that we obtained on segments [1-18] and [19-31] (both in the embryo stage). We also found similarities at the larva stage (which is close to our inferred [32-40] segment).



**Figure 3.8** – Posterior probability of a change-point occurring at time  $t$  as a function of time integrated on  $K$  (left) and posterior distribution for  $K$  (right) for the full (Full) and tree-structured (Tree) models.



**Figure 3.9** – Graphical representation of posterior edge probability matrix for each segment of the best segmentation with 5 segments. The width of an edge is proportional to its posterior probability. Edges with probability higher than 0.5 are coloured in blue. Edges with probability lower than 0.2 were not represented.

### 3.7.2 Functional MRI Data

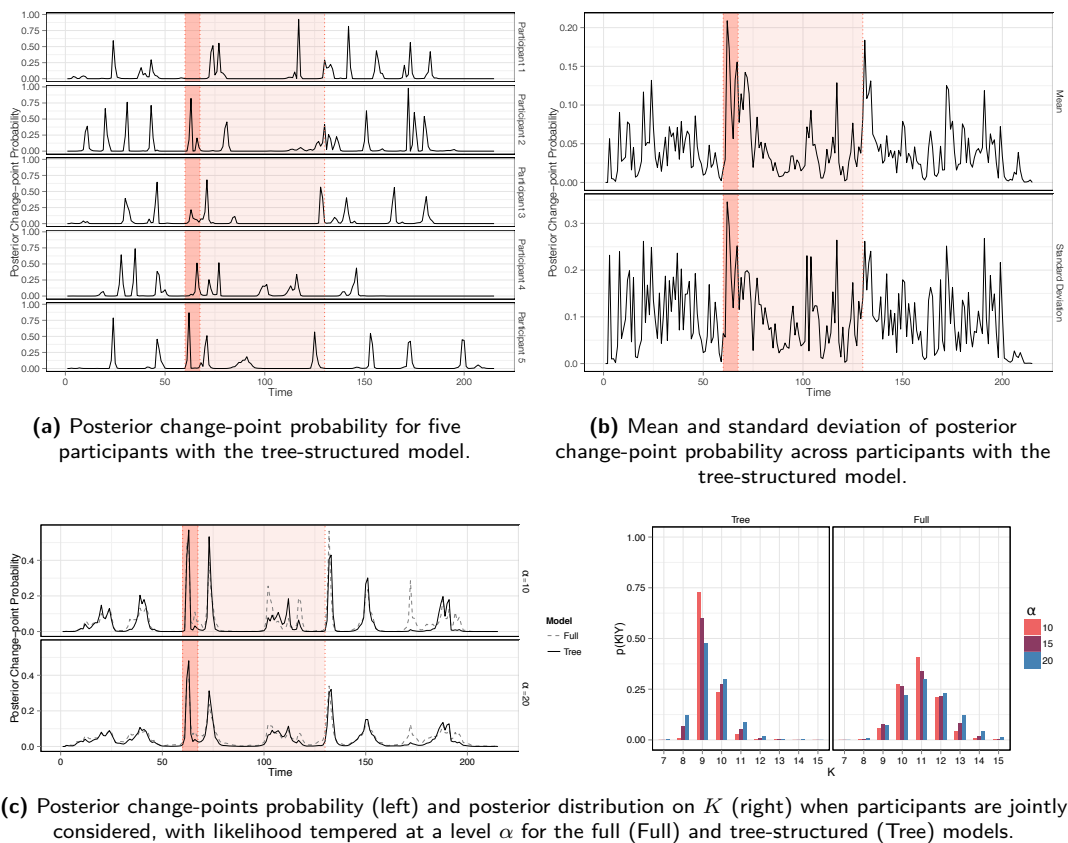
Functional magnetic resonance imaging (fMRI) is commonly used in neuroscience to study the neural basis of perception, cognition, and emotion by detecting changes associated with blood flow. This second application focuses on fMRI data collected by [Cribben et al. \(2012\)](#). We give a brief description of the experiment but we refer the reader to their article for a more detailed description. Twenty participants were submitted to an anxiety-inducing experiment. Before scanning, participants were told that they would have two minutes to prepare a speech on a subject given to them during scanning. Afterwards, they would have to give their speech in front of expert judges, but they had a “small chance” not to be selected. The subject of the speech was given after two minutes of recording. After two minutes of preparation, participants were told that they would not have to give the speech. The recording continued for two minutes afterwards. A series of 215 images at two-second intervals were acquired during the experiment. [Cribben et al. \(2012\)](#) preprocessed the data and determined five regions of interest (ROIs) in the brain on which the signals were averaged. Thus, we have  $p = 5$  and  $N = 215$ , for  $U = 20$  participants. We standardised the data across all participants.

Each participant can be analysed individually by using the same approach as in the previous application. To analyse all participants together, we make the assumption that the dependence structure between the different ROIs of the brain is the same across participants, while being allowed to vary throughout time. Nonetheless, on a given temporal segment, therefore for a given structure, parameters are independently drawn for each participant, so that the likelihood on a segment  $r$  can be written as

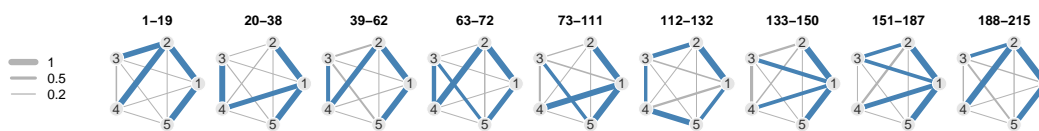
$$p(y^r) = \sum_{T \in \mathcal{T}} \prod_{u=1}^U \left[ \int \prod_{t \in r} p(y^{t,u} | \theta_u) p(\theta_u | T) d\theta_u \right] \quad (3.13)$$

where  $y^{t,u}$  stands for the vector of observations at time  $t$  for participant  $u$ . The distribution  $p(\theta_u | T)$  and  $p(y^{t,u} | \theta_u)$  are respectively taken to be normal-Wishart and Gaussian distributions, as in the individual model. In practice, when we tried to perform the inference of the joint model, we were faced with numerical issues, occurring at different levels. The summation over trees was problematic for some segments, especially the largest one. Indeed, we are summing very small quantities and the product over participants in  $p(y|T)$  brings us to deal with quantities of the order of machine precision. Moreover, while searching for the best segmentation can be achieved through  $\log(A) = [\log(A_{s,t})]_{1 \leq s, t \leq N+1}$ , integrating over segmentations requires the actual computation of matrix  $A$ . Thus, the exponentiation of the segment log-likelihood matrix leads to other numerical issues.





**Figure 3.10** – Change-point location for the fMRI data. During the dark red interval, the subject of the speech was revealed to participants, who prepared their speech during the light red interval. This preparation is ended by a statement saying that they would not have to give the speech.



**Figure 3.11** – Graphical representation of posterior edge probability matrix for each segment of the best segmentation with  $K = 9$  segments on fMRI data with non-tempered likelihood. The width of an edge is proportional to its posterior probability. Edges with probability higher than 0.5 are coloured in blue.

Our pragmatic answer to these issues was to consider a tempered version of the likelihood given in (3.13):

$$p_\alpha^*(y^r) = \sum_{T \in \mathcal{T}} \prod_{u=1}^U \left[ \int \prod_{t \in r} p(y^{t,u} | \theta_u) p(\theta_u | T) d\theta_u \right]^{1/\alpha}, \quad (3.14)$$

with  $\alpha > 1$ . Tempering the likelihood does not change the mode of the posterior distribution on  $m$ , if the matrix  $a$  giving prior segment weights is tempered similarly. By doing this, we are actually reducing the effective sample size: a very big  $\alpha$  would yield a posterior distribution on  $m$  close to the prior.

Figures 3.10a and 3.10b sum up the results obtained participant per participant for change-point location. They vary a lot across participants, as shown by the five given examples, as well as the mean and standard deviation curves. The left panel of Figure 3.10c shows the posterior probability of observing a change-point when participants are jointly considered with a tree-structured model or with a non-structured model, with likelihood tempered at  $\alpha = U/2 = 10$  and  $\alpha = U = 20$ . For both values of  $\alpha$ , the profiles are quite similar, with an expected more peaked behaviour for  $\alpha = 10$ . The strongest peak is observed during the announcement of the speech topic. The right panel of Figure 3.10c gives the posterior distribution of  $K$  for both models and for different values of  $\alpha$ . We observe flatter distributions for the full model, with a mode at 11 segments. In the tree-structured model, 9 segments are selected. For this value of  $K$ , we looked at the best segmentation and computed the posterior edge probability matrices for its segments. A graphical representation of the results is given in Figure 3.11. Cribben et al. (2012) retrieved 8 segments from these data. There is no clear correspondence between our segmentation and theirs. A remark that can nonetheless be made is that, in our case, each change-point is associated with a clear change in the topology of the network. These structure changes are less obvious in (Cribben et al., 2012). This might be a consequence of our model explicitly modelling the structure, thus encouraging change-points to mark out abrupt changes in structure rather than in parameters.

## 3.8 Discussion

In this paper, we showed how exact Bayesian inference could be achieved for change-points in the structure of a multivariate time-series with careful specification of prior distributions. Essentially, prior distributions have to factorise over both segments and edges. For the sake of clarity, we assumed that, within a segment  $r$ , observations  $Y^t$  were independent conditionally on  $T$  and  $\theta$ . While convenient and leading to comfortable formulas, this independence assumption is hardly realistic in many applied situations, including those that we have considered here. Yet, time dependency could be considered within segments, as long as  $p(y^r | T)$  still factorises over the edges of  $T$ . One could for instance consider using the work of Siracusa (2009) to achieve this. Trees would then be used to model the dependences between two consecutive times instead of instantaneous dependences.

The framework that we have described is also convenient for Bayesian model comparison. When one is faced with an alternative in modelling, Bayes factors between two models are easily obtained, as fully marginal likelihood can be computed exactly and efficiently. For instance, the question of whether changes should be allowed in the mean of a Gaussian distribution or not can be addressed by computing  $p(y)$  in both cases and by looking at





their ratio. This is by no mean specific to our approach, but exact computation makes it completely straightforward.

The exactness of the inference also creates a comfortable framework to precisely study the effect of the prior distribution on segmentations. Once again, as the inference does not rely on stochastic integration, the impact of prior specification could be evaluated at low cost and in an exact manner.

We finish this discussion by mentioning numerical issues. As explained in Section 3.7, when the number of observations increases, we have to deal with elementary probabilities that differ from several order of magnitudes. Because the summations over the huge spaces of both segmentations and trees are carried out in an exact manner, these quantities have to be added to each others, resulting in numerical errors. Obviously, no naive implementation would work and some of these errors can be avoided with careful and skilful programming. At this stage, this is still not sufficient and the likelihood tempering approach that we propose is not satisfying. Further numerical improvements could be considered such as the systematic ordering of the terms when computing a determinant in a recursive way.

The R code used in the simulations and the applications is available from the authors upon request. A package will soon be available from the Comprehensive R Archive Network.

## Acknowledgements .....

We thank Ivor Cribben (Alberta School of Business, Canada) for kindly providing the fMRI data. We also thank Sarah Ouadah (AgroParisTech, INRA, Paris, France) for fruitful discussions.

# 4

## Extensions

---

<b>4.1</b>	<b>Covariates</b>	<b>83</b>
4.1.1	Multivariate Multiple Linear Regression	83
4.1.2	Copulas: a Pragmatical Approach	88
<b>4.2</b>	<b>Temporal Dependence</b>	<b>90</b>
4.2.1	Temporal Interaction Models	92
4.2.2	Bayesian Inference of Tree-structured TIM	93
<b>4.3</b>	<b>Prior Distribution on Segmentations</b>	<b>96</b>
4.3.1	A Different Kind of Prior	97
4.3.2	Transferring Knowledge from One Dataset to Another	100

---

In the previous chapters, we used distinctive features pertaining to the sets of spanning trees and segmentations, and leading to tractable inference in various models. A blatant similarity between these sets is that they are both made of elements that are intrinsically modular. A spanning tree is a collection of edges, just as a segmentation is collection of segments. Modularity is actually one of the key features allowing for an efficient browse through these large collections. The algorithms used to sum over spanning trees and segmentations assume that the function to be summed up factorises according to these basic units, and express the global constraints on the collections of units in a smart algebraic fashion. A spanning tree on  $p$  vertices is a collection of  $p - 1$  edges, with the additional constraint that the resulting graph is connected. This connectivity assumption can be encoded as a condition on the incidence matrix corresponding to the collection of edges, and through this property, the summation can be performed by computing a single determinant of size  $p - 1$ . Similarly, in the algorithm developed by Rigail et al. (2012) for segmentations, a crucial ingredient is that the matrix giving the weights of all segments is a strictly upper triangular matrix. The  $K$ -th power of this matrix is used to sum over  $\mathcal{M}_K$  and all irrelevant terms in the summation are set to zero thanks to this property.



All of this is easy enough to notice when we are presented with both examples, but is not quite helpful in finding other situations for which such algebraic tricks can be conjured. Nonetheless, it can further be noticed that maximisation and summation problems are closely linked in this matter, in the sense that they are often jointly tractable or not. Indeed, the maximum (or minimum) spanning tree problem can be solved through Kruskal's or Prim's algorithms, and classic dynamic programming can be used to obtain the best segmentation. This has to do with the fact that max-sum and sum-product problems can often be seen as two particular instances of a more generic problem expressed with abstract operators. This link could be useful to dig out other interesting situations in which exact inference is tractable.

As our goal was to perform Bayesian inference on models involving spanning trees and segmentations, we had to consider distributions on these sets. Their modularity obviously lead to the exponential family with a binary vector indexed by all possible units (either edges or segments) as sufficient statistics. The normalising constant, or partition functions, of these distributions could then be computed through the aforementioned algorithms, as the factorisation assumptions are obviously satisfied. Of course, this is what motivated our considering these particular sets in the first place. For many distributions in the exponential family, computing the normalising constant is not possible in a reasonable amount of time. There is for instance no explicit form for this constant in the case of decomposable graphs.

Models were then built so as to preserve this factorisation property for posterior distributions. If  $Q$  either denotes a tree or a segmentation, the distribution  $\pi$  of the observations conditionally on  $Q$  was assumed to factorise over the elements of  $Q$ . So was the prior distribution on  $\pi$  (or on its parameters  $\theta$ ). Under these assumptions, expressed through (hyper) Markov properties, the integration over  $\pi$  or  $\theta$  could be performed locally and separately from the summation on  $Q$ . We merely assumed that these low-dimensional integrals could be computed through conjugacy and explicit formulas, for the sake of exact inference. If one has to stray from conjugate priors, numerical or stochastic integration can be considered, as well as approximation techniques such as Laplace's approximation (see Schwaller & Robin, 2015).

Among the various assumptions that we have made in the models, some were fundamental for the proper functioning of the approach, and some were merely convenient, leaving room for some interesting tweaks. This chapter is meant to provide some extensions and perspectives to the work presented before.

The first section explains how covariates can be taken into account in what has been previously presented. When observations are assumed to be normally distributed, adapting the model is straightforward. We show that conjugate priors for the multivariate multiple linear regression induce hyperdistributions that can be used as a basis to build a compatible family of hyperdistributions. We also derive an expression for the marginal likelihood of the observations on any subset of vertices. Whenever observations cannot be modelled by a multivariate normal distribution, we suggested a heuristic approach based on copulas, that we illustrate on microbial ecology data.

In the second section, we consider a different set of graphs for which exact structure inference is possible, namely the set of directed trees. This set inherits many of the interesting properties of its undirected counter-part, as all directed trees can be obtained from an undirected tree by choosing a root. We use these graphs to introduce temporal

dependence between observations. The independence assumption is indeed hardly realistic in many applied situations. Our suggestion is to use the Temporal Independence Models (TIMs) introduced by [Siracusa \(2009\)](#), for which structure inference can also be performed with cubic complexity with respect to the number of vertices, as long as possible structures are restricted to directed trees. More than introducing temporal dependence, we actually exhibit a slightly different example of discrete set on which summation is possible through algebra.

Finally, we investigate prior specification on segmentations. We make out a non-exhaustive bestiary of priors fitting in our framework. We also describe a different kind of prior in which taking into account an initial guess on the segmentation, or information coming from a previous study, is straightforward. These prior distributions still factorise over segments, but the contribution of a segment to a given segmentation depends on its position within the segmentation. Previously, the position of a segment had no influence on its contribution. The algorithm used to perform the inference has to be adapted to this new kind of prior distribution.

## 4.1 Covariates

In some situations, observations of a set of variables come with observations of associated variables that, while not being of direct interest, have to be taken into account in the analysis. The latter are often called covariates. We were drawn to such problems through a collaboration with Corinne Vacher (UMR Biogeco, Biodiversity, Genes & Communities, Université de Bordeaux) on microbial ecology data.

Plan-inhabiting micro-organisms can interact with each other forming complex interaction networks. These interactions can either be direct (predation, parasitism, mutualism, *etc*) or mediated by the environment. The first kind of interactions are especially interesting when trying to understand how a new microbial species might integrate in an existing ecosystem. The microbial ecologists we worked with were interested in better understanding powdery mildew, one of the most common tree diseases in European forests caused by a fungus called *Erysiphe albitoides*. To that end, they measured the abundances of the microbial species that could be found on 120 leaves taken from three different oak trees. For each leaf, many covariates were also measured, including the distance of the leaf to the base of the branch, to the trunk or to the ground. Our goal was to infer the network giving the direct interactions between species while taking these covariates into account.

We show that, for multivariate normal distributions, including covariates in the framework developed in Chapters 2 and 3 is straightforward. Nonetheless, this is basically made possible by specific properties of normal distributions. In the general case, the hyper Markov property for the prior on the distribution of the observations is hard to retain while including covariates. For the count data described above, we used an approach based on copulas as a pragmatical solution. The results were given in an article published in the journal *Microbial Ecology* ([Jakuschkin et al., 2016](#)).

### 4.1.1 Multivariate Multiple Linear Regression

We consider a setting where we have  $N$  observations of  $p$  real-valued variables of interest and  $k$  real-valued covariates. These observations are denoted by  $(\mathbf{Y}^{(n)}, \mathbf{X}^{(n)}) \in \mathbf{R}^p \times \mathbf{R}^k$ ,



$1 \leq n \leq N$  and gathered in matrices  $\mathbf{Y}$  and  $\mathbf{X}$ :

$$\mathbf{Y} = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(N)} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(n)} \end{pmatrix}.$$

For  $U \subset V$ , we let  $\mathbf{Y}_U$  and  $\mathbf{B}_U$  respectively denote matrices  $\mathbf{Y}$  and  $\mathbf{B}$  restricted to the columns corresponding to the variables in  $U$ . Our variables of interest are now the univariate series given by column vectors  $\mathbf{Y}_i$ ,  $i \in V$ , for which we only have one observation, along with the associated covariates. The sample space we are working on is

$$\mathcal{X} = \bigotimes_{i \in V} \mathcal{X}_i = \bigotimes_{i \in V} \mathbf{R}^N.$$

A distribution  $\pi$  on  $\mathcal{X}$  can be seen as a distribution on  $\mathbf{R}^{Np}$ , which is obviously isomorphic to  $\mathcal{X}$ . Similarly, if  $\pi$  is a distribution on  $\mathcal{X}$  and  $A$  is a subset of  $V$  of size  $a$ , the marginal distribution  $\pi_A$  is basically a distribution on  $\mathbf{R}^{ap}$ .

Let us consider the regression model given by

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \quad \text{vec}(\mathbf{E}) \sim \mathcal{N}(0, \Lambda^{-1} \otimes I_N),$$

where “vec” is the linear transformation stacking the columns of a matrix on top of one another, and  $\otimes$  stands for the Kronecker product of two matrices. Matrix  $\mathbf{B}$  is a  $k \times p$  real-valued matrix giving the regression coefficients of the model. Matrix  $\Lambda$  is the precision matrix of size  $p \times p$  describing the interactions between the different series. This model can be rewritten as

$$\text{vec}(\mathbf{Y}) | \mathbf{X}, \mathbf{B}, \Lambda \sim \mathcal{N}(\text{vec}(\mathbf{X}\mathbf{B}), \Lambda^{-1} \otimes I_N). \quad (4.1)$$

We remind  $\text{vec}(\mathbf{X}\mathbf{B}) = (I_p \otimes \mathbf{X})\text{vec}(\mathbf{B})$ .

Any prior distribution on  $(\mathbf{B}, \Lambda)$  induces a hyperdistribution  $\rho$  on the distribution  $\pi(\cdot | \mathbf{X})$  of  $\mathbf{Y}$ . Classic results on conjugate priors for regression models suggest this choice of prior distributions:

$$\begin{aligned} \Lambda &\sim \mathcal{W}(\alpha, \Psi), \\ \text{vec}(\mathbf{B}) | \Lambda &\sim \mathcal{N}(\text{vec}(\mathbf{B}_0), \Lambda^{-1} \otimes K_0^{-1}). \end{aligned} \quad (4.2)$$

where we remind that  $\mathcal{W}(\alpha, \Psi)$  is the Wishart distribution with  $\alpha$  degrees of freedom and parametric matrix  $\Psi$ . In the following, we use the shorthands

$$\begin{aligned} \beta &= \text{vec}(\mathbf{B}), & \beta_A &= \text{vec}(\mathbf{B}_A), & \forall A \subseteq V, \\ \beta^0 &= \text{vec}(\mathbf{B}_0), & \beta_A^0 &= \text{vec}((\mathbf{B}_0)_A), & \forall A \subseteq V, \\ \epsilon &= \text{vec}(\mathbf{E}). \end{aligned}$$

#### 4.1.1.a Hyper Markov Property

We now show that any distribution for  $(\mathbf{B}, \Lambda)$  of the form given in (4.2) induces a hyperdistribution that is strong hyper Markov with respect to the complete graph, and can therefore be used to build a compatible family of strong hyper Markov hyperdistributions.

**Proposition 4.1.** Let  $R \subset V$  and  $S = V \setminus R$ , of respective sizes  $r$  and  $s = p - r$ . Then

$$\text{vec}(\mathbf{Y}_R) | \mathbf{X}, \beta, \Lambda \sim \mathcal{N}((I_r \otimes \mathbf{X})\beta_R, (\Lambda_{R \bullet S})^{-1} \otimes I_N), \quad (4.3)$$

$$\text{vec}(\mathbf{Y}_S) | \mathbf{Y}_R = \mathbf{y}_r, \mathbf{X}, \beta, \Lambda \sim \mathcal{N}(\mu_{S \bullet R}, \Lambda_{SS}^{-1} \otimes I_N), \quad (4.4)$$

where  $[\Lambda_{RR}, \Lambda_{SR}, \Lambda_{RS}, \Lambda_{SS}]$  is a partitioning of  $\Lambda$  according to  $(R, S)$  and where

$$\Lambda_{R \bullet S} = \Lambda_{RR} - \Lambda_{RS} \Lambda_{SS}^{-1} \Lambda_{SR},$$

$$\mu_{S \bullet R} = (I_s \otimes \mathbf{X})\beta_S - (\Lambda_{SS}^{-1} \Lambda_{SR} \otimes I_N) (\text{vec}(\mathbf{y}_R) - (I_r \otimes \mathbf{X})\beta_R).$$

*Proof.* Let  $R \subset V$  and  $S = V \setminus R$  of size  $r$  and  $s = p - r$  respectively. Without loss of generality, we suppose that  $\beta$  and  $\Lambda$  can be written as

$$\beta = \begin{pmatrix} \beta_R \\ \beta_S \end{pmatrix}; \quad \Lambda = \begin{pmatrix} \Lambda_{RR} & \Lambda_{RS} \\ \Lambda_{SR} & \Lambda_{SS} \end{pmatrix}.$$

The mean vector of  $\text{vec}(\mathbf{Y})$  can be expressed using  $\beta_R$  and  $\beta_S$ :

$$\text{vec}(\mathbf{X}\mathbf{B}) = (I_p \otimes \mathbf{X})\beta = \begin{pmatrix} I_r \otimes \mathbf{X} & 0 \\ 0 & I_s \otimes \mathbf{X} \end{pmatrix} \begin{pmatrix} \beta_R \\ \beta_S \end{pmatrix} = \begin{pmatrix} (I_r \otimes \mathbf{X})\beta_R \\ (I_s \otimes \mathbf{X})\beta_S \end{pmatrix}.$$

If  $\Sigma := \Lambda^{-1}$  and considering a partitioning of  $\Sigma$  similar to  $\Lambda$ , the covariance and precision matrices of  $\text{vec}(\mathbf{Y})$  are respectively given by

$$\begin{aligned} (\Lambda^{-1} \otimes I_N)^{-1} &= \Lambda \otimes I_N = \begin{pmatrix} \Lambda_{RR} \otimes I_N & \Lambda_{RS} \otimes I_N \\ \Lambda_{SR} \otimes I_N & \Lambda_{SS} \otimes I_N \end{pmatrix}; \\ \Lambda^{-1} \otimes I_N &= \Sigma \otimes I_N = \begin{pmatrix} \Sigma_{RR} \otimes I_N & \Sigma_{RS} \otimes I_N \\ \Sigma_{SR} \otimes I_N & \Sigma_{SS} \otimes I_N \end{pmatrix}. \end{aligned}$$

and  $\Sigma_{RR}^{-1} = \Lambda_{R \bullet S}$ . Classic results on the marginal and conditional distributions of a multivariate normal distribution conclude this proof.  $\square$

**Proposition 4.2.** Let  $R \subset V$  and  $S = V \setminus R$ . Then

$$(\beta_R, \Lambda_{R \bullet S}) \perp\!\!\!\perp (\beta_S + (\Lambda_{SS}^{-1} \Lambda_{SR} \otimes I_k) \beta_R, \Lambda_{SS}, \Lambda_{RS}). \quad (4.5)$$

*Proof.* The probability density function of  $(\beta, \Lambda)$  is given by

$$f(\beta, \Lambda) \propto |\Lambda|^{(\alpha - p + k - 1)/2} \exp\left(-\frac{1}{2}(\beta - \beta^0)^\top (\Lambda \otimes K_0)(\beta - \beta^0)\right) \exp\left(-\frac{1}{2}\text{tr}(\Psi\Lambda)\right). \quad (4.6)$$

Let us define

$$\beta_{S \bullet R} := \beta_S + (\Lambda_{SS}^{-1} \Lambda_{SR} \otimes I_k) \beta_R.$$

$\beta_{S \bullet R}^0$  is defined similarly. We want to make the following change of variables in (4.6):

$$(\beta_R, \beta_S, \Lambda_{RR}, \Lambda_{SS}, \Lambda_{RS}) \mapsto (\beta_R, \beta_{S \bullet R}, \Lambda_{R \bullet S}, \Lambda_{SS}, \Lambda_{RS})$$

The Jacobian matrix of this change of variables can be written as

$$\begin{pmatrix} \beta_R & \beta_{S \bullet R} & (\Lambda_{R \bullet S}, \Lambda_{SS}, \Lambda_{RS}) \\ \beta_R & I_{rk} & \Lambda_{SS}^{-1} \Lambda_{SR} \otimes I_k & 0 \\ \beta_S & 0 & I_{sk} & 0 \\ (\Lambda_{RR}, \Lambda_{SS}, \Lambda_{RS}) & 0 & * & Z \end{pmatrix} = J. \quad (4.7)$$

So  $|J| = |Z|$  and Massam & Neher (1997) showed that  $|Z| = 1$ . So we only need to express (4.6) as a function of  $(\beta_R, \beta_{S\bullet R}, \Lambda_{R\bullet S}, \Lambda_{SS}, \Lambda_{RS})$ .

$$\begin{aligned} f(\beta, \Lambda) &\propto |\Lambda_{R\bullet S}|^{(\alpha-p+k-1)/2} \exp\left(-\frac{1}{2}(\beta_R - \beta_R^0)^\top (\Lambda_{R\bullet S} \otimes K_0)(\beta_R - \beta_R^0)\right) \exp\left(-\frac{1}{2}\text{tr}(\Psi_{RR}\Lambda_{R\bullet S})\right) \\ &\quad \cdot |\Lambda_{SS}|^{(\alpha-p+k)/2} \exp\left(-\frac{1}{2}(\beta_{S\bullet R} - \beta_{S\bullet R}^0)^\top (\Lambda_{SS} \otimes K_0)(\beta_{S\bullet R} - \beta_{S\bullet R}^0)\right) \\ &\quad \cdot \exp\left(-\frac{1}{2}(\text{tr}(\Psi_{RR}\Lambda_{RS}\Lambda_{SS}^{-1}\Lambda_{SR}) + 2\text{tr}(\Psi_{SR}\Lambda_{RS}) + \text{tr}(\Psi_{SS}\Lambda_{SS}))\right) \end{aligned} \quad (4.8)$$

Hence the independence of  $(\beta_R, \Lambda_{R\bullet S})$  and  $(\beta_{S\bullet R}, \Lambda_{SS}, \Lambda_{RS})$ . We also get the distribution of  $(\beta_R, \Lambda_{R\bullet S})$ :

$$\begin{aligned} \Lambda_{R\bullet S} &\sim \mathcal{W}(\alpha + k - r, \Psi_{RR}), \\ \beta|\Lambda_{R\bullet S} &\sim \mathcal{N}(\beta_R^0, \Lambda_{R\bullet S}^{-1} \otimes K_0^{-1}). \end{aligned} \quad (4.9)$$

□

**Proposition 4.3.** *The hyper distribution  $\rho$  on  $\pi(\cdot|\mathbf{X})$  induced by the joint distribution of  $(\Lambda, \beta)$  is strong hyper Markov with respect to the complete graph on  $V$ .*

*Proof.* We have to show that, for  $R \subseteq V$  and  $S = V \setminus R$ , it holds that

$$\pi_R(\cdot|\mathbf{X}) \perp\!\!\!\perp \pi_{S|R}(\cdot|\mathbf{X}). \quad (4.10)$$

These distributions are given by Proposition 4.1 and it is enough to notice that

$$\begin{aligned} \mu_{S\bullet R} &= (I_S \otimes \mathbf{X})(\beta_S + (\Lambda_{SS}^{-1}\Lambda_{SR} \otimes I_k)\beta_R) - (\Lambda_{SS}^{-1}\Lambda_{SR} \otimes I_k)\text{vec}(\mathbf{y}_R) \\ &= (I_S \otimes \mathbf{X})\beta_{S\bullet R} - (\Lambda_{SS}^{-1}\Lambda_{SR} \otimes I_k)\text{vec}(\mathbf{y}_R) \end{aligned}$$

to get (4.10) from Proposition 4.2. □

We can therefore put Proposition 1.9 to good use and build a compatible family of strong hyper Markov hyperdistributions from the hyperdistribution  $\rho$  induced by the distribution on  $(\mathbf{B}, \Lambda)$  given in (4.2).

#### 4.1.1.b Marginal Likelihood

Now that we have our compatible family of strong hyper Markov hyperdistributions, we have to compute the marginal likelihood of the observations on any subset of  $V$ . For  $A \subseteq V$ , it can be written as

$$p(\mathbf{Y}_A|\mathbf{X}) = \int \pi_A(\mathbf{Y}_A|\mathbf{X})\rho_A(\pi_A)d\nu_A(\pi_A).$$

We are in fact mainly interested in subsets of size 1 and 2 but we derive the formula for any subset.

**Proposition 4.4.** *The marginal likelihood of  $\mathbf{Y}$  is given by*

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{\pi^{\frac{pN}{2}}} \frac{\Gamma_p\left(\frac{\alpha+N}{2}\right)}{\Gamma_p\left(\frac{\alpha}{2}\right)} \frac{|K_0|^{\frac{p}{2}}}{|K_N|^{\frac{p}{2}}} \frac{|\Psi|^{\frac{\alpha}{2}}}{|\Psi^{(N)}|^{\frac{\alpha+N}{2}}}$$

where

$$\begin{aligned}\Psi^{(N)} &:= \Psi + \mathbf{B}_0^\top K_0 \mathbf{B}_0 + \mathbf{B}_N^\top K_N \mathbf{B}_N + \mathbf{Y}^\top \mathbf{Y}, \\ K_N &:= K_0 + \mathbf{X}^\top \mathbf{X}, \\ \mathbf{B}_N &:= K_N^{-1} (\mathbf{X}^\top \mathbf{Y} + K_0 \mathbf{B}_0).\end{aligned}\tag{4.11}$$

*Proof.* We have that

$$\begin{aligned}p(\mathbf{Y}, \Lambda, \mathbf{B} | \mathbf{X}) &= \frac{|\Lambda|^{\frac{\alpha-p+k+N-1}{2}} |K_0|^{\frac{p}{2}} |\Psi|^{\frac{\alpha}{2}}}{2^{\frac{\alpha p}{2}} (2\pi)^{\frac{p(N+k)}{2}} \Gamma_p\left(\frac{\alpha}{2}\right)} \\ &\cdot \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda [(\mathbf{Y} - \mathbf{X}\mathbf{B})^\top (\mathbf{Y} - \mathbf{X}\mathbf{B}) + (\mathbf{B} - \mathbf{B}_0)^\top K_0 (\mathbf{B} - \mathbf{B}_0) + \Psi]\right)\right).\end{aligned}$$

If  $K_N$  and  $\mathbf{B}_N$  are defined as in 4.11, then

$$\begin{aligned}(\mathbf{Y} - \mathbf{X}\mathbf{B})^\top (\mathbf{Y} - \mathbf{X}\mathbf{B}) + (\mathbf{B} - \mathbf{B}_0)^\top K_0 (\mathbf{B} - \mathbf{B}_0) \\ = (\mathbf{B} - \mathbf{B}_N)^\top K_N (\mathbf{B} - \mathbf{B}_N) + \mathbf{B}_N^\top K_N \mathbf{B}_N + \mathbf{B}_0^\top K_0 \mathbf{B}_0 + \mathbf{Y}^\top \mathbf{Y}.\end{aligned}$$

In the right-hand side of the equation above, only the first term depends on  $\mathbf{B}$  and the corresponding term in  $p(\mathbf{Y}, \Lambda, \mathbf{B} | \mathbf{X})$  is a Gaussian kernel. We therefore get that

$$\begin{aligned}p(\mathbf{Y}, \Lambda | \mathbf{X}) &= \frac{|\Lambda|^{\frac{\alpha-p+N-1}{2}} |\Psi|^{\frac{\alpha}{2}} |K_0|^{\frac{p}{2}}}{2^{\frac{\alpha p}{2}} (2\pi)^{\frac{pN}{2}} \Gamma_p\left(\frac{\alpha}{2}\right) |K_N|^{\frac{p}{2}}} \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda [\mathbf{B}_N^\top K_N \mathbf{B}_N + \mathbf{B}_0^\top K_0 \mathbf{B}_0 + \mathbf{Y}^\top \mathbf{Y} + \Psi]\right)\right) \\ &= \frac{|\Lambda|^{\frac{\alpha-p+N-1}{2}} |\Psi|^{\frac{\alpha}{2}} |K_0|^{\frac{p}{2}}}{2^{\frac{\alpha p}{2}} (2\pi)^{\frac{pN}{2}} \Gamma_p\left(\frac{\alpha}{2}\right) |K_N|^{\frac{p}{2}}} \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda \Psi^{(N)}\right)\right).\end{aligned}$$

In the equation above, we recognise that

$$|\Lambda|^{\frac{\alpha-p+N-1}{2}} \exp\left(-\frac{1}{2} \text{tr}\left(\Lambda \Psi^{(N)}\right)\right)$$

is the unnormalised density of a Wishart distribution with  $\alpha + N$  degrees of freedom and parametric matrix  $\Psi^{(N)}$ , hence the sought-after result.  $\square$

**Proposition 4.5.** For  $A \subseteq V$  of size  $a$ , the marginal likelihood of  $\mathbf{Y}_A$  is given by

$$p(\mathbf{Y}_A | \mathbf{X}) = \frac{1}{\pi^{\frac{aN}{2}}} \frac{\Gamma_a\left(\frac{\alpha-p+a+N}{2}\right)}{\Gamma_a\left(\frac{\alpha-p+a}{2}\right)} \frac{|K_0|^{\frac{a}{2}} |\Psi_{AA}|^{\frac{\alpha-p+a}{2}}}{|K_N|^{\frac{a}{2}} |\Psi_{AA}^{(N)}|^{\frac{\alpha-p+a+N}{2}}}.$$

*Proof.* It follows from (4.9) and Proposition 4.4.  $\square$

We remind that network inference in tree-structured graphical models relies on a posterior edge weight matrix given by

$$\omega_{i,j} = \beta_{i,j} \frac{p(\mathbf{Y}_i, \mathbf{Y}_j)}{p(\mathbf{Y}_i)p(\mathbf{Y}_j)}, \quad \forall \{i, j\} \in \mathcal{P}_2(V),$$





where  $\beta$  is a prior edge weight matrix. From Proposition 4.5, we get that

$$\frac{p(\mathbf{Y}_i, \mathbf{Y}_j | \mathbf{X})}{p(\mathbf{Y}_i | \mathbf{X})p(\mathbf{Y}_j | \mathbf{X})} \propto \frac{|\Psi_{i,i}^{(N)}|^{\frac{\alpha-p+1+N}{2}}}{|\Psi_{i,i}^{(N)}|^{\frac{\alpha-p+1}{2}}} \frac{|\Psi_{j,j}^{(N)}|^{\frac{\alpha-p+1+N}{2}}}{|\Psi_{j,j}^{(N)}|^{\frac{\alpha-p+1}{2}}} \frac{|\Psi_{\{i,j\}\{i,j\}}|^{\frac{\alpha-p+2}{2}}}{|\Psi_{\{i,j\}\{i,j\}}^{(N)}|^{\frac{\alpha-p+2+N}{2}}}.$$

For network inference, knowing  $p(\mathbf{Y}_i, \mathbf{Y}_j | \mathbf{X})/p(\mathbf{Y}_i | \mathbf{X})p(\mathbf{Y}_j | \mathbf{X})$  up to a constant identical for all edges is enough. For segmentation however, the remaining terms have to be included as they depend on the number of observations and the covariates, which vary depending on the segment that is considered.

## 4.1.2 Copulas: a Pragmatical Approach

We now come back to the reason we considered covariates in the first place, that is to say the microbial ecology data gathered on oak trees that we described at the beginning of this chapter. These data were obtained by high-throughput sequencing of the DNA that could be found on the collected leaves. For each of the  $N$  leaves, the abundance of a microbial species was given by a number of reads. We are therefore dealing with multivariate count data. If we consider that all covariates are real-valued for the sake of simplicity, we are now in a situation where  $(Y^{(n)}, X^{(n)}) \in \mathbf{N}^p \times \mathbf{R}^k$ ,  $1 \leq n \leq N$ .  $\mathbf{Y}$  and  $\mathbf{X}$  are defined as in the previous section.

### 4.1.2.a A Model for Multivariate Count Data based on Copulas

Our first idea to model multivariate count data with covariates was to rely on univariate generalized linear models (GLMs) weaved together with a copula. We wanted to remain as close as possible to the Gaussian case, and we hoped to get similar results when using Gaussian copulas. But we will see that the strong hyper Markov property that we showed for multivariate multiple regression does not hold in this particular model.

For  $i \in V$ , we model the abundance of species  $i$  on all leaves conditionally on  $\mathbf{X}$  as follows:

$$\begin{cases} \mathbf{Y}_i^{(n)} & \sim \mathcal{P}(\lambda_i^n), \\ \log(\lambda_i^n) & = \mathbf{X}^{(n)} \mathbf{B}_i, \end{cases} \quad \{\mathbf{Y}_i^{(n)}\}_{1 \leq n \leq N} \text{ independent} \mid \mathbf{X}. \quad (4.12)$$

We could also have modelled  $\mathbf{Y}_i | \mathbf{X}$  with a negative binomial GLM. The dependence between the different  $\mathbf{Y}_i$  is then handled by a Gaussian copula. Let  $\Phi^{-1}$  denote the inverse cumulative density of the univariate standard normal distribution, and  $\psi_p^\Lambda$  the probability density function of the multivariate normal distribution with mean vector zero and partial correlation matrix  $\Lambda$ . The probability density function of the Gaussian copula  $C_p^\Lambda$  with partial correlation matrix  $\Lambda$  is given by

$$c_p^\Lambda(u) = \psi_p^\Lambda(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_p)), \quad \forall u \in [0; 1]^p.$$

If we let  $\mathfrak{P}(\cdot | \lambda)$  denote the cdf of the Poisson distribution with parameter  $\lambda$ , the distribution of  $\mathbf{Y}$  can be written as

$$p(\mathbf{y} | \Lambda, \mathbf{B}, \mathbf{X}) = \prod_{n=1}^N c_p^\Lambda \left( \mathfrak{P}(y_1^{(n)} | \lambda_1^n), \dots, \mathfrak{P}(y_p^{(n)} | \lambda_p^n) \right). \quad (4.13)$$

The distributions of the form given in (4.13) do not form a strong meta Markov model with respect to the complete graph on  $V$ . We will therefore not be able to build a compatible

family of strong hyper Markov hyperdistributions for this model. Indeed, let  $A \subseteq V$  and  $B = V \setminus A$ . Evaluating  $p(\mathbf{y}_B | \mathbf{y}_A, \Lambda, \mathbf{B}, \mathbf{X})$  requires more information about what happens on  $A$  than just  $\mathbf{y}_A$ . We actually need to know  $\mathfrak{P}(\cdot | \lambda_i^n)$ ,  $i \in A$ ,  $1 \leq n \leq N$ . This means that we cannot mimic what we have done in the multiple linear regression framework.

#### 4.1.2.b Pragmatical Approach

All our attempts to model multivariate count data while taking covariates into account ended up leading to models that were not strong meta Markov. We therefore decided to use the copula-based model described above, while dropping the Bayesian paradigm for the univariate part of the model involving GLMs. In so doing, we betrayed our committed stance on fully exact Bayesian inference, but this moral lapse brought us back to the simple framework of tree-structured copulas described in Section 2.4.1.a.

The approach splits up in two steps. The first one deals with covariates by estimating the regression coefficients in the univariate GLMs described in (4.12). For all  $i \in V$ , we obtain an estimation  $\hat{\mathbf{B}}_i$  of  $\mathbf{B}_i$ . The effect of covariates is then eliminated by computing the Pearson residuals associated to each fitted model. For Poisson GLMs, they are given by

$$R_i^{(n)} = \frac{\mathbf{Y}_i^{(n)} - \hat{\lambda}_i^{(n)}}{\sqrt{\hat{\lambda}_i^{(n)}}}, \quad \hat{\lambda}_i^{(n)} = \exp(\mathbf{X}^{(n)} \hat{\mathbf{B}}_i), \quad \forall i \in V, \forall n \in \llbracket 1; N \rrbracket.$$

We let  $\hat{F}_i$  denote the empirical cdf of  $\{R_i^{(n)}\}_{n=1}^N$  and we define

$$\tilde{\mathbf{Y}}_i^{(n)} := \hat{F}_i(R_i^{(n)}), \quad \forall i \in V, \forall n \in \llbracket 1; N \rrbracket.$$

Having (somehow) dealt with covariates in the first step, we come back to Bayesian inference. We consider that  $\{\tilde{\mathbf{Y}}^{(n)}\}_{n=1}^N$  is an independent sample drawn from a distribution on  $[0; 1]^p$  with uniform marginal distributions. The second step is a direct application of Section 2.4.1.a. The model that we used is the following:

$$\begin{aligned} T &\sim \xi, \\ \Lambda_{i,j} | T &\sim \begin{cases} \delta_1 & \text{if } i = j, \\ \mathcal{U}([-1; 1]) & \text{if } \{i, j\} \in E_T, \\ \delta_0 & \text{otherwise,} \end{cases} \quad \{\Lambda_{i,j}\}_{1 \leq i, j \leq p} \text{ independent } | T, \\ \tilde{\mathbf{Y}} | \Lambda &\sim C_p^\Lambda. \end{aligned} \tag{4.14}$$

In this model, network inference can be performed as described in Chapter 2. The results obtained for  $\tilde{\mathbf{Y}}$  are interpreted as a proxy for the desired results on  $\mathbf{Y} | \mathbf{X}$ .

The main flaw of this pragmatical approach is that no uncertainty is taken into account at the regression level, as residuals of the GLMs are plugged in the tree-copula model through their empirical cdfs. We are trudging troubled waters mixing both frequentist and Bayesian paradigms. But what is lost in clarity and exactness is partially made up for in flexibility. Indeed, both the univariate GLMs and the copula can be tailored to the specific needs pertaining to a given modelling problem. If need be, each GLM can be individually adapted, and so can the bivariate copulas describing the pairwise interactions in  $\tilde{\mathbf{Y}}$ .



### 4.1.2.c Results for Microbial Ecology Data

We briefly present the results obtained on the microbial ecology data through the copula approach described above. Data were collected from 120 leaves taken from three oak trees presenting different observed susceptibility to powdery mildew. DNA was extracted from these leaves and sequenced. After standard preprocessing, for each leaf, we were given the number of reads per species or, more accurately, operational taxonomic unit (OTU). We were also given the distance of the leaf to the base of the branch, to the trunk and to the ground, as well as the orientation (either south-west or north-east) of the branch.

A subset of the most abundant OTUs was used for the analysis. These OTUs represented at least 0.5% of the sequences for at least one of the three oak trees. We retained 114 OTUs, among which 48 were bacterial and 66 were fungal (including the pathogen responsible for powdery mildew, *Erysiphe alphitoides*). In most cases, the number of reads for each OTU was fitted with a negative binomial GLM with environmental variables as predictors and the total number of reads per sample (log-transformed) as an offset. Poisson GLMs were used for a few OTUs. The analysis was performed for each tree separately, as well as for all trees together with an additional covariate indicating the origin of the leaf.

The network inference model was chosen as presented in (4.14), with  $\xi$  put to the uniform distribution on  $\mathcal{T}$ . Posterior edge probabilities were computed and the prior appearance probability of each edge was set to 0.5 as described in Section 2.4.3. For each edge, the sign of the interaction was determined by looking at the maximum a posteriori estimate for the partial correlation of the corresponding bivariate copula.

Figure 4.1 shows the posterior probability matrices obtained when all trees are jointly considered and for each individual tree. The former is more contrasted as a result of the grouped analysis working with more data. The block-structure that can more or less be seen on all four matrices indicates that intra-kingdom interactions (fungus to fungus or bacteria to bacteria) are more frequent than inter-kingdom interactions. Of particular interest were the interactions of *Erysiphe alphitoides* with other OTUs on the susceptible tree, as the main goal of the study was to get a better understanding of the microbial mechanisms behind powdery mildew. Truncating the posterior edge probability matrix at the level set for prior appearance probability, namely 0.5, we found that *E. alphitoides* was involved in 26 interactions, among which 13 with fungal OTUs (6 positive, 7 negative) and 13 with bacterial OTUs (6 positive, 7 negative). These interactions are represented in Figure 4.2. All but one of these interactions were previously considered likely by microbial ecologists.

## 4.2 Temporal Dependence

In the models described in Chapters 2 and 3, observations are assumed to be independent given the parameters. As discussed in Section 3.8, this assumption is hardly realistic in many practical cases. Let us take the example of the fMRI data treated in Section 3.7.2. Such data are typically recorded with high temporal resolution. Here, an image was taken every two seconds, but it is not unusual to sample at higher frequencies. With observations this close to one another in time, it would be desirable for the state of the brain of a participant at time  $t$  to directly influence on the state of its brain at time  $t+1$ . In order to be entirely satisfactory, the dependence model should preserve some factorisation property with respect to edges, so that we can keep using the Matrix-Tree theorem to efficiently integrate on structures.

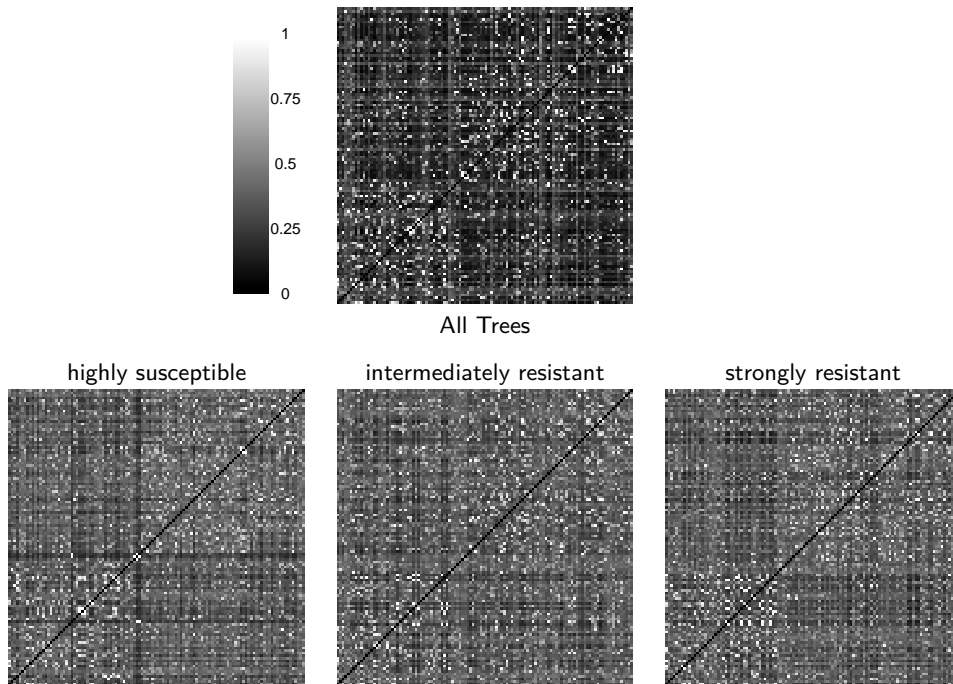


Figure 4.1 – Posterior edge probability matrix.

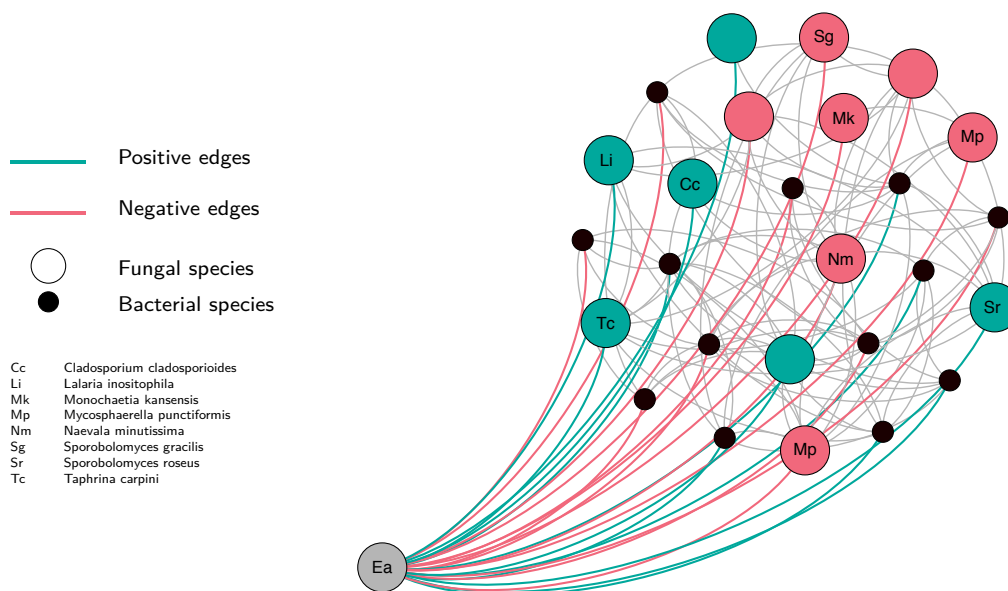


Figure 4.2 – Interactions of *Erysiphe alphitoides* (Ea) with other OTUs on the susceptible oak tree. The names of the fungal OTUs that could be assigned to species level are indicated.



Directed graphical models prove to be more suited to this situation than their undirected counter-parts. Siracusa (2009) introduced Temporal Interaction Models as a way to include structured temporal dependence in the modelling of multivariate time-series. When the time dependence structure of these models is taken to be a directed tree, a directed version of the Matrix-Tree theorem can be used to perform calculation in the same manner as in the undirected case.

When considering the segmentation model, the factorability assumption of Rigail et al. (2012) is not violated as long as there is no temporal dependence between segments. Within segments however, there is no problem introducing dependence if we are able to compute the marginal likelihood.

#### 4.2.1 Temporal Interaction Models

Let  $y = \{y^t\}_{t=1}^N$  be a realisation of a multivariate random process  $Y = \{Y^t\}_{t=1}^N$  of dimension  $p \geq 2$ . For  $1 \leq t \leq N$ ,  $Y^t = (Y_1^t, \dots, Y_p^t)$  is a multivariate random variable taking values in a product of measurable spaces  $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$ . We let  $\mathfrak{D}$  denote the set of DAGs on  $V$ . For  $D \in \mathfrak{D}$  and  $i \in V$ , we let  $\text{pa}(i, D)$  denote the parents of vertex  $i$  in  $D$ . When there is no ambiguity on the DAG, we drop  $D$  from the notation.

Directed graphical models are expressed in terms of conditional distributions rather than marginal distributions. A conditional distribution is a collection of distributions indexed by the values that can be taken by the conditioning variables. Such a collection is called a *Markov kernel*.

**Definition 4.1.** Let  $(\mathfrak{X}, \mathcal{A})$  and  $(\mathfrak{B}, \mathcal{B})$  be two measurable spaces. A Markov kernel from  $\mathfrak{X}$  to  $\mathfrak{B}$  is a map  $\kappa : \mathfrak{X} \times \mathfrak{B} \rightarrow [0; 1]$  such that,

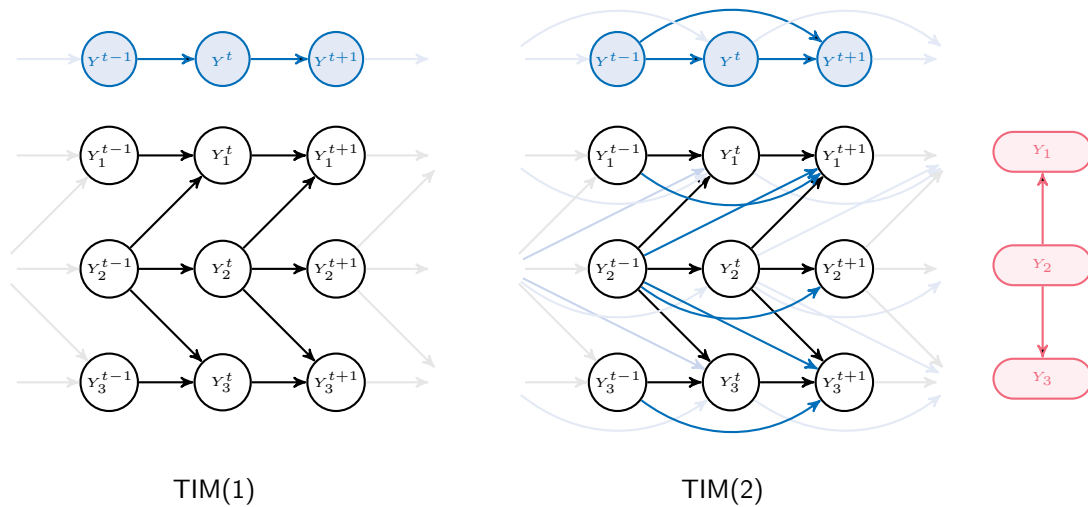
- for all  $B \in \mathcal{B}$ , the map  $x \mapsto \kappa(x, B)$  is  $\mathcal{A}$ -measurable,
- for all  $x \in \mathfrak{X}$ , the map  $B \mapsto \kappa(x, B)$  is a probability measure on  $(\mathfrak{B}, \mathcal{B})$ .

We begin by describing Temporal Interaction Models with unitary lag or TIM(1). Such models are given by a DAG  $D$  and a set of kernels  $\kappa = \{\kappa_{i|\text{pa}(i,D)}\}_{i \in V}$ , where, for  $i \in V$ ,  $\kappa_{i|\text{pa}(i,D)}$  is a kernel from  $\mathcal{X}_{\text{pa}(i,D) \cup \{i\}}$  to  $\mathcal{X}_i$ . We also need a set of initial conditions  $y^0 \in \mathcal{X}$ . The likelihood of  $y$  is then given by

$$p(y|D, \kappa, y^0) = \prod_{t=1}^N \prod_{i \in V} \kappa_{i|\text{pa}(i)}(y_{\text{pa}(i) \cup \{i\}}^{t-1}, y_i^t).$$

The kernels in  $\kappa$  describe the distribution of  $Y^t|Y^{t-1}$  and this distribution factorises according to  $D$ , so that  $Y_i^t$  depends on  $Y_i^{t-1}$  and  $Y_{\text{pa}(i)}^{t-1}$ . In a parametric framework, these kernels would just be a set of parameters  $\{\theta_{i|\text{pa}(i)}\}$ , where  $\theta_{i|\text{pa}(i)}$  specifies the distribution of  $Y_i^t|Y_i^{t-1}, Y_{\text{pa}(i)}^{t-1}$ . We refer to  $D$  as the dependence structure of the model.

Temporal Interaction Models with an arbitrary lag  $\ell$ , or TIM( $\ell$ ) in short, are obtained similarly, but  $Y^t$  is assumed to depend on  $(Y^{t-\ell}, \dots, Y^{t-1})$ . For  $i \in V$ ,  $\kappa_{i|\text{pa}(i)}$  is now a kernel from  $(\mathcal{X}_{\text{pa}(i) \cup \{i\}})^\ell$  to  $\mathcal{X}_i$ . We also need to provide extended initial conditions  $(y^{-\ell+1}, \dots, y^0)$ . The graphical representation of an example of TIM(1) and TIM(2) sharing the same dependence structure is given in Figure 4.3.



**Figure 4.3** – Graphical representation of a TIM(1) and TIM(2) model with tree dependence structure. The temporal and dependence structures are respectively depicted in blue and red.

## 4.2.2 Bayesian Inference of Tree-structured TIM

We limit our attention to TIMs whose dependence structures are directed trees, *i.e.* connected DAGs in which each vertex has at most one parent. We remind that  $\vec{\mathcal{T}}$  denotes the set of such DAGs. Performing Bayesian inference in these models requires to define prior distributions on  $D$  and  $\kappa$ . If these priors factorise properly, the marginal likelihood of the observations, integrated on both  $D$  and  $\kappa$ , can be obtained in polynomial time.

### 4.2.2.a Prior Distribution on $D$

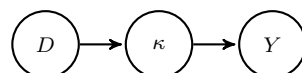
As  $D \in \vec{\mathcal{T}}$ , for  $i \in V$ ,  $\text{pa}(i)$  is either empty or reduced to a single vertex. A distribution on  $\vec{\mathcal{T}}$  can therefore be given by a  $p \times p$  matrix  $\beta$  through

$$p(D) = \frac{1}{\vec{Z}(\beta)} \prod_{i \in V} \beta_{\text{pa}(i), i}, \quad \forall D \in \vec{\mathcal{T}},$$

where,  $\forall i \in V$ ,  $\beta_{\emptyset, i} = \beta_{i, i}$ . Contrary to the undirected case,  $\beta$  is not necessarily symmetric. The normalising constant  $\vec{Z}(\beta)$  is given by

$$\vec{Z}(\beta) = \sum_{D \in \vec{\mathcal{T}}} \prod_{i \in V} \beta_{\text{pa}(i, D), i}.$$

The sum in  $\vec{Z}(\beta)$  involves  $p^{p-1}$  terms. Indeed, we have seen that there are  $p^{p-2}$  undirected spanning trees on  $p$  vertices, and each of these trees can be oriented in  $p$  different ways by



**Figure 4.4** – Hierarchical model for structure inference in TIM.

choosing a root and directing all edges away from this root. As stated by Siracusa (2009), the summation over the trees rooted at a vertex  $r$  can be performed in  $O(p^3)$  time thanks to the directed version of the Matrix-Tree theorem. We define the Laplacian matrix associated to a directed edge weight matrix  $\beta$  as the matrix whose general term is given by

$$\vec{\Delta}_{i,j}(\beta) = \begin{cases} -\beta_{i,j} & \text{if } i \neq j, \\ \sum_{l=1}^p \beta_{l,j} - \beta_{i,i} & \text{if } i = j. \end{cases}$$

If we let  $\vec{\mathcal{T}}_r$ ,  $r \in V$ , denote the set of directed trees rooted at  $r$ , Theorem 1.7 directly yields that

$$\vec{Z}_r(\beta) := \sum_{D \in \vec{\mathcal{T}}_r} \prod_{(i,j) \in \mathcal{E}_D} \beta_{i,j} = |\vec{\Delta}^{\{r\}}(\beta)|, \quad \forall r \in V,$$

where we remind that  $\vec{\Delta}^{\{r\}}(\beta)$  is the matrix obtained from  $\vec{\Delta}(\beta)$  by removing row and column  $r$ . Using this formula to compute  $\vec{Z}(\beta)$  yields a time complexity of  $O(p^4)$ , as  $\vec{Z}(\beta) = \sum_r \beta_{r,r} \vec{Z}_r(\beta)$ . It can in fact be obtained in  $O(p^3)$  time.

**Proposition 4.6** (Koo et al., 2007, Prop. 1). *Let  $\vec{\Delta}(\beta)$  be the Laplacian matrix associated to a directed edge weight matrix  $\beta$ . We denote  $\vec{\Delta}^*(\beta)$  the matrix obtained by replacing the first row in  $\vec{\Delta}(\beta)$  by  $[\beta_{1,1}, \dots, \beta_{p,p}]$ . Then it holds that*

$$\vec{Z}(\beta) = |\vec{\Delta}^*(\beta)|.$$

#### 4.2.2.b Prior Distribution on $\kappa$

The distribution on  $\kappa$ , conditionally on  $D$ , is taken so that the kernels in  $\{\kappa_{i|\text{pa}(i,D)}\}_{i \in V}$  are mutually independent. The distribution for each individual kernel is also chosen once and for all dependence structures. In a parametric framework, Siracusa (2009) refer to these properties as *parameter independence* and *parameter modularity*. Under these assumptions, the distribution of  $\kappa$  conditionally on  $D$  is specified for all  $D \in \vec{\mathcal{T}}$  through a collection of distributions for individual kernels  $\{\kappa_{i|j}\}_{(i,j) \in V^2}$  where  $\kappa_{i|j}$  is a kernel from  $(\mathcal{X}_{\{i,j\}})^\ell$ ,  $\ell$  being the lag of the model, to  $\mathcal{X}_i$ . By convention,  $\kappa_{i|i}$  is the kernel associated to  $\text{pa}(i) = \emptyset$ . In a Gaussian setting, these kernels are linear regressions models with  $Y_i^t$  as response variable and  $(Y_i^{t-1}, Y_j^{t-1})$  as predictor variables, and the prior distributions on the regression coefficients and the covariance matrix of the noise are drawn independently between the different models.

Assume that, for  $(i,j) \in V^2$  the distribution of  $\kappa_{i|j}$  admits a density  $\varsigma_{i|j}$  with respect to some measure  $\nu_{i|j}$  on the kernels from  $(\mathcal{X}_{\{i,j\}})^\ell$  to  $\mathcal{X}_i$ . Then the density of the distribution of  $\kappa|K$  with respect to the appropriate product measure is given by

$$\varsigma(\kappa|D) = \prod_{i \in V} \varsigma_{i|\text{pa}(i)}(\kappa_{i|\text{pa}(i,D)}).$$

When the prior on  $\kappa$  is chosen of this form, the posterior distribution on  $D$  continues to factorises on directed edges and can be expressed as

$$p(D|y) = \frac{1}{\vec{Z}(\omega)} \prod_{i \in V} \omega_{\text{pa}(i),i}, \quad \forall D \in \vec{\mathcal{T}},$$

where, by convention,  $\omega_{\emptyset,i} = \omega_{i,i}$ , and  $\forall(i,j) \in V^2$ ,

$$\omega_{i,j} = \beta_{i,j} \int \left( \prod_{t=1}^N \kappa_{j|i}(y_{i,j}^{t-\ell}, \dots, y_{i,j}^{t-1}, y_j^t) \right) \varsigma_{j|i}(\kappa_{j|i}) d\nu_{j|i}(\kappa_{j|i}).$$

Matrix  $\omega$  contains the posterior weights associated to every possible directed edges that can be borrowed by a tree in  $\vec{\mathcal{T}}$ .

The kernels in  $\kappa$  define a distribution for  $(Y_1, \dots, Y_p)$ , so that  $\varsigma$  induces a hyperdistribution on the set of distributions on  $\mathcal{X}^N = \bigotimes_{i \in V} \mathcal{X}_i^N$ . In Chapter 1, we have not defined hyper Markov properties for DAGs, but in this case, parameter independence is actually an other name for the strong hyper Markov property with respect to a DAG.

**Definition 4.2.** *Let  $D$  be a DAG on  $V$ . A hyperdistribution  $\rho$  over a random distribution  $\pi$  is said to be strong hyper Markov with respect to  $D$ , if it holds that  $\{\pi_{v|pa(v)}\}_{v \in V}$  are mutually independent under  $\rho$ .*

#### 4.2.2.c Computing Posterior Edge Probabilities

We now give a proof a theorem, similar to Theorem 2.3, on the computation of posterior edge probabilities in the case of directed trees. This result is mentioned in (Koo et al., 2007). In this section, we drop  $\omega$  from the notations and  $\vec{\Delta}$ ,  $\vec{\Delta}^*$ ,  $\vec{Z}$  are used as shorthands for  $\vec{\Delta}(\omega)$ ,  $\vec{\Delta}^*(\omega)$ ,  $\vec{Z}(\omega)$ . We begin by proving the following lemma on the derivatives of  $\vec{Z}$ .

**Lemma 4.1.** *Let  $\vec{M}$  be the matrix whose general term is given by*

$$\vec{M}_{k,l} = \begin{cases} [(\vec{\Delta}^*)^{-1}]_{1,l} & \text{if } k = l, \\ [(\vec{\Delta}^*)^{-1}]_{l,l} - (1 - \delta_{1k}) [(\vec{\Delta}^*)^{-1}]_{k,l} & \text{if } l \neq k. \end{cases}$$

Then, it holds that, for  $\{k,l\} \in V^2$ ,

$$\frac{\partial \vec{Z}}{\partial \omega_{k,l}} = \vec{M}_{k,l} \cdot |\vec{\Delta}^*|.$$

*Proof.* By Proposition 4.6,  $\vec{Z} = |\vec{\Delta}^*|$ . We assume that  $\vec{Z} \neq 0$ , so that  $\vec{\Delta}^*$  is non-singular and

$$\frac{\partial |\vec{\Delta}^*|}{\partial \vec{\Delta}_{i,j}^*} = |\vec{\Delta}^*| \cdot [(\vec{\Delta}^*)^{-1}]_{i,j}, \quad \forall(i,j) \in V^2.$$

For  $k \neq l$  and  $k \neq 1$ , the only elements of  $\vec{\Delta}^*$  that depend on  $\omega_{k,l}$  are  $\vec{\Delta}_{k,l}^* = -\omega_{k,l}$  and  $\vec{\Delta}_{l,l}^* = \sum_{u \neq l} \omega_{u,l}$ , so that

$$\frac{\partial \vec{Z}}{\partial \omega_{k,l}} = \frac{\partial |\vec{\Delta}^*|}{\partial \omega_{k,l}} = \sum_{(i,j) \in V^2} \frac{\partial |\vec{\Delta}^*|}{\partial \vec{\Delta}_{i,j}^*} \frac{\partial \vec{\Delta}_{i,j}^*}{\partial \omega_{k,l}} = |\vec{\Delta}^*| \left( [(\vec{\Delta}^*)^{-1}]_{l,l} - [(\vec{\Delta}^*)^{-1}]_{k,l} \right).$$

The result is similarly obtained for  $k = 1$ ,  $k \neq l$  and  $k = l$ . □





**Theorem 4.1.** *Posterior edge probabilities can be computed for all edges at once in  $O(p^3)$  time from the posterior edge weight matrix  $\omega$  as*

$$P(\{k, l\} \in \mathcal{E}_D | y) = \omega_{k,l} \cdot \vec{M}_{k,l},$$

where  $\vec{M}$  is defined as in Lemma 4.1

*Proof.* This proof follows the outline of the proof written for Theorem 2.3 in the case of undirected trees. Let  $(k, l)$  be a directed edge. We define

$$\vec{Z}^+ := \sum_{\substack{D \in \vec{\mathcal{T}} \\ (k,l) \in \mathcal{E}_D}} \prod_{i \in V} \omega_{\text{pa}(i,D),i}, \quad \vec{Z}^- := \sum_{\substack{D \in \vec{\mathcal{T}} \\ (k,l) \notin \mathcal{E}_D}} \prod_{i \in V} \omega_{\text{pa}(i,D),i},$$

so that  $\vec{Z} = \vec{Z}^+ + \vec{Z}^-$  and  $P(\{k, l\} \in \mathcal{E}_D | y) = \vec{Z}^+ / \vec{Z}$ . Using Lemma 4.1, as  $\vec{Z}^-$  does not depend on  $\omega_{k,l}$ , we have that

$$\frac{\partial \vec{Z}^+}{\partial \omega_{k,l}} = \frac{\partial \vec{Z}}{\partial \omega_{k,l}} = \vec{M}_{k,l} \cdot |\vec{\Delta}^*|. \quad (4.15)$$

Using Corollary 1.8.1, we also get that

$$\begin{aligned} \vec{Z}^+ &= \sum_{r \in v} \omega_{r,r} \sum_{\substack{D \in \vec{\mathcal{T}}_r \\ (k,l) \in \mathcal{E}_D}} \prod_{(i,j) \in \mathcal{E}_D} \omega_{i,j} \\ &= \omega_{k,l} \sum_{r \in v} \omega_{r,r} \left| |\vec{\Delta}^{\{k,l\},\{r,l\}}| \right| \\ &= \omega_{k,l} \cdot L_{k,l} \end{aligned} \quad (4.16)$$

where  $\left| |\vec{\Delta}^{\{k,l\},\{r,l\}}| \right|$  stands for the absolute value of the determinant of the matrix obtained from  $\vec{\Delta}$  by removing rows  $k$  and  $l$  and columns  $r$  and  $l$ . As  $\left| |\vec{\Delta}^{\{k,l\},\{r,l\}}| \right|$ ,  $r \in V$ , does not depend on  $\omega_{k,l}$ , neither does  $L_{k,l}$ , and

$$\frac{\partial \vec{Z}^+}{\partial \omega_{k,l}} = L_{k,l}. \quad (4.17)$$

Combining (4.15), (4.16) and (4.17) yields

$$\vec{Z}^+ = \omega_{k,l} \cdot \vec{M}_{k,l} \cdot |\vec{\Delta}^*|, \quad P(\{k, l\} \in \mathcal{E}_D | y) = \omega_{k,l} \cdot \vec{M}_{k,l}. \quad (4.18)$$

The complexity of the computation is dominated by the complexity of the inversion of  $\vec{\Delta}^*$  required to compute  $\vec{M}$ , and all posterior edge probabilities can therefore be obtained in  $O(p^3)$  time.  $\square$

### 4.3 Prior Distribution on Segmentations

In Chapter 3, we computed, among other things, the posterior probability of a change-point occurring at a given time  $t$ . Assuming that the number of segments is equal to  $K$ , we are

actually evaluating the posterior probability associated to the subset of  $\mathcal{M}_K$  made of the segmentations having a change-point at time  $t$ . This subset was denoted by  $\mathcal{B}_K(t)$  in Chapter 3. The posterior probability given to this subset obviously depends on the prior distribution that was set on  $\mathcal{M}_K$ . We remind that we limited our attention to the distributions on  $\mathcal{M}_K$  that factorise over the segments. Such distributions are given through a upper triangular matrix  $a$  with non-negative terms by

$$p(m) = \frac{1}{C_K(a)} \prod_{r \in m} a_r, \quad \forall m \in \mathcal{M}_K. \quad (4.19)$$

Under such prior distributions, the prior probability of  $\mathcal{B}_K(t)$  can be computed exactly as described in Section 3.4.1 by replacing  $A$  by  $a$ . A uniform prior on  $\mathcal{M}_K$  can be obtained by setting all terms above the diagonal in  $a$  to 1. But matrix  $a$  can be used to enforce additional constraints on the allowed segmentations. For instance, segments with length smaller than  $L$  can be forbidden by putting the terms in  $a = (a_{s,t})$  such that  $t = s + l$ ,  $1 \leq l \leq L$ , to zero. Similarly, it is possible to enforce that all segments have an even length by settings all terms corresponding to odd segments to zero. In  $a$ , each superdiagonal corresponds to segments of a given length, so that designing a matrix allowing for a list of given lengths is straightforward. In the aforementioned examples, whenever all non-zero terms are set to one,  $a$  induces a uniform distribution on its support. But we are by no means limited to binary values. Segmentations with segments of similar lengths can for instance be favoured *a priori* by setting

$$a_{s,t} = |t - s|^d, \quad 1 \leq s < t \leq N + 1, \quad d > 0. \quad (4.20)$$

The greater  $d$  is taken, the more we tip the scale in favour of homogeneous segmentations in terms of segment length. Similar results can be obtained when power functions are replaced with any strictly increasing non-negative function on  $\mathbf{R}^+$ . On the contrary, heterogeneous segmentations are favoured when  $a_{s,t}$  is a decreasing function of the length of segment  $\llbracket s; t \rrbracket$ . Obviously it is perfectly conceivable to combine (4.20) with hard constraints to obtain complex prior distributions. Figure 4.5 illustrates the different examples that we have given on the segmentations of  $\llbracket 1; 20 \rrbracket$ .

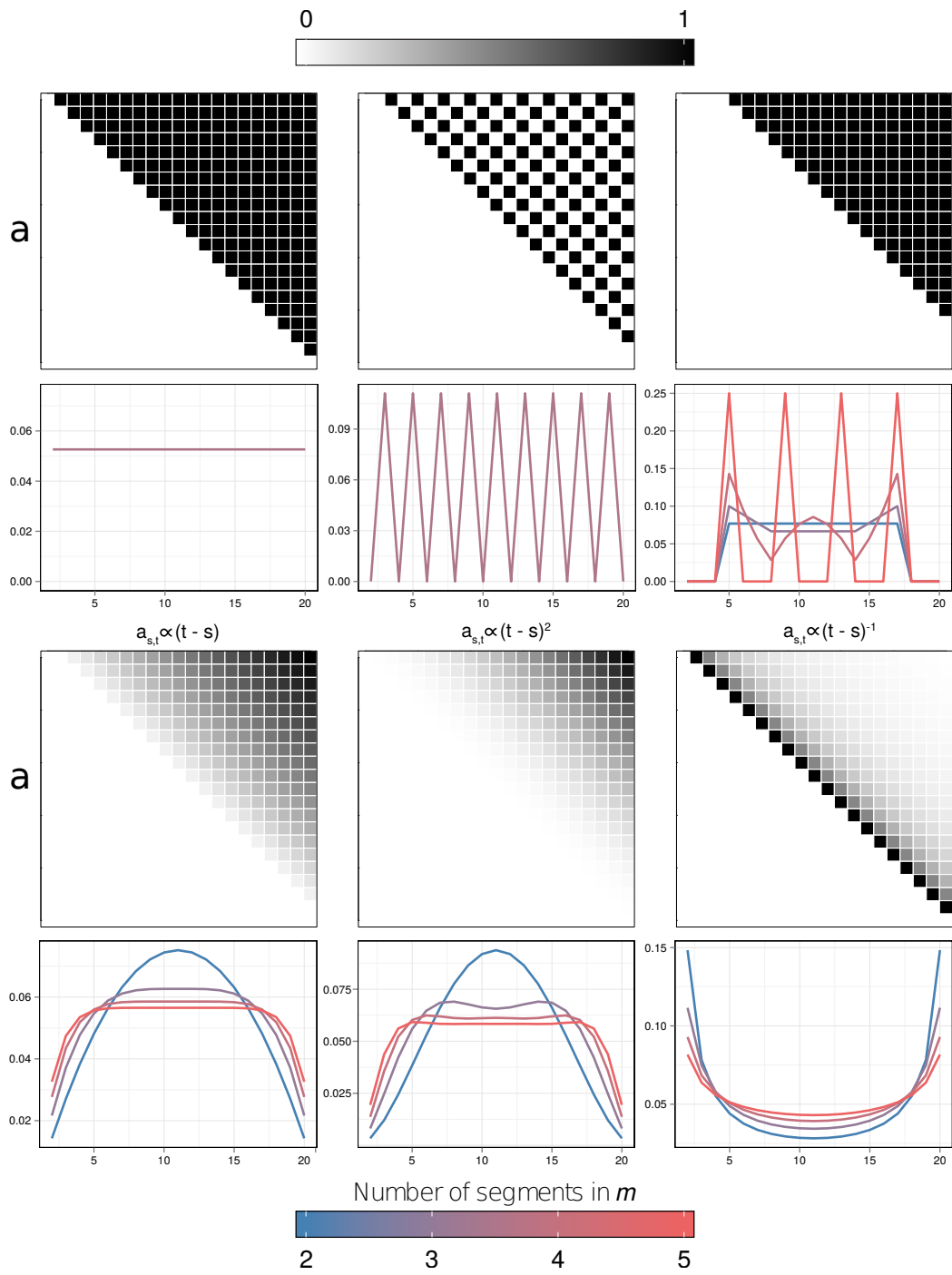
### 4.3.1 A Different Kind of Prior

Now imagine that we are given a segmentation  $\tilde{m}$ , either from previous analyses or from expert elicitation, in which we believe to some extent. Our wish would be to specify a prior distribution on  $m$  that concentrates around this initial guess. This is possible to some extent with the prior distributions we have considered so far. We consider a different kind of prior for  $m$  in which it is straightforward, as well as the consequences of this change on the inference.

For  $m \in \mathcal{M}_K$  and  $1 \leq k \leq K$ , we let  $\ell_k(m) = |r_k(m)|/N$ , where  $r_k(m)$  denotes the  $k$ -th segment of  $m$  and  $|r_k(m)|$  denotes its length. When  $a$  is chosen as in (4.20), the prior distribution on  $m$  can be written as

$$\begin{aligned} p(m) &\propto \prod_{k=1}^K \ell_k(m)^d, & \forall m \in \mathcal{M}_K, \\ &\propto f(\ell_1(m), \dots, \ell_K(m) | d + 1, \dots, d + 1), \end{aligned}$$





**Figure 4.5** – Prior probability of observing a change-point as a function of time, for different choices of prior matrices  $a$  (represented above each set of curves) and for a number of segments  $K \in [2; 5]$ . Probabilities are normalised by the number of change-points of the corresponding segmentations, *i.e.*  $K - 1$ .

where  $f(\cdot|d+1, \dots, d+1)$  stands for the density of the Dirichlet distribution on

$$\mathcal{L}_K = \left\{ \ell \in [0; 1]^K \mid \sum_{k=1}^K \ell_k = 1 \right\}$$

with parameters  $(d+1, \dots, d+1)$ . In a general Dirichlet distribution, the parameters are not assumed to be identical. In our context, that would mean that the position of a segment within a segmentation influences its contribution to the probability of the segmentation, with a distribution of the form

$$p(m) \propto \prod_{k=1}^K \ell_k(m)^{d_k}. \quad (4.21)$$

If  $(d'_1, \dots, d'_K) = (d_1 + 1, \dots, d_K + 1)$ , the distribution of  $(\ell_1(m), \dots, \ell_K(m))$  is basically a discretised version of  $\mathcal{D}(d'_1, \dots, d'_K)$ . Indeed, it can be seen as a distribution on  $\mathcal{L}_K$  whose density  $f_N$  is proportional to

$$f_N(\ell_1, \dots, \ell_K | d_1, \dots, d_K) \propto f\left(\frac{\lfloor N\ell_1 \rfloor}{N}, \dots, \frac{\lfloor N\ell_K \rfloor}{N} \mid d'_1, \dots, d'_K\right), \quad \forall \ell \in \mathcal{L}_K. \quad (4.22)$$

Now, if  $\ell \sim \mathcal{D}(d'_1, \dots, d'_K)$ , the mean and mode of  $\ell$  are respectively given by

$$\mathbf{E}[\ell_k] = \frac{d'_k}{\sum_{k=1}^K d'_k}, \quad \ell_k^{\max} = \frac{d_k}{\sum_{k=1}^K d_k}, \quad \forall k \in \llbracket 1; K \rrbracket.$$

Therefore, if we are given a prior guess  $\tilde{m}$  on the segmentation, the vector  $d'$  can be taken proportional to  $\ell(\tilde{m})$ , so that the Dirichlet distribution underpinning the distribution of  $m$  has an expected value corresponding to  $\tilde{m}$ .

Using such prior distributions for  $m$  in the general model described in Chapter 3 requires some adjustments. We now have to consider one weighted segment likelihood matrix  $A$  per position within the segmentation. The matrix linked with the  $k$ -th segment is denoted by  $A^{(k)}$ , and its general term is given by

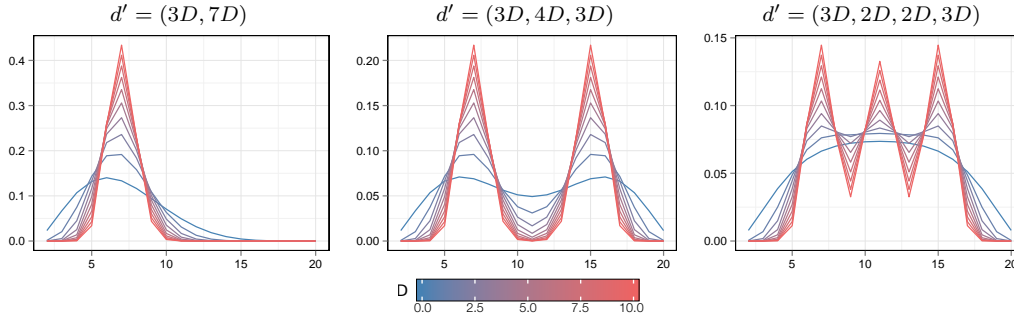
$$A_{s,t}^{(k)} = \begin{cases} a_{s,t}^{(k)} \cdot p(y^{\llbracket s;t \rrbracket}) & \text{if } s < t, \\ 0 & \text{otherwise,} \end{cases} \quad a_{s,t}^{(k)} = \begin{cases} |t-s|^{d_k} & \text{if } s < t, \\ 0 & \text{otherwise,} \end{cases} \quad \forall k \in \llbracket 1; K \rrbracket.$$

For  $1 \leq k_1 \leq k_2 \leq K$ , we let  $a^{(k_1:k_2)} := \prod_{k=k_1}^{k_2} a^{(k)}$  and  $A^{(k_1:k_2)} := \prod_{k=k_1}^{k_2} A^{(k)}$ . The latter are meant to replace the powers of matrix  $A$  involved in the different results of Section 3.4. The normalising constant in (4.22) is equal to  $[a^{(1:K)}]_{1,N+1}$  and the marginal likelihood of the observations is therefore given by

$$p(y|K) = \frac{[A^{(1:K)}]_{1,N+1}}{[a^{(1:K)}]_{1,N+1}}.$$

Similarly, the posterior probabilities associated to

$$\begin{aligned} \mathcal{B}_{K,k}(t) &= \{m \in \mathcal{M}_K \mid \tau_k = t\}, & \mathcal{B}_K(t) &= \bigcup_{k=1}^K \mathcal{B}_{K,k}(t), \\ \mathcal{S}_{K,k}(\llbracket s; t \rrbracket) &= \{m \in \mathcal{M}_K \mid r_k = \llbracket s; t \rrbracket\}, & \mathcal{S}_K(\llbracket s; t \rrbracket) &= \bigcup_{k=1}^K \mathcal{S}_{K,k}(\llbracket s; t \rrbracket), \end{aligned}$$



**Figure 4.6** – Prior probability of observing a change-point as a function of time, for different values of  $d$ . Probabilities are normalised by  $|d| - 1$ .

are now given by

$$B_{K,k}(t) = \frac{[A^{(1:k)}]_{1,t} [A^{(k+1:K)}]_{t,N+1}}{[A^{(1:K)}]_{1,N+1}}, \quad B_K(t) = \sum_{k=1}^{K-1} B_{K,k}(t),$$

$$S_{K,k}(\llbracket s; t \rrbracket) = \frac{[A^{(1:k-1)}]_{1,s} A_{s,t}^{(k)} [A^{(k+1:K-)}]_{t,N+1}}{[A^{(1:K)}]_{1,N+1}}, \quad S_K(\llbracket s; t \rrbracket) = \sum_{k=1}^K S_{K,k}(\llbracket s; t \rrbracket).$$

As these probabilities can all be computed from the first line of matrices  $\{A^{(1:k)}\}_{k=1}^K$  and the last column of matrices  $\{A^{(k:K)}\}_{k=1}^K$ , we retain a complexity of  $O(KN^2)$ . Prior probabilities are likewise obtained through matrices  $\{a^{(k_1:k_2)}\}_{1 \leq k_1 \leq k_2 \leq K}$  (see Figure 4.6 for some examples).

### 4.3.2 Transferring Knowledge from One Dataset to Another

We begin this section with a practical example. Genome annotation is a segmentation problem in which one tries to delimit the regions of a DNA sequence that are transcribed to produce RNA, also called genes, from intergenic regions. Different technologies are currently available to measure gene expression along a genome.

DNA microarrays have been used for almost twenty years. Roughly speaking, these microarrays consist of a collection of short single-strand DNA sequences, or probes, fixed and arranged on a solid support. Each probe is represented a great number of time. The RNA sequences contained in a biological sample of interest are retro-transcribed into single-strand DNA sequences and marked with a fluorescent compound. These DNA sequences are put on the chip and a property of DNA called hybridisation brings them together with probe sequences if their nucleotidic sequences are matching. After washing off unhybridised sequences, the fluorescence associated to each group of probes is measured and used as a proxy for gene expression level at the *locus* of the probe. The number of probes is usually small with respect to the length of the sequence of interest. This technology therefore produces low-resolution continuous measurements of gene expression.

More recent technologies are based on high-throughput sequencing, also called next-generation sequencing or NGS. They are referred to RNA-seq methods. In such methods, the RNA sequences of a biological sample are cut into small pieces and sequenced. After

aligning all these sequences on a reference genome, we get a profile indicating the number of times each nucleotide has been seen in an RNA sequence, so that gene expression is given by count data, at a much higher resolution than microarray data.

Although RNA-seq methods are progressively overshadowing microarrays, a lot of data are available from times preceding the advent of NGS. Imagine a situation in which we have both microarray and RNA-seq data. Ideally, the microarray data could be used to provide some kind of prior information for the analysis of the finer RNA-seq data. This is actually possible in the framework that we have described.

Let  $y^1, y^2$  be realisations of two random processes  $Y^1, Y^2$  with  $N_1, N_2$  time-points respectively. Each random process is modelled with a segmentation model, for which we assume that the number of segments is known to be  $K$ :

$$\begin{aligned}
 p(m^1|d^1) &\propto \prod_{k=1}^K \ell_k(m)^{d_k^1}, & p(m^2|d^2) &\propto \prod_{k=1}^K \ell_k(m)^{d_k^2}, \\
 p(\theta^1|m^1, \alpha^1) &= \prod_{r \in m^1} p(\theta_r^1|\alpha^1), & p(\theta^2|m^2, \alpha^2) &= \prod_{r \in m^2} p(\theta_r^2|\alpha^2), \\
 p(y^1|m^1, \theta^1) &= \prod_{r \in m^1} p(y^{1,r}), & p(y^2|m^2, \theta^2) &= \prod_{r \in m^2} p(y^{2,r}).
 \end{aligned}$$

The marginal likelihood of  $y^1$  is given by

$$p(y^1|d^1, \alpha^1) \propto \sum_{m \in \mathcal{M}_K(\llbracket 1; N_1 \rrbracket)} \prod_{k=1}^K \ell_k(m)^{d_k^1} p(y^{1,r}|\alpha^1).$$

Let  $\hat{d}^1$  be the value of  $d^1$  maximising this expression. We are therefore dealing with the first dataset in a frequentist paradigm. Taking  $d^2 = \hat{d}^1$ , we then use what we have learned on the distribution of  $m^1$  to specify a prior distribution on  $m^2$  and perform classical Bayesian inference on the second dataset. A schematic representation of this procedure is given in Figure 4.7. The key element is that a given Dirichlet distribution can be discretised at any level  $N \in \mathbf{N}$  to obtain a distribution on  $\mathcal{M}_K(\llbracket 1; N \rrbracket)$ . This procedure is thus allowed by the specific form that we have chosen for the prior distribution of  $m$ . It would not be possible

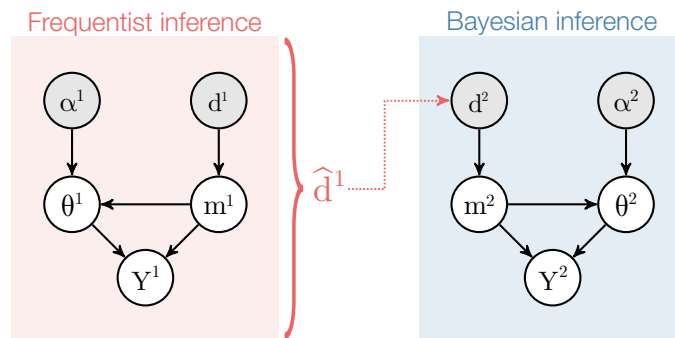
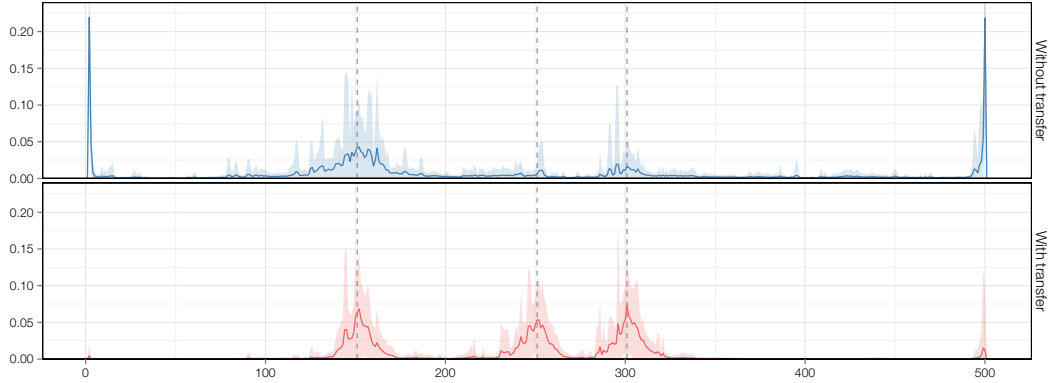


Figure 4.7 – Schematic representation of the knowledge transfer procedure.



**Figure 4.8** – Posterior probability of observing a change-point for the second dataset without (top) and with (down) knowledge transfer from the first dataset. The curve represents the mean value obtained from 50 samples and the ribbon gives the standard deviation.

with a prior of the form  $p(m) = \prod_{r \in m} a_r$ .

If we were to perform both steps of the procedure on the same dataset, we would be considering what is called an empirical Bayes approach. Such methods are straying from standard Bayesian inference, as information is drawn from the data before actually performing the inference. There is no proper theoretical basis behind such approaches, and all qualms raised by the facts that data are used twice are well founded.

We finish this section with a toy example on synthetic data in the context of gene annotation. We took datasets of respective sizes  $N_1 = 100$  and  $N_2 = 500$ . In both cases, the true segmentation was chosen with segment lengths proportional to  $(0.3, 0.2, 0.1, 0.4)$ . Datasets were simulated as follows:

$$\begin{aligned} \mu &= (20, 10, 20, 10), & \lambda &= (5, 4, 5, 4), \\ \sigma &= 2.5, & Y^{2,t} &\sim \mathcal{P}(\lambda_k), \quad \forall t \in r_k. \\ Y^{1,t} &\sim \mathcal{N}(\mu_k, \sigma^2), \quad \forall t \in r_k, \end{aligned}$$

For the inference, the prior distributions on  $\mu$ ,  $\sigma$  and  $\lambda$  were chosen as in (Rigail et al., 2012). On the first dataset, the prior distribution on segmentations was set to the uniform. On the second one, it was either set to the uniform or using  $\hat{d}^1$ , which was estimated through numerical optimisation. Figure 4.8 shows the posterior probability of observing a change-point for the second dataset on 50 repetitions and for both prior distribution on segmentations. As expected, on this very simple example, we notice that there is an interest in transferring knowledge from the first dataset to the second. By doing so, we are able to detect the smallest of the four segments that is otherwise overlooked.

# 5

## En Bref

---

*This chapter, written in French, is a standalone and substantial abstract of the remainder of the dissertation.*

Que l'on parle de réseaux de régulation de gènes, de réseaux informatiques ou encore de réseaux sociaux, il est indéniable que le concept de réseau est central dans de nombreux domaines. On peut définir un réseau comme un système d'entités inter-connectées. Dans les problèmes qui nous intéressent, le réseau n'est pas connu. Il ne s'agit donc pas d'étudier les propriétés d'un réseau que l'on observe, mais plutôt de travailler sur un réseau latent. Dans un premier temps, nous nous intéressons à ce qui est communément appelé l'*inférence de réseaux*, dans lequel le but est de reconstruire ce réseau latent à partir d'observations individuelles sur chacune des entités impliquées. Lorsque ces observations sont organisées en série temporelle, il est souvent peu réaliste de supposer que le réseau est stationnaire. Nous concentrons notre attention sur le cas où le réseau subit un certain nombre de brusques changements au cours du temps. La détermination de ces points de rupture est un problème de *segmentation*.

Le formalisme naturel dans lequel traiter les problèmes impliquant des réseaux est celui des *modèles graphiques*. Un graphe est un ensemble de sommets reliés par des arêtes, qui peuvent être dirigées ou non. Dans un modèle graphique, les relations de dépendance et d'indépendance conditionnelles vérifiées par une distribution sont représentées à l'aide d'un graphe. Ces propriétés abstraites deviennent ainsi plus facilement appréhendables. Dans le cadre des modèles graphiques, l'inférence de réseaux consiste donc en l'inférence du graphe décrivant le modèle graphique. Le graphe est alors lui-même un paramètre du modèle général.

Nous avons décidé de considérer les problèmes d'inférence de réseaux et de segmentation d'un point de vue bayésien. Il est donc nécessaire de définir des distributions *a priori* pour les différents paramètres impliqués, qu'ils soient à valeurs discrètes, tels que le graphe ou la segmentation, ou à valeurs continues. En choisissant ces distributions avec soin, il est possible d'effectuer une inférence bayésienne exacte dans le cadre des modèles graphiques en un temps



polynomial par rapport au nombre de variables. Cette complexité est en particulier permise par une restriction sur l'espace des graphes explorés. En effet, une sommation efficace sur l'ensemble des graphes non-dirigés, connectés et acycliques, aussi appelés arbres couvrants, est permise par un résultat algébrique connu sous le nom de théorème arbre-matrice. Il s'agit d'une voie différente des méthodes basées sur l'échantillonnage, telles que les approches de type *Markov Chain Monte Carlo* (MCMC), qui sont couramment utilisées dans le contexte de l'inférence de réseaux. L'échantillonnage au sein de l'ensemble des graphes est dans notre cas remplacée par une exploration exhaustive d'un sous-ensemble de graphes.

## Chapitre 1

Le premier chapitre fournit le contexte dans lequel s'inscrivent les chapitres suivants. La première partie est consacrée aux modèles graphiques, plus particulièrement aux modèles graphiques non-dirigés. Elle se fonde principalement sur les travaux de Dawid & Lauritzen (1993), Lauritzen (1996) et Byrne (2011). Les modèles dirigés sont évoqués bien que n'étant pas au centre des travaux présentés ici. La seconde présente deux résultats d'algèbre utiles pour la suite.

### Modèles Graphiques

Formellement, un graphe non-dirigé est donné par un ensemble de sommets  $V$ , que l'on prend ici égal à  $\{1, \dots, p\}$ , et par un ensemble d'arêtes  $E$  contenu dans l'ensemble des parties de  $V$  de taille 2, dénoté  $\mathcal{P}_2(V)$ . On note  $G = (V, E)$ . Pour tout graphe  $G$ , on note également  $E_G$  les arêtes de  $G$ . Un chemin  $\gamma$  entre deux sommets distincts  $\alpha, \beta$  d'un graphe  $G$  est une suite  $\alpha = \gamma_0, \dots, \gamma_n = \beta, n \geq 1$ , telle que, pour tout  $1 \leq i \leq n, \{\gamma_{i-1}, \gamma_i\} \in E_G$ .

**Définition** (Séparation). *Soit  $G$  un graphe non-dirigé et  $(A, B, S)$  un triplet de sous-ensembles de  $V$ . L'ensemble  $S$  sépare  $A$  de  $B$  si tout chemin joignant un sommet de  $A$  à un sommet de  $B$  dans  $G$  intersecte  $S$ .*

La notion de séparation est fondamentale pour établir le lien entre graphe et indépendance conditionnelle. Soit  $X = (X_i)_{i \in V}$  un vecteur aléatoire à valeurs dans un espace produit  $\mathcal{X} = \bigotimes_{i \in V} \mathcal{X}_i$ . Pour tout sous-ensemble  $A$  de  $V$ ,  $X_A$  dénote  $(X_i)_{i \in A}$ .

**Propriété** (Markov globale). *Soit  $G$  un graphe. La distribution de  $X$  satisfait la propriété de Markov (globale) par rapport à  $G$  si, pour tout triplet  $(A, B, S)$  de sous-ensembles disjoints de  $V$ , on a*

$$S \text{ sépare } A \text{ de } B \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B | X_S.$$

Ainsi, lorsque qu'une distribution satisfait la propriété de Markov par rapport à un graphe  $G$ , ses propriétés d'indépendance conditionnelle sont représentées par le graphe. Il n'y a cependant qu'une implication entre séparation et indépendance conditionnelle, et toute distribution est par exemple Markov par rapport au graphe complet sur  $V$ . Pour tout graphe  $G$ ,  $\mathcal{M}(G)$  dénote l'ensemble des distributions sur  $\mathcal{X}$  qui sont Markov par rapport à  $G$ .

**Définition** (Modèle graphique). *Un modèle graphique est un couple formé d'un graphe  $G$ , appelé structure du modèle, et d'une famille de distributions  $\mathcal{F}$  telle que  $\mathcal{F} \subseteq \mathcal{M}(G)$ .*

Les notions de propriété de Markov et de modèle graphique peuvent être adaptées au cas des graphes dirigés acycliques, plus connus sous l'acronyme anglais de *DAG*.

L'inférence de la structure d'un modèle graphique dans un cadre bayésien suppose que la distribution des observations ainsi que le graphe donnant la structure du modèle soient eux-mêmes des objets aléatoires. Dawid & Lauritzen (1993) et Byrne (2011) ont respectivement étendu la propriété de Markov aux distributions sur les distributions et sur les graphes. Ces extensions permettent de donner un cadre théorique complet à l'inférence de réseaux.

## Algèbre et Algorithmique

L'ensemble des arbres couvrants sur  $V$ , dénoté  $\mathcal{T}$ , est de cardinal  $p^{p-2}$ , bien inférieur au cardinal de l'ensemble des graphes non-dirigés qui est de  $2^{p(p-1)/2}$ , tout en restant conséquent. Cependant, sous une hypothèse de factorisation sur les arêtes, il est possible d'intégrer une fonction définie sur  $\mathcal{T}$  en un temps polynomial.

**Théorème** (arbre-matrice). *Soit  $\omega = (\omega_{i,j})_{1 \leq i,j \leq p}$  une matrice symétrique à valeurs positives, de diagonale nulle. Soit  $\Delta$  la matrice laplacienne associée à  $\omega$ , donnée par*

$$\Delta_{i,j} = \begin{cases} -\omega_{i,j} & \text{if } i \neq j, \\ \sum_{k \in V} \omega_{k,j} & \text{if } i = j. \end{cases}$$

Alors tous les cofacteurs de  $\Delta$  sont égaux à

$$\Sigma := \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \omega_{i,j}.$$

Ce résultat est attribué, sous différentes formes, à Kirchhoff (1847) ou Cayley (1889). Le calcul de  $\Sigma$  est donc possible en  $O(p^3)$  en tant que le déterminant d'une matrice de taille  $p-1$ . Une version généralisée de ce résultat (Chaiken, 1982) permet de relâcher la contrainte de connexité sur les graphes considérés, et ainsi de passer des arbres aux forêts. Le théorème arbre-matrice et sa version généralisée sont utilisés au Chapitre 2 dans le calcul des probabilités d'apparition d'arêtes *a posteriori*. Nous fournissons une nouvelle preuve d'un résultat du à Kirshner (2007).

Le second résultat énoncé dans cette partie concerne les partitions de  $\llbracket 1; N \rrbracket$ ,  $N \geq 1$ , en sous-ensembles de la forme  $\llbracket i; j \rrbracket$ , appelés segments. De telles partitions sont appelées segmentations. L'ensemble des segmentations d'un segment  $\llbracket i; j \rrbracket$  en un nombre  $K$  de sous-segments est noté  $\mathcal{M}_K(\llbracket i; j \rrbracket)$ . Le résultat suivant permet d'intégrer une fonction sur les segmentations factorisant sur les segments.

**Théorème** (Rigail et al., 2012). *Soit  $f$  une fonction définie par*

$$f(m) = \prod_{\llbracket s; t \rrbracket \in m} A_{s,t}, \quad \forall 1 \leq K \leq N, \forall 1 \leq i < j \leq N+1, \forall m \in \mathcal{M}_K(\llbracket i; j \rrbracket),$$

où  $A$  est une matrice triangulaire supérieure à valeurs positives, de diagonale nulle et de taille  $N+1$ . Alors

$$F_{i,j}^K := \sum_{m \in \mathcal{M}_K(\llbracket i; j \rrbracket)} f(m) = [A^K]_{i,j}.$$

À partir de ce résultat, il est possible de calculer de nombreuses quantités dans les modèles de détection de rupture.

## Chapitre 2

Ce chapitre s'intéresse à l'inférence de réseaux. Comme annoncé dans l'introduction, l'inférence est réalisée dans un cadre bayésien. L'espace des graphes explorés est restreint aux arbres couvrants afin de permettre une inférence exacte. Chow & Liu (1968) sont parmi les premiers à avoir étudié l'inférence de distributions dont la structure de dépendance est arborescente. Depuis, de nombreux travaux se sont penchés sur le sujet. On citera notamment ceux de Meilă & Jordan (2001), Meilă & Jaakkola (2006), Kirshner (2007) et Lin et al. (2009). Cependant, la connexité et l'absence de cycles sont des contraintes fortes inhérentes aux arbres couvrants, qui ont le défaut d'être peu réalistes dans de nombreux cas d'application. Localement, de nombreux réseaux ont cependant une structure arborescente, tout en ne satisfaisant pas ces contraintes globales. C'est pourquoi, au lieu de chercher à apprendre le meilleur arbre, nous nous sommes intéressés à la probabilité d'apparition de structures locales, telles que les arêtes, dans un arbre aléatoire. En moyennant sur l'espace des arbres couvrants, il est ainsi possible d'effacer en partie les contraintes individuelles de chaque arbre.

La contribution majeure de ce chapitre est la construction d'un cadre bayésien complet pour l'inférence de modèles graphiques à structure d'arbre. Ce cadre est fondé sur les travaux de Dawid & Lauritzen (1993) généralisant la propriété de Markov aux distributions sur des espaces de distributions elles-mêmes Markov par rapport à un graphe. De telles distributions sont appelées *hyperdistributions* pour plus de clarté. La propriété hyper Markov forte sur les hyperdistributions permet de définir une distribution *a priori* sur la distribution des observations (ou sur la loi des paramètres de la distribution dans un cadre paramétrique) respectant la structure du modèle graphique. Lorsque ce choix de distribution est combiné à une distribution *a priori* sur les arbres couvrants qui factorise sur les arêtes, il est en fait possible d'intégrer à la fois sur la distribution des observations et sur la structure de manière exacte et efficace.

Nous fournissons également une nouvelle preuve d'un résultat énoncé par Kirshner (2007) sur le calcul des probabilités d'apparition d'arêtes *a posteriori*. Un calcul direct à partir du théorème arbre-matrice conduit à une complexité en  $O(p^5)$  pour le calcul de ces probabilités, mais on montre ici qu'il est en fait possible de se ramener à une complexité en  $O(p^3)$ .

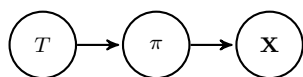
Une étude simulatoire permet de quantifier empiriquement l'impact de l'hypothèse arborescente sur l'inférence de réseaux dans le cas où le réseau à récupérer n'est pas un arbre. Nous observons des performances similaires à celles obtenues par une inférence sur l'espace des graphes dirigés acycliques. Une application sur des données de cytométrie en flux est également présentée, pour laquelle on remarque un léger avantage des graphes dirigés acycliques sur les arbres.

Un paquet R implémentant l'approche développée dans ce chapitre est disponible sur le CRAN sous le nom de **saturnin**.

### Modèle

Reprenant les notations du chapitre précédent,  $X = (X_i)_{i \in V}$  est un vecteur aléatoire à valeurs dans un espace produit  $\mathcal{X} = \bigotimes_{i \in V} \mathcal{X}_i$ . Pour tout arbre  $T \in \mathcal{T}$ , on considère le modèle graphique  $m_T = (T, \mathcal{M}(T))$ , où l'on rappelle que  $\mathcal{M}(T)$  désigne l'ensemble des distributions sur  $\mathcal{X}$  qui sont Markov par rapport à  $T$ . Le modèle global consiste à tirer un arbre  $T$  dans  $\mathcal{T}$  selon une certaine loi  $\xi$ . Une distribution  $\pi$  est ensuite tirée dans le modèle graphique  $m_T$  selon une hyperdistribution  $\rho_T$  définie sur  $\mathcal{M}(T)$ . Enfin,  $X$  est tiré selon  $\pi$ .

Une représentation graphique de ce modèle est donnée ci-dessous.



$$\begin{aligned} T &\sim \xi, \\ \pi|T &\sim \rho^T, \\ \mathbf{X}|\pi &\sim \pi. \end{aligned}$$

Une spécification complète de ce modèle nécessite de donner une distribution *a priori* sur les arbres  $\xi$  et une collection d'hyperdistribution  $\{\rho^T\}_{T \in \mathcal{T}}$  comme distribution *a priori* de  $\pi|T$ . Cette collection peut en fait être définie, sous certaines conditions, en donnant les hyperdistributions marginales sur la loi de  $(X_i, X_j)$  pour tout  $\{i, j\} \in \mathcal{P}_2(V)$ .

### Distributions *a priori* sur $T$ et $\pi$

La distribution *a priori* sur  $T$  est prise de la forme

$$\xi(T) = \frac{1}{Z(\beta)} \prod_{\{i,j\} \in E_T} \beta_{ij}, \quad \forall T \in \mathcal{T}, \quad (5.1)$$

avec  $\beta$  une matrice symétrique à valeurs positives, de diagonale nulle. En prenant  $\xi$  sous cette forme, la constante de normalisation  $Z(\beta) = \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} \beta_{ij}$  peut être calculée grâce au théorème arbre-matrice. L'idée est maintenant de choisir  $\{\rho^T\}_{T \in \mathcal{T}}$  de telle sorte à conserver cette factorisation pour la distribution *a posteriori*  $\xi(\cdot|\mathbf{x})$ . Par la formule de Bayes, on obtient que  $\xi(T|\mathbf{x}) \propto \xi(T)p(\mathbf{x}|T)$ , où  $p(\mathbf{x}|T)$  est la vraisemblance marginale des observations donnée par

$$p(\mathbf{x}|T) = \int \pi(\mathbf{x})\rho^T(\pi)d\pi, \quad \forall T \in \mathcal{T}, \forall \mathbf{x} \in \mathcal{X}.$$

Pour un arbre donné  $T$ ,  $\rho^T$  doit donc être choisie de telle manière que  $p(\mathbf{x}|T)$  reste factorisable sur les arêtes. Ce choix doit de plus être effectué de manière cohérente entre les différents arbres afin d'obtenir une écriture similaire à (5.1) pour  $\xi(\cdot|\mathbf{x})$ . Ceci est effectué en construisant une famille compatible d'hyperdistributions fortement hyper Markov (Dawid & Lauritzen, 1993).

Pour tout couple  $(A, B)$  de sous-ensembles de  $V$ , on note  $\pi_A$  la distribution marginale obtenue à partir de  $\pi$  sur les variables  $\mathbf{X}_A$  et  $\pi_{B|A}$  la collection de distributions conditionnelles de  $\mathbf{X}_B|\mathbf{X}_A$  sous  $\pi$ . On note également  $\rho_A^T$  l'hyperdistribution induite par  $\rho^T$  sur  $\pi_A$  et  $\rho_{B|A}^T$  la collection d'hyperdistributions induites par  $\rho^T$  sur  $\pi_{B|A}$ .

**Définition.**  $\rho^T$  est dite *fortement hyper Markov par rapport à  $T$*  si, pour tous  $A, B \subset V$  tels que  $A \cap B$  est complet et sépare  $A$  de  $B$ ,  $\{\pi_{B|A}, \pi_{A|B}, \pi_{A \cap B}\}$  sont mutuellement indépendants sous  $\rho^T$ .

De telles hyperdistributions ont des propriétés intéressantes vis-à-vis de  $p(\cdot|T)$ .

**Proposition 1** (Dawid & Lauritzen, 1993). *Si  $\rho^T$  est fortement hyper Markov par rapport à  $T$ , alors  $p(\cdot|T)$  est Markov par rapport à  $T$ .*

Cela signifie que  $p(\mathbf{x}|T)$  peut être factorisée sur les arêtes de  $T$ .

$$p(\mathbf{x}|T) = \prod_{i \in V} p(\mathbf{x}_i|T) \prod_{\{i,j\} \in E_T} \frac{p(\mathbf{x}_i, \mathbf{x}_j|T)}{p(\mathbf{x}_i|T)p(\mathbf{x}_j|T)}, \quad \forall \mathbf{x} \in \mathcal{X}. \quad (5.2)$$

Il faut maintenant choisir toutes les hyperdistributions  $\rho^T$  de telle manière qu'elles soient hyper Markov par rapport à leur arbre respectif et que les intégrales locales utilisées en (5.2) ne dépendent pas de  $T$ .

**Proposition 5.1** (Dawid & Lauritzen, 1993). *Soit  $\rho$  une hyperdistribution générale sur  $\mathcal{F}$  telle que, pour tout  $A \subset V$ ,  $\pi_A \perp\!\!\!\perp \pi_{V \setminus A|A}$  sous  $\rho$ . Alors, pour tout arbre  $T \in \mathcal{T}$ , il existe une unique hyperdistribution  $\rho^T$  sur  $\mathcal{F}_T$  fortement hyper Markov par rapport à  $T$  et telle que, pour toute arête  $\{i, j\} \in E_T$ ,*

$$\rho_{ij}^T = \rho_{ij}. \quad (5.3)$$

$\{\rho^T\}_{T \in \mathcal{T}}$  est une famille compatible d'hyperdistributions.

On requiert en fait que l'hyperdistribution  $\rho$  servant de base à la famille compatible soit fortement hyper Markov par rapport au graphe complet. Si  $\{\rho^T\}_{T \in \mathcal{T}}$  est construite ainsi, la Proposition 5.1 garantit que  $\rho^T$  est fortement hyper Markov par rapport à  $T$  pour tout  $T \in \mathcal{T}$  et que les vraisemblances locales intégrées  $p(\mathbf{x}_i, \mathbf{x}_j|T)$  et  $p(\mathbf{x}_i|T)$  ne dépendent en fait pas de  $T$ . Ces termes peuvent être calculés une fois pour toutes et utilisés pour tous les arbres. Ainsi, on a

$$\xi(T|\mathbf{x}) = \frac{1}{Z(\omega)} \prod_{\{i,j\} \in E_T} \omega_{i,j}, \quad \forall T \in \mathcal{T},$$

$$\omega_{i,j} = \beta_{i,j} \frac{p(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_i)p(\mathbf{x}_j)}, \quad \forall \{i, j\} \in \mathcal{P}_2(V).$$

Dans la pratique, les intégrales locales permettant d'obtenir  $\omega$  peuvent être calculées simplement dans un cadre conjugué, mais peuvent aussi être obtenues par des méthodes d'intégration numérique ou stochastique. Trois cas particuliers sont détaillés, à savoir le cas où  $\mathbf{X}$  suit une distribution multinomiale, une distribution gaussienne ou une copule.

## Probabilité d'apparition d'arête

La probabilité d'apparition d'une arête  $\{k, l\}$  *a posteriori* est donnée par

$$P(\{k, l\} \in E_T|\mathbf{x}) = \sum_{\substack{T \in \mathcal{T}, \\ E_T \ni \{k, l\}}} \xi(T|\mathbf{x}) = 1 - \frac{1}{Z(\omega)} \sum_{\substack{T \in \mathcal{T}, \\ E_T \not\ni \{k, l\}}} \prod_{\{i,j\} \in E_T} \omega_{i,j}.$$

Une utilisation directe du théorème arbre-matrice fournit  $Z(\omega)$ . La seconde somme peut être calculée de manière similaire en appliquant le théorème à une matrice de poids  $\omega^{(kl)}$  obtenue à partir de  $\omega$  en mettant le poids  $\omega_{k,l} = \omega_{l,k}$  à zéro. Effectué de cette manière, le calcul de  $\{P(\{k, l\} \in E_T|\mathbf{x})\}_{\{k,l\} \in \mathcal{P}_2(V)}$  est de complexité  $O(p^5)$ . Ces probabilités peuvent en fait être calculées en temps cubique.

**Théorème** (Kirshner, 2007). *Soit  $u$  un sommet de  $V$ . On note  $\Delta^{\{u\}}$  la matrice obtenue à partir de la matrice laplacienne  $\Delta$  associée à  $\omega$  en enlevant la ligne et le colonne correspondant à  $u$  et  $Q = (\Delta^{\{u\}})^{-1}$ . Soit  $M$  la matrice dont le terme général est donné par*

$$M_{k,l} = \begin{cases} (Q_{k,k} + Q_{l,l} - 2Q_{k,l}) & \text{if } k \neq u, l \neq u, \\ Q_{k,k} & \text{if } l = u, \\ Q_{l,l} & \text{if } k = u. \end{cases}$$

Alors, pour toute arête  $\{k, l\} \in \mathcal{P}_2(V)$ ,  $P(\{k, l\} \in E_T|\mathbf{x}) = \omega_{k,l} \cdot M_{k,l}$ .

Ainsi, le calcul de  $\{P(\{k, l\} \in E_T|\mathbf{x})\}_{\{k,l\} \in \mathcal{P}_2(V)}$  revient au calcul de l'inverse d'une matrice de taille  $p - 1$ , de complexité  $O(p^3)$ .

## Chapitre 3

La question à laquelle se propose de répondre ce chapitre est celle de la détection de brusques changements, aussi appelés ruptures, dans la structure de dépendance d'une série temporelle multivariée. On parle aussi de problème de segmentation. En se fondant sur le cadre développé au chapitre précédent, le modèle graphique décrivant la structure de dépendance à un instant donné est supposé arborescent. En tirant partie de cette hypothèse et en utilisant le résultat de [Rigaill et al. \(2012\)](#) énoncé au Chapitre 1, nous montrons qu'il est possible de calculer un certain nombre de quantités d'intérêt vis-a-vis du problème posé, telles que la probabilité *a posteriori* d'observer un point de rupture à un temps donné, ou encore la loi *a priori* du nombre de segments, en un temps polynomial par rapport à la longueur de la série et au nombre de variables. Lorsque la segmentation est donnée, nous fournissons également un moyen d'évaluer la probabilité *a posteriori* qu'une arête (ou que le graphe dans son intégralité) conserve le même statut au cours du temps.

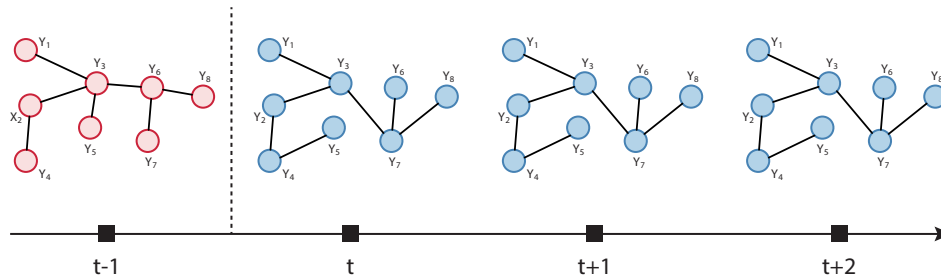
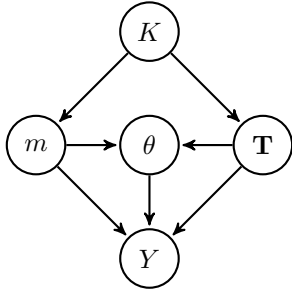


Figure 5.1 – Illustration du problème de détection de ruptures dans la structure d'un modèle graphique.

### Modèle

On suppose que les observations  $\{y^t\}_{t=1}^N$  sont une réalisation d'un processus aléatoire multivarié  $\{Y^t\}_{t=1}^N$  de dimension  $p \geq 2$ . Pour  $1 \leq t \leq N$ ,  $Y^t = (Y_1^t, \dots, Y_p^t)$  est un vecteur aléatoire à valeurs dans un espace produit  $\mathcal{X} = \bigotimes_{i=1}^p \mathcal{X}_i$ . Pour tout intervalle temporel  $r \subseteq \llbracket 1; N \rrbracket$ ,  $Y^r := \{Y^t\}_{t \in r}$  dénote le processus restreint à  $r$ . Le modèle suppose qu'il existe une partition  $m$  de  $\llbracket 1; N \rrbracket$  en intervalles  $r$  telle que, sur chaque intervalle  $r$  de  $m$ ,  $Y^r$  est indépendant et identiquement distribué selon un modèle graphique à structure d'arbre. Ce modèle appartient à la famille des modèles à partition décrite par [Barry & Hartigan \(1992\)](#). Si  $m$  possède  $K$  segments  $r_1, \dots, r_K$ , on note respectivement  $\mathbf{T} = \{T_k\}_{k=1}^K$  et  $\theta = \{\theta_r\}_{r \in m}$  les arbres et paramètres donnant le modèle graphique de chaque segment. Pour  $r \in m$ ,  $\kappa(r|m)$  indique la position du segment  $r$  dans la segmentation  $m$ . Les observations  $\{Y^t\}_{t \in r}$  sont indépendantes, identiquement distribuées et suivent un modèle graphique de structure  $T_{\kappa(r|m)}$  et de paramètres  $\theta_r$ . La structure et les paramètres relatifs à chaque segment sont également indépendants et identiquement distribués, selon un modèle de la forme décrite dans le Chapitre 2. La distribution de chacun des arbres de  $\mathbf{T}$  est donnée à travers une matrice de poids d'arêtes  $b$ . La distribution *a priori* sur  $m$  est choisie comme factorisant sur les segments, conformément à l'hypothèse de factorisation requise par [Rigaill et al. \(2012\)](#). Elle est donnée par une matrice triangulaire supérieure  $a$  de taille  $N + 1$ , dans laquelle  $a_{s,t}$  donne le poids du segment  $\llbracket s; t \rrbracket$ . Une description complète du modèle est donnée en Figure 5.2.



$$\begin{aligned}
p(m|K) &= \frac{1}{C_K(a)} \prod_{r \in m} a_r, \\
p(\mathbf{T}|K) &= \prod_{k=1}^K p(T_k) = \frac{1}{Z(b)^K} \prod_{k=1}^K \prod_{\{i,j\} \in E_{T_k}} b_{ij}, \\
p(\theta|m, \mathbf{T}) &= \prod_{r \in m} p(\theta_r | T_{\kappa(r|m)}), \\
p(y|m, \theta, \mathbf{T}) &= \prod_{r \in m} \prod_{t \in r} p(y^t | T_{\kappa(r|m)}, \theta_r).
\end{aligned}$$

Figure 5.2 – Modèle de détection de rupture.

## Inférence

Dans ce modèle, la vraisemblance marginale des observations, conditionnellement à  $K$ , est donnée par

$$p(y|K) = \sum_{m \in \mathcal{M}_K} \sum_{\mathbf{T} \in \mathcal{T}^K} \int p(y, m, \theta, \mathbf{T}|K) d\theta. \quad (5.4)$$

L'intégration sur les paramètres discrets nécessite de sommer sur un ensemble de taille  $|\mathcal{M}_K| \cdot |\mathcal{T}^K| = \binom{N-1}{K-1} \cdot p^{K(p-2)} \approx \left(\frac{Np^{p-2}}{K}\right)^K$ . Ce calcul peut cependant être effectué en un temps polynomial en utilisant le théorème de sommation sur les segmentations énoncé au Chapitre 1 sur une matrice de terme général

$$A_{s,t} = \begin{cases} a_{s,t} \cdot p(y^{\llbracket s,t \rrbracket}) & \text{si } s < t, \\ 0 & \text{sinon,} \end{cases}$$

où  $p(y^{\llbracket s,t \rrbracket})$  est la vraisemblance du segment  $\llbracket s,t \rrbracket$ , intégrée sur l'arbre et les paramètres. Ces vraisemblances sont obtenues grâce au théorème arbre-matrice. Une matrice de poids d'arêtes *a posteriori*  $\omega^{\llbracket s,t \rrbracket}$  est calculée par segment. On a alors

$$p(y^{\llbracket s,t \rrbracket}) = \frac{Z(\omega^{\llbracket s,t \rrbracket})}{Z(b)} \cdot \prod_{i \in V} p(y_i^{\llbracket s,t \rrbracket}).$$

Le résultat de Rigauil et al. (2012) indique alors que la vraisemblance marginale donnée en (5.4) est donnée par

$$p(y|K) = \frac{[A^K]_{1,N+1}}{C_K(a)},$$

la constante de normalisation  $C_K(a)$  étant elle-même égale à  $[a^K]_{1,N+1}$ . Cette vraisemblance marginale permet en particulier de calculer la loi *a posteriori* du nombre de segments  $K$ , puisque  $p(K|y) \propto p(K)p(y|K)$ . Il est ainsi possible d'obtenir un estimateur du maximum *a posteriori* pour  $K$ .

D'autres quantités peuvent être calculées à partir des puissances de la matrice  $A$ . La probabilité *a posteriori*  $B_{K,k}(t)$  que le  $k$ -ème des  $K-1$  points de rupture d'une segmentation

à  $K$  segments intervienne à l'instant  $t$  peut par exemple être exprimée comme

$$B_{K,k}(t) = \frac{[A^k]_{1,t}[A^{K-k}]_{t,N+1}}{[A^K]_{1,N+1}}.$$

La probabilité qu'un segment donné apparaisse dans  $m$  peut être obtenue de manière similaire. Une formule exacte est également donnée pour la probabilité d'apparition *a posteriori* au cours du temps.

Il n'est en revanche pas possible de calculer de la même manière la probabilité qu'une arête conserve le même statut tout au long de la série temporelle lorsque  $m$  est aléatoire. En effet, ce calcul nécessite d'intégrer sur des sous-ensembles de segmentations en contradiction directe avec l'hypothèse de factorisation sur les segments. Ce calcul devient néanmoins possible dès que la segmentation est fixée.

## Simulations et applications

Une étude simulatoire a été effectuée afin d'étudier le comportement de notre méthode, notamment lorsque l'hypothèse arborescente n'est pas vérifiée par les graphes servant à générer les données. Nous nous sommes placés dans le cadre classique des modèles graphiques gaussiens. Nous avons comparé notre modèle à celui obtenu en n'imposant aucune structure sur la matrice de précision. Les résultats semblent indiquer que l'hypothèse arborescente pénalise très peu notre approche en termes de segmentation quand la densité des réseaux reste faible. Dans tous les cas, l'inférence semble plus stable dans le modèle que nous avons décrit, en comparaison avec le modèle non-structuré, tout en permettant l'inférence de la structure.

Nous avons également appliqué notre approche à des données d'expression de gènes récupérées au cours du cycle de vie de la drosophile (Arbeitman et al., 2002) et concernant onze gènes impliqués dans le développement des muscles des ailes. Les résultats obtenus semblent cohérents avec les différents stades de la morphogénèse observés chez la drosophile. Une seconde application à des données d'imagerie par résonance magnétique fonctionnelle (Cribben et al., 2012) est également présentée.

## Chapitre 4

Dans ce chapitre, nous tentons de dégager les conditions génériques sous lesquelles les techniques d'inférence bayésiennes exactes présentées dans ce manuscrit peuvent s'appliquer. Nous présentons ensuite diverses extensions et perspectives aux travaux des deux chapitres précédents. Celles-ci se distinguent en trois parties.

La première section s'intéresse à l'intégration de covariables dans les problèmes d'inférence de réseaux et de segmentation. La motivation vient d'un jeu de données issu de l'écologie microbienne où l'abondance d'un certain nombre d'espèces a été mesurée sur des feuilles d'arbres. Ces données sont accompagnées de plusieurs covariables, telles que l'arbre dont proviennent les feuilles, ou leur position dans la canopée. Il semble pertinent d'intégrer cette information à la procédure d'inférence de réseaux, dans la mesure où ce sont généralement les interactions directes entre espèces que les écologues cherchent à mettre en évidence. Dans le cas où les observations peuvent être modélisées par une loi normale multivariée, l'intégration



de covariables peut être effectuée à peu de frais. Nous montrons par exemple que le cadre de la régression linéaire multiple multivariée permet de conserver les propriétés de Markov nécessaires au bon fonctionnement des résultats algébriques utilisés dans l'inférence.

L'exemple ayant motivé notre intérêt pour la question des covariables concernait des données de comptage. Le passage au modèle linéaire généralisé n'est dans ce cas pas sans frais, et empêche de poursuivre dans la même direction. C'est pourquoi nous proposons une approche pragmatique fondée sur les copules. L'abondance de chacune des espèces est modélisée individuellement par un modèle linéaire généralisé utilisant les covariables comme variables explicatives. Les paramètres de ces modèles sont ajustés de manière classique. Les résidus de Pearson correspondant à chaque espèce sont calculés et ramenés à l'intervalle  $[0; 1]$  grâce à leur fonction de répartition empirique. L'inférence du réseau sous-jacent est ensuite effectuée sur ces fractiles en utilisant des distributions définies sur  $[0; 1]^p$  dont les marginales sont uniformes, aussi appelées copules. Cette approche, bien qu'ayant le mérite d'être entièrement générique, n'est bien évidemment qu'un pis-aller, dans la mesure où l'incertitude sur les paramètres de régression n'est pas du tout prise en compte.

Dans la seconde section, nous expliquons comment introduire de la dépendance temporelle dans le modèle de segmentation décrit au Chapitre 3. En effet, l'hypothèse d'indépendance faite dans ce modèle est souvent peu vraisemblable en pratique. En ce qui concerne l'intégration sur l'ensemble des segmentations, la seule indépendance nécessaire est celle des observations entre les segments, afin que la loi a posteriori sur  $m$  factorise sur les segments. De ce point de vue là, rien n'empêche d'introduire de la dépendance temporelle au sein des segments. Notre suggestion est d'utiliser les modèles d'indépendance temporelle (abrégé par TIM, d'après l'acronyme anglais correspondant) introduit par Siracusa (2009) afin de gérer la dépendance temporelle intra-segment. Ce sont des modèles graphiques dirigés, dans lesquels la dépendance d'un instant au suivant est décrite par un graphe dirigé. Si  $\text{pa}(i)$  désigne l'ensemble des parents du sommet  $i$  dans le graphe décrivant la dépendance, alors  $Y_i^t$  dépend de  $Y_{\text{pa}(i)}^{t-l}, \dots, Y_{\text{pa}(i)}^{t-1}$ , où  $l$  est la latence du modèle. Un exemple est donné Figure 5.3. Si l'on suppose que la structure de dépendance est un arbre, il est en fait possible d'utiliser une version dirigée du théorème arbre-matrice pour effectuer l'inférence. Les hypothèse d'indépendance sur les paramètres doivent également être adaptées à cette nouvelle configuration.

Enfin, la dernière partie se penche un peu plus sur la question de la distribution *a priori* sur la segmentation  $m$  du modèle donné Figure 5.2. Une liste non-exhaustive de contraintes pouvant être encodées dans la matrice  $a$  des poids de segments est donnée. Il est par exemple possible d'interdire les segments de longueur inférieure à une certaine taille, ou encore de

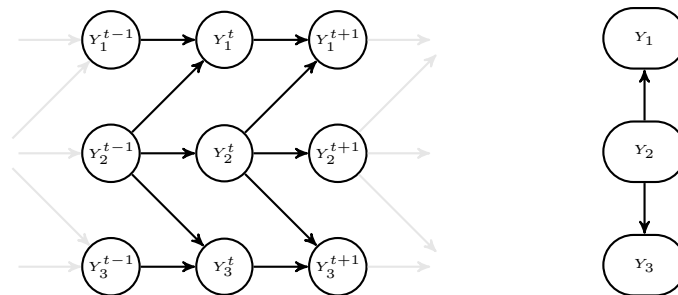


Figure 5.3 – Représentation graphique d'un modèle d'indépendance temporelle de latence 1 à structure de dépendance arborescente.

favoriser les segmentations dont les segments sont de taille homogène. Nous décrivons également une autre forme de distributions *a priori* pour laquelle quelques ajustements sont nécessaires dans l'inférence. Ces distributions permettent de facilement intégrer de la connaissance provenant d'un autre jeu de données dans l'analyse.



## Published Articles and Preprints

- L. Schwaller and S. Robin. Exact bayesian inference for off-line change-point detection in tree-structured graphical models. *Statistics and Computing*, pages 1–15, 2016. ISSN 1573-1375. doi: 10.1007/s11222-016-9689-3
- B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen erysiphe alphi-toides. *Microbial Ecology*, pages 1–11, 2016. ISSN 1432-184X. doi: 10.1007/s00248-016-0777-x. URL <http://dx.doi.org/10.1007/s00248-016-0777-x>
- C. Vacher, A. Tamaddoni-Nezhad, S. Kamenova, N. Peyrard, Y. Moalic, R. Sabbadin, L. Schwaller, J. Chiquet, M. A. Smith, J. Vallance, V. Fievet, B. Jakuschkin, and D. A. Bohan. Chapter One - Learning Ecological Networks from Next-Generation Sequencing Data. In G. Woodward and D. A. Bohan, editors, *Ecosystem Services: From Biodiversity to Society, Part 2*, volume 54 of *Advances in Ecological Research*, pages 1 – 39. Academic Press, 2016. doi: <http://dx.doi.org/10.1016/bs.aecr.2015.10.004>. URL <http://www.sciencedirect.com/science/article/pii/S0065250415000331>
- L. Schwaller, S. Robin, and M. Stumpf. Bayesian Inference of Graphical Model Structures Using Trees. *ArXiv e-prints*, 2015
- L. Schwaller and S. Robin. Apprentissage de réseaux par agrégation bayésienne d’arbres couvrants. *Revue d’Intelligence Artificielle*, 29(2):153–172, 2015. doi: 10.3166/ria.29.153-172. URL <http://dx.doi.org/10.3166/ria.29.153-172>

# Bibliography

- R. Almond and A. Kong. Optimality issues in constructing a Markov tree from graphical models. *Journal of Computational and Graphical Statistics*, 1993.
- M. N. Arbeitman, E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. a. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. Gene expression during the life cycle of *Drosophila melanogaster*. *Science (New York, N. Y.)*, 297(5590):2270–2275, 2002. ISSN 1095-9203. doi: 10.1126/science.1072152.
- A. Atay-Kayis and H. Massam. A Monte Carlo method to compute the marginal likelihood in non decomposable graphical Gaussian models. *Biometrika*, 92:317–335, 2005.
- I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989. ISSN 00928240. doi: 10.1007/BF02458835.
- D. Barber and A. Cemgil. Graphical Models for Time-Series. *IEEE Signal Processing Magazine*, (January), 2010. ISSN 1053-5888. doi: 10.1109/MSP.2010.938028.
- D. Barry and J. A. Hartigan. Product Partition Models for Change Point Problems. *The Annals of Statistics*, 20(1):260–279, 1992. ISSN 0090-5364. doi: 10.1214/aos/1176348521.
- J. G. Broida and S. G. Williamson. *A Comprehensive Introduction To Linear Algebra*. Addison-Wesley, Redwood City, Calif., 1989. ISBN 9780201500653. URL <http://isbnplus.org/9780201500653>.
- L. Burger and E. Van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology*, 6(1), 2010. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000633.
- S. Byrne. *Hyper and Structural Markov Laws for Graphical Models*. PhD thesis, 2011.
- S. Byrne and A. P. Dawid. Structural markov graph laws for bayesian model uncertainty. *Ann. Statist.*, 43(4):1647–1681, 08 2015. doi: 10.1214/15-AOS1319. URL <http://dx.doi.org/10.1214/15-AOS1319>.
- F. Caron, A. Doucet, and R. Gottardo. On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2):579–595, 2012. ISSN 09603174. doi: 10.1007/s11222-011-9248-x.
- A. Cayley. A theorem on trees. *Quarterly Journal of Mathematics*, 23:376–378, 1889.
- S. Chaiken. A Combinatorial Proof of the All Minors Matrix Tree Theorem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):319–329, 1982.

- D. M. Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554, 2002. ISSN 15324435. doi: 10.1162/153244303321897717. URL [http://dl.acm.org/citation.cfm?id=944933&delimiter="026E30F\\$npapers2://publication/uuid/0A738057-E5F9-40E5-A716-58D8BCDCCC8F](http://dl.acm.org/citation.cfm?id=944933&delimiter=).
- C. Chow and C. Liu. Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.
- I. Cribben, R. Haraldsdottir, L. Y. Atlas, T. D. Wager, and M. a. Lindquist. Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage*, 61(4):907–920, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.03.070.
- P. Dawid and S. Lauritzen. Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, 21(3):1272–1317, 1993.
- F. Dondelinger, S. Lèbre, and D. Husmeier. Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Machine Learning*, 90(2):191–230, 2013. ISSN 08856125. doi: 10.1007/s10994-012-5311-x.
- P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006. ISSN 09603174. doi: 10.1007/s11222-006-8450-8.
- P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(4):589–605, 2007. ISSN 13697412. doi: 10.1111/j.1467-9868.2007.00601.x.
- E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Nonparametric Bayesian learning of switching linear dynamical systems. *Advances in neural information processing systems*, 21:457–464, 2009.
- J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. ISSN 14654644. doi: 10.1093/biostatistics/kxm045.
- N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003. ISSN 08856125. doi: 10.1023/A:1020249912095.
- N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 139–147, 1998. doi: 10.1111/j.1469-7580.2008.00962.x.
- M. Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistics*, pages 333–353, 1990.
- D. Geiger and D. Heckerman. A Characterization of the Dirichlet Distribution Through Global and Local Parameter Independence. *The Annals of Statistics*, pages 1344–1369, 1997.
- D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440, 2002.

- P. J. Green and A. Thomas. Sampling decomposable graphs using a markov chain on junction trees. *Biometrika*, 100(1):91–110, 2013. doi: 10.1093/biomet/ass052.
- M. Grzegorzczak and D. Husmeier. Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. *Bioinformatics*, 27(5):693–699, 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq711.
- J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. 1971.
- D. Heckerman and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pages 20–197, 1995.
- B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen *erysiphe alphitoides*. *Microbial Ecology*, pages 1–11, 2016. ISSN 1432-184X. doi: 10.1007/s00248-016-0777-x. URL <http://dx.doi.org/10.1007/s00248-016-0777-x>.
- F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Publishing Company, Incorporated, 2nd edition, 2007. ISBN 9780387682815.
- G. Kirchhoff. Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. *Annalen der Physik*, 148(12): 497–508, 1847. ISSN 1521-3889. doi: 10.1002/andp.18471481202.
- S. Kirshner. Learning with Tree-Averaged Densities and Distributions. *Advances in Neural Information Processing Systems 2008*, 20:761–768, 2007.
- M. Kolar, L. Song, A. Ahmed, and E. P. Xing. *Estimating time-varying networks*, volume 4. 2010. ISBN 0001409107. doi: 10.1214/09-AOAS308.
- T. Koo, a. Globerson, X. Carreras, and M. Collins. Structured prediction models via the matrix-tree theorem. *Proc. of EMNLP-CoNLL*, pages 141–150, 2007. URL [http://acl.ldc.upenn.edu/D/D07/D07-1015.pdf%\\$delimitter"026E30F\\$npapers2://publication/uuid/507BF5D8-B9DF-4CE2-B3E4-1EA6DCCE27A3](http://acl.ldc.upenn.edu/D/D07/D07-1015.pdf%$delimitter).
- J. Kuipers, G. Moffa, and D. Heckerman. Addendum on the scoring of gaussian directed acyclic graphical models. *Ann. Statist.*, 42(4):1689–1691, 08 2014. doi: 10.1214/14-AOS1217. URL <http://dx.doi.org/10.1214/14-AOS1217>.
- S. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.
- S. Lèbre, J. Becq, F. Devaux, M. P. H. Stumpf, and G. Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC systems biology*, 4:130, 2010. ISSN 1752-0509. doi: 10.1186/1752-0509-4-130.
- Y. Lin, S. Zhu, D. D. Leet, and B. Taskar. Learning Sparse Markov Network Structure via Ensemble-of-Trees Models. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009*, volume 5, pages 360–367, 2009.
- D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232, 1995. ISSN 03067734. doi: 10.2307/1403615.
- H. Massam and E. Neher. On Transformations and Determinants of Wishart Variables on Symmetric Cones 1. *Journal of Theoretical Probability*, 10(4):867–902, 1997.

- M. Meilä. *Learning with Mixtures of Trees*. PhD thesis, Massachusetts Institute of Technology, 1999.
- M. Meilä and T. Jaakkola. Tractable bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92, 2006.
- M. Meilä and M. I. Jordan. Learning with Mixtures of Trees. *The Journal of Machine Learning Research*, 1:1–48, 2001.
- N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006. ISSN 00905364. doi: 10.1214/009053606000000281.
- J. Moussouris. Gibbs and Markov Random Systems with Constraints. *Journal of statistical physics*, 10(1):11–33, 1974. ISSN 0022-4715. doi: 10.1007/BF01011714.
- K. Murphy and S. Mian. Modelling gene expression data using dynamic bayesian networks. Technical report, 1999.
- R. B. Nelsen. *An Introduction to Copulas (Springer series in statistics)*. 2006. ISBN 0387286594. doi: 10.1080/00401706.2000.10486066.
- T. Niinimäki, P. Parviainen, and M. Koivisto. Partial order mcmc for structure discovery in bayesian networks. In F. G. Cozman and A. Pfeffer, editors, *UAI*, pages 557–564. AUAI Press, 2011. ISBN 978-0-9749039-7-2.
- P. Parviainen and M. Koivisto. Exact Structure Discovery in Bayesian Networks with Less Space. *Uai*, pages 436–443, 2009. ISSN 15324435.
- J. Pearl and A. Paz. Graphoids: a graph-based logic for reasoning about relevance relations. *Artificial Intelligence II*, pages 357–63, 1987.
- G. Rigaiil, E. Lebarbier, and S. Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4): 917–929, 2012. ISSN 0960-3174. doi: 10.1007/s11222-011-9258-8.
- J. Robinson and A. Hartemink. Learning non-stationary dynamic Bayesian networks. *The Journal of Machine Learning ...*, 11:3647–3680, 2010. ISSN 1532-4435.
- A. Roverato. Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.
- K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science (New York, N.Y.)*, 308: 523–529, 2005. ISSN 0036-8075. doi: 10.1126/science.1105809.
- L. Schwaller and S. Robin. Apprentissage de réseaux par agrégation bayésienne d’arbres couvrants. *Revue d’Intelligence Artificielle*, 29(2):153–172, 2015. doi: 10.3166/ria.29.153-172. URL <http://dx.doi.org/10.3166/ria.29.153-172>.
- L. Schwaller and S. Robin. Exact bayesian inference for off-line change-point detection in tree-structured graphical models. *Statistics and Computing*, pages 1–15, 2016. ISSN 1573-1375. doi: 10.1007/s11222-016-9689-3.



- L. Schwaller, S. Robin, and M. Stumpf. Bayesian Inference of Graphical Model Structures Using Trees. *ArXiv e-prints*, 2015.
- M. R. Siracusa. Tractable Bayesian Inference of Time-Series Dependence Structure. *Proceedings of the 23th International Conference on Artificial Intelligence and Statistics*, 5: 528–535, 2009.
- D. Spiegelhalter and S. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990. ISSN 00283045. doi: 10.1002/net.3230200507.
- R. Tibshirani. Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(3):273–282, 2011. ISSN 13697412. doi: 10.1111/j.1467-9868.2011.00771.x.
- C. Vacher, A. Tamaddoni-Nezhad, S. Kamenova, N. Peyrard, Y. Moalic, R. Sabbadin, L. Schwaller, J. Chiquet, M. A. Smith, J. Vallance, V. Fievet, B. Jakuschkin, and D. A. Bohan. Chapter One - Learning Ecological Networks from Next-Generation Sequencing Data. In G. Woodward and D. A. Bohan, editors, *Ecosystem Services: From Biodiversity to Society, Part 2*, volume 54 of *Advances in Ecological Research*, pages 1 – 39. Academic Press, 2016. doi: <http://dx.doi.org/10.1016/bs.aecr.2015.10.004>. URL <http://www.sciencedirect.com/science/article/pii/S0065250415000331>.
- A. V. Werhli, M. Grzegorzcyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics (Oxford, England)*, 22(20):2523–31, Oct. 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl391.
- X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. *Proceedings of the 24th International Conference on Machine Learning (2007)*, 227 (m):1055–1062, 2007. ISSN 1595937935. doi: 10.1145/1273496.1273629.
- J. C. York and D. Madigan. Bayesian methods for estimating the size of a closed population. Technical Report 234, 1992.
- W. Zhao, E. Serpedin, and E. R. Dougherty. Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, 22(17): 2129–2135, 2006. ISSN 13674803. doi: 10.1093/bioinformatics/btl364.
- S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Machine Learning*, 80:295–319, 2010.



## Titre : Inférence bayésienne exacte dans les modèles graphiques : inférence de réseaux à structure arborescente et segmentation

**Mots clefs :** arbre couvrant, inférence bayésienne, inférence de réseaux, modèles graphiques, segmentation, théorème arbre-matrice.

**Résumé :** Cette thèse porte sur l'inférence de réseaux. Le cadre statistique naturel à ce genre de problèmes est celui des modèles graphiques, dans lesquels les relations de dépendance et d'indépendance conditionnelles vérifiées par une distribution multivariée sont représentées à l'aide d'un graphe. Il s'agit alors d'apprendre la structure du modèle à partir d'observations portant sur les sommets. Nous considérons le problème d'un point de vue bayésien. Nous avons également décidé de nous concentrer sur un sous-ensemble de graphes permettant d'effectuer l'inférence de manière exacte et efficace, à savoir celui des arbres couvrants. Il est en effet possible d'intégrer une fonction définie sur les arbres couvrants en un temps cubique par rapport au nombre de variables à la condition que cette fonction factorise selon les arêtes, et ce malgré le cardinal super-exponentiel de cet ensemble. En choisissant les distributions *a priori* sur la structure et les paramètres du modèle de ma-

nière appropriée, il est possible de tirer parti de ce résultat pour l'inférence de modèles graphiques arborescents. Nous proposons un cadre formel complet pour cette approche.

Nous nous intéressons également au cas où les observations sont organisées en série temporelle. En faisant l'hypothèse que la structure du modèle graphique latent subit un certain nombre de brusques changements, le but est alors de retrouver le nombre et la position de ces points de rupture. Il s'agit donc d'un problème de segmentation. Sous certaines hypothèses de factorisation, l'exploration exhaustive de l'ensemble des segmentations est permise et, combinée aux résultats sur les arbres couvrants, permet d'obtenir, entre autres, la distribution *a posteriori* des points de ruptures en un temps polynomial à la fois par rapport au nombre de variables et à la longueur de la série.

## Title: Exact Bayesian Inference in Graphical Models: Tree-structured Network Inference and Segmentation

**Keywords:** Bayesian inference, graphical models, Matrix-Tree theorem, network inference, segmentation, spanning tree.

**Abstract:** In this dissertation we investigate the problem of network inference. The statistical framework tailored to this task is that of graphical models, in which the (in)dependence relationships satisfied by a multivariate distribution are represented through a graph. We consider the problem from a Bayesian perspective and focus on a subset of graphs making structure inference possible in an exact and efficient manner, namely spanning trees. Indeed, the integration of a function defined on spanning trees can be performed with cubic complexity with respect to number of variables under some factorisation assumption on the edges, in spite of the super-exponential cardinality of this set. A careful choice of prior distributions on both graphs and distribution parameters allows to use this result for network

inference in tree-structured graphical models, for which we provide a complete and formal framework.

We also consider the situation in which observations are organised in a multivariate time-series. We assume that the underlying graph describing the dependence structure of the distribution is affected by an unknown number of abrupt changes throughout time. Our goal is then to retrieve the number and locations of these change-points, therefore dealing with a segmentation problem. Using spanning trees and assuming that segments are independent from one another, we show that this can be achieved with polynomial complexity with respect to both the number of variables and the length of the series.