SOCIETE FRANÇAISE DE STATISTIQUE

Journal de la Société Française de Statistique

Vol. 160 No. 2 (2019) arXiv: 1504.02723

Closed-form Bayesian inference of graphical model structures by averaging over trees

Loïc Schwaller[1] , Stéphane Robin[2] and Michael Stumpf[3]

Abstract: We consider the inference of the structure of an undirected graphical model in a Bayesian framework. To avoid convergence issues and highly demanding Monte Carlo sampling, we focus on exact inference. More specifically we aim at achieving the inference with closed-form posteriors, avoiding any sampling step. To this aim, we restrict the set of considered graphs to mixtures of spanning trees. We investigate under which conditions on the priors – on both tree structures and parameters – closed-form Bayesian inference can be achieved. Under these conditions, we derive a fast an exact algorithm to compute the posterior probability for an edge to belong to the tree model using an algebraic result called the Matrix-Tree theorem. We show that the assumption we have made does not prevent our approach to perform well on synthetic and flow cytometry data.

Résumé : Nous nous intéressons à l'inférence de la structure d'un modèle graphique non orienté dans une cadre bayésien. Pour éviter de recourir à des méthodes de Monte-Carlo coûteuses et aux problèmes de convergence associés, nous nous concentrons sur des méthodes exactes. Plus précisément, nous menons l'inférence au moyen de lois a posteriori explicites, évitant ainsi toute étape d'échantillonnage. Dans ce but, nous restreignons l'espace des graphes à des mélanges d'arbres recouvrants. Nous étudions sous quelles condition sur les lois a priori – à la fois sur les arbres et sur les paramètres – une inférence bayésienne exacte peut être menée. Dans ce cadre, nous proposons un algorithme exact et rapide permettant de calculer la probabilité a posteriori pour qu'une arête appartienne au graphe, en utilisant un résultat algébrique connu sous le nom de théorème Arbre-Matrice. Nous montrons que la restriction aux arbres n'empêche pas d'obtenir de bons résultats aussi bien sur des données simulées que sur des données issues de cytométrie de flux.

Keywords: graphical models, hyper Markov, matrix-tree theorem, spanning trees
Mots-clés : arbres recouvrants, hyper-Markov, modèles graphiques, théorème arbre-matrice

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, among others. The Bayesian framework is a convenient way to perform the inference of the structure while taking into account the uncertainty on the parameters. This requires to define a full model and, more specifically,

[1] Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands
E-mail: loic.schwaller@ens-lyon.org

[2] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France
E-mail: stephane.robin@agroparistech.fr

[3] Centre for Integrative Systems Biology and Bioinformatics, Imperial College London, London, United Kingdom
E-mail: m.stumpf@imperial.ac.uk

a prior distribution on graphical models (therefore on graphs themselves) as well as a prior on parameters. For the inference to remain efficient, these priors need to combine nicely.

Regardless of whether we consider directed or undirected graphs, their sheer number make them difficult to deal with. Limiting ourselves to Bayesian approaches, 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 and Koller, 2003; Niinimäki et al., 2016) or decomposable graphs (Green and Thomas, 2013). The decomposability assumption for undirected graphical models, also called Markov random fields, is commonly made , although some interest has been devoted to the less easy to handle non-decomposable graphs (Roverato, 2002; Atay-Kayis and 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. This motivates our choice to focus on closed-form Bayesian inference whenever possible.

In this paper, we refer to closed-form Bayesian inference, as Bayesian inference that does not rely on a sample from the posterior distribution. For classical parameters, the usual way is to resort to conjugate priors. Theoretically, closed-form posterior distributions on graphs can be computed, but the combinatorial complexity becomes prohibitive as soon as there are more than thirty or so variables of interest (Parviainen and Koivisto, 2009). For larger problems, closed-form approaches can be considered at the price of a restriction on the structure space. When a subset of graphs is considered, it becomes possible to get access to the full posterior distribution on graphs, provided that the integration over the whole space of graphs can be achieved with a reasonable computational burden. In that perspective, trees have been of particular interest as a subset of both decomposable graphs and DAGs (Chow and Liu, 1968; Meilă and Jordan, 2001; Meilă and Jaakkola, 2006; Kirshner, 2007; Lin et al., 2009; Burger and Van Nimwegen, 2010).

In this paper, we consider tree-based structure inference and we discuss under which conditions closed-form Bayesian inference can be achieved. Broadly speaking, we are interested in the posterior distribution of the unknown tree structure T given the data D while integrating over parameter distribution π:

$$p(T|D) = \frac{p(T,D)}{p(D)} = \frac{p(T)\int p(\pi|T)p(D|\pi)d\pi}{\sum_T p(T)\int p(\pi|T)p(D|\pi)d\pi}.$$

Our first contribution is to provide a well-defined fully Bayesian framework for graphical model inference based on trees in which fast and efficient inference is possible. We use the work of Dawid and Lauritzen (1993) on hyper Markov laws to define priors on distribution parameter that can easily be marginalised over to get $p(D|T) = \int p(D|\pi)p(\pi|T)d\pi$. This framework spares us from requiring likelihood equivalence between Markov-equivalent directed tree models, like Meilă and Jaakkola (2006) did building on the work of Heckerman and Chickering (1995). We also point out that it fits within the recent work of Byrne and 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 (Meilă and Jaakkola, 2006) and Gaussian distributions.

Our second contribution focuses on structure inference as opposed to Meilă and Jaakkola (2006) and Kirshner (2007) who were more interested in the joint distribution of the observations.

More specifically, we focus on local features such as the posterior mean and variance of the degree of a given node or the posterior probability for two given nodes to be connected. The latter amounts to compute

$$P(\{k,l\} \in T|D) = \frac{\sum_{T \ni \{k,l\}} p(T)p(D|T)}{\sum_T p(T)p(D|T)}$$

where the sum in the numerator is restricted to the trees containing an edge between nodes k and l. Most works on tree-structured graphical model inference rely on an algebraic result called the Matrix-Tree theorem that allow to compute the sums above as determinants. 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. We also derive a closed-form expression for the entropy of the posterior distribution on trees.

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. Indeed, the 'true' graph is unlikely to be a spanning tree, so computing a maximum a posteriori (MAP) estimate of the whole graph would for instance yield a systematically wrong answer. However, our approach is not designed to assess the global structure all at once but to separately assess a collection of local features of the graph (typically, edges). The rationale is that the inference of such features is weakly affected by the restriction to spanning tree. In the simulation study, we demonstrate that, as long as edge inference is concerned, the tree-based approach provides similar results as this obtained when considering a larger class of graphs, but with a dramatic reduction of the computational time.

An R-language package **saturnin** implementing the approach presented here is available from the Comprehensive R Archive Network at `https://cran.r-project.org/web/packages/saturnin/`.

In Section 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 3. Section 4 deals with the inference of the model. Integrations with respect to distributions and structures are respectively discussed in Sections 4.1 and 4.2. The simulation study and its results are described in Section 5. An application to flow cytometry data is presented in Section 6.

2. Background & model

2.1. *Markov properties & graphical models*

Let $V = \{1,...,p\}$ and let $\mathbf{X} = (X_1,...,X_p)$ be a random vector indexed by V and taking values in a product space $\mathscr{X} = \bigotimes_{i=1}^{p} \mathscr{X}_i$. We let $\mathscr{F}$ denote the set of distributions on $\mathscr{X}$. For any subset A of V, $\mathbf{X}_A$ stands for the subvector of $\mathbf{X}$ indexed by A. We also let $\mathscr{P}_2(V)$ denote the subsets of V of size 2. For $E \subseteq \mathscr{P}_2(V)$, $G = (V,E)$ is the undirected graph with vertices V and edges E. In the following, the notation of Dawid and 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 $V = A \cup B$, if the subgraph induced by G on $A \cap B$ is complete (i.e. fully connected) and if $A \cap B$ separates A from B (i.e. any path from A to B goes through $A \cap B$). When 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. For graphs, the decomposability property is equivalent to the chordality property (see Lauritzen, 1996).

Definition 1. *A distribution $\pi \in \mathscr{F}$ is said to be Markov with respect to (w.r.t.) a decomposable graph G if, for any decomposition (A,B) of G, it holds that $\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_{A\cap B}$ under π.*

Proposition 1. *(Hammersley and Clifford, 1971) Let $\pi \in \mathscr{F}$. If π is a positive distribution (for all $\mathbf{x} \in \mathscr{X}$, $\pi(\mathbf{x}) > 0$), being Markov w.r.t. a decomposable graph G is equivalent to the existence of a factorisation of π on the (maximal) cliques of G.*

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, denoting π_i and π_{ij}, respectively, for $i \neq j$ both in V:

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

(see e.g. Meilă and Jaakkola, 2006). Such distributions will be referred to as tree distributions in the following.

Definition 2. *A graphical model $m_G := (G, \mathscr{F}_G)$ is given by a decomposable graph G and a family of distributions $\mathscr{F}_G \subseteq \mathscr{F}$ that are Markov w.r.t. G.*

Let $m_G = (G, \mathscr{F}_G)$ be a graphical model. To avoid any confusion, distributions on a set of distributions will be called hyperdistributions. For $\pi \in \mathscr{F}_G$ and $A, B \subseteq V$, we let π_A denote the marginal distribution obtained from π on the variables $\mathbf{X}_A$, and $\pi_{B|A}$ denote the collection of conditional distributions of $\mathbf{X}_B | \mathbf{X}_A$ under π. If ρ is a hyperdistribution on $\mathscr{F}_G$, we also let ρ_A and $\rho_{B|A}$ respectively denote the marginal hyperdistribution induced by ρ on π_A and the collection of hyperdistributions induced by ρ on $\pi_{B|A}$.

Definition 3. *A hyperdistribution ρ is said to be strong hyper Markov w.r.t. G if, for any decomposition (A,B) of G, it holds that $\pi_A \perp\!\!\!\perp \pi_{B|A}$ under ρ.*

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

2.2. Model for Bayesian inference of graphical models based on trees

Let $\mathscr{T}$ denote the set of spanning trees on V. For any tree $T \in \mathscr{T}$, we consider a graphical model $m_T = (T, \mathscr{F}_T)$ with a family of positive distributions $\mathscr{F}_T \subseteq \mathscr{F}$ Markov w.r.t. T. As we consider a Bayesian framework, we need to define prior distributions for T and for π conditionally on T. This is dealt with in Section 3. The full Bayesian model is hierarchically described as follows.

We first draw a random tree T^* in the set of spanning trees, then a distribution π in $\mathscr{F}_T$ and finally $\mathbf{X}$ according to π. 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 hyper-distributions 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. This choice of priors and the fact that we only consider trees as possible structures make the inference of the graph in our model tractable in an exact manner.

3. Priors on tree structures & distributions

The cardinality of $\mathscr{T}$ is p^{p-2}. Thus, restraining possible structures to spanning trees still leaves us with a large collection of graphical models to consider. Nonetheless, a suitable choice of priors on tree structures and parameters leads to a tractable situation. Meilă and Jaakkola (2006) define 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 using algebra. We will make use of strong hyper Markov hyperdistributions (Dawid and Lauritzen, 1993) to define our priors, but the idea is basically the same. Let $D = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ be an independent sample of size $n \geq 1$ drawn from $\mathbf{X}$. Our goal is to define a prior distribution ξ on (T, π) such that the posterior distribution on trees $\xi(\cdot|D)$ factorises over the edges, *i.e.*

$$\xi(T|D) = \frac{1}{Z} \prod_{\{i,j\} \in E_T} \omega_{ij}, \qquad \forall T \in \mathscr{T}, \tag{1}$$

where $\omega = (\omega_{ij})_{(i,j) \in V^2}$ is a symmetric matrix with non-negative values and $Z = \sum_{T \in \mathscr{T}} \prod_{\{i,j\} \in E_T} \omega_{ij}$ is a normalising constant. Both ω and Z obviously depend on the data D, but we drop the dependence in the notations for the sake of clarity.

3.1. Prior on tree structures

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

$$\xi(T) = \frac{1}{Z_0} \prod_{\{i,j\} \in E_T} \beta_{ij}, \qquad \forall T \in \mathscr{T}. \tag{2}$$

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

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,

$$p(D|T) = \int_{\mathscr{F}_T} p(D|\pi)p(\pi|T)d\pi. \tag{3}$$

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

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

Proposition 2. *(Dawid and Lauritzen, 1993, Prop. 5.6) If ρ^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 $D_i = \{x_i^{(1)}, \ldots, x_i^{(n)}\}$ be the observed data restricted to X_i. The integral given in (3) can then be rewritten as

$$p(D|T) = \int \pi(D)\rho^T(\pi)d\pi = \prod_{i \in V} p(D_i|T) \prod_{\{i,j\} \in E_T} \frac{p(D_i, D_j|T)}{p(D_i|T)p(D_j|T)} \tag{4}$$

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

$$\begin{aligned} p(D_i, D_j|T) &= \int \pi_{ij}(D_i, D_j)\rho_{ij}^T(\pi_{ij})d\pi_{ij}; \\ p(D_i|T) &= \int \pi_i(D_i)\rho_i^T(\pi_i)d\pi_i. \end{aligned} \tag{5}$$

The calculation of these integrals will be addressed in Section 4.1.

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

$$\pi_A \perp\!\!\!\perp \pi_{V \setminus A|A}. \tag{6}$$

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

Proposition 3. *(Dawid and Lauritzen, 1993, §6.2) For any tree $T \in \mathscr{T}$, there exists a unique hyperdistribution ρ^T on $\mathscr{F}_T$ that is strong hyper Markov w.r.t. T and such that, for every edge $\{i,j\} \in E_T$, $\rho_{ij}^T = \rho_{ij}$. The collection $\{\rho^T\}_{T \in \mathscr{T}}$ is said to be a (hyper) compatible family of strong hyper Markov hyperdistributions.*

Proposition 3 guarantees that all ρ^T are strong hyper Markov w.r.t. T. By Proposition 2, for all $T \in \mathscr{T}$, the marginal likelihood under ρ^T is Markov w.r.t. T. Moreover, the compatibility of the family $\{\rho^T\}_{T \in \mathscr{T}}$ makes the dependence on T in the local marginal distributions given in (5) irrelevant. They can be computed once and for all for every $\{i,j\} \in \mathscr{P}_2(V)$. With this choice of hyperdistributions, the factorisation property needed for the posterior tree distribution (Eq. 1) is satisfied with

$$\omega_{ij} = \beta_{ij} \frac{p(D_i, D_j)}{p(D_i)p(D_j)}, \qquad \forall (i,j) \in V^2. \tag{7}$$

A full description of the model is given in Figure 1.

Proposition 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 $\{\rho_{ij}\}_{\mathscr{P}_2(V)}$, 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 $\{\rho_{ij}\}_{\{i,j\}\in\mathscr{P}_2(V)}$ is obtained from a fully specified hyperdistribution ρ. 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 ρ_{ij} (Dawid and Lauritzen, 1993, Prop. 3.16), meaning that ρ_{ij} is strong hyper Markov w.r.t. the graph on $\{i,j\}$ where vertices i and j are connected.

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 and Liu (1968) described an algorithm that can be used to get the tree distribution maximizing the likelihood of discrete multivariate data in the frequentist equivalent of the model given 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 $p(\mathbf{x}|D) = \sum_{T\in\mathscr{T}} p(\mathbf{x}|T)\xi(T|D)$ (Meilă and Jaakkola, 2006). In some other situations, the dependence structure 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 β maximising 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). In their approach, the parameters of the models are estimated with plug-in estimators. Even if the distribution on trees cannot be called a prior in the traditional sense, 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, for any edge $\{k,l\}$,

$$P(\{k,l\} \in E_{T^*}|D,\xi) = \sum_{T\in\mathscr{T}:E_T\ni\{k,l\}} \xi(T|D). \tag{8}$$

The previous section shows that achieving this requires two things. First, we have to get access to ω by computing local marginal likelihoods, which amounts to integrating w.r.t. π (Section 4.1). Then comes in the integration over the set of trees, that can be performed exactly using an algebraic result called the Matrix-Tree theorem (Section 4.2).

$$\begin{aligned} T \,|\, \beta &\sim \xi; \\ \pi \,|\, T,\rho &\sim \rho^T; \\ \mathbf{X} \,|\, \pi &\sim \pi. \end{aligned}$$

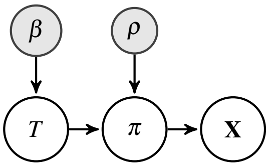


Figure 1: Compatible strong hyper Markov tree model.

4.1. Integration with respect to π

Thanks to the strong hyper Markov property required for the hyperdistributions, the integration on π can be performed locally and the compatibility ensures that these local integrated quantities can be passed from one tree graphical model to another whenever they are needed. Thus, the integrations are always carried out on sets of bivariate distributions, with $p(p+1)/2$ 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 settings where the local integrated likelihood terms can be computed exactly by using conjugate priors for the local distributions.

4.1.1. Tree-Structured Copulas

Let us assume that $\mathscr{X} = [0,1]^p$. 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 let $\mathscr{U}$ denote the uniform distribution on $[0,1]$ and we assume that, for all $i \in V$, $X_i \sim \mathscr{U}$. 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 ρ_i for π_i is then a Dirac distribution concentrated on $\mathscr{U}$, denoted by $\delta_{\mathscr{U}}$. Defining a compatible family of hyperdistributions requires that we consider pairwise hyperdistributions with marginals equal to $\delta_{\mathscr{U}}$. Such hyperdistributions are in fact defined on bivariate copulas.

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

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

Let us remark that if $\psi(0) = \infty$, $\psi^{[-1]} = \psi^{-1}$. The cdf of the Archimedean copula generated by ψ is given by $C_\psi(x_i,x_j) = \psi^{[-1]}(\psi(x_i) + \psi(x_j))$. Function ψ is said to be a generator of the copula C_ψ. 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 as an example.

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 Υ between θ and the distributions π_{ij} on (X_i,X_j). A pairwise hyperdistribution ρ_{ij} for π_{ij} is then defined by any distribution κ for θ through the identity $\rho_{ij}(\pi_{ij}) = \kappa\left(\Upsilon^{-1}(\pi_{ij})\right)$ and the integrated pairwise distribution $p(x_i,x_j)$ is given by

$$p(x_i,x_j) = \int_\Theta \frac{\partial^2 C_\theta}{\partial x_i \partial x_j}(x_i,x_j)\kappa(\theta)d\theta, \qquad \forall (x_i,x_j) \in [0,1]^2. \tag{9}$$

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

The integrals given in (9) 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 Gumbel copulas, a numerical or Monte Carlo integration is required. 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 4.1.3.

4.1.2. Multinomial Distributions

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

$$\Theta = \left\{ \theta \in [0;1]^{|\mathscr{X}|} \;\middle|\; \sum_{\mathbf{x}\in\mathscr{X}} \theta(\mathbf{x}) = 1 \right\}.$$

This is the set of multinomial distributions on $\mathscr{X}$. It happens that the conjugate Dirichlet distribution is satisfying the condition given in (6) necessary to build a compatible family of strong hyper Markov hyperdistributions. Let $\lambda = (\lambda(\mathbf{x}))_{\mathbf{x}\in\mathscr{X}}$ be a family of positive numbers indexed by $\mathscr{X}$. For $\theta \in \Theta$, we let $\mathscr{D}(\lambda)$ denote the Dirichlet distribution, with density$f(\theta|\lambda) \propto \prod_{\mathbf{x}\in\mathscr{X}} \theta(\mathbf{x})^{\lambda(\mathbf{x})-1}$.

Proposition 4. *(Dawid and Lauritzen, 1993, Lemma 7.2) Let $A \subseteq V$ and $B = V \setminus A$. For all $\mathbf{x}_A \in \mathscr{X}_A$, we define $\lambda_A(\mathbf{x}_A) := \sum_{\mathbf{y},\mathbf{y}_A=\mathbf{x}_A} \lambda(\mathbf{y})$. If $\theta \sim \mathscr{D}(\lambda)$, then $\theta_A \sim \mathscr{D}(\lambda_A)$ and $\theta_A \perp\!\!\!\perp \theta_{B|A}$.*

It results from the fact that, if $\{Y_k\}_{k=1}^K$ are independent random variables distributed as $\Gamma(\lambda_k, \theta)$ respectively and if $V := \sum_{k=1}^K Y_k$, then $(Y_1/V, \ldots, Y_K/V) \sim \mathscr{D}(\lambda)$. Proposition 4 states that any λ gives rise to a hyperdistribution ρ on the multinomial family of distributions from which we can build a family of compatible strong hyper Markov hyperdistributions and that the marginal hyperdistributions are also Dirichlet distributed. The conjugacy can then be used locally to compute ω. These hyperdistributions were referred to as hyper-Dirichlet laws in (Dawid and Lauritzen, 1993, §7.2.2).

As mentioned in Section 3.2, specifying a full set of hyperparameters λ is in fact not necessary to define the family of hyperdistributions $\{\rho^T\}_{T\in\mathscr{T}}$. We only need a consistent family of $\{\lambda_{ij}\}_{(i,j)\in V^2}$, in the sense that, for $(i,j,k) \in V^3$, λ_{ij} and λ_{ik} should induce the same λ_i. A possibility is to use an equivalent sample size N and to set, for all $(i,j) \in V^2$, $\lambda_{ij} := N/r_i r_j$ and $\lambda_i := N/r_i$. If all $\mathscr{X}_i$ are of equal size r, one can choose $N = r^2/2$ so that all λ_{ij} 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 do not belong to hyper-Dirichlet hyperdistributions (York and Madigan, 1992). For an edge $\{i, j\}$, we let λ'_{ij} denote the updated hyperparameters for the edge $\{i, j\}$ given by $\lambda'_{ij}(\ell,\ell') = \lambda_{ij}(\ell,\ell') + \sum_{k=1}^n \delta_{x_i^k,\ell}\delta_{x_j^k,\ell'}$, $\forall(\ell,\ell') \in \mathscr{X}_i \times \mathscr{X}_j$, where $\delta_{x,\ell} = 1$ if

$x = \ell$ and 0 otherwise. The matrix ω defined in (7) is then given by (Meilă and Jaakkola, 2006)

$$\omega_{ij} = \beta_{ij} \prod_{\ell \in \mathscr{X}_i} \frac{\Gamma(\lambda_i(\ell))}{\Gamma(\lambda_i'(\ell))} \prod_{\ell' \in \mathscr{X}_j} \frac{\Gamma(\lambda_j(\ell'))}{\Gamma(\lambda_j'(\ell'))} \prod_{(\ell,\ell') \in \mathscr{X}_i \times \mathscr{X}_j} \frac{\Gamma(\lambda_{i,j}'(\ell,\ell'))}{\Gamma(\lambda_{i,j}(\ell,\ell'))}$$

where Γ denotes the gamma function. If $R = \max_{i \in V} r_i$, computing ω requires $O(np^2R^2)$ operations (Meilă and Jaakkola, 2006).

4.1.3. *Gaussian Distributions*

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

$$\Lambda \sim \mathscr{W}(\alpha,\Phi), \qquad\qquad \mu|\Lambda \sim \mathscr{N}(\nu,(\lambda\Lambda)^{-1}),$$

where $\mathscr{W}(\alpha,\Phi)$ stands for the Wishart distribution with $\alpha > p-1$ degrees of freedom and positive-definite parametric matrix Φ. Geiger and Heckerman (2002) showed that the normal-Wishart distribution satisfies the parameter independence property given in (6). They further proved that this property coerces the distribution to be normal-Wishart whenever $p \geq 3$. It can thus be used to build a compatible family of strong hyper Markov hyperdistributions. Moreover, for any partitioning (A,B) of V, $\mathbf{X}_A \sim \mathscr{N}(\mu_A, \left(\Lambda_A - \Lambda_{AB}\Lambda_B^{-1}\Lambda_{AB}^T\right)^{-1})$ and $(\mu_A, \Lambda_A - \Lambda_{AB}\Lambda_B^{-1}\Lambda_{AB}^T)$ is also normal-Wishart-distributed with parameters $(\nu_A, \lambda, \alpha - p + l, \Phi_A - \Phi_{AB}\Phi_B^{-1}\Phi_{AB}^T)$ 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 $(\nu',\lambda',\alpha',\Phi')$, applying classical Bayesian updating formulæ. The locally updated hyperparameters are then derived from the globally updated ones and

$$p(D_i,D_j) \propto \frac{|\Phi_{\{i,j\}}|^{\frac{\alpha-p+2}{2}}}{|\Phi'_{\{i,j\}}|^{\frac{\alpha'-p+2}{2}}}, \qquad p(D_i) \propto \frac{|\Phi_i|^{\frac{\alpha-p+1}{2}}}{|\Phi'_i|^{\frac{\alpha'-p+1}{2}}}, \tag{10}$$

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 and Heckerman (2002).

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

4.2. *Integration with respect to T*

We assume that we have knowledge of ω. The computation of ω has a typical complexity of $O(np^2)$. The complexities mentionned in this section leave out this prior computation step. Given

ω, 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 borrow edge $\{k,l\}$. Because we are only considering trees, these summations can be efficiently performed.

Let $\omega = (\omega_{ij})_{(i,j)\in V^2}$ be a symmetric weight matrix such that, for all $i \in V$, $\omega_{ii} = 0$, and with non-negative off-diagonal terms. The weight of a graph $G = (V,E_G)$ is defined as the product of the weights of its edges, $\omega_G := \prod_{\{i,j\}\in E_G} \omega_{ij}$. The Laplacian $\Delta = (\Delta_{ij})_{(i,j)\in V^2}$ of ω is given by $\Delta_{ij} = -\omega_{ij}$ if $i \neq j$ and $\Delta_{ii} = \sum_{j\in V} \omega_{ij}$ for $i \in V$. For $U \subseteq V$, we defined Δ^U as the matrix obtained from Δ by removing the rows and columns corresponding to U, with rows and columns indexed by $V \setminus U$.

Theorem 1 (Chaiken, 1982)**.** *Let Δ be the Laplacian of a weight matrix ω. Then all minors $|\Delta^{\{u\}}|$, $u \in V$, are equal and $|\Delta^{\{u\}}| = \sum_{T\in\mathscr{T}} \omega_T$.*

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 (All Minors Matrix-Tree theorem, Chaiken, 1982)**.** *Let Δ be the Laplacian of a weight matrix ω and $U \subseteq V$. Let $\mathscr{F}_U$ 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 $|\Delta^U| = \sum_{F\in\mathscr{F}_U} \omega_F$.*

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 $\mathscr{F}_U$. If U is taken equal to a single vertex, then the forests in F_U only have one connected component which is a tree and we get Theorem 1. This theorem will be used in the proof of the following result that was first stated by Kirshner (2007).

Theorem 3 (Kirshner, 2007)**.** *Let ω be defined as in (7) and Δ be the associated Laplacian. Let u be a vertex in V. We define matrices Q and M respectively by*

$$Q_{kl} = \begin{cases} \left[\left(\Delta^{\{u\}}\right)^{-1}\right]_{kl} & \text{if } k,l \neq u, \\ 0 & \text{otherwise}, \end{cases} \tag{11}$$

$$M_{kl} = Q_{kk} + Q_{ll} - 2Q_{kl}. \tag{12}$$

Then, for all $\{k,l\} \in \mathscr{P}_2(V)$,

$$P(\{k,l\} \in E_{T^*}|D,\xi) = \omega_{kl} \cdot M_{kl} \tag{13}$$

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.

Proof. Let $\{k,l\}$ be an edge in $\mathscr{P}_2(V)$. Let Z, Z_{kl}^+ and Z_{kl}^- respectively denote the sums of ω_T over the sets $\mathscr{T}$, $\{T \in \mathscr{T} : \{k,l\} \in E_T\}$ and $\{T \in \mathscr{T} : \{k,l\} \notin E_T\}$. It is immediate to see that $Z = Z_{kl}^+ + Z_{kl}^-$. Lemma 3 of (Meilă and Jaakkola, 2006) states that $\frac{\partial Z}{\partial \omega_{kl}} = M_{kl}|\Delta^{\{u\}}| = M_{kl}Z$ where M is defined as in (12). It is easy to see that Z_{kl}^- can be obtained by applying Theorem 1 to a weight matrix equal to ω except for the terms ω_{kl} and ω_{lk} that are set to 0. This means that Z_{kl}^{-} does not depend on ω_{kl} and $\frac{\partial Z}{\partial \omega_{kl}} = \frac{\partial Z_{kl}^{+}}{\partial \omega_{kl}}$.

We then use Theorem 2 to get an expression of Z_{kl}^{+}. Indeed, there is a one-to-one correspondence between the set of forests rooted in k and l (denoted by $\mathscr{F}_{\{k,l\}}$) and the set of trees borrowing edge $\{k,l\}$. Going from one to the other is just a matter of adding or removing edge $\{k,l\}$. Then, by Theorem 2,

$$Z_{kl}^{+} = \omega_{kl} \sum_{F \in \mathscr{F}_{\{k,l\}}} \omega_F = \omega_{kl} \cdot |\Delta^{\{k,l\}}|. \tag{14}$$

$|\Delta^{\{k,l\}}|$ does not depend on ω_{kl} since the only terms of Δ that depend on ω_{kl} are Δ_{kl}, Δ_{lk}, Δ_{kk}, Δ_{ll} and these terms are all withdrawn in $\Delta^{\{k,l\}}$. Therefore,

$$|\Delta^{\{k,l\}}| = \frac{\partial Z_{kl}^{+}}{\partial \omega_{kl}} = \frac{\partial Z}{\partial \omega_{kl}} = M_{kl} \cdot Z. \tag{15}$$

Combining (14) and (15) with the fact that $P(\{k,l\} \in E_T | D, \xi) = Z_{kl}^{+}/Z$, we get the claimed result. □

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

4.2.1. *Posterior moments of node degrees*

The aim of structure inference is to decipher the dependency structure of a set of random variable. In this perspective, the degree of vertex k (i.e. its number of neighbors) in the graph informs us about the centrality of the corresponding variable X_k in the system. Denoting N_k this degree, we can easily derive the posterior mean of N_k from the end of the proof of Theorem 3 as

$$\mathbf{E}[N_k|D] = \sum_{l \neq k} Z_{kl}^{+}/Z = \sum_{l \neq k} P(\{k,l\} \in E_{T^*} | D, \xi)$$

The posterior variance of N_k can also be computed for all vertices with total complexity $O(p^3)$. The proof of this result is based on the following lemma giving some the second-order derivatives of the normalising constant Z.

Lemma 1. *Let ω be defined as in (7) and Δ be the associated Laplacian. Let u be a vertex in V and Q defined as in Theorem 3. For $k \in V$, let $M^{(k)}$ be the matrix whose general term is given by*

$$M_{l_1 l_2}^{(k)} = M_{kl_1} M_{kl_2} - M_{k,l_1,l_2}^2,$$

where $M_{l_1 l_2}$ is defined as in (12) and $M_{k,l_1,l_2} := Q_{kk} + Q_{l_1 l_2} - Q_{kl_1} - Q_{kl_2}$. Then, for $k, l_1, l_2 \in V$ that are pairwise distinct, it holds that

$$\frac{\partial^2 Z}{\partial \omega_{kl_1} \partial \omega_{kl_2}} = Z \cdot M_{l_1 l_2}^{(k)}.$$

The proof of this lemma is given in the Appendix.

Theorem 4. *Let ω be defined as in (7) and Δ be the associated Laplacian. Let u be a vertex in V and Q be defined as in Theorem 3. Then, for all $k \in V$, we let $\mathbf{V}(N_k|D)$ denote the posterior variance of N_k and it holds that*

$$\mathbf{V}(N_k|D) = \mathbf{E}[N_k|D]\,(1-\mathbf{E}[N_k|D]) + \sum_{\substack{l_1 \neq k, l_2 \neq k \\ l_1 \neq l_2}} \omega_{k,l_1}\omega_{k,l_2} M^{(k)}_{l_1,l_2}. \tag{16}$$

Proof. We have that

$$\mathbf{E}[N_k^2|D] = \sum_{\substack{l_1 \neq k \\ l_2 \neq k}} \mathbf{E}[\mathbf{1}_{\{k,l_1\}}\mathbf{1}_{\{k,l_2\}}|D].$$

Let $l_1, l_2 \in V \setminus \{k\}$ such that $l_1 \neq l_2$. There is a one-to-one correspondence between the set of trees borrowing edges $\{k,l_1\}$ and $\{k,l_2\}$, and the forests rooted in $\{k,l_1,l_2\}$. Using Theorem 2 and Lemma 1, we deduce that

$$\mathbf{E}[\mathbf{1}_{\{k,l_1\}}\mathbf{1}_{\{k,l_2\}}|D] = \omega_{k,l_1}\omega_{k,l_2} M^{(k)}_{l_1,l_2}$$

by a reasoning similar to the one used in the proof of Theorem 3. The expression given in (16) is then easily derived. □

Theorem 4 shows that the posterior variance for the degree of all vertices can be obtain directly at virtually no extra cost once posterior edge probabilities have been computed, since both computations rely on the inversion of the same matrix.

4.2.2. Posterior entropy

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 5. *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 $(P(\{k,l\} \in E_{T^*}|D,\xi))_{\{k,l\}\in\mathscr{P}_2(V)}$ which can both be computed with complexity $O(p^3)$ through Theorems 1 & 3. Indeed,

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

□

4.2.3. *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 $\mathscr{T}$ 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 $\{k,l\} \in \mathscr{P}_2(V)$ and the event $\mathscr{E}_{kl} := \{T : \{k,l\} \in E_T\}$. We let p^0_{kl} and p_{kl} respectively denote the prior and posterior probabilities of event $\mathscr{E}_{kl}$. These probabilities are obtained through Theorem 3.

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

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

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

$$p(D) = \lambda_{kl}\, p(D|\mathscr{E}_{kl}) + (1-\lambda_{kl})\, p(D|\overline{\mathscr{E}}_{kl}).$$

We are now interested in the posterior distribution of ε_{kl}.

Proposition 6.

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

Proof.

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

□

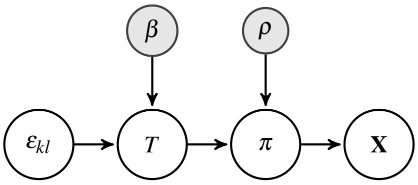


Figure 2: Model with variable ε_{kl} explicitly controlling the status of edge $\{k,l\}$ in T.

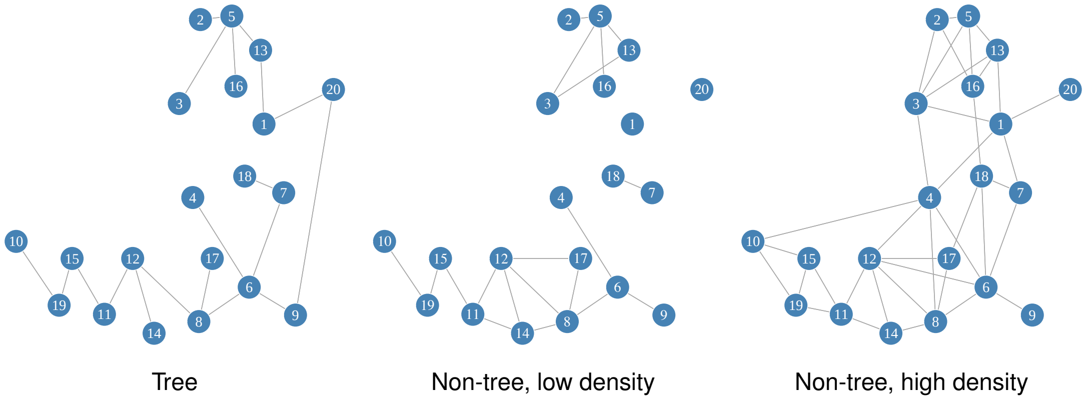
The computation of $P(\varepsilon_{kl} = 1|D)$ for all edges can be achieved in $O(p^2)$ time from the posterior edge probability matrix $\{p_{kl}\}_{\{k,l\}\in\mathscr{P}_2(V)}$. We can notice that $P(\varepsilon_{kl} = 1|D)$ is a strictly increasing function of p_{kl}. When the initial prior on trees ξ is uniform and all λ_{kl} are taken equal, the order induced on the edges by $\{P(\varepsilon_{kl} = 1|D)\}_{\{k,l\}\in\mathscr{P}_2(V)}$ 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.

5. Behavior on synthetic data

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. (2016) and implemented in the BEANDisco software. Computations for our approach were performed with the R package **saturnin**.

5.1. *Simulation scheme*

We have chosen three networks with $p = 20$ vertices. The first one is a spanning tree. The second and third graphs are not spanning trees and respectively have as many and twice as many edges as the first one. These graphs are shown in Figure 3. We then simulated data according to a multinomial model with $\mathscr{X}_i = \{1,2,3\}$ for $i \in V$. For each graph G, we have chosen a DAG D with skeleton equal to G. We let $\mathrm{par}(i,D)$ stands for the set made of the parents of vertex i in DAG D. For $\mathbf{X} \in \{1,2,3\}^p$, we let $N_i^D(r;\mathbf{X}) := \left|\left\{j \in \mathrm{par}(i,D) : X_j = r\right\}\right|$ denote the number of parents of vertex i in D taking value $r \in \{1,2,3\}$ in $\mathbf{X}$. Then, conditionally on D, we used the



Tree

Non-tree, low density

Non-tree, high density

Figure 3: Gold standard networks in the simulation study.

following distribution for $\mathbf{X}$: $P(X_i = r) = 1/3$ if $\text{par}(i,D) = \varnothing$ and

$$P\left(X_i = r | X_{\text{par}(i,D)}\right) \propto \eta \cdot N_i^D(r;\mathbf{X}) + 1 \qquad\qquad \text{if } \text{par}(i,D) \neq \varnothing.$$

As the variables at root vertices are drawn uniformly, it can be shown that all variables are marginally uniformly distributed by a symmetry argument. Here, η was set to 0.5. For $n = 25$, 50, 75, 100 and 200, we generated 100 samples of size n.

We then considered the Multinomial/Dirichlet framework described in Section 4.1.2, setting the prior on trees ξ to the uniform and the equivalent prior sample size N to $3^2/2 = 4.5$ (see Section 4.1.2). For each data set, we computed

— the MAP tree structure through a Maximal Spanning tree algorithm (Prim, 1957) applied to ω;

— 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 set to $\lambda_{kl} = 1/2$ (see Section 4.2.3);

— an estimation of the matrix of posterior edge appearance probabilities in a random DAG obtained by MCMC sampling (Niinimäki et al., 2016). 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 sampling was performed for one minute on each dataset. The direction of the edges of the sampled DAGs was not taken into account to get empirical frequencies for all undirected edges.

The accuracy of the inference was evaluated against the true undirected 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.

5.2. Results

Comparison with MAP Figure 4 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 penalised 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 favour of considering the whole posterior distribution on trees instead of the MAP is presented in Figure 5. For all three simulation scenarios, posterior tree distributions are not really peaked around their modes, especially for small samples. The second most probable tree is always very close to the MAP. Moreover, the entropy of the posterior distribution on trees behaves similarly across all simulation scenarios.

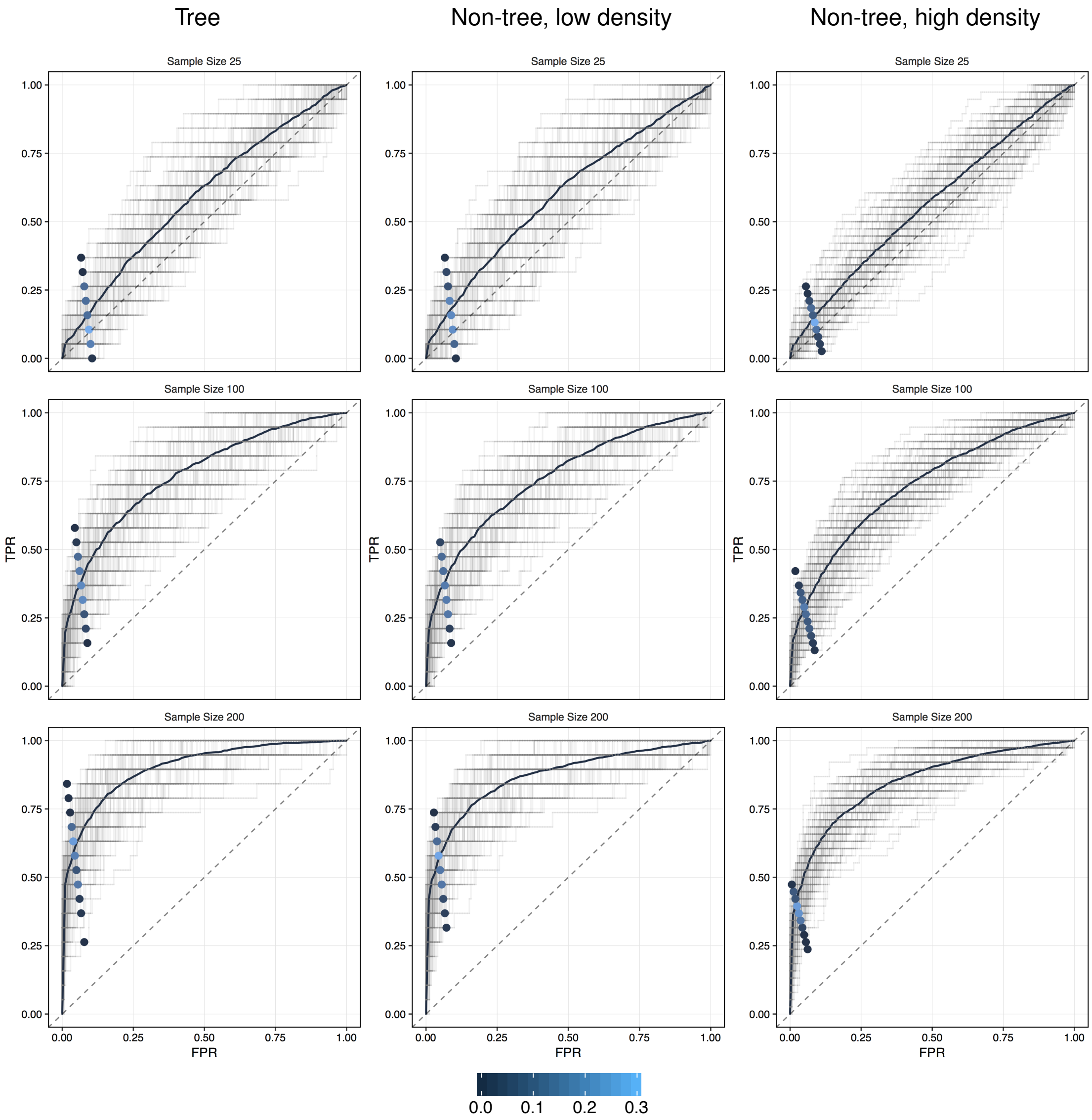


Figure 4: ROC curves for the posterior edge probabilities and (TPR, FPR) scores for the MAP estimate on data sets of size 25, 100 and 200 (from top to bottom). For the ROC curves, the mean curve is plotted in bold line. The color of a (TPR, FPR) point expresses its frequency within the 100 samples.

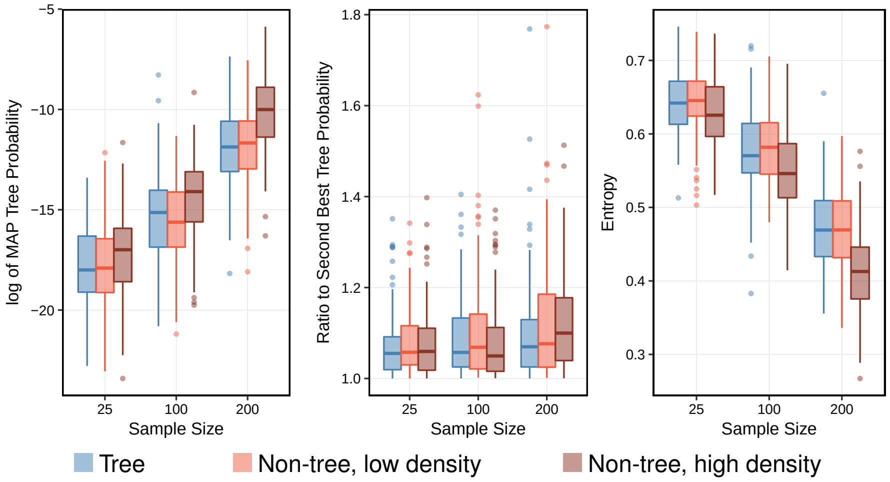


Figure 5: Posterior probability of the MAP tree, ratio to the posterior probability of the second best tree and entropy of the posterior tree distribution (normalised by the entropy of the uniform distribution on $\mathscr{T}$, *i.e.* $(p-2)\log(p)$).

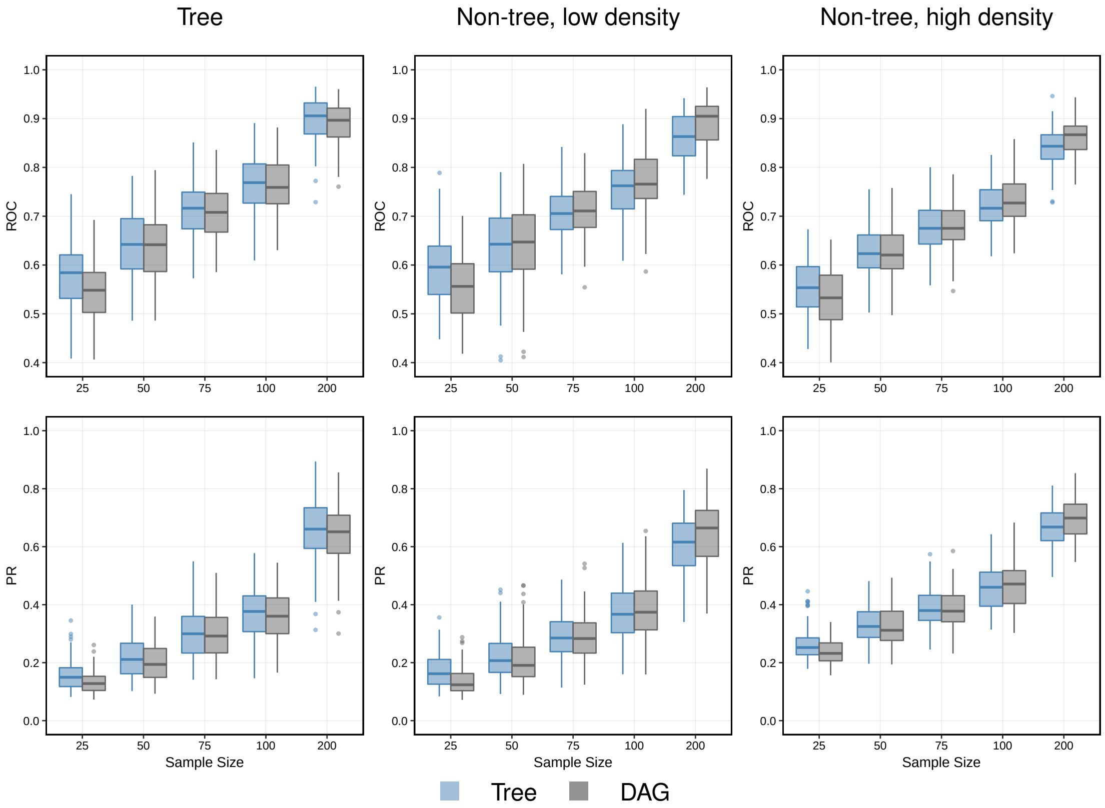


Figure 6: Area under the ROC (top) & PR (bottom) curves computed for the output of our approach and of the MCMC sampling algorithm in the set of DAGs.

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 6. Both algorithms seem to perform equally well in all three scenarios. 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 favour of considering an inference within the broader class of DAG structures.

Running time We conclude this section on synthetic data by mentioning running times. For $p = 25$, 50 and 75, we respectively observed average running times of 11, 206 and 1393 seconds for the MCMC approach on DAGs and of 0.2, 1 and 2.2 seconds for our method. While retaining similar accuracy to the algorithm based on MCMC sampling in the space of DAGs 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.

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 inference techniques. They are related to the Raf cellular signalling network, which is involved in many different biological 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 were measured by flow cytometry. The generally accepted structure of the Raf pathway is given in Figure 7, but the true structure of this network is not fully understood, despite considerable experimental and theoretical efforts, and may be more subtle. The undirected skeleton of this network will, however, be used as the gold standard network in our study.

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 of size $n = 100$ from the data provided by Sachs et al. (2005). We discretised 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), as described 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.

6.2. Results

With the DAG approach, we got average areas under the ROC and PR curves of 0.767 and 0.725 respectively (with standard deviation of 0.068 and 0.070). With trees, we respectively got 0.729 and 0.690 for these areas (with standard deviation of 0.047 and 0.051). The DAG approach seems to perform better than our inference based on trees. These results qualify those of the previous

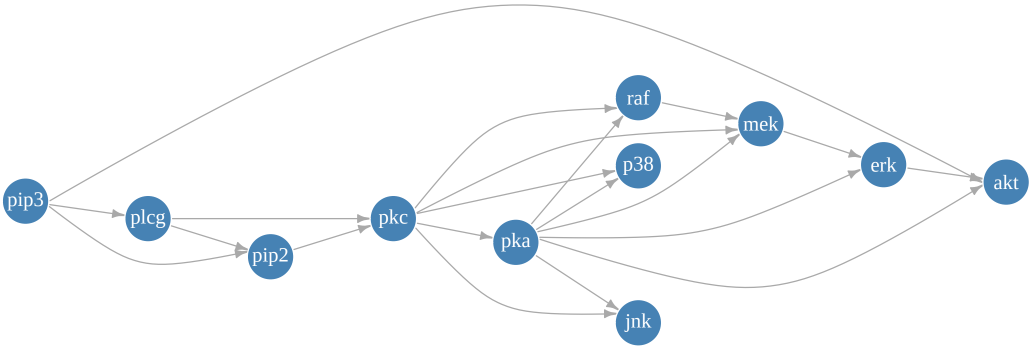
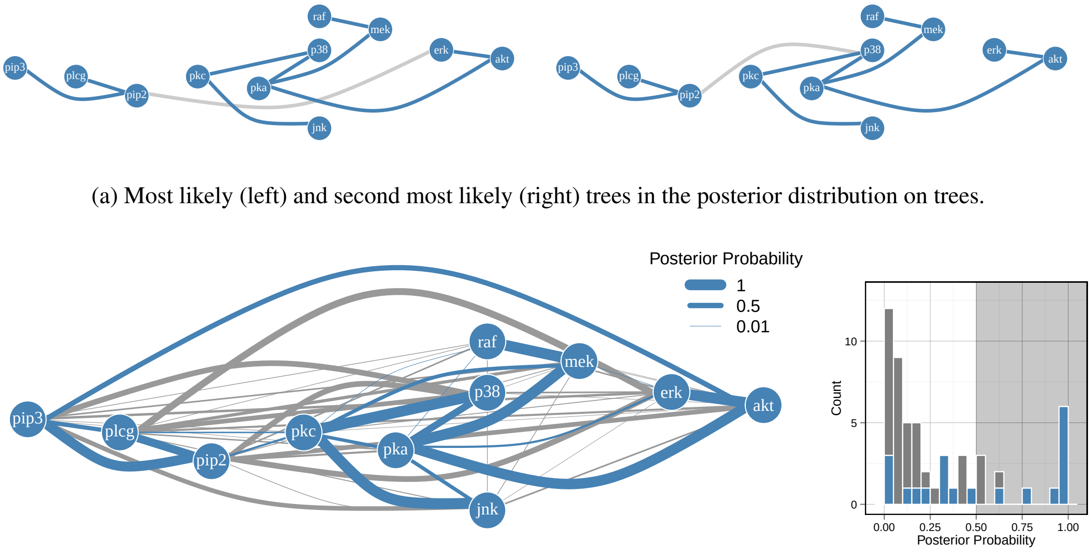


Figure 7: Raf pathway.



(a) Most likely (left) and second most likely (right) trees in the posterior distribution on trees.

(b) Posterior probabilities for the edges in the tree model (with change of prior probability to $\lambda_{kl} = 1/2$ for all edges).

Figure 8: Graphical representation of the results obtained on one of the five data sets. The edges of the golden standard network are colored in blue.

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 study, the gold standard network against which the accuracy of the inference is assessed here, shown in Figure 7, is not perfectly known and may still differ quite considerably from the truth.

Figure 8 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 8a. 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 8b). 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 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 on the accuracy.

Acknowledgements

The authors thank Sophie DONNET for her helpful comments and remarks.

Appendix

Proof of Lemma 1. Let $\overline{Q}$ be the matrix obtained from Q when row and column u are removed. Notice that $\overline{Q}=\left[\Delta^{\{u\}}\right]^{-1}$. For convenience, we also let $R:=\Delta^{\{u\}}=\overline{Q}^{-1}$. The rows and columns of $\overline{Q}$ and R are indexed by $\overline{V}:=V\setminus\{u\}$.

Let k,l_1,l_2 be pairwise distinct vertices in V. Using Theorem 1 and Lemma 3 of (Meilă and Jaakkola, 2006), we get that

$$\begin{aligned}\frac{\partial^2 Z}{\partial\omega_{kl_1}\partial\omega_{kl_2}} &= \frac{\partial^2|R|}{\partial\omega_{kl_1}\partial\omega_{kl_2}}\\ &= \frac{\partial}{\partial\omega_{kl_1}}\left(|R|\cdot M_{kl_2}\right)\\ &= |R|\cdot\left[M_{kl_1}M_{kl_2}+\frac{\partial M_{kl_2}}{\partial\omega_{kl_1}}\right]\end{aligned}$$

Assume that $u \notin \{k, l_1, l_2\}$. Then $M_{kl_2} = \overline{Q}_{kk} + \overline{Q}_{l_2 l_2} - 2\overline{Q}_{k,2}$ and

$$\frac{\partial \overline{Q}_{kk}}{\partial \omega_{kl_1}} = \sum_{i,j \in \overline{V}} \frac{\partial \overline{Q}_{kk}}{\partial R_{ij}} \frac{\partial R_{ij}}{\partial \omega_{kl_1}} = - \sum_{i,j \in \overline{V}} \overline{Q}_{ki} \overline{Q}_{jk} \frac{\partial R_{ij}}{\partial \omega_{kl_1}} = - \left(\overline{Q}_{kk} - \overline{Q}_{kl_1} \right)^2$$

where the last identity is obtained by noticing that the only terms of $R = \Delta^{\{u\}}$ that depend on ω_{kl_1} are R_{kl_1}, R_{l_1k}, R_{kk} and $R_{l_1l_1}$. We similarly obtain that

$$\begin{aligned} \frac{\partial \overline{Q}_{l_2 l_2}}{\partial \omega_{kl_1}} &= - \left(\overline{Q}_{l_1 l_2} - \overline{Q}_{kl_2} \right)^2 , \\ \frac{\partial \overline{Q}_{kl_2}}{\partial \omega_{kl_1}} &= \left(\overline{Q}_{kk} - \overline{Q}_{kl_1} \right) \left(\overline{Q}_{l_1 l_2} - \overline{Q}_{kl_2} \right) . \end{aligned}$$

Putting all pieces together, we get

$$\begin{aligned} \frac{\partial^2 Z}{\partial \omega_{kl_1} \partial \omega_{kl_2}} &= |R| \cdot \left[M_{kl_1} M_{kl_2} - \left(\overline{Q}_{kk} - \overline{Q}_{kl_1} - \overline{Q}_{kl_2} + \overline{Q}_{l_1 l_2} \right)^2 \right] , \\ &= Z \cdot M^{(k)}_{l_1 l_2} . \end{aligned}$$

The cases $k = u$ and $l_2 = u$ are dealt with similarly. □

References

Atay-Kayis, A. and Massam, H. (2005). A Monte Carlo method to compute the marginal likelihood in non decomposable graphical Gaussian models. *Biometrika*, 92:317–335.

Burger, L. and Van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology*, 6(1).

Byrne, S. and Dawid, A. P. (2015). Structural markov graph laws for Bayesian model uncertainty. *Ann. Statist.*, 43(4):1647–1681.

Chaiken, S. (1982). A Combinatorial Proof of the All Minors Matrix Tree Theorem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):319–329.

Chow, C. and Liu, C. (1968). Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, 21(3):1272–1317.

Friedman, N. and Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125.

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440.

Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a markov chain on junction trees. *Biometrika*, 100(1):91–110.

Hammersley, J. M. and Clifford, P. (1971). Markov field on finite graphs and lattices.

Heckerman, D. and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pages 20–197.

Kirshner, S. (2007). Learning with Tree-Averaged Densities and Distributions. *Advances in Neural Information Processing Systems 2008*, 20:761–768.

Kuipers, J., Moffa, G., and Heckerman, D. (2014). Addendum on the scoring of gaussian directed acyclic graphical models. *Ann. Statist.*, 42(4):1689–1691.

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

Lin, Y., Zhu, S., Leet, D. D., and Taskar, B. (2009). 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.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232.

Meilă, M. (1999). *Learning with Mixtures of Trees*. PhD thesis, Massachusetts Institute of Technology.

Meilă, M. and Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92.

Meilă, M. and Jordan, M. I. (2001). Learning with Mixtures of Trees. *The Journal of Machine Learning Research*, 1:1–48.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics.

Niinimäki, T., Parviainen, P., and Koivisto, M. (2016). Structure Discovery in Bayesian Networks by Sampling Partial Orders. *Journal of Machine Learning Research*, 17(57):1–47.

Parviainen, P. and Koivisto, M. (2009). Exact Structure Discovery in Bayesian Networks with Less Space. *Uai*, pages 436–443.

Prim, R. C. (1957). Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, 36(6):1389–1401.

Roverato, A. (2002). 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.

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

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics (Oxford, England)*, 22(20):2523–31.

York, J. C. and Madigan, D. (1992). Bayesian methods for estimating the size of a closed population. Technical Report 234, Department of Statistics, University of Washington, Seattle.