HAL
open science

# Exact Bayesian inference for off-line change-point detection in tree-structured graphical models

Loic Schwaller, Stephane Robin

## To cite this version:

Loic Schwaller, Stephane Robin. Exact Bayesian inference for off-line change-point detection in tree-structured graphical models. Statistics and Computing, 2017, 27 (5), pp.1331-1345. 10.1007/s11222-016-9689-3 . hal-01555266

**HAL Id: hal-01555266**

**https://hal.science/hal-01555266**

Submitted on 3 Jul 2017

**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.

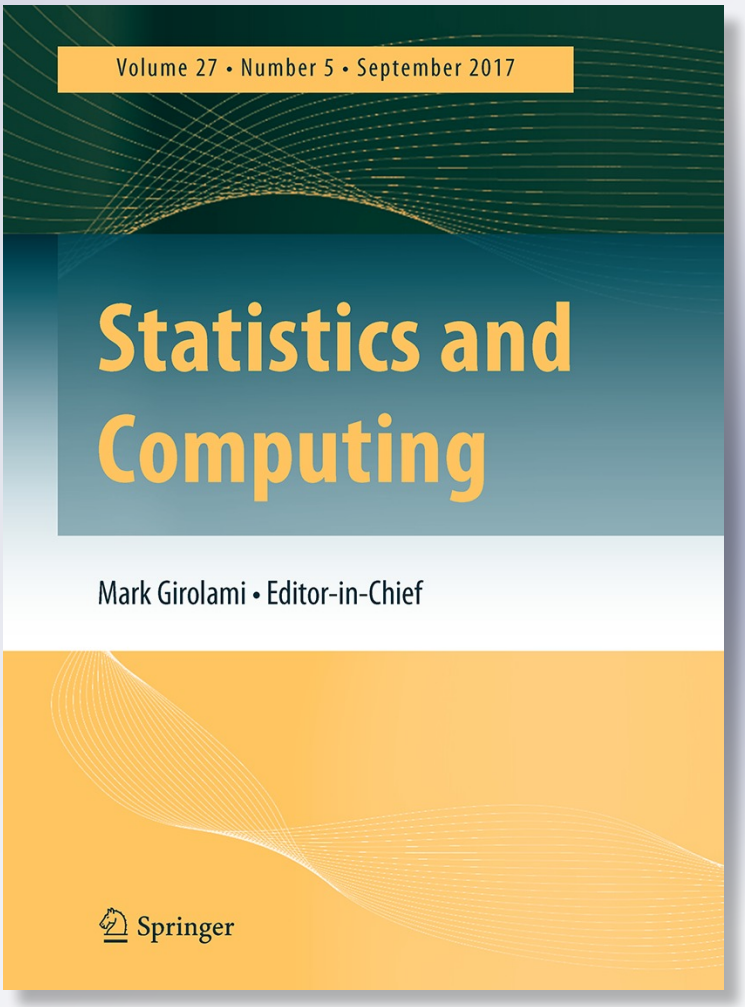
# *Exact Bayesian inference for off-line change-point detection in tree-structured graphical models*

**L. Schwaller & S. Robin**

**Statistics and Computing**

ISSN 0960-3174
Volume 27
Number 5

Stat Comput (2017) 27:1331-1345
DOI 10.1007/s11222-016-9689-3




Springer

Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media New York. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com”.

Springer

Author's personal copy

Stat Comput (2017) 27:1331–1345
DOI 10.1007/s11222-016-9689-3

# Exact Bayesian inference for off-line change-point detection in tree-structured graphical models

**L. Schwaller**[1] · **S. Robin**[1]

Received: 25 March 2016 / Accepted: 30 July 2016 / Published online: 11 August 2016
© Springer Science+Business Media New York 2016

**Abstract** 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. We demonstrate that it is possible to perform exact Bayesian inference whenever one considers a simple class of undirected graphs called spanning trees as possible structures. We are then able to integrate on the graph and segmentation spaces at the same time by combining classical dynamic programming with algebraic results pertaining to spanning trees. In particular, we show that quantities such as posterior distributions for change-points or posterior edge probabilities over time can efficiently be obtained. We illustrate our results on both synthetic and experimental data arising from biology and neuroscience.

**Keywords** Change-point detection · Exact Bayesian inference · Graphical model · Multivariate time-series · Spanning tree

## 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 and 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 and Hartemink 2010; Lèbre et al. 2010; Grzegorczyk and 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 and 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 and Hartigan 1992). We assume that the observed data $\{y^t\}_{t=1,\ldots,N}$ are a realisa-

**Electronic supplementary material** The online version of this article (doi:10.1007/s11222-016-9689-3) contains supplementary material, which is available to authorized users.

✉ L. Schwaller
loic.schwaller@agroparistech.fr

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France

Author's personal copy



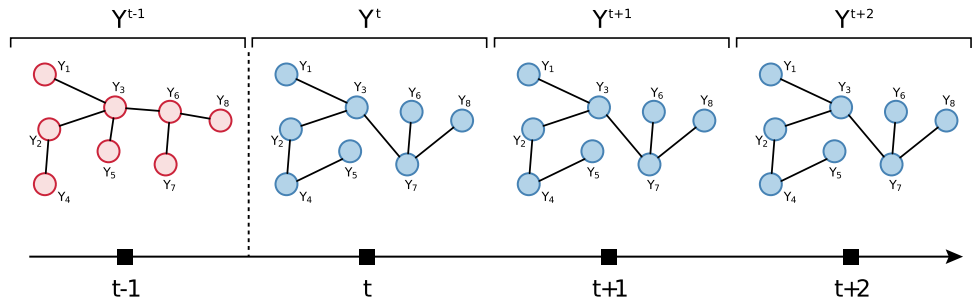



**Fig. 1** Illustration of the change-point detection problem in the tree structure of a graphical model

tion of a process $\{Y^t\}_{t=1,\dots,N}$ where, for $1 \leqslant t \leqslant T$, $Y^t$ is a random vector with dimension $p \geqslant 2$. If $m$ is a segmentation of $\{1, \dots, T\}$ with change-points $1 = \tau_0 < \tau_1 < \cdots < \tau_{K-1} < \tau_K = N$, the model has the general form

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

where $G_r$ and $\theta_r$ respectively stand for the graph describing the dependence structure and the distribution parameters on segment $r$. The parameters $(G_r, \theta_r)$ are assumed to be independent between segments. This model is illustrated in Fig. 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; Rigaill et al. 2012) and on-line detection (Fearnhead and Liu 2007; Caron et al. 2012) of change-points can be performed exactly and efficiently in this model thanks to dynamic programming. Xuan and 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 (see Fig. 1 for examples). 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 $[\![s; t]\!]$, defined by

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

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 Sect. 2, we provide some background on graphical models and product partition models. In particular, we give a more detailed presentation of the results of Rigaill 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 Sect. 3. Section 4 enumerates a list of quantities of interest that can be computed in this model, while Sect. 5 deals with edge and graph status comparison, when the segmentation is known. Sections 6 and 7 respectively present the

Author's personal copy



simulation study and the applications to both biological and neuroscience data.

## 2 Background

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

### 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 and Hartigan 1992) work under the assumption that the observations can be divided in independent adjacent segments. Thus, if $m$ is a partition of $[\![1; N]\!]$, 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 $K \geqslant 1$, we let $\mathcal{M}_K$ denote the set made of the partitions of $[\![1; N]\!]$ into $K$ segments. The cardinality of this set is $\binom{N-1}{K-1}$. More generally, we let $\mathcal{M}_K([\![s; t[\![)$ denote the partitions of any interval $[\![s; t[\![$ into $K$ segments. In order to get the marginal likelihood of $y$ conditionally on $K$, one has to integrate out both $m$ and $\theta = \{\theta_r\}_{r\in m}$:

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

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

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

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. Rigaill et al. (2012) introduced a matrix containing the weighted likelihood of all possible segments, whose general term is given by

$$A_{s,t} = \begin{cases} a_{[\![s;t[\![} \cdot p(y^{[\![s;t[\![}) & \text{if } 1 \leqslant s < t \leqslant N+1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

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 1** (*Rigaill et al. 2012*)

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

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

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

*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 Sect. 4.1). Classical Bayesian selection criteria for $K$ are also given. One can notice that $C_K(a)$ can be recovered by applying Proposition 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 and Liu (2007) showed that on-line and exact inference is also tractable in this model.

### 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.

Author's personal copy



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 $\theta$ (Meilă and Jaakkola 2006; Schwaller et al. 2015). The distribution on trees is taken to be factorised on the edges,

$$p(T) = \frac{1}{Z(b)} \prod_{\{i,j\} \in E_T} b_{ij}, \tag{3}$$

where $b_{ij}$ are non-negative edge weights and

$$Z(b) := \sum_{T \in \mathcal{T}} \prod_{\{i,j\} \in E_T} b_{ij} \tag{4}$$

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

$$p(y_i, y_j) = \int p(y_i, y_j|\theta_{ij}) d\theta_{ij},$$
$$p(y_i) = \int p(y_i|\theta_{ij}) d\theta_i. \tag{5}$$

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 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 (4) and $\omega$ is the posterior edge weight matrix whose general term is given by*

$$\omega_{ij} := b_{ij} \frac{p(y_i, y_j)}{p(y_i) p(y_j)}. \tag{6}$$

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

*Proof*

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

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. □

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 Model and properties

Sections 2.1 and 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

Author's personal copy



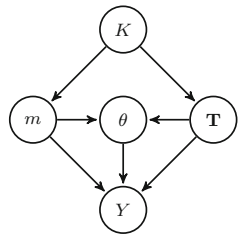



**Fig. 2** Global graphical model

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.

### 3.1 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 $p \geqslant 2$. For $1 \leqslant t \leqslant N$, $Y^t = (Y_1^t, \ldots, Y_p^t)$ 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 $m$. Our model can then be written as follows:

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

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 (1) and (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 Sect. 2.2. A graphical representation of this model is given in Fig. 2.

### 3.2 Factorisation properties

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

$$
p(y|K) = \sum_{m \in \mathcal{M}_K} \sum_{\mathbf{T} \in \mathcal{T}^K} \int p(y, m, \theta, \mathbf{T}|K) d\theta. \tag{7}
$$

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 Sect. 2.

**Proposition 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*

$$
\omega_{ij}^{(r)} := b_{ij} \frac{p(y_i^r, y_j^r)}{p(y_i^r) p(y_j^r)}. \tag{8}
$$

*$p(y_i^r, y_j^r)$ and $p(y_i^r)$ are local integrals on $\theta$ computed on edges and vertices as defined in (5).*

*Proof* For any segmentation $m \in \mathcal{M}_K$ of $[\![1; N]\!]$ 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 $r$,

$$
p(y^r) = \sum_{T \in \mathcal{T}} p(T) \int \left( \prod_{t \in r} p(y^t | T, \theta) \right) p(\theta|T) d\theta. \tag{9}
$$

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

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

is computed, Proposition 1 can be used to gain access to $p(y|K)$ in $O(KN^2)$ time.

Author's personal copy



Computing matrix $A$ requires to integrate the likelihood over tree structure $T$ and parameters $\theta$ for all possible segments $r \subseteq [\![1; N]\!]$. On each segment, we fall back to the tree-structured model described in Sect. 2.2 and the integrated likelihood can be expressed using the local terms computed on vertices and edges that were defined in (5). Indeed, for $r \subset [\![1; N]\!]$, $p(y^r)$ is obtained through Proposition 2 applied to $\omega^{(r)}$ (defined in (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 $\{\omega^{(r)}\}_{r \subseteq [\![1;N]\!]}$, hence the total complexity. □

The components of 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 Sect. 3.1. This integration comes down to remove node $\theta$ in the global graphical model displayed in Fig. 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.

## 4 Quantities of interest

### 4.1 Change-point location

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

$$\mathcal{B}_{K,k}(t) := \{m \in \mathcal{M}_K | \tau_k = t\},$$

$$\mathcal{B}_K(t) := \bigcup_{k=1}^{K} \mathcal{B}_{K,k}(t),$$

$$\mathcal{S}_{K,k}\left([\![s; t]\!]\right) := \{m \in \mathcal{M}_K | r_k = [\![s; t]\!]\}$$

$$\mathcal{S}_K\left([\![s; t]\!]\right) := \bigcup_{k=1}^{K} \mathcal{S}_{K,k}\left([\![s; t]\!]\right).$$

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

Rigaill et al. (2012) showed that, with the convention that $\left[A^0\right]_{t1,t2} = 1$ for all $1 \leqslant t_1 < t_2 \leqslant N+1$, these probabilities could be expressed as

$$B_{K,k}(t) = \frac{\left[A^k\right]_{1,t}\left[A^{K-k}\right]_{t,N+1}}{\left[A^k\right]_{1,N+1}},$$

$$B_K(t) = \sum_{k=1}^{K-1} B_{K,k}(t),$$

$$S_{K,k}([\![s; t]\!]) = \frac{\left[A^{k-1}\right]_{1,s} A_{s,t}\left[A^{K-k}\right]_{t,N+1}}{\left[A^k\right]_{1,N+1}},$$

$$S_K([\![s; t]\!]) = \sum_{k=1}^{K} S_{K,k}\left([\![s; t]\!]\right).$$

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

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

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

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

### 4.2 Number of segments

The posterior distribution on $K$ can also be derived from Proposition 1.

**Proposition 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 1, $p(y|K) = [A^K]_{1,N+1}/C_K(a)$. □

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

$$\hat{K}_1 = \arg\max_K p(K|y),$$

$$\hat{K}_2 = K(\arg\max_m p(m|y)).$$

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

Author's personal copy



### 4.3 Posterior edge probability

For any segment $r \subseteq [\![1; N]\!]$, 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 $T_1, \ldots, T_K$. 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 5** *The instant posterior probability of edge $\{i, j\}$ at time $t$ is given by*

$$\mathbf{p}_{ij}^K(t) = \sum_{r \ni t} S_K(r) P(\{i, j\} \in E_T | y^r). \tag{10}$$

$\{p_{ij}^K(t)\}_{\substack{1 \leqslant i, j \leqslant p \\ 1 \leqslant t \leqslant N}}$ *can be computed in* $O\left(\max(K, p^3) N^2\right)$ *time from $A$ and* $\{\omega^{(r)}\}_{r \subseteq [\![1; N]\!]}$.

*Proof* This formula is similar to the one giving the posterior mean of the signal in (Rigaill 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. $\{S_K(r)\}_{r \in [\![1; N]\!]}$ is obtained with complexity $O(KN^2)$ and $\{P(\{i, j\} \in E_T | y^r)\}_{r \in [\![1; N]\!]}$ with complexity $O(p^3 N^2)$, and that gives an upper bound on total complexity. □

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 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.

## 5 Edge status and structure comparisons

We now turn to the specific case where $m$ is known and has $K$ segments $(r_1, \ldots, r_K)$. 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.

### 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

$$\begin{aligned}
\mathcal{E}_{ij}^+ &= \{\mathbf{T} = (T_1, \ldots, T_K) | \forall k \in [\![1; K]\!], \{i, j\} \in E_{T_k}\}, \\
\mathcal{E}_{ij}^- &= \{\mathbf{T} = (T_1, \ldots, T_K) | \forall k \in [\![1; K]\!], \{i, j\} \notin E_{T_k}\}, \\
\mathcal{E}_{ij} &= \mathcal{E}_{ij}^+ \cup \mathcal{E}_{ij}^-,
\end{aligned}$$

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 $(q_0^-, \overline{q_0}, q_0^+)$ respectively denote the prior probabilities of $\mathcal{E}_{ij}^-, \overline{\mathcal{E}_{ij}}$ and $\mathcal{E}_{ij}^+$. These probabilities can be written as

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

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

Posterior probabilities $(q^-, \overline{q}, q^+)$ 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:

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

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 $(\lambda^-, \overline{\lambda}, \lambda^+)$ and explicitly controlling the status of edge $\{i, j\}$ in all trees:

Author's personal copy



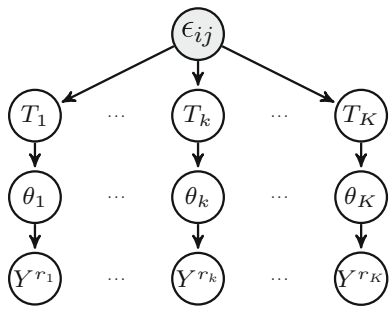



**Fig. 3** Model for edge status comparison

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

We obtain the model described in Fig. 3, in which

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

**Proposition 6** *The vector of posterior probabilities for $\epsilon_{ij}$ is proportional to* $\left(\lambda^{-}\frac{q^{-}}{q_0^{-}}, \overline{\lambda}\frac{\overline{q}}{\overline{q_0}}, \lambda^{+}\frac{q^{+}}{q_0^{+}}\right)$.

*Proof* We have that

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

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

### 5.2 Structure comparison

The same reasoning can be applied for the global event

$$\mathcal{E} = \{\mathbf{T} = (T_1, \ldots, T_K) | \exists T \in \mathcal{T}, \quad \forall k \in [\![1; K]\!], T_k = T\},$$

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

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

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

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

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

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

where $\bigodot$ 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 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 6. □

## 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?

### 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

Author's personal copy



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 Erdös-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 as 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

$$\begin{aligned} &\{\Lambda_r\}_{r \in m} \text{ i.i.d.}, && \Lambda_r \sim \mathcal{W}(\alpha, \phi), \\ &\{Y_t\}_{t=1}^N \text{ independent}, && Y^t \sim \mathcal{N}(\mathbf{0}_p, \Lambda_r), \quad \forall t \in r. \end{aligned} \tag{11}$$

where $\mathcal{W}(\alpha, \phi)$ stands for the Wishart distribution with $\alpha$ degrees of freedom and scale matrix $\phi$. The second one is the corresponding model with tree-structure assumption, as described in Sect. 3.1, and given by

$$\begin{aligned} &\{T_k\}_{k=1}^K \text{ i.i.d.}, && T_k \sim \mathcal{U}(\mathcal{T}), \\ &\{\Lambda_r\}_{r \in m} \text{ independent}, && \Lambda_r \sim h\mathcal{W}(\alpha, \phi, T_{\kappa(r|m)}) \\ &\{Y_t\}_{t=1}^N \text{ independent}, && Y^t \sim \mathcal{N}(\mathbf{0}_p, \Lambda_r), \quad \forall t \in r, \end{aligned} \tag{12}$$

where we let $h\mathcal{W}(\alpha, \phi, T)$ denote the hyper-Wishart distribution based on $\mathcal{W}(\alpha, \phi)$ and with structure $T$ (Schwaller et al. 2015). In both cases, we set $\alpha = p + 10$ and $\phi = (\alpha - p - 1) \cdot \mathbf{I}_p$, 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 $[\![1; 10]\!]$. Results for other prior distributions on $K$ are presented in the supplementary material.

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 Erdös-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 $r$.

### 6.2 Results

*Change-point location* 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 (Fig. 4). In both cases, change-points are hardly retrieved in the high-density Erdös-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 ($K = 10$, displayed in the supplementary material) 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 Fig. 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 Erdös-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 $t \in [\![1; N]\!]$, we computed the posterior edge probability matrix defined in (10) for $K = 4$. Figure 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 Sect. 5. We set the prior probability to change status at $\overline{\lambda} = 0.5$ 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

Author's personal copy



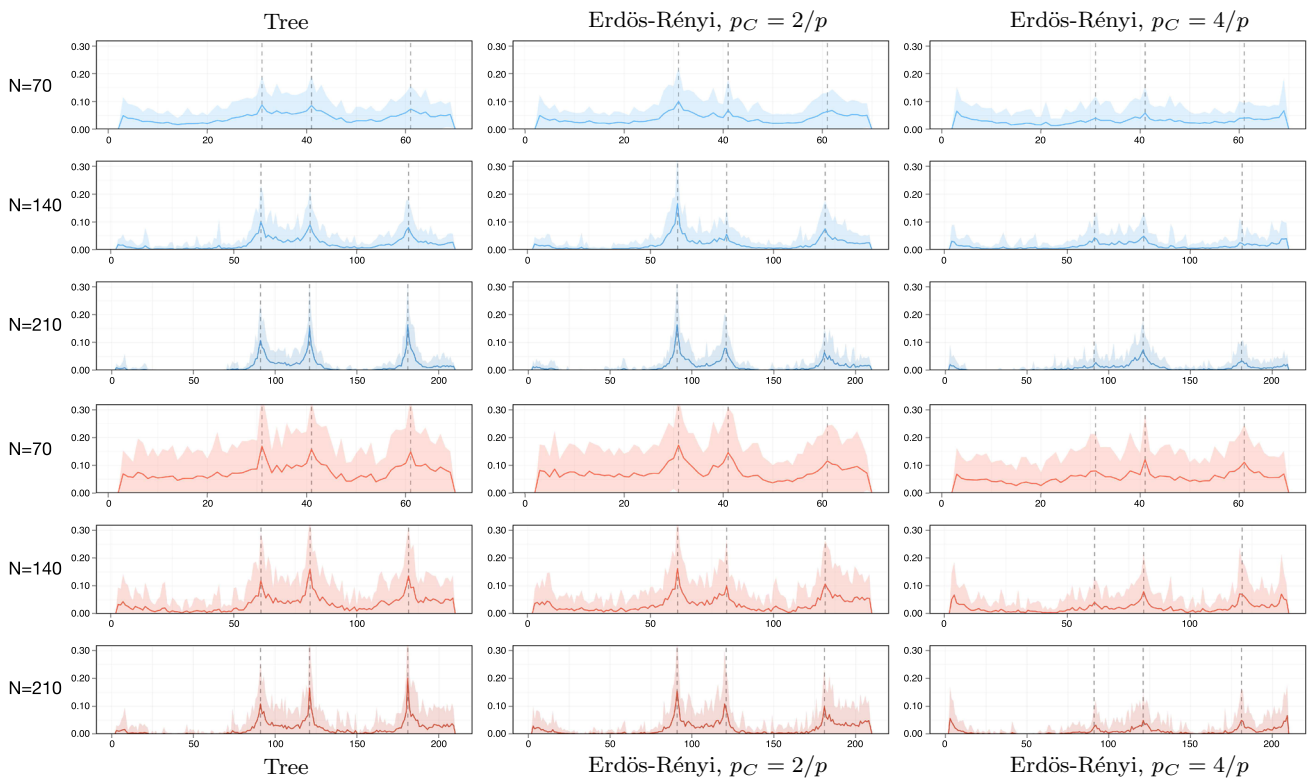



**Fig. 4** Posterior probability of observing a change-point for the tree-structured model (*blue*) and for the full model (*red*). The *curve* represents the mean value obtained from the 100 samples and the ribbon gives the standard deviation. (Color figure online)

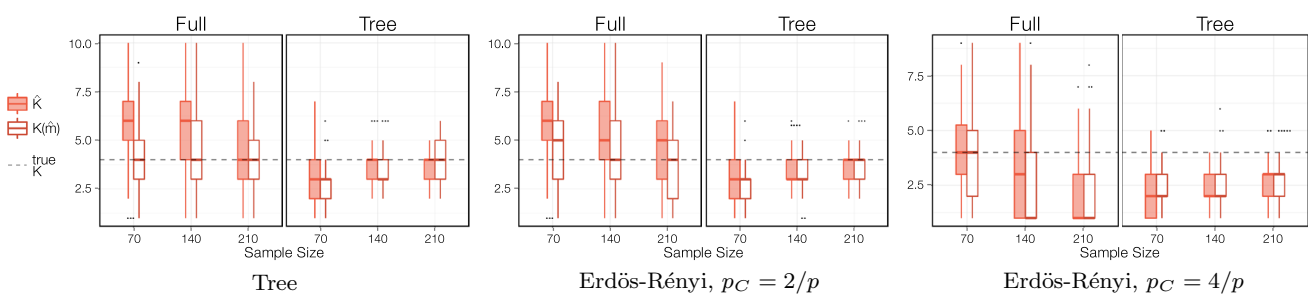



**Fig. 5** Boxplot of $\hat{K} = \arg\max_K p(K|y)$ and $K(\hat{m}) = K(\arg\max_m p(m|y))$ against sample size $N$ for the full model (full) and the tree-structured model (tree)

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 (Fig. 7).

## 7 Applications

### 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 Sect. 6, we obtained poor results (Fig. 8a), probably because of the small number of time-

Author's personal copy



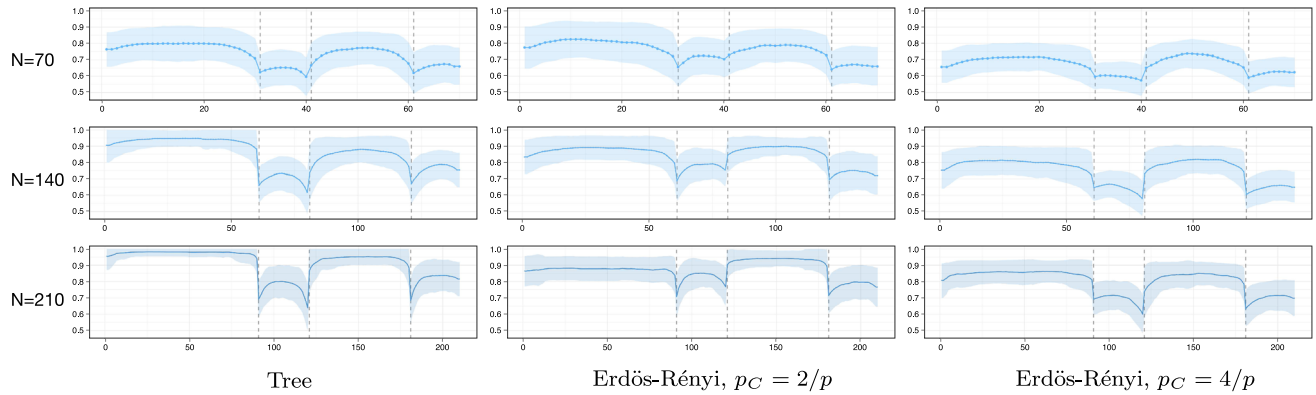



**Fig. 6** Area under the ROC curve computed for the posterior edge probability matrix $\left[\mathbf{p}_{ij}^{K}(t)\right]_{i,j=1}^{p}$ with respect to the true adjacency matrix at time $t$. We set $K$ to the true number of segments ($K = 4$). The *curve* represents the mean value obtained from the 100 samples and the ribbon gives the standard deviation

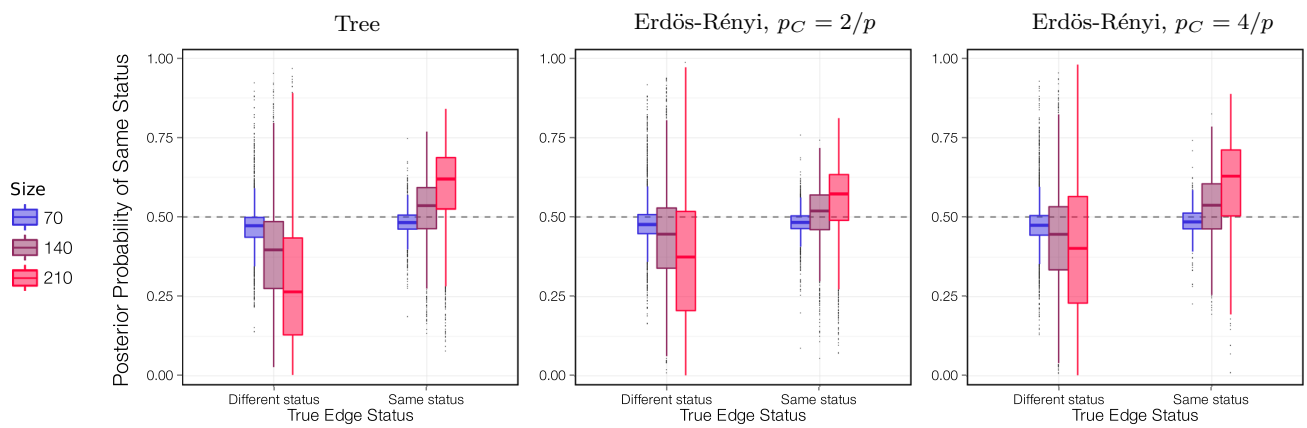



**Fig. 7** Boxplot of the posterior probability for an edge to have the same status throughout the time-series. Edges were separated according to their true status (either identical in all graphs or not). Each boxplot aggregates the results for all edges with a given status and all datasets. (Color figure online)

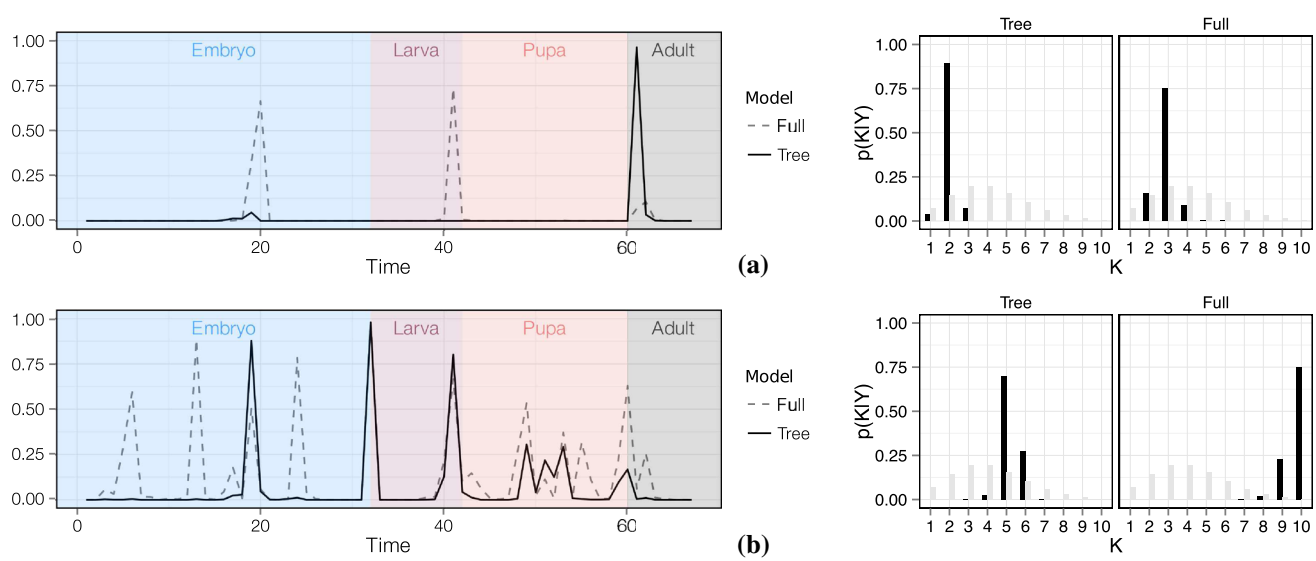



**Fig. 8** Posterior probability of a change-point occurring at time $t$ as a function of time integrated on $K$ (left) and posterior distribution for $K$ (right) for the full (full) and tree-structured (tree) models. **a** Naive prior. **b** Data-driven prior

Author's personal copy



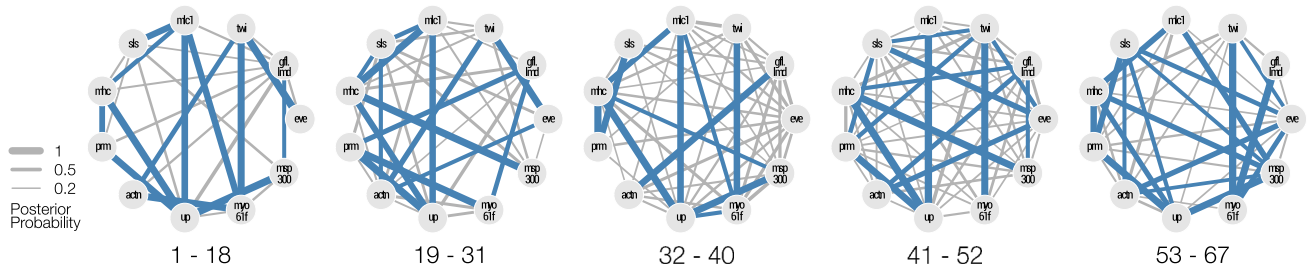



**Fig. 9** Graphical representation of posterior edge probability matrix for each segment of the best segmentation with 5 segments. The width of an edge is proportional to its posterior probability. Edges with probability higher than 0.5 are coloured in *blue*. Edges with probability lower than 0.2 were not represented. (Color figure online)

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 expectancy $(\mathbf{0}_p, \Sigma_y)$. We then obtained the results given in Fig. 8b. 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 (Fig. 8b). 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 pulpa stage. These results are nonetheless encouraging. For each segment $r$ of $\hat{m}_5$, we computed the posterior edge probability matrix given by $(P(\{i, j\} \in E_T|y^r))_{1 \leqslant i,j \leqslant p}$. 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 Sect. 5. We give a graphical representation of the results in Fig. 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 Fig. 8b 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).

### 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

$$p(y^r) = \sum_{T \in \mathcal{T}} \prod_{u=1}^{U} \left[ \int \prod_{t \in r} p(y^{t,u}|\theta_u) p(\theta_u|T) d\theta_u \right] \tag{13}$$

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

Author's personal copy



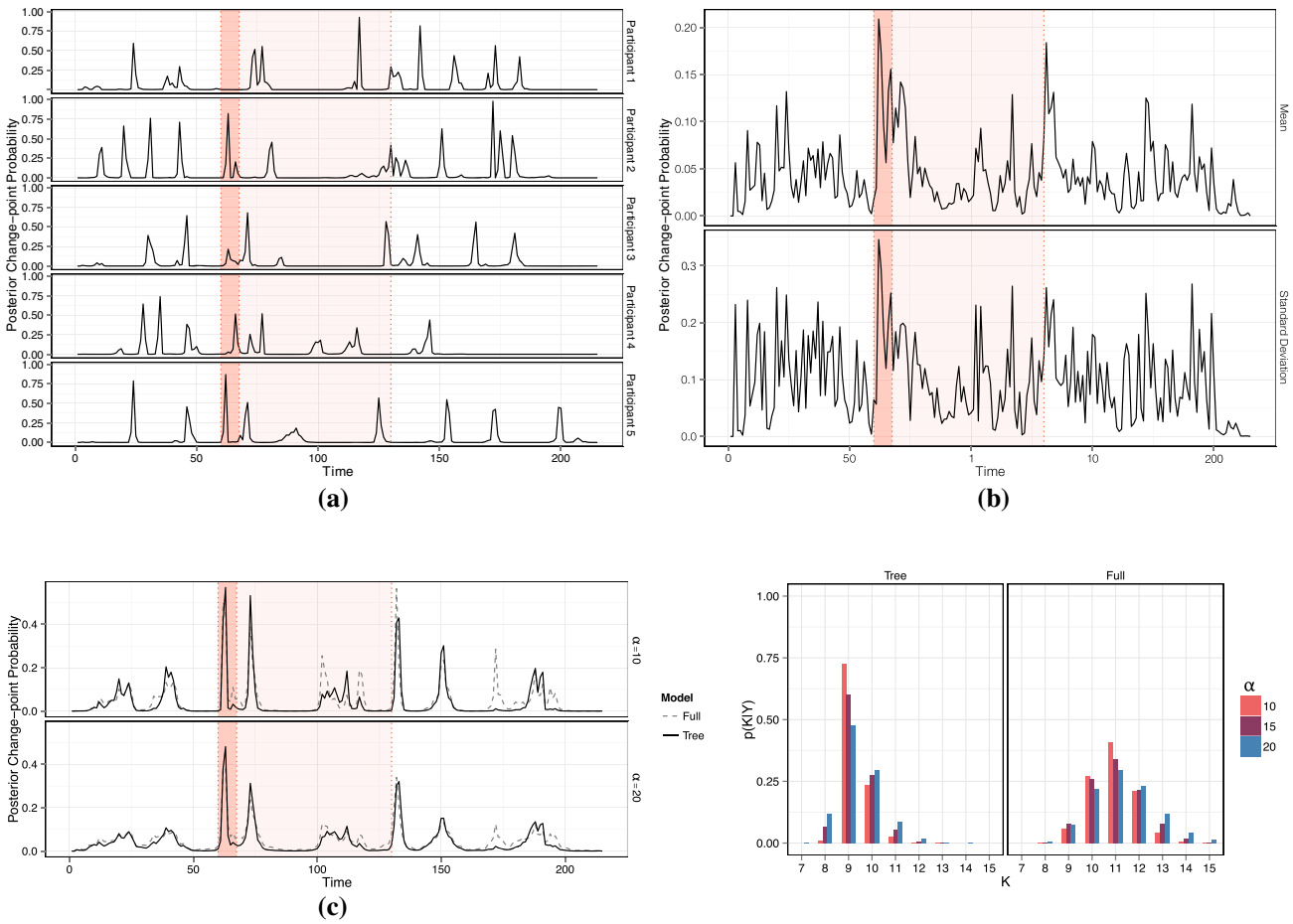



**Fig. 10** Change-point location for the fMRI data. During the dark red interval, the subject of the speech was revealed to participants, who prepared their speech during the light red interval. This preparation is ended by a statement saying that they would not have to give the speech. **a** Posterior change-point probability for five participants with the tree-structured model. **b** Mean and standard deviation of posterior change-point probability across participants with the tree-structured model. **c** Posterior change-points probability (left) and posterior distribution on $K$ (right) when participants are jointly considered, with likelihood tempered at a level $\alpha$ for the full (full) and tree-structured (tree) models. (Color figure online)

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 $\log(A) = [\log(A_{s,t})]_{1\leqslant s,t\leqslant N+1}$, integrating over segmentations requires the actual computation of matrix $A$. Thus, the exponentiation of the segment log-likelihood matrix leads to other numerical issues.

Our pragmatic answer to these issues was to considered a tempered version of the likelihood given in (13):

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

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.

Figure 10a, b 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 Fig. 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 Fig. 10c gives the posterior distrib-

Author's personal copy



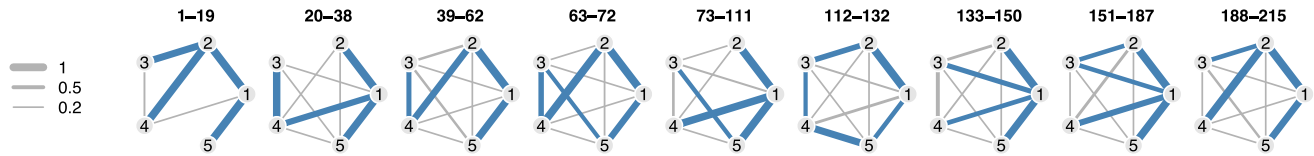



**Fig. 11** Graphical representation of posterior edge probability matrix for each segment of the best segmentation with $K = 9$ segments on fMRI data with non-tempered likelihood. The width of an edge is proportional to its posterior probability. Edges with probability higher than 0.5 are coloured in *blue*. (Color figure online)

ution 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 Fig. 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.

## 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 Sect. 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 other, 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** The authors thank Ivor Cribben (Alberta School of Business, Canada) for kindly providing the fMRI data. They also thank Sarah Ouadah (AgroParisTech, INRA, Paris, France) for fruitful discussions.

## References

Arbeitman, M.N., Furlong, E.E.M., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, Ma., Scott, M.P., Davis, R.W., White, K.P.: Gene expression during the life cycle of *Drosophila melanogaster*. Science **297**(5590), 2270–2275 (2002). doi:10.1126/science.1072152. (ISSN: 1095-9203)

Auger, I.E., Lawrence, C.E.: Algorithms for the optimal identification of segment neighborhoods. Bull. Math. Biol. **51**(1), 39–54 (1989). doi:10.1007/BF02458835. (ISSN: 00928240)

Barber, D., Cemgil, A.: Graphical models for time-series. IEEE Signal Process. Mag. (2010). doi:10.1109/MSP.2010.938028**(ISSN: 1053-5888)**

Barry, D., Hartigan, J.A.: Product partition models for change point problems. Ann. Stat. **20**(1), 260–279 (1992). doi:10.1214/aos/1176348521. (ISSN: 0090-5364)

Caron, F., Doucet, A., Gottardo, R.: On-line changepoint detection and parameter estimation with application to genomic data. Stat. Com-

Author's personal copy



put. **22**(2), 579–595 (2012). doi:10.1007/s11222-011-9248-x. (ISSN: 09603174)

Cribben, I., Haraldsdottir, R., Atlas, L.Y., Wager, T.D., Lindquist, M.A.: Dynamic connectivity regression: determining state-related changes in brain connectivity. NeuroImage **61**(4), 907–920 (2012). doi:10.1016/j.neuroimage.2012.03.070. (ISSN: 10538119)

Dondelinger, F., Lèbre, S., Husmeier, D.: Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. Mach. Learn. **90**(2), 191–230 (2013). (ISSN: 08856125)

Fearnhead, P.: Exact and efficient Bayesian inference for multiple changepoint problems. Stat. Comput. **16**(2), 203–213 (2006). doi:10.1007/s11222-006-8450-8. (ISSN: 09603174)

Fearnhead, P., Liu, Z.: On-line inference for multiple changepoint problems. J. R. Stat. Soc. Ser B Stat. Methodol. **69**(4), 589–605 (2007). doi:10.1111/j.1467-9868.2007.00601.x. (ISSN: 13697412)

Fox, E., Sudderth, E., Jordan, M., Willsky, A.: Nonparametric Bayesian learning of switching linear dynamical systems. Adv. Neural Inf. Process. Syst. **21**, 457–464 (2009)

Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 139–147, (1998). doi:10.1111/j.1469-7580.2008.00962.x

Grzegorczyk, M., Husmeier, D.: 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). doi:10.1093/bioinformatics/btq711. (ISSN: 1367-4803)

Kolar, M., Song, L., Ahmed, A., Xing, E.P.: Estimating time-varying networks, **4**. (2010). doi:10.1214/09-AOAS308 (ISBN: 0001409107)

Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996). (ISBN: 0-19-852219-3)

Lèbre, S., Becq, J., Devaux, F., Stumpf, M.P.H., Lelandais, G.: Statistical inference of the time-varying structure of gene-regulation networks. BMC Syst. Biol. **4**, 130 (2010). doi:10.1186/1752-0509-4-130. (ISSN: 1752-0509)

Meilă, M., Jaakkola, T.: Tractable bayesian learning of tree belief networks. Stat. Comput. **16**(1), 77–92 (2006)

Murphy, K., Mian, S.: Modelling Gene Expression Data Using Dynamic Bayesian Networks. Technical report (1999)

Rigaill, G., Lebarbier, E., Robin, S.: Exact posterior distributions and model selection criteria for multiple change-point detection problems. Stat. Comput. **22**(4), 917–929 (2012). doi:10.1007/s11222-011-9258-8. (ISSN: 0960-3174)

Robinson, J., Hartemink, A.: Learning non-stationary dynamic Bayesian networks. J. Mach. Learn. Res. **11**, 3647–3680 (2010). (ISSN: 1532-4435)

Schwaller, L., Robin, S., Stumpf, M.: Bayesian Inference of Graphical Model Structures Using Trees. http://arxiv.org/abs/1504.02723 (2015)

Siracusa, M.R.: Tractable Bayesian inference of time-series dependence structure. In: Proceedings of the 23th International Conference on Artificial Intelligence and Statistics, vol. 5, pp. 528–535. (2009)

Xuan, X., Murphy, K.: Modeling changing dependency structure in multivariate time series. In: Proceedings of the 24th International Conference on Machine Learning (2007), vol. 227, issue id m, pp. 1055–1062, (2007). doi:10.1145/1273496.1273629**(ISSN: 1595937935)**

Zhao, W., Serpedin, E., Dougherty, E.R.: Inferring gene regulatory networks from time series data using the minimum description length principle. Bioinformatics **22**(17), 2129–2135 (2006). doi:10.1093/bioinformatics/btl364. (ISSN: 13674803)

Zhou, S., Lafferty, J., Wasserman, L.: Time varying undirected graphs. Mach. Learn. **80**, 295–319 (2010)