Mach Learn (2013) 90:191–230
DOI 10.1007/s10994-012-5311-x

Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure

Frank Dondelinger · Sophie Lèbre · Dirk Husmeier

Received: 16 September 2011 / Revised: 19 March 2012 / Accepted: 15 June 2012 /
Published online: 18 July 2012

Abstract The proper functioning of any living cell relies on complex networks of gene regulation. These regulatory interactions are not static but respond to changes in the environment and evolve during the life cycle of an organism. A challenging objective in computational systems biology is to infer these time-varying gene regulatory networks from typically short time series of transcriptional profiles. While homogeneous models, like conventional dynamic Bayesian networks, lack the flexibility to succeed in this task, fully flexible models suffer from inflated inference uncertainty due to the limited amount of available data. In the present paper we explore a semi-flexible model based on a piecewise homogeneous dynamic Bayesian network regularized by gene-specific inter-segment information sharing. We explore different choices of prior distribution and information coupling and evaluate their performance on synthetic data. We apply our method to gene expression time series obtained during the life cycle of *Drosophila melanogaster*, and compare the predicted segmentation with other state-of-the-art techniques. We conclude our evaluation with an ap-

Editor: James Cussens.

Software: R code for all models described in this paper is available from http://www.bioss.ac.uk/students/frankd.html, and will be made available as an R package on the Comprehensive R Archive Network (CRAN) in the near future.

F. Dondelinger · D. Husmeier
Biomathematics and Statistics Scotland, JCMB, Edinburgh EH9 3JZ, UK

F. Dondelinger
Institute for Adaptive and Neural Computation, The University of Edinburgh, Edinburgh EH8 9AB, UK
e-mail: frankd@bioss.ac.uk

S. Lèbre
LSIIT, UMR 7005, Université de Strasbourg, 67412 Illkirch, France
e-mail: slebre@unistra.fr

D. Husmeier (✉)
School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QW, UK
e-mail: dirk.husmeier@glasgow.ac.uk

plication to synthetic biology, where the objective is to predict an in vivo regulatory network of five genes in *Saccharomyces cerevisiae* subjected to a changing environment.

Keywords Dynamic Bayesian networks · Hierarchical Bayesian models · Multiple changepoint processes · Reversible jump Markov chain Monte Carlo · Gene expression time series · Systems and synthetic biology

1 Introduction

One of the challenging problems in the field of systems biology is the inference of gene regulatory networks from high-throughput transcriptomic profiles, as obtained e.g. with microarrays or next generation sequencing. While protein interactions can be measured directly with various high-throughput assays (e.g. yeast-2-hybrid or phage display), gene regulatory interactions involve several intermediate steps related to the formation, activation and complex formation of transcription factors (e.g. via phosphorylation or dimerization). These processes are not observable at the transcriptional level. For that reason the inference of interactions has to be based on indirect noisy measurements of mRNA concentrations (a proxy for gene activity), rendering the problem of regulatory network reconstruction more difficult than for proteins. Various statistical techniques aim to perform network inference on this data, and the reconstructed regulation networks can reveal how the genes and the proteins they code for interact. However, many of the regulatory interactions in the cell vary in time. During the development and growth of an organism, some genes and pathways are more active during the early stages, but show practically no activity during the later stages, or vice-versa. *Drosophila melanogaster*, for instance, goes through several developmental stages, from embryo to larva to pupa to adult. Genes involved in wing muscle development would naturally fulfill different roles during the embryonal phase, when no wings are present, than they do in the adult fly, when the wings have fully developed. Another instance in which the gene regulatory network varies in time is in reaction to an environmental trigger, such as the type of growth substrate. Such a trigger can enhance or prevent the interactions of certain genes, which in turn can have repercussions for the whole gene network.

We are therefore presented with the problem of inferring a regulatory network from a series of discrete measurements or observations in time, where the structure of the network is subject to potential change. Moreover, we may not always know at which stage structural changes are likely to occur, as the underlying processes may be time-delayed or dependent on unobservable external factors. To extend conventional reverse engineering methods, which only aim to infer a single immutable regulatory network, our work builds on recent research in combining dynamic Bayesian networks (DBNs) with multiple changepoint processes (Robinson and Hartemink 2009, 2010; Grzegorczyk and Husmeier 2009, 2011; Lèbre 2007; Lèbre et al. 2010; Kolar et al. 2009). Below, we will briefly review the state of the art and the shortcomings of existing methods that we aim to address.

The standard assumption underlying DBNs is that time-series have been generated from a homogeneous Markov process. This assumption is too restrictive, as discussed above, and can potentially lead to erroneous conclusions. While there have been various efforts to relax the homogeneity assumption for undirected graphical models (Talih and Hengartner 2005; Xuan and Murphy 2007), relaxing this restriction in DBNs is a more recent research topic (Robinson and Hartemink 2009, 2010; Grzegorczyk and Husmeier 2009, 2011; Ahmed and Xing 2009; Lèbre 2007; Lèbre et al. 2010; Kolar et al. 2009). At present, none of the proposed methods is without its limitations, leaving room for further methodological

innovation. The method proposed in Ahmed and Xing (2009) and Kolar et al. (2009) is non-Bayesian. This requires certain regularization parameters to be optimized "externally", by applying information criteria (like AIC or BIC), cross-validation or bootstrapping. The first approach is suboptimal, the latter approaches are computationally expensive.[1] In the present paper we therefore follow the Bayesian paradigm, as in Robinson and Hartemink (2009, 2010), Grzegorczyk and Husmeier (2009, 2011), Lèbre (2007) and Lèbre et al. (2010). These approaches also have their limitations. The method proposed in Grzegorczyk and Husmeier (2009, 2011) assumes a fixed network structure and only allows the interaction parameters to vary with time. This assumption is too rigid when looking at processes where changes in the overall regulatory network structure are expected, e.g. in morphogenesis or embryogenesis. The method proposed in Robinson and Hartemink (2009, 2010) requires a discretization of the data, which incurs an inevitable information loss. These limitations are addressed in Lèbre (2007) and Lèbre et al. (2010), where the authors propose a method for continuous data that allows network structures associated with different nodes to change with time in different ways. However, this high flexibility causes potential problems when applied to time series with a low number of measurements, as typically available from systems biology, leading to overfitting or inflated inference uncertainty.

The objective of the present paper is to propose a novel model that addresses the methodological shortcomings of the three Bayesian methods mentioned above, and to demonstrate its viability by application to gene expression time series from *Drosophila melanogaster* and *Saccharomyces cerevisiae*. Unlike Robinson and Hartemink (2009, 2010), our model is continuous and therefore avoids the information loss inherent in a discretization of the data. We further improve on the model in Robinson and Hartemink (2009, 2010) by allowing for different penalties for changing edges and non-edges in the network, and by allowing different nodes in the networks to have different penalty terms. Unlike Grzegorczyk and Husmeier (2009, 2011), our model allows the network structure to change among segments, leading to greater model flexibility. As an improvement on Lèbre (2007) and Lèbre et al. (2010), our model introduces information sharing among time series segments, which provides an essential regularization effect. We have applied the model to reconstruct two regulatory networks: a network of genes involved in wing muscle development during the life cycle of *Drosophila melanogaster* (Arbeitman et al. 2002), and an engineered network from synthetic biology, consisting of five genes in *Saccharomyces cerevisiae* (Cantone et al. 2009).

The present paper follows on from two earlier conference papers of ours (Dondelinger et al. 2010; Husmeier et al. 2010). In Dondelinger et al. (2010), we compared two different information coupling paradigms: global information coupling and sequential information coupling. Global information coupling is appropriate when there is no natural sequential order of the time series segments, such as for segments derived from different experimental conditions. Sequential information sharing, which we investigated in more detail in Husmeier et al. (2010) and in the present paper, is appropriate for modelling a temporal developmental process, such as those related to morphogenesis, where changes to the network structure happen sequentially.

The present paper extends Dondelinger et al. (2010) and Husmeier et al. (2010) in several respects. Firstly, restricted by a strict page limit, our earlier papers were rather terse. The present paper provides a more comprehensive exposition of the methodology, which is self-contained. Secondly, we have explored different versions of information coupling (hard versus soft) and functional forms of the prior (exponential versus binomial). In Husmeier

[1] See Larget and Simon (1999) for a demonstration of the higher computational costs of bootstrapping over Bayesian approaches based on MCMC.

et al. (2010), not all combinations of strength versus functional form were investigated, and we have completed these combinations in our present work. Thirdly, we have improved the MCMC scheme. In our earlier work, a standard Metropolis-Hastings-Green (RJMCMC) sampler was employed. In the present work we have identified several scenarios where this sampler is bound to fail, and we propose a new type of MCMC proposal move. We show that these new moves avoid the convergence problems encountered with the original sampler, leading to a substantial improvement in mixing. Fourthly, the Bayesian hierarchical models that we propose depend on various hyperparameters. As opposed to our earlier work, we have investigated the influence of the higher level hyperparameters. To this end, we have first carried out a set of simulation studies for the proposed model. To substantiate our findings, we have then additionally carried out semi-analytical investigations for a simplified scenario, in which the computation of the marginal likelihood is tractable (see Sect. 5.2). Fifthly, we have rerun all our earlier simulations to understand the effect of model choice, unconfounded by MCMC mixing problems, and we have improved the interpretation of the results for the real-world problems.

We note that while we were extending our earlier work of Husmeier et al. (2010), a somewhat related paper has been published: Wang et al. (2011). While methodologically similar, there is an important difference in the application and inference, though. The objective of Wang et al. (2011) is online parameter estimation via particle filtering, with applications e.g. in tracking. This is a different scenario from most systems biology applications, where an interaction structure is typically learnt off-line after completion of a series of high-throughput experiments. Unlike Wang et al. (2011), our work thus follows other applications of DBNs in systems biology (Robinson and Hartemink 2009, 2010; Grzegorczyk and Husmeier 2009, 2011; Lèbre 2007; Lèbre et al. 2010; Kolar et al. 2009) and aims to infer the model structure by marginalizing out the parameters in closed form. To paraphrase this: while inference in Wang et al. (2011) is based on a filter, inference in our work is based on a smoother.

Our paper is organized as follows. Section 2 reviews the non-homogeneous DBN on which our work is based. Section 3 describes the methodological innovation of Bayesian regularization via information coupling. Section 4 describes the implementation of our method and the setup of the simulation studies. Section 5 discusses results obtained on synthetic data, with an investigation of the influence of the hyperparameters. Section 6 describes and interprets two real-world applications, related to morphogenesis in *Drosophila melanogaster* and synthetic biology in *Saccharomyces cerevisiae*. The paper concludes in Sect. 7 with a general discussion and summary.

2 Background: non-homogeneous DBNs

This section summarizes the auto regressive time-varying DBN proposed in Lèbre (2007) and Lèbre et al. (2010). A similar model was proposed in Punskaya et al. (2002). The idea is to combine the Bayesian regression model of Andrieu and Doucet (1999) with multiple changepoint processes and pursue Bayesian inference with reversible jump Markov chain Monte Carlo (RJMCMC) (Green 1995). We call this method TVDBN (Time-Varying Dynamic Bayesian Network).

The model is based on the first-order Markov assumption. This assumption is not critical, though, and a generalization to higher orders, as pursued in Punskaya et al. (2002), is straightforward. The value that a node in the graph takes on at time t is determined by the values that the node's parents (i.e. potential regulators, see below) take on at the previous

time point, $t-1$. More specifically, the conditional probability of the observation associated with a node at a given time point is a conditional Gaussian distribution, where the conditional mean is a linear weighted sum of the parent values at the previous time point, and the interaction parameters and parent sets depend on the time series segment. The latter dependence adds extra flexibility to the model and thereby relaxes the homogeneity assumption. The interaction parameters, the variance parameters, the number of potential parents, the location of changepoints demarcating the time series segments, and the number of changepoints are given (conjugate) prior distributions in a hierarchical Bayesian model. For inference, all these quantities are sampled from the posterior distribution with RJMCMC. Note that a complete specification of all node-parent configurations determines the structure of a regulatory network: each node receives incoming directed edges from each node in its parent set. In what follows, we will refer to nodes as genes and to the network as a gene regulatory network. The method is not restricted to molecular systems biology, though.

2.1 Graph

Let p be the number of observed genes, and let $\boldsymbol{x} = (x_i(t))_{1\le i\le p, 1\le t\le N}$ be the expression values measured at N time points. $\boldsymbol{G}^h$ represents a directed graph, i.e. the network defined by a set of directed edges among the p genes. $\boldsymbol{G}_i^h$ is the subnetwork associated with target gene i, determined by the set of its parents, i.e. the nodes with a directed edge feeding into gene i; these are the potential regulators of the target gene. The meaning of the superscript h is explained in the next section.

2.2 Multiple changepoint process

The set of regulatory relationships among the genes, defined by $\boldsymbol{G}^h$, may vary across time, which we model with a multiple changepoint process. For each target gene i, an unknown number k_i of changepoints define k_i+1 non-overlapping segments. Segment $h = 1, .., k_i+1$ starts at changepoint ξ_i^{h-1} and stops before ξ_i^h, where $\boldsymbol{\xi}_i = (\xi_i^0, \ldots, \xi_i^{h-1}, \xi_i^h, \ldots, \xi_i^{k_i+1})$ with $\xi_i^{h-1} < \xi_i^h$. To delimit the bounds, two pseudo-changepoints are introduced: $\xi_i^0 = 2$ and $\xi_i^{k_i+1} = N+1$. Thus vector $\boldsymbol{\xi}_i$ has length $|\boldsymbol{\xi}_i| = k_i+2$. The set of changepoints is denoted by $\boldsymbol{\xi} = (\boldsymbol{\xi}_i)_{1\le i\le p}$. This changepoint process induces a partition of the time series, $\boldsymbol{x}_i^h = (x_i(t))_{\xi_i^{h-1}\le t<\xi_i^h}$, with different network structures $\boldsymbol{G}_i^h$ associated with the different segments $h \in \{1, \ldots, k_i+1\}$. Identifiability is satisfied by ordering the changepoints based on their position in the time series. We define $\boldsymbol{G}_i = \{\boldsymbol{G}_i^h\}_{1\le h\le k_i+1}$ and $\boldsymbol{G} = \{\boldsymbol{G}_i\}_{1\le i\le p}$.

2.3 Regression model

For each gene i, the random variable $X_i(t)$ refers to the expression of gene i at time t. Within any segment h, the expression of gene i depends on the p gene expression values measured at the previous time point through a regression model defined by (a) a set of s_i^h parents denoted by $\boldsymbol{G}_i^h = \{j_1, \ldots, j_{s_i^h}\} \subseteq \{1, \ldots, p\}$, $|\boldsymbol{G}_i^h| = s_i^h$, and (b) a set of parameters $(\boldsymbol{a}_i^h,\ \sigma_i^h)$ where $\boldsymbol{a}_i^h = (a_{ij}^h)_{0\le j\le p}$, $a_{ij}^h \in \mathbb{R}$ and $\sigma_i^h > 0$. For all $j \neq 0$, $a_{ij}^h = 0$ if $j \notin \boldsymbol{G}_i^h$. For each gene i, for each time point t in segment h ($\xi_i^{h-1} \le t < \xi_i^h$), the random variable $X_i(t)$ depends on the p variables $\{X_j(t-1)\}_{1\le j\le p}$ according to

$$X_i(t) = a_{i0}^h + \sum\nolimits_{j\in\boldsymbol{G}_i^h} a_{ij}^h\, X_j(t-1) + \varepsilon_i^h(t) \tag{1}$$

where the noise $\varepsilon_i^h(t)$ is assumed to be Gaussian with mean 0 and variance $(\sigma_i^h)^2$, $\varepsilon_i^h(t) \sim N(0,(\sigma_i^h)^2)$. We define $\boldsymbol{a}_i = (\boldsymbol{a}_i^h)_{1\le h\le k_i+1}$, $\boldsymbol{a} = (\boldsymbol{a}_i)_{0\le i\le p}$, $\boldsymbol{\sigma_i^2} = (\sigma_i^h)_{1\le h\le k_i+1}^2$ and $\boldsymbol{\sigma^2} = (\boldsymbol{\sigma_i^2})_{0\le i\le p}$.

2.4 Prior

The $k_i + 1$ segments are delimited by k_i changepoints, where k_i is distributed a priori as a truncated Poisson random variable with mean λ and maximum $\overline{k} = N - 2$:

$$P(k_i|\lambda) \propto \frac{\lambda^{k_i}}{k_i!}\mathbb{1}_{\{k_i\le\overline{k}\}}; \qquad P(\boldsymbol{k}|\lambda) = \prod_{i=1}^{p} P(k_i|\lambda) \tag{2}$$

where $\boldsymbol{k} = (k_1,\ldots,k_p)$. Conditional on k_i changepoints, the changepoint position vector $\boldsymbol{\xi}_i = (\xi_i^0,\xi_i^1,\ldots,\xi_i^{k_i+1})$ takes non-overlapping integer values, which we take to be uniformly distributed a priori. There are $(N-2)$ possible positions for the k_i changepoints, thus vector $\boldsymbol{\xi}_i$ has prior density:

$$P(\boldsymbol{\xi}_i|k_i) = 1\Big/\binom{N-2}{k_i} = \frac{k_i!(N-2-k_i)!}{(N-2)!} \tag{3}$$

For each gene i, for each segment h, the number s_i^h of parents for node i follows a truncated Poisson distribution with mean Λ and maximum $\overline{s} = 5$:

$$P\big(s_i^h\big|\Lambda\big) \propto \frac{\Lambda^{s_i^h}}{s_i^h!}\mathbb{1}_{\{s_i^h\le\overline{s}\}} \tag{4}$$

Conditional on s_i^h, the prior for the parent set $\boldsymbol{G}_i^h$ is a uniform distribution over all parent sets with cardinality s_i^h,

$$P\big(\boldsymbol{G}_i^h\big|s_i^h\big) = 1\Big/\binom{p}{s_i^h} = \frac{s_i^h!(p-s_i^h)!}{p!} \tag{5}$$

The overall prior on the network structures is given by marginalization:

$$P\big(\boldsymbol{G}_i^h\big|\Lambda\big) = \sum\nolimits_{s_i^h=0}^{\overline{s}} P\big(\boldsymbol{G}_i^h\big|s_i^h\big)P\big(s_i^h\big|\Lambda\big) \tag{6}$$

Conditional on the parent set $\boldsymbol{G}_i^h$ of size s_i^h, the $s_i^h + 1$ regression coefficients form a subset of $\boldsymbol{a}_i^h$ denoted by $\boldsymbol{a}_{\boldsymbol{G}_i^h} = (a_{i0}^h, (a_{ij}^h)_{j\in\boldsymbol{G}_i^h})$. They are assumed zero-mean multivariate Gaussian with covariance matrix $(\sigma_i^h)^2\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h}$,

$$P\big(\boldsymbol{a}_i^h\big|\boldsymbol{G}_i^h,\sigma_i^h\big) = \big|2\pi\big(\sigma_i^h\big)^2\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h}\big|^{-\frac{1}{2}}\exp\left(-\frac{\boldsymbol{a}_{\boldsymbol{G}_i^h}^{\dagger}\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h}^{-1}\boldsymbol{a}_{\boldsymbol{G}_i^h}}{2(\sigma_i^h)^2}\right) \tag{7}$$

where $|.|$ denotes the determinant of a matrix, the symbol $\dagger$ denotes matrix transposition, $\boldsymbol{\Sigma}_{\boldsymbol{G}_i^h} = \delta^{-2}\boldsymbol{D}_{\boldsymbol{G}_i^h}^{\dagger}\boldsymbol{D}_{\boldsymbol{G}_i^h}$ and $\boldsymbol{D}_{\boldsymbol{G}_i^h}$ is the $(\xi_i^h - \xi_i^{h-1}) \times (s_i^h + 1)$ matrix whose first column is a vector of 1's (for the constant in model (1)) and each $(j+1)^{th}$ column contains the observed values $(x_j(t))_{\xi_i^{h-1}-1\le t<\xi_i^h-1}$ for each factor gene j in $\boldsymbol{G}_i^h$. This so-called g-prior was also

used in Andrieu and Doucet (1999) and is motivated in Zellner (1986). Finally, the conjugate prior for the variance $(\sigma_i^h)^2$ is the inverse gamma distribution, $P((\sigma_i^h)^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following Lèbre (2007) and Lèbre et al. (2010), we set the hyper-hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that give a vague distribution. The terms λ and Λ can be interpreted as the expected number of changepoints and parents, respectively, and δ^2 is the expected signal-to-noise ratio. These hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family:

$$P(\Lambda) = P(\lambda) = \mathcal{G}a(0.5, 1) = \Lambda^{-0.5}\frac{\exp(-\Lambda)}{\Gamma(0.5)} \tag{8}$$

and

$$P\left(\delta^2\right) = \mathcal{IG}(2, 0.2) = \delta^{-6}\frac{0.04\exp(-\frac{0.2}{\delta^2})}{\Gamma(2)} \tag{9}$$

2.5 Posterior

Equation (1) implies that

$$\begin{aligned} &P\left(\boldsymbol{x}_i^h \middle| \boldsymbol{G}_i^h, \boldsymbol{a}_i^h, \sigma_i^h\right) \\ &\quad = \left(\sqrt{2\pi}\sigma_i^h\right)^{-\text{length}(\boldsymbol{x}_i^h)} \exp\left(-\frac{(\boldsymbol{x}_i^h - \boldsymbol{D}_{\boldsymbol{G}_i^h}\boldsymbol{a}_{\boldsymbol{G}_i^h})^\dagger(\boldsymbol{x}_i^h - \boldsymbol{D}_{\boldsymbol{G}_i^h}\boldsymbol{a}_{\boldsymbol{G}_i^h})}{2(\sigma_i^h)^2}\right) \end{aligned} \tag{10}$$

where $\text{length}(\boldsymbol{x}_i^h)$ is the length of the time series segment h. From Bayes' theorem, the posterior is given by the following equation, where all prior distributions have been defined above:

$$\begin{aligned} &P\left(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \boldsymbol{a}, \boldsymbol{\sigma^2}, \lambda, \Lambda, \delta^2 \middle| \boldsymbol{x}\right) \\ &\quad \propto P\left(\delta^2\right) P(\lambda) P(\Lambda) \prod_{i=1}^{p} P(k_i|\lambda) P(\boldsymbol{\xi}_i|k_i) \prod_{h=1}^{k_i} P\left(\boldsymbol{G}_i^h \middle| \Lambda\right) \\ &\qquad \times P\left(\left[\sigma_i^h\right]^2\right) P\left(\boldsymbol{a}_i^h \middle| \boldsymbol{G}_i^h, \left[\sigma_i^h\right]^2, \delta^2\right) P\left(\boldsymbol{x}_i^h \middle| \boldsymbol{G}_i^h, \boldsymbol{a}_i^h, \left[\sigma_i^h\right]^2\right) \end{aligned} \tag{11}$$

An attractive feature of the chosen model is that the integration over the parameters $\boldsymbol{a}$ and $\boldsymbol{\sigma^2}$ in the posterior distribution of Eq. (11) is analytically tractable:

$$\begin{aligned} &P\left(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \lambda, \Lambda, \delta^2 \middle| \boldsymbol{x}\right) \\ &\quad = \int\int P\left(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \boldsymbol{a}, \boldsymbol{\sigma^2}, \lambda, \Lambda, \delta^2 \middle| \boldsymbol{x}\right) d\boldsymbol{a} d\boldsymbol{\sigma^2} \\ &\quad \propto P\left(\delta^2\right) P(\lambda) P(\Lambda) \prod_{i=1}^{p} \int\int P\left(k_i, \boldsymbol{\xi}_i, \boldsymbol{G}_i, \boldsymbol{a}_i, \boldsymbol{\sigma}_i^2, \boldsymbol{x}_i \middle| \lambda, \Lambda, \delta^2\right) d\boldsymbol{a}_i d\boldsymbol{\sigma}_i^2 \end{aligned} \tag{12}$$

For each gene i, the joint distribution for k_i, $\boldsymbol{\xi}_i$, $\boldsymbol{G}_i$, $\boldsymbol{a}_i$, $\boldsymbol{\sigma}_i^2$, $\boldsymbol{x}_i$ conditional on hyperparameters λ, Λ, δ^2, is integrated over the parameters $\boldsymbol{a}_i$ (normal distribution) and $\boldsymbol{\sigma}_i^2$ (inverse gamma distribution). Solving this integral (for details see Lèbre et al. 2010), the following

expression is obtained:

$$\int\int P\big(k_i, \boldsymbol{\xi}_i, \boldsymbol{G}_i, \boldsymbol{a}_i, \boldsymbol{\sigma}_i^2, \boldsymbol{x}_i \big| \lambda, \Lambda, \delta^2\big) d\boldsymbol{a}_i d\boldsymbol{\sigma}_i^2$$

$$= C_\lambda \lambda^{k_i} \frac{(N-2-k_i)!}{(N-2)!} \prod_{h=1}^{k_i+1} \left\{ \frac{(p-s_i^h)!}{p!} C_\Lambda \Lambda^{s_i^h} P\big(\boldsymbol{x}_i^h \big| \boldsymbol{G}_i^h, \delta^2\big) \right\} \tag{13}$$

where C_λ, C_Λ are the normalization constants required by the truncation of the Poisson distribution (2) and (4) and where

$$P\big(\boldsymbol{x}_i^h \big| \boldsymbol{G}_i^h, \delta^2\big) = \big(\delta^2+1\big)^{-\frac{s_i^h+1}{2}} \frac{(\frac{\gamma_0}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})} \Gamma\left(\frac{\upsilon_0 + \text{length}(\boldsymbol{x}_i^h)}{2}\right)$$

$$\times \left(\frac{\gamma_0 + (\boldsymbol{x}_i^h)^\dagger \boldsymbol{P}_i^h \boldsymbol{x}_i^h}{2}\right)^{-\frac{\upsilon_0+\text{length}(\boldsymbol{x}_i^h)}{2}} \tag{14}$$

where the matrices $\boldsymbol{P}_i^h$ and $\boldsymbol{M}_i^h$ are defined as follows, with $\boldsymbol{I}$ referring to the identity matrix of size length $(\boldsymbol{x}_i^h)$:

$$\boldsymbol{P}_i^h = \boldsymbol{I} - \boldsymbol{D}_{\boldsymbol{G}_i^h} \boldsymbol{M}_i^h \boldsymbol{D}_{\boldsymbol{G}_i^h}^\dagger, \tag{15}$$

$$\boldsymbol{M}_i^h = \frac{\delta^2}{\delta^2+1} \big(\boldsymbol{D}_{\boldsymbol{G}_i^h}^\dagger \boldsymbol{D}_{\boldsymbol{G}_i^h}\big)^{-1} \tag{16}$$

The number of changepoints $\boldsymbol{k}$ and their location, $\boldsymbol{\xi}$, the network structure $\boldsymbol{G}$ and the hyperparameters λ, Λ and δ^2 can be sampled from the posterior distribution $P(\boldsymbol{k}, \boldsymbol{\xi}, \boldsymbol{G}, \lambda, \Lambda, \delta^2 | \boldsymbol{x})$ with a reversible jump MCMC (Green 1995) scheme detailed in the next subsection.

2.6 RJMCMC scheme

Four different update moves are proposed: birth of a new changepoint (B); death (removal) of an existing changepoint (D); shift of a changepoint to a different time-point (S); and update of the network topology within the segments (N). These moves occur with probabilities b_{k_i} for B, d_{k_i} for D, u_{k_i} for S and v_{k_i} for N, depending only on the current number of changepoints k_i and satisfying $b_{k_i} + d_{k_i} + u_{k_i} + v_{k_i} = 1$. The changepoint birth and death moves represent changes from, respectively, k_i to k_i+1 segments and k_i to k_i-1 segments. In order to preserve the restriction on the number of changepoints, some probabilities are set to 0: $d_0 = u_0 = 0$ and $b_{\overline{k}} = 0$. Otherwise, following Green (1995), these probabilities are chosen as follows,

$$b_{k_i} = c \min\left\{1, \frac{P(k_i+1|\lambda)}{P(k_i|\lambda)}\right\}, \qquad d_{k_i+1} = c \min\left\{1, \frac{P(k_i|\lambda)}{P(k_i+1|\lambda)}\right\} \tag{17}$$

where $P(k_i|\lambda)$ is the prior distribution for the number of changepoints defined in Eq. (2) and the constant c is chosen to be smaller than $1/4$ so that network structure updates and changepoint position shifts are proposed more frequently than births and deaths of changepoints. This improves mixing and convergence with respect to changepoint positions and network structures within the different segments. Shifting of a changepoint is proposed with

probability $u_{k_i} = (1 - b_{k_i} - d_{k_i+1})/3$, and updating of the network structure within each segment is proposed with probability $v_{k_i} = 1 - (b_{k_i} + d_{k_i} + u_{k_i})$.

Following Green (1995), the RJMCMC acceptance probability of a changepoint birth is equal to $\min\{1, R\}$ where the acceptance ratio R reads as follows:

$$R = (\text{likelihood ratio}) \times (\text{prior ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian}) \tag{18}$$

The product of the likelihood and the prior ratio is the posterior ratio which is derived from Eq. (12). The computation of the proposal ratio and the Jacobian depends on the choice for the various moves designed to sample the time-varying network distribution. We briefly describe below the chosen moves and their associated acceptance ratio. A complete description of the computation of the acceptance ratio for each move can be found in Lèbre et al. (2010).

Let $\boldsymbol{\xi}_i$ be the current changepoint vector containing k_i changepoints. For a changepoint birth move, a new changepoint position $\xi^\star$ is sampled uniformly from the available positions. The new changepoint is within an existing segment $h^\star$ of the target gene i, $\xi_i^{h^\star - 1} < \xi^\star < \xi_i^{h^\star}$. Let us denote by $h_L^\star$ and $h_R^\star$ the segments to the left and to the right of the new changepoint respectively and by $\boldsymbol{x}_i^{h^\star} = (\boldsymbol{x}_i^{h_L^\star}, \boldsymbol{x}_i^{h_R^\star})$ the observed values for gene i in those segments. One of $h_L^\star$ and $h_R^\star$ is chosen with equal probability. That segment retains the current network topology $\boldsymbol{G}_i^{h^\star}$ of segment $h^\star$, and an entirely new topology is sampled from the prior defined in Eq. (6) for the other segment. Let us denote by $s^\star$ the number of edges of the new topology. The Jacobian is equal to 1 and the prior ratio is computed from the probability of choosing a new changepoint position and a new network structure for the new segment. Then the birth of the proposed changepoint is accepted with probability $A(\boldsymbol{\xi}_i^+ | \boldsymbol{\xi}_i) = \min\{1, R(\boldsymbol{\xi}_i^+ | \boldsymbol{\xi}_i)\}$, with

$$\begin{aligned} R\big(\boldsymbol{\xi}_i^+ \big| \boldsymbol{\xi}_i\big) = {} & \frac{1}{(\delta^2+1)^{(s^\star+1)/2}} \frac{(\frac{\gamma_0}{2})^{\upsilon_0/2}}{\Gamma(\frac{\upsilon_0}{2})} \frac{\Gamma_{h_L^\star} \Gamma_{h_R^\star}}{\Gamma_{h^\star}} \left(\frac{\upsilon_0 + (\boldsymbol{x}_i^{h^\star})^\dagger \boldsymbol{P}_i^{h^\star} \boldsymbol{x}_i^{h^\star}}{2} \right)^{\frac{1}{2}(\upsilon_0 + \xi_i^{h^\star} - \xi_i^{h^\star - 1})} \\ & \times \left(\frac{\upsilon_0 + (\boldsymbol{x}_i^{h_L^\star})^\dagger \boldsymbol{P}_i^{h_L^\star} \boldsymbol{x}_i^{h_L^\star}}{2} \right)^{-\frac{1}{2}(\upsilon_0 + \xi_i^{h_L^\star} - \xi_i^{h_L^\star - 1})} \\ & \times \left(\frac{\upsilon_0 + (\boldsymbol{x}_i^{h_R^\star})^\dagger \boldsymbol{P}_i^{h_R^\star} \boldsymbol{x}_i^{h_R^\star}}{2} \right)^{-\frac{1}{2}(\upsilon_0 + \xi_i^{h_R^\star} - \xi_i^{h_R^\star - 1})} \end{aligned} \tag{19}$$

For details see Lèbre et al. (2010). Here $\boldsymbol{\xi}_i^+$ refers to the proposed changepoint vector after adding the new changepoint $\xi^\star$ to the current vector $\boldsymbol{\xi}_i$ and for all h in $\{1, \ldots, k_{i+1}\}$, $\Gamma_h = \Gamma(\frac{\upsilon_0 + \xi_i^h - \xi_i^{h-1}}{2})$, and all other quantities are defined in Sect. 2.5.

For a changepoint death move, an existing changepoint in the current configuration is selected uniformly at random. The two segments adjacent to this changepoint are proposed to be merged into one segment, which will conserve the network structure of one of the two segments (selected with equal probability). Let us denote by $\boldsymbol{\xi}_i^-$ the proposed changepoint vector after removing the selected changepoint from the current vector $\boldsymbol{\xi}_i$. The acceptance ratio of the changepoint death move is equal to the inverse of the changepoint birth acceptance ratio $R(\boldsymbol{\xi}_i | \boldsymbol{\xi}_i^-)$ for proposing a change from $\boldsymbol{\xi}_i^-$ to $\boldsymbol{\xi}_i$, given in Eq. (19). Therefore the acceptance probability of a changepoint death move is,

$$A\big(\boldsymbol{\xi}_i^- \big| \boldsymbol{\xi}_i\big) = \min\big\{1, \big(R\big(\boldsymbol{\xi}_i \big| \boldsymbol{\xi}_i^-\big)\big)^{-1}\big\} \tag{20}$$

Proposed shifts in changepoint positions are accepted using a standard Metropolis-Hastings step (Hastings 1970) where a change is accepted with probability $\min\{1, R\}$ where $R =$ (posterior ratio) $\times$ (proposal ratio). The new changepoint vector $\tilde{\boldsymbol{\xi}}_i$ is obtained by replacing ξ_i^h with $\tilde{\xi}_i^h$ such that the absolute value $|\xi_i^h - \tilde{\xi}_i^h| = 1$. The posterior ratio is obtained from Eq. (12). Let us denote by $\mathcal{Q}(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i)$ the probability of shifting changepoint ξ_i^h to $\tilde{\xi}_i^h$ in the current changepoint vector $\boldsymbol{\xi}^i$ (and reciprocally for $\mathcal{Q}(\boldsymbol{\xi}_i|\tilde{\boldsymbol{\xi}}_i)$), then the changepoint shift is accepted with probability $A(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i) = \min\{1, R(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i)\}$ where,

$$R(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i) = \left(\frac{(\gamma_0 + (\tilde{\boldsymbol{x}}_i^h)^\dagger \tilde{\boldsymbol{P}}_i^h \tilde{\boldsymbol{x}}_i^h)^{(\upsilon_0+\tilde{\xi}_i^h-\xi_i^{h-1})}(\gamma_0 + (\tilde{\boldsymbol{x}}_i^{h+1})^\dagger \tilde{\boldsymbol{P}}_i^{h+1} \tilde{\boldsymbol{x}}_i^{h+1})^{(\upsilon_0+\xi_i^{h+1}-\tilde{\xi}_i^h)}}{(\gamma_0 + (\boldsymbol{x}_i^h)^\dagger \boldsymbol{P}_i^h \boldsymbol{x}_i^h)^{(\upsilon_0+\xi_i^h-\xi_i^{h-1})}(\gamma_0 + (\boldsymbol{x}_i^{h+1})^\dagger \boldsymbol{P}_i^{h+1} \boldsymbol{x}_i^{h+1})^{(\upsilon_0+\xi_i^{h+1}-\xi_i^h)}}\right)^{1/2} \times \frac{\Gamma(\frac{\upsilon_0+\tilde{\xi}_i^h-\xi_i^{h-1}}{2})\Gamma(\frac{\upsilon_0+\xi_i^{h+1}-\tilde{\xi}_i^h}{2})}{\Gamma(\frac{\upsilon_0+\xi_i^h-\xi_i^{h-1}}{2})\Gamma(\frac{\upsilon_0+\xi_i^{h+1}-\xi_i^h}{2})} \frac{\mathcal{Q}(\boldsymbol{\xi}_i|\tilde{\boldsymbol{\xi}}_i)}{\mathcal{Q}(\tilde{\boldsymbol{\xi}}_i|\boldsymbol{\xi}_i)}, \tag{21}$$

where $\tilde{\boldsymbol{x}}_i^h$ and $\tilde{\boldsymbol{x}}_i^{h+1}$ refer to the expression levels for gene i observed in phase h and $h+1$ of the new changepoint vector $\tilde{\boldsymbol{\xi}}_i$, and $\tilde{\boldsymbol{P}}_i^h$ and $\tilde{\boldsymbol{P}}_i^{h+1}$ are the projection matrices built from $\tilde{\boldsymbol{x}}_i^h$ and $\tilde{\boldsymbol{x}}_i^{h+1}$ as defined in Eq. (15), and all other quantities are as defined in Sect. 2.5. See Lèbre et al. (2010) for the derivation of this equation.

Finally, network structure updates within segments invoke a second RJMCMC scheme, which was adapted from the model selection approach of Andrieu and Doucet (1999). When such a move is chosen, for each segment successively, we consider either the birth or death of an edge. For an edge birth move, a new edge is selected uniformly at random from the set of possible edges. For an edge death move, an edge to be removed is selected uniformly at random from the set of existing edges. The edge birth and death moves represent changes from s_i^h to $s_i^h + 1$ or $s_i^h - 1$ parents in the regression model. The probabilities of choosing these moves, $b_{s_i^h}$ and $d_{s_i^h}$ respectively, are defined as follows,

$$b_{s_i^h} = C_{s_i^h} \min\left\{1, \frac{P_{\bar{s}}(s_i^h + 1)}{P_{\bar{s}}(s_i^h)}\right\} \quad \text{and} \quad d_{s_i^h} = C_{s_i^h} \min\left\{1, \frac{P_{\bar{s}}(s_i^h - 1)}{P_{\bar{s}}(s_i^h)}\right\} \tag{22}$$

where $C_{s_i^h}$ is a normalization constant dependent on s_i^h, and set to ensure that $b_{s_i^h} + d_{s_i^h} = 1$. Additionally, we define $b_0 = 1$, $d_0 = 0$, $b_{\bar{s}} = 0$ and $d_{\bar{s}} = 1$. The acceptance ratio $R(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)$ for the new set of $\tilde{s}_i^h$ parents $\tilde{\boldsymbol{G}}_i^h$ (which corresponds to $\boldsymbol{G}_i^h$ with a parent added or removed) is computed according to Eq. (18). Using Eqs. (4) and (5), the edge birth prior ratio becomes

$$R_{prior} = \frac{P(\tilde{\boldsymbol{G}}_i^h|\tilde{s}_i^h)}{P(\boldsymbol{G}_i^h|s_i^h)} \frac{P(\tilde{s}_i^h|\Lambda)}{P(s_i^h|\Lambda)} \tag{23}$$

and the proposal ratio becomes

$$R_{proposal} = \frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)} \tag{24}$$

where $\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)$ is the proposal probability of parent set $\tilde{\boldsymbol{G}}_i^h$ given parent set $\boldsymbol{G}_i^h$, which is defined as follows:

$$\mathcal{Q}\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = b_{|\boldsymbol{G}_i^h|}\delta\big(\big|\tilde{\boldsymbol{G}}_i^h\big|, \big|\boldsymbol{G}_i^h\big| + 1\big)\mathcal{Q}^+\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) + d_{|\boldsymbol{G}_i^h|}\delta\big(\big|\tilde{\boldsymbol{G}}_i^h\big|, \big|\boldsymbol{G}_i^h\big| - 1\big)\mathcal{Q}^-\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) \tag{25}$$

with $\delta(x, y)$ being the Kronecker delta function. $\mathcal{Q}^+(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h) = 1/(p - |\tilde{\boldsymbol{G}}_i^h|)$ is the proposal probability of an edge birth move, and $\mathcal{Q}^-(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h) = 1/|\tilde{\boldsymbol{G}}_i^h|$ is the proposal probability of an edge death move. The Jacobian equals 1. Then using Eq. (14) for the likelihood ratio, the Metropolis-Hastings acceptance ratio for an edge move becomes

$$R\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = \frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)} \frac{P(\tilde{s}_i^h|\Lambda)}{P(s_i^h|\Lambda)} \frac{P(\tilde{\boldsymbol{G}}_i^h|\tilde{s}_i^h)}{P(\boldsymbol{G}_i^h|s_i^h)} \frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h, \delta^2)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h, \delta^2)} \tag{26}$$

Note that the prior ratio and the proposal ratio cancel out, and hence the edge move acceptance ratio is equal to the likelihood ratio, that is,

$$R\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = \frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h, \delta^2)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h, \delta^2)} \tag{27}$$

Finally, the probability of accepting an edge move is,

$$A\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big) = \min\big\{1, R\big(\tilde{\boldsymbol{G}}_i^h \big| \boldsymbol{G}_i^h\big)\big\} \tag{28}$$

The sampling scheme for updating the hyperparameters δ^2, λ and Λ is described in Lèbre (2007) and Lèbre et al. (2010). Together the four moves B, D, S and N allow the generation of samples from probability distributions defined on unions of spaces of different dimensions for both the number of changepoints k_i and the number of parents s_i^h within each segment h for gene i.

3 Model improvement: information coupling between segments

Allowing the network structure to change between segments leads to a highly flexible model. However, this approach faces a conceptual and a practical problem. The *practical* problem is potential model over-flexibility. If subsequent changepoints are close together, network structures have to be inferred from short time series segments. This will almost inevitably lead to overfitting (in a maximum likelihood context) or inflated inference uncertainty (in a Bayesian context). The *conceptual* problem is the underlying assumption that structures associated with different segments are a priori independent. While this may be true in some circumstances (e.g. if a drug treatment leads to a drastic, rather than gradual, change), in most cases this assumption is not realistic. For instance, for the evolution of a gene regulatory network during embryogenesis, we would assume that the network evolves gradually and that networks associated with adjacent time intervals are a priori similar.

To address these problems, we propose four methods of information sharing among time series segments, as illustrated in Figs. 1 and 2. The first method is based on hard information coupling between the nodes, using the exponential distribution proposed in Werhli and Husmeier (2008). The second scheme uses the same exponential distribution, but replaces the hard by a soft information coupling scheme. The third and fourth scheme are also based on hard and soft information coupling, respectively, but use a binomial distribution with a conjugate beta prior.

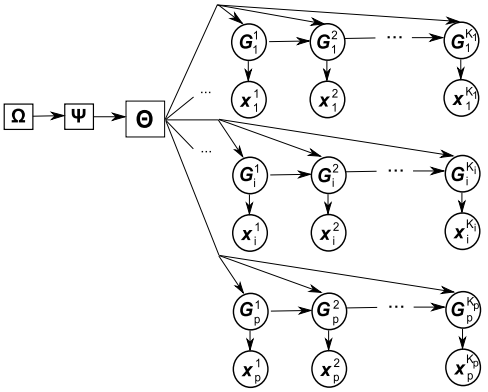

Fig. 1 Hierarchical Bayesian model for inter-segment and hard inter-node information coupling. Hard coupling among nodes i is achieved by a common hyperparameter $\boldsymbol{\Theta}$ regulating the strength of the coupling between structures associated with adjacent segments, $\boldsymbol{G}_i^h$ and $\boldsymbol{G}_i^{h+1}$. This corresponds to the models in Sect. 3.2, with $\boldsymbol{\Theta} = \{\beta\}$, $\boldsymbol{\Psi} = [0, 20]$, and no $\boldsymbol{\Omega}$, and Sect. 3.4, with $\boldsymbol{\Theta} = \{a, b\}$, $\boldsymbol{\Psi} = \{\alpha, \overline{\alpha}, \gamma, \overline{\gamma}\}$, and $\boldsymbol{\Omega} = \{1, 2, \ldots, 100\}$

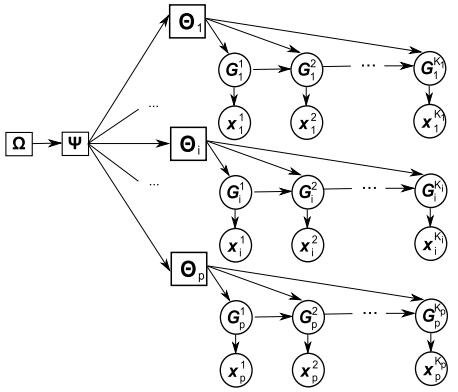

Fig. 2 Hierarchical Bayesian model for inter-segment and soft inter-node information coupling. Soft coupling among nodes i is achieved by node-specific hyperparameters $\boldsymbol{\Theta}_i$ regulating the strength of the coupling between structures associated with adjacent segments, $\boldsymbol{G}_i^h$ and $\boldsymbol{G}_i^{h+1}$, coupled via level-2 hyperparameters $\boldsymbol{\Psi}$. This corresponds to the model in Sect. 3.3, with $\boldsymbol{\Theta}_i = \{\beta_i\}$, $\boldsymbol{\Psi} = \kappa$, and $\boldsymbol{\Omega} = \lambda_\kappa = 10$, and Sect. 3.5, with $\boldsymbol{\Theta}_i = \{a_i, b_i\}$, $\boldsymbol{\Psi} = \{\alpha, \overline{\alpha}, \gamma, \overline{\gamma}\}$, and $\boldsymbol{\Omega} = \{1, 2, \ldots, 100\}$

3.1 Hard versus soft information coupling of nodes

As noted above, we propose to share information about the network structure among the different time series segments that result from the changepoint process. The strength of these couplings is governed by the hyperparameters associated with the information sharing prior. We represent these hyperparameters collectively by $\boldsymbol{\Theta}$. However, another level of coupling is possible, coupling genes (nodes in the network) rather than time series segments.

Recall from Sect. 2 that each node in the network is associated with a random variable $X_i(t)$ that represents the gene expression level of gene i at time t. Under the regression model in Eq. (1), the regulators for gene i are independent of the structure of the rest of

the network. Once we bring in information sharing, however, there is a set of hyperparameters that could conceivably be shared among different nodes; namely $\boldsymbol{\Theta}$. We address this by proposing two different ways of sharing $\boldsymbol{\Theta}$: Hard coupling, where the information sharing prior has the same hyperparameters $\boldsymbol{\Theta}$ for all nodes (with hyperprior having level-2 hyperparameters $\boldsymbol{\Psi}$); and soft coupling, where the information sharing prior has node-specific hyperparameters $\boldsymbol{\Theta}_i$, with common level-2 hyperparameters $\boldsymbol{\Psi}$. In both cases we have a prior on $\boldsymbol{\Psi}$ with level-3 hyperparameters $\boldsymbol{\Omega}$. See Figs. 1 and 2 for an illustration of hard versus soft information coupling of nodes.

In the following sub-sections, we will describe the different information sharing schemes in more detail.

3.2 Hard information coupling based on an exponential prior

Denote by $K_i := k_i + 1$ the total number of partitions in the time series associated with node i, and recall that each time series segment $\boldsymbol{x}_i^h$ is associated with a separate subnetwork $\boldsymbol{G}_i^h$, $1 \leq h \leq K_i$. We modify the prior from Eq. (6) by imposing a prior distribution $P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1}, \beta)$ on the structures, and the joint probability distribution factorizes according to a Markovian dependence:

$$\begin{aligned} &P\big(\boldsymbol{x}_i^1, \ldots, \boldsymbol{x}_i^{K_i}, \boldsymbol{G}_i^1, \ldots, \boldsymbol{G}_i^{K_i}, \beta\big) \\ &\quad = P\big(\boldsymbol{x}_i^1|\boldsymbol{G}_i^1\big)P\big(\boldsymbol{G}_i^1\big)P(\beta)\prod_{h=2}^{K_i} P\big(\boldsymbol{x}_i^h\big|\boldsymbol{G}_i^h\big)P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1}, \beta\big) \end{aligned} \tag{29}$$

Similar to Werhli and Husmeier (2008) we define

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1}, \beta\big) = \frac{\exp(-\beta|\boldsymbol{G}_i^h - \boldsymbol{G}_i^{h-1}|)}{Z(\beta, \boldsymbol{G}_i^{h-1})} \tag{30}$$

for $h \geq 2$, where β is a hyperparameter that defines the strength of the coupling between $\boldsymbol{G}_i^h$ and $\boldsymbol{G}_i^{h-1}$, and $|.|$ denotes the Hamming distance. For $h = 1$, $P(\boldsymbol{G}_i^h)$ is given by (6). The denominator $Z(\beta, \boldsymbol{G}_i^{h-1})$ in (30) is a normalizing constant, also known as the partition function: $Z(\beta, \boldsymbol{G}_i^{h-1}) = \sum_{\boldsymbol{G}_i^h \in \mathbb{G}} e^{-\beta|\boldsymbol{G}_i^h - \boldsymbol{G}_i^{h-1}|}$ where $\mathbb{G}$ is the set of all valid subnetwork structures. If we ignore any fan-in restriction that might have been imposed a priori (via $\overline{s}$ in Eq. (4)), then the expression for the partition function can be simplified: $Z(\beta, \boldsymbol{G}_i^{h-1}) \approx \prod_{j=1}^{p} Z_j(\beta, e_{ij}^{h-1})$, where e_{ij}^h is a binary variable indicating the presence or absence of a directed edge from node j to node i in time series segment h, and $Z_j(\beta, e_{ij}^{h-1}) = \sum_{e_{ij}^h=0}^{1} e^{-\beta|e_{ij}^h - e_{ij}^{h-1}|} = 1 + e^{-\beta}$. Note that this expression no longer depends on $\boldsymbol{G}_i^{h-1}$, and hence

$$Z\big(\beta, \boldsymbol{G}_i^{h-1}\big) = Z(\beta) = \big(1 + e^{-\beta}\big)^p \tag{31}$$

Inserting this expression into (30) gives:

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1}, \beta\big) = \frac{\exp(-\beta|\boldsymbol{G}_i^h - \boldsymbol{G}_i^{h-1}|)}{(1 + e^{-\beta})^p} \tag{32}$$

It is straightforward to integrate the proposed model into the RJMCMC scheme of Lèbre (2007) and Lèbre et al. (2010), which we have summarized in Sect. 2.6. When proposing a

new network structure $\boldsymbol{G}_i^h \to \tilde{\boldsymbol{G}}_i^h$ for segment h, the prior probability ratio in Eq. (23) has to be replaced by $\frac{P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta)}{P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta)}$, leading to the acceptance probability

$$A\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big) = \min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta)\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta)\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{33}$$

This equation is equivalent to Eq. (28), with the prior probabilities in Eq. (23) replaced by those in Eq. (32). Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ is defined in Eqs. (24) and (25). An additional MCMC step is introduced for sampling the hyperparameter β from the posterior distribution. For a proposal move $\beta \to \tilde{\beta}$ with symmetric proposal probability $\mathcal{Q}(\tilde{\beta}|\beta) = \mathcal{Q}(\beta|\tilde{\beta})$ we get the following acceptance probability:

$$A(\tilde{\beta}|\beta) = \min\left\{\frac{P(\tilde{\beta})}{P(\beta)}\prod_{i=1}^{p}\prod_{h=2}^{K_i}\frac{\exp(-\tilde{\beta}|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{\exp(-\beta|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}\frac{(1+e^{-\beta})^p}{(1+e^{-\tilde{\beta}})^p},1\right\} \tag{34}$$

where in our study the hyperprior $P(\beta)$ was chosen as the uniform distribution on the interval $[0, 20]$.

3.3 Soft information coupling based on an exponential prior

We modify the model defined in (29) by making the hyperparameter β, which defines the prior coupling strength between structures associated with adjacent segments, node-dependent: $\beta \to \beta_i$, and

$$P\big(\boldsymbol{x}_i^1,\ldots,\boldsymbol{x}_i^{K_i},\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^K,\beta_i\big) = P\big(\boldsymbol{x}_i^1\big|\boldsymbol{G}_i^1\big)P\big(\boldsymbol{G}_i^1\big)\prod_{h=2}^{K_i}P\big(\boldsymbol{x}_i^h\big|\boldsymbol{G}_i^h\big)P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},\beta_i\big)P(\beta_i) \tag{35}$$

with

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},\beta_i\big) = \frac{\exp(-\beta_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{Z(\beta_i,\boldsymbol{G}_i^{h-1})} = \frac{\exp(-\beta_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{(1+e^{-\beta_i})^p} \tag{36}$$

where by analogy with the previous section, $Z(\beta_i,\boldsymbol{G}_i^{h-1}) \approx (1+e^{-\beta_i})^p$. To introduce soft information coupling between the subnetworks, we choose a hierarchical structure for the prior distribution on the hyperparameters β_i. At the first level, the hyperparameters are given a common gamma prior:

$$P(\beta_i) = P(\beta_i|\kappa,\rho) = \beta_i^{\kappa-1}\frac{\exp(-\beta_i/\rho)}{\rho^\kappa\Gamma(\kappa)} \tag{37}$$

with shape parameter $\kappa > 0$ and scale parameter $\rho > 0$. Recall that the gamma distribution has mean $\mu = \kappa\rho$ and variance $\sigma^2 = \kappa\rho^2$. We elect to set the scale parameter $\rho = 0.1$ fixed. The shape parameter κ is given a vague exponential prior:

$$P(\kappa|\lambda_\kappa) = \lambda_\kappa\exp(-\kappa/\lambda_\kappa) \tag{38}$$

with $\lambda_\kappa = 10$ to reflect our prior ignorance. This choice of prior has the following motivation. The coupling strength between the substructures is defined by the coefficient of variation

$\sigma/\mu = 1/\sqrt{\kappa}$, with smaller coefficients corresponding to stronger coupling strengths, and a zero coefficient ($\kappa \to \infty$) reducing to the hard coupling scheme discussed in the previous section. By inferring the shape parameter κ from the data, starting from a vague yet proper prior distribution, we determine if the coupling strength should be strong or weak.

It is straightforward to adapt the RJMCMC scheme of the previous section. When proposing a new network structure $\boldsymbol{G}_i^h \to \tilde{\boldsymbol{G}}_i^h$ for segment h, the prior probability ratio in Eq. (23) has to be replaced by the ratio $\frac{P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta_i)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)}{P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta_i)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)}$, leading to the equivalent of the acceptance probability in Eq. (28):

$$A\big(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h\big) = \min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\boldsymbol{G}_i^{h+1}|\tilde{\boldsymbol{G}}_i^h,\beta_i)P(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\boldsymbol{G}_i^{h+1}|\boldsymbol{G}_i^h,\beta_i)P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1},\beta_i)\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{39}$$

Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ defined in Eqs. (24) and (25). When proposing new hyperparameters $\tilde{\beta}_i$ from a symmetric proposal distribution $\mathcal{Q}(\tilde{\beta}_i|\beta_i) = \mathcal{Q}(\beta_i|\tilde{\beta}_i)$ we get the following acceptance probability:

$$A(\tilde{\beta}_i|\beta_i) = \min\left\{\frac{P(\tilde{\beta}_i|\rho,\kappa)}{P(\beta_i|\rho,\kappa)}\prod_{h=2}^{K_i}\frac{\exp(-\tilde{\beta}_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}{\exp(-\beta_i|\boldsymbol{G}_i^h-\boldsymbol{G}_i^{h-1}|)}\left(\frac{1+e^{-\beta_i}}{1+e^{-\tilde{\beta}_i}}\right)^p,1\right\} \tag{40}$$

An additional sampling step is needed for the shape parameter κ of the level-2 hyperprior. Drawing a new shape parameter $\tilde{\kappa}$ from a symmetric proposal distribution $\mathcal{Q}(\tilde{\kappa}|\kappa)$, the acceptance probability is given by

$$A(\tilde{\kappa}|\kappa) = \min\left\{\frac{\exp(-\tilde{\kappa}/\lambda_\kappa)}{\exp(-\kappa/\lambda_\kappa)}\prod_{i=1}^{p}\frac{P(\beta_i|\tilde{\kappa},\rho)}{P(\beta_i|\kappa,\rho)},1\right\} \tag{41}$$

3.4 Hard information coupling based on a binomial prior

An alternative way of information sharing among segments and nodes is by using a binomial prior:

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},a,b\big) = a^{N_1^1[h,i]}(1-a)^{N_1^0[h,i]}b^{N_0^0[h,i]}(1-b)^{N_0^1[h,i]} \tag{42}$$

where we have defined the following sufficient statistics: $N_1^1[h,i]$ is the number of edges in $\boldsymbol{G}_i^{h-1}$ that are matched by an edge in $\boldsymbol{G}_i^h$, $N_1^0[h,i]$ is the number of edges in $\boldsymbol{G}_i^{h-1}$ for which there is no edge in $\boldsymbol{G}_i^h$, $N_0^1[h,i]$ is the number of edges in $\boldsymbol{G}_i^h$ for which there is no edge in $\boldsymbol{G}_i^{h-1}$, and $N_0^0[h,i]$ is the number of coinciding non-edges in $\boldsymbol{G}_i^{h-1}$ and $\boldsymbol{G}_i^h$. Since the hyperparameters are shared, the joint distribution can be expressed as:

$$P\big(\{\boldsymbol{G}_i^h\}\big|a,b\big) = \prod_{i=1}^{p}P\big(\boldsymbol{G}_i^1\big)\prod_{h=2}^{K_i}P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},a,b\big) = a^{N_1^1}(1-a)^{N_1^0}b^{N_0^0}(1-b)^{N_0^1}\prod_{i=1}^{p}P\big(\boldsymbol{G}_i^1\big) \tag{43}$$

where we have defined $N_k^l = \sum_{i=1}^{p}\sum_{h=2}^{K_i}N_k^l[h,i]$, and the right-hand side follows from Eq. (42). The conjugate prior for the hyperparameters a,b is a beta distribution,

$$P(a,b|\alpha,\overline{\alpha},\gamma,\overline{\gamma}) \propto a^{(\alpha-1)}(1-a)^{(\overline{\alpha}-1)}b^{(\gamma-1)}(1-b)^{(\overline{\gamma}-1)} \tag{44}$$

which using Bayes' rule leads to the (beta) posterior distribution:

$$P\big(a,b\big|\alpha,\overline{\alpha},\gamma,\overline{\gamma},\big\{\boldsymbol{G}_i^h\big\}\big)\propto a^{(\alpha+N_1^1-1)}(1-a)^{(\overline{\alpha}+N_1^0-1)}b^{(\gamma+N_0^0-1)}(1-b)^{(\overline{\gamma}+N_0^1-1)} \tag{45}$$

This allows the hyperparameters to be integrated out in closed form:

$$\begin{aligned}&P\big(\big\{\boldsymbol{G}_i^h\big\}\big|\alpha,\overline{\alpha},\gamma,\overline{\gamma}\big)\\&\quad=\int\int P\big(\big\{\boldsymbol{G}_i^h\big\}\big|a,b\big)P(a,b|\alpha,\overline{\alpha},\gamma,\overline{\gamma})da\,db\\&\quad\propto\frac{\Gamma(\alpha+\overline{\alpha})}{\Gamma(\alpha)\Gamma(\overline{\alpha})}\frac{\Gamma(N_1^1+\alpha)\Gamma(N_1^0+\overline{\alpha})}{\Gamma(N_1^1+\alpha+N_1^0+\overline{\alpha})}\frac{\Gamma(\gamma+\overline{\gamma})}{\Gamma(\gamma)\Gamma(\overline{\gamma})}\frac{\Gamma(N_0^0+\gamma)\Gamma(N_0^1+\overline{\gamma})}{\Gamma(N_0^0+\gamma+N_0^1+\overline{\gamma})}\end{aligned} \tag{46}$$

The level-2 hyperparameters $\alpha,\overline{\alpha},\gamma,\overline{\gamma}$, which can be interpreted as fictitious prior observations due to the conjugacy of the prior, are given a discrete uniform hyperprior over $\{1,2,\ldots,100\}$. The MCMC scheme of Sect. 2.6 has to be modified as follows. When proposing a new network structure for node i and segment h, $\boldsymbol{G}_i^h\rightarrow\tilde{\boldsymbol{G}}_i^h$, the structures $\boldsymbol{G}_i^h$ and $\tilde{\boldsymbol{G}}_i^h$ enter the prior probability ratio in Eq. (23) via the expression $P(\{\boldsymbol{G}_i^h\}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})$. The prior probability ratio becomes $\frac{P(\{\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}{P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}$, leading to the acceptance probability

$$A\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big)=\min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\{\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{47}$$

This equation is equivalent to Eq. (28), with the prior probabilities in Eq. (23) replaced by those in Eq. (46). Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ defined in Eqs. (24) and (25). From Fig. 1, it becomes clear that as a consequence of integrating out the hyperparameters, all network structures become interdependent, and information about the structures is contained in the sufficient statistics N_1^1,N_1^0,N_0^1,N_0^0. A new proposal move for the level-2 hyperparameters is added to the existing RJMCMC scheme of Sect. 2.6. New values for the level-2 hyperparameters α are proposed from a uniform distribution over the support of $P(\alpha)$. For a move $\alpha\rightarrow\tilde{\alpha}$, the acceptance probability is:

$$A(\tilde{\alpha}|\alpha)=\min\left\{\frac{P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\tilde{\alpha},\overline{\alpha},\gamma,\overline{\gamma})}{P(\{\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}\}_{i=1}^p|\alpha,\overline{\alpha},\gamma,\overline{\gamma})},1\right\} \tag{48}$$

and similarly for $\overline{\alpha}$, γ and $\overline{\gamma}$.

3.5 Soft information coupling based on a binomial prior

We can relax the information sharing scheme from a hard to a soft coupling by introducing node-specific hyperparameters a_i,b_i that are softly coupled via a common level-2 hyperprior, $P(a_i,b_i|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\propto a_i^{(\alpha-1)}(1-a_i)^{(\overline{\alpha}-1)}b_i^{(\gamma-1)}(1-b_i)^{(\overline{\gamma}-1)}$ as illustrated in Fig. 2:

$$P\big(\boldsymbol{G}_i^h\big|\boldsymbol{G}_i^{h-1},a_i,b_i\big)=(a_i)^{N_1^1[h,i]}(1-a_i)^{N_1^0[h,i]}(b_i)^{N_0^0[h,i]}(1-b_i)^{N_0^1[h,i]} \tag{49}$$

This leads to a straightforward modification of Eq. (43)—replacing a, b by a_i, b_i—from which we get as an equivalent to (46), using the definition $N_k^l[i] = \sum_{h=2}^{K_i} N_k^l[h,i]$:

$$P\big(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}\big|\alpha,\overline{\alpha},\gamma,\overline{\gamma}\big) \propto \frac{\Gamma(\alpha+\overline{\alpha})}{\Gamma(\alpha)\Gamma(\overline{\alpha})}\frac{\Gamma(N_1^1[i]+\alpha)\Gamma(N_1^0[i]+\overline{\alpha})}{\Gamma(N_1^1[i]+\alpha+N_1^0[i]+\overline{\alpha})} \times \frac{\Gamma(\gamma+\overline{\gamma})}{\Gamma(\gamma)\Gamma(\overline{\gamma})}\frac{\Gamma(N_0^0[i]+\gamma)\Gamma(N_0^1[i]+\overline{\gamma})}{\Gamma(N_0^0[i]+\gamma+N_0^1[i]+\overline{\gamma})} \tag{50}$$

As in Sect. 3.4, we extend the RJMCMC scheme from Sect. 2.6 so that when proposing a new network structure, $\boldsymbol{G}_i^h \to \tilde{\boldsymbol{G}}_i^h$, the prior probability ratio in Eq. (23) has to be replaced by: $\frac{P(\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}{P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})}$, leading to the equivalent of the acceptance probability in Eq. (28):

$$A\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big) = \min\left\{\frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h)P(\boldsymbol{G}_i^1,\ldots,\tilde{\boldsymbol{G}}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^h,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)},1\right\} \tag{51}$$

Note that $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h)$ is short for $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ which is defined in Eq. (14) and the proposal ratio $\frac{\mathcal{Q}(\boldsymbol{G}_i^h|\tilde{\boldsymbol{G}}_i^h)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h)}$ defined in Eqs. (24) and (25). In addition, we have to add a new level-2 hyperparameter update move: when proposing a level-2 hyperparameter $\alpha \to \tilde{\alpha}$, where the prior and proposal probabilities are the same as in Sect. 3.4, the acceptance probability becomes:

$$A(\tilde{\alpha}|\alpha) = \min\left\{\prod_{i=1}^{p}\frac{P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}|\tilde{\alpha},\overline{\alpha},\gamma,\overline{\gamma})}{P(\boldsymbol{G}_i^1,\ldots,\boldsymbol{G}_i^{K_i}|\alpha,\overline{\alpha},\gamma,\overline{\gamma})},1\right\} \tag{52}$$

and similarly for $\overline{\alpha}$, γ and $\overline{\gamma}$.

3.6 Improved MCMC scheme

The various information sharing priors that we have introduced in the previous Sects 3.2, 3.3, 3.4, 3.5 share the characteristic that they encourage the networks of all segments to be similar to each other.[2] When applying the MCMC scheme from Lèbre et al. (2010), summarized in Sect. 2.6, adapted to our prior as discussed above, this can lead to the following curious effect. On simulated data where the network structure is the same for all segments we found that the network reconstruction accuracy deteriorated when we increased the coupling strength between the structures. The results will be presented below, in Sect. 5 and Fig. 4. These findings appear counter-intuitive, given that increasing the coupling strength brings the prior more in line with the truth (the perfect prior would have infinitely strong coupling). However, it is easily seen that increasing the coupling strength adversely affects the mixing of the Markov chains. Consider a set of identical network structures which, at an initial stage of the MCMC simulations, are all poor at explaining the data. We now visit a

[2]Note that the binomial information sharing prior (Sects. 3.4 and 3.5) can in principle encourage either similarity or dissimilarity depending on the hyperparameters a and b. As discussed in Sect. 5, we had originally envisaged setting the level-2 hyperparameters $\overline{\alpha}$ and $\overline{\gamma}$ equal to 1 to enforce similarity, but Fig. 8 demonstrates that this constraint is too restrictive.

segment and propose a modification of the network structure associated with it. This modification introduces a mismatch between the structures and is, hence, discouraged by the prior. For strong coupling this discouragement might outweigh the gain in the likelihood that would result from a better structure. The structures thus remain identical, which in turn will tend to increase the coupling strength. The MCMC simulation thus gets trapped in a suboptimal state of the configuration space (local optimum).

To deal with this problem, we have implemented an alternative MCMC scheme where changes are applied to multiple segments. The new moves will propose changes to the network structure in more than one segment, and we will hence refer to them as multi-segment moves. Note that the moves for proposing new changepoint configurations are unaffected by these modifications. The multi-segment moves are presented as target-node specific (i.e. they presuppose a choice of target node i). However, they can be generalized for inference over the whole network by simply picking a target node at random. Given a node, the proposal move consists of two steps: (1) Pick one of p possible parents for the target node i. (2) For each segment h of the K_i segments, flip the edge status (changing an edge to a non-edge or vice-versa) between the parent node and the target node with probability q. In our simulations, we set $q = \frac{1}{2}$ so that flipping the edge status and conserving it are equally likely outcomes. It is straightforward to adapt this parameter during the burn-in phase. This means that the probability of proposing a new set of structures $\tilde{\boldsymbol{G}}_i$ given the set of network structures $\boldsymbol{G}_i$ using the multi-segment move is:

$$\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = \frac{1}{p2^{K_i}} \tag{53}$$

where $\boldsymbol{G}_i = \{\boldsymbol{G}_i^h\}_{1\leq h\leq K_i}$ as before.

We now derive the acceptance ratio for multi-segment moves. We define $R_{prior}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ to be the ratio of the prior probabilities of the original set $\boldsymbol{G}_i$ and the proposed set $\tilde{\boldsymbol{G}}_i$. Let $R_{likelihood}(\tilde{\boldsymbol{G}}_i^h|\boldsymbol{G}_i^h) = \frac{P(\boldsymbol{x}_i^h|\tilde{\boldsymbol{G}}_i^h,\delta^2)}{P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)}$ be the likelihood ratio of the original and proposed network structures for segment h and target node i, where the likelihood $P(\boldsymbol{x}_i^h|\boldsymbol{G}_i^h,\delta^2)$ is defined in Eq. (14) of Sect. 2.6. Note that the changes introduced by multi-segment moves are equivalent to a sequence of add and remove edge moves applied to individual segments, so that this ratio remains unchanged. Then the acceptance ratio for multi-segment moves can be expressed as:

$$R(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = R_{prior}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)\prod_{h=1}^{K_i} R_{likelihood}\big(\tilde{\boldsymbol{G}}_i^h\big|\boldsymbol{G}_i^h\big) \tag{54}$$

where $R_{prior}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = \frac{P(\tilde{\boldsymbol{G}}_i)}{P(\boldsymbol{G}_i)}$. The form of $P(\boldsymbol{G}_i)$ depends on our choice of prior. If segments are independent, then $P(\boldsymbol{G}_i) = \prod_{h=1}^{K_i} P(\boldsymbol{G}_i^h)$, where $P(\boldsymbol{G}_i^h)$ is the prior from Eq. (6), with a Poisson distribution on the number of parents. If we want to use information sharing between segments, then the prior for segment h depends on segment $h-1$, so that $P(\boldsymbol{G}_i) = P(\boldsymbol{G}_i^1)\prod_{h=2}^{K_i} P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1})$, where $P(\boldsymbol{G}_i^h|\boldsymbol{G}_i^{h-1})$ could be any of the information sharing priors introduced in Sect. 3. Finally, $R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ is the Hastings ratio:

$$R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = \frac{\mathcal{Q}(\boldsymbol{G}_i|\tilde{\boldsymbol{G}}_i)}{\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)} \tag{55}$$

where $\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ is defined in Eq. (53). Since the proposal probability $\mathcal{Q}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i)$ is independent of the set of network structures $\boldsymbol{G}_i$, the multi-segment moves are symmetric, and we obtain that $R_{proposal}(\tilde{\boldsymbol{G}}_i|\boldsymbol{G}_i) = 1$.

We have explored an alternative proposal scheme consisting of two moves: (1) a move proposing network structures where an edge has been set identical in all segments, and (2) the move described above, which corresponds to a random perturbation of an edge. However, we found that including the first kind of proposal move adversely affected mixing and convergence in simulations where the true network structure presented differences among segments. These network structures are less likely to be proposed when both moves are included. Details can be found in Dondelinger (2012).

4 Implementation and simulations

We have implemented our model in R, based on code from Lèbre (2007) and Lèbre et al. (2010). The network structure, the changepoints and the hyperparameters are sampled from the posterior distribution using RJMCMC as described in Sects. 2.6 and 3.6. We ran the MCMC chains until we were satisfied that convergence was reached. Then we sampled 1000 network and changepoint configurations in intervals of 200 RJMCMC steps. By marginalization and under the assumption of convergence, this represents a sample from the posterior distribution in Eq. (12). By further marginalization, we get the posterior probabilities of all gene regulatory interactions, which defines a ranking of the interactions in terms of posterior confidence. We use the potential scale reduction factor (PSRF) (Gelman and Rubin 1992), computed from the within-chain and between-chain variances of marginal edge posterior probabilities, as a convergence diagnostic. The usual threshold for sufficient convergence lies at PSRF ≤ 1.1. In our simulations, we extended the burn-in phase until a value of PSRF ≤ 1.05 was reached.

For the study on simulated data, and the synthetic biology data, the true interaction network is known. Therefore, varying the threshold on this ranking allows us to construct the Receiver Operating Characteristic (ROC) curve (plotting the sensitivity or recall[3] against the complementary specificity[4]) and the precision-recall (PR) curve (plotting the precision[5] against the recall), and to assess the network reconstruction accuracy in terms of the areas under these graphs (AUROC and AUPRC, respectively); see Davis and Goadrich (2006). These two measures are widely used in the systems biology literature to quantify the overall network reconstruction accuracy (Prill et al. 2010), with larger values indicating a better prediction performance overall.

5 Evaluation on simulated data

5.1 Comparative evaluation of network reconstruction and hyperparameter inference

The purpose of the simulation study is two-fold. Firstly, we want to carry out a comparative evaluation of the proposed Bayesian regularization schemes for a controlled scenario in

[3]The *sensitivity* or *recall* denotes the fraction of true interaction that have been recovered.

[4]The *specificity* denotes the fraction of spurious interactions that have been successfully avoided.

[5]The *precision* is the fraction of predicted interactions that are correct.

which the true network structure is known. Secondly, we want to assess the Bayesian inference scheme and test the viability of the proposed MCMC samplers. To focus on the task of network reconstruction, we keep the changepoints fixed at their true values. The inference of the changepoints will be investigated later, on the real gene expression time series (see Fig. 12).

The simulation set-up we chose was as follows. We randomly generated 10 networks with 10 nodes each. A Poisson distribution with mean $\lambda_{parents} = 3$ was used to determine the number of parents for each node. We simulated changes in the network structure by producing 4 different network segments, where a Poisson distribution with mean $\lambda_{changes} \in \{0.25, 0.5, 1\}$ was used to determine the number of changes per node. The changes were then applied uniformly at random to edges and non-edges in the previous segment. For each segment h, we generated a time series of length 15 using a linear regression model:

$$\boldsymbol{x}(t) = \boldsymbol{W}^h \boldsymbol{x}(t-1) + \boldsymbol{\epsilon} \tag{56}$$

where $\boldsymbol{x}(t)$ is the 10×1 vector of observations at time t and $\boldsymbol{W}^h = \{w_{ij}^h\}$ is the 10×10 matrix of segment-specific regression weights for each edge. We chose the regression weights such that $w_{ij}^h = 0$ if there is no edge between node i and node j in the network structure for segment h, and $w_{ij}^h \sim N(0, 1)$ otherwise. We added Gaussian observation noise $\epsilon_i \sim N(0, 1)$ independently for each observation of node i.

First, we consider the scenario of homogeneous time series in which the regulatory network structure does not change (although the regression coefficients associated with each edge may change between segments). This is the situation in which the proposed Bayesian regularization scheme should achieve the strongest boost in the network reconstruction accuracy. We indeed found this conjecture confirmed in our simulations, as demonstrated in Fig. 3 (0 % changes). We would also assume that high values of the hyperparameter β should lead to the best network reconstruction accuracy, as this corresponds to the tightest tying between adjacent structures. However, repeating the MCMC simulations initially did not confirm this conjecture; see Figs. 4(c) and 4(d). As discussed in Sect. 3.6, the observed mismatch was a consequence of poor mixing and convergence for large hyperparameter values, which is endemic to the naive extension of the MCMC sampler from Lèbre et al. (2010). Repeating the simulations with the novel MCMC scheme proposed in Sect. 3.6 leads to the graphs of Figs. 4(a) and 4(b). Here, the network reconstruction accuracy no longer deteriorates with increasing hyperparameters, indicating that the mixing and convergence problems have been averted.

Another question we investigated is whether the sampled values of the hyperparameters concur with those that optimize the network reconstruction accuracy. While the hyperparameter β of the exponential prior does indeed tend to higher values, the situation is different for the hyperparameters a and b of the binomial prior. The top panels in Fig. 5 show the network reconstruction accuracy in terms of AUROC and AUPRC scores for several fixed values of the hyperparameters a and b. As expected, the peak performance is reached for the highest values, as no mismatch between the structures implies that tight coupling is consistent with the data. The centre panels of Fig. 5 show the posterior distribution of the hyperparameters that was obtained with the conventional MCMC proposal scheme adapted from Lèbre et al. (2010) and described in Sect. 2.6. There is an obvious mismatch between the high-posterior probability region and the region of hyperparameters that optimize the network reconstruction. This provides more evidence that the sampler adapted for segment coupling from Lèbre et al. (2010) suffers from mixing and convergence problems. The bottom panels of Fig. 5 show the marginal posterior distributions of the hyperparameters inferred in the

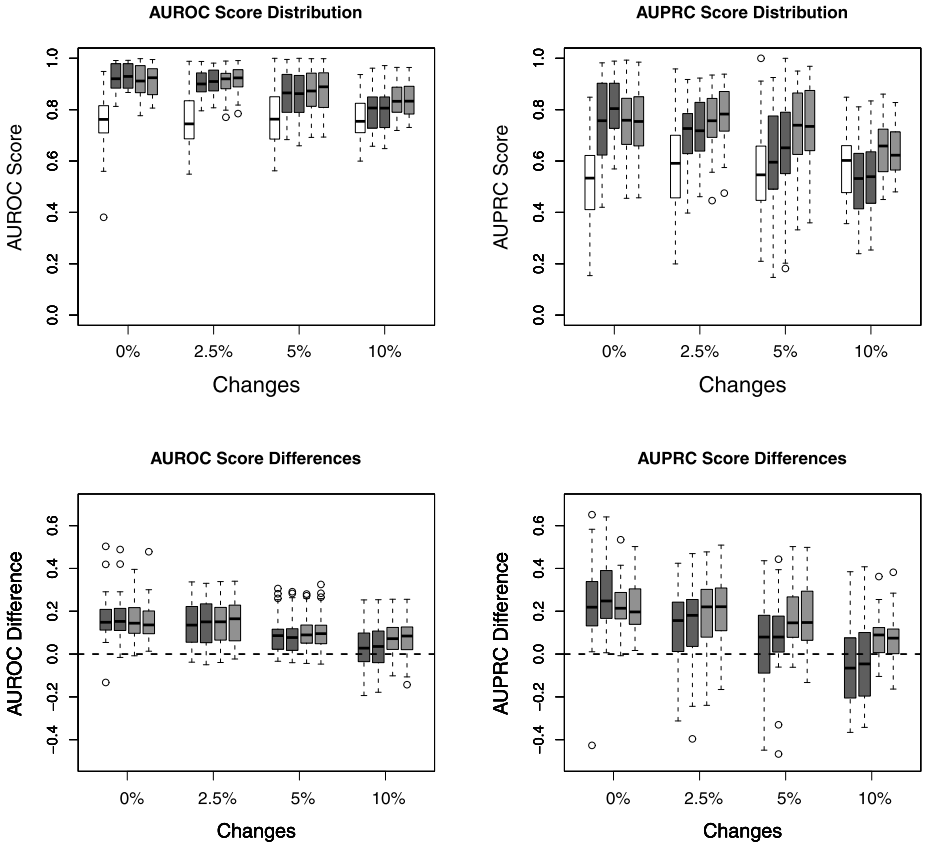

Fig. 3 Evaluation of AUROC and AUPRC network reconstruction scores for the five methods, TVDBN-0 (*white*), TVDBN-Exp-hard (*dark grey, left*), TVDBN-Exp-soft (*dark grey, right*), TVDBN-Bino-hard (*light grey, left*), TVDBN-Bino-soft (*light grey, right*). *Top row*: The *boxplots* show the distributions of the reconstruction scores. *Bottom row*: The *boxplots* show the difference of the AUROC and AUPRC reconstruction scores to TVDBN-0; larger differences indicate better performance with information sharing. All simulations were repeated for 10 independent data sets with 4 network segments each. Structure changes were applied to the segments sequentially, changing between 0–10 % of the edges with each new segment. A paired t-test shows that for 0 % changes, the difference to TVDBN-0 was significant for all methods ($p < 0.05$). For >0 % changes, the difference to TVDBN-0 was significant ($p < 0.05$) except for the difference in AUPRC scores for TVDBN-Exp-hard for 5 % changes ($p = 0.08$) and TVDBN-Exp-hard and TVDBN-Exp-soft for 10 % changes ($p = \{0.07, 0.18\}$). In all *plots*, the *horizontal bar of the boxplot* shows the median, the *box margins* show the 25th and 75th percentiles, the *whiskers* indicate data within 2 times the interquartile range, and *circles* are outliers. See Table 1 for hyperparameter settings

MCMC simulations with the novel multi-segment proposal move introduced in Sect. 3.6. It is seen that, unlike the centre panels in Fig. 5, and as a consequence of the different proposal scheme, the high posterior probability region now concurs with the region of maximum network reconstruction accuracy. This agreement suggests that the novel MCMC sampler leads to a significant improvement in mixing and convergence, in corroboration of our conjecture in Sect. 3.6.

Table 1 List of different information sharing (IS) priors for the TVDBN (Time-Varying Dynamic Bayesian Network), the equation where they were defined, and the most common hyperparameter settings that were used, or hyperparameter ranges if they are inferred. Only the highest level hyperparameters in the Bayesian hierarchy are shown

Name	Prior	Section	Equation	Hyperparameters
TVDBN-0	Poisson (No IS)	2.4	4, 6	$\Lambda = 3$
TVDBN-Exp-hard	Exponential Hard IS	3.2	30	$\beta \in [0, 20]$
TVDBN-Exp-soft	Exponential Soft IS	3.3	38	$\lambda_\kappa = 10$
TVDBN-Bino-hard	Binomial Hard IS	3.4	45, 46	$\alpha, \overline{\alpha}, \gamma, \overline{\gamma} \in \{1, 2, \ldots, 100\}$
TVDBN-Bino-soft	Binomial Soft IS	3.5	50	$\alpha, \overline{\alpha}, \gamma, \overline{\gamma} \in \{1, 2, \ldots, 100\}$

Next, we turn our attention to varying network structures. We varied the percentage of edges that change from segment to segment between 2.5 % to 10 %.[6] A significant improvement in the network reconstruction accuracy can be achieved over the unregularized method, as shown in the bottom panels of Fig. 3. However, the magnitude of the improvement in the scores decreases as the number of changes between adjacent segments increases. This is plausible: as we introduce more structural changes between adjacent networks, we would expect to gain less benefit from information sharing. We note that the degradation in performance seems to be stronger for the exponential prior than for the binomial prior.

We investigated whether the inferred hyperparameters coincide with the optimal reconstruction performance for the case where 10 % of the edges in the network change between adjacent segments. There are two effects to be traded off. Hyperparameter values that are too low will not bring about any improvement over the uncoupled unregularized scenario. Hyperparameter values that are too high will not allow the network structure to change with time. We would therefore expect to find some optimal finite range of hyperparameter values, $0 < \beta < \infty$ and $0 < a, b < 1$. This has in fact been borne out in our simulations. Figure 6 shows the network reconstruction accuracy in terms of AUROC and AUPRC scores for different values of the hyperparameters a, b. The best network reconstruction accuracy is obtained when b, which governs consistency among non-interactions, is high ($\geq$0.9), while a, which controls agreement among interactions, is reduced to a range around its uninformative setting $a \approx 0.5$. The bottom panel of Fig. 6 shows that the inferred posterior distribution is consistent with these ranges, and that the Bayesian inference scheme thus optimizes the network reconstruction accuracy. A slightly different picture emerges for the exponential prior, though. Figures 7(a)–7(b) show the AUROC and AUPRC scores for different values of β, indicating a clear peak in the network reconstruction accuracy for finite $0 < \beta < \infty$. This peak does not coincide with the high posterior probability range of β, as shown in Fig. 7(c). Only when increasing the data set size by a factor of 4 does the Bayesian inference scheme succeed in optimizing the network reconstruction accuracy in the sense that the high posterior probability region now coincides with the range of the highest AUROC/AUPRC scores. The obvious question to ask is whether this trend is another artifact of poor MCMC convergence/mixing. To this end we have devised a simplified model for which the posterior distribution can be computed in closed form. Our analysis, which we present in Sect. 5.2, re-

[6] Because our simulation was set up so that we had on average 3 regulatory interactions per node, this corresponds to a change of between 8.25 % and 33 % of the original interactions.

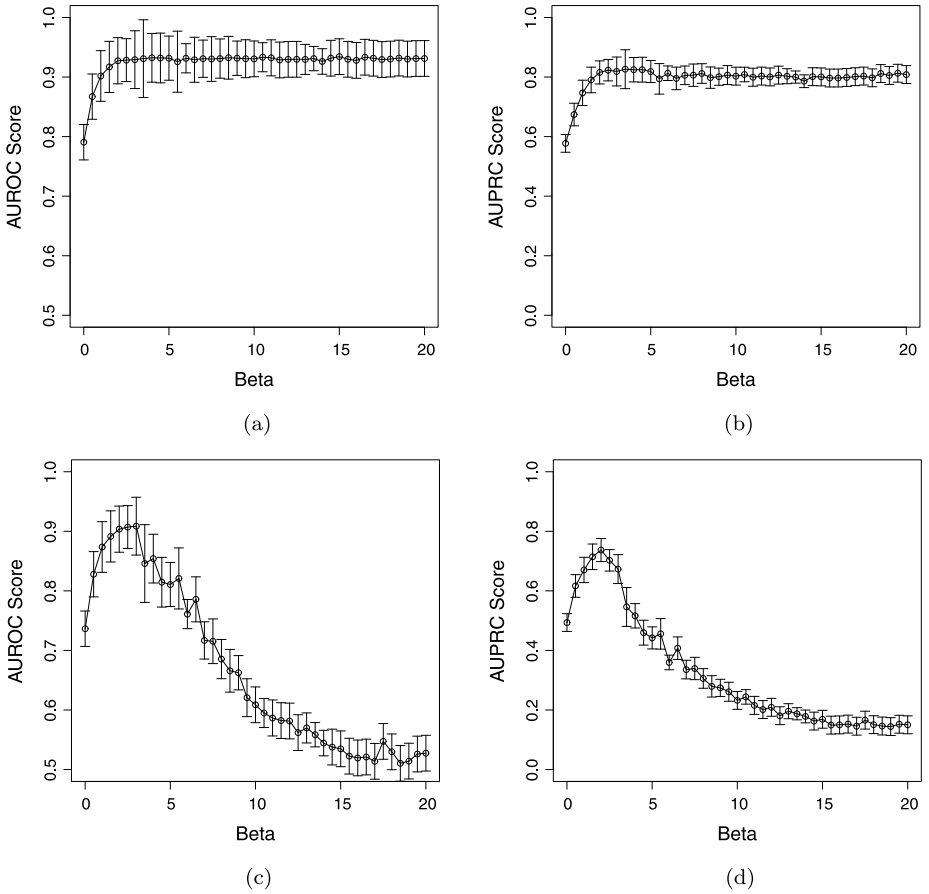

Fig. 4 Results for the exponential prior with hard coupling on the simulated data without mismatch among the structures. Panel (**a**) shows the AUROC scores for different values of the hyperparameter β. Panel (**b**) shows a corresponding plot for the AUPRC scores. The simulations were repeated on 10 independent data instantiations of time series length 60. The *error bars* show the standard error. The results were obtained with the novel MCMC sampler, described in Sect. 3.6. Panels (**c**) and (**d**) show the results from corresponding simulations with the old MCMC sampler adapted from Lèbre et al. (2010) and described in Sect. 2.6. The reconstruction performance deteriorates with larger values of the hyperparameter, as a consequence of poor MCMC mixing and convergence

produces the results from this simulation study, suggesting that the suboptimal performance of the Bayesian inference scheme is intrinsic to the chosen form of the prior.[7]

Returning to the binomial prior, we finally investigated the influence of the level-2 hyperparameters $\alpha, \overline{\alpha}, \gamma$, and $\overline{\gamma}$. Recall that owing to the conjugacy of the prior, these values can be interpreted as fictitious prior observation counts. Our initial idea was to keep the mismatch hyperparameters fixed at $\overline{\alpha} = \overline{\gamma} = 1$, while putting a vague uniform distribution over

[7]We note that the results for the exponential prior seem to be at odds with those reported in Husmeier et al. (2010). The reason is that in Husmeier et al. (2010) we had selected, by a fluke, a more restrictive prior on the hyperparameter: $\beta \in [0, 5]$. As our discussion in Sect. 5.2 shows, this setting boosts the network reconstruction performance.

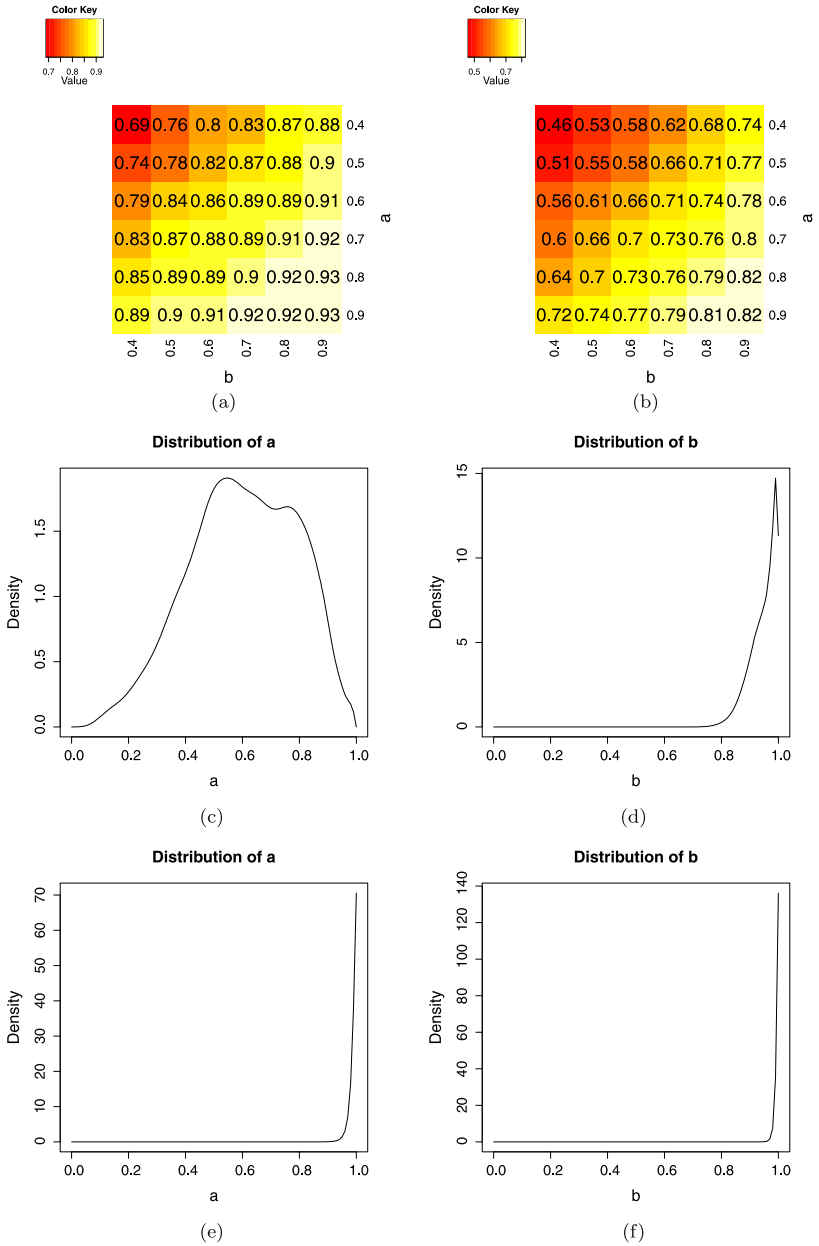

Fig. 5 Results for the binomial prior with hard coupling on the simulated data without mismatch among the structures. Panel (**a**) shows the AUROC scores for different values of the hyperparameters a and b. Panel (**b**) shows a corresponding plot for the AUPRC scores. Panels (**c**) and (**d**) show the marginal posterior distribution of the hyperparameters a and b, as obtained with the MCMC sampler adapted from Lèbre et al. (2010) and described in Sect. 2.6. Panels (**e**) and (**f**) show the marginal posterior distribution of the hyperparameters a and b, as obtained with the new MCMC sampler proposed in Sect. 3.6. The marginal distributions of a and b are obtained from the sampled values of the level-2 hyperparameters α, $\overline{\alpha}$, γ, $\overline{\gamma}$ and from the sampled networks using a kernel density estimator with the beta distribution from Eq. (45). The level-2 hyperparameters were given a uniform prior over the discrete set $\{1, 2, \ldots, 100\}$

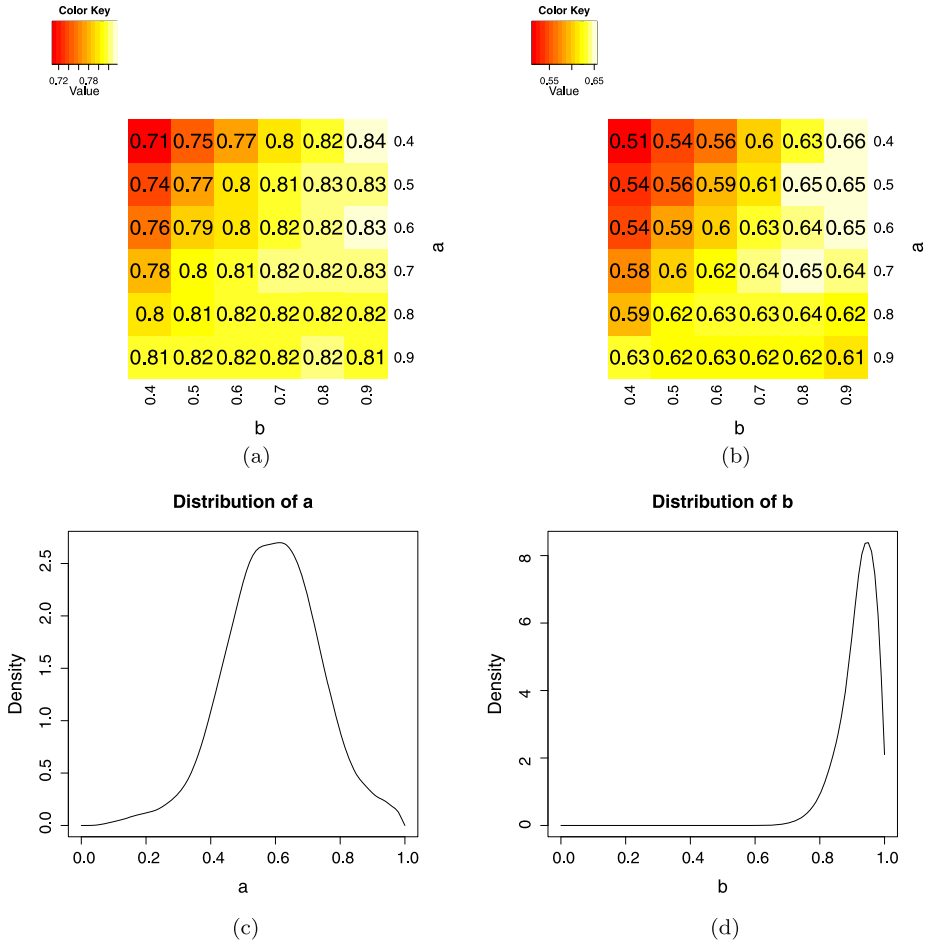

Fig. 6 Results for the binomial prior with hard coupling on the simulated data with mismatch among the structures. Panel (**a**) shows the AUROC scores for different values of the hyperparameters a and b. Panel (**b**) shows a corresponding plot for the AUPRC scores. Panels (**c**) and (**d**) show the marginal posterior distribution of the hyperparameters a and b, as obtained with the novel MCMC sampler proposed in Sect. 3.6. The marginal distributions of a and b were obtained from the sampled values of the level-2 hyperparameters α, $\overline{\alpha}$, γ, $\overline{\gamma}$ and from the sampled networks using a kernel density estimator with the beta distribution from Eq. (45). The level-2 hyperparameters were given a uniform prior over the discrete set $\{1, 2, \ldots, 100\}$

the set $\{1, 2, \ldots, 100\}$ as a prior on the match hyperparameters α and γ. The rationale behind this choice is that the regularization scheme is intended to encourage similarity rather than dissimilarity between adjacent network structures. However, repeating the MCMC simulations for different values of the level-2 hyperparameters revealed that the setting $\overline{\alpha} = \overline{\gamma} = 1$ is too restrictive and that the network reconstruction accuracy can be improved by relaxing this constraint (see Fig. 8).

The findings of our simulation study can be summarized as follows. A naive extension of the MCMC sampler of Lèbre et al. (2010), as described in Sect. 3.6, leads to a poor network reconstruction accuracy for high values of the hyperparameters; this problem can be resolved with the novel proposal scheme introduced in Sect. 3.6. With this new proposal

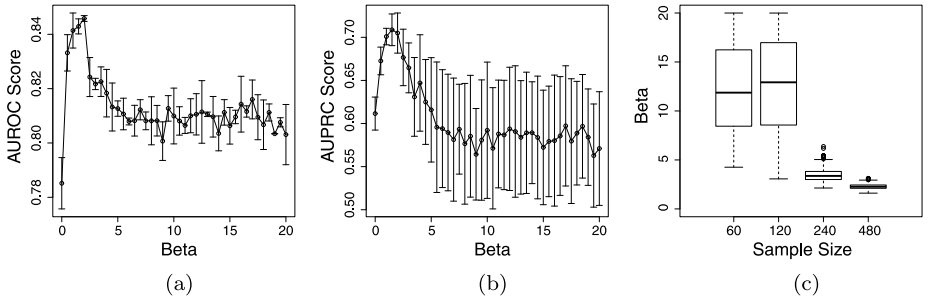

Fig. 7 Results for the exponential prior with hard coupling on the simulated data with mismatch among the structures. Panel (**a**) shows the AUROC scores and their standard deviations for different values of the hyperparameter β. Panel (**b**) shows a corresponding plot for the AUPRC scores. Panel (**c**) shows box plot representations of the inferred posterior distribution of β, for different sample sizes, using the MCMC scheme from Sect. 3.6. The *horizontal bar* shows the median, the *box margins* show the 25th and 75th percentiles, the *whiskers* indicate data within 2 times the interquartile range, and *circles* are outliers. The simulations were repeated on 10 independent data instantiations of time series length $n = 60$

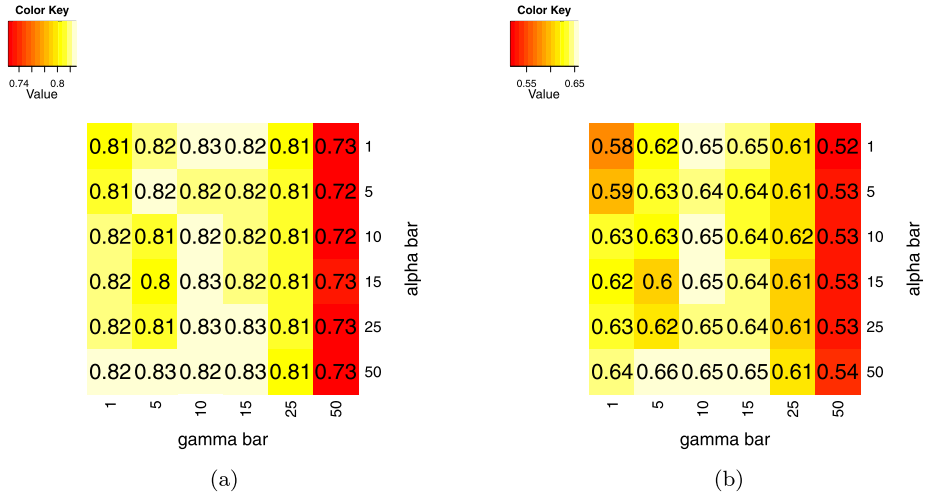

Fig. 8 Results for the binomial prior with hard coupling on the simulated data with mismatch among the structures: dependence of the reconstruction accuracy on the higher-level hyperparameters. Panel (**a**) shows the AUROC scores for different values of the level-2 hyperparameters $\overline{\alpha}$ and $\overline{\gamma}$. Panel (**b**) shows a corresponding plot for the AUPRC scores. The results indicate that setting $\overline{\alpha} = \overline{\gamma} = 1$ is over-restrictive and that the reconstruction accuracy improves as a consequence of employing a non-informative prior

scheme, information sharing with the binomial prior leads to a significant improvement in the network reconstruction accuracy in all cases, while information sharing with the exponential prior leads to a significant improvement when the true network structures are sufficiently similar. A detailed analysis of hyperparameter inference shows that the Bayesian inference scheme is consistent for the binomial prior in the sense that the high posterior probability region of the hyperparameters concurs with the one that optimizes the network reconstruction accuracy. For the exponential prior, this consistency is only given when the data set size is sufficiently large; otherwise a more restrictive hyperprior (i.e. prior on β) is needed. On the other hand, a restrictive setting for the level-2 hyperparameters of the bino-

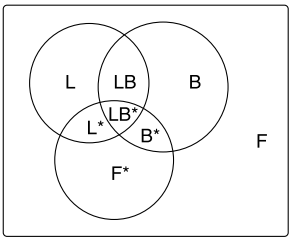

Fig. 9 Illustration of a hypothetical network scenario, where edges fall into several categories. Edges in sets L, LB, L^* and LB^* are true edges, which means they are included in the network corresponding to the current time series segment. Edges in sets L and LB are 'true positives' in that they contribute a score $A > 1$ to the likelihood. Edges in sets L^* and LB^* are 'false negatives', which contribute the neutral score of 1 to the likelihood. The edges in sets F^* and B^* are 'false positives', which contribute a score $A^* > A > 1$ to the likelihood. The edges in sets LB, LB^*, B^* and B are consistent with the prior network, all those in the complementary sets are not found in the prior network. Edges in set F are neither included in the network associated with the current segment, nor can they be found in the prior network. They also don't contribute any score to the log likelihood (i.e. they contribute a neutral score of 1 to the likelihood). An overview can be found in Table 2

mial prior is counter-productive, and better network reconstruction scores are obtained with a non-informative hyperprior.

5.2 Closed-form inference for the exponential prior

The results in Fig. 7 indicated that for the exponential prior, the Bayesian inference scheme might fail to find the hyperparameters that optimize the network reconstruction accuracy. Our conjecture is that this is not a consequence of poor mixing and convergence of the MCMC sampler, but intrinsic to the Bayesian inference scheme *per se*. As a demonstration, we reproduce the observation from Fig. 7 with a simpler model for which a closed-form expression of the posterior distribution of the hyperparameter can be derived. We consider the scenario depicted in Fig. 9, where edges of a hypothetical network can be divided into different categories, depending on whether or not they are true, supported by the data, or included in the prior network. An overview of the notation is presented in Table 2. With the simplifying assumption of posterior independence of the edges, the likelihood is given by

$$P(\boldsymbol{x}|\boldsymbol{G}) = A^{(n_L+n_{LB})} A^{*(n_{B^*}+n_{F^*})} \tag{57}$$

where n_S counts the number of elements in set S for network $\boldsymbol{G}$, and the symbols denoting the sets have been defined in Table 2. Assuming a uniform prior on β, the posterior distribution of the hyperparameter becomes:

$$\begin{aligned} P(\beta|\boldsymbol{x}) \propto P(\boldsymbol{x},\beta) &= \sum_{\boldsymbol{G}} P(\boldsymbol{x}|\boldsymbol{G}) P(\boldsymbol{G}|\beta) P(\beta) \\ &\propto \frac{1}{Z(\beta)} \sum_{\boldsymbol{G}} P(\boldsymbol{x}|\boldsymbol{G}) \exp\bigl(-\beta \bigl|\boldsymbol{G} - \boldsymbol{G}^0\bigr|\bigr) \end{aligned} \tag{58}$$

Table 2 Likelihood and prior scores for the edges contained in the sets defined in Fig. 9. The product of the prior and the likelihood defines the rank of the edge; the truth indicator is shown in the second column

Set	True edge	Supported by the data	Supported by the prior	Likelihood	Prior	Number of edges
L	yes	yes	no	A	$e^{-\beta}$	N_L
LB	yes	yes	yes	A	1	N_{LB}
LB^*	yes	no	yes	1	1	N_{LB^*}
L^*	yes	no	no	1	$e^{-\beta}$	N_{L^*}
B	no	no	yes	1	1	N_B
B^*	no	yes	yes	A^*	1	N_{B^*}
F^*	no	yes	no	A^*	$e^{-\beta}$	N_{F^*}
F	no	no	no	1	$e^{-\beta}$	N_F

where $\boldsymbol{G}^0$ represents our prior knowledge. Inserting (57) into (58) we get, with Eq. (31) for $Z(\beta)$ and under the assumption of a uniform prior on β:

$$
\begin{aligned}
P(\beta|\boldsymbol{x}) \propto \frac{1}{(1+e^{-\beta})^N} & \sum_{n_L=0}^{N_L}\sum_{n_{LB}=0}^{N_{LB}}\sum_{n_B=0}^{N_B}\sum_{n_F=0}^{N_F}\sum_{n_{L^*}=0}^{N_{L^*}}\sum_{n_{LB^*}=0}^{N_{LB^*}}\sum_{n_{B^*}=0}^{N_{B^*}}\sum_{n_{F^*}=0}^{N_{F^*}} \\
& \times \binom{N_L}{n_L}\binom{N_{LB}}{n_{LB}}\binom{N_B}{n_B}\binom{N_F}{n_F}\binom{N_{L^*}}{n_{L^*}}\binom{N_{LB^*}}{n_{LB^*}}\binom{N_{B^*}}{n_{B^*}}\binom{N_{F^*}}{n_{F^*}} \\
& \times A^{(n_L+n_{LB})}A^{*(n_{B^*}+n_{F^*})} \\
& \times \exp\big(-\beta[n_L+n_F+N_{LB}-n_{LB}+N_B-n_B \\
& + n_{L^*}+n_{F^*}+N_{LB^*}-n_{LB^*}+N_{B^*}-n_{B^*}]\big)
\end{aligned}
\tag{59}
$$

A plot of (59) is shown in Fig. 10. The optimal network reconstruction in terms of AUROC and AUPRC scores is achieved for a finite value of $\beta \approx 1$. The effect of the data set size is emulated by varying the settings of the parameters entering the likelihood. For small values of A and A^*, corresponding to small data sets, the posterior probability increases monotonically in β, and the Bayesian inference scheme intrinsically fails to find the range of hyperparameters that optimizes the network reconstruction accuracy. When we increase the data set size, this mismatch disappears, and the two regions concur. These findings are consistent with those presented in Fig. 7 and suggest that the observed mismatch is a genuine inference feature rather than an MCMC artifact.

To further analyse this effect, we have investigated the values of A and A^* for which the posterior distribution shows a peak for a finite value of β. Analytically, this corresponds to finding values for A and A^* such that the equation $\frac{dP(\beta|\boldsymbol{x})}{d\beta} = 0$ has a solution. Unfortunately, it is non-trivial to determine the existence of a solution analytically; we have therefore resorted to numerically calculating $\frac{dP(\beta|\boldsymbol{x})}{d\beta}$ for $\beta = 20$. At $\beta = 0$, we have $\frac{dP(\beta|\boldsymbol{x})}{d\beta} > 0$; therefore, if $\frac{dP(\beta|\boldsymbol{x})}{d\beta} < 0$ at $\beta = 20$, this indicates that the distribution has a peak on the interval $[\beta, 20]$. On the other hand, under the assumption of unimodality, $\frac{dP(\beta|\boldsymbol{x})}{d\beta} > 0$ at $\beta = 20$ indicates that the marginal posterior probability of β increases monotonically with β. The results of this analysis are shown in Fig. 11, which shows a clear phase shift towards distributions with a peak as A and A^* increase.

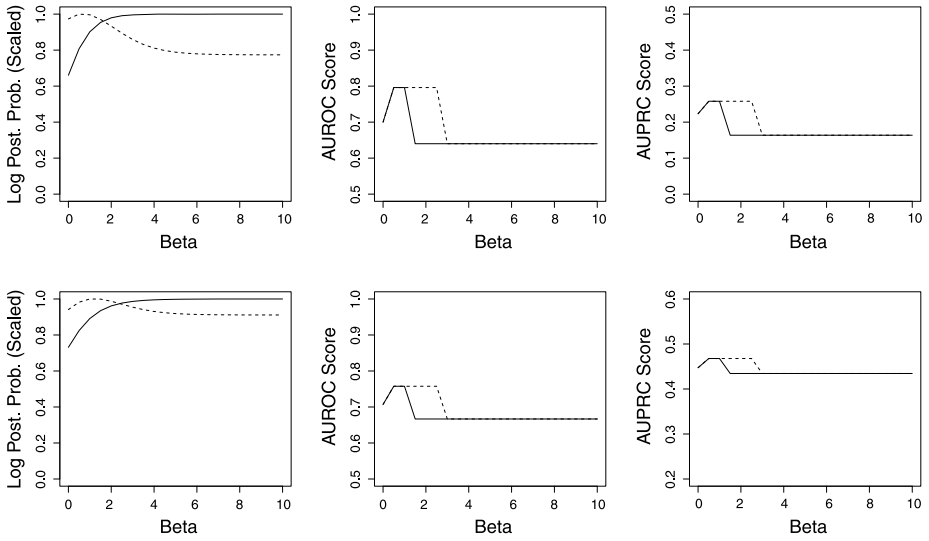

Fig. 10 Results for the simplified model with exponential prior. The *leftmost column* shows the marginal posterior distribution of β, computed from Eq. (59). The *middle column* shows the AUROC score as β varies. The *rightmost column* shows the AUPRC score as β varies. *Solid line*: $A = 2$, $A^* = 4$, *dashed line*: $A = 12$, $A^* = 14$. The *top* and *bottom rows* correspond to two different settings of the set sizes. *Top row*: $\{L : 15, LB : 0, B : 40, F : 60, L^* : 0, LB^* : 10, B^* : 25, F^* : 0\}$. *Bottom row*: $\{L : 15, LB : 20, B : 10, F : 25, L^* : 0, LB^* : 10, B^* : 20, F^* : 0\}$

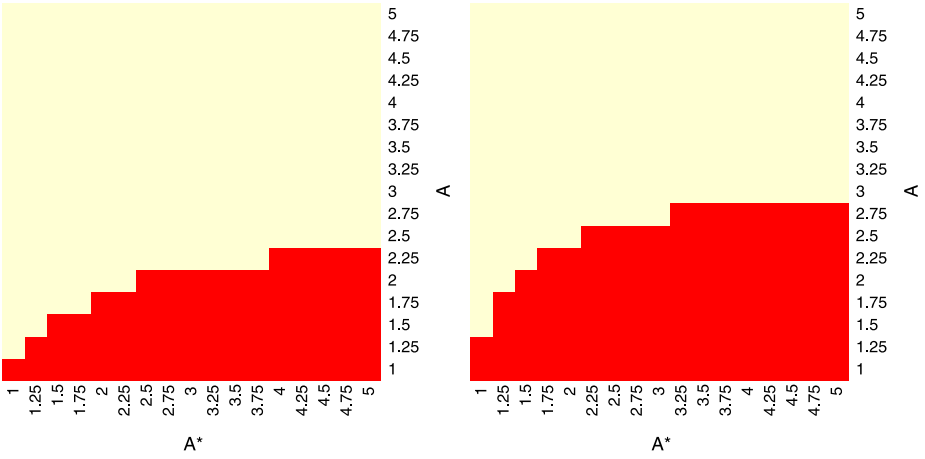

Fig. 11 Existence of a peak in the posterior distribution of β for the simplified model with exponential prior. The *two plots* show values of A and A^* for which the marginal posterior probability of β monotonically increases as β increases (*red tiles*), and those where the posterior probability decreases for high β (*white tiles*), indicating the existence of a peak in the distribution. We used the same settings of the set sizes as in Fig. 10. *Left*: $\{L : 15, LB : 0, B : 40, F : 60, L^* : 0, LB^* : 10, B^* : 25, F^* : 0\}$. *Right*: $\{L : 15, LB : 20, B : 10, F : 25, L^* : 0, LB^* : 10, B^* : 20, F^* : 0\}$

What does this analysis entail for the general applicability of the exponential prior? It is clear that when the data set size is too small, then the marginal posterior distribution of β will be biased towards high values. The exact definition of "too small" will crucially depend on the nature of the dataset. Given that we have shown in Sect. 5.1 that the binomial prior avoids this weakness and outperforms the exponential prior in terms of network reconstruction accuracy, we would recommend that this form of information sharing prior be used in preference of the exponential prior.

6 Real-world applications

6.1 Morphogenesis in *Drosophila melanogaster*

During its life-cycle, *Drosophila melanogaster* undergoes four major stages of morphogenesis: embryo, larva, pupa and adult. Arbeitman et al. (2002) obtained a gene expression time series covering all four stages. We have applied our methods to a subset of this gene expression time series consisting of eleven genes involved in wing muscle development. First, we investigated whether the changepoints inferred by our methods correspond to the known transitions between stages. Figure 12(a) shows the posterior probabilities of inferred changepoints for any gene using TVDBN-0 (unregularized by information sharing, see Table 1), while Figs. 12(c)–12(d) show the posterior probabilities for the information sharing methods. We compared this performance to the method proposed in Ahmed and Xing (2009), using the authors' software package TESLA (Fig. 12(b)). In addition, Robinson and Hartemink (2009) used a discrete non-homogeneous DBN to analyse the same data set, and a plot corresponding to Fig. 12(b) can be found in their paper.

An analysis of the results suggests that our non-homogeneous DBN methods are generally more successful than TESLA. We recover changepoints for all three transitions (embryo $\rightarrow$ larva, larva $\rightarrow$ pupa, and pupa $\rightarrow$ adult). As shown in Fig. 12(b), the last transition, pupa $\rightarrow$ adult, is less clearly detected with TESLA, and it is completely absent in Robinson and Hartemink (2009). Furthermore, TESLA and our method both detect additional changepoints during the embryo stage, which are missing in Robinson and Hartemink (2009). It is not implausible that additional transitions at the gene regulatory network level should occur within one morphogenic phase. One would expect that a complex gene regulatory network is unlikely to transition into a new phase all at once, and some pathways might have to undergo activational changes earlier in preparation for the morphogenic transition. However, a failure to detect a known transition represents a shortcoming of a method, and so we can say that in this aspect, our model appears to outperform the two alternative approaches.

In addition to the changepoints, we have inferred network structures for the morphogenic stages of embryo, larva, pupa and adult (see Fig. 13). An objective assessment of the reconstruction accuracy is not feasible due to the limited existing biological knowledge and the absence of a gold standard. However, our reconstructed networks show many similarities with the networks discovered by Robinson and Hartemink (2009), Guo et al. (2007) and Zhao et al. (2006). For instance, we recover the interaction between two genes, *eve* and *twi*. This interaction is also reported in Guo et al. (2007) and Zhao et al. (2006), while Robinson and Hartemink (2009) seem to have missed it. We also recover a cluster of interactions among the genes *myo61f*, *msp300*, *mhc*, *prm*, *mlc1* and *up* during all morphogenic phases. This result is not implausible, as all genes (except *up*) belong to the myosin family. However, unlike Robinson and Hartemink (2009), we find that *actn* also participates as a regulator in this cluster. There is some indication of this in Zhao et al. (2006), where *actn* is found to regulate *prm*. We have further validated our reconstructed networks using genetic and protein

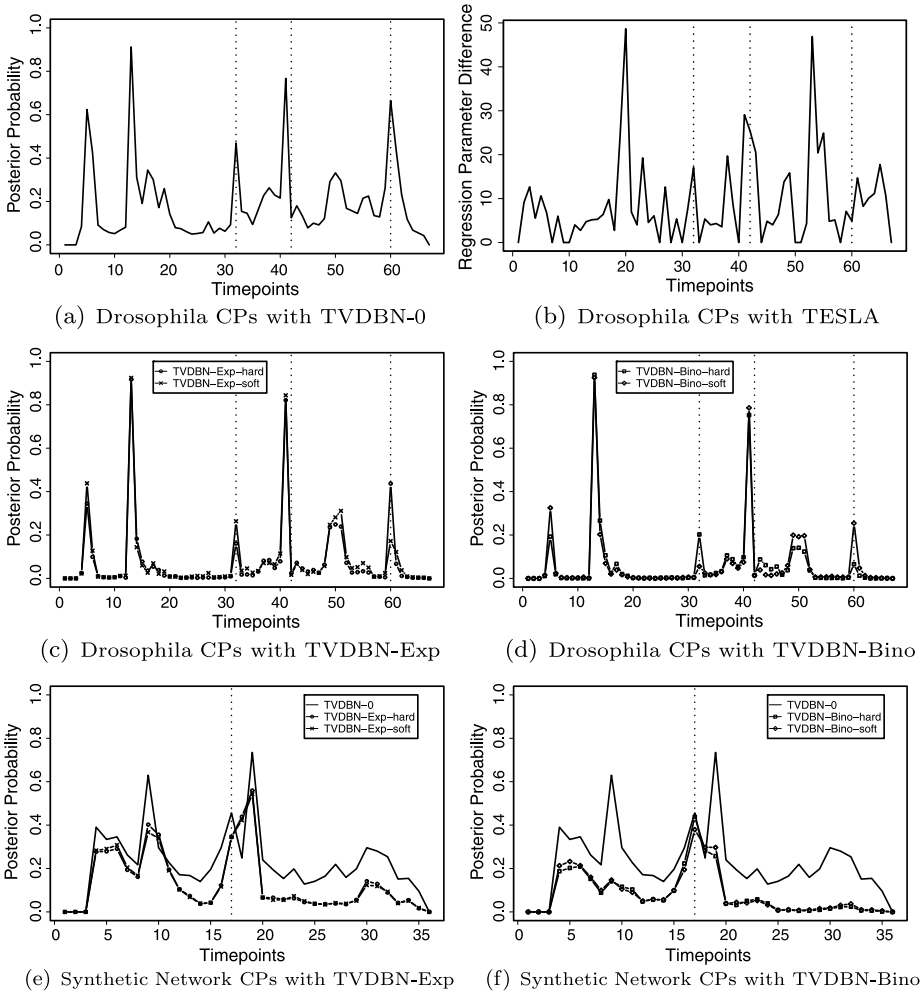

Fig. 12 Changepoints inferred from gene expression time series related to morphogenesis in *Drosophila melanogaster*, and synthetic biology in *Saccharomyces cerevisiae* (yeast). (**a**): TVDBN-0 changepoints for Drosophila (no information sharing). (**b**): TESLA, L1-norm of the difference of the regression parameter vectors associated with two adjacent time points plotted against time. (**c**) and (**d**): TVDBN changepoints for Drosophila with information sharing; the method is indicated by the legend. (**e**) and (**f**): TVDBN changepoints for the synthetic gene regulatory network in yeast. All figures using TVDBN plot the posterior probability of a changepoint occurring for any node at a given time (ordinate) against time (abscissa). In (**a**)–(**d**), the *vertical dotted lines* indicate the three morphogenic transitions, while in (**e**) and (**f**) the *line* indicates the boundary between the "switch on" (galactose) and "switch off" (glucose) phases

interactions recorded in the FLIGHT database (Sims et al. 2006). We found that a number of the inferred interactions over all segments correspond to interactions that have been reported in the literature. Some of these result from indirect interactions, where the intermediate gene is missing in the data. Table 3 gives an overview of the identified interactions with references to the biological literature.

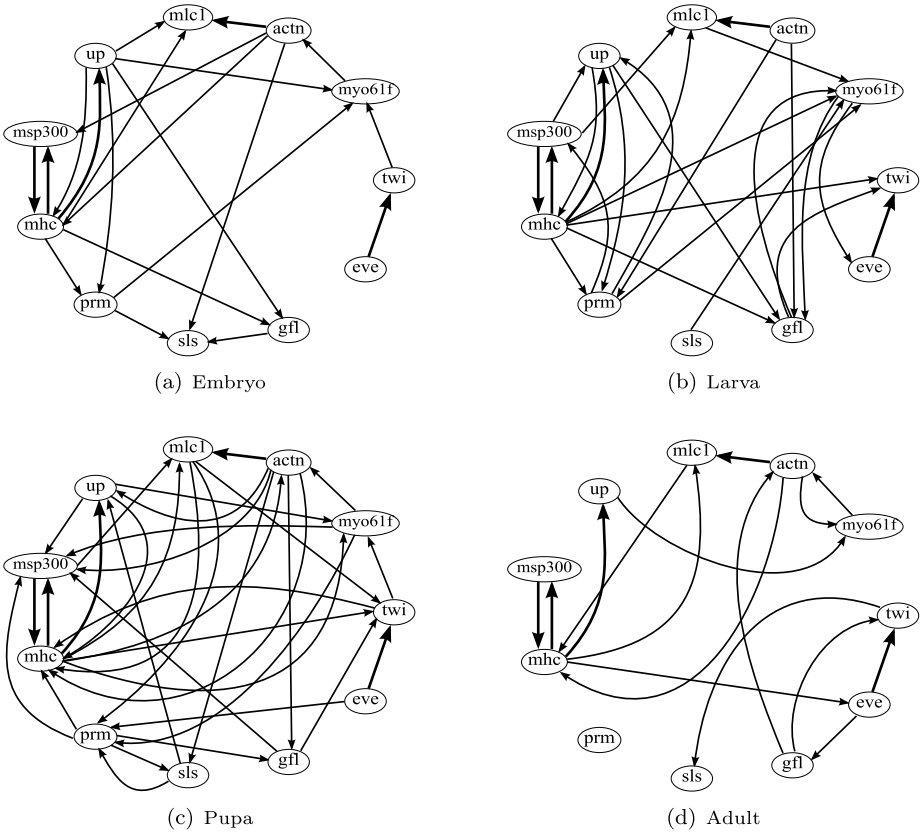

Fig. 13 Gene regulatory networks inferred from gene expression time series related to morphogenesis in *Drosophila melanogaster*, using TVDBN-Bino-hard. The networks were obtained by applying a threshold of 0.25 to the marginal posterior probabilities of the gene interactions. We have reconstructed a network for each morphological phase; interactions that were consistent across all four phases are marked in *bold*

6.2 Synthetic biology in *Saccharomyces cerevisiae*

Synthetic biology is a rapidly developing and highly topical discipline that aims to combine the biological sciences and engineering (Andrianantoandro et al. 2006). One of its aims is to design new gene regulatory networks in living cells. We make use of these endeavours by using gene expression time series obtained *in vivo* from cells with a known gene regulatory network structure to objectively assess the network reconstruction accuracy. Our work is based on Cantone et al. (2009), where the authors constructed a synthetic regulatory network with 5 genes in *Saccharomyces cerevisiae* (yeast). Then they measured gene expression time series with RT-PCR for 16 and 21 time points under two experimental conditions, related to the carbon source: galactose ("switch on"), and glucose ("switch off"). The authors applied two established state-of-the-art methods from computational systems biology to reconstruct the known underlying network from these time series. One is based on ODEs: ordinary differential equations (TSNI), the other is based on conventional DBNs (Banjo); see Cantone et al. (2009) for details. Both methods are optimization-based and only output a single network. By comparison with the known network, the authors calculated the

Table 3 Reconstructed interactions in the *Drosophila melanogaster* wing muscle development network that have been validated using the FLIGHT database (Sims et al. 2006)

Interaction	References	Interaction	Notes
actn ↔ *mhc*	Homyk and Emerson (1988); Nongthomba et al. (2003),; Montana and Littleton (2004)	Protein	Via missing gene *wupA*
actn → *up*	Homyk and Emerson (1988); Nongthomba et al. (2003)	Protein	Via missing gene *wupA*
eve → *twi*	Parkhurst and Ish-Horowicz (1991)	Protein	Via missing gene *RpIII40*
up ↔ *mhc*	Homyk and Emerson (1988); Nongthomba et al. (2003); Montana and Littleton (2004)	Protein	Direct interaction
actn → *msp300*	Formstecher et al. (2005)	Gene	Via missing gene *TSG101* or missing gene *Hrs*
actn → *sls*	Sanchez et al. (1999)	Gene	Direct Interaction
actn → *prm*	Formstecher et al. (2005)	Gene	Via missing gene *exo70*
prm ↔ *sls*	Sanchez et al. (1999); Formstecher et al. (2005)	Gene	Via missing gene *exo70* and present gene *actn*
sls → *up*	Sanchez et al. (1999); Formstecher et al. (2005)	Protein and Gene	Via missing gene *Act88F*

precision (proportion of predicted regulatory interactions in the network that are correct) and recall (proportion of predicted true interactions) scores. Figure 14 shows the true networks, the reconstructed networks for TSNI and Banjo, as well as the reconstructed networks using TVDBN-Bino-hard, where we have applied a threshold of 0.75 to the inferred marginal posterior probabilities of the gene interactions to obtain absence/presence values for the edges.[8]

In our study, we merged the time series from the two experimental conditions under exclusion of the boundary point,[9] and applied the non-homogeneous DBNs from Table 1. Figures 12(e) and 12(f) show the inferred marginal posterior probabilities of potential changepoints. The salient changepoint is at the boundary between the "switch on" (galactose) and "switch off" (glucose) phases, confirming that the true changepoint is consistently identified. However, in the absence of information sharing, we observe additional spurious changepoints. These changepoints are successfully suppressed with the proposed Bayesian information-coupling schemes, with the binomial prior having a slightly stronger regularizing effect than the exponential one.

As described in Sect. 4, the Bayesian inference scheme provides a ranking of the potential gene interactions in terms of their marginal posterior probabilities. From this ranking we computed the precision-recall curves (Davis and Goadrich 2006) shown in Fig. 15. By using information sharing, our non-homogeneous DBN outperforms Banjo and TSNI both in the "switch on" and the "switch off" phase. The information sharing methods also perform better than TVDBN-0 on the "switch off" data, but are slightly worse on

[8]Note that while our TVDBN methods are in principle capable of inferring the type of interaction (activation or inhibition) by sampling regression weights, we have not investigated this for the purpose of this paper. Therefore in Fig. 14, the arrows in the networks reconstructed using TVDBN-Bino-hard only record the presence or absence of an interaction, and not its type.

[9]When merging two time series $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$, only the pairs $x_i \to x_j$ and $y_i \to y_j$ are presented to the DBN, while the pair $x_m \to y_1$ is excluded due to the obvious discontinuity.

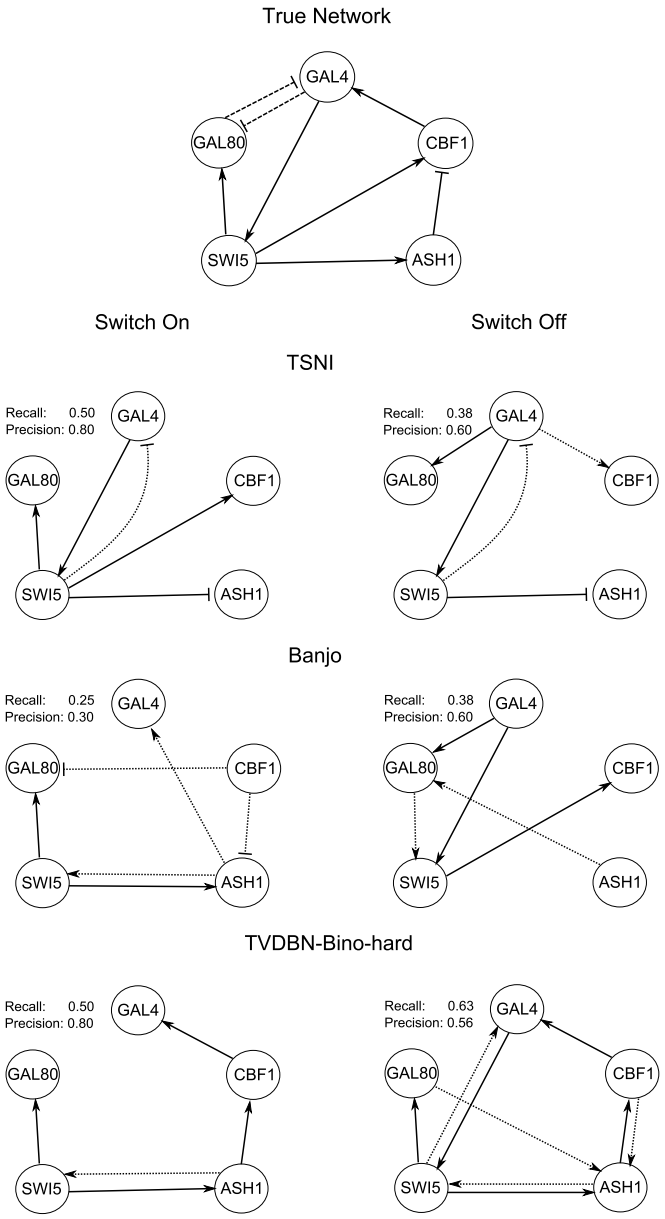

Fig. 14 True and reconstructed networks for a synthetic biology gene regulatory network in *Saccharomyces cerevisiae* (yeast). *Top row*: True network as described in Cantone et al. (2009). *2nd row*: Networks reconstructed using TSNI, a method based on ordinary differential equations (ODEs). *3rd row*: Networks reconstructed using Banjo, a conventional DBN. *Bottom row*: Networks reconstructed using TVDBN-Bino-hard, applying a threshold of 0.75 on the marginal posterior probabilities of gene interactions to obtain an absence/presence value for each edge. All reconstructed networks were reconstructed from two gene expression time series obtained with RT-PCR in two experimental conditions, reflecting the switch in the carbon source from galactose ("switch on") to glucose ("switch off"). The *dashed lines* in the true network indicate protein-protein regulation. The *dotted lines* in the reconstructed networks indicate false positive gene interactions. The networks found by Banjo and TSNI are reproduced from Cantone et al. (2009)

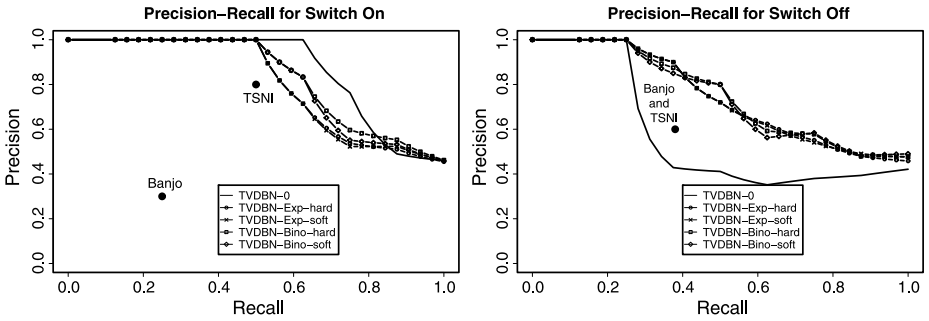

Fig. 15 Reconstruction of a gene regulatory network designed with synthetic biology in *Saccharomyces cerevisiae*. The network was reconstructed from two gene expression time series obtained with RT-PCR in two experimental conditions, reflecting the switch in the carbon source from galactose ("switch on") to glucose ("switch off"). The reconstruction accuracy of the methods proposed in Sect. 3 and Table 1, where the legend is explained, is shown in terms of precision (*vertical axis*)–recall (*horizontal axis*) curves. Results were averaged over 10 independent MCMC simulations. For comparison, fixed precision/recall scores are shown for two state-of-the-art methods, as reported in Cantone et al. (2009): Banjo, a conventional DBN, and TSNI, a method based on ordinary differential equations (ODEs)

the "switch on" data. Cantone et al. (2009) showed that in general, the reconstruction accuracy on the "switch off" data is poorer than on the "switch on" data. This lends credence to our results, suggesting that the proposed Bayesian regularization and information sharing schemes substantially improve the gene network reconstruction accuracy on the poorer time series segment, at the cost of a slightly degraded performance on the stronger one. Overall, the effect of information sharing is a performance improvement, as shown by the average areas under the PR curves, averaged over both phases ("switch on and off"): TVDBN-0 = 0.68, TVDBN-Exp-hard = 0.74, TVDBN-Exp-soft = 0.74, TVDBN-Bino-hard = 0.76, TVDBN-Bino-soft = 0.75.

We complete our investigation of the yeast network by providing an analysis of the network reconstruction performance (in terms of average area under the PR curve) as the hyperparameters vary. This is analogous to the evaluation we performed in Sect. 5.1 on simulated data. The results are shown in Fig. 16. As expected, higher values of the hyperparameter β, which correspond to stronger coupling, result in a better performance (Fig. 16(a)). Figure 16(b) shows the effect of different values for κ in Eq. (37). There is no discernible trend, which suggests that the strength of the coupling scheme does not matter much for this application, and that when moving closer to the hard coupling scheme (higher κ while keeping the mean μ of the gamma distribution fixed), the network reconstruction performance does not change significantly. The results obtained with the binomial prior demonstrate that, for this application, encouraging agreement related to the presence of interactions is more important than agreement related to the absence of interactions (Fig. 16(c)). Figure 16(d) confirms that our sampled hyperparameters a and b are in the correct range for optimal network reconstruction.

7 Discussion

In the present paper we have addressed some of the challenges encountered in systems biology when attempting to reconstruct gene regulatory networks from gene expression time series. We have looked at the case where the network structure may change

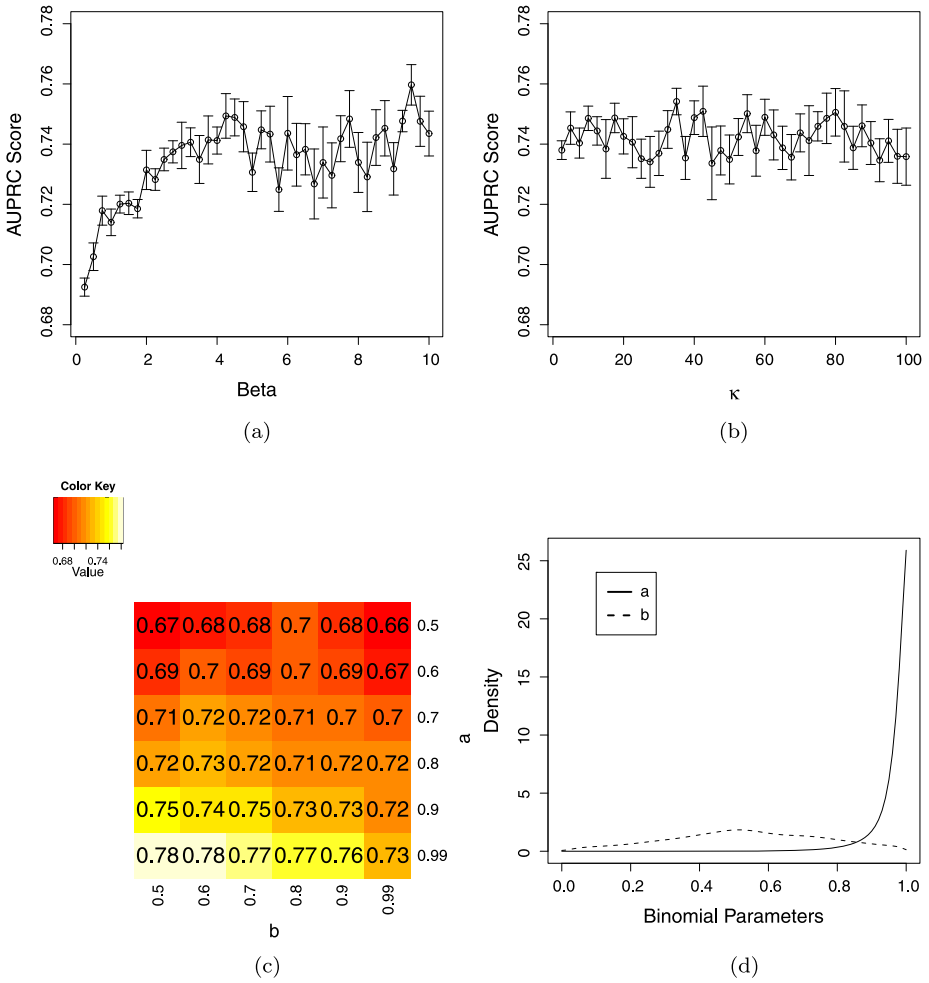

Fig. 16 Effect of the hyperparameters on the reconstruction of a known gene regulatory network from synthetic biology in yeast. The reconstruction accuracy is measured in terms of the average area under the precision–recall curve (AUPRC). Results were averaged over 10 independent MCMC simulations. (**a**): Variation of the hyperparameter β for the exponential information sharing prior with hard coupling. (**b**): Variation of the level-2 hyperparameter κ for the exponential prior with soft coupling, where the mean of the gamma distribution is kept fixed at $\mu = 5$. (**c**): Variation of hyperparameters a and b for the binomial prior. (**d**): Sampled distributions of hyperparameters a and b for the binomial prior with hard coupling. These distributions were obtained from the sampled values of the level-2 hyperparameters α, $\overline{\alpha}$, γ, $\overline{\gamma}$ using a kernel density estimator with the beta distribution from Eq. (45)

over time due to developmental or environmental causes. To deal with this situation, we have developed a non-homogeneous DBN, which has various advantages over existing schemes: it does not require the data to be discretized (as opposed to Robinson and Hartemink 2009, 2010); it allows the network structure to change with time (as opposed to Grzegorczyk and Husmeier 2009, 2011); it includes four different regularization schemes based on inter-time segment information sharing (as opposed to Lèbre 2007;

Lèbre et al. 2010); and it allows all hyperparameters to be inferred from the data via a consistent Bayesian inference scheme (as opposed to Ahmed and Xing 2009).

We note that the model of Robinson and Hartemink (2009, 2010) is conceptually similar to our exponential information sharing prior with hard coupling described in Sect. 3.2. By including three alternative information sharing schemes, we have extended the model of Robinson and Hartemink (2009, 2010) in two further respects:

(1) We allow for different penalties between edges and non-edges. The method in Robinson and Hartemink (2009, 2010) simply penalizes the number of different edges, i.e. the Hamming distance, between two adjacent structures. This corresponds to the approach taken for the exponential prior in Sects. 3.2 and 3.3. The inclusion of an extra edge leads to the same penalty as the deletion of an existing edge. This might not always be appropriate. Removing a rate-limiting reaction step of a critical signalling pathway is a more substantial change than including some redundant bypass pathway. Our two models based on the binomial prior (Sects. 3.4 and 3.5) allow for that by introducing different prior penalties for the deviation between edges and for the deviation between non-edges. In Sect. 5.1 we have experimentally shown that an information sharing approach based on different penalties for edges and non-edges can outperform the simpler approach when the number of changes among segments is small, but non-zero.
(2) We allow for different nodes of the network to have different penalty terms. The model in Robinson and Hartemink (2009, 2010) has a single hyperparameter for penalizing differences between structures: λ_s. This might not be appropriate if different subnetworks are conserved to a different degree. For instance, we would assume that molecular network substructures related to generic functionality, e.g. to maintain an essential baseline metabolism, are conserved to a greater extent than more peripheral pathways. By introducing node-dependent hyperparameters, the priors described in Sects. 3.3 and 3.5 generalize the approach in Robinson and Hartemink (2009, 2010) by allowing different parts of the network to be conserved during the temporal process to a different extent.

A further difference to Robinson and Hartemink (2009, 2010) merits some additional discussion. In our model, the changepoints are node-dependent. This gives us extra model flexibility, which is biologically motivated: on infection of an organism by a pathogen, genes involved in defence pathways are likely to be up-regulated, while others are not. Hence, it is plausible that different genes respond to changes in the environment differently, and this is directly incorporated in our model. In Robinson and Hartemink (2010), node-specific changepoints can be obtained indirectly: the calculation of the sufficient statistics for computing the marginal likelihood depends on the intervals during which each parent set is active. The marginal likelihood is recomputed for epochs, where an epoch is the union of consecutive time intervals during which a node-dependent substructure does not change. Since these unions of sets can be different for different nodes, the model does allow different changepoint sets to be associated with different nodes. However, there is a considerable price to pay for that: a changepoint in Robinson and Hartemink (2010) is intrinsically associated with a structure change, whereas in our model, a changepoint can be related to either a structure or a parameter change, or both. This gives us extra model flexibility, which is important for systems biology: when adapting to environmental change, several molecular interactions in signalling pathways may be up- or down-regulated, rather than switched on or off altogether.

An evaluation on simulated data has demonstrated that the proposed Bayesian regularization and information sharing schemes lead to an improved performance over Lèbre (2007) and Lèbre et al. (2010). We have carried out a comparative evaluation of four different information coupling schemes: a binomial versus an exponential prior, and hard versus soft

information coupling. This comparison has revealed that the binomial prior allows for more consistent inference of the right level of information sharing, while the exponential prior tends to enforce overly-strong information sharing. The difference between hard and soft information coupling seems negligible in the scenarios we investigated. A detailed investigation of the hyperparameter inference has allowed us to improve the MCMC sampler for better convergence, and to explore the limitations of the exponential information sharing prior.

The application of our method to gene expression time series taken during the life cycle of *Drosophila melanogaster* has revealed better agreement with known morphogenic transitions than the methods of Robinson and Hartemink (2009, 2010) and Ahmed and Xing (2009), and we have been able to identify several gene and protein interactions that are known from the literature. In an application to data from a topical study in synthetic biology (Cantone et al. 2009), our methods have outperformed two established network reconstruction methods from computational systems biology, and information sharing has allowed us to reconstruct the true underlying gene network with higher overall precision and recall than would have been possible without it.

We have investigated the performance of our methods on datasets which arise from gene regulatory networks with temporal changes in the structure of the network. There are several special cases of this situation which merit further discussion. The simplest case occurs when the changes of the underlying process are limited to parameter changes, and the true structure of the network remains constant. We have shown in Sect. 5.1 that our methods can deal with this situation effectively thanks to information sharing among segments. A more complicated case could involve a reoccurring event that causes certain gene interactions to switch on or off, leading to repeated network structures. For example, in a circadian clock system such as Locke et al. (2006), Pokhilko et al. (2010), the absence of sunlight might deactivate the interaction between two genes in the network, causing its structure to change from A to B.[10] If gene expression levels are measured both during the day and at night for three days, then we will observe a sequence like ABABAB. While our methods can in principle represent repeated segments, the multiple changepoint process was not designed with this in mind. A better model for repeated segments might be a Hidden Markov Model (HMM), where each hidden state corresponds to a network structure, and transitions between states correspond to changes in the structure, in the same vein as applied to changing tree structures in phylogeny (Husmeier and McGuire 2003). The disadvantage of using HMMs is that they impose a geometric distribution on the segment lengths, and in that respect our changepoint process is more flexible. To have the same flexibility with HMMs, model extensions along the lines of hierarchical HMMs or HMMs with weighting times could be pursued, as known from speech processing, but this would come at significantly increased computational costs. Hence, this approach only appears to make sense if there is strong prior indication that repetitions occur.

An interesting topic for future work is to investigate other functional forms of the information sharing mechanism. In our work, we have investigated four different models, based on an exponential versus binomial distribution, with or without gene-specific hyperparameters. It has recently come to our attention that Wang et al. (2011) have experimented with a different approach, which effectively combines our exponential prior with an additional factor that encourages network sparsity. Sparsity in our model is encouraged by the truncated Poisson prior of Eq. (4), as explained in the paragraph under Eq. (30). It would be

[10] Note that our definition of a deactivated gene interaction includes interactions that no longer occur because one of the interacting genes is no longer expressed.

interesting to explore the effect of the additional factor used in Eq. (7) of Wang et al. (2011) in the context of gene network reconstruction.

Reconstructing gene regulatory networks from transcriptional profiles remains a challenging problem, which a flurry of ongoing methodological developments in the computational systems biology community are trying to address. We believe that our paper adds a valuable contribution to this field, by presenting a consistent and flexible Bayesian model for the case where the network structures change over time.

Acknowledgements Most of the work was carried out while Dirk Husmeier was employed at Biomathematics and Statistics Scotland, and the work was supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). This work was partly funded by EU FP7 grant "Timet". Frank Dondelinger's PhD research is partly funded by the Engineering and Physical Sciences Research Council (EPSRC).

References

Ahmed, A., & Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, *106*, 11878–11883.

Andrianantoandro, E., Basu, S., Karig, D., & Weiss, R. (2006). Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology*, *2*(1), E1–E14.

Andrieu, C., & Doucet, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, *47*(10), 2667–2676.

Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R., & White, K. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, *297*(5590), 2270–2275.

Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., & Cosma, M. P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, *137*(1), 172–181.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning* (p. 240). New York: ACM.

Dondelinger, F. (2012). *A machine learning approach to reconstructing signalling pathways and interaction networks in biology*. PhD thesis, University of Edinburgh (in preparation).

Dondelinger, F., Lebre, S., & Husmeier, D. (2010). Heterogeneous continuous dynamic Bayesian networks with flexible structure and inter-time segment information sharing. In *Proceedings of the 27th international conference on machine learning (ICML)*.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. (2005). Protein interaction mapping: a Drosophila case study. *Genome Research*, *15*(3), 376.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Grzegorczyk, M., & Husmeier, D. (2009). Non-stationary continuous dynamic Bayesian networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 22, pp. 682–690).

Grzegorczyk, M., & Husmeier, D. (2011). Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning*, *83*, 355–419.

Guo, F., Hanneke, S., Fu, W., & Xing, E. (2007). Recovering temporally rewiring networks: a model-based approach. In *Proceedings of the 24th international conference on machine learning* (p. 328). New York: ACM.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Homyk, T. Jr, & Emerson, C. Jr (1988). Functional interactions between unlinked muscle genes within haploinsufficient regions of the Drosophila genome. *Genetics*, *119*(1), 105.

Husmeier, D., & McGuire, G. (2003). Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution*, *20*(3), 315–337.

Husmeier, D., Dondelinger, F., & Lèbre, S. (2010). Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks. In J. Lafferty (Ed.), *Proceedings of the twenty-fourth annual conference on neural information processing systems (NIPS)* (Vol. 23, pp. 901–909). New York: Curran Associates.

Kolar, M., Song, L., & Xing, E. (2009). Sparsistent learning of varying-coefficient models with structural changes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 22, pp. 1006–1014).

Larget, B., & Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, *16*(6), 750–759.

Lèbre, S. (2007). *Stochastic process analysis for genomics and dynamic Bayesian networks inference*. PhD thesis, Université d'Evry-Val-d'Essonne, France.

Lèbre, S., Becq, J., Devaux, F., Lelandais, G., & Stumpf, M. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, *4*, 130.

Locke, J., Kozma-Bognár, L., Gould, P., Fehér, B., Kevei, E., Nagy, F., Turner, M., Hall, A., & Millar, A. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology*, *2*(1), 59.

Montana, E., & Littleton, J. (2004). Characterization of a hypercontraction-induced myopathy in Drosophila caused by mutations in *mhc*. *The Journal of Cell Biology*, *164*(7), 1045.

Nongthomba, U., Cummins, M., Clark, S., Vigoreaux, J., & Sparrow, J. (2003). Suppression of muscle hypercontraction by mutations in the myosin heavy chain gene of *Drosophila melanogaster*. *Genetics*, *164*(1), 209.

Parkhurst, S., & Ish-Horowicz, D. (1991). *WIMP*, a dominant maternal-effect mutation, reduces transcription of a specific subset of segmentation genes in Drosophila. *Genes & Development*, *5*(3), 341.

Pokhilko, A., Hodge, S., Stratford, K., Knox, K., Edwards, K., Thomson, A., Mizuno, T., & Millar, A. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, *6*(1), 416.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., & Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE*, *5*(2), e9202.

Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, *50*(3), 747–758.

Robinson, J. W., & Hartemink, A. J. (2009). Non-stationary dynamic Bayesian networks. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 21, pp. 1369–1376). San Mateo: Morgan Kaufmann.

Robinson, J., & Hartemink, A. (2010). Learning non-stationary dynamic Bayesian networks. *Journal of Machine Learning Research*, *11*, 3647–3680.

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Roeder, L., Euzenat, J., Rechenmann, F., & Jacq, B. (1999). Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an internet database. *Nucleic Acids Research*, *27*(1), 89.

Sims, D., Bursteinas, B., Gao, Q., Zvelebil, M., & Baum, B. (2006). FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets. *Nucleic Acids Research*, *34*(suppl 1), D479.

Talih, M., & Hengartner, N. (2005). Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society B*, *67*(3), 321–341.

Wang, Z., Kuruoglu, E., Yang, X., Xu, Y., & Huang, T. (2011). Time varying dynamic Bayesian network for non-stationary events modeling and online inference. *IEEE Transactions on Signal Processing*, *4*(59), 1553.

Werhli, A. V., & Husmeier, D. (2008). Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *Journal of Bioinformatics and Computational Biology*, *6*(3), 543–572.

Xuan, X., & Murphy, K. (2007). Modeling changing dependency structure in multivariate time series. In Z. Ghahramani (Ed.), *Proceedings of the 24th annual international conference on machine learning (ICML 2007)* (pp. 1055–1062). New York: Omnipress.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques* (pp. 233–243). Amsterdam: Elsevier.

Zhao, W., Serpedin, E., & Dougherty, E. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, *22*(17), 2129.