Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 4 2012 *Article* 7

A Non-Homogeneous Dynamic Bayesian Network with Sequentially Coupled Interaction Parameters for Applications in Systems and Synthetic Biology

Marco Grzegorczyk, *Department of Statistics, TU Dortmund University*
Dirk Husmeier, *School of Mathematics and Statistics, University of Glasgow, and Biomathematics and Statistics Scotland (BioSS), Edinburgh*

Recommended Citation:
Grzegorczyk, Marco and Husmeier, Dirk (2012) "A Non-Homogeneous Dynamic Bayesian Network with Sequentially Coupled Interaction Parameters for Applications in Systems and Synthetic Biology," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 4, Article 7.
10.1515/1544-6115.7

A Non-Homogeneous Dynamic Bayesian Network with Sequentially Coupled Interaction Parameters for Applications in Systems and Synthetic Biology

Marco Grzegorczyk and Dirk Husmeier

Abstract

An important and challenging problem in systems biology is the inference of gene regulatory networks from short non-stationary time series of transcriptional profiles. A popular approach that has been widely applied to this end is based on dynamic Bayesian networks (DBNs), although traditional homogeneous DBNs fail to model the non-stationarity and time-varying nature of the gene regulatory processes. Various authors have therefore recently proposed combining DBNs with multiple changepoint processes to obtain time varying dynamic Bayesian networks (TV-DBNs). However, TV-DBNs are not without problems. Gene expression time series are typically short, which leaves the model over-flexible, leading to over-fitting or inflated inference uncertainty. In the present paper, we introduce a Bayesian regularization scheme that addresses this difficulty. Our approach is based on the rationale that changes in gene regulatory processes appear gradually during an organism's life cycle or in response to a changing environment, and we have integrated this notion in the prior distribution of the TV-DBN parameters. We have extensively tested our regularized TV-DBN model on synthetic data, in which we have simulated short non-homogeneous time series produced from a system subject to gradual change. We have then applied our method to real-world gene expression time series, measured during the life cycle of *Drosophila melanogaster*, under artificially generated constant light condition in *Arabidopsis thaliana*, and from a synthetically designed strain of *Saccharomyces cerevisiae* exposed to a changing environment.

KEYWORDS: gene regulatory networks, non-stationary gene expression time series, time varying dynamic Bayesian networks, Bayesian regularization, Bayesian multiple changepoint processes, reversible jump Markov chain Monte Carlo

Author Notes: Marco Grzegorczyk is funded by the German Research Foundation (DFG, GR3853/ 1-1). Parts of the work were carried out while Dirk Husmeier was employed at Biomathematics & Statistics Scotland and funded by RESAS and EU FP7 grant ”Timet”.

Marco Grzegorczyk, Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany.

Dirk Husmeier, School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QW, UK, and Biomathematics and Statistics Scotland (BioSS), Edinburgh EH9 3JZ, UK.

1 Introduction

Molecular pathways consisting of interacting proteins underlie the major functions of living cells, and a central goal of molecular biology is to understand the regulatory mechanisms of gene transcription and protein synthesis. In recent years, there has been considerable interest in learning the structure of gene regulatory networks from transcriptomic time series with dynamic Bayesian networks (DBNs) (Ong *et al.* (2002); Husmeier (2003); Smith *et al.* (2002)), which is motivated by related applications to the inference of neural networks from electrochemical spike train data (Smith *et al.* (2006); Echtermeyer *et al.* (2009)) and follows up on the successful reconstruction of protein signalling pathways with static Bayesian networks (Sachs *et al.* (2005)).

The standard assumption underlying DBNs is that of homogeneity: temporal processes and the time-series they generate are assumed to be governed by a homogeneous Markov relation, while regulatory interactions and signal transduction processes in the cell are usually adaptive and change in response to external stimuli. A method that does not allow for this non-stationarity is too restrictive and can potentially lead to artifacts and 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); Kolar *et al.* (2009); Song *et al.* (2009); Lèbre *et al.* (2010)). Various authors have proposed relaxing the homogeneity assumption by complementing the traditional homogeneous DBN with a Bayesian multiple changepoint process (MCP) (Robinson and Hartemink (2009, 2010); Grzegorczyk and Husmeier (2009, 2011); Lèbre *et al.* (2010)).

Each time series segment defined by two demarcating changepoints is associated with separate node-specific DBN parameters, and in this way the conditional probability distributions are allowed to vary from segment to segment. An attractive feature of this approach is that under certain regularity conditions, most notably parameter independence and conjugacy of the prior, the parameters can be integrated out in closed form in the likelihood. The inference task thus reduces to sampling the network structure as well as the number and location of changepoints from the posterior distribution, which can be effected with reversible jump Markov chain Monte Carlo (RJMCMC) (Green (1995)), as in Robinson and Hartemink (2009) and Lèbre *et al.* (2010), or with dynamic programming (Fearnhead (2006)), as in Grzegorczyk and Husmeier (2011).

Although the combination of DBNs with MCPs leads to a model with a considerable degree of flexibility, such an approach is not without problems.

Gene expression time series, which are typically produced with postgenomic high-throughput platforms like microarrays or RNA-seq, tend to be short. The consequence is that the information in the data is usually insufficient for proper inference of all degrees of freedom, leading to over-fitting or inflated inference uncertainty.

A Bayesian approach to this problem is to regularize the model with an informative parameter prior. In fact, an inspection of the nature of biological processes suggests that the common assumption of parameter independence is questionable. Consider the cellular processes during an organism's development (morphogenesis) or its adaptation to changing environmental conditions. The assumption of a homogeneous process with constant parameters is obviously over-restrictive in that it fails to allow for the non-stationary nature of the processes. However, complete parameter independence is over-flexible in that it ignores the evolutionary aspect of developmental and adaptation processes. Given a regulatory network at a given time interval in an organism's life cycle, it is unrealistic to assume that at an adjacent time interval, nature has completely reinvented the regulatory pathways. Instead, we would assume that the interaction strengths at any interval in time are, overall, similar to those at the previous time interval. The objective of the present article is to describe how this idea can be implemented in the model, and which adaptations are required for the inference scheme.

There are various articles in the recent bioinformatics and machine learning literature that present somewhat related ideas. Lèbre *et al.* (2010) propose a DBN with a time-varying structure in the context of biological processes that act on long time scales during an organism's life cycle. However, the resulting high degree of model flexibility causes over-fitting problems when inference is based on the short gene expression time series that are typically available from postgenomic high-throughput platforms. To remedy this, Husmeier *et al.* (2010) introduce information sharing between adjacent time series segments by penalizing structural changes in the time varying network with a restrictive prior. While this has proven to considerably improve the network reconstruction accuracy, the model is based on independent parameter vectors that lack a mechanism to encourage a gradual evolution of the interaction weights. The model of Robinson and Hartemink (2009, 2010) is based on similar principles as the one in Husmeier *et al.* (2010), with the additional assumption that the data have been discretized.

There are also various articles from the signal processing community that are related to our work. The model proposed in Punskaya *et al.* (2002) is akin to the one of Lèbre *et al.* (2010), with the flexibility to learn Markov chains of order higher than one. Information is only shared among different parameter vectors via a common scalar scale hyperparameter, though, which does not provide the sort of more explicit information sharing motivated by our discussion above. Unlike Punskaya *et al.* (2002), Lèbre *et al.* (2010) and our present work, which are based

on a switching piecewise homogeneous autoregressive process, the models in Andrieu *et al.* (2003) and Wang *et al.* (2011) are based on continuously time-varying autoregressive processes in a filtering context, 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 the experiments, and inference is more naturally placed in a smoothing context. In the same vein as other recent applications of DBNs in systems biology (Robinson and Hartemink (2009, 2010); Grzegorczyk and Husmeier (2009, 2011); Lèbre *et al.* (2010); Husmeier *et al.* (2010)), our work thus follows this latter paradigm and aims to infer the model structure based on the posterior distribution conditional on the whole data, by marginalizing out the parameters in closed form.

This article is organized as follows. Section 2 provides a brief motivation for our approach, and the methodological details are explained in Section 3. Section 4 provides an overview of various benchmark data on which we have tested the proposed scheme, and in Section 5 we describe how we have applied the method in practice. The results on the benchmark data are presented in Section 6, and the paper concludes in Section 7 with a final discussion.

2 A simple illustration

We use the following notational convention. Lower case bold-face letters represent vectors, upper case bold-face letters represent matrices, and the superscript "T" denotes vector/matrix transposition. Our notation does not distinguish between random variables and their realizations. Hence $\mathbf{x}$ and $\mathbf{X}$ are different quantities, the former being a vector, and the latter being a matrix. In particular, note that $\mathbf{x}$ is *not* a realization of $\mathbf{X}$.

We consider a simple linear regression

$$y = \mathbf{w}^{\mathsf{T}}\mathbf{x} + \varepsilon, \tag{1}$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is a vector of (regression or interaction) parameters, y is the observed target variable, ε is additive Gaussian iid noise, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and the superscript "$\mathsf{T}$" denotes matrix transposition, with the convention that the untransposed vector is a column vector. Given a training set $\mathcal{D} = \{(\mathbf{x}_t, y_t), t = 1, \ldots, T\}$, we collect the targets in the vector $\mathbf{y} = (y_1, \ldots, y_T)^{\mathsf{T}}$ and define the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. The likelihood is given by

$$\begin{aligned} P(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{t=1}^{T} P(y_t|\mathbf{x}_t, \mathbf{w}) &= \left(\frac{1}{2\pi\sigma^2}\right)^{T/2} \exp\left(-\sum_{t=1}^{T} \frac{(y_t - \mathbf{w}^{\mathsf{T}}\mathbf{x}_t)^2}{2\sigma^2}\right) \\ &= \mathcal{N}(\mathbf{X}^{\mathsf{T}}\mathbf{w}, \sigma^2\mathbf{I}), \end{aligned} \tag{2}$$

where $\mathbf{I}$ denotes the unit matrix. We assume a Gaussian distribution with mean vector, $\mathbf{m}_0$ and covariance matrix $\mathbf{\Sigma}_0$ for the regression parameters

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_0). \tag{3}$$

From Bayes' rule

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{X}, \mathbf{w})P(\mathbf{w})}{P(\mathbf{y}|\mathbf{X})}, \tag{4}$$

and the application of standard Gaussian integrals (see, e.g., Bishop (2006), Section 3.3) we get for the posterior distribution of the regression parameters:

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{m}, \mathbf{\Sigma}), \tag{5}$$

where $\mathbf{m} = \mathbf{\Sigma}(\mathbf{\Sigma}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\mathbf{X}\mathbf{y})$ and $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}\mathbf{X}^\mathsf{T}$. Let us now assume that we have a changepoint, $\tau \in \{1, \ldots, T-1\}$, that divides the data into two subsets:

$$\mathcal{D}_1 = \{(\mathbf{x}_t, y_t), t = 1, \ldots, \tau\}; \qquad \mathcal{D}_2 = \{(\mathbf{x}_t, y_t), t = \tau + 1, \ldots, T\}.$$

Both subsets are modelled with the linear model of Equation (1), but with different regression parameter vectors, $\mathbf{w}_1$ and $\mathbf{w}_2$. Introducing the subsequent definitions $\mathbf{y}_1 = (y_1, \ldots, y_\tau)^\mathsf{T}, \mathbf{y}_2 = (y_{\tau+1}, \ldots, y_T)^\mathsf{T}$, $\mathbf{X}_1 = (\mathbf{x}_1, \ldots, \mathbf{x}_\tau)$, and $\mathbf{X}_2 = (\mathbf{x}_{\tau+1}, \ldots, \mathbf{x}_T)$ and imposing the prior $P(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_0)$ of Equation (3) onto both $\mathbf{w}_1$ and $\mathbf{w}_2$, we get for the posterior distributions:

$$\begin{aligned} P(\mathbf{w}_1|\mathbf{y}_1, \mathbf{X}_1) &= \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1) \\ P(\mathbf{w}_2|\mathbf{y}_2, \mathbf{X}_2) &= \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2), \end{aligned} \tag{6}$$

where

$$\mathbf{m}_i = \mathbf{\Sigma}_i(\mathbf{\Sigma}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\mathbf{X}_i\mathbf{y}_i), \tag{7}$$

and

$$\mathbf{\Sigma}_i^{-1} = \mathbf{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}_i\mathbf{X}_i^\mathsf{T} \tag{8}$$

for $i \in \{1, 2\}$. Imposing the same prior from Equation (3), $P(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_0)$, onto both $\mathbf{w}_1$ and $\mathbf{w}_2$ is reasonable if we believe that the processes on the two sides of the changepoint are genuinely independent. However, if we believe that the second process, for $t > \tau$, is a modified version of the first process, for $t \leq \tau$, then it appears more reasonable to make use of the inference results obtained from the first

process. As a first idea, let us adopt the approach of standard Bayesian sequential analysis (see, e.g., Carlin and Louis (2009)), and take the posterior distribution of the regression parameters from the preceding segment as the prior for the current segment and replace Equation (3) by the left-hand side of Equation (6):

$$P(\mathbf{w}_2) \;=\; \mathcal{N}(\mathbf{m}_0, \mathbf{\Sigma}_0) = \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1). \tag{9}$$

Using Bayes' rule (see Equation (4) and standard transformations of Gaussians (Bishop (2006)) we get for the posterior distribution of $\mathbf{w}_2$:

$$P(\mathbf{w}_2|\mathbf{y}_2, \mathbf{X}_2) = \mathcal{N}(\mathbf{m}_\star, \mathbf{\Sigma}_\star), \tag{10}$$

where

$$\mathbf{m}_\star = \mathbf{\Sigma}_\star(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 + \sigma^{-2}\mathbf{X}_2\mathbf{y}_2),$$

and

$$\mathbf{\Sigma}_\star^{-1} = \mathbf{\Sigma}_1^{-1} + \sigma^{-2}\mathbf{X}_2\mathbf{X}_2^\mathsf{T}.$$

However, inserting the expressions for $\mathbf{m}_1$ and $\mathbf{\Sigma}_1$ from Equations (7-8) into these expressions leads to Equation (5), which means: the posterior distribution of the regression parameters in the second segment is identical to the posterior distribution that would have been obtained if no breakpoint had been introduced. To understand this apparent paradox, which is a standard result of Bayesian sequential analysis, note that the prior of Equation (10) couples the regression parameters, $\mathbf{w}_2$, to the mean from the previous segment, $\mathbf{m}_1$, with a coupling strength given by the Fisher information matrix, $\mathbf{\Sigma}_1^{-1}$. Coupling to $\mathbf{m}_1$ is consistent with our prior notion that interactions and regulatory processes are similar between different segments. However, the coupling strength depends on the nature of the processes investigated, whereas the Fisher information matrix depends on the number of observations and, hence, the experimental design. For illustration, consider gene regulatory processes in a cell during meiosis. There are certain similarities and differences between the interaction strengths before and after cell division, but these differences depend on the nature of the biological processes (meiosis), not on the number of measurements taken before cell division (experimental design). To address this inconsistency, we replace Equation (9) by

$$P(\mathbf{w}_2) \;=\; \mathcal{N}(\mathbf{m}_1, \lambda\sigma^2\mathbf{C}) \tag{11}$$

for some matrix $\mathbf{C}$ with multiplicative scalar λ. Note that the explicit dependence of the covariance matrix on the noise variance, σ^2, is a common approach in Bayesian

modelling (see e.g., Sections 3.3-3.4 in Gelman *et al.* (2004)), as it leads to a fully conjugate prior in both the regression parameters and the noise variances that allows both parameter groups to be integrated out analytically in the marginal likelihood. The choice of the matrix, $\mathbf{C}$, may be guided by our prior knowledge about the nature of the studied processes. In the absence of specific prior knowledge, we set $\mathbf{C}$ to the unit matrix and infer λ from the posterior distribution, based on some vague prior.[1]

Inserting Equation (11) into Equation (4) and following the same derivations as those leading to Equation (10), we get

$$P(\mathbf{w}_2|\mathbf{y}_2, \mathbf{X}_2) \;=\; \mathcal{N}(\mathbf{m}_\star, \sigma^2\mathbf{\Sigma}_\star), \tag{12}$$

with

$$\mathbf{m}_\star = \mathbf{\Sigma}_\star([\lambda\mathbf{C}]^{-1}\mathbf{m}_1 + \mathbf{X}_2\mathbf{y}_2),$$

and

$$\mathbf{\Sigma}_\star^{-1} = [\lambda\mathbf{C}]^{-1} + \mathbf{X}_2\mathbf{X}_2^\mathsf{T}.$$

3 Application to dynamic Bayesian networks

3.1 Fixed changepoints

We now generalize this scheme to time-varying dynamic Bayesian networks (TV-DBNs) along the lines proposed in Lèbre *et al.* (2010). Consider a set of N nodes, $g \in \{1, \ldots, N\}$, in a network, $\mathcal{M} = \{\boldsymbol{\pi}_1(\mathcal{M}), \ldots, \boldsymbol{\pi}_N(\mathcal{M})\}$, where $\boldsymbol{\pi}_g(\mathcal{M})$ denotes the parents of node g in $\mathcal{M}$, that is the set of nodes with a directed edge pointing to g. In this article we follow Grzegorczyk and Husmeier (2011) and assume that the regulatory network structure, $\mathcal{M}$, is fixed. While it is straightforward to allow $\mathcal{M}$ to vary with time, as in Lèbre *et al.* (2010), this flexibility would not be appropriate for the type of problem we consider in our applications (Subsection 6.3: cellular response to fast environmental change). Let $y_{g,t}$ denote the realization of the random variable associated with node g at time $t \in \{1, \ldots, T\}$, and let $\mathbf{x}_{\pi_g(\mathcal{M}),t}$ denote the vector of realizations of the random variables associated with the parents of node g, $\boldsymbol{\pi}_g(\mathcal{M})$, at the previous time point, $(t-1)$, and including a constant element equal to 1 (for the intercept). We consider N sets of $(K_g - 1)$ node-specific

[1] Setting $\mathbf{C}$ to the unit matrix penalizes the Euclidean distance between $\mathbf{w}_2$ and $\mathbf{m}_1$. An alternative approach is to penalize the Mahalanobis distance, i.e., $\mathbf{C}^{-1} = \mathbf{X}\mathbf{X}^\mathsf{T}$. This choice of prior is also motivated by maximum entropy arguments (Zellner (1986)). However, as opposed to the unit matrix, $\mathbf{C}^{-1} = \mathbf{X}\mathbf{X}^\mathsf{T}$ is not guaranteed to be non-singular.

changepoints, $\boldsymbol{\tau}_g = \{\tau_{g,h}\}_{1 \leq h \leq (K_g-1)}$, $1 \leq g \leq N$, which define K_g node-specific segments and which, for now, we assume to be fixed. We define

$$\begin{aligned} \mathbf{y}_{g,h} &= (y_{g,(\tau_{g,h}+1)}, \ldots, y_{g,\tau_{g,(h+1)}})^{\mathsf{T}} \\ \mathbf{X}_{\pi_g(\mathcal{M}),h} &= (\mathbf{x}_{\pi_g(\mathcal{M}),(\tau_{g,h}+1)}, \ldots, \mathbf{x}_{\pi_g(\mathcal{M}),\tau_{g,(h+1)}}), \end{aligned}$$

and apply the linear Gaussian regression model defined in Equations (1-2):

$$\begin{aligned} & P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g(\mathcal{M}),h}, \mathbf{w}_{g,h}, \sigma_{g,h}) \\ &= \prod_{t=\tau_{g,h}+1}^{\tau_{g,(h+1)}} P(y_{g,t}|\mathbf{x}_{\pi_g(\mathcal{M}),t}, \mathbf{w}_{g,h}, \sigma_{g,h}) \\ &= \left(\frac{1}{2\pi\sigma_{g,h}^2}\right)^{T_{g,h}/2} \exp\left(-\frac{1}{2}\sum_{t=\tau_{g,h}+1}^{\tau_{g,(h+1)}} \frac{(y_{g,t} - \mathbf{x}_{\pi_g(\mathcal{M}),t}^{\mathsf{T}}\mathbf{w}_{g,h})^2}{2\sigma_{g,h}^2}\right) \\ &= \mathcal{N}(\mathbf{X}_{\pi_g(\mathcal{M}),h}^{\mathsf{T}}\mathbf{w}_{g,h}, \sigma_{g,h}^2\mathbf{I}), \end{aligned} \tag{13}$$

where $T_{g,h} = \tau_{g,(h+1)} - \tau_{g,h}$.

For the prior on $\mathbf{w}_{g,h}$ we adapt Equation (11) in the standard way (see also Andrieu and Doucet (1999)):

$$P(\mathbf{w}_{g,h}|\mathbf{m}_{g,(h-1)}, \sigma_{g,h}, \lambda_g) \;=\; \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{m}_{g,(h-1)}, \lambda_g\sigma_{g,h}^2\mathbf{C}_{g,h}). \tag{14}$$

The motivation for the explicit dependence of the covariance matrix on the noise variance, $\sigma_{g,h}^2$, has been discussed below Equation (11). For the posterior distribution we get, in direct adaptation of Equation (12):

$$P(\mathbf{w}_{g,h}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g(\mathcal{M}),h}, \sigma_{g,h}, \lambda_g) = \mathcal{N}(\mathbf{m}_{g,h}, \sigma_{g,h}^2\boldsymbol{\Sigma}_{g,h}), \tag{15}$$

where

$$\boldsymbol{\Sigma}_{g,h}^{-1} = [\lambda_g\mathbf{C}_{g,h}]^{-1} + \mathbf{X}_{\pi_g(\mathcal{M}),h}\mathbf{X}_{\pi_g(\mathcal{M}),h}^{\mathsf{T}},$$

$\mathbf{m}_{g,0} = \mathbf{0}$, and for $h \geq 1$:

$$\mathbf{m}_{g,h} = \boldsymbol{\Sigma}_{g,h}([\lambda_g\mathbf{C}_{g,h}]^{-1}\mathbf{m}_{g,(h-1)} + \mathbf{X}_{\pi_g(\mathcal{M}),h}\mathbf{y}_{g,h}). \tag{16}$$

We obtain the marginal likelihood by application of standard results for Gaussian integrals; see e.g., Section 2.3.2 and Appendix B in Bishop (2006):

$$\begin{aligned} & P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g(\mathcal{M}),h}, \sigma_{g,h}, \lambda_g) \\ &= \int P(\mathbf{y}_{g,h}|\mathbf{w}_{g,h}, \mathbf{X}_{\pi_g(\mathcal{M}),h}, \sigma_{g,h}, \lambda_g) P(\mathbf{w}_{g,h}|\mathbf{X}_{\pi_g(\mathcal{M}),h}, \sigma_{g,h}, \lambda_g) d\mathbf{w}_{g,h} \\ &= \mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h}, \sigma_{g,h}^2\tilde{\boldsymbol{\Sigma}}_{g,h}), \end{aligned} \tag{17}$$

where

$$\begin{aligned}\tilde{\mathbf{m}}_{g,h} &= \mathbf{X}^{\mathsf{T}}_{\pi_g(\mathcal{M}),h}\mathbf{m}_{g,(h-1)} \\ \tilde{\mathbf{\Sigma}}_{g,h} &= \mathbf{I} + \lambda_g \mathbf{X}^{\mathsf{T}}_{g,h}\mathbf{C}_{g,h}\mathbf{X}_{g,h}.\end{aligned}$$

In the absence of prior knowledge, we set $\mathbf{C}_{g,h} = \mathbf{I}$. So far, we have assumed that $\sigma_{g,h}$ and λ_g are fixed. We now relax this constraint and impose a conjugate gamma prior on $\sigma_{g,h}^{-2}$ and λ_g^{-1}:

$$P(\sigma_{g,h}^{-2}|\alpha_\sigma,\beta_\sigma) = \mathrm{Gam}(\sigma_{g,h}^{-2}|\alpha_\sigma,\beta_\sigma) = \frac{[\beta_\sigma]^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)}[\sigma_{g,h}^{-2}]^{\alpha_\sigma-1}\exp(-\beta_\sigma\sigma_{g,h}^{-2}) \quad (18)$$

$$P(\lambda_g^{-1}|\alpha_\lambda,\beta_\lambda) = \mathrm{Gam}(\lambda_g^{-1}|\alpha_\lambda,\beta_\lambda) = \frac{[\beta_\lambda]^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)}[\lambda_g^{-1}]^{\alpha_\lambda-1}\exp(-\beta_\lambda\lambda_g^{-1}). \quad (19)$$

We set $\alpha_\sigma = \beta_\sigma = \nu/2$ and note that the integral resulting from the marginalization over $\sigma_{g,h}^{-2}$ has a closed-form solution; see, e.g., Section 2.3.7 in Bishop (2006):

$$\begin{aligned}&P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g(\mathcal{M}),h},\lambda_g) \quad (20)\\ &= \int_0^\infty P(\mathbf{y}_{g,h}|\sigma_{g,h},\mathbf{X}_{\pi_g(\mathcal{M}),h},\lambda_g)P(\sigma_{g,h}^{-2}|\nu/2,\nu/2)d\sigma_{g,h}^{-2} \\ &= \int_0^\infty \mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h},\sigma_{g,h}^2\tilde{\mathbf{\Sigma}}_{g,h})\mathrm{Gam}(\sigma_{g,h}^{-2}|\nu/2,\nu/2)d\sigma_{g,h}^{-2} \\ &= \frac{\Gamma(T_{g,h}/2+\nu/2)}{\Gamma(\nu/2)}\frac{1}{(\pi\nu)^{T_{g,h}/2}|\tilde{\mathbf{\Sigma}}_{g,h}|^{1/2}}\left(1+\frac{\Delta_{g,h}^2}{\nu}\right)^{-(T_{g,h}+\nu)/2} \\ &= \frac{\Gamma(T_{g,h}/2+\nu/2)}{\Gamma(\nu/2)}\frac{\nu^{\nu/2}}{(\pi)^{T_{g,h}/2}|\tilde{\mathbf{\Sigma}}_{g,h}|^{1/2}}\left(\nu+\Delta_{g,h}^2\right)^{-(T_{g,h}+\nu)/2},\end{aligned}$$

with the squared Mahalanobis distance

$$\Delta_{g,h}^2 = (\mathbf{y}_{g,h}-\tilde{\mathbf{m}}_{g,h})^{\mathsf{T}}\tilde{\mathbf{\Sigma}}_{g,h}^{-1}(\mathbf{y}_{g,h}-\tilde{\mathbf{m}}_{g,h}). \quad (21)$$

This is a multivariate Student t-distribution (see, e.g., Subsection 2.3.7 in Bishop (2006)). Note that the application of the matrix inversion theorem (see, e.g., Appendix C in Bishop (2006)) gives:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_{g,h}^{-1} &= [\mathbf{I}+\lambda_g\mathbf{X}^{\mathsf{T}}_{\pi_g(\mathcal{M}),h}\mathbf{C}_{g,h}\mathbf{X}_{\pi_g(\mathcal{M}),h}]^{-1} \quad (22)\\ &= \mathbf{I}-\mathbf{X}^{\mathsf{T}}_{\pi_g(\mathcal{M}),h}([\lambda_g\mathbf{C}_{g,h}]^{-1}+\mathbf{X}_{\pi_g(\mathcal{M}),h}\mathbf{X}^{\mathsf{T}}_{\pi_g(\mathcal{M}),h})^{-1}\mathbf{X}_{\pi_g(\mathcal{M}),h}.\end{aligned}$$

The expression of the marginal likelihood in Equation (20) thus subsumes Equations (10-11) in Andrieu and Doucet (1999) as a limiting case for $\mathbf{m}_{g,(h-1)} = \mathbf{0}$. Note that $\mathbf{X}_{\pi_g(\mathcal{M}),h}$ is a matrix of dimension $|\pi_g(\mathcal{M})|$-by-$T_{g,h}$, where $|\pi_g(\mathcal{M})|$ is the cardinality of the parent set of node g (recall that this includes a fixed parent node with constant value 1 for the bias). Thus, it can be seen from Equation (22) that the computation of $\tilde{\mathbf{\Sigma}}_{g,h}^{-1}$ does not require the inversion of a $T_{g,h}$-by-$T_{g,h}$ matrix, but only the inversion of a $|\pi_g(\mathcal{M})|$-by-$|\pi_g(\mathcal{M})|$ matrix. For sufficiently rich data we would expect $T_{g,h} \gg |\pi_g(\mathcal{M})|$, and Equation (22) thus provides a simple way to substantially reduce the computational costs.

For updating the noise variances, $\{\sigma_{g,h}^2\}$, and coupling hyperparameters, $\{\lambda_g\}$, $1 \leq g \leq N$, $1 \leq h \leq K_g$, with a Gibbs sampling scheme (see Section 3.3) note that

$$
\begin{aligned}
&P(\mathbf{y}_{g,\boldsymbol{\tau}_g}, \mathbf{w}_{g,\boldsymbol{\tau}_g}, \boldsymbol{\sigma}^2_{g,\boldsymbol{\tau}_g}, \lambda_g | \mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}) \\
&= P(\lambda_g^{-1}|\alpha_\lambda, \beta_\lambda) \\
&\prod_h P(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma) P(\mathbf{y}_{g,h}|\mathbf{w}_{g,h}, \sigma_{g,h}^2, \mathbf{X}_{\pi_g(\mathcal{M}),h}) P(\mathbf{w}_{g,h}|\sigma_{g,h}^2, \lambda_g, \mathbf{m}_{g,(h-1)}) \\
&= \mathrm{Gam}(\lambda_g^{-1}|\alpha_\lambda, \beta_\lambda) \\
&\prod_h \mathrm{Gam}(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma) \mathcal{N}(\mathbf{y}_{g,h}|\mathbf{m}_{g,h}, \sigma_{g,h}^2 \mathbf{\Sigma}_{g,h}) \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{m}_{g,(h-1)}, \lambda_g \sigma_{g,h}^2 \mathbf{C}_{g,h}) \\
&\propto [\lambda_g^{-1}]^{\alpha_\lambda - 1} e^{-\beta_\lambda \lambda_g^{-1}} \prod_h [\sigma_{g,h}^{-2}]^{\alpha_\sigma - 1} e^{-\beta_\sigma \sigma_{g,h}^{-2}} \\
&\frac{1}{\sqrt{(2\pi\lambda_g\sigma_{g,h}^2)^{|\pi_g(\mathcal{M})|}|\mathbf{C}_{g,h}|}} \cdot e^{-\frac{1}{2\lambda_g\sigma_{g,h}^2}[\mathbf{w}_{g,h}-\mathbf{m}_{g,(h-1)}]^\mathsf{T}\mathbf{C}_{g,h}^{-1}[\mathbf{w}_{g,h}-\mathbf{m}_{g,(h-1)}]} \\
&\frac{1}{\sqrt{(2\pi\sigma_{g,h}^2)^{T_{g,h}}}} \cdot e^{-\frac{1}{2\sigma_{g,h}^2}[\mathbf{y}_{g,h}-\mathbf{X}_{\pi_g(\mathcal{M}),h}^\mathsf{T}\mathbf{w}_{g,h}]^\mathsf{T}[\mathbf{y}_{g,h}-\mathbf{X}_{\pi_g(\mathcal{M}),h}^\mathsf{T}\mathbf{w}_{g,h}]},
\end{aligned}
$$

where $\mathbf{y}_{g,\boldsymbol{\tau}_g} = (\mathbf{y}_{g,1}, \ldots, \mathbf{y}_{g,K_g})$, $\mathbf{w}_{g,\boldsymbol{\tau}_g} = (\mathbf{w}_{g,1}, \ldots, \mathbf{w}_{g,K_g})$, $\boldsymbol{\sigma}^2_{g,\boldsymbol{\tau}_g} = (\sigma^2_{g,1}, \ldots, \sigma^2_{g,K_g})$, and $\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g} = (\mathbf{X}_{\pi_g(\mathcal{M}),1}, \ldots, \mathbf{X}_{\pi_g(\mathcal{M}),K_g})$ are the segmentations implied by the changepoint set, $\boldsymbol{\tau}_g$, $|\pi_g(\mathcal{M})|$ is the cardinality of the parent set $\boldsymbol{\pi}_g(\mathcal{M})$ of node g, while $|.|$ denotes the determinant operator for matrices. On collecting all the terms that depend on λ_g^{-1} and normalization this gives:

$$
\begin{aligned}
&P(\lambda_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g}, \mathbf{w}_{g,\boldsymbol{\tau}_g}, \boldsymbol{\sigma}^2_{g,\boldsymbol{\tau}_g}, \mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}) \\
&= \mathrm{Gam}\left(\lambda_g^{-1} \middle| \alpha_\lambda + \frac{K_g|\pi_g(\mathcal{M})|}{2}, \beta_\lambda + B_{\lambda,g}^{1:K_g}\right),
\end{aligned}
\tag{23}
$$

where K_g is the number of segments for node g, and

$$B_{\lambda,g}^{1:K_g} = \frac{1}{2}\sum_{h=1}^{K_g}\frac{1}{\sigma_{g,h}^2}[\mathbf{w}_{g,h} - \mathbf{m}_{g,(h-1)}]^{\mathsf{T}}\mathbf{C}_{g,h}^{-1}[\mathbf{w}_{g,h} - \mathbf{m}_{g,(h-1)}]$$

For the inverse variances, $\sigma_{g,h}^{-2}$, we could in principle follow the same procedure and then use Gibbs sampling. However, a computationally more efficient way is to proceed as in Andrieu and Doucet (1999) and use Equation (17) instead of Equation (13), i.e., to use a collapsed Gibbs sampler in which the interaction parameters, $\mathbf{w}_{g,h}$, have been integrated out. From Equations (17-18) we obtain:

$$\begin{aligned}
&P(\mathbf{y}_{g,h}, \sigma_{g,h}^{-2}|\mathbf{X}_{\pi_g(\mathcal{M}),h}, \lambda_g) = \mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h}, \sigma_{g,h}^2\tilde{\mathbf{\Sigma}}_{g,h})\mathrm{Gam}(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma)\\
&\propto \frac{1}{\sqrt{\sigma_{g,h}^{2T_{g,h}}|\tilde{\mathbf{\Sigma}}_{g,h}|}}\exp\left(-\frac{\Delta_{g,h}^2}{2\sigma_{g,h}^2}\right)[\sigma_{g,h}^{-2}]^{\alpha_\sigma-1}\exp(-\beta_\sigma\sigma_{g,h}^{-2})\\
&\propto [\sigma_{g,h}^{-2}]^{\frac{T_{g,h}}{2}+\alpha_\sigma-1}\exp\left(-\left[\frac{\Delta_{g,h}^2}{2} + \beta_\sigma\right]\sigma_{g,h}^{-2}\right), \qquad (24)
\end{aligned}$$

where $\Delta_{g,h}^2$ was defined in Equation (21) and depends on λ_g via Equation (22). By normalization we get:

$$P(\sigma_{g,h}^{-2}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g(\mathcal{M}),h}, \lambda_g) \;=\; \mathrm{Gam}\left(\sigma_{g,h}^{-2}\middle|\frac{T_{g,h}}{2}+\alpha_\sigma, \frac{\Delta_{g,h}^2}{2}+\beta_\sigma\right). \qquad (25)$$

The previous discussion and Equations (24-25) follow Andrieu and Doucet (1999) and Lèbre *et al.* (2010) and assume that there is a separate noise variance, $\sigma_{g,h}^2$, associated with each segment, h, for each node, g. Other choices could be considered. For example, in our studies we obtained improved results when using a common variance shared by all segments, $\sigma_{g,h}^2 = \sigma_g^2 \,\forall\, h$, with

$$P(\sigma_g^{-2}|\alpha_\sigma, \beta_\sigma) \;=\; \mathrm{Gam}(\sigma_g^{-2}|\alpha_\sigma, \beta_\sigma) \;=\; \frac{[\beta_\sigma]^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)}[\sigma_g^{-2}]^{\alpha_\sigma-1}\exp(-\beta_\sigma\sigma_g^{-2}). \qquad (26)$$

Equations (24-25) will then change as follows:

$$\begin{aligned}
&P(\mathbf{y}_{g,\boldsymbol{\tau}_g}, \sigma_g^{-2}|\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}, \lambda_g)\\
&= \mathrm{Gam}(\sigma_g^{-2}|\alpha_\sigma, \beta_\sigma)\prod_{h=1}^{K_g}\mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h}, \sigma_g^2\tilde{\mathbf{\Sigma}}_{g,h})\\
&\propto [\sigma_g^{-2}]^{\frac{T_g}{2}+\alpha_\sigma-1}\exp\left(-\sigma_g^{-2}\left[\beta_\sigma + \sum_{h=1}^{K_g}\frac{\Delta_{g,h}^2}{2}\right]\right),
\end{aligned}$$

and

$$P(\sigma_g^{-2}|\mathbf{y}_{g,\boldsymbol{\tau}_g}, \mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}, \lambda_g) = \mathrm{Gam}\left(\sigma_g^{-2}\left|\frac{T_g}{2}+\alpha_\sigma, \beta_\sigma + \sum_{h=1}^{K_g}\frac{\Delta_{g,h}^2}{2}\right.\right), \quad (27)$$

where $T_g = \sum_{h=1}^{K_g} T_{g,h}$, and $\Delta_{g,h}^2$ depends on λ_g via Equations (21-22). A comparison between Equation (25) and Equation (27) leads to the intuitive result that we obtain the posterior distribution of σ_g^{-2} from the one of $\sigma_{g,h}^{-2}$ by summing the sufficient statistics in the gamma distribution over all segments.

Note that using a common variance, σ_g^2, implies changes in Equations (15) and (20). Denote by $\tilde{\boldsymbol{\Sigma}}_{g,\boldsymbol{\tau}_g}$ a matrix with block structure, in which the matrices $\tilde{\boldsymbol{\Sigma}}_{g,h}$ ($h = 1, \ldots, K_g$) are arranged along the diagonal, and all other entries are 0. In modification of Equations (15) and (20) we now get:

$$P(\mathbf{w}_{g,h}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g(\mathcal{M}),h}, \sigma_g, \lambda_g) = \mathcal{N}(\mathbf{m}_{g,h}, \sigma_g^2\boldsymbol{\Sigma}_{g,h}), \quad (28)$$

and

$$\begin{aligned} &P(\mathbf{y}_{g,\boldsymbol{\tau}_g}|\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}, \lambda_g) \quad (29)\\ &= \frac{\Gamma(T_g/2+\nu/2)}{\Gamma(\nu/2)}\frac{\nu^{\nu/2}}{(\pi)^{T_g/2}|\tilde{\boldsymbol{\Sigma}}_{g,\boldsymbol{\tau}_g}|^{1/2}}\left(\nu+\Delta_{g,\boldsymbol{\tau}_g}^2\right)^{-(T_g+\nu)/2}. \end{aligned}$$

The determinant of a product of matrices is the product of the determinants of the individual matrices; hence $|\tilde{\boldsymbol{\Sigma}}_{g,\boldsymbol{\tau}_g}| = \prod_h |\tilde{\boldsymbol{\Sigma}}_{g,h}|$. The inverse of a block matrix with quadratic submatrices along the diagonal is a block matrix of inverted submatrices along the diagonal; see, e.g., Equations (2.76) and (2.77) in Bishop (2006). This leads to a decomposition of $\Delta_{g,\boldsymbol{\tau}_g}^2$ as a sum over contributions from all segments, h:

$$\Delta_{g,\boldsymbol{\tau}_g}^2 \;=\; \sum\nolimits_h (\mathbf{y}_{g,h}-\tilde{\mathbf{m}}_{g,h})^\mathsf{T}\tilde{\boldsymbol{\Sigma}}_{g,h}^{-1}(\mathbf{y}_{g,h}-\tilde{\mathbf{m}}_{g,h}) \;=\; \sum\nolimits_h \Delta_{g,h}^2. \quad (30)$$

In the same vein, we can split the hyperparameter, λ_g, into a set of different hyperparameters associated with different segments. In particular, it is reasonable to use a different value for the first segment, which does not couple the network to any previous segment: $\lambda_{g,h} = \delta_g$ for $h = 1$ and $\lambda_{g,h} = \lambda_g$ for $2 \leq h \leq K_g$. The hyperparameter, δ_g, can be interpreted as *signal-to-noise hyperparameter* (Lèbre *et al.*, 2010), and we refer to the hyperparameter λ_g as *coupling hyperparameter*, since its inverse, λ_g^{-1}, can be interpreted as the *coupling strength* between adjacent time series segments. In modification of Equations (27) and (29) we then obtain:

$$P(\sigma_g^{-2}|\mathbf{y}_{g,\boldsymbol{\tau}_g}, \mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}, \lambda_g, \delta_g) = \mathrm{Gam}\left(\sigma_g^{-2}\left|\frac{T_g}{2}+\alpha_\sigma, \beta_\sigma + \frac{\Delta_{g,\boldsymbol{\tau}_g}^2}{2}\right.\right), \quad (31)$$

and

$$P(\mathbf{y}_{g,\boldsymbol{\tau}_g}|\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g},\lambda_g,\delta_g) \tag{32}$$
$$= \frac{\Gamma(T_g/2+\nu/2)}{\Gamma(\nu/2)}\frac{\nu^{\nu/2}}{(\pi)^{T_g/2}|\tilde{\boldsymbol{\Sigma}}_{g,\boldsymbol{\tau}_g}|^{1/2}}\left(\nu+\Delta^2_{g,\boldsymbol{\tau}_g}\right)^{-(\frac{T_g+\nu}{2})},$$

where $\Delta^2_{g,\boldsymbol{\tau}_g}$ was defined in Equation (30), and we now have to distinguish:

$$\tilde{\boldsymbol{\Sigma}}_{g,h} = \begin{cases} \mathbf{I}+\delta_g\mathbf{X}^{\mathsf{T}}_{g,h}\mathbf{C}_{g,h}\mathbf{X}_{g,h}, & h=1 \\ \mathbf{I}+\lambda_g\mathbf{X}^{\mathsf{T}}_{g,h}\mathbf{C}_{g,h}\mathbf{X}_{g,h}, & h\geq 2 \end{cases}. \tag{33}$$

Moreover, Equation (28) changes as follows:

$$P(\mathbf{w}_{g,h}|\mathbf{y}_{g,h},\mathbf{X}_{\pi_g(\mathcal{M}),h},\sigma_g,\lambda_g,\delta_g) = \mathcal{N}(\mathbf{m}_{g,h},\sigma^2_g\boldsymbol{\Sigma}_{g,h}), \tag{34}$$

where

$$\boldsymbol{\Sigma}^{-1}_{g,h} = \begin{cases} [\delta_g\mathbf{C}_{g,h}]^{-1}+\mathbf{X}_{\pi_g(\mathcal{M}),h}\mathbf{X}^{\mathsf{T}}_{\pi_g(\mathcal{M}),h}, & h=1 \\ [\lambda_g\mathbf{C}_{g,h}]^{-1}+\mathbf{X}_{\pi_g(\mathcal{M}),h}\mathbf{X}^{\mathsf{T}}_{\pi_g(\mathcal{M}),h}, & h\geq 2 \end{cases},$$

$\mathbf{m}_{g,0}=\mathbf{0}$, and for $h\geq 1$:

$$\mathbf{m}_{g,h} = \boldsymbol{\Sigma}_{g,h}([\lambda_g\mathbf{C}_{g,h}]^{-1}\mathbf{m}_{g,(h-1)}+\mathbf{X}_{\pi_g(\mathcal{M}),h}\mathbf{y}_{g,h}).$$

With the prior distributions $P(\lambda_g^{-1}|\alpha_\lambda,\beta_\lambda) = \mathrm{Gam}(\lambda_g^{-1}|\alpha_\lambda,\beta_\lambda)$ and $P(\delta_g^{-1}|\alpha_\delta,\beta_\delta) = \mathrm{Gam}(\delta_g^{-1}|\alpha_\delta,\beta_\delta)$ we can go through the same derivations as those leading to Equation (23), and we get

$$P(\lambda_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g},\mathbf{w}_{g,\boldsymbol{\tau}_g},\sigma^2_g,\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}) \tag{35}$$
$$= \mathrm{Gam}\left(\lambda_g^{-1}\middle|\alpha_\lambda+\frac{(K_g-1)|\pi_g(\mathcal{M})|}{2},\beta_\lambda+B^{2:K_g}_{\lambda,g}\right),$$

with

$$B^{2:K_g}_{\lambda,g} = \frac{1}{2}\sum_{h=2}^{K_g}\frac{1}{\sigma^2_g}[\mathbf{w}_{g,h}-\mathbf{m}_{g,(h-1)}]^{\mathsf{T}}\mathbf{C}^{-1}_{g,h}[\mathbf{w}_{g,h}-\mathbf{m}_{g,(h-1)}],$$

and

$$P(\delta_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g},\mathbf{w}_{g,\boldsymbol{\tau}_g},\sigma^2_g,\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}) \tag{36}$$
$$= \mathrm{Gam}\left(\delta_g^{-1}\middle|\alpha_\delta+\frac{|\pi_g(\mathcal{M})|}{2},\beta_\delta+B^1_{\lambda,g}\right),$$

with

$$B^1_{\lambda,g} = \frac{1}{2\sigma_g^2}\mathbf{w}_{g,1}^{\mathsf{T}}\mathbf{C}_{g,1}^{-1}\mathbf{w}_{g,1}$$

where in the latter equation we have assumed a prior distribution of the form of Equation (14) with mean $\mathbf{m}_{g,0} = \mathbf{0}$, covariance matrix $\mathbf{C}_{g,1}$, and $\sigma^2_{g,1} = \sigma^2_g$ for the parameters, $\mathbf{w}_{g,1}$:

$$P(\mathbf{w}_{g,1}|\mathbf{0}, \sigma_g, \delta_g) \;=\; \mathcal{N}(\mathbf{w}_{g,1}|\mathbf{0}, \delta_g\sigma_g^2\mathbf{C}_{g,1}).$$

3.2 Variable changepoints

So far, we have assumed that the node-specific changepoints, $\boldsymbol{\tau}_g$ are fixed, but it is straightforward to make them variable. To this end, we need to decide on a prior distribution. Two alternative forms have been compared in the literature (Fearnhead (2006)). The first approach, adopted in Lèbre *et al.* (2010), is based on a truncated Poisson prior on the number of changepoints, (K_g-1), and an explicit specification of $P(\boldsymbol{\tau}_g|(K_g-1))$, e.g., the uniform distribution. The second alternative, pursued in Grzegorczyk and Husmeier (2011) and used in the present work, is based on a point process, where the distribution of the distance between two successive points is a negative binomial distribution. The mathematical details of the point process prior have been relegated to Appendix I.

3.3 Inference

Given the data, $\mathcal{D} = \{y_{g,t}\}, 1 \leq g \leq N, 1 \leq t \leq T$, the ultimate objective is to infer the network structure, $\mathcal{M} = \{\boldsymbol{\pi}_1(\mathcal{M}), \ldots, \boldsymbol{\pi}_N(\mathcal{M})\}$, from the marginal posterior distribution, $P(\mathcal{M}|\mathcal{D})$. The other variable quantities are nuisance parameters, which are marginalized over; these are the changepoint sets, $\{\boldsymbol{\tau}_g\}$, the interaction parameters, $\mathbf{w}_{g,h}$, the noise variances, $\sigma^2_{g,h}$, the signal-to-noise hyperparameters, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ and the coupling hyperparameters, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$. Our model also depends on various higher-level hyperhyperparameters that are fixed; these are the hyperparameters of the changepoint prior as well as the hyperhyperparameters of the gamma distributions: $\{\alpha_\sigma, \beta_\sigma, \alpha_\delta, \beta_\delta, \alpha_\lambda, \beta_\lambda\}$. To avoid notational opacity we do not make them explicit in the following equations, but we briefly investigate their influence in the simulation sections. We pursue inference based on the partially collapsed Gibbs sampler used in Andrieu and Doucet (1999) and Lèbre *et al.* (2010). When choosing a common variance, $\sigma^2_{g,h} = \sigma^2_g \,\forall h$, as we did in our

simulations, we have:

$$P(\mathcal{M}|\mathcal{D},\{\boldsymbol{\tau}_g\},\boldsymbol{\lambda},\boldsymbol{\delta}) \propto P(\mathcal{D}|\mathcal{M},\{\boldsymbol{\tau}_g\},\boldsymbol{\lambda},\boldsymbol{\delta})P(\mathcal{M}) = P(\mathcal{M})\prod_g P(\mathbf{y}_{g,\boldsymbol{\tau}_g}|\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g},\lambda_g,\delta_g), \quad (37)$$

and

$$P(\{\boldsymbol{\tau}_g\}|\mathcal{D},\mathcal{M},\boldsymbol{\lambda},\boldsymbol{\delta}) \propto P(\mathcal{D}|\mathcal{M},\{\boldsymbol{\tau}_g\},\boldsymbol{\lambda},\boldsymbol{\delta})P(\{\boldsymbol{\tau}_g\}) = P(\{\boldsymbol{\tau}_g\})\prod_g P(\mathbf{y}_{g,\boldsymbol{\tau}_g}|\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g},\lambda_g,\delta_g), \quad (38)$$

where $P(\mathcal{M})$ is the prior distribution on network structures, e.g., a uniform distribution subject to a fan-in restriction. Note that the expressions for $P(\mathbf{y}_{g,\boldsymbol{\tau}_g}|\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g},\lambda_g,\delta_g)$, which are given by Equation (32), have been obtained by marginalizing over $\mathbf{w}_{g,h}$ and σ_g^2 ("collapsed" Gibbs steps).

We sample from Equation (37) with the improved structure MCMC scheme proposed in Subsection 2.3.1 in Grzegorczyk and Husmeier (2011). That is, for sampling the network topology we use the structure MCMC algorithm from Madigan and York (1995) and improve this standard sampler for Bayesian networks by the novel single edge operation move ("the flip move") proposed in Grzegorczyk and Husmeier (2011).

For applications where the true changepoint sets are unknown we have to sample the system of node-specific changepoint sets, $\{\boldsymbol{\tau}_g\}$, from Equation (38), and we borrow the RJMCMC scheme introduced by Green (1995). The later RJMCMC sampling scheme is based on single changepoint birth, changepoint death, and changepoint re-allocation moves, and was also employed in Lèbre *et al.* (2010) and Grzegorczyk and Husmeier (2011) among others. As there is an independent changepoint set, $\boldsymbol{\tau}_g$, for each individual gene, g, and the graph, $\mathcal{M}$, of a dynamic Bayesian network is also fully specified by N independent node-specific parent node sets, $\boldsymbol{\pi}_g$ ($g = 1,\ldots,N$), we perform for each gene, g ($g = 1,\ldots,N$), one single changepoint move on $\boldsymbol{\tau}_g$ and one single edge move on $\boldsymbol{\pi}_g$ in each single MCMC iteration.

To sample the signal-to-noise hyperparameters, $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_N)$, and the coupling hyperparameters, $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_N)$, from the posterior distribution, we need to resort to uncollapsed Gibbs steps:

$$P(\boldsymbol{\lambda}|\mathcal{D},\{\boldsymbol{\tau}_g\},\mathcal{M}) = \prod_g P(\lambda_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g},\mathbf{w}_{g,\boldsymbol{\tau}_g},\sigma_g^2,\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}) \quad (39)$$

$$P(\boldsymbol{\delta}|\mathcal{D},\{\boldsymbol{\tau}_g\},\mathcal{M}) = \prod_g P(\delta_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g},\mathbf{w}_{g,\boldsymbol{\tau}_g},\sigma_g^2,\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}), \quad (40)$$

where $P(\lambda_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g},\mathbf{w}_{g,\boldsymbol{\tau}_g},\sigma_g^2,\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g})$ is given by Equation (35) and $P(\delta_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g},\mathbf{w}_{g,\boldsymbol{\tau}_g},\sigma_g^2,\mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g})$ is given by Equation (36). "Uncollapsing" requires the noise variances, σ_g^2, and the interaction parameters, $\mathbf{w}_{g,h}$, to be sampled

Symbol	Explanation
g	the g-th network node ($g = 1, \ldots, N$)
K_g	the number of segments for node g
h	the h-th time segment ($h = 1, \ldots, K_g$)
$\mathcal{M}$	the network structure, $\mathcal{M} = \{\pi_1(\mathcal{M}), \ldots, \pi_N(\mathcal{M})\} = \{\pi_1, \ldots, \pi_N\}$
σ_g^2	the noise variance hyperparameter for node g
δ_g	the signal-to-noise hyperparameter for node g
λ_g	the coupling hyperparameter for node g
π_g	the parent node set of node g
$\boldsymbol{\tau}_g$	the changepoint set for node g
$\mathbf{w}_{g,h}$	the interaction parameter vector for the h-th segment of node g
$\mathbf{y}_{g,h}$	the target values of node g in segment h
$\mathbf{X}_{\pi_g,h}$	the design matrix for segment h of node g
$\mathbf{y}_{g,\boldsymbol{\tau}_g}$	the vectors of target values, $(\mathbf{y}_{g,1}, \ldots, \mathbf{y}_{g,K_g})$, implied by $\boldsymbol{\tau}_g$
$\mathbf{w}_{g,\boldsymbol{\tau}_g}$	the interaction parameter vectors, $(\mathbf{w}_{g,1}, \ldots, \mathbf{w}_{g,K_g})$, implied by $\boldsymbol{\tau}_g$
$\mathbf{X}_{\pi_g,\boldsymbol{\tau}_g}$	the design matrices, $(\mathbf{X}_{\pi_g,1}, \ldots, \mathbf{X}_{\pi_g,K_g})$, implied by $\boldsymbol{\tau}_g$

Table 1: **Overview of parameters and hyperparameters.**

from the corresponding posterior distributions; these are given by Equations (31) and (34).

3.4 Pseudo code for the MCMC sampling scheme

In this subsection we provide pseudo-code for the MCMC sampling scheme, described in Subsection 3.3. In a first step, the Markov chain must be initialized with a network structure, $\mathcal{M}^{(0)} = (\pi_1^{(0)}(\mathcal{M}), \ldots, \pi_N^{(0)}(\mathcal{M}))$, a system of node-specific changepoint sets, $\boldsymbol{\tau}^{(0)} = \{\boldsymbol{\tau}_1^{(0)}, \ldots, \boldsymbol{\tau}_N^{(0)}\}$, the signal-to-noise hyperparameters, $\delta_1^{(0)}, \ldots, \delta_N^{(0)}$, and the coupling hyperparameters, $\lambda_1^{(0)}, \ldots, \lambda_N^{(0)}$. For notational convenience, we do not explicitly indicate in the pseudo code that the parent node sets, π_g ($g = 1, \ldots, N$), depend on the overall network structure, $\mathcal{M}$. Table 1 contains a brief summary of all the relevant parameters and hyperparameters. After the initialization, the MCMC algorithm can be started. In each MCMC iteration a new parent

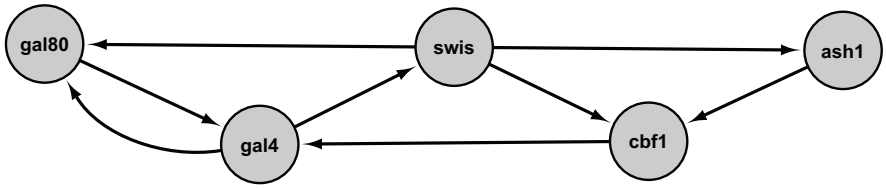

Figure 1: **The topology of the *Saccharomyces cerevisiae* network.** The figure shows the *Saccharomyces cerevisiae* network, as designed in Cantone *et al.* (2009). It consists of 5 nodes (proteins), namely gal4, gal80, cbf1, swis, and ash1, and there are 8 directed edges, which represent gene interactions.

node set, π_g, a new changepoint set, $\boldsymbol{\tau}_g$, and novel hyperparameters, λ_g and δ_g, are sampled for each gene, g, independently. Pseudo code for the MCMC algorithm provided in Table 2.

4 Data

4.1 Simulated network data

We use two different mechanisms to generate non-homogeneous dynamic network data. For a given network topology, e.g., the synthetically generated *Saccharomyces cerevisiae* network (Cantone *et al.* (2009)) shown in Figure 1 or the RAF signal transduction pathway (Sachs *et al.* (2005)) shown in Figure 2, we generate data from a multiple changepoint linear regression model with time-dependent regression parameters. For the simulation study reported in Subsection 6.1, we generate time series with $T = 41$ observations with three equidistant changepoints at $t = 11$, $t = 21$, and $t = 31$, where the regulatory relationships (i.e., the regression parameters) change. In the first scenario, we rotate the interaction parameter vectors from the previous time series segment at each of the three changepoints. The motivation for this rotation is to emulate systematic changes of the regulatory processes, e.g., in plant metabolism during the diurnal cycle of day and night or in organisms during morphogenesis. Since the systematic rotation of the interaction parameter vector is quite a substantial perturbation, we also consider a second scenario, in which we use random perturbations of increasing amplitude ϵ to vary the interaction parameter vectors. In order to keep the signal-to-noise ratio (SNR) constant, the focus in our simulation study is on additive white noise. In a gene-wise manner we add Gaussian distributed noise variables to the gene-specific observations

MCMC algorithm: In each MCMC iteration i:
For each gene $g = 1, \ldots, N$:

- **Network update:** Randomly choose a single-edge operation move (edge deletion, edge addition, or parent flip move) and perform a Metropolis-Hastings MCMC step to propose/sample a new parent set, π_g (see Equations (37) and (32) with $\boldsymbol{\tau}_g = \boldsymbol{\tau}_g^{(i-1)}$, $\lambda_g = \lambda_g^{(i-1)}$, and $\delta_g = \delta_g^{(i-1)}$). For details on the design of these moves see Grzegorczyk and Husmeier (2011). If the move is accepted, set $\pi_g^{(i)} = \pi_g$; otherwise leave the parent set unchanged, $\pi_g^{(i)} = \pi_g^{(i-1)}$.

- **Segmentation update:** Skip this step if changepoints are chosen fixed. Otherwise, randomly select a changepoint birth, death or re-allocation move, and perform a Metropolis-Hastings MCMC step to propose/sample a new changepoint set, $\boldsymbol{\tau}_g$ (see Equations (38) and (32) with $\pi_g = \pi_g^{(i)}$, $\lambda_g = \lambda_g^{(i-1)}$ and $\delta_g = \delta_g^{(i-1)}$). For details on the design of the moves see Grzegorczyk and Husmeier (2011). If the move is accepted, set $\boldsymbol{\tau}_g^{(i)} = \boldsymbol{\tau}_g$; otherwise set $\boldsymbol{\tau}_g^{(i)} = \boldsymbol{\tau}_g^{(i-1)}$.

- **Hyperparameter update:** Re-sample the hyperparameters with collapsed Gibbs sampling steps:

  - Conditional on $\lambda_g^{(i-1)}$, $\delta_g^{(i-1)}$, $\boldsymbol{\tau}_g^{(i)}$, $\pi_g^{(i)}$, sample the variance hyperparameter, $\sigma_g^{(i)}$, from $P(\sigma_g^{-2}|\mathbf{y}_{g,\boldsymbol{\tau}_g^{(i)}}, \mathbf{X}_{\pi_g^{(i)},\boldsymbol{\tau}_g^{(i)}}, \lambda_g^{(i-1)}, \delta_g^{(i-1)})$ (see Equation (31)), and then sample interaction hyperparameter vectors, $\mathbf{w}_{g,\boldsymbol{\tau}_g^{(i)}}^{(i)} = (\mathbf{w}_{g,1}^{(i)}, \ldots, \mathbf{w}_{g,K_g}^{(i)})$, from $P(\mathbf{w}_{g,h}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g^{(i)},h}, \sigma_g^{(i)}, \lambda_g^{(i-1)}, \delta_g^{(i-1)})$ (see Equation (34)).
  - Sample $\lambda_g^{(i)}$ from $P(\lambda_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g^{(i)}}, \mathbf{w}_{g,\boldsymbol{\tau}_g^{(i)}}^{(i)}, \sigma_g^{(i)}, \mathbf{X}_{\pi_g^{(i)},\boldsymbol{\tau}_g^{(i)}})$ (see Equation (35)).
  - Sample $\delta_g^{(i)}$ from $P(\delta_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g^{(i)}}, \mathbf{w}_{g,\boldsymbol{\tau}_g^{(i)}}^{(i)}, \sigma_g^{(i)}, \mathbf{X}_{\pi_g^{(i)},\boldsymbol{\tau}_g^{(i)}})$ (see Equation (36)).

- **Set** $\mathcal{M}^{(i)} = (\pi_1^{(i)}, \ldots, \pi_N^{(i)})$ and $\boldsymbol{\tau}^{(i)} = \{\boldsymbol{\tau}_1^{(i)}, \ldots, \boldsymbol{\tau}_N^{(i)}\}$.

Table 2: **Pseudo code for the MCMC sampling scheme.**

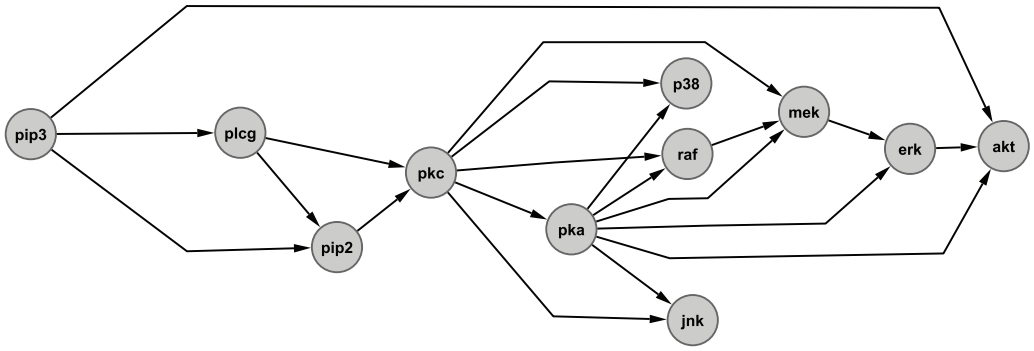

Figure 2: **The topology of the RAF pathway.** The figure shows the RAF network, as reported in Sachs *et al.* (2005). It consists of 11 nodes (proteins), namely pip3, plcg, pip2, pkc, p38, raf, pka, jnk, mek, erk, and akt, and there are 20 directed edges, which represent protein interactions.

to reach a pre-specified signal-to-noise ratio (SNR). We distinguish three signal-to-noise ratios SNR$= 10$ (weak noise), SNR$= 3$ (moderate noise), and SNR$= 1$ (strong noise). The mathematical details of the two data generating mechanisms have been relegated to Appendix II.

4.2 Gene expression time series from circadian rhythms in *Arabidopsis thaliana*

Microarray gene expression time series related to the study of circadian regulation in plants were measured in *Arabidopsis thaliana*. *Arabidopsis thaliana* seedlings, grown under artificially controlled T_e-hour-light/T_e-hour-dark cycles, were transferred to constant light and harvested at 12-13 time points in τ-hour intervals. From these seedlings, RNA was extracted and assayed on Affymetrix GeneChip oligonucleotide arrays. The data were background-corrected and normalized according to standard procedures[2], using GeneSpring© software (Agilent Technologies). Four individual time series, which differed with respect to the pre-experiment entrainment condition and the harvesting intervals: $T_e \in \{10, 12, 14\}$ and $\tau \in \{2, 4\}$, were measured. For an overview see Table 3. The data, with detailed information about the experimental protocols, can be obtained from Edwards *et al.* (2006), Grzegorczyk *et al.* (2008), and Mockler *et al.* (2007).

[2] We used RMA rather than GCRMA for reasons discussed in Lim *et al.* (2007).

	Segment 1	Segment 2	Segment 3	Segment 4
Source	Mockler et al.(2007)	Edwards et al. (2006)	Grzegorcyk et al. (2008)	Grzegorcyk et al. (2008)
Time points	12	13	13	13
Time interval	4h	4h	2h	2h
Pre-experimental entrainment	12h:12h light:dark cycle	12h:12h light:dark cycle	10h:10h light:dark cycle	14h:14h light:dark cycle
Measurements	Constant light	Constant light	Constant light	Constant light
Laboratory	Kay Lab	Millar Lab	Millar Lab	Millar Lab

Table 3: **Gene expression time series segments for *Arabidopsis thaliana*.** The table contains an overview of the experimental conditions under which each of the gene expression experiments was carried out.

We focus our analysis on 9 genes that are involved in circadian regulation[3], and we arrange the four individual time series successively, as indicated in Table 3, so as to obtain one single time series consisting of four segments, where each of the segments has been measured under different experimental conditions.[4] The motivation for this concatenation is given by the following consideration. Since the processes of circadian regulation that the 9 genes are involved in are the same, it makes sense to aim to infer the underlying gene regulatory network structure from a combination of all four time series. On the other hand, the detailed nature and strength of the gene interactions may well be influenced by the changes in the experimental and pre-experimental entrainment conditions (see Table 3), rendering these changes natural candidates for changepoints in our TV-DBN model.

4.3 Gene expression network data from morphogenesis in *Drosophila melanogaster* (fruit fly)

The gene expressions in *Drosophila melanogaster* cells were sampled at $T = 67$ time steps during four different morphogenetical stages of life: the embryonic, larval, pupa, and adult stage (Arbeitman *et al.* (2002)). Since these phases cover

[3]These 9 circadian genes are LHY, TOC1, CCA1, ELF4, ELF3, GI, PRR9, PRR5, and PRR3.

[4]In the merged time series the gene expression values at the first time point of a time series segment have no relation with the expression values at the last time point of the preceding segment, so that the corresponding boundary time points have to be appropriately removed from the data. This ensures that for all pairs of consecutive time points a proper conditional dependence relation determined by the nature of the regulatory cellular processes is given; see Subsection 4.4 for a mathematical treatment.

time periods of different lengths, gene expression profiles were collected at non-equidistant time-points. The three main morphogenetical transitions occur at time points $t = 31$ (embryonic to larval), $t = 41$ (larval to pupal), and $t = 59$ (pupal to adult) (Arbeitman *et al.* (2002)). Like other researchers (Zhao *et al.* (2006); Robinson and Hartemink (2009)) we focus our analysis on $N = 11$ genes involved in growth and muscle development: EVE, GFL, TWI, MLC1, SLS, MHC, PRM, ACTN, UP, MYP61F, and MSP300. The data set is available from Robinson and Hartemink (2009), and we standardized the observations of each gene independently with a z-score transformation.

4.4 Synthetic biology in *Saccharomyces cerevisiae* (yeast)

While *systems biology* aims to develop a formal understanding of biological processes via the development of quantitative mathematical models, *synthetic biology* aims to use such models to design unique biological circuits (synthetic networks) in the cell able to perform specific tasks. Conversely, data from synthetic biology can be utilized to assess the performance of models from systems biology. Cantone *et al.* (2009) synthetically generated a network of five genes in *Saccharomyces cerevisiae* (yeast), depicted in Figure 1. The network was obtained from synthetically designed yeast cells grown with different carbon sources: galactose ("switch on") or glucose ("switch off"). Cantone *et al.* (2009) obtained in vivo data with quantitative real-time RT-PCR in intervals of 20 minutes up to 5 hours for the first, and in intervals of 10 minutes up to 3 hours for the second condition. For our study we arranged the two time series successively to one single time series. Since the first time point of the "switch off" time series has no relation with the expression values at the last time point of the preceding "switch on" time series, the first time point of the second series was appropriately removed to ensure that for all pairs of consecutive time points a proper conditional dependence relation is given.[5] Finally, the merged time series was standardized via a log transformation and a subsequent mean standardization.

[5]When merging two time series $(\mathbf{D}^{(1)}_{.,1}, \ldots, \mathbf{D}^{(1)}_{.,T_1})$ and $(\mathbf{D}^{(2)}_{.,1}, \ldots, \mathbf{D}^{(2)}_{.,T_2})$, where $\mathbf{D}^{(i)}_{.,t}$ is the t-th observation (vector) of the i-th time series ($i = 1, 2$), to one single time series $(\mathbf{D}^{(1)}_{.,1}, \ldots, \mathbf{D}^{(1)}_{.,T_1}, \mathbf{D}^{(2)}_{.,1}, \ldots, \mathbf{D}^{(2)}_{.,T_2})$ of length $T_1 + T_2$ only the pairs $\mathbf{D}^{(1)}_{.,t-1} \to \mathbf{D}^{(1)}_{.,t}$ $(t = 2, \ldots, T_1)$ and $\mathbf{D}^{(2)}_{.,t-1} \to \mathbf{D}^{(2)}_{.,t}$ $(t = 2, \ldots, T_2)$ can be used for inference. The boundary pair $\mathbf{D}^{(1)}_{.,T_1} \to \mathbf{D}^{(2)}_{.,1}$ has to be excluded.

5 Details on data analysis

5.1 Hyperparameter settings

We want to compare the proposed coupled (regularized) time-varying dynamic Bayesian network (TV-DBN) model, which was presented in Section 3, with the conventional uncoupled (unregularized) TV-DBN model akin to Lèbre *et al.* (2010). For both TV-DBN models under comparison the gene- and segment-specific interaction parameter vectors, $\mathbf{w}_{g,h}$, are assumed to be multivariate Gaussian distributed. For the uncoupled TV-DBN model (akin to Lèbre *et al.* (2010)) we have:

$$P(\mathbf{w}_{g,h}|\sigma_{g,h},\delta_g) = \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{0},\delta_g\sigma_{g,h}^2\mathbf{C}_{g,h}), \tag{41}$$

where δ_g is the signal-to-noise hyperparameter (Lèbre *et al.* (2010)). For the coupled TV-DBN model, proposed in Section 3, we have:

$$P(\mathbf{w}_{g,h}|\mathbf{m}_{g,h-1},\sigma_{g,h},\delta_g,\lambda_g) = \begin{cases} \mathcal{N}(\mathbf{w}_{g,1}|\mathbf{m}_{g,0}=\mathbf{0},\delta_g\sigma_{g,h}^2\mathbf{C}_{g,h}), & h=1 \\ \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{m}_{g,h-1},\lambda_g\sigma_{g,h}^2\mathbf{C}_{g,h}), & h\geq 2 \end{cases}, \tag{42}$$

where δ_g is the signal-to-noise hyperparameter, λ_g is the coupling hyperparameter, and $\mathbf{m}_{g,h-1}$ depends on the preceding segment via Equation (16) for $h \geq 2$.

For both models we assume that $\mathbf{C}_{g,h}$ is the unit matrix, $\mathbf{I}$, and that the variances are gene-specific but shared by the gene-specific segments, symbolically $\sigma_{g,h}^2 = \sigma_g^2$ for all h ($g = 1,\ldots,N$), as we observed slightly improved results for this setting in a preliminary study with simulated data. We then obtain

$$P(\mathbf{w}_{g,h}|\sigma_g,\delta_g) = \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{0},\delta_g\sigma_g^2\mathbf{I}) \tag{43}$$

for the uncoupled TV-DBN model, and:

$$P(\mathbf{w}_{g,h}|\mathbf{m}_{g,h-1},\sigma_{g,h},\delta_g,\lambda_g) = \begin{cases} \mathcal{N}(\mathbf{w}_{g,1}|\mathbf{0},\delta_g\sigma_g^2\mathbf{I}), & h=1 \\ \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{m}_{g,h-1},\lambda_g\sigma_g^2\mathbf{I}), & h\geq 2 \end{cases} \tag{44}$$

for the proposed coupled TV-DBN model.

For both models under comparison, the prior distribution of the inverse variance hyperparameters, σ_g^{-2}, is assumed to be a Gamma distribution with hyperhyperparameters α_σ and β_σ (see Equation (26)). For both TV-DBN models we set $\alpha_\sigma = \beta_\sigma = \nu/2$ with $\nu = 0.01$ ($g = 1,\ldots,N$). This yields:

$$P(\sigma_g^{-2}) = Gam(\sigma_g^{-2}|\alpha_\sigma = 0.005, \beta_\sigma = 0.005). \tag{45}$$

For the conventional unregularized TV-DBN model (akin to Lèbre *et al.* (2010)) we follow Lèbre *et al.* (2010) and assume the signal-to-noise hyperparameters, δ_g, to

be inverse Gamma distributed with hyperhyperparameters, $\alpha_\delta = 2$ and $\beta_\delta = 0.02$. We obtain:

$$P(\delta_g^{-1}|\alpha_\delta, \beta_\delta) = Gam(\delta_g^{-1}|\alpha_\delta = 2, \beta_\delta = 0.02). \tag{46}$$

With $\mathbf{m}_{g,h-1} = \mathbf{0}$ for all h, $\mathbf{C}_{g,h} = \mathbf{I}$, and $\sigma_{g,h}^2 = \sigma_g^2$ it follows from Equation (23):

$$\begin{aligned} &P(\delta_g^{-1}|\mathbf{y}_{g,\boldsymbol{\tau}_g}, \mathbf{w}_{g,\boldsymbol{\tau}_g}, \sigma_g^2, \mathbf{X}_{\pi_g(\mathcal{M}),\boldsymbol{\tau}_g}) \\ &= \mathrm{Gam}\left(\delta_g^{-1} \middle| \alpha_\delta + \frac{K_g|\pi_g(\mathcal{M})|}{2}, \beta_\delta + \frac{1}{2}\sum_{h=1}^{K_g} \frac{1}{\sigma_g^2} \mathbf{w}_{g,h}^\mathsf{T} \mathbf{w}_{g,h}\right) \end{aligned} \tag{47}$$

For the proposed regularized TV-DBN model the mean vectors of the first segment are zero vectors and the hyperparameters δ_g ($g = 1, \ldots, N$) correspond to the signal-to-noise hyperparameters. We assume exactly the same inverse Gamma prior distribution as for the uncoupled TV-DBN model:

$$P(\delta_g^{-1}|\alpha_\delta, \beta_\delta) = Gam(\delta_g^{-1}|\alpha_\delta = 2, \beta_\delta = 0.02). \tag{48}$$

The mean vectors of the subsequent time series segments, $\mathbf{m}_{g,h-1}$ ($h \geq 2$), depend on the results from the preceding time series segment via Equations (14) and (16). For each gene, g, the inverse, λ_g^{-1}, of the coupling hyperparameter, λ_g, is the coupling strength, and we employ independent $Gam(\lambda_g^{-1}|\alpha_\lambda, \beta_\lambda)$ prior distributions for λ_g^{-1}. In the absence of any genuine prior knowledge we set $\alpha_\lambda = \beta_\lambda = \gamma$ and vary the hyperhyperparameter, $\gamma \in \{0.3, 1, 3, 10\}$, to obtain an impression of the robustness of the proposed Bayesian regularization scheme.

For the point process prior on the changepoint sets we employ various hyperparameter settings, as explained in more detail in Appendix I.

5.2 Convergence diagnostics and criterions for the network reconstruction accuracy

We pursue inference based on the partially collapsed Gibbs sampler, described in Subsection 3.3, and the output is a sample of network structures from the posterior distribution. We apply standard convergence diagnostics, based on trace plot diagnostics (Giudici and Castelo (2003)) and the potential scale reduction factor (Gelman and Rubin (1992)), as explained in Appendix I. We found that the PSRF's of all individual edges were below 1.2 for simulation lengths of 50,000 MCMC steps (if the changepoints are known) and 100,000 MCMC steps (if the changepoints are unknown and have to be inferred from the data); see Appendix I for details.

From the network structures, which are sampled from Equation 37, we can obtain a ranking of the gene interactions based on their marginal posterior probabilities. For the simulated data from Subsection 4.1 and the in vivo *Saccharomyces cerevisiae* data from Subsection 4.4 the true interaction networks are known so that the ranking of the gene interactions can be employed to obtain the Receiver Operating Characteristic (ROC) and precision-recall (PR) curves (Davis and Goadrich (2006); Prill *et al.* (2010)). These curves can be numerically integrated to get the areas under the curves (AUC) for both (AUROC or AUPRC, respectively) as a global measure of network reconstruction accuracy (with larger values indicating a better performance). The technical details on these two criterions have been relegated to Appendix I.

5.3 The objectives of our data analysis

In Subsection 6.1 our objective is to monitor the network reconstruction accuracy on a series of increasingly strong violations of the prior assumption inherent in Equations (14) and (16). To this end, we generate synthetic network data, as explained in Subsection 4.1, and we reverse-engineer the underlying network by sampling network structures from the posterior distribution (see Equation (37)). In this study we assume the three changepoints to be known, while the other nuisance hyperparameters[6], are sampled with MCMC (see Section 3), and marginalized out. In the absence of proper gold standard networks for the *Drosophila melanogaster* and the *Arabidopsis thaliana* gene expression time series, our focus in Subsection 6.2 is on quantifying the strength of the information coupling and its effect on the network inference. First, we compute the correlations between the segment-specific interaction parameter vectors (sampled from Equation (28)) for the uncoupled and for the coupled TV-DBN. For comparing the correlations of the two TV-DBN models we require an invariant segmentation for both biological systems. For *Arabidopsis thaliana* there are four individual time series, which have been measured under different external conditions, and we can arrange them successively to obtain the segmentation indicated in Table 3. For the *Drosophila melanogaster* data we take the well-established morphogenetical phases from the literature, as shown in Figure 7. The other nuisance hyperparameters can be sampled with MCMC and marginalized out again. In a subsequent study, we investigate the effect of the coupling hyperparameter, λ_g, on the network inference. For fixed coupling hyperparameters, λ_g, in Equation (14) we sample network structures with MCMC, while the other nuisance hyperparameters are sampled with MCMC and marginalized out again. For the

[6]These nuisance hyperparameters are the variances, σ_g^2, the coupling hyperparameters, λ_g, and the signal-to-noise hyperparameters, δ_g ($g = 1, \ldots, N$).

Saccharomyces cerevisiae time series from Subsection 4.4 we have a gold standard network (see Figure 1) and our focus is on monitoring the network reconstruction accuracy in dependence on the inferred segmentations. That is, in Subsection 6.3 we follow an unsupervised approach and assume that the changepoint between the two time series is unknown. To infer different segmentations of the time series we employ different hyperparameters of the point process prior on the changepoint sets; see Subsection 3.2 and Appendix I. In this final study the full potential of the novel coupled TV-DBN model is exploited, as the node-specific changepoint sets are also sampled with MCMC.

6 Results

6.1 Results on simulated data

We evaluated the proposed Bayesian regularization scheme on simulated data, and the first objective was to investigate how the network reconstruction accuracy depends on the SNR and the hyperhyperparameters α_λ and β_λ of Equation (19).

Figure 3 shows that for homogeneous data, in which the true interaction parameters do not change with time, the proposed Bayesian regularization scheme consistently outperforms the unregularized model without information sharing between segments. Our next objective was to monitor the network reconstruction accuracy on a series of increasingly strong violations of the prior assumption on which Equation (14) is based. We effected these violations by systematically rotating the interaction parameter vectors associated with the nodes by a prespecified angle every time we move from one time series segment into the next; see Appendix II for the mathematical details. The motivation for this rotation is to emulate systematic changes of the regulatory processes, e.g., in plant metabolism during the diurnal cycle of day and night or in organisms during morphogenesis. The results are shown in Figure 4 and exhibit a trend that one would intuitively expect: as the angle gets larger, the improvement achieved with our model increasingly diminishes, until for an angle of 90° there is no significant difference (for the higher SNRs) or even a slight deterioration (SNR=1). The study allows us to quantify how robust the proposed method is. For low and medium noise (SNR=10,3), it achieves a significant improvement up to an angle of 22.5°. The improvement turns out to be most pronounced for the medium noise level (SNR=3). When the noise level is low (SNR=10), then the network reconstruction accuracy obtained with the unregularized method is already quite high (AUROC > 0.9), which intrinsically leaves less room for improvement. When the noise level is high (SNR=1), then the network reconstruction accuracy (AUROC ≈ 0.6) is close to random expectation (AUROC $=$

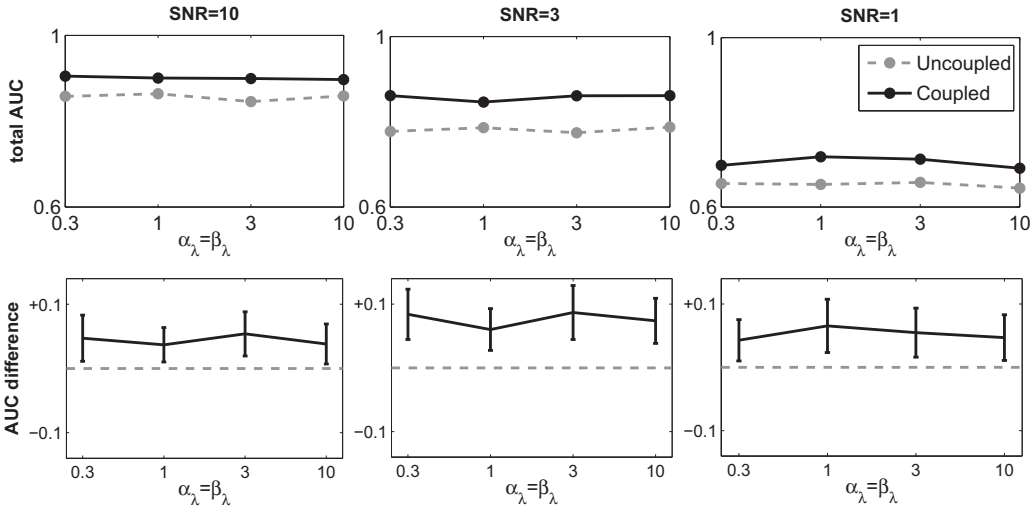

Figure 3: **Network reconstruction accuracy for homogeneous data, simulated from the *Saccharomyces cerevisiae* network shown in Figure 1.** The figure shows the dependence of the network reconstruction accuracy in terms of AUROC scores, as defined in Subsection 5.2, on the signal-to-noise ratio (SNR) and the hyperhyperparameters α_λ and β_λ from Equation (19). The latter vary along the horizontal axis, while the columns represent three different SNR values of the additive white noise (from left to right: 10, 3, and 1). The top row shows the absolute values of the AUROC scores (dashed line: conventional uncoupled TV-DBN model; solid line: proposed coupled TV-DBN model), while the bottom row shows the differences between the proposed regularized TV-DBN and the conventional unregularized TV-DBN. The columns correspond to different signal-to-noise ratios (SNR) of the additive white noise, with the values (from left to right): SNR=10,3,1. All simulations were repeated on 25 independent data instantiations, with error bars indicating confidence intervals (two-sided 95%-quantiles). A similar figure with AUPRC scores can be found as Figure 14 in Appendix IV.

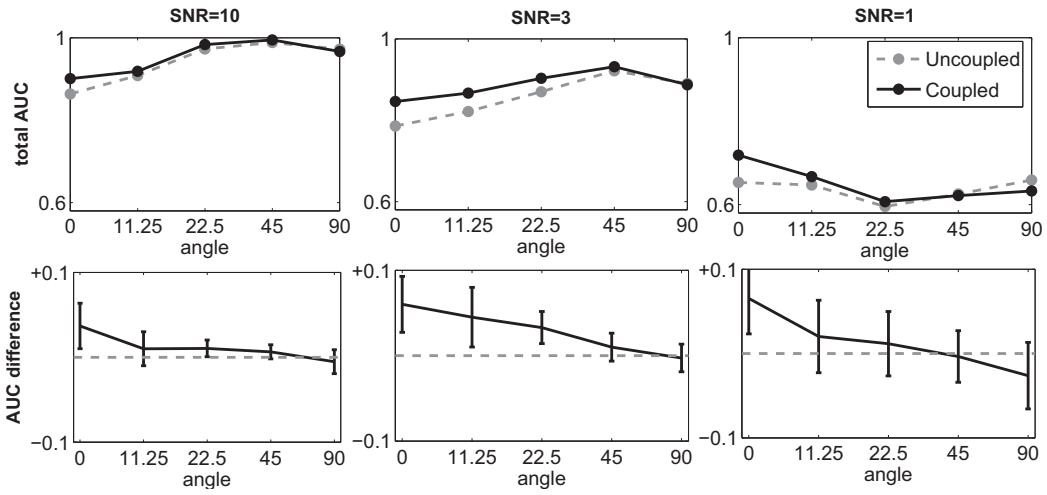

Figure 4: **Network reconstruction accuracy for non-homogeneous data, simulated from the *Saccharomyces cerevisiae* network shown in Figure 1.** The figure shows the network reconstruction accuracy in terms of AUROC scores, as defined in Subsection 5.2, and demonstrates how the proposed Bayesian regularization scheme is affected by increasing violations of the prior assumption inherent in Equations (14) and (16). The abscissa represents the angle by which the parameter vectors are rotated as one moves from a time series segment into the adjacent one; see Subsection 4.1 and Appendix II for details. The hyperhyperparameters were fixed with values of $\alpha_\lambda = \beta_\lambda = 1$, and the columns represent three different SNR levels (from left to right: 10, 3, and 1). The top row shows the absolute values of the AUROC scores (dashed line: conventional uncoupled TV-DBN model; solid line: proposed regularized TV-DBN model), while the bottom row shows the differences between the proposed regularized TV-DBN and the conventional unregularized TV-DBN. All simulations were repeated on 25 independent data instantiations, with error bars indicating confidence intervals (two-sided 95%-quantiles). A similar figure with AUPRC scores can be found as Figure 15 in Appendix IV.

0.5), and the fluctuations over different data instantiations (error bars) are large. We note that except for particularly noisy data (SNR$= 1$) small and moderate rotations (up to $45°$) increase the average AUROC scores for both models under comparison. This trend is plausible, in that perturbing interaction parameters by rotation diminishes the chances of false negatives (i.e., not recovering true interactions) related to fixed small interaction parameters.

We also investigate to which extent the posterior distribution of the coupling hyperparameters, λ_g, depends on (i) the rotation angles, i.e., the strength of the violation of the prior assumption inherent in Equations (14) and (16), and on (ii) the choice of the hyperhyperparameters α_λ and β_λ. Figure 5 shows histograms of the coupling hyperparameters sampled with MCMC for each combination of rotation angle ($\{0°$,$11.25°$,$22.5°$,$45°$,$90°\}$) and hyperhyperparameter setting ($\alpha_\lambda = \beta_\lambda = 0.3, 1, 3, 10$). The histograms in Figure 5 reveal that the posterior average depends on the rotation angle. That is, the greater the angle, i.e. the stronger the deviations between the segment-specific interaction parameter vectors, the higher the posterior mean of the coupling hyperparameters, i.e. the weaker the coupling strength among segments. This trend suggests that the proposed coupled TV-DBN model infers plausible coupling hyperparameters: When the true interaction parameters are similar between segments, low hyperparameter values, $log(\lambda_g) < 0$, are inferred, corresponding to strong coupling between the segments. Conversely, when the true interaction parameters differ between segments, high hyperparameter values, $log(\lambda_g) > 0$, are inferred, corresponding to weak coupling between the segments. Figure 5 also demonstrates that the hyperhyperparameters $\alpha_\lambda = \beta_\lambda$ primarily have an effect on the variances of the posterior distributions with high hyperhyperparameter values yielding more strongly peaked distributions.

To ascertain that the same trends can also be obtained for other network structures, we also compare both methods on synthetic data generated from the RAF network topology shown in Figure 2. For this study we re-use the systematic rotation of the interaction parameter vector approach (see Subsection 4.1 and Appendix II) to simulate data from the RAF network. Since our earlier findings suggest that the hyperhyperparameters do not have a strong effect on the inference results (see Figure 3), we restrict our attention on the hyperhyperparameter setting $\alpha_\lambda = \beta_\lambda = 1$ in this study. It can be seen from Figure 6 that the results for the RAF network are very similar to those obtained earlier for the smaller *Saccharomyces cerevisiae* network (see Figure 4). For the RAF network there are even more significant trends in favor of the proposed regularized scheme, as the increased network size intrinsically leads to a decreased variance of the AUROC scores. For all three signal-to-noise ratio (SNR) levels, the proposed method achieves a significant (two-sided t-test p value < 0.05) improvement up to an angle of $22.5°$. The posterior distributions of the coupling hyperparameters show exactly the same trends, as

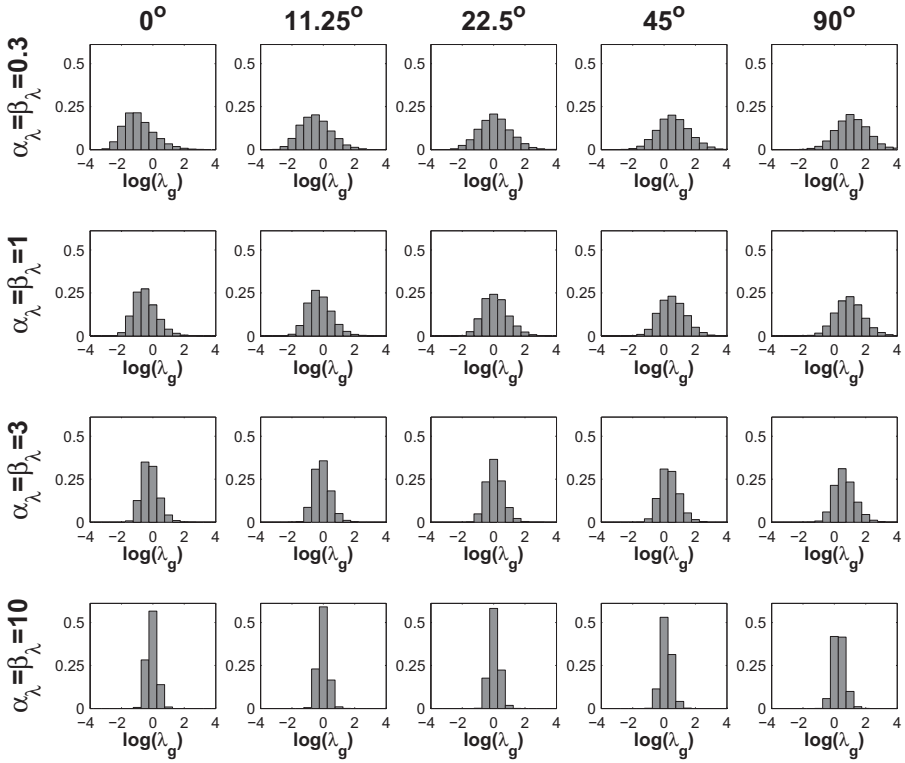

Figure 5: **Posterior distribution of the (logarithmic) coupling hyperparameters, λ_g ($g = 1, \ldots, N$), of the proposed coupled TV-DBN model for synthetic *Saccharomyces cerevisiae* data, which were simulated using the rotation approach and the signal-to-noise ratio SNR$= 3$.** The figure is arranged as a matrix, where the rows correspond to the hyperhyperparameters $\alpha_\lambda = \beta_\lambda$, and the columns correspond to the Euler angle by which the interaction parameter vectors are rotated as one moves from a time series segment into the adjacent one (see Equations (58-59) in Appendix II). Each histogram was obtained by merging the sampled coupling hyperparameters, λ_g, of all genes g ($g = 1, \ldots, N$) from 25 independent data instantiations.

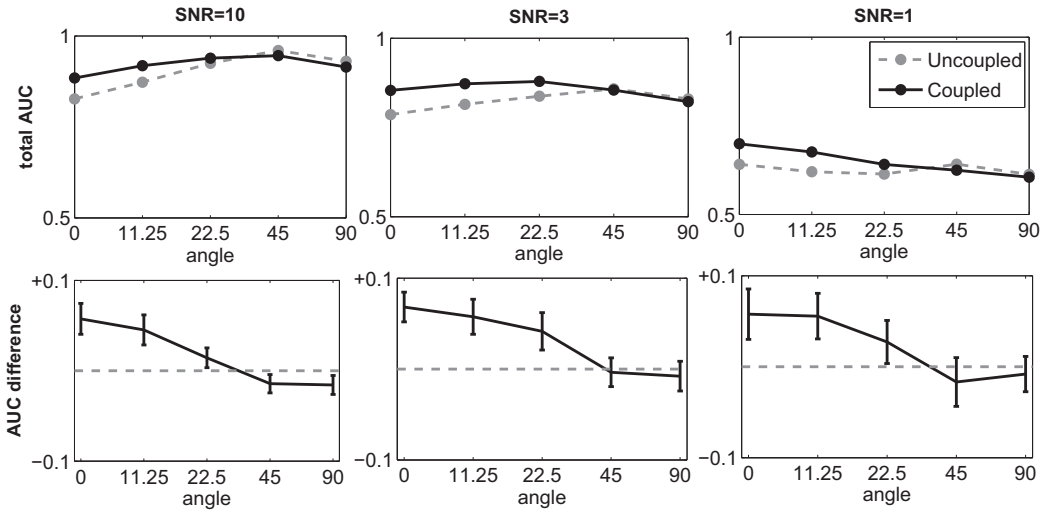

Figure 6: **Network reconstruction accuracy for non-homogeneous data, simulated from the RAF network shown in Figure 2.** The figure shows how the proposed Bayesian regularization scheme is affected by increasing violations of the prior assumption inherent in Equations (14) and (16). In each panel the abscissa represents the angle by which the interaction parameter vectors are rotated as one moves from a time series segment into the adjacent one; see Subsection 4.1 and Appendix II for details. The hyperhyperparameters were set to $\alpha_\lambda = \beta_\lambda = 1$, and each column represents a signal-to-noise ratio (SNR) level of the additive white noise. The top row shows the absolute values of the AUROC scores, as defined in Subsection 5.2, while the bottom row shows the mean differences between the proposed regularization scheme and the unregularized method. All simulations were repeated on 25 independent data instantiations, with error bars in the bottom row, indicating the bounds of 95% t-test confidence intervals for the mean differences.

	Study no. 1	**Study no. 2**	**Study no. 3**
Network structure	Yeast net (see Figure 1)	Yeast net (see Figure 1)	RAF net (see Figure 2)
True regulatory mechanisms	Rotation with increasing angle (see Section 4.1)	Perturbation with increasing amplitude (see Section 4.1)	Rotation with increasing angle (see Section 4.1)
Hyperhyper-parameter variation for homogeneous data	**Yes** Figure 3 (AUCROC) Figure 14 (AUCPR)	**Yes** Figure 12 (AUCROC) Figure 18 (AUCPR)	**No**
Network reconstruction for $\alpha_\lambda = \beta_\lambda = 1$	**Yes** Figure 4 (AUCROC) Figure 15 (AUCPR)	**Yes** Figure 13 (AUCROC) Figure 19(AUCPR)	**Yes** Figure 6 (AUCROC) Figure 16 (AUCPR)

Table 4: **An overview of all our studies on simulated network data.** In all three studies three signal-to-noise ratios – SNR$= 10$, 3, and 1 – were chosen. For testing the robustness with respect to the hyperhyperparameter variation we set the values of $\alpha_\lambda = \beta_\lambda$ to 0.3, 1, 3, and 10. Figures 3, 4, and 6 can be found in Section 6.1. All other figures have been relegated to Appendix III or Appendix IV.

already observed for the *Saccharomyces cerevisiae* network structure. That is, as in Figure 5 the posterior average of λ_g gets higher (i.e. the coupling gets weaker) with increasingly strong violations of the prior assumption inherent in Equations (14) and (16).

We also repeated the simulation study on the *Saccharomyces cerevisiae* network for an alternative data generating mechanisms with random perturbations of increasing amplitude, as explained in detail in Subsection 4.1 and Appendix II. We obtained very similar results, which have been relegated to Appendix III (see Figures 12 and 13). Finally, we note that quantifying the network reconstruction accuracy in terms of AUPRC scores (Davis and Goadrich (2006)) rather than AUROC scores also gives very similar trends, as briefly reported in Appendix IV. An overview of all our studies on simulated network data is given in Table 4. This table also provides references to all figures that contain empirical results for simulated data.

6.2 Information coupling in *Drosophila melanogaster* and *Arabidopsis thaliana*

The first question we wanted to investigate is the effect the proposed Bayesian regularization scheme has on the inference of the interaction parameters. We employ two data sets for this study: The *Arabidopsis thaliana* gene expression time series from Subsection 4.2 and the *Drosophila melanogaster* gene expression time series from Subsection 4.3. For the *Arabidopsis thaliana* data each segment corresponds to a short time series experiment (see Table 3). Although the experimental

conditions for the four segments differ with respect to pre-experimental entrainment conditions and harvesting times (see Table 3), a critical factor related to the diurnal effect of absence versus presence of light (Salome and McClung (2004)) was artificially set to be constant. For the *Arabidopsis thaliana* data, we therefore expect the segment-specific interaction parameters to show a considerable degree of correlation[7]. On the other hand, for the *Drosophila melanogaster* data, where the four segments correspond to different morphogenetical stages, the molecular regulatory interactions related to the insect's anatomy and physiology (wings) are expected to undergo a series of changes during morphogenesis (Arbeitman *et al.* (2002)), and we expect the segment-specific interaction parameters to show a substantially lower amount of correlation[8]. Hence, we have two time series, consisting of 4 segments each, with complementary types (strengths) of non-homogeneity, for which we can separately study the effect of the proposed Bayesian regularization scheme on the interaction parameters. For our study, we ran MCMC simulations for both the regularized and unregularized TV-DBN model, where we sampled the interaction parameter vectors from the posterior distribution of Equation (28). We repeated the regularized data analyses with three different choices of the hyperhyperparameters α_λ and β_λ in Equation (19) to see how critical an effect they have on the inference results. We sampled the network structure from the posterior distribution of Equation (37). We analyzed the *Drosophila melanogaster* time series twice, for two different structure inference scenarios. In the first data analysis, we selected a network structure from the literature (Zhao *et al.* (2006)) and kept it fixed. In the second data analysis, we sampled the network structure from the posterior distribution of Equation (37).

Figure 7 shows the results for the *Drosophila melanogaster* data. For both structure inference scenarios, we computed a measure of the posterior average correlation between interaction parameters associated with adjacent segments of the four established morphogenetical phases (embryo, larva, pupa, and adult) by sampling the interaction vectors from the posterior distribution with MCMC, agglomerating them into long vectors, and computing the correlation coefficients from these vectors.[9] In consistency with our expectation, it can been seen from the leftmost

[7]That is, for the four *Arabidopsis thaliana* time series we expect the segment-specific regulatory relationships to be very similar, and thus the interaction parameters of the TV-DBN model to be strongly correlated.

[8]That is, for the four morphogenetical stages in *Drosophila melanogaster* we expect the segment-specific regulatory relationships to be different, and thus the interaction parameters of the TV-DBN model to be only moderately correlated.

[9]The mathematical details are as follows: For T graphs, $\mathcal{M}_{(1)}, \ldots, \mathcal{M}_{(T)}$, which are either sampled from Equation (37) or kept fixed (i.e., $\mathcal{M}_{(i)} = \mathcal{M}$), we sample all gene- and segment-specific interaction parameter vectors, $\mathbf{w}_{g,h}$, from Equation (28). For each graph, $\mathcal{M}_{(i)}$, and each segment, h, we arrange the N gene-specific network interaction parameter vectors successively, $\mathbf{w}_{(i),h} :=$

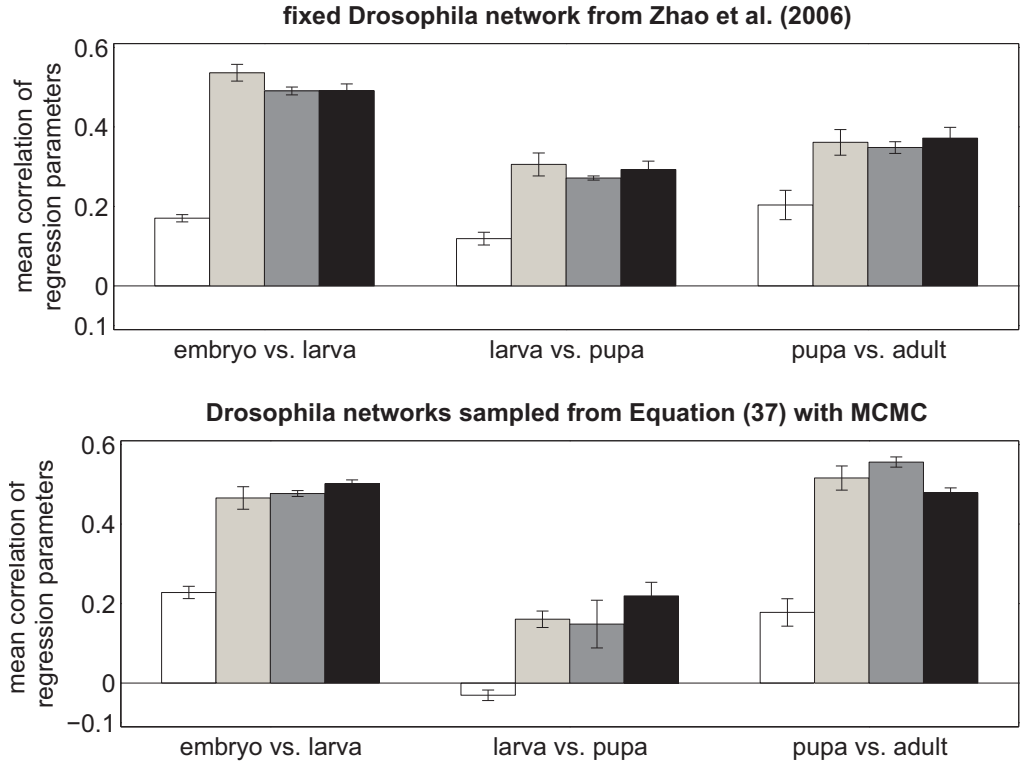

Figure 7: **Similarity of the interaction parameters for adjacent morphogenetical phases of *Drosophila melanogaster*.** The figure shows histograms of our measure of the posterior average correlation for adjacent morphogenetical phases of *Drosophila melanogaster*. These adjacent morphogenetical phases are indicated by the columns. *Left column:* Embryo to larva; *centre column:* larva to pupa; *right column:* pupa to adult. The top row shows the results from MCMC simulations where the structure of the gene regulatory network was taken from the literature (Zhao *et al.* (2006)) and kept fixed. The bottom row shows the results from simulations in which the network structure was sampled from the posterior distribution, see Equation (37), with MCMC. Each panel contains a histogram that shows the measure of the posterior average correlation between interaction parameters associated with adjacent time series segments. The leftmost *white* bars show the values obtained with the conventional unregularized TV-DBN. The *grey* bars to the right show the average correlation coefficients obtained with the proposed coupled TV-DBNs, where the grey shading indicates the setting of the hyperhyperparameters α_λ and β_λ in Equation (19). *Light grey*: $\alpha_\lambda = \beta_\lambda = 0.3$; *dark grey*: $\alpha_\lambda = \beta_\lambda = 1$; *black*: $\alpha_\lambda = \beta_\lambda = 10$. For each setting 5 independent MCMC simulations have been performed and have led to very similar results, indicated by small error bars reflecting one standard deviation of the average posterior correlation measure.

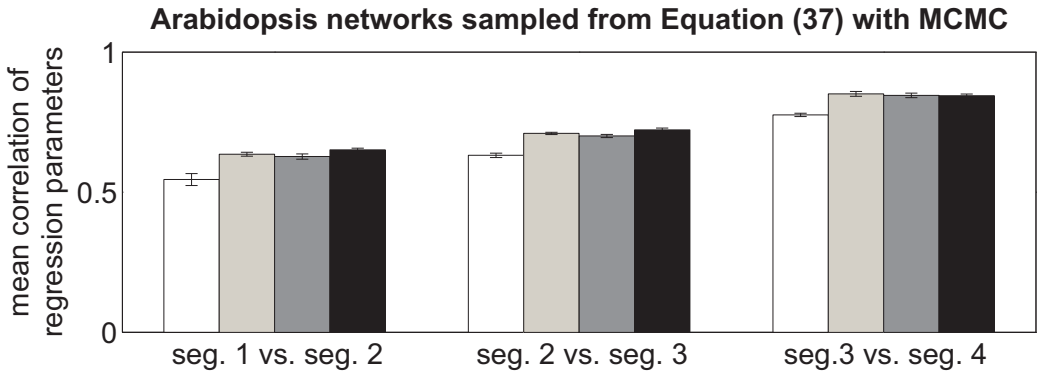

Figure 8: **Similarity of the interaction parameters for adjacent time series of *Arabidopsis thaliana*.** The figure shows histograms of our measure of the posterior average correlation for adjacent *Arabidopsis thaliana* time series. The four time series segments are indicated by the column. *Left column:* first time series segment versus second time series segment; *centre column:* second segment versus third segment; *right column:* third segment versus fourth segment. Details on the segmentation can be found in Table 3 in Subsection 4.2. The network structures were sampled from the posterior distribution, see Equation (37), with MCMC. Each panel contains a histogram that shows the measure of the posterior average correlation between interaction parameters associated with adjacent time series segments, for four different methods. The leftmost *white* bars show the values obtained with the conventional unregularized TV-DBN. The *grey* bars to the right show the posterior average correlation obtained with the proposed coupled TV-DBNs, where the grey shading indicates the setting of the hyperhyperparameters α_λ and β_λ in Equation (19). *Light grey*: $\alpha_\lambda = \beta_\lambda = 0.3$; *dark grey*: $\alpha_\lambda = \beta_\lambda = 1$; *black*: $\alpha_\lambda = \beta_\lambda = 10$. For each hyperhyperparameter setting 5 independent MCMC simulations have been performed and have led to very similar results, indicated by small error bars reflecting one standard deviation of the average posterior correlation measure. The results suggest that the similarity between interaction parameters in different time serie segments increases as a result of the proposed Bayesian information coupling scheme, and that this scheme is fairly robust with respect to a variation of the hyperhyperparameters.

white bars in Figure 7 that the correlation of the interaction parameters is relatively weak for the conventional unregularized TV-DBN model. Our simulations with the novel regularized TV-DBN model lead to increased average similarities between the interaction parameters from the four morphogenetical stages as a consequence of information coupling between the time series segments. For *Arabidopsis thaliana* we also computed our posterior average correlation measure and observed a similar effect, as seen from Figure 8, although, as mentioned above, the correlations are already larger from the outset. This increase in the correlation between the segment-specific interaction parameters points to a shrinkage effect, as one would expect from a Bayesian hierarchical model, in the sense of the well-known "Stein and Lindley effect" (Stein (1955); Lindley (1962)). This shrinkage has the potential to improve the inference for time series segments that are fairly short, as we will demonstrate in the next section. We note that for both biological systems, a perfect correlation, which would be biologically unrealistic due the different biological conditions (different morphogenetical stages in *Drosophila melanogaster*, different external conditions in *Arabidopsis thaliana*) is avoided. We also investigated the dependence of the results on the choice of hyperhyperparameters α_λ and β_λ in Equation (19), and were reassured by the fact that they appear to have only a minor influence (see Figures 7 and 8).

The next question we wanted to address is what influence the proposed Bayesian regularization scheme has on the inference of the network topology. To this end, we inferred network topologies for fixed coupling hyperparameter, λ_g, in Equation (14). For various strengths of the coupling hyperparameter, λ_g, in Equation (14) we computed the posterior probabilities of the gene-specific subnetwork structures, i.e., the set of regulatory genes. The results for *Drosophila melanogaster* are shown in Figure 9, and those for the *Arabidopsis thaliana* data are shown in Figure 10. For both biological systems there are some some genes, like TWI (*Drosophila melanogaster*) or PRR3 (*Arabidopsis thaliana*), that are not affected by the Bayesian regularization scheme, but for others, like ACTN (*Drosophila melanogaster*) or TOC1 (*Arabidopsis thaliana*), there are clear transitions and bifurcation points as λ_g and, hence, the strength of the information coupling is varied. This suggests that the proposed Bayesian regularization scheme has a clear influence on the inferred network structure, which is the entity that is, ultimately, the focus of interest.

$(\mathbf{w}^{\mathsf{T}}_{(i),g=1,h},\ldots,\mathbf{w}^{\mathsf{T}}_{(i),g=N,h})^{\mathsf{T}}$. For each segment, $h = 1,\ldots,4$, these vectors, $\{\mathbf{w}_{(i),h}|i = 1,\ldots,T\}$, can be agglomerated into one single long vector: $\mathbf{w}_h := (\mathbf{w}^{\mathsf{T}}_{(1),h},\ldots,\mathbf{w}^{\mathsf{T}}_{(T),h})^{\mathsf{T}}$. Our measure of similarity is the Pearson correlation coefficient between two vectors, $\mathbf{w}_h$ and $\mathbf{w}_{h+1}$, corresponding to adjacent segments.

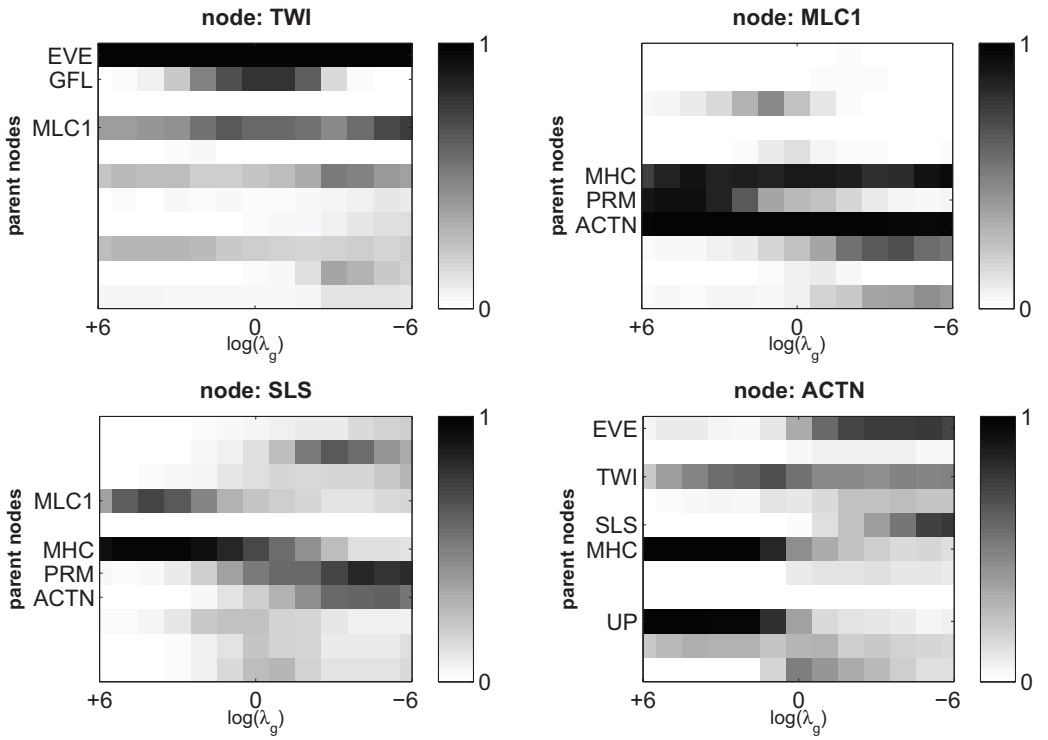

Figure 9: **Influence of the Bayesian regularization scheme on the gene regulatory network structure for the *Drosophila melanogaster* data.** Each of the four panels, which refer to different target genes (shown on top), shows the posterior probability of the subnetwork topology as a function of the coupling hyperparameter, λ_g, in Equation (14). The coupling hyperparameter, λ_g, is represented by the horizontal axes, and the vertical axes represent the potential parent nodes (genes) for the selected target genes. Labels on the vertical axes appear only for those parent nodes (genes) that yield the highest posterior probabilities. The grey shading is explained by the legends, ranging from white, indicating a posterior probability of zero, to black, which indicates a posterior probability of 1. Some genes, like TWI, are not affected by the Bayesian regularization scheme in the sense that their associated subnetworks (i.e., their regulators) are indifferent with respect to a variation of λ_g. However, for other genes, like ACTN, there is a clear transition from one subnetwork topology to another as λ_g and, hence, the strength of the information coupling is varied.

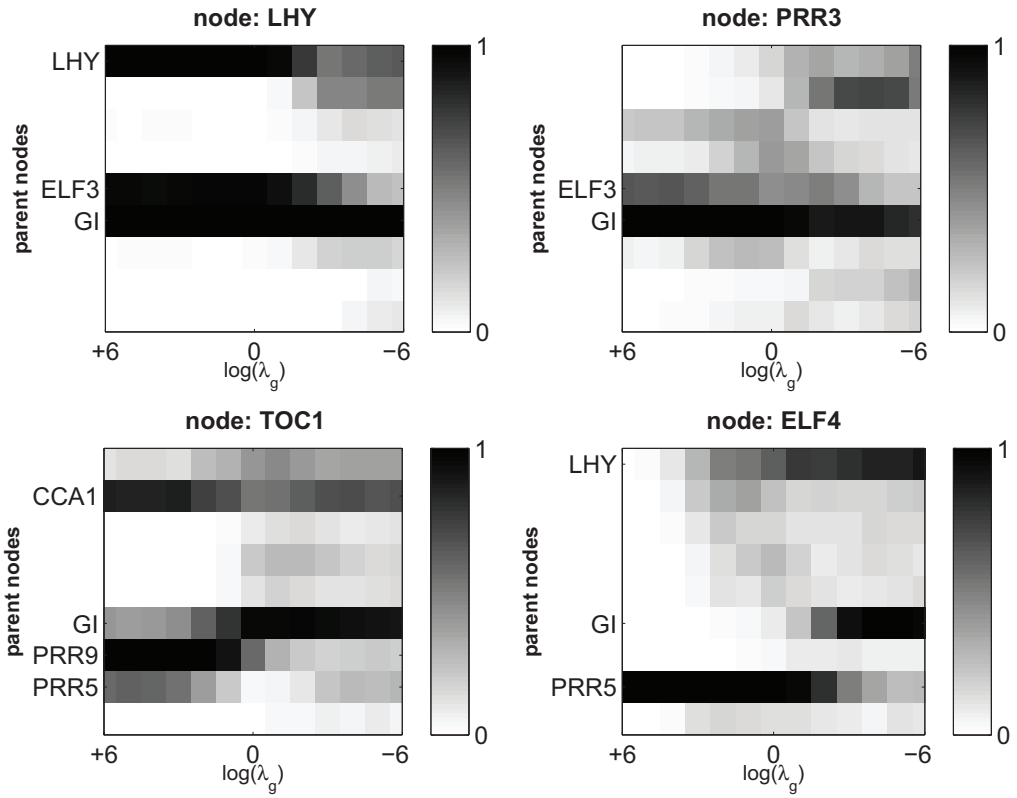

Figure 10: **Influence of the Bayesian regularization scheme on the gene regulatory network structure for the *Arabidopsis thaliana* data.** Each of the four panels, which refer to different target genes (shown on top), shows the posterior probability of the subnetwork topology as a function of the strength of the coupling hyperparameter, λ_g, in Equation (14). The strength of the coupling hyperparameter, λ_g, is represented by the horizontal axes, and the vertical axes represent the potential parent nodes (genes) for the selected target genes. Labels on the vertical axes appear only for those parent nodes (genes) that yield the highest posterior probabilities. The grey shading is explained by the legends, ranging from white, indicating a posterior probability of zero, to black, which indicates a posterior probability of 1. Some genes, like PRR3, are not affected by the Bayesian regularization scheme in the sense that their associated subnetworks (i.e., their regulators) are indifferent with respect to a variation of λ_g. However, for other genes, such as TOC1 and ELF4, there are clear transitions from one subnetwork topology to another as λ_g and, hence, the strength of the information coupling is varied.

The critical question one would finally like to ask is whether the effect on network structure inference leads to an actual improvement in the gene regulatory network reconstruction. Unfortunately, for the *Drosophila melanogaster* and the *Arabidopsis thaliana* data this evaluation is infeasible owing to the absence of a proper gold standard. Several authors aim to pursue an evaluation without gold standard by arguing for the biological plausibility of subsets of inferred interactions. However, such an approach inevitably suffers from a certain selection bias and is somewhat subject to subjective interpretation. We have therefore elected to abandon this route and rather use a real-world in vivo study from synthetic biology, for which we do have a gold standard, in Subsection 6.3.

6.3 Gene regulation in *Saccharomyces cerevisiae*

In the previous two subsections, we have evaluated our Bayesian regularization scheme on two types of data: synthetic and real-world time series. The advantage of the synthetic data (see Subsection 4.1) is the fact that the true network structure is known and that we can therefore objectively score the network reconstruction accuracy. However, the disadvantage is that the data come from a simulation study that oversimplifies the complexity of true biological systems. This shortcoming is addressed with the real gene expression time series obtained during the life cycle of *Drosophila melanogaster* (see Subsection 4.3). However, as we have just argued, the absence of a gold standard renders an objective evaluation of the network reconstruction accuracy infeasible. The purpose of the present section is to evaluate the network reconstruction accuracy on a data set from synthetic biology, which combines the advantages of the two previous sections: having a known gold-standard *and* resulting from an in vivo (i.e., real-world) study. Given that synthetic biology is a new emerging research field, though, the size of the known true gene regulatory network is inevitably quite small.

We tested the proposed coupled time varying dynamic Bayesian network (TV-DBN) model on gene expression profiles from *Saccharomyces cerevisiae* (yeast), using the data from Cantone *et al.* (2009). The authors synthetically designed living yeast cells so that they contained a known gene regulatory subnetwork composed of 5 genes (shown in Figure 1). They measured the gene expression levels of these genes *in vivo* with RT-PCR at 37 timepoints over 8 hours. In about the middle of this time period, the environment was changed by switching the carbon source from galactose to glucose. In our study, we tried to reconstruct the known gene regulatory subnetwork from the RT-PCR time courses with two types of time-varying DBNs: the conventional uncoupled TV-DBN model without information sharing among the segments, and the Bayesian coupling scheme proposed in the present paper, referred to as coupled TV-DBN model. We sampled the network

structure, the number of changepoints and the location of changepoints with the MCMC scheme described in Section 3.3. We repeated the MCMC simulations for different hyperparameter values, p, of the negative binomial prior on the number of changepoints; see Section 3.2 and Appendix I. In Appendix V we briefly investigate the posterior distribution of the gene-specific numbers of changepoints for six different hyperparameter values, p, of the negative binomial prior (see Equation (51) in Appendix I). Figure 20) in Appendix V shows the posterior distribution of the gene-specific numbers of changepoints, which were inferred with different hyperparameters, p. The gene-specific posterior means of the numbers of changeponts, $\overline{K}_g$ ($g = 1, \ldots, 5$), along with the averages over all five genes, $\overline{K}$, and the mean AUC(-ROC) values can be found in Table 5 in Appendix V. For each of the five genes in *Saccharomyces cerevisiae* the average number of changepoints increases in the hyperparameter p. It appears that there are fewer changepoints for gene GAL4 than for the other four genes. That is, the number of changepoints for gene GAL4 peaks at 0 up to $p = 0.05$, indicating that the regulation of GAL4 changes less substantially between the switch-off and the switch-on time series than the regulation of the other genes. Figure 11 shows the average AUC scores plotted against the posterior mean of the number of changepoints, $\overline{K}$.[10] The best performance is obtained for $\overline{K} \approx 1$, which reflects the imposed environment change related to the switch of the carbon source from galactose to glucose. The value of $\overline{K} = 0$ corresponds to the traditional homogeneous DBN, for which the network reconstruction is significantly worse. Much larger values of $\overline{K}$ render the model over-flexible, which is reflected by a decline in the AUC scores. However, this degradation is much worse for the unregularized (uncoupled) TV-DBN than for the proposed regularized (coupled) TV-DBN. While large values of $\overline{K}$ can in principle be ruled out by an informative restrictive prior, the availability of prior knowledge about changepoint scenarios cannot always be assumed, and the appropriate figure of merit for a comparative evaluation in Figure 11 is the area under the curve rather than its maximum. From that perspective, our study suggests that the Bayesian coupling scheme that we have proposed renders the performance of time-varying dynamic Bayesian networks (TV-DBNs) more robust with respect to a misspecification of the prior and redeems the effect of over-fitting as a consequence of potential model over-flexibility.

[10]The posterior mean numbers of changepoints and the mean AUC scores are averages over 5 independent MCMC simulations on the *Saccharomyces cerevisiae* data.

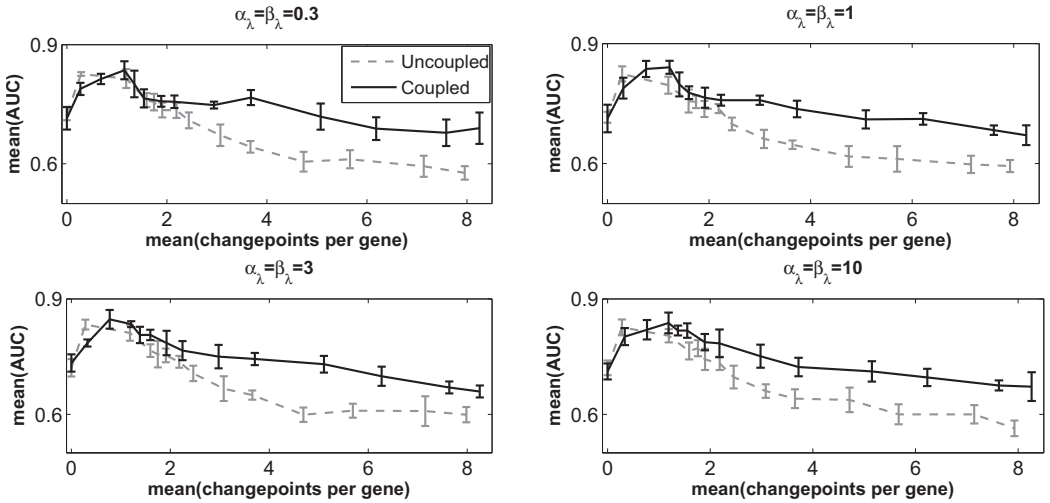

Figure 11: **Network reconstruction accuracy for *Saccharomyces cerevisiae*.** The graphs show the network reconstruction accuracy (ordinate) plotted against the average posterior mean number of changepoints per gene (abscissa) for the conventional uncoupled TV-DBN model (dashed line) and the proposed coupled TV-DBN model (solid line). The network reconstruction accuracy (ordinate) is quantified in terms of mean AUROC scores, as defined in Subsection 5.2, averaged over 5 independent MCMC simulations, with the vertical bars indicating standard deviations. The four panels show the results obtained with four settings of the hyperhyperparameters α_λ and β_λ defined in Equation (19), with values of $\alpha_\lambda = \beta_\lambda \in \{0.3, 1, 3, 10\}$. For similar plots with AUPRC scores, see Appendix III.

7 Conclusion

Time-varying dynamic Bayesian networks (TV-DBNs) have recently been proposed as a promising tool for inferring gene regulatory networks from non-stationary gene expression time series. Modelling time-varying dynamic Bayesian networks (TV-DBNs) with a multiple changepoint process, as pursued by various authors (Robinson and Hartemink (2009, 2010); Grzegorczyk and Husmeier (2009, 2011); Lèbre *et al.* (2010)), is popular due to the fact that conditional on the changepoints, the marginal likelihood can be computed in closed form. All previous studies (Robinson and Hartemink (2009, 2010); Grzegorczyk and Husmeier (2009, 2011); Lèbre *et al.* (2010)) compute the marginal likelihood from a typically uninformative parameter prior that is the same for all time series segments. This approach ignores the fact that many systems, e.g., regulatory processes and signalling pathways in the

cell, evolve gradually, e.g., during an organism's development (morphogenesis) or in adaptation to changing environmental conditions. In the present work, we have proposed a Bayesian regularization scheme in which the parameters associated with separate time series segments are *a priori* encouraged to be similar. Avoiding the fallacies of a Bayesian filter, we have introduced a coupling hyperparameter that is shared among the segments and that is itself inferred from the data in a Bayesian sense. Our assessment on simulated data in Subsection 6.1 has revealed that the proposed regularized TV-DBN model leads to an improvement in the network reconstruction accuracy, and we have quantified the robustness with respect to a systematic evolution of the interaction parameters (via node-specific vector rotations in parameter space). In Subsection 6.2 we have demonstrated on gene expression time series from *Drosophila melanogaster* and *Arabidopsis thaliana* that the proposed coupled TV-DBN model yields more strongly correlated interaction parameters for segmented time series than the conventional uncoupled TV-DBN model. Furthermore, it could be shown that the coupling strength between segments can have a substantial effect on the inferred network structures. Finally, on real data from RT-PCR experiments in *Saccharomyces cerevisiae*, we have demonstrated in Subsection 6.3 the increased robustness of the proposed coupled TV-DBN model with respect to a variation in the (unknown) number of changepoints.

Our future work will seek an extension of the proposed coupled TV-DBN model in two respects. First, for the *Drosophila melanogaster* data, an increased flexibility of the model would be adequate, to allow the network structure to vary between different morphogenetic stages, as in Lèbre *et al.* (2010). To this end, we will aim to integrate the proposed parameter coupling scheme into the method of Lèbre *et al.* (2010). Second, for the *Arabidopsis thaliana* data, a nonsequential hierarchical coupling of the interaction parameters would be more appropriate, as the different externally controlled experimental conditions lack a natural temporal order. To address this, we will complement the proposed sequential information coupling scheme with a hierarchical, non-sequential one.

Appendices

Appendix I contains some algorithmic implementation details, Appendix II is devoted to the technical details on the data generating mechanisms for simulated network data, Appendix III presents an additional study on simulated network data, and finally, Appendix IV shows that all our findings on the network reconstruction accuracies can also be produced when the area under the precision recall curve rather than the area under the ROC curve is employed for evaluation.

Appendix I - Implementation details

The point process prior for the changepoint sets $\boldsymbol{\tau}_g$

In Section 3.2 we did not specify the prior, $P(\{\boldsymbol{\tau}_g\})$, on the gene-specific changepoint sets, $\boldsymbol{\tau}_g = \{\tau_{g,1},\ldots,\tau_{g,K_g-1}\}$ $(g = 1,\ldots,N)$, explicitly. Here, we provide the details: We assume that the gene-specific changepoint sets, $\boldsymbol{\tau}_g$, are independently distributed, $P(\{\boldsymbol{\tau}_g\}) = \prod_{g=1}^{N} P(\boldsymbol{\tau}_g)$, and we follow Fearnhead (2006) and employ a point process prior to model the distances between successive changepoints for each gene, g $(g = 1,\ldots,N)$. In the point process model $s(t)$ $(t = 1,2,3,\ldots)$ denotes the prior probability that there are t time points between two successive changepoints, $\tau_{g,j-1}$ and $\tau_{g,j}$, on the discrete interval $\{2,\ldots,T-1\}$. The prior probability, $P(\boldsymbol{\tau}_g)$, of the changepoint set of gene g, $\boldsymbol{\tau}_g = \{\tau_{g,1},\ldots,\tau_{g,K_g-1}\}$, containing K_g-1 changepoints, $\tau_{g,j}$ $(j = 1,\ldots,K_g-1)$ with $1 < \tau_{g,j-1} < \tau_{g,j} < T$ $(j = 2,\ldots,K_g-1)$, is:

$$\begin{aligned} P(\boldsymbol{\tau}_g) &= P(\tau_{g,1},\ldots,\tau_{g,K_g-1}) \\ &= s_0(\tau_{g,1}) \left(\prod_{j=2}^{K_g-1} s(\tau_{g,j} - \tau_{n,j-1}) \right) (1 - S(\tau_{g,K_g} - \tau_{g,K_g-1})), \end{aligned} \tag{49}$$

where $\tau_{g,0} = 1$ and $\tau_{g,K_g} = T$ are two pseudo changepoints, $s_0(.)$ is the prior distribution of the first changepoint, $\tau_{g,1}$, and

$$S(t) = \sum_{s=1}^{t} s(t); \quad S_0(t) = \sum_{s=1}^{t} s_0(t) \tag{50}$$

are the cumulative distribution functions corresponding to $s(.)$ and $s_0(.)$. For $s(.)$ we follow Fearnhead (2006) and use the probability mass function of the negative binomial distribution[11] NBIN(p,k) with hyperparameters p and k:

$$s(t) = \left(\begin{array}{c} t-1 \\ k-1 \end{array} \right) p^k (1-p)^{t-k}. \tag{51}$$

In a point process model on the positive *and* negative integers the probability mass function of the first changepoint, $\tau_{g,1} \in \{2,\ldots,T-1\}$, is a mixture of k negative binomial distributions:

$$s_0(\tau_{g,1}) = \frac{1}{k} \sum_{i=1}^{k} \left(\begin{array}{c} (\tau_{g,1}-1)-1 \\ i-1 \end{array} \right) p^i (1-p)^{(\tau_{g,1}-1)-i}. \tag{52}$$

[11] Note that the negative binomial distribution can be seen as a discrete version of the Gamma distribution.

For our analysis of the synthetic biology data from *Saccharomyces cerevisiae* (see Subsection 6.3) we used a fixed value for k ($k = 1$) and we varied the hyperparameter p. In the first instance, we started with six different values: $p \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ that cover the whole area from zero changepoints (per gene) to about 8 changepoints per gene. Then, to shed more light onto the more interesting area with fewer changepoints per gene, we additionally employed: $p \in \{0.001, 0.01, 0.02, 0.025, 0.03, 0.04, 0.05, 0.075, 0.15\}$.

MCMC sampling and convergence diagnostics

For the simulated network data from Subsection 4.1, the expression data from *Arabidopsis thaliana* from Subsection 4.2, and the expression data from *Drosophila melanogaster* from Subsection 4.3, we assume that the true changepoints are known (supervised approach) and we set the lengths of the burn-in and the sampling phase to 25,000 MCMC iterations each. In each single iteration *all* gene-specific variances σ_g^2 and signal-to-noise hyperparameters, δ_g (and the coupling hyperparameter λ_g for the proposed regularized TV-DBN), are re-sampled and the performance of one basic operation on the parent node set, $\boldsymbol{\pi}_g$, is proposed for *each* gene, g ($g = 1, \ldots, N$). In our implementation there are three basic operations on $\boldsymbol{\pi}_g$: (i) adding a new parent node to $\boldsymbol{\pi}_g$, (ii) deleting one of the parent nodes from $\boldsymbol{\pi}_g$, and (iii) substituting a parent node from $\boldsymbol{\pi}_g$ for another one, as proposed in Grzegorczyk and Husmeier (2011). During the sampling phase of length 25,000 we sampled every 250 iterations to obtain a network sample of size 100. From this sample a marginal edge posterior probability can be computed for each individual network edge.

For the synthetic biology data from *Saccharomyces cerevisiae* we assume that the gene-specific changepoint sets, $\boldsymbol{\tau}_g$, are unknown and also have to be inferred via MCMC. We therefore doubled the lengths of the burn-in and the sampling phase to 50,000 MCMC iterations each, and in each single MCMC step we additionally performed one move on the changepoint set, $\boldsymbol{\tau}_g$, for *each* gene, g. In our implementation there are three types of moves: (i) a changepoint birth adds a new changepoint to $\boldsymbol{\tau}_g$, (ii) a changepoint death move removes a changepoint from $\boldsymbol{\tau}_g$, and (iii) a changepoint re-allocation move substitutes a changepoint from $\boldsymbol{\tau}_g$ for another one.

We applied the standard diagnostic based on trace plots (Giudici and Castelo (2003))and potential scale reduction factors (Gelman and Rubin (1992)) to determine appropriate MCMC simulation lengths. In particular for the real data from *Saccharomyces cerevisiae* we started five MCMC simulations from different network, $\mathcal{M}$, and changepoint set, $\boldsymbol{\tau}_g$ ($g = 1, \ldots, N$), initializations for half a dozen point process hyperparameters $p = 0, 0.1, 0.2 \ldots, 0.5$, in Equation (51) to assess

convergence. MCMC convergence was monitored in terms of the potential scale reduction factors (PSRFs) based on the marginal edge posterior probabilities. For the above mentioned MCMC run lengths we observed a sufficient degree of convergence ($PSFR < 1.2$ for *all* individual edges). Because of the computational costs this convergence diagnostic could not be determined for every MCMC simulation that was employed in our study. However, we assume that the MCMC simulations with hyperparameters $p = 0, 0.1, 0.2 \ldots, 0.5$, are representative, and since we consistently observed a sufficient degree of convergence, according to the above mentioned criterion, we concluded that the selected MCMC run lengths also ensure convergence for other hyperparameters p. In Section 6 we consistently report results of those MCMC runs that were seeded with empty parent sets ($\boldsymbol{\pi}_g = \emptyset$ for all g), empty changepoint sets ($\boldsymbol{\tau}_g = \emptyset$ for all g) and the following hyperparameters: $\sigma_g^2 = \lambda_g = \delta_g = 1$ for all g. Details on further (hyper-)hyperparameter settings can be found in Subsection 5.1.

Network reconstruction accuracy

The network reconstruction accuracy can be evaluated as follows: Let $\mathcal{M}^{\diamond}(n,j) = 1$ indicate that the true graph possesses the edge $X_n \rightarrow X_j$, while $\mathcal{M}^{\diamond}(n,j) = 0$ indicates that there is no edge from X_n to X_j. For both Bayesian network models network structures, $\mathcal{M}_1, \mathcal{M}_2, \ldots$, can be sampled from Equation (37) and a marginal edge posterior probability, $e_{n,j}$, can be computed for every edge, $\mathcal{M}^{\diamond}(n,j)$. For $\zeta \in [0,1]$ we define $E(\zeta) := \{\mathcal{M}(n,j) | e_{n,j} \geq \zeta\}$ as the set of all edges $\mathcal{M}(n,j)$ whose posterior probabilities exceed the threshold, ζ. Since the true edges are known, for each $E(\zeta)$ the number of true positive $TP[\zeta]$, false positive edges $FP[\zeta]$, true negative edges $TN[\zeta]$, and false negative edges $FN[\zeta]$ can be counted. From this we can compute the true positive rate, $TPR[\zeta] = TP[\zeta]/(TP[\zeta] + FN[\zeta])$ (also called *recall* or *sensitivity*), the false positive rate, $FPR[\zeta] = FP[\zeta]/(TN[\zeta] + FP[\zeta])$, and the precision, $PRE[\zeta] = TP[\zeta]/(TP[\zeta] + FP[\zeta])$. Plotting the $TPR[\zeta]$ values (y-axis) against the corresponding $FPR[\zeta]$ values (x-axis) and connecting neighboring points by linear interpolation gives the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) is a quantitative measure that can be obtained by numerically integrating the ROC curve on the interval $[0,1]$; larger AUC values indicate a better network reconstruction accuracy, where 1 indicates perfect prediction, whereas 0.5 corresponds to a random expectation. An alternative score of the network reconstruction accuracy can be obtained by numerically integrating the Precision-Recall (PR) curve. PR curves can be obtained as follows: (i) The $PRE[\zeta]$ values (y-axis) are plotted against the corresponding $TPR[\zeta]$ values (x-axis). (ii) Unlike the ROC curve, neighboring points cannot be connected by

straight lines and a nonlinear interpolation is required[12]. In our implementation we use the interpolation scheme described in Davis and Goadrich (2006). (iii) As the precision is not defined for TP=0 and FP=0 (PRE=0/0), we integrate the PR curve on the interval $[(1/E), 1]$ where E is the number of edges of the true graph, $\mathcal{M}^{\diamond}$.

Appendix II: Simulated network data

In this section we describe two data generating mechanisms, which can be used to generate non-homogeneous dynamic network data, $\mathcal{D} = \{y_{g,t}\}$, $1 \leq g \leq N$, $1 \leq t \leq T$, where N is the number of network nodes and T denotes the number of time points. For a given network topology, e.g., the network shown in Figure 1, we generate data from a multiple changepoint linear regression model with time-dependent regression parameters. For our simulation studies we randomly sample the first time point at $t = 1$ from independent Gaussian distributions with mean 0 and variance $\sigma^2 = 0.025$, before we generate data for 40 subsequent time points ($t = 2, \ldots, 41$). We assume that there are three equidistant changepoints, where the interaction parameters are perturbed.

For a given network topology with N genes let $\mathbf{D}_{.,t} = (y_{1,t}, \ldots, y_{N,t})^{\mathsf{T}}$ denote the column vector of realizations at time point t ($t = 1, \ldots, 41$). The complete data set, $\mathcal{D}$, is then a N-by-41 data matrix, symbolically: $\mathcal{D} = (\mathbf{D}_{.,1}, \ldots, \mathbf{D}_{.,41})$. We initialize the genes at time point $t = 1$ independently and randomly: $y_{g,1} = \epsilon_{g,1}$ ($g = 1, \ldots, N$), where $\epsilon_{1,1}, \ldots \epsilon_{N,1}$ are independent and identically distributed (iid) Gaussian (noise) variables with expectation 0 and variance σ^2, symbolically $\epsilon_{g,1} \sim \mathcal{N}(0, \sigma^2)$. The subsequent values $y_{g,t}$ ($g = 1, \ldots, N$; $t = 2, \ldots, 41$) are then generated as follows:

$$y_{g,t} = \tilde{w}_{g,0} + \mathbf{D}_{.,t-1}^{\mathsf{T}} \tilde{\mathbf{w}}_g + \epsilon_{g,t}, \tag{53}$$

where the $\epsilon_{g,t}$ variables are iid $\mathcal{N}(0, \sigma^2)$ (dynamic) noise variables, $\tilde{w}_{g,0}$ is an intercept, and $\tilde{\mathbf{w}}_g = (w_{g,1}, \ldots, w_{g,N})^{\mathsf{T}}$ is the gene-specific vector of interaction parameters with $w_{g,i} = 0$ if the i-th gene is not a parent of gene g ($i \notin \boldsymbol{\pi}_g$). For convenience we introduce the notations, $\mathbf{w}_g = (\tilde{w}_{g,0}, \tilde{\mathbf{w}}_g^{\mathsf{T}})^{\mathsf{T}}$ and $\mathbf{Y}_{.,t-1} = (1, \mathbf{D}_{.,t-1}^{\mathsf{T}})^{\mathsf{T}} = (1, y_{1,t-1}, \ldots, y_{N,t-1})^{\mathsf{T}}$, so that Equation (53) can be re-written compactly:

$$y_{g,t} = \mathbf{Y}_{.,t-1}^{\mathsf{T}} \mathbf{w}_g + \epsilon_{g,t}. \tag{54}$$

For our simulation study we require non-homogeneous time series, i.e., interaction parameter vectors, $\mathbf{w}_g$, that depend on time. We therefore modify Equation (54):

$$y_{g,t} = \mathbf{Y}_{.,t-1}^{\mathsf{T}} \mathbf{w}_{g,H(t)} + \epsilon_{g,t}, \tag{55}$$

[12]The linear interpolation has to be done in terms of the true positives (TPs) and false positives (FPs), and this corresponds to a nonlinear interpolation in the precision recall representation.

where the vectors of interaction parameters, $\mathbf{w}_{g,H(t)}$, depend on t. This yields a third subscript index for the components in $\mathbf{w}_{g,H(t)}$: $\mathbf{w}_{g,H(t)} = (w_{g,0,H(t)}, w_{g,1,H(t)}, \ldots, w_{g,N,H(t)})^{\mathsf{T}}$, where the first index g refers to the gene, the last one, $H(t)$, indicates the segment, and the second index, $0, \ldots, N$, corresponds to the potential parent nodes. $H(t)$ is a step function that indicates the time series segment to which time point t belongs. In our setting (i.e., $T = 41$ time points and three equidistant changepoints) there are four time series segments, $h = 1, \ldots, 4$, and the step function, $H(t)$, is given by:

$$H(t) = \begin{cases} 1, & 2 \leq t \leq 11 \\ 2, & 12 \leq t \leq 21 \\ 3, & 22 \leq t \leq 31 \\ 4, & 32 \leq t \leq 41 \end{cases}. \tag{56}$$

Scenario 1: Parameter vector rotations

The initial values of the node-specific interaction parameter vectors, $\mathbf{w}_{g,1}$ ($g = 1, \ldots, N$), for the first segment ($\{t : H(t) = 1\}$) are sampled independently from iid standard Gaussian distributions and normalized such that each vector, $\mathbf{w}_{g,1}$, is of Euclidean norm 1:

$$|\mathbf{w}_{g,1}|_2 = \sqrt{\sum_{i=0}^{N} (w_{g,i,1})^2} = 1. \tag{57}$$

At each of the three changepoints, $t = 11$, $t = 21$, and $t = 31$, we rotate the interaction parameter vectors from the previous time series segment, $\mathbf{w}_{g,h-1}$, to obtain a new vector, $\mathbf{w}_{g,h}$ ($h = 2, \ldots, 4$), of Euclidean norm 1. As explained in the main text in Subsection 4.1, the motivation for this rotation is to emulate systematic changes of the regulatory processes.

For interaction parameter vectors with 2 or 3 non-zeros entries, the rotation is effectively either 2- or 3-dimensional. For a 2-dimensional rotation $(x, y)^T$ only one single angle, θ (with respect to the x-axis), is required and we (independently) flip unbiased coins to decide for each interaction parameter vector, $\mathbf{w}_{g,h-1}$ ($g = 1, \ldots, N$), at each changepoint, $t = 11$, $t = 21$, and $t = 31$, whether the corresponding rotation is clockwise ($\theta < 0$) or counterclockwise ($\theta > 0$). The 2-dimensional rotation can be implemented straightforwardly:

$$\mathbf{w}_{g,h} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \mathbf{w}_{g,h-1}. \tag{58}$$

Rotations in three dimensions $(x, y, z)^T$ are specified by three Euler angles ϕ, θ, and ψ, each angle corresponding to an orthogonal rotation axis. The 3-dimensional rotation can also be implemented using a simple matrix operation: $\mathbf{w}_{g,h} = \mathbf{R} \cdot \mathbf{w}_{g,h-1}$ where the rotation matrix, $\mathbf{R}$, is given by:

$$\mathbf{R} = \begin{pmatrix} \cos(\theta)\cos(\psi) & -\cos(\phi)\sin(\psi) + \sin(\phi)\sin(\theta)\cos(\psi) & \sin(\phi)\sin(\psi) + \cos(\phi)\sin(\theta)\cos(\psi) \\ \cos(\theta)\sin(\psi) & \cos(\phi)\cos(\psi) + \sin(\phi)\sin(\theta)\sin(\psi) & -\sin(\phi)\cos(\psi) + \cos(\phi)\sin(\theta)\sin(\psi) \\ -\sin(\theta) & \sin(\phi)\cos(\theta) & \cos(\phi)\cos(\theta) \end{pmatrix}. \quad (59)$$

Again, at each of the three changepoints we (independently) flip three coins for each interaction parameter vector, $\mathbf{w}_{g,h-1}$ ($g = 1, \ldots, N$), to determine for the three rotation axes whether the corresponding movement is clockwise (negative angle) or counterclockwise (positive angle).

The rotation approach ensures that the Euclidean norms of the segment- and node- specific interaction parameter vectors, $\mathbf{w}_{g,h}$, do not change, $|\mathbf{w}_{g,h}|_2 = 1$ ($g = 1, \ldots, N$; $h = 1, \ldots, 4$). In our simulation studies we employed five different rotation angles, $\alpha \in \{0°, 11.25°, 22.5°, 45°, 90°\}$ and for 3-dimensional rotations we set $|\phi| = |\theta| = |\psi| = \alpha$ so that the three Euler angles have equal absolute values α, but randomly and independently sampled signs.

The number of non-zero entries in the interaction parameter vectors, $\mathbf{w}_{g,h}$, is equal to $|\boldsymbol{\pi}_g| + 1$ where $\boldsymbol{\pi}_g$ is the parent node set of gene g, $|.|$ denotes the cardinality and one additional parameter is required as intercept. Consequently, for the *Saccharomyces cerevisiae* network shown in Figure 1 all node-specific interaction parameter vectors consist of either 2 or 3 non-zero entries and all required rotations are 2- or- 3-dimensional. For the RAF network shown in Figure 2 three nodes are exceptions to that rule: Node PIP3 has no parents so that the vectors $\mathbf{w}_{PIP3,h}$ have only one non-zero entry, namely the intercept. Since a 1-dimensional rotation cannot be defined, we assume a zero intercept and exclude PIP3 from the rotation. Moreover, there are two nodes (MEK and AKT) with three parent nodes in the RAF network. To avoid a more complex 4-dimensional rotation, we simply assume that the intercepts of these two nodes are zero. Without intercept the vectors of interaction parameters $\mathbf{w}_{MEK,h}$ and $\mathbf{w}_{AKT,h}$ are 3-dimensional and the rotation can be done in three dimensions, as described above.

Scenario 2: Random perturbations

In addition to the systematic parameter perturbations corresponding to rotations in vector space we also consider random perturbations. In this second scenario we use random perturbations of increasing amplitude ϵ to vary the interaction parameter vectors.

As in scenario no. 1, the initial values of the node-specific interaction parameter vectors, $\mathbf{w}_{g,1}$ ($g = 1, \ldots, N$), for the first segment ($\{t : H(t) = 1\}$) are

sampled independently from iid standard Gaussian distributions and normalized such that each vector, $\mathbf{w}_{g,1}$, is of Euclidean norm 1: $|\mathbf{w}_{g,1}|_2 = 1$. At the three changepoints, $t = 11$, $t = 21$, and $t = 31$, we change the interaction parameter vector from the previous time series segment, $\mathbf{w}_{g,h-1}$, as follows: We independently re-sample the non-zero interaction parameters from iid Gaussian $\mathcal{N}(0,1)$ distributions to obtain a new interaction parameter vector instantiation, $\mathbf{w}_{g,new}$, which we normalize such that it has the Euclidean norm (amplitude) ϵ, symbolically:

$$\mathbf{w}_{g,new} \leftarrow \frac{\epsilon}{|\mathbf{w}_{g,new}|_2} \cdot \mathbf{w}_{g,new}. \tag{60}$$

This new vector with Euclidean norm (amplitude), $|\mathbf{w}_{g,new}|_2 = \epsilon$, is added to the vector of the previous segment, $\mathbf{w}_{g,h-1}$, with amplitude $|\mathbf{w}_{g,h-1}|_2 = 1$. Re-normalization yields the vector of interaction parameters for the new time series segment, h:

$$\mathbf{w}_{g,h} := \frac{\mathbf{w}_{g,h-1} + \mathbf{w}_{g,new}}{|\mathbf{w}_{g,h-1} + \mathbf{w}_{g,new}|_2}. \tag{61}$$

In our simulations we implement the perturbations at the three changepoints with five different amplitudes, $\epsilon \in \{0, 0.125, 0.25, 0.5, 1\}$.

Additive and dynamic noise

For our simulation study we implement both dynamic and additive noise, but our focus is on additive white noise with the objective to keep the signal-to-noise ratio (SNR) constant such that it can be controlled and specified. In a dynamical system, dynamic noise systematically increases the variances of the signals for subsequent time points. E.g., from Equations (53-55) it can be seen that adding (dynamic) noise (via $\epsilon_{g,t}$) at time point t increases the expected variance of the variables at time point t, D_t, which serve as signals for the next time point $t+1$. That is, strong dynamic noise injections increase the variances of the variables in D_t and the signal-to-noise ratio gets weaker over time. Therefore, we use relatively weak dynamic noise levels. We employ Gaussian noise variables, $\epsilon_{g,t}$, with variance $\sigma^2 = 0.025$ in Equations (53-55) and also for initializing the domain variables at the first time point, $t = 1$.

On the other hand, additive white noise can be employed without noise inflation. We add noise to each individual data instantiation, generated as described above (see Scenario 1 and Scenario 2), such that a pre-specified signal-to-noise ratio (SNR) is reached. That is, having generated a time series, $\mathcal{D}$, from one of the two scenarios, we consider it as "noise-free", and in a gene-wise manner we add white noise. That is, for each node, g, we compute the standard deviation, s_g, of its last 40

observations, $\mathbf{D}_{g,2:41} := (d_{g,2}, \ldots, d_{g,41})$, and we add iid Gaussian noise with zero mean and standard deviation $\text{SNR}^{-1}\ s_g$ to each individual observation, $d_{g,t}$, where SNR is the pre-defined signal-to-noise ratio level. That is, in a gene-wise manner we compute:

$$s_g := \frac{1}{SNR}\sqrt{\frac{1}{40}\sum_{t=2}^{41}\left(d_{g,t} - \{\frac{1}{40}\sum_{i=2}^{41} d_{g,i}\}\right)^2}, \tag{62}$$

and for each individual observation, $d_{g,t}$ ($t = 2, \ldots, 41$), of gene g we substitute $d_{g,t}$ for $d_{g,t} + u_{g,t}$ where $u_{g,2}, \ldots, u_{g,41}$ are realizations of iid $\mathcal{N}(0, s_g^2)$ Gaussian variables. In our simulation study we distinguish three signal-to-noise ratios SNR$=10$ (weak noise), SNR$=3$ (moderate noise), and SNR$=1$ (strong noise).

Appendix III: Further empirical results

As an extension to the study on simulated data reported in Subsection 6.1, we also generate synthetic network data from a different data generating mechanism for the *Saccharomyces cerevisiae* network topology shown in Figure 1. As described in Appendix II, we employ random perturbations of increasing amplitude to alter the interaction parameter vectors at the changepoints. The results are very similar to those reported in Subsection 6.1: For homogeneous network data, see Figure 12, we observe that the performance of the proposed regularized model is robust with respect to a variation of the hyperhyperparameters, and it can be seen from Figure 13 that the regularized model performs systematically better than the standard unregularized model.

For the non-homogeneous network data, shown in Figure 13, the results are even slightly more in favor of the proposed regularization scheme than those obtained with the vector rotation approach (see Figure 4 in Subsection 6.1). In particular, for the strong (SNR=10) and moderate (SNR=3) signal-to-noise ratios the proposed regularization scheme consistently outperforms the unregularized model for 4 of 5 amplitudes, $\epsilon \in \{0, 0.125, 0.25, 0.5\}$, of the perturbation. This finding is plausible, since the alternative data generating mechanism tends to yield less substantial perturbations and therefore more gradually evolving interaction parameters. The proposed regularization scheme can profit from that, as it systematically shares information between the segment-specific interaction parameters.

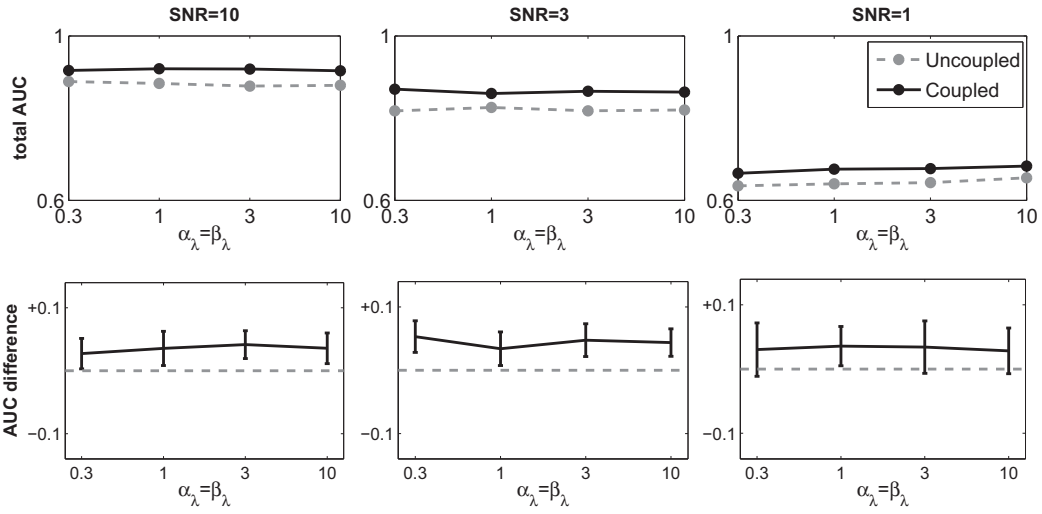

Figure 12: **Additional study: Variation of the hyperhyperparameters for homogeneous data, simulated from the *Saccharomyces cerevisiae* network.** The homogeneous data sets from the *Saccharomyces cerevisiae* network, shown in Figure 1, have been generated as described in Subsection 4.1 and Appendix II. The interaction parameters are kept fixed ($\varepsilon = 0$), and the hyperhyperparameters α_λ and β_λ from Equation (19) vary along the horizontal axis. The network reconstruction accuracy has been evaluated in terms of areas under the ROC curve (AUROC). All simulations were repeated for 25 independent data instantiations, with error bars indicating 95% t-test confidence bounds for the mean differences. The three columns represent SNR levels. The top row shows the absolute AUROC scores, and the bottom row shows the mean difference between the proposed regularized method and the unregularized method.

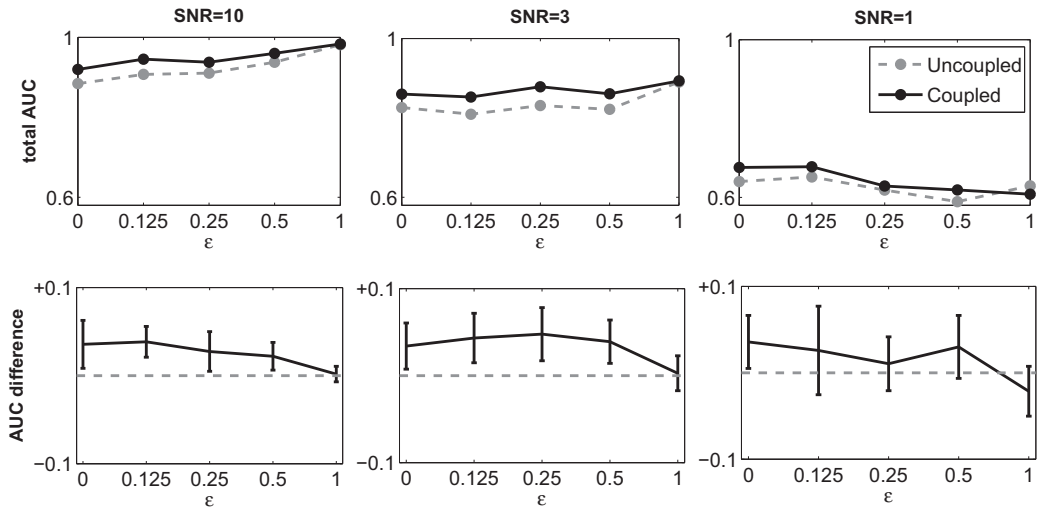

Figure 13: **Additional study: Non-homogeneous data with increasing amplitude of the perturbations, simulated from the *Saccharomyces cerevisiae* network.** The non-homogeneous data sets from the *Saccharomyces cerevisiae* network, shown in Figure 1, have been generated using random perturbations with increasing amplitudes, ε, as described in Subsection 4.1 and Appendix II. The abscissa represents the amplitude, ε, of the random perturbation that has been added to the interaction parameter vectors, while the hyperhyperparameters are kept fixed: $\alpha_\lambda = \beta_\lambda = 1$. The network reconstruction accuracy has been evaluated in terms of areas under the ROC curve (AUROC). All simulations were repeated for 25 independent data instantiations, with error bars indicating 95% t-test confidence bounds for the mean differences. The three columns represent SNR levels. The top row shows the absolute AUROC scores, and the bottom row shows the mean difference between the proposed regularized method and the unregularized method.

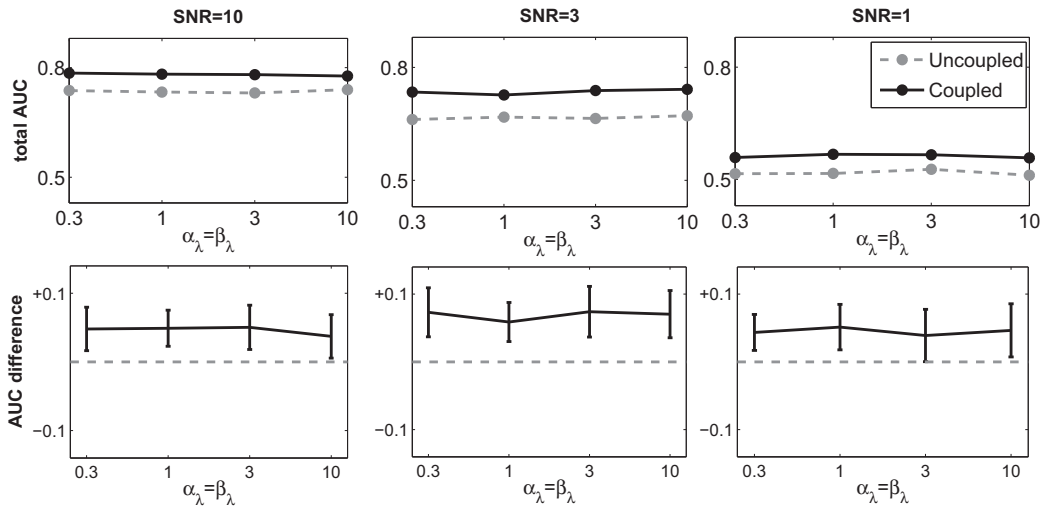

Figure 14: **Network reconstruction accuracy evaluated by areas under the *precision recall curve* for homogeneous data, simulated from the *Saccharomyces cerevisiae* network.** Unlike in Figure 3 the network reconstruction accuracy has been evaluated according to the area under the precision recall curve; see the caption of Figure 3 for further details.

Appendix IV - Another evaluation criterion: The area under the precision recall curve

In this fourth appendix, we show that all our findings on the network reconstruction accuracies can be reproduced with another popular network reconstruction accuracy criterion. Two scores have been established for evaluating the network reconstruction accuracy in systems biology research. The area under the receiver operator characteristic curve (AUROC) and the area under the precision recall curve (AUCPR). A comparison of these two criterions can be found in Davis and Goadrich (2006). These two criterions have been used for evaluating the results of the regularly held DREAM network reconstruction challenge (see, e.g., Prill *et al.* (2010)). Therefore, we evaluated our network inference results with both criterions independently, and we found that both criterions yield very similar results.

Appendix V - Posterior distribution of the number of changepoints for the *Saccharomyces cerevisiae* data.

In this fifth appendix we briefly investigate the posterior distribution of the number of changepoints for the *Saccharomyces cerevisiae* data. Figure 20 shows the

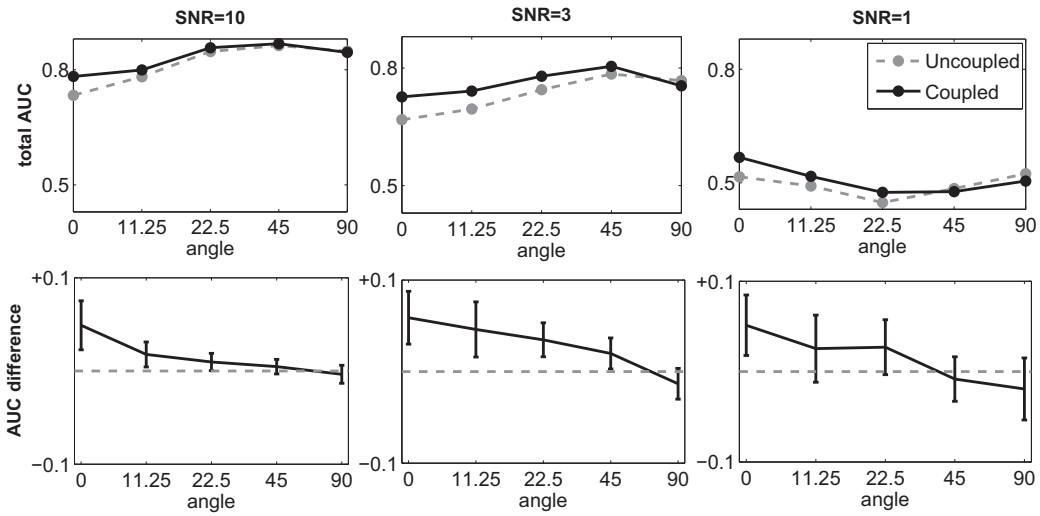

Figure 15: **Network reconstruction accuracy evaluated by areas under the *precision recall curve* for non-homogeneous data from the *Saccharomyces cerevisiae* network.** Unlike in Figure 4 the network reconstruction accuracy has been evaluated according to the area under the precision recall curve; see the caption of Figure 4 for further details.

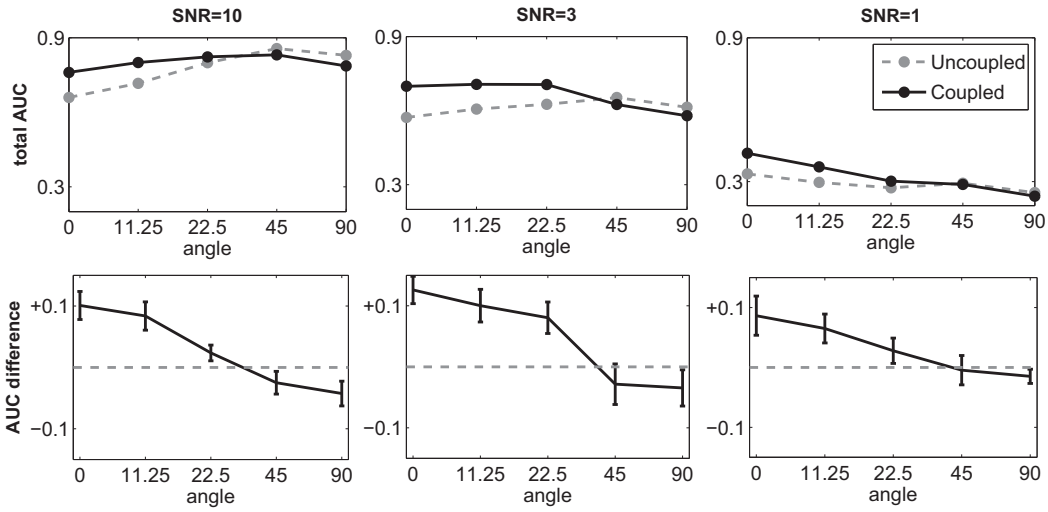

Figure 16: **Network reconstruction accuracy in terms of areas under the *precision recall curve* for non-homogeneous data, from the RAF network.** Unlike in Figure 6 the network reconstruction accuracy has been evaluated according to the area under the precision recall curve; see the caption of Figure 6 for further details.

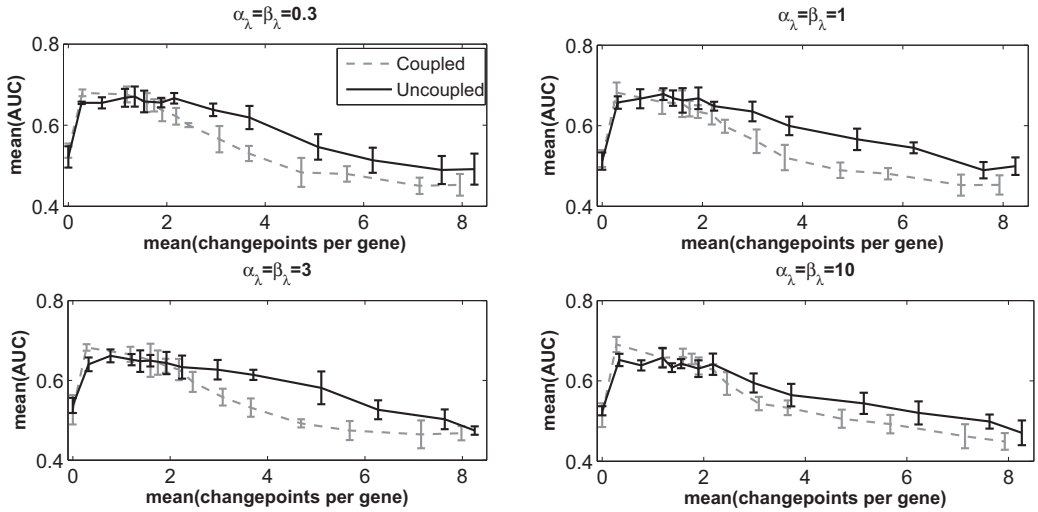

Figure 17: **Network reconstruction accuracy for the in vivo *Saccharomyces cerevisiae* network data in terms of areas under the *precision recall curve*.** Unlike in Figure 11 the network reconstruction accuracy has been evaluated according to the area under the precision recall curve; see the caption of Figure 11 for further details.

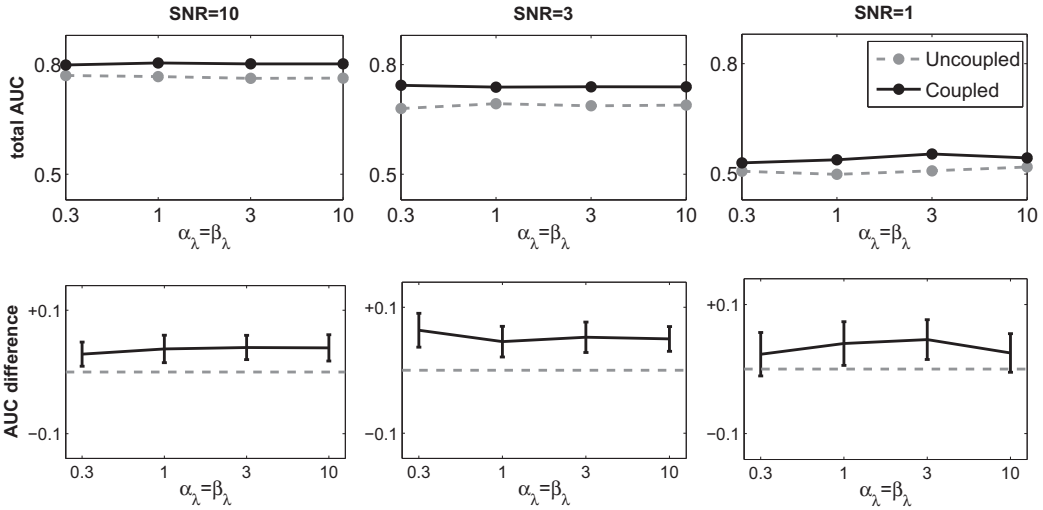

Figure 18: **Network reconstruction accuracy in terms of areas under the *precision recall curve* for homogeneous data from the *Saccharomyces cerevisiae* network.** Unlike in Figure 12 the network reconstruction accuracy has been evaluated according to the area under the precision recall curve; see the caption of Figure 12 for further details.

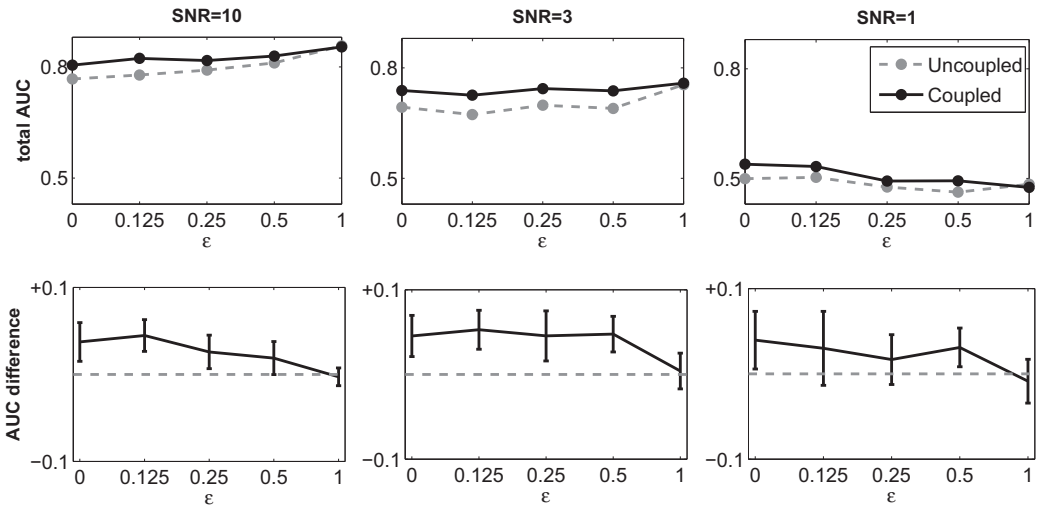

Figure 19: **Network reconstruction accuracy in terms of areas under the *precision recall curve* for non-homogeneous data with increasing amplitudes of the perturbations from the *Saccharomyces cerevisiae* network.** Unlike in Figure 13 the network reconstruction accuracy has been evaluated according to the area under the precision recall curve; see the caption of Figure 13 for further details.

posterior distribution of the gene-specific numbers of changepoints for six different hyperparameter values p of the negative binomial prior (see Equation (51) in Appendix I). The corresponding gene-specific posterior means of the numbers of changeponts, $\overline{K}_g$ ($g = 1, \ldots, 5$), along with the averages over all five genes, $\overline{K}$, and the mean AUC(-ROC) values can be found in Table 5.

	$\overline{K}_{CBF1}$	$\overline{K}_{GAL4}$	$\overline{K}_{SWI5}$	$\overline{K}_{GAL80}$	$\overline{K}_{ASH1}$	$\overline{K}$	mean(AUC)
$p = 0$	0.00	0.00	0.00	0.00	0.00	0.00	0.71
$p = 0.025$	1.91	0.20	1.91	1.07	1.84	1.38	0.80
$p = 0.05$	3.25	0.72	2.71	1.99	2.54	2.24	0.76
$p = 0.1$	4.54	2.49	4.26	3.55	3.69	3.71	0.74
$p = 0.2$	6.64	5.62	6.78	6.00	6.30	6.27	0.71
$p = 0.4$	8.31	8.10	8.37	8.17	8.32	8.25	0.67

Table 5: **Posterior mean number of gene-specific changepoints for the proposed coupled TV-DBN model applied to the *Saccharomyces cerevisiae* data.** We assume the segmentation of the gene expression time series to be unknown. The gene-specific changepoints can be inferred for different hyperparameters p of the negative binomial prior of the partition model (see Equation (51) in Appendix I for details). The table shows the posterior mean number of gene-specific changepoints, $\overline{K}_g$ ($g = 1, \ldots, 5$), for $\alpha_\lambda = \beta_\lambda = 1$ and different hyperparameters $p = 0$, $p = 0.025$, $p = 0.5$, $p = 0.1$, $p = 0.2$, and $p = 0.4$. The last two columns show the mean over all five genes, $\overline{K}$, and the mean AUC value, mean(AUC). All means were computed from five independent MCMC simulations on the *Saccharomyces cerevisiae* data. For both the conventional uncoupled and the proposed coupled TV-DBN models, the mean(AUC) values and the posterior mean numbers of changepoints, $\overline{K}$, can be computed for different (hyper-)hyperparameters $\alpha_\lambda = \beta_\lambda$ and p, and have been plotted against each other in Figure 11.

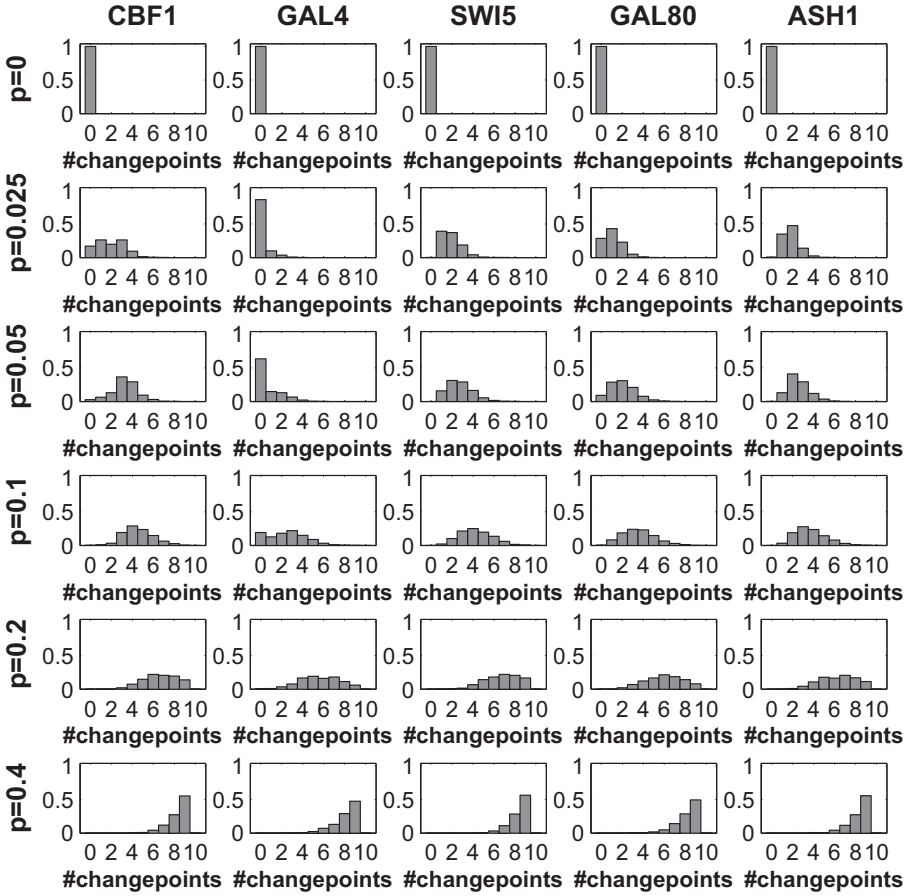

Figure 20: **Posterior distribution of the number of changepoints for the *Saccharomyces cerevisiae* data, obtained with the coupled TV-DBN model using the hyperhyperparameters** $\alpha_\lambda = \beta_\lambda = 1$**.** For the *Saccharomyces cerevisiae* data we assume the segmentation to be unknown. The gene-specific changepoints can be inferred with different hyperparameters, p, of the negative binomial prior of the partition model (see Equation (51) in Appendix I for details). The figure is arranged as a matrix, where each row corresponds to the hyperparameter p ($p = 0, 0.025, 0.5, 0.1, 0.2, 0.4$) of the negative binomial distribution and each column corresponds to one of the five genes in *Saccharomyces cerevisiae*. Each panel gives a histogram of the posterior distribution of the number of changepoints, which was obtained by combining the MCMC samples from five independent MCMC simulations on the *Saccharomyces cerevisiae* data. We note that the maximal cardinality of the changepoint sets, $\boldsymbol{\tau}_g$ ($g = 1, \ldots, 5$), was restricted to 9, i.e., $K_g = 10$ segments.

References

Ahmed, A. and 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.

Andrieu, C., Davy, M. and Doucet, A. (2003) Efficient particle filtering for jump Markov systems. application to time-varying autoregressions. *Signal Processing, IEEE Transactions on*, **51**, 1762–1770.

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

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

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. Springer, Singapore.

Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D. and Cosma1, M. P. (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

Carlin, B. and Louis, T. (2009) *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, 3rd edition.

Davis, J. and Goadrich, M. (2006) The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd international conference on Machine Learning*, pp. 233–240. ACM, New York, NY, USA.

Echtermeyer, C., Smulders, T. V. and Smith, V. A. (2009) Causal pattern recovery from neural spike train data using the snap shot score. *Journal of Computational Neuroscience*.

Edwards, K. D., Anderson, P. E., Hall, A., Salathia, N. S., Locke, J. C., Lynn, J. R., Straume, M., Smith, J. Q. and Millar, A. J. (2006) Flowering locus C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. *The Plant Cell*, **18**, 639–650.

Fearnhead, P. (2006) Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, **16**, 203–213.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004) *Bayesian Data Analysis*. Chapman and Hall/CRC, London, 2nd edition.

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

Giudici, P. and Castelo, R. (2003) Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, **50**, 127–158.

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

Grzegorczyk, M. and Husmeier, D. (2009) Non-stationary continuous dynamic Bayesian networks. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I. and Culotta, A. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pp. 682–690.

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

Grzegorczyk, M., Husmeier, D., Edwards, K., Ghazal, P. and Millar, A. (2008) Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. *Bioinformatics*, **24**, 2071–2078.

Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.

Husmeier, D., Dondelinger, F. and Lèbre, S. (2010) Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks. In Lafferty, J. e. a. (ed.), *Proceedings of the twenty-fourth annual conference on Neural Information Processing Systems (NIPS)*, volume 23, pp. 901–909. Curran Associates.

Kolar, M., Song, L. and Xing, E. (2009) Sparsistent learning of varying-coefficient models with structural changes. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I. and Culotta, A. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pp. 1006–1014.

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

Lim, W., Wang, K., Lefebvre, C. and Califano, A. (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**, i282–i288.

Lindley, D. (1962) Discussion on the article by Stein. *Journal of the Royal Society, Series B*, **24**, 265–296.

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

Mockler, T., Michael, T., Priest, H., Shen, R., Sullivan, C., Givan, S., McEntee, C., Kay, S. and Chory, J. (2007) The diurnal project: Diurnal and circadian expression profiling, model-based pattern matching and promoter analysis. *Cold Spring Harbor Symposia on Quantitative Biology*, **72**, 353–363.

Ong, I., Glasner, J. and Page, D. (2002) Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics*, **18**, S241–S248.

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

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

Robinson, J. W. and Hartemink, A. J. (2009) Non-stationary dynamic Bayesian networks. In Koller, D., Schuurmans, D., Bengio, Y. and Bottou, L. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pp. 1369–1376. Morgan Kaufmann Publishers.

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

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

Salome, P. and McClung, C. (2004) The Arabidopsis thaliana clock. *Journal of Biological Rhythms*, **19**, 425–435.

Smith, V. A., Jarvis, E. D. and Hartemink, A. J. (2002) Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, **18**, S216–S224. (ISMB02 special issue).

Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J. and Jarvi, E. D. (2006) Computational inference of neural information flow networks. *PLoS Computational Biology*, **2**, 1436–1449.

Song, L., Kolar, M. and Xing, E. (2009) Time-varying dynamic Bayesian networks. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I. and Culotta, A. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 22, pp. 1732–1740.

Stein, C. (1955) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, pp. 197–206. Berkeley University Press.

Talih, M. and 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**, 321–341.

Wang, S., Cui, L., Cheng, S., Zhai, S., Yeary, M. and Wu, Q. (2011) Noise adaptive LDPC decoding using particle filtering. *Communications, IEEE Transactions on*, **59**, 913–916.

Xuan, X. and Murphy, K. (2007) Modeling changing dependency structure in multivariate time series. In Ghahramani, Z. (ed.), *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007)*, pp. 1055–1062. Omnipress.

Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques*, pp. 233–243. Elsevier.

Zhao, W., Serpedin, E. and Dougherty, E. R. (2006) Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, **22**, 2129–2135.