Bayesian regularization of non-homogeneous dynamic Bayesian networks by globally coupling interaction parameters

Marco Grzegorczyk
Department of Statistics
TU Dortmund University
44221 Dortmund, Germany
grzegorczyk@statistik.tu-dortmund.de

Dirk Husmeier
School of Mathematics and Statistics,
University of Glasgow
Glasgow G12 8QW, UK
dirk.husmeier@glasgow.ac.uk

Abstract

To relax the homogeneity assumption of classical dynamic Bayesian networks (DBNs), various recent studies have combined DBNs with multiple changepoint processes. The underlying assumption is that the parameters associated with time series segments delimited by multiple changepoints are *a priori* independent. Under weak regularity conditions, the parameters can be integrated out in the likelihood, leading to a closed-form expression of the marginal likelihood. However, the assumption of prior independence is unrealistic in many real-world applications, where the segment-specific regulatory relationships among the interdependent quantities tend to undergo gradual evolutionary adaptations. We therefore propose a Bayesian coupling scheme to introduce systematic information sharing among the segment-specific interaction parameters. We investigate the effect this model improvement has on the network reconstruction accuracy in a reverse engineering context, where the objective is to learn the structure of a gene regulatory network from temporal gene expression profiles.

1 Introduction

There is considerable interest in structure learning of dynamic Bayesian networks (DBNs), with a variety of applications in computational systems biology. However, the standard assumption underlying DBNs

– that time-series have been generated from a homogeneous Markov process – is too restrictive in many applications 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), and Xuan and Murphy (2007)), relaxing this restriction in DBNs is a more recent research topic (Ahmed and Xing (2009), Kolar *et al.* (2009), Lèbre *et al.* (2010), Robinson and Hartemink (2010), and Grzegorczyk and Husmeier (2011)). Various authors have proposed relaxing the homogeneity assumption by complementing the traditional homogeneous DBN with a Bayesian multiple changepoint process (Lèbre *et al.* (2010), Robinson and Hartemink (2010), and Grzegorczyk and Husmeier (2011)). 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 Lèbre *et al.* (2010) and Robinson and Hartemink (2010), or with dynamic programming (Fearnhead, 2006), as in Grzegorczyk and Husmeier (2011).

In many real-word applications, the assumption of parameter independence is questionable, though. 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 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 adaptation processes. Given a regulatory network at some time interval in an organism's life cycle, it is unrealistic to assume that at the adjacent time intervals, nature has reinvented different regulatory circuits from scratch. Instead, we would assume that the knowledge of the interaction strengths at other time intervals will improve the inference of the interaction strengths associated with the given time interval, especially for sparse data. In what follows, we will describe how this idea can be implemented in the model, and which adaptations are required for the inference scheme.

There are various articles from the signal processing community that are related to our work. Our hierarchical Bayesian model structure is similar to the one proposed in Punskaya *et al.* (2002). However, conditionally on the hyperparameters, the regression parameters in Punskaya *et al.* (2002) are independent, and there is no explicit conditional dependence relation of the form we propose below. Like the model in Punskaya *et al.* (2002), our model is based on a switching piecewise homogeneous autoregressive process, whereas the models in Andrieu *et al.* (2003), Moulines *et al.* (2005), and Wang *et al.* (2011) are based on continuously time varying autoregressive processes. Like our paper, Moulines *et al.* (2005) and Wang *et al.* (2011) introduce information sharing between consecutive regression parameter vectors; this is only achieved indirectly in Andrieu *et al.* (2003) via a nonlinear transformation into the space of complex-valued poles. Moulines *et al.* (2005) is a theoretical non-Bayesian paper on error bounds under a Lipschitz condition. A closer relative to our paper is the method of Wang *et al.* (2011), whose objective is on-line parameter estimation via particle filtering, with applications e.g. in tracking. This is a different scenario from most systems biology applications, where an interaction structure is typically learnt off-line after completion of the experiments. Unlike Wang *et al.* (2011), our work thus follows other applications of DBNs in systems biology (Lèbre *et al.* (2010), Robinson and Hartemink (2010), and Grzegorczyk and Husmeier (2011)) and aims to infer the model structure by marginalizing out the parameters in closed form. To paraphrase this: while inference in Wang *et al.* (2011) is based on filtering, inference in our work is based on smoothing.

2 Bayesian linear regression

Consider a simple linear regression

$$f(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}, \qquad y = f(\mathbf{x}) + \varepsilon \qquad (1)$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is a vector of (interaction) parameters, f is the function value, y is

the observed target variable, and ε is additive Gaussian iid noise: $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. 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

$$P(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{X}^{\mathsf{T}}\mathbf{w}, \sigma_n^2\mathbf{I}) \qquad (2)$$

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

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

With Bayes' rule,

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = P(\mathbf{y}|\mathbf{X}, \mathbf{w})P(\mathbf{w})/P(\mathbf{y}|\mathbf{X}) \qquad (4)$$

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

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

where $\mathbf{m}_N = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\mathbf{m}_0 + \sigma_n^{-2}\mathbf{X}\mathbf{y})$ and $\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \sigma_n^{-2}\mathbf{X}\mathbf{X}^{\mathsf{T}}$. Let us now assume that we have a set of changepoints $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{K-1}\}$ with $1 \leq \tau_j \leq T-1$ that divide the data into K subsets:

$$\mathcal{D}_h = \{(\mathbf{x}_t, y_t), t = \tau_{h-1}, \ldots, \tau_h - 1\} \qquad (6)$$

All subsets are modelled with the linear model of (1), but with different parameter vectors $\mathbf{w}_h$ ($h = 1, \ldots, K$). Introducing the subsequent definitions $\mathbf{y}_h = (y_{\tau_{h-1}}, \ldots, y_{\tau_h-1})^{\mathsf{T}}$, and $\mathbf{X}_h = (\mathbf{x}_{\tau_{h-1}}, \ldots, \mathbf{x}_{\tau_h-1})$, and imposing the prior of (3) onto each $\mathbf{w}_h$, we get for the posterior distributions:

$$P(\mathbf{w}_h|\mathbf{y}_h, \mathbf{X}_h, \mathbf{m}_0) = \mathcal{N}(\mathbf{m}_N[h], \boldsymbol{\Sigma}_N[h]) \qquad (7)$$

where $\mathbf{m}_N[h] = \boldsymbol{\Sigma}_N[h](\boldsymbol{\Sigma}_0^{-1}\mathbf{m}_0 + \sigma_n^{-2}\mathbf{X}_h\mathbf{y}_h)$ and $\boldsymbol{\Sigma}_N^{-1}[h] = \boldsymbol{\Sigma}_0^{-1} + \sigma_n^{-2}\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}}$. For a fixed prior (3), e.g. $\mathbf{m}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \mathbf{I}$, where $\mathbf{I}$ is the unit matrix, the parameter vectors $\mathbf{w}_h$ are conditionally independent. To introduce information sharing among the segments, we can add an extra layer to the Bayesian hierarchy and turn $\mathbf{m}_0$ into a random vector, which is given a conjugate Gaussian prior distribution with mean vector $\mathbf{m}_{\dagger}$ and covariance matrix $\boldsymbol{\Sigma}_{\dagger}$, $P(\mathbf{m}_0|\mathbf{m}_{\dagger}, \boldsymbol{\Sigma}_{\dagger}) = \mathcal{N}(\mathbf{m}_{\dagger}, \boldsymbol{\Sigma}_{\dagger})$; see e.g. Section 3.6 in Gelman *et al.* (2004). Sampling of the parameters and hyperparameters from the posterior distribution follows a Gibbs sampling strategy. Given $\mathbf{m}_0$, we can sample the parameter vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$ from (7). Given $\{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$, the sufficient statistics $\mathbf{m}_{\star} = \boldsymbol{\Sigma}_{\star}(\boldsymbol{\Sigma}_{\dagger}^{-1}\mathbf{m}_{\dagger} + \boldsymbol{\Sigma}_0^{-1}[\sum_{h=1}^K \mathbf{w}_h])$, and $\boldsymbol{\Sigma}_{\star} = (\boldsymbol{\Sigma}_{\dagger}^{-1} + K\boldsymbol{\Sigma}_0^{-1})^{-1}$ can be computed, and $\mathbf{m}_0$ can be re-sampled from its posterior distribution

$$P(\mathbf{m}_0|\{\mathbf{w}_h\}_{h=1,\ldots,K}) = \mathcal{N}(\mathbf{m}_{\star}, \boldsymbol{\Sigma}_{\star}) \qquad (8)$$

3 Application to dynamic Bayesian networks

3.1 Fixed changepoints

We now generalize this coupling scheme for the interaction parameter prior distributions to non-homogeneous dynamic Bayesian networks (NH-DBNs) along the lines proposed in Lèbre *et al.* (2010). We restrict our NH-DBN to first-order Markov dynamics, noting that a generalization to higher order Markov dependencies, as included in Punskaya *et al.* (2002), is straightforward. Consider a set of N nodes $g \in \{1, \ldots, N\}$ in a network $\mathcal{M} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N\}$, where $\boldsymbol{\pi}_g$ denotes the parents of node g, that is the set of nodes with a directed edge pointing to g. We follow Grzegorczyk and Husmeier (2011) and assume that the regulatory network structure $\mathcal{M}$ is fixed over time. 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 our real-world applications (see Subsections 4.2 and 4.3), for which developmental (morphogenetical) changes can be excluded. Let $y_{g,t}$ denote the realization of the random variable associated with node g at time $t \in \{1, \ldots, T\}$, and let $\mathbf{x}_{\pi_g,t}$ denote the vector of realizations of the random variables associated with the parents of node g, π_g, at the previous time point, $(t-1)$, and including a constant element equal to 1 (for the bias or intercept). We consider N sets of $(K_g - 1)$ node-specific changepoints $\boldsymbol{\tau}_g = \{\tau_{g,h}\}_{1 \leq h \leq (K_g-1)}$, $1 \leq g \leq N$, which for now we assume to be fixed, with $T_{g,h} = \tau_{g,(h+1)} - \tau_{g,h}$. We define $\mathbf{y}_{g,h} = (y_{g,(\tau_{g,h}+1)}, \ldots, y_{g,\tau_{g,(h+1)}})^{\mathsf{T}}$, and $\mathbf{X}_{\pi_g,h} = (\mathbf{x}_{\pi_g,(\tau_{g,h}+1)}, \ldots, \mathbf{x}_{\pi_g,\tau_{g,(h+1)}})$ and apply the linear Gaussian regression model defined in (1-2):

$$P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \mathbf{w}_{g,h}, \sigma_{g,h}^2) = \mathcal{N}(\mathbf{X}_{\pi_g,h}^{\mathsf{T}}\mathbf{w}_{g,h}, \sigma_{g,h}^2\mathbf{I})$$

For the prior on $\mathbf{w}_{g,h}$ we use:

$$P(\mathbf{w}_{g,h}|\mathbf{m}_g, \sigma_{g,h}^2, \delta_g) = \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{m}_g, \delta_g\sigma_{g,h}^2\mathbf{C}_{g,h}) \tag{9}$$

where δ_g can be interpreted as a gene-specific "signal-to-noise" hyperparameter. Unlike other authors (Andrieu and Doucet (1999), Punskaya *et al.* (2002), and Lèbre *et al.* (2010)), we do not fix $\mathbf{m}_g$ in equation (9), but leave this hyperparameter variable, with its own prior distribution (a hyperprior)

$$P(\mathbf{m}_g|\mathbf{m}_\dagger, \boldsymbol{\Sigma}_\dagger) = \mathcal{N}(\mathbf{m}_\dagger, \boldsymbol{\Sigma}_\dagger) \tag{10}$$

with mean vector $\mathbf{m}_\dagger$ and covariance matrix $\boldsymbol{\Sigma}_\dagger$. This follows exactly the principle illustrated for the Bayesian linear regression model in Section 2. Note that when $\mathbf{m}_g$ is fixed, the $\mathbf{w}_{g,h}$'s are conditionally independent, or d-separated in the parlance of probabilistic graphical models. Hence, there is no information coupling between them. When $\mathbf{m}_g$ is flexible,

d-separation is lost, and the $\mathbf{w}_{g,h}$'s become dependent or "coupled", as a consequence of the marginalization over $\mathbf{m}_g$. We therefore refer to the proposed model, which provides an essential regularization effect, as the "coupled" model.

For the posterior distribution we get, in direct adaptation of (5):

$$P(\mathbf{w}_{g,h}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g,h}, \sigma_{g,h}^2, \delta_g, \mathbf{m}_g) = \mathcal{N}(\mathbf{m}_{g,h}^\star, \sigma_{g,h}^2\boldsymbol{\Sigma}_{g,h}^\star) \tag{11}$$

where $\mathbf{m}_{g,h}^\star = \boldsymbol{\Sigma}_{g,h}^\star([\delta_g\mathbf{C}_{g,h}]^{-1}\mathbf{m}_g + \mathbf{X}_{\pi_g,h}\mathbf{y}_{g,h})$ and $\boldsymbol{\Sigma}_{g,h}^\star = \left([\delta_g\mathbf{C}_{g,h}]^{-1} + \mathbf{X}_{\pi_g,h}\mathbf{X}_{\pi_g,h}^{\mathsf{T}}\right)^{-1}$. 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):

$$P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \sigma_{g,h}^2, \delta_g, \mathbf{m}_g) \tag{12}$$
$$= \int P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \sigma_{g,h}^2, \mathbf{w}_{g,h})P(\mathbf{w}_{g,h}|\sigma_{g,h}^2, \delta_g, \mathbf{m}_g)d\mathbf{w}_{g,h}$$
$$= \mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h}, \sigma_{g,h}^2\tilde{\boldsymbol{\Sigma}}_{g,h})$$

where $\tilde{\boldsymbol{\Sigma}}_{g,h} = \mathbf{I} + \delta_g\mathbf{X}_{\pi_g,h}^{\mathsf{T}}\mathbf{C}_{g,h}\mathbf{X}_{\pi_g,h}$, and $\tilde{\mathbf{m}}_{g,h} = \mathbf{X}_{\pi_g,h}^{\mathsf{T}}\mathbf{m}_g$. Note that the application of the matrix inversion theorem (e.g. Bishop, Appendix C) gives:

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

So far, we have assumed that $\sigma_{g,h}$ and δ_g are fixed. We now relax this constraint and impose conjugate gamma priors on $\sigma_{g,h}^{-2}$ and δ_g^{-1}:

$$P(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma) = \mathrm{Gam}(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma) \tag{13}$$

$$P(\delta_g^{-1}|\alpha_\delta, \beta_\delta) = \mathrm{Gam}(\delta_g^{-1}|\alpha_\delta, \beta_\delta) \tag{14}$$

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-from solution; see e.g. Section 2.3.7 in Bishop (2006):

$$P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \delta_g) \tag{15}$$
$$= \int_0^\infty P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \sigma_{g,h}^2, \delta_g)P(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma)d\sigma_{g,h}^{-2}$$
$$= \int_0^\infty \mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h}, \sigma_{g,h}^2\tilde{\boldsymbol{\Sigma}}_{g,h})Gam(\sigma_{g,h}^{-2}|\nu/2, \nu/2)d\sigma_{g,h}^{-2}$$
$$= \frac{\Gamma(T_{g,h}/2 + \nu/2)\nu^{\nu/2}}{\Gamma(\nu/2)(\pi)^{T_{g,h}/2}|\tilde{\boldsymbol{\Sigma}}_{g,h}|^{1/2}}\left(\nu + \Delta_{g,h}^2\right)^{-(T_{g,h}+\nu)/2}$$

with the squared Mahalanobis distance

$$\Delta_{g,h}^2 = (\mathbf{y}_{g,h} - \tilde{\mathbf{m}}_{g,h})^{\mathsf{T}}\tilde{\boldsymbol{\Sigma}}_{g,h}^{-1}(\mathbf{y}_{g,h} - \tilde{\mathbf{m}}_{g,h}) \tag{16}$$

This is a multivariate Student t-distribution with ν degrees of freedom. For updating the noise variances $\sigma_{g,h}^2$ and the signal-to-noise hyperparameters δ_g with a Gibbs sampling scheme (see Section 3.3) note that

$$\delta_g^{-1}|(\mathbf{y}_{g,.}, \mathbf{w}_{g,.}, \boldsymbol{\sigma}_{g,.}^2, \mathbf{X}_{\pi_g,.}) \sim \mathrm{Gam}(\alpha_\delta + A_\delta, \beta_\delta + B_\delta) \tag{17}$$

with $B_\delta = \frac{1}{2}\sum_h \frac{1}{\sigma_{g,h}^2}[\mathbf{w}_{g,h} - \mathbf{m}_g]^\mathsf{T}\mathbf{C}_{g,h}^{-1}[\mathbf{w}_{g,h} - \mathbf{m}_g]$, and $A_\delta = \frac{K_g k_g}{2}$, where K_g is the number of segments for node g, k_g is the cardinality of the parent set, $\boldsymbol{\pi}_g$, and $\mathbf{y}_{g,.} := (\mathbf{y}_{g,1}^\mathsf{T}, \ldots, \mathbf{y}_{g,K_g}^\mathsf{T})^\mathsf{T}$, $\mathbf{w}_{g,.} := (\mathbf{w}_{g,1}^\mathsf{T}, \ldots, \mathbf{w}_{g,K_g}^\mathsf{T})^\mathsf{T}$, $\boldsymbol{\sigma}_{g,.}^2 := (\sigma_{g,1}^2, \ldots, \sigma_{g,K_g}^2)$, and $\mathbf{X}_{\pi_g,.} := (\mathbf{X}_{\pi_g,1}, \ldots, \mathbf{X}_{\pi_g,K_g})$; for a derivation see supplementary material.

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 use the marginal likelihood (12) instead of (9), i.e. to use a collapsed Gibbs sampler in which the interaction parameters $\mathbf{w}_{g,h}$ have been integrated out. From (12) and (13) we obtain (see supplementary material):

$$\sigma_{g,h}^{-2}|(\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g,h}, \delta_g) \sim \text{Gam}\left(\alpha_\sigma + A_\sigma, \beta_\sigma + B_\sigma\right) \tag{18}$$

where $B_\sigma = \frac{\Delta_{g,h}^2}{2}$, and $A_\sigma = \frac{T_{g,h}}{2}$. $\Delta_{g,h}^2$ was defined in (16) and depends on the hyperparameter δ_g via (12). The previous discussions 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 study we obtained better results when using a common variance shared by all segments: $\sigma_{g,h}^2 = \sigma_g^2 \,\forall\, h$. Equation (18) will then change as follows:

$$\sigma_g^{-2}|(\mathbf{y}_{g,.}, \mathbf{X}_{\pi_g,.}, \delta_g) \sim \text{Gam}\left(\alpha_\sigma + A_\sigma^\star, \beta_\sigma + B_\sigma^\star\right) \tag{19}$$

where $A_\sigma^\star = \frac{T_g}{2}$, $B_\sigma^\star = \sum_{h=1}^{K_g} \frac{\Delta_{g,h}^2}{2}$, $T_g = \sum_{h=1}^{K_g} T_{g,h}$, and $\Delta_{g,h}^2$ was defined in (16) and depends on δ_g via (12). A comparison between (18) and (19) leads to the intuitive result that we can 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 (11) and (15). Denote by $\tilde{\mathbf{m}}_{g,.}$ the accumulated vector $(\tilde{\mathbf{m}}_{g,1}^\mathsf{T}, \ldots, \tilde{\mathbf{m}}_{g,K_g}^\mathsf{T})^\mathsf{T}$, and denote by $\tilde{\boldsymbol{\Sigma}}_{g,.}$ a matrix with block structure, in which the matrices $\tilde{\boldsymbol{\Sigma}}_{g,h}$ are arranged along the diagonal, and all other entries are 0. In modification of equations (11) and (15) we now get:

$$P(\mathbf{w}_{g,h}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g,h}, \sigma_g) = \mathcal{N}(\mathbf{m}_{g,h}^\star, \sigma_g^2 \boldsymbol{\Sigma}_{g,h}^\star) \tag{20}$$

$$P(\mathbf{y}_{g,.}|\mathbf{X}_{\pi_g,.}, \delta_g) = \frac{\Gamma(T_g/2 + \nu/2)\nu^{\nu/2}}{\Gamma(\nu/2)(\pi)^{T_g/2}|\tilde{\boldsymbol{\Sigma}}_{g,.}|^{1/2}}\left(\nu + \Delta_g^2\right)^{-\frac{T_g+\nu}{2}} \tag{21}$$

where with definition (16) and by exploiting the block structure of $\tilde{\boldsymbol{\Sigma}}_{g,.}$ we get:

$$\Delta_g^2 = (\mathbf{y}_{g,.} - \tilde{\mathbf{m}}_{g,.})^\mathsf{T}\tilde{\boldsymbol{\Sigma}}_{g,.}^{-1}(\mathbf{y}_{g,.} - \tilde{\mathbf{m}}_{g,.}) = \sum_h \Delta_{g,h}^2 \tag{22}$$

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 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. For space restrictions, the mathematical details have been relegated to the supplementary material.

3.3 Inference

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

$$P(\mathcal{M}|\mathcal{D}, \{\boldsymbol{\tau}_g\}, \boldsymbol{\delta}) \propto P(\mathcal{M}) \prod_g \prod_h P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \delta_g) \tag{23}$$

$$P(\{\boldsymbol{\tau}_g\}|\mathcal{D}, \boldsymbol{\delta}, \mathcal{M}) \propto \prod_g P(\boldsymbol{\tau}_g) \prod_h P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \delta_g) \tag{24}$$

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 $P(\mathbf{y}_{g,h}|\mathbf{X}_{\pi_g,h}, \delta_g)$, which are given by (21), have been obtained by marginalizing over $\mathbf{w}_{g,h}$ and σ_g^2 ("collapsed" Gibbs steps). We sample from (23) with the improved structure MCMC scheme proposed in Grzegorczyk and Husmeier (2011), and from (24) with the RJMCMC scheme of Lèbre *et al.* (2010). To sample the signal-to-noise hyperparameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ from the posterior distribution, we need to resort to an uncollapsed Gibbs step:

$$P(\boldsymbol{\delta}|\mathcal{D}, \{\boldsymbol{\tau}_g\}, \mathcal{M}) = \prod_g P(\delta_g^{-1}|\mathbf{y}_{g,.}, \mathbf{w}_{g,.}, \boldsymbol{\sigma}_{g,.}^2, \mathbf{X}_{\pi_g,.}) \tag{25}$$

where $P(\delta_g^{-1}|\mathbf{y}_{g,.}, \mathbf{w}_{g,.}, \boldsymbol{\sigma}_{g,.}^2, \mathbf{X}_{\pi_g,.})$ is given by (17). "Uncollapsing" requires the interaction and noise parameters to be sampled from the corresponding posterior distributions; these are given by (19-20). In the proposed coupled NH-DBN these MCMC steps are followed by "uncollapsed" Gibbs sampling steps: For each gene g the interaction parameters $\mathbf{w}_{g,1}, \ldots, \mathbf{w}_{g,K_g}$ are sampled from (11). Conditional on the sampled parameter vectors, the hyperparameter $\mathbf{m}_g$ in (9) can be re-sampled from the posterior distribution

$$\mathbf{m}_g|(\mathbf{w}_{g,1}, \ldots, \mathbf{w}_{g,K_g}) \sim \mathcal{N}(\mathbf{m}_{\star,g}, \boldsymbol{\Sigma}_{\star,g}) \qquad (26)$$

which depends on the sufficient statistics:

$$\boldsymbol{\Sigma}_{\star,g} := (\boldsymbol{\Sigma}_\dagger^{-1} + K_g \boldsymbol{\Sigma}_0^{-1})^{-1} \qquad (27)$$

$$\mathbf{m}_{\star,g} := \boldsymbol{\Sigma}_{\star,g}(\boldsymbol{\Sigma}_\dagger^{-1}\mathbf{m}_\dagger + \boldsymbol{\Sigma}_0^{-1}[\sum_{h=1}^{K_g} \mathbf{w}_{g,h}]) \quad (28)$$

where $\boldsymbol{\Sigma}_0 := \delta_g \sigma_{g,h}^2 \mathbf{C}_{g,h}$ (see Section 3.6 in Gelman *et al.* (2004)).

4 Data

4.1 Simulated data from the RAF pathway

For the RAF pathway, shown in Figure 1a), we generate non-homogeneous dynamic expression data. We assume that we have a time series with four segments $h = 1, \ldots, 4$, which consist of 10 observations each, and that the network interaction parameters vary from segment to segment. We assume that there is a global parameter vector $\mathbf{w}_{g,\star}$ with amplitude (Euclidean norm) 1, $|\mathbf{w}_{g,\star}|_2 = 1$, for each interaction between a node, g, and its parent nodes in π_g, where the latter are defined by the graph in Figure 1a). Segment-specific parameter vectors $\mathbf{w}_{g,h}$ ($h = 1, \ldots, 4$) can then be obtained by adding iid random noise vectors $\tilde{\mathbf{w}}_{g,h}$ to the global vector $\mathbf{w}_{g,\star}$. The similarity between the four segment-specific parameter vectors depends on the amplitude ε of the random vectors $\tilde{\mathbf{w}}_{g,h}$. Re-normalization ensures that the segment-specific interaction parameters $\mathbf{w}_{g,h}$ have amplitude 1 independently of ε:

$$\mathbf{w}_{g,h} = \frac{\mathbf{w}_{g,\star} + \varepsilon\tilde{\mathbf{w}}_{g,h}}{|\mathbf{w}_{g,\star} + \varepsilon\tilde{\mathbf{w}}_{g,h}|_2} \qquad (29)$$

We sample the elements of the global parameter vectors $\mathbf{w}_{g,\star}$ ($g = 1, \ldots, N$) and the random noise vectors $\tilde{\mathbf{w}}_{g,h}$ ($g = 1, \ldots, N$; $h = 1, \ldots, 4$) from iid $\mathcal{N}(0,1)$ distributions and we normalize the vectors $\mathbf{w}_{g,\star}$ and $\tilde{\mathbf{w}}_{g,h}$ to Euclidean norm 1. Having computed all the interaction parameter vectors $\mathbf{w}_{g,h}$ from equation (29), the data can be generated straighforwardly: We sample observations for the first time point, $t = 1$, from iid $\mathcal{N}(0, 0.025)$ distributions, before we generate data for

40 subsequent time points. The complete data set $\mathcal{D}$ is then an 11-by-41 matrix, where for $t = 2, \ldots, 41$ the t-th observation of node g, $\mathcal{D}_{g,t}$, is given by:

$$\mathcal{D}_{g,t} = (1, \mathcal{D}_{\pi_g,t-1}^\mathsf{T})\mathbf{w}_{g,H(t)} + u_{g,t} \qquad (30)$$

where $\mathcal{D}_{\pi_g,t-1}$ is the vector of observations of the parent nodes of g at the previous time point $t - 1$, the function $H(.)$ indicates the segment ($H(t) = 1$ for $t = 2, \ldots, 11$, $H(t) = 2$ for $t = 12, \ldots, 21$, etc.), and the $u_{g,t}$ are iid $N(0, 0.025)$ distributed dynamic noise variables. We also include white observational noise with the objective to control the signal-to-noise ratio (SNR). Here, we add noise in a gene-wise manner: For each g we compute its standard deviation s_g and add iid $\mathcal{N}(0, (SNR)^{-1}s_g)$ distributed noise to each individual observation of g.

4.2 Circadian rhythms in *A. thaliana*

Four gene expression time series related to the study of circadian regulation in plants were measured in *Arabidopsis thaliana*. Arabidopsis 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 ($\tau \in 2, 4$) intervals. For an overview see Table 1 in the supplementary material. The data and the experimental protocols are available from Edwards *et al.* (2006), Grzegorczyk *et al.* (2008), and Mockler *et al.* (2007). We focus on nine genes that are involved in circadian regulation (LHY, TOC1, CCA1, ELF4, ELF3, GI, PRR9, PRR5, and PRR3) and we arrange the four individual time series successively, so as to obtain one single time series, where each of the four segments has been measured under constant light after different T_e-hour-light/T_e-hour-dark pre-experimental entrainment conditions.

4.3 Synthetic biology in *S. cerevisiae*

Cantone *et al.* (2009) synthetically designed a network of five genes in *Saccharomyces cerevisiae* (yeast), depicted in Figure 3a). The authors measured expression levels of these genes *in vivo* with quantitative real-time PCR at 37 time points over 8 hours. In about the middle of this time period, they changed the environment by switching the carbon source from galactose ("switch on") to glucose ("switch off"). We removed the two measurements that were taken during the washing steps, i.e while the glucose (galactose) medium was removed and the fresh new galactose (glucose) containing medium was added, before we re-arranged the two time series successively to one single time series. Since the first time point after the washing period of the "switch off" time series has then 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 also appropriately removed to

ensure that for all pairs of consecutive time points a proper conditional dependence relation is given. The merged time series was standardized via a log transformation and a subsequent mean standardization.

5 Simulation study

We want to compare the proposed coupled (regularized) NH-DBN with the conventional uncoupled (unregularized) NH-DBN akin to Lèbre *et al.* (2010). We assume that the gene-specific variances are shared by all segments: $\sigma^2_{g,h} = \sigma^2_g$. The prior distributions of the inverse variances σ^{-2}_g are assumed to be Gamma distributed (see (13)), and we set $\alpha_\sigma = \beta_\sigma = \nu/2$ with $\nu = 0.01$ ($g = 1, \ldots, N$). For the inverse Gamma distributed hyperparameters δ_g we follow Lèbre *et al.* (2010) and set $\alpha_\delta = 2$ and $\beta_\delta = 0.02$ in (14). The gene- and segment-specific interaction parameter vectors $\mathbf{w}_{g,h}$ are assumed to be multivariate Gaussian distributed according to (9), and we set $\mathbf{C}_{g,h} = \mathbf{I}$. In the conventional NH-DBN, $\mathbf{m}_g = \mathbf{0}$ is fixed, which yields, from equation (9): $\mathbf{w}_{g,h}|(\sigma_g, \delta_g) \sim \mathcal{N}(\mathbf{0}, \delta_g\sigma^2_g\mathbf{I})$. For the proposed coupled NH-DBN, $\mathbf{m}_g$ is flexible, with prior distribution $\mathcal{N}(\mathbf{m}_\dagger, \mathbf{\Sigma}_\dagger)$ – see equation (9) – where we set $\mathbf{m}_\dagger = \mathbf{0}$ and $\mathbf{\Sigma}_\dagger = \mathbf{I}$.
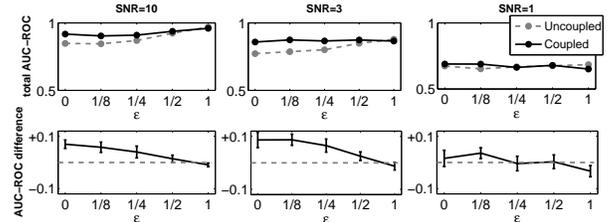
We aim to monitor the network reconstruction accuracy on a series of increasingly strong violations of the prior assumption inherent in (10). To this end, we generate synthetic data, as explained in Subsection 4.1, and we reverse-engineer the RAF pathway by sampling networks from the posterior distribution (23) for fixed changepoints. In the absence of a gold standard for the *A. thaliana* data, described in Subsection 4.2, our focus is on quantifying the strength of the information coupling between the time series segments and the influence this coupling has on the regulatory network reconstruction. For the *S. cerevisiae* data from Subsection 4.3 the true regulatory network is known (see Figure 3a)), and our focus is on monitoring the network reconstruction accuracy. We follow an unsupervised approach and assume that the changepoints segmenting the time series are unknown. To infer different segmentations we employ different hyperparameters of the point process prior on the changepoint sets; see supplementary material for details.

We pursue inference based on the partially collapsed Gibbs sampler, described in Subsection 3.3, and the output is a sample of networks from the posterior distribution. We applied standard convergence diagnostics, based on trace plots (Giudici and Castelo, 2003) and the potential scale reduction factor (Gelman and Rubin, 1992), and found that the PSRF's of all individual edges were below 1.1 for simulation lengths of 100,000 MCMC steps; see supplementary material. From the sampled networks we can obtain a ranking of the gene interactions based on their marginal posterior
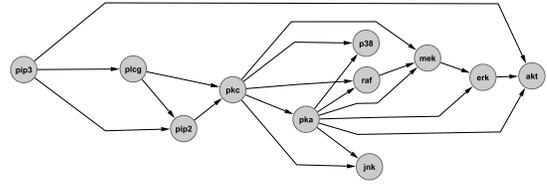
(a) The RAF network (Sachs *et al.*, 2005)

(b) Network reconstruction accuracy for the RAF network

Figure 1: **Reconstruction of the RAF pathway from simulated expression data. (a)** The RAF pathway, as reported in Sachs *et al.* (2005). Panel **(b)** monitors the network reconstruction accuracy in terms of AUC-ROC scores for the conventional uncoupled (dotted gray) and the proposed coupled (solid black lines) NH-DBN and demonstrates how the proposed regularization scheme is affected by increasing violations of the prior assumption inherent in equation (10). Simulated data were generated as described in Subsection 4.1. The global parameter vector with amplitude 1 was perturbed in a segment-wise manner by a random perturbation of amplitude ε (abscissa); see equation (29). The columns represent the SNR levels 10, 3, and 1. The top row shows the absolute values of the AUC-ROC scores, while the bottom row shows the differences between the proposed coupled and the conventional uncoupled NH-DBN. All simulations were repeated on 25 independent data instantiations, with error bars indicating two-sided 95% confidence intervals. A similar plot with AUC-PR scores is provided in the supplementary material.

probabilities. If the true network is known, this ranking can be employed to obtain the ROC and precision-recall (PR) curves (Davis and Goadrich (2006), and Prill *et al.* (2010)). These curves can be numerically integrated to get the areas under the curves (AUC) for both (AUC-ROC or AUC-PR, respectively) as a global measure of network reconstruction accuracy (with larger values indicating a better performance); see supplementary material for details.

6 Results

We first evaluated the proposed Bayesian regularization scheme on simulated data. We took the RAF network from Sachs *et al.* (2005), see Figure 1a), and generated synthetic non-homogeneous time series from a multiple changepoint linear regression model, as explained in Subsection 4.1. Our objective was to monitor the network reconstruction accuracy on a series of
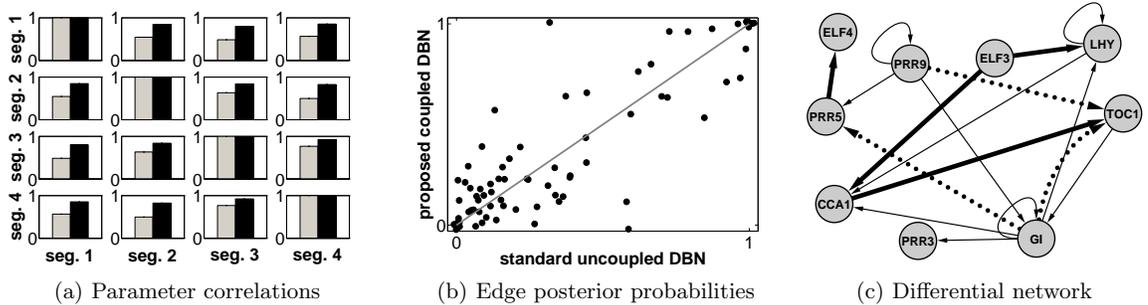
(a) Parameter correlations (b) Edge posterior probabilities (c) Differential network

Figure 2: **Inference for the _A. thaliana_ gene expression time series. a)** Histograms of the average similarities (correlations) of the interaction parameters, sampled from the posterior distribution with MCMC between four time series segments, indicated by the rows and columns. Details on the segmentation can be found in Subsection 4.2. The networks were sampled from the posterior distribution, see (23), with MCMC. Each panel contains a histogram that shows the average similarity of the interaction parameters among segments for the conventional uncoupled (grey) and the proposed coupled (black bars) NH-DBN; see main text for details on our similarity measure. **b)** Scatter plot of the marginal edge posterior probabilities inferred with the uncoupled (horizontal) and the coupled (vertical axis) NH-DBN. **c)** The (differential) network prediction that can be obtained when the threshold 0.75 is imposed on the edge posterior probabilities. Thin black edges indicate interactions that are inferred with both NH-DBNs. Three edges (dotted) are inferred with the conventional NH-DBN only while four edges (bold) are inferred with the coupled NH-DBN only.
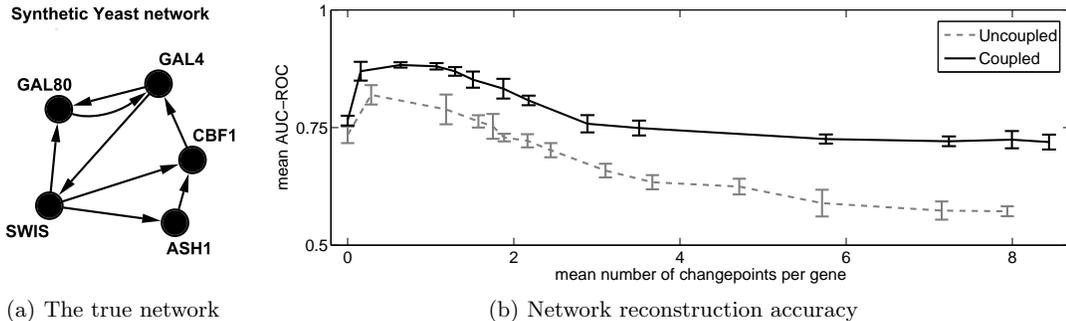

(a) The true network (b) Network reconstruction accuracy

Figure 3: **Gene network reconstruction accuracy for the _S. cerevisiae_ data.** Cantone _et al._ (2009) designed the network shown in panel a) and measured in vivo gene expression levels with RT-PCR. The graph shows the network reconstruction accuracy (ordinate) plotted against the mean number of changepoints per gene (abscissa) for the conventional uncoupled (dashed) and the proposed coupled NH-DBN (solid line). The network reconstruction accuracy is quantified in terms of mean AUC-ROC scores, averaged over 5 MCMC simulations, with the vertical bars indicating standard errors. A similar plot with AUC-PR scores is provided in the supplementary material.

increasingly strong violations of the prior assumption inherent in equation (10). The results are shown in Figure 1b). For the low signal-to-noise ratio (SNR=1) there is no significant difference between the models. However, owing to the high noise level, the network reconstruction accuracy is close to random expectation in that case (AUC-ROC = 0.5). For low (SNR=10) and moderate (SNR=3) noise levels, the proposed coupled NH-DBN increasingly outperforms the conventional NH-DBN as the amplitude of the perturbation ε of the parameter vectors decreases. In particular, for perturbations of $\varepsilon < 1/4$ the performance improvement of the proposed regularized NH-DBN over the conventional NH-DBN is clearly significant. We used the gene expression time series from _A. thaliana_, described in Subsection 4.2, to investigate which effect

the proposed Bayesian coupling scheme has on the inference of the interaction parameters. To focus on the relevant task, the regulatory network reconstruction, we kept the changepoints fixed at their known true values. We compared the correlations of the segment-specific interaction parameter vectors for the uncoupled and for the coupled NH-DBN. During the RJM-CMC simulation, we sampled for each segment h the interaction parameters from equation (11), agglomerated the interaction parameters for each h into a long vector $\mathbf{v}_h$, and computed as a similarity measure the correlation coefficient between pairwise different vectors $\mathbf{v}_{h_1}$ and $\mathbf{v}_{h_2}$ ($h_1 \neq h2$). The results are shown in Figure 2a) and suggest that the proposed Bayesian regularization scheme increases the average similarity between the interaction parameters from the four time

473

series. This is a shrinkage effect that one would expect from a Bayesian hierarchical model, in the sense of the well-known "Stein and Lindley effect" (Stein (1955), and Lindley (1962)), and it has the potential to improve the inference for time series segments that are fairly short, as we demonstrate below. Our results also indicate that the proposed Bayesian regularization scheme avoids a complete coupling, corresponding to a perfect correlation. This would be unrealistic, as the four time series segments were subject to different pre-entrainment conditions, which are known to influence the regulatory relationships (Johnson *et al.* (2003), and McClung (2006)). Figure 2b) shows a scatter plot of the marginal edge posterior probabilities from both NH-DBNs. Despite a certain correlation, there are several edges for which different posterior probabilities have been inferred. To more clearly demonstrate the effect of the proposed regularization scheme on the network reconstruction, Figure 2c) shows a network possessing only those edges whose posterior probability exceeds the threshold of 0.75 for at least one of the two NH-DBNs. It can be seen that the proposed Bayesian regularization scheme has a clear influence on the inferred structure. We queried the biological literature and found evidence for at least three of the four gene interactions that were inferred with the proposed coupled NH-DBN only (i.e. 75%): $CCA1 \rightarrow TOC1$ (Alabadi *et al.*, 2001) as well as $ELF3 \rightarrow CCA1$ and $ELF3 \rightarrow LHY$ (Kikis *et al.*, 2005). On the other hand, we only found evidence for one out of the three interactions that were solely predicted with the conventional NH-DBN (corresponding to 33%); this is the feedback loop $GI \leftrightarrow TOC1$, reported in (Locke *et al.*, 2005). We acknowledge that this evaluation is somewhat subjective and susceptible to a certain selection bias, which is the inevitable consequence of the absence of a gold-standard. For that reason, we also tested the methods on gene expression profiles from *S. cerevisiae*, described in Subsection 4.3. Here we do know the true regulatory network, shown in Figure 3a), so that we can objectively compare the network reconstruction accuracy of the two NH-DBNs. With both NH-DBNs we ran RJMCMC simulations for various hyperparameter values of the negative binomial prior on the number of changepoints (see supplementary material), and we sampled the networks, the number of changepoints and the location of changepoints with the RJMCMC scheme described in Section 3. Figure 3b) shows the average AUC-ROC scores plotted against the mean posterior mode[1] of the number of changepoints, $\overline{K}$. It is clearly seen that the proposed coupled NH-DBN yields a systematically better

network reconstruction accuracy than the conventional non-homogeneous NH-DBN. The best performance of the novel coupled NH-DBN is given 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-ROC scores. Interestingly, this decline is less pronounced for the proposed regularized NH-DBN model than for the conventional NH-DBN model, indicating increased robustness with respect to a variation of the prior assumptions on the time series segmentation.

7 Conclusion

Modelling non-homogeneous dynamic Bayesian networks (NH-DBNs) with a multiple changepoint process is popular due to the fact that conditional on the changepoints, the marginal likelihood can be computed in closed form. To our knowledge, all previous studies, including Lèbre *et al.* (2010), Robinson and Hartemink (2010), and Grzegorczyk and Husmeier (2011), compute the marginal likelihood under the assumption of parameter independence and the same independent parameter prior distributions for all time series segments. These approaches ignore the fact that many systems, e.g. regulatory networks and signalling pathways in the cell, adapt to changing internal and external conditions *gradually*. To allow for information sharing among separate time series segments we have proposed a novel regularized NH-DBN with a coupling mechanism in the sense that *a priori* the interaction parameters associated with separate time series segments are encouraged to be similar. Our empirical assessment on simulated data has revealed that the proposed method leads to an improvement in the network reconstruction accuracy. We have quantified the effect of the regularization for gene expression time series from *A. thaliana*. For time series from RT-PCR experiments in *S. cerevisiae*, we have demonstrated that the novel NH-DBN also yields a better network reconstruction accuracy than the conventional uncoupled NH-DBN, and that it leads to increased robustness with respect to a variation of the prior assumptions about the temporal heterogeneity.

---

[1] For each gene, the mean of the posterior distribution of the number of changepoints was determined, and these values were averaged over all genes.

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.

Alabadi, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Mas, P. and Kay, S. A. (2001) Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science*, **293**, 880–883.

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.

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.

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.

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. (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.

Johnson, C., Elliott, J. and Foster, R. (2003) Entrainment of circadian programs. *Chronobiology International*, **20**, 741–774.

Kikis, E., Khanna, R. and Quail, P. (2005) ELF4 is a phytochrome-regulated component of a negative-feedback loop involving the central oscillator components CCA1 and LHY. *Plant J.*, **44**, 300–313.

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**.

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

Locke, J., Southern, M., Kozma-Bognar, L., Hibberd, V., Brown, P., Turner, M. and Millar, A. (2005) Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular Systems Biology*, **1**, (online).

McClung, C. R. (2006) Plant circadian rhythms. *Plant Cell*, **18**, 792–803.

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.

Moulines, E., Priouret, P. and Roueff, F. (2005) On recursive estimation for time varying autoregressive processes. *Annals of Statistics*, **33**, 2610–2654.

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. (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.

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.