Heterogeneous Continuous Dynamic Bayesian Networks with Flexible Structure and Inter-Time Segment Information Sharing

Frank Dondelinger FRANKD@BIOSS.AC.UK
Biomathematics and Statistics Scotland, JCMB, EH9 3JZ, Edinburgh, UK
Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh

Sophie Lèbre SOPHIE.LEBRE@LSIIT-CNRS.UNISTRA.FR
Université de Strasbourg, LSIIT - UMR 7005, 67412 Illkirch, France

Dirk Husmeier DIRK@BIOSS.AC.UK
Biomathematics and Statistics Scotland, JCMB, EH9 3JZ, Edinburgh, UK

Abstract

Classical dynamic Bayesian networks (DBNs) are based on the homogeneous Markov assumption and cannot deal with heterogeneity and non-stationarity in temporal processes. Various approaches to relax the homogeneity assumption have recently been proposed. The present paper aims to improve the shortcomings of three recent versions of heterogeneous DBNs along the following lines: (i) avoiding the need for data discretization, (ii) increasing the flexibility over a time-invariant network structure, (iii) avoiding over-flexibility and overfitting by introducing a regularization scheme based in inter-time segment information sharing. The improved method is evaluated on synthetic data and compared with alternative published methods on gene expression time series from *Drosophila melanogaster*.

1. Introduction

There has recently been considerable interest in structure learning of dynamic Bayesian networks (DBNs), with a variety of applications in signal processing and computational biology; see e.g. (Robinson & Hartemink, 2009) and (Grzegorczyk & Husmeier, 2009). The standard assumption underlying DBNs is that time-series have been generated from a homogeneous Markov process. This assumption is too restrictive in many applications and can potentially lead to erroneous conclusions. In the recent past, various research efforts have therefore addressed this issue and proposed more flexible models.

(Robinson & Hartemink, 2009) proposed a discrete heterogeneous DBN, which allows for different structures in different segments of the time series, with a regularization term penalizing differences among the structures. (Grzegorczyk & Husmeier, 2009) proposed a continuous heterogeneous DBN, in which only the parameters are allowed to vary, with a common network structure providing information sharing among the time series segments. (Lèbre, 2007) proposed an alternative continuous heterogeneous DBN, which is more flexible in that it allows the network structure to vary among the segments. The model proposed in (Ahmed & Xing, 2009) and (Kolar et al., 2009) can be regarded as a heterogeneous DBN where inference is based on sparse L1-regularized regression (LASSO) of the interaction parameters, and a second L1 regularization term penalizes differences between networks associated with different segments.

Parameter estimation in (Ahmed & Xing, 2009) and (Kolar et al., 2009) is based on penalized maximum likelihood for fixed regularization parameters, which are optimized using BIC or cross-validation. In the present paper, we follow (Robinson & Hartemink, 2009), (Grzegorczyk & Husmeier, 2009) and (Lèbre, 2007) to infer the network structure, the interaction parameters, as well as the number and location of changepoints in a Bayesian framework by sampling

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

them from the posterior distribution with RJMCMC (Green, 1995). The objective of our paper is to propose a model that addresses the principled shortcomings of the three Bayesian methods mentioned above. Unlike (Robinson & Hartemink, 2009), our model is continuous and therefore avoids the information loss inherent in a discretization of the data. Unlike (Grzegorczyk & Husmeier, 2009), our model allows the network structure to change among segments, leading to greater model flexibility. As an improvement on (Lèbre, 2007), our model introduces information sharing among time series segments, which provides an essential regularization effect.

2. Background

This paragraph summarizes briefly the heterogeneous DBN proposed in (Lèbre, 2007). The model is based on the first-order Markov assumption. This assumption is not critical, though, and a generalization to higher orders is straightforward. The value that a node in the graph takes on at time t is determined by the values that the node's parents take on at the previous time point, $t-1$, as well as the time series segment. More specifically, the conditional probability of the observation associated with a node at a given time point is a conditional Gaussian distribution, where the conditional mean is a linear weighted sum of the parent values at the previous time point, and the weights themselves depend on the time series segment. The latter dependence adds extra flexibility to the model and thereby relaxes the homogeneity assumption. The interaction weights, the variance parameters, the number of potential parents, the location of changepoints demarcating the time series segments, and the number of changepoints are given (conjugate) prior distributions in a hierarchical Bayesian model. For inference, all these quantities are sampled from the posterior distribution with RJMCMC. Note that a complete specification of all node-parent configurations determines the structure of a regulatory network: each node receives incoming directed edges from each node in its parent set. In what follows, we will refer to nodes as genes and to the network as a gene regulatory network. The method is not restricted to molecular systems biology, though.

2.1. Model

Multiple changepoints. Let p be the number of observed genes, whose expression values $y = \{y_i(t)\}_{1\leq i\leq p, 1\leq t\leq N}$ are measured at N time points. $\mathcal{M}$ represents a directed graph, i.e. the network defined by a set of directed edges among the p genes. $\mathcal{M}_i$ is the subnetwork associated with target gene i, determined by the set of its parents (nodes with a directed edge feeding into gene i). The regulatory relationships among the genes, defined by $\mathcal{M}$, may vary across time, which we model with a multiple changepoint process. For each target gene i, an unknown number k_i of changepoints define k_i+1 non-overlapping segments. Segment $h = 1, .., k_i+1$ starts at changepoint ξ_i^{h-1} and stops before ξ_i^h, where $\xi_i = (\xi_i^0, ..., \xi_i^{h-1}, \xi_i^h, ..., \xi_i^{k_i+1})$ with $\xi_i^{h-1} < \xi_i^h$. To delimit the bounds, $\xi_i^0 = 2$ and $\xi_i^{k_i+1} = N+1$. Thus vector ξ_i has length $|\xi_i| = k_i + 2$. The set of changepoints is denoted by $\xi = \{\xi_i\}_{1\leq i\leq p}$. This changepoint process induces a partition of the time series, $y_i^h = (y_i(t))_{\xi_i^{h-1}\leq t<\xi_i^h}$, with different structures $\mathcal{M}_i^h$ associated with the different segments $h \in \{1, \ldots, k_i+1\}$. Identifiability is satisfied by ordering the changepoints based on their position in the time series.

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

$$Y_i(t) \ = \ a_{i0}^h + \sum\nolimits_{j\in\mathcal{M}_i^h} a_{ij}^h \ Y_j(t-1) \ + \varepsilon_i(t) \qquad (1)$$

where the noise $\varepsilon_i(t)$ is assumed to be Gaussian with mean 0 and variance $(\sigma_i^h)^2$, $\varepsilon_i(t) \sim N(0, (\sigma_i^h)^2)$.

2.2. Prior

The k_i+1 segments are delimited by k_i changepoints, where k_i is distributed a priori as a truncated Poisson random variable with mean λ and maximum $\overline{k} = N-2$: $P(k_i|\lambda) \propto \frac{\lambda^{k_i}}{k_i!}\mathbb{1}_{\{k_i\leq\overline{k}\}}$. [1]

Conditional on k_i changepoints, the changepoint positions vector $\xi_i = (\xi_i^0, \xi_i^1, ..., \xi_i^{k_i+1})$ takes non-overlapping integer values, which we take to be uniformly distributed a priori. There are $(N-2)$ possible positions for the k_i changepoints, thus vector ξ_i has prior density $P(\xi_i|k_i) = 1/\binom{N-2}{k_i}$. For all genes i and all segments h, the number s_i^h of parents for node i

[1] A restrictive Poisson prior encourages sparsity of the network, and is therefore comparable to a sparse exponential prior, or an approach based on the LASSO.

follows a truncated Poisson distribution with mean Λ and maximum $\overline{s} = 5$:

$$P(s_i^h|\Lambda) \propto \frac{\Lambda^{s_i^h}}{s_i^h!}\mathbb{1}_{\{s_i^h \leq \overline{s}\}}. \quad (2)$$

Conditional on s_i^h, the prior for the parent set $\mathcal{M}_i^h$ is a uniform distribution over all parent sets with cardinality s_i^h: $P(\mathcal{M}_i^h \,||\mathcal{M}_i^h| = s_i^h) = 1/\binom{p}{s_i^h}$. The overall prior on the network structures is given by marginalization:

$$P(\mathcal{M}_i^h|\Lambda) = \sum_{s_i^h=1}^{\overline{s}} P(\mathcal{M}_i^h|s_i^h)P(s_i^h|\Lambda) \quad (3)$$

Conditional on the parent set $\mathcal{M}_i^h$ of size s_i^h, the $s_i^h + 1$ regression coefficients, denoted by $a_{\mathcal{M}_i^h} = (a_{i0}^h, (a_{ij}^h)_{j\in\mathcal{M}_i^h})$, are assumed zero-mean multivariate Gaussian with covariance matrix $(\sigma_i^h)^2\Sigma_{\mathcal{M}_i^h}$,

$$P(a_i^h|\mathcal{M}_i^h, \sigma_i^h) = |2\pi(\sigma_i^h)^2\Sigma_{\mathcal{M}_i^h}|^{-\frac{1}{2}} \exp\left(-\frac{a_{\mathcal{M}_i^h}^\dagger \Sigma_{\mathcal{M}_i^h}^{-1} a_{\mathcal{M}_i^h}}{2(\sigma_i^h)^2}\right),$$

where the symbol $\dagger$ denotes matrix transposition, $\Sigma_{\mathcal{M}_i^h} = \delta^{-2}D_{\mathcal{M}_i^h}^\dagger(y)D_{\mathcal{M}_i^h}(y)$ and $D_{\mathcal{M}_i^h}(y)$ is the $(\xi_i^h - \xi_i^{h-1}) \times (s_i^h+1)$ matrix whose first column is a vector of 1 (for the constant in model (1)) and each $(j+1)^{th}$ column contains the observed values $(y_j(t))_{\xi_i^{h-1}\leq t<\xi_i^h}$ for all factor gene j in $\mathcal{M}_i^h$. Finally, the conjugate prior for the variance $(\sigma_i^h)^2$ is the inverse gamma distribution, $P((\sigma_i^h)^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following (Lèbre, 2007), we set the hyper-hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that give a vague distribution. The terms λ and Λ can be interpreted as the expected number of changepoints and parents, respectively, and δ^2 is the expected signal-to-noise ratio. These hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family: $P(\Lambda) = P(\lambda) = \mathcal{G}a(0.5, 1)$ and $P(\delta^2) = \mathcal{IG}(2, 0.2)$.

2.3. Posterior

Equation (1) implies that

$$P(y_i^h|\xi_i^{h-1}, \xi_i^h, \mathcal{M}_i^h, a_i^h, \sigma_i^h) = \left(\sqrt{2\pi}\sigma_i^h\right)^{-(\xi_i^h - \xi_i^{h-1})} \exp\left(-\frac{(y_i^h - D_{\mathcal{M}_i^h}(y)a_{\mathcal{M}_i^h})^\dagger (y_i^h - D_{\mathcal{M}_i^h}(y)a_{\mathcal{M}_i^h})}{2(\sigma_i^h)^2}\right). \quad (4)$$

From Bayes theorem, the posterior is given by

$$P(k, \xi, \mathcal{M}, a, \sigma^2, \lambda, \Lambda, \delta^2|y) \propto \quad (5)$$
$$P(\delta^2)P(\lambda)P(\Lambda)\prod_{i=1}^p P(k_i|\lambda)P(\xi_i|k_i)\prod_{h=1}^{k_i} P(\mathcal{M}_i^h|\Lambda)$$
$$P([\sigma_i^h]^2)P(a_i^h|\mathcal{M}_i^h, [\sigma_i^h]^2, \delta^2)P(y_i^h|\xi_i^{h-1}, \xi_i^h, \mathcal{M}_i^h, a_i^h, [\sigma_i^h]^2)$$

where all prior distributions have been defined above.

2.4. Inference

An attractive feature of the chosen model is that the marginalization over the parameters a and σ in the posterior distribution of (5) is analytically tractable:

$$P(k,\xi,\mathcal{M},\lambda,\Lambda,\delta^2|y) = \int P(k,\xi,\mathcal{M},a,\sigma,\lambda,\Lambda,\delta^2|y)da d\sigma \quad (6)$$

See (Lèbre, 2007) for details and an explicit expression. The number of changepoints and their location, k, ξ, the network structure $\mathcal{M}$ and the hyperparameters $\lambda, \Lambda, \delta^2$ can be sampled from the posterior distribution $P(k, \xi, \mathcal{M}, \lambda, \Lambda, \delta^2|y)$ with RJMCMC. A detailed description can be found in (Lèbre, 2007).

3. Model Improvement

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

To address these problems, we propose two methods of information sharing among time series segments. The first method is based on the hierarchical Bayesian model of (Werhli & Husmeier, 2008). However, rather than sharing information hierarchically – comparing all network structures to a central latent structure – we share information sequentially: a network structure is a priori assumed to be similar to the adjacent ones. The second method uses information from all the other segments to define a prior distribution on the edges for a given segment. We will investigate the relative merits of these two information sharing schemes below.

3.1. Sequential information sharing

Denote by $K_i := k_i + 1$ the number of partitions associated with node i, and recall that each time series segment y^h is associated with a separate subnetwork $\mathcal{M}_i^h$, $1 \leq h \leq K_i$. We impose a prior distribution $P(\mathcal{M}_i^h|\mathcal{M}_i^{h-1}, \beta_i)$ on the structures, and the joint probability distribution factorizes according to a

Markovian dependence:

$$P(y^1,\ldots,y^{K_i},\mathcal{M}_i^1,\ldots,\mathcal{M}_i^{K_i},\beta_i) = \prod_{h=1}^{K_i} P(y^h|\mathcal{M}_i^h)P(\mathcal{M}_i^h|\mathcal{M}_i^{h-1},\beta_i)P(\beta_i) \quad (7)$$

Similar to (Werhli & Husmeier, 2008) we define

$$P(\mathcal{M}_i^h|\mathcal{M}_i^{h-1},\beta_i) = \frac{\exp(-\beta_i|\mathcal{M}_i^h-\mathcal{M}_i^{h-1}|)}{Z_i(\beta_i,\mathcal{M}_i^{h-1})} \quad (8)$$

for $h \geq 2$, where β_i is a hyperparameter that defines the strength of the coupling between $\mathcal{M}_i^h$ and $\mathcal{M}_i^{h-1}$. For $h = 1$, $P(\mathcal{M}_i^h)$ is given by (3). The denominator $Z_i(\beta_i,\mathcal{M}_i^{h-1})$ in (8) is a normalizing constant, also known as the partition function:

$$Z_i(\beta_i) = \sum\nolimits_{\mathcal{M}_i^h\in\mathbb{M}_i} e^{-\beta_i|\mathcal{M}_i^h-\mathcal{M}_i^{h-1}|} \quad (9)$$

where $\mathbb{M}_i$ is the set of all valid subnetwork structures. If we ignore any fan-in restriction that might have been imposed a priori (via $\overline{s}$), then the expression for the partition function can be simplified: $Z_i(\beta_i) \approx \prod_i \prod_j Z_{ij}(\beta_i)$ where

$$Z_{ij}(\beta_i) = \sum\nolimits_{e_{ij}^h=0}^{1} e^{-\beta_i|e_{ij}^h-e_{ij}^{h-1}|} = 1+e^{-\beta_i} \quad (10)$$

and hence

$$Z_i = \left(1+e^{-\beta_i}\right)^p \quad (11)$$

Inserting (11) into (8) gives:

$$P(\mathcal{M}_i^h|\mathcal{M}_i^{h-1},\beta_i) = \frac{\exp(-\beta_i|\mathcal{M}_i^h-\mathcal{M}_i^{h-1}|)}{(1+e^{-\beta_i})^p} \quad (12)$$

It is straightforward to integrate the proposed model into the RJMCMC scheme of (Lèbre, 2007). When proposing a new network structure $\mathcal{M}_i^h \rightarrow \tilde{\mathcal{M}_i}^h$ for segment h, the prior probability ratio has to be replaced by the following one:

$$\frac{P(\mathcal{M}_i^{h+1}|\tilde{\mathcal{M}_i}^h,\beta_i)P(\tilde{\mathcal{M}_i}^h|\mathcal{M}_i^{h-1},\beta_i)}{P(\mathcal{M}_i^{h+1}|\mathcal{M}_i^h,\beta_i)P(\mathcal{M}_i^h|\mathcal{M}_i^{h-1},\beta_i)} = \frac{\exp[-\beta_i(|\mathcal{M}_i^{h+1}-\tilde{\mathcal{M}}_i^h|+|\tilde{\mathcal{M}}_i^h-\mathcal{M}_i^{h-1}|)]}{\exp[-\beta_i(|\mathcal{M}_i^{h+1}-\mathcal{M}_i^h|+|\mathcal{M}_i^h-\mathcal{M}_i^{h-1}|)]} \quad (13)$$

An additional MCMC step is introduced for sampling the hyperparameters β_i from the posterior distribution. For a proposal move $\beta_i \rightarrow \tilde{\beta}_i$ with symmetric proposal probability $Q(\tilde{\beta}_i|\beta_i) = Q(\beta_i|\tilde{\beta}_i)$ we get the following acceptance probability:

$$A(\tilde{\beta}_i|\beta_i) = \min\left\{\prod_{h=2}^{K_i}\frac{\exp(-\tilde{\beta}_i|\mathcal{M}_i^h-\mathcal{M}_i^{h-1}|)}{\exp(-\beta_i|\mathcal{M}_i^h-\mathcal{M}_i^{h-1}|)}\frac{\left(1+e^{-\beta_i}\right)^p P(\tilde{\beta}_i)}{\left(1+e^{-\tilde{\beta}_i}\right)^p P(\beta_i)},1\right\} \quad (14)$$

where in our study the hyperprior $P(\beta_i)$ was chosen as the uniform distribution on the interval $[0,10]$. Note that the scheme proposed in (Robinson & Hartemink, 2009) can be regarded as a special case of the one we propose, with two simplifications: (1) Changepoints are not allowed to vary between nodes. (2) The common hyperparameter $\beta_i = \beta\ \forall i$ has to be chosen by the user in advance and is not inferred from the data.

3.2. Global information sharing

We investigate an alternative scheme, based on ideas presented in (Ferrazzi et al., 2008). Let $e_{ij}^h \in \{0,1\}$ denote the indicator variable for a directed edge from node i to node j in the hth network (i.e. the network corresponding to the hth section of the time series), and let $\theta_{ij} \in [0,1]$ denote the probability that the node pair (i,j) is connected by a directed edge. We assume that for a given node pair (i,j), the edge indicator variables $\{e_{ij}^h\}$ are iid distributed,

$$P(e_{ij}^h|\theta_{ij}) = (\theta_{ij})^{e_{ij}^h}(1-\theta_{ij})^{1-e_{ij}^h} \quad (15)$$

with a conjugate beta prior on the parameters θ_{ij}:

$$P(\theta_{ij}) = \frac{\Gamma(\alpha_{ij}+\overline{\alpha_{ij}})}{\Gamma(\alpha_{ij})\Gamma(\overline{\alpha_{ij}})}\theta_{ij}^{\alpha_{ij}-1}(1-\theta_{ij})^{\overline{\alpha_{ij}}-1} \quad (16)$$

where α_{ij} and $\overline{\alpha_{ij}}$ are hyperparameters. Given the subnetworks $\mathcal{M}_i^{\tilde{h}}$ in all segments $\tilde{h}$ different from the current segment h, the prior probability of the subnetwork in the current segment, $\mathcal{M}_i^h$, is

$$P(\mathcal{M}_i^h|\{\mathcal{M}_i^{\tilde{h}}\}_{\tilde{h}\neq h}) = \prod\nolimits_j P(e_{ij}^h|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}) \quad (17)$$

$$P(e_{ij}^h|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}) = \int P(e_{ij}^h|\theta_{ij})P(\theta_{ij}|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h})d\theta_{ij}$$

where

$$P(\theta_{ij}|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}) \propto P(\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}|\theta_{ij})P(\theta_{ij}) \quad (18)$$

We introduce the following sufficient statistics: B_{ij}^h is the number of networks in segments different from the current segment h in which the node pair (i,j) is connected by a directed edge. Conversely, $\overline{B_{ij}^h}$ is the size of the complement set, i.e. the number of networks in segments different from the current segment h without an edge from node i to node j. Obviously, $B_{ij}^h + \overline{B_{ij}^h} = K_i - 1$, and

$$P(\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}|\theta_{ij}) = \theta_{ij}^{B_{ij}^h}(1-\theta_{ij})^{\overline{B_{ij}^h}} \quad (19)$$

Inserting (19) and (16) into (18) leads to:

$$P(\theta_{ij}|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}) = \frac{\Gamma(\alpha_{ij}+B_{ij}^h+\overline{\alpha_{ij}}+\overline{B_{ij}^h})}{\Gamma(B_{ij}^h+\alpha_{ij})\Gamma(\overline{B_{ij}^h}+\overline{\alpha_{ij}})}\theta_{ij}^{B_{ij}^h+\alpha_{ij}-1}(1-\theta_{ij})^{\overline{B_{ij}^h}+\overline{\alpha_{ij}}-1} \quad (20)$$

Inserting (15) and (20) into (17) yields:

$$P(e_{ij}^h|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}) = \frac{\Gamma(\alpha_{ij}+B_{ij}^h+\overline{\alpha_{ij}}+\overline{B_{ij}^h})}{\Gamma(B_{ij}^h+\alpha_{ij})\Gamma(\overline{B_{ij}^h}+\overline{\alpha_{ij}})}$$
$$\int(\theta_{ij})^{B_{ij}^h+e_{ij}^h+\alpha_{ij}-1}(1-\theta_{ij})^{\overline{B_{ij}^h}+\overline{e_{ij}^h}+\overline{\alpha_{ij}}-1}d\theta_{ij}$$
$$= \frac{\Gamma(\alpha_{ij}+B_{ij}^h+\overline{\alpha_{ij}}+\overline{B_{ij}^h})}{\Gamma(B_{ij}^h+\alpha_{ij})\Gamma(\overline{B_{ij}^h}+\overline{\alpha_{ij}})}$$
$$\frac{\Gamma(B_{ij}^h+\alpha_{ij}+e_{ij}^h)\Gamma(\overline{B_{ij}^h}+\overline{\alpha_{ij}}+\overline{e_{ij}^h})}{\Gamma(\alpha_{ij}+B_{ij}^h+\overline{\alpha_{ij}}+\overline{B_{ij}^h}+1)} \quad (21)$$

where we have defined $\overline{e_{ij}^h} = 1-e_{ij}^h$. Using $\Gamma(x+1) = x\Gamma(x)$, this expression can be simplified:

$$P(e_{ij}^h = 1|\{e_{ij}^{\tilde{h}}\}_{\tilde{h}\neq h}) = \frac{\alpha_{ij}+B_{ij}^h}{\alpha_{ij}+B_{ij}^h+\overline{\alpha_{ij}}+\overline{B_{ij}^h}} \quad (22)$$

The MCMC scheme is identical to the one described in (Lèbre, 2007), except that $P(\mathcal{M}_i^h|\{\mathcal{M}_i^{\tilde{h}}\}_{\tilde{h}\neq h})$ has to be used as the prior on $\mathcal{M}_i^h$, which is obtained by inserting (22) into (17). In our study, we have set $\alpha_{ij} = \overline{\alpha_{ij}} = 1$, in which case $P(\theta_{ij})$ in (16) reduces to the uniform distribution over the unit interval. One can extend this scheme by imposing a hyperprior on α_{ij} and $\overline{\alpha_{ij}}$, and sampling these hyperparameters from the posterior distribution with MCMC – this is the subject of our future work.

4. Setup and Diagnostics

The methods described in this paper have been implemented in R, based on code from (Lèbre, 2007), and the software is available from the authors upon request. Our program sets up an RJMCMC simulation to sample the network structure, the changepoints and the hyperparameters from the posterior distribution. As a convergence diagnostic we monitor the potential scale reduction factor (PSRF) (Gelman & Rubin, 1992), computed from the within-chain and between-chain variances of marginal edge posterior probabilities. Values of PSRF$\leq$1.1 are usually taken as indication of sufficient convergence. In our simulations, we extended the burn-in phase until a value of PSRF<= 1.05 was reached, and then sampled 1000 network and changepoint configurations in intervals of 200 RJMCMC steps. From these samples we compute the marginal posterior probabilities of all potential interactions, which defines a ranking of the edges in the recovered network. For the synthetic simulation study (see below), the gold standard (i.e. the true interaction network) is known. Therefore, by varying the threshold on the rank, we can construct the Receiver Operating Characteristic, or ROC curve (plotting the sensitivity or recall against the complementary specificity), and the precision-recall or PR curve (plotting the precision against the recall). To assess and succinctly score the network reconstruction accuracy, we follow a three-prong approach and compute three figures of merit that have been widely applied in the literature: the area under the ROC curve (AUROC), the area under the PR-curve (AUPRC), and the true positive rate at a fixed false positive rate of 5% (TPFP5).

5. Data

5.1. Synthetic data

We generated synthetic time series, each consisting of $K = 10$ segments of length 50, as follows. Random networks $\mathcal{M}^h$, $1 \leq h \leq K$, are generated stochastically, with the number of edges drawn from a Poisson distribution. Each directed edge from node j (the parent) to node i (the child) has a weight a_{ij}^h that determines the interaction strength, drawn from a Normal distribution. The signal associated with node i at time t, $y_i(t-1)$, evolves according to the heterogeneous first-order Markov process of equation (1). Denote by $\mathbf{A}^h$ the matrix of all interaction strengths a_{ij}^h. To ensure stationarity of the time series, we tested if all eigenvalues of $\mathbf{A}^h$ had a modulus ≤ 1, and removed edges randomly until this condition was met. The networks $\mathcal{M}^h$ that generated the time series consisted of 10 nodes, with on average 3 parents per node. To simulate a sequence of networks separated by changepoints, we sampled Δn_h from a Poisson distribution and then randomly changed Δn_h edges between $\mathcal{M}^h$ and $\mathcal{M}^{h+1}$, leaving the total number of edges unchanged. The parameter of the Poisson distribution, which determines the average number of changes between adjacent structures, $\mathcal{M}^h$ and $\mathcal{M}^{h+1}$, was varied, as described in more detail in Section 6.1.

5.2. Gene expression times course during morphogenesis in Drosophila

We also applied our method to the developmental gene expression time series for *Drosophila melanogaster* (fruit fly), obtained by (Arbeitman et al., 2002). Expression values of 4028 genes were measured with microarrays at 67 time points during the Drosophila life cycle, which contains the four distinct phases of embryo, larva, pupa and adult. In our study we concentrated on a subset of 11 genes that regulate muscle development. This dataset has also been used in (Guo et al., 2007), (Zhao et al., 2006) and (Robinson & Hartemink, 2009).

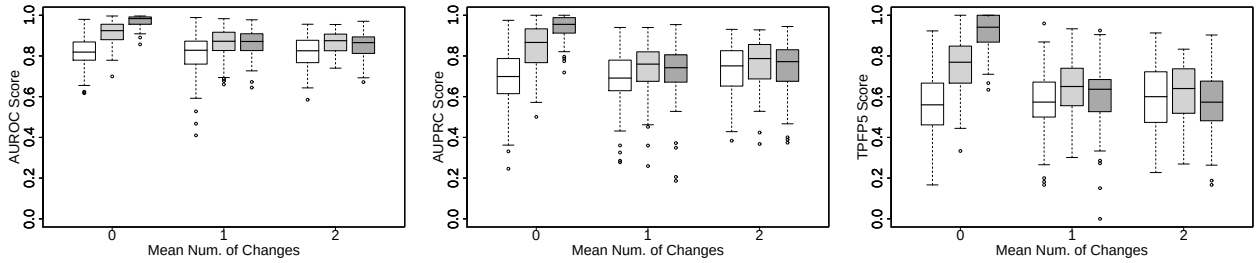

Figure 1. Network reconstruction accuracy, measured with three scoring schemes, as discussed in Section 4. Left panel: AUROC; centre panel: AUPRC; right panel: TPFP5. The boxplots show the distributions of these scores, where the horizontal bar shows the median, the box margins show the 25th and 75th percentiles, the whiskers indicate data within 2 times the interquartile range, and circles are outliers. The grey shading indicates the method. Unshaded boxes: HetDBN-0. Light shading: HetDBN-SI. Dark shading: HetDBN-GI. The numbers on the horizontal axes indicate the average number of network structure changes per node between adjacent time series segments. A paired t-test showed that all differences are significant at the 5% level except for the following. AUROC: HetDBN-GI versus HetDBN-SI, 1 change; AUPRC: HetDBN-0 versus HetDBN-GI, 2 changes; TPFP5: HetDBN-0 versus HetDBN-GI, 1 and 2 changes.

6. Results and Discussion

6.1. Experiments on simulated data

We compared the network reconstruction accuracy of three models: the heterogeneous DBN proposed in (Lèbre, 2007) (HetDBN-0), the heterogeneous DBN with the sequential information sharing scheme proposed in Section 3.1 (HetDBN-SI), and the heterogeneous DBN with the global information sharing scheme proposed in Section 3.2 (HetDBN-GI). The methods were applied to the synthetic data described in Section 5.1. We repeated the simulations for each experimental setting on 10 independent data instantiations, and scored the network reconstruction accuracy with three separate measures, as discussed in Section 4. We investigated how the average number of changes in network structure between adjacent segments affects the performance. Figure 1 shows boxplots of the score distributions. To test for significance of the discerned trends, we carried out a paired t-test; see the caption of Figure 1, and the supplementary material[2] for a table. When there are no changes in the network structure, information sharing results in a considerable performance improvement, and HetDBN-GI outperforms HetDBN-SI. The latter finding is plausible, as HetDBN-GI utilizes information from all the segments, whereas HetDBN-SI only utilizes information from the adjacent segments. When the number of edge changes between segments increases, information sharing achieves a less substantial, yet still significant improvement over HetDBN-0. Also, the performance between the two approaches is inverted, with HetDBN-SI slightly yet significantly outperforming HetDBN-GI. Again, this result is plausible. Larger differences among network structures imply that, per se, less is gained from information sharing. Also, given a segment, a network associated with a remote segment will on average have accumulated a larger number of structure differences than a network associated with a close segment; this explains the superiority of the sequential (HetDBN-SI) over the global (HetDBN-GI) information sharing scheme. To investigate the trend more thoroughly, we reduced the computational costs of the MCMC simulations by reducing the network complexity to 1 target node and 20 potential parents, and keeping the hyperparameters fixed. We then carried out simulations over an extended range of average structure differences. The results are shown in Figure 2 - for space restrictions, we only show the AUROC scores. For small numbers of differences among the network structures associated with different segments, information sharing results in a considerable performance improvement over HetDBN-0. The amount of improvement degrades as the differences among structures increase. For small differences, HetDBN-GI tends to outperform HetDBN-SI. This trend is inverted when the difference among network structures increases. These results thus confirm the patterns found in Figure 1, which have been discussed above.

[2]The supplementary material can be found at: http://www.bioss.ac.uk/staff/dirk/Supplements/

6.2. Gene networks related to morphogenesis in the Drosophila life cycle

The top panel in Figure 3 shows the marginal posterior probability of changepoints during the life cycle of *Drosophila melanogaster*, inferred with the proposed method HetDBN-SI from the gene expression time series described in Section 5.2. For a comparison, we applied the method proposed in (Ahmed & Xing,

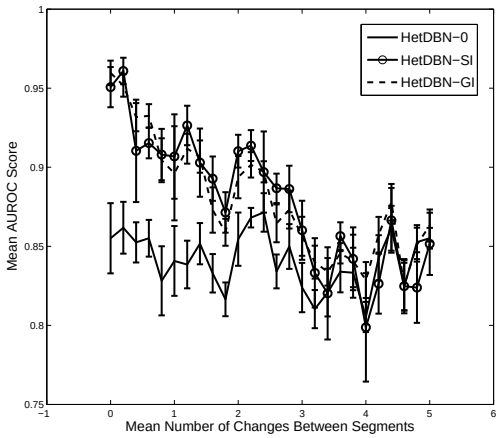

Figure 2. Network reconstruction accuracy for different methods. The plots show mean AUROC scores (vertical axis) plotted against the average number of network structure changes per node between adjacent time series segments (horizontal axis). Mean values and standard errors were obtained from 10 independent time series.

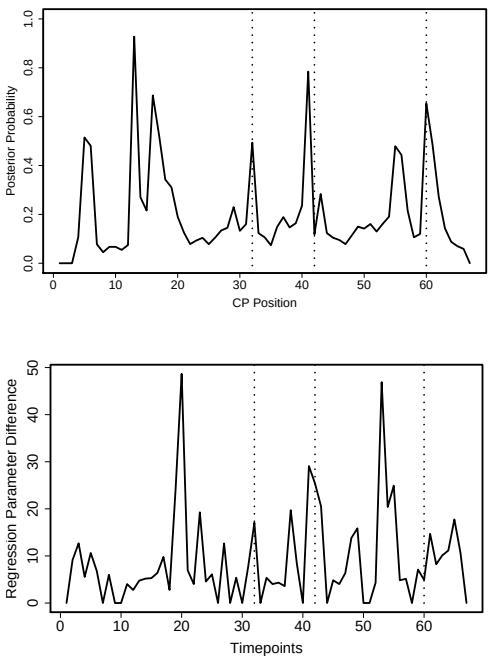

Figure 3. Changepoints during morphogenesis in *Drosophila melanogaster*. Top panel: HetDBN-SI, posterior probability of a changepoint occurring for any node at a given time (vertical axis) plotted against time (horizontal axis). Bottom panel: TESLA, L1-norm of the difference of the regression parameter vectors associated with two adjacent time points (vertical axis) plotted against time (horizontal axis). The vertical dotted lines indicate the three morphogenic transitions.

2009), using the authors' software package TESLA. Note that this model depends on various regularization parameters, which were optimized by maximizing the BIC score, as in (Ahmed & Xing, 2009). The results are shown in the bottom panel of Figure 3, where the graph shows the L1-norm of the difference of the regression parameter vectors associated with adjacent time points. (Robinson & Hartemink, 2009) applied their discrete heterogeneous DBN to the same data set, and a plot corresponding to the top panel of Figure 3 can be found in their paper. A comparison of these plots suggests that our method is the only one that clearly detects all three morphogenic transitions: embryo → larva, larva → pupa, and pupa → adult. The bottom panel of Figure 3 indicates that the last transition, pupa → adult, is less clearly detected with TESLA, and it is completely missing in (Robinson & Hartemink, 2009). Both our method, HetDBN-SI, as well as TESLA detect additional transitions during the embryo stage, which are missing in (Robinson & Hartemink, 2009). We would argue that a complex gene regulatory network is unlikely to transit into a new morphogenic phase all at once, and some pathways might have to undergo activational changes earlier in preparation for the morphogenic transition. As such, it is not implausible that additional transitions at the gene regulatory network level occur. However, a failure to detect known morphogenic transitions can clearly be seen as a shortcoming of a method, and on these grounds our model appears to outperform the two alternative ones.

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

important change in the role of *twi*. This gene does not have an important role as a regulator during the early phases, but functions as a regulator of five other genes during the adult phase: *mlc1, gfl, actn, msp300* and *sls*. The absence of a regulatory role for *twi* during the earlier phases is consistent with (Elgar et al., 2008), who found that another regulator, *mef2* (not included in the dataset) controls the expression of *mlc1*, *actn* and *msp300* during early development.

7. Conclusions

We have proposed a novel heterogeneous DBN, which has various advantages over existing schemes: it does not require the data to be discretized (as opposed to (Robinson & Hartemink, 2009)); it allows the network structure to change with time (as opposed to (Grzegorczyk & Husmeier, 2009)); it includes a regularization scheme based on inter-time segment information sharing (as opposed to (Lèbre, 2007)); and it allows all hyperparameters to be inferred from the data via a consistent Bayesian inference scheme (as opposed to (Ahmed & Xing, 2009)). An evaluation on synthetic data has demonstrated an improved performance over (Lèbre, 2007). The application of our method to gene expression time series taken during the life cycle of *Drosophila melanogaster* has revealed better agreement with known morphogenic transitions than the methods of (Robinson & Hartemink, 2009) and (Ahmed & Xing, 2009), and we have detected changes in gene regulatory interactions that are consistent with independent biological findings. We have carried out a comparison between two alternative paradigms of information sharing – global versus sequential – and we have discussed the relative merits and shortcomings.

References

Ahmed, Amr and Xing, Eric P. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106:11878–11883, 2009.

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

Elgar, S.J., Han, J., and Taylor, M.V. mef2 activity levels differentially affect gene expression during Drosophila muscle development. *Proceedings of the National Academy of Sciences*, 105(3):918, 2008.

Ferrazzi, Fulvia, Rinaldi, S., Parikh, A., Shaulsky, G., Zupan, Blaz, and Bellazzi, Riccardo. Population models to learn Bayesian networks from multiple gene expression experiments. http://www.labmedinfo.org/biblio/author/326, 2008.

Gelman, A. and Rubin, D.B. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

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

Grzegorczyk, Marco and Husmeier, Dirk. 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. 2009.

Guo, F., Hanneke, S., Fu, W., and Xing, E.P. Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 328. ACM, 2007.

Kolar, Mladen, Song, Le, and Xing, Eric. 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. 2009.

Lèbre, S. *Stochastic process analysis for Genomics and Dynamic Bayesian Networks inference.* PhD thesis, Université d'Evry-Val-d'Essonne, France, 2007.

Robinson, Joshua W and Hartemink, Alexander J. 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, 2009.

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

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