Computational Statistics manuscript No.
(will be inserted by the editor)

Supplementary material for our Computational Statistics paper: ’Modelling non-stationary dynamic gene regulatory processes with the BGM model’

Marco Grzegorczyk · Dirk Husmeier · Jörg Rahnenführer

Received: date / Accepted: date

Abstract This paper provides the technical details, which for space restrictions could not be included in the main paper ’Modelling non-stationary dynamic gene regulatory processes with the BGM model’ submitted to the *Computational Statistics* journal. The three sections of this paper are organized as follows: In Section 1 we provide details about the Gaussian *BGe* scoring metric for static Bayesian networks as developed by Geiger and Heckerman [1]. The *BGe* scoring metric for dynamic Bayesian networks is described in detail in Section 2. Section 3 is an extended version of the methodology section of our main paper. We note that Subsections 3.1 and 3.2 have been modified slightly by adding references to the equations provided in Sections 1 and 2 of this supplementary paper. Subsection 3.3 is an substantially extended version of Section 2.3 (main paper) and provides all details of the changepoint process from Green’s RJMCMC paper [2].
Availability: This supplementary paper is available from:
http://www.statistik.tu-dortmund.de/cost2010.html

1 The Gaussian BGe scoring metric for static Bayesian networks

This section describes the linear Gaussian BGe scoring metric (Bayesian metric for Gaussian networks having score equivalence) for static Bayesian networks as developed by Geiger and Heckerman [1]. Given a data set **D** with *m* observations of the

M. Grzegorczyk and J. Rahnenführer
Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
E-mail: grzegorczyk@statistik.tu-dortmund.de
E-mail: rahnenfuehrer@statistik.tu-dortmund.de

D. Husmeier
Biomathematics and Statistics Scotland, JCMB, The King’s Buildings, Edinburgh EH93JZ, UK, E-mail: dirk@bioss.sari.ac.uk

variables $X_1,\ldots,X_N$:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{1,1} & \mathbf{D}_{1,2} & \ldots & \mathbf{D}_{1,m-1} & \mathbf{D}_{1,m} \\ \mathbf{D}_{2,1} & \mathbf{D}_{2,2} & \ldots & \mathbf{D}_{2,m-1} & \mathbf{D}_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{D}_{N,1} & \mathbf{D}_{N,2} & \ldots & \mathbf{D}_{N,m-1} & \mathbf{D}_{N,m} \end{pmatrix} \tag{1}$$

so that $\mathbf{D}_{n,j}$ denotes the jth realization of the nth node X_n, and the jth column of $\mathbf{D}$: $\mathbf{D}_{.,j} = (\mathbf{D}_{1,j},\ldots,\mathbf{D}_{N,j})^T$ is the jth realization vector of the variables. The Gaussian BGe model assumes that the observation vectors $\mathbf{D}_{.,j}$ $(j = 1,\ldots,m)$ are a random sample from a multivariate Gaussian distribution $\mathscr{N}(\mu,\Sigma)$ with an unknown mean vector μ and an unknown covariance matrix Σ. The prior joint distribution of μ and $\mathbf{W} = \Sigma^{-1}$ is supposed to be the normal-Wishart distribution, that is, the conditional distribution of μ given $\mathbf{W}$ is $\mathscr{N}(\mu_0,(v\cdot\mathbf{W})^{-1})$ with $v > 0$, and the marginal distribution of $\mathbf{W}$ is a Wishart distribution with $\alpha > N+1$ degrees of freedom and precision matrix $\mathbf{T}_0$, denoted $\mathscr{W}(\alpha,\mathbf{T}_0)$. The condition $\alpha > N+1$ ensures that the second moments of the posterior distribution are finite (see also Eq. (26) in [1]). Geiger and Heckerman show that the marginal likelihood $P(\mathbf{D}|\mathscr{G})$ of the data $\mathbf{D}$ given a graph $\mathscr{G}$ can then – under fairly weak conditions of parameter independence and parameter modularity – be computed in closed form. We define:

$$\mathbf{T}_{\mathbf{D},m} := \mathbf{T}_0 + \mathbf{S}_{\mathbf{D},m} + \frac{v\cdot m}{v+m}(\mu_0 - \overline{\mathbf{D}_m})(\mu_0 - \overline{\mathbf{D}_m})^T \tag{2}$$

where

$$\overline{\mathbf{D}_m} := \frac{1}{m}\sum_{j=1}^{m}\mathbf{D}_{.,j} \tag{3}$$

is the mean of the m realization vectors and

$$\mathbf{S}_{\mathbf{D},m} := \sum_{j=1}^{m}(\mathbf{D}_{.,j} - \overline{\mathbf{D}_m})\cdot(\mathbf{D}_{.,j} - \overline{\mathbf{D}_m})^T \tag{4}$$

$\mathbf{T}_0$, μ_0, α, and v are hyperparameters of the normal-Wishart prior and have to be specified in advance. $\mathbf{T}_0$ is an N-by-N matrix, μ_0 is a N-by-1 column vector, and v and α are 1-dimensional and usually referred to as total prior precision parameters. Furthermore, we set:

$$c(n,\alpha) := \left\{ 2^{\alpha\cdot n/2}\cdot\pi^{n\cdot(n-1)/4}\cdot\prod_{i=1}^{n}\Gamma\left(\frac{\alpha+1-i}{2}\right)\right\}^{-1} \tag{5}$$

The marginal likelihood can then be computed as follows ([1]):

$$P(\mathbf{D}|\mathscr{G}) = \prod_{n=1}^{N}\Psi(\mathbf{D}_n^{\pi_n}) = \prod_{n=1}^{N}\frac{P(\mathbf{D}^{\{X_n,\pi_n\}}|\mathscr{G}_F(\{X_n,\pi_n\}))}{P(\mathbf{D}^{\{\pi_n\}}|\mathscr{G}_F(\pi_n))} \tag{6}$$

where X_n is the nth variable, π_n is the parent set of X_n in the graph $\mathscr{G}$, $\mathbf{D}^{\{X_n,\pi_n\}}$ and $\mathbf{D}^{\{\pi_n\}}$ are the data submatrices corresponding to the realizations of the variables in the sets $\{X_n,\pi_n\}$ and $\{\pi_n\}$ only, and $\mathscr{G}_F(\{X_n,\pi_n\})$ and $\mathscr{G}_F(\pi_n)$ correspond to so-called *full graphs* for the variable subsets $\{X_n, \pi_n\}$ and $\{\pi_n\}$, that is, to subgraphs with the maximal number of edges so that the subgraphs do not impose any independence restrictions on these subsets of variables.

The marginal likelihood of the data subset $\mathbf{D}^{\{S\}} \subset \mathbf{D}$ corresponding to the m realizations of the $N^\dagger$-dimensional subset $S \subset \{X_1, \dots, X_N\}$ of the N variables given a full graph $\mathscr{G}_F(S)$ for the sub-domain S can be computed as follows ([1]):

$$P(\mathbf{D}^S|\mathscr{G}_F(S)) = (2\pi)^{-\frac{N^\dagger \cdot m}{2}} \cdot \left\{ \frac{v}{v+m} \right\}^{N^\dagger/2} \cdot \frac{c(N^\dagger, \alpha)}{c(N^\dagger, \alpha+m)} \tag{7}$$
$$\cdot det(\mathbf{T}_0^S)^{\frac{\alpha}{2}} \cdot det(\mathbf{T}_{\mathbf{D},m}^S)^{-\frac{\alpha+m}{2}}$$

where $det(\mathbf{T}_0^S)$ and $det(\mathbf{T}_{\mathbf{D},m}^S)$ denote the determinants of the submatrices $\mathbf{T}_0^S$ and $\mathbf{T}_{\mathbf{D},m}^S$ consisting only of those $N^\dagger$ rows and columns that correspond to variables in the subset S. $\mathbf{T}_{\mathbf{D},m}$ was defined in Eq. (2), and $c(N^\dagger, \alpha)$ and $c(N^\dagger, \alpha+m)$ can be computed with Eq. (5).

2 The Gaussian BGe scoring metric for dynamic Bayesian networks

We now consider the case that instead of independent observations, time series data have been collected for the domain: $(X_1(t), \dots X_N(t))_{t=1,\dots,m}$, and that we have a (1st-order) Markovian dependence structure. In this case, dynamic Bayesian networks (DBNs) can be employed. In DBNs each edge corresponds to an interaction with a time delay τ; e.g. for $\tau = 1$ an edge pointing from X_i to X_j means that the realization $x_{j,t}$ of X_j at time point t is influenced by the realization $x_i(t-1)$ of X_i at the previous time point $t-1$. This can be taken into consideration in the context of the Gaussian BGe model by building new data matrices – one for each domain variable – from the original data matrix of size N-by-m given in Eq. (1). For dynamic data the columns do not represent independent (steady-state) observations: the tth column of $\mathbf{D}$ is the realization of the variables at time point t ($t = 1, \dots, m$). We note that the score equivalence aspect of the *BGe* model is not required for dynamic Bayesian networks, because edge reversals are not permissible. However, formulating the models in terms of the *BGe* score is advantageous in case one intends to adapt the framework proposed in the main paper to non-linear static Bayesian networks along the line of [5].

In principle, there are two alternatives which can be used, and it depends on whether or not 'self-feedback loops', that is edges having the same node as starting and end point, should be allowed in the network. Here, we allow for 'self-feedback loops', and we build the following N matrices of size $(N+1)$-by-$(m-1)$ from the (time series) data matrix given in Eq. (1) :

$$\mathbf{D}(n) = \begin{pmatrix} \mathbf{D}_{1,1} & \mathbf{D}_{1,2} & \dots & \mathbf{D}_{1,m-1} \\ \mathbf{D}_{2,1} & \mathbf{D}_{2,2} & \dots & \mathbf{D}_{2,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{D}_{N,1} & \mathbf{D}_{N,2} & \dots & \mathbf{D}_{N,m-1} \\ \mathbf{D}_{n,2} & \mathbf{D}_{n,3} & \dots & \mathbf{D}_{n,m} \end{pmatrix} \tag{8}$$

$n = 1,\dots,N$. That is, we obtain $\mathbf{D}(n)$ by deleting the last column of $\mathbf{D}$ and adding a novel row $(\mathbf{D}_{n,2},\dots,\mathbf{D}_{n,m})$, i.e. the nth row of $\mathbf{D}$ shifted leftwards by 1, as the $(N+1)$-th row. We can identify the $(N+1)$-th row with a new domain variable X_{N+1}. This new variable is the nth domain variable with a time shift of size $\tau = 1$ and we note that the novel data matrices $\mathbf{D}(n)$ consist of observations for $N+1$ domain variables so that the hyperparameters T_0 and μ_0 have to be an $(N+1)$-by-$(N+1)$ matrix and an $(N+1)$-by-1 column vector, respectively, here. As before we can compute the matrix $\mathbf{T}_{\mathbf{D}(n)}$ for each data set $\mathbf{D}(n)$, and we replace Eq. (6) by:

$$P(\mathbf{D}|\mathscr{G}) = \prod_{n=1}^{N} \Psi(\mathbf{D}_n^{\pi_n}) \tag{9}$$

where

$$\Psi(\mathbf{D}_n^{\pi_n}) = \frac{P(\mathbf{D}(n)^{\{X_{N+1},\pi_n\}}|\mathscr{G}_F(\{X_{N+1},\pi_n\})}{P(\mathbf{D}(n)^{\{\pi_n\}}|\mathscr{G}_F(\pi_n))} \tag{10}$$

Eq. (7) has to be replaced by:

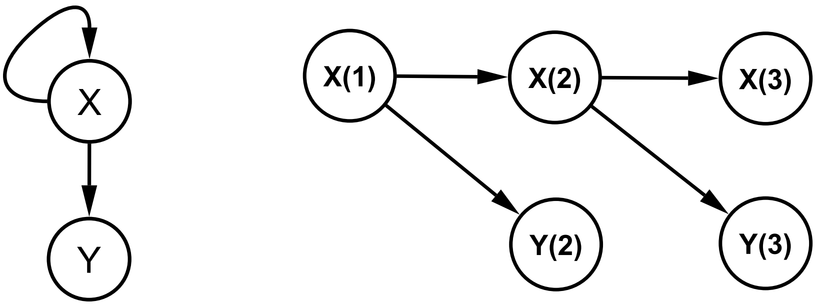
$$\begin{aligned} P(\mathbf{D}(n)^S|\mathscr{G}_F(S)) \;=\; & (2\pi)^{-\frac{N^\dagger\cdot(m-1)}{2}} \cdot \left\{\frac{\nu}{\nu+(m-1)}\right\}^{N^\dagger/2} \cdot \frac{c(N^\dagger,\alpha)}{c(N^\dagger,\alpha+(m-1))} \\ & \cdot det(\mathbf{T}_0^S)^{\frac{\alpha}{2}} \cdot det(\mathbf{T}^S_{\mathbf{D}(n),(m-1)})^{-\frac{\alpha+(m-1)}{2}} \end{aligned} \tag{11}$$

where $\mathscr{G}_F(S)$ is a full graph for the domain variable subset S of cardinality $N^\dagger$ and $\mathbf{T}_0^S$ and $\mathbf{T}^S_{\mathbf{D}(n),(m-1)}$ are sub-matrices as explained in Section 1.

3 Methodology

3.1 The dynamic BGe network (duplicated from the main paper)

Dynamic Bayesian networks (DBNs) are flexible models for representing probabilistic relationships among variables $X_1,\dots,X_N$. The graph $\mathscr{G}$ of a DBN describes the relationships among the variables, which have been measured at equidistant time points $t = 1,\dots,m$, in the form of conditional probability distributions. An edge pointing from X_i to X_j means that the realisation of X_j at time point t, symbolically:



(a) recurrent network (b) unfolded dynamic network (DBN)

Fig. 1 State space graph and corresponding dynamic Bayesian network. (a) Recurrent state space graph containing two nodes. Node X has a recurrent self-loop and acts as a regulator of node Y. (b) The bipartite graph structure (DBN) imposed by the graph $\mathscr{G}$ in panel (a).

$X_j(t)$, is influenced by the realisation of X_i at time point $t-1$, symbolically: $X_i(t-1)$. $\pi_n = \pi_n(\mathscr{G})$ denotes the parent node set of node X_n in $\mathscr{G}$, i.e. the set of all nodes from which an edge points to node X_n in $\mathscr{G}$. In principle, each node can be its own parent node in DBNs. Such self-loops $X_n(t-1) \to X_n(t)$ model autocorrelations, and it depends on the application whether or not they should be allowed. Alternatively, self-loops can be ruled out altogether to focus on a gene's interactions with other genes. We note that a DBN is based on a bipartite graph structure between two time steps t and $t+1$ so that the acyclicity constraint – known from static Bayesian networks – is guaranteed to be satisfied. The bipartite graph structure of DBNs is illustrated graphically in Fig. 1. The figure shows the state space representation (a) and the bipartite graph structure of the corresponding DBN (b). The network consists of two interacting nodes X and Y. Node X regulates node Y, and X also has a regulatory self-loop acting back on itself.

Given a data set $\mathbf{D}$, where $\mathbf{D}_{n,t}$ and $\mathbf{D}_{\pi_n,t}$ are the tth realisations $X_n(t)$ and $\pi_n(t)$ of X_n and π_n, respectively, DBNs are based on the following homogeneous Markov chain expansion:

$$P(\mathbf{D}|\mathscr{G},\theta) = \prod_{n=1}^{N}\prod_{t=2}^{m} P(X_n(t) = \mathbf{D}_{n,t}|\pi_n(t-1) = \mathbf{D}_{\pi_n,t-1}, \theta_n) \tag{12}$$

where θ is the total parameter vector, composed of subvectors θ_n. θ_n specifies the nth local conditional distribution $P(X_n(t)|\pi_n(t-1),\theta_n)$ in the factorisation. We note that in DBNs with time lag $\tau = 1$ the first time point $t = 1$ cannot be employed for computing the likelihood in Eq. (12), since the realisations of potential parent nodes at the previous time point $t = 0$ are unknown. The BGe model [1] from Section 2 specifies the distributional form $P(\mathbf{D}|\mathscr{G},\theta) = P(\mathbf{D}|\mathscr{G},(\mu,\mathbf{W}^{-1}))$ as a multivariate Gaussian distribution with expectation vector μ and precision matrix $\mathbf{W}$, and assumes a normal-Wishart distribution as prior distribution $P(\mu,\mathbf{W}|\mathscr{G})$. The local probability distributions $P(X_n(t)|\pi_n(t-1),\theta_n)$ are then given by conditional linear Gaussian distributions. Under fairly weak conditions imposed on the parameter vector θ and the prior distribution $P(\theta)$, the parameters can be integrated out analytically as shown in Section 2 and the marginal likelihood satisfies the same expansion rule as the DBN with fixed parameters [1]:

$$P(\mathbf{D}|\mathscr{G}) = \int P(\mathbf{D}|\mathscr{G},\theta)P(\theta|\mathscr{G})d\theta = \prod_{n=1}^{N} \Psi(\mathbf{D}_n^{\pi_n}) \tag{13}$$

where

$$\Psi(\mathbf{D}_n^{\pi_n}) = \int \prod_{t=2}^{m} P(X_n(t) = \mathbf{D}_{n,t}|\pi_n(t-1) = \mathbf{D}_{\pi_n,t-1},\theta_n)P(\theta_n|\pi_n)d\theta_n \tag{14}$$

and $\mathbf{D}_n^{\pi_n} := \{(\mathbf{D}_{n,t},\mathbf{D}_{\pi_n,t-1}) : 2 \leq t \leq m\}$ denotes the subset of the data pertaining to node X_n and its parent set π_n. For the Gaussian BGe model the (local) factors $\Psi(\mathbf{D}_n^{\pi_n})$ in Eq. (14) can be computed in closed-form according to Eqn. (10) and (11) in Section 2.

A sample of graphs from the posterior distribution $P(\mathscr{G}|\mathbf{D})$ can be obtained with Markov chain Monte Carlo (MCMC) simulations. The structure MCMC algorithm of [6] generates a sample of graphs as follows: A new candidate graph $\mathscr{G}_{i+1}$ is randomly drawn out of the set of graphs $\mathscr{N}(\mathscr{G}_i)$ that can be reached from the current graph $\mathscr{G}_i$ by deletion or addition of a single edge, and the proposed graph $\mathscr{G}_{i+1}$ is accepted with probability $A(\mathscr{G}_{i+1}|\mathscr{G}_i) = min\{R,1\}$ where

$$R = \frac{P(\mathbf{D}|\mathscr{G}_{i+1})P(\mathscr{G}_{i+1})}{P(\mathbf{D}|\mathscr{G}_i)P(\mathscr{G}_i)} \cdot \frac{|\mathscr{N}(\mathscr{G}_i)|}{|\mathscr{N}(\mathscr{G}_{i+1})|} \tag{15}$$

otherwise the chain is left unchanged: $\mathscr{G}_{i+1} := \mathscr{G}_i$. For each edge the fraction of sampled graphs that contain this edge is an estimator of its (marginal) posterior probability. If the true network is known, the reconstruction accuracy can for example be measured in terms of receiver operator characteristic (ROC) curves (e.g. [4]). We assume that $e_{ij} = 1$ indicates that there is an edge from X_i to X_j in the graph, while $e_{ij} = 0$ indicates that this edge is not present. BNs infer marginal posterior probabilities $\widehat{e_{ij}}$ for each edge e_{ij}.

Let $\varepsilon(\theta) = \{e_{ij}|\widehat{e_{ij}} > \theta\}$ denote the set of edges whose probabilities exceed a given threshold θ. Given θ the number of true positive (TP), false positive (FP), and false negative (FN) edge findings can be counted, and the *sensitivity* $S = TP/(TP+FN)$ and the *inverse specificity* $I = FP/(TN+FP)$ can be computed. This procedure can be repeated for several values of θ and the ensuing sensitivities can be plotted against the corresponding inverse specificities. This gives the ROC curve. Larger areas under the curve (AUC) indicate a better network reconstruction performance, where AUC= 1 is an upper limit, while AUC= 0.5 corresponds to random expectation. An alternative and more intuitive criteria is given by $(FP|TP = 10)$ counts: For each MCMC output a threshold ψ is imposed on the inferred edge posterior probabilities such that 10 true positive (TP) edges are extracted and the corresponding number of false positive (FP) edges, symbollicaly $(FP|TP = 10)$, exceeding the threshold ψ, is counted.

3.2 The dynamic Bayesian Gaussian Mixture (BGM$_D$) Bayesian network model (duplicated from the main paper)

In the Gaussian BGe model the local distributions $P(X_n(t)|\pi_n(t-1),\theta_n)$ are conditional linear Gaussian distributions so that only linear relationships among variables can be inferred. We generalise the BGe model by the introduction of a latent allocation vector $\mathbf{V}$, which assigns the data points to K different mixture components, where K is inferred from the data by applying changepoint birth and death moves, along the line of the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm of [2]. As in the original BGM model [3], conditional on the latent vector $\mathbf{V}$, a separate BGe score can be computed for each of the K mixture components.
The allocation vector $\mathbf{V}$ of size m describes the allocation of the time points $t = 2,\dots,m$ to the K compartments. $\mathbf{V}(t)=k$ denotes that time point t $(2 \leq t \leq m)$ is allocated to the k-th compartment $(1 \leq k \leq K)$, and $\mathbf{D}^{(\mathbf{V},k)}$ denotes all time points that are allocated to compartment k.
The posterior probability of $\mathscr{G}$, $\mathbf{V}$ and K is proportional to the joint distribution:

$$P(\mathscr{G},\mathbf{V},K|\mathbf{D}) = \frac{P(\mathscr{G},\mathbf{V},K,\mathbf{D})}{P(\mathbf{D})} \propto P(\mathscr{G},\mathbf{V},K,\mathbf{D}) \tag{16}$$

and the joint distribution can be factorised as follows in the BGM$_D$ model:

$$P(\mathscr{G},\mathbf{V},K,\mathbf{D}) = P(K)P(\mathbf{V}|K)P(\mathscr{G})P(\mathbf{D}|\mathscr{G},\mathbf{V},K) \tag{17}$$

where

$$P(\mathbf{D}|\mathscr{G},\mathbf{V},K) = \prod_{k=1}^{K} P(\mathbf{D}^{(\mathbf{V},k)}|\mathscr{G}) = \prod_{k=1}^{K}\prod_{n=1}^{N} \Psi(\mathbf{D}_n^{(\mathbf{V},k),\pi_n}) \tag{18}$$

and $\mathbf{D}_n^{(\mathbf{V},k),\pi_n} := \{(\mathbf{D}_{n,t},\mathbf{D}_{\pi_n,t-1})|t \in \{2,\dots,m\} : \mathbf{V}(t)=k\}$ denotes the set of realisations of node X_n and its parent set π_n for those time points that have been allocated to the kth component. It can be seen from these equations that $\mathbf{V}$ acts as a filter which divides the data $\mathbf{D}$ into K different compartments $\mathbf{D}^{(\mathbf{V},k)}$ $(k=1\dots K)$, for which separate independent BGe scores can be computed in closed-form using Eqns. (13) and (14). The BGM$_D$ counterpart of Eq. (14) is given by:

$$\Psi(\mathbf{D}_n^{(\mathbf{V},k),\pi_n}) = \int \prod_{t:\mathbf{V}(t)=k} P(X_n(t)=\mathbf{D}_{n,t}|\pi_n(t-1)=\mathbf{D}_{\pi_n,t-1},\theta_n)P(\theta_n|\pi_n)d\theta_n \tag{19}$$

When a data compartment $\mathbf{D}^{(\mathbf{V},k)}$ is empty, then we set the factors $\Psi(\mathbf{D}_n^{(\mathbf{V},k),\pi_n})$ equal to 1 $(n=1,\dots,N)$. The $\Psi(\mathbf{D}_n^{(\mathbf{V},k),\pi_n})$ terms that correspond to non-empty data (sub)sets $\mathbf{D}_n^{(\mathbf{V},k),\pi_n}$ can be computed with Eqn. (10) and (11) from Section 2. The data set $\mathbf{D}_n^{\pi_n}$ in Eq. (10) has to be replaced by the subset $\mathbf{D}_n^{(\mathbf{V},k),\pi_n}$, and $\mathbf{D}(n)^S$ in Eq. (11) has to be replaced by the subset $\left(\mathbf{D}_n^{(\mathbf{V},k),\pi_n}(n)\right)^S$.
For $P(\mathscr{G})$ we take a uniform distribution over all graph structures subject to a fan-in restriction of $|\pi_n| \leq 3$, and for $P(K)$ we take a truncated Poisson distribution with $\lambda = 1$ restricted to $1 \leq K \leq K^\star$ as prior. In our applications we set $K^\star = 10$, i.e.

we restrict the maximal number of compartments K to 10. We note that the MCMC inference scheme, which we will discuss in the next subsection, does not sample $\mathbf{V}$ directly, but is based on local modifications of $\mathbf{V}$ based on changepoint birth, death and reallocation moves. That is, different from the free allocation in the BGM model [3], we here elect to change the assignment of data points to components via a changepoint process [2]. This reduces the complexity of the allocation space and incorporates our prior knowledge that adjacent time points are likely to be assigned to the same component. We identify K with $K-1$ changepoints: $b_1,\dots,b_{K-1}$ on the continuous interval $[2,m]$, and for notational convenience we introduce the pseudo-changepoints $b_0 = 2$ and $b_K = m$. The observation at time point t is assigned to the kth component, symbolically $\mathbf{V}(t) = k$, if $b_{k-1} \leq t < b_k$. Following [2] we assume that the changepoints are distributed as the even-numbered order statistics of $L := 2(K-1)+1$ points $u_1,\dots,u_L$ uniformly and independently distributed on the interval $[2,m]$. The motivation for this prior, instead of taking $K-1$ uniformly distributed points, is to encourage *a priori* equal spacings between changepoints, i.e. to discourage (too) short segments.

3.3 MCMC inference (extended version of the main paper)

We now describe an MCMC inference algorithm that can be used to obtain a sample $\{\mathscr{G}^i, \mathbf{V}^i, K^i\}_{i=1,\dots,I}$ from the posterior distribution $P(\mathscr{G},\mathbf{V},K|\mathbf{D})$. Our algorithm combines the structure MCMC algorithm for Bayesian networks [6] with the changepoint model (e.g. see [2]), and draws on the fact that conditional on the allocation vector $\mathbf{V}$, separate BGe scores $P(\mathbf{D}^{(\mathbf{V},k)}|\mathscr{G})$ can be computed for the K data compartments. Note that this approach is equivalent to the idea underlying the allocation sampler [7]. The resulting algorithm is effectively an RJMCMC sampling scheme in the discrete space of network structures and latent allocation vectors, where the Jacobian in the acceptance criterion is always 1 and can be omitted. With probability $p = 0.5$ we perform a traditional structure MCMC move on the current graph $\mathscr{G}^i$ and leave the latent vector $\mathbf{V}$ and the number of mixture components K unchanged, symbolically: $\mathbf{V}^{i+1} = \mathbf{V}^i$ and $K^{i+1} = K^i$. A new candidate graph $\mathscr{G}^{i+1}$ is randomly drawn out of the set of graphs $\mathscr{N}(\mathscr{G}^i)$ that can be reached from the current graph $\mathscr{G}^i$ by deletion or addition of one single edge. The proposed graph $\mathscr{G}^{i+1}$ is accepted with probability:

$$A(\mathscr{G}^{i+1}|\mathscr{G}^i) = min\left\{1, \frac{P(\mathbf{D}|\mathscr{G}^{i+1},\mathbf{V}^i,K^i)}{P(\mathbf{D}|\mathscr{G}^i,\mathbf{V}^i,K^i)} \cdot \frac{P(\mathscr{G}^{i+1})}{P(\mathscr{G}^i)} \cdot \frac{|\mathscr{N}(\mathscr{G}^i)|}{|\mathscr{N}(\mathscr{G}^{i+1})|}\right\} \tag{20}$$

where $|.|$ is the cardinality, and the marginal likelihood terms have been specified in Eq. (18). The graph is left unchanged, symbolically $\mathscr{G}^{i+1} := \mathscr{G}^i$, if the move is not accepted. We note that the network reconstruction will be based on the marginal posterior probabilities of the individual edges, which can be estimated for each edge from the MCMC sample $\mathscr{G}^1,\dots,\mathscr{G}^I$ by the fraction of graphs in the sample that contain the edge of interest:

$$\widehat{E_{i,j}} = \sum_{k=1}^{I} I_{i,j}(\mathscr{G}^k) \tag{21}$$

where $I_{i,j}(.)$ is the indicator function with $I_{i,j}(\mathscr{G}^k) = 1$ if there is an edge from X_i to X_j in $\mathscr{G}^k$.

With the complementary probability $1-p$ we leave the graph unchanged: $\mathscr{G}^{i+1} = \mathscr{G}^i$, and perform a move on $(\mathbf{V}^i, K^i)$. We change the current number of components K^i via a changepoint birth or death move, or the allocation vector $\mathbf{V}^i$ by a changepoint reallocation move along the lines of the Reversible Jump Markov chain Monte Carlo algorithm (RJMCMC) algorithm [2].

The changepoint birth (death) move increases (decreases) K^i by 1 and may also have an effect on $\mathbf{V}^i$. The changepoint reallocation move leaves K^i unchanged and may have an effect on $\mathbf{V}^i$. If with probability $(1-p)$ a changepoint move on $(K^i, \mathbf{V}^i)$ is performed, we randomly draw the move type. Under fairly mild regularity conditions (ergodicity), the MCMC sampling scheme converges to the desired posterior distribution if the equation of detailed balance is fulfilled [2]. The condition of detailed balance implies that for each move a complementary move is defined, and that the acceptance probability depends on the proposal probability of the complementary move. The moves presented below are designed such that there is a unique complementary death move for each birth move and vice-versa. Moreover, each reallocation move can be reversed by a single (complementary) reallocation move. The acceptance probabilities for these three changepoint moves $(K^i, \mathbf{V}^i) \rightarrow (K^{i+1}, \mathbf{V}^{i+1})$ are of the following form [2]:

$$A = min\left\{1, \frac{P(\mathbf{D}|\mathscr{G}^i, \mathbf{V}^{i+1}, K^{i+1})}{P(\mathbf{D}|\mathscr{G}^i, \mathbf{V}^i, K^i)} \times R \times B\right\} \tag{22}$$

where $R = P(\mathbf{V}^{i+1}|K^{i+1})P(K^{i+1})/P(\mathbf{V}^i|K^i)P(K^i)$ is the prior probability ratio, and B is the inverse proposal probability ratio. The exact form of these factors depends on the move type. (i) For a changepoint reallocation (r) we randomly select one of the existing changepoints $b_j \in \{b_1, \ldots, b_{K-1}\}$, and the replacement value $b_j^\dagger$ is drawn from a uniform distribution on $[b_{j-1}, b_{j+1}]$ where $b_0 = 2$ and $b_K = m$. Hence, the proposal probability ratio is one, the prior probabilities $P(K^{i+1}) = P(K^i)$ cancel out, and the remaining prior probability ratio $P(\mathbf{V}^{i+1}|K^{i+1})/P(\mathbf{V}^i|K^i)$ can be obtained from p.720 in Green's RJMCMC paper [2]:

$$R_r = \frac{(b_{j+1} - b_j^\dagger)(b_j^\dagger - b_{j-1})}{(b_{j+1} - b_j)(b_j - b_{j-1})}, \quad B_r = 1 \tag{23}$$

If there is no changepoint ($K^i = 1$) the move is rejected and the Markov chain is left unchanged. (ii) If a changepoint birth move (b) on K^i is proposed, the location of the new changepoint $b^\dagger$ is randomly drawn from a uniform distribution on the interval $[2, m]$; the proposal probability for this move is $b_K^i/(m-2)$, where b_{K^i} is the (K^i-dependent) probability of selecting a birth move. The reverse death move, which is selected with probability $d_{(K^i+1)}$, consists in discarding randomly one of the $(K^i - 1) + 1 = K^i$ changepoints. The inverse proposal probability ratio is thus given by $B = d_{(K^i+1)}(m-2)/b_{K^i}K^i$. The prior probability ratio is given by the expression at the bottom of p.720 in Green's RJMCMC paper [2] slightly modified to allow for the fact that K components correspond to $K-1$ changepoints, and we obtain:

$$R_b = \frac{P(K^i+1)}{P(K^i)} \frac{2K^i(2K^i+1)}{(m-2)^2} \frac{(b_{j+1}-b^\dagger)(b^\dagger-b_j)}{(b_{j+1}-b_j)}, \; B_b = \frac{d_{(K^i+1)}(m-2)}{b_{K^i}K^i} \tag{24}$$

For $K^i = K^\star$ the birth of a new changepoint is invalid and the Markov chain is left unchanged.

We note that the proposal probabilities b_K and $d_{(K+1)}$ for birth and death moves can be chosen as follows:

$$b_K = c \cdot min\left\{1, \frac{P(K+1)}{P(K)}\right\}, \; d_{(K+1)} = c \cdot min\left\{1, \frac{P(K)}{P(K+1)}\right\} \tag{25}$$

where c is a constant that can be choosen as large as possible subject to the constraint $b_K + d_K \leq 0.9$ for $K = 1, \ldots, K^\star$. This choice yields both a certain acceptance rate of the MCMC sampling scheme [2] and a simple prior probability (Hastings) ratio. From $b_{K^i} \cdot P(K^i) = d_{(K^i+1)} \cdot P(K^i+1)$ it follows that the ratio $d_{(K^i+1)}/b_{K^i}$ in the expression R_b cancels out against the prior ratio $P(K^i+1)/P(K^i)$ in the expression B_b, and the prior probability ratio simplifies to:

$$R_b B_b = \frac{2(2K^i+1)}{(m-2)} \frac{(b_{j+1}-b^\dagger)(b^\dagger-b_j)}{(b_{j+1}-b_j)} \tag{26}$$

(iii) A changepoint death move (d) is the reverse of the birth move, and we obtain:

$$R_d B_d = \frac{(m-2)}{2(2K^i-1)} \frac{(b_{j+1}-b_j)}{(b_{j+1}-b^\dagger)(b^\dagger-b_j)} \tag{27}$$

References

1. Geiger, D., Heckerman, D.: Learning Gaussian networks. Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence pp. 235–243 (1994)
2. Green, P.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995)
3. Grzegorczyk, M., Husmeier, D., Edwards, K., Ghazal, P., Millar, A.: Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics **24**, 2071–2078 (2008)
4. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics **19**, 2271–2282 (2003)
5. Ko, Y., Zhai, C., Rodriguez-Zas, S.: Inference of gene pathways using Gaussian mixture models. In: BIBM International Conference on Bioinformatics and Biomedicine, pp. 362–367. Fremont, CA (2007)
6. Madigan, D., York, J.: Bayesian graphical models for discrete data. International Statistical Review **63**, 215–232 (1995)
7. Nobile, A., Fearnside, A.: Bayesian finite mixtures with an unknown number of components: The allocation sampler. Statistics and Computing **17**(2), 147–162 (2007)