BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 18 2008, pages 2071–2078
doi:10.1093/bioinformatics/btn367

Systems biology

Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler

Marco Grzegorczyk[1,2,*], Dirk Husmeier[2,3,*], Kieron D. Edwards[4], Peter Ghazal[2,5] and Andrew J. Millar[1,2]

[1]School of Biological Sciences, The University of Edinburgh, Swann Building, King's Buildings, Edinburgh EH9 3JR, [2]Centre for Systems Biology at Edinburgh (CSBE), Darwin Building, King's Buildings, Edinburgh EH9 3JU, [3]Biomathematics and Statistics Scotland (BioSS), JCMB, King's Buildings, Edinburgh EH9 3JZ, [4]Advanced Technologies (Cambridge) Ltd, Cambridge CB4 0WA and [5]Division of Pathway Medicine (DPM), Medical School, The University of Edinburgh, Chancellor's Buildings, Edinburgh EH16 4SB, UK

Received on May 17, 2008; revised on July 9, 2008; accepted on July 14, 2008

Advance Access publication July 28, 2008

Associate Editor: Trey Ideker

ABSTRACT

Method: The objective of the present article is to propose and evaluate a probabilistic approach based on Bayesian networks for modelling non-homogeneous and non-linear gene regulatory processes. The method is based on a mixture model, using latent variables to assign individual measurements to different classes. The practical inference follows the Bayesian paradigm and samples the network structure, the number of classes and the assignment of latent variables from the posterior distribution with Markov Chain Monte Carlo (MCMC), using the recently proposed allocation sampler as an alternative to RJMCMC.

Results: We have evaluated the method using three criteria: network reconstruction, statistical significance and biological plausibility. In terms of network reconstruction, we found improved results both for a synthetic network of known structure and for a small real regulatory network derived from the literature. We have assessed the statistical significance of the improvement on gene expression time series for two different systems (viral challenge of macrophages, and circadian rhythms in plants), where the proposed new scheme tends to outperform the classical BGe score. Regarding biological plausibility, we found that the inference results obtained with the proposed method were in excellent agreement with biological findings, predicting dichotomies that one would expect to find in the studied systems.

Availability: Two supplementary papers on theoretical (T) and experimental (E) aspects and the datasets used in our study are available from http://www.bioss.ac.uk/associates/marco/supplement/

Contact: marco@bioss.ac.uk, dirk@bioss.ac.uk

1 INTRODUCTION

The ultimate objective of systems biology is the elucidation of the regulatory networks and signalling pathways of the cell. The ideal approach would be the deduction of a detailed mathematical description of the entire system in terms of a set of coupled non-linear differential equations. As high-throughput measurements at the cell level are inherently stochastic and most kinetic rate constants cannot be measured directly, the parameters of the system would have to be estimated from the data. Unfortunately, multiple parameter sets of non-linear systems of differential equations can offer equally plausible solutions, and standard optimization techniques in high-dimensional multimodal parameter spaces are not robust and do not provide a reliable indication of the confidence intervals. Most importantly, model selection would be impeded by the fact that more complex pathway models would always provide a better explanation of the data than less complex ones, rendering this approach intrinsically susceptible to over-fitting.

To assist the elucidation of regulatory network structures, probabilistic machine learning methods based on Bayesian networks can be employed, as proposed in the seminal paper by Friedman *et al.* (2000). In a nutshell, the idea is to simplify the mathematical description of the biological system by replacing coupled differential equations by simple conditional probability distributions of a standard form such that the unknown parameters can be integrated out analytically. This results in a scoring function (the 'marginal likelihood') of closed form that depends only on the structure of the regulatory network and avoids the over-fitting problem referred to above. Novel fast Markov Chain Monte Carlo (MCMC) algorithms, like Grzegorczyk and Husmeier (2008), can be applied to systematically search the space of network structures for those that are most consistent with the data. To obtain the closed form expression of the marginal likelihood referred to above, two probabilistic models with their respective conjugate prior distributions have been employed in the past: the multinomial distribution with the Dirichlet prior, leading to the so-called BDe score (Cooper and Herskovits, 1992), and the linear Gaussian distribution with the normal-Wishart prior, leading to the BGe score (Geiger and Heckerman, 1994). These approaches are restricted in that they either require the data to be discretized (BDe) or can only capture linear regulatory relationships (BGe). A non-linear non-discretized model based on heteroscedastic regression has been proposed by Imoto *et al.* (2003). However, this approach no longer allows the marginal likelihood to be obtained in closed form and requires a restrictive approximation (the Laplace approximation)

*To whom correspondence should be addressed.

to be adopted. Another non-linear model based on node-specific Gaussian mixture models has been proposed in Ko *et al.* (2007). Again, the marginal likelihood is intractable. The authors resort to the Bayesian information criterion (BIC) of Schwarz (1978) for model selection, which is only a good approximation to the marginal likelihood in the limit of very large datasets. In the present article we propose a non-linear generalization of the BGe score, which is motivated by the fact that any probability distribution can, in principle, be approximated arbitrarily closely by a mixture model. Our method is based on recent work by Nobile and Fearnside (2007), who proposed the allocation sampler as an alternative to the computationally expensive approach of reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). We will describe the method in Section 2. We have evaluated our approach on a set of synthetic and real-world datasets described in Section 3 according to criteria outlined in Section 4. The results are presented in Section 5 and discussed in Section 6. A concluding summary of the proposed method and the results can be found in Section 7.

2 METHOD

We focus on the most important methodological aspects. A more detailed representation can be found in Supplementary Material T.

2.1 Bayesian network methodology

Static Bayesian networks (BNs) are interpretable and flexible models for representing probabilistic relationships between interacting variables. At a qualitative level, the graph of a BN describes the relationships between the domain variables in the form of conditional independence relations. At a quantitative level, local relationships between variables are described by conditional probability distributions. Formally, a BN is defined by a graph $\mathcal{G}$, a family of conditional probability distributions F, and their parameters q, which together specify a joint distribution over the domain variables.

The graph $\mathcal{G}$ of a BN consists of a set of N nodes (variables) $X_1,...,X_N$ and a set of directed edges between these nodes. The *parent set* of node X_n, symbolically π_n, is defined as the set of all parent nodes of X_n, that is, the set of nodes from which an edge points to X_n in $\mathcal{G}$. The structure of a static BN is defined to be a *directed acyclic graph* (DAG), that is, a directed graph without any cycles of directed edges (loops). It is due to this acyclicity constraint that the joint probability distribution in BNs can be uniquely factorized as follows:

$$P(X_1,...,X_N)=\prod_{n=1}^{N}P(X_n|\pi_n) \tag{1}$$

Stochastic models for Bayesian networks (Friedman *et al.*, 2000) specify the distributional form of the local probability distributions $P(X_n|\pi_n)$. Given data $\mathcal{D}$ and a parametric model, (DAGs), $\mathcal{G}$ can be scored with respect to their posterior probabilities:

$$P(\mathcal{G}|\mathcal{D})=\frac{P(\mathcal{G},\mathcal{D})}{P(\mathcal{D})}=\frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{\sum_{\mathcal{G}^\star}P(\mathcal{D}|\mathcal{G}^\star)P(\mathcal{G}^\star)}, \tag{2}$$

where $P(\mathcal{D}|\mathcal{G})$ is the marginal likelihood and $P(\mathcal{G})$ is the prior distribution over the space of graphs. For two stochastic models BDe and BGe a closed-form solution can be derived for the likelihood $P(\mathcal{D}|\mathcal{G})$ (Cooper and Herskovits, 1992; Geiger and Heckerman, 1994).

When time series data $(X_{1,t},\dots X_{N,t})_{t=1,...,m}$ have been collected, dynamic Bayesian networks (DBNs) can be employed. In DBNs edges correspond to interactions with a time delay τ; e.g. for $\tau=1$ an edge pointing from X_i to X_j means that the realization of X_j at time point t is influenced by the realization of X_i at the previous time point $t-1$. In DBNs parameters are tied such that the transition probabilities between time slices $t-1$ and t are the same for all t, resulting in a homogeneous Markovian dependence. Because of the time delay of interactions and the bipartite graph structure thus imposed, the acyclicity of the underlying graph $\mathcal{G}$ is automatically guaranteed, and Equation (1) is replaced by:

$$P(X_{1,t},...,X_{N,t})=\prod_{n=1}^{N}P(X_{n,t}|\pi_{n,t-1}) \tag{3}$$

where $\pi_{n,t-1}$ denotes the parent set of X_n at the previous time point $t-1$. For more details see Friedman *et al.* (1998).

MCMC methods can be used for sampling DAGs $\mathcal{G}$ from the posterior distribution $P(\mathcal{G}|\mathcal{D})$. The structure MCMC approach of Madigan and York (1995) generates a sample of graphs $\mathcal{G}_1,...,\mathcal{G}_T$ as follows: given a graph $\mathcal{G}_i$, a new candidate graph $\mathcal{G}_{i+1}$ is proposed with probability:

$$Q(\mathcal{G}_{i+1}|\mathcal{G}_i)=\left\{\begin{array}{ll}\frac{1}{|\mathcal{N}(\mathcal{G}_i)|}, & \mathcal{G}_{i+1}\in\mathcal{N}(\mathcal{G}_i)\\ 0\quad\quad, & \mathcal{G}_{i+1}\notin\mathcal{N}(\mathcal{G}_i)\end{array}\right\} \tag{4}$$

where $\mathcal{N}(\mathcal{G}_i)$ denotes the *neighbourhood* of $\mathcal{G}_i$, that is the collection of all valid graphs that can be reached from $\mathcal{G}_i$ by deletion, addition or reversal of one single edge of the current graph $\mathcal{G}_i$, and $|\mathcal{N}(\mathcal{G}_i)|$ is the cardinality of this collection. We note that all neighbour graphs $\mathcal{G}_{i+1}$ have to be acyclic when non-dynamic BNs are employed. The graph $\mathcal{G}_{i+1}$ is accepted with probability:

$$A(\mathcal{G}_{i+1}|\mathcal{G}_i)=min\left\{1,\frac{P(\mathcal{D}|\mathcal{G}_{i+1})P(\mathcal{G}_{i+1})}{P(\mathcal{D}|\mathcal{G}_i)P(\mathcal{G}_i)}\cdot\frac{|\mathcal{N}(\mathcal{G}_i)|}{|\mathcal{N}(\mathcal{G}_{i+1})|}\right\} \tag{5}$$

otherwise the chain is left unchanged, symbolically $\mathcal{G}_{i+1}:=\mathcal{G}_i$. The Markov chain $\{\mathcal{G}_i\}$ converges to the posterior distribution $P(\mathcal{G}|\mathcal{D})$ (Madigan and York, 1995). If a *fan-in* restriction is imposed on the cardinality of the parent sets, all graphs possessing a node with more than fan-in parent nodes have to be excluded from the graph neighbourhoods. Structure MCMC generates a graph sample $\{\mathcal{G}_1,\dots,\mathcal{G}_T\}$, from which posterior probabilities of edges can be computed. We focus on *undirected edges* for independent data and *directed edges* for time-dependent data. There is an undirected edge between X_i and X_j $(i<j)$ in $\mathcal{G}$, if $\mathcal{G}$ possesses either the edge $X_i\rightarrow X_j$ or the edge $X_i\leftarrow X_j$. Likewise, there is a directed edge from X_i to X_j $(i\neq j)$ in $\mathcal{G}$, if $\mathcal{G}$ possesses the edge $X_i\rightarrow X_j$. An estimator for the posterior probabilities of an edge is given by the fraction of graphs in the sample that contain the edge of interest. When the true graph for the domain is known, the concept of *receiver operator characteristic* (ROC) curves and *area under receiver operator characteristic* (AUROC) values can be used to evaluate the global network reconstruction accuracy of BN inference (see e.g. Husmeier (2003) for details). An alternative and more intuitive criteria is given by (TP|FP=5) counts: for each MCMC output a threshold ψ is imposed on the inferred edge posterior probabilities such that five false positive (FP) edges are extracted and the corresponding number of true positive (TP) edges, symbolically (TP|FP=5), exceeding the threshold ψ, is counted (Werhli *et al.*, 2006).

2.2 Gaussian mixture Bayesian network model

We assume that we have either m independent and identically distributed (iid) observations (BNs) or $m+1$ time-dependent observations with a homogeneous first-order Markovian dependence structure (DBNs) for the variables $X_1,\dots,X_N$. This gives a dataset matrix of size N-by-m, where $\mathcal{D}_j$ $(j=1,\dots,m)$ is the j-th observation of the N nodes. The allocation vector $\vec{\mathcal{V}}$ of size m defines an allocation of the m observations to $\mathcal{K}$ mixture components: $\vec{\mathcal{V}}(j)=k$ means that the j-th observation is allocated to the k-th component. $\mathcal{D}^{(\vec{\mathcal{V}},k)}$ denotes the data subset consisting of all observations allocated to the k-th component by $\vec{\mathcal{V}}$ $(1\leq k\leq\mathcal{K})$. The joint posterior probability of a graph $\mathcal{G}$, an allocation vector $\vec{\mathcal{V}}$, and $\mathcal{K}$ mixture components can be factorized as follows:

$$\begin{aligned}P(\mathcal{G},\vec{\mathcal{V}},\mathcal{K}|\mathcal{D})&=\frac{P(\mathcal{G},\vec{\mathcal{V}},\mathcal{K},\mathcal{D})}{P(\mathcal{D})}\propto P(\mathcal{G},\vec{\mathcal{V}},\mathcal{K},\mathcal{D})\\&=P(\mathcal{K})P(\vec{\mathcal{V}}|\mathcal{K})P(\mathcal{G})P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})\end{aligned} \tag{6}$$

where

$$P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})=\prod_{k=1}^{\mathcal{K}}P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G}) \quad (7)$$

In Equation (7) the likelihood terms $P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})$ for the data subsets $\mathcal{D}^{(\vec{\mathcal{V}},k)}$ given the same graph $\mathcal{G}$ can be computed independently with the BGe scoring metric of Geiger and Heckerman (1994), as derived and discussed in Supplementary Materials T. If no observation is allocated to the k-th component ($\mathcal{D}^{(\vec{\mathcal{V}},k)}=\emptyset$), then $P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})$ is equal to 1. As we do not have any prior knowledge about the graph topology we assume a uniform prior distribution on graphs for the real gene expression data, $P(\mathcal{G})=const$. For the synthetic Raf-Mek-Erk network data we employ a more restrictive graph prior (see Supplementary Materials T and E). For the prior on $\mathcal{K}$, $P(\mathcal{K})$, we take a truncated Poisson distribution with parameter $\lambda=1$ restricted to $1\leq\mathcal{K}\leq\mathcal{K}_{\mathrm{MAX}}$. This prior is known to be suitable for finite mixture models (Nobile, 2005). We further assume that the probability distribution of the allocation vector $\vec{\mathcal{V}}$ conditional on $\mathcal{K}$ is given by:

$$P(\vec{\mathcal{V}}|\mathcal{K},\vec{p})=\prod_{k=1}^{\mathcal{K}}p_k^{n_k} \quad (8)$$

where $\vec{p}=(p_1,\ldots,p_{\mathcal{K}})^T$ with $\sum_{k=1}^{\mathcal{K}}p_k=1$ are the non-negative mixture weights, and n_k is the number of observations allocated to the k-th mixture component by $\vec{\mathcal{V}}$. The prior on the mixture weights $\vec{p}=(p_1,\ldots,p_{\mathcal{K}})^T$ is chosen to be a Dirichlet distribution, $P(\vec{p})=Dir(\alpha_1,\ldots,\alpha_{\mathcal{K}})$, with hyperparameters $\vec{\alpha}=(\alpha_1,\ldots,\alpha_{\mathcal{K}})^T$. This prior is conjugate, and the marginal probability of $\vec{\mathcal{V}}$ conditional on $\mathcal{K}$ is thus given by

$$P(\vec{\mathcal{V}}|\mathcal{K})=\int P(\vec{\mathcal{V}}|\mathcal{K},\vec{p})P(\vec{p})d\vec{p}=Dir(n_1+\alpha_1,\ldots,n_{\mathcal{K}}+\alpha_{\mathcal{K}}) \quad (9)$$

2.3 Gaussian mixture allocation MCMC inference

The new Gaussian mixture allocation MCMC sampling scheme (BGM) generates a sample from the joint posterior distribution $P(\mathcal{G},\mathcal{K},\vec{\mathcal{V}}|\mathcal{D})$ given in Equation (6) and comprises six different types of moves in the state-space $[\mathcal{G},\mathcal{K},\vec{\mathcal{V}}]$. The first move type is a structure MCMC single edge operation on the graph $\mathcal{G}$ while the number of components $\mathcal{K}$ and the allocation vector $\vec{\mathcal{V}}$ are left unchanged. According to Equation (4), a new graph $\tilde{\mathcal{G}}$ is proposed, and the new state $[\tilde{\mathcal{G}},\mathcal{K},\vec{\mathcal{V}}]$ is accepted or rejected according to Equation (5) where the likelihood terms $P(\mathcal{D}|\mathcal{G})$ in Equation (5) have to be replaced by the $P(\mathcal{D}|\mathcal{G},\mathcal{K},\vec{\mathcal{V}})$ terms given in Equation (7). The five other move types are adapted from Nobile and Fearnside (2007) and operate on $\vec{\mathcal{V}}$ or on $\mathcal{K}$ and $\vec{\mathcal{V}}$. If there are $\mathcal{K}>2$ mixture components, then moves of the type M1 and M2 can be used to re-allocate some observations from one component k_1 to another one k_2. That is, a new allocation vector $\vec{\mathcal{V}}^*$ is proposed while $\mathcal{G}$ and $\mathcal{K}$ are left unchanged. The Ejection move type proposes an increase in the number of mixture components by one and simultaneously tries to re-allocate some observations to fill the new component. More precisely, it randomly selects a mixture component and tries to re-allocate some of its observations to the newly proposed component $\mathcal{K}+1$, while $\mathcal{G}$ is left unchanged. The Absorption move is complementary to the Ejection move and decreases the number of mixture components by one. It randomly selects two mixture components and deletes one of them after having reallocated all of its observations to the other component. The acceptance probabilities for M1, M2, Ejection and Absorption moves are of the same functional form:

$$A=\left\{1,\frac{P(\vec{\mathcal{V}}^*|\mathcal{K})}{P(\vec{\mathcal{V}}|\mathcal{K})}\cdot\frac{P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})}{P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K}^*)}\cdot\frac{Q(\vec{\mathcal{V}}^*|\vec{\mathcal{V}})}{Q(\vec{\mathcal{V}}|\vec{\mathcal{V}}^*)}\cdot\frac{P(\mathcal{K}^*)}{P(\mathcal{K})}\right\} \quad (10)$$

where the likelihood terms have been specified in Equation (7), the proposal probabilities $Q(.|.)$ depend on the move type (M1, M2, Ejection or Absorption), and $\mathcal{K}^*=\mathcal{K}$ for M1 and M2 moves $\mathcal{K}^*=\mathcal{K}+1$ for Ejection moves, and $\mathcal{K}^*=\mathcal{K}-1$ for Absorption moves. See Supplementary Material T and Nobile and Fearnside (2007) for details. Finally, the sixth move type uses Gibbs sampling to re-allocate a single observation by sampling its new allocation from the corresponding univariate conditional distribution, while leaving $\mathcal{K}$ and the other components of $\vec{\mathcal{V}}$ unchanged.

3 DATA

We have evaluated the proposed method on synthetic data generated from a widely studied protein signalling network and on gene expression time series from two different biological systems. For details of the simulation studies see Supplementary Material E.

3.1 Synthetic data

For a comparative evaluation study, Werhli *et al.* (2006) generated synthetic datasets for the Raf-Mek-Erk signalling pathway presented in Sachs *et al.* (2005), which consists of 11 nodes representing phosphorylated proteins and 20 directed edges. As Werhli *et al.* (2006) assigned Gaussian regulatory mechanisms with varying (randomly sampled) parameters, we can generate Gaussian mixture network data as follows: for obtaining data with $\mathcal{K}=1,\ldots,5$ components we randomly selected $\mathcal{K}$ of the original datasets, sampled the same number of observations $m/\mathcal{K}$ from each and merged these observations to a single dataset of size m. For each of 16 combinations of m ($m=30,60,120,180$) and $\mathcal{K}$ we generated five datasets by applying this procedure. Additionally, for each $\mathcal{K}=1,\ldots,5$ we generated five further datasets along this line Werhli *et al.* (2006) with $m=480$ observations each.

3.2 Bone marrow-derived macrophages

Interferons (IFNs) play a pivotal role in the innate and adaptive mammalian immune response against infection, and central research efforts, therefore, aim to elucidate their regulatory interactions (Honda *et al.*, 2006). For the present study, we have applied our method to gene expression time series from bone marrow-derived macrophages, which were sampled at 24×30 min time intervals. The macrophages were subjected to three external conditions: (1) infection with Cytomegalovirus (CMV), (2) treatment with Interferon Gamma (IFNγ) and (3) infection with Cytomegalovirus after pretreatment with IFNγ (CMV+IFNγ). To obtain the gene expression profiles, samples derived from the macrophages were hybridized to Agilent mouse genome arrays. Samples were co-hybridized with a pooled common control RNA. Expression levels were obtained in the form of $\log_2$ scale signal intensity ratios between the sample and the pooled control RNA. Differential dye-label incorporation between the two samples on each array was corrected by applying a within-array, non-linear, loess normalization to the ratios. Global non-biological variations between ratio distributions were corrected by applying median-absolute-deviation between-array normalization. We focus on time series of the Interferon regulatory factors (Irfs) 1, 2 and 3 (which we write as Irf1, Irf2 and Irf3, respectively), as a gold standard network for the interactions between these factors can be derived from the literature (Darnell *et al.*, 1994; Raza *et al.*, 2008): Irf2 $\leftrightarrow$ Irf1 $\leftarrow$ Irf3. The Irfs are the key regulators in the response of the macrophage cell to pathogens. They mediate the cellular signalling that leads to a transcriptional response to the initial binding events on the surface of the cell.

3.3 Circadian regulation in *Arabidopsis thaliana*

We have also applied our method to two gene expression time series from *A.thaliana* cells, which were sampled at 13×2 h time intervals with Affymetrix microarray chips, and robust multi-array (RMA) normalized. The expressions were measured twice independently

under experimentally generated constant light condition, but differed with respect to the prehistories. In the first experiment, T_{20}, the plant was entrained in a 10h:10h light/dark cycle, while the plant in the second experiment, T_{28}, was entrained in 14h:14h light/dark cycle. Our analysis focuses on nine genes, namely LHY, CCA1, TOC1, ELF4, ELF3, GI, PRR9, PRR5 and PRR3, which are known to be involved in circadian regulation (Mas, 2008; Salome and McClung, 2004).

4 EVALUATION

To evaluate the proposed method (BGM), we have compared it with Bayesian learning of homogeneous BNs using the standard BGe score, as described in Geiger and Heckerman (1994). We have applied a 3-fold evaluation procedure. First, we use static synthetic network data to show that if the data are, in fact, of a heterogeneous nature, BGM achieves improved network reconstruction results. Second, using gene expression time series from bone marrow-derived macrophages, we focus on a small subsystem of the IFN pathway whose biology is well understood, and we demonstrate that BGM leads to a better pathway reconstruction. Third, we consider a larger set of circadian genes from *A.thaliana.* Since the true network structure in this case is not known, we apply two standard methods from statistics for the evaluation: Bayes factors and predictive distributions. We briefly describe these methods in the remainder of this section. The mathematical details can be found in Supplementary Material T. We want to compare two competing hypotheses. According to the null hypothesis H_0, the conventional homogeneous DBN (BGe) is the adequate model. We want to compare this with the alternative hypothesis H_1 that the proposed non-homogeneous DBN (BGM) provides the right description of the system. We want to pursue a Bayesian approach, according to which the decision between the two hypotheses is based on the Bayes factor: $P(\mathcal{D}|H_1)/P(\mathcal{D}|H_0)$. Note that the two hypotheses are nested, and that $P(\mathcal{D}|H_0)=P(\mathcal{D}|\mathcal{K}{=}1,H_1)$. We can therefore follow Huelsenbeck *et al.* (2004) and calculate the Bayes factor using the Savage-Dickey ratio (Verdinelli and Wasserman, 1995):

$$\frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_0)}=\frac{P(K=1|H_1)}{P(\mathcal{K}=1|\mathcal{D},H_1)} \tag{11}$$

where $\mathcal{K}$ is the number of mixture components (segments). The validity of Equation (11) can easily be proven from Bayes rule:

$$P(\mathcal{K}=1|\mathcal{D},H_1)=\frac{P(\mathcal{D}|\mathcal{K}=1,H_1)P(\mathcal{K}=1|H_1)}{P(\mathcal{D}|H_1)} \tag{12}$$

As an alternative procedure, we adopt an approach based on the predictive distribution promoted in Vehtari and Lampinen (2002). However, as opposed to the authors we do not resort to a cross-validation procedure, but exploit the fact that in our experiments gene expressions were obtained under different experimental conditions: CMV, IFN$_\gamma$ and CMV+IFN$_\gamma$ (macrophages) or T_{20} and T_{28} (circadian genes), respectively. Denote by $\mathcal{D}$ the gene expression data obtained under a condition used for training. Denote by $\tilde{\mathcal{D}}$ the gene expression data obtained from a separate experiment under a different condition. We can then base the hypothesis test on a comparison of the predictive distributions $P(\tilde{\mathcal{D}}|\mathcal{D},H_1)$ and $P(\tilde{\mathcal{D}}|\mathcal{D},H_0)$. Note that these distributions measure how well new independent test data $\tilde{\mathcal{D}}$ can be predicted under the two hypotheses, using the training data $\mathcal{D}$. As before, $\mathcal{K}$ denotes the number of mixture components, $\vec{\mathcal{V}}$ denotes the allocation vector, $\mathcal{G}$ denotes the graph, and let $\vec{q}$ denote the vector of parameters associated with $\mathcal{G}$. We get the following expression for the predictive distribution:

$$P(\tilde{\mathcal{D}}|\mathcal{D},H_i)=\sum_{\mathcal{K},\vec{\mathcal{V}},\mathcal{G}}\int P(\tilde{\mathcal{D}}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q},H_i)P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q}|\mathcal{D},H_i)d\vec{q} \tag{13}$$

A possible approach is to approximately sample $[\mathcal{K},\vec{\mathcal{V}},\mathcal{G}]$ and $\vec{q}$ from the posterior distribution $P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q}|\mathcal{D},H_i)$ with MCMC and to approximate the integral in Equation (13) by a sum over this sample. A better method is to use the expansion $P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q}|\mathcal{D},H_i)=P(\vec{q}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\mathcal{D},H_i)P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G}|\mathcal{D},H_i)$ and draw on the fact that

$$\Psi(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\tilde{\mathcal{D}})=\int P(\tilde{\mathcal{D}}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q},H_i)P(\vec{q}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\mathcal{D},H_i)d\vec{q} \tag{14}$$

can be calculated analytically (Geiger and Heckerman, 1994) and Supplementary Material T). Inserting (14) in (13) yields:

$$P(\tilde{\mathcal{D}}|\mathcal{D},H_i)=\sum_{\mathcal{K},\vec{\mathcal{V}},\mathcal{G}}\Psi(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\tilde{\mathcal{D}})P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G}|\mathcal{D},H_i) \tag{15}$$

which in practice is computed from a sample $\{[\mathcal{K}_1,\vec{\mathcal{V}}_1,\mathcal{G}_1],\ldots,[\mathcal{K}_T,\vec{\mathcal{V}}_T,\mathcal{G}_T]\}$ approximately drawn from the posterior distribution $P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G}|\mathcal{D},H_i)$ with MCMC:

$$P(\tilde{\mathcal{D}}|\mathcal{D},H_i)=\frac{1}{T}\sum_{i=1}^{T}\Psi(\mathcal{K},\vec{\mathcal{V}}_i,\mathcal{G}_i,\tilde{\mathcal{D}}) \tag{16}$$

The computation of $\Psi(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\tilde{\mathcal{D}})$ in (14) requires only a minor modification of the standard BGe score discussed in Geiger and Heckerman (1994). The vector $\vec{\mathcal{V}}$ acts as a filter dividing the data into different categories, for which separate BGe scores are computed. For instance, if we have 2 states, 10 time points and $\vec{\mathcal{V}}=[1111122222]$, then separate BGe scores are computed for the first five and the last five time points. The computation of the BGe score is modified by the fact that the prior distribution $P(\vec{q}|\mathcal{G},H_i)$ is replaced by the posterior distribution $P(\vec{q}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\mathcal{D},H_i)$. This results in a straightforward modification of the score as follows: in Equation (13) of Geiger and Heckerman (1994), those training data that correspond to the corresponding state k, $\{\mathcal{D}_j\in\mathcal{D}|\vec{\mathcal{V}}(j)=k\}$, are included in the conditioning part of the distribution, and the sufficient statistics are adjusted accordingly. We note that BDe and BGM cannot be compared in terms of predictive distributions, as the required data discretization (BDe) is not part of the BN model. That is, while BGM and BGe model the same datasets $\mathcal{D}$ and $\tilde{\mathcal{D}}$, BDe is based on their discretized counterparts, resulting from some (heuristic) pre-processing.

5 RESULTS

The mean AUROC values and the mean (TP|FP=5) counts for assessing the reconstruction of the Raf-Mek-Erk pathway from the synthetic data, described in Section 3.1, are represented as histograms in Figures 1 and 2. It can be seen that BGM performs significantly better than BDe and BGe for almost all combinations of $\mathcal{K}$ and m. Only if there is either one single component or a small sample size ($m=30$), there is no (significant) difference between BGM and BGe. In particular for $\mathcal{K}=1$, BGM assigns all observations to one single component, and so does not differ

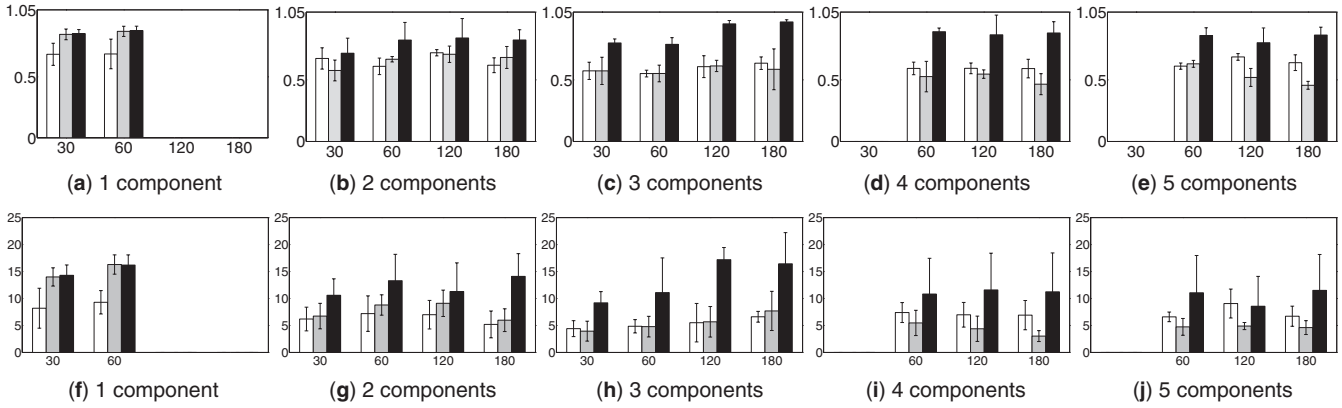


Fig. 1. Raf-Mek-Erk network reconstruction accuracy for synthetic data. Histograms of the network reconstruction accuracy for different combinations of $\mathcal{K}_{\text{TRUE}}$ ($\mathcal{K}_{\text{TRUE}} = 1, \ldots, 5$) and sample size m ($m = 30, 60, 120, 180$) assessed in terms of mean AUROC values (panels (a–e)) and (TP|FP = 5) counts (panels (f–j)) derived from undirected edges. White bars refer to BDe, grey bars refer to BGe, and black bars refer to BGM. The SDs are indicated by vertical black lines.

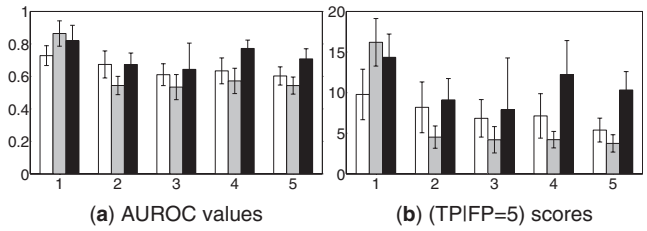


Fig. 2. Raf-Mek-Erk network reconstruction accuracy for synthetic data with m = 480. Histograms of the network reconstruction accuracy for $\mathcal{K}_{\text{TRUE}} = 1, \ldots, 5$ assessed in terms of mean AUROC values (**a**) and (TP|FP= 5) counts (**b**) derived from undirected edges. White bars refer to BDe, grey bars refer to BGe and black bars refer to BGM. The SDs are indicated by vertical black lines.

from BGe. Figure 3 reveals that BGM infers for each number of components $\mathcal{K}_{\text{TRUE}}$ the correct number of components for the synthetic data with $m = 480$ observations. Histograms of the numbers of inferred components for the synthetic data with fewer data points m are provided in Figure 1 in Supplementary Material E.

A comparison of Figures 1 and 2 reveals that the reconstruction accuracy is slightly worse for the datasets with $m = 480$ observations. This finding might appear counter intuitive, as larger datasets contain more information and should therefore lead to better performances. However, our finding is consistent with the fact that increased dataset sizes lead to likelihood landscapes that are more rugged and, hence, result in increased mixing and convergence problems. This shortcoming of the structure MCMC sampler by Madigan and York (1995) has already been reported (e.g. see Grzegorczyk and Husmeier, 2008).

For the macrophage gene expression time series, BGM infers $\mathcal{K} = 2$ components for the conditions CMV and IFNγ, while for the third condition (CMV+IFNγ) most of the sampled states consist of $\mathcal{K} = 1$ component only, as shown in Figure 4. The fraction of sampled states for which two observations i and j are allocated to the same component k ($1 \leq k \leq \mathcal{K}$) can be used as a connectivity measure $C(i, j)$. Figure 5 displays the resulting connectivity matrices graphically as heat matrices. From the heat matrices the same systematic trend can be observed for the three conditions. The first part (observations no. 2–6) and the last part of the three time series (observations no. 8–25) are allocated to different components. For condition CMV (IFNγ) the allocation of observation no. 7 (no. 9) is not fixed, that is, the allocation changes during the MCMC simulation. For CMV + IFNγ, whose number of components peaks at $\mathcal{K} = 1$ (Fig. 4), the separation between the two parts is less pronounced, though consistent with the other results. To understand whether BGM also leads to a better network reconstruction accuracy, we compare the mean posterior probabilities of the true and false edges of BGM in Figure 6 with those obtained from BGe and BDe. For the IFNγ condition (Fig. 6b) it becomes obvious that BGM has performed substantially better than BGe and BDe. For the other two conditions the difference between the posterior means for the true and the false edges is also best for BGM, but the difference is less pronounced [BDe: 0.24, BGe: 0.24, BGM: 0.39 (CMV) and BDe: -0.42 , BGe: 0.09, BGM: 0.14 (CMV + IFNγ)]. Since it appears that the three conditions do not lead to systematic deviations between the expression profiles of Irf1, Irf2 and Irf3, we treat the three experiments as independent replications and compute predictive probabilities, as discussed in Section 4. The predictive probabilities for BGM are much higher than those of BGe (Table 2). This finding provides further evidence that BGM does not overfit the data but outputs results that can be confirmed by independent replications. The BGM/BGe Bayes factors are: 36.45 (CMV), 2.73 (IFNγ) and 0.71 (CMV + IFNγ). This finding is consistent with Figure 4, where the peaks for CMV and IFNγ are at 2, while CMV + IFNγ peaks at 1. Gene expression time series plots and scatter plots for the three Irf factors can be found in Supplementary Material E.

For both *A.thaliana* gene expression time series (see Supplementary Material E) the number of components inferred with BGM peaks at 2 (Fig. 7a and b). The heat matrices shown in Figure 7a and b appear to be of a similar structure, but subject to a translation along the main diagonal. More precisely, it appears that the transition from the first to the second component is shifted by 2–3 time points (4–6 h). Compared with BGe the Bayes factors are in favour of BGM: 5.66 (T_{20}) and 9.41 (T_{28}). The predictive probabilities are given in Table 1 and confirm the improved generalization performance. Further plots for the *Arabidopsis* data

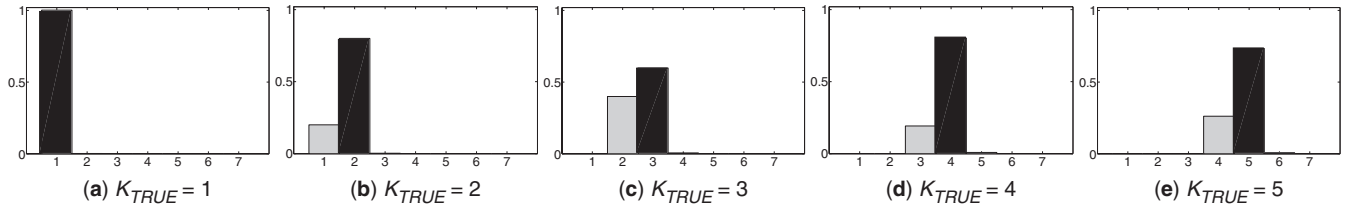


Fig. 3. Histograms of the numbers of BGM components for synthetic Gaussian data with m = 480 observations. The posterior probabilities (vertical axis) of the number of components $\mathcal{K}$ (horizontal axis) have been estimated from the MCMC trajectories. For $\mathcal{K}_{\mathrm{TRUE}} = 1,\ldots,5$ the MCMC trajectories for the five datasets have been merged.

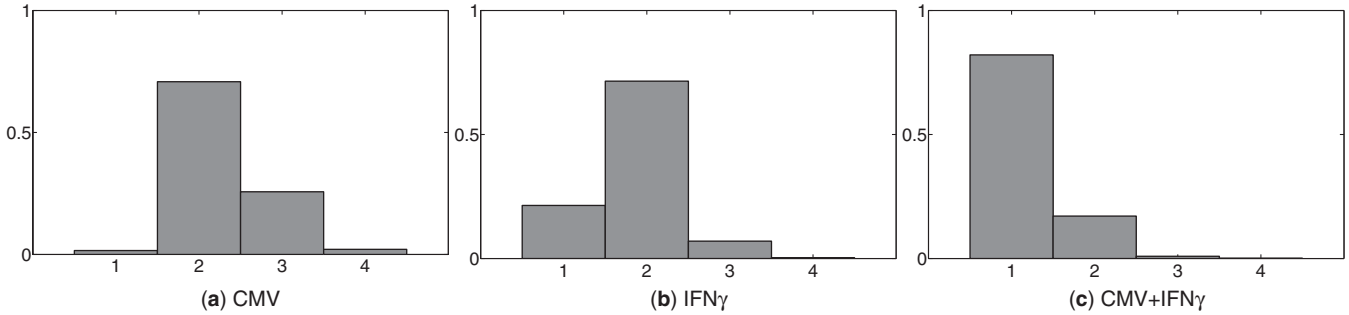


Fig. 4. Histograms of the numbers of BGM components for macrophage gene expression time series. For each experimental condition the posterior probability (vertical axis) of the number of components $\mathcal{K}$ (horizontal axis) have been estimated from the MCMC trajectories.

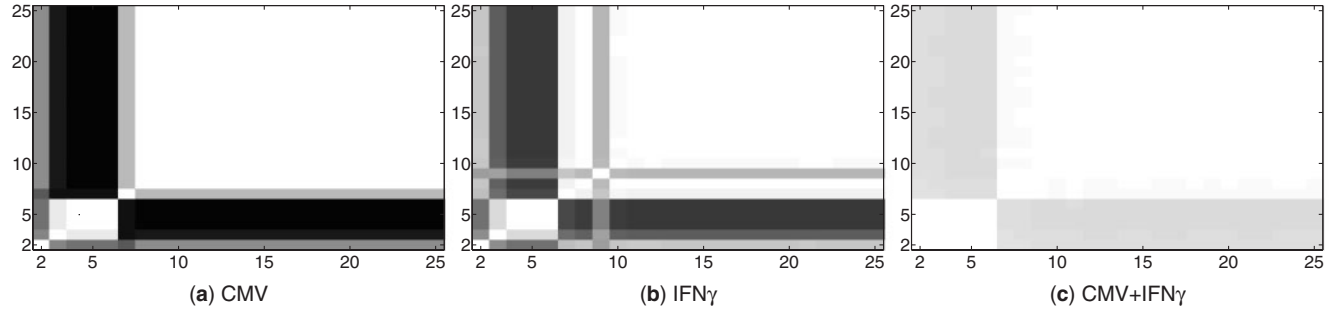


Fig. 5. Graphical presentation of the temporal connectivity structure for the macrophage gene expression data. The figure shows heatmap representations that indicate the estimated posterior probability of two time points being assigned to the same state (component). The probabilities are represented by a grey shading, where white corresponds to a probability of 1, and black corresponds to a probability of 0. The numbers on the axes represent the time points of the time course experiment. The analysis was repeated for all three experimental conditions CMV, IFNγ and CMV + IFNγ, as explained in the text.

are provided in Supplementary Material E, and the inference results are discussed in more detail in Section 6.

6 DISCUSSION

The results for the synthetic data generated from the Raf-Mek-Erk pathway of Sachs *et al.* (2005) show that the proposed BGM scheme consistently outperforms the conventional BGe and BDe metrics in terms of global network reconstruction accuracy (Figs 1 and 2). This confirms that BGM is superior when the data stem from a mixture distribution, and that the proposed sampling scheme (allocation MCMC) renders the inference, which is more complex than for the conventional case, practically viable. Furthermore, histograms of the number of inferred mixture components (Fig. 3) reveal that BGM succeeds in inferring the correct number of components. To assess whether BGM achieves any improvement for real biological applications, we applied it to gene expression data obtained from two different platforms (Agilent and Affymetrix) for two different systems: macrophages challenged with viral infection, and circadian rhythms in plants.

For macrophages challenged with CMV or pretreated with IFNγ, BGM tends to infer a two-stage process (Fig. 4). This two-stage process reflects a state change in the host macrophage brought about by infection (CMV) or immune activation (IFNγ), and can be found in all three experimental conditions (Fig. 5). Interestingly, though, this state change is less pronounced in the combined condition CMV+IFNγ (Fig. 4c), where the Bayes factor does not support the more flexible heterogeneous model (see the previous section).

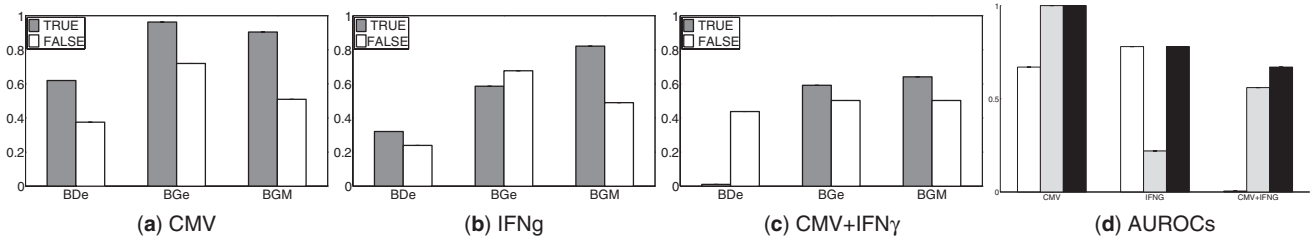


Fig. 6. Reconstructing the regulatory network of the Irfs. (a–c): Mean posterior probabilities (vertical axis) of true and false edges in the Irf regulatory network, inferred with BDe, BGe and BGM (horizontal axis) from the macrophage gene expression time series. According to the biological literature the true edges are: $Irf1 \to Irf2$, $Irf2 \to Irf1$ and $Irf3 \to Irf1$, while the edges $Irf1 \to Irf3$, $Irf2 \to Irf3$ and $Irf3 \to Irf2$ are spurious. In (**d**) an AUROC histogram plot is given. For each of the three conditions the histogram shows bars of the BDe (white), BGe (grey) and BGM (black) AUROC values. It can be seen that BGM is never inferior to BDe or BGe in terms of AUROC scores, but BGM outperforms (i) BDe for conditions CMV and CMV + IFN_γ and (ii) BGe for conditions IFN_γ and CMV + IFN_γ.

Table 1. Logarithmic predictive probabilities for the *A.thaliana* data: $log_e(P(\tilde{\mathcal{D}}|\mathcal{D},H_0))$ (BGE) and $log_e(P(\tilde{\mathcal{D}}|\mathcal{D},H_1))$ (BGM)

$\mathcal{D}$	H_i	$\tilde{\mathcal{D}}=T_{20}$	$\tilde{\mathcal{D}}=T_{28}$
T_{20}	BGe	–	−64.29 (±0.29)
	BGM	–	−53.69 (±0.42)
T_{28}	BGe	−63.93 (±0.22)	–
	BGM	−54.78 (±0.63)	–

The SDs of the logarithmic probabilities are given in brackets.

Table 2. Logarithmic predictive probabilities for the macrophage data: $log_e(P(\tilde{\mathcal{D}}|\mathcal{D},H_0))$ (BGe) and $log_e(P(\tilde{\mathcal{D}}|\mathcal{D},H_1))$ (BGM)

		$\tilde{\mathcal{D}}=\mathcal{D}_{TEST}$		
$\mathcal{D}=\mathcal{D}_{TRAIN}$	*Model*	CMV	IFNγ	CMV and IFNγ
CMV	BGe	–	−76.01 (±0.07)	−45.26 (±0.03)
	BGM	–	−63.63 (±0.02)	−33.80 (±0.38)
IFNγ	BGe	−56.78 ±0.05	–	−57.30 (±0.05)
	BGM	−39.62 ±0.02	–	−42.69 (±0.11)
CMV +	BGe	−37.76 (±0.08)	−69.19 ±0.06	–
IFNγ	BGM	−21.67 (±0.33)	−53.26 ±0.51	–

The SDs of the logarithmic probabilities are given in brackets.

This observation is consistent with the known biological responses of macrophages to simultaneous infection by virus (mCMV) *and* immune (IFNγ) activation. It suggests that upon dual challenge with both an infection and immune activation (CMV + IFN_γ) signalling leads to a pronounced singular response. This is in agreement with observations of cooperation between viral and immune signalling in effective vigorous anti-viral state within the host macrophage, as discussed in Benedict *et al.* (2001).

For the *A.thaliana* gene expression time series, BGM also infers a two-stage process (Fig. 7). In this application, the two stages are most likely related to the diurnal nature of the dark/light cycle. We have applied our method to two sets of plant samples, which were subjected to different prehistories, related to different lengths of the artificial, experimentally controlled light/dark cycle. Although the two-stage nature of the process is preserved, the state co-allocation posterior probabilities, shown in the heatmap of Figure 7, points to a phase shift of about 4–6 h as a consequence of the increased day length. This phase shift is biologically plausible and indeed expected. It can be explained by the early phase of entrainment that is required to elicit a phase delay that matches the 24-h period of the wild-type plants to the longer light/dark cycle (T_{28}), compared to the later phase of entrainment required to elicit a phase advance to match the shorter light/dark cycle (T_{20}) (Johnson *et al.*, 2003).

We anticipate that a non-linear and non-homogeneous generalization of (Bayesian) networks will have broader general utility for reconstructing regulatory networks in systems biology. In this regard there is increasing interest in the development of new statistical methods, as exemplified by the recent and related work of (Lèbre, 2008). Our article complements this work and constitutes a natural generalization of the BGe score of (Geiger and Heckerman, 1994) by applying the ideas of mixture models and allocation sampling presented in Nobile and Fearnside (2007). This is an advantage over the work of Ko *et al.* (2007). While the latter model is more flexible owing to the fact that different nodes can have different breakpoints, it leaves the computation of the marginal likelihood intractable. The authors resort to BIC (Schwarz, 1978) as a crude approximation to the marginal likelihood. However, this approximation is only valid in the limit of very large datasets, and BIC is known to be over-regularized in many practical applications. For a more detailed theoretical comparison between BGM and the approaches of Lèbre (2008), and Ko *et al.* (2007) see Supplementary Material E. The evaluation of the proposed BGM approach on synthetic benchmark data and the novel application to two real biological scenarios provide an encouraging demonstration of the viability of the proposed BGM method.

7 CONCLUSION

We have proposed a non-linear and non-homogeneous generalization of the BGe score for Bayesian networks (BGM). BGM is based on a mixture model, using latent variables to assign individual measurements to different classes. The practical inference follows the Bayesian paradigm and samples the graph, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the allocation sampler of Nobile and Fearnside (2007) as an alternative to RJMCMC (Green, 1995). We have evaluated BGM using three criteria: network reconstruction, statistical significance and agreement with

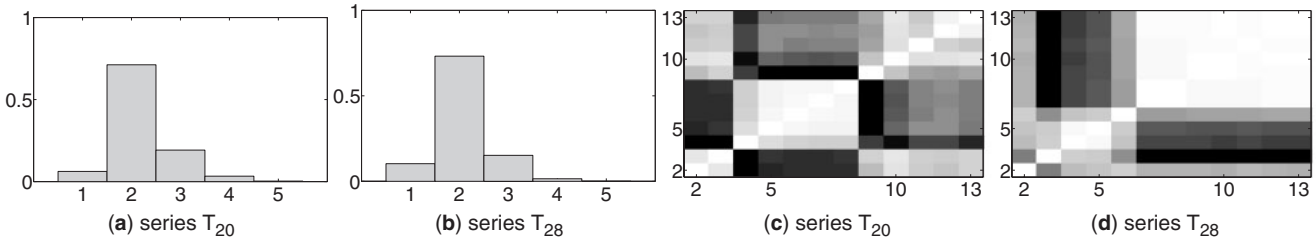


Fig. 7. Results of BGM analysis of nine circadian genes in *A*.thaliana. Two independent experiments under constant light condition were conducted. In experiment T_{20} (T_{28}) *A*. thaliana was entrained in a 10h:10h (14h:14h) dark/light cycle. (**a**) and (**b**) show the estimated posterior probabilities (vertical axis) of the number of BGM components $\mathcal{K}$ (horizontal axis). (**c**) and (**d**) show the heat map representations of the temporal connectivities, as explained in the caption of Figure 5. A comparison between the two panels reveals a phase shift of about 2–3 time points (4–6 h.) between the different entrainments T_{20} and T_{28}.

intrinsic biological features. In terms of network reconstruction, we found improved results both for a synthetic network of known structure (Figs 1 and 2) and for a small real regulatory network derived from the literature (Fig. 6). For assessing the statistical significance of the improvement, we computed two scores: Bayes factors and predictive distributions. We applied these scores to gene expression time series obtained on different platforms (Agilent and Affymetrix) for two different systems (viral challenge of macrophages and circadian rhythms in plants), where BGM tended to outperform BGe (Tables 1 and 2). Interestingly, we found that when the improvement obtained with BGM was significant, the posterior distribution peaked at two latent classes (Figs 4 and 7). This result provides excellent agreement with intrinsic dichotomies that we expect to find in these systems, related to the dichotomy between the healthy and diseased state of the cell, and the diurnal contrast between light and darkness.

ACKNOWLEDGEMENTS

We are grateful to T. Forster, S. Watterson, K. Robertson, and P. Dickinson for discussions and assistance with the handling and interpretation of the data, and to D. Nutter for technical support with computational issues.

Funding: M.G. is supported by the Biotechnology and Biological Sciences Research Council (BBSRC) and by the Engineering and Physical Sciences Research Council (EPSRC). D.H. is supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). Experimental work on Arabidopsis was supported by BBSRC grant G19886 to A.J.M. The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CISB) funded by BBSRC and EPSRC, reference BB/D019621/1.

Conflict of Interest: none declared.

REFERENCES

Benedict,C.A. *et al.* (2001) Lymphotoxins and cytomegalovirus cooperatively induce interferon-b establishing host-virus détente. *Immunity*, **15**, 617–626.

Cooper,G.F. and Herskovits,E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, **9**, 309–347.

Darnell,J. *et al.* (1994) Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science*, **264**, 1415–1421.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Friedman,N. *et al.* (1998) Learning the structure of dynamic Bayesian probabilistic networks. In Cooper,G.F. and Moral,S. (eds) *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann Publishers, San Francisco, CA, pp. 139–147.

Geiger,D. and Heckerman,D. (1994) Learning Gaussian networks. In López de Mántaras,R. and Poole,D. (eds) *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann Publishers, Seattle, Washington, USA, pp. 235–243.

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

Grzegorczyk,M. and Husmeier,D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Mach. Learn.*, **71**, 265–305.

Honda,K. *et al.* (2006) Type I Inteferon gene induction by the Interferon regulatory factor family of transcription factors. *Immunity*, **25**, 349–360.

Huelsenbeck,J. *et al.* (2004) Bayesian phylogenetic model selection using reversible jump Markov chain monte carlo. *Mol. Biol. Evol.*, **21**, 1123–1133.

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

Imoto,S. *et al.* (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Computat. Biol.*, **1**, 231–252.

Johnson,C. *et al.* (2003) Entrainment of circadian programs. *Chronobiol. Int.*, **20**, 741–774.

Ko,Y. *et al.* (2007) Inference of gene pathways using Gaussian mixture models. In Xia,J. (eds) *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'07)*. IEEE Computer Society, Los Alamitos, CA, USA, pp. 362–367.

Lèbre,S. (2008) *Analyse de processus stochastiques pour la génomique : étude du modèle MTD et inférence de réseaux bayésiens dynamiques*. Ph.D. thesis, Université d'Evry-Val-d'Essonne, Paris, France.

Madigan,D. and York,J. (1995) Bayesian graphical models for discrete data. *Int. Stat. Rev.*, **63**, 215–232.

Mas,P. (2008) Circadian clock function in Arabidopsis thaliana: time beyond transcription. *Trends Cell Biol.*, **18**, 273–281.

Nobile,A. (2005) Bayesian finite mixtures: a note on prior specification and posterior computation. *Technical report*. Department of Statistics, University of Glasgow, UK.

Nobile,A. and Fearnside,A. (2007) Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Stat. Comput.*, **17**, 147–162.

Raza,S. *et al.* (2008) A logic based diagram of signalling pathways central to macrophage activation. *BMC Syst. Biol.*, **2**. Article 36.

Sachs,K. *et al.* (2005) Protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

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

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Vehtari,A. and Lampinen,J. (2002) Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput.*, **14**, 2439–2468.

Verdinelli,I. and Wasserman,L. (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.*, **90**, 614–618.

Werhli,A.V. *et al.* (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**, 2523–2531.

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–9

Supplementary paper on experimental aspects (E): Implementation details and supplementary figures for the article: Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler

Marco Grzegorczyk[a,b], Dirk Husmeier[a,b], Kieron Edwards[e], Peter Ghazal[a,c] and Andrew J. Millar[a,d]

[a] Centre for Systems Biology at Edinburgh (CSBE), United Kingdom
[b] Biomathematics and Statistics Scotland (BioSS), Edinburgh, United Kingdom
[c] Division of Pathway Medicine (DPM), University of Edinburgh, United Kingdom
[d] Institute of Molecular Plant Sciences, University of Edinburgh, United Kingdom
[e] Advanced Technologies Cambridge, United Kingdom

ABSTRACT

Article: In the article we propose a non-linear and non-homogeneous generalization of the classical BGe score for Bayesian networks. The method is based on a mixture model, using latent variables to assign individual measurements to different classes. The practical inference follows the Bayesian paradigm and samples the network structure, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the recently proposed allocation sampler as an alternative to RJMCMC.

Supplementary material: Due to space restrictions of the article we provide some additional information as supplementary material. The implementation details of all applied algorithms are given in Section 1. Additional figures and tables are provided in Section 2. The computational complexity of the proposed BGM algorithm is briefly discussed in Section 3. Finally, in Section 4 we provide a theoretical comparison with two related approaches by Lèbre (2008) and Ko *et al.* (2007).

Availability: This supplementary paper on experimental aspects (E) is available from
http://www.bioss.ac.uk/associates/marco/supplement/E.pdf
A separate supplementary paper on theoretical aspects (T) providing a more detailed presentation of the mathematical methodology is available from
http://www.bioss.ac.uk/associates/marco/supplement/T.pdf
The data sets used in our study are available from
http://www.bioss.ac.uk/associates/marco/supplement/

Contact: marco@bioss.ac.uk, dirk@bioss.ac.uk

1 IMPLEMENTATION DETAILS

We implemented structure MCMC according to the presentations given in Madigan and York (1995), and in all experimental applications we used the following settings: For structure MCMC we set the burn-in length to 1,000,000 and then collected 500 graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_{500}\}$ by sampling every 2000 iterations. For BGM we set the probability for a structure MCMC move to 0.5. And the probabilities of the other four move types, which all leave the graph $\mathcal{G}$ unchanged, are set to 0.125. The maximal number of components $\mathcal{K}_{MAX}$ was set to 10 and we note that this upper limit was never reached during any MCMC simulation. Equal to the structure MCMC setting we set the burn-in length to 1,000,000 and then collected 500 states $\left\{[\mathcal{G}_1, \mathcal{K}_1, \vec{\mathcal{V}}_1], \ldots, [\mathcal{G}_{500}, \mathcal{K}_{500}, \vec{\mathcal{V}}_{500}]\right\}$ each consisting of a graph $\mathcal{G}_i$, a number of mixture components $\mathcal{K}_i$, and an allocation vector $\vec{\mathcal{V}}_i$.

Following Werhli *et al.* (2006) we restricted the fan-in to 3 and employed the graph prior $P(\mathcal{G})$ given in Eq. (3) of the supplementary paper on theoretical aspects (T) when analysing the synthetic Gaussian data. This guarantees that our results for $\mathcal{K} = 1$ are comparable to those of Werhli *et al.* (2006). But the graph prior employed by Werhli *et al.* (2006) yields an intrinsic penalty for complex networks (see Subsection 1.1 of the supplementary paper on theoretical aspects (T)). Therefore and as we did not have any biological prior knowledge about the interactions in the macrophage and the Arabidopsis domain, the analysis of the gene expression data was performed with a uniform prior over graphs instead, i.e. every graph was set to be equally likely a priori. Furthermore, we decided not to restrict the fan-in for these relatively small domains with $N = 3$ (macrophage) and $N = 9$ (Arapidopsis) nodes only.

For the time series we did not allow for *self-loops*, that is we did not allow that a node can be its own parent node, by restricting the graph's neighbourhoods in Eq. (7) of the supplementary paper on theoretical aspects (T) correspondingly. We decided to exclude self-loops, as they mainly capture degradation processes which are not of interest for modelling the regulatory interactions between genes.

Finally we note that we always performed two independent structure MCMC runs for each inference model: BDe, BGe, and BGM on every data set. Following Friedman and Koller (2003), we

started the structure MCMC simulations with the BDe metric and the BGe metric from the following initialisations: (i) As uninformed initialization the first structure MCMC run was always seeded by an empty graph without any edges. (ii) To obtain an informed initialization we always performed a greedy search algorithm and seeded the second structure MCMC run with the most likely graph outputted by greedy search. We initialised the proposed BGM algorithm with the same graph as the corresponding BGe structure MCMC simulation, and the number of mixture components was set to $\mathcal{K} = 1$ so that all observations were allocated to the same single mixture component at the beginning of the MCMC simulation. We note that BGM inference with the restriction $\mathcal{K} = 1$ is equivalent to structure MCMC inference with the BGe scoring metric. Hence, a greedy search based on BGe can be seen as a greedy search based on the BGM model under the constraint that there is exactly one mixture component, symbolically: $\mathcal{K} = 1$. We note that it may be advisable to initialise the BGM algorithm not only with a graph found by a greedy search algorithm based on BGe but also with an allocation vector outputted by a classification or cluster algorithm. In our experiments we deliberately avoided to employ a more informative initialisation for BGM to demonstrate that BGM succeeds in inferring the true relationships - and especially the mixture components - independently of the initialisation.

Edge posterior probability scatter plots and trace plot diagnostics, e.g. of the number of edges of the sampled graphs or of their logarithmic scores, were used to assess convergence. Except for the synthetic Raf-Mek-Erk pathway data with $m = 480$ data points (where some MCMC simulations did not converge satisfactorily) we could see from the edge posterior probabilities that the total MCMC run-length of 2,000,000 for relatively small domains (between $N = 3$ and $N = 11$ nodes) had led to a satisfactory degree of convergence (Pearson correlation coefficients greater than 0.98) for all three inference models (BGe, BDe, and BGM). Therefore we report only the results of the empty-seeded runs in the article and point out that we had some convergence problems for the Raf-Mek-Erk pathway data with $m = 480$ data points.

The hyperparameters of the BGe and BGM models (see Section 4.1 of the supplementary paper on theoretical aspects (T)) were set as follows: $v = 1$, $\alpha = N + 2$, $\vec{\mu_0} = (0, \ldots, 0)^T$ and $T_0 = 0.5 \cdot I_{N,N}$ where $I_{N,N}$ is the N-by-N identity matrix and N is the number of domain variables. The choices of T_0 and $\vec{\mu_0}$ ensure that we are not explicitly biasing our inference to any particular edge (Friedman *et al.*, 2000). They reflect a prior belief where all N domain variables (genes) are identically and independently standard Gaussian distributed (with mean 0 and variance 1). The effective sample size parameters v and α were set to small values, as this ensures that the weight of the prior distribution (induced by T_0 and $\vec{\mu_0}$) is as uninformative as possible subject to the constraint that the resulting covariance matrix $T_{\mathcal{D}}$ (see Section 4.1 of the supplementary paper on theoretical aspects (T)) is non-singular (Geiger and Heckerman, 1994). The prior parameters for the BDe model were selected as in Giudici and Castelo (2003) to ensure (i) that the prior is uninformative (total prior decision was set to 1) and (ii) that equal marginal likelihoods are given to equivalent DAGs. See Giudici and Castelo (2003) for further details.

2 SUPPLEMENTARY FIGURES AND TABLES

This second section provides additional figures and tables, which - due to space limitations - could not be included in the main paper. Most of the captions are self-explanatory, but some further explanations are given in the text. Figure 1 shows histograms of the posterior probabilities of the number of MCMC inferred mixture components for the synthetic Raf-Mek-Erk pathway data with $30 \leq m \leq 180$ data points. It can be seen that the proposed BGM model tends to infer the correct number of mixture components for these data sets. There are only 3 out of 16 combinations of $\mathcal{K}$ and m for which an incorrect number of components was inferred, namely: $(\mathcal{K} = 3, m = 60)$, $(\mathcal{K} = 5, m = 60)$, and $(\mathcal{K} = 5, m = 180)$. Especially for the data sets with $\mathcal{K}_{TRUE} = 5$ mixture components it appears that this inaccuracy of the BGM inference is due to the fact that there are only few observations per mixture component, namely $m_i = 12$ for $m = 60$ and $m_i = 36$ for $m = 180$, so that the posterior probability landscape may be relatively flat around the true regulatory relationships. It can be seen from Figure 3 in the main paper that the BGM inference on the number of mixture components becomes more accurate when more data points ($m = 480$) are available.

The time series of the analysed Interferon regulatory factors (Irf1, Irf2, and Irf3) and scatter plots of the three Irf genes are shown in Figures 2 and 3. In both plots symbols indicate to which mixture component the observations were allocated. Concrete allocations were obtained by imposing thresholds on the connectivity matrices, whereby for each condition (CMV, IFN$_\gamma$, and CMV+IFN$_\gamma$) the threshold was selected such that an allocation consistent with the trends indicated by the corresponding heat matrix (shown in Figure 5 of the main paper) was obtained. From the time series and the scatter plots it appears that the inferred mixture components differ with respect to the marginal distributions of the three Irf genes; especially in Figure 3 most of the observations allocated to the same component tend to appear as clusters of points in the scatter plots.

The directed edge posterior probability estimates for the Interferon regulatory factor domain derived from BDe, BGe and BGM inference are given in Table 1. A concrete network prediction can be obtained from the estimates in Table 1 by imposing a threshold and extracting those edges only whose posterior probability estimate exceeds the predefined threshold. The AUROC scores resulting from the posterior probability estimates in Table 1 - under the assumption that the true regulatory relationships are as follows: $Irf2 \leftrightarrow Irf1 \rightarrow Irf3$ (Darnell *et al.* (1994) and Raza *et al.* (2008)) - are shown in Figure 6 panel (d) of the main paper.

The time series of the nine circadian genes in Arabidopsis thaliana are shown in Figure 4. Obviously all these genes have a strong 24hr circadian rhythm, and interestingly it can also be seen that the light:dark entrainment shifts the gene expression profiles. For most of the circadian genes the dashed line (T_{28} corresponding to 14h:14h entrainment) seems to be shifted by approximately 2 hours compared to the solid line (T_{20} corresponding to 10h:10h entrainment). This is in agreement with the BGM inference result where heat maps (see panels (c) and (d) in Figure 7 of the main paper) also indicate a time shift. Although the time lags differ (4-6 hours instead of 2 hours) it seems that the general trend, i.e. a time shift, has been captured by the proposed BGM model.

The directed edge posterior probability estimates for the circadian

genes in Arabidopsis thaliana are given in Table 3 (T_{20}) and Table 4 (T_{28}). As explained above, concrete network predictions can be obtained from these estimates by imposing an arbitrary threshold on these posterior probability estimates. Furthermore, to illustrate graphically that the light:dark entrainment has an effect on the regulatory relationships, an edge posterior probability estimates scatter plot T_{20} versus T_{28} is given in Figure 5. Interestingly, it appears that the edge posterior probabilities are slightly different but do no completely differ; especially the edges with the highest posterior probability (around 1) are almost the same for both time series T_{20} and T_{28}. The Pearson correlation coefficient is equal to 0.84. Scatter plots of the directed edge posterior probabilities obtained by BGM inference versus BGe inference are shown in Figure 6. It can be seen from the two panels that the posterior probabilities are correlated and do not differ drastically. The Pearson correlations are equal to 0.94 (T_{20}) and 0.93 (T_{28}).

3 COMPUTATIONAL COMPLEXITY AND PERFORMANCE OF THE BGM ALGORITHM

The computational complexity of the proposed BGM algorithm depends on the number of network nodes N and the number of observations m. The computational complexity related to N is the same as for standard Bayesian network inference based on either the BGe or the BDe scoring metric. As the number of domain nodes N increases, convergence and mixing of the MCMC simulations become poorer, and the posterior distributions become more diffuse. To deal with the diffuse posteriors, the analysis of networks should focus on conserved subnetworks and network features, as discussed in Friedman *et al.* (2000). To improve mixing and convergence of the MCMC simulations, improved and alternative proposal scheme have been introduced; see Friedman and Koller (2003) and Grzegorczyk and Husmeier (2008). These aspects have already been investigated in the literature before, and we therefore do not revisit them.

The additional complexity of the proposed BGM algorithm is also related to the data set size m, as each new data point is associated with a separate allocation variable, that is a new component of the allocation vector $\vec{\mathcal{V}}$. To investigate how well our model scales up as m increases, we have also run simulations on larger synthetic Gaussian data sets with $m = 480$ data points, and we found that the computational costs do not increase substantially.

The BGM inference results suggest that the number of components in the heterogeneous data can be learned more accurately than with the smaller data set (see Figure 3 in the main paper and Figure 1 in this supplementary paper); however, the network reconstruction accuracy appears to slightly deteriorate (see Figure 1 and Figure 2 in the main paper). This finding might be counter-intuitive, as a larger data set contains more information and should therefore lead to a better performance. However, our finding is consistent with the fact that increased data set sizes lead to likelihood landscapes that are more rugged and, hence, result in increased mixing and convergence problems; see Figure 7 in Grzegorczyk and Husmeier (2008). When learning conventional Bayesian networks based on the BGe and BDe scoring metrics this problem can be addressed, e.g. by improving the MCMC proposal moves, as reported in Grzegorczyk and Husmeier (2008). Unfortunately, this approach is not applicable to the proposed BGM model, as the reassignment of allocation variables requires a computationally expensive re-computation of the scores on which the proposal distributions depend. We therefore have to resort to classical structure MCMC Madigan and York (1995), which scales up less favourably to larger systems; see the discussions in Grzegorczyk and Husmeier (2008). This problem can in principle be alleviated by the development of improved MCMC sampling schemes – akin to the improvement of MCMC schemes for conventional Bayesian networks (Grzegorczyk and Husmeier, 2008) – but the practical implementation needs to be left for future research.

4 GENERAL DISCUSSIONS AND RELATED WORK

Bayesian networks provide an abstract and simplified representation of regulatory networks and signalling pathways, which is certainly not appropriate when trying to resolve the detailed structure of a specific pathway. There is a clear trade-off between model complexity and inference accuracy/computational complexity. Bayesian networks based on the BDe and BGe scoring metric are of a simple form, but allow the marginal likelihood to be computed analytically. More complex models along the line we discuss below sacrifice inference accuracy and resort to measures that are only reliable in the limit of very large data sets, like the Laplace approximation or, worse, the Bayesian information criterion BIC (Schwarz, 1978). Computing marginal likelihoods for even more accurate models based on differential equations have been attempted, but the computational costs are so high that this approach is restricted to model selection from a very small set of candidate pathways (Vyshemirsky and Girolami, 2008). We therefore hold the view that simpler models, like Bayesian networks using BGe (Geiger and Heckerman, 1994), still play an important role in systems biology.

In principle, one could obtain a model that is more flexible than the proposed BGM method by selecting the components and allocations for each domain variable separately, and originally we intended to implement our BGM model along this line. But unfortunately it turned out that the BGe scoring metric by Geiger and Heckerman (1994) is not consistent with a model where each variable has different (independent) breakpoints. E.g. Ko *et al.* (2007) also apply a mixture of Bayesian networks model to infer gene regulatory networks from expression data. In fact, the model of Ko *et al.* (2007) is more flexible than our BGM model, with node-specific Gaussian mixture models and, hence, node-specific breakpoints. However, the inference procedure is less sound in that the marginal likelihood is intractable. The authors resort to the Bayesian information criterion BIC for model selection, which is only a good approximation to the marginal likelihood in the limit of very large data sets. In more detail: Our BGM model is based on the BGe scoring metric by Geiger and Heckerman (1994) so that the (component-wise) precision matrices of the whole network are taken into consideration when computing local scores. That is, the BGM model is based on correlations conditional on the whole domain (network). The approach of Ko *et al.* (2007) decomposes the whole network into local subnetworks (each consisting of a single domain node and its parent nodes only), and the local scores are computed from the precision matrices of these subnetworks only without taking the dependency structure of the complete system, that is, the precision matrix of the whole network, into consideration. The main shortcoming of the approach of Ko *et al.* (2007) is that model selection and inference do not use

	BDe				**BGe**				**BGM**		
	CMV				**CMV**				**CMV**		
	Irf_1	Irf_2	Irf_3		Irf_1	Irf_2	Irf_3		Irf_1	Irf_2	Irf_3
Irf_1	—	**0.89**	0.04	Irf_1	—	**1.00**	0.83	Irf_1	—	**1.00**	0.84
Irf_2	**0.63**	—	0.91	Irf_2	**0.91**	—	0.83	Irf_2	**0.86**	—	0.40
Irf_3	**0.33**	0.18	—	Irf_3	**0.98**	0.51	—	Irf_3	**0.86**	0.29	—
	IFN$_\gamma$				**IFN$_\gamma$**				**IFN$_\gamma$**		
	Irf_1	Irf_2	Irf_3		Irf_1	Irf_2	Irf_3		Irf_1	Irf_2	Irf_3
Irf_1	—	**0.18**	0.67	Irf_1	—	**0.75**	0.79	Irf_1	—	**0.94**	0.79
Irf_2	**0.05**	—	0.03	Irf_2	**0.34**	—	0.80	Irf_2	**0.77**	—	0.37
Irf_3	**0.73**	0.02	—	Irf_3	**0.67**	0.44	—	Irf_3	**0.75**	0.30	—
	CMV+IFN$_\gamma$				**CMV+IFN$_\gamma$**				**CMV+IFN$_\gamma$**		
	Irf_1	Irf_2	Irf_3		Irf_1	Irf_2	Irf_3		Irf_1	Irf_2	Irf_3
Irf_1	—	**0.02**	0.39	Irf_1	—	**0.77**	0.80	Irf_1	—	**0.80**	0.80
Irf_2	**0.01**	—	0.02	Irf_2	**0.34**	—	0.37	Irf_2	**0.44**	—	0.37
Irf_3	**0.01**	0.90	—	Irf_3	**0.66**	0.34	—	Irf_3	**0.68**	0.33	—

Table 1. Macrophage data: Inferred posterior probabilities of directed edges for each combination of experimental condition (CMV, IFN$_\gamma$, and CMV+IFN$_\gamma$) and BN inference procedure (BDe, BGe, and BGM). In each of the nine subtables the (i,j)-th cell contains the marginal posterior probability for an edge from Irf_i to Irf_j $(i, j = 1, \ldots, 3)$.

data	CMV	IFN$_\gamma$	CMV+IFN$_\gamma$
BDe	0.67	**0.78**	0.00
BGe	**1.00**	0.22	0.56
BGM	**1.00**	**0.78**	**0.67**

Table 2. Macrophage data: AUROC values. For each of the three macrophage data sets the table shows the BDe, BGe and BGM AUROC values computed from the directed edge relation features. The highest AUROC values for each data set are set in bold.

genes	LHY	CCA1	TOC1	ELF4	ELF3	GI	PRR9	PRR5	PRR3
LHY	—	1.00	0.53	0.37	0.43	0.35	0.19	0.15	0.35
CCA1	0.94	—	0.48	0.36	0.51	0.40	0.32	0.13	0.40
TOC1	0.08	0.15	—	0.28	0.47	0.09	0.28	0.15	0.33
ELF4	0.16	0.13	0.18	—	0.25	0.04	0.94	0.19	0.23
ELF3	0.09	0.15	0.08	0.13	—	0.04	0.53	0.15	0.15
GI	0.99	0.99	0.88	0.48	0.27	—	0.33	0.97	0.98
PRR9	0.49	0.26	0.20	0.43	0.26	1.00	—	0.90	0.19
PRR5	0.07	0.09	0.42	0.63	0.22	0.99	0.14	—	0.18
PRR3	0.11	0.15	0.11	0.14	0.24	0.06	0.17	0.16	—

Table 3. Arabidopsis thaliana T_{20} data: Inferred posterior probabilities of directed edges. The estimates were obtained with BGM inference for time series T_{20} (10h:10h light:dark entrainment). The (i,j)-th cell contains the marginal posterior probability of an edge from the gene in the i-th row to the gene in the j-th column.

genes	LHY	CCA1	TOC1	ELF4	ELF3	GI	PRR9	PRR5	PRR3
LHY	—	1.00	0.65	0.71	0.39	0.13	0.44	0.23	0.51
CCA1	0.92	—	0.40	0.39	0.61	0.16	0.35	0.51	0.26
TOC1	0.12	0.06	—	0.24	0.40	0.10	0.60	0.18	0.28
ELF4	0.09	0.11	0.14	—	0.23	0.05	0.44	0.08	0.08
ELF3	0.10	0.08	0.10	0.17	—	0.55	0.53	0.07	0.10
GI	1.00	1.00	0.75	0.63	0.30	—	0.16	0.89	0.92
PRR9	0.20	0.42	0.12	0.15	0.24	0.99	—	0.90	0.11
PRR5	0.18	0.13	0.62	0.37	0.24	0.92	0.21	—	0.65
PRR3	0.31	0.12	0.12	0.17	0.25	0.04	0.13	0.09	—

Table 4. Arabidopsis thaliana T_{28} data: Inferred posterior probabilities of directed edges. The estimates were obtained with BGM inference for time series T_{28} (14h:14h light:dark entrainment). The (i,j)-th cell contains the marginal posterior probability of an edge from the gene in the i-th row to the gene in the j-th column.

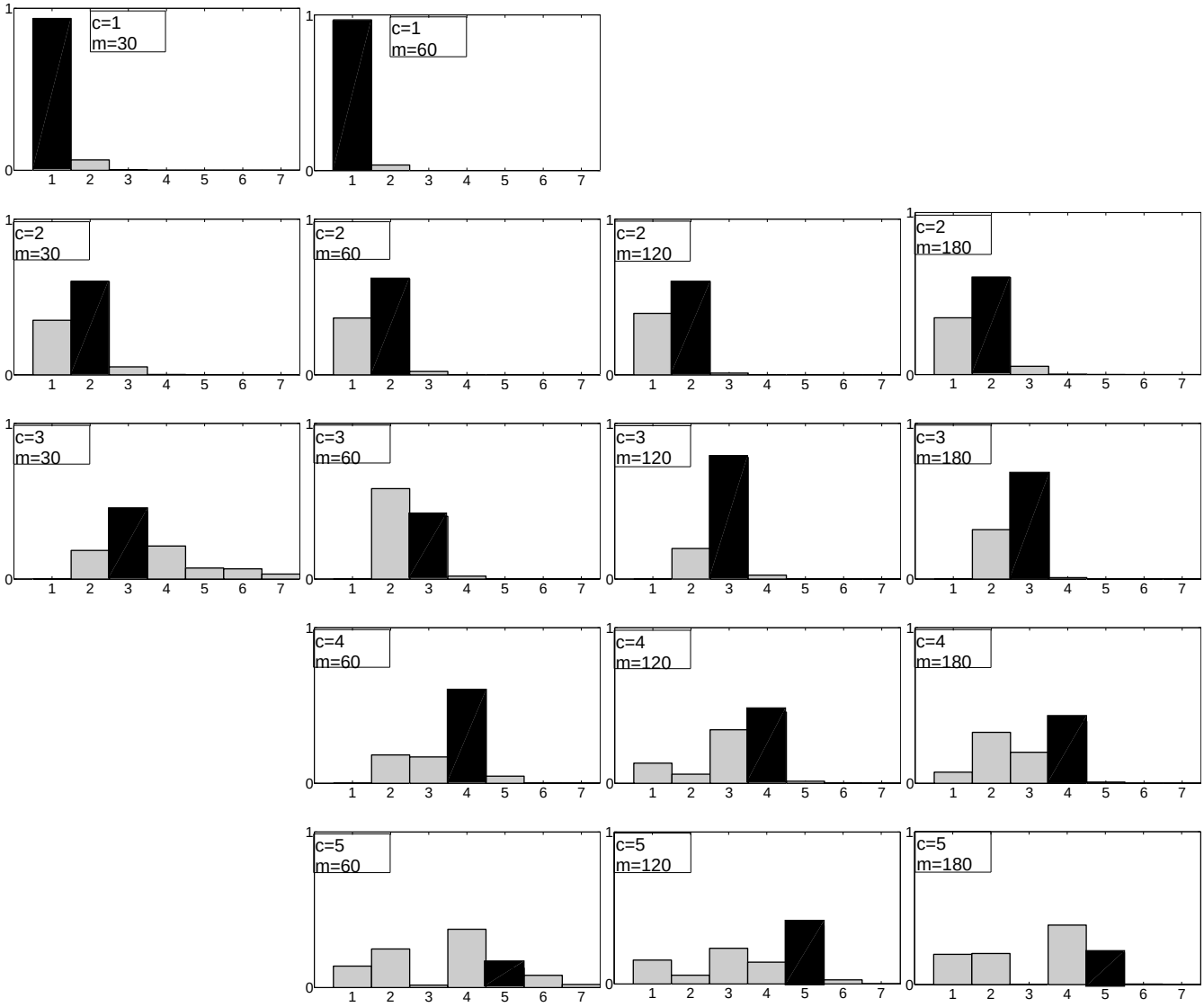
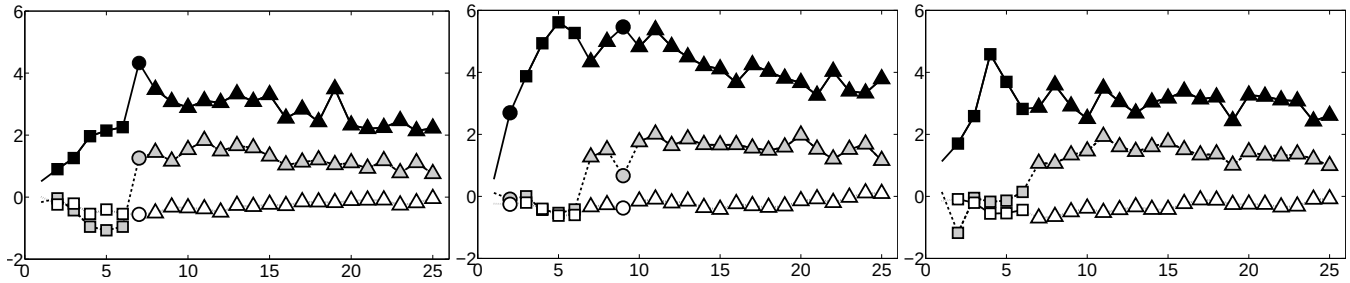


Fig. 1. Synthetic Gaussian data: Histograms of the number of inferred mixture components. For each considered combination of true components ($1 \leq \mathcal{K}_{TRUE} \leq 5$) and sample size m a histogram of the number of BGM-inferred components is shown. In each histogram the vertical axes represent posterior probabilities estimated with MCMC whereby the BGM MCMC trajectories have been merged across the 5 independent replications. From the histograms it can be seen that the posterior distribution of the number of mixture components $\mathcal{K}$ inferred with BGM tends to peak at the correct number (indicated by black bars) for $\mathcal{K} \leq 4$. Only for the combination $\mathcal{K} = 3$ and $m = 60$ the posterior distribution of the number of inferred components wrongly peaks at $\mathcal{K} = 2$. For $\mathcal{K}_{TRUE} = 5$ (last row) the posterior distribution of the number of inferred components becomes flat and does not peak at the correct number of components for $m = 60$ and $m = 180$.

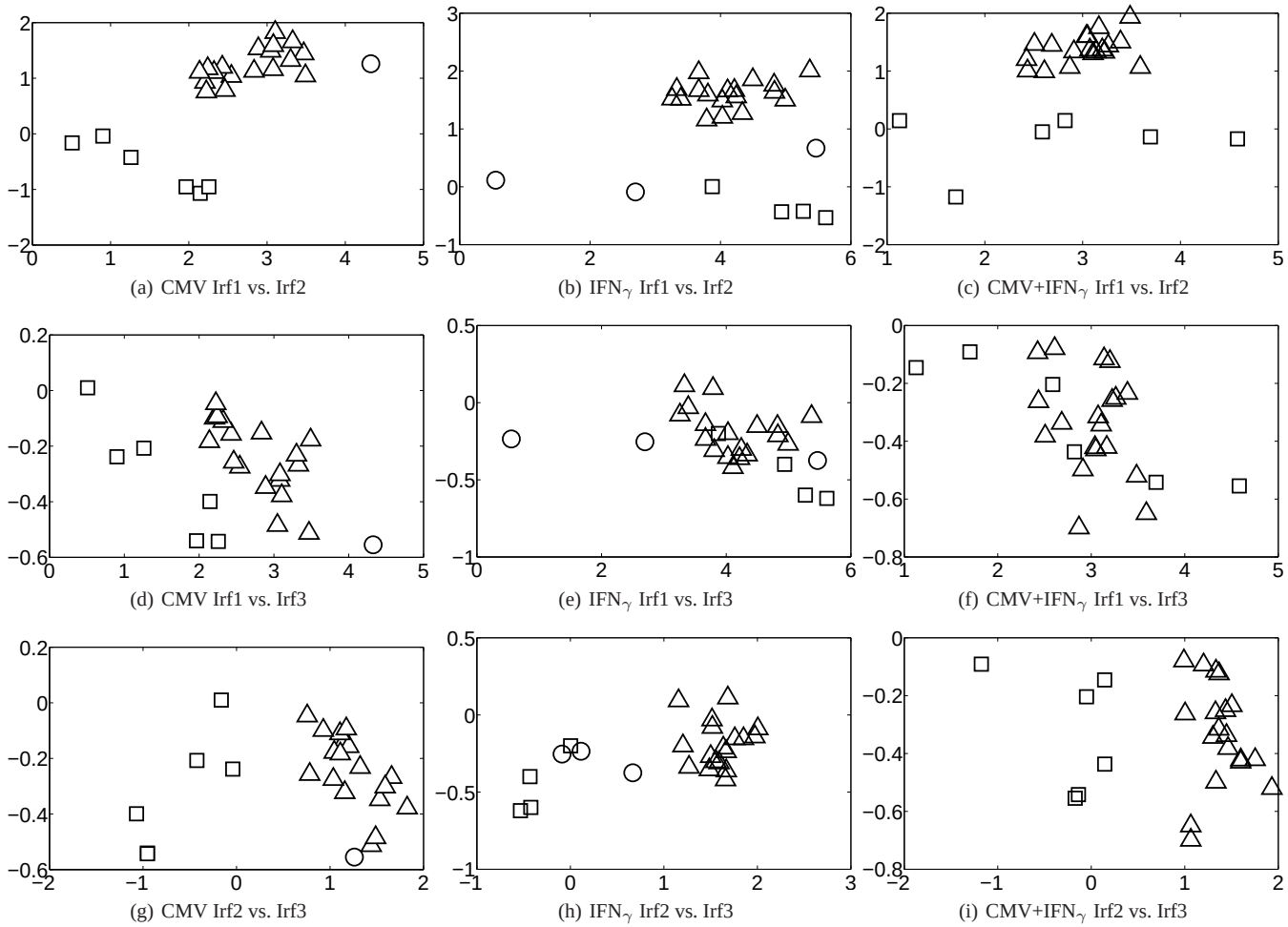
a proper Bayesian network scoring metric based on the marginal likelihood, such as BGe or BDe, but the Bayesian information criterion (BIC). BIC is known to be a crude approximation to the proper BGe score, which in many practical applications is strongly over-regularized, especially when the data are sparse. Additionally, instead of sampling the network structure, the number of components, and the allocation of the observations from the joint posterior distribution with Markov chain Monte Carlo (MCMC), as in our work, the approach proposed in Ko *et al.* (2007) is based on a heuristic optimisation scheme that fails to take the intrinsic inference uncertainty into account.

To understand why the marginal likelihood of the BGe scoring metric (Geiger and Heckerman, 1994) becomes intractable for the variable-specific change point model, consider the following example. Let there be three domain nodes X, Y, and Z and the network structure $Y \leftarrow X \rightarrow Z$ whereby the dependency $Y \leftarrow X$ is modelled by one single component, symbolically: $\vec{V}_Y(i) = 1$ for all observations i, and the dependency $X \rightarrow Z$ is modelled by two components, symbolically: $\vec{V}_Z(i) \in \{1, 2\}$ for all observations i. The BGe score of Geiger and Heckerman (1994) is based on the



(a) CMV (b) IFNγ (c) CMV+IFNγ

Fig. 2. Macrophage data: Gene expression time series of the Interferon regulatory factors. Black symbols: Irf1; grey symbols: Irf2; and white symbols: Irf3. Concrete allocations were obtained by imposing thresholds on the connectivity matrices, whereby for each condition the threshold was selected such that an allocation consistent with the trends indicated by the corresponding heat matrix shown in Figure 5 of the main paper was obtained. The different symbols (triangles, circles, squares) along the time series indicate which observations are then assigned to the same mixture component by the proposed inference scheme.

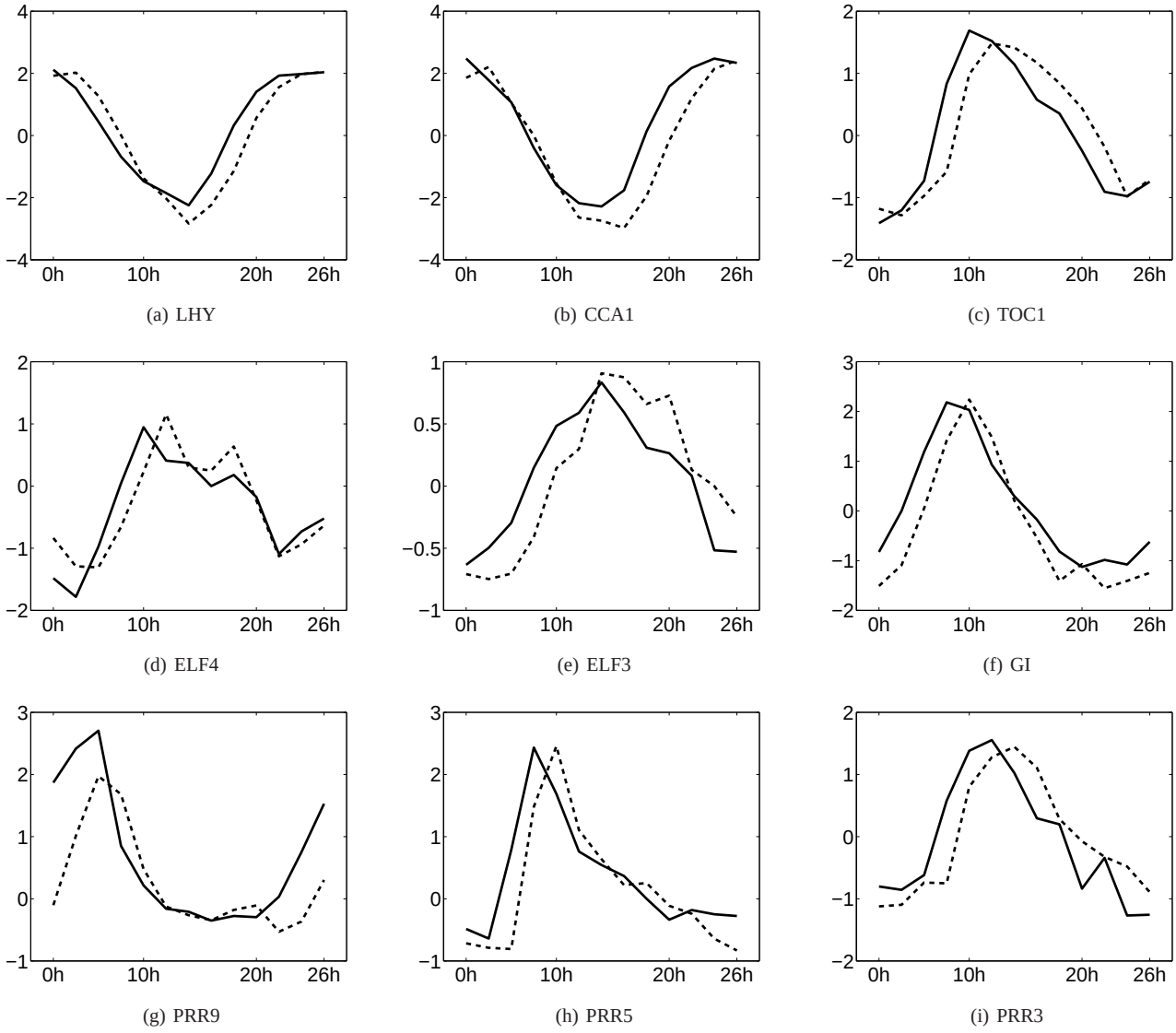


(a) CMV Irf1 vs. Irf2 (b) IFN$_\gamma$ Irf1 vs. Irf2 (c) CMV+IFN$_\gamma$ Irf1 vs. Irf2

(d) CMV Irf1 vs. Irf3 (e) IFN$_\gamma$ Irf1 vs. Irf3 (f) CMV+IFN$_\gamma$ Irf1 vs. Irf3

(g) CMV Irf2 vs. Irf3 (h) IFN$_\gamma$ Irf2 vs. Irf3 (i) CMV+IFN$_\gamma$ Irf2 vs. Irf3

Fig. 3. Macrophage data: Scatter plots for the macrophage data. The figure shows scatter plots of the collected Irf gene expression data. For each condition (CMV, IFN$_\gamma$ and CMV+IFN$_\gamma$.) there is a column with three panels showing the scatter plots for the three Irf gene pairs (Irf1 vs. Irf2, Irf1 vs. Irf3, and Irf2 vs. Irf3). The symbols (rectangles, triangles, and circles) indicate to which component the data points are allocated according to Figure 2. See caption of Figure 2 for more details.



(a) LHY (b) CCA1 (c) TOC1

(d) ELF4 (e) ELF3 (f) GI

(g) PRR9 (h) PRR5 (i) PRR3

Fig. 4. Gene expression time series of nine circadian genes in Arabidopsis thaliana. For each of the selected nine circadian clock-regulated genes there is a plot of two time series. The solid lines refer to the measurements of time series T_{28} (14:14 light:dark entrainment) and the dashed lines refer to the measurements of series T_{20} (10:10 light:dark entrainment). It can be clearly seen that varying the entrainment leach to a phase shift of the gene expression profiles. For most of the circadian genes the dashed line (T_{28}) seems to be shifted by 2h compared to the solid line (T_{20}).

precision matrix of the whole network (see Eq. (11), Eq. (15), and Eq. (24) in Geiger and Heckerman (1994)). Therefore, when we compute the local score of node Y conditional on X, we will have to (i) consider the precision matrix of the whole network domain (X, Y, and Z) and (ii) extract the relevant submatrix consisting of those rows and columns corresponding to X and Y. But since the relationship between X and Z is modelled by a mixture of two components and so depends on $\vec{V}_Z$, the precision matrix of the whole network also depends on the allocation vector $\vec{V}_Z$, and the precision matrix entries of the submatrix corresponding to Y and Z are different for the two components of $\vec{V}_Z$. Especially, this implies that the whole network (X, Y, and Z) does not have a multivariate Gaussian distribution but must be represented as a mixture of two multivariate Gaussian distributions.

Consequently, if there are local probability distributions which are modelled with more than one component, then the precision matrix of the whole domain has to be computed from a mixture of multivariate Gaussians. More precisely, it holds: If the local probability distribution of X_i is modelled according to the allocation vector $\vec{V}_i$ where $\vec{V}_i$ consists of c_i different mixture components ($i = 1, \dots n$), then the precision matrix of the whole domain ($X_1, \dots X_n$) consists of up to $c = \prod_{i=1}^{n} c_i$ mixture components, and for each of the c different realisations of ($\vec{V}_1, \dots \vec{V}_n$) there is a multivariate Gaussian distribution with a different precision matrix.

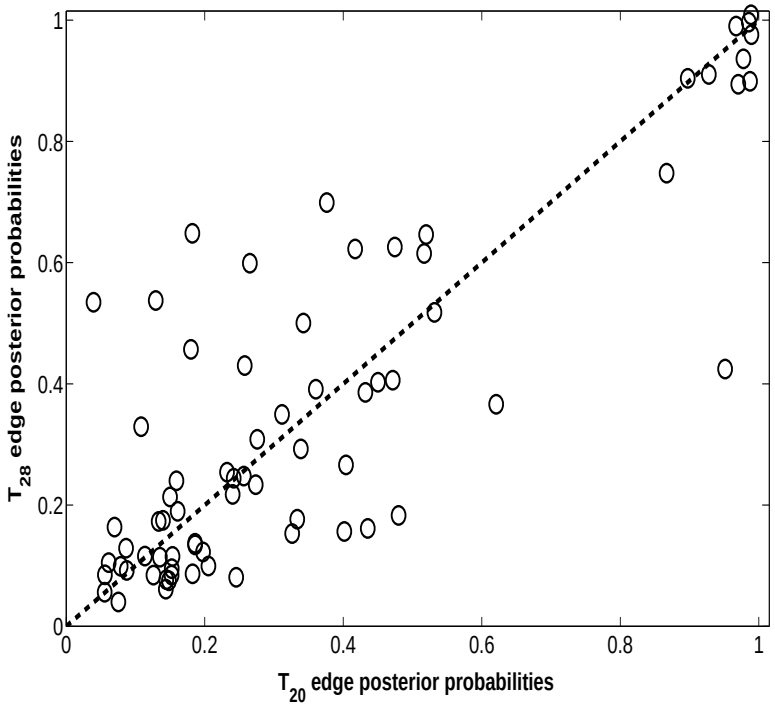
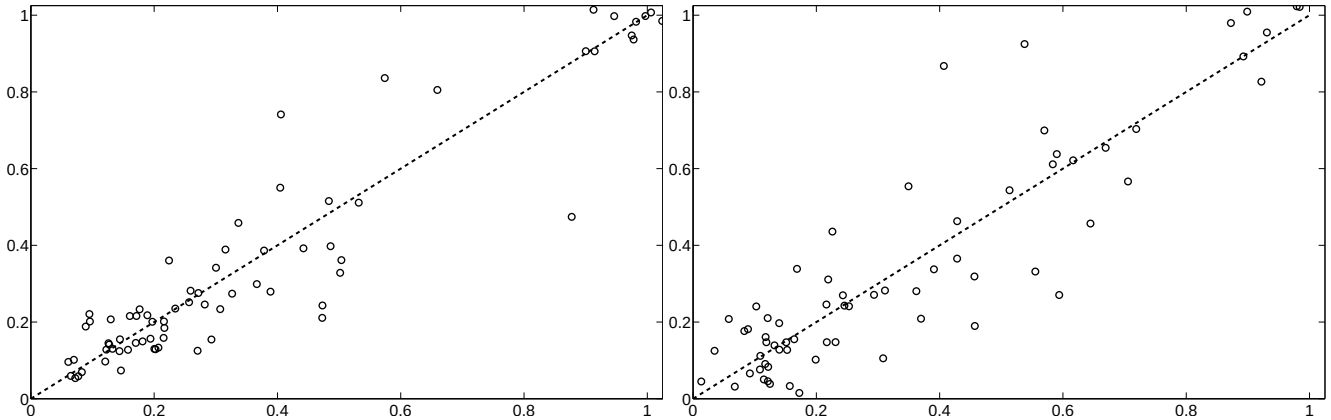


Fig. 5. Arabidopsis thaliana data. Scatter plot of edge posterior probabilities: T_{20} (horizontal axis) versus T_{28} (vertical axis). The Pearson correlation coefficient is equal to 0.84. The coordinates of all points were randomly slightly perturbed to visualize clusters of points.



(a) time series T_{20} (b) time series T_{28}

Fig. 6. Arabidopsis thaliana data. Edge posterior probabilities BGe versus BGM. For both time series T_{20} (panel (a)) and T_{28} (panel (b)) the edge posterior probabilities of BGM (horizontal axis) have been plotted against the edge posterior probabilities of BGe (vertical axis). The Pearson correlation coefficients are equal to 0.94 (T_{20}) and 0.93 (T_{28}). The coordinates of all points were randomly slightly perturbed to visualize clusters of points.

Finally, we note that it has recently come to our attention that closely related work has been carried out in Lèbre (2008). The main differences are as follows. The present work has been motivated by the attempt to find a non-linear generalization of the BGe model, using a mixture distribution and the allocation sampler. The regionality, that is, the segmentation of the time series into consecutive segments has come out of the inference automatically, that is, it is purely data-driven. The breakpoint model applied in Lèbre (2008) imposes this structure onto the model a priori. While this is a useful assumption in most cases, it is more restricted in terms of modelling non-linear distributions. Also, if the regionality assumption is valid, it is straightforward to include it as prior knowledge in our model via a Markovian dependence between the latent variables. In fact, this approach could be regarded as a

generalization of the breakpoint model, as discussed in Lehrach (2007). The second difference is that Lèbre (2008) allows the model to learn different graphs between different breakpoints, while in our approach the graph is constrained to remain unchanged. While this makes the approach of Lèbre (2008) more flexible, it implies that there is no sharing of information between different breakpoints. To rephrase this: while the method of Lèbre (2008) infers the breakpoint structure from the whole data set, it infers a graph associated with a breakpoint only from the subset of the data assigned to the respective segment. Note that time series available for contemporary microarray studies are usually limited to a few dozen time points. Further decreasing the effective sample set size will inevitably increase the vagueness of the posterior distribution. By allowing for certain information sharing between the segments, our approach alleviates this problem. In other words, by assuming that the graph remains unchanged, and only allowing the distributions of the parameters associated with the interactions to vary between segments, the inference uncertainty is considerably reduced.

REFERENCES

Darnell, J., Kerr, I. and Stark, G. (1994) Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science*, **264**, 1415–1421.

Friedman, N. and Koller, D. (2003) Being Bayesian about network structure. *Machine Learning*, **50**, 95–126.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Geiger, D. and Heckerman, D. (1994) Learning Gaussian networks. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 235–243.

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

Grzegorczyk, M. and Husmeier, D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265–305.

Ko, Y., Zhai, C. and Rodriguez-Zas, S. (2007) Inference of gene pathways using Gaussian mixture models. In *BIBM International Conference on Bioinformatics and Biomedicine*, pp. 362–367. Fremont, CA.

Lèbre, S. (2008) *Analyse de processus stochastiques pour la génomique : étude du modèle MTD et inférence de réseaux bayésiens dynamiques*. Ph.D. thesis, Université d'Evry-Val-d'Essonne.

Lehrach, W. (2007) *Bayesian machine learning methods for predicting protein-peptide interactions and detecting mosaic structures in DNA sequence alignments*. Ph.D. thesis, University of Edinburgh.

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

Raza, S., Robertson, K., Lacaze, P., Page, D., Enright, A., Ghazal, P. and Freeman, T. (2008) A logic based diagram of signalling pathways central to macrophage activation. *BMC Systems Biology*, **2**. Article 36.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Vyshemirsky, V. and Girolami, M. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833–839.

Werhli, A. V., Grzegorczyk, M. and Husmeier, D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**, 2523–2531.

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–11

Supplementary paper on theoretical aspects (T) for the article: Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler

Marco Grzegorczyk[a,b], Dirk Husmeier[a,b], Kieron Edwards[e], Peter Ghazal[a,c] and Andrew J. Millar[a,d]

[a] Centre for Systems Biology at Edinburgh (CSBE), United Kingdom
[b] Biomathematics and Statistics Scotland (BioSS), Edinburgh, United Kingdom
[c] Division of Pathway Medicine (DPM), University of Edinburgh, United Kingdom
[d] Institute of Molecular Plant Sciences, University of Edinburgh, United Kingdom
[e] Advanced Technologies Cambridge, United Kingdom

ABSTRACT

Article: In the article we propose a non-linear and non-homogeneous generalization of the classical BGe score for Bayesian networks. The method is based on a mixture model, using latent variables to assign individual measurements to different classes. The practical inference follows the Bayesian paradigm and samples the network structure, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the recently proposed allocation sampler as an alternative to RJMCMC.

Supplementary material: Due to space restrictions of the article we provide some additional information as supplementary material. This supplementary paper on theoretical aspects (T) presents more details of the mathematical theory. Section 1 gives a detailed overview to Bayesian network methodology. Section 3 deals with the proposed BGM model and the corresponding MCMC sampling scheme. The BGe scoring metric and its straightforward extension to the new BGM model is discussed in Section 4. Section 5 deals with predictive probabilities for BGe and BGM. The implementation details of all applied algorithms and additional figures and tables are available as a separate supplementary paper on experimental aspects (E).

Availability: This supplementary paper on theoretical aspects (T) is available from
http://www.bioss.ac.uk/associates/marco/supplement/T.pdf
A separate supplementary paper on experimental aspects (E) with the implementation details and additional figures and tables is available from
http://www.bioss.ac.uk/associates/marco/supplement/E.pdf
The data sets used in our study are available from
http://www.bioss.ac.uk/associates/marco/supplement/

Contact: marco@bioss.ac.uk, dirk@bioss.ac.uk

1 BAYESIAN NETWORK METHODOLOGY

This first section of this supplementary paper on theoretical aspects (T) gives a more detailed introduction to standard Bayesian network inference. The first subsection describes the Bayesian network model, the second summarizes the structure MCMC sampling scheme for Bayesian networks developed by Madigan and York (1995). Additional information on edge posterior probabilities, ROC curves and AUROC values is given in the third subsection.

1.1 Bayesian networks

Static Bayesian networks (BNs) are interpretable and flexible models for representing probabilistic relationships between interacting variables. At a qualitative level, the graph of a BN describes the relationships between the domain variables in the form of conditional independence relations. At a quantitative level, local relationships between variables are described by conditional probability distributions. Formally, a BN is defined by a graph $\mathcal{G}$, a family of conditional probability distributions F, and their parameters $\vec{q}$, which together specify a joint distribution over the domain variables.

The graph $\mathcal{G}$ of a BN consists of a set of N nodes (variables) $X_1, ..., X_N$ and a set of directed edges between these nodes. The *directed edges* indicate dependence relations. If there is a directed edge pointing from node X_i to node X_j, then X_i is called a *parent* (node) of X_j, and X_j is called a *child* (node) of X_i. The *parent set* of node X_n, symbolically π_n, is defined as the set of all parent nodes of X_n, that is, the set of nodes from which an edge points to X_n in $\mathcal{G}$. We say that a node X_n is *orphaned* if it has an empty parent set: $\pi_n = \emptyset$. If a node X_k can be reached by following a *path* of directed edges starting at node X_i, then X_k is called a *descendant* of X_i. The structure of a Bayesian network is defined to be a *directed acyclic graph*, that is, a directed graph in which no node can be its own descendant. Graphically this means that there are no cycles of directed edges (loops) in DAGs. It is due to the acyclicity that the joint probability distribution in BNs can be factorised as follows:

$$P(X_1, ..., X_N) = \prod_{n=1}^{N} P(X_n|\pi_n) \quad (1)$$

For further details, see Jensen (1996). Thus, DAGs imply sets of conditional independence assumptions for BNs, and so factorisations of the joint probability distribution in which each node depends on its parent nodes only. But more than one DAG can imply exactly the same set of conditional independencies, and if two DAGs assert the same set of conditional independence assumptions, those DAGs are said to be *equivalent*. This relation of graph equivalence imposes a set of *equivalence classes* over DAGs. The DAGs within an equivalence class have the same underlying undirected graph, but may disagree on the direction of some of the edges. (Verma and Pearl, 1990) prove that two DAGs are equivalent if and only if they have the same *skeleton* and the same set of *v-structures*. The skeleton of a directed acyclic graph (DAG) is defined as the undirected graph which results from ignoring all edge directions. And a v-structure denotes a configuration $X_i \rightarrow X_n \leftarrow X_k$ of two directed edges converging on the same node X_n without an edge between X_i and X_k (Chickering, 1995).
Although Bayesian networks (BNs) are based on DAGs, it is important to note that not all directed edges in a BN can be interpreted causally. Like a BN, a *causal network* is mathematically represented by a DAG. However, the edges in a causal network have a stricter interpretation: the parents of a variable are its immediate causes. In the presentation of a causal network it is meaningful to make the *causal Markov assumption* (Pearl, 2000): Given the values of a variable's immediate causes, it is independent of its earlier causes. Under this assumption, a causal network can be interpreted as a BN in that it satisfies the corresponding Markov independencies. However, the reverse does not hold.
The probability models for BNs we will consider in this paper lead to the same scores for equivalent DAGs, so that only equivalence classes can be learnt from data. Chickering (1995) shows that equivalence classes of DAGs can be uniquely represented using *completed partially directed acyclic graphs* (CPDAGs). A CPDAG contains the same skeleton as the original DAG, but possesses both directed and undirected edges. Every directed edge $X_i \rightarrow X_j$ of a CPDAG denotes that all DAGs of this class contain this edge, while every undirected edge $X_i - X_j$ in this CPDAG-representation denotes that some DAGs contain the directed edge $X_i \rightarrow X_j$, while others contain the oppositely orientated edge $X_i \leftarrow X_j$. An algorithm that takes as input a DAG, and outputs the CPDAG representation of the equivalence class to which that DAG belongs, can be found in (Chickering, 2002).

Stochastic models for Bayesian networks (Friedman *et al.*, 2000) specify the distributional form F and the parameters q of the local probability distributions $P(X_n|\pi_n)$ $(n = 1, ..., N)$. They assert a distribution to each domain node X_n conditional on its parent set π_n, whereby the parent sets are implied through the underlying DAG. The local probability distributions together specify the joint probability distribution of all domain variables $P(X_1, ..., X_N)$ (see Eq. (1)). Consequently, given data $\mathcal{D}$ these parametric models can be used to score DAGs $\mathcal{G}$ with respect to their posterior probabilities $P(\mathcal{G}|\mathcal{D}, F, q)$. We assume that the data matrix $\mathcal{D}$ is of size N-by-m and each of the m columns corresponds to an independent realisation of the domain $X_1, \ldots, X_N$. $\mathcal{D}_{i,j}$ is the j-th observation of the i-th domain node X_i.

Neglecting the family of probability distributions F and their parameters $\vec{q}$, we have for the posterior probability $P(\mathcal{G}|\mathcal{D})$ of a DAG $\mathcal{G}$ given the data matrix $\mathcal{D}$:

$$P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{G},\mathcal{D})}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\mathcal{G}) \cdot P(\mathcal{G})}{\sum_{\mathcal{G}^\star \in \Omega} P(\mathcal{D}|\mathcal{G}^\star) \cdot P(\mathcal{G}^\star)}, \qquad (2)$$

whereby $P(\mathcal{G})$ $(\mathcal{G} \in \Omega)$ is the prior probability over the space Ω of all possible DAGs over the domain $X_1, ..., X_N$. $P(\mathcal{D}|\mathcal{G})$ is the marginal likelihood, that is the probability of the graph $\mathcal{G}$ given the data matrix $\mathcal{D}$. A commonly used graph prior $P(\mathcal{G})$ $(\mathcal{G} \in \Omega)$ is a uniform distribution over Ω. Another graph prior is given by:

$$P(\mathcal{G}) = \frac{1}{\Pi} \prod_{n=1}^{N} \left(\begin{array}{c} N-1 \\ |\pi_n| \end{array} \right)^{-1} \qquad (3)$$

where Π is a normalization constant, and $|\pi_n|$ is the cardinality of the parent set π_n. The graph prior given in Eq. (3) implicitly assumes that the cardinalities of the parent sets for each domain node are uniformly distributed and, hence, includes a penalty for complex networks (Friedman and Koller, 2003).
There are two major stochastic models for which certain regularity conditions can be satisfied, so that a closed-form solution can be derived for the likelihood $P(\mathcal{D}|\mathcal{G})$ by analytical integration. See Geiger and Heckerman (1994) and Heckerman (1999) for further details. The posterior probability $P(\mathcal{G}|\mathcal{D})$ (see Eq. (2)) has a modular form:

$$P(\mathcal{G}|\mathcal{D}) \;=\; \frac{1}{Z_c} \prod_{n=1}^{N} \exp(\psi[X_n, \pi_n|\mathcal{D}]) \qquad (4)$$

Here, Z_c is a normalization factor, and $\psi[X_n, \pi_n|\mathcal{D}]$ are local scores that are computed from the data $\mathcal{D}$ and depend on the parent sets π_n implied through the DAG $\mathcal{G}$. The local scores $\psi[.]$ are defined by the employed probability model. The two major stochastic models, leading to a closed-form solution, are 1) the linear Gaussian model with a Normal-Wishart distribution as the conjugate prior (BGe-model), and 2) the multinomial distribution with a Dirichlet prior (BDe-model). A comparison of these models in the context of reverse engineering gene regulatory networks can be found in Friedman *et al.* (2000). In this article we focus on an non-homogeneous extension of the BGe-model. See Geiger and Heckerman (1994) or Grzegorczyk *et al.* (2008) for more detailed presentations of the BGe model for Bayesian networks.

When instead of m independent observations for the domain $X_1, \ldots, X_m$ time series data $(X_{1,t}, \ldots X_{N,t})_{t=1,\ldots,m}$ have been collected, *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 realisation $x_{j,t}$ of X_j at time point t is influenced by the realisation $x_{i,t-1}$ of X_i at the previous time point $t-1$. In DBNs parameters are tied such that the transition probabilities between time slices $t-1$ and t are the same for all t, that is, DBNs are homogeneous Markov models. Because of the time delay of interactions the acyclicty of the underlying graph $\mathcal{G}$ is not required, and Eq. (1) is replaced by:

$$P(X_{1,t}, ..., X_{N,t}) = \prod_{n=1}^{N} P(X_{n,t}|\pi_{n,t-1}) \qquad (5)$$

where $\pi_{n,t-1}$ denotes the parent set of X_n at the previous time point $t-1$. Accordingly, the DBN counterpart of Eq. (4) is given by:

$$P(\mathcal{G}|\mathcal{D}) \;=\; \frac{1}{Z_c}\prod_{n=1}^{N}\exp(\psi[X_{n,t},\pi_{n,t-1}|\mathcal{D}]) \tag{6}$$

We note that no realisations for the potential parent nodes of the domain variables $X_{i,1}$ at the first time point ($t=1$) are available. Consequently the first observations for $X_{1,1},\ldots,X_{1,m}$ at time point $t=1$ cannot be included when computing likelihoods for DBNs. That is, for time series of length m the effective sample size that can be used for the computation of DBN likelihoods is equal to $m-1$.

1.2 Structure MCMC sampling of Bayesian networks

In the context of static Bayesian networks (BNs) Different Markov chain Monte Carlo (MCMC) methods have been proposed for sampling directed acyclic graphs (DAGs) $\mathcal{G}$ from the posterior distribution $P(\mathcal{G}|\mathcal{D})$ (Madigan and York (1995), Friedman and Koller (2003), or Grzegorczyk and Husmeier (2008). The structure MCMC approach of Madigan and York (1995) generates a sample of DAGs $\mathcal{G}_1,...,\mathcal{G}_T$ from the posterior distribution by a Metropolis Hastings sampler in the space of DAGs. Given a DAG $\mathcal{G}_i$, in a first step a new DAG $\mathcal{G}_{i+1}$ is proposed with the following proposal probability $Q(\mathcal{G}_{i+1}|\mathcal{G}_i)$:

$$Q(\mathcal{G}_{i+1}|\mathcal{G}_i)=\left\{\begin{array}{cc}\frac{1}{|\mathcal{N}(\mathcal{G}_i)|} & ,\mathcal{G}_{i+1}\in\mathcal{N}(\mathcal{G}_i)\\ 0 & ,\mathcal{G}_{i+1}\notin\mathcal{N}(\mathcal{G}_i)\end{array}\right\} \tag{7}$$

where $\mathcal{N}(\mathcal{G}_i)$ denotes the *neighbourhood* of $\mathcal{G}_i$, that is, the collection of all DAGs that can be reached from $\mathcal{G}_i$ by deletion, addition or reversal of one single edge of the current graph $\mathcal{G}_i$, and $|\mathcal{N}(\mathcal{G}_i)|$ is the cardinality of this collection. We note that the new graph $\mathcal{G}_{i+1}$ has to be acyclic, so it has to be checked which edges can be added to $\mathcal{G}_i$ and which edges can be reversed in $\mathcal{G}_i$ without violating the acyclicity-constraint. In the Metropolis Hastings algorithm the proposed graph $\mathcal{G}_{i+1}$ is accepted with the acceptance probability: $A(\mathcal{G}_{i+1}|\mathcal{G}_i)$ =min$\{1,R(\mathcal{G}_{i+1}|\mathcal{G}_i)\}$, where

$$\begin{aligned}R(\mathcal{G}_{i+1}|\mathcal{G}_i) \;&:=\; \frac{P(\mathcal{G}_{i+1}|\mathcal{D})}{P(\mathcal{G}_i|\mathcal{D})}\cdot\frac{Q(\mathcal{G}_i|\mathcal{G}_{i+1})}{Q(\mathcal{G}_{i+1}|\mathcal{G}_i)} \\ &=\; \frac{P(\mathcal{D}|\mathcal{G}_{i+1})\cdot P(\mathcal{G}_{i+1})}{P(\mathcal{D}|\mathcal{G}_i)\cdot P(\mathcal{G}_i)}\cdot\frac{|\mathcal{N}(\mathcal{G}_i)|}{|\mathcal{N}(\mathcal{G}_{i+1})|}\end{aligned} \tag{8}$$

while the Markov chain is left unchanged, symbolically $\mathcal{G}_{i+1}:=\mathcal{G}_i$, if the new graph $\mathcal{G}_{i+1}$ is not accepted. $\{\mathcal{G}_i\}$ is then a Markov chain in the space of DAGs whose Markov transition kernel $\mathcal{T}(\tilde{\mathcal{G}}|\mathcal{G})$ for a move from $\mathcal{G}$ to $\tilde{\mathcal{G}}$ is given by the product of the proposal probability and the acceptance probability for $\mathcal{G}\neq\tilde{\mathcal{G}}$:

$$\mathcal{T}(\tilde{\mathcal{G}}|\mathcal{G})=Q(\tilde{\mathcal{G}}|\mathcal{G})\cdot A(\tilde{\mathcal{G}}|\mathcal{G}) \tag{9}$$

and

$$\mathcal{T}(\mathcal{G}|\mathcal{G})=1-\sum_{\tilde{\mathcal{G}}\in\mathcal{N}(\mathcal{G})}Q(\tilde{\mathcal{G}}|\mathcal{G})\cdot A(\tilde{\mathcal{G}}|\mathcal{G}).$$

Per construction it is guaranteed that the Markov transition kernel satisfies the equation of detailed balance:

$$\frac{P(\tilde{\mathcal{G}}|\mathcal{D})}{P(\mathcal{G}|\mathcal{D})}=\frac{\mathcal{T}(\tilde{\mathcal{G}}|\mathcal{G})}{\mathcal{T}(\mathcal{G}|\tilde{\mathcal{G}})} \tag{10}$$

Under ergodicity, that is a sufficient condition for the Markov chain $\{\mathcal{G}_i\}$ to converge, the posterior distribution $P(\mathcal{G}|\mathcal{D})$ is the stationary distribution:

$$P(\tilde{\mathcal{G}}|\mathcal{D})=\sum_{\mathcal{G}}\mathcal{T}(\tilde{\mathcal{G}}|\mathcal{G})\cdot P(\tilde{\mathcal{G}}|\mathcal{D}). \tag{11}$$

The structure MCMC sampling scheme for static Bayesian networks (BNs) can be straightforwardly modified in order to sample dynamic Bayesian networks (DBNs). For (static) BNs the *neighbourhood* of a DAG $\mathcal{G}$ in Eq. (7) is defined as the collection of all DAGs that can be reached from $\mathcal{G}$ by deletion, addition or reversal of one single edge. For DBNs we define that the neighbourhood of a (not-necessarily acyclic) directed graph is the collection of all (not necessarily acyclic) directed graphs that can be reached from $\mathcal{G}$ either by deletion or by addition of one single edge.

A reasonable approach adopted in most Bayesian network applications is to impose a limit on the cardinality of the parent sets. This limit is referred to as the *fan-in*. The practical advantage of the restriction on the maximum number of edges converging on a node is a reduction of the computational complexity, which improves the convergence. Fan-in restrictions can be justified in the context of biological expression data, as many experimental results have shown that the expression of a gene is usually controlled by a comparatively small number of active regulator genes, while on the other hand regulator-genes seem to be nearly unrestricted in the number of genes they regulate. The imputation of a fan-in restriction leads to a further reduction of the graph's neighbourhoods: Graphs that contain nodes with too many parents, that is more than the fan-in value, have to be removed from the respective neighbourhoods.

1.3 Posterior probability of edges and AUROC diagnostics

Structure MCMC can be used to generate a graph sample $\mathcal{G}_1,...,\mathcal{G}_T$, and usually the next step is to compute posterior probabilities of edges. We focus on *undirected edges* for independent data (BNs) and *directed edges* for time-dependent (DBN) data. There is an undirected edge between X_i and X_j ($i<j$) in $\mathcal{G}$ if it possesses either the edge $X_i\rightarrow X_j$ or the edge $X_i\leftarrow X_j$, and there is a directed edge from X_i to X_j ($i\neq j$) in the graph $\mathcal{G}$ if it possesses the edge $X_i\rightarrow X_j$. An estimator for the posterior probabilities of an edge F is given by the fraction of graphs in the sample that contain the edge of interest:

$$\widehat{P(F|\mathcal{D})}=\frac{1}{T}\sum_{t=1}^{T}I_F(\mathcal{G}_t) \tag{12}$$

where I_F is a binary indicator variable over the space of graphs, which is 1 if the edge F is present in the DAG, and 0 otherwise.

When the true graph or at least a gold-standard graph for the domain is known, the concept of *ROC curves* and *AUROC values* can be used to evaluate the network reconstruction accuracy of the Bayesian network inference. We assume that $e_{ij}=1$ indicates that there is an (directed/undirected) edge between X_i and X_j in the true graph, while $e_{ij}=0$ indicates that this edge is not given in the true graph. Bayesian network inference outputs a posterior probability

estimate $\widehat{P(F_{ij}|\mathcal{D})}$ for each edge e_{ij}.
Let $\epsilon(\theta) = \left\{e_{ij}|\widehat{P(F_{ij}|\mathcal{D})} > \theta\right\}$ denote the set of all edges whose posterior probabilitiy estimates exceed a given threshold θ. Given θ the number of true positive (TP), false positive (FP), and false negative (FN) edge (relation) feature findings can be counted, and the *sensitivity* $S = TP/(TP + FN)$ and the *inverse specificity* $I = FP/(TN + FP)$ can be computed. But rather than selecting an arbitrary value for the threshold θ, 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 *receiver operator charcteristic* (ROC) curve. A quantitative measure for the learning performance can be obtained by integrating the ROC curve so as to obtain the area under the ROC curve, which is usually referred to as AUROC_1 value. We note that larger AUROC_1 values indicate a better learning performance, whereby 1 is an upper limit and corresponds to a perfect estimator, while 0.5 corresponds to a random estimator.
An alternative and more intuitive criteria is given by $(TP|FP = 5)$ counts: For each MCMC output a threshold ψ is imposed on the inferred edge posterior probabilities such that 5 false positive (FP) edges are extracted and the corresponding number of true positive (TP) edges, symbolically $(TP|FP = 5)$, exceeding the threshold ψ, is counted (Werhli *et al.*, 2006).

2 THE GAUSSIAN MIXTURE APPROACH FOR BAYESIAN NETWORKS

In this section we motivate the proposed Gaussian mixture approach for Bayesian networks (BGM). The BGM model is based on the idea that the joint probability distribution $P(X_1, \ldots, X_N)$ can be replaced by a mixture distribution:

$$P(X_1, \ldots, X_N|\mathcal{K}, \vec{q}) = \sum_{k=1}^{\mathcal{K}} \lambda_k P(X_1, \ldots, X_N|\vec{q}_k) \tag{13}$$

whose number of mixture components $\mathcal{K}$, mixture weights $\vec{\lambda} = (\lambda_1, \ldots, \lambda_{\mathcal{K}})^T$, and mixture components' parameters in the vector $\vec{q} = \left(\vec{q}_1^T, \ldots, \vec{q}_k^T\right)^T$ are regarded as unknowns.

The local probability distributions $P(X_n|\pi_n)$ in Eq. (1) can then be factorised accordingly, and we obtain:

$$P(X_1, \ldots, X_N|\mathcal{K}, \vec{q}) = \sum_{k=1}^{\mathcal{K}} \lambda_k \prod_{n=1}^{N} P(X_n|\pi_n, \vec{q}_k) \tag{14}$$

Moreover, we assume that independent priors can be assigned to the parameters in $\vec{q}$:

$$P(q|\mathcal{K}, \vec{\phi}) = \prod_{k=1}^{\mathcal{K}} P(\vec{q}_k|\vec{\phi}_k) \tag{15}$$

where $\vec{\phi}_k$ is the set of hyperparameters for the prior distribution of the parameters $\vec{q}_k$ of the k-th mixture component, and $\vec{\phi} = \left(\vec{\phi}_1^T, \ldots, \vec{\phi}_{\mathcal{K}}^T\right)^T$.

In classical Bayesian network approaches the marginal likelihood of a data set $\mathcal{D}$ given a graph $\mathcal{G}$ is the integral over the parameter space:

$$P(\mathcal{D}|\mathcal{G}) = \int P(\mathcal{D}, \vec{q}|\mathcal{G}) d\vec{q} \tag{16}$$

$$= \int P(\mathcal{D}|\vec{q}, \mathcal{G}) P(\vec{q}|\mathcal{G}) d\vec{q} \tag{17}$$

and a closed-form solution for the BDe and BGe model can be derived under two fairly weak assumptions. *Parameter independence* means that the prior distribution $P(\vec{q}|\mathcal{G})$ of the unknown parameters $\vec{q}$ can be factorised into a product of N subsets of parameters $\vec{q}_{(n)}$ each associated with a local probability distribution:

$$P(\vec{q}|\mathcal{G}) = \prod_{n=1}^{N} P(\vec{q}_{(n)}|\mathcal{G}) \tag{18}$$

whereby $\vec{q}_{(n)}$ consists of those parameters required for parameterising the local probability distribution X_n given graph $\mathcal{G}$.

Parameter modularity means that the probability of the parameter subset $\vec{q}_n$ in the local probability distribution $P(\vec{q}_{(n)}|\mathcal{G})$ depends on the parent variables π_n of X_n in $\mathcal{G}$ only. That is, for $n = 1, \ldots, N$ it holds:

$$P(\vec{q}_{(n)}|\mathcal{G}) = P(\vec{q}_{(n)}|\pi_n) \tag{19}$$

Let $\mathcal{D}(n, .)$ denote the observations of the n-th domain node X_n in the data $\mathcal{D}$, and $\mathcal{D}(\pi_n, .)$ denotes the observations of X_n's parent nodes π_n in $\mathcal{D}$. Under the assumption of parameter independence the likelihood can be factorised according to Eq. (1):

$$P(\mathcal{D}|\mathcal{G}, \vec{q}) = \prod_{n=1}^{N} P(X_n = \mathcal{D}(n, .)|\pi_n = \mathcal{D}(\pi_n, .), \vec{q}_{(n)}) \tag{20}$$

Inserting Eq. (18), Eq. (19), and Eq. (20) in Eq. (16) yields:

$$P(\mathcal{D}|\mathcal{G}) =$$

$$\prod_{n=1}^{N} \int P(X_n = \mathcal{D}(n, .)|\vec{q}_n, \pi_n = \mathcal{D}(\pi_n, .)) P(\vec{q}_{(n)}|\pi_n) d\vec{q}_{(n)}$$

This can be straightforwardly extended to the BGM model when the assumptions of parameter independence and parameter modularity are extended with respect to a mixture model approach. For $k = 1, \ldots \mathcal{K}$ it can be assumed that:

$$P(\vec{q}_k|\mathcal{G}) = \prod_{n=1}^{N} P(\vec{q}_{k,(n)}|\mathcal{G}) \tag{21}$$

and

$$P(\vec{q}_{k,(n)}|\mathcal{G}) = P(\vec{q}_{k,(n)}|\pi_n) \tag{22}$$

where $\vec{q}_{k,(n)}$ consists of those parameters required for parameterising the local probability distribution of X_n given a graph $\mathcal{G}$, in which the parent set of X_n is π_n, in the k-th mixture distribution.

For the Gaussian mixture model the likelihood then factorises as follows:

$$P(\mathcal{D}|\mathcal{G}, \mathcal{K}, \vec{q}) = \sum_{k=1}^{\mathcal{K}} \lambda_k \prod_{n=1}^{N} P(X_n = \mathcal{D}(n, .)|\pi_n = \mathcal{D}(\pi_n, .), \vec{q}_{k,(n)}) \tag{23}$$

what in turn can be interpreted as a mixture of Bayesian network BGe likelihoods (see Eq. (20)) where $\vec{q}_k = \left(\vec{q}_{k,(1)}^T, \ldots, \vec{q}_{k,(n)}^T\right)^T$ is the parameter vector associated with the k-th Bayesian network model in the mixture distribution.

3 GAUSSIAN MIXTURE ALLOCATION MCMC INFERENCE

The third section of this supplementary paper on theoretical aspects (T) deals with the novel non-linear and non-homogeneous generalization of the classical BGe score for Bayesian networks. Gaussian mixture allocation MCMC inference (BGM) is based on a mixture model, using latent variables to assign individual measurements (observations of the domian) to different classes (mixture components). The practical inference follows the Bayesian paradigm and samples the graph structure, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the recently proposed allocation sampler (Nobile and Fearnside, 2007) as an alternative to RJMCMC. In the first subsection we present the new BGM model. Subsequently, in the second subsection we describe the BGM sampling scheme in detail. Finally, in the third subsection we discuss all different MCMC move types in detail.

3.1 Gaussian mixture Bayesian network model

We assume that we have either m independent and identically distributed (iid) observations (BNs) or m time dependent observations with a homogeneous first-order Markovian dependence structure (DBNs) for the variables $X_1, \dots X_N$. This gives a data set matrix of size N-by-m where $\mathcal{D}_{\cdot,j}$ $(j = 1, \dots, m)$ is the j-th observation of the N nodes. The allocation vector $\vec{\mathcal{V}}$ of size m defines an allocation of the m observations to $\mathcal{K}$ mixture components: $\vec{\mathcal{V}}(j) = k$ means that the j-th observation is allocated to the k-th component. $\mathcal{D}^{(\vec{\mathcal{V}},k)}$ denotes the data subset consisting of all observations allocated to the k-th component by $\vec{\mathcal{V}}$ $(1 \leq k \leq \mathcal{K})$. We assume that the joint posterior probability of a graph $\mathcal{G}$, an allocation vector $\vec{\mathcal{V}}$, and $\mathcal{K}$ mixture components can be factorised as follows:

$$\begin{aligned} P(\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K}|\mathcal{D}) &= \frac{P(\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K}, \mathcal{D})}{P(\mathcal{D})} \propto P(\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K}, \mathcal{D}) \qquad (24) \\ &= P(\mathcal{K}) \cdot P(\vec{\mathcal{V}}|\mathcal{K}) \cdot P(\mathcal{G}) \cdot P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K}) \end{aligned}$$

where

$$P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K}) = \prod_{k=1}^{\mathcal{K}} P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G}) \qquad (25)$$

In Eq. (25) the likelihood terms $P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})$ for the data subsets $\mathcal{D}^{(\vec{\mathcal{V}},k)}$ given the same graph $\mathcal{G}$ can be computed independently with the BGe scoring metric (Geiger and Heckerman, 1994). If no observation is allocated to the k-th component ($\mathcal{D}^{(\vec{\mathcal{V}},k)} = \emptyset$), $P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})$ is equal to 1. Following Nobile and Fearnside (2007) we assume as prior on $\mathcal{K}$ the Poisson distribution with parameter $\lambda = 1$ restricted to $1 \leq \mathcal{K} \leq \mathcal{K}_{MAX}$ and that the probability distribution of the allocation vector $\vec{\mathcal{V}}$ conditional on $\mathcal{K}$ is given by:

$$P(\vec{\mathcal{V}} = \vec{v}|\mathcal{K}, p) = \prod_{k=1}^{\mathcal{K}} p_k^{n_k} \qquad (26)$$

where $\vec{p} = (p_1, \dots, p_{\mathcal{K}})$ with $\sum_{k=1}^{\mathcal{K}} p_k = 1$ are the non-negative mixture weights, and n_k is the number of observations allocated to the k-th mixture component by $\vec{\mathcal{V}}$. The prior on the mixture weights $\vec{p} = (p_1, \dots, p_{\mathcal{K}})^T$ is chosen to be a Dirichlet distribution $Dir(\alpha_1, \dots, \alpha_{\mathcal{K}})$ with hyperparameters $\vec{\alpha} = (\alpha_1, \dots \alpha_{\mathcal{K}})^T$ so that the posterior probability of $\vec{\mathcal{V}}$ conditional on $\mathcal{K}$ is given by $Dir(n_1 + \alpha_1, \dots, n_{\mathcal{K}} + \alpha_{\mathcal{K}})$:

$$\begin{aligned} P(\vec{\mathcal{V}}|\mathcal{K}) &= \int d\vec{p} P(\vec{\mathcal{V}} = \vec{v}|\mathcal{K}, \vec{p}) \cdot P(\vec{p}) \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + m)} \cdot \prod_{k=1}^{\mathcal{K}} \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)} \end{aligned}$$

where $\alpha_0 = \alpha_1 + \ldots + \alpha_{\mathcal{K}}$.

We know that in DBNs the variables at time point $t-1$ are potential parent nodes of the variables a time point t. And we know that the effective number of observations (sample size) for dynamic Bayesian networks is therefore equal to $m - 1$, as no observations for the potential parent nodes of the domain variables at time point $t = 1$ are available. Bearing this in mind, we interpret the allocation vector $\vec{\mathcal{V}}$ for DBNs as follows: For $t = 1, \dots, m-1$, $\vec{\mathcal{V}}(t) = k$ means that the domain variables $X_{i,t+1}$ at time point $t + 1$, whose potential parent nodes are the domain variables $X_{1,t}, \dots, X_{N,t}$ at time point t, are allocated to the k-th mixture component. From this point of view the m-th (last) entry of the allocation vector is redundant and can be excluded from all operations that may change its value. Therefore, for the remainder of this paper we assume that the length of the allocation vectors $\vec{\mathcal{V}}$ are decreased by 1 ($m - 1$ instead of m) when they correspond to dynamic Bayesian network (DBN) models.

3.2 MCMC inference

The new Gaussian mixture Allocation MCMC sampling scheme (BGM) generates a sample from the joint posterior distribution $P(\mathcal{G}, \mathcal{K}, \vec{\mathcal{V}}|\mathcal{D})$ given in Eq. (24) and comprises five different types of moves in the state-space $[\mathcal{G}, \mathcal{K}, \vec{\mathcal{V}}]$. The first move type is a classical structure MCMC single edge operation on the graph $\mathcal{G}$ while the number of components $\mathcal{K}$ and the allocation vector $\vec{\mathcal{V}}$ are left unchanged. According to Eq. (7) a new graph $\tilde{\mathcal{G}}$ is proposed, and the new state $[\tilde{\mathcal{G}}, \mathcal{K}, \vec{\mathcal{V}}]$ is accepted according to Eq (8) where the likelihood terms $P(\mathcal{D}|\mathcal{G})$ in Eq. (8) have to be replaced by $P(\mathcal{D}|\mathcal{G}, \mathcal{K}, \vec{\mathcal{V}})$ terms given in Eq. (25). The four other move types are adapted from Nobile and Fearnside (2007) and operate on $\vec{\mathcal{V}}$ or on $\mathcal{K}$ and $\vec{\mathcal{V}}$. If there are $\mathcal{K} > 2$ mixture components, then moves of the type M1 and M2 can be used to re-allocate some observations from one component k to another one $\tilde{k}$. That is, a new allocation vector $\vec{\mathcal{V}}^*$ is proposed while $\mathcal{G}$ and $\mathcal{K}$ are left unchanged. The EA move type changes $\mathcal{K}$ and $\vec{\mathcal{V}}$. An ejection EA move proposes to increase the number of mixture components by 1 and simultaneously tries to re-allocate some observations to fill the new component. More precisely, it randomly selects a mixture component and tries to re-allocate some of its observations to the newly proposed component $\mathcal{K} + 1$ while $\mathcal{G}$ is left unchanged. Absorption EA moves are complementary to ejection EA moves and decrease the number of mixture components by 1. An EA absorption move randomly selects two mixture components and deletes one of them after having re-allocated all its observations to the other component. The acceptance probabilities for M1, M2, EA ejection,

and EA absorption moves are of the same functional form:

$$A = \left\{ 1, \frac{P(\vec{\mathcal{V}}^*|\mathcal{K}^*)}{P(\vec{\mathcal{V}}|\mathcal{K})} \cdot \frac{P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}}^*,\mathcal{K}^*)}{P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})} \cdot \frac{Q(\vec{\mathcal{V}}^*|\vec{\mathcal{V}})}{Q(\vec{\mathcal{V}}|\vec{\mathcal{V}}^*)} \cdot \frac{P(\mathcal{K}^*)}{P(\mathcal{K})} \right\} \tag{27}$$

where the likelihood terms have been specified in Eq. (25), the proposal probabilities $Q(.|.)$ depend on the move type (M1, M2, EA), and $\mathcal{K}^* = \mathcal{K}$ for M1 and M2 moves, and $\mathcal{K}^* \in \mathcal{K}-1, \mathcal{K}+1$ for EA moves. Finally, the Gibbs move re-allocates only one single observation by sampling its new allocation from the coressponding Boltzmann distribution (see Nobile and Fearnside (2007)) while leaving $\mathcal{K}$ and $\vec{\mathcal{V}}$ unchanged. The next subsection discusses all BGM moves in detail.

3.3 Moves for BGM in detail

Before the MCMC simulation is started, probabilities p_i $(i = 1,\ldots,5)$ with $p_1+\ldots+p_5 = 1$ must be predefined with which one of these move types (structure, M1, M2, Gibbs, EA) is selected. The classical structure MCMC move type (Madigan and York (1995)) changes the graph $\mathcal{G}$ and leaves the number of components $\mathcal{K}$ and the allocation vector $\vec{\mathcal{V}}$ unchanged. The other move types are immediately adopted from Nobile and Fearnside (2007).

3.3.1 Structure MCMC Move on the graph $\mathcal{G}$: The first move type is a standard structure MCMC move in the graph space. It proposes to change the current graph $\mathcal{G}$ by adding, deleting or reversing a single edge as explained in detail in Section 1. The acceptance probability for a move from $\left[\mathcal{G},\mathcal{K},\vec{\mathcal{V}}\right]$ to $[\tilde{\mathcal{G}},\mathcal{K},\vec{\mathcal{V}}]$ is given by: $A = \min\{1,R\}$ where

$$\begin{aligned} R &= \frac{P(\tilde{\mathcal{G}},\mathcal{K},\vec{\mathcal{V}}|\mathcal{D})}{P(\mathcal{G},\mathcal{K},\vec{\mathcal{V}}|\mathcal{D})} \cdot \frac{Q(\mathcal{G}|\tilde{\mathcal{G}})}{Q(\tilde{\mathcal{G}}|\mathcal{G})} \\ &= \frac{P(\mathcal{K}) \cdot P(\vec{\mathcal{V}}|\mathcal{K}) \cdot P(\tilde{\mathcal{G}}) \cdot P(\mathcal{D}|\tilde{\mathcal{G}},\vec{\mathcal{V}},\mathcal{K})}{P(\mathcal{K}) \cdot P(\vec{\mathcal{V}}|\mathcal{K}) \cdot P(\mathcal{G}) \cdot P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})} \cdot \frac{Q(\mathcal{G}|\tilde{\mathcal{G}})}{Q(\tilde{\mathcal{G}}|\mathcal{G})} \\ &= \frac{P(\tilde{\mathcal{G}})}{P(\mathcal{G})} \cdot \prod_{k=1}^{\mathcal{K}} \frac{P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\tilde{\mathcal{G}})}{P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})} \cdot \frac{|\mathcal{N}(\mathcal{G})|}{|\mathcal{N}(\tilde{\mathcal{G}})|} \end{aligned} \tag{28}$$

where $Q(\mathcal{G}|\tilde{\mathcal{G}})$ and $Q(\tilde{\mathcal{G}}|\mathcal{G})$ are the proposal probabilities for moves from $\tilde{\mathcal{G}}$ to $\mathcal{G}$ and vice-versa, $\mathcal{N}(\mathcal{G})$ and $\mathcal{N}(\tilde{\mathcal{G}})$ are the sets of neighbour graphs of $\mathcal{G}$ and $\tilde{\mathcal{G}}$, and Eq. (25) was used for factorising the likelihoods $P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\tilde{\mathcal{G}})$ and $P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})$.

3.3.2 Gibbs Move on the allocation vector $\vec{\mathcal{V}}$: If there is one component only, symbolically $\mathcal{K} = 1$, select another move type. Otherwise randomly select an observation i among the m available and determine to which component k $(1 \leq k \leq \mathcal{K})$ this observation currently belongs. For each mixture component $\tilde{k} = 1,\ldots,\mathcal{K}$ replace the i-th entry of the allocation vector $\vec{\mathcal{V}}$ by component $\tilde{k}$ to obtain $\vec{\mathcal{V}}(i \leftarrow \tilde{k})$ $(\tilde{k} = 1,\ldots,\mathcal{K})$. We note that $\vec{\mathcal{V}}(i \leftarrow k)$ is equal to the current allocation vector $\vec{\mathcal{V}}$. Subsequently, sample the new allocation vector $\vec{\mathcal{V}}^*$ from the full conditional distribution: For $\tilde{k} = 1,\ldots,\mathcal{K}$:

$$P(\vec{\mathcal{V}}^* = \vec{\mathcal{V}}(i \leftarrow \tilde{k})) := \frac{P(\mathcal{G},\vec{\mathcal{V}}(i \leftarrow \tilde{k}),\mathcal{K}|\mathcal{D})}{\sum_{k^*=1}^{\mathcal{K}} P(\mathcal{G},\vec{\mathcal{V}}(i \leftarrow k^*),\mathcal{K}|\mathcal{D})} \tag{29}$$

whereby it can be shown that the ratio on the right is equal to:

$$\frac{P(\vec{\mathcal{V}}(i \leftarrow \tilde{k})|\mathcal{K}) \cdot \prod_{j \in k,\tilde{k}} P(\mathcal{D}^{(\vec{\mathcal{V}}(i \leftarrow \tilde{k},j)}|\mathcal{G})}{\sum_{k^*=1}^{\mathcal{K}} \left\{ P(\vec{\mathcal{V}}(i \leftarrow k^*)|\mathcal{K}) \cdot \prod_{j \in k,k^*} P(\mathcal{D}^{(\vec{\mathcal{V}}(i \leftarrow k^*),j)}|\mathcal{G}) \right\}}$$

See Nobile and Fearnside (2007) for further details on this systematic sweep Gibbs move.

3.3.3 The M1 Move on the allocation vector $\vec{\mathcal{V}}$: If there is one component only, symbolically $\mathcal{K} = 1$, select a different type of move. Otherwise randomly select two mixture components k and $\tilde{k}$ among the $\mathcal{K}$ available. Draw a random number $\tilde{p}$ from a Beta distribution whose parameters are equal to the corresponding hyperparameters α_k and $\alpha_{\tilde{k}}$ of the Dirichlet prior on the mixture weights. Re-allocating each observation currently belonging to the k-th or $\tilde{k}$-th component to component k with probability $\tilde{p}$ or to component $\tilde{k}$ with probability $1-\tilde{p}$ gives the new allocation vector $\vec{\mathcal{V}}^*$. Nobile and Fearnside (2007) show that for M1 proposal probabilities holds:

$$\frac{Q(\vec{\mathcal{V}}^*|\vec{\mathcal{V}})}{Q(\vec{\mathcal{V}}|\vec{\mathcal{V}}^*)} = \left\{ \frac{P(\vec{\mathcal{V}}^*|\mathcal{K})}{P(\vec{\mathcal{V}}|\mathcal{K})} \right\}^{-1}$$

so that the corresponding terms in Eq. (27) cancel out. Furthermore, as the number of components $\mathcal{K}$ is not changed either, all that remains to compute is the likelihood ratio: $\frac{P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}}^*,\mathcal{K})}{P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})}$. For M1 moves all except the k-th and the $\tilde{k}$-th factor cancel out from the ratio when the likelihoods are factorised according to Eq. (25). Hence the acceptance probability for an M1 move from $[\mathcal{G},\mathcal{K},\vec{\mathcal{V}}]$ to $[\mathcal{G},\mathcal{K},\vec{\mathcal{V}}^*]$ is given by:

$$A = min\left\{ 1, \frac{P(\mathcal{D}^{(\vec{\mathcal{V}}^*,k)}|\mathcal{G})}{P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})} \cdot \frac{P(\mathcal{D}^{(\vec{\mathcal{V}}^*,\tilde{k})}|\mathcal{G})}{P(\mathcal{D}^{(\vec{\mathcal{V}},\tilde{k})}|\mathcal{G})} \right\} \tag{30}$$

See Nobile and Fearnside (2007) for further details on the M1 move.

3.3.4 The M2 Move on the allocation vector $\vec{\mathcal{V}}$: If there is one component only, symbolically $\mathcal{K} = 1$, select a different move type. Otherwise randomly select two mixture components k and $\tilde{k}$ among the $\mathcal{K}$ available and then randomly select a group of observations allocated to component k and attempt to re-allocate them to component $\tilde{k}$. If the k-th component is empty the move fails outright. Otherwise draw a random number u from a uniform distribution on $1,\ldots,n_k$ where n_k is the number of observation allocated to the k-th component. Subsequently, randomly select u observations from the n_k in component k and allocate the selected observations to component $\tilde{k}$ to obtain the new allocation vector $\vec{\mathcal{V}}^*$. As $\mathcal{K}$ is not changed and all except the k-th and the $\tilde{k}$-th factor cancel out from the ratio when the likelihoods are factorised according to Eq. (25), the acceptance probability for an M2 move from $[\mathcal{G},\mathcal{K},\vec{\mathcal{V}}]$ to $[\mathcal{G},\mathcal{K},\vec{\mathcal{V}}^*]$ is given by:

$$A = \left\{ 1, \frac{P(\vec{\mathcal{V}}^*|\mathcal{K})}{P(\vec{\mathcal{V}}|\mathcal{K})} \cdot \frac{\prod_{j \in k,\tilde{k}} P(\mathcal{D}^{(\vec{\mathcal{V}}^*,j)}|\mathcal{G})}{\prod_{j \in k,\tilde{k}} P(\mathcal{D}^{(\vec{\mathcal{V}},j)}|\mathcal{G})} \cdot \frac{Q(\vec{\mathcal{V}}^*|\vec{\mathcal{V}})}{Q(\vec{\mathcal{V}}|\vec{\mathcal{V}}^*)} \right\} \tag{31}$$

Nobile and Fearnside (2007) show that for the proposal probability ratio holds:

$$\frac{Q(\vec{\mathcal{V}}^*|\vec{\mathcal{V}})}{Q(\vec{\mathcal{V}}|\vec{\mathcal{V}}^*)} = \frac{n_k}{n_{\tilde{k}}+u} \cdot \frac{n_k! \cdot n_{\tilde{k}}!}{(n_k-u)! \cdot (n_{\tilde{k}}+u)!} \tag{32}$$

where n_k and $n_{\tilde{k}}$ are the numbers of observations allocated to the k-th and $\tilde{k}$-th component by $\vec{\mathcal{V}}$. See Nobile and Fearnside (2007) for further details on the M2 move.

3.3.5 EA (ejection/absorption) moves on the number of components $\mathcal{K}$ and the allocation vector $\vec{\mathcal{V}}$: If there is only one component, symbolically $\mathcal{K} = 1$, then an ejection move has to be performed. If the maximal number of components is currently given, symbolically $\mathcal{K} = \mathcal{K}_{MAX}$, then an absorption move has to be performed. If $1 < \mathcal{K} < \mathcal{K}_{MAX}$ then perform an ejection move with probability 0.5 and otherwise an absorption move.

The ejection move
Randomly select a mixture component k $(1 \leq k < \mathcal{K})$ as the ejecting component. Make a draw p_E from a $Beta(a,a)$ distribution and re-allocate each observation currently allocated to component k in the vector $\vec{\mathcal{V}}$ with probability p_E to a new (rejected) component with label $\mathcal{K}+1$. Subsequently swap the labels of the new (rejected) mixture component $\mathcal{K}+1$ with a randomly choosen mixture component label $\tilde{k}$ including the label $\mathcal{K}+1$ of the ejected component itself $(1 \leq \tilde{k} \leq \mathcal{K}+1)$ to obtain the allocation vector $\vec{\mathcal{V}}^*$. Nobile and Fearnside (2007) show that the acceptance probability for an EA ejection move from $[\mathcal{G}, \mathcal{K}, \vec{\mathcal{V}}]$ to $[\mathcal{G}, \mathcal{K}^*, \vec{\mathcal{V}}^*]$ is given by $A = \{1, R\}$ where:

$$R = \frac{P(\vec{\mathcal{V}}^*|\mathcal{K}^*)}{P(\vec{\mathcal{V}}|\mathcal{K})} \cdot \frac{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}^*, \mathcal{K}^*)}{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K})} \cdot \frac{Q([\vec{\mathcal{V}}^*, \mathcal{K}^*]|[\vec{\mathcal{V}}, \mathcal{K}])}{Q([\vec{\mathcal{V}}, \mathcal{K}]|[\vec{\mathcal{V}}^*, \mathcal{K}^*])} \cdot \frac{P(\mathcal{K}^*)}{P(\mathcal{K})}$$

and $\mathcal{K}^* = \mathcal{K}+1$.

Nobile and Fearnside (2007) show that for the ratio of the proposal probabilities holds:

$$\frac{Q([\vec{\mathcal{V}}^*, \mathcal{K}]|[\vec{\mathcal{V}}, \mathcal{K}])}{Q([\vec{\mathcal{V}}, \mathcal{K}]|[\vec{\mathcal{V}}^*, \mathcal{K}^*])} = p_E \cdot \frac{\Gamma(a)^2}{\Gamma(2a)} \cdot \frac{\Gamma(2a+n_k)}{\Gamma(a+n_w^*)\Gamma(a+n_k^*)}$$

where $w = \tilde{k}$ if $\tilde{k} \neq k$, and $w = \mathcal{K}+1$ if $\tilde{k} = k$, n_k is the number of observations allocated to the k-th component in $\vec{\mathcal{V}}$, n_w^* and n_k^* are the numbers of observations allocated to the w-th and k-th component by $\vec{\mathcal{V}}^*$. Furthermore, it holds: $p_E = 0.5$ if $\mathcal{K} = 1$, $p_E = 2$ if $\mathcal{K} = \mathcal{K}_{MAX} - 1$, and $p_E = 1$ otherwise. For the likelihood ratio holds:

$$\frac{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}^*, \mathcal{K}^*)}{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K})} = \frac{P(\mathcal{D}^{(\vec{\mathcal{V}}^*,k)}|\mathcal{G}) \cdot P(\mathcal{D}^{(\vec{\mathcal{V}}^*,w)}|\mathcal{G})}{P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G})}$$

where $w = \tilde{k}$ if $\tilde{k} \neq k$, and $w = \mathcal{K}+1$ if $\tilde{k} = k$. Following Nobile and Fearnside (2007) the parameter a of the $Beta(a,a)$ distribution can be selected by numerically solving the following equation:

$$\frac{\Gamma(2a)}{\Gamma(a)} \cdot \frac{\Gamma(a+n_k)}{\Gamma(2a+n_k)} = 0.1$$

whereby a lookup table was used in our BGM implementation. See Nobile and Fearnside (2007) for further details.

The absorption move
Randomly select a mixture component k $(1 \leq k \leq \mathcal{K})$ as the absorbing component and another component $\tilde{k}$ $(1 \leq \tilde{k} \leq \mathcal{K}$ with $\tilde{k} \neq k)$ as the disappearing component. Re-allocate all observations currently allocated to the disappearing component $\tilde{k}$ by $\vec{\mathcal{V}}$ to component k to obtain the new allocation vector $\vec{\mathcal{V}}^*$. Then delete the (empty) component $\tilde{k}$ to obtain the new number of components $\mathcal{K}^* = \mathcal{K}-1$.

Nobile and Fearnside (2007) show that the acceptance probability for an EA absorption move from $[\mathcal{G}, \mathcal{K}, \vec{\mathcal{V}}]$ to $[\mathcal{G}, \mathcal{K}^*, \vec{\mathcal{V}}^*]$ is given by $A = \{1, R\}$ where:

$$R = \frac{P(\vec{\mathcal{V}}^*|\mathcal{K}^*)}{P(\vec{\mathcal{V}}|\mathcal{K})} \cdot \frac{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}^*, \mathcal{K}^*)}{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K})} \cdot \frac{Q([\vec{\mathcal{V}}^*, \mathcal{K}^*]|[\vec{\mathcal{V}}, \mathcal{K}])}{Q([\vec{\mathcal{V}}, \mathcal{K}]|[\vec{\mathcal{V}}^*, \mathcal{K}^*])} \cdot \frac{P(\mathcal{K}^*)}{P(\mathcal{K})}$$

and $\mathcal{K}^* = \mathcal{K}-1$.

Nobile and Fearnside (2007) show that for the ratio of the proposal probabilities holds:

$$\frac{Q([\vec{\mathcal{V}}^*, \mathcal{K}]|[\vec{\mathcal{V}}, \mathcal{K}])}{Q([\vec{\mathcal{V}}, \mathcal{K}]|[\vec{\mathcal{V}}^*, \mathcal{K}^*])} = p_A \cdot \frac{\Gamma(2a)}{\Gamma(a)^2} \cdot \frac{\Gamma(a+n_{\tilde{k}})\Gamma(a+n_k)}{\Gamma(2a+n_k^*)}$$

where n_k^* is the number of observations allocated to the k-th component in $\vec{\mathcal{V}}^*$, n_k and $n_{\tilde{k}}$ are the numbers of observations allocated to the k-th and $\tilde{k}$-th component by $\vec{\mathcal{V}}$. Furthermore, it holds: $p_A = 0.5$ if $\mathcal{K} = \mathcal{K}_{MAX}$, $p_A = 2$ if $\mathcal{K} = 2$, and $p_A = 1$ otherwise. For the likelihood ratio holds:

$$\frac{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}^*, \mathcal{K}^*)}{P(\mathcal{D}|\mathcal{G}, \vec{\mathcal{V}}, \mathcal{K})} = \frac{P(\mathcal{D}^{(\vec{\mathcal{V}}^*,k)}|\mathcal{G})}{P(\mathcal{D}^{(\vec{\mathcal{V}},k)}|\mathcal{G}) \cdot P(\mathcal{D}^{(\vec{\mathcal{V}},\tilde{k})}|\mathcal{G})}$$

4 BGE SCORE AND EXTENSION TO BGM

This section deals with the standard BGe scoring metric (Bayesian metric for Gaussian networks having score equivalence) for Bayesian networks. The first subsection focuses on BGe for static data (independent observations of the domain) and dynamic data (time series of the domain). The formula for the closed-form solution of the marginal likelihood are given. In the second subsection we explain how to expand BGe to the proposed BGM model and provide all necessary formula.

4.1 BGe

Given a data set $\mathcal{D}$ with m observations of the domain $X_1, \ldots, X_N$, let $\mathcal{D}_{i,j}$ denote the j-th observation of the i-th domain node X_i, and let $\mathcal{D}_{\cdot,j} = (\mathcal{D}_{1,j}, \ldots, \mathcal{D}_{N,j})^T$ denote the j-th observation vector of the domain. The BGe model (Geiger and Heckerman (1994)) assumes that the set of observation vectors $\mathcal{D}_{\cdot,j}$ $(j = 1, \ldots, m)$ is a random sample from a multivariate Gaussian normal distribution $\mathcal{N}(\vec{\mu}, \Sigma)$ with an unknown mean vector $\vec{\mu}$ and an unknown covariance matrix Σ. The prior joint distribution of $\vec{\mu}$ and $W = \Sigma^{-1}$ is supposed to be the normal-Wishart distribution, that is, the conditional distribution of $\vec{\mu}$ given W is $\mathcal{N}(\vec{\mu}_0, v \cdot W)$ such that $v > 0$, and the marginal distribution of W is a Wishart distribution with $\alpha > N+1$ degrees of freedom and precision matrix T_0, denoted $\mathcal{W}(\alpha, T_0)$. The condition $\alpha > N+1$ ensures that the second moments of the posterior distribution are finite (see also Eq. (26) in Geiger and Heckerman (1994)). Geiger and Heckerman (1994) show that the likelihood (score) $P(\mathcal{D}|\mathcal{G})$ of the data $\mathcal{D}$ given

a graph $\mathcal{G}$ can then - under fairly weak conditions - be computed as follows: We define:

$$T_{\mathcal{D}} := T_0 + S_{\mathcal{D}} + \frac{vm}{v+m}(\vec{\mu}_0 - \overline{\mathcal{D}})(\vec{\mu}_0 - \overline{\mathcal{D}})^T \qquad (33)$$

where

$$\overline{\mathcal{D}} := \frac{1}{m}\sum_{j=1}^{m} \mathcal{D}_{.,j} \qquad (34)$$

is the mean of the m observation vectors and

$$S_{\mathcal{D}} := \sum_{j=1}^{m} (\mathcal{D}_{.,j} - \overline{\mathcal{D}}) \cdot (\mathcal{D}_{.,j} - \overline{\mathcal{D}})^T \qquad (35)$$

Furthermore, we set:

$$c(n,\alpha) := \left\{ 2^{\alpha \cdot n/2} \cdot \pi^{n\cdot(n-1)/4} \cdot \prod_{i=1}^{n} \Gamma(\frac{\alpha+1-i}{2}) \right\}^{-1} \qquad (36)$$

The likelihood can then be computed as follows (Geiger and Heckerman (1994)):

$$P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{N} \frac{P(\mathcal{D}^{\{X_i,\pi_i\}}|\mathcal{G}_F(\{X_i,\pi_i\})}{P(\mathcal{D}^{\{\pi_i\}}|\mathcal{G}_F(\pi_i))} \qquad (37)$$

where X_i is the i-th domain variable, π_i is the parent set of the i-th domain variable X_i in the graph $\mathcal{G}$, $\mathcal{D}^{\{X_i,\pi_i\}}$ and $\mathcal{D}^{\{\pi_i\}}$ are the data submatrices corresponding to the observations for the domain variables in the sets $\{X_i,\pi_i\}$ and $\{\pi_i\}$ only, and $\mathcal{G}_F(\{X_i,\pi_i\})$ and $\mathcal{G}_F(\pi_i)$ correspond to so called *full graphs* for the domain subsets $\{X_i,\pi_i\}$ and $\{\pi_i\}$, that is, to subgraphs with maximal number of edges so that the subgraphs do not impose any independency restrictions on the variables.
The likelihood of the data subset $\mathcal{D}^{\{S\}} \subset \mathcal{D}$ corresponding to the m observations of the n-dimensional subset $S \subset \{X_1,\ldots,X_N\}$ of the N domain variables given a full graph $\mathcal{G}_F(S)$ for the sub-domain S can be computed as follows (Geiger and Heckerman (1994)):

$$\begin{aligned} P(\mathcal{D}^S|\mathcal{G}_F(S)) \quad &= \quad (2\pi)^{-\frac{n\cdot m}{2}} \cdot \left\{\frac{v}{v+m}\right\}^{n/2} \cdot \frac{c(n,\alpha)}{c(n,\alpha+m)} \\ &\quad \cdot det(T_0^S)^{\frac{\alpha}{2}} \cdot det(T_{\mathcal{D}}^S)^{-\frac{\alpha+m}{2}} \end{aligned}$$

where T_0, α, and v are hyperparameters that have to be specified, and $det(T_0^S)$ and $det(T_{\mathcal{D}}^S)$ denote the determinants of the submatrices T_0^S and $T_{\mathcal{D}}^S$ consisting only of those rows and columns that correspond to variables in the subset S. $T_{\mathcal{D}}$ was defined in Eq. (33), and $c(n,\alpha)$ and $c(n,\alpha+m)$ can be computed with Eq. (36).

When (instead of independent observations) time series data $(X_{1,t},\ldots X_{N,t})_{t=1,\ldots,m}$ have been collected for the domain, 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 realisation $x_{j,t}$ of X_j at time point t is influenced by the realisation $x_{i,t-1}$ of X_i at the previous time point $t-1$. This can be taken into consideration in the context of BGe by building new matrices from the original data matrix of size N-by-m:

$$\mathcal{D} = \begin{pmatrix} \mathcal{D}_{1,1} & \mathcal{D}_{1,2} & \ldots & \mathcal{D}_{1,m-1} & \mathcal{D}_{1,m} \\ \mathcal{D}_{2,1} & \mathcal{D}_{2,2} & \ldots & \mathcal{D}_{2,m-1} & \mathcal{D}_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathcal{D}_{N,1} & \mathcal{D}_{N,2} & \ldots & \mathcal{D}_{N,m-1} & \mathcal{D}_{N,m} \end{pmatrix} \qquad (38)$$

We build the following matrices of size $(N+1)$-by-$(m-1)$:

$$\mathcal{D}(i) = \begin{pmatrix} \mathcal{D}_{1,1} & \mathcal{D}_{1,2} & \ldots & \mathcal{D}_{1,m-1} \\ \mathcal{D}_{2,1} & \mathcal{D}_{2,2} & \ldots & \mathcal{D}_{2,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{D}_{N,1} & \mathcal{D}_{N,2} & \ldots & \mathcal{D}_{N,m-1} \\ \mathcal{D}_{i,2} & \mathcal{D}_{i,3} & \ldots & \mathcal{D}_{i,m} \end{pmatrix} \qquad (39)$$

$i = 1,\ldots,N$. That is, we obtain $\mathcal{D}(i)$ by deleting the last column of $\mathcal{D}$ and adding the row $(\mathcal{D}_{i,2},\ldots,\mathcal{D}_{i,m})$, i.e. the i-th row of $\mathcal{D}$ shifted leftwards by 1, as the $(N+1)$-th row. For convenience, we identify the $(N+1)$-th row with a new domain variable X_{N+1}.

Finally, we replace Eq. (37) by:

$$P(\mathcal{D}|\mathcal{G}) = \prod_{i=1}^{N} \frac{P(\mathcal{D}(i)^{\{X_{N+1},\pi_i\}}|\mathcal{G}_F(\{X_{N+1},\pi_i\})}{P(\mathcal{D}(i)^{\{\pi_i\}}|\mathcal{G}_F(\pi_i))} \qquad (40)$$

4.2 BGM

The results of the last subsection can be straightforwardly extended to the BGM model by factorising the likelihood $P_{BGM}(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K})$ according to Eq. (25). The (static) BGM counterpart of Eq. (37) is given by:

$$P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K}) = \prod_{k=1}^{\mathcal{K}}\prod_{i=1}^{N} \frac{P(\mathcal{D}^{(\vec{\mathcal{V}},k),\{X_i,\pi_i\}}|\mathcal{G}_F(\{X_i,\pi_i\})}{P(\mathcal{D}^{(\vec{\mathcal{V}},k),\{\pi_i\}}|\mathcal{G}_F(\pi_i))} \qquad (41)$$

where $\tilde{\mathcal{D}}^{(\vec{\mathcal{V}},k),S}$ is the data subset of $\mathcal{D}$ which is restricted to those rows that correspond to variables in S and to those columns that have been assigned to component k by the allocation vector $\vec{\mathcal{V}}$.

The (dynamic) BGM counterpart of Eq. (40) is given by:

$$P(\mathcal{D}|\mathcal{G},\vec{\mathcal{V}},\mathcal{K}) = \prod_{k=1}^{\mathcal{K}}\prod_{i=1}^{N} \frac{P(\mathcal{D}(i)^{(\vec{\mathcal{V}},k),\{X_{N+1},\pi_i\}}|\mathcal{G}_F(\{X_{N+1},\pi_i\})}{P(\mathcal{D}(i)^{(\vec{\mathcal{V}},k),\{\pi_i\}}|\mathcal{G}_F(\pi_i))} \qquad (42)$$

where $\mathcal{D}(i)^{(\vec{\mathcal{V}},k),S}$ is the data subset of $\mathcal{D}(i)$ which is restricted to those rows that correspond to variables in S and to those columns that have been assigned to component k by the allocation vector $\vec{\mathcal{V}}$.

5 PREDICTIVE PROBABILITIES

In this section we describe how to compute predictive probabilities for Bayesian networks. The first subsection deals with BGe, and the second subsection fosuses on the novel BGM model. Finally, in the third subsection we give a brief summary of error propagation.

5.1 Predictive probabilities for BGe

We assume that we have a training data set $\mathcal{D}$ of size N-by-m and an independent test data set $\tilde{\mathcal{D}}$ of size N-by-$\tilde{m}$ for the domain $X_1,\ldots,X_N$. As before, $\mathcal{D}_{i,j}$ and $\tilde{\mathcal{D}}_{i,j}$ correspond to the j-th

observation of the i-th domain node X_i in $\mathcal{D}$ and $\tilde{\mathcal{D}}$ respectively. We merge both data sets row-wise to obtain a new data set $\mathcal{D}^*$ of size N-by-$(m+\tilde{m})$, and we define:

$$S_{\mathcal{D},\mathcal{D}^*} := \sum_{j=1}^{m}(\mathcal{D}_{.,j}-\overline{\mathcal{D}^*})\cdot(\mathcal{D}_{.,j}-\overline{\mathcal{D}^*})^T+\sum_{j=1}^{\tilde{m}}(\tilde{\mathcal{D}}_{.j}-\overline{\mathcal{D}^*})\cdot(\tilde{\mathcal{D}}_{.j}-\overline{\mathcal{D}^*})^T \tag{43}$$

where

$$\overline{\mathcal{D}^*} = \frac{1}{m+\tilde{m}}\left(\sum_{i=1}^{m}\mathcal{D}_{.,j}+\sum_{i=1}^{\tilde{m}}\tilde{\mathcal{D}}_{.j}\right) \tag{44}$$

Let $S \subset \{X_1,\ldots,X_N\}$ denote an n-dimensional subset of the N domain variables. The predictive probability $P := P(\tilde{\mathcal{D}}^{\{S\}}|\mathcal{D}^{\{S\}},\mathcal{G}_F(S))$ for the data subset $\tilde{\mathcal{D}}^{\{S\}} \subset \tilde{\mathcal{D}}$ conditional on the subset $\mathcal{D}^{\{S\}} \subset \mathcal{D}$ and a full graph $\mathcal{G}_F(S)$ for the sub-domain S can then be factorised using Eq. (15) of Geiger and Heckerman (1994):

$$P = \prod_{j=1}^{\tilde{m}} P(\tilde{\mathcal{D}}_{.j}^{\{S\}}|\tilde{\mathcal{D}}_{.1}^{\{S\}},\ldots,\tilde{\mathcal{D}}_{.j-1}^{\{S\}},\mathcal{D}_{.1}^{\{S\}},\ldots,\mathcal{D}_{.m}^{\{S\}},\mathcal{G}_F(S)) \tag{45}$$

where $\mathcal{D}_{.,j}^{\{S\}}$ and $\tilde{\mathcal{D}}_{.,j}^{\{S\}}$ denote the j-th observation of the n-dimensional sub-domain S in $\mathcal{D}$ and $\tilde{\mathcal{D}}$ respectively.
And in analogy to Eq. (15) in Geiger and Heckerman (1994) it can be derived:

$$\begin{aligned} P &= (2\pi)^{-\frac{n\cdot\tilde{m}}{2}}\cdot\left(\frac{v+m}{v+m+\tilde{m}}\right)^{n/2}\cdot\frac{c(n,\alpha+m)}{c(n,\alpha+m+\tilde{m})} \\ &\quad\cdot det(T_{\mathcal{D}}^S)^{\frac{\alpha+m}{2}}\cdot det(T_{\mathcal{D},\tilde{\mathcal{D}}}^S)^{-\frac{\alpha+m+\tilde{m}}{2}} \end{aligned} \tag{46}$$

where α and v are hyperparameters that have to be specified, $T_{\mathcal{D}}^S$ is the submatrix of $T_{\mathcal{D}}$ (see Eq. (33)) imposed by the subset S of the domain variables, that is, the submatrix consisting only of those rows and columns that correspond to variables in S. $c(n,\alpha+m)$ and $c(n,\alpha+m+\tilde{m})$ can be computed from Eq. (36). $T_{\mathcal{D},\tilde{\mathcal{D}}}$ is given by:

$$T_{\mathcal{D},\tilde{\mathcal{D}}} := T_0 + S_{\mathcal{D},\tilde{\mathcal{D}}} + \frac{v\cdot(m+\tilde{m})}{v+m+\tilde{m}}\cdot(\vec{\mu}_0-\overline{\mathcal{D}^*})(\vec{\mu}_0-\overline{\mathcal{D}^*})^T \tag{47}$$

and $T_{\mathcal{D},\tilde{\mathcal{D}}}^S$ is the submatrix of $T_{\mathcal{D},\tilde{\mathcal{D}}}$ imposed by the subset S of the domain variables. That is the submatrix consisting only of those rows and columns that correspond to variables in S.

Predictive probabilities $P(\tilde{\mathcal{D}}|\mathcal{D})$ can be computed for static and dynamic Bayesian networks with the BGe scoring metric. Here we focus on the predictive distribution for dynamic Bayesian networks (DBNs), and we show that they can be estimated from a sample $\{\mathcal{G}_1,\ldots,\mathcal{G}_T\}$ approximately drawn from the posterior distribution $P(\mathcal{G}|\mathcal{D})$ with MCMC.
As before, denote by $\mathcal{G}$ the graph, and let $\vec{q}$ denote the vector of parameters associated with $\mathcal{G}$. We get the following expression for the predictive distribution:

$$P(\tilde{\mathcal{D}}|\mathcal{D}) = \sum_{\mathcal{G}}\int d\vec{q}P(\mathcal{D}|\mathcal{G},\vec{q})\cdot P(\mathcal{G},\vec{q}|\mathcal{D}) \tag{48}$$

A possible approach is to approximately sample graphs $\{\mathcal{G}_i\}$ and $\{\vec{q}_i\}$ from the posterior distribution $P(\mathcal{G},\vec{q}|\mathcal{D})$ with MCMC and to approximate the integral in Eq. (48) by a sum over this sample. A better method is to use the expansion $P(\mathcal{G},\vec{q}|\mathcal{D}) = P(\vec{q}|\mathcal{G},\mathcal{D})\cdot P(\mathcal{G}|\mathcal{D})$ and draw on the fact that

$$\Psi(\mathcal{G},\tilde{\mathcal{D}}) = \int d\vec{q}P(\tilde{\mathcal{D}}|\mathcal{G},\vec{q})\cdot P(\vec{q}|\mathcal{G},\mathcal{D}) \tag{49}$$

can be calculated analytically. Inserting Eq. (49) in Eq. (48) yields:

$$P(\tilde{\mathcal{D}}|\mathcal{D}) = \sum_{\mathcal{G}}\Psi(\mathcal{G},\tilde{\mathcal{D}})\cdot P(\mathcal{G}|\mathcal{D}) \tag{50}$$

which in practice is computed from a sample $\{\mathcal{G}_1,\ldots,\mathcal{G}_T\}$ approximately drawn from the posterior distribution $P(\mathcal{G}|\mathcal{D})$ with MCMC:

$$P(\tilde{\mathcal{D}}|\mathcal{D}) = \frac{1}{T}\sum_{i=1}^{T}\Psi(\mathcal{G}_i,\tilde{\mathcal{D}}) \tag{51}$$

Consequently, an estimator for the predictive probability is given by:

$$\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D}) = \frac{1}{T}\sum_{i=1}^{T}P_{BGe}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G}_i) \tag{52}$$

and the probabilities $P_{BGe}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G})$ are given by:

$$P_{BGe}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G}) = \prod_{i=1}^{N}\frac{P(\tilde{\mathcal{D}}(i)^{\{X_{N+1},\pi_i\}}|\mathcal{D}(i)^{\{X_{N+1},\pi_i\}},\mathcal{G}_F(\{X_{N+1},\pi_i\}))}{P(\tilde{\mathcal{D}}(i)^{\{\pi_i\}}|\mathcal{D}(i)^{\{\pi_i\}},\mathcal{G}_F(\pi_i))} \tag{53}$$

In Eq. (53) π_i is the parent set of variable X_i in $\mathcal{G}$ and $\mathcal{G}_F(\{X_{N+1},\pi_i\})$ and $\mathcal{G}_F(\pi_i)$ are full graphs for the corresponding subsets.

5.2 Predictive probabilities for BGM

The results of the last subsection can be straightforwardly extended to the dynamic BGM model when $m = \tilde{m}$ and a one-to-one correspondence between the observations in $\mathcal{D}$ and $\tilde{\mathcal{D}}$, e.g. implied by identical time points, is given. We assume that we have a sample $\{[\mathcal{G}_1,\vec{\mathcal{V}}_1,\mathcal{K}_1],\ldots,[\mathcal{G}_T,\vec{\mathcal{V}}_T,\mathcal{K}_T]\}$ approximately drawn from the posterior distribution $P(\mathcal{G},\vec{\mathcal{V}},\mathcal{K}|\mathcal{D})$ with MCMC. The BGM analogon of Eq. (52) is then given by:

$$\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D}) = \frac{1}{T}\sum_{i=1}^{T}P_{BGM}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G}_i,\vec{\mathcal{V}}_i,\mathcal{K}_i) \tag{54}$$

where the probabilities $P_{BGM}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G},\vec{\mathcal{V}},\mathcal{K})$ can be factorised according to Eq.(25):

$$P_{BGM}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G},\vec{\mathcal{V}},\mathcal{K}) = \prod_{k=1}^{\mathcal{K}}\prod_{i=1}^{N}\frac{P(\tilde{\mathcal{D}}(i)^{(\vec{\mathcal{V}},k),\{X_{N+1},\pi_i\}}|\mathcal{D}(i)^{(\vec{\mathcal{V}},k),\{X_{N+1},\pi_i\}},\mathcal{G}_F(\{X_{N+1},\pi_i\})}{P(\tilde{\mathcal{D}}(i)^{(\vec{\mathcal{V}},k),\{\pi_i\}}|\mathcal{D}(i)^{(\vec{\mathcal{V}},k),\{\pi_i\}},\mathcal{G}_F(\pi_i))} \tag{55}$$

where $\tilde{\mathcal{D}}(i)^{(\vec{\mathcal{V}},k),S}$ and $\mathcal{D}(i)^{(\vec{\mathcal{V}},k),S}$ are data subsets of $\tilde{\mathcal{D}}(i)$ and $\mathcal{D}(i)$, respectively, which are restricted to those rows that correspond to variables in S and to those columns that have been assigned to component k by the allocation vector $\vec{\mathcal{V}}$. The probabilities in the numerator and denominator of each factor can be computed using a modified version of Eq. (46). That is, each predictive probability $P := P(\tilde{\mathcal{D}}^{(\vec{\mathcal{V}},k),\{S\}}|\mathcal{D}^{(\vec{\mathcal{V}},k),\{S\}},\mathcal{G}_F(S))$ for the data subset $\tilde{\mathcal{D}}^{(\vec{\mathcal{V}},k),\{S\}} \subset \tilde{\mathcal{D}}$ conditional on the subset $\mathcal{D}^{(\vec{\mathcal{V}},k),\{S\}} \subset \mathcal{D}$ and a full graph $\mathcal{G}_F(S)$ for the n-dimensional subdomain $S \subset \{X_1,\ldots,X_N\}$ can be factorised using Eq. (15) of Geiger and Heckerman (1994):

$$P = (2\pi)^{-\frac{n\cdot\tilde{m}_k}{2}} \cdot \left(\frac{v+m_k}{v+m_k+\tilde{m}_k}\right)^{n/2} \cdot \frac{c(n,\alpha+m_k)}{c(n,\alpha+m_k+\tilde{m}_k)} \cdot det(T_{\mathcal{D}}^{(\vec{\mathcal{V}},k),S})^{\frac{\alpha+m_k}{2}} \cdot det(T_{\mathcal{D},\tilde{\mathcal{D}}}^{(\vec{\mathcal{V}},k),S})^{-\frac{\alpha+m_k+\tilde{m}_k}{2}} \quad (56)$$

where α and v are hyperparameters that have to be specified, m_k and $\tilde{m}_k$ are the numbers of observations that are allocated to the k-th mixture component by $\vec{\mathcal{V}}$, $c(n,\alpha+m_k)$ and $c(n,\alpha+m_k+\tilde{m}_k)$ can be computed from Eq. (36). $T_{\mathcal{D}}^{(\vec{\mathcal{V}},k),S}$ and $T_{\mathcal{D},\tilde{\mathcal{D}}}^{(\vec{\mathcal{V}},k),S}$ can be computed using Eq. (33) and Eq. (47) after having replaced $\mathcal{D}$ and $\tilde{\mathcal{D}}$ by the data subsets $\mathcal{D}^{(\vec{\mathcal{V}},k),S}$ and $\tilde{\mathcal{D}}^{(\vec{\mathcal{V}},k),S}$, m and $\tilde{m}$ by m_k and $\tilde{m}_k$, $\vec{\mu_0}$ by the subvector $\vec{\mu_0}^S$ consisting of those entries only corresponding to variables in S, and T_0 by the submatrix T_0^S consisting of those rows and columns only corresponding to variables in S. We note that the means in Eq. (34) and Eq. (44) and the covariances in Eq. (35) and Eq. (43) are then n-dimensional, that is, restricted to the variables in S. Furthermore, m and $\tilde{m}$ are replaced by m_k and $\tilde{m}_k$, as the means and covariances are computed for the subset of observations that are allocated to the k-th mixture component by $\vec{\mathcal{V}}$.

Finally, we note that Eq. (54) can be derived in analogy to Eq. (52) in the last subsection.
For BGM we get the following expression for the predictive probabilities:

$$P(\tilde{\mathcal{D}}|\mathcal{D}) = \sum_{\mathcal{K},\vec{\mathcal{V}},\mathcal{G}} \int P(\tilde{\mathcal{D}}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q})P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q}|\mathcal{D})d\vec{q} \quad (57)$$

and we can use the expansion

$$P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q}|\mathcal{D}) = P(\vec{q}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\mathcal{D}) \cdot P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G}|\mathcal{D})$$

and draw on the fact that

$$\Psi(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\tilde{\mathcal{D}}) = \int d\vec{q}P(\tilde{\mathcal{D}}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\vec{q}) \cdot P(\vec{q}|\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\mathcal{D}) \quad (58)$$

can be calculated analytically. Inserting Eq. (58) in Eq. (57) yields:

$$P(\tilde{\mathcal{D}}|\mathcal{D}) = \sum_{\mathcal{K},\vec{\mathcal{V}},\mathcal{G}} \Psi(\mathcal{K},\vec{\mathcal{V}},\mathcal{G},\tilde{\mathcal{D}}) \cdot P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G}|\mathcal{D}) \quad (59)$$

which in practice is computed from a sample $\{[\mathcal{K}_1,\vec{\mathcal{V}}_1,\mathcal{G}_1],\ldots,[\mathcal{K}_T,\vec{\mathcal{V}}_T,\mathcal{G}_T]\}$ approximately drawn from the posterior distribution $P(\mathcal{K},\vec{\mathcal{V}},\mathcal{G}|\mathcal{D})$ with MCMC:

$$P(\tilde{\mathcal{D}}|\mathcal{D}) = \frac{1}{T}\sum_{i=1}^{T} \Psi(\mathcal{K}_i,\vec{\mathcal{V}}_i,\mathcal{G}_i,\tilde{\mathcal{D}}) \quad (60)$$

5.3 Error propagation

The standard deviations of the estimators $\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D})$ in Eq. (52) and $\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D})$ in Eq. (54) are given by:

$$\sigma\left\{\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D})\right\} = \left(\frac{1}{T\cdot(T-1)}\sum_{i=1}^{T}\left(P_{BGe}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G}_i) - \widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D})\right)^2\right)^{1/2}$$

$$\sigma\left\{\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D})\right\} = \left(\frac{1}{T\cdot(T-1)}\sum_{i=1}^{T}\left(P_{BGM}(\tilde{\mathcal{D}}|\mathcal{D},\mathcal{G}_i,\vec{\mathcal{V}}_i,\mathcal{K}_i) - \widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D})\right)^2\right)^{1/2}$$

Applying the statistical rules of error propagation ($\sigma(f(x)) = f'(x)\cdot\sigma(x)$) for the $log_e(.)$ transformation we obtain that the standard deviations of the logarithmic predictive probability estimators $log_e(\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D}))$ and $log_e(\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D}))$ are given by:

$$\sigma\left\{log_e(\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D}))\right\} = \frac{\sigma\left\{\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D})\right\}}{\widehat{P_{BGe}}(\tilde{\mathcal{D}}|\mathcal{D})}$$

$$\sigma\left\{log_e(\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D}))\right\} = \frac{\sigma\left\{\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D})\right\}}{\widehat{P_{BGM}}(\tilde{\mathcal{D}}|\mathcal{D})}$$

REFERENCES

Chickering, D. M. (1995) A transformational characterization of equivalent Bayesian network structures. *International Conference on Uncertainty in Artificial Intelligence (UAI)*, **11**, 87–98.

Chickering, D. M. (2002) Learning equivalence classes of Bayesian network structures. *Journal of Machine Learning Research*, **2**, 445–498.

Friedman, N. and Koller, D. (2003) Being Bayesian about network structure. *Machine Learning*, **50**, 95–126.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

Geiger, D. and Heckerman, D. (1994) Learning Gaussian networks. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 235–243.

Grzegorczyk, M. and Husmeier, D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265–305.

Grzegorczyk, M., Husmeier, D. and Werhli, A. (2008) Reverse engineering gene regulatory networks with various machine learning methods. In Emmert-Streib, F. and Dehmer, M. (eds.), *Analysis of Microarray Data. A Network-Based Approach*. Wiley-WCH.

Heckerman, D. (1999) A tutorial on learning with Bayesian networks. In Jordan, M. I. (ed.), *Learning in Graphical Models*, Adaptive Computation and Machine Learning, pp. 301–354. MIT Press, Cambridge, Massachusetts.

Jensen, F. V. (1996) *An Introduction to Bayesian Networks*. UCL Press, London, England.

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

Nobile, A. and Fearnside, A. (2007) Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, **17**, 147–162.

Pearl, J. (2000) *Causality: Models, Reasoning and Intelligent Systems*. Cambridge University Press, London, UK.

Verma, T. and Pearl, J. (1990) Equivalence and synthesis of causal models. *In: Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, **6**, 220–227.

Werhli, A. V., Grzegorczyk, M. and Husmeier, D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**, 2523–2531.