

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 Grzegorzcyk^{1,5,*}, Dirk Husmeier^{2,5,*}, Kieron D. Edwards³, Peter Ghazal^{4,5} and Andrew J. Millar^{1,5}

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

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

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 $P(X_n | \pi_n)$ ($n = 1, \dots, N$). 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 $P(X_1, \dots, X_N)$ (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, \dots, 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} :

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

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, \dots, 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:

$$P(\mathcal{G}) = \frac{1}{\Pi} \prod_{n=1}^N \binom{N-1}{|\pi_n|}^{-1} \quad (3)$$

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:

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

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 $\psi[\cdot]$ 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 a non-homogeneous extension of the BGe-model. See Geiger and Heckerman (1994) or Grzegorzczak *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, \dots, X_m time series data $(X_{1,t}, \dots, X_{N,t})_{t=1, \dots, m}$ 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 acyclicity of the underlying graph \mathcal{G} is not required, and Eq. (1) is replaced by:

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

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:

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

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 $X_{1,1}, \dots, X_{1,m}$ 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 Grzegorzczak and Husmeier (2008). The structure MCMC approach of Madigan and York (1995) generates a sample of DAGs $\mathcal{G}_1, \dots, \mathcal{G}_T$ 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 $Q(\mathcal{G}_{i+1}|\mathcal{G}_i)$:

$$Q(\mathcal{G}_{i+1}|\mathcal{G}_i) = \begin{cases} \frac{1}{|\mathcal{N}(\mathcal{G}_i)|} & , \mathcal{G}_{i+1} \in \mathcal{N}(\mathcal{G}_i) \\ 0 & , \mathcal{G}_{i+1} \notin \mathcal{N}(\mathcal{G}_i) \end{cases} \quad (7)$$

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

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

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 $\mathcal{G} \neq \tilde{\mathcal{G}}$:

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

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:

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

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:

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

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 $\mathcal{G}_1, \dots, \mathcal{G}_T$, 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:

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

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 $P(\widehat{F_{ij}|\mathcal{D}})$ for each edge e_{ij} .

Let $\epsilon(\theta) = \{e_{ij} | P(\widehat{F_{ij}|\mathcal{D}}) > \theta\}$ denote the set of all edges whose posterior probability 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 characteristic* (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 AUROC₁ value. We note that larger AUROC₁ 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 $P(X_1, \dots, X_N)$ can be replaced by a mixture distribution:

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

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

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

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

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

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

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 $\vec{\phi} = (\vec{\phi}_1^T, \dots, \vec{\phi}_{\mathcal{K}}^T)^T$.

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:

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

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

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:

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

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 $n = 1, \dots, N$ it holds:

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

Let $\mathcal{D}(n, \cdot)$ denote the observations of the n -th domain node X_n in the data \mathcal{D} , and $\mathcal{D}(\pi_n, \cdot)$ 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):

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

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

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

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 $k = 1, \dots, \mathcal{K}$ it can be assumed that:

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

and

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

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:

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

what in turn can be interpreted as a mixture of Bayesian network BGe likelihoods (see Eq. (20)) where $\vec{q}_k = (\vec{q}_{k, (1)}^T, \dots, \vec{q}_{k, (n)}^T)^T$ 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 domain) 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:

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

where

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

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:

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

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

We know that in DBNs the variables at time point $t - 1$ are potential parent nodes of the variables at 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 $\vec{\mathcal{G}}$ is proposed, and the new state $[\vec{\mathcal{G}}, \mathcal{K}, \vec{\mathcal{V}}]$ is accepted according to Eq (8) where the likelihood terms $P(\mathcal{D} | \mathcal{G})$ in Eq. (8) have to be replaced by $P(\mathcal{D} | \vec{\mathcal{G}}, \mathcal{K}, \vec{\mathcal{V}})$ 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:

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

where the likelihood terms have been specified in Eq. (25), the proposal probabilities $Q(\cdot|\cdot)$ 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 corresponding 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 ($i = 1, \dots, 5$) with $p_1 + \dots + p_5 = 1$ 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 $[\mathcal{G}, \mathcal{K}, \vec{\mathcal{V}}]$ to $[\tilde{\mathcal{G}}, \mathcal{K}, \vec{\mathcal{V}}]$ is given by: $A = \min\{1, R\}$ where

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

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 $\tilde{k} = 1, \dots, \mathcal{K}$

replace the i -th entry of the allocation vector $\vec{\mathcal{V}}$ by component \tilde{k} to obtain $\vec{\mathcal{V}}(i \leftarrow \tilde{k})$ ($\tilde{k} = 1, \dots, \mathcal{K}$). 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 $\tilde{k} = 1, \dots, \mathcal{K}$:

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

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:

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

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 $1, \dots, n_k$ 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 $[\mathcal{G}, \mathcal{K}, \vec{V}]$ to $[\mathcal{G}, \mathcal{K}, \vec{V}^*]$ is given by:

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

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

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

where n_k and $n_{\tilde{k}}$ are the numbers of observations allocated to the k -th and \tilde{k} -th component by \vec{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{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{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 chosen 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{V}^* . Nobile and Fearnside (2007) show that the acceptance probability for an EA ejection move from $[\mathcal{G}, \mathcal{K}, \vec{V}]$ to $[\mathcal{G}, \mathcal{K}^*, \vec{V}^*]$ is given by $A = \{1, R\}$ where:

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

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

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

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

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{V} , n_w^* and $n_{\tilde{k}}^*$ are the numbers of observations allocated to the w -th and \tilde{k} -th component by \vec{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:

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

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{V} to component k to obtain the new allocation vector \vec{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 $[\mathcal{G}, \mathcal{K}, \vec{V}]$ to $[\mathcal{G}, \mathcal{K}^*, \vec{V}^*]$ is given by $A = \{1, R\}$ where:

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

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

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

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

where n_k^* is the number of observations allocated to the k -th component in \vec{V}^* , n_k and $n_{\tilde{k}}$ are the numbers of observations allocated to the k -th and \tilde{k} -th component by \vec{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:

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

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, \dots, X_N , let $\mathcal{D}_{i,j}$ denote the j -th observation of the i -th domain node X_i , and let $\mathcal{D}_{.,j} = (\mathcal{D}_{1,j}, \dots, \mathcal{D}_{N,j})^T$ denote the j -th observation vector of the domain. The BGe model (Geiger and Heckerman (1994)) assumes that the set of observation vectors $\mathcal{D}_{.,j}$ ($j = 1, \dots, m$) 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:

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

where

$$\bar{\mathcal{D}} := \frac{1}{m} \sum_{j=1}^m \mathcal{D}_{\cdot,j} \quad (34)$$

is the mean of the m observation vectors and

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

Furthermore, we set:

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

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

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

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 $S \subset \{X_1, \dots, X_N\}$ 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)):

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

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 $(X_{1,t}, \dots, X_{N,t})_{t=1, \dots, m}$ 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 :

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

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

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

$i = 1, \dots, N$. That is, we obtain $\mathcal{D}(i)$ by deleting the last column of \mathcal{D} and adding the row $(\mathcal{D}_{i,2}, \dots, \mathcal{D}_{i,m})$, 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:

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

4.2 BGM

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

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

where $\mathcal{D}^{(\vec{V}, k), S}$ 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{V} .

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

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

where $\mathcal{D}(i)^{(\vec{V}, k), S}$ 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{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 focuses 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, \dots, 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:

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

where

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

Let $S \subset \{X_1, \dots, X_N\}$ 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):

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

where $\mathcal{D}_{\cdot,j}^{\{S\}}$ and $\tilde{\mathcal{D}}_{\cdot,j}^{\{S\}}$ 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:

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

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:

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

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 $\{\mathcal{G}_1, \dots, \mathcal{G}_T\}$ 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:

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

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

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

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

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

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

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

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

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

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

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

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 $\{[\mathcal{G}_1, \vec{V}_1, \mathcal{K}_1], \dots, [\mathcal{G}_T, \vec{V}_T, \mathcal{K}_T]\}$ approximately drawn from the posterior distribution $P(\mathcal{G}, \vec{V}, \mathcal{K} | \mathcal{D})$ with MCMC. The BGM analogon of Eq. (52) is then given by:

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

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

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

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 $S \subset \{X_1, \dots, X_N\}$ can be factorised using Eq. (15) of Geiger and Heckerman (1994):

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

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). $T_{\tilde{\mathcal{D}}}^{(\vec{\mathcal{V}}, k), S}$ 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 $\{[\mathcal{K}_1, \vec{\mathcal{V}}_1, \mathcal{G}_1], \dots, [\mathcal{K}_T, \vec{\mathcal{V}}_T, \mathcal{G}_T]\}$ 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 $\log_e(\cdot)$ 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.
- Grzegorzcyk, 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.
- Grzegorzcyk, 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., Grzegorzcyk, 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.