Being Bayesian about learning Gaussian Bayesian networks from incomplete data

Grzegorczyk, Marco

Published in:
International Journal of Approximate Reasoning

DOI:
10.1016/j.ijar.2023.108954

International Journal of Approximate Reasoning 160 (2023) 108954

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Being Bayesian about learning Gaussian Bayesian networks from incomplete data

Marco Grzegorczyk

Bernoulli Institute, Groningen University, Nijenborgh 9, 9747 AG, Groningen, Netherlands

ARTICLE INFO

Article history:
Received 25 April 2023
Accepted 5 June 2023
Available online 19 June 2023

Keywords:
Gaussian Bayesian networks
BGe score
Incomplete data
Markov Chain Monte Carlo (MCMC)
Conditional Gaussians

ABSTRACT

We propose a Bayesian model averaging (BMA) approach for inferring the structure of Gaussian Bayesian networks (BNs) from incomplete data, i.e. from data with missing values. Our method builds on the 'Bayesian metric for Gaussian networks having score equivalence' (BGe score) and we make the assumption that the unobserved data points are 'missing completely at random'. We present a Markov Chain Monte Carlo sampling algorithm that allows for simultaneously sampling directed acyclic graphs (DAGs) as well as the values of the unobserved data points. We empirically cross-compare the network reconstruction accuracy of the new BMA approach with two non-Bayesian approaches for dealing with incomplete BN data, namely the classical structural Expectation Maximisation (EM) approach and the more recently proposed node average likelihood (NAL) method. For the empirical evaluation we use synthetic data from a benchmark Gaussian BN and real wet-lab protein phosphorylation data from the RAF signalling pathway.

1. Introduction

Bayesian networks (BNs) are a flexible and powerful statistical tool, not only for describing but also for learning the dependence relations among interacting variables [1,2]. In BNs the variables are considered to be the nodes of a directed acyclic graph (DAG) whose edges encode the conditional dependency relations among them. Inferring the DAG from data, which is often referred to as 'BN structure learning', is a challenging task. This is, because the number of possible DAGs grows super-exponentially in the number of nodes [3], and the acyclicity constraint does not allow this task to be decomposed and to be solved in parallel. In the literature many different approaches for BN structure learning have been proposed. It can be distinguished between so called 'constraint-based' approaches (see e.g. [4–6]) and 'score-based' approaches. The latter group of score-based approaches features methods that search for the 'best' (highest scoring) DAG (see e.g. [7–10]) as well as Markov Chain Monte Carlo sampling methods that aim at Bayesian model averaging (BMA) by generating DAG posterior samples (see e.g. [11–15]). Also hybrid BN structure learning approaches have been proposed; see, e.g., the recent works by Scutari et al. [16] and Kuipers et al. [17] which combine constraint- and score-based approaches. For an impressively exhaustive recent review of 61 different BN network structure learning algorithms we refer to the work by Kitson et al. [18].

The works cited above all have in common that they assume the data to be complete without any missing values. However, in practical applications the data might be incomplete what renders the above approaches inapplicable. The BN structure learning task gets much more challenging when data are incomplete. For incomplete data not only the network

E-mail address: m.a.grzegorczyk@rug.nl.

https://doi.org/10.1016/j.ijar.2023.108954

structure but also the missing data points have to be inferred from the observed data points. A classical but still widely-applied approach for learning Bayesian networks from incomplete data is the so called structural EM algorithm invented by Friedman [19]. The structural EM searches for the best DAG in terms of a penalized likelihood inside an Expectation Maximisation (EM) algorithm [20]. Another conceptually easier approach is to employ penalized node-average log-likelihoods (NALs) which can be computed from locally complete observations. In the context of discrete Bayesian networks the NAL approach has been invented by Balov [21] and recently been extended to Gaussian Bayesian networks by Bodewes and Scutari [22]. In particular, it has been shown [21,22] that the NAL approach is consistent and competitive to the structural EM. In Section 4 we briefly review the structural EM and the NAL approach. Over the years, of course, also many other BN structure learning methods for incomplete data have been proposed in the literature. An overview can be found in a work by Scutari [23]. Without any claim to completeness, we would like to mention here four more approaches, namely (i) the Bayesian structural EM from Friedman [24] that searches for the DAG that maximizes a Bayesian score rather than a penalized likelihood, (ii) the variational Bayesian EM approach from Beal and Ghahramani [25] which approximates the marginal likelihood in the presense of missing data, (iii) the auxiliary variable approach by Adel and de Campos [26], and (iv) the novel 'anytime algorithm' (k-MAX) from Scanagatta et al. [27]. Although a variety of approaches for BN structure learning from incomplete data has been proposed in the literature, to the best of our knowledge, all the proposed methods have in common that they search for one single 'best' (highest-scoring) DAG. In this paper we follow a different route and propose a Bayesian model averaging (BMA) approach for inferring Gaussian Bayesian networks from incomplete data. Our new Bayesian method builds on the well-known 'Bayesian metric for Gaussian networks having score equivalence' (BGe score) of Geiger and Heckerman [28] and we present a Markov Chain Monte Carlo (MCMC) sampling algorithm that extends the structure MCMC sampler [11,12], so as to allow for simultaneously sampling directed acyclic graphs (DAGs) as well as the values of the unobserved data points from the posterior distribution.

This paper is organized as follows: In Section 2 we review the Bayesian approach for learning the structure of Gaussian BNs using the classical structure Markov Chain Monte Carlo (MCMC) sampling technique. As the standard approach assumes that the data are complete, we show in Section 3 how to extend the Bayesian framework to incomplete data. The newly proposed BMA approach and its MCMC sampling scheme allows Gaussian BNs to be posterior sampled from incomplete data. In Section 4 we briefly review two competing non-Bayesian methods for handling missing data in BNs. In Section 5 we present the results of an empirical study in which we compare the performances of the competing approaches in terms of their network reconstruction accuracies. In Section 6 we conclude with a short discussion.[1]

2. Bayesian network (BN) learning from complete data

In this section we review Gaussian Bayesian networks (BNs) and the classical Bayesian model averaging (BMA) approach of learning BNs from data. We make the assumption that the data are complete; i.e. do not have any missing data points. In Section 3 we extend the Bayesian approach such that it can also be used for inferring BNs from incomplete data.

2.1. Bayesian networks (BNs) and directed acyclic graphs (DAGs)

BNs use directed acyclic graphs (DAGs) to describe the dependencies among random variables $X_1, \ldots, X_n$. Each variable X_i becomes a node of the DAG and the directed edges among the n nodes encode the conditional (in-)dependence relations. X_k is called a parent node of X_i if there is a directed edge $X_k \rightarrow X_i$ from X_k to X_i, and we let $\boldsymbol{\pi}_{\mathcal{G}}(i)$ denote the set of all parent nodes of X_i implied by the DAG $\mathcal{G}$. Hence, we have $X_k \in \boldsymbol{\pi}_{\mathcal{G}}(i)$ if and only if X_k is a parent of X_i in $\mathcal{G}$. Moreover, we call X_k an ancestor node of X_i if there is a directed path (i.e. a sequence of directed edges) leading from X_k to X_i, symbolically $X_k \rightarrow \ldots \rightarrow X_i$. The acyclicity constraint bans directed paths of the form $X_i \rightarrow \ldots \rightarrow X_i$, which are referred to as cycles. In an acyclic graph no node X_i can be its own ancestor.

In BNs any given DAG $\mathcal{G}$ implies conditional (in-)dependence relations such that the joint distribution factorizes into a product of n local conditional distributions of the form:

$$p(X_1, \ldots, X_n|\mathcal{G}) = \prod_{i=1}^{n} p(X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)) \tag{1}$$

We call a DAG complete if it possesses the maximal number of edges; i.e. $n(n-1)/2$ edges. A complete DAG $\mathcal{G}^C$ encodes that the n random variables are pairwise mutually dependent, i.e. that there are no conditional independence relations. In the absence of conditional (in-)dependence relations, the probability chain rule implies that there are $n!$ possible factorizations:

$$p(X_1, \ldots, X_n|\mathcal{G}^C_\tau) = \prod_{i=1}^{n} p(X_{\tau(i)}|X_{\tau(1)}, \ldots, X_{\tau(i-1)}) \tag{2}$$

[1] The title of this work was inspired by the title of a work by Friedman and Koller [13].

where $\tau(.)$ can be any of the $n!$ permutations of the integers $\{1,\ldots,n\}$. That is, there are $n!$ complete DAGs and each can be identified with a permutation $\tau(.)$. The parent node sets are: $\boldsymbol{\pi}_{\mathcal{G}_\tau^C}(\tau(i)) = \{X_{\tau(1)},\ldots,X_{\tau(i-1)}\}$.

With regard to Section 3.2 we notice the following: For each possible local conditional distribution $X_i|\boldsymbol{\pi}(i)$, where $\boldsymbol{\pi}(i) \subset \{X_1,\ldots,X_n\}\backslash\{X_i\}$ is an arbitrary parent set, we can find a permutation τ such that $X_i|\boldsymbol{\pi}(i)$ appears as local conditional distribution of the complete DAG $\mathcal{G}_\tau^C$.[2]

2.1.1. Learning Bayesian networks (BNs) from data

The goal of BN structure learning is to infer DAGs from data. As different DAGs impose different conditional (in-)dependence relations, and so different factorizations in Equation (1), the marginal likelihood $P(\mathcal{D}|\mathcal{G})$ of observed data $\mathcal{D}$ depends on the specific DAG $\mathcal{G}$. We assume that the available data set $\mathcal{D}$ consists of N observations, $\mathcal{D} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, where each observation $\mathbf{x}_j \in \mathbb{R}^n$ is a realisation of the random vector $\mathbf{X} := (X_1,\ldots,X_n)^\mathsf{T}$.

Following the Bayesian paradigm, we get for the DAG posterior distribution:

$$P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})} \tag{3}$$

where $P(\mathcal{G})$ is the prior probability for DAG $\mathcal{G}$ and $P(\mathcal{D}) = \sum_{\mathcal{G}^\star} P(\mathcal{D}|\mathcal{G}^\star)P(\mathcal{G}^\star)$ is a normalization constant with the sum being across all possible DAGs $\mathcal{G}^\star$. Since $P(\mathcal{D})$ does not depend on $\mathcal{G}$, we have:

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

In the absence of genuine prior knowledge about the DAG, we assume all DAGs to be equally likely apriori, symbolically $P(\mathcal{G}^\star) = c$ for all $\mathcal{G}^\star$.

In Section 2.2 we briefly review the Gaussian BGe score from Geiger and Heckerman [28–31], which allows the marginal likelihood $P(\mathcal{D}|\mathcal{G})$ to be computed analytically. Markov Chain Monte Carlo (MCMC) sampling techniques can then be used to generate DAG samples from the posterior in Equation (3). From the sampled DAGs the marginal posterior probabilities of so called 'edge features' can be estimated. For example, an estimator for the marginal posterior probability that there is an edge connection between the nodes X_i and X_j is the proportion of sampled DAGs that have these two nodes connected (either via the edge $X_i \rightarrow X_j$ or via the edge $X_i \leftarrow X_j$). For more details on the MCMC algorithm, edge features and their marginal posterior probabilities we refer to Section 2.3.

2.1.2. DAG equivalence classes and score equivalence

Different DAGs can impose the same conditional (in-)dependence relations among $X_1,\ldots,X_n$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ that impose the same conditional (in-)dependence relations are called equivalent, and we have:

$$p(X_1,\ldots,X_n|\mathcal{G}_1) = \prod_{i=1}^{n} p(X_i|\boldsymbol{\pi}_{\mathcal{G}_1}(i)) = \prod_{i=1}^{n} p(X_i|\boldsymbol{\pi}_{\mathcal{G}_2}(i)) = p(X_1,\ldots,X_n|\mathcal{G}_2)$$

DAGs therefore fall into equivalence classes, such that the DAGs within each equivalence class encode the same conditional (in-)dependence relations among the nodes [32]. For example, there are $n!$ complete DAGs $\mathcal{G}_\tau^C$. They are equivalent to each other, as each complete DAG implies that there are no conditional (in-)dependence relations among the nodes, i.e. that the n variables are pairwise mutually dependent.

The existence of DAG equivalence classes makes it challenging to statistically model BNs. As two equivalent DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ state the same about the conditional (in-)dependence relations, it is required that they yield the same (marginal) likelihood, $P(\mathcal{D}|\mathcal{G}_1) = P(\mathcal{D}|\mathcal{G}_2)$. A modelling approach that fulfills this requirement is said to yield 'score-equivalence'. To the best of our knowledge, there are only two Bayesian approaches that yield score equivalence: the discrete BDe score from Madigan and York [11] and the Gaussian BGe score from Geiger and Heckerman [28]. Our focus is on the Gaussian BGe score, and we briefly review it in Section 2.2.

Chickering [33] shows that two DAGs are equivalent if and only if they have the same skeleton and the same v-structures and that the DAG equivalence classes can be represented by 'completed partially directed acyclic graphs' (CPDAGs).[3] The DAGs within one equivalence class and the corresponding CPDAG share the same skeleton. But unlike the DAGs, the CPDAG possesses a mixture of directed and undirected edges. A directed edge $X_i \rightarrow X_j$ in a CPDAG indicates that all DAGs within the equivalence class agree on this edge direction, while an undirected edge $X_i - X_j$ in a CPDAG indicates that the DAGs within the equivalence class have the corresponding two nodes connected but disagree about the edge direction, i.e. some DAGs have the edge $X_i \rightarrow X_j$ while others have the oppositely oriented edge $X_i \leftarrow X_j$.

[2] Let $j := |\boldsymbol{\pi}(i)|$ denote the cardinality of the parent set $\boldsymbol{\pi}(i)$, then select any permutation τ with: $\{X_{\tau(1)},\ldots,X_{\tau(j)}\} = \boldsymbol{\pi}(i)$, and $X_{\tau(j+1)} = X_i$.

[3] The skeleton of a DAG $\mathcal{G}$ is obtained by replacing all directed edges by undirected edges. A v-structure is the constellation of two edges converging on a node $X_k \rightarrow X_i \leftarrow X_l$ without any edge between the parents X_k and X_l.

2.2. The Gaussian BGe score

The 'Bayesian metric for Gaussian networks having score equivalence' (short: BGe score) from Geiger and Heckerman [28–31] can be used for modelling BNs among Gaussian distributed random variables. We briefly review the BGe score in this subsection. A more detailed exposition of the BGe score can be found in Section **S1** of the supplementary paper.

2.2.1. Gaussian likelihood and parameter mapping

Given any DAG $\mathcal{G}$, it is assumed that each observation $\mathbf{x}_j \in \mathbb{R}^n$ is a realisation of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$, which has a multivariate Gaussian distribution whose parameters are coherent with $\mathcal{G}$:

$$\mathbf{X}|\mathcal{G} \sim \mathcal{N}_n(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}) \tag{4}$$

where $\boldsymbol{\mu}^{\mathcal{G}} \in \mathbb{R}^n$ is the expectation vector, and $\boldsymbol{\Sigma}^{\mathcal{G}}$ is the positive definite covariance matrix. The $\mathcal{G}$ upper scripts on the parameters indicate that the multivariate Gaussian in Equation (4) must imply the factorization given in Equation (1). The conditional distributions on the right hand side of Equation (1) then refer to univariate conditional Gaussians $X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)$:

$$X_i|\boldsymbol{\pi}_{\mathcal{G}}(i) \sim \mathcal{N}\left(\boldsymbol{\mu}^{\mathcal{G}}_{X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)}, \boldsymbol{\Sigma}^{\mathcal{G}}_{X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)}\right) \qquad (i = 1, \ldots, n)$$

where $\boldsymbol{\mu}^{\mathcal{G}}_{X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)} \in \mathbb{R}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}_{X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)} > 0$ are the expectation and the variance of the conditional Gaussian $X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)$. There is a one-to-one mapping between the parameters of the conditional Gaussians and the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$ of the joint Gaussian distribution; see Section **S1** of the supplementary paper. In Section 3.2 we exploit this mapping when sampling covariance matrices $\boldsymbol{\Sigma}^{\mathcal{G}}$ and expectation vectors $\boldsymbol{\mu}^{\mathcal{G}}$ that are coherent with a DAG $\mathcal{G}$.

2.2.2. Parameter prior distributions

Given N independent and complete realizations of the random vector $\mathbf{X}$ and any DAG $\mathcal{G}$, we have for the likelihood:

$$\mathbf{X}_j|\mathcal{G} \sim \mathcal{N}_n(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}) \qquad (j = 1, \ldots, N) \tag{5}$$

where the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$ must be such that they imply the factorization from Equation (1). On the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of any complete DAG $\mathcal{G}^{\mathcal{C}}_{\tau}$ Geiger and Heckerman impose the fully conjugate normal-Wishart prior

$$\begin{aligned} \boldsymbol{\mu}|\boldsymbol{\Sigma} &\sim \mathcal{N}_n(\boldsymbol{\mu}_0, \alpha_\mu^{-1}\boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma}^{-1} &\sim \mathcal{W}_n(\alpha_w, \mathbf{R}) \end{aligned} \tag{6}$$

where $\mathcal{W}_n(\alpha_w, \mathbf{R})$ denotes the n-dimensional Wishart distribution with $\alpha_w > n - 1$ degrees of freedom and positive definite parametric matrix $\mathbf{R}$. Geiger and Heckerman show that a sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from this normal-Wishart also determines the parameters $(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ of *all* possible DAGs $\mathcal{G}$ and that the prior yields a marginal likelihood that is score-equivalent.[4] Moreover, they show that two of the conditions for score equivalence, namely the likelihood modularity and the parameter modularity condition, imply that the conditional Gaussian distributions $X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)$ do not depend on the overall DAG $\mathcal{G}$. That is, if two DAGs imply the same parent set $\boldsymbol{\pi}(i)$ for X_i then their factorizations both feature the same conditional Gaussian distribution $X_i|\boldsymbol{\pi}(i)$ with the same prior (and posterior) parameters.

Although Geiger and Heckerman show that one sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ determines the parameters $(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ of *all* possible DAGs $\mathcal{G}$, they do not provide an explicit algorithm for deriving $(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ from $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This is because concrete parameter instantiations are not required for computing the marginal likelihood. In Section 3.2 we show how to extract $(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ from $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We require the parameters $(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ when dealing with incomplete data sets. Our algorithm exploits that two DAGs that imply the same parent set $\boldsymbol{\pi}(i)$ for X_i must have the same conditional Gaussian distribution $X_i|\boldsymbol{\pi}(i)$ with the same prior (and posterior) parameters. For more details we refer to Section 3.2.

2.2.3. Parameter posterior distributions and the marginal likelihood

The Gaussian likelihood from Equation (5) in combination with the conjugate normal-Wishart prior from Equation (6) yields the following normal-Wishart posterior distribution for the parameters of any complete DAG $\mathcal{G}^{\mathcal{C}}_{\tau}$:

$$\begin{aligned} \boldsymbol{\mu}|(\boldsymbol{\Sigma}, \mathcal{D}) &\sim \mathcal{N}_n\left(\boldsymbol{\mu}^{\diamond},\ (\alpha_\mu + N)^{-1}\boldsymbol{\Sigma}\right) \\ \boldsymbol{\Sigma}^{-1}|\mathcal{D} &\sim \mathcal{W}_n(\alpha_w + N, \mathbf{T}) \end{aligned} \tag{7}$$

where $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is the data, $\boldsymbol{\mu}^{\diamond} := \frac{\alpha_\mu \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}_N}{\alpha_\mu + N}$, $\bar{\mathbf{x}}_N := \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j$, and

[4] For the Gaussian likelihood from Equation (5) Geiger and Heckerman even show that only the normal-Wishart prior from Equation (6) fulfills the assumptions that are required for score-equivalence; i.e. no other prior can yield a score-equivalent marginal likelihood.

$$\mathbf{T} := \mathbf{R} + \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}}_N)(\mathbf{x}_j - \bar{\mathbf{x}}_N)^\mathsf{T} + \frac{\alpha_\mu N}{\alpha_\mu + N} (\boldsymbol{\mu}_0 - \bar{\mathbf{x}}_N)(\boldsymbol{\mu}_0 - \bar{\mathbf{x}}_N)^\mathsf{T}$$

Each posterior sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of the parameters of a complete DAG determines posterior parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$ of any DAG $\mathcal{G}$. The marginal likelihood $P(\mathcal{D}|\mathcal{G})$ of any DAG $\mathcal{G}$ can be computed analytically and is called the BGe score of $\mathcal{G}$ [28–31]; cf. Section **S1** of the supplementary paper.

2.3. Bayesian Model Averaging (BMA)

2.3.1. The structure MCMC sampler

For sampling DAGs from the posterior distribution in Equation (3), we use the 'structure MCMC' Metropolis-Hastings (MH) sampling scheme from Madigan and York [11] and we implement it using the efficient algorithms from Giudici and Castelo [12]. Let $N(\mathcal{G})$ denote the 'neighbourhood' of $\mathcal{G}$, i.e. the set of all DAGs that can be reached from $\mathcal{G}$ by adding, deleting or reversing one single edge. Given the current DAG $\mathcal{G}$, we propose to move to a randomly selected neighbour DAG $\mathcal{G}^\star \in N(\mathcal{G})$. The acceptance probability of the move is:

$$A(\mathcal{G}, \mathcal{G}^\star) = \min \left\{ 1, \frac{p(\mathcal{D}|\mathcal{G}^\star)}{p(\mathcal{D}|\mathcal{G})} \cdot \frac{p(\mathcal{G}^\star)}{p(\mathcal{G})} \cdot \mathrm{HR}(\mathcal{G}, \mathcal{G}^\star) \right\} \tag{8}$$

where $p(\mathcal{D}|\mathcal{G}^\star)$ and $p(\mathcal{D}|\mathcal{G}^\star)$ are marginal likelihoods that can be computed analytically, $p(\mathcal{G}^\star)$ and $p(\mathcal{G}^\star)$ are the DAG prior probabilities, and the Hastings ratio is $\mathrm{HR}(\mathcal{G}, \mathcal{G}^\star) = \frac{|N(\mathcal{G})|}{|N(\mathcal{G}^\star)|}$ with $|.|$ denoting the cardinality of the DAG neighbourhood sets $N(\mathcal{G})$ and $N(\mathcal{G}^\star)$, respectively.

If the move is accepted, we exchange $\mathcal{G}$ by $\mathcal{G}^\star$, otherwise we keep $\mathcal{G}$ unchanged.

2.3.2. Marginal posterior probabilities of edge features

Given a DAG posterior sample $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(R)}$, we estimate the marginal posterior probabilities of edge features. As DAGs fall into equivalence classes, we first replace each DAG $\mathcal{G}^{(r)}$ by its CPDAG $\mathcal{G}_\diamond^{(r)}$, where the latter can also feature undirected edges $X_i - X_j$. Since we later average across the sampled CPDAGs, we interpret $X_i - X_j$ as bidirectional edge $X_i \leftrightarrow X_j$. We then estimate the marginal posterior probability of any directed edge $X_i \rightarrow X_j$ by the proportion of sampled DAGs whose CPDAGs possess this edge:

$$\hat{p}_{i,j} := \frac{1}{R} \sum_{r=1}^{R} \mathcal{I}_{X_i \rightarrow X_j}(\mathcal{G}_\diamond^{(r)}) \tag{9}$$

where the indicator function $\mathcal{I}_{X_i \rightarrow X_j}(.)$ is equal to 1 if the $\mathcal{G}_\diamond^{(r)}$ has either the directed edge $X_i \rightarrow X_j$ or the undirected edge $X_i - X_j$, as we decided to interpret $X_i - X_j$ as bidirectional $X_i \leftrightarrow X_j$, and $\mathcal{I}_{X_i \rightarrow X_j}(\mathcal{G}_\diamond^{(r)}) = 0$ otherwise.

2.3.3. Network reconstruction accuracy

We assess and cross-compare the network reconstruction accuracy using two complementary approaches, namely AUROC scores and relative structural Hamming distance (rSHD) scores. Unlike for AUROC scores, for rSHD scores a theshold has to be imposed on the marginal edges posterior probabilities, so as to obtain a concrete network prediction. Here we employ the theshold $\psi = 0.5$. For detailed descriptions of our AUROC and rSHD scores we refer to Section **S1** of the supplementary paper.

3. Network learning from incomplete data

We now extend the Bayesian approach from Section 2 such that Gaussian BNs can also be inferred from incomplete data. We assume that the individual data points are 'missing completely at random' (MCAR) and we introduce a new Markov Chain Monte Carlo (MCMC) algorithm that allows for sampling DAGs and the missing data points altogether from the posterior.

3.1. Sampling DAGs in the presence of missing data points

The Bayesian approach for learning BNs from Section 2 assumes that the data $\mathcal{D}$ are complete without missing values. To the best of our knowledge, no Bayesian modelling averaging (BMA) approach for learning BNs from incomplete data has been proposed yet. We here fill the gap by proposing a new MCMC sampling scheme for inferring BNs from incomplete data.

We recall that we have assumed that the data set is of the form $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})^\mathsf{T} \in \mathbb{R}^n$ is the j-th observation of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$. We here assume that data points are 'missing completely at random' (MCAR) and we refer to Rubin [34] for the formal definitions of three different patterns of missingness. The

assumed MCAR mechanism implies that each individual $x_{i,j}$ has the same probability $p_{miss} \in [0,1]$ for being absent. The data $\mathcal{D}$ can then be thought of as consisting of two parts: the observed data $\mathcal{D}_{obs}$ and the missing data $\mathcal{D}_{miss}$ and the posterior distribution is of the form $P(\mathcal{G}, \mathcal{D}_{miss}|\mathcal{D}_{obs})$. For generating posterior samples we propose the following MCMC sampling scheme, which consists of three consecutive steps:

1. Given any values for the missing data points from $\mathcal{D}_{miss}$, we can think of $\mathcal{D} := \{\mathcal{D}_{miss}, \mathcal{D}_{obs}\}$ as a complete data set. Hence, we can perform the structure MCMC Metropolis Hastings sampling steps from Section 2.3.1 to sample DAGs $\mathcal{G}$ from the posterior distribution $P(\mathcal{G}|\mathcal{D})$.
2. Conditional on complete data $\mathcal{D} := \{\mathcal{D}_{miss}, \mathcal{D}_{obs}\}$ and the DAG $\mathcal{G}$, we can posterior sample the model parameters from $P(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}|\mathcal{D}, \mathcal{G})$. In Section 3.2 we propose a new algorithm for this sampling step.
3. Given the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$ and the observed data $\mathcal{D}_{obs}$, we can posterior sample the missing data from $P(\mathcal{D}_{miss}|\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}, \mathcal{D}_{obs})$. In Section 3.3 we show that this refers to sampling from conditional Gaussian distributions.

The proposed MCMC algorithm can be classified as a 'Metropolis-Hastings within Gibbs MCMC sampling scheme'. Metropolis Hastings (MH) moves are employed to sample graphs from $P(\mathcal{G}|\mathcal{D})$, where $\mathcal{D} := \{\mathcal{D}_{miss}, \mathcal{D}_{obs}\}$. The MH moves employ the marginal likelihood (BGe score), $P(\mathcal{D}|\mathcal{G})$, i.e. they are marginalized over all possible model parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$. Then, the network parameters are sampled from $P(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}|\mathcal{D}, \mathcal{G})$ via a partially collapsed Gibbs sampling step; cf. Section 3.2. Finally, the missing data are sampled from the full conditional distribution $P(\mathcal{D}_{miss}|\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}, \mathcal{D}_{obs})$; cf. Section 3.3.

3.2. Sampling posterior parameters given a DAG and complete data

Let $\mathcal{D} := \{\mathcal{D}_{obs}, \mathcal{D}_{miss}\}$ be complete data, where the missing part $\mathcal{D}_{miss}$ has been filled with sampled values (cf. Section 3.3). For the complete DAGs posterior samples of the parameters can be generated by sampling from Equation (7). The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are then coherent with complete DAGs, but they are not coherent with the DAG $\mathcal{G}$. However, we can extract the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$ that are coherent with $\mathcal{G}$. We proceed as follows:

(S1) Sample $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ from the posterior distribution in Equation (7). These parameters are coherent with the complete DAGs. They do not imply the conditional (in-)dependence relations implied by the DAG $\mathcal{G}$.

(S2) Recall that $\boldsymbol{\pi}_{\mathcal{G}}(i)$ denotes the parent set of variable X_i implied by $\mathcal{G}$. Given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from step (S1), we compute the n univariate conditional Gaussian distributions (cf. Section 2.2.1):

$$X_i|\boldsymbol{\pi}_{\mathcal{G}}(i) \sim \mathcal{N}\left(\boldsymbol{\mu}_{X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)}, \boldsymbol{\Sigma}_{X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)}\right) \qquad (i = 1, \ldots, n)$$

(S3) Given the parameters of the n univariate conditional Gaussians $X_i|\boldsymbol{\pi}_{\mathcal{G}}(i)$, we use the recursive formula of Shachter and Kenley [35] to compute the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$ which are coherent with $\mathcal{G}$, i.e. which imply the factorization in Equation (1). We have the relationship $\boldsymbol{\mu}^{\mathcal{G}} = \boldsymbol{\mu}$, but it holds $\boldsymbol{\Sigma}^{\mathcal{G}} \neq \boldsymbol{\Sigma}$, unless $\mathcal{G}$ is a complete DAG.

In Section **S1** of the supplementary paper we provide more detailed explanations of why the algorithm generates posterior samples from $P(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}|\mathcal{D}, \mathcal{G})$.

3.3. Sampling the missing data points

Given the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$, the random vector $\mathbf{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ has the multivariate Gaussian distribution $\mathbf{X}|\mathcal{G} \sim \mathcal{N}_n(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$. In case of missing data, $\mathbf{X}$ consists of two parts: the observed subvector $\mathbf{X}_{obs}$ and the complementary unobserved subvector $\mathbf{X}_{miss}$. Given $\mathbf{X}_{obs} = \mathbf{x}_{obs}$, the missing values $\mathbf{X}_{miss}$ have the following conditional Gaussian distribution:

$$\mathbf{X}_{miss}|(\mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}) \sim \mathcal{N}\left(\boldsymbol{\mu}^{\mathcal{G}}_{miss|obs}, \boldsymbol{\Sigma}^{\mathcal{G}}_{miss|obs}\right) \tag{10}$$

with

$$\boldsymbol{\mu}^{\mathcal{G}}_{miss|obs} := \boldsymbol{\mu}^{\mathcal{G}}_{miss} + \boldsymbol{\Sigma}^{\mathcal{G}}_{miss,obs}\left\{\boldsymbol{\Sigma}^{\mathcal{G}}_{obs,obs}\right\}^{-1}\left(\mathbf{x}_{obs} - \boldsymbol{\mu}^{\mathcal{G}}_{obs}\right)$$

$$\boldsymbol{\Sigma}^{\mathcal{G}}_{miss|obs} := \boldsymbol{\Sigma}^{\mathcal{G}}_{miss,miss} - \boldsymbol{\Sigma}^{\mathcal{G}}_{miss,obs}\left\{\boldsymbol{\Sigma}^{\mathcal{G}}_{obs,obs}\right\}^{-1}\boldsymbol{\Sigma}^{\mathcal{G}}_{obs,miss}$$

where the subscripts 'obs' and 'miss' refer to the subvectors and submatrices that only contain the rows and columns that belong to observed or missing data points.

More generally, the data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are a random sample of $\mathbf{X}$. And each individual data vector $\mathbf{x}_j$ consists of an observed subvector $\mathbf{x}_{j,obs}$ and the complementary unobserved subvector $\mathbf{x}_{j,miss}$.[5] Conditional on the parameters $\boldsymbol{\mu}^{\mathcal{G}}$ and $\boldsymbol{\Sigma}^{\mathcal{G}}$,

[5] In each $\mathbf{x}_j \in \mathbb{R}^n$ we assumed the i-th element $x_{i,j}$ to be missing with probability p_{miss}.

Table 1
Pseudo code. MCMC algorithm to generate a posterior sample of DAGs when the data are incomplete. To allow for a burn-in period we remove the first samples (e.g. the first 50%) and we thin out the remaining samples (e.g. by the factor $\xi = 2000$ ('ECOLI70') or $\xi = 1000$ ('RAF')) to reduce the auto-correlation.

- **Fix the tuning parameter** $q \in [0, 1]$.
  q is the probability for re-sampling the missing data points.
- **Initialization**
  - Initialize the DAG $\mathcal{G}^{[0]}$. E.g. start with a DAG without edges.
  - Initialize the missing data $\mathcal{D}_{miss}^{[0]}$. E.g. for each X_i compute the empirical mean $\bar{x}_i$ of the observed X_i values and set every missing value of X_i to $\bar{x}_i$. Then merge $\mathcal{D}^{[0]} := \{\mathcal{D}_{miss}^{[0]}, \mathcal{D}_{obs}\}$
- **MCMC iterations**: For $t = 1, \ldots, T$:
  - Apply a structure MCMC move on the DAG, i.e. generate

    $$\mathcal{G}^{[t]} \sim P(\mathcal{G}|\mathcal{D}^{[t-1]})$$

    See Section 2.3.1 for the details of this Metropolis-Hastings move.
  - Draw a random number $u \in [0, 1]$
    - If $u < q$ generate a parameter sample

      $$\boldsymbol{\mu}_{[\star]}^{\mathcal{G}}, \boldsymbol{\Sigma}_{[\star]}^{\mathcal{G}} \sim P(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}|\mathcal{D}^{[t-1]}, \mathcal{G}^{[t]})$$

      See Section 3.2 for the details of this MCMC move.
      Use the parameters for re-sampling the missing data

      $$\mathcal{D}_{miss}^{[t]} \sim P(\mathcal{D}_{miss}|\boldsymbol{\mu}_{[\star]}^{\mathcal{G}}, \boldsymbol{\Sigma}_{[\star]}^{\mathcal{G}}, \mathcal{D}_{obs})$$

      See Section 3.3 for the details of this MCMC move.
      Then update the data and set $\mathcal{D}^{[t]} := \{\mathcal{D}_{miss}^{[t]}, \mathcal{D}_{obs}\}$
    - If $u > q$, leave the data unchanged; i.e. set $\mathcal{D}^{[t]} = \mathcal{D}^{[t-1]}$.
- **Output** $\mathcal{G}^{[1]}, \ldots, \mathcal{G}^{[T]}$.

the N data vectors are stochastically independent, so that the missing values $\mathbf{x}_{j,miss}$ can be sampled separately for each j. That is, we sample from $P(\mathcal{D}_{miss}|\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}, \mathcal{D}_{obs})$ by looping through the observations $j = 1, \ldots, N$ and sampling the missing values of the j-th observation $\mathbf{x}_{j,miss}$ from the conditional Gaussian $\mathbf{X}_{j,miss}|(\mathbf{X}_{j,obs} = \mathbf{x}_{j,obs}, \boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$. The latter conditional Gaussian was defined in Equation (10). If a data vector $\mathbf{x}_j$ features either only missing or only observed values, we skip it in the loop.

3.4. The MCMC sampling algorithm

To generate a posterior sample of DAGs from incomplete data we use the MCMC algorithm outlined in the pseudo code in Table 1. For the BGe score we use a weak uninformative parameter prior by setting the hyperparameters in Equation (6) to: $\boldsymbol{\mu}_0 = \mathbf{0}$, $\mathbf{R} = \mathbf{I}$, $\alpha_\mu = 1$, and $\alpha_w = n$.

Since the structure MCMC sampler is based on single edge operations, the trajectory of DAGs is strongly auto-correlated. Before re-sampling the missing data points $\mathcal{D}_{miss}$, we would like the DAG to have changed 'sufficiently'. Therefore, we implement the MCMC scheme such that it performs a structure MCMC move in every iteration, while the re-sampling of $\mathcal{D}_{miss}$ is only performed with probability $q \in [0, 1]$. To allow for a burn-in period we withdraw the first 50% of the samples and we thin-out the remaining samples by the factor ξ by keeping only every ξ-th sample.

In a pre-study we performed convergence diagnoses on a few data sets. For different numbers of iterations T and different probabilities q we checked for convergence by running independent MCMC simulations on the same data set. Some example convergence diagnoses can be found in Section **S2** of the supplementary paper. For the ECOLI70 data (see Section 5.1) we eventually ran MCMC simulations with $T = 4,000,000$ MCMC iterations and with $q = 0.01$. For the RAF data (see Section 5.2) we eventually ran MCMC simulations with $T = 2,000,000$ MCMC iterations and with $q = 0.05$. For both data types we withdrew the first 50% of the samples and thinned out the remaining samples such that we were left with $R = 1000$ posterior samples.

4. Competing methods

We compare the proposed Bayesian model averaging (BMA) approach with two non-Bayesian approaches that were recently cross-compared in a study by Bodewes and Scutari [22]. The classical approach for learning BNs from incomplete data has been introduced by Friedman [19] and is often referred to as 'structural EM algorithm'. It makes use of the Expectation Maximization (EM) algorithm [20] which iterates between an expectation (E) and a maximization (M) step till convergence is reached. Loosely speaking, in the context of BN structure learning from incomplete data the structural EM algorithm proceeds as follows: Given the observed data and the current 'best' DAG, the E step employs belief propagation techniques [36,37] to approximate the expected sufficient statistics. These expected sufficient statistics can be thought of as

if they would have computed from a complete (or completed) data set. Given the expected sufficient statistics from the E step, the M step makes use of network learning techniques (such as Tabu hill-climbing [38]) to find the 'best' DAG for the current expected sufficient statistic values. Another computationally more efficient EM approach is to avoid the computation of the expected sufficient statistics and instead to complete the data set by imputing the expected values of the missing values. Thereby the imputed expectations are conditional on the observed values in the same observation. In the literature these two approaches are often referred to as the 'soft' (work with the expected sufficient statistics) and the 'hard' (compute and assign expected values) structural EM algorithm. For more mathematical details we refer to Chapter 19 of the textbook by Koller and Friedman [39]. An empirical comparison and guidelines for when to use the 'soft' or 'hard' EM can be found in the work by Ruggieri et al. [40].

Another computational efficient non-Bayesian way to deal with missing data is to employ so called node-average likelihoods (NALs). For discrete BNs Balov [21] introduced NALs and showed that it is a consistent method for dealing with missing data. In a more recent work, Bodewes and Scutari [22] have proven that NALs are also consistent when applied in Gaussian and conditional Gaussian BNs. Basically the key idea of NALs is as follows: Every graph $\mathcal{G}$ implies the node-specific parent sets $\pi_{\mathcal{G}}(i)$ and the joint distribution factorizes into a product of local conditional Gaussian distributions of $X_i|\pi_{\mathcal{G}}(i)$ (cf. Equation (1)). The Maximum Likelihood (ML) estimators $\hat{\theta}_{i,\pi_{\mathcal{G}}(i)}$ of the parameters of the conditional Gaussians can only be estimated based on that data subset $\mathcal{D}_{i,\pi_{\mathcal{G}}(i)} \subset \mathcal{D}$ for which the value of X_i and the parents in $\pi_{\mathcal{G}}(i)$ are available. To take into account that the sample sizes vary with the node and its parent set, the log-likelihood is replaced by the node average log-likelihood (NAL):

$$\bar{l}(\mathcal{G}|\mathcal{D}) := \sum_{i=1}^{n} \frac{1}{|\mathcal{D}_{i,\pi_{\mathcal{G}}(i)}|} \sum_{x_i \in \mathcal{D}_{i,\pi_{\mathcal{G}}(i)}} \log\{p(x_i|\pi_{\mathcal{G}}(i), \hat{\theta}_{i,\pi_{\mathcal{G}}(i)})\}$$

where $|\mathcal{D}_{i,\pi_{\mathcal{G}}(i)}| \leq N$ is the cardinality of the data subset $\mathcal{D}_{i,\pi_{\mathcal{G}}(i)}$.

The structural EM algorithm [19,24] as well as the NAL approach [22] both score DAGs in terms of penalized log-likelihoods. For example, in terms of the node average log-likelihood the penalized likelihood of a DAG $\mathcal{G}$ is:

$$S_P(\mathcal{G}|\mathcal{D}) = \bar{l}(\mathcal{G}|\mathcal{D}) - \lambda_{N,n} \cdot |\theta_{\mathcal{G}}|$$

where $\lambda_{N,n}$ is the penalty per parameter, and $|\theta_{\mathcal{G}}| = \sum_{i=1}^{n} |\hat{\theta}_{i,\pi_{\mathcal{G}}(i)}|$ is the number of parameters implied by $\mathcal{G}$. We follow [22] and implement the two competing methods (EM and NAL) with four different penalty strengths $\lambda_{N,n} = \frac{\log(N)}{2n}$, which refers to the BIC criterion [41] and $\lambda_{N,n} = \frac{N^{-\alpha}}{n}$ with $\alpha \in \{0.1, 0.25, 0.4\}$.

Both approaches (EM and NAL) have been implemented by Bodewes and Scutari [22]. In their R software a hill-climbing algorithm with tabu list [38] is used for finding the DAG that maximizes the penalized log-likelihood. For our evaluation study in Section 5 we make use of this software and we also re-use the tuning parameters from the earlier study by Bodewes and Scutari.

5. Empirical results

Our empirical evaluation and method comparison consists of two parts. In Section 5.1 we reconstruct a benchmark Gaussian BN from synthetic data, and in Section 5.2 we reconstruct the RAF protein signalling pathway from real protein phosphorylation data [42].

We generated data sets with different sample sizes N and we distinguished different average fractions p_{miss} of missing completely at random (MCAR) data by deleting each individual observation $x_{i,j}$ (= j-th observation of X_i) with probability p_{miss}. For the structural EM algorithm and the NAL approach we employed the **R** software implementations from [22] and we used both with four different penalty parameters $\lambda_{n,N}$; see Section 4 for more details. For the new BMA approach we generated posterior samples using the MCMC algorithm from Section 3.4 and we used the posterior sampled DAGs to compute the marginal edge posterior probabilities, cf. Equation (9). Our **R** implementation of the BMA approach is available from GitHub.[6]

5.1. Data from synthetic Gaussian network ('ECOLI70')

The ECOLI70 network is implemented in the 'bnlearn' R package [43–45] and it had already been used by Bodewes and Scutari [22] to cross-compare the performances of the structural EM and the NAL approach for Gaussian networks.[7] The ECOLI70 network consists of $n = 46$ nodes and like its name suggests it features 70 directed edges among them. The advantage of this first study was that the true DAG is known, so that the network reconstruction accuracy could

[6] https://github.com/MarcoAndreas/BMA.

[7] It has been adapted from the 'GeneNet' R package by Schäfer and Strimmer [46].

Table 2
Average AUROC scores for the ECOLI70 network. For each sample size N we generated 10 independent data sets. From each data set we then randomly removed data points, so as to achieve different average fractions of missing values p_{miss}. From the incomplete data sets we inferred the networks (CPDAGs) with the structural **EM**, the **NAL** approach and the new Bayesian model averaging (**BMA**) approach from Section 3. For EM and NAL we distinguish four different penalty parameters $\lambda_{n,N}$ (cf. Section 4). The AUROC results for larger samples sizes N can be found in Table 3.

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
10	0	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	**0.71**
	0.05	0.65	0.64	0.64	0.64	0.65	0.65	0.65	0.65	**0.71**
	0.1	0.63	0.63	0.62	0.62	0.62	0.62	0.63	0.63	**0.70**
	0.2	0.60	0.59	0.59	0.59	0.60	0.60	0.60	0.60	**0.68**
	0.4	0.48	0.48	0.48	0.48	0.46	0.46	0.46	0.46	**0.54**
25	0	0.66	0.64	0.64	0.64	0.66	0.65	0.64	0.65	**0.74**
	0.05	0.64	0.62	0.63	0.63	0.64	0.64	0.63	0.63	**0.74**
	0.1	0.63	0.60	0.60	0.61	0.61	0.61	0.61	0.61	**0.74**
	0.2	0.60	0.58	0.58	0.59	0.60	0.60	0.60	0.60	**0.70**
	0.4	0.55	0.55	0.55	0.55	0.60	0.60	0.60	0.60	**0.69**
50	0	0.68	0.64	0.65	0.65	0.68	0.64	0.65	0.65	**0.82**
	0.05	0.66	0.63	0.62	0.63	0.65	0.66	0.66	0.66	**0.82**
	0.1	0.66	0.62	0.62	0.63	0.62	0.62	0.62	0.62	**0.79**
	0.2	0.62	0.60	0.60	0.60	0.63	0.63	0.63	0.63	**0.76**
	0.4	0.59	0.57	0.56	0.57	0.60	0.60	0.60	0.60	**0.71**
100	0	0.74	0.65	0.68	0.72	0.74	0.65	0.67	0.71	**0.90**
	0.05	0.72	0.63	0.66	0.69	0.67	0.66	0.66	0.66	**0.89**
	0.1	0.71	0.60	0.63	0.67	0.64	0.65	0.65	0.64	**0.87**
	0.2	0.67	0.58	0.60	0.63	0.64	0.64	0.64	0.64	**0.82**
	0.4	0.62	0.56	0.55	0.59	0.63	0.63	0.63	0.63	**0.73**

Table 3
Table 2 continued. Average AUROC scores for the ECOLI70 network for larger sample sizes. See caption of Table 2 for more information.

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
0.25k	0	0.78	0.71	0.75	0.79	0.79	0.69	0.74	0.79	**0.93**
	0.05	0.76	0.69	0.74	0.79	0.65	0.63	0.63	0.66	**0.93**
	0.1	0.75	0.66	0.72	0.76	0.65	0.65	0.65	0.65	**0.92**
	0.2	0.73	0.62	0.68	0.73	0.65	0.65	0.65	0.65	**0.90**
	0.4	0.67	0.56	0.63	0.68	0.64	0.64	0.64	0.64	**0.84**
0.5k	0	0.82	0.74	0.80	0.88	0.82	0.74	0.81	0.87	**0.96**
	0.05	0.80	0.73	0.79	0.84	0.64	0.62	0.62	0.67	**0.96**
	0.1	0.77	0.71	0.76	0.83	0.64	0.64	0.64	0.65	**0.95**
	0.2	0.75	0.67	0.74	0.78	0.66	0.67	0.67	0.67	**0.94**
	0.4	0.71	0.61	0.67	0.74	0.65	0.65	0.66	0.66	**0.89**
1k	0	0.83	0.78	0.85	0.87	0.82	0.78	0.85	0.87	**0.97**
	0.05	0.83	0.77	0.84	0.86	0.61	0.59	0.59	0.72	**0.97**
	0.1	0.81	0.76	0.82	0.84	0.63	0.63	0.63	0.66	**0.97**
	0.2	0.77	0.71	0.79	0.84	0.66	0.66	0.66	0.66	**0.97**
	0.4	0.72	0.64	0.73	0.78	0.66	0.66	0.66	0.66	**0.95**
2.5k	0	0.86	0.82	0.90	0.88	0.87	0.82	0.90	0.87	**0.98**
	0.05	0.85	0.81	0.87	0.88	0.61	0.62	0.64	0.82	**0.98**
	0.1	0.85	0.79	0.88	0.86	0.63	0.63	0.64	0.65	**0.98**
	0.2	0.80	0.77	0.84	0.83	0.67	0.67	0.67	0.68	**0.98**
	0.4	0.76	0.74	0.80	0.81	0.67	0.67	0.67	0.67	**0.94**
5k	0	0.90	0.91	0.90	0.90	0.89	0.88	0.88	0.89	**0.98**
	0.05	0.90	0.84	0.89	0.88	0.64	0.65	0.71	0.87	**0.98**
	0.1	0.89	0.85	0.89	0.88	0.61	0.61	0.61	0.72	**0.99**
	0.2	0.83	0.84	0.88	0.88	0.64	0.64	0.64	0.67	**0.98**
	0.4	0.78	0.75	0.82	0.82	0.69	0.69	0.69	0.69	**0.96**
10k	0	0.91	0.91	0.90	0.89	0.89	0.90	0.89	0.89	**1.00**
	0.05	0.93	0.90	0.90	0.88	0.67	0.65	0.74	0.86	**0.97**
	0.1	0.87	0.91	0.93	0.85	0.60	0.60	0.60	0.85	**0.97**
	0.2	0.85	0.85	0.88	0.88	0.64	0.64	0.65	0.73	**0.99**
	0.4	0.77	0.77	0.83	0.82	0.65	0.65	0.65	0.66	**0.94**

Table 4
Average AUROC scores for the RAF pathway network. For each sample sizes $N \in \{100, 250, 500, 1000, 3530\}$ we randomly sub-sampled 10 independent data sets. From each data set we then randomly removed data points, so as to achieve five different average fractions of missing values p_{miss}. From the incomplete data sets we inferred the networks (CPDAGs) with the structural **EM**, the **NAL** approach and the new Bayesian model averaging (**BMA**) approach from Section 3. For EM and NAL we distinguished four different penalty parameters $\lambda_{n,N}$ (cf. Section 4).

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
100	0	0.60	0.60	0.60	0.60	0.60	0.60	0.61	0.61	**0.65**
	0.05	0.59	0.57	0.60	0.60	0.54	0.54	0.55	0.57	**0.64**
	0.1	0.60	0.60	0.61	0.61	0.55	0.55	0.55	0.57	**0.65**
	0.2	0.58	0.56	0.60	0.59	0.54	0.52	0.53	0.54	**0.64**
	0.4	0.57	0.57	0.58	0.58	0.54	0.56	0.53	0.55	**0.62**
250	0	0.63	0.62	0.63	0.60	0.62	0.61	0.62	0.60	**0.69**
	0.05	0.63	0.62	0.63	0.60	0.56	0.57	0.60	0.56	**0.68**
	0.1	0.62	0.62	0.62	0.60	0.55	0.54	0.54	0.55	**0.68**
	0.2	0.62	0.59	0.62	0.59	0.54	0.54	0.54	0.53	**0.64**
	0.4	0.61	0.61	0.62	0.58	0.51	0.51	0.54	0.51	**0.65**
500	0	0.64	0.64	0.65	0.60	0.62	0.63	0.65	0.60	**0.70**
	0.05	0.63	0.64	0.64	0.60	0.58	0.58	0.58	0.57	**0.69**
	0.1	0.64	0.64	0.63	0.61	0.52	0.52	0.54	0.56	**0.70**
	0.2	0.64	0.64	0.65	0.61	0.51	0.52	0.52	0.53	**0.71**
	0.4	0.61	0.63	0.63	0.59	0.52	0.52	0.55	0.53	**0.66**
1k	0	0.63	0.67	0.66	0.62	0.65	0.66	0.66	0.62	**0.72**
	0.05	0.66	0.66	0.65	0.61	0.55	0.56	0.57	0.59	**0.72**
	0.1	0.66	0.68	0.65	0.61	0.56	0.57	0.58	0.57	**0.73**
	0.2	0.65	0.65	0.64	0.60	0.54	0.55	0.57	0.57	**0.72**
	0.4	0.62	0.65	0.62	0.56	0.53	0.53	0.51	0.53	**0.66**
all	0	0.63	0.69	0.63	0.63	0.62	0.69	0.64	0.63	**0.75**
	0.05	0.65	0.69	0.65	0.63	0.58	0.58	0.59	0.63	**0.76**
	0.1	0.66	0.68	0.65	0.63	0.57	0.58	0.61	0.61	**0.75**
	0.2	0.66	0.67	0.63	0.57	0.51	0.51	0.54	0.58	**0.75**
	0.4	0.66	0.67	0.63	0.56	0.53	0.53	0.52	0.54	**0.71**

be objectively assessed and cross-compared. However, a disadvantage was that there is no mismatch between the data generating process and the methods that were used for inference, since both are Gaussian BNs. The network learning task might therefore not be representative for typical real-world applications.

For the empirical evaluation we generated data sets with 10 different sample sizes ranging from $N = 10$ to $N = 10,000$ ($10k$) observations. By generating 10 data sets per sample size N, we obtained 100 data sets. For each we implemented five different average fractions of missing data $p_{miss} \in \{0, 0.05, 0.1, 0.2, 0.4\}$, yielding a total of 500 data sets. Table 2 provides the method-specific average AUROC scores for each of the 50 combinations of N and p_{miss}, with each average being across 10 independent data sets. As expected the mean AUROC scores increase in the sample size N and they decrease in the average fraction of missing data p_{miss}. Most importantly, it can be seen that the new BMA method (last column) yields consistently the highest average AUROC scores. To assess the variability in the individual AUROC differences, we computed the p-values of two-sample t-tests for paired samples. When comparing the means of the 8 competing methods with the mean of the BMA method, all p-values were below 0.05, indicating that each individual AUROC difference has a clear sign in favour of BMA which does not only stem from random fluctuations. This shows the superiority of the BMA approach over the two classical methods for dealing with incomplete data.

In Appendix A we compare the models in terms of relative structural Hamming distance (rSHD) scores. The results can be found in Tables A.5 and A.6. It can be seen that the rSHD scores are in good agreement with the AUROC scores. However, it can also be seen that the small sample sizes ($N \leq 50$) yield rSHD scores greater than 1, indicating that the predicted networks are even worse than predicting an empty network without any edges (with $rSHD = 1$). We refer to Appendix A for more details.

5.2. Protein phosphorylation data from signalling pathway (RAF)

A potential disadvantage of our study from Section 5.1 was that Gaussian data perfectly match the assumption of Gaussian data that is made by the network reconstruction methods. The results might thus not be representative for real-world applications. In particular, the BMA method re-samples the missing data points form conditional Gaussian distributions (cf. Section 3.3). This sampling step might be sensitive to deviations from Gaussianity, leading perhaps to erroneous results for non-Gaussian data. To this end, we performed a second study with real protein phosphorylation data from the RAF protein signalling pathway. By extensive flow cytometry experiments, Sachs et al. [42] measured the phosphorylation sites of $n = 11$ key proteins of the RAF signalling cascade. In our second study we focused on the $N_{all} = 3530$ observational data points

from [42]. That is, we ignored the measurements from intervention experiments, in which specific proteins were either inhibited of activated.

The advantage of real wet-lab measurements comes with the disadvantage that the true underlying DAG is not fully known. That is, there is some uncertainty about the true DAG. Hence, an objective evaluation of the network reconstruction accuracies was hindered. However, Sachs et al. [42] reported a gold-standard network of the RAF pathway, which we here used as proxy for the true DAG. The topology of this gold standard network can be found in Section **S3** of the supplementary material.

We distinguished the sample sizes $N \in \{100, 250, 500, 1000, 3530\}$ and crossed them with the five fractions of missing data $p_{miss} \in \{0, 0.05, 0.1, 0.2, 0.4\}$. Since we here could not generate independent data sets, we generated random subsamples of size N from the whole data set with $N_{all} = 3530$ observations. Table 4 shows the method-specific average AUROC scores for each of the 25 combinations of N and p_{miss}, with each average being again across 10 data sets. Like for the ECOLI70 data, the AUROC scores increase in N and they decrease in p_{miss}, and the BMA method (last column) consistently yields the largest AUROC scores. Again we computed the p-values of two sample t-tests for paired samples. Although a few individual p-values are rather high, the majority of p-values is lower than 0.05. In Appendix A we report the corresponding relative structural Hamming distance (rSHD) scores; see Table A.7. For the RAF pathway the rSHD scores deviate from the AUROC results. The structural EM algorithm and the new BMA approach achieve very similar rSHD scores. The differences tend to be non-significant in terms of t-tests for paired samples. Since the rSHD scores were generally rather high ($rSHD \geq 0.74$), we conclude that the RAF pathway topology and in particular the edge types ('undirected' vs. 'directed') can hardly be predicted from observational measurements. We note that the AUROC scores (of up to 0.75) seemed much more satisfactory. An important difference is that we interpreted undirected CPDAG edges as bidirectional edges when computing AUROC scores. This interpretation could explain the higher accuracy, since the CPDAG of the true gold standard RAF pathway contains 17 undirected and only 3 directed edges. Unlike for the rSHD scores, in the context of AUROC scores we interpret this as 17 bidirectional and 3 unidirectional edges; i.e. we interpret it as 47 directed edges in total.

5.3. Additional Bayesian Model Averaging (BMA) results

Additional results for the new BMA approach on the ECOLI70 data can be found in Section **S2** of the supplementary paper. In particular, we computed the rSHD scores for different edge score thresholds ψ, and we ran additional MCMC simulations with a more restrictive graph prior as well as with a more restrictive model parameter prior. The results suggest that a larger edge score threshold can be advantageous for smaller sample sizes ($N = 100$), while for larger sample size ($N = 1000$) $\psi = 0.5$ seems a good choice. The different Bayesian network priors did not affect the convergence rates, but had an effect on the network reconstruction accuracy. While the more restrictive graph prior led to better AUROC scores for small sample sizes, the more restrictive model parameter prior leads to consistently worse AUROC scores. For more details and for interpretations of these results we refer to Section **S2** of the supplementary material.

We also explored the computational costs of the proposed Bayesian Model Averaging (BMA) approach. The detailed results are provided in Section **S2** of the supplementary material. Our main finding is that the MCMC simulation based model inference is computationally more expensive than the inference for the non-Bayesian approaches (NAL and EM). Our run time analyses suggest that the sampling of the missing values can become a computational bottleneck. For large networks with many nodes n and many observations N it has to be looped through the N observations, and for each observation $j = 1, \ldots, N$ a specific conditional Gaussian distribution of the large n-dimensional Gaussian distributions has to be computed, so as to sample the missing data points of observation j from it. Henceforth, up to N conditional Gaussians of n-dimensional Gaussians have to be computed for sampling the missing data.

6. Conclusions and discussion

In this paper we have proposed a new Bayesian model averaging (BMA) approach for learning Gaussian Bayesian networks (BNs) from incomplete data. To the best of our knowledge, this is the first work to propose a fully Bayesian approach for handling missing data in BNs. The new method builds on the Gaussian BGe score [28,29] and extends the classical structure MCMC sampler for posterior sampling DAGs [11,12] by two new MCMC moves that allow the missing data to be posterior sampled. Like earlier proposed methods for handling missing data in BNs [22], the new approach assumes that the data points are missing completely at random (MCAR). In two empirical evaluation studies we have compared the new BMA approach with two alternative non-Bayesian approaches, namely the classical structural EM [19] and the more recently developed NAL approach [22]. The two competing approaches had recently been cross-compared in a method comparison study by Bodewes and Scutari [22]. On benchmark Gaussian BN as well as for real protein phosphorylation data from the RAF protein signalling pathway the new BMA method led to higher network reconstruction accuracies than the two competitors. A potential limitation of the BMA approach is that it would not be easy to adapt the method for discrete BNs (or for conditional Gaussian BNs). For sampling the missing values it is looped through the independent observations of the data set, and for each observation the missing values have to be sampled conditional on the available observed values. For the Gaussian BGe score and any given DAG this can easily be accomplished, because for any possible constellation of missing and available values the missing values can always be sampled from a multivariate conditional Gaussian distribution. From a conceptual perspective the same approach could also be followed for discrete BNs. But unlike for Gaussian BNs, the

required conditional distributions would no longer be of well-known forms, rendering the approach numerically much less practical for discrete BNs.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Structural Hamming distances

In this appendix we assess and cross-compare the network learning methods in terms of the relative structural Hamming distance (rSHD). A description of rSHD scores can be found in Section **S1** of the supplementary material. To extract a concrete network prediction from the marginal edges scores of the new Bayesian modelling averaging (BMA) approach, we employed the standard threshold $\psi = 0.5$. That is, only edges whose scores exceeded $\psi = 0.5$ were assumed to be present. If two oppositely oriented edges (e.g. $A \rightarrow B$ and $A \leftarrow B$) both had a score larger than $\psi = 0.5$, we concluded that there is an undirected edge between A and B. On the other hand, if only one of the two edges (i.e. either $A \rightarrow B$ or $A \leftarrow B$) had a score larger than $\psi = 0.5$, we concluded that the corresponding edge is unidirectional.

The rSHD results for the ECOLI70 data are in agreement with the corresponding AUROC results, i.e. the new Bayesian modelling approach (BMA) yields almost consistently the best (=lowest) rSHD scores; see Tables A.5–A.6. Only in 3 out of 50 cases the BMA does not yield the lowest rSHD, and in these three cases the rSHD differences between BMA and the best non-Bayesian method seems almost neglectable. As an empty network (without any edges) would yield the rSHD score equal to 1, we conclude that rSHD scores higher than 1 indicate an 'unsatisfactory' learning performance. Since only very few of the rSHDs in Table A.5 are slightly below 1, we conclude that low sample sizes of up to $N = 100$ do not allow the true CPDAG of the ECOLI70 network to be properly inferred. However, it can be seen from Table A.6 that higher sample sizes N lead to better performances. As expected, the rSHD scores decrease in the sample size N and increase in the fraction of missing data point p_{miss}.

Table A.7 shows the rSHD scores for the RAF pathway data. Unlike for the AUROC scores in Table 4, we here do not see a clear trend in favour of the new Bayesian model averaging approach (BMA). However, as even the best rSHD scores are still rather high (rSHD≈ 0.75), we conclude that the CPDAG and in particular the edge types ('directed' vs. 'undirected') cannot be properly learned from the available observational protein phosphorylation data. This finding is in agreement with

Table A.5
Average rSHD scores for the ECOLI70 network. This table refers to Table 2 but compares the methods in terms of the rSHD scores rather than AUROC scores. For each sample size N we generated 10 independent data sets. From each data set we then randomly removed data points, so as to achieve different average fractions of missing values p_{miss}. From the incomplete data sets we inferred the networks (CPDAGs) with the structural **EM**, the **NAL** approach and the new Bayesian model averaging (**BMA**) approach. For EM and NAL we distinguished different penalty parameters $\lambda_{n,N}$. The rSHD scores for larger samples sizes N can be found in Table A.6.

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
10	0	**1.14**	1.14	1.14	1.14	1.14	1.15	1.14	1.15	1.30
	0.05	**1.20**	1.20	1.21	1.20	1.21	1.21	1.21	1.21	1.22
	0.1	1.25	1.26	1.26	1.26	1.24	1.24	1.24	1.24	**1.21**
	0.2	1.33	1.33	1.34	1.31	1.30	1.30	1.30	1.30	**1.09**
	0.4	1.49	1.49	1.49	1.49	1.24	1.29	1.19	1.29	**1.04**
25	0	1.93	2.59	2.57	2.52	1.93	2.57	2.57	2.50	**1.63**
	0.05	2.14	2.69	2.62	2.63	2.18	2.29	2.28	2.28	**1.64**
	0.1	2.32	2.79	2.73	2.71	1.92	2.01	2.00	2.00	**1.51**
	0.2	2.43	2.86	2.87	2.80	1.66	1.68	1.68	1.68	**1.42**
	0.4	2.73	2.99	2.99	2.97	1.33	1.34	1.34	1.34	**1.42**
50	0	1.65	4.70	4.37	3.53	1.93	4.62	4.26	3.58	**1.17**
	0.05	1.97	4.78	4.52	4.10	3.15	3.55	3.52	3.45	**1.26**
	0.1	2.30	4.90	4.67	4.31	2.75	2.89	2.89	2.86	**1.29**
	0.2	3.15	5.03	4.88	4.66	2.13	2.17	2.16	2.17	**1.39**
	0.4	3.75	5.36	5.21	4.98	1.62	1.63	1.63	1.63	**1.46**
100	0	1.26	6.17	3.85	1.91	1.34	6.54	4.83	2.65	**0.85**
	0.05	1.44	7.12	5.08	2.47	4.59	5.34	5.24	4.98	**0.89**
	0.1	1.57	7.62	6.05	3.07	3.91	4.11	4.09	4.03	**0.96**
	0.2	2.42	8.24	6.99	4.56	2.82	2.87	2.87	2.85	**1.07**
	0.4	3.99	8.73	7.98	6.16	1.81	1.82	1.81	1.81	**1.47**

Table A.6
Table A.5 continued. Average rSHD scores for the ECOLI70 network for larger sample sizes. See caption of Table A.5 for more information. This table refers to Table 2 but assesses the methods in terms of rSHD rather than AUROC scores.

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
0.25k	0	0.98	3.89	1.85	0.88	1.01	4.95	2.31	0.95	**0.66**
	0.05	1.13	4.92	2.22	0.97	6.22	7.64	7.21	5.98	**0.67**
	0.1	1.32	5.80	2.59	1.11	5.64	5.92	5.84	5.61	**0.72**
	0.2	1.66	7.70	4.15	1.44	4.00	4.04	4.03	4.00	**0.73**
	0.4	2.83	10.15	6.46	2.43	2.44	2.43	2.43	2.43	**0.98**
0.5k	0	0.79	2.56	1.08	**0.49**	0.82	3.31	1.07	0.53	0.50
	0.05	0.95	3.21	1.19	0.59	7.69	9.01	8.18	5.95	**0.49**
	0.1	1.16	3.71	1.55	0.61	6.85	7.15	6.97	6.25	**0.56**
	0.2	1.44	5.59	1.91	1.04	4.82	4.87	4.84	4.75	**0.61**
	0.4	2.24	8.32	3.52	1.19	2.85	2.86	2.85	2.84	**0.79**
1k	0	0.80	1.74	0.73	0.55	0.83	1.87	0.75	0.53	**0.41**
	0.05	0.81	2.15	0.78	0.57	9.15	10.39	9.25	2.90	**0.43**
	0.1	1.00	2.38	0.89	0.67	8.11	8.42	8.03	6.36	**0.47**
	0.2	1.38	3.85	1.18	0.70	5.76	5.78	5.75	5.56	**0.45**
	0.4	2.23	6.57	1.78	0.93	3.29	3.29	3.29	3.28	**0.55**
2.5k	0	0.70	1.08	0.51	0.52	0.72	1.08	0.55	0.55	**0.34**
	0.05	0.91	1.24	0.61	0.50	9.15	9.68	7.52	0.89	**0.34**
	0.1	0.85	1.58	0.65	0.56	9.40	9.41	8.83	6.77	**0.33**
	0.2	1.43	2.12	0.96	0.73	6.81	6.85	6.75	6.01	**0.39**
	0.4	2.02	3.66	1.04	0.99	4.01	4.01	4.00	3.98	**0.51**
5k	0	0.57	0.62	0.49	0.45	0.65	0.72	0.58	0.47	**0.23**
	0.05	0.75	1.16	0.50	0.52	8.90	8.38	5.11	0.55	**0.25**
	0.1	0.70	1.13	0.50	0.52	10.4	10.4	9.59	4.56	**0.26**
	0.2	1.24	1.26	0.81	0.55	7.68	7.68	7.54	6.06	**0.25**
	0.4	1.93	2.54	1.20	0.73	4.40	4.40	4.40	4.44	**0.44**
10k	0	0.53	0.53	0.49	0.50	0.70	0.68	0.59	0.49	**0.17**
	0.05	0.44	0.63	0.47	0.48	8.58	8.44	3.75	0.57	**0.24**
	0.1	1.15	0.70	0.36	0.57	11.0	11.1	10.1	1.10	**0.28**
	0.2	1.45	1.37	0.73	0.48	8.37	8.36	8.15	4.21	**0.21**
	0.4	2.21	1.98	1.27	0.77	4.98	4.98	4.95	4.82	**0.44**

Table A.7
Average rSHD scores for the RAF pathway network. This table refers to Table 4 but compares the methods in terms of the rSHD scores rather than AUROC scores. For each sample size N we randomly sub-sampled 10 independent data sets. From each data set we then randomly removed data points, so as to achieve different average fractions of missing values p_{miss}. From the incomplete data sets we inferred the networks (CPDAGs) with the structural **EM**, the **NAL** approach and the new Bayesian model averaging (**BMA**) approach. For EM and NAL we distinguished four different penalty parameters $\lambda_{n,N}$.

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
100	0	1.06	1.15	0.98	0.89	1.08	1.24	1.02	0.92	**0.91**
	0.05	1.01	1.17	0.95	0.91	1.51	1.67	1.46	1.29	**0.90**
	0.1	1.00	1.16	1.00	0.88	1.77	1.87	1.71	1.48	**0.94**
	0.2	0.97	1.10	0.93	**0.85**	1.89	1.83	1.80	1.77	0.88
	0.4	1.12	1.22	**0.91**	0.91	1.40	1.43	1.49	1.39	0.95
250	0	0.93	0.95	**0.81**	0.86	0.99	1.02	0.82	0.87	0.82
	0.05	0.95	0.95	0.85	0.87	1.56	1.59	1.34	1.10	**0.83**
	0.1	0.98	1.01	0.85	0.87	1.79	1.79	1.55	1.27	**0.84**
	0.2	0.97	0.99	**0.85**	0.87	2.09	2.12	2.04	1.83	0.86
	0.4	0.93	1.00	**0.83**	0.88	1.74	0.75	1.72	1.72	0.90
500	0	0.87	0.81	**0.80**	0.85	0.89	0.84	0.81	0.85	0.82
	0.05	0.88	0.85	**0.80**	0.86	1.54	1.51	1.32	1.03	0.82
	0.1	0.97	0.86	0.79	0.85	1.75	1.73	1.42	1.10	**0.78**
	0.2	0.86	0.85	0.83	0.90	2.11	2.01	1.90	1.58	**0.82**
	0.4	0.99	0.89	**0.86**	0.88	1.85	1.86	1.84	1.90	0.90
1k	0	0.94	0.82	**0.79**	0.86	0.93	0.81	0.80	0.86	0.82
	0.05	0.93	0.81	**0.79**	0.85	1.67	1.59	1.32	0.97	0.83
	0.1	0.84	0.79	0.78	0.84	1.81	1.76	1.56	1.14	**0.75**
	0.2	0.94	0.79	**0.77**	0.88	1.96	1.98	1.77	1.47	0.78
	0.4	0.88	0.83	0.81	0.88	2.10	2.05	2.08	1.98	**0.78**

(*continued on next page*)

Table A.7 (*continued*)

N	p_{miss}	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	BMA
all	0	1.10	0.76	0.80	0.80	1.05	0.77	0.80	0.80	**0.75**
	0.05	1.01	**0.75**	0.78	0.80	1.69	1.58	1.23	0.97	0.79
	0.1	1.00	**0.75**	0.78	0.82	1.90	1.73	1.40	0.92	0.80
	0.2	0.91	0.76	0.79	0.87	1.95	1.91	1.84	1.16	**0.74**
	0.4	0.86	**0.76**	0.78	0.89	2.13	2.11	2.11	1.98	0.81

the results reported in [47]. Given the unsatisfactory learning performance of all methods, we would argue that for the RAF data the cross-method comparison in terms of rSHD scores is only of limited informative value.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijar.2023.108954.

References

[1] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco, CA, USA, 1988.
[2] R. Neapolitan, Probabilistic Reasoning in Expert Systems: Theory and Algorithms, CreateSpace, Scotts Valley, CA, USA, 1989.
[3] D.M. Chickering, Learning Bayesian networks is NP-complete, in: D. Fisher, H.J. Lenz (Eds.), Learning from Data: Artificial Intelligence and Statistics, vol. 5, Springer, New York, 1996, pp. 121–130.
[4] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, Springer Verlag, New York, 2001.
[5] M. Kalisch, P. Bühlmann, Estimating high-dimensional directed acyclic graphs with the PC-algorithm, J. Mach. Learn. Res. 8 (2007) 613–636.
[6] D. Marella, P. Vicard, Bayesian network structural learning from complex survey data: a resampling based approach, Stat. Methods Appl. 31 (2022) 981–1013.
[7] R.R. Bouckaert, Properties of Bayesian belief network learning algorithms, in: R.L. de Mántaras, D. Poole (Eds.), UAI'94: Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 1994, pp. 102–109.
[8] D. Chickering, D. Geiger, D. Heckerman, Learning Bayesian networks: search methods and experimental results, in: Proceedings of Fifth Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence in Statistics, Ft. Lauderdale, FL, 1995, pp. 112–128.
[9] J. Cussens, Bayesian network learning with cutting planes, in: F. Cozman, A. Pfeffer (Eds.), UAI 2011: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2011, pp. 153–160.
[10] A.C. Constantinou, Y. Liu, N.K. Kitson, K. Chobtham, Z. Guo, Effective and efficient structure learning with pruning and model averaging strategies, Int. J. Approx. Reason. 151 (2022) 292–321.
[11] D. Madigan, J. York, Bayesian graphical models for discrete data, Int. Stat. Rev. 63 (1995) 215–232.
[12] P. Giudici, R. Castelo, Improving Markov chain Monte Carlo model search for data mining, Mach. Learn. 50 (2003) 127–158.
[13] N. Friedman, D. Koller, Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks, Mach. Learn. 50 (2003) 95–126.
[14] M. Grzegorczyk, D. Husmeier, Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move, Mach. Learn. 71 (2008) 265–305.
[15] J. Kuipers, G. Moffa, Partition MCMC for inference on acyclic digraphs, J. Am. Stat. Assoc. 112 (2017) 282–299.
[16] M. Scutari, C.E. Graafland, J.M. Gutiérrez, Who learns better Bayesian network structures: constraint-based, score-based or hybrid algorithms?, in: International Conference on Probabilistic Graphical Models, PMLR, 2018, pp. 416–427.
[17] J. Kuipers, P. Suter, G. Moffa, Efficient sampling and structure learning of Bayesian networks, J. Comput. Graph. Stat. 31 (2022) 639–650.
[18] N.K. Kitson, A.C. Constantinou, Z. Guo, Y. Liu, K. Chobtham, A survey of Bayesian network structure learning, arXiv:2109.11415 [abs], 2021.
[19] N. Friedman, Learning belief networks in the presence of missing values and hidden variables, in: D.H. Fisher (Ed.), Proceedings of the Fourteenth International Conference on Machine Learning (ICML), Morgan Kaufmann, Nashville, Tennessee, USA, 1997, pp. 125–133.
[20] A.P. Dempster, N.M. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1977) 1–38.
[21] N. Balov, Consistent model selection of discrete Bayesian networks from incomplete data, Electron. J. Stat. 7 (2013) 1047–1077.
[22] T. Bodewes, M. Scutari, Learning Bayesian networks from incomplete data with the node-average likelihood, Int. J. Approx. Reason. 138 (2021) 145–160.
[23] M. Scutari, Bayesian network models for incomplete and dynamic data, Stat. Neerl. 74 (2020) 397–419.
[24] N. Friedman, The Bayesian structural EM algorithm, in: G.F. Cooper, S. Moral (Eds.), Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI), Morgan Kaufmann, Madison, Wisconsin, USA, 1998, pp. 129–138.
[25] M. Beal, Z. Ghahramani, The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures, in: J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West (Eds.), Proceedings of the 7th Valencia International Meeting, Oxford, UK, 2003, pp. 453–464.
[26] T. Adel, C. de Campos, Learning Bayesian networks with incomplete data by augmentation, in: S. Satinder, S. Markovitch (Eds.), Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017, pp. 1684–1690.
[27] M. Scanagatta, G. Corani, M. Zaffalon, J. Yoo, U. Kang, Efficient learning of bounded-tree width Bayesian networks from complete and incomplete data sets, Int. J. Approx. Reason. 95 (2018) 152–166.
[28] D. Geiger, D. Heckerman, Parameter priors for directed acyclic graphical models and the characterization of several probability distributions, Ann. Stat. 30 (2002) 1412–1440.
[29] J. Kuipers, G. Moffa, D. Heckerman, Addendum on the scoring of Gaussian directed acyclic graphical models, Ann. Stat. 42 (2014) 1689–1691.
[30] D. Heckerman, D. Geiger, Learning Bayesian networks: a unification for discrete and Gaussian domains, in: Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95), Morgan Kaufmann, San Francisco, CA, 1995, pp. 274–284.
[31] D. Geiger, D. Heckerman, Learning Gaussian networks, in: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, 1994, pp. 235–243.
[32] D.M. Chickering, A transformational characterization of equivalent Bayesian network structures, in: International Conference on Uncertainty in Artificial Intelligence (UAI), vol. 11, 1995, pp. 87–98.
[33] D.M. Chickering, Learning equivalence classes of Bayesian-network structures, J. Mach. Learn. Res. 2 (2002) 445–498.

[34] D. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.
[35] R.D. Shachter, R. Kenley, Gaussian influence diagrams, Manag. Sci. 35 (1989) 527–550.
[36] G.R. Shafer, P.P. Shenoy, Probability propagation, Ann. Math. Artif. Intell. 2 (1990) 327–351.
[37] S.L. Lauritzen, The EM algorithm for graphical association models with missing data, Comput. Stat. Data Anal. 19 (1995) 191–201.
[38] R. Bouchaert, Bayesian belief networks: from construction to inference, Ph.D. thesis, University of Utrecht, 1995.
[39] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, Adaptive Computation and Machine Learning Series, MIT Press, Cambridge, MA, USA, ISBN 9780262013192, 2009.
[40] A. Ruggieri, F. Stranieri, F. Stella, M. Scutari, Hard and soft EM in Bayesian network learning from incomplete data, Algorithms 13 (2020).
[41] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[42] K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, G. Nolan, Protein-signaling networks derived from multiparameter single-cell data, Science 308 (2005) 523–529.
[43] M. Scutari, Learning Bayesian networks with the bnlearn R package, J. Stat. Softw. 35 (2010) 1–22.
[44] M. Scutari, Bayesian network constraint-based structure learning algorithms: parallel and optimized implementations in the bnlearn R package, J. Stat. Softw. 77 (2017) 1–20.
[45] M. Scutari, J.-B. Denis, Bayesian Networks: With Examples in R, Chapman & Hall, Boca Raton, FL, 2014.
[46] J. Schäfer, K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol. 4 (2005) 32.
[47] A.V. Werhli, M. Grzegorczyk, D. Husmeier, Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks, Bioinformatics 22 (2006) 2523–2531.