Hierarchically Coupled Gaussian Bayesian Networks

Chang Ma CHANG.MA@RUG.NL
Bernoulli Institute
University of Groningen
Groningen, NL

Marco Grzegorczyk M.A.GRZEGORCZYK@RUG.NL
Bernoulli Institute
University of Groningen
Groningen, NL

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

Gaussian Bayesian networks are probabilistic graphical models for learning network structures from continuous data. Standard approaches typically assume that all observations arise from a homogeneous population. In many applications, however, the data originate from multiple related but heterogeneous groups that share a common network structure while having group-specific network parameters. Pooling all observations and inferring a single set of parameters ignores between-group heterogeneity and may lead to biased inference. Conversely, estimating group-specific parameters independently substantially increases model complexity and inferential uncertainty.

We introduce a hierarchical prior that enables information sharing across groups. The proposed prior preserves group-specific parameters while introducing shrinkage by coupling the group-specific mean vectors and covariance structures through shared global hyperparameters. The corresponding coupling strengths are inferred from the data, allowing the degree of information sharing to adapt automatically to the observed heterogeneity across groups. We develop a Markov chain Monte Carlo (MCMC) algorithm for posterior inference and demonstrate empirically that the proposed approach improves both network structure learning and predictive performance.

Keywords: Gaussian Bayesian networks, hierarchical Bayesian modelling, structure learning, heterogeneous groups, information coupling, Markov chain Monte Carlo

1. Introduction

Bayesian networks (BNs) are probabilistic graphical models that use directed acyclic graphs (DAGs) to represent conditional dependencies among random variables. Each node in a DAG corresponds to a variable, while directed edges encode probabilistic relationships. BNs have become important tools in statistics, machine learning, and artificial intelligence. Comprehensive introductions to BNs are available in standard textbooks (Pearl, 1988; Neapolitan, 1989; Koller and Friedman, 2009; Scutari and Denis, 2014).

BN structure learning algorithms aim to infer DAGs from data. Since the search space grows super-exponentially with the number of nodes (Chickering, 1996), this problem is computationally challenging. Consequently, the development of efficient BN structure learning algorithms has received considerable attention; see, for example, Scutari et al. (2018, 2019);

© 2026 C. Ma & M. Grzegorczyk.

Scutari (2020); Kitson et al. (2023). In sparse-data settings, there is often substantial uncertainty regarding the underlying DAG. Within a Bayesian framework, this uncertainty can be addressed by sampling DAGs from the posterior distribution and averaging over them, an approach commonly referred to as Bayesian model averaging (Constantinou et al., 2022). Examples of Bayesian BN structure learning algorithms include Madigan and York (1995); Giudici and Castelo (2003); Friedman and Koller (2003); Grzegorczyk and Husmeier (2008); Kuipers and Moffa (2017); Nikzad et al. (2025).

Beyond the standard BN framework, tailored approaches have been developed for incomplete data (Friedman, 1997; Ruggieri et al., 2020; Bodewes and Scutari, 2021; Grzegorczyk, 2023), ordinal data (Luo et al., 2021; Castelletti, 2024; Grzegorczyk, 2024; Scauda et al., 2025), and hybrid data (Chickering et al., 1995; Scutari, 2020; Grzegorczyk, 2025). These extensions substantially broaden the applicability of BNs, but they retain the assumption that all observations arise from a single homogeneous data-generating process.

Many real-world applications involve observations arising from multiple related but heterogeneous groups rather than from a single homogeneous population. In such settings, it may be assumed that the groups share a common underlying network structure but that the network parameters vary across groups (Grzegorczyk and Husmeier, 2011). A straightforward strategy is to pool all observations and to infer a single parameter set across groups. However, this ignores heterogeneity and may lead to biased inference. Conversely, inferring group-specific parameters independently can become unstable when the number of observations per group is limited. As a principled compromise, we propose a new hierarchical prior for learning Gaussian BNs (GBNs) from multiple related datasets (groups). The proposed prior couples the group-specific mean vectors and covariance matrices through shared global hyperparameters, thereby encouraging similar groups to borrow strength from each other while still allowing for group-specific deviations. The mean vectors and covariance matrices are coupled separately with different strengths, each controlled by a dedicated coupling parameter. Both coupling strength parameters are inferred from the data.

Several approaches have been proposed for learning BNs from related datasets. One of the earliest examples is Bayesian multinets, which allow different network structures under different contexts or hypotheses (Geiger and Heckerman, 1996). More recently, Azzimonti et al. (2021) proposed a Bayesian score for learning common network structures from related discrete datasets. Subsequently, Scutari et al. (2022) developed a frequentist mixed-effects approach for learning GBNs from related datasets. Our work differs from these approaches by introducing a fully Bayesian hierarchical framework for learning GBNs from related datasets. The proposed framework enables adaptive information sharing across groups through jointly inferred hierarchical priors on both the group-specific mean vectors and covariance structures.

2. Methods

2.1. Bayesian Networks (BNs)

Bayesian networks (BNs) use directed acyclic graphs (DAGs) to represent conditional independencies among random variables. The n nodes of the DAG correspond to the variables $X_1, \ldots, X_n$, and directed edges encode direct conditional dependencies. Node X_j is called a parent of X_i if there is an edge $X_j \rightarrow X_i$, and it is called an ancestor of X_i if X_i can be

reached from X_j by following a sequence of directed edges, symbolically $X_j \to \ldots \to X_i$. By definition, a node cannot be its own ancestor, so cyclic paths of directed edges, such as $X_i \to \ldots \to X_i$, are excluded. The set of parent nodes of X_i is denoted by $\mathbf{\Pi}_i$. A DAG $\mathcal{G}$ is fully characterized by its n parent sets $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_n$. Given $\mathcal{G}$, the joint density factorizes according to the parent sets implied by $\mathcal{G}$:

$$p(X_1, \ldots, X_n \mid \mathcal{G}) = \prod_{i=1}^{n} p(X_i \mid \mathbf{\Pi}_i). \tag{1}$$

Two DAGs are equivalent if they imply the same conditional independencies, and equivalent DAGs form an equivalence class. DAG equivalence classes can be represented by completed partially directed acyclic graphs (CPDAGs) (Chickering, 2002). Two DAGs are equivalent if and only if they share the same skeleton and v-structures. A CPDAG contains directed edges shared by all DAGs in the equivalence class and undirected edges whose orientation is not identifiable from conditional independencies.

The random variables $X_1, \ldots, X_n$ can be stacked into a random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$. A dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ consists of m independent realizations, where $\mathbf{x}_i$ denotes the i-th observation of $\mathbf{X}$. The posterior distribution of a DAG $\mathcal{G}$ is given by

$$p(\mathcal{G} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{G}) p(\mathcal{G})}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \mathcal{G}) p(\mathcal{G}), \tag{2}$$

where $p(\mathcal{D} \mid \mathcal{G})$ is the marginal likelihood, $p(\mathcal{G})$ is the prior over DAGs, and $p(\mathcal{D})$ is a normalizing constant. Using Markov chain Monte Carlo (MCMC) sampling, DAGs can be sampled from the posterior distribution, yielding a sample $\mathcal{G}_1, \ldots, \mathcal{G}_R$.

The marginal likelihood $p(\mathcal{D} \mid \mathcal{G})$ is often referred to as the *score* of $\mathcal{G}$. For certain model classes, this score can be computed in closed form. Two widely used scoring metrics are the *BDe scoring metric* for categorical variables (Heckerman et al., 1995; Chickering et al., 1995; Heckerman and Geiger, 1995), and the *BGe scoring metric* for Gaussian variables (Geiger and Heckerman, 2002; Kuipers et al., 2014; Heckerman and Geiger, 1995; Geiger and Heckerman, 1994). Both metrics allow the marginal likelihood to be computed analytically and ensure *score equivalence*, meaning that equivalent DAGs receive the same score.

2.2. Gaussian Bayesian Networks (GBNs)

In Gaussian Bayesian networks (GBNs), the random vector $\mathbf{X} \in \mathbb{R}^n$ is assumed to follow an n-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. For the dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of m independent observations, the likelihood is given by

$$p(\mathcal{D} \mid \boldsymbol{\mu}, \mathbf{\Sigma}) = \prod_{i=1}^{m} \mathcal{N}_n(\mathbf{x}_i \mid \boldsymbol{\mu}, \mathbf{\Sigma}), \tag{3}$$

where $\mathcal{N}_n(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the density of the n-variate Gaussian distribution at $\mathbf{x}$. A key feature of GBNs is that the covariance matrix $\mathbf{\Sigma}$ needs to be consistent with a given DAG. Although the Gaussian distribution is naturally parameterized by its covariance matrix

$\boldsymbol{\Sigma}$, conjugate Bayesian inference is more conveniently formulated in terms of the precision matrix $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$.

For any complete DAG $\mathcal{G}_C$ (i.e. a fully connected DAG consistent with a node ordering), a fully conjugate prior for the mean vector $\boldsymbol{\mu}$ and the precision matrix $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ is given by the Normal–Wishart distribution:

$$\mathbf{W} \mid \mathcal{G}_C \sim \mathcal{W}_n\big(\alpha_w, (\alpha_w - n - 1)\mathbf{T}_0\big), \text{ and } \boldsymbol{\mu} \mid (\mathbf{W}, \mathcal{G}_C) \sim \mathcal{N}_n\big(\boldsymbol{\mu}_0, (\alpha_\mu \mathbf{W})^{-1}\big), \tag{4}$$

where $\mathbf{T}_0 \in \mathbb{R}^{n\times n}$, $\boldsymbol{\mu}_0 \in \mathbb{R}^n$, $\alpha_w > n+1$, and $\alpha_\mu > 0$ are fixed hyperparameters. The matrix $(\alpha_w - n - 1)\mathbf{T}_0$ is the rate matrix of the Wishart distribution, such that

$$p(\mathbf{W} \mid \mathcal{G}_C) \propto |\mathbf{W}|^{(\alpha_w - n - 1)/2} \exp\big(-\tfrac{1}{2}\operatorname{tr}\big((\alpha_w - n - 1)\mathbf{T}_0\mathbf{W}\big)\big)\,.$$

Under this parametrization, the prior expectation of the covariance matrix is given by $\mathbb{E}[\boldsymbol{\Sigma} \mid \mathcal{G}_C] = \mathbb{E}[\mathbf{W}^{-1} \mid \mathcal{G}_C] = \mathbf{T}_0$, so that $\mathbf{T}_0$ is the prior covariance matrix.

2.3. GBNs for Multi-Group Data

Suppose that the observations are partitioned into H groups indexed by $h = 1, \ldots, H$. Let $\mathcal{D}_h = \{\mathbf{x}_{h,1}, \ldots, \mathbf{x}_{h,m_h}\}$ denote the dataset of group h consisting of m_h independent observations, where $\sum_{h=1}^{H} m_h = m$. Assume that all groups share a common DAG $\mathcal{G}$, while the parameters are allowed to differ across groups. For each group h,

$$\mathbf{x}_{h,l} \sim \mathcal{N}_n(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \qquad (l = 1, \ldots, m_h), \tag{5}$$

where $\boldsymbol{\mu}_h \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_h \in \mathbb{R}^{n\times n}$ denote the group-specific mean vector and covariance matrix, respectively. For complete DAGs $\mathcal{G}_C$, independent and identically distributed (iid) priors can be assigned to the group-specific parameter pairs $(\mathbf{W}_h, \boldsymbol{\mu}_h)$:

$$\mathbf{W}_h \mid (\mathcal{G}_C) \sim \mathcal{W}_n\big(\alpha_w, (\alpha_w - n - 1)\mathbf{T}_0\big), \tag{6}$$

$$\boldsymbol{\mu}_h \mid (\mathbf{W}_h, \mathcal{G}_C) \sim \mathcal{N}_n\big(\boldsymbol{\mu}_0, (\alpha_\mu \mathbf{W}_h)^{-1}\big), \tag{7}$$

where $\mathbf{W}_h = \boldsymbol{\Sigma}_h^{-1}$ and the hyperparameters are defined as in Section 2.2. Under these iid priors, the parameters are independent across groups conditional on the shared hyperparameters. For each group h, define the empirical mean vector and scatter matrix as

$$\bar{\mathbf{x}}_{[h]} := \frac{1}{m_h}\sum_{l=1}^{m_h} \mathbf{x}_{h,l}, \qquad \mathbf{S}_{[h]} := \sum_{l=1}^{m_h} (\mathbf{x}_{h,l} - \bar{\mathbf{x}}_{[h]})(\mathbf{x}_{h,l} - \bar{\mathbf{x}}_{[h]})^\top. \tag{8}$$

Conditional on a complete DAG $\mathcal{G}_C$, integrating out $\mathbf{W}_h$ and $\boldsymbol{\mu}_h$ under the prior (6)–(7) yields the marginal likelihood

$$p(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\Theta}) = \prod_{h=1}^{H} \left[\pi^{-\frac{n m_h}{2}} \left(\frac{\alpha_\mu}{\alpha_\mu + m_h}\right)^{\frac{n}{2}} \frac{\Gamma_n\big(\frac{\alpha_w + m_h}{2}\big)}{\Gamma_n\big(\frac{\alpha_w}{2}\big)} \frac{\det(\mathbf{T}_0^\dagger)^{\frac{\alpha_w}{2}}}{\det(\mathbf{T}_0^\dagger + \mathbf{T}_{[h]}^{\boldsymbol{\mu}_0})^{\frac{\alpha_w + m_h}{2}}} \right], \tag{9}$$

where $\mathbf{T}_0^\dagger := (\alpha_w - n - 1)\mathbf{T}_0$, $\Gamma_n(\cdot)$ denotes the multivariate gamma function, and

$$\mathbf{T}_{[h]}^{\boldsymbol{\mu}_0} = \mathbf{S}_{[h]} + \frac{\alpha_\mu m_h}{\alpha_\mu + m_h}(\bar{\mathbf{x}}_{[h]} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_{[h]} - \boldsymbol{\mu}_0)^\top. \tag{10}$$

Here, $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu\}$ denotes the set of hyperparameters.[1] For an arbitrary DAG $\mathcal{G}$, the marginal likelihood can be computed using the factorization of Geiger and Heckerman (Geiger and Heckerman, 1994, 2002; Kuipers et al., 2014):

$$p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} \frac{p(\mathcal{D}^{\{X_i, \boldsymbol{\Pi}_i\}} \mid \mathcal{G}_C^{\{X_i, \boldsymbol{\Pi}_i\}}, \boldsymbol{\Theta})}{p(\mathcal{D}^{\{\boldsymbol{\Pi}_i\}} \mid \mathcal{G}_C^{\{\boldsymbol{\Pi}_i\}}, \boldsymbol{\Theta})}. \tag{11}$$

For a variable subset $L \subseteq \{X_1, \ldots, X_n\}$, the marginal likelihood $p(\mathcal{D}^L \mid \mathcal{G}_C^L, \boldsymbol{\Theta})$ is obtained from (9) by restricting the model to the variables in L. Specifically, the dimension n is replaced by $n_L := |L|$, $\Gamma_n(\cdot)$ is replaced by $\Gamma_{n_L}(\cdot)$, all matrices are replaced by their principal submatrices indexed by L, and the empirical quantities $\bar{\mathbf{x}}_{[h]}$ and $\mathbf{S}_{[h]}$ are computed using only the variables in L.

2.4. GBNs with Hierarchically Coupled Group-Specific Parameters

The priors in (6)–(7) treat the groups independently and do not allow for information sharing across them. This can lead to unstable parameter inference when the number of observations per group is limited. To address this limitation, we propose a hierarchical prior that induces coupling between the group-specific precision matrices $\mathbf{W}_h$ and mean vectors $\boldsymbol{\mu}_h$ through shared global hyperparameters. Specifically, we assume that the precision matrices are coupled through a common random rate matrix $\mathbf{T}_0$, with coupling strength controlled by $\alpha_w > n + 1$, and that the mean vectors are coupled through a shared centre $\boldsymbol{\mu}_0$, with coupling strength controlled by $\alpha_\mu > 0$. For the complete DAGs $\mathcal{G}_C$ we have

$$\mathbf{W}_h \mid (\mathbf{T}_0, \alpha_w, \mathcal{G}_C) \sim \mathcal{W}_n\big(\alpha_w,\ \alpha_w \mathbf{T}_0\big), \tag{12}$$

$$\boldsymbol{\mu}_h \mid (\mathbf{W}_h, \boldsymbol{\mu}_0, \alpha_\mu, \mathcal{G}_C) \sim \mathcal{N}_n\big(\boldsymbol{\mu}_0,\ (\alpha_\mu \mathbf{W}_h)^{-1}\big), \tag{13}$$

with hyperpriors

$$\mathbf{T}_0 \sim \mathcal{W}_n\big(\alpha_0,\ \alpha_0 \mathbf{I}_n\big), \tag{14}$$

$$\boldsymbol{\mu}_0 \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n). \tag{15}$$

In contrast to the standard homogeneous GBN from Section 2.2 and the uncoupled multi-group GBN from Section 2.3, the parameters $\mathbf{T}_0$ and $\boldsymbol{\mu}_0$ are treated here as random quantities. $\mathbf{T}_0$ represents a shared global baseline structure around which the group-specific precision matrices are centered. We adopt the rate parametrization $\alpha_w \mathbf{T}_0$ rather than the conventional parametrization $(\alpha_w - n - 1)\mathbf{T}_0$. The new parametrization places relatively more emphasis on the shared matrix $\mathbf{T}_0$ (for comparable values of α_w) and thereby induces stronger coupling between the group-specific precision matrices.[2] Similarly, $\boldsymbol{\mu}_0$ represents a global mean around which the group-specific mean vectors $\boldsymbol{\mu}_h$ are shrunk. The parameters α_w and α_μ control the strength of coupling across groups. Specifically, larger values of α_w

1. The special case $H = 1$ and $m_1 = m$ yields the marginal likelihood of the homogeneous GBN from Section 2.2. A common choice of the hyperparameters is $\mathbf{T}_0 = \mathbf{I}$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\alpha_w = n + 2$, and $\alpha_\mu = 1$.
2. For the homogeneous and uncoupled multi-group GBNs, the rate parametrization $\alpha_w \mathbf{T}_0$ typically yields inferior performance, as it assigns more weight to a fixed, often arbitrarily chosen matrix $\mathbf{T}_0$; for example, $\mathbf{T}_0 = \mathbf{I}$. In the proposed coupled multi-group GBN, $\mathbf{T}_0$ is inferred from the data and acts as a shared baseline covariance structure, up to scaling, through which information is exchanged across groups.

concentrate the precision matrices $\mathbf{W}_h$ more tightly around the shared structure induced by $\mathbf{T}_0$, while larger values of α_μ shrink the group-specific mean vectors $\boldsymbol{\mu}_h$ more strongly towards the global mean $\boldsymbol{\mu}_0$. Conditionally on the hyperparameters $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu\}$, the marginal likelihood retains the same closed-form expression as in (9)–(11), except that $\mathbf{T}_0^\dagger = (\alpha_w - n - 1)\mathbf{T}_0$ is replaced by $\mathbf{T}_0^\dagger = \alpha_w \mathbf{T}_0$ to reflect the modified parametrization.

2.5. Conditional Distributions for Complete DAGs

Owing to conjugacy, all required conditional distributions remain available in closed form for complete DAGs $\mathcal{G}_C$. Derivations are given in Appendix A. For each group h we have

$$\mathbf{W}_h \mid (\mathcal{D}, \boldsymbol{\mu}_0, \alpha_\mu, \mathbf{T}_0, \alpha_w, \mathcal{G}_C) \;\sim\; \mathcal{W}_n\Big(\alpha_w + m_h,\; \alpha_w \mathbf{T}_0 + \mathbf{T}_{[h]}^{\boldsymbol{\mu}_0}\Big), \tag{16}$$

$$\boldsymbol{\mu}_h \mid (\mathcal{D}, \mathbf{W}_h, \boldsymbol{\mu}_0, \alpha_\mu, \mathcal{G}_C) \;\sim\; \mathcal{N}_n\Big(\boldsymbol{\mu}_{[h]}^\star,\; \big((\alpha_\mu + m_h)\mathbf{W}_h\big)^{-1}\Big), \tag{17}$$

where $\mathbf{T}_{[h]}^{\boldsymbol{\mu}_0}$ is defined in (10), and $\boldsymbol{\mu}_{[h]}^\star := \frac{\alpha_\mu \boldsymbol{\mu}_0 + m_h \bar{\mathbf{x}}_{[h]}}{\alpha_\mu + m_h}$. The corresponding full conditional distributions for the global parameters are

$$\mathbf{T}_0 \mid (\mathbf{W}_{1:H}, \alpha_w) \;\sim\; \mathcal{W}_n\Big(\alpha_0 + H\alpha_w,\; \alpha_0 \mathbf{I}_n + \alpha_w \sum_{h=1}^H \mathbf{W}_h\Big), \tag{18}$$

$$\boldsymbol{\mu}_0 \mid (\boldsymbol{\mu}_{1:H}, \mathbf{W}_{1:H}, \alpha_\mu) \;\sim\; \mathcal{N}_n\Big(\boldsymbol{\mu}_0^\diamond,\; \big(\mathbf{I}_n + \alpha_\mu \sum_{h=1}^H \mathbf{W}_h\big)^{-1}\Big), \tag{19}$$

where $\mathbf{W}_{1:H} := \{\mathbf{W}_1, \ldots, \mathbf{W}_H\}$, $\boldsymbol{\mu}_{1:H} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H\}$, and

$$\boldsymbol{\mu}_0^\diamond = \big(\mathbf{I}_n + \alpha_\mu \sum_{h=1}^H \mathbf{W}_h\big)^{-1} \big(\alpha_\mu \sum_{h=1}^H \mathbf{W}_h \boldsymbol{\mu}_h\big). \tag{20}$$

We have for the expectation $\mathbb{E}[\mathbf{T}_0 \mid \mathbf{W}_{1:H}, \alpha_w] = (\alpha_0 + H\alpha_w)\left(\alpha_0 \mathbf{I}_n + \alpha_w \sum_{h=1}^H \mathbf{W}_h\right)^{-1}$.

2.6. Metropolis-Hastings Sampling for Arbitrary DAGs

For arbitrary DAGs there are no closed-form conditional distributions and we need to resort to Metropolis–Hastings (MH) sampling. However, we use the complete DAG conditional distributions from Section 2.5 to construct efficient 2-step proposal kernels. Given the current parameters $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu\}$, we sample auxiliary precision matrices $\mathbf{W}_{1:H}^\star$ from Eq. (16) and subsequently a new candidate matrix $\mathbf{T}_0^\star$ from Eq. (18). The auxiliary variables are used solely to define the proposal distribution and are discarded immediately afterwards. This proposal mechanism defines a transition kernel $q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0)$. According to the MH criterion, we accept the new candidate matrix $\mathbf{T}_0^\star$ with probability $A = \min\{1, R\}$, with

$$R = \frac{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star)}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})} \times \frac{p(\mathbf{T}_0^\star)}{p(\mathbf{T}_0)} \times \frac{q_C(\mathbf{T}_0 \mid \mathbf{T}_0^\star)}{q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0)} = \ldots = \frac{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star)}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})} \times \frac{p(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\Theta})}{p(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\Theta}^\star)}$$

where $\boldsymbol{\Theta}^\star := \{\mathbf{T}_0^\star, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu\}$, and the latter equality is shown in Appendix B.

Similarly, given the current parameters $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu\}$, we sample auxiliary parameters $\mathbf{W}^\star_{1:H}$ and $\boldsymbol{\mu}^\star_{1:H}$ from Eqs. (16) and (17), and subsequently a new candidate vector $\boldsymbol{\mu}^\star_0$ from Eq. (19). This proposal mechanism defines another transition kernel $q_C(\boldsymbol{\mu}^\star_0 \mid \boldsymbol{\mu}_0)$, and we have the MH acceptance probability $A = \min\{1, R\}$, with

$$R = \frac{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star)}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})} \times \frac{p(\boldsymbol{\mu}^\star_0)}{p(\boldsymbol{\mu}_0)} \times \frac{q_C(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}^\star_0)}{q_C(\boldsymbol{\mu}^\star_0 \mid \boldsymbol{\mu}_0)} = \ldots = \frac{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star)}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})} \times \frac{p(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\Theta})}{p(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\Theta}^\star)}$$

where $\boldsymbol{\Theta}^\star := \{\mathbf{T}_0, \boldsymbol{\mu}^\star_0, \alpha_w, \alpha_\mu\}$, and the latter equality is shown in Appendix C.

To infer appropriate coupling strengths, we treat the hyperparameters $\alpha_\mu > 0$ and $\alpha_w > n + 1$ as unknown and infer them from the data. We define $\tilde{\alpha}_w := \alpha_w - n - 1$, and impose Gamma priors on α_μ and $\tilde{\alpha}_w$, $\alpha_\mu \sim \text{Gamma}(a_\mu, b_\mu)$ and $\tilde{\alpha}_w \sim \text{Gamma}(a_w, b_w)$, where b_μ and b_w denote the rate parameters. As $\tilde{\alpha}_w > 0$, we have $\alpha_w > n + 1$. The MH sampling steps for both parameters are described in Appendix D.

2.7. Markov Chain Monte Carlo Sampling Algorithm

To generate posterior samples, we employ a structure MCMC sampler for Bayesian networks (Madigan and York, 1995; Giudici and Castelo, 2003) and augment it with the MH parameter updates described in Section 2.6. Given the current state $[\mathcal{G}, \boldsymbol{\Theta}]$, where $\mathcal{G}$ is a DAG and $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\}$, the MCMC algorithm alternates between two steps.

Step 1: Given $[\mathcal{G}, \boldsymbol{\Theta}]$, let $N(\mathcal{G})$ denote the neighbourhood of $\mathcal{G}$, i.e. all DAGs obtained by adding, deleting, or reversing a single edge. Keep $\boldsymbol{\Theta}$ fixed, and propose a new DAG $\mathcal{G}^\star \in N(\mathcal{G})$ uniformly at random. The acceptance probability is

$$A = \min\left\{1, \frac{p(\mathcal{D} \mid \mathcal{G}^\star, \boldsymbol{\Theta})}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})} \cdot \frac{p(\mathcal{G}^\star)}{p(\mathcal{G})} \cdot \frac{|N(\mathcal{G})|}{|N(\mathcal{G}^\star)|}\right\}. \tag{21}$$

Step 2: Given $[\mathcal{G}, \boldsymbol{\Theta}]$, keep the DAG $\mathcal{G}$ fixed, and update the model parameters by performing one Metropolis–Hastings update for each component of $\boldsymbol{\Theta}$.

2.8. Performance Measures

Repeating the two sampling steps from Section 2.7 yields a Markov chain with the desired posterior distribution as stationary distribution. We run the algorithm for T iterations, discard the first $T/2$ iterations as burn-in, and thin the remaining samples by a factor ξ, resulting in $R = \left\lfloor \frac{T}{2\xi} \right\rfloor$ posterior samples $[\mathcal{G}^{(1)}, \boldsymbol{\Theta}^{(1)}], \ldots, [\mathcal{G}^{(R)}, \boldsymbol{\Theta}^{(R)}]$.

We follow Friedman and Koller (2003) and compute marginal edge posterior probabilities

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

where $\mathcal{I}_{X_i \to X_j}(\mathcal{G}) = 1$ if the CPDAG associated with $\mathcal{G}$ contains either the directed edge $X_i \to X_j$ or the undirected edge $X_i - X_j$, and $\mathcal{I}_{X_i \to X_j}(\mathcal{G}) = 0$ otherwise.[3] Each posterior

3. Each sampled DAG is first converted to its CPDAG. If the CPDAG contains the undirected edge $X_i - X_j$, we treat both edge orientations as present for evaluation purposes. Accordingly, undirected edges in the true CPDAG are also treated as bidirectional.

edge probability $\hat{p}_{i,j}$ serves as a score for predicting the presence of an edge. Imposing a threshold $\psi \in [0,1]$ on the edge probabilities $\hat{p}_{i,j}$ yields a predicted network $\hat{\mathcal{G}}_\psi$, which contains only those edges whose posterior probabilities exceed ψ. Evaluating all distinct threshold values yields a sequence of precision–recall pairs, where precision denotes the proportion of predicted edges that are correct and recall denotes the proportion of true CPDAG edges that are recovered. Connecting these points yields the precision–recall curve. The **area under this curve (AUPRC)**, with $0 \leq \text{AUPRC} \leq 1$, is a widely used measure of network reconstruction accuracy (Davis and Goadrich, 2006), with larger values indicating better performance.

For each posterior sample $[\mathcal{G}^{(r)}, \boldsymbol{\Theta}^{(r)}]$, we then draw posterior samples of the corresponding group-specific Gaussian parameters $\{(\boldsymbol{\mu}_1^{(r)}, \boldsymbol{\Sigma}_1^{(r)}), \ldots, (\boldsymbol{\mu}_H^{(r)}, \boldsymbol{\Sigma}_H^{(r)})\}$, where each covariance matrix $\boldsymbol{\Sigma}_h^{(r)}$ is consistent with $\mathcal{G}^{(r)}$, as described in Appendix E. For a new dataset $\mathcal{D}^\star = \{\mathcal{D}_1^\star, \ldots, \mathcal{D}_H^\star\}$, where $\mathcal{D}_h^\star = \{\mathbf{x}_{h,1}^\star, \ldots, \mathbf{x}_{h,m_h^\star}^\star\}$, we estimate the **geometric-mean predictive probability** via Monte Carlo integration:

$$\hat{p}(\mathcal{D}^\star \mid \mathcal{D}) := \left[\frac{1}{R} \sum_{r=1}^{R} \prod_{h=1}^{H} \prod_{i=1}^{m_h^\star} \mathcal{N}_n(\mathbf{x}_{h,i}^\star \mid \boldsymbol{\mu}_h^{(r)}, \boldsymbol{\Sigma}_h^{(r)}) \right]^{1/m^\star}, \text{ where } m^\star = \sum_{h=1}^{H} m_h^\star.$$

2.9. Implementation Details

In the absence of prior knowledge regarding the underlying network structure, a uniform graph prior over all DAGs $\mathcal{G}$ was adopted, thereby avoiding an unwarranted preference for sparse network structures. For the homogeneous and uncoupled multi-group GBNs, we employ weakly informative Normal–Wishart priors. Specifically, we set the prior covariance matrix to $\mathbf{T}_0 = \mathbf{I}_n$ and the global mean to $\boldsymbol{\mu}_0 = \mathbf{0}$, where $\mathbf{I}_n$ denotes the identity matrix and $\mathbf{0}$ the zero vector. A relatively diffuse prior specification is obtained by setting the equivalent sample sizes to $\alpha_w = n + 2$ and $\alpha_\mu = 1$.
For the hierarchically coupled multi-group GBN, the parameters $\mathbf{T}_0$ and $\boldsymbol{\mu}_0$ are treated as unknown and assigned the hyperpriors $\mathbf{T}_0 \sim \mathcal{W}_n(\alpha_0, \alpha_0 \mathbf{I}_n)$, and $\boldsymbol{\mu}_0 \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$, with $\alpha_0 = n + 2$, resulting in a relatively diffuse prior specification. For the coupling strength parameters, we impose the following Gamma hyperpriors $\alpha_\mu \sim \text{Gamma}(a_\mu = 1, b_\mu = 1)$, and $\tilde{\alpha}_w \sim \text{Gamma}(a_w = 2, b_w = 0.5)$, where $\alpha_w = \tilde{\alpha}_w + n + 1$. This yields the prior expectations $\mathbb{E}[\alpha_\mu] = 1$ and $\mathbb{E}[\alpha_w] = n + 5$.
All empirical results reported in this paper are based on $T = 200{,}000$ MCMC iterations. As a conservative default that consistently yielded satisfactory convergence in our empirical studies, the first $T/2$ iterations are discarded as burn-in. A fixed thinning factor $\xi = 100$ is applied, resulting in a posterior sample size of $R = 1000$.

3. Data

We evaluate model performance using both simulated and real-world datasets.

Simulated Network Data: We assume that the variables $X_1, \ldots, X_n$ are arranged in topological order, yielding $\zeta = n(n-1)/2$ possible edges $X_j \to X_i$ with $j < i$. For

each edge, an independent Bernoulli indicator is drawn with success probability $p = 20/\zeta$. Consequently, the expected number of included edges equals $\zeta p = 20$. This defines a common directed acyclic graph structure shared across all groups. To construct the group-specific covariance matrices, we first consider a zero-mean GBN in which each observation $l = 1, \ldots, m_h$ in group h satisfies

$$x_{i,h,l} = \sum_{j \in \mathbf{\Pi}_i} b_{i,j,h}\, x_{j,h,l} + \epsilon_{i,h,l}, \tag{22}$$

where $x_{i,h,l}$ denotes the value of X_i, $\mathbf{\Pi}_i \subset \{X_1, \ldots, X_{i-1}\}$ is the parent set of X_i, and $\epsilon_{i,h,l} \sim \mathcal{N}(0, \sigma_{i,h}^2)$ is Gaussian noise. For each node X_i, let $\mathbf{b}_i$ denote the generic parameter vector consisting of regression coefficients $b_{i,j}$ followed by the noise standard deviation σ_i. The entries of $\mathbf{b}_i$ are sampled independently from Unif$(1, 2)$. Random signs are assigned to the first $|\mathbf{\Pi}_i|$ entries, after which each vector is normalized to unit Euclidean norm, $\mathbf{b}_i \leftarrow \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|_2}$. To generate similar covariance structures, we adapt an idea from Grzegorczyk and Husmeier (2013): The generic parameter vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n$ define a baseline GBN whose covariance matrix is $\mathbf{\Sigma}$ (Shachter and Kenley, 1989). We then introduce group-specific perturbations. For each group h, we generate deviation vectors $\mathbf{d}_{1,h}, \ldots, \mathbf{d}_{n,h}$ in the same manner and normalize them to unit Euclidean norm. Subsequently, we define

$$\mathbf{w}_{i,h} := \frac{\mathbf{b}_i + \alpha\, \mathbf{d}_{i,h}}{\|\mathbf{b}_i + \alpha\, \mathbf{d}_{i,h}\|_2}, \qquad (i = 1, \ldots, n), \tag{23}$$

where $\alpha \geq 0$ controls the degree of heterogeneity across groups. The entries of $\mathbf{w}_{i,h}$ define the group-specific regression coefficients and noise standard deviation in (22); specifically, the first $|\mathbf{\Pi}_i|$ components correspond to the regression coefficients $b_{i,j,h}$, while the final component defines $\sigma_{i,h} := |w_{i,h,|\mathbf{\Pi}_i|+1}|$, thereby ensuring $\sigma_{i,h} > 0$. Larger values of α produce less similar parameter vectors across groups.

For each group h, the parameter vectors $\mathbf{w}_{1,h}, \ldots, \mathbf{w}_{n,h}$ define a covariance matrix $\mathbf{\Sigma}_h$. The resulting covariance matrices $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_H$ are perturbations of the baseline covariance matrix $\mathbf{\Sigma}$, with their similarity controlled by α. In addition, we generate datasets with *independent* group-specific parameters by setting $\mathbf{b}_i = \mathbf{0}$ and $\alpha = 1$, so that $\mathbf{w}_{i,h} := \frac{\mathbf{d}_{i,h}}{\|\mathbf{d}_{i,h}\|_2} = \mathbf{d}_{i,h}$.

In the same vein, we sample a baseline mean vector, $\boldsymbol{\mu} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$, and group-specific deviation vectors, $\boldsymbol{\tau}_h \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$, and define

$$\boldsymbol{\mu}_h := \frac{\boldsymbol{\mu} + \alpha \boldsymbol{\tau}_h}{\sqrt{1 + \alpha^2}},$$

so that $\boldsymbol{\mu}_h \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$. For the *independent* parameter scenario, we set $\boldsymbol{\mu}_h := \boldsymbol{\tau}_h$. Finally, conditional on these group-specific means and covariance matrices, we sample the group-specific observations from $\mathbf{x}_{h,l} \sim \mathcal{N}_n(\boldsymbol{\mu}_h, \mathbf{\Sigma}_h) \quad (h = 1, \ldots, H, \;\; l = 1, \ldots, m_h)$.

We set $n = 10$ and $H = 5$, and combine $\alpha \in \{0, 0.125, 0.25, 0.5, 0.75\}$ and the *independent* parameter scenario with $m_h \in \{20, 40, 80\}$ yielding $\sum_h m_h = m \in \{100, 200, 400\}$. For each of the 18 combinations, we generate 25 sets of related groups. From each setting, one dataset per group is sampled, yielding a total of 450 simulated multi-group datasets.

Benchmark Multi-Group Datasets: We consider the following benchmark datasets: **Iris** $(n = 4, m = 150, H = 3)$, **Fish Market** $(n = 6, m = 159, H = 7)$, Wisconsin **Breast**

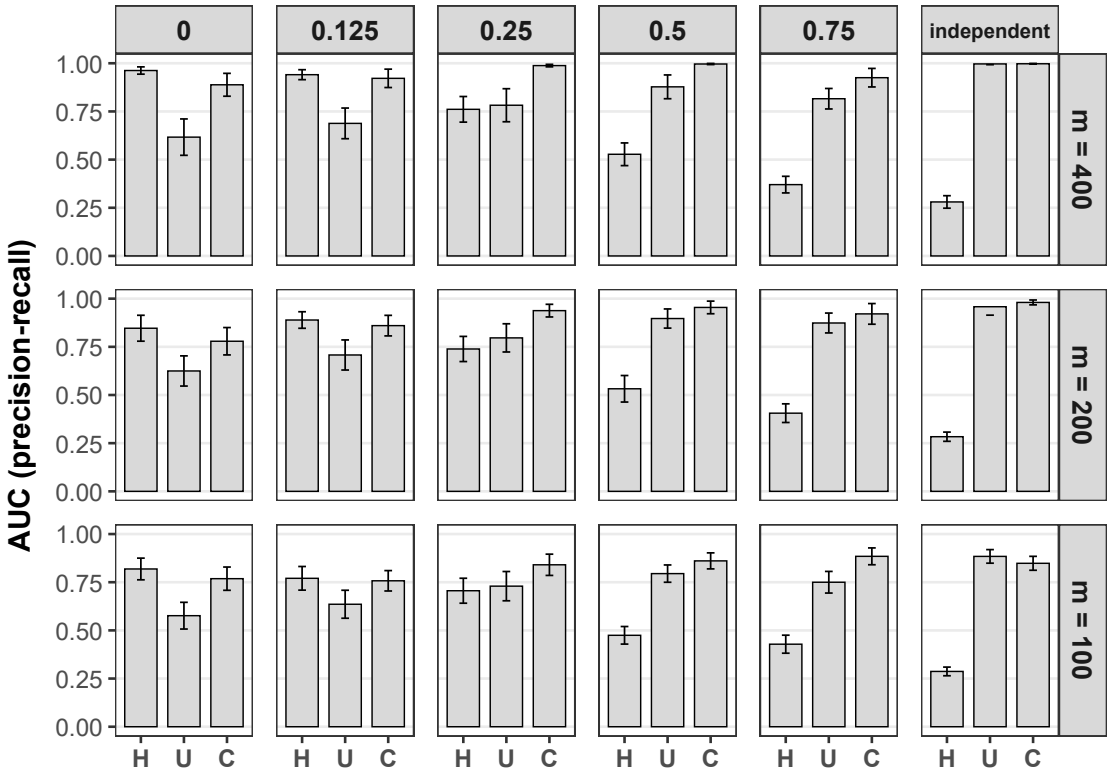


Figure 1: Network reconstruction accuracy for simulated multi-group data. For sample sizes $m \in \{100, 200, 400\}$ and six parameter scenarios, data were generated as described in Section 3. Parameter similarity decreases from left ($\alpha = 0$) to right (*independent*). Panels show the mean area under the precision–recall curve (AUPRC) for the homogeneous (H), the uncoupled multi-group (U), and the proposed coupled multi-group (C) GBN. Error bars indicate 95% confidence intervals.

Cancer ($n = 10, m = 569, H = 2$), **Wine** ($n = 13, m = 178, H = 3$), **Vehicle** Silhouettes ($n = 18, m = 813, H = 3$), and **Parkinsons** ($n = 22, m = 195, H = 2$). For each dataset, the groups correspond to the class labels provided with the data. Observations containing missing values were removed. Within each dataset, all variables were standardized using z-score normalization prior to train/test splits.

4. Results

On the simulated multi-group network data, we evaluate the network reconstruction accuracy of the models. Figure 1 presents the mean area under the precision–recall curve (AUPRC) together with error bars. As expected, the homogeneous GBN performs best for homogeneous data ($\alpha = 0$) but poorly as the data become increasingly heterogeneous ($\alpha \geq 0.5$), whereas the uncoupled multi-group GBN exhibits the opposite trend. Over-

all, the proposed coupled multi-group GBN performs best. In particular, for moderate heterogeneity ($0.25 \leq \alpha \leq 0.75$), it tends to outperform both competitors, and paired t-tests suggest that the improvements in favor of the new coupled model are significant in most settings ($p < 0.05$). For the remaining values of α, the proposed coupled multi-group GBN tends to perform similarly to the best-performing competitor. Specifically, its performance is comparable to that of the homogeneous GBN for homogeneous ($\alpha = 0$) and mildly heterogeneous data ($\alpha = 0.125$), and to that of the uncoupled multi-group GBN in the independent-parameter setting. This suggests that the new hierarchical prior is able to adapt automatically between the two limiting models. Interestingly, the mean AUPRCs of the two multi-group GBNs increase as the data become more heterogeneous. One possible explanation is that, under stronger heterogeneity, true edges tend to receive stronger support in at least some groups. In homogeneous settings, some edges may instead receive only weak support across all groups, which makes their identification more difficult.

On the real-world datasets, we assess model performance using posterior predictive probabilities. For each dataset, we repeatedly selected 25%, 50%, 75%, and 90% of the observations for training and computed the geometric-mean of the posterior predictive probabilities of the held-out test observations. Figure 2 shows boxplots of the log predictive probabilities over 50 replicates. The results are consistent with the network reconstruction results obtained for multi-group data with moderate heterogeneity. Overall, the proposed coupled GBN achieves the highest predictive probabilities, whereas the homogeneous GBN performs worst. These findings suggest that the proposed hierarchical prior can also improve predictive performance in heterogeneous real-world data.

5. Conclusion

We introduced a Bayesian hierarchical extension of Gaussian Bayesian networks (GBNs) for multi-group data that enables adaptive information sharing across groups through joint coupling of mean vectors and covariance structures. Posterior inference is performed using structure MCMC combined with Metropolis–Hastings updates for the hierarchical model parameters. Our empirical results show that hierarchical coupling improves both network reconstruction accuracy and predictive performance, particularly when the groups exhibit moderate similarity. At the same time, the proposed approach remains robust in the limiting cases where groups have either identical or independent network parameters.

A key advantage of the proposed Bayesian framework over competing approaches, such as the mixed-effects model of Scutari et al. (2022), is its modularity. As a result, our methodology can be readily integrated with recently proposed Bayesian frameworks for learning GBNs from incomplete, ordinal, hybrid, and mixture data; see the quartet of papers Grzegorczyk (2023, 2024, 2025, 2026).

Our immediate next research goal is to incorporate the hierarchical prior into mixture GBNs along the lines of Grzegorczyk et al. (2008); Grzegorczyk (2026). Rather than assuming fixed and known groups, we aim to learn mixture GBNs from heterogeneous data with unknown group memberships. This extension will enable simultaneous inference of the latent component assignments and the component-specific model parameters while allowing information to be shared across mixture components through the hierarchical prior. The resulting model will adaptively infer the degree of similarity between component-specific

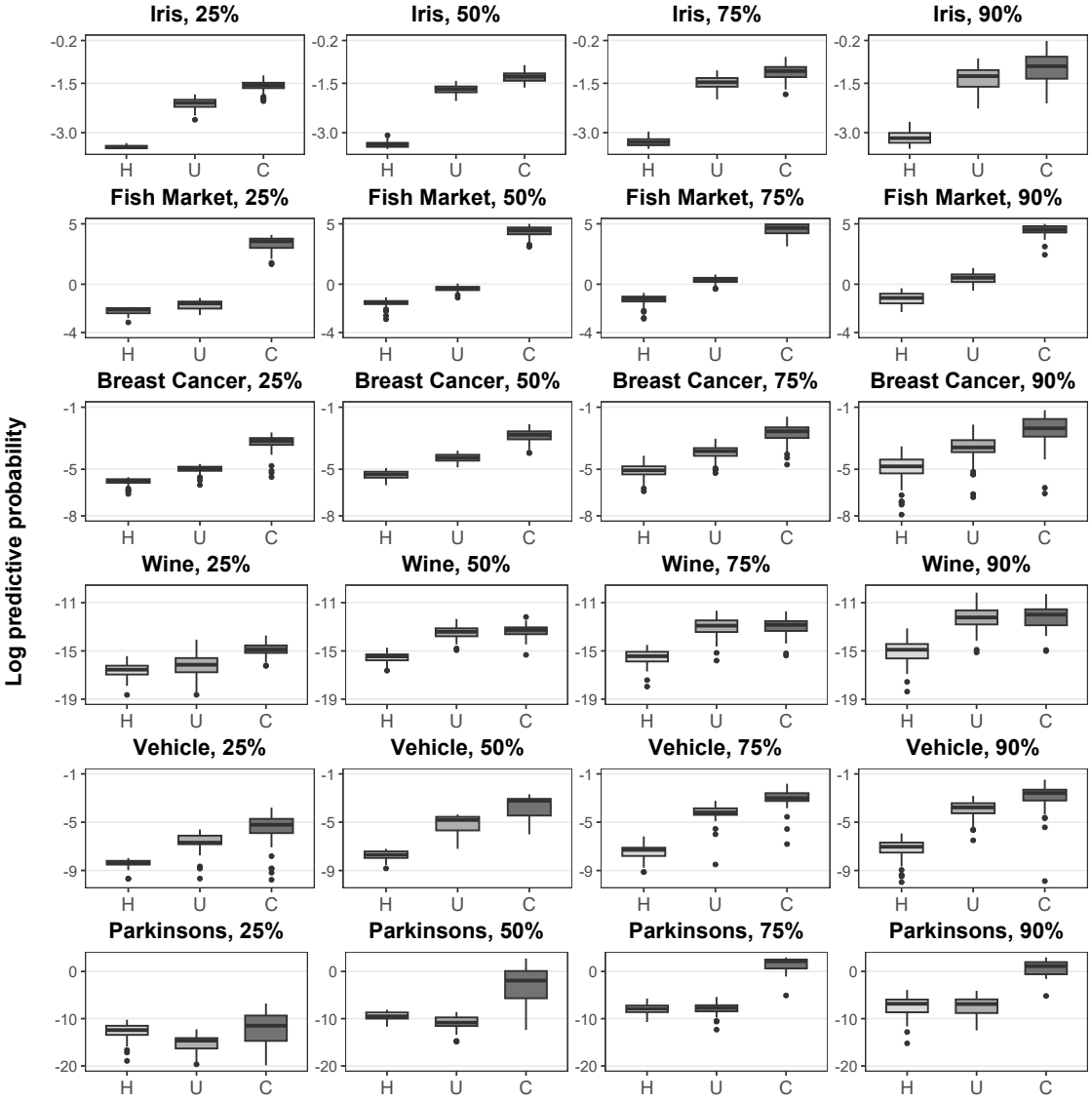


Figure 2: Predictive probabilities for real-world data. We repeatedly selected 25%–90% of the observations for training and computed the predictive probability of the held-out observations. The boxplots correspond to the homogeneous (H), uncoupled (U), and new coupled (C) GBNs and show the log predictive probabilities.

parameters, thereby providing a bridge between full-covariance (Grzegorczyk et al., 2008) and tied-covariance (Grzegorczyk, 2026) mixture GBNs.

Acknowledgments

C.M. is supported by China Scholarship Council (CSC) under Grant number 202206730055.

References

L. Azzimonti, G. Corani, and M. Scutari. A Bayesian hierarchical score for structure learning from related data sets. *International Journal of Approximate Reasoning*, 134:183–196, 2021.

T. Bodewes and M. Scutari. Learning Bayesian networks from incomplete data with the node-average likelihood. *International Journal of Approximate Reasoning*, 138:145–160, 2021.

F. Castelletti. Learning Bayesian networks: A Copula approach for mixed-type data. *Psychometrika*, 89:658–686, 2024.

D. M. Chickering. Learning Bayesian networks is NP-complete. In D. Fisher and H. J. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics*, volume 5, pages 121–130, New York, 1996. Springer.

D. M. Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

D.M. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian networks: Search methods and experimental results. *Machine Learning*, 20:112–128, 1995.

A. C. Constantinou, Y. Liu, N. K. Kitson, K. Chobtham, and Z. Guo. Effective and efficient structure learning with pruning and model averaging strategies. *International Journal of Approximate Reasoning*, 151:292–321, 2022.

J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd international conference on Machine Learning*, pages 233–240, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2.

N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In Douglas H. Fisher, editor, *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 125–133, Nashville, Tennessee, USA, 1997. Morgan Kaufmann.

N. Friedman and D. Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–126, 2003.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 235–243, San Francisco, CA., 1994. Morgan Kaufmann.

D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82(1–2):45–74, 1996.

D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5): 1412–1440, 2002.

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

M. Grzegorczyk. Being Bayesian about learning Gaussian Bayesian networks from incomplete data. *International Journal of Approximate Reasoning*, 160:108954, 2023.

M. Grzegorczyk. Being Bayesian about learning Bayesian networks from ordinal data. *International Journal of Approximate Reasoning*, 170:109205, 2024.

M. Grzegorczyk. Being Bayesian about learning Bayesian networks from hybrid data. *International Journal of Approximate Reasoning*, 187:109549, 2025.

M. Grzegorczyk. A BGe score for tied-covariance mixtures of Gaussian Bayesian networks. *International Journal of Approximate Reasoning*, 198:109771, 2026.

M. Grzegorczyk and D. Husmeier. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71(2-3):265–305, 2008.

M. Grzegorczyk and D. Husmeier. Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning*, 83(3):355–419, 2011.

M. Grzegorczyk and D. Husmeier. Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. *Machine Learning*, 91(1):105–154, 2013.

M. Grzegorczyk, D. Husmeier, K. Edwards, P. Ghazal, and A. Millar. Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. *Bioinformatics*, 24(18):2071–2078, 2008.

D. Heckerman and 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)*, pages 274–82, San Francisco, CA, 1995. Morgan Kaufmann.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:245–274, 1995.

N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham. A survey of Bayesian network structure learning. *Artificial Intelligence Review*, 56:8721–8814, 2023.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, USA, 2009. ISBN: 9780262013192.

J. Kuipers and G. Moffa. Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112:282–299, 2017.

J. Kuipers, G. Moffa, and D. Heckerman. Addendum on the scoring of Gaussian directed acyclic graphical models. *The Annals of Statistics*, 42(4):1689–1691, 2014.

X.G. Luo, G. Moffa, and J. Kuipers. Learning Bayesian networks from ordinal data. *Journal of Machine Learning Research*, 22:1–44, 2021.

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

R.E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. CreateSpace, Scotts Valley CA, USA, 1989.

D. Nikzad, A. Zhilkin, J. Harviainen, J. Kuipers, G. Moffa, and M. Koivisto. Scaling up Bayesian DAG sampling, 2025. URL https://arxiv.org/abs/2510.25254.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, USA, 1988.

A. Ruggieri, F. Stranieri, F. Stella, and M. Scutari. Hard and soft EM in Bayesian network learning from incomplete data. *Algorithms*, 13(12), 2020.

M. Scauda, J. Kuipers, and G. Moffa. A latent causal inference framework for ordinal variables, 2025. URL https://arxiv.org/abs/2502.10276.

M. Scutari. Bayesian network models for incomplete and dynamic data. *Statistica Neerlandica*, 74:397–419, 2020.

M. Scutari and J-B. Denis. *Bayesian Networks: With Examples in R*. Chapman & Hall, Boca Raton, FL, 2014.

M. Scutari, C. E. Graafland, and J. M. Gutiérrez. Who learns better Bayesian network structures: Constraint-based, score-based or hybrid algorithms? In *International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 416–427, 2018.

M. Scutari, C.E. Graafland, and J.M. Gutiérrez. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.

M. Scutari, C. Marquis, and L. Azzimonti. Using mixed-effects models to learn Bayesian networks from related data sets. In *International Conference on Probabilistic Graphical Models*, volume 186 of *Proceedings of Machine Learning Research*, pages 73–84, 2022.

R. D. Shachter and R. Kenley. Gaussian influence diagrams. *Management Science*, 35: 527–550, 1989.

Appendix A. Conditional Distributions for Complete DAGs

This appendix derives the conditional distributions presented in Section 2.5. Throughout, we use the rate parametrization of the Wishart distribution

$$p(\mathbf{W} \mid \nu, \mathbf{R}) \propto |\mathbf{W}|^{(\nu-n-1)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\mathbf{R}\mathbf{W})\right\}.$$

We omit terms that do not depend on the parameter currently being updated.

Collapsed conditional of $\mathbf{W}_h$. For group h, we derive the conditional distribution of $\mathbf{W}_h$ with $\boldsymbol{\mu}_h$ integrated out. The likelihood is

$$p(\mathcal{D}_h \mid \boldsymbol{\mu}_h, \mathbf{W}_h) \propto |\mathbf{W}_h|^{m_h/2} \exp\left\{-\frac{1}{2}\sum_{l=1}^{m_h}(\mathbf{x}_{h,l}-\boldsymbol{\mu}_h)^\top \mathbf{W}_h(\mathbf{x}_{h,l}-\boldsymbol{\mu}_h)\right\}.$$

The relevant prior distributions are

$$\begin{aligned}\mathbf{W}_h \mid (\mathbf{T}_0, \alpha_w) &\sim \mathcal{W}_n(\alpha_w, \alpha_w\mathbf{T}_0),\\ \boldsymbol{\mu}_h \mid (\mathbf{W}_h, \boldsymbol{\mu}_0, \alpha_\mu) &\sim \mathcal{N}_n\left(\boldsymbol{\mu}_0, (\alpha_\mu \mathbf{W}_h)^{-1}\right).\end{aligned}$$

Combining the likelihood with the conditional prior of $\boldsymbol{\mu}_h$, the quadratic terms involving $\boldsymbol{\mu}_h$ can be written as

$$\begin{aligned}&\sum_{l=1}^{m_h}(\mathbf{x}_{h,l}-\boldsymbol{\mu}_h)^\top \mathbf{W}_h(\mathbf{x}_{h,l}-\boldsymbol{\mu}_h) + \alpha_\mu(\boldsymbol{\mu}_h-\boldsymbol{\mu}_0)^\top\mathbf{W}_h(\boldsymbol{\mu}_h-\boldsymbol{\mu}_0)\\ &\quad= \operatorname{tr}\left(\mathbf{T}_{[h]}^{\boldsymbol{\mu}_0}\mathbf{W}_h\right) + (\alpha_\mu+m_h)(\boldsymbol{\mu}_h-\boldsymbol{\mu}_{[h]}^\star)^\top\mathbf{W}_h(\boldsymbol{\mu}_h-\boldsymbol{\mu}_{[h]}^\star),\end{aligned}$$

where

$$\boldsymbol{\mu}_{[h]}^\star = \frac{\alpha_\mu\boldsymbol{\mu}_0 + m_h\bar{\mathbf{x}}_{[h]}}{\alpha_\mu+m_h}$$

and

$$\mathbf{T}_{[h]}^{\boldsymbol{\mu}_0} = \mathbf{S}_{[h]} + \frac{\alpha_\mu m_h}{\alpha_\mu+m_h}(\bar{\mathbf{x}}_{[h]}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}_{[h]}-\boldsymbol{\mu}_0)^\top.$$

Integrating the resulting Gaussian kernel with respect to $\boldsymbol{\mu}_h$ and combining the remaining terms with the Wishart prior for $\mathbf{W}_h$ gives

$$p(\mathbf{W}_h \mid \mathcal{D}_h, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu) \propto |\mathbf{W}_h|^{(\alpha_w+m_h-n-1)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\alpha_w\mathbf{T}_0 + \mathbf{T}_{[h]}^{\boldsymbol{\mu}_0}\right)\mathbf{W}_h\right]\right\}.$$

Recognizing the Wishart kernel yields

$$\mathbf{W}_h \mid (\mathcal{D}_h, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_w, \alpha_\mu) \sim \mathcal{W}_n\left(\alpha_w+m_h, \alpha_w\mathbf{T}_0 + \mathbf{T}_{[h]}^{\boldsymbol{\mu}_0}\right).$$

Full conditional of $\boldsymbol{\mu}_h$.
Collecting all terms depending on $\boldsymbol{\mu}_h$, we obtain

$$p(\boldsymbol{\mu}_h \mid \mathbf{W}_h, \mathcal{D}_h, \boldsymbol{\mu}_0, \alpha_\mu) \propto \exp\{-\tfrac{1}{2}Q_h(\boldsymbol{\mu}_h)\},$$

where

$$Q_h(\boldsymbol{\mu}_h) = \sum_{l=1}^{m_h}(\mathbf{x}_{h,l}-\boldsymbol{\mu}_h)^\top\mathbf{W}_h(\mathbf{x}_{h,l}-\boldsymbol{\mu}_h) + \alpha_\mu(\boldsymbol{\mu}_h-\boldsymbol{\mu}_0)^\top\mathbf{W}_h(\boldsymbol{\mu}_h-\boldsymbol{\mu}_0).$$

Expanding and collecting terms gives

$$Q_h(\boldsymbol{\mu}_h) = (\alpha_\mu+m_h)\boldsymbol{\mu}_h^\top\mathbf{W}_h\boldsymbol{\mu}_h - 2(\alpha_\mu\boldsymbol{\mu}_0 + m_h\bar{\mathbf{x}}_{[h]})^\top\mathbf{W}_h\boldsymbol{\mu}_h + \text{const.}$$

Completing the square gives

$$Q_h(\boldsymbol{\mu}_h) = (\alpha_\mu + m_h)(\boldsymbol{\mu}_h - \boldsymbol{\mu}^\star_{[h]})^\top \mathbf{W}_h (\boldsymbol{\mu}_h - \boldsymbol{\mu}^\star_{[h]}) + \text{const},$$

with

$$\boldsymbol{\mu}^\star_{[h]} = \frac{\alpha_\mu \boldsymbol{\mu}_0 + m_h \bar{\mathbf{x}}_{[h]}}{\alpha_\mu + m_h}.$$

Thus,

$$\boldsymbol{\mu}_h \mid (\cdots) \sim \mathcal{N}_n\Big(\boldsymbol{\mu}^\star_{[h]},\ \big((\alpha_\mu + m_h)\mathbf{W}_h\big)^{-1}\Big).$$

Full conditional of $\mathbf{T}_0$.
The prior is $\mathbf{T}_0 \sim \mathcal{W}_n(\alpha_0, \alpha_0 \mathbf{I}_n)$. As a function of $\mathbf{T}_0$, we have

$$p(\mathbf{W}_h \mid \mathbf{T}_0, \alpha_w) \propto |\mathbf{T}_0|^{\alpha_w/2} \exp\left\{-\frac{1}{2}\operatorname{tr}(\alpha_w \mathbf{T}_0 \mathbf{W}_h)\right\}.$$

Multiplying the likelihood contributions over $h = 1, \ldots, H$ and combining them with the prior gives

$$p(\mathbf{T}_0 \mid \{\mathbf{W}_h\}, \alpha_w) \propto |\mathbf{T}_0|^{(\alpha_0 - n - 1)/2} |\mathbf{T}_0|^{H\alpha_w/2} \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\alpha_0 \mathbf{I}_n + \alpha_w \sum_{h=1}^{H} \mathbf{W}_h\right)\mathbf{T}_0\right]\right\}.$$

Therefore,

$$p(\mathbf{T}_0 \mid \cdots) \propto |\mathbf{T}_0|^{(\alpha_0 + H\alpha_w - n - 1)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\alpha_0 \mathbf{I}_n + \alpha_w \sum_{h=1}^{H} \mathbf{W}_h\right)\mathbf{T}_0\right]\right\}.$$

Recognizing the Wishart kernel yields

$$\mathbf{T}_0 \mid (\cdots) \sim \mathcal{W}_n\left(\alpha_0 + H\alpha_w,\ \alpha_0 \mathbf{I}_n + \alpha_w \sum_{h=1}^{H} \mathbf{W}_h\right).$$

Full conditional of $\boldsymbol{\mu}_0$.
The prior is $\boldsymbol{\mu}_0 \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$. Collecting all terms depending on $\boldsymbol{\mu}_0$ gives

$$p(\boldsymbol{\mu}_0 \mid \{\boldsymbol{\mu}_h\}, \{\mathbf{W}_h\}, \alpha_\mu) \propto \exp\left\{-\frac{1}{2} R(\boldsymbol{\mu}_0)\right\},$$

where

$$R(\boldsymbol{\mu}_0) = \boldsymbol{\mu}_0^\top \boldsymbol{\mu}_0 + \alpha_\mu \sum_{h=1}^{H} (\boldsymbol{\mu}_h - \boldsymbol{\mu}_0)^\top \mathbf{W}_h (\boldsymbol{\mu}_h - \boldsymbol{\mu}_0).$$

Expanding and collecting terms gives

$$R(\boldsymbol{\mu}_0) = \boldsymbol{\mu}_0^\top \left(\mathbf{I}_n + \alpha_\mu \sum_{h=1}^{H} \mathbf{W}_h\right) \boldsymbol{\mu}_0 - 2\left(\alpha_\mu \sum_{h=1}^{H} \mathbf{W}_h \boldsymbol{\mu}_h\right)^\top \boldsymbol{\mu}_0 + \text{const}.$$

Completing the square yields a Gaussian kernel with precision

$$\mathbf{I}_n + \alpha_\mu \sum_{h=1}^{H} \mathbf{W}_h$$

and mean

$$\boldsymbol{\mu}_0^\diamond = \left(\mathbf{I}_n + \alpha_\mu \sum_{h=1}^{H} \mathbf{W}_h \right)^{-1} \left(\alpha_\mu \sum_{h=1}^{H} \mathbf{W}_h \boldsymbol{\mu}_h \right).$$

Hence,

$$\boldsymbol{\mu}_0 \mid (\cdots) \sim \mathcal{N}_n \left(\boldsymbol{\mu}_0^\diamond, \ \left(\mathbf{I}_n + \alpha_\mu \sum_{h=1}^{H} \mathbf{W}_h \right)^{-1} \right).$$

Appendix B. Hastings-Ratio for Sampling the Global Covariance

Recall that we defined $\mathbf{W}_{1:H} := \{\mathbf{W}_1, \ldots, \mathbf{W}_H\}$, and that $\mathcal{G}_C$ denotes a complete DAG. During the update of $\mathbf{T}_0$, the remaining hyperparameters $\boldsymbol{\mu}_0$, α_μ, and α_w are held fixed. Given the current value $\mathbf{T}_0$, we first sample auxiliary group-specific precision matrices from the complete-DAG posterior with the group-specific mean vectors integrated out,

$$\mathbf{W}_{1:H}^\star \sim p(\mathbf{W}_{1:H} \mid \mathbf{T}_0, \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w).$$

Conditional on these auxiliary variables, a proposed value $\mathbf{T}_0^\star$ is drawn from

$$\mathbf{T}_0^\star \sim p(\mathbf{T}_0 \mid \mathbf{W}_{1:H}^\star, \alpha_w).$$

Under the hierarchical model,

$$p(\mathbf{T}_0 \mid \mathbf{W}_{1:H}, \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w) = p(\mathbf{T}_0 \mid \mathbf{W}_{1:H}, \alpha_w),$$

since, conditional on $\mathbf{W}_{1:H}$, the data and the remaining parameters provide no additional information about $\mathbf{T}_0$. Since the auxiliary variables $\mathbf{W}_{1:H}^\star$ are discarded, the resulting proposal density for $\mathbf{T}_0^\star$ is the marginal transition kernel

$$q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0) := \int p(\mathbf{W}_{1:H} \mid \mathbf{T}_0, \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w)\, p(\mathbf{T}_0^\star \mid \mathbf{W}_{1:H}, \alpha_w)\ d\mathbf{W}_{1:H}. \tag{24}$$

Here and in the following, posterior distributions involving $\mathbf{W}_{1:H}$ are understood to be marginalized over the group-specific mean vectors $\boldsymbol{\mu}_{1:H}$.
Define the marginal posterior distribution of $\mathbf{T}_0$ under the complete DAG by

$$\pi_C(\mathbf{T}_0) := p\left(\mathbf{T}_0 \mid \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right).$$

The transition kernel in (24) is reversible with respect to π_C. To see this, consider

$$\begin{aligned}
\pi_C(\mathbf{T}_0) q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0) &= \int \pi_C(\mathbf{T}_0) p\left(\mathbf{W}_{1:H} \mid \mathbf{T}_0, \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right) p\left(\mathbf{T}_0^\star \mid \mathbf{W}_{1:H}, \alpha_w\right)\ d\mathbf{W}_{1:H} \\
&= \int p\left(\mathbf{T}_0, \mathbf{W}_{1:H} \mid \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right) p\left(\mathbf{T}_0^\star \mid \mathbf{W}_{1:H}, \alpha_w\right)\ d\mathbf{W}_{1:H} \\
&= \int p\left(\mathbf{W}_{1:H} \mid \mathcal{D}, \mathcal{G}_C, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right) p\left(\mathbf{T}_0 \mid \mathbf{W}_{1:H}, \alpha_w\right) p\left(\mathbf{T}_0^\star \mid \mathbf{W}_{1:H}, \alpha_w\right)\ d\mathbf{W}_{1:H}.
\end{aligned}$$

Interchanging $\mathbf{T}_0$ and $\mathbf{T}_0^\star$ leaves the integrand unchanged. Consequently,

$$\pi_C(\mathbf{T}_0) q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0) = \pi_C(\mathbf{T}_0^\star) q_C(\mathbf{T}_0 \mid \mathbf{T}_0^\star), \tag{25}$$

and hence

$$\frac{q_C(\mathbf{T}_0 \mid \mathbf{T}_0^\star)}{q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0)} = \frac{\pi_C(\mathbf{T}_0)}{\pi_C(\mathbf{T}_0^\star)}. \tag{26}$$

The desired target distribution conditional on the current DAG $\mathcal{G}$ is

$$\pi_{\mathcal{G}}(\mathbf{T}_0) := p\left(\mathbf{T}_0 \mid \mathcal{D}, \mathcal{G}, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right).$$

Therefore, the Metropolis–Hastings acceptance ratio is

$$R = \frac{\pi_{\mathcal{G}}(\mathbf{T}_0^\star)}{\pi_{\mathcal{G}}(\mathbf{T}_0)} \times \frac{q_C(\mathbf{T}_0 \mid \mathbf{T}_0^\star)}{q_C(\mathbf{T}_0^\star \mid \mathbf{T}_0)}. \tag{27}$$

Using (26), this becomes

$$R = \frac{\pi_{\mathcal{G}}(\mathbf{T}_0^\star)}{\pi_{\mathcal{G}}(\mathbf{T}_0)} \times \frac{\pi_C(\mathbf{T}_0)}{\pi_C(\mathbf{T}_0^\star)}. \tag{28}$$

The two posterior densities satisfy

$$\begin{aligned} \pi_{\mathcal{G}}(\mathbf{T}_0) &\propto p\left(\mathcal{D} \mid \mathcal{G}, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right) \cdot\ p(\mathbf{T}_0), \\ \pi_C(\mathbf{T}_0) &\propto p\left(\mathcal{D} \mid \mathcal{G}_C, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right) \cdot p(\mathbf{T}_0). \end{aligned}$$

Substitution into (28) gives

$$R = \frac{p(\mathbf{T}_0^\star) \cdot p\left(\mathcal{D} \mid \mathcal{G}, \mathbf{T}_0^\star, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)}{p(\mathbf{T}_0) \cdot p\left(\mathcal{D} \mid \mathcal{G}, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)} \times \frac{p(\mathbf{T}_0) \cdot p\left(\mathcal{D} \mid \mathcal{G}_C, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)}{p(\mathbf{T}_0^\star) \cdot p\left(\mathcal{D} \mid \mathcal{G}_C, \mathbf{T}_0^\star, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)}. \tag{29}$$

The prior terms $p(\mathbf{T}_0)$ and $p(\mathbf{T}_0^\star)$ cancel, yielding

$$R = \frac{p\left(\mathcal{D} \mid \mathcal{G}, \mathbf{T}_0^\star, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)}{p\left(\mathcal{D} \mid \mathcal{G}_C, \mathbf{T}_0^\star, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)} \times \frac{p\left(\mathcal{D} \mid \mathcal{G}_C, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)}{p\left(\mathcal{D} \mid \mathcal{G}, \mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\right)}. \tag{30}$$

The proposed value $\mathbf{T}_0^\star$ is accepted with probability $A = \min\{1, R\}$.

Appendix C. Hastings-Ratio for Sampling the Global Mean Vector

Recall that $\mathbf{W}_{1:H} := \{\mathbf{W}_1, \ldots, \mathbf{W}_H\}$, $\boldsymbol{\mu}_{1:H} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H\}$, and let $\mathcal{G}_C$ denote a complete DAG. During the update of $\boldsymbol{\mu}_0$, the remaining hyperparameters $\mathbf{T}_0$, α_μ, and α_w are held fixed. Given the current $\boldsymbol{\mu}_0$, we first sample auxiliary group-specific precision matrices from

$$\mathbf{W}_{1:H}^\star \sim p\left(\mathbf{W}_{1:H} \mid \boldsymbol{\mu}_0, \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right),$$

where the group-specific mean vectors have been integrated out. Conditional on $\mathbf{W}_{1:H}^\star$, auxiliary group-specific mean vectors are sampled from

$$\boldsymbol{\mu}_{1:H}^\star \sim p\left(\boldsymbol{\mu}_{1:H} \mid \mathbf{W}_{1:H}^\star, \boldsymbol{\mu}_0, \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right).$$

Since the first distribution is marginalized over $\boldsymbol{\mu}_{1:H}$, the two sequential sampling steps jointly define a draw from

$$p(\mathbf{W}_{1:H}, \boldsymbol{\mu}_{1:H} \mid \boldsymbol{\mu}_0, \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w)\,.$$

Conditional on these auxiliary variables, a proposed global mean vector is drawn from

$$\boldsymbol{\mu}_0^\star \sim p\left(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_{1:H}^\star, \mathbf{W}_{1:H}^\star, \alpha_\mu\right).$$

Under the complete-DAG hierarchical model, the full conditional distribution satisfies

$$p\left(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_{1:H}, \mathbf{W}_{1:H}, \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) = p\left(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_{1:H}, \mathbf{W}_{1:H}, \alpha_\mu\right).$$

Since the auxiliary variables $\mathbf{W}_{1:H}^\star$ and $\boldsymbol{\mu}_{1:H}^\star$ are discarded, the resulting proposal density for $\boldsymbol{\mu}_0^\star$ is the marginal transition kernel

$$q_C(\boldsymbol{\mu}_0^\star \mid \boldsymbol{\mu}_0) := \int p\left(\mathbf{W}_{1:H}, \boldsymbol{\mu}_{1:H} \mid \boldsymbol{\mu}_0, \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) \times p\left(\boldsymbol{\mu}_0^\star \mid \boldsymbol{\mu}_{1:H}, \mathbf{W}_{1:H}, \alpha_\mu\right) \; d\boldsymbol{\mu}_{1:H}\, d\mathbf{W}_{1:H}.$$

Define the marginal posterior distribution of $\boldsymbol{\mu}_0$ under the complete DAG by

$$\pi_C(\boldsymbol{\mu}_0) := p\left(\boldsymbol{\mu}_0 \mid \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right).$$

For brevity, we also define $\tilde{p}(\boldsymbol{\mu}_0) := p\left(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_{1:H}, \mathbf{W}_{1:H}, \alpha_\mu\right)$.

The transition kernel is reversible with respect to π_C. Indeed,

$$\begin{aligned}
\pi_C(\boldsymbol{\mu}_0) q_C(\boldsymbol{\mu}_0^\star \mid \boldsymbol{\mu}_0) &= \int \pi_C(\boldsymbol{\mu}_0) p\left(\mathbf{W}_{1:H}, \boldsymbol{\mu}_{1:H} \mid \boldsymbol{\mu}_0, \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) \times \tilde{p}(\boldsymbol{\mu}_0^\star) \; d\boldsymbol{\mu}_{1:H}\, d\mathbf{W}_{1:H} \\
&= \int p\left(\boldsymbol{\mu}_0, \mathbf{W}_{1:H}, \boldsymbol{\mu}_{1:H} \mid \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) \times \tilde{p}(\boldsymbol{\mu}_0^\star) \; d\boldsymbol{\mu}_{1:H}\, d\mathbf{W}_{1:H} \\
&= \int p\left(\mathbf{W}_{1:H}, \boldsymbol{\mu}_{1:H} \mid \mathcal{D}, \mathcal{G}_C, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) \times \tilde{p}(\boldsymbol{\mu}_0) \times \tilde{p}(\boldsymbol{\mu}_0^\star) \; d\boldsymbol{\mu}_{1:H}\, d\mathbf{W}_{1:H}.
\end{aligned}$$

The final integrand is symmetric in $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_0^\star$. Consequently,

$$\pi_C(\boldsymbol{\mu}_0) q_C(\boldsymbol{\mu}_0^\star \mid \boldsymbol{\mu}_0) = \pi_C(\boldsymbol{\mu}_0^\star) q_C(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_0^\star), \tag{31}$$

and hence

$$\frac{q_C(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_0^\star)}{q_C(\boldsymbol{\mu}_0^\star \mid \boldsymbol{\mu}_0)} = \frac{\pi_C(\boldsymbol{\mu}_0)}{\pi_C(\boldsymbol{\mu}_0^\star)}. \tag{32}$$

The desired target distribution conditional on the current DAG $\mathcal{G}$ is

$$\pi_{\mathcal{G}}(\boldsymbol{\mu}_0) := p\left(\boldsymbol{\mu}_0 \mid \mathcal{D}, \mathcal{G}, \mathbf{T}_0, \alpha_\mu, \alpha_w\right).$$

Therefore, the Metropolis–Hastings acceptance ratio is

$$R = \frac{\pi_{\mathcal{G}}(\boldsymbol{\mu}_0^\star)}{\pi_{\mathcal{G}}(\boldsymbol{\mu}_0)} \times \frac{q_C(\boldsymbol{\mu}_0 \mid \boldsymbol{\mu}_0^\star)}{q_C(\boldsymbol{\mu}_0^\star \mid \boldsymbol{\mu}_0)} = \frac{\pi_{\mathcal{G}}(\boldsymbol{\mu}_0^\star)}{\pi_{\mathcal{G}}(\boldsymbol{\mu}_0)} \times \frac{\pi_C(\boldsymbol{\mu}_0)}{\pi_C(\boldsymbol{\mu}_0^\star)}. \tag{33}$$

Since the posterior densities satisfy

$$\pi_{\mathcal{G}}(\boldsymbol{\mu}_0) \propto p\left(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\mu}_0, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) \cdot p(\boldsymbol{\mu}_0), \tag{34}$$

$$\pi_C(\boldsymbol{\mu}_0) \propto p\left(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\mu}_0, \mathbf{T}_0, \alpha_\mu, \alpha_w\right) \cdot p(\boldsymbol{\mu}_0). \tag{35}$$

we obtain

$$R = \frac{p(\boldsymbol{\mu}_0^\star) \cdot p\left(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\mu}_0^\star, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)}{p(\boldsymbol{\mu}_0) \cdot p\left(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\mu}_0, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)} \times \frac{p(\boldsymbol{\mu}_0) \cdot p\left(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\mu}_0, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)}{p(\boldsymbol{\mu}_0^\star) \cdot p\left(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\mu}_0^\star, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)}.$$

The prior terms $p(\boldsymbol{\mu}_0)$ and $p(\boldsymbol{\mu}_0^\star)$ cancel, yielding

$$R = \frac{p\left(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\mu}_0^\star, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)}{p\left(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\mu}_0^\star, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)} \times \frac{p\left(\mathcal{D} \mid \mathcal{G}_C, \boldsymbol{\mu}_0, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)}{p\left(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\mu}_0, \mathbf{T}_0, \alpha_\mu, \alpha_w\right)}. \tag{36}$$

The proposed value $\boldsymbol{\mu}_0^\star$ is accepted with probability $A = \min\{1, R\}$.

Appendix D. Metropolis-Hastings Sampling of the Coupling Strengths

For the two coupling strength scalar parameters $\alpha_w > n+1$ and $\alpha_\mu > 0$ we also implement MH sampling steps. We define $\tilde{\alpha}_w = \alpha_w - n - 1$ and impose the Gamma prior distributions from Section 2.6. Given the current parameters $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\}$, a new candidate value $\alpha_w^\star$ for α_w is proposed. We sample the new values as follows:

$$\tilde{\alpha}_w^\star = \tilde{\alpha}_w + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, s_w^2),$$

and propose $\alpha_w^\star = \tilde{\alpha}_w^\star + n + 1$. As the proposal distribution is symmetric, no Hastings correction term appears and we accept $\alpha_w^\star$ with probability

$$A = \min\left\{1, \frac{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star) \cdot p(\tilde{\alpha}_w^\star)}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}) \cdot p(\tilde{\alpha}_w)}\right\},$$

where $\boldsymbol{\Theta}^\star = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w^\star\}$, and $p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})$ and $p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star)$ denote marginal likelihoods. If $\tilde{\alpha_w}^\star \leq 0$ (equivalently, $\alpha_w^\star \leq n+1$), the proposal is automatically rejected because the Gamma prior assigns zero probability outside its support.

Accordingly, given $\boldsymbol{\Theta} = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu, \alpha_w\}$, a new candidate value $\alpha_\mu^\star$ for α_μ is proposed by sampling

$$\alpha_\mu^\star = |\alpha_\mu + \varepsilon|, \qquad \varepsilon \sim \mathcal{N}(0, s_\mu^2).$$

The absolute-value transformation yields a symmetric folded Gaussian random-walk proposal, so no Hastings correction term is required. We accept $\alpha_\mu^\star$ with probability

$$A = \min\left\{1, \frac{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star) \cdot p(\alpha_\mu^\star)}{p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}) \cdot p(\alpha_\mu)}\right\},$$

where $\boldsymbol{\Theta}^\star = \{\mathbf{T}_0, \boldsymbol{\mu}_0, \alpha_\mu^\star, \alpha_w\}$, and $p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta}^\star)$ and $p(\mathcal{D} \mid \mathcal{G}, \boldsymbol{\Theta})$ denote marginal likelihoods. For simplicity, both MH updates use the same proposal standard deviation $s_w = s_\mu = 0.1$.

Appendix E. Posterior Sampling of Group-Specific GBN Parameters

For computing the posterior predictive probabilities in Section 2.8, posterior samples of the Gaussian parameters are required conditional on a sampled DAG. The key idea is to independently sample one local conditional distribution for each node from the complete-DAG

posterior and then assemble these local conditional distributions into a globally consistent Gaussian Bayesian network by recursively reconstructing the covariance matrix and the mean vector.
Suppose that a DAG $\mathcal{G}$ and hyperparameter values $\boldsymbol{\Theta}$ are given. Let $\mathcal{G}_C$ denote a complete DAG that is consistent with the topological ordering of $\mathcal{G}$. The following procedure generates one posterior sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. Generate n independent joint posterior samples from the complete-DAG posterior,
$$(\boldsymbol{\mu}_C^{(i)}, \mathbf{W}_C^{(i)}) \sim p(\boldsymbol{\mu}, \mathbf{W} \mid \mathcal{D}, \mathcal{G}_C, \boldsymbol{\Theta}), \qquad (i = 1, \ldots, n).$$

2. For $i = 1, \ldots, n$ use the i-th draw $(\boldsymbol{\mu}_C^{(i)}, \mathbf{W}_C^{(i)})$ to compute the conditional distribution
$$X_i \mid \mathbf{X}_{\Pi_i} \sim \mathcal{N}\left(m_i + \boldsymbol{\beta}_i^\top \mathbf{X}_{\Pi_i}, v_i\right),$$
where $\mathbf{X}_{\Pi_i}$ denotes the vector of random variables corresponding to the parents of X_i. The regression coefficients $\boldsymbol{\beta}_i$ and conditional variance v_i are computed from the corresponding submatrices of
$$\boldsymbol{\Sigma}_C^{(i)} = (\mathbf{W}_C^{(i)})^{-1}.$$
The corresponding intercept is $m_i = (\boldsymbol{\mu}_C^{(i)})_i - \boldsymbol{\beta}_i^\top (\boldsymbol{\mu}_C^{(i)})_{\boldsymbol{\Pi}_i}$, where $(\boldsymbol{\mu}_C^{(i)})_i$ and $(\boldsymbol{\mu}_C^{(i)})_{\boldsymbol{\Pi}_i}$ originate from the i-th complete-DAG posterior draw $\boldsymbol{\mu}_C^{(i)}$.

3. Reconstruct the covariance matrix $\boldsymbol{\Sigma}$ (or equivalently the precision matrix $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$) from the sampled collection of regression coefficients and conditional variances obtained in Step 2, using the recursive construction of Shachter and Kenley (1989).

4. Construct the mean vector $\boldsymbol{\mu}$ recursively according to the topological ordering implied by $\mathcal{G}$. For every root node,
$$\boldsymbol{\mu}_i := m_i = (\boldsymbol{\mu}_C^{(i)})_i.$$
For every non-root node,
$$\boldsymbol{\mu}_i := m_i + \boldsymbol{\beta}_i^\top \boldsymbol{\mu}_{\boldsymbol{\Pi}_i} = (\boldsymbol{\mu}_C^{(i)})_i + \boldsymbol{\beta}_i^\top \left(\boldsymbol{\mu}_{\boldsymbol{\Pi}_i} - (\boldsymbol{\mu}_C^{(i)})_{\boldsymbol{\Pi}_i}\right),$$
where $\boldsymbol{\mu}_{\boldsymbol{\Pi}_i}$ denotes the recursively constructed means μ_j of the parent nodes $X_j \in \boldsymbol{\Pi}_i$.
Notice that the parent means $\boldsymbol{\mu}_{\boldsymbol{\Pi}_i}$ differ from $(\boldsymbol{\mu}_C^{(i)})_{\boldsymbol{\Pi}_i}$. $(\boldsymbol{\mu}_C^{(i)})_{\boldsymbol{\Pi}_i}$ is a subvector of the i-th complete-DAG posterior draw $(\boldsymbol{\mu}_C^{(i)})$. The elements of $\boldsymbol{\mu}_{\boldsymbol{\Pi}_i}$ are the recursively constructed parent means $\boldsymbol{\mu}_j$ $(X_j \in \boldsymbol{\Pi}_i)$, which are implied by independently sampled local conditional distributions and need to be constructed in topological order.

In our application, this generic procedure is applied to every posterior sample $(\mathcal{G}^{(r)}, \boldsymbol{\Theta}^{(r)})$ and every group $h = 1, \ldots, H$ using the group-specific data $\mathcal{D}_h$. Thus, the generic parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ become $(\boldsymbol{\mu}_h^{(r)}, \boldsymbol{\Sigma}_h^{(r)})$, while the independently generated complete-DAG posterior draws are denoted by $(\boldsymbol{\mu}_{C,h}^{(r,i)}, \mathbf{W}_{C,h}^{(r,i)})$. Repeating the above procedure for all groups and all MCMC iterations yields the posterior samples
$$(\boldsymbol{\mu}_1^{(r)}, \boldsymbol{\Sigma}_1^{(r)}), \ldots, (\boldsymbol{\mu}_H^{(r)}, \boldsymbol{\Sigma}_H^{(r)}), \qquad (r = 1, \ldots, R).$$