Received: 00 0000
DOI: 10.1111/stan.12136

SPECIAL ISSUE ARTICLE

WILEY

Improving nonhomogeneous dynamic Bayesian networks with sequentially coupled parameters

Mahdi Shafiee Kamalabad | Marco Grzegorczyk

Johann Bernoulli Institute, Groningen University, 9747 AG Groningen, Netherlands

Correspondence
Marco Grzegorczyk, Johann Bernoulli Institute, Groningen University, 9747 AG Groningen, Netherlands.
Email: m.a.grzegorczyk@rug.nl

In systems biology, nonhomogeneous dynamic Bayesian networks (NH-DBNs) have become a popular modeling tool for reconstructing cellular regulatory networks from postgenomic data. In this paper, we focus our attention on NH-DBNs that are based on Bayesian piecewise linear regression models. The new NH-DBN model, proposed here, is a generalization of an earlier proposed model with sequentially coupled network interaction parameters. Unlike the original model, our novel model possesses segment-specific coupling parameters, so that the coupling strengths between parameters can vary over time. Thereby, to avoid model overflexibility and to allow for some information exchange among time segments, we globally couple the segment-specific coupling (strength) parameters by a hyperprior. Our empirical results on synthetic and on real biological network data show that the new model yields better network reconstruction accuracies than the original model.

KEYWORDS
Bayesian modeling, dynamic Bayesian network, network reconstruction, parameter coupling, sequential coupling, systems biology

1 | INTRODUCTION

Dynamic Bayesian networks (DBNs) are a popular class of models for learning the dependencies between random variables from temporal data. In DBNs, a dependency between two variables X and Y is typically interpreted in terms of a regulatory interaction with a time delay. A directed edge from variable X to variable Y, symbolically $X \rightarrow Y$, indicates that the value of variable Y at

any time point t depends on the realization of X at the previous time point $t-1$. Typically, various variables $X_1, \ldots, X_k$ have a regulatory effect on Y, and the relationship between $X_1, \ldots, X_k$ and Y can be represented by a regression model that takes the time delay into account. For example, if the time delay is one time point,

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \cdots + \beta_k x_{k,t-1} + u_t \quad (t = 2, \ldots, T),$$

where T is the number of time points, y_t is the value of Y at time point t, $x_{i,t-1}$ is the value of covariate X_i at time point $t-1$, $\beta_0, \ldots, \beta_k$ are regression coefficients, and u_t is the "unexplained" noise at time point t.

In DBN applications, there are N domain variables $Y_1, \ldots, Y_N$, and the goal is to infer the covariates of each variable Y_i. As the covariates can be learned for each Y_i separately, DBN learning can be thought of as learning the covariates for a set of target variables $\{Y_1, \ldots, Y_N\}$. There are N regression tasks, and in the ith regression model, Y_i is the target variable, and the remaining $N-1$ variables take the role of the potential covariates. The goal is to infer a covariate set $\pi_i \subset \{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_N\}$ for each Y_i.

From the covariate sets $\pi_1, \ldots, \pi_N$, a network can be extracted. The network shows all regulatory interactions among the variables $Y_1, \ldots, Y_N$. An edge $Y_j \to Y_i$ indicates that Y_j is a covariate of Y_i, that is, $Y_j \in \pi_i$. In the terminology of DBNs, Y_j is then called a regulator of Y_i. All variables in π_i are regulators of Y_i $(i = 1, \ldots, N)$.

The traditional (homogeneous) DBN is based on the assumption that the regression coefficients $\beta_0, \ldots, \beta_K$ stay constant across all time points $(t = 2, \ldots, T)$. Especially in systems biology, where one important application is to learn gene regulatory networks from gene expression time series, this assumption is often not appropriate. Many gene regulatory processes are subject to temporal changes, for example, implied by cellular or experimental conditions.

It has therefore been proposed by Lébre, Becq, Devaux, Lelandais, and Stumpf (2010) to replace the linear model by a piecewise linear regression model, where the data points are divided into H temporal segments (by changepoints). The data points within each segment h are modeled by a linear model with segment-specific regression coefficients $\beta_{h,0}, \ldots, \beta_{h,K}$ $(h = 1, \ldots, H)$. If the segmentation is not known, it can be inferred from the data (e.g., by employing a changepoint process). Replacing the linear model by a piecewise linear model yields a nonhomogeneous DBN (NH-DBN) model.

The problem of the piecewise linear regression approach is that the regression coefficients have to be learned separately for each segment. As the available time series are usually rather short, this approach comes with inflated inference uncertainties. A short time series is divided into shorter segments, and for each of the segments, the regression coefficients have to be learned. To avoid model overflexibility, two parameter coupling schemes were proposed: The segment-specific regression coefficients can be coupled globally (Grzegorczyk & Husmeier, 2013) or sequentially (Grzegorczyk & Husmeier, 2012). In this paper, we put our focus on the sequential coupling scheme and show how it can be improved. In the model of Grzegorczyk and Husmeier (2012), the coefficients of any segment are encouraged to be similar to those of the previous segment by setting the prior expectations of the coefficients in segment $h+1$ to the posterior expectations of the coefficients from segment h. A coupling parameter λ_c regulates the coupling strength (i.e., the similarity of the regression coefficients). A drawback of the model is that all pairs of neighboring segments $(h-1, h)$ are coupled with the same coupling parameter λ_c and thus with the same strength.

We present a new and improved sequentially coupled model that addresses this bottleneck. Our generalized sequentially coupled model possesses segment-specific coupling parameters. For

each pair of neighboring segments $(h-1, h)$, there is then a segment-specific continuous coupling strength λ_h $(h = 2, \ldots, H)$. The segment-specific coupling parameter λ_h has to be inferred from the data points within segment h $(h = 1, \ldots, H)$. Because some segments might be rather short and uninformative, we impose a hyperprior onto the second hyperparameter of the coupling parameter prior. This allows for information exchange among the segment-specific coupling strengths $\lambda_2, \ldots, \lambda_H$.

2 | MATHEMATICAL DETAILS

We consider piecewise linear regression models where the random variable Y is the target and the random variables $X_1, \ldots, X_k$ are the covariates. We assume that T data points $\mathcal{D}_1, \ldots, \mathcal{D}_T$ are available and that the subscript index $t \in \{1, \ldots, T\}$ refers to T equidistant time points. Each data point $\mathcal{D}_t$ contains a value of the target Y and the corresponding values of the k covariates. We assume further that the T data points are allocated to H disjunct segments, $h \in \{1, \ldots, H\}$. Segment h contains T_h consecutive data points with $\sum_{h=1}^{H} T_h = T$. Within each individual segment h, we apply a Bayesian linear regression model with a segment-specific regression coefficient vector $\boldsymbol{\beta}_h = (\beta_{h,0}, \ldots, \beta_{h,k})^{\mathsf{T}}$. Let $\mathbf{y}_h$ be the target vector of length T_h and let $\mathbf{X}_h$ be the T_h-by-$(k+1)$ design matrix for segment h, which includes a first column of ones for the intercept. For each segment $h = 1, \ldots, H$, we assume a Gaussian likelihood,

$$\mathbf{y}_h | (\boldsymbol{\beta}_h, \sigma^2) \sim \mathcal{N}(\mathbf{X}_h \boldsymbol{\beta}_h, \sigma^2 \mathbf{I}), \tag{1}$$

where $\mathbf{I}$ denotes the T_h-by-T_h identity matrix, and σ^2 is the noise variance parameter, which is shared among segments. We impose an inverse gamma prior on σ^2, $\sigma^{-2} \sim \text{GAM}(\alpha_\sigma, \beta_\sigma)$. In the forthcoming subsections, we will present different model instantiations with different prior distributions for the segment-specific regression coefficient vectors $\boldsymbol{\beta}_h$ $(h = 1, \ldots, H)$.

2.1 | The original sequential coupling scheme

In the sequentially coupled piecewise linear regression model, proposed by Grzegorczyk and Husmeier (2012), it is assumed that the regression coefficient vectors $\boldsymbol{\beta}_h$ have the following Gaussian prior distributions:

$$\boldsymbol{\beta}_h | \left(\sigma^2, \lambda_u, \lambda_c\right) \sim \begin{cases} \mathcal{N}\left(\mathbf{0}, \sigma^2 \lambda_u \mathbf{I}\right) & \text{if } h = 1 \\ \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{h-1}, \sigma^2 \lambda_c \mathbf{I}\right) & \text{if } h > 1 \end{cases} \quad (h = 1, \ldots, H), \tag{2}$$

where $\mathbf{0}$ is the zero vector of length $k+1$, $\mathbf{I}$ denotes the $(k+1)$-by-$(k+1)$ identity matrix, and $\tilde{\boldsymbol{\beta}}_{h-1}$ is the posterior expectation of $\boldsymbol{\beta}_{h-1}$. That is, only the first segment $h = 1$ gets an uninformative prior expectation, namely the zero vector, whereas the subsequent segments $h > 1$ obtain informative prior expectations, stemming from the preceding segment $h-1$. We follow Grzegorczyk and Husmeier (2012) and refer to $\lambda_u \in \mathbb{R}^+$ as the signal-to-noise ratio parameter and to $\lambda_c \in \mathbb{R}^+$ as the coupling parameter. A low (high) signal-to-noise ratio parameter λ_u yields a peaked (vague) prior for $h = 1$ in Equation (2), and thus, the distribution of $\boldsymbol{\beta}_1$ is peaked (diffuse) around the zero vector. A low (high) coupling parameter λ_c yields a peaked (vague) prior for $h > 1$ in Equation (2) and thus a strong (weak) coupling of $\boldsymbol{\beta}_h$ to the posterior expectation $\tilde{\boldsymbol{\beta}}_{h-1}$ from the preceding segment. We note that reemploying the variance parameter σ^2 in Equation (2) yields a fully conjugate prior in both groups of parameters $\boldsymbol{\beta}_h$ $(h = 1, \ldots, H)$ and σ^2 (see, e.g., Sections 3.3 and 3.4; Gelman, Carlin, Stern, & Rubin, 2004) with the marginal likelihood given below in Equation (10).

The posterior distribution of $\boldsymbol{\beta}_h$ ($h = 1, \ldots, H$) can be computed in closed form (Grzegorczyk & Husmeier, 2012), as follows:

$$\boldsymbol{\beta}_h | \left(\mathbf{y}_h, \sigma^2, \lambda_u, \lambda_c\right) \sim \begin{cases} \mathcal{N}\left(\left[\lambda_u^{-1}\mathbf{I} + \mathbf{X}_1^\top \mathbf{X}_1\right]^{-1} \mathbf{X}_1^\top \mathbf{y}_1, \sigma^2\left(\lambda_u^{-1}\mathbf{I} + \mathbf{X}_1^\top \mathbf{X}_1\right)^{-1}\right) & \text{if } h = 1 \\ \mathcal{N}\left(\left[\lambda_c^{-1}\mathbf{I} + \mathbf{X}_h^\top \mathbf{X}_h\right]^{-1} \left(\lambda_c^{-1}\tilde{\boldsymbol{\beta}}_{h-1} + \mathbf{X}_h^\top \mathbf{y}_h\right), \sigma^2\left(\lambda_c^{-1}\mathbf{I} + \mathbf{X}_h^\top \mathbf{X}_h\right)^{-1}\right) & \text{if } h \geq 2, \end{cases} \tag{3}$$

and the posterior expectations in Equation (3) are the prior expectations used in Equation (2), as follows:

$$\tilde{\boldsymbol{\beta}}_{h-1} := \begin{cases} \left[\lambda_u^{-1}\mathbf{I} + \mathbf{X}_1^\top \mathbf{X}_1\right]^{-1} \mathbf{X}_1^\top \mathbf{y}_1 & \text{if } h = 2 \\ \left[\lambda_c^{-1}\mathbf{I} + \mathbf{X}_{h-1}^\top \mathbf{X}_{h-1}\right]^{-1} \left(\lambda_c^{-1}\tilde{\boldsymbol{\beta}}_{h-2} + \mathbf{X}_{h-1}^\top \mathbf{y}_{h-1}\right) & \text{if } h \geq 3. \end{cases} \tag{4}$$

Grzegorczyk and Husmeier (2012) assigned inverse gamma priors to the parameters λ_u and λ_c, as follows:

$$\lambda_u^{-1} \sim \text{GAM}(\alpha_u, \beta_u) \tag{5}$$

$$\lambda_c^{-1} \sim \text{GAM}(\alpha_c, \beta_c). \tag{6}$$

The fully sequentially coupled model is then fully specified, and we will refer to it as the $\mathcal{M}_{0,0}$ model. A graphical model representation for the relationships in the first segment $h = 1$ is provided in Figure 1, whereas Figure 2 shows a graphical model representation for the segments $h > 1$. The posterior distribution of the $\mathcal{M}_{0,0}$ model fulfills

$$\begin{aligned} p\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \lambda_c | \mathbf{y}_1, \ldots, \mathbf{y}_H\right) &\propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \lambda_c\right) \\ &\propto \prod_{h=1}^{H} p\left(\mathbf{y}_h | \sigma^2, \boldsymbol{\beta}_h\right) \cdot p\left(\boldsymbol{\beta}_1 | \sigma^2, \lambda_u\right) \cdot \prod_{h=2}^{H} p\left(\boldsymbol{\beta}_h | \sigma^2, \lambda_c\right) \cdot p(\sigma^2) \cdot p(\lambda_u) \cdot p(\lambda_c). \end{aligned} \tag{7}$$

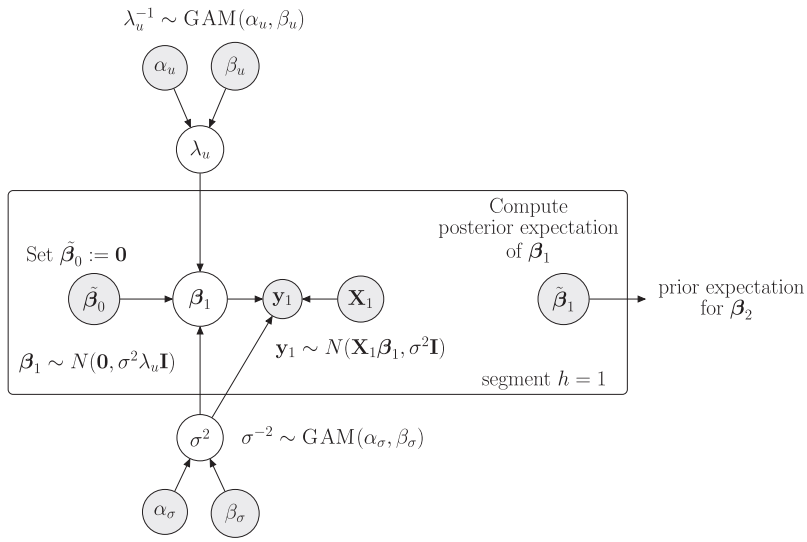

FIGURE 1 Graphical model of the probabilistic relationships in the first segment, $h = 1$. Parameters that have to be inferred are represented by white circles. The observed data ($\mathbf{y}_1$ and $\mathbf{X}_1$) and the fixed hyperparameters are represented by gray circles. All nodes in the plate are specific for the first segment. The posterior expectation $\tilde{\boldsymbol{\beta}}_1$ is computed and then treated like a fixed hyperparameter vector when used as input for segment $h = 2$

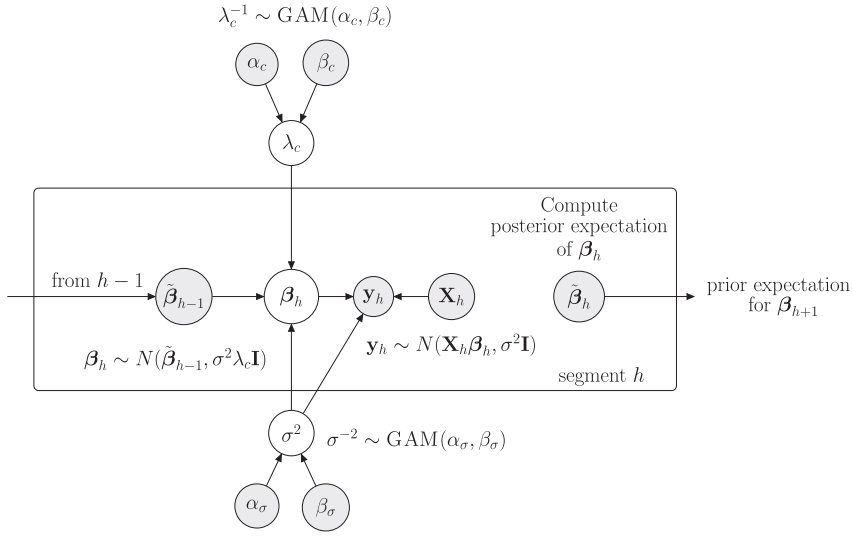

FIGURE 2 Graphical model of the probabilistic relationships within and between segments $h > 1$ for the $\mathcal{M}_{0,0}$ model from Grzegorczyk and Husmeier (2012). Parameters that have to be inferred are represented by white circles. The observed data and the fixed hyperparameters are represented by gray circles. All nodes in the plate are specific for segment h. The posterior expectation $\tilde{\boldsymbol{\beta}}_{h-1}$ of the regression coefficient vector from the previous segment $h-1$ is treated like a fixed hyperparameter vector. The posterior expectation $\tilde{\boldsymbol{\beta}}_h$ is computed and forwarded as a fixed hyperparameter vector to the subsequent segment $h+1$

Like the regression coefficient vectors $\boldsymbol{\beta}_h$, whose full conditional distributions have been provided in Equation (3), the parameters λ_u and λ_c can also be sampled from their full conditional distributions.

$$\begin{aligned}\lambda_c^{-1}|\left(\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\boldsymbol{\beta}_1,\ \ldots,\boldsymbol{\beta}_H,\sigma^2,\lambda_u,\lambda_c\right) &\sim \mathrm{GAM}\\ &\left(\alpha_c+\tfrac{(H-1)\cdot(k+1)}{2},\beta_c+\tfrac{1}{2}\sigma^{-2}\cdot\sum_{h=2}^{H}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})\right)\\ \lambda_u^{-1}|\left(\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\boldsymbol{\beta}_1,\ \ldots,\boldsymbol{\beta}_H,\sigma^2,\lambda_u,\lambda_c\right) &\sim \mathrm{GAM}\left(\alpha_u+\tfrac{1\cdot(k+1)}{2},\beta_u+\tfrac{1}{2}\sigma^{-2}\cdot\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1\right).\end{aligned} \tag{8}$$

For the parameter σ^2, a collapsed Gibbs sampling step, with $\boldsymbol{\beta}_h$ ($h = 1, \ldots, H$) integrated out, can be used, as follows:

$$\begin{aligned}\sigma^{-2}|\left(\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\lambda_u,\lambda_c\right) &\sim \mathrm{GAM}\left(\alpha_\sigma+\frac{1}{2}\cdot T,\beta_\sigma+\frac{1}{2}\cdot\sum_{h=1}^{H}(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}\mathbf{C}_h^{-1}(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1})\right),\\ \text{where}\ \ \tilde{\boldsymbol{\beta}}_0 &:= \mathbf{0}\ \ \text{and}\ \ \mathbf{C}_h := \begin{cases}\mathbf{I}+\lambda_u\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}} & \text{if}\ \ h=1\\ \mathbf{I}+\lambda_c\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}} & \text{if}\ \ h>1\end{cases}.\end{aligned} \tag{9}$$

For the marginal likelihood, with $\boldsymbol{\beta}_h$ ($h = 1, \ldots, H$) and σ^2 integrated out, the marginalization rule from section 2.3.7 of Bishop (2006) can be applied:

$$\begin{aligned}p(\mathbf{y}_1,\ \ldots,\mathbf{y}_H|\lambda_u,\lambda_c) = \frac{\Gamma\left(\frac{T}{2}+a_\sigma\right)}{\Gamma(a_\sigma)}\cdot\frac{\pi^{-T/2}\cdot(2b_\sigma)^{a_\sigma}}{\left(\prod_{h=1}^{H}\det(\mathbf{C}_h)\right)^{1/2}}\\ \cdot\left(2b_\sigma+\sum_{h=1}^{H}(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}\mathbf{C}_h^{-1}(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1})\right)^{-\left(\frac{T}{2}+a_\sigma\right)},\end{aligned} \tag{10}$$

where the matrices $\mathbf{C}_h$ were defined below Equation (9). For the derivations of the full conditional distributions in Equations 8 and 9 and the marginal likelihood in Equation (10), we refer to Grzegorczyk and Husmeier (2012).

2.2 | The improved sequential coupling scheme proposed here

We propose to generalize the sequentially coupled model from Section 2.1 by introducing segment-specific coupling parameters λ_h ($h = 2, \ldots, H$). This yields the new prior distributions,

$$\boldsymbol{\beta}_h | \left(\sigma^2, \lambda_u, \lambda_2, \ldots, \lambda_H\right) \sim \begin{cases} \mathcal{N}\left(\mathbf{0}, \sigma^2 \lambda_u \mathbf{I}\right) & \text{if } h = 1 \\ \mathcal{N}\left(\tilde{\boldsymbol{\beta}}_{h-1}, \sigma^2 \lambda_h \mathbf{I}\right) & \text{if } h > 1, \end{cases} \tag{11}$$

where $\tilde{\boldsymbol{\beta}}_{h-1}$ is again the posterior expectation of $\boldsymbol{\beta}_{h-1}$. For notational convenience, we now introduce two new definitions, namely $\lambda_1 := \lambda_u$ and $\tilde{\boldsymbol{\beta}}_0 := \mathbf{0}$. We can then compactly write: For $h = 2, \ldots, H$,

$$\tilde{\boldsymbol{\beta}}_{h-1} := \left[\lambda_{h-1}^{-1}\mathbf{I} + \mathbf{X}_{h-1}^{\mathsf{T}}\mathbf{X}_{h-1}\right]^{-1}\left(\lambda_{h-1}^{-1}\tilde{\boldsymbol{\beta}}_{h-2} + \mathbf{X}_{h-1}^{\mathsf{T}}\mathbf{y}_{h-1}\right). \tag{12}$$

We show in the next subsection that $\tilde{\boldsymbol{\beta}}_{h-1}$, defined in Equation (12), is the posterior expectation of $\boldsymbol{\beta}_{h-1}$; compare Equation (14). For the parameter λ_u, we reuse the inverse gamma prior with hyperparameters α_u and β_u. For the first segment $h = 1$, we thus have the same probabilistic relationships for the original model; compare the graphical model representation in Figure 1. For the parameters λ_h, we assume that they are inverse gamma distributed, $\lambda_h^{-1} \sim \text{GAM}(\alpha_c, \beta_c)$, where α_c is fixed and β_c is a free hyperparameter. A free β_c allows for information exchange among the segments $h = 2, \ldots, H$. We impose a gamma hyperprior onto β_c, symbolically $\beta_c \sim \text{GAM}(a,b)$.

We refer to the improved model as the $\mathcal{M}_{1,1}$ model. A graphical model representation of the relationships within and between segments $h > 1$ is provided in Figure 3. The posterior of the $\mathcal{M}_{1,1}$ model is

$$\begin{aligned} p\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \lambda_2, \ldots, \lambda_H, \beta_c | \mathbf{y}_1, \ldots, \mathbf{y}_H\right) &\propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \lambda_2, \ldots, \lambda_H, \beta_c\right) \\ &\propto \prod_{h=1}^{H} p\left(\mathbf{y}_h | \sigma^2, \boldsymbol{\beta}_h\right) \cdot p\left(\boldsymbol{\beta}_1 | \sigma^2, \lambda_u\right) \cdot \prod_{h=2}^{H} p\left(\boldsymbol{\beta}_h | \sigma^2, \lambda_h\right) \\ &\quad \cdot p(\sigma^2) \cdot p(\lambda_u) \cdot \prod_{h=2}^{H} p(\lambda_h | \beta_c) \cdot p(\beta_c). \end{aligned} \tag{13}$$

2.3 | Full conditional distributions of the improved sequentially coupled model

In this subsection, we derive the full conditional distributions for the $\mathcal{M}_{1,1}$ model, proposed in Section 2.2. For the derivations, we exploit that the full conditional densities are proportional to the posterior density and thus proportional to the factorized joint density in Equation (13). From the shape of the densities, we can conclude what the full conditional distributions are. For notational convenience, let $\{\lambda_h\}_{h\geq 2}$ denote the set of coupling parameters $\lambda_2, \ldots, \lambda_H$ and let $\{\lambda_k\}_{k\neq h}$ denote the set of coupling parameters $\lambda_2, \ldots, \lambda_{h-1}, \lambda_{h+1}, \ldots, \lambda_H$ with parameter λ_h left out.

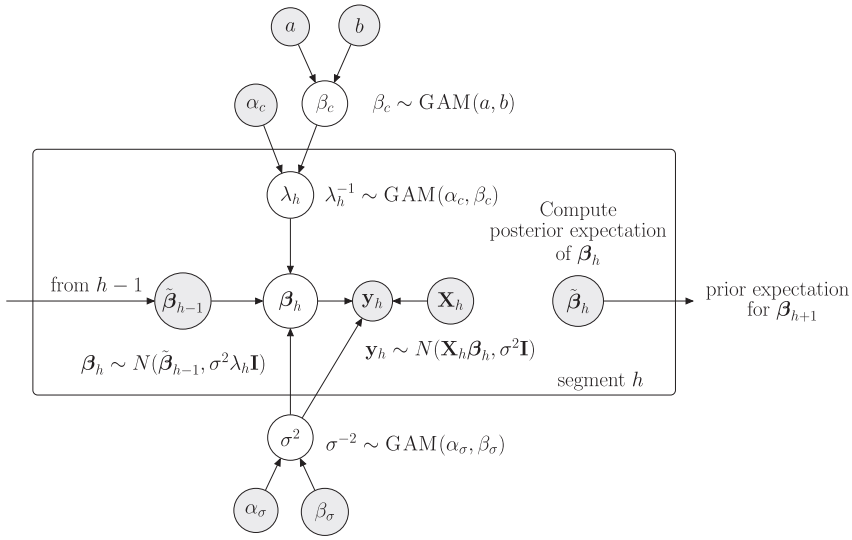

FIGURE 3 Graphical model of the probabilistic relationships within and between segments $h > 1$ for the proposed $\mathcal{M}_{1,1}$ model. Parameters that have to be inferred are represented by white circles. The observed data and the fixed hyperparameters are represented by gray circles. All nodes in the plate are specific for segment h. Unlike the original $\mathcal{M}_{0,0}$ model, whose graphical model is shown in Figure 2, the $\mathcal{M}_{1,1}$ model has a specific coupling parameter λ_h for each segment $h > 1$. Furthermore, there is a new gamma hyperprior onto the second parameter of the inverse gamma prior on λ_h. The hyperprior allows for information exchange among the segment-specific coupling parameters $\lambda_2, \ldots, \lambda_H$

The full conditional distribution of $\boldsymbol{\beta}_h$ $(h = 1, \ldots, H)$ can be derived as follows. With $\lambda_1 := \lambda_u$ and $\tilde{\boldsymbol{\beta}}_0 := \mathbf{0}$, we have

$$
\begin{aligned}
p\left(\boldsymbol{\beta}_h|\mathbf{y}_1, \ldots, \mathbf{y}_H, \{\boldsymbol{\beta}_k\}_{k\neq h}, \sigma^2, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c\right) &\propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c\right) \\
&\propto p\left(\boldsymbol{\beta}_h|\lambda_h, \sigma^2\right) \cdot p\left(\mathbf{y}_h|\boldsymbol{\beta}_h, \sigma^2\right) \\
&\propto e^{-\frac{1}{2}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})^\mathsf{T}\left(\lambda_h\sigma^2\mathbf{I}\right)^{-1}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})} \\
&\quad \cdot e^{-\frac{1}{2}(\mathbf{y}_h-\mathbf{X}_h\boldsymbol{\beta}_h)^\mathsf{T}(\sigma^2\mathbf{I})^{-1}(\mathbf{y}_h-\mathbf{X}_h\boldsymbol{\beta}_h)} \\
&\propto e^{-\frac{1}{2}\sigma^{-2}\cdot\left(\lambda_h^{-1}\boldsymbol{\beta}_h^\mathsf{T}\boldsymbol{\beta}_h+2\lambda_h^{-1}\boldsymbol{\beta}_h^\mathsf{T}\tilde{\boldsymbol{\beta}}_{h-1}+\boldsymbol{\beta}_h^\mathsf{T}(\mathbf{X}_h^\mathsf{T}\mathbf{X}_h)\boldsymbol{\beta}_h-2\boldsymbol{\beta}_h^\mathsf{T}\mathbf{X}_h^\mathsf{T}\mathbf{y}_h\right)} \\
&\propto e^{-\frac{1}{2}\sigma^{-2}\cdot\left(\boldsymbol{\beta}_h^\mathsf{T}\left(\lambda_h^{-1}\cdot\mathbf{I}+\mathbf{X}_h^\mathsf{T}\mathbf{X}_h\right)\boldsymbol{\beta}_h-2\cdot\boldsymbol{\beta}_h^\mathsf{T}\left(\lambda_h^{-1}\tilde{\boldsymbol{\beta}}_{h-1}+\mathbf{X}_h^\mathsf{T}\mathbf{y}_h\right)\right)} \\
&\propto e^{-\frac{1}{2}\cdot\boldsymbol{\beta}_h^\mathsf{T}\left(\sigma^{-2}\left[\lambda_h^{-1}\cdot\mathbf{I}+\mathbf{X}_h^\mathsf{T}\mathbf{X}_h\right]\right)\boldsymbol{\beta}_h+\boldsymbol{\beta}_h^\mathsf{T}\left(\sigma^{-2}\left(\lambda_h^{-1}\tilde{\boldsymbol{\beta}}_{h-1}+\mathbf{X}_h^\mathsf{T}\mathbf{y}_h\right)\right)},
\end{aligned}
$$

and from the shape of the latter density, it follows for the following full conditional distribution:

$$
\begin{aligned}
&\boldsymbol{\beta}_h|\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \{\boldsymbol{\beta}_k\}_{k\neq h}, \sigma^2, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c\right) \\
&\qquad \sim \mathcal{N}\left(\left[\lambda_h^{-1}\mathbf{I}+\mathbf{X}_h^\mathsf{T}\mathbf{X}_h\right]^{-1}\left(\lambda_h^{-1}\tilde{\boldsymbol{\beta}}_{h-1}+\mathbf{X}_h^\mathsf{T}\mathbf{y}_h\right),\ \sigma^2\left[\lambda_h^{-1}\mathbf{I}+\mathbf{X}_h^\mathsf{T}\mathbf{X}_h\right]^{-1}\right). \quad (14)
\end{aligned}
$$

We note that the posterior expectation in Equation (14) is identical to the one we used in Equation (12).

For the full conditional distributions of the segment-specific coupling parameters λ_h ($h = 2, \ldots, H$), we get

$$\begin{aligned}
p\left(\lambda_h|\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \{\lambda_k\}_{k\neq h}, \beta_c\right) &\propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c\right) \\
&\propto p(\lambda_h|\beta_c) \cdot p\left(\boldsymbol{\beta}_h|\sigma^2, \lambda_h\right) \\
&\propto \frac{\beta_c^{\alpha_c}}{\Gamma(\alpha_c)}\left(\lambda_h^{-1}\right)^{\alpha_c-1} e^{-\beta_c\lambda_h^{-1}} \\
&\quad \cdot (2\pi)^{-\frac{k+1}{2}} \frac{1}{\sqrt{\det(\lambda_h\sigma^2\mathbf{I})}} e^{-\frac{1}{2}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}(\lambda_h\sigma^2\mathbf{I})^{-1}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})} \\
&\propto \left(\lambda_h^{-1}\right)^{\alpha_c-1} e^{-\beta_c\lambda_h^{-1}} \\
&\quad \cdot \lambda_h^{-(k+1)/2} e^{-\frac{1}{2}\lambda_h^{-1}\sigma^{-2}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})} \\
&\propto \left(\lambda_h^{-1}\right)^{\alpha_c+\frac{k+1}{2}-1} \cdot e^{-\lambda_h^{-1}\left(\beta_c+\frac{1}{2}\sigma^{-2}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}(\boldsymbol{\beta}_h-\tilde{\boldsymbol{\beta}}_{h-1})\right.},
\end{aligned}$$

and it follows from the shape of the following full conditional density:

$$\begin{aligned}
&\lambda_h^{-1}|\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \{\lambda_k\}_{k\neq h}, \beta_c\right) \\
&\qquad \sim \text{GAM}\left(\alpha_c + \frac{(k+1)}{2}, \beta_c + \frac{1}{2}\sigma^{-2}(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_{h-1})\right). \quad (15)
\end{aligned}$$

For the full conditional distribution of λ_u, we get in a similar way

$$\begin{aligned}
p\left(\lambda_u|\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \{\lambda_h\}_{h\geq 2}, \beta_c\right) &\propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c\right) \\
&\propto p(\lambda_u) \cdot p\left(\boldsymbol{\beta}_1|\sigma^2, \lambda_u\right) \\
&\propto \left(\lambda_u^{-1}\right)^{\alpha_u-1} e^{-\beta_u\lambda_u^{-1}} \cdot \frac{1}{\sqrt{\det(\lambda_u\sigma^2\mathbf{I})}} e^{-\frac{1}{2}\boldsymbol{\beta}_1^{\mathsf{T}}(\lambda_u\sigma^2\mathbf{I})^{-1}\boldsymbol{\beta}_1} \\
&\propto \left(\lambda_u^{-1}\right)^{\alpha_u-1} \cdot e^{-\beta_u\lambda_u^{-1}} \cdot \lambda_u^{-(k+1)/2} e^{-0.5\lambda_u^{-1}\sigma^{-2}\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1} \\
&\propto \left(\lambda_u^{-1}\right)^{(k+1)/2+\alpha_u-1} \cdot e^{-\lambda_u^{-1}(\beta_u+0.5\sigma^{-2}\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1)}.
\end{aligned}$$

The full conditional density has the shape of the inverse gamma distribution in Equation (8), that is, the following full conditional distribution of λ_u stays unchanged:

$$\lambda_u^{-1}|\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \{\lambda_h\}_{h\geq 2}, \beta_c\right) \sim \text{GAM}\left(\alpha_u + \frac{1\cdot(k+1)}{2}, \beta_u + \frac{1}{2}\sigma^{-2}\cdot\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{\beta}_1\right). \quad (16)$$

The new hyperparameter β_c can also be sampled from its full conditional distribution. The shape of

$$\begin{aligned}
p\left(\beta_c|\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2\lambda_u, \{\lambda_h\}_{h\geq 2}\right) &\propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2\lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c\right) \\
&\propto p(\beta_c) \cdot \prod_{h=2}^{H} p(\lambda_h|\beta_c) \\
&\propto \frac{b^a}{\Gamma(a)} \cdot \beta_c^{a-1} \cdot e^{-b\beta_c} \cdot \prod_{h=2}^{H}\left(\frac{\beta_c^{\alpha_c}}{\Gamma(\alpha_c)} \cdot \lambda_h^{\alpha_c-1} \cdot e^{-\beta_c\lambda_h^{-1}}\right) \\
&\propto \beta_c^{a+(H-1)\alpha_c-1} \cdot e^{-\left(b+\sum_{h=2}^{H}\lambda_h^{-1}\right)\beta_c}
\end{aligned}$$

implies the following full conditional distribution of β_c:

$$\beta_c|\left(\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\boldsymbol{\beta}_1,\ \ldots,\boldsymbol{\beta}_H,\sigma^2,\lambda_u,\{\lambda_h\}_{h\geq 2}\right)\sim \text{GAM}\left(a+(H-1)\cdot\alpha_c,b+\sum_{h=2}^{H}\lambda_h^{-1}\right). \tag{17}$$

For the noise variance parameter σ^2, we follow Grzegorczyk and Husmeier (2012) and implement a collapsed Gibbs sampling step (with the $\boldsymbol{\beta}_h$ integrated out). We have

$$\begin{aligned} p\left(\mathbf{y}_h|\sigma^2,\lambda_h\right) &= \int p\left(\mathbf{y}_h,\boldsymbol{\beta}_h|\sigma^2,\lambda_h\right)d\boldsymbol{\beta}_h = \int p\left(\mathbf{y}_h|\boldsymbol{\beta}_h,\sigma^2,\lambda_h\right)p\left(\boldsymbol{\beta}_h|\sigma^2,\lambda_h\right)d\boldsymbol{\beta}_h \\ &= \int p\left(\mathbf{y}_h|\boldsymbol{\beta}_h,\sigma^2\right)p\left(\boldsymbol{\beta}_h|\sigma^2,\lambda_h\right)d\boldsymbol{\beta}_h. \end{aligned}$$

A standard rule for Gaussian integrals (see, e.g., section 2.3.2 in Bishop, 2006) implies:

$$\mathbf{y}_h|\left(\sigma^2,\lambda_h\right)\sim\mathcal{N}\left(\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1},\sigma^2\left[\mathbf{I}+\lambda_h\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}}\right]\right). \tag{18}$$

With $\lambda_1 := \lambda_u$, $\tilde{\boldsymbol{\beta}}_0 := \mathbf{0}$, and using the marginal likelihood from Equation (18), we have

$$\begin{aligned} p\left(\sigma^2|\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\lambda_u,\{\lambda_h\}_{h\geq 2},\beta_c\right) &\propto p\left(\sigma^2,\lambda_u,\{\lambda_h\}_{h\geq 2},\beta_c|\mathbf{y}_1,\ \ldots,\mathbf{y}_H\right) \\ &\propto p\left(\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\sigma^2,\lambda_u,\{\lambda_h\}_{h\geq 2},\beta_c\right) \\ &\propto \left(\prod_{h=1}^{H}p\left(\mathbf{y}_h|\sigma^2,\lambda_h\right)\right)\cdot p(\sigma^2)\cdot p(\lambda_u)\cdot\prod_{h=2}^{H}p(\lambda_h|\beta_c)\cdot p(\beta_c) \\ &\propto \left(\prod_{h=1}^{H}p\left(\mathbf{y}_h|\sigma^2,\lambda_h\right)\right)\cdot p(\sigma^2) \\ &\propto (\sigma^{-2})^{0.5\sum_{h-1}^{H}T_h}\exp\left\{-0.5\sigma^{-2}\sum_{h=1}^{H}\left(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1}\right)^{\mathsf{T}}\left(\mathbf{I}+\lambda_h\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}}\right)^{-1}\left(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1}\right)\right\} \\ &\quad\cdot(\sigma^{-2})^{\alpha_\sigma-1}\exp\left\{-\beta_\sigma\sigma^{-2}\right\} \\ &\propto \exp\left\{-\sigma^{-2}\left(\beta_\sigma+0.5\cdot\sum_{h=1}^{H}\left(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1}\right)^{\mathsf{T}}\left(\mathbf{I}+\lambda_h\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}}\right)^{-1}\left(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1}\right)\right)\right\} \\ &\quad\cdot(\sigma^{-2})^{\alpha_\sigma+0.5\cdot T-1}. \end{aligned}$$

The shape of the latter density implies the following collapsed Gibbs sampling step (with the $\boldsymbol{\beta}_h$ integrated out):

$$\sigma^{-2}|\left(\mathbf{y}_1,\ \ldots,\mathbf{y}_H,\lambda_u,\{\lambda_h\}_{h\geq 2},\beta_c\right)\sim\text{GAM}\left(\alpha_\sigma+\frac{1}{2}\cdot T,\beta_\sigma+\frac{1}{2}\sum_{h=1}^{H}(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1})^{\mathsf{T}}\left(\mathbf{I}+\lambda_h\mathbf{X}_h\mathbf{X}_h^{\mathsf{T}}\right)^{-1}(\mathbf{y}_h-\mathbf{X}_h\tilde{\boldsymbol{\beta}}_{h-1})\right), \tag{19}$$

where $\lambda_1 := \lambda_u$ and $\tilde{\boldsymbol{\beta}}_0 := \mathbf{0}$.

For the marginal likelihood, with $\boldsymbol{\beta}_h$ ($h = 1,\ \ldots,H$) and σ^2 integrated out, again the marginalization rule from section 2.3.7 of Bishop (2006) can be applied. For the improved model, the

following marginalization rule implies:

$$p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H | \lambda_u, \{\lambda_h\}_{h\geq 2}\right) = \frac{\Gamma\left(\frac{T}{2} + a_\sigma\right)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-\frac{T}{2}} (2b_\sigma)^{a_\sigma}}{\left(\prod_{h=1}^{H} \det(\mathbf{C}_h)\right)^{1/2}} \left(2b_\sigma + \sum_{h=1}^{H} (\mathbf{y}_h - \mathbf{X}_h \tilde{\boldsymbol{\beta}}_{h-1})^\top \mathbf{C}_h^{-1} (\mathbf{y}_h - \mathbf{X}_h \tilde{\boldsymbol{\beta}}_{h-1})\right)^{-\left(\frac{T}{2} + a_\sigma\right)}, \quad (20)$$

where $\lambda_1 := \lambda_u$, $\tilde{\boldsymbol{\beta}}_0 := \mathbf{0}$, and $\mathbf{C}_h := \mathbf{I} + \lambda_h \mathbf{X}_h \mathbf{X}_h^\top$,

2.4 | Models "in between" the original and the improved sequentially coupled model

In the last subsection, we have proposed an improved sequentially coupled model, $\mathcal{M}_{1,1}$. Compared with the original model $\mathcal{M}_{0,0}$ from Section 2.1, we have proposed two modifications: (a) to replace the shared coupling parameter λ_c by segment-specific coupling parameters $\{\lambda_h\}_{h\geq 2}$, and (b) to impose a hyperprior onto the hyperparameter β_c of the inverse gamma prior on the coupling parameters $\{\lambda_h\}_{h\geq 2}$. To shed more light onto the relative merits of the two individual modifications, we also define the two "in-between" models where only one of the two modifications is implemented.

The first "in-between" model $\mathcal{M}_{1,0}$ does not introduce segment-specific coupling parameters but places a hyperprior onto the hyperparameter β_c of the inverse gamma prior on the shared coupling parameter λ_c. A graphical model representation for $\mathcal{M}_{1,0}$ is shown in Figure 4. The posterior distribution of the $\mathcal{M}_{1,0}$ model is an extension of Equation (7), as follows:

$$p\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \lambda_c, \beta_c | \mathbf{y}_1, \ldots, \mathbf{y}_H\right) \propto \prod_{h=1}^{H} p\left(\mathbf{y}_h | \sigma^2, \boldsymbol{\beta}_h\right) \cdot p\left(\boldsymbol{\beta}_1 | \sigma^2, \lambda_u\right) \cdot \prod_{h=2}^{H} p\left(\boldsymbol{\beta}_h | \sigma^2, \lambda_c\right) \cdot p(\sigma^2) \cdot p(\lambda_u) \cdot p(\lambda_c | \beta_c) \cdot p(\beta_c).$$

The modification does neither change the earlier defined full conditional distributions from Section 2.1 nor the marginal likelihood in Equation (10). The only difference is that β_c has become a free parameter that, thus, must now be sampled too. For the full conditional distribution of β_c in the $\mathcal{M}_{1,0}$ model, we have

$$\begin{aligned} p\left(\beta_c | \mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2 \lambda_u, \lambda_c\right) &\propto p\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2 \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c | \mathbf{y}_1, \ldots, \mathbf{y}_H\right) \\ &\propto p(\lambda_c | \beta_c) \cdot p(\beta_c) \\ &\propto \frac{\beta_c^{\alpha_c}}{\Gamma(\alpha_c)} \cdot \lambda_c^{\alpha_c - 1} \cdot e^{-\beta_c \lambda_c^{-1}} \cdot \frac{b^a}{\Gamma(a)} \cdot \beta_c^{a-1} \cdot e^{-b\beta_c} \\ &\propto \beta_c^{a+\alpha_c - 1} \cdot e^{-\left(b + \lambda_c^{-1}\right)\beta_c}. \end{aligned}$$

This implies for the full conditional distribution:

$$\beta_c | \left(\mathbf{y}_1, \ldots, \mathbf{y}_H, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \lambda_c\right) \sim \text{GAM}\left(a + \alpha_c, b + \lambda_c^{-1}\right). \quad (21)$$

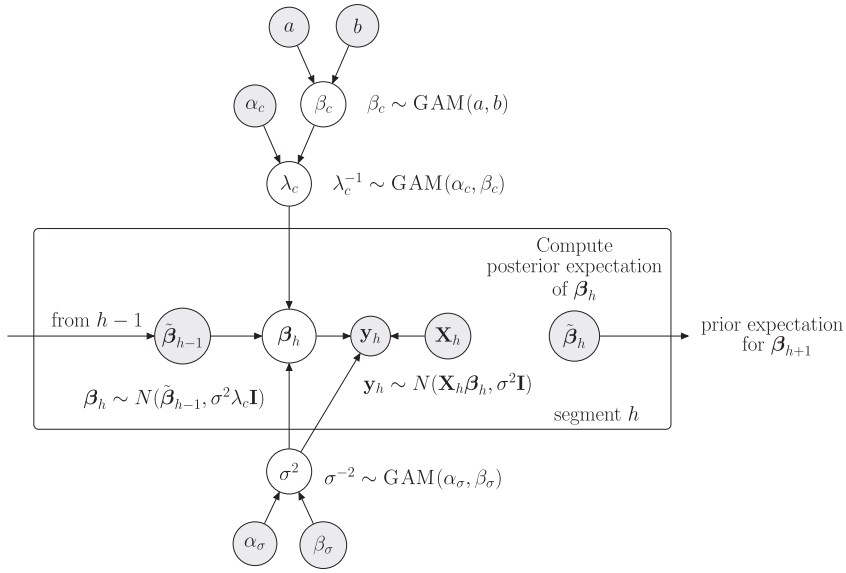

FIGURE 4 Graphical model for the first "in-between" model $\mathcal{M}_{1,0}$. See caption of Figure 2 for the terminology. The model is an extension of the original sequentially coupled model, whose graphical model is shown in Figure 2. Unlike the original $\mathcal{M}_{0,0}$ model, the $\mathcal{M}_{1,0}$ has a free hyperparameter, β_c, with a gamma hyperprior

The second "in-between" model $\mathcal{M}_{0,1}$ does make use of segment-specific coupling parameters $\{\lambda_h\}_{h\geq 2}$ but keeps the hyperparameter β_c of the inverse gamma priors on the parameters $\{\lambda_h\}_{h\geq 2}$ fixed. This yields that the segment-specific coupling parameters $\lambda_2, \ldots, \lambda_H$ are independent a priori. A graphical model representation is shown in Figure 5. The posterior distribution of the second "in-between" model $\mathcal{M}_{0,1}$ is a simplified version of Equation (13), as follows:

$$p\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \sigma^2, \lambda_u, \{\lambda_h\}_{h\geq 2} | \mathbf{y}_1, \ldots, \mathbf{y}_H\right) \propto \prod_{h=1}^{H} p\left(\mathbf{y}_h | \sigma^2, \boldsymbol{\beta}_h\right) \cdot p\left(\boldsymbol{\beta}_1 | \sigma^2, \lambda_u\right)$$
$$\cdot \prod_{h=2}^{H} p\left(\boldsymbol{\beta}_h | \sigma^2, \lambda_h\right) \cdot p(\sigma^2) \cdot p(\lambda_u) \cdot \prod_{h=2}^{H} p(\lambda_h),$$

and the modification (i.e., fixing λ_c) does not change either the full conditional distributions in Equations (14)–(16) and (19) or the marginal likelihood in Equation (20). The only difference is that λ_c is kept fixed and will not be inferred from the data. The corresponding Gibbs sampling step [see Equation (21)] is never performed.

2.5 | Learning the covariate set

In typical applications, the covariates have to be inferred from the data. That is, there is a set of potential covariates, and the subset, relevant for the target Y, has to be found.

Let $X_1, \ldots, X_n$ be a set of *potential* covariates for the target variable Y, and let $\mathcal{D}_1, \ldots, \mathcal{D}_T$ be equidistant and temporally ordered data points. Each $\mathcal{D}_t$ contains a target value y_t and the values $x_{1,t-1}, \ldots, x_{n,t-1}$ of the n potential covariates. A priori, we assume that all covariate sets

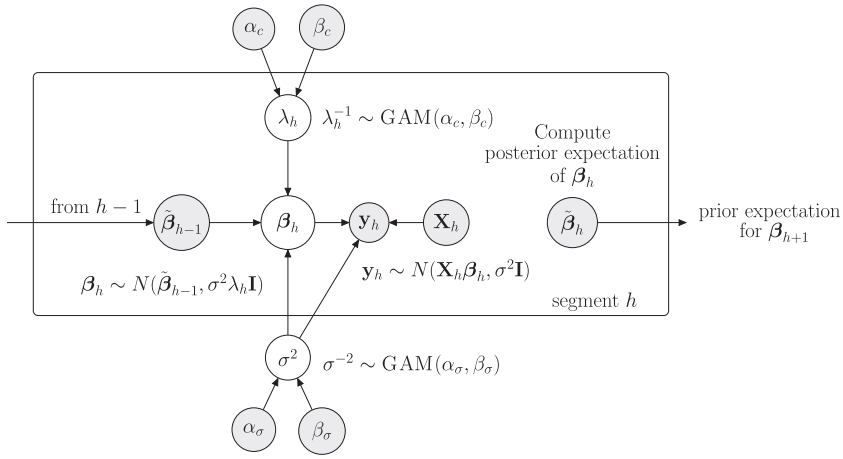

FIGURE 5 Graphical model for the second "in-between" model $\mathcal{M}_{0,1}$. See caption of Figure 3 for the terminology. The model is similar to the proposed improved sequentially coupled model, whose graphical model is shown in Figure 3. Unlike the proposed $\mathcal{M}_{1,1}$ model, the $\mathcal{M}_{0,1}$ model has a fixed hyperparameter β_c

$\pi \subset \{X_1, \ldots, X_n\}$ with up to three covariates are equally likely, whereas all parent sets with more than three elements have zero prior probability.[1]

$$p(\pi) = \begin{cases} \frac{1}{c} & \text{if} |\pi| \leq 3 \\ 0 & \text{if} |\pi| > 3 \end{cases}, \quad \text{where} \quad c = \sum_{i=0}^{3} \binom{n}{i}.$$

We make the assumption that there cannot be more than three covariates ($|\pi| \leq 3$) with regard to our applications in the field of gene network inference, and we note that this assumption might be inappropriate for other applications. However, in the context of gene regulatory networks, this assumption is very common. The known topologies of gene regulatory networks suggest that there are rarely genes that are regulated by more than three regulators. The assumption is thus biologically reasonable and has the advantage that it reduces the complexity of the problem and the computational costs of the Markov chain Monte Carlo (MCMC)-based model inference.

Given a fixed segmentation into the segments $h = 1, \ldots, H$, for each possible covariate set π, the piecewise linear regression models can be applied. We focus our attention on the improved sequentially coupled model $\mathcal{M}_{1,1}$ from Section 2.2, but we note that the MCMC algorithm can also be used for generating samples for the competing models ($\mathcal{M}_{0,0}$, $\mathcal{M}_{1,0}$, and $\mathcal{M}_{0,1}$). Only the marginal likelihood expressions have to be replaced in the acceptance probabilities.

Using the marginal likelihood from Equation (20), we obtain the following for the posterior of the $\mathcal{M}_{1,1}$ model:

$$p(\pi, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c | \mathcal{D}_1, \ldots, \mathcal{D}_T) \propto p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H | \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c, \pi\right) \cdot p(\pi) \cdot p(\lambda_u) \cdot \prod_{h=2}^{H} p(\lambda_h|\beta_c) \cdot p(\beta_c). \tag{22}$$

Given λ_u, $\{\lambda_h\}_{h\geq 2}$, and β_c, the Metropolis–Hastings algorithm can be used to sample the covariate set π. We implement three moves: In the deletion move (D), we randomly select one $X_i \in \pi$ and remove it from π. In the addition move (A), we randomly select one $X_i \notin \pi$ and add it to π. In the exchange move (E), we randomly select one $X_i \in \pi$ and replace it by a randomly

[1]To be consistent with earlier studies, we assume all covariate sets containing up to three covariates to be equally likely.

selected $X_j \notin \pi$. Each move yields a new covariate set $\pi^\star$, and we propose to replace the current π by $\pi^\star$. When randomly selecting the move type, the acceptance probability for the proposed move is

$$A(\pi, \pi^\star) = \min\left\{1, \frac{p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H | \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c, \pi^\star\right)}{p\left(\mathbf{y}_1, \ldots, \mathbf{y}_H | \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c, \pi\right)} \cdot \frac{p(\pi^\star)}{p(\pi)} \cdot HR\right\},$$

$$\text{where} \quad HR = \begin{cases} \frac{|\pi|}{n-|\pi^\star|} & \text{for(D)} \\ \frac{n-|\pi|}{|\pi^\star|} & \text{for(A)} \\ 1 & \text{for(E)} \end{cases},$$

n is the number of potential covariates $X_1, \ldots, X_n$, and $|.|$ denotes the cardinality.

2.6 | Learning the segmentation

If the segmentation of the data is unknown, we can also infer it from the data. We assume that a changepoint set $\boldsymbol{\tau} := \{\tau_1, \ldots, \tau_{H-1}\}$ with $1 \leq \tau_h < T$ divides the data points $\mathcal{D}_1, \ldots, \mathcal{D}_T$ into disjunct segments $h = 1, \ldots, H$ covering $T_1, \ldots, T_H$ consecutive data points, where $\sum_{h=1}^{H} T_h = T$. Data point $\mathcal{D}_t$ $(1 \leq t \leq T)$ is in segment h if $\tau_{h-1} < t \leq \tau_h$, where $\tau_0 := 1$ and $\tau_H := T$. A priori, we assume that the distances between changepoints are geometrically distributed with hyperparameter $p \in (0, 1)$ and that there cannot be more $H = 10$ segments.[2] This implies the prior density:

$$p(\boldsymbol{\tau}) = \begin{cases} (1-p)^{\tau_H - \tau_{H-1}} \cdot \prod_{h=1}^{H-1} p \cdot (1-p)^{\tau_h - \tau_{h-1} - 1} & \text{if } |\boldsymbol{\tau}| \leq 9 \ \text{ (i.e., } H \leq 10) \\ 0 & \text{if } |\boldsymbol{\tau}| > 9 \ \text{ (i.e., } H > 10). \end{cases} \tag{23}$$

Let $\mathbf{y}_{\boldsymbol{\tau}} := \{\mathbf{y}_1, \ldots, \mathbf{y}_H\}$ denote the segmentation, implied by the changepoint set $\boldsymbol{\tau}$.

Again, we focus on the improved sequentially coupled model $\mathcal{M}_{1,1}$, and just note that the MCMC algorithm requires only minor adaptions when used for the three competing models, namely the original sequentially coupled model $\mathcal{M}_{0,0}$ (see Section 2.1) and the two "in-between" models $\mathcal{M}_{1,0}$ and $\mathcal{M}_{0,1}$ (see Section 2.4).

Using the marginal likelihood from Equation (20), the posterior of the improved model takes the following form:

$$p(\pi, \boldsymbol{\tau}, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c | \mathcal{D}_1, \ldots, \mathcal{D}_T) \propto p\left(\mathbf{y}_{\boldsymbol{\tau}} | \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c, \pi, \boldsymbol{\tau}\right) \cdot p(\pi) \cdot p(\boldsymbol{\tau}) \cdot p(\lambda_u) \cdot \prod_{h=2}^{H} p(\lambda_h | \beta_c) \cdot p(\beta_c). \tag{24}$$

For sampling the changepoint sets, we also implement three Metropolis–Hastings moves. Each move proposes to replace the current changepoint set $\boldsymbol{\tau}$ by a new changepoint set $\boldsymbol{\tau}^\star$, and $\boldsymbol{\tau}^\star$ implies a new data segmentation $\mathbf{y}_{\boldsymbol{\tau}}^\star := \{\mathbf{y}_1^\star, \ldots, \mathbf{y}_{H^\star}^\star\}$. The new segmentation $\mathbf{y}_{\boldsymbol{\tau}}^\star$ contains new segments h that have not been in $\mathbf{y}_{\boldsymbol{\tau}}$, symbolically $\mathbf{y}_h^\star \notin \mathbf{y}_{\boldsymbol{\tau}}$, and for each of those segments h, we sample a new segment-specific coupling parameter from the prior, $\lambda_h^\star \sim \text{INV-GAM}(\alpha_c, \beta_c)$. For all

[2]The assumption that there cannot be more than $H = 10$ segments is made with regard to our applications. Gene expression time series are often rather short; the gene expressions in yeast, described in Section 3.2, have been measured over $T = 33$ time points only. Restricting the number of segments H avoids segmentations whose individual segments are very short and uninformative.

other segments, we do not change the segment-specific coupling parameters. Let $\{\lambda_h^\star\}_{h\geq 2}$ denote the set of coupling parameters associated with the new segmentation $\mathbf{y}_{\boldsymbol{\tau}}^\star$.

In the birth move (B), we propose to set a new changepoint at a randomly selected location. The new changepoint set $\boldsymbol{\tau}^\star$ then contains $H^\star = H + 1$ changepoints. The new changepoint is located in a segment h and divides it into two consecutive subsegments h and $h+1$. For both, we resample the segment-specific coupling parameters $\lambda_h^\star, \lambda_{h+1}^\star \sim \text{INV-GAM}(\alpha_c, \beta_c)$. In the death move (D), we randomly select one changepoint $\tau \in \boldsymbol{\tau}$ and delete it. The new changepoint set $\boldsymbol{\tau}^\star$ then contains $H^\star = H - 1$ changepoints. Removing a changepoint yields that two adjacent segments h and $h+1$ are merged into one single segment h, and we sample $\lambda_h^\star \sim \text{INV-GAM}(\alpha_c, \beta_c)$. In the reallocation move (R), we randomly select one changepoint $\tau \in \boldsymbol{\tau}$ and propose to reallocate it to a randomly selected position in between the two surrounding changepoints. The reallocated changepoint yields new bounds for two consecutive segments h (whose upper bound changes) and $h+1$ (whose lower bound changes), and for both segments, we resample the coupling parameters $\lambda_h^\star, \lambda_{h+1}^\star \sim \text{INV-GAM}(\alpha_c, \beta_c)$.

When randomly selecting the move type, the acceptance probabilities for the move from the changepoints set $\boldsymbol{\tau}$ with segmentation $\mathbf{y}_{\boldsymbol{\tau}} := \{\mathbf{y}_1, \ldots, \mathbf{y}_H\}$ and coupling parameters $\{\lambda_h\}_{h\geq 2}$ to the changepoint set $\boldsymbol{\tau}^\star$ with the new segmentation $\mathbf{y}_{\boldsymbol{\tau}}^\star := \{\mathbf{y}_1^\star, \ldots, \mathbf{y}_{H^\star}^\star\}$ and the new coupling parameters $\{\lambda_h^\star\}_{h\geq 2}$ are as follows:

$$A\left(\left[\boldsymbol{\tau}, \{\lambda_h\}_{h\geq 2}\right], \left[\boldsymbol{\tau}^\star, \left\{\lambda_h^\star\right\}_{h\geq 2}\right]\right) = \min\left\{1, \frac{p\left(\mathbf{y}_{\boldsymbol{\tau}^\star} | \lambda_u, \left\{\lambda_h^\star\right\}_{h\geq 2}, \beta_c, \pi, \boldsymbol{\tau}^\star\right)}{p\left(\mathbf{y}_{\boldsymbol{\tau}} | \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c, \pi, \boldsymbol{\tau}\right)} \cdot \frac{p(\boldsymbol{\tau}^\star)}{p(\boldsymbol{\tau})} \cdot HR\right\},$$

$$\text{where } HR = \begin{cases} \frac{T-1-|\boldsymbol{\tau}^*|}{|\boldsymbol{\tau}|} & \text{for (B)} \\ \frac{|\boldsymbol{\tau}^*|}{T-1-|\boldsymbol{\tau}|} & \text{for (D)} \\ 1 & \text{for (R)} \end{cases},$$

T is the number of data points, and $|.|$ denotes the cardinality. We note that the prior ratio $\frac{p(\{\lambda_h^\star\}_{h\geq 2})}{p(\{\lambda_h\}_{h\geq 2})}$ has canceled with the inverse proposal ratio (HR) for resampling the coupling parameters for the new segments.

To adapt the MCMC algorithm to the competing models, the marginal likelihood expressions in the acceptance probability have to be replaced. Moreover, for the two models ($\mathcal{M}_{0,0}$ and $\mathcal{M}_{1,0}$) with a shared coupling parameter λ_c, we follow Grzegorczyk and Husmeier (2012) and implement the three changepoint moves such that they do not propose to resample λ_c.

2.7 | MCMC inference

For model inference, we use an MCMC algorithm. For the posterior distribution of the improved sequentially coupled model $\mathcal{M}_{1,1}$, described in Section 2.2, we have

$$p(\pi, \boldsymbol{\tau}, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c | \mathcal{D}_1, \ldots, \mathcal{D}_T) \propto p\left(\mathbf{y}_{\boldsymbol{\tau}} | \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c, \pi, \boldsymbol{\tau}\right) \cdot p(\pi) \cdot p(\boldsymbol{\tau}) \cdot p(\lambda_u) \cdot \prod_{h=2}^{H} p(\lambda_h | \beta_c) \cdot p(\beta_c). \tag{25}$$

We initialize all entities, for example, $\pi = \{\}$, $\boldsymbol{\tau} = \{\}$, $\lambda_u = 1$, $\lambda_h = 1$ for $h > 1$, and $\beta_c = 1$, before we iterate between Gibbs and Metropolis–Hastings sampling steps:

Gibbs sampling part: We keep the covariate set π and the changepoint set $\boldsymbol{\tau}$ fixed, and we successively resample the parameters λ_u, $\lambda_h (h = 2, \ldots, H)$, and β_c from their full conditional distributions. Although the parameters σ^2 and $\boldsymbol{\beta}_h$ $(h = 1, \ldots, H)$ do not appear in the posterior above, the full conditionals of λ_u and $\lambda_2, \ldots, \lambda_H$ depend on instantiations of σ^2 and $\boldsymbol{\beta}_h$. The latter parameters thus have to be sampled first but can then be withdrawn at the end of the Gibbs sampling part. The full conditional distributions have been derived in Section 2.3. With $\lambda_1 := \lambda_u$ and $\tilde{\boldsymbol{\beta}}_0 = \mathbf{0}$, we have

(G.1) $\sigma^{-2} \sim \text{GAM}\left(\alpha_\sigma + \frac{1}{2} \cdot T, \beta_\sigma + \frac{1}{2} \sum_{h=1}^{H} \left(\mathbf{y}_h - \mathbf{X}_h \tilde{\boldsymbol{\beta}}_{h-1}\right)^\mathsf{T} \left(\mathbf{I} + \lambda_h \mathbf{X}_h \mathbf{X}_h^\mathsf{T}\right)^{-1} \left(\mathbf{y}_h - \mathbf{X}_h \tilde{\boldsymbol{\beta}}_{h-1}\right)\right)$

(G.2) $\boldsymbol{\beta}_h \sim \mathcal{N}\left(\left[\lambda_h^{-1}\mathbf{I} + \mathbf{X}_h^\mathsf{T}\mathbf{X}_h\right]^{-1} \left(\lambda_h^{-1}\tilde{\boldsymbol{\beta}}_{h-1} + \mathbf{X}_h^\mathsf{T}\mathbf{y}_h\right),\ \sigma^2\left[\lambda_h^{-1}\mathbf{I} + \mathbf{X}_h^\mathsf{T}\mathbf{X}_h\right]^{-1}\right) \quad (h = 1, \ldots, H)$

(G.3) $\lambda_u^{-1} \sim \text{GAM}\left(\alpha_u + \frac{1 \cdot (k+1)}{2}, \beta_u + \frac{1}{2}\sigma^{-2} \cdot \boldsymbol{\beta}_1^\mathsf{T}\boldsymbol{\beta}_1\right)$

(G.4) $\lambda_h^{-1} \sim \text{GAM}\left(\alpha_c + \frac{(k+1)}{2}, \beta_c + \frac{1}{2}\sigma^{-2}(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_{h-1})^\mathsf{T}(\boldsymbol{\beta}_h - \tilde{\boldsymbol{\beta}}_{h-1})\right) \quad (h = 1, \ldots, H)$

(G.5) $\beta_c \sim \text{GAM}\left(a + (H-1) \cdot \alpha_c, b + \sum_{h=2}^{H} \lambda_h^{-1}\right)$.

We note that each Gibbs step yields parameter updates and that the subsequent full conditional distributions are always based on the newest parameter instantiations (sampled in the previous steps).

Metropolis–Hastings sampling part: We keep λ_u, $\lambda_h (h = 2, \ldots, H)$, and β_c fixed, and we perform one Metropolis–Hastings move on the covariate set π and one Metropolis–Hastings step on the changepoint set $\boldsymbol{\tau}$.

(M.1) We propose to change the covariate set $\pi \rightarrow \pi^\star$ by adding, deleting, or exchanging one covariate. The new covariate set is accepted with probability $A(\pi, \pi^\star)$; see Section 2.5 for details. If accepted, we replace $\pi \leftarrow \pi^\star$. If rejected, we leave π unchanged.

(M.2) We propose to change the changepoint set $\boldsymbol{\tau} \rightarrow \boldsymbol{\tau}^\star$ by adding, deleting, or reallocating one changepoint. Along with the changepoint set, we propose to update coupling parameters, $\{\lambda_h\}_{h\geq 2} \rightarrow \{\lambda_h^\star\}_{h\geq 2}$. The new state is accepted with probability $A([\boldsymbol{\tau}, \{\lambda_h\}_{h\geq 2}], [\boldsymbol{\tau}^\star, \{\lambda_h\}_{h\geq 2}^\star])$; see Section 2.6 for details. If accepted, we replace $\boldsymbol{\tau} \leftarrow \boldsymbol{\tau}^\star$ and $\{\lambda_h\}_{h\geq 2} \leftarrow \{\lambda_h^\star\}_{h\geq 2}$. If rejected, we leave $\boldsymbol{\tau}$ and $\{\lambda_h\}_{h\geq 2}$ unchanged.

The MCMC algorithm, consisting of seven sampling steps (G.1–5) and (M.1–2), yields a posterior sample, as follows:

$$\left\{\pi^{(r)}, \boldsymbol{\tau}^{(r)}, \lambda_u^{(r)}, \{\lambda_h\}_{h\geq 2}^{(r)}, \beta_c^{(r)}\right\} \sim p(\pi, \boldsymbol{\tau}, \lambda_u, \{\lambda_h\}_{h\geq 2}, \beta_c | \mathcal{D}_1, \ldots, \mathcal{D}_T) \qquad (r = 1, \ldots, R).$$

We adapt the MCMC inference algorithm for the three alternative models ($\mathcal{M}_{0,0}$, $\mathcal{M}_{1,0}$, and $\mathcal{M}_{0,1}$) from Sections 2.1 and 2.4. To this end, we modify the Gibbs and Metropolis–Hastings steps as outlined in Sections 2.4–2.6.

2.8 | Learning dynamic networks

Dynamic network models are used for learning the regulatory interactions among variables from time series data. The standard assumption is that all regulatory interactions are subject to a time lag $\xi \in \mathbb{N}$. Here, we assume that the time lag has the standard value $\xi = 1$. We further assume that the values of n random variables $Y_1, \ldots, Y_n$ have been measured at T equidistant time points $t = 1, \ldots, T$. Let $\mathbf{D}$ denote the n-by-T data matrix with $\mathbf{D}_{i,t}$ being the observed value of Y_i at time point t. The piecewise linear regression models, described in the previous subsections, can then be applied to each variable $Y_i (i = 1, \ldots, n)$ separately.

In the ith regression model Y_i is the target, and the potential covariates are the $\tilde{n} := n - 1$ remaining variables $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$. Given the time lag ξ, the number of data points, which can be used for the regression model, reduces from T to $\tilde{T} := T - \xi$. For each target Y_i, we have the data points $\mathcal{D}_{i,1}, \ldots, \mathcal{D}_{i,\tilde{T}}$, and each data point $\mathcal{D}_{i,t}$ ($t = 1, \ldots, \tilde{T}$) contains a target value $\mathbf{D}_{i,t+\xi}$ (i.e., the value of Y_i at time point $t + \xi$) and the values of the ñ potential covariates: $\mathbf{D}_{1,t}, \ldots, \mathbf{D}_{i-1,t}, \mathbf{D}_{i+1,t}, \ldots, \mathbf{D}_{N,t}$ (i.e., the values of $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$ at the shifted time point t).

The system of all n covariate sets $\{\pi_i\}_{i=1,\ldots,n}$, where π_i is the covariate set for Y_i, can be thought of as a network. There is an edge $Y_j \to Y_i$ in the network if Y_j is a covariate for Y_i, symbolically $Y_j \in \pi_i$. There is no edge from Y_j to Y_i in the network if $Y_j \notin \pi_i$. We represent the resulting network in the form of a n-by-n adjacency matrix $\mathcal{N}$ whose elements are binary, $\mathcal{N}_{j,i} \in \{0, 1\}$. $\mathcal{N}_{j,i} = 1$ indicates that there is an edge from X_j to X_i (i.e., $X_j \in \pi_i$).

For each Y_i, we can generate a posterior sample, as described in Section 2.7. For each Y_i, we then extract the covariate sets, $\pi_i^{(1)}, \ldots, \pi_i^{(R)}$, from the sample and use the covariate sets to build a sample of adjacency matrices $\mathcal{N}^{(1)}, \ldots, \mathcal{N}^{(R)}$, where $\mathcal{N}_{j,i}^{(r)} = 1$ if $X_j \in \pi_i^{(r)}$ ($i,j \in \{1, \ldots, n\}; r \in \{1, \ldots, R\}$). The mean of the adjacency matrices

$$\hat{\mathcal{N}} := \frac{1}{R} \sum_{r=1}^{R} \mathcal{N}^{(r)}$$

yields estimates of the marginal posterior probabilities that the individual edges are present. For example, $\hat{\mathcal{N}}_{j,i} \in [0, 1]$ is an estimate for the marginal probability that there is an edge from X_j to X_i (i.e., $X_j \in \pi_i$). By imposing a threshold $\psi \in [0, 1]$ on the edge probabilities, we get a concrete network prediction. The predicted network contains all edges $X_j \to X_i$ whose probability to be present is equal to or higher than ψ ($\hat{\mathcal{N}}_{j,i} \geq \psi$).

For applications where the true network is known, we can build the n-by-n adjacency matrix of the true network $\mathcal{T}$ with $\mathcal{T}_{j,i} \in \{0, 1\}$ and $\mathcal{T}_{j,i} = 1$ if and only if the true network contains the edge $X_j \to X_i$. For each $\psi \in [0, 1]$, we can then compute the recall $\mathcal{R}(\psi)$ and the precision $\mathcal{P}(\psi)$ of the predicted network, as follows:

$$\mathcal{R}(\psi) = \frac{\left|\{X_j \to X_i | \mathcal{T}_{j,i} = 1, \hat{\mathcal{N}}_{j,i} \geq \psi\}\right|}{\left|\{X_j \to X_i | \mathcal{T}_{j,i} = 1\}\right|}$$

$$\mathcal{P}(\psi) = \frac{\left|\{X_j \to X_i | \mathcal{T}_{j,i} = 1, \hat{\mathcal{N}}_{j,i} \geq \psi\}\right|}{\left|\{X_j \to X_i | \hat{\mathcal{N}}_{j,i} \geq \psi\}\right|}.$$

The curve $\{(\mathcal{R}(\psi), \mathcal{P}(\psi)) | 0 \leq \psi \leq 1\}$ is the precision recall curve (Davis & Goadrich, 2006). The area under the precision recall curve (AUC), which can be obtained by numerical integration, is a popular measure for the network reconstruction accuracy. The higher the AUC, the higher the accuracy of the predicted network.

2.9 | Technical details of our simulation study

Table 1 provides an overview to the four models $\mathcal{M}_{i,j}$ ($i,j \in \{0, 1\}$) under comparison. We reuse the hyperparameters from the works by Lèbre et al. (2010) and Grzegorczyk and Husmeier (2012), namely

$$\sigma^{-2} \sim \text{GAM}(\alpha_\sigma = 0.005, \beta_\sigma = 0.005) \quad \text{and} \quad \lambda_u^{-1} \sim \text{GAM}(\alpha_u = 2, \beta_u = 0.2).$$

TABLE 1 Overview to the four model instantiations that we cross-compare

Model	Coupling parameter(s) for $h \geq 2$	Hyperparameter	Graphical model	see Section
$\mathcal{M}_{0,0}$	shared: $\lambda_c \sim \text{GAM}(\alpha_c, \beta_c)$	β_c fixed	see Figure 2	2.1
$\mathcal{M}_{1,0}$	shared: $\lambda_c \sim \text{GAM}(\alpha_c, \beta_c)$	$\beta_c \sim \text{GAM}(a, b)$	see Figure 4	2.4
$\mathcal{M}_{0,1}$	segment specific: $\lambda_h \sim \text{GAM}(\alpha_c, \beta_c)$	β_c fixed	see Figure 5	2.4
$\mathcal{M}_{1,1}$	segment specific: $\lambda_h \sim \text{GAM}(\alpha_c, \beta_c)$	$\beta_c \sim \text{GAM}(a, b)$	see Figure 3	2.2

Note. $\mathcal{M}_{0,0}$ is the sequentially coupled model from Grzegorczyk and Husmeier (2012). In this paper, we propose the improved $\mathcal{M}_{1,1}$ model, featuring two modifications. We also include the "in-between" models ($\mathcal{M}_{0,1}$ and $\mathcal{M}_{1,0}$) with only one modification incorporated. The first subscript of $\mathcal{M}$ indicates whether there is a hyperprior on β_c (0 = no,1 = yes). The second subscript of $\mathcal{M}$ indicates whether the model has segment-specific coupling parameters λ_h (0 = no,1 = yes).

For the models without hyperprior ($\mathcal{M}_{0,0}$ and $\mathcal{M}_{0,1}$), we further set

$$\lambda_c^{-1}, \lambda_h^{-1} \sim \text{GAM}(\alpha_c = 2, \beta_c = 0.2)$$

while we use the following for the models with hyperprior ($\mathcal{M}_{1,0}$ and $\mathcal{M}_{1,1}$):

$$\lambda_c^{-1}, \lambda_h^{-1} \sim \text{GAM}(\alpha_c = 2, \beta_c) \;\text{ with }\; \beta_c \sim \text{GAM}(a = 0.2, b = 1) \;\text{ so that }\; E[\beta_c] = \frac{a}{b} = 0.2$$

For the four models, we run the MCMC algorithms with 100,000 iterations. Withdrawing the first 50% of the samples ("burn-in phase") and thinning out the remaining 50,000 samples (from the "sampling phase") by the factor 100 yield $R = 500$ samples from each posterior. To check for convergence, we applied diagnostics based on trace plot and potential scale reduction factor diagnostics (see, e.g., Gelman & Rubin, 1992). All diagnostics indicated perfect convergence for the above setting.

3 | DATA

3.1 | Synthetic RAF pathway data

For our cross-method comparison, we generate synthetic network data from the RAF pathway, as reported by Sachs, Perez, Pe'er, Lauffenburger, and Nolan (2005). The RAF pathway shows the regulatory interactions among the following $n = 11$ proteins: Y_1: PIP3, Y_2: PLCG, Y_3: PIP2, Y_4: PKC, Y_5: PKA, Y_6: JNK, Y_7: P38, Y_8: RAF, Y_9: MEK, Y_{10}: ERK, and Y_{11}: AKT. There are 20 regulatory interactions (directed edges) in the RAF pathway. We extract the true 11-by-11 adjacency matrix $\mathcal{T}$ where $\mathcal{T}_{j,i} = 1$ if there is an edge from the jth to the ith protein. We get

$$\mathcal{T} = \begin{pmatrix} 0&1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0 \\ 0&0&1&1&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&1&1&0&0&1&1 \\ 1&1&0&0&0&0&0&1&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0 \end{pmatrix}.$$

The covariate set of variable Y_i is then $\pi_i = \{Y_j : \mathcal{T}_{j,i} = 1\}$. We follow Grzegorczyk and Husmeier (2012) and generate synthetic data sets with $H = 4$ segments having m data points each. For each

Y_i, we thus require four segment-specific regression coefficient vectors $\boldsymbol{\beta}_{i,1}, \ldots, \boldsymbol{\beta}_{i,4}$, with each being of length $|\pi_i| + 1$. Given those vectors, we generate 11-by-$(4m + 1)$ data matrices $\mathbf{D}$, where $\mathbf{D}_{i,t}$ is the value of Y_i at t. Let $\mathbf{D}_{.,t}$ denote the tth column of $\mathbf{D}$ (i.e., the values of the variables at t), and let $\mathbf{D}_{\pi_i,t}$ denote the subvector of $\mathbf{D}_{.,t}$ containing only the values of the $|\pi_i|$ covariates of Y_i. We randomly sample the values of $\mathbf{D}_{.,1}$ from independent Gaussian distributions with mean 0 and variance $\sigma^2 = 0.025$, before we successively generate data for the next time points. $\mathbf{D}_{i,t}$ $(i = 1, \ldots, 11; t = 2, \ldots, 4m + 1)$ is generated as follows:

$$\mathbf{D}_{i,t} = \left(1, \mathbf{D}^{\mathsf{T}}_{\pi_i,t-1}\right) \cdot \boldsymbol{\beta}_{i,H(t)} + \epsilon_{i,t}, \tag{26}$$

where the $\epsilon_{i,t}$ are independently $\mathcal{N}(0, \sigma^2)$ distributed noise variables, and $H(t)$ is a step function, indicating the segment to which time point t belongs, as follows:

$$H(t) = \begin{cases} 1, & 2 \le t \le m + 1 \\ 2, & m + 2 \le t \le 2m + 1 \\ 3, & 2m + 2 \le t \le 3m + 1 \\ 4, & 3m + 2 \le t \le 4m + 1 \end{cases}.$$

We sample the regression coefficient vectors for the first segment $h = 1$, $\boldsymbol{\beta}_{i,1}$ $(i = 1, \ldots, 11)$, from independent standard Gaussian distributions, and normalize each vector to Euclidean norm 1: $\boldsymbol{\beta}_{i,1} \leftarrow \frac{\boldsymbol{\beta}_{i,1}}{|\boldsymbol{\beta}_{i,1}|}$. For the segments $h > 1$, we change the vector from the previous segment, $\boldsymbol{\beta}_{i,h-1}$, as follows:

(U) either we change the regression coefficients drastically by flipping their signs, $\boldsymbol{\beta}_{i,h} = (-1) \cdot \boldsymbol{\beta}_{i,h-1}$; we then say that segment h is uncoupled ("U") from segment $h - 1$, that is, $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ are very dissimilar;

(C) or we change the regression coefficients moderately. To this end, we first sample the entries of a new vector $\boldsymbol{\beta}_{i,\star}$, from independent standard $\mathcal{N}(0, 1)$ Gaussians, and normalize $\boldsymbol{\beta}_{i,\star}$ to Euclidean norm ϵ, $\boldsymbol{\beta}_{i,\star} \leftarrow \epsilon \cdot \frac{\boldsymbol{\beta}_{i,\star}}{|\boldsymbol{\beta}_{i,\star}|}$, where ϵ is a tuning parameter. Then, we add the new vector to the vector $\boldsymbol{\beta}_{i,h-1}$ and renormalize the result to Euclidean norm 1, as follows:

$$\boldsymbol{\beta}_{i,h} := \frac{\boldsymbol{\beta}_{i,h-1} + \boldsymbol{\beta}_{i,\star}}{|\boldsymbol{\beta}_{i,h-1} + \boldsymbol{\beta}_{i,\star}|}.$$

We then say that segment h is coupled ("C") to segment $h - 1$, that is, $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ are similar.

We use the symbolic notation "$C - U - C$" to indicate that segment 2 is coupled to segment 1, segment 3 is uncoupled from segment 2, and segment 4 is coupled to segment 3. In our simulation study, we consider all possible scenarios "$S_2 - S_3 - S_4$" with $S_h \in \{C, U\}$ $(h = 2, 3, 4)$, where $S_h = U$ $(S_h = C)$ indicates that segment h is uncoupled from (coupled to) segment $h - 1$, that is, the regression coefficient vectors $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ are dissimilar (similar).

For coupled segments ("C"), the parameter ϵ regulates how similar the vectors $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ are. For our first study, we set $\epsilon = 0.25$. In a follow-up study, we then investigate the effect of ϵ and vary this parameter $(\epsilon \in \{0, 0.25, 0.5, 1\})$.

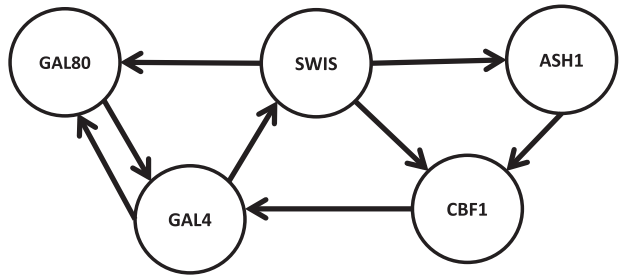

FIGURE 6 The true yeast network, as synthetically designed by Cantone et al. (2009)

3.2 | Yeast gene expression data

Cantone et al. (2009) synthetically generated a small network of $n = 5$ genes in *Saccharomyces cerevisiae* (yeast), depicted in Figure 6. The five genes are: *CBF*1, *GAL*4, *SWI*5, *GAL*80, and *ASH*1. The network among those genes was obtained from synthetically designed yeast cells grown with different carbon sources: galactose ("switch on") or glucose ("switch off"). Cantone et al. (2009) obtained in vivo data with quantitative real-time polymerase chain reaction (RT-PCR) in intervals of 20 min up to 5 h for the first and of 10 min up to 3 h for the second condition. This led to the sample sizes $T_1 = 16$ ("switch on") and $T_2 = 21$ ("switch off"). We follow Grzegorczyk and Husmeier (2012) and preprocess the data, $(\mathbf{D}^{(1)}_{.,1}, \ldots, \mathbf{D}^{(1)}_{.,16})$ and $(\mathbf{D}^{(2)}_{.,1}, \ldots, \mathbf{D}^{(2)}_{.,21})$, where $\mathbf{D}^{(c)}_{.,t}$ is the tth observation (vector) of the cth condition ($c = 1, 2$), as follows: For both conditions, we withdraw the initial measurements $\mathbf{D}^{(1)}_{.,1}$ and $\mathbf{D}^{(2)}_{.,1}$, as they were taken while extant glucose (galactose) was washed out and new galactose (glucose) was supplemented. This leaves us with the data vectors $\mathbf{D}^{(1)}_{.,2}, \ldots, \mathbf{D}^{(1)}_{.,16}, \mathbf{D}^{(2)}_{.,2}, \ldots, \mathbf{D}^{(2)}_{.,21}$, which we standardize via a log transformation and a subsequent genewise mean standardization (to mean 0). We also take into account that $\mathbf{D}^{(2)}_{.,2}$ has no proper relation with $\mathbf{D}^{(1)}_{.,16}$. For each target gene Y_i with covariate set π_i, we therefore only use $\tilde{T} = T_1 + T_2 - 4 = 33$ target $\mathbf{D}^{(1)}_{i,3}, \ldots, \mathbf{D}^{(1)}_{i,16}, \mathbf{D}^{(2)}_{i,3}, \ldots, \mathbf{D}^{(2)}_{i,21}$ and covariate values $\mathbf{D}^{(1)}_{\pi_i,2}, \ldots, \mathbf{D}^{(1)}_{\pi_i,15}, \mathbf{D}^{(2)}_{\pi_i,2}, \ldots, \mathbf{D}^{(2)}_{\pi_i,20}$.

4 | EMPIRICAL RESULTS

4.1 | Results for synthetic RAF pathway data

In our first empirical evaluation study, we cross-compare the network reconstruction accuracies of the four models $\mathcal{M}_{i,j}$ ($i, j \in \{0, 1\}$), listed in Table 1, on synthetic RAF pathway data, generated as described in Section 3.1. In this study, we assume the data segmentation into $H = 4$ segments to be known. We can then set the three changepoints at the right locations and do not perform MCMC moves on the changepoint set $\boldsymbol{\tau}$. That is, we keep $\boldsymbol{\tau}$ fixed during the MCMC simulations. The corresponding moves on $\boldsymbol{\tau}$, described in Section 2.6, are skipped.

In our first study, we generate data for eight different coupling scenarios of the form "$S_2 - S_3 - S_4$" with $S_i \in \{U, C\}$, where $X_h = U$ indicates that segment h is uncoupled from segment $h - 1$ (i.e., $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ are dissimilar), and $S_h = C$ indicates that segment h is coupled to segment $h - 1$ (i.e., $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ are similar). For the technical details, we refer to Section 3.1.

First, we perform a sanity check for the proposed $\mathcal{M}_{1,1}$ model: We investigate whether it actually infers different coupling parameters for the segments and whether the segment-specific coupling parameter distributions are consistent with the underlying coupling schemes of the form "$S_2 - S_3 - S_4$". For uncoupled segments with $S_h = U$ the coupling parameters should on

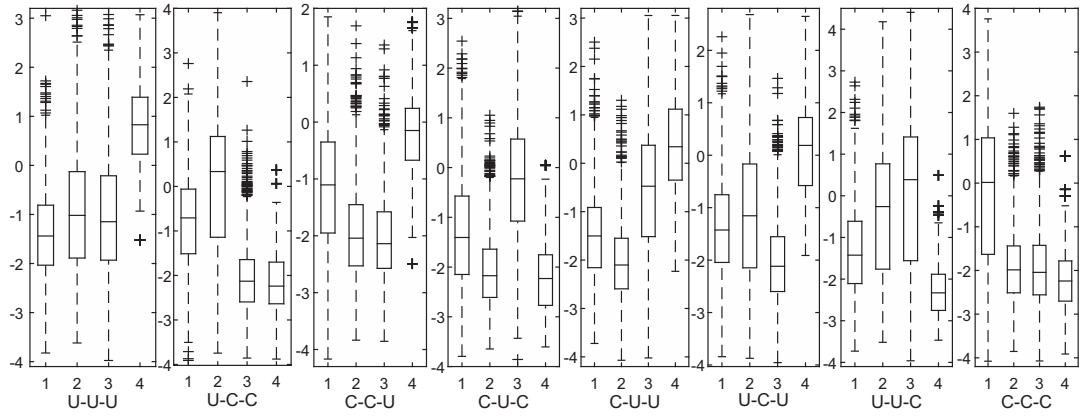

FIGURE 7 Boxplots of the logarithmic segment-specific coupling parameters for the RAF pathway data. We generated data with $H = 4$ segments and $m = 10$ data points per segment and distinguished eight coupling scenarios of the type: "$S_2 - S_3 - S_4$" with $S_h \in \{U, C\}$. For each scenario, there is a panel with four boxplots. The boxplots indicate how the logarithmic sampled coupling parameters $\log(\lambda_1) := \log(\lambda_u)$, $\log(\lambda_2)$, $\log(\lambda_3)$, and $\log(\lambda_4)$ are distributed for the given scenario. For a compact representation, we decided to merge all samples taken for $N = 11$ variables from independent Markov chain Monte Carlo simulations on 25 independent data instantiations. In each panel, the hth boxplot from the left refers to $\log(\lambda_h)$. Our focus is on λ_h with $h > 1$, and it can be seen that the coupling parameters for coupled segments ($S_h = C$) are lower than for uncoupled segments ($S_h = U$). For $m = 5$ data points per scenario, we observed the same trends (boxplots not shown)

average be greater than for coupled segments with $S_h = C$ ($h = 2, 3, 4$). Figure 7 shows boxplots of the inferred segment-specific coupling parameters $\lambda_1 (= \lambda_u)$, λ_2, λ_3, and λ_4. Focusing on λ_h with $h = 2, 3, 4$, it can be seen from the boxplots that the coupling parameters for coupled segments (where $S_h = C$) are consistently lower than for uncoupled segments (where $S_h = U$).

The AUC results, provided in Figure 8, show that the proposed generalized model ($\mathcal{M}_{1,1}$) shows, overall, the best performance. It is always among the best models and never performs substantially worse than any other model. On the other hand, the proposed $\mathcal{M}_{1,1}$ model outperforms its competitors for some settings, especially for scenarios where two out of three segments with $h > 1$ are coupled to the previous segment. For the scenarios "$C - C - U$", "$U - C - C$", and "$C - U - C$", the proposed model performs better than the three other models. When comparing the two in-between models ($\mathcal{M}_{1,0}$ and $\mathcal{M}_{0,1}$) with the original $\mathcal{M}_{0,0}$ model from Grzegorczyk and Husmeier (2012), it becomes obvious that imposing a hyperprior on β_c, as implemented in the $\mathcal{M}_{1,0}$ model, almost consistently improves the AUC scores, whereas making the coupling parameter segment specific, as done in the $\mathcal{M}_{0,1}$ model, can lead to deteriorations of the AUC scores. We draw the conclusion that replacing the coupling parameter λ_c by segment-specific parameters $\lambda_2, \ldots, \lambda_4$ is counterproductive, unless this modification is combined with a hyperprior so that information can be shared among the segment-specific coupling strengths. Just imposing a hyperprior, which then only allows us to adjust the prior for the coupling strength parameter λ_c (in light of the data), also improves the network reconstruction accuracy, but the improvement is slightly minor to the improvement that can be achieved by implementing both modifications together, as proposed in this paper.

In a follow-up study, we then had a closer look at the eight scenarios and varied the tuning parameter $\epsilon \in \{0, 0.25, 0.5, 1\}$ for each of them. With the parameter ϵ, the similarity of the regression parameters (i.e., the coupling strength between coupled segments) can be adjusted. The greater ϵ, the weaker the similarity of the regression coefficient $\boldsymbol{\beta}_{i,h}$ and $\boldsymbol{\beta}_{i,h-1}$ of coupled

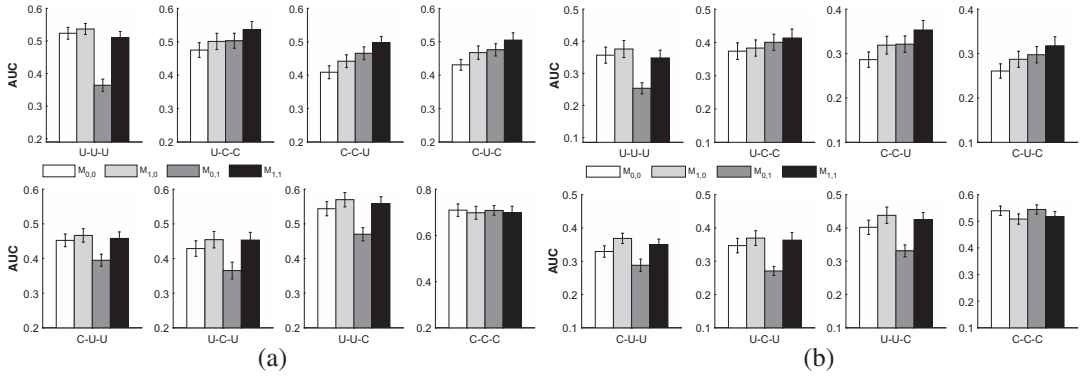

FIGURE 8 Network reconstruction accuracy for the RAF pathway data. For the RAF pathway, we generated synthetic data with $H = 4$ segments and with $m \in \{5, 10\}$ data points per segment. For both m, we distinguished eight coupling scenarios of the type "$S_2 - S_3 - S_4$" (with $S_h \in \{U, C\}$). For coupled ("C") segments, we set the parameter ϵ to 0.25; see Section 3.1 for details. The histogram bars correspond to the model-specific average area under the precision recall curve (AUC) values, averaged across 25 Markov chain Monte Carlo simulations. The error bars correspond to standard deviations

segments; see Section 3.1 for the mathematical details. As an example, Figure 9 shows the results for the scenario "$C - C - U$", which belongs to the scenarios where the proposed $\mathcal{M}_{1,1}$ model was found to outperform its competitors (see Figure 8). We can see from the AUC results in Figure 9 that $\epsilon = 0$ and $\epsilon = 0.5$ yield the same trends, as observed earlier for $\epsilon = 0.25$; see Figure 8. However, for the highest ϵ ($\epsilon = 1$), $\mathcal{M}_{1,0}$ and $\mathcal{M}_{1,1}$ perform equally well and better than the other two models ($\mathcal{M}_{0,0}$ and $\mathcal{M}_{0,1}$). The explanation for this finding is most likely as follows: The similarity of the coupled regression coefficients decreases in ϵ. Hence, for $\epsilon = 1$, even the regression parameters for the two coupled segments get very dissimilar. Thus, $\epsilon = 1$ implies that all four segment-specific regression coefficients $\boldsymbol{\beta}_{i,h}$ ($h = 1, \ldots, 4$) are dissimilar, and there is no more need for segment-specific coupling parameters. The reason why for $\epsilon = 1$, the original $\mathcal{M}_{0,0}$ model and the $\mathcal{M}_{0,1}$ model are inferior to the models that possess hyperpriors is probably the following: The models $\mathcal{M}_{0,0}$ and $\mathcal{M}_{0,1}$ cannot adjust the prior of the coupling parameter (in light of the data). As a consequence, they are likely to overpenalize dissimilar regression coefficients (i.e., high coupling parameters) through the prior.

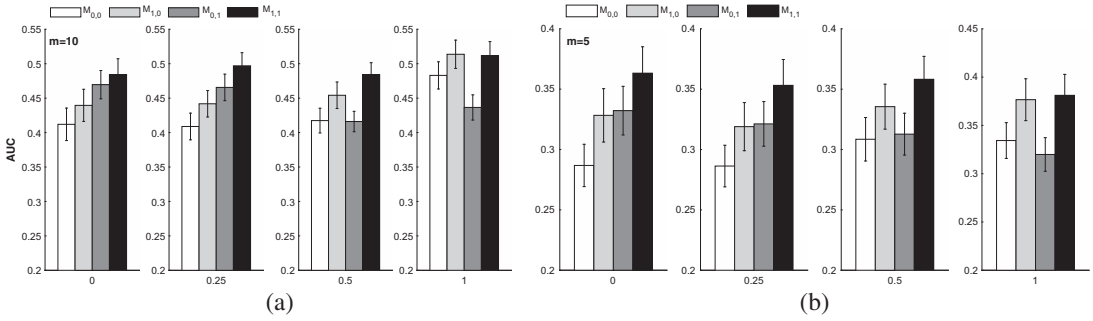

FIGURE 9 Network reconstruction accuracy for the RAF pathway data. Unlike in Figure 8, here, we focus on the scenario "$C - C - U$" and vary the tuning parameter $\epsilon \in \{0, 0.25, 0.5, 1\}$. Like in Figure 8, the model-specific bars and error bars correspond to the average area under the precision recall curve (AUC) values and their standard deviations

4.2 | Results for the yeast gene expression data

In this subsection, we cross-compare the network reconstruction accuracies of the four models $\mathcal{M}_{i,j}$ $(i,j = 0,1)$, listed in Table 1, on the yeast gene expression data, described in Section 3.2. For this application, we infer the data segmentation (i.e., the changepoint set $\boldsymbol{\tau}$) along with the network structure from the data.

To vary the segmentations and especially the number of segments (i.e., the number of changepoints in $\boldsymbol{\tau}$), we implement the four models with eight different hyperparameters $p \in \{0.00625, 0.0125, 0.025, 0.05, 0.1, 0.2, 0.4, 0.8\}$ of the geometric prior on the distance between changepoints; see Equation (23) in Section 2.6 for the mathematical details. We note that the hyperparameters are of the form $p = 0.1 \cdot 2^i$ with $i = -4, -3, \ldots, 3$.

Figure 10 shows histograms of the inferred posterior distributions of the numbers of segments H for the proposed $\mathcal{M}_{1,1}$ model from Section 2.2. It can be clearly seen that the inferred

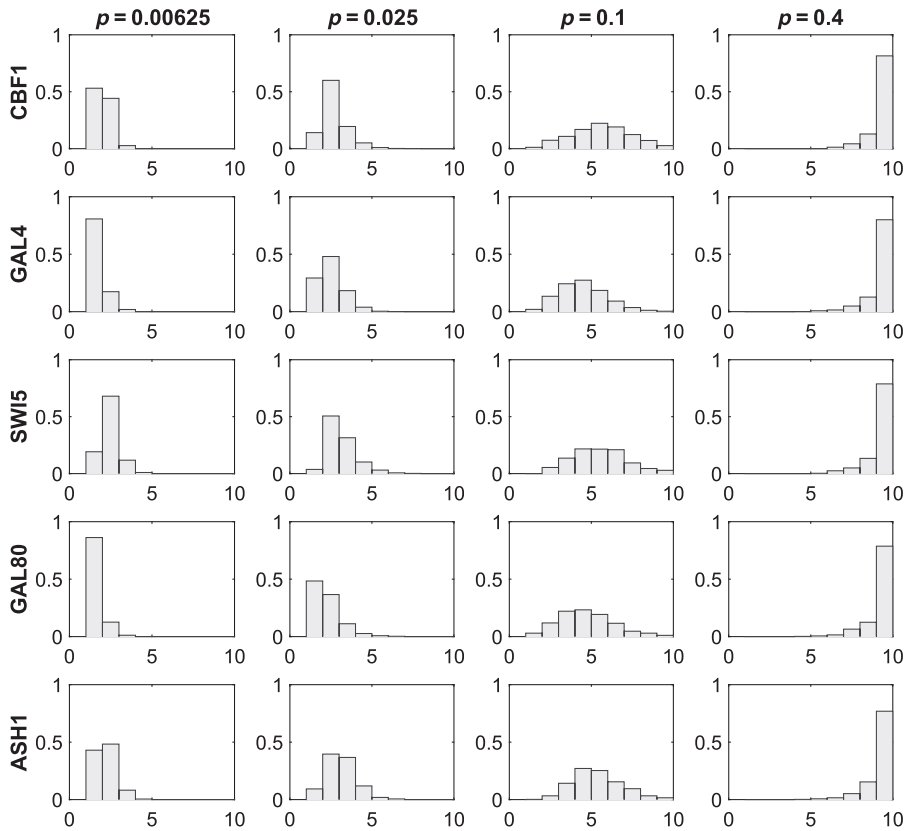

FIGURE 10 Posterior distribution of the numbers of segments H for the yeast data from Section 3.2. We implemented the models with different hyperparameters p for the geometric distribution on the distance between changepoints. For the proposed $\mathcal{M}_{1,1}$ model, the histograms show how the posterior distributions of the numbers of segments H vary with the hyperparameter p. Each row refers to a gene of the yeast network, and the four columns refer to the hyperparameters $p \in \{0.00625, 0.025, 0.1, 0.4\}$. For each number of segments $1 \leq H \leq 10$, the bars give the relative frequencies with which the data of the corresponding gene were segmented into H segments in the posterior samples. The relative frequencies are averaged over 25 independent Markov chain Monte Carlo simulations

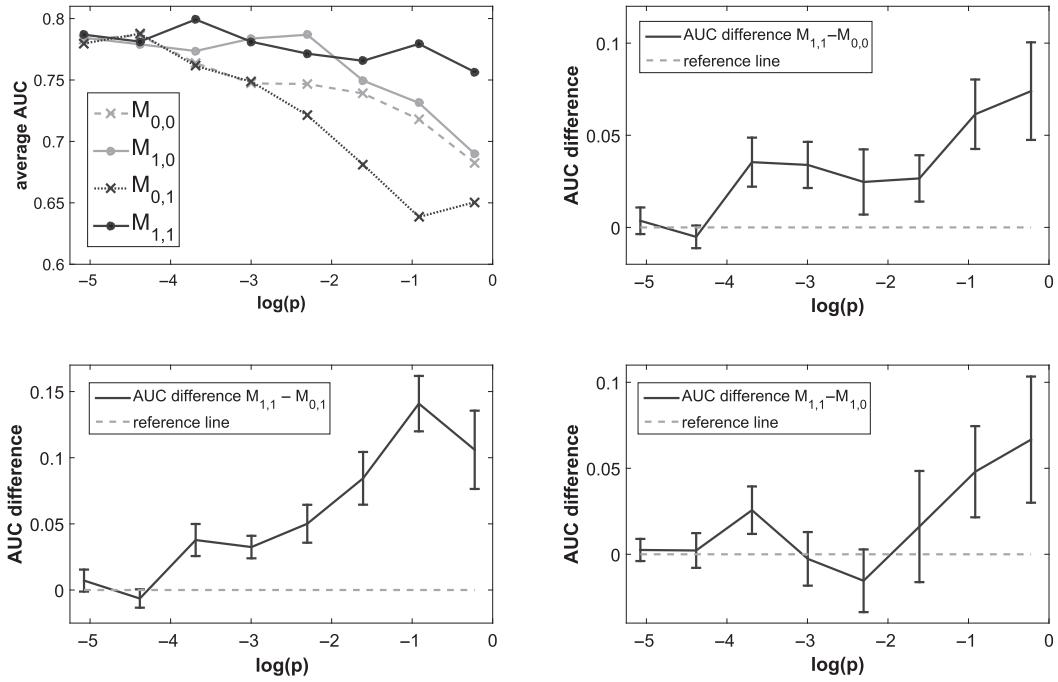

FIGURE 11 Network reconstruction accuracy for the yeast gene expression data from Section 3.2. For our study, we implemented the four models, listed in Table 1, with eight different hyperparameters $p = 0.1 \cdot 2^i$ ($i \in \{-4, -3, \ldots, 3\}$) for the geometric distribution on the distance between changepoints; see Equation (23). The upper left panel shows the average area under the precision recall curve (AUC) scores, averaged across 25 Markov chain Monte Carlo simulations. The other panels show the relative area under the precision recall curve differences in favor of the $\mathcal{M}_{1,1}$ model, with error bars indicating 95% t-test confidence intervals

segmentations strongly depend on the hyperparameter p. For the lowest p, the data are rarely divided into more than $H = 2$ segments, whereas the posterior for $p = 0.4$ peaks at the imposed maximum of $H = 10$ segments. For the other three models, we observed almost identical trends (histograms not shown in this paper).

Figure 11 shows how the empirical network reconstruction accuracy, quantified in terms of the average AUC values, varies with the hyperparameter p. From the upper left panel of Figure 11, which shows the average AUC scores, the following trends can be observed:

- The $\mathcal{M}_{0,1}$ model (which has segment-specific coupling parameters $\lambda_2, \ldots, \lambda_H$ but does not couple them by a gamma hyperprior) performs worse than the original sequentially coupled model from Grzegorczyk and Husmeier (2012). Only for very small hyperparameters $p \leq 0.05$ (i.e., when few changepoints are learned) the two models $\mathcal{M}_{0,0}$ and $\mathcal{M}_{0,1}$ reach approximately the same AUC scores. For higher hyperparameters, the introduction of segment-specific coupling parameters has a counterproductive effect and is thus not recommended.
- The $\mathcal{M}_{1,0}$ model, which leaves the coupling parameter λ_c shared among segments and only imposes a hyperprior to adjust the prior on λ_c (in light of the data), performs better than the original model from Grzegorczyk and Husmeier (2012). For very small hyperparameters $p \leq 0.0125$ (i.e., when very few changepoints are learned), the two models $\mathcal{M}_{0,0}$ and $\mathcal{M}_{1,0}$ reach approximately the same AUC scores. For the six higher hyperparameters $p > 0.0125$,

the introduction of the hyperprior consistently leads to improved network reconstruction accuracies.

- The $\mathcal{M}_{1,1}$ model, which we propose in this paper, has both modifications implemented: It has segment-specific coupling parameters $\lambda_2, \ldots, \lambda_H$ and couples those parameters by a gamma hyperprior; see Section 2.2 for a detailed model description. The combination of both modifications yields that the $\mathcal{M}_{1,1}$ model performs best overall. Its performance even stays stable for rather large hyperparameters p, where the AUC scores of the other three models substantially diminish. The AUC difference plots in Figure 11 show that most of the improvements are statistically significant in terms of paired t tests. In particular, the $\mathcal{M}_{1,1}$ model performs better than the original $\mathcal{M}_{0,0}$ model, proposed by Grzegorczyk and Husmeier (2012), for six out of eight hyperparameters p, namely for all $p > 0.0125$.

Based on our network reconstruction accuracy results for the real gene expression data, we conclude that the performance of the proposed generalized model ($\mathcal{M}_{1,1}$) is superior to the performance of the original sequentially coupled model ($\mathcal{M}_{0,0}$). The empirical results obtained with the two "in-between" models ($\mathcal{M}_{1,0}$ and $\mathcal{M}_{0,1}$) suggest further that the main contribution stems from the hyperprior. The capability to adjust the coupling parameter prior in light of the data boosts the performance. The model $\mathcal{M}_{0,1}$, whose segment-specific coupling parameters are not coupled by a hyperprior, led to decreased AUC results. The results, obtained for the yeast gene expression data, are hence in agreement with the earlier results obtained for synthetic network data; see Section 4.1.

5 | CONCLUSIONS

In this paper, we have proposed an improved version of the NH-DBN model, proposed by Grzegorczyk and Husmeier (2012). Unlike the original $\mathcal{M}_{0,0}$ model, our new $\mathcal{M}_{1,1}$ model possesses segment-specific coupling (strength) parameters and a hyperprior on the coupling parameter priors; see Section 2.2 for the mathematical details. Replacing the shared coupling parameter λ_c by segment-specific coupling parameters $\lambda_1, \ldots, \lambda_H$ increases the model flexibility, whereas the new hyperprior is allowing information exchange among segments and adjusting the coupling parameter prior(s) in light of the data. Our empirical evaluation studies on synthetic RAF pathway data (see Section 4.1) and on yeast gene expression data (see Section 4.2) have shown that the new $\mathcal{M}_{1,1}$ model leads to improved network reconstruction accuracies. To gain more insight into the merits of the two individual modifications, we also compared it with the performances of the two "in-between" models ($\mathcal{M}_{1,0}$ and $\mathcal{M}_{0,1}$), which we defined to be subject to only one of the two modifications; see Table 1 for an overview to the four models $\mathcal{M}_{i,j}$ ($i,j \in \{0,1\}$) under comparison. Overall, the proposed $\mathcal{M}_{1,1}$ has reached the highest network reconstruction accuracies among the four models. The $\mathcal{M}_{0,1}$ model, which we defined to have segment-specific coupling parameters but no hyperprior, performed worse than the original $\mathcal{M}_{0,0}$ model. The $\mathcal{M}_{1,0}$ model, which we defined to have a shared coupling parameter with a hyperprior, performed better than the original $\mathcal{M}_{0,0}$ model and for some scenarios comparable with the proposed $\mathcal{M}_{1,1}$ model. This shows that the major part of the improvement, achieved with the proposed $\mathcal{M}_{1,1}$, stems from imposing a hyperprior onto the coupling parameter prior(s).

To put it in a nutshell, our empirical results show that the model variant $\mathcal{M}_{1,1}$ reaches, overall, the highest network reconstruction accuracies. With regard to future applications, we therefore recommend giving precedence to this model. Moreover, our results for the yeast gene

expression data (see Figure 11) also suggest that only the $\mathcal{M}_{1,1}$ model is robust with respect to the changepoint process hyperparameter. The network reconstruction accuracies of the other models deteriorate, as the number of inferred changepoints increases. Only the network reconstruction accuracy of the $\mathcal{M}_{1,1}$ model stays high, even if the data are divided into short (uninformative) segments. This is a very important property for applications where the underlying segmentation is unknown and has to be inferred from the data. The number of inferred changepoints (i.e., the data segmentation) strongly depends on the changepoint process hyperparameter (see Figure 10). In the absence of any genuine prior knowledge, the changepoint process hyperparameter can easily be misspecified. Nonrobust models will then output biased results that might lead to erroneous conclusions.

Our future work will aim to transfer the concept of segment-specific coupling (strength) parameters to the globally coupled NH-DBN model from Grzegorczyk and Husmeier (2013). Unlike the sequential coupling mechanism, which requires a temporal ordering of the segments, the global coupling mechanism treats all segments as interchangeable units. When segmenting a single time series by changepoints, the assumption of interchangeable segments is often not appropriate. However, there are other applications in systems biology where data stem from different experiments so that the segments might contain data from different experimental conditions. The segments then do not have any natural order, and the sequential coupling scheme should be replaced by the global coupling scheme.

REFERENCES

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Singapore: Springer.

Cantone, I., Marucci, L., Iorio, F., Ricci, M., Belcastro, V., Bansal, M., & Cosma, M. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, *137*(1), 172–181.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning,* USA, 233–240. https://doi.acm.org/10.1145/1143844.1143874

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data analysis* (2nd ed.). London, UK: Chapman and Hall/CRC.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Grzegorczyk, M., & Husmeier, D. (2012). A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology. *Statistical Applications in Genetics and Molecular Biology*, *11*(4).

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

Lèbre, S., Becq, J., Devaux, F., Lelandais, G., & Stumpf, M. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, *4*(130).

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., & Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), 523–529.

How to cite this article: Shafiee Kamalabad M, Grzegorczyk M. Improving nonhomogeneous dynamic Bayesian networks with sequentially coupled parameters. *Statistica Neerlandica*. 2018;1–25. https://doi.org/10.1111/stan.12136