


RESEARCH

Open Access



A new Bayesian piecewise linear regression model for dynamic network reconstruction

Mahdi Shafiee Kamalabad^{1,2} and Marco Grzegorzcyk^{3*} 

From Computational Intelligence methods for Bioinformatics and Biostatistics Bergamo, Italy. 4–6 September 2019

*Correspondence:
m.a.grzegorzcyk@rug.nl
³ Bernoulli Institute,
Groningen University,
Nijenborgh 9, 9747
AG Groningen, The
Netherlands
Full list of author information
is available at the end of the
article

Abstract

Background: Linear regression models are important tools for learning regulatory networks from gene expression time series. A conventional assumption for non-homogeneous regulatory processes on a short time scale is that the network structure stays constant across time, while the network parameters are time-dependent. The objective is then to learn the network structure along with changepoints that divide the time series into time segments. An uncoupled model learns the parameters separately for each segment, while a coupled model enforces the parameters of any segment to stay similar to those of the previous segment. In this paper, we propose a new consensus model that infers for each individual time segment whether it is coupled to (or uncoupled from) the previous segment.

Results: The results show that the new consensus model is superior to the uncoupled and the coupled model, as well as superior to a recently proposed generalized coupled model.

Conclusions: The newly proposed model has the uncoupled and the coupled model as limiting cases, and it is able to infer the best trade-off between them from the data.

Keywords: Bayesian piece-wise linear regression, Gene regulatory networks, Network reconstruction, Segment-wise parameter coupling

Background

Non-homogeneous dynamic Bayesian networks have become a popular tool for learning the structures of cellular regulatory networks from gene expression and protein concentration data. The traditional (homogeneous) dynamic Bayesian network models assume the network parameters to stay constant across time. This can lead to biased results and wrong conclusions, as cellular regulatory processes can change in time. It was therefore proposed to combine dynamic Bayesian network models with Bayesian changepoint processes, see, e.g., [1–3]. Then a multiple changepoint process is used to divide the temporal data into disjoint segments, and the data within each segment are modelled



© The Author(s), 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

by linear regression models. For most cellular processes on a short time scale it is not realistic to assume that the network structure changes over time. The network structure is therefore usually assumed to stay unchanged and only the network parameters are assumed to time-varying. As a motivation for this assumption consider a gene regulatory network, in which an edge from gene Z_i to gene Z_j , $Z_i \rightarrow Z_j$, typically would indicate that gene Z_i codes for a transcription factor that can bind to the promoter of gene Z_j , so that Z_j 's transcription is initiated. The ability to bind to the promoter (= the edge connection) is unlikely to change within a short time period, whereas the extent of binding (= the network interaction parameter) can undergo quick temporal changes. Regarding our two real-life applications to *S. cerevisiae* (yeast) and *A. thaliana* (plant) gene expression data, the assumption of a fixed network structure therefore seems more faithful.

The uncoupled model, akin to the models proposed by Lèbre et al. [1] and Dondelinger et al. [3], learns the segment-specific network parameters for each segment separately. To allow for information-sharing with respect to the network parameters, models with globally [4] and sequentially [5] coupled network parameters were proposed. As sequential information-sharing seems more suitable for temporal time segments, we focus here on the sequential coupling. The underlying idea is that the network parameters of each segment should be enforced to stay similar to those of the previous segment. Grzegorzyc and Husmeier [5] proposed a coupled model, in which the posterior expectations of the network parameters of segment h are used as prior expectations for the next segment $h + 1$. The strength of the coupling, i.e. the variance of the network parameter priors, is regulated by a coupling parameter. Although it was shown that this is very useful for applications where the network parameters stay similar over time, the fully coupled model has the drawback that it enforces coupling and does not feature any possibility for uncoupling. In this paper we therefore propose a partially segment-wise coupled model, which can be seen as a consensus model between the uncoupled and the fully coupled model. Discrete binary indicator variables δ_h indicate for each segment h whether it is coupled to the previous segment ($\delta_h = 1$) or uncoupled from it ($\delta_h = 0$). Along with the network structure and the data segmentation the values of those indicator variables are inferred from the data. The new partially coupled model reaches the original models in the limit: If it couples all segments ($\delta_h = 1$ for all $h > 2$), it becomes the fully coupled model. If it uncouples all segments ($\delta_h = 0$ for all h), it becomes the uncoupled model.

In our earlier work [6] we have proposed a new generalized fully coupled model. While the fully coupled model from [5] couples all neighbouring segments $(h - 1, h)$ with the same coupling strength $\lambda \in \mathbb{R}^+$, the generalised (fully) coupled model from [6] uses for each pair of neighbouring segments $(h - 1, h)$ a segment-specific coupling strength parameter $\lambda_h \in \mathbb{R}^+$. This leads to a higher model flexibility, but like the coupled model the generalized coupled model still does not allow for uncoupling. In our comparative evaluation study, we will compare the new partially coupled model with the three competing models: the uncoupled model, the (fully) coupled model, and the generalized (fully) coupled model.

In recent works alternative model refinements have been proposed [7, 8]. These models distinguish coupled from uncoupled network edges rather than distinguishing coupled from uncoupled time segments. The partially non-homogeneous model from Shafiee Kamalabad et al. [7] builds on the idea that only some network parameters (i.e.

some edges) might be subject to changes, while other network parameters (i.e. edges) might stay constant. The model has been designed for analysing data that have been measured under different experimental conditions, so that it does not allow the segmentation of a time series to be inferred. The non-homogeneous model from Shafiee Kamalabad and Grzegorzcyk [8] distinguishes between two groups of edges: (i) edges that are fully coupled among all segments and (ii) edges that are uncoupled among all segments. The new model that we propose here is conceptual related, but complementary in that it replaces the concept of partially coupled edges by the concept of partially coupled time segments.

We note that network reconstruction is a topical research field in the computational biology literature and that many different network reconstruction approaches have been proposed over the years. However, most of the proposed models do not focus on non-homogeneous regulatory processes but rely on a homogeneity of the regulatory processes. For some applications this assumption of homogeneity can be too restrictive; compare, e.g., our data applications. In response to one of the reviewers of our paper, we here briefly discuss a few recently proposed network reconstruction methods. Vignes et al. [9] investigated and compared a wide variety of methods, ranging from Bayesian networks to penalised linear regression based models and proposed a meta-analysis based on Fisher's Inverse Chi-Square meta-test for combining different approaches. Huang et al. [10] proposed to apply Bayesian model averaging for linear regression methods. The method uses a closed form solution to compute the edge posterior probabilities within a hybrid framework of Bayesian model averaging and linear regression. Xing et al. [11] proposed a Candidate Auto Selection algorithm based on the pairwise mutual information and breakpoint detection. With a greedy search algorithm it is searched for the best network topology. Unlike the above mentioned models, Fan et al. [12] propose to impose a prior on the topology information in their inference process. Incorporating this prior information can partially compensate for the lack of reliable data. They then developed a Bayesian group lasso with spike and slab prior approach based on non-parametric models. Xu et al. [13] propose to employ a series of linear regression problems to model the relationship between the network nodes. They use an efficient variational Bayes method for optimization and inference of the unknown network parameters.

Methods

Learning dynamic networks with time-varying parameters

Consider N random variables Z_1, \dots, Z_N that are the nodes of a network. Let \mathbf{D} denote an N -by- $(T + 1)$ data matrix, whose N rows correspond to the variables and whose $T + 1$ columns correspond to time points $t = 1, \dots, T + 1$. The element in the i th row and t th column, $\mathbf{D}_{i,t}$, is the value of Z_i at time point t . For temporal data it is typically assumed that the regulatory interactions are subject to a lag of one time point. For example, an edge $Z_i \rightarrow Z_j$ indicates that $\mathbf{D}_{j,t+1}$ (Z_j at $t + 1$) depends on $\mathbf{D}_{i,t}$ (Z_i at t). The variable Z_i is then called a parent (node) of Z_j .

Because of the lag, there is no need for any acyclicity constraint, and for each node Z_j ($j = 1, \dots, N$) the parent nodes can be learned separately. This has computational advantages, since the 'network learning task' can be separated into N independent

‘parent learning tasks’. Henceforth, when a computer cluster is available, the N parent sets can be learned in parallel, so that the inference algorithms scale-up well.

A popular method is to apply linear regression, where $Y := Z_j$ is the response and $\{Z_1, \dots, Z_{j-1}, Z_{j+1}, \dots, Z_N\} =: \{X_1, \dots, X_n\}$ are potential covariates (with $n := N - 1$). Because of the lag, $T + 1$ time points yield T observations for the linear regression model. Each observation \mathcal{D}_t ($t \in \{1, \dots, T\}$) consists of a response value $Y = \mathbf{D}_{j,t+1}$ and the shifted covariate values: $X_1 = \mathbf{D}_{1,t}, \dots, X_{j-1} = \mathbf{D}_{j-1,t}, X_j = \mathbf{D}_{j+1,t}, \dots, X_n = \mathbf{D}_{N,t}$, where $n = N - 1$.

Having inferred a covariate set π_j for each Z_j , a network is built by merging the covariate sets: $\mathcal{G} := \{\pi_1, \dots, \pi_N\}$. There is the edge $Z_i \rightarrow Z_j$ in \mathcal{G} if and only if $Z_i \in \pi_j$.

As the same linear regression approaches are used for each Z_j , we describe the models using a general terminology: Let Y be the response and let X_1, \dots, X_n be the covariates of the linear regression model.

To allow for time-dependent regression coefficients, a piece-wise linear regression model can be used. Changepoints $\tau := \{\tau_1, \dots, \tau_{H-1}\}$ with $1 \leq \tau_h < T$ divide the observations $\mathcal{D}_1, \dots, \mathcal{D}_T$ into disjoint segments $h = 1, \dots, H$ containing T_1, \dots, T_H consecutive data points, so that: $\sum T_h = T$. Observation \mathcal{D}_t ($1 \leq t \leq T$) belongs to segment h if $\tau_{h-1} < t \leq \tau_h$, where $\tau_0 := 1$ and $\tau_H := T$ are two pseudo changepoints.

We assume all covariate sets $\pi \subset \{X_1, \dots, X_n\}$ with up to $\mathcal{F} = 3$ covariates to be equally likely a priori, $p(\pi) = c$, while parent sets with more than \mathcal{F} covariates get a zero prior probability (‘fan-in restriction’). Further we assume that the distance between changepoints is geometrically distributed with hyperparameter $p \in (0, 1)$, so that

$$p(\tau) = (1 - p)^{\tau_H - \tau_{H-1} - 1} \cdot \prod_{h=1}^{H-1} p \cdot (1 - p)^{\tau_h - \tau_{h-1} - 1} = (1 - p)^{(T-1) - (H-1)} \cdot p^{H-1}$$

With $\mathbf{y} = \mathbf{y}_\tau := \{\mathbf{y}_1, \dots, \mathbf{y}_H\}$ being the set of segment-specific response vectors, implied by the changepoint set τ , the posterior distribution takes the form:

$$p(\pi, \tau, \theta | \mathbf{y}) \propto p(\pi) \cdot p(\tau) \cdot p(\theta | \pi, \tau) \cdot p(\mathbf{y} | \pi, \tau, \theta) \tag{1}$$

where $\theta = \theta(\pi, \tau)$ denotes the set of all model parameters, including segment-specific parameters as well as parameters that are shared among segments.

In the following subsections we assume $\pi \subset \{X_1, \dots, X_n\}$ and the segmentation $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_H\}$, induced by τ , to be fixed, and we do not make π and τ explicit anymore. Without loss of generality, we assume that π contains the first k covariates: $\pi := \{X_1, \dots, X_k\}$. For fixed π and τ , Eq. (1) reduces to:

$$p(\theta | \mathbf{y}) \propto p(\theta) \cdot p(\mathbf{y} | \theta)$$

A generic Bayesian piece-wise linear regression model

Consider a Bayesian linear regression model, where Y is the response and X_1, \dots, X_k are the covariates. We assume that T observations $\mathcal{D}_1, \dots, \mathcal{D}_T$ have been made at equidistant time points and that the data can be subdivided into disjoint segments $h \in \{1, \dots, H\}$, where segment h contains T_h data points and has a segment-specific regression coefficient vector \mathbf{w}_h . Let \mathbf{y}_h be the response vector and \mathbf{X}_h be the design matrix for segment

h , where each \mathbf{X}_h includes a first column of 1's for the intercept. For each segment $h = 1, \dots, H$ we assume a Gaussian likelihood:

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

where \mathbf{I} is the identity matrix, and σ^2 is a noise variance parameter that is shared among all segments. We impose an inverse Gamma prior on σ^2 , $\sigma^{-2} \sim \text{GAM}(\alpha_\sigma, \beta_\sigma)$, and we assume that the vectors \mathbf{w}_h have Gaussian priors:

$$\mathbf{w}_h | (\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \sigma^2) \sim \mathcal{N}(\boldsymbol{\mu}_h, \sigma^2 \boldsymbol{\Sigma}_h) \tag{3}$$

where $\boldsymbol{\mu}_h$ is a $(k+1)$ -dimensional vector, and $\boldsymbol{\Sigma}_h$ is a positive definite $(k + 1)$ -by- $(k + 1)$ matrix. Re-using the parameter σ^2 in Eq. (3), yields a fully-conjugate prior in both \mathbf{w}_h and σ^2 (see, e.g., Sections 3.3 and 3.4 in Gelman [14]). Figure 1 shows a graphical model representation of this generic model. For notational convenience we define:

$$\boldsymbol{\theta} := \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_H; \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_H\}$$

The full conditional distribution of \mathbf{w}_h is (cp. Section 3.3 in [15]):

$$\mathbf{w}_h | (\mathbf{y}_h, \sigma^2, \boldsymbol{\theta}) \sim \mathcal{N}\left(\left[\boldsymbol{\Sigma}_h^{-1} + \mathbf{X}_h^\top \mathbf{X}_h\right]^{-1} \left(\boldsymbol{\Sigma}_h^{-1} \boldsymbol{\mu}_h + \mathbf{X}_h^\top \mathbf{y}_h\right), \sigma^2 \left(\boldsymbol{\Sigma}_h^{-1} + \mathbf{X}_h^\top \mathbf{X}_h\right)^{-1}\right) \tag{4}$$

and the segment-specific marginal likelihoods with \mathbf{w}_h integrated out are:

$$\mathbf{y}_h | (\sigma^2, \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{X}_h \boldsymbol{\mu}_h, \sigma^2 \mathbf{C}_h(\boldsymbol{\theta})) \tag{5}$$

where $\mathbf{C}_h(\boldsymbol{\theta}) := \mathbf{I} + \mathbf{X}_h \boldsymbol{\Sigma}_h \mathbf{X}_h^\top$ (cp. Section 3.3 in [15]). From Eq. (5) we get:

$$p(\sigma^2 | \mathbf{y}, \boldsymbol{\theta}) \propto p(\sigma^2) \cdot \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \boldsymbol{\theta}) = (\sigma^{-2})^{a_\sigma + \frac{1}{2} \cdot T - 1} e^{-\sigma^{-2} (b_\sigma + \frac{1}{2} \cdot \Delta^2(\boldsymbol{\theta}))}$$

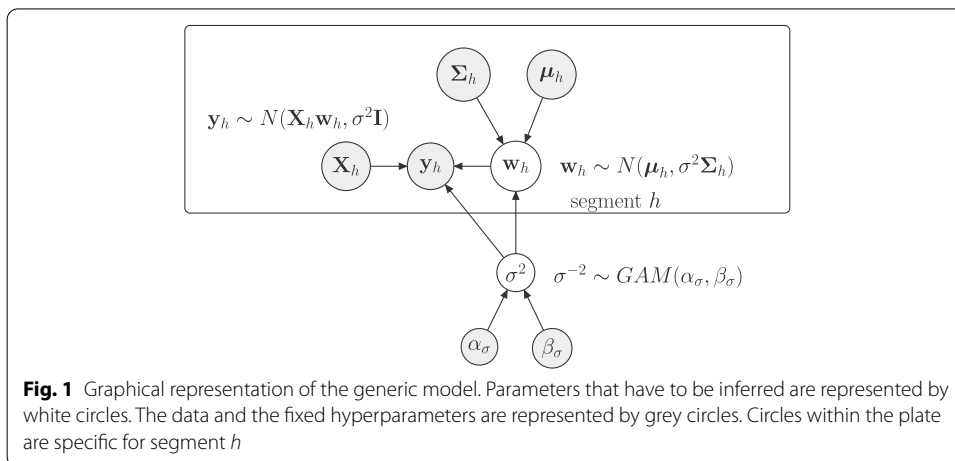


Fig. 1 Graphical representation of the generic model. Parameters that have to be inferred are represented by white circles. The data and the fixed hyperparameters are represented by grey circles. Circles within the plate are specific for segment h

where $\mathbf{y} := \{\mathbf{y}_1, \dots, \mathbf{y}_H\}$ and $\Delta^2(\boldsymbol{\theta}) := \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h)^\top \mathbf{C}_h(\boldsymbol{\theta})^{-1} (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h)$. The shape of $p(\sigma^2 | \mathbf{y}, \boldsymbol{\theta})$ implies:

$$\sigma^{-2} | (\mathbf{y}, \boldsymbol{\theta}) \sim \text{GAM} \left(\alpha_\sigma + \frac{1}{2} \cdot T, \beta_\sigma + \frac{1}{2} \cdot \Delta^2(\boldsymbol{\theta}) \right) \tag{6}$$

For the marginal likelihood, with \mathbf{w}_h ($h = 1, \dots, H$) and σ^2 integrated out, we apply the rule from Section 2.3.7 of Bishop [15]:

$$p(\mathbf{y} | \boldsymbol{\theta}) = \frac{\Gamma(\frac{T}{2} + a_\sigma)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-T/2} \cdot (2b_\sigma)^{a_\sigma}}{\left(\prod_{h=1}^H \det(\mathbf{C}_h(\boldsymbol{\theta})) \right)^{1/2}} \cdot \left(2b_\sigma + \Delta^2(\boldsymbol{\theta}) \right)^{-\left(\frac{T}{2} + a_\sigma\right)} \tag{7}$$

When all parameters in $\boldsymbol{\theta}$ are fixed, the marginal likelihood of the piece-wise linear regression model can be computed in closed form. In typical models the (hyper-)hyperparameters in $\boldsymbol{\theta}$ depend on hyperparameters with their own hyperprior distributions. From now on we will only include the free hyperparameters in $\boldsymbol{\theta}$. In the following subsections we describe four possible model instantiations, namely: the uncoupled model (M1), the coupled model (M2), the newly proposed partially coupled model (M3), and the generalized coupled model (M4). In the forthcoming subsections we will introduce

Table 1 List of mathematical symbols

Symbol	Description	Prior distribution
N	Total number of nodes (genes)	–
n	Number of potential parent nodes, here $n = N - 1$	–
h	Data segment h	–
H	Total number of data segments	–
k	Number of covariates in covariate set	–
t	Data point t	–
σ^2	Noise variance parameter	$\sigma^{-2} \sim \text{GAM}(\alpha_\sigma, \beta_\sigma)$
λ_c	Coupling strength parameter, $h > 1$	$\lambda_c^{-1} \sim \text{GAM}(\alpha_c, \beta_c)$
λ_u	SNR parameter, $h = 1$	$\lambda_u^{-1} \sim \text{GAM}(\alpha_u, \beta_u)$
λ_h	h th coupling strength parameter (M4 model)	$\lambda_h^{-1} \sim \text{GAM}(\alpha_c, \beta_c)$
δ_h	h th coupling indicator variable (M3 model)	$\delta_h \sim \text{BER}(p), p \sim \text{BETA}(a, b)$
T	Total number of data points	–
T_h	Number of data points in segment h	–
D_i	i th data point	–
Z_i	i th network node	–
$\boldsymbol{\pi}_i$	Parent (covariate) set of i th node, Z_i	$p(\boldsymbol{\pi} \leq 3) = c, p(\boldsymbol{\pi} > 3) = 0$
$\boldsymbol{\tau}$	Changepoint set	$p(\boldsymbol{\tau}) = (1 - p)^{(T-1) - (H-1)} \cdot p^{H-1}$
τ_h	Changepoint h	–
X_i	i th covariate	–
\mathbf{X}_h	Design matrix of segment h	–
\mathbf{y}_h	Response vector of segment h	$\mathbf{y}_h (\mathbf{w}_h, \sigma^2) \sim \mathcal{N}(\mathbf{X}_h \mathbf{w}_h, \sigma^2 \mathbf{I})$
\mathbf{w}_h	Regression coefficient vector of segment h	$\mathbf{w}_h (\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \sigma^2) \sim \mathcal{N}(\boldsymbol{\mu}_h, \sigma^2 \boldsymbol{\Sigma}_h)$
$\tilde{\mathbf{w}}_{h-1}$	Posterior expectation of \mathbf{w}_{h-1}	–

further mathematical symbols. For convenience, Table 1 lists the mathematical symbols that we will use in this paper.

Model M1: the uncoupled model

A standard approach, akin to the models of Lèbre et al. [1] and Dondelinger et al. [3], is to set $\mu_h = \mathbf{0}$ and to assume that the matrices Σ_h are diagonal matrices $\Sigma_h = \lambda_u \mathbf{I}$, where the parameter $\lambda_u \in \mathbf{R}^+$ is shared among segments and assumed to be inverse Gamma distributed, $\lambda_u^{-1} \sim GAM(\alpha_u, \beta_u)$. In the supplementary material we provide a graphical model representation of the uncoupled model (M1). Using the notation of the generic model, we have:

$$\theta = \{\lambda_u\}, \quad \mathbf{C}_h(\lambda_u) = \mathbf{I} + \lambda_u \mathbf{X}_h \mathbf{X}_h^T, \quad \Delta^2(\lambda_u) := \sum_{h=1}^H \mathbf{y}_h^T \mathbf{C}_h(\lambda_u)^{-1} \mathbf{y}_h \tag{8}$$

For the posterior distribution of the uncoupled model we have:

$$p(\mathbf{w}, \sigma^2, \lambda_u | \mathbf{y}) \propto p(\sigma^2) \cdot p(\lambda_u) \cdot \prod_{h=1}^H p(\mathbf{w}_h | \sigma^2, \lambda_u) \cdot \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \mathbf{w}_h) \tag{9}$$

where $\mathbf{w} := \{\mathbf{w}_1, \dots, \mathbf{w}_H\}$. From Eq. (9) it follows for the full conditional distribution of λ_u :

$$\begin{aligned} p(\lambda_u | \mathbf{y}, \mathbf{w}, \sigma^2) &\propto p(\lambda_u) \cdot \prod_{h=1}^H p(\mathbf{w}_h | \sigma^2, \lambda_u) \\ &\propto (\lambda_u^{-1})^{a_u + \frac{H \cdot (k+1)}{2}} \cdot \exp \left\{ -\lambda_u^{-1} \left(b_u + \frac{1}{2} \sigma^{-2} \sum_{h=1}^H \mathbf{w}_h^T \mathbf{w}_h \right) \right\} \end{aligned}$$

and the shape of the latter density implies:

$$\lambda_u^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2) \sim GAM \left(\alpha_u + \frac{H \cdot (k+1)}{2}, \beta_u + \frac{1}{2} \sigma^{-2} \sum_{h=1}^H \mathbf{w}_h^T \mathbf{w}_h \right) \tag{10}$$

Since the full conditional distribution of λ_u depends on σ^2 and \mathbf{w} , those parameters have to be sampled first. From Eq. (6) a value of σ^2 can be sampled via a collapsed Gibbs-sampling step, with the \mathbf{w}_h 's being integrated out. Subsequently, given σ^2 , Eq. (4) can be used to sample the vectors \mathbf{w}_h 's. Finally, for each λ_u sampled from Eq. (10) the marginal likelihood, $p(\mathbf{y} | \lambda_u)$, can be computed by plugging in the expressions from Eq. (8) into Eq. (7).

Model M2: the (fully) coupled model

The (fully) coupled model, proposed by Grzegorzcyk and Husmeier [5], uses the posterior expectation of \mathbf{w}_{h-1} as prior expectation for \mathbf{w}_h . Only the first segment $h = 1$ has an uninformative prior:

$$\mathbf{w}_h \sim \begin{cases} \mathcal{N}(\mathbf{0}, \sigma^2 \lambda_u \mathbf{I}) & \text{if } h = 1 \\ \mathcal{N}(\tilde{\mathbf{w}}_{h-1}, \sigma^2 \lambda_c \mathbf{I}) & \text{if } h > 1 \end{cases} \tag{11}$$

where $\tilde{\mathbf{w}}_{h-1}$ is the posterior expectation of \mathbf{w}_{h-1} (cp. Eq. (4)):

$$\tilde{\mathbf{w}}_{h-1} := \begin{cases} [\boldsymbol{\Sigma}_1^{-1} + \mathbf{X}_1^T \mathbf{X}_1]^{-1} (\mathbf{X}_1^T \mathbf{y}_1) & \text{if } h = 2 \\ \left[\boldsymbol{\Sigma}_{h-1}^{-1} + \mathbf{X}_{h-1}^T \mathbf{X}_{h-1} \right]^{-1} (\lambda_c^{-1} \tilde{\mathbf{w}}_{h-2} + \mathbf{X}_{h-1}^T \mathbf{y}_{h-1}) & \text{if } h > 2 \end{cases}$$

The parameter λ_c has been called the ‘coupling parameter’ and it has been assumed that it has an inverse Gamma prior distribution, $\lambda_c^{-1} \sim GAM(\alpha_c, \beta_c)$. Using the notation from the generic model (see Fig. 1), we note that Eq. (11) corresponds to:

$$\boldsymbol{\mu}_h = \begin{cases} \mathbf{0} & \text{if } h = 1 \\ \tilde{\mathbf{w}}_{h-1} & \text{if } h > 1 \end{cases}, \quad \boldsymbol{\Sigma}_h = \begin{cases} \lambda_u \mathbf{I} & \text{if } h = 1 \\ \lambda_c \mathbf{I} & \text{if } h > 1 \end{cases},$$

$$\mathbf{C}_h(\boldsymbol{\theta}) = \begin{cases} \mathbf{I} + \lambda_u \mathbf{X}_h \mathbf{X}_h^T & \text{if } h = 1 \\ \mathbf{I} + \lambda_c \mathbf{X}_h \mathbf{X}_h^T & \text{if } h > 1 \end{cases}, \quad \boldsymbol{\theta} = \{\lambda_u, \lambda_c\}, \quad \Delta^2(\boldsymbol{\theta}) = \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \tilde{\mathbf{w}}_{h-1})^T \mathbf{C}_h(\boldsymbol{\theta})^{-1} (\mathbf{y}_h - \mathbf{X}_h \tilde{\mathbf{w}}_{h-1})$$

with $\tilde{\mathbf{w}}_0 := \mathbf{0}$, $\lambda_u^{-1} \sim GAM(\alpha_u, \beta_u)$ and $\lambda_c^{-1} \sim GAM(\alpha_c, \beta_c)$. As $\tilde{\mathbf{w}}_{h-1}$ is treated like a fixed hyperparameter when used as input for segment h , we exclude the parameters $\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_{H-1}$ from $\boldsymbol{\theta}$.

In the supplementary material we provide a graphical model representation of the coupled M2 model. For the posterior we have:

$$p(\mathbf{w}, \sigma^2, \lambda_u, \lambda_c | \mathbf{y}) \propto p(\sigma^2) \cdot p(\lambda_u) \cdot p(\lambda_c) \cdot p(\mathbf{w}_1 | \sigma^2, \lambda_u) \cdot \prod_{h=2}^H p(\mathbf{w}_h | \sigma^2, \lambda_c) \cdot \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \mathbf{w}_h) \tag{12}$$

In analogy to the derivations in the previous subsection one can derive (cp. [5]):

$$\lambda_u^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2, \lambda_c) \sim GAM\left(\alpha_u + \frac{1 \cdot (k + 1)}{2}, \beta_u + \frac{1}{2} \sigma^{-2} D_u^2\right) \tag{13}$$

$$\lambda_c^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2, \lambda_u) \sim GAM\left(\alpha_c + \frac{(H - 1) \cdot (k + 1)}{2}, \beta_c + \frac{1}{2} \sigma^{-2} D_c^2\right) \tag{14}$$

where $D_u^2 := \mathbf{w}_1^T \mathbf{w}_1$ and $D_c^2 := \sum_{h=2}^H (\mathbf{w}_h - \tilde{\mathbf{w}}_{h-1})^T (\mathbf{w}_h - \tilde{\mathbf{w}}_{h-1})$.

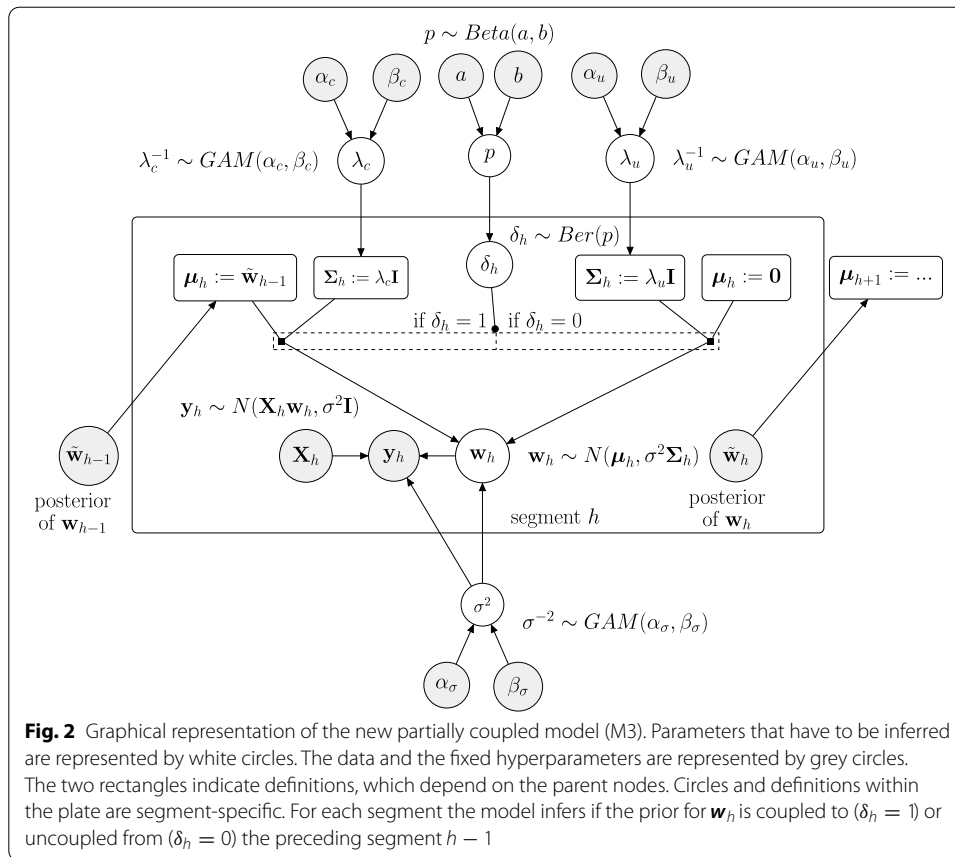
For each $\boldsymbol{\theta} = \{\lambda_u, \lambda_c\}$ the marginal likelihood, $p(\mathbf{y} | \lambda_u, \lambda_c)$, can be computed by plugging the expressions $\mathbf{C}_h(\boldsymbol{\theta})$ and $\Delta^2(\boldsymbol{\theta})$ into Eq. (7).

Model M3: the new partially segment-wise coupled model

We propose a new ‘consensus’ model between the M1 and the M2 model. The new model (M3) allows each segment $h > 1$ either to coupled top or to uncouple from the preceding segment $h - 1$. We use an uninformative prior for the first segment $h = 1$, and for all segments $h > 1$ we introduce a binary variable δ_h which indicates whether segment h is coupled to ($\delta_h = 1$) or uncoupled from ($\delta_h = 0$) the preceding segment $h - 1$:

$$\mathbf{w}_h \sim \begin{cases} \mathcal{N}(\mathbf{0}, \sigma^2 \lambda_u \mathbf{I}) & \text{if } h = 1 \\ \mathcal{N}(\delta_h \cdot \tilde{\mathbf{w}}_{h-1}, \sigma^2 \lambda_c^{\delta_h} \lambda_u^{1-\delta_h} \mathbf{I}) & \text{if } h > 1 \end{cases} \tag{15}$$

where $\tilde{\mathbf{w}}_{h-1}$ is the posterior expectation of \mathbf{w}_{h-1} . The new priors from Eq. (15) yield for $h \geq 2$ the following posterior expectations (cp. Eq. (4)):



$$\tilde{\mathbf{w}}_{h-1} = \left(\lambda_c^{-\delta_{h-1}} \lambda_u^{-(1-\delta_{h-1})} \mathbf{I} + \mathbf{X}_{h-1}^T \mathbf{X}_{h-1} \right)^{-1} \left(\delta_{h-1} \lambda_c^{-1} \tilde{\mathbf{w}}_{h-2} + \mathbf{X}_{h-1}^T \mathbf{y}_{h-1} \right)$$

with $\tilde{\mathbf{w}}_0 := \mathbf{0}$, $\delta_1 := 0$, we have in the generic model notation:

$$\boldsymbol{\mu}_h = \delta_h \tilde{\mathbf{w}}_{h-1}, \quad \boldsymbol{\Sigma}_h = \lambda_c^{\delta_h} \lambda_u^{1-\delta_h} \mathbf{I}, \quad \boldsymbol{\theta} = \{ \lambda_u, \lambda_c, \{ \delta_h \}_{h \geq 2} \}, \quad \mathbf{C}_h(\boldsymbol{\theta}) = \mathbf{I} + \lambda_c^{\delta_h} \lambda_u^{1-\delta_h} \mathbf{X}_h \mathbf{X}_h^T$$

We assume the binary variables $\delta_2, \dots, \delta_H$ to have a Bernoulli prior distributions, $\delta_h \sim \text{BER}(p)$, with a joint hyperparameter $p \in [0, 1]$ having a Beta hyperprior distribution, $p \sim \text{BETA}(a, b)$. We note that

- $\delta_h = 0$ ($h \geq 2$) gives model M1 with $P(\mathbf{w}_h) = \mathcal{N}(\mathbf{0}, \lambda_u \sigma^2 \mathbf{I})$ for all h
- $\delta_h = 1$ ($h \geq 2$) gives model M2 with $P(\mathbf{w}_h) = \mathcal{N}(\tilde{\mathbf{w}}_{h-1}, \lambda_c \sigma^2 \mathbf{I})$ for $h \geq 2$.
- The new partially segment-wise coupled model infers the variables δ_h ($h \geq 2$) from the data. It searches for the best trade-off between the models M1 and M2.

A graphical model presentation of the partially coupled model is shown in Fig. 2. For $\delta_h \sim \text{BER}(p)$ with $p \sim \text{BETA}(a, b)$ the joint marginal density of $\{ \delta_h \}_{h \geq 2}$ is:

$$p(\{\delta_h\}_{h \geq 2}) = \int p(p) \prod_{h=2}^H p(\delta_h|p) dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma\left(a + \sum_{h=2}^H \delta_h\right) \Gamma\left(b + \sum_{h=2}^H (1 - \delta_h)\right)}{\Gamma(a+b+(H-1))} \tag{16}$$

For the posterior distribution of the partially segment-wise coupled model we get:

$$p(\mathbf{w}, \sigma^2, \lambda_u, \lambda_c, \{\delta_h\}_{h \geq 2} | \mathbf{Y}) \propto p(\sigma^2) \cdot p(\lambda_u) \cdot p(\lambda_c) \cdot p(\{\delta_h\}_{h \geq 2}) \cdot p(\mathbf{w}_1 | \sigma^2, \lambda_u) \cdot \prod_{h=2}^H p(\mathbf{w}_h | \sigma^2, \lambda_u, \lambda_c, \delta_h) \cdot \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \mathbf{w}_h)$$

For the full conditional distributions of λ_u and λ_c we have:

$$p(\lambda_u | \mathbf{y}, \mathbf{w}, \sigma^2, \lambda_c, \{\delta_h\}_{h \geq 2}) \propto p(\lambda_u) \cdot \prod_{h: \delta_h=0} p(\mathbf{w}_h | \sigma^2, \lambda_u)$$

$$p(\lambda_c | \mathbf{y}, \mathbf{w}, \sigma^2, \lambda_u, \{\delta_h\}_{h \geq 2}) \propto p(\lambda_c) \cdot \prod_{h: \delta_h=1} p(\mathbf{w}_h | \sigma^2, \lambda_c)$$

where $\delta_1 := 0$ fixed. And it follows from the shapes of the densities:

$$\lambda_u^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2, \lambda_c, \{\delta_h\}_{h \geq 2}) \sim \text{GAM}\left(\alpha_u + \frac{H_u \cdot (k+1)}{2}, \beta_u + \frac{1}{2} \sigma^{-2} D_u^2\right)$$

$$\lambda_c^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2, \lambda_u, \{\delta_h\}_{h \geq 2}) \sim \text{GAM}\left(\alpha_c + \frac{H_c \cdot (k+1)}{2}, \beta_c + \frac{1}{2} \sigma^{-2} D_c^2\right)$$

where $H_c = \sum_h \delta_h$ is the number of coupled segments, $H_u = \sum_h (1 - \delta_h)$ is the number of uncoupled segments, so that $H_c + H_u = H$, and

$$D_u^2 := \sum_{h: \delta_h=0} \mathbf{w}_h^T \mathbf{w}_h, \quad D_c^2 := \sum_{h: \delta_h=1} (\mathbf{w}_h - \tilde{\mathbf{w}}_{h-1})^T (\mathbf{w}_h - \tilde{\mathbf{w}}_{h-1})$$

For each parameter instantiation $\boldsymbol{\theta} = \{\lambda_u, \lambda_c, \{\delta_h\}_{h \geq 2}\}$ the marginal likelihood, $p(\mathbf{y} | \boldsymbol{\theta})$, can be computed with Eq. (7), where $\mathbf{C}_h(\boldsymbol{\theta})$ was defined above, and

$$\Delta^2(\boldsymbol{\theta}) = \sum_{h=1}^H (\mathbf{y}_h - \delta_h \mathbf{X}_h \tilde{\mathbf{w}}_{h-1})^T \left[\mathbf{I} + \lambda_c^{\delta_h} \lambda_u^{1-\delta_h} \mathbf{X}_h \mathbf{X}_h^T \right]^{-1} (\mathbf{y}_h - \delta_h \mathbf{X}_h \tilde{\mathbf{w}}_{h-1})$$

We have for each binary variable δ_k ($k = 2, \dots, H$):

$$p(\delta_k = 1 | \lambda_u, \lambda_c, \{\delta_h\}_{h \neq k}, \mathbf{y}) \propto p(\mathbf{y} | \lambda_u, \lambda_c, \{\delta_h\}_{h \neq k}, \delta_k = 1) \cdot p(\{\delta_h\}_{h \neq k}, \delta_k = 1)$$

so that its full conditional distribution is:

$$\delta_k | (\lambda_u, \lambda_c, \{\delta_h\}_{h \neq k}, \mathbf{y}) \sim \text{BER} \left(\frac{p(\mathbf{y} | \lambda_u, \lambda_c, \{\delta_h\}_{h \neq k}, \delta_k = 1) \cdot p(\{\delta_h\}_{h \neq k}, \delta_k = 1)}{\sum_{j=0}^1 p(\mathbf{y} | \lambda_u, \lambda_c, \{\delta_h\}_{h \neq k}, \delta_k = j) \cdot p(\{\delta_h\}_{h \neq k}, \delta_k = j)} \right)$$

Each δ_k ($k > 1$) can therefore be sampled with a collapsed Gibbs sampling step, where $\{\mathbf{w}_h\}$, σ^2 and \mathbf{p} have been integrated out.

Model M4: the generalised (fully) coupled model

In [6] we proposed to generalise the (fully) coupled model (i.e. the M2 model) by introducing a segment-specific coupling parameter λ_h for each segment $h > 2$. This yields:

$$\mathbf{w}_h \sim \begin{cases} \mathcal{N}(\mathbf{0}, \sigma^2 \lambda_u \mathbf{I}) & \text{if } h = 1 \\ \mathcal{N}(\tilde{\mathbf{w}}_{h-1}, \sigma^2 \lambda_h \mathbf{I}) & \text{if } h > 1 \end{cases} \tag{17}$$

where $\tilde{\mathbf{w}}_{h-1}$ is the posterior expectation of \mathbf{w}_{h-1} . For the parameters λ_h we have assumed that they are inverse Gamma distributed, $\lambda_h^{-1} \sim GAM(\alpha_c, \beta_c)$, with hyperparameters α_c and β_c . In the supplementary material we provide a graphical model representation of the M4 model. Recalling the generic notation and setting $\tilde{\mathbf{w}}_0 := \mathbf{0}$ and $\lambda_1 := \lambda_u$, Eq. (17) gives:

$$\begin{aligned} \boldsymbol{\mu}_h &= \tilde{\mathbf{w}}_{h-1}, \quad \boldsymbol{\Sigma}_h = \lambda_h \mathbf{I}, \quad \mathbf{C}_h(\boldsymbol{\theta}) = \mathbf{I} + \lambda_h \mathbf{X}_h \mathbf{X}_h^\top, \quad \boldsymbol{\theta} = \{\lambda_u, \{\lambda_h\}_{h \geq 2}\}, \\ \text{and } \Delta^2(\boldsymbol{\theta}) &= \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \tilde{\mathbf{w}}_{h-1})^\top \mathbf{C}_h(\boldsymbol{\theta})^{-1} (\mathbf{y}_h - \mathbf{X}_h \tilde{\mathbf{w}}_{h-1}) \end{aligned}$$

For the posterior we have:

$$\begin{aligned} p(\mathbf{w}, \sigma^2, \lambda_u, \{\lambda_h\}_{h \geq 2} | \mathbf{y}) &\propto p(\sigma^2) \cdot p(\lambda_u) \cdot \left(\prod_{h=2}^H p(\lambda_h) \right) \\ &\cdot p(\mathbf{w}_1 | \sigma^2, \lambda_u) \cdot \prod_{h=2}^H p(\mathbf{w}_h | \sigma^2, \lambda_h) \cdot \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \mathbf{w}_h) \end{aligned} \tag{18}$$

For $k = 2, \dots, H$ it follows:

$$\begin{aligned} \lambda_k^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2, \lambda_u, \{\lambda_h\}_{h \neq k}) &\sim GAM\left(\alpha_c + \frac{(k+1)}{2}, \beta_c + \frac{1}{2} \sigma^{-2} D_k^2\right) \\ \text{and } \lambda_u^{-1} | (\mathbf{y}, \mathbf{w}, \sigma^2, \{\lambda_h\}_{h \geq 2}) &\sim GAM\left(\alpha_u + \frac{(k+1)}{2}, \beta_u + \frac{1}{2} \sigma^{-2} D_u^2\right) \end{aligned}$$

where $D_u^2 := \mathbf{w}_1^\top \mathbf{w}_1$ and $D_k^2 := (\mathbf{w}_k - \tilde{\mathbf{w}}_{k-1})^\top (\mathbf{w}_k - \tilde{\mathbf{w}}_{k-1})$.

For each $\boldsymbol{\theta} = \{\lambda_u, \{\lambda_h\}_{h \geq 2}\}$ the marginal likelihood, $p(\mathbf{y} | \{\lambda_u, \{\lambda_h\}_{h \geq 2}\})$, can be computed with Eq. (7); using the expressions $\mathbf{C}_h(\boldsymbol{\theta})$ and $\Delta^2(\boldsymbol{\theta})$ defined above.

Unlike the proposed partially coupled M3 model, the generalized coupled M4 model does not feature any mechanism to uncouple neighbouring segments. Like the fully coupled M2 model, the M4 model has been designed such that it has to couple all neighbouring segments. The only advantage over the M2 model is that the the M4 model introduces segment-specific coupling parameters, so that the coupling strength(s) can vary over time.

Reversible jump Markov chain Monte Carlo inference

We use Reversible Jump Markov Chain Monte Carlo simulations to generate posterior samples $\{\boldsymbol{\pi}^{(w)}, \boldsymbol{\tau}^{(w)}, \boldsymbol{\theta}^{(w)}\}_{w=1, \dots, W}$. In each iteration we re-sample the parameters in $\boldsymbol{\theta}$ from their full conditional distributions (Gibbs sampling), and we perform two Metropolis-Hastings moves; one on the covariate set $\boldsymbol{\pi}$ and one on the changepoint set $\boldsymbol{\tau}$. For the four models (M1–M4) Eq. (1) takes the form:

$$p(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\theta} | \mathbf{y}) \propto \begin{cases} p(\boldsymbol{\pi})p(\boldsymbol{\tau})p(\lambda_u) \cdot p(\mathbf{y} | \boldsymbol{\pi}, \boldsymbol{\tau}, \lambda_u) & \text{M1} \\ p(\boldsymbol{\pi})p(\boldsymbol{\tau})p(\lambda_u) \cdot p(\lambda_c) \cdot p(\mathbf{y} | \boldsymbol{\pi}, \boldsymbol{\tau}, \lambda_u, \lambda_c) & \text{M2} \\ p(\boldsymbol{\pi})p(\boldsymbol{\tau})p(\lambda_u) \cdot p(\lambda_c) \cdot p(\{\delta_h\}_{h \geq 2}) \cdot p(\mathbf{y} | \boldsymbol{\pi}, \boldsymbol{\tau}, \lambda_u, \lambda_c, \{\delta_h\}_{h \geq 2}) & \text{M3} \\ p(\boldsymbol{\pi})p(\boldsymbol{\tau})p(\lambda_u) \cdot \left(\prod_{h=2}^H p(\lambda_h)\right) \cdot p(\mathbf{y} | \boldsymbol{\pi}, \boldsymbol{\tau}, \lambda_u, \{\lambda_h\}_{h \geq 2}) & \text{M4} \end{cases}$$

All likelihood terms, $p(\mathbf{y} | \dots)$, are marginalized over σ^2 and $\{\mathbf{w}_h\}$ and for the new M3 model also the Bernoulli parameter p has been integrated out.

For the models M1–M2 the dimension of $\boldsymbol{\theta}$ does not depend on $\boldsymbol{\tau}$, while for the models M3–M4 the dimension of $\boldsymbol{\theta}$ *does* depend on $\boldsymbol{\tau}$. The M3 model has a discrete parameter $\delta_h \in \{0, 1\}$ and the M4 model has a continuous parameter $\lambda_h \in \mathbb{R}^+$ for each $h > 1$.

The model-specific full conditional distributions for the Gibbs sampling steps have been provided above. For sampling $\boldsymbol{\pi}$ we implement 3 moves: covariate ‘removal (R)’, ‘addition (A)’, and ‘exchange (E)’. Each move proposes to replace $\boldsymbol{\pi}$ by a new covariate set $\boldsymbol{\pi}^*$ having one covariate more (A) or less (R) or exchanged (E). When randomly selecting the move type and the involved covariate(s), we get for all models the acceptance probability:

$$A(\boldsymbol{\pi} \rightarrow \boldsymbol{\pi}^*) = \min \left\{ 1, \frac{p(\mathbf{y} | \boldsymbol{\pi}^*, \dots)}{p(\mathbf{y} | \boldsymbol{\pi}, \dots)} \cdot \frac{p(\boldsymbol{\pi}^*)}{p(\boldsymbol{\pi})} \cdot HR_{\boldsymbol{\pi}} \right\}$$

with the Hastings Ratios: $HR_{\boldsymbol{\pi},R} = \frac{|\boldsymbol{\pi}|}{n - |\boldsymbol{\pi}^*|}$, $HR_{\boldsymbol{\pi},A} = \frac{n - |\boldsymbol{\pi}|}{|\boldsymbol{\pi}^*|}$, $HR_{\boldsymbol{\pi},E} = 1$

For sampling $\boldsymbol{\tau}$ we also implement 3 move types: changepoint ‘birth (R)’, ‘death (D)’, and ‘re-allocation (R)’ moves. Each move proposes to replace $\boldsymbol{\tau}$ by a new changepoint set $\boldsymbol{\tau}^*$ having one changepoint added (B) or deleted (D) or re-allocated (R). When randomly selecting the move type, the involved changepoint and the new changepoint location, we get for M1 and M2:

$$A(\boldsymbol{\tau} \rightarrow \boldsymbol{\tau}^*) = \min \left\{ 1, \frac{p(\mathbf{y} | \boldsymbol{\tau}^*, \dots)}{p(\mathbf{y} | \boldsymbol{\tau}, \dots)} \cdot \frac{p(\boldsymbol{\tau}^*)}{p(\boldsymbol{\tau})} \cdot HR_{\boldsymbol{\tau}} \right\}$$

where $HR_{\boldsymbol{\tau},B} = \frac{T - 1 - |\boldsymbol{\tau}|}{|\boldsymbol{\tau}^*|}$, $HR_{\boldsymbol{\tau},D} = \frac{|\boldsymbol{\tau}|}{T - 1 - |\boldsymbol{\tau}^*|}$, $HR_{\boldsymbol{\tau},R} = 1$

For the models M3 (proposed here) and the model M4 from [6] the changepoint moves also affect the numbers of parameters in $\{\delta_h\}_{h \geq 2}$ and $\{\lambda_h\}_{h \geq 2}$, respectively. For all segments that stay identical we keep the parameters unchanged. For all new segments we re-sample the corresponding parameters. For the new model M3 we flip coins to get candidates for the involved δ_h ’s. This yields:

$$A([\boldsymbol{\tau}, \{\delta_h\}] \rightarrow [\boldsymbol{\tau}^*, \{\delta_h\}^*]) = \min \left\{ 1, \frac{p(\mathbf{y}|\boldsymbol{\tau}^*, \{\delta_h\}^*, \dots) p(\boldsymbol{\tau}^*) p(\{\delta_h\}^*)}{p(\mathbf{y}|\boldsymbol{\tau}, \{\delta_h\}, \dots) p(\boldsymbol{\tau}) p(\{\delta_h\})} \cdot HR_{\boldsymbol{\tau}} \cdot c_{\boldsymbol{\tau}} \right\}$$

where $c_{\boldsymbol{\tau},B} = 2$ for birth, $c_{\boldsymbol{\tau},D} = 1/2$ for death, and $c_{\boldsymbol{\tau},R} = 1$ for re-allocation moves. For the model M4 we follow [6] and re-sample the involved λ_h 's from their priors $p(\lambda_h)$. We obtain:

$$A([\boldsymbol{\tau}, \{\lambda_h\}] \rightarrow [\boldsymbol{\tau}^*, \{\lambda_h\}^*]) = \min \left\{ 1, \frac{p(\mathbf{y}|\boldsymbol{\tau}^*, \{\lambda_h\}^*, \dots) p(\boldsymbol{\tau}^*)}{p(\mathbf{y}|\boldsymbol{\tau}, \{\lambda_h\}, \dots) p(\boldsymbol{\tau})} \cdot HR_{\boldsymbol{\tau}} \right\}$$

Note that the additional factor $c_{\boldsymbol{\tau}} := \frac{p(\{\lambda_h\})}{p(\{\lambda_h\}^*)}$ of the Hastings ratio has been canceled with the prior ratio $\frac{p(\{\lambda_h\}^*)}{p(\{\lambda_h\})}$.

Edge scores and areas under precision recall curves (AUC)

For a network with N variables Z_1, \dots, Z_N we infer N separate regression models. For each Z_i we get a sample $\{\boldsymbol{\pi}_i^{(w)}, \boldsymbol{\tau}_i^{(w)}, \boldsymbol{\theta}_i^{(w)}\}_{w=1, \dots, W}$ from the i th posterior. From the covariate sets we form a sample of graphs $\mathcal{G}^{(w)} = \{\boldsymbol{\pi}_1^{(w)}, \dots, \boldsymbol{\pi}_N^{(w)}\}_{w=1, \dots, W}$. For each edge $Z_i \rightarrow Z_j$ the edge posterior probability (edge score) is:

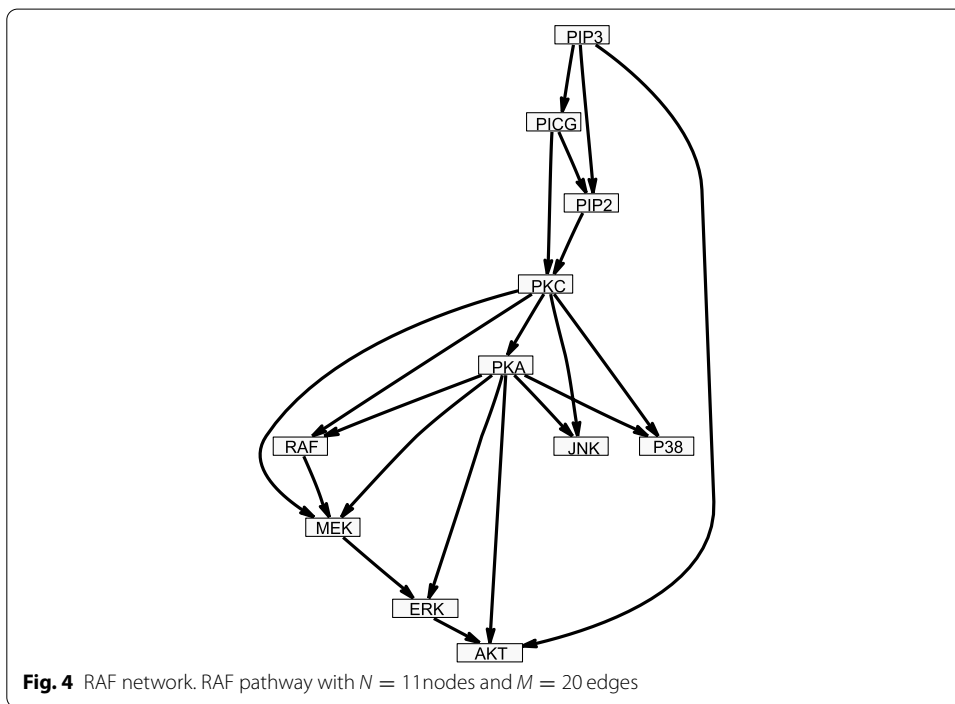
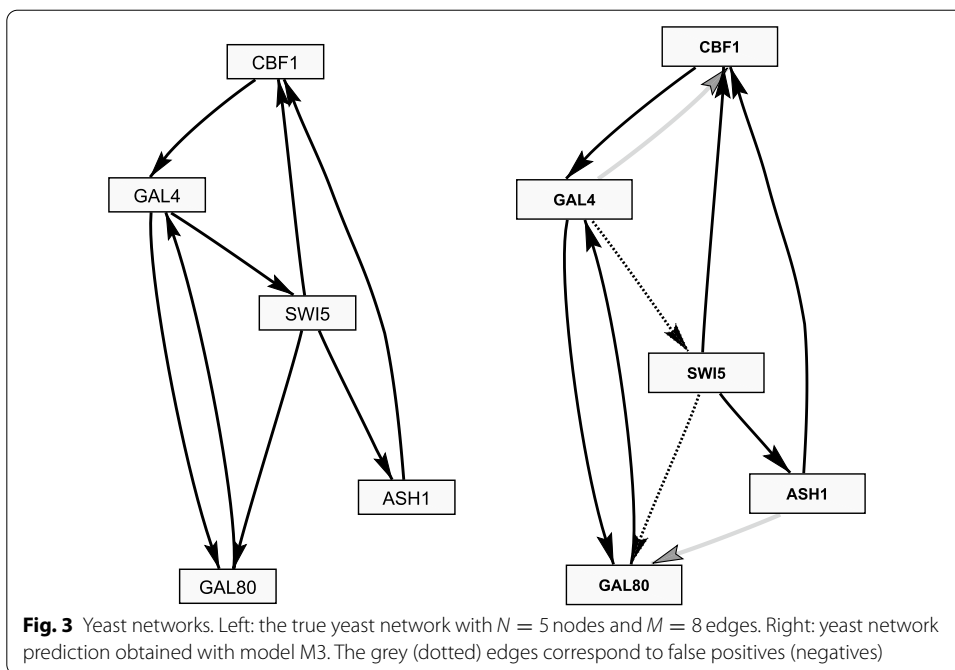
$$\hat{e}_{i,j} = \frac{1}{W} \sum_{w=1}^W I_{i \rightarrow j}(\mathcal{G}^{(w)}) \quad \text{where} \quad I_{i \rightarrow j}(\mathcal{G}^{(w)}) = \begin{cases} 1 & \text{if } X_i \in \boldsymbol{\pi}_j^{(w)} \\ 0 & \text{if } X_i \notin \boldsymbol{\pi}_j^{(w)} \end{cases}$$

If the true network is known and has M edges, we can quantify the network reconstruction accuracy. For each threshold $\xi \in [0, 1]$ we extract the n_{ξ} edges whose scores $\hat{e}_{i,j}$ exceed ξ , and we count the number of true positives T_{ξ} among them. Plotting the precisions $P_{\xi} := T_{\xi}/n_{\xi}$ against the recalls $R_{\xi} := T_{\xi}/M$, gives the precision-recall curve. We refer to the area under the curve as AUC value.

Hyperparameter settings and simulation details

The hyperparameters of the priors and hyperpriors of the four NH-DBN models (M1–M4) have to be specified in advance, and we note that the hyperparameter setting can have an effect on the resulting posterior distributions and so on the network reconstruction results. Selecting appropriate hyperparameters is therefore a crucial task. In the absence of genuine prior knowledge (e.g. from experts or from the literature), we re-use the rather uninformative (and thus generic) parameter settings from earlier publications. Re-using those hyperparameters also has the advantage that our empirical results can be compared with earlier reported results. More specifically, we proceed as follows:

For the models M1, M2 and M4 we re-use the hyperparameters from the earlier works by Lèbre et al. [1], Grzegorzcyk and Husmeier [5], and Shafiee Kamalabad and Grzegorzcyk [6]: $\sigma^{-2} \sim GAM(\alpha_{\sigma} = \nu, \beta_{\sigma} = \nu)$ with $\nu = 0.005$, $\lambda_u^{-1} \sim GAM(\alpha_u = 2, \beta_u = 0.2)$, and $\lambda_c^{-1} \sim GAM(\alpha_c = 3, \beta_c = 3)$. For the new partially coupled model M3 we use the same setting with the extension: $\delta_h \sim BER(p)$ with $p \sim BETA(a = 1, b = 1)$, which seems to be a very natural choice. For the M3 model we also tested several alternative hyperparameter settings, but we did not observe significantly deviating results, indicating that the M3 model is rather robust with respect to the hyperparameter settings. For



more thorough studies on how the hyperparameter setting affects the network reconstruction results, we refer to the work by Grzegorzyc and Husmeier [5].

For all models M1–M4 we run each reversible jump Markov chain Monte Carlo simulation for $V = 100,000$ iterations. Setting the burn-in phase to $0.5V$ (50%) and thinning out by the factor 10 during the sampling phase, yields $W = 0.5V/10 = 5000$ samples

from each posterior. To check for convergence, we compared the samples of independent simulations, using standard trace plot diagnostics as well as scatter plots of the estimated edge scores. For most of the data sets, analysed here, the diagnostics indicated almost perfect convergence already after $V = 10,000$ iterations; see Fig. 7a for an example.

Data

Synthetic network data

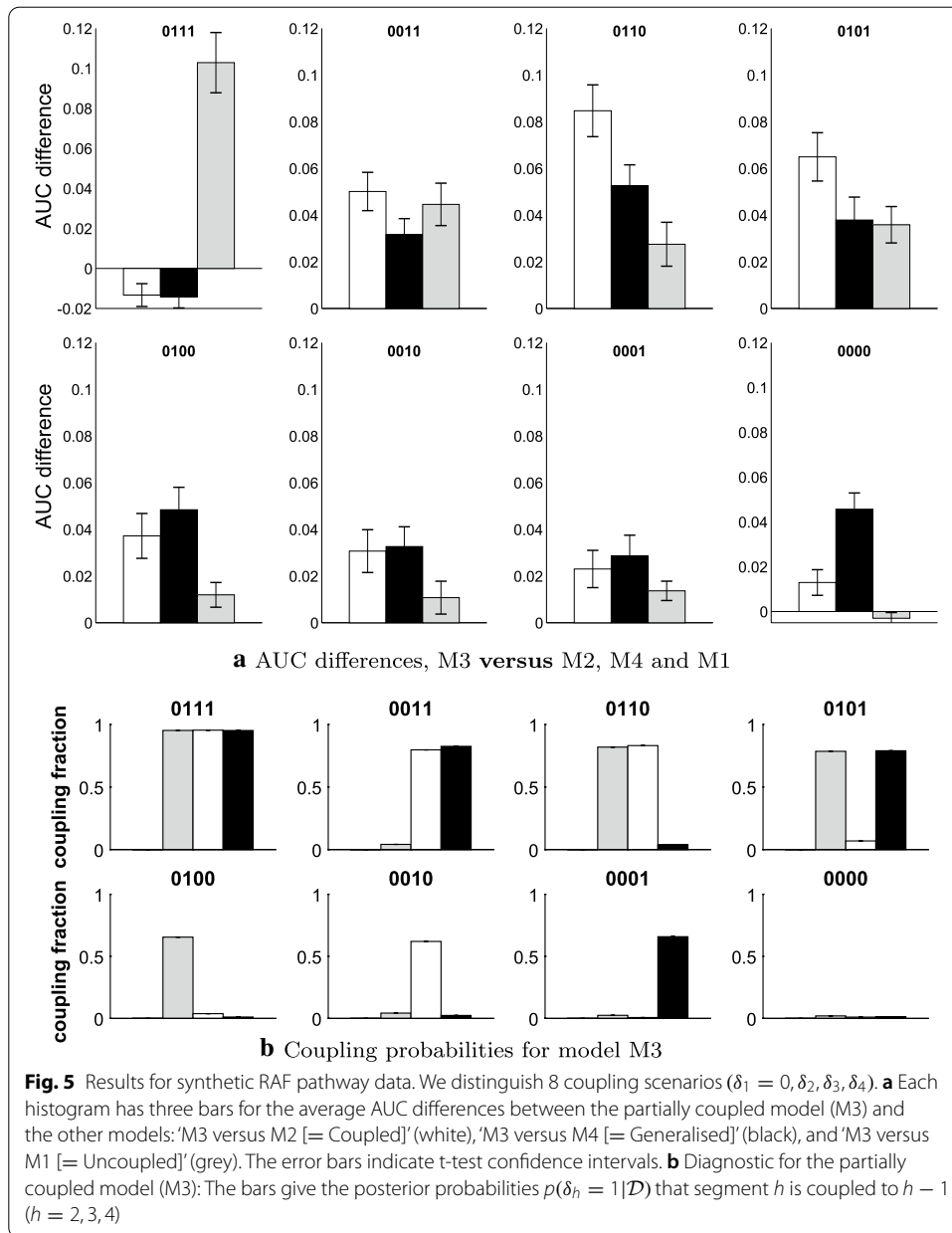
For model comparisons we generated various synthetic network data sets. We report here on two studies with realistic network topologies, shown in Figs. 3 and 4. In both studies we assumed the data segmentation to be known. Hence, we kept the change-points in τ fixed at their right locations and did not perform reversible jump Markov chain Monte Carlo moves on τ .

Study 1 For the RAF pathway with $N = 11$ nodes and $M = 20$ edges, shown in Fig. 4 and taken from Sachs et al. [16], we generated data with $H = 4$ segments having $m = 10$ data points each. For each node Z_i and its parent nodes in π_i we sampled the regression coefficients for $h = 1$ from standard Gaussian distributions and collected them in a vector \mathbf{w}_1^i which we normalised to Euclidean norm 1, $\mathbf{w}_1^i \leftarrow \mathbf{w}_1^i / |\mathbf{w}_1^i|$. For the segments $h = 2, 3, 4$ we use: $\mathbf{w}_h^i = \mathbf{w}_{h-1}^i$ ($\delta_h = 1$, coupled) or $\mathbf{w}_h^i = -\mathbf{w}_{h-1}^i$ ($\delta_h = 0$, uncoupled). The design matrices \mathbf{X}_h^i contain a first column of 1's for the intercept and the segment-specific values of the parent nodes, shifted by one time point. To the segment-specific values of Z_i : $\mathbf{z}_h^i = \mathbf{X}_h^i \mathbf{w}_h^i$ we element-wise added Gaussian noise with standard deviation $\sigma = 0.05$. For all coupling scenarios $(\delta_2, \delta_3, \delta_4) \in \{0, 1\}^3$, we generated 25 data sets having different regression coefficients.

Study 2 This study is similar to the first one with three changes: (i) We used the yeast network with $N = 5$ nodes and $M = 8$ edges, shown in the left panel of Fig. 3 and taken from Cantone et al. [17]. (ii) Again we generated data with $H = 4$ segments, but we varied the number of time points per segment $m \in \{2, 3, \dots, 12\}$. (iii) We focused on one scenario: For each node Z_i and its parent nodes in π_i we generated two vectors \mathbf{w}_\diamond^i and \mathbf{w}_\star^i with standard Gaussian distributed entries. We re-normalised the first vector to Euclidean norm 1, $\mathbf{w}_\diamond^i \leftarrow \mathbf{w}_\diamond^i / |\mathbf{w}_\diamond^i|$, and the 2nd vector to norm 0.5, $\mathbf{w}_\star^i \leftarrow 0.5 \cdot \mathbf{w}_\star^i / |\mathbf{w}_\star^i|$. We set $\mathbf{w}_1^i = \mathbf{w}_2^i = \mathbf{w}_\diamond^i$ so that the segments $h = 2$ and $h = 3$ are coupled, and $\mathbf{w}_3^i = \mathbf{w}_4^i = (\mathbf{w}_\diamond^i + \mathbf{w}_\star^i) / (|\mathbf{w}_\diamond^i + \mathbf{w}_\star^i|)$, so that the segments $h = 3$ and $h = 4$ are coupled, while the coupling between $h = 3$ and $h = 2$ is 'moderate'. For each m we generated 25 data matrices with different regression coefficients.

Yeast gene expression data

Cantone et al. [17] synthetically designed a network in *S. cerevisiae* (yeast) with $N = 5$ genes, and measured gene expression data under galactose- and glucose-metabolism: 16 measurements were taken in galactose and 21 measurements were taken in glucose, with 20 minutes intervals in between measurements. Although the network is small, it is an ideal benchmark data set: The network structure is known, so that network reconstruction methods can be cross-compared on real wet-lab data. We follow Grzegorzczuk and Husmeier and pre-process the data as described in [5]. The true network structure is shown in the left panel of Fig. 3. As an example, a network prediction obtained with the



partially coupled model (M3) is shown in the right panel. For the prediction we extracted the 8 edges with the highest scores.

Arabidopsis gene expression data

The circadian clock in *Arabidopsis thaliana* optimizes the gene regulatory processes with respect to the daily dark:light cycles (photo periods). In four experiments *Arabidopsis* plants were entrained in different dark:light cycles, before gene expression data were measured under constant light condition over 24- and 48-h time intervals. We follow Grzegorzcyk and Husmeier [5] and merge the four time series to one single data set

with $T = 47$ data points and focus our attention on the $N = 9$ core genes: LHY, TOC1, CCA1, ELF4, ELF3, GI, PRR9, PRR5, and PRR3.

Results

In this section we present the results of a comparative evaluation study, in which we compare the performance of the new partially coupled model (M3) with the competing models M1, M2 and M4. Throughout this section we use the new M3 model as reference model.

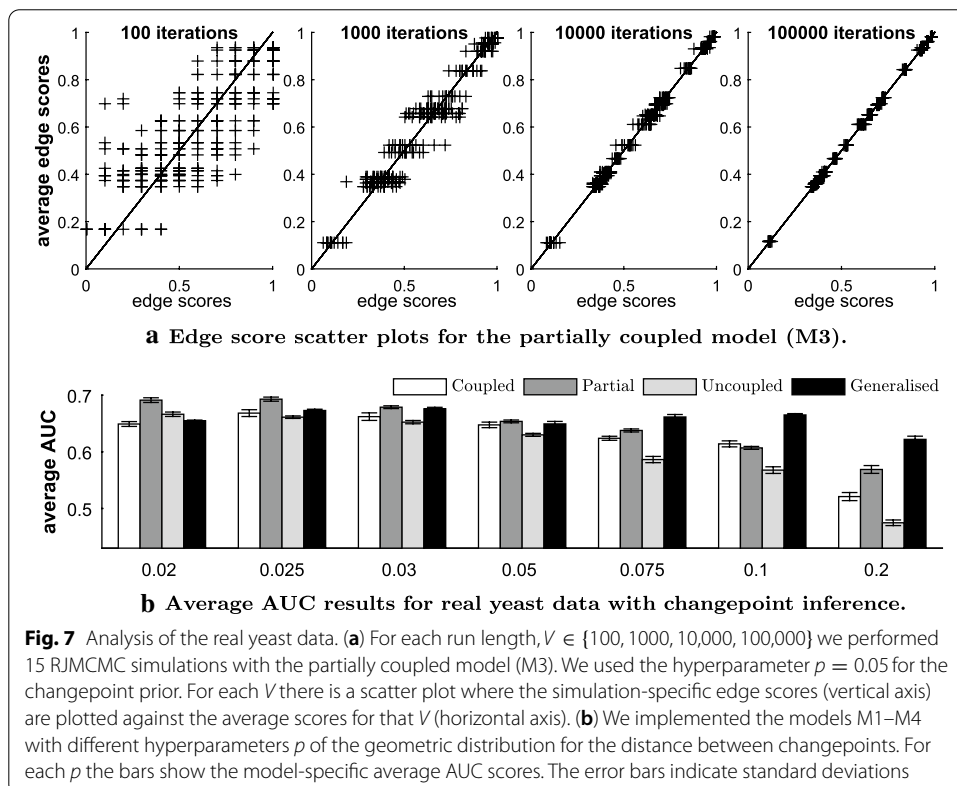
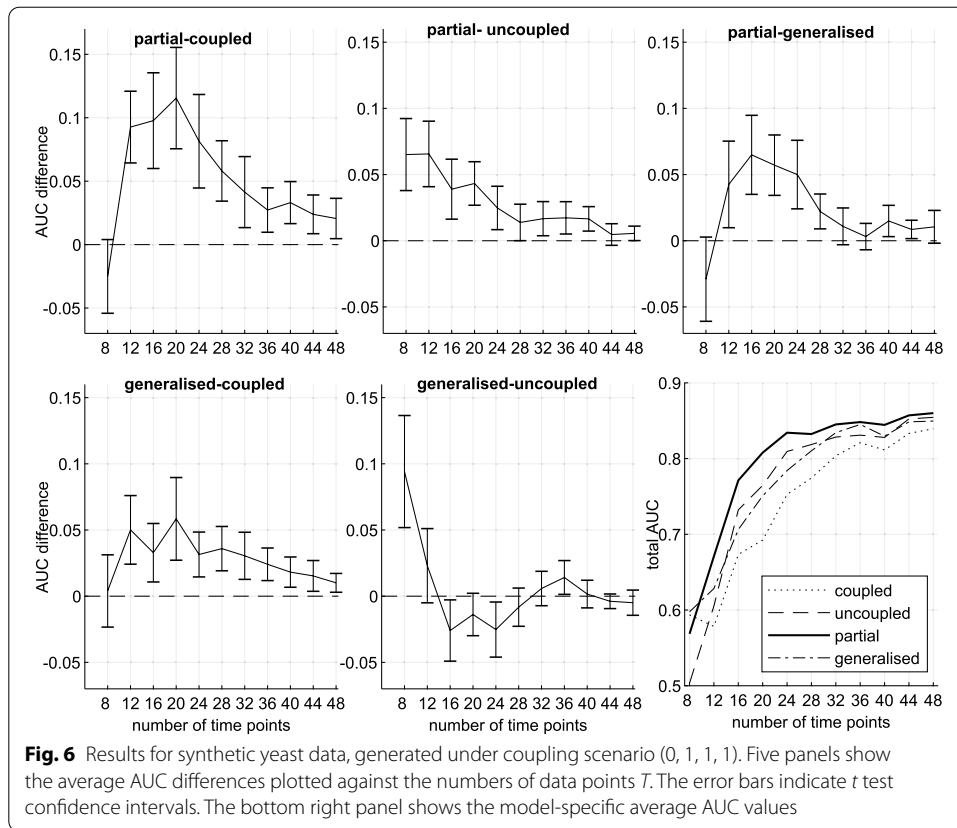
Results for synthetic network data

We start with the RAF-pathway for which we generated network data for 8 different coupling scenarios. Figure 5a compares the network reconstruction accuracies in terms of average AUC value differences. For 6 out of 8 scenarios the three AUC differences are clearly and significantly in favour of M3. Not surprisingly, for the two extreme scenarios, where all segments $h \geq 2$ are either coupled ('0111') or uncoupled ('0000'), M3 performs slightly worse than the fully coupled models (M2 and M4) or the uncoupled model (M1), respectively. But unlike the uncoupled model (M1) for coupled data ('0111'), and unlike the coupled models (M2 and M4) for uncoupled data ('0000'), the partially coupled model (M3) never performs significantly worse than the respective 'gold-standard' model. For the partially coupled model, Fig. 5b shows the posterior probabilities that the segments $h = 2, 3, 4$ are coupled. The trends are in good agreement with the true coupling mechanism. Model M3 correctly infers whether the regression coefficients stay similar (identical) or change (substantially). The generalised coupled model (M4) can only adjust the segment-specific coupling strengths, but has no option to uncouple. Like the coupled model (M2), it fails when the parameters are subject to drastic changes. When comparing the coupled model (M2) with the generalised coupled model (M4), we see that M2 performs better when only one segment is coupled, while the new M4 model is superior to M2 if two segments are coupled, see the scenarios '0011', '0110', and '0101'.

For the yeast network we generated data corresponding to a '0101' coupling scheme and the change of the parameters (from the 2nd to the 3rd segment) is less drastic than for the RAF pathway data. Figure 6 shows how the AUC differences vary with the number of time points T , where $T = 4m$ and m is the number of data points per segment. For sufficiently many data points the effect of the prior diminishes and all models yield high AUC values (see bottom right panel). There are then no significant differences between the AUC values anymore. However, for the lower sample sizes again the new partially coupled model (M3) performs clearly best. For $12 \leq m \leq 28$ model M3 is significantly superior to all other models and for $30 \leq T \leq 40$ it still significantly outperforms the uncoupled (M1) and the coupled (M2) model. The performance of the generalised model (M4) is comparable to the performance of the uncoupled model. For moderate sample sizes ($12 \leq T \leq 44$) model M4 is significantly better than the fully coupled model (M2).

Results for yeast gene expression data

For the yeast gene expression data we assume the changepoint(s) to be unknown and we infer the segmentation from the data. Figure 7a shows convergence diagnostics for the



(See figure on next page.)

Fig. 8 Results for real yeast data with fixed changepoints. We imposed $K \in \{1, \dots, 5\}$ changepoints and kept them fixed. K changepoints yield $H = K + 1$ segments. For each K we used the first changepoint to separate the two parts of the time series (galactose vs. glucose metabolism). Successively we located the next changepoint in the middle of the longest segment to divide it into 2 segments, until K changepoints were set. **a** show the model-specific average total AUC scores with error bars indicating standard deviations. **b** shows the AUC score differences in favour of the partially coupled model (M3). Here the error bars indicate t-test confidence intervals

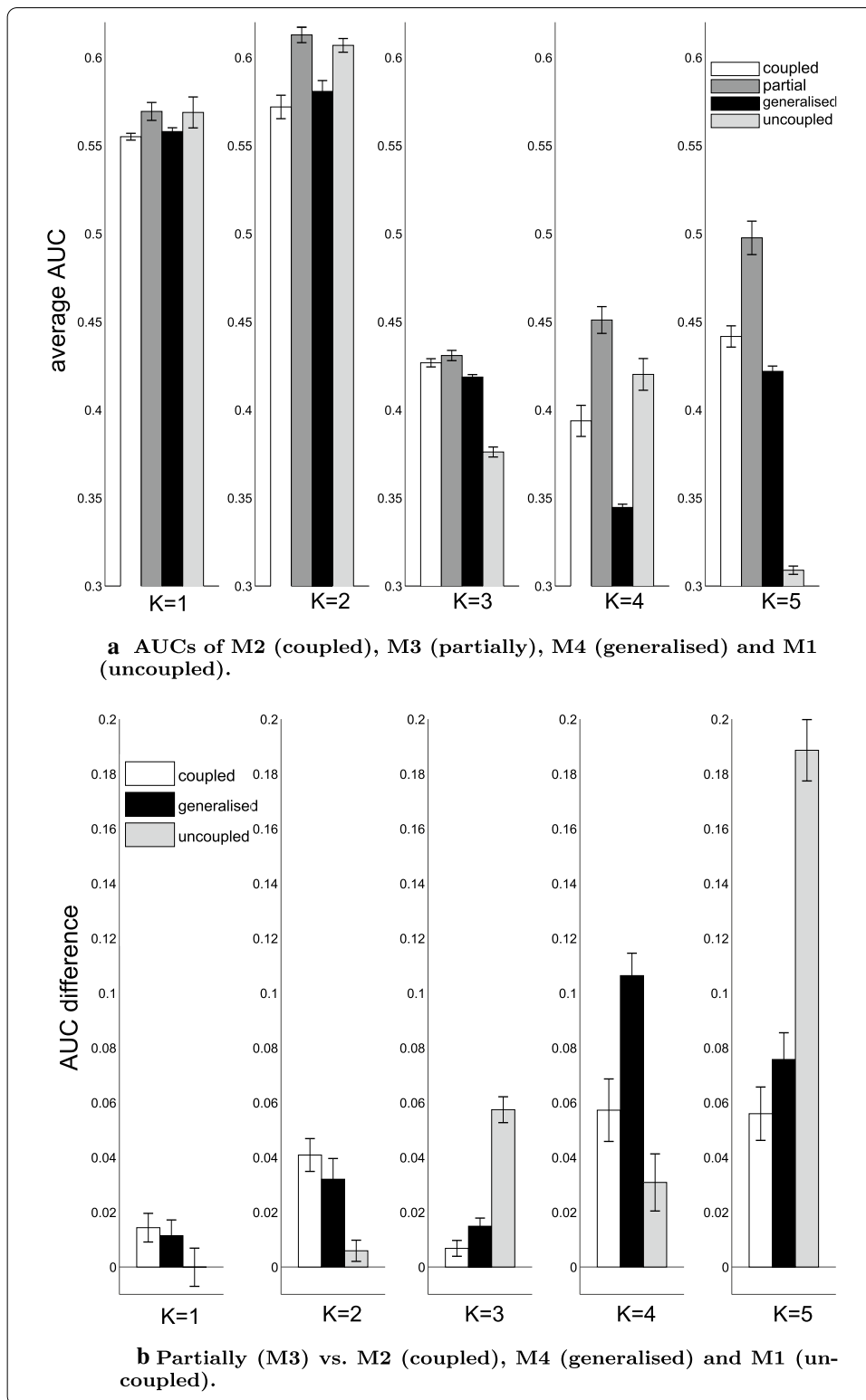
partially coupled model (M3). It can be seen from the scatter plots that $V = 10,000$ RJM-CMC iterations yield already almost perfect convergence. The edge scores of 15 independent MCMC runs are almost identical to each other.

The average AUC scores of the models M1–M4 are shown in Fig. 7b. Since the number of inferred changepoints grows with the hyperparameter p of the geometric distribution on the distance between changepoints, we implemented the models with different p 's. The uncoupled model is superior to the coupled model for the lowest p ($p = 0.02$) only, but becomes more and more inferior to the coupled model, as p increases. This result is consistent with the finding in Grzegorzczuk and Husmeier [5] and can be explained as follows: As the hyperparameter of the changepoint prior $p \in (0, 1)$ increases, the number of inferred data segments H grows so that the individual data segments $h = 1, \dots, H$ get shorter. The individual segments h then cover less data points and are thus less informative. The coupling scheme allows for information-sharing among segments. The information content of large segments is sufficient for inference, so that coupling does not provide any noteworthy advantage. But for short (uninformative) segments information coupling improves the inference certainty, as coupling allows for the incorporation of information from the preceding segment(s). Therefore the potential improvement that can be gained by coupling grows with the hyperparameter p .

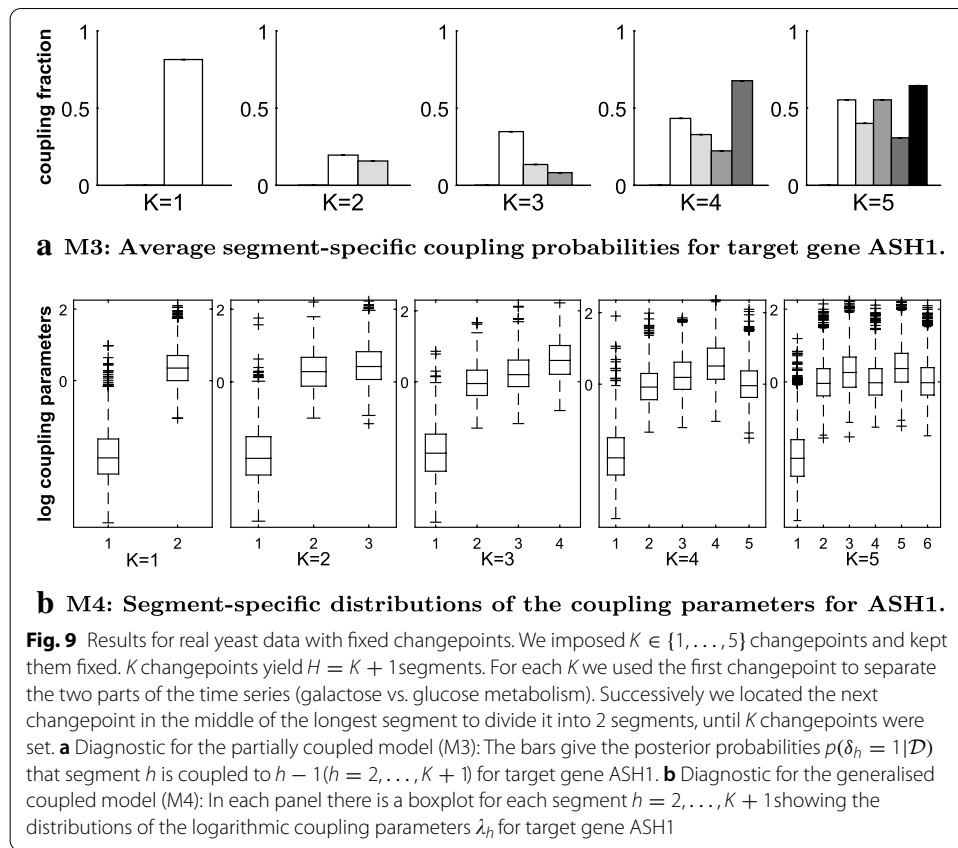
The new partially coupled model (M3) performs consistently better than the uncoupled and the coupled model (M1–M2). The only exemption occurs for $p = 0.1$ where the coupled model (M2) appears to perform slightly (but not significantly) better than M3. For p 's up to $p = 0.05$ the fully coupled (M2) and the generalised fully coupled model (M4) perform approximately equally well. However, for the three highest p 's the M4 model performs better than the coupled model (M2) and even outperforms the new partially coupled model (M3). While the performances of the models M1–M3 decrease with the number of changepoints, the performance of the model M4 stays rather robust.

Subsequently, we re-analysed the yeast data with $K = 1, \dots, 5$ fixed changepoints. Figure 8a, b shows the average AUC scores and the AUC score differences in favour of the partially coupled model (M3). Panel (a) reveals that the new partially coupled model (M3) reaches again the highest network reconstruction accuracy. Panel (b) shows that the superiority of M3 is significant, with only one exemption: For $K = 1$ the uncoupled model M1 does not perform worse than the partially coupled model (M3).

Subsequently, we also investigated the segment-specific coupling posterior probabilities $p(\delta_h = 1 | \mathcal{D})$ ($h = 2, \dots, H = K + 1$) for the new partially coupled model (M3) and the posterior distributions of the coupling parameters $\lambda_u, \lambda_2, \dots, \lambda_{K+1}$ for the generalised model (M4), but we could not find clear trends for any gene. As an example, we provide the results for gene ASH1 in Fig. 9a, b. Panel (a) shows that the coupling posterior



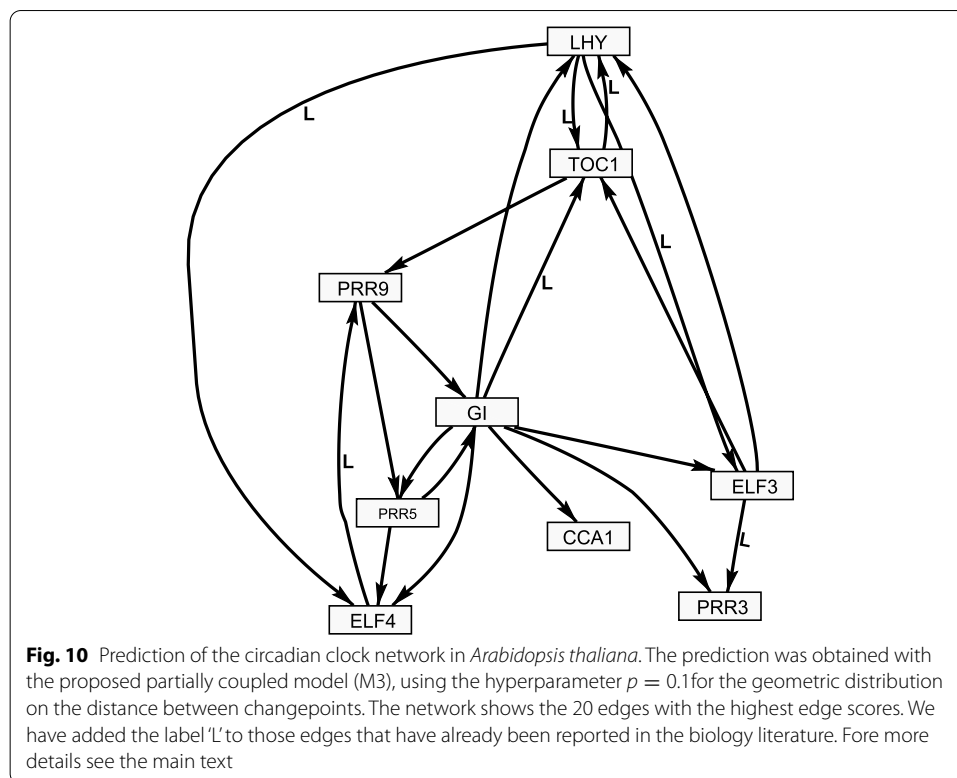
probabilities of model M3 do not have a clear pattern. However, it becomes obvious that the partially coupled model makes use of segment-wise switches between the uncoupled and the coupled approach. Panel (b) shows that the distributions of the segment-specific



coupling parameters, $\lambda_2, \dots, \lambda_{K+1}$, of model M4 stay rather similar among segments. This explains why the generalised coupled model (M4) is not superior to the fully coupled model (M2).

Application to Arabidopsis gene expression data

For the Arabidopsis gene expression data we cannot objectively compare the network reconstruction accuracies of the four models, since the true circadian clock network is not known. We therefore only applied the new partially coupled model (M3), which we had found to be the best model in our earlier studies. Figure 10 shows the Arabidopsis network, which was reconstructed using the hyperparameter $p = 0.1$ for the geometric distribution on the distance between changepoints. To obtain a network prediction, we extracted the 20 edges with the highest edge scores. Although a proper evaluation of the network prediction is beyond the scope of this paper, we note that several features of the network are consistent with the plant biology literature. E.g. the feedback loop between *LHY* and *TOC1* is the most important key feature of the circadian clock network (see, e.g., the work by Locke et al. [18]). Many of the other predicted edges have been reported in more recent works. E.g. the edges $LHY \rightarrow ELF3$, $LHY \rightarrow ELF4$, $GI \rightarrow TOC1$, $ELF3 \rightarrow PRR3$ and $ELF4 \rightarrow PRR9$ can all be found in the circadian clock network (hypothesis) of Herrero et al. [19].



Discussion and conclusions

We have proposed a new Bayesian piece-wise linear regression model for reconstructing regulatory networks from gene expression time series. The new partially coupled model (M3), whose graphical model representation is given in Fig. 2, is a consensus model between the uncoupled model (M1) and the fully coupled model (M2). In the uncoupled model (M1) the segment-specific regression coefficients have to be learned for each segment separately. In the fully coupled model (M2) each segment is compelled to be coupled to the previous one. The new partially coupled model (M3) combines features of the uncoupled and the fully coupled model, and it can infer for each individual time segment whether it is coupled to (or uncoupled from) the preceding segment.

We have cross-compared the new model (M3) with the two established models (M1–M2) as well as with the generalised coupled model (M4) that makes use of segment-specific coupling parameters [6]. In our data applications, the new partially coupled model (M3) reached significantly better network reconstruction accuracies than its competitors (M1, M2, and M4).

In an earlier work [6], we found that the performances of the fully coupled model (M1) and of the generalised fully coupled model (M4) can be improved by imposing additional hyperpriors on the hyperparameters of the coupling strength parameter. In our future work we will therefore investigate whether either the use of hyperpriors or the use of segment specific continuous (coupling/SNR) parameters along the lines of the M4 model can improve the new partially coupled model (M3). Moreover, in our future work we will also try to combine the concept of partially coupled time segments of the proposed model (M3) with the recently proposed concept of partially coupled edges [8]. The

combination of both concepts will yield a highly flexible novel NH-DBN model, in which each individual network edge is partially segment-wise coupled. We will empirically test whether this new hybrid model leads to improved network reconstruction results or whether it suffers from model over-flexibility.

Abbreviations

DBN: Dynamic Bayesian network; NH-DBN: Non-homogeneous dynamic Bayesian network; MCMC: Markov chain Monte Carlo; RJMCMC: Reversible jump Markov chain Monte Carlo; SNR: Signal-to-noise ratio; AUC: Areas under precision recall curve.

Supplementary Information

The online version supplementary material available at <https://doi.org/10.1186/s12859-021-03998-9>.

Additional file 1. Graphical model representations of the three competing models are provided as additional files. Figure 11 shows a graphical model representation of the M1 model. Figure 12 shows a graphical model representation of the M2 model. Figure 13 shows a graphical model representation of the M4 model.

Acknowledgements

Not applicable.

About this supplement

This article has been published as part of BMC Bioinformatics Volume 22, Supplement 2 2021: 15th and 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2018-19). The full contents of the supplement are available at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-22-supplement-2>.

Authors' contributions

Both authors contributed equally to the methodological work and both authors. MSK performed the computational work and drafted the manuscript. MG supervised the project and revised the draft version of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available in the figshare repository, <https://figshare.com/s/96f578777aa6b43f3638>

We note that the data stem from earlier publications [5, 17].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Prof. Cobbenhagenlaan 225, 5037 DB Tilburg, The Netherlands. ² Jheronimus Academy of Data Science, Sint Janssingel 92, 5211 DA 's-Hertogenbosch, The Netherlands. ³ Bernoulli Institute, Groningen University, Nijenborgh 9, 9747 AG Groningen, The Netherlands.

Received: 14 January 2021 Accepted: 5 February 2021

Published: 26 April 2021

References

1. Lèbre S, Becq J, Devaux F, Lelandais G, Stumpf MPH. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst Biol.* 2010;4:130.
2. Grzegorzyc M, Husmeier D. Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. *Bioinformatics.* 2011;27(5):693–9.
3. Dondelinger F, Lèbre S, Husmeier D. Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Mach Learn.* 2012;90:191–230.

4. Grzegorzczak M, Husmeier D. Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. *Mach Learn.* 2013;91:105–54.
5. Grzegorzczak M, Husmeier D. A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology. *Stat Appl Genet Mol Biol SAGMB.* 2012;11(4) (Article 7).
6. Shafiee Kamalabad M, Grzegorzczak M. Improving nonhomogeneous dynamic Bayesian networks with sequentially coupled parameters. *Stat Neerl.* 2018;72(3):281–305.
7. Shafiee Kamalabad M, Heberle AM, Thedieck K, Grzegorzczak M. Partially non-homogeneous dynamic Bayesian networks based on Bayesian regression models with partitioned design matrices. *Bioinformatics.* 2019;35(12):2108–17.
8. Shafiee Kamalabad M, Grzegorzczak M. Non-homogeneous dynamic Bayesian networks with edge-wise sequentially coupled parameters. *Bioinformatics.* 2020;36(4):1198–207.
9. Vignes M, Vandel J, Allouche D, Ramadan-Alban N, Cierco-Ayrolles C, Schiex T, Mangin B, De Givry S. Gene regulatory network reconstruction using Bayesian networks, the Dantzig selector, the Lasso and their meta-analysis. *PLoS ONE.* 2011;6(12):29165.
10. Huang X, Zi Z. Inferring cellular regulatory networks with Bayesian model averaging for linear regression (BMALR). *Mol Biol Syst.* 2014;10(8):2023–30.
11. Xing L, Guo M, Liu X, Wang C, Wang L, Zhang Y. An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. *BMC Genom.* 2017;18(9):17–30.
12. Fan Y, Wang X, Peng Q. Inference of gene regulatory networks using Bayesian nonparametric regression and topology information. *Comput Math Methods Med.* 2017;2017:8307530.
13. Xu S, Zhang C-X, Wang P, Zhang J. Variational Bayesian complex network reconstruction. *CoRR* 2018.
14. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian data analysis.* 2nd ed. London: Chapman and Hall/CRC; 2004.
15. Bishop CM. *Pattern recognition and machine learning.* Singapore: Springer; 2006.
16. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Protein-signaling networks derived from multiparameter single-cell data. *Science.* 2005;308:523–9.
17. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D, Cosma MP. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell.* 2009;137:172–81.
18. Locke JCW, Kozma-Bognár L, Gould PD, Fehér B, Kevei E, Nagy F, Turner MS, Hall A, Millar AJ. Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Mol Syst Biol.* 2006;2(1):59.
19. Herrero E, Kolmos E, Bujdosó N, Yuan Y, Wang M, Berns MC, Uhlworm H, Coupland G, Saini R, Jaskolski M, Webb A, Concalves J, Davis SJ. EARLY FLOWERING4 recruitment of EARLY FLOWERING3 in the nucleus sustains the *Arabidopsis* circadian clock. *Plant Cell.* 2012;24(2):428–43.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

