Biostatistics (2011), **12**, 4, *pp.* 682–694
doi:10.1093/biostatistics/kxr009
Advance Access publication on May 5, 2011

CD

Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression

EDWARD R. MORRISSEY

Systems Biology Centre, University of Warwick, Coventry House, CV4 7AL, Coventry, UK

MIGUEL A. JUÁREZ*

School of Mathematics and Statistics, University of Sheffield, Hicks Building, S3 7RH, Sheffield, UK
m.juarez@sheffield.ac.uk

KATHERINE J. DENBY

Warwick Life Sciences and Systems Biology Centre, University of Warwick, Wellesbourne, CV35 9EF, Coventry, UK

NIGEL J. BURROUGHS

Systems Biology Centre, University of Warwick, Coventry House, CV4 7AL, Coventry, UK

SUMMARY

We propose a semiparametric Bayesian model, based on penalized splines, for the recovery of the time-invariant topology of a causal interaction network from longitudinal data. Our motivation is inference of gene regulatory networks from low-resolution microarray time series, where existence of nonlinear interactions is well known. Parenthood relations are mapped by augmenting the model with kinship indicators and providing these with either an overall or gene-wise hierarchical structure. Appropriate specification of the prior is crucial to control the flexibility of the splines, especially under circumstances of scarce data; thus, we provide an informative, proper prior. Substantive improvement in network inference over a linear model is demonstrated using synthetic data drawn from ordinary differential equation models and gene expression from an experimental data set of the *Arabidopsis thaliana* circadian rhythm.

Keywords: Circadian clock; Gibbs variable selection; Markov process prior; Nonlinear gene regulatory networks; P-Splines regression; time course gene expression data.

1. INTRODUCTION

Our objective is the inference of gene regulatory networks (GRNs) from time series data; specifically, inferring the gene regulatory kinships for a particular process. To this end, we can conceptualize a GRN

*To whom correspondence should be addressed.

as a directed graph, with its nodes representing genes and the edges gene–gene regulation. Bayesian networks (BNs) have been used previously in gene network determination (Friedman *and others*, 2000). However, it is well known that biological processes have feedback loops and thus the validity of BNs is questionable when modeling such systems. Dynamic Bayesian networks (DBNs) have been proposed for modeling time course (longitudinal) gene expression data (Zou and Conzen, 2005). These can be thought of as "unfolding" a BN for every time point, and when folding back the network, selfregulation and cliques may be obtained. Formally, a DBN is characterized by a set of conditional relations, $p(y^{t+1}|\boldsymbol{y}^t)$. In the case of an (auto)regression-based DBN, these relations can be written as $y_i^{t+1} = f_i(\boldsymbol{y}^t) + \varepsilon_i^{t+1}$, where y_i^t is the expression measurement of gene $i = 1, \ldots, G$ at time $t = 1, \ldots, T$, $\boldsymbol{y}^t = y_1^t, y_2^t, \ldots, y_G^t$ and ε_i^t is an idiosyncratic error term. The functional forms of the interactions, $f_i(\cdot)$, are usually unknown and typically nonlinear due to the complex biochemistry behind gene regulation. Whether or not $\partial f_i(\boldsymbol{y}^t)/\partial y_j^t \equiv 0$ defines the topology of the network. The interaction topology is key in GRN, as it determines the causal relations in the gene regulatory dynamics for a given biological process. Although gene regulatory relationships can change in time, especially when dealing with varying experimental conditions (Ahmed and Xing, 2009), we assume that the data have been produced in controlled conditions and thus regulation can be suitably captured with a time-invariant network topology.

A flexible way of including unknown nonlinearities, and thus avoiding model selection issues, is to use a semiparametric specification by letting the interactions be described by spline functions. The use of splines in the estimation of GRNs has been advanced by *i.a.* Gustafsson *and others* (2005) and Kim *and others* (2004). A fundamental problem when using spline regression is knot selection that greatly influences the curve fitting. One efficient solution is to select a few well-placed knots for a given spline degree. This requires determining both the optimal number and the position of the knots, which is typically addressed by means of a transdimensional Monte Carlo Markov chain (MCMC) scheme (Ferreira *and others*, 2008; Denison *and others*, 2002) or by cross validation (Ruppert, 2002; Friedman, 1991). The efficiency gained in the modeling may be offset by mixing problems in the sampler, due mainly to the vast space that must be explored and the associated computational problems, or by the unwieldy amount of comparisons required for cross validation.

Our approach avoids such issues by relying on P-splines (Eilers and Marx, 1996; Lang and Brezger, 2004), which are characterized by specifying a rather large number of evenly spaced knots. Then, in order to avoid overfitting and also to control for the effective number of parameters to be estimated, a penalty that shrinks the spline coefficients toward the origin is specified. Such a penalty depends crucially on a so-called smoothness parameter. In this paper, we propose a fully Bayesian setup for dealing with this smoothness parameter and discuss the implications of alternative prior specifications for this key model component.

The proposed model is presented in Section 2, where we also discuss the prior specification. Implementation is briefly described in Section 3. Section 4 illustrates the application of our model to 3 examples, where we reconstruct the corresponding networks and assess their accuracy. Conclusions and possible extensions are given in Section 5. Data sets and Matlab code used in the paper are available in the supplementary material available at *Biostatistics* online and in http://majuarez.staff.shef.ac.uk/materials/index.html.

2. The model

Let y_g^t denote the gene expression level of gene $g = 1, \ldots, G$, measured at time $t = 1, \ldots, T$. We propose to model it as $y_g^t = \eta_g^t + \varepsilon_g^t$, where η_g^t is the predictor and ε_g^t is an idiosyncratic error term, centered at zero. We assume that η_g^t is determined by some unknown subset of the genes at the previous time point, and that the error terms are Gaussian and independent for all genes and time points. Thus, we

can write it as

$$y_g^t = \eta_g^t(\boldsymbol{y}^{t-1}; \boldsymbol{\theta}_g) + \varepsilon_g^t, \quad \varepsilon_g^t \sim N(\varepsilon_g^t|0, \lambda_g) \quad \text{ind.}, \tag{2.1}$$

where $\boldsymbol{y}^t = \{y_1^t, \ldots, y_G^t\}$, $\boldsymbol{\theta}_g$ is a set of parameters indexing $\eta_g^t(\cdot;\cdot)$ and $\lambda_g^{-1} = \text{Var}(\varepsilon_g^t)$.

In order to accommodate nonlinearities, the regulatory relationships are modeled by

$$\eta_g^t = f_{g1}(y_1^{t-1}) + f_{g2}(y_2^{t-1}) + \ldots + f_{gG}(y_G^{t-1}) + \mu_g, \tag{2.2}$$

where μ_g is a gene-specific constant term and $f_{gi}(y_i) = \sum_{k=1}^{M} \beta_{ik}^g B_{ik}(y_i)$. Here, $\{B_{ik}(y_i)\}$ are M B-spline basis functions of degree l defined over the set of r evenly spaced knots, $\boldsymbol{\kappa}_i = \{\kappa_{i1}, \ldots, \kappa_{ir}\}$, with $\min\{y_i\} = \kappa_{i1} < \kappa_{i2} < \ldots < \kappa_{ir} = \max\{y_i\}$, and $M = r + l$. By defining the spline design row vectors $X_j^t \in \mathbb{R}^M$, such that $X_j^t(k) = B_{jk}(y_j^t)$, we can rewrite the predictor in (2.1) as $\eta_g^t = X_1^{t-1}\boldsymbol{\beta}_{1g} + \ldots + X_G^{t-1}\boldsymbol{\beta}_{Gg} + \mu_g$, with $\boldsymbol{\beta}_{jg} = \{\beta_{j1}^g, \ldots, \beta_{jM}^g\} \in \mathbb{R}^M$ a column vector of coefficients for $j = 1, \ldots, G$. If $\|\boldsymbol{\beta}_{jg}\| \approx 0$, there is negligible influence of gene j on gene g, and thus the "link" $j \to g$ is off. If the link is on, then we say that j is a "parent" of g.

Stacking the bases and the coefficients into $X^t = \{X_1^t, \ldots, X_G^t\} \in \mathbb{R}^{MG}$ and $\boldsymbol{\beta}_g = \{\boldsymbol{\beta}_{1g}, \ldots, \boldsymbol{\beta}_{Gg}\} \in \mathbb{R}^{MG}$, respectively, and after further stacking the equations over time, we have

$$\boldsymbol{y}_g = \boldsymbol{\mu}_g + \mathcal{X}\boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_g, \quad g = 1, \ldots, G, \tag{2.3}$$

where $\boldsymbol{\mu}_g = \mu_g \boldsymbol{\iota}_T'$, with $\boldsymbol{\iota}_T$ a row vector of ones of size T and $\mathcal{X} = \{X^1, X^2, \ldots, X^T\}'$ a bases matrix of size $T \times MG$. This model is unidentifiable given that every potential parent spline contributes with its own constant term. To correct for this, we add the identifiability restriction $\boldsymbol{\iota}_T \times (\mathcal{X}\boldsymbol{\beta}_g) = 0$. We describe its implementation within the sampling scheme in the supplementary material available at *Biostatistics* online.

As it stands to estimate the $2 + M \times G$ parameters of each spline-regression component in (2.3) would require in excess of this number of data points per gene. If the number of time measurements is relatively small, one would need to select a rather small number of knots, thus effectively reducing the capacity of the splines to capture nonlinearities. We address this issue by performing a Gibbs variable selection as in Smith and Kohn (1996). The model is augmented with the indicators γ_{jg}, such that $\tilde{\boldsymbol{\beta}}_{jg} = \gamma_{jg} \times \boldsymbol{\beta}_{jg}$, where $\gamma_{jg} = 1$ if the link is on and $\gamma_{ij} = 0$ if the link is off and substituting these new coefficients into the model.

The practical advantage of augmenting with the indicators is that it allows us to make inference about the network topology, now parameterized by the connectivity matrix, $\Gamma = \{\gamma_{jg}\}$.

2.1 *The prior*

We use conditionally conjugate priors where suitable, which simplifies the sampling algorithm. We take particular care when specifying a shrinkage or penalty prior for the spline coefficients, as this determines the smoothness of the functional form fitted.

Precisions. We use conjugate, i.i.d. gamma priors, $\text{Ga}(\lambda_g|a_\lambda, b_\lambda)$, on the gene precisions, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_G\}$,

$$\pi(\boldsymbol{\lambda}) = \prod_{g=1}^{G} \frac{b_\lambda^{a_\lambda}}{\Gamma[a_\lambda]} \lambda_g^{a_\lambda - 1} \exp[-b_\lambda \lambda_g]. \tag{2.4}$$

Constant term. An independent Gaussian prior, $N(\boldsymbol{\mu}|\mathbf{0}, \tau_\mu I)$, for the gene-specific constant, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_G\}$

$$\pi(\boldsymbol{\mu}) = \left(\frac{\tau_\mu}{2\pi}\right)^{G/2} \exp\left[-\frac{\tau_\mu}{2}\boldsymbol{\mu}'\boldsymbol{\mu}\right]. \tag{2.5}$$

Network structure. We provide 2 alternatives for modeling the network topology. The first is to define the "overall network connectivity," ρ, as $P[\gamma_{jg} = 1] = \rho$ and complement it with a Beta prior, $\text{Be}(\rho|a_\rho, b_\rho)$. The full specification is then,

$$\pi(\gamma_{jg}|\rho) = \rho^{\gamma_{jg}}(1-\rho)^{1-\gamma_{jg}}, \quad g, j = 1, \ldots, G, \tag{2.6}$$

$$\pi(\rho) = [\text{B}(a_\rho, b_\rho)]^{-1}\rho^{a_\rho - 1}(1-\rho)^{b_\rho - 1} \quad 0 < \rho < 1. \tag{2.7}$$

It is well known that GRNs often present hub-like structures where a handful of genes control the regulation process almost completely and the rest of the genes have very few children, if any (see, e.g. Seo *and others*, 2009, and references therein). One can capture such features by allowing for "parent-wise connectivity," $P[\gamma_{jg} = 1] = \rho_j$ and complementing it with independent priors, that is,

$$\pi(\gamma_{jg}|\rho_j) = \rho_j^{\gamma_{jg}}(1-\rho_j)^{1-\gamma_{jg}}, \quad g = 1, \ldots, G, \tag{2.8}$$

$$\pi(\rho_j) = [\text{B}(a_\rho, b_\rho)]^{-1}\rho_j^{a_\rho - 1}(1-\rho_j)^{b_\rho - 1} \quad j = 1, \ldots, G. \tag{2.9}$$

The hyperparameters $\{a_\rho, b_\rho\}$, convey our prior knowledge about the connectivity of the network and can be set accordingly. For general purposes, we recommend setting both equal to $1/2$, as this is the reference prior for a Bernoulli experiment (Bernardo and Smith, 1994). If biological knowledge of the process demands it, it is straightforward to fix any link to be deterministically on (off) by setting $\gamma_{rl} = 1(0)$, modifying the prior accordingly.

Spline coefficients. We use a second-order Markov process prior on the coefficients $\boldsymbol{\beta}_{jg}$ to shrink them toward the origin.

$$\pi(\boldsymbol{\beta}_{jg}|\tau_{jg}) = N(\boldsymbol{\beta}_{jg}|\mathbf{0}, \tau_{jg}K), \tag{2.10}$$

where τ_{jg} are the smoothness parameters addressed below. The structure of the covariance matrix, $K = \{K_{kl}\}$, is constructed from the second-order differences between adjacent coefficients, that is, $\beta_k = 2\beta_{k-1} - \beta_{k-2}$, omitting link identifiers for simplicity (see supplementary material available at *Biostatistics* online). The prior for the 2 remaining coefficients, $\{\beta_1, \beta_2\}$, is discussed below.

Smoothness parameters. In the case of small data sets, the specification of the smoothness parameters, τ_{jg}, becomes crucial as these largely determine the fitting of the spline to the data. In the limit, when $\tau_{jg} \to 0$, an interpolating spline is fitted, while as $\tau_{jg} \to \infty$ a straight line is rendered.

A conditionally conjugate prior is the product of independent gamma distributions, $\text{Ga}\cdot|a_\tau, b_\tau)$. This specification concentrates mass around a_τ/b_τ and has a relatively large right tail for small values of b_τ. It is common to find in the literature $a_\tau = b_\tau$ and set to quite small values, for example, 0.001. This indeed is quite flat over a large range of τ, but has a mode at zero effectively giving relative importance to rougher curves and thus favoring overfitting when the data are only weakly informative. On the other hand, if mass is carried toward larger values of τ—thus favoring smoother curves—the gamma distribution tails off quite quickly to the left and experiences difficulties capturing nonlinearities, (see, e.g. Jullion and Lambert, 2007).

In order to obtain a more flexible prior specification, while retaining the conditional conjugacy, we also tried a gamma scale mixture of gammas. The resulting gamma–gamma distribution (Bernardo and Smith,

1994, p. 120; Zellner, 1971, p. 376), can achieve a larger spread than the gamma and also has a heavier right tail. It may also not have any finite moments for certain parameter values. Despite these desirable characteristics, we found that the heavy right tail of this prior, combined with the flatness of the likelihood in regions where τ is very large can lead to identifiability issues. This can be understood since there exists a threshold value, $\tau^\star$, for which the fit of the spline is practically linear and thus indistinguishable for any $\tau > \tau^\star$.

This lead us to propose an inverted Pareto prior, $\text{Ip}(\cdot|a_\tau, b_\tau)$:

$$\pi(\tau_{jg}|a_\tau, b_\tau) = \frac{a_\tau}{b_\tau}\left(\frac{\tau_{jg}}{b_\tau}\right)^{a_\tau - 1}, \quad \tau_{jg} \leqslant b_\tau, a_\tau > 0. \tag{2.11}$$

We restrict $a_\tau \geqslant 1$, to prevent concentration of mass near the origin. Setting $a_\tau = 1$ is tantamount to putting a uniform prior on $(0, b_\tau)$. The prior is concave for $1 < a_\tau \leqslant 2$ and convex for $a_\tau \geqslant 2$, gathering mass closer to b_τ as a_τ grows, thus favoring smoother curves. Values of $a_\tau > 3$ allocate too much mass close to b_τ and thus are not advisable, unless there is prior evidence for high levels of linearity. The cutoff value b_τ can be interpreted as that level of τ after which the likelihood is numerically invariant, that is, the fitted curve is practically linear.

2.2 *Posterior propriety*

In most of our intended applications, we will have a limited number of time measurements compared to the number of genes. Given that an improper prior will yield an improper posterior if the number of parents for any given gene exceeds T/M (see the supplementary material available at *Biostatistics* online), we construct a proper prior by supplying (2.10) with an independent specification for the first 2 coefficients,

$$\pi(\beta_1, \beta_2) = N(\beta_1|0, k_1)N(\beta_2|0, k_2). \tag{2.12}$$

To approximate the behavior of the improper prior, we could let $k_1, k_2 \to 0$. In situations where the data are scarce, we do not recommend this, as it will affect the stability of the posterior (Sun and Speckman, 2008). In our applications, we set $k_1 = k_2 = \tau_0$.

3. IMPLEMENTATION

3.1 *P-splines model algorithm*

As there is no closed-form expression for the posterior numerical methods are needed. We propose a Metropolis-within-Gibbs scheme that leads to a dramatic decrease in autocorrelation of the chain, compared to a Gibbs move. Details are given in the supplementary material available at *Biostatistics* online.

3.2 *A linear model*

In order to compare the network retrieval power of the splines model, we constructed a fully parametric, linear AR(1) model

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G} \beta_{jg} y_j^t + \varepsilon_g^t, \tag{3.1}$$

with the same prior specification as above, deleting the irrelevant terms.

4. Illustrations and applications

First, we analyze 2 synthetic, discrete time data sets where the data generation mechanism and the topology of the network are known. Second, we examine a synthetic data set comprising discrete time measurements drawn from a continuous time ordinary differential equation (ODE) model of a circadian clock. For our last example, we use microarray gene expression data from the *Arabidopsis thaliana* circadian clock. Details on the prior parameters specification are given in the supplementary material available at *Biostatistics* online.

4.1 *Discrete time synthetic networks*

In order to assess the network topology recovery power of our model, we produced 2 synthetic, first-order autoregressive processes. One has only linear and the second a number of nonlinear relations. In the nonlinear case, all the functional relations were produced using Hill functions, except for the self-interactions that are linear. In both cases, we set $G = 16$, $T = 40$, and $\rho \approx 0.1$.

When the topology of the network is known, we can use the receiver operating characteristic (ROC) curve to assess graphically the retrieval performance of a model. A more formal comparison can be carried out by calculating the area under the ROC curve (AUC) and the mean cross entropy (MxE). For the linear data set, the AUC were 0.999 for the fully parametric model and 0.998 with the splines; and when fitting the nonlinear data set, we obtained 0.728 and 0.912, respectively. In the linear network, the MxE was 0.042 when fitting the parametric model and 0.064 when fitting the splines; with Hill interactions the values were 0.41 and 0.22, respectively. Thus, using these scores network topology retrieval from the splines model is almost as good as that from the linear when the interactions are linear and outperforms it when nonlinearities are present (ROC curves are shown in the supplementary material available at *Biostatistics* online).

To further understand the differences between the inferred networks under either model, we plot in Figure 1, the partial and full reconstructions for gene 8's trace in the nonlinear data set, along with the posterior of the corresponding smoothing parameter. Both models provide similar predictions, as illustrated by the full reconstructions that are practically undistinguishable (Figure 1(d)). However, the way this fit is achieved varies significantly. As expected, both models have a very similar fit for the self-regulation (Figure 1(a)). As the self-interaction is linear, the splines model fits it by allocating most of the posterior mass of the corresponding smoothness parameter toward high values, depicted by the solid line in Figure 1(e). Gene 8 has one parent with a nonlinear interaction and the splines model is capable of reproducing the Hill functional relationship quite precisely (Figure 1(b)), by allocating almost all posterior mass toward small values of the corresponding smoothness parameter, shown in Figure 1(e) (dot-dashed line). Obviously, the linear model cannot accommodate such behavior and may include spurious parents in order to compensate for the lack of fit, as in this case, illustrated in Figure 1(c). In contrast, the splines model does not predict Gene 5 as a parent (solid line in Figure 1(c)). Notice the mass allocation of the self-regulation link (solid) in Figure 1(e): it is basically drawn from the prior (dashed), illustrating that our specification is adequate for linear relations to be reproduced accurately.

When the network topology—that is, the biological model—is fixed, we can compare the fit of alternative statistical models using formal tools. We calculated the deviance information criterion (DIC) obtained from the different data/model combinations used in this paper by fixing the network topology to those links with posterior probability larger than 0.8 (Table 1). In the linear data case, both models produced similar estimates of the connectivity matrix and therefore their AUC and MxE scores are quite close to each other. However, the DIC indicates that the linear model is preferred to the splines, mainly due to the costs associated with the additional complexity of the splines model, unnecessary for this data set. In contrast, the DIC from the nonlinear data set favors the splines model, granting the increase in model complexity.

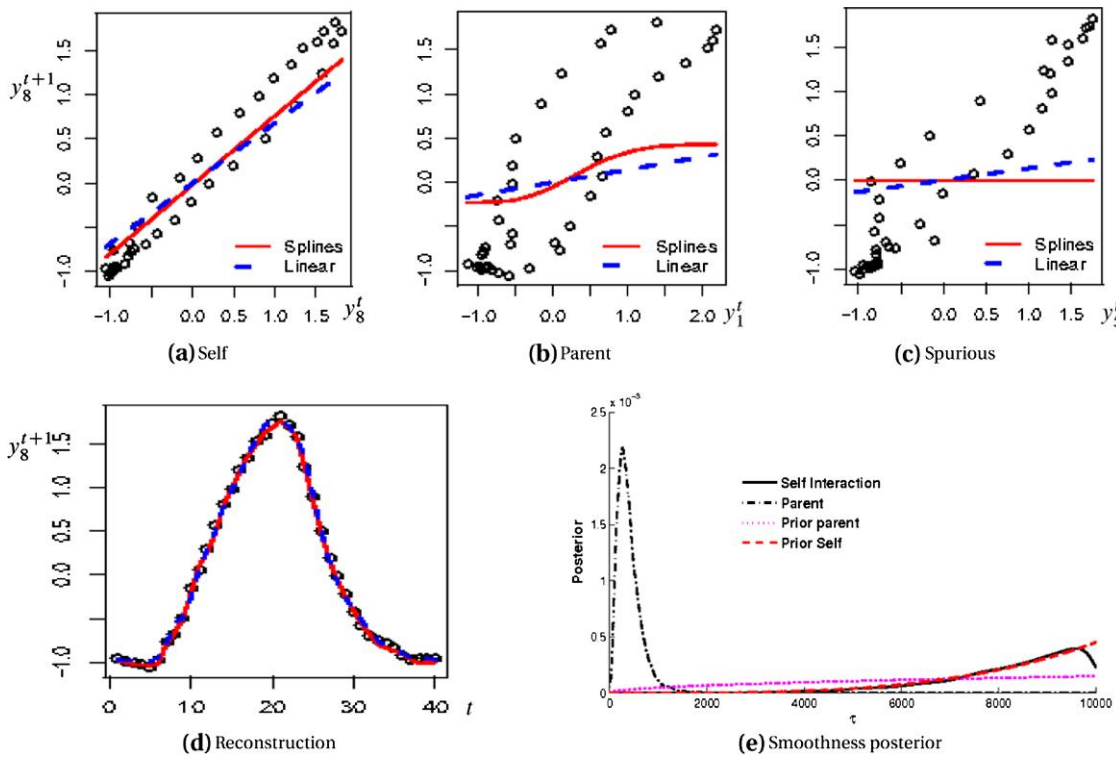


Fig. 1. Partial and full reconstructions of Gene 8's trace using splines (solid) and linear (dashed) models. The circles represent a scatter plot of the expression values of Gene 8 against 3 potential parents (Genes 8, 1, and 5) at the previous time point. (a) Both models capture the linear self-regulation. (b) The true parent is predicted in both models, while splines is able to reproduce the Hill functional relationship. (c) The linear model predicts one spurious parent. (d) Trace reconstruction from both models is almost identical. (e) Marginal posterior distributions of the smoothing parameter for the self-regulation (solid, prior dashed) and nonlinear parent (dot-dashed, prior dotted).

Table 1. *Comparing model fit with a fixed network topology. DIC values obtained from the data sets used in the paper when fitting the linear and the splines models. The network topology is fixed by selecting those links with posterior probability above* 0.8. *The model preferred by DIC is highlighted in bold font*

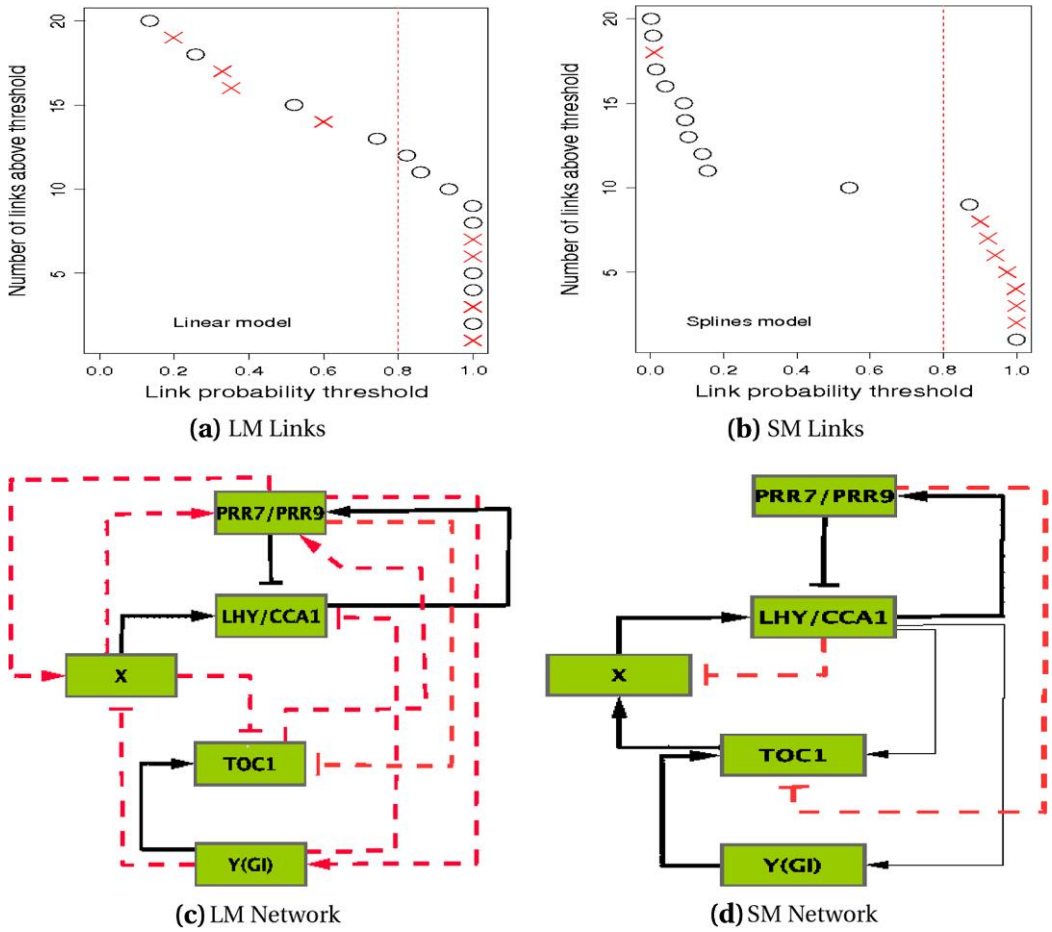
Data set	Linear model	Splines model
Synthetic linear	$\mathbf{2.02 \times 10^4}$	2.56×10^4
Synthetic Hill	3.01×10^4	$\mathbf{3.64 \times 10^3}$
ODE data	2.44×10^5	$\mathbf{2.34 \times 10^5}$
Microarray data	9.95×10^3	$\mathbf{1.93 \times 10^3}$

4.2 *Biological GRN: the plant Circadian clock*

In the following sections, we focus on a partially known GRN, specifically the plant *A. thaliana* circadian clock. Locke *and others* (2006) developed an ODE model of the clock, which we use below for generating synthetic observations. The current working biological model is due to McClung (2008). Both models include nodes *X* and *Y*, representing genes that are thought to be involved in the circadian clock, but

whose identity remains unknown. These network models are shown schematically and further explained in the supplementary material available at *Biostatistics* online.

Differential equation data. We generated data from the ODE model fixing the light source to be permanently on. The data were then subsampled, logged, and standardized. The resulting data set has 50 time points with a time spacing of 1 h. We present the results obtained using the parent-wise connectivity structure (2.8)–(2.9). In order to interpret the output, rather than examining the ROC curves, we analyze the inferred network at a given threshold. This is more convenient given that there are only a few genes and therefore a more detailed comparison with the true network is possible. We plot the number of links included in the predicted network against the posterior link probability when fitting the linear model, Figure 2(a), and when using the splines model, Figure 2(b). We use a cross (circle) for a correctly



(a) LM Links **(b)** SM Links **(c)** LM Network **(d)** SM Network

Fig. 2. Network topology inference on the ODE circadian clock data. (a) The number of links predicted to be present in the network versus posterior link probabilities estimated when using the linear model (LM) and (b) with the splines model (SM). Crosses (circles) represent correctly (incorrectly) predicted links. (c) The network obtained with a threshold of 0.8 using the LM and (d) when using the SM. Solid lines represent correct predictions, dashed lines incorrect predictions, and thin lines correct predictions, but with the wrong sign.

(incorrectly) predicted link; for instance, the predicted network with the splines model using a threshold of 0.85 would have 9 links (circles and crosses with link probability above 0.85 in Figure 2(a)), 7 out of these correct. It is apparent that the splines model produces a better separation in the link probabilities, classifying all but one link into 2 populations: a low probability (below 0.2) and a high probability (above 0.8) group. This contrasts with the linear model Figure 2(a) where almost 40% of the links are in the ambiguous region between 0.2 and 0.8.

Using 0.8 as the threshold value, we plot the reconstructed networks for both models in Figure 2(c) and (d). The inferred network for splines (Figure 2(d)) contains all correct links from the network, except for the TOC1–Y link. There are 2 spurious links (LHY/CCA1–X and PRR7/PRR9–TOC1) and 2 links with incorrect signs (LHY/CCA1–TOC1 and LHY/CCA1–Y). In addition to only having half of the correct links, the inferred network for the linear model adds a large number of spurious links (false positives). Again, we found cases where the splines model correctly predicts a single parent using a nonlinear interaction, whereas the linear model predicts that link but adds extra spurious links to improve the fit (not shown). Moreover, the DIC of the estimated network topology from the splines model is smaller than that from the linear model at threshold 0.8 (see third row of Table 1).

Experimental data. We use our methods on gene expression time series for *Arabidopsis* leaves generated using microarrays and analyze the output using the parent-wise connectivity structure. The separation of posterior link probabilities into groups is no longer as pronounced as in the synthetic data—see Figure 3(a) and (b). This may be due to the combination of a high level of noise and fewer time points. The networks inferred by each model, using a threshold of 0.8, are shown in Figure 3(c) and (d). All links predicted by the splines model appear in the linear model reconstruction. However, the linear model predicts an additional 2 parents for ELF4 and an additional 3 parents for LUX. Among those additional links are TOC1–ELF4 and TOC1–LUX, which while we have marked as correct on the plot (for consistency with the current working model), are probably incorrect. Those links were included in the accepted model as an indication that TOC1 regulates some gene (X) that in turn regulates LHY, but neither of the genes are predicted to regulate LHY. Furthermore, from the previous examples, it is clear that the linear model tends to add spurious parents.

Although only mild nonlinearities are found, the DIC score indicates a better fit from the splines model than the linear alternative for the given threshold (last row of Table 1), suggesting that even mild departures from linearity can have an important effect in model fit.

For network reconstruction, we have used a subjective choice for the posterior probability threshold. Moreover, as our reconstruction is based solely on the individual link marginal probabilities, possible correlations between these are disregarded. In order to provide a graphical representation of the uncertainty in our network retrieval, in Figure 4, we plot a heatmap with the distribution of the number of parents for each gene in the clock (left), together with a heatmap with the marginal link probabilities of its top 4 potential parents (right). These complementary sources of information render a picture of the uncertainty in the retrieval of the network topology. For instance, the splines model predicts one parent for LHY with very high probability and there is only one potential parent with high marginal probability, suggesting a very confident prediction; in contrast, the linear model predicts 1 or 2 parents with a mild probability (and 3 with a very slight probability), while one of the potential parents has a high marginal probability, there are 2 more with intermediate marginal probabilities, suggesting an ambiguity in the identity of the second potential parent. Overall, there is a shift to the left of the distribution for the number of parents from the splines model compared to the linear model, strengthening the evidence for overfitting in the latter model. Likewise, marginal link probabilities for the splines model seem to be higher over a smaller number of potential parents, thus suggesting a decrease in the uncertainty in topology retrieval compared to the linear model.

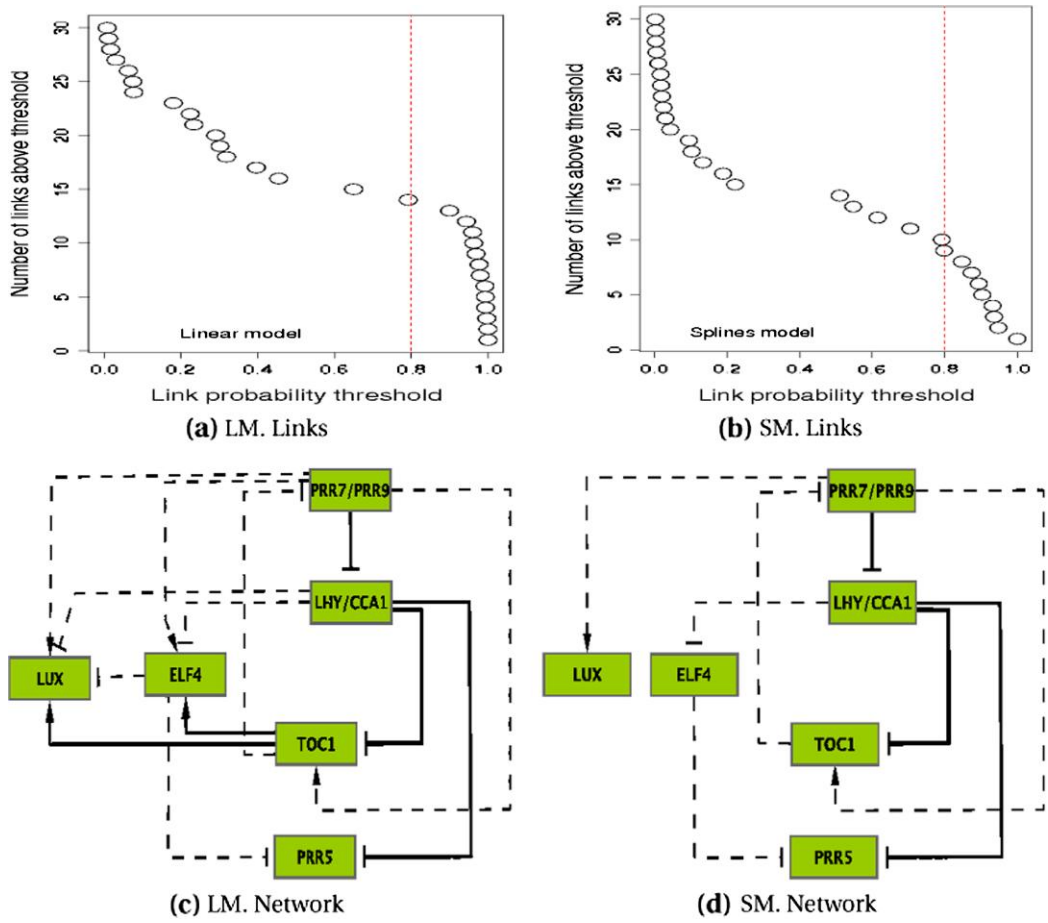


Fig. 3. Network topology inference on the circadian clock microarray data. (a) The number of links predicted to be present in the network versus posterior link probabilities estimated when using the LM and (b) with the SM. (c) The network obtained with a threshold of 0.8 using the LM and (d) when using the SM. Solid lines represent inferred links that are included in the currently accepted model (most of which have been experimentally confirmed) and dashed lines inferred links absent in the accepted model (though not necessarily incorrect).

5. Discussion

We have presented a fully Bayesian implementation of *P*-spline based inference of a DBN within a sparse connectivity context. Our motivation is the inference of GRNs from longitudinal data, for instance, from microarray time series data. Despite being capable of measuring up to tens of thousands of genes simultaneously, currently available microarray time series are typically shorter than 20 time points. This introduces significant problems for analysis and modeling, particularly as it limits the complexity of the models that can be used. We addressed this issue through use of spike-and-slab type priors that limit the connectivity of the GRN. Within this context, we are able to increase regression model complexity, designing a method for exploring whether nonlinear regulatory mechanisms are present in time series data.

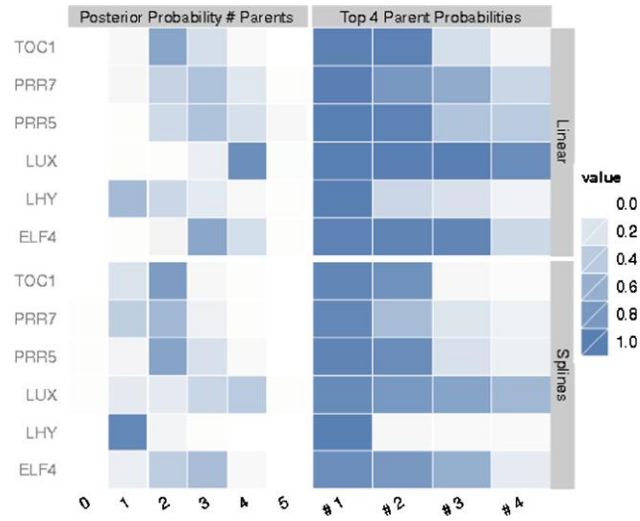


Fig. 4. Uncertainty in *Arabidopsis* circadian clock gene network reconstruction. On the left, a heatmap of the distribution of the number of parents for each gene in the clock, estimated using microarray data with the linear (top) and splines (bottom) models. On the right, a heatmap with the marginal probabilities of the top 4 potential parents.

Our model successfully identified nonlinear interactions on simulated data (both discrete time and ODE models), while the corresponding DIC scores favored the estimated network topologies with the splines model, for a given threshold, over the linear alternative. On simulated data with nonlinear interactions, the inferred GRN under a linear model typically acquired additional parents, these incorrectly predicted parents improved the fit to a similar quality to that achieved by the P-splines model. The P-splines model also enhances network sparsity since an additional parent under a splines regression model incurs a greater penalty than a parent with a linear functional dependence model given the higher (model) complexity; thus even when the links are actually linear there is stronger control on the number of parents. This compares to artificially imposed parent number penalization, for instance, through an arbitrary weighting $\exp(-n)$ for n parents, as in Kim *and others* (2004).

Assessing the uncertainty in network topology retrieval is an active field of research. We provide a graphical representation that combines the information on the distribution on the number of parents for each gene, with the marginal posterior probabilities of the most probable regulators. In our example using microarray data from *A. thaliana* leaves, joint inspection of these estimates suggests that the splines model provides a more accurate network reconstruction compared to the linear model.

Use of splines in inference requires handling of their functional flexibility. We recommend that the number of knots is much smaller than the number of time points; here, we presented results using 10 knots for a time series with 40–50 time points. We found that doubling the number of knots (20) gave severe problems in the mixing of the chain, while using a smaller number (7) gave similar results. We also use a prior on the coefficients that effectively controls the spline curvature. This entails choice of the value of the smoothing parameter τ; previous authors have optimized and fixed it before estimating the regression. We propose a fully Bayesian approach, inferring it concomitantly with the regression and performed a sensitivity analysis to confirm our prior is sufficiently weak, further confirming that linear relations can be retrieved. Network connectivity and spline smoothness were regression/gene specific; this allowed for both heterogeneity in the nonlinearity and the number of parents across the network. We presented the results for parent-wise connectivity and the proposed Beta prior parameters that we expect to

be appropriate for data sets similar as those used in this paper. Moreover, performing a sensitivity analysis by modifying these values in the region $(1/2, 2)$, and also restricting to an overall connectivity did not affect the results significantly in our examples. However, we have found that when the number of genes increases significantly with respect to the number of time points there might not be enough information for using the parent-wise prior and we suggest using an overall/global connectivity model.

Our P-splines model can be extended and modified for specific purposes. First, we model only direct, first-order filiation. One can extend the present model for allowing higher degree interactions, for example, by using tensor product splines. The main hindrance would then be the combinatorial growth of the topology space, and efficient methods for exploring it must be devised. Second, spline coefficient shrinkage can be performed in a number of ways. Additional constraints can be used, including a further term on the prior for the spline coefficients, $N(\boldsymbol{\beta}|\mathbf{0}, \omega H)$, with H derived from the first-order differences of adjacent coefficients. This effectively penalizes large first-order differences and favors less jagged curves, depending on the value of $\omega > 0$. Additionally, the shape of the functional form the spline may take can also be further restricted. For instance, many gene regulatory effects are monotonic. Extending the model to include monotonicity restrictions is feasible by providing such information through a prior (Ansley *and others*, 1993). Finally, the splines model can be utilized to infer the functional form of the regulation, and coupled with current biological knowledge, serve as a basis of a tailor-made parametric model.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors gratefully acknowledge the assistance of Kathryn Richardson. *Conflict of Interest:* None declared.

Funding

Engineering and Physical Sciences Research Council through the WSB-DTC (EP/d508150/1 to E.R.M.); M.A.J. carried out the research while working in WSBC and was funded by Biotechnology and Biological Sciences Research Council through the SysMO initiative (BB/ff003498/1 to M.A.J); The experimental data were provided by K.J.D. through the PRESTA project (BB/f005806/1 to K.J.D.).

References

Ahmed, A. and Xing, E. P. (2009). Recovering time-varying network dependencies in social and biological studies. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11878–11883.

Ansley, C. F., Kohn, R. and Wong, C. M. (1993). Nonparametric spline regression with prior information. *Biometrika* **80**, 75–88.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Chichester: Wiley.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalised likelihood (with discussion). *Statistical Science* **11**, 89–121.

Ferreira, J. T. A. S., Juárez, M. A. and Steel, M. F. J. (2008). Directional log-spline distributions. *Bayesian Analysis* **3**, 267–315.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. AND PE'ER, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**, 601–620.

GUSTAFSSON, M., HÖRQUIST, M. AND LOMBARDI, A. (2005). Constructing and analysing a large-scale gene-to-gene regulatory network—Lasso-constrained inference and biological validation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**, 254–261.

JULLION, A. AND LAMBERT, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis* **51**, 2542–2558.

KIM, S. Y., IMOTO, S. AND MIYANO, S. (2004). Dynamic Bayesian network and nonparametric regression for nonlinear modelling of gene networks from time series gene expression data. *Biosystems* **75**, 57–65.

LANG, S. AND BREZGER, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

LOCKE, J. C. W., KOZMA-BOGNR, L., GOULD, P. D., FEHÉR, B., KEVEI, E., NAGY, F., TURNER, M. S., HALL, A. AND MILLAR, A. J. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology* **2**, 59.

MCCLUNG, C. R. (2008). Comes a time. *Current Opinion in Plant Biology* **11**, 514–520.

RUPPERT, D. (2002). Selecting the number of knots for penalised splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

SEO, C. H., KIM, J. R., KIM, M. S. AND CHO, K. H. (2009). Hub genes with positive feedbacks function as master switches in developmental gene regulatory networks. *Bioinformatics* **25**, 1898–1904.

SMITH, M. AND KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–343.

SUN, D. AND SPECKMAN, P. (2008). Bayesian hierarchical linear mixed models for additive smoothing splines. *Annals of the Institute of Statistical Mathematics* **60**, 499–517.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

ZOU, M. AND CONZEN, S. D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**, 71–79.

[*Received August 26, 2010; revised February 16, 2011; accepted for publication February 27, 2011*]