Inferring the topology of a non-linear sparse gene regulatory network using fully Bayesian spline regression

Edward R. Morrissey, Miguel A. Juárez*, Nigel J. Burroughs
Systems Biology Centre
University of Warwick, UK

Katherine J. Denby
Systems Biology Centre and HRI
University of Warwick, UK

Abstract

We propose a semi-parametric Bayesian model, based on penalised splines, for the recovery of the topology of an interaction network from longitudinal data. Our motivation is inference of gene regulatory networks from low resolution microarray time series ($10 - 50$ time points). 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 and analyse sensitivity. The posterior is analytically intractable and numerical methods are needed. A Metropolis-within-Gibbs sampler is proposed, with a novel Metropolis-Hastings step for sampling the topology and the spline coefficients simultaneously. We also construct a linear model for comparison purposes. Model fit is illustrated using synthetic data drawn from ODE models and and gene expression from an experimental data set of the *Arabidopsis thaliana* circadian rhythm.

KEYWORDS: Circadian clock, Gibbs variable selection, Markov process prior, non-linear gene regulatory networks, P-Splines regression, time course gene expression data.

1 INTRODUCTION

Modern DNA sequencing technology has enabled the genome of many organisms to be determined ranging from small unicellular genomes such as those of bacteria and yeast, to complex large genomes, such as for humans (see *e.g.* `www.ncbi.nlm.nih.gov` and `www.ebi.ac.uk`). Knowledge of the genome is fundamental, given that gene/protein interactions are responsible for cell (and tissue) function. A key component of this interaction process is gene regulation, specifically the degree to which a gene is expressed, when, and for what duration. Gene expression is the conversion of the (sequence) information in the gene into messenger RNA (mRNA) by a process termed *transcription* and then to a functional protein by *translation*, resulting in the phenotypic expression of the gene. Gene expression, quantified as the amount of mRNA, can be measured by a number of technologies, the most common being microarrays although sequencing (RNA-seq) is becoming increasingly popular. Despite knowledge of the genome sequence, a key issue is identifying what events and which proteins control expression of individual genes. This information cannot presently be determined from the genome sequence alone.

Gene regulation is a complex process, the actual regulators being proteins, or complexes of proteins that bind just upstream of the gene and regulate the rate of that gene's transcription. Various processes can regulate the amount of active protein at a given time, regulation that depends on the nature of the organism and the specific gene. Specifically, regulation can occur through the rate of transcription, modification of the mRNA (*e.g.* splicing),

*Corresponding author: Miguel A. Juárez, University of Warwick, Systems Biology Centre, Coventry House, CV4 7AL, Coventry, UK. Email: `m.a.juarez@warwick.ac.uk`

decay rate of the mRNA, the rate of translation, the folding efficiency of the protein, the decay rate of the protein and through a number post-translational modifications, such as phosphorylation, that affect the activity of the protein, *i.e.* its capability of carrying out specific functions. Regulation of a gene's transcription rate is then an integration of all these processes applied to its regulators, coupled with the kinetics and combinatorics of those regulators in binding to the gene's regulatory sequences. In this manner, by mapping gene regulation back to genes, we obtain a vast interconnecting gene regulatory network (GRN). Many biotechnical and medical problems have at their core a GRN, for instance, the immune system (Gilchrist *et al.*, 2006), the mode of action of antibiotics (Kohanski *et al.*, 2008) and crop yield (Yin *et al.*, 2004).

The complexity of gene regulation, and the complex association of steps from transcription through to functional protein make inference of gene-gene regulatory interactions from mRNA data difficult; essentially we only have transcriptional changes, both in the regulator and regulatee from which to infer regulatory interactions. The catalogue of post-transcriptional changes mentioned above are extremely difficult to measure on a global basis, and a partial analysis can only be performed on a small number of select proteins. From a modelling standpoint, a GRN can be thought of as a directed graph, with the nodes representing genes and the edges regulatory relationships (Gardner *et al.*, 2000; Hache *et al.*, 2009; Hartemik, 2005; Ronen *et al.*, 2002). One of the main obstacles when retrieving the interaction structure, *i.e.* the topology, of a GRN is the combinatorial growth of the network topology space with respect to the number of genes considered. Further, it has been proven that finding the maximum a posteriori (MAP) network is an NP-hard problem (Shimony, 1994). Thus, several approximations to estimating the topology of the network have been proposed, for instance, Chen and Zheng (2009), Ahn *et al.* (2009) and Schäfer and Strimmer (2005) use (partial) correlations as an association measure and select the significant edges by controlling for the false discovery rate; while Faith *et al.* (2007) employs a variant of the Kullback-Liebler divergence as a measure of association and selects the edges by testing the estimated divergences against a background distribution of the scores related to the relevant node. Reviews are provided *i.a.* Bansal *et al.* (2007), Jaffrezic and Tosser-Klopp (2009) and Yu *et al.* (2004).

In any living organism, the cell has to cope with a plethora of environmental conditions, and thus has readily available adaptation mechanisms. For any specific process, typically only a small fraction of the genome is active, *e.g.* the SOS pathway in *Escherichia coli* (Courcelle *et al.*, 2001). This modularity and specialisation gives the regulatory network a key characteristic that is vital in inference, essentially the network is sparse with considerably fewer than the $G(G-1)$ possible interactions on G genes.

Using time series data, our main objective is inference of GRNs. In particular, we are interested in inferring the gene regulatory kinships for a given process. When stated in a graphical manner such relationships are represented as edges and the genes as nodes; these edges can be directed or not, and feedback loops and cliques may or may not be allowed. One very well known example of such graphs is a directed acyclic graph, characterised by directed edges and a tree structure (Bang-Jensen and Gutin, 2009; Jensen and Nielsen, 2007). From a Bayesian perspective, which will be followed in this paper, these are usually estimated using Bayesian Networks (Cowell *et al.*, 1999; Lauritzen, 1996).

Bayesian networks (BN) have been used previously in gene network determination (Friedman, 2004; Friedman *et al.*, 2000; Hongqiang *et al.*, 2005). However, it is well known that biological processes have feedback loops and thus the validity of BNs is questionable when modelling such systems. Dynamic Bayesian networks (DBN) have been proposed for modelling time course (longitudinal) gene expression data (Cao and Zhao, 2008; Kim *et al.*, 2003; Murphy and Mian, 1999; Perrin *et al.*, 2003; Zou and Conzen, 2005). These can be thought of as "unfolding" a BN for every time point and when folding back the network self-regulation and cliques may be obtained. Formally, a DBN is characterised by a set of conditional relations, $\mathrm{p}\left(y^{t+1} \mid \boldsymbol{y}^{t}\right)$. In the case of a regression based DBN these relations can be written as $y_i^{t+1} = f_i\left(\boldsymbol{y}^t\right) + \varepsilon_i^{t+1}$, where y_i^t is the measurement of gene $i = 1, \ldots, G$, at time $t = 1, \ldots, T$, $\boldsymbol{y}^t = \left\{y_1^t, y_2^t, \ldots, y_G^t\right\}$ and ε_i^t is an idiosyncratic error term. The functional forms of the interactions,

$f_i(\cdot)$, are usually unknown. Whether or not $\partial f_i\left(\boldsymbol{y}^t\right)/\partial y_j^t \equiv 0$ 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. In turn, the estimated network can be used to propose new experiments —*e.g.* gene knockouts— to further understand said process. This programme of experimenting-modelling-hypothesising-experimenting is at the heart of Systems Biology.

Irrespective of the actual form of $f_i(\cdot)$, DBN are typically heavily parameterised. If we consider a network of G genes, we need to estimate G^2 interactions (edges), in addition to any other parameters involved in the model. A common approach is to assume the simplest form of interaction, specifically a linear form. In biochemical reaction modelling, regulatory relations are frequently modelled as systems of ordinary differential equations (ODEs), with monotonic functional interactions, and thus linearity is justified as a first order Taylor expansion (di Bernardo *et al.*, 2005; Bonneau *et al.*, 2006; Gardner *et al.*, 2003). Frequently, however, the linear assumption does not hold in practice. This may be due to a large spacing between measurements, thus rendering the linear approximation invalid. Another reason is that some of the interactions are indeed highly non-linear, given the extensive set of processes involved in protein expression, and the fact that regulators often work together, either synergistically, for instance through binding to form a larger complex, or antagonistically, for instance competing for binding to overlapping regulatory regions on the genome. Further, even when linearity is appropriate, its range may be limited through saturation effects (*e.g.* unlimited amounts of mRNA cannot be produced), or there may be a minimum amount of any given regulator required in order for a reaction to take place or to be detectable by the measurement device. Linear specifications of the network interactions are incapable of capturing such effects. Specific forms of nonlinearity for the interactions can be assigned (power, exponential, Michaelis-Menten, etc.); however, misspecification of the actual shape may yield a spurious estimate of the network topology.

A flexible way of including unknown non-linearities, and thus avoiding model selection issues, is to use a semi-parametric specification by letting the interactions be described by spline functions. There is a vast literature on spline curve smoothing (Denison *et al.*, 1998; Dierckx, 1993; Fan and Gijbels, 1996; Wahba, 1990; Wand and Jones, 1995) and spline regression (Biller, 2000; Green and Silverman, 1994; Marx and Eilers, 1998; Wu and Zhang, 2006). Within GRNs, smoothing of discretely observed gene expression time series with splines to aid in the retrieval of GRNs has been advanced by *e.g.* Gustafsson *et al.* (2005), Opgen-Rhein and Strimmer (2006), Toyoshiba *et al.* (2004), while Kim *et al.* (2004) use spline regression for GRN inference. One fundamental problem when using spline regression is knot selection which greatly influences the curve fitting. One efficient solution is to select a few well placed knots for a given spline degree. This implies determining both the optimal number and position of the knots, which is typically addressed by means of a trans-dimensional MCMC scheme (Denison *et al.*, 2002; DiMatteo *et al.*, 2001; Ferreira *et al.*, 2008) or by cross-validation (Friedman, 1991; Ruppert, 2002). The efficiency gained in the modelling 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 (Brezger and Lang, 2008; Eilers and Marx, 1996; Lang and Brezger, 2004), which are characterised by specifying a rather large number of evenly spaced knots. Then, in order to avoid over-fitting and also to control for the effective number of parameters to be estimated, a penalty that shrinks the spline coefficients towards the origin is specified. Such a penalty depends crucially on a so-called smoothness parameter. Imoto *et al.* (2002) propose a semi-parametric P-spline regression model for GRN retrieval, optimising the smoothness parameter using a modified BIC and then performing a greedy search on the network topology space. Imoto and Konishi (2003) discuss the use of a modified AIC for optimising the smoothing parameter in a similar context. In this paper, we propose a fully Bayesian set up for dealing with this smoothness parameter and discuss the implications of alternative prior specifications for this key model component.

Nonlinear interaction models entail a larger number of parameters compared to the linear case. Given the underdetermined nature of network inference typically presented by microarray data on small time series data

sets, this increase in complexity needs careful handling. Our solution is to impose sparsity on the network, as is biologically justified, through use of a spike-and-slab prior. We augment our model with kinship indicator variables, and determine the network topology by carrying out a Gibbs variable selection procedure. This leads to a dramatic improvement in power, and thus the availability of information on many of the inferred links. Through careful construction of a prior to control spline complexity, we achieve a balance between limited information and model complexity.

Given the significant interest in estimating GRNs, there is a vast growing literature on this problem. Most of the approaches dealing with longitudinal GRNs are based on autoregressive models, and given that the number of time measurements is rather small, an AR(1) specification is pervasive. For instance, Opgen-Rhein and Strimmer (2007) uses a linear AR(1) specification for the regulatory interactions and shrink the AR coefficients using a James-Stein type estimator. Network retrieval is performed by identifying the significant partial correlations by means of a local false discovery rate. Lèbre (2009) proceeds in a very similar fashion, from the same linear setup partial regression coefficients are tested to reduce the space of possible links. The final network is obtained by performing a further t-test on the estimated regression coefficients of the restricted network. In contrast, our splines model is semi-parametric, allowing for non-linear interactions within a context of simultaneous topology (edge) selection achieved through use of Gibbs variable selection.

The model proposed is presented in Section 2, where we also discuss the prior specification. Given that it is common to specify an improper prior for this kind of model, we provide sufficient conditions for posterior propriety, and also provide an alternative, proper prior. The resulting posterior distribution is intractable analytically and in Section 3 we provide an MCMC scheme for sampling the posterior with a novel Metropolis-Hastings step which improves mixing and convergence of the chain. Section 4 illustrates the application of our model to three examples, where we reconstruct the corresponding networks and assess their accuracy. We also provide some guidelines for calibrating the prior. Conclusions and possible extensions are given in Section 5. Data sets and Matlab code used in the paper are available upon request.

2 THE MODEL

Let y_g^t denote the gene expression measurement of gene $g = 1, \ldots, G$, at time $t = 1, \ldots, T$. We propose to model it as

$$y_g^t = \eta_g^t + \varepsilon_g^t \ , \tag{1}$$

where η_g^t is the predictor and ε_g^t is an idiosyncratic error term, centred 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 (1) as

$$y_g^t = \eta_g\big(\boldsymbol{y}^{t-1}; \boldsymbol{\theta}_g\big) + \varepsilon_g^t \ , \qquad \varepsilon_g^t \sim \mathrm{N}\big(\varepsilon_g^t \mid 0, \lambda_g\big) \quad \text{ind} \ . \tag{2}$$

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

In order to accommodate nonlinearities, we use a flexible, non-parametric setting for the mean level, based on B-splines (Eilers and Marx, 1996). Thus,

$$\eta_g^t = f_{g1}\big(y_1^{t-1}\big) + f_{g2}\big(y_2^{t-1}\big) + \cdots + f_{gG}\big(y_G^{t-1}\big) + \mu_g \ , \tag{3}$$

where

$$f_{gi}(y_i) = \sum_{k=1}^{M} \beta_{ik}^g B_{ik}(y_i) \ .$$

Here, μ_g is a gene-specific constant term, $M = r + l$ is the number of spline basis functions, $B_{ik}(y_i)$, of degree l, defined over the set, $\boldsymbol{\kappa}_i = \{\kappa_{i1}, \ldots, \kappa_{ir}\}$, of r evenly spaced knots,

$$\min\{y_i\} = \kappa_{i1} < \kappa_{i2} < \cdots < \kappa_{ir} = \max\{y_i\} .$$

The basis functions B_{ik} are nonzero only in a domain spanned by $2 + l$ (adjacent) knots. By defining the spline design row vectors $X_j^t \in \mathbb{R}^M$, such that $X_j^t(k) = B_{jk}\left(y_j^t\right)$, we can rewrite the predictor as

$$\eta_g^t = X_1^{t-1}\boldsymbol{\beta}_{1g} + \cdots + X_G^{t-1}\boldsymbol{\beta}_{Gg} + \mu_g$$

with $\boldsymbol{\beta}_{jg} = \left\{\beta_{j1}^g, \ldots, \beta_{jM}^g\right\} \in \mathbb{R}^M$ 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 deem the *link* from j to g as off. If the link is on, then we say that j is a *parent* of g.

Stacking the bases and the coefficients into $X^t = \left\{X_1^t, \ldots, X_G^t\right\} \in \mathbb{R}^{MG}$ and $\boldsymbol{\beta}_g = \{\boldsymbol{\beta}_{1g}, \ldots, \boldsymbol{\beta}_{Gg}\} \in \mathbb{R}^{MG}$, respectively, we can express the model as $y_g^{t+1} = \mu_g + X^t\boldsymbol{\beta}_g + \varepsilon_g^t$ and after further stacking the equations over time we have,

$$\boldsymbol{y}_g = \boldsymbol{\mu}_g + \mathcal{X}\,\boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_g , \qquad g = 1, \ldots, G , \tag{4}$$

where $\boldsymbol{\mu}_g = \mu_g \boldsymbol{\iota}_T'$, with $\boldsymbol{\iota}_T$ a row vector of ones of size T and $\mathcal{X} = \left\{X^1, X^2, \ldots, X^T\right\}'$ 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 Section 3.

As it stands, (4) would require at least $M \times G$ data points per gene to be estimated. 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 non-linearities. Furthermore, in genetics applications data are typically obtained from high throughput methods, such as microarrays, providing measurements from hundreds to thousands of genes at every time point, while the actual number of time measurements, T, is generally not more than a few dozen. It is then necessary to introduce some degree of sparseness into the link structure, *i.e.* restrict the number of potential parents, in order to carry out any estimation. Also, it is well known that in any given biological process, typically only a handful of genes are responsible for gene activity, and thence this restriction arises naturally. We accommodate this idea by performing a Gibbs variable selection as in Smith and Kohn (1996). The model is augmented with the indicators γ_{jg}, such that

$$\widetilde{\boldsymbol{\beta}}_{jg} = \gamma_{jg} \times \boldsymbol{\beta}_{jg} \qquad \text{where} \qquad \gamma_{jg} = \begin{cases} 1 & \text{the link is on} \\ 0 & \text{the link is off} \end{cases} .$$

and substituting these new coefficients into the model. From a Bayesian perspective this can also be thought of as a spike-and-slab prior specification (Ishwaran and Rao, 2005; Mitchell and Beauchamp, 1988) on the spline coefficients. The practical advantage of augmenting with the indicators, γ_{jg}, is that it allows us to make inference about the network topology, now parameterised by the connectivity matrix, $\Gamma = \{\gamma_{jg}\}$.

2.1 THE PRIOR

Here we complete the specification of our Bayesian model by formalising the prior specification. Conditionally conjugate priors are used 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. It will become apparent that in the cases where the data is scarce,

this part of the prior is crucial and therefore we discuss some alternatives and perform associated sensitivity analyses. We assume that the data have been standardised (zero mean and unit variance for each gene), so we need not introduce gene specific scalings in the prior.

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

$$\pi(\boldsymbol{\lambda}) = \prod_{g=1}^{G} \frac{b_\lambda^{a_\lambda}}{\Gamma[a_\lambda]} \lambda_g^{a_\lambda - 1} \exp[-b_\lambda \lambda_g] \ . \tag{5}$$

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

$$\pi(\boldsymbol{\mu}) = \left(\frac{\tau_\mu}{2\pi}\right)^{G/2} \exp\left[-\frac{\tau_\mu}{2} \boldsymbol{\mu}' \boldsymbol{\mu}\right] \ . \tag{6}$$

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

$$\pi(\gamma_{jg} \mid \rho) = \rho^{\gamma_{jg}} (1-\rho)^{1-\gamma_{jg}} \ , \quad g, j = 1, \ldots, G \ , \tag{7}$$

$$\pi(\rho) = [\text{B}(a_\rho, b_\rho)]^{-1} \ \rho^{a_\rho - 1}(1-\rho)^{b_\rho - 1} \quad 0 < \rho < 1 \ . \tag{8}$$

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

$$\pi(\gamma_{jg} \mid \rho_j) = \rho_j^{\gamma_{jg}} (1-\rho_j)^{1-\gamma_{jg}} \ , \quad g = 1, \ldots, G \ , \tag{9}$$

$$\pi(\rho_j) = [\text{B}(a_\rho, b_\rho)]^{-1} \ \rho_j^{a_\rho - 1}(1-\rho_j)^{b_\rho - 1} \quad j = 1, \ldots, G \ . \tag{10}$$

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 the prior on the coefficients $\boldsymbol{\beta}_{jg}$ to shrink them towards the origin. One option is to penalise using a Gaussian process prior as in Speckman and Sun (2003). However, we follow Eilers and Marx (1996), specifying a second order Markov process prior

$$\pi(\boldsymbol{\beta}_{jg} \mid \tau_{jg}) = \text{N}(\boldsymbol{\beta}_{jg} \mid \mathbf{0}, \tau_{jg} K) \ . \tag{11}$$

Where τ_{jg} are the smoothness parameters addressed below. The structure of the covariance matrix, K, is constructed from the second order differences between adjacent coefficients, *i.e.*, $\beta_k = 2\beta_{k-1} - \beta_{k-2}$, omitting link identifiers for simplicity. Thus,

$$K(M, M-2) = 1 \ , \qquad K(M-2, M-1) = -4 \ , \qquad K(M, M) = 1 \ ,$$
$$K(M, M-1) = -2 \ , \qquad K(M-1, M-1) = 5 \ ;$$

and for all $i, j \in \{3, \dots M-2\}$,

$$K[i,j] = \begin{cases} 0 & |i-j| > 2 \\ -4 & |i-j| = 1 \\ 1 & |i-j| = 2 \\ 6 & |i-j| = 0 \end{cases} .$$

The two remaining coefficients are given a constant (improper) prior $\pi(\beta_1, \beta_2) \propto 1$.

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. Various alternatives have been proposed for dealing with the smoothing parameters (Belitz and Lang, 2008; Kohn *et al.*, 1991; Speckman and Sun, 2003; Wand, 1999). Following Fahrmeir and Lang (2001) and Lang and Brezger (2004), we perform a full Bayesian analysis and specify priors for the smoothing parameters.

The conditionally conjugate prior is the product of independent gamma distributions, $\text{Ga}(\cdot \mid a_\tau, b_\tau)$. 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, *e.g.* 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 favouring over-fitting when the data is only weakly informative. On the other hand, if mass is carried towards larger values of τ —thus favouring smoother curves— the gamma distribution tails off quite quickly to the left and experiences difficulties capturing non-linearities, (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 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, $\text{Ip}(\cdot \mid a_\tau, b_\tau)$:

$$\pi(\tau_{jg} \mid a_\tau, b_\tau) = \frac{a_\tau}{b_\tau} \left(\frac{\tau_{jg}}{b_\tau} \right)^{a_\tau - 1} , \qquad \tau_{jg} \leq b_\tau \, , \; a_\tau > 0 \, . \tag{12}$$

We restrict $a_\tau \geq 1$, to prevent concentration of mass near the origin. Setting $a_\tau = 1$ is tantamount to putting a uniform prior on $(0, b_\tau)$. If $a_\tau > 1$, the prior behaves as a power function and gathers mass closer to b_τ as a_τ grows, thus favouring 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 cut-off value b_τ can be interpreted as that level of τ after which the likelihood is numerically invariant, *i.e.* the fitted curve is practically linear. Determining the value of b_τ could be done, for instance, by cross-validation if feasible; otherwise, a sensitivity analysis should be carried out.

The resulting conditional posterior is amenable to a Gibbs step, requiring sampling from a truncated distribution. Details are given in Section 3.

2.1.1 POSTERIOR PROPRIETY

Fahrmeir and Kneib (2009) discuss conditions for posterior propriety using this covariance structure and different alternatives for the smoothing parameters within the context of structured additive models. We provide a result justifying an extension of this prior for GRN inference. The proof is standard and therefore omitted.

Theorem 1.
Consider the longitudinal data set $Y = \{y_g^t\}$, consisting of $g = 1, \ldots, G$ genes measured at times $t = 1, \ldots, T$, modelled as (4) and with prior given by (5)–(12). Denote by $G^\star$ the number of parents of gene g. Let

$$\mathcal{K}_g = \text{blkdiag}\left[\tau_{1g} K, \tau_{2g} K, \ldots, \tau_{G^\star g} K\right] \qquad \textit{and} \qquad \Psi_g = \mathcal{X}_g' \mathcal{X}_g + \mathcal{K}_g .$$

Where $\mathcal{X}_g$ is the design submatrix conformable to $G^\star$. Then, the posterior distribution of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G, \boldsymbol{\lambda}\}$ is proper if Ψ_g is positive definite for every g and $M \times G^\star < T$.

Given that in most of our applications we will only have a limited number of time measurements compared to the number of genes, this leads to an improper posterior if the prior was not proper, since the number of parents for any given gene only needs to exceed T/M. To construct a proper prior we supply (11) with an independent specification for the first two coefficients,

$$\pi(\beta_1, \beta_2) = \text{N}(\beta_1 \mid 0, k_1)\ \text{N}(\beta_2 \mid 0, k_2)\ . \tag{13}$$

Including these into the covariance structure we have,

$$K(1,1) = (1 + k_1/\tau)\,, \quad K(1,2) = -2\,, \quad K(1,3) = 1\,, \quad K(2,2) = (5 + k_2/\tau)\,, \quad K(2,3) = 1\ .$$

for the appropriate smoothness parameter, τ. To approximate the behaviour of the improper prior, we could let $k_1, k_2 \to 0$. In situations where the data is scarce, we do not recommend this, as it will affect the stability of the posterior (Lambert *et al.*, 2005; Sun and Speckman, 2008). In our applications, we set $k_1 = k_2 = \tau_0$.

3 IMPLEMENTATION

3.1 P-SPLINES MODEL ALGORITHM

Combining the likelihood (4) with the prior (5)–(13) and letting Θ denote all the model parameters we obtain,

$$\pi(\Theta \mid \mathcal{X}, Y) \propto \left[\prod_g \text{N}_T\left(\boldsymbol{y}_g \mid \boldsymbol{\mu}_g + \mathcal{X}\,\widetilde{\boldsymbol{\beta}}_g, \lambda_g I_T\right)\right] \left[\prod_g \pi(\boldsymbol{\beta}_g \mid \boldsymbol{\tau}_g)\pi(\boldsymbol{\tau}_g)\pi(\lambda_g)\pi(\mu_g)\pi(\boldsymbol{\gamma}_g \mid \rho_g)\pi(\rho_g)\right] ,$$

where I_T is the identity matrix of size T, $\boldsymbol{\gamma}_g = \{\gamma_{1g}, \ldots, \gamma_{Gg}\}$ and $\boldsymbol{\tau}_g = \{\tau_{1g}, \ldots, \tau_{Gg}\}$. As there is no closed form expression for the posterior numerical methods are needed. We propose a Metropolis-within-Gibbs scheme which is drafted below.

Precisions The full conditional of λ_g, $g = 1, \ldots, G$ is given by

$$\pi(\lambda_g \mid -) \propto \lambda_g^{T/2 + a_\lambda - 1} \exp\left[-\lambda_g \left(b_\lambda + \frac{1}{2}\left(\boldsymbol{y}_g - \boldsymbol{\mu}_g - \mathcal{X}\widetilde{\boldsymbol{\beta}}_g\right)'\left(\boldsymbol{y}_g - \boldsymbol{\mu}_g - \mathcal{X}\widetilde{\boldsymbol{\beta}}_g\right)\right)\right]$$

which is the kernel of a gamma distribution.

Constant term μ_g is conditionally Gaussian, with mean and precision

$$m_g = \frac{\bar{y}_g - \bar{\mathcal{X}}\tilde{\boldsymbol{\beta}}_g}{\lambda_g + \tau_\mu / T} \qquad \text{and} \qquad \tau'_\mu = \tau_\mu + T\,\lambda_g \ ,$$

respectively, where $\bar{y}_g = T^{-1}\sum_t y_g^t$ and $\bar{\mathcal{X}} = T^{-1}\sum_t \mathcal{X}$.

Connectivity If (9)–(10) is used, the full conditionals for the gene-wise connectivities, ρ_g, are obtained as

$$\pi(\rho_g \mid -) \propto \rho_g^{S_g + a_\rho - 1}(1-\rho_g)^{G + b_\rho - S_g - 1} \qquad \text{with} \qquad S_g = \sum_i^G \gamma_{gi} \quad \text{for } g = 1, \ldots, G \ ,$$

and are sampled from a $\text{Be}(\rho_g \mid S_g + a_\rho, G + b_\rho - S_g)$.

If (7)–(8) is used instead, the overall connectivity, ρ, is sampled from a $\text{Be}\big(\rho \mid S + a_\rho, G^2 + b_\rho - S\big)$, with $S = \sum_{g=1}^G S_g$.

Smoothness parameters When the corresponding link is on, the full conditional is given by

$$\pi(\tau_{jg} \mid -) \propto \tau_{jg}^{(M-2)/2 + a_\tau - 1} \exp\left[-\tau_{jg}\,\frac{1}{2}\tilde{\boldsymbol{\beta}}'_{jg} K\,\tilde{\boldsymbol{\beta}}_{jg}\right] , \qquad 0 < \tau_{jg} < b_\tau$$

and can be sampled from a truncated gamma distribution (Damien and Walker, 2001; Gentle, 2003) with parameters $\Big\{(M-2)/2 + a_\tau, \tilde{\boldsymbol{\beta}}'_{jg} K\,\tilde{\boldsymbol{\beta}}_{jg}/2\Big\}$. An observation is drawn from the prior when the link is off.

Spline Coefficients and link probabilities The update of the spline coefficients and indicator variables is performed as a block. Specifically, the update of a given indicator variable γ_{jg} and all the coefficients of the regression for gene g, $\boldsymbol{\beta}_g$, are performed simultaneously. In practice, as the regression is sparse, only a few links are actually present drastically reducing this computation. At every iteration, the individual link indicator γ_{jg} is turned on (off) if it is off(on) and the associated coefficient, $\boldsymbol{\beta}_g \in \mathbb{R}^{MG}$, for present links (on) is proposed from the joint conditional, schematically:

$$\gamma : 0 \to 1 \qquad \text{and} \qquad \boldsymbol{\beta} : \boldsymbol{\beta}^a \to \boldsymbol{\beta}^b$$

with acceptance probability

$$\alpha = \min\left\{\frac{\pi\big(\tilde{\boldsymbol{\beta}}^b\big)}{\pi\big(\tilde{\boldsymbol{\beta}}^a\big)}\frac{q(\boldsymbol{\beta}^a \mid \gamma^a)q(\gamma^a)}{q(\boldsymbol{\beta}^b \mid \gamma^b)q(\gamma^b)},\ 1\right\} \ ,$$

where the subscripts have been omitted for clarity. γ is proposed symmetrically, thus $q(\gamma^a)/q(\gamma^b) = 1$. For $q(\boldsymbol{\beta} \mid \gamma)$ we use the proposal

$$q(\boldsymbol{\beta} \mid -) = N(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$$

with

$$\Sigma_{\boldsymbol{\beta}} = \big[\lambda_g \mathcal{X}'_g \mathcal{X}_g + \Upsilon_g\big] \qquad \text{and} \qquad \mu_{\boldsymbol{\beta}} = \lambda_g(\boldsymbol{y}_g - \boldsymbol{\mu}_g)\mathcal{X}_g\,\Sigma_{\boldsymbol{\beta}}^{-1} \ ,$$

where Υ_g is the block diagonal penalty (precision) matrix, calculated by multiplying each block in $\mathcal{K}_g$ times the corresponding τ_{jg}. Note that, as only the coefficients with non zero indicator variable are updated, $\mathcal{X}_g$,

$\boldsymbol{y}_g$ and Υ_g are adjusted to only include the appropriate elements. Substituting this in the Hastings ratio gives

$$\frac{\rho}{1-\rho}\,\tau_0\tau_{jg}^{(M-2)/2}\frac{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^{b}\Sigma_{\boldsymbol{\beta}}^{-1b}\mu_{\boldsymbol{\beta}}^{b}\right]\left|\Sigma_{\boldsymbol{\beta}}^{b}\right|^{1/2}}{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^{a}\Sigma_{\boldsymbol{\beta}}^{-1a}\mu_{\boldsymbol{\beta}}^{a}\right]\left|\Sigma_{\boldsymbol{\beta}}^{a}\right|^{1/2}}\,.$$

The opposite move (switching an indicator variable off) can be performed using the reciprocal of the ratio above.

In order to enforce the identifiability restriction, at each step we calculate $\bar{m}_g = \boldsymbol{\iota}_T \times \left[\mathcal{X}\ \tilde{\boldsymbol{\beta}}_g\right]$, for every gene, subtract it from the splines and add it to the constant term, μ_g.

Our sampler exploits the conditional independence structure of the model. We constructed a parallel scheme where the calculation for each parent is assigned to a CPU-node, these communicating only when the overall connectivity is updated and for sample recording. Gains in computation times can potentially be up to n-fold, with n the number of CPU-nodes used.

3.2 A LINEAR MODEL

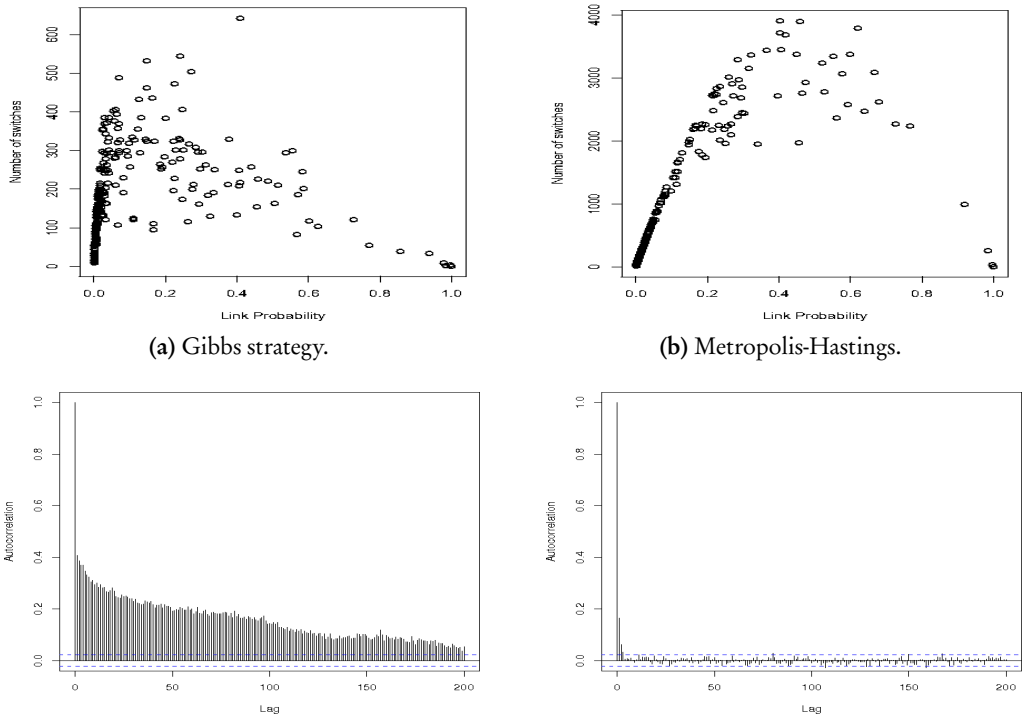
In order to compare the network retrieval power of the splines model, we constructed a fully parametric, linear AR(1) model

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G}\beta_{jg}\,y_j^t + \varepsilon_g^t\,, \tag{14}$$

with the same prior specification as above, deleting the irrelevant terms. We would like to emphasize that (14) is used for comparison purposes only. Although the basic interaction dynamics are similar to those used in *e.g.* Lèbre (2009); Opgen-Rhein and Strimmer (2006), the way network topology estimation is carried out varies significantly. Clearly, the latter models outperform our implementation in terms of speed; however, our Bayesian formulation is capable of providing measures of variability on all model parameters, including the network topology. Further, by using an identical prior structure (up to the relevant terms), we can focus on non-linear departures alone.

3.3 IMPROVING CONVERGENCE

The last move in Section 3 leads to a dramatic decrease in autocorrelation of the Markov chain, compared to a Gibbs move. Indeed, a common approach in these cases is to use a full Gibbs specification, with a full conditional Bernoulli distribution on the γ_{jg} and a full conditional Gaussian for the coefficients $\boldsymbol{\beta}_{jg}$. The latter requires the introduction of a so-called pseudo-prior which needs to be tuned to improve the mixing of the chain (Dellaportas *et al.*, 2000; Ntzoufras, 2002; O'Hara and Sillanpää, 2009). In order to assess the gains in mixing, we implemented a full Gibbs sampler for (14). When chain mixing is compared, the advantage of our MH update becomes apparent as illustrated in Figure 1, obtained by running both samplers on the non-linear synthetic data described in Section 4.1. The top panels plot the number of times that link was switched during the MCMC run against the posterior probability of the link being present. One would expect that links with probabilities around 1/2 would change more often, as in Figure 1b. However, the Gibbs strategy tends to mix more slowly, as shown in Figure 1a. Although the MH step is more computationally demanding, the benefit brought about by the improved mixing of the chain, quantified by the reduction in autocorrelation (ACF), offsets this cost easily (compare Figure 1c with Figure 1d). Given that the parameter space of the splines model is much larger than the linear one, the benefits of using this move, compared to the full Gibbs alternative are expected to be even greater.



(a) Gibbs strategy.

(b) Metropolis-Hastings.

(c) ACF λ_{15}, Gibbs strategy.

(d) ACF λ_{15}, Metropolis-Hastings.

Figure 1. *Chain mixing comparison of the Gibbs and MH strategies. Top panels plot the number of state changes of a link during the MCMC run against its posterior probability. The bottom panels show the autocorrelation function (ACF) for a single link's precision (gene 15).*

4 ILLUSTRATIONS AND APPLICATIONS

Here we illustrate the implementation and performance of our P-splines regression model with three examples. First we analyse two synthetic, discrete time data sets where the data generation mechanism and the topology of the network are known. Secondly, we examine a synthetic data set comprising discrete time measurements drawn from a continuous time ODE model of a circadian clock. For our last example, we use microarray gene expression data from the *Arabidopsis thaliana* circadian clock.

In all our applications, we include a slight modification of the structure of the network topology to that described in Section 2.1. We know from the context that each gene has a decay term, corresponding to mRNA decay. We include this information in the prior by fixing $\gamma_{gg} = 1$. As we also know that this decay is close to linear, we set the shape of the inverted Pareto prior for these smoothing parameters to thrice the value used for the rest, *i.e.* $a_{\tau_{gg}} = 3 \times a_{\tau_{ij}}, i \neq j$.

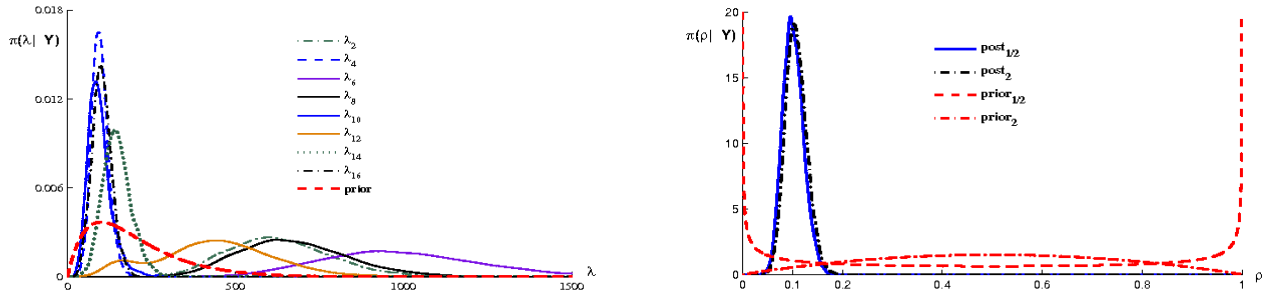
The splines and linear models were fitted using the overall and gene-wise connectivity specifications. Throughout, 13 bases were used. Prior parameters were set to $\{a_\rho, b_\rho\} = \{1/2, 1/2\}$, $\{a_\lambda, b_\lambda\} = \{2, 0.01\}$, $\tau_\mu = 1/4$, $\tau_0 = 0.25$ and $\{a_\tau, b_\tau\} = \{1.5, 10^4\}$. We ran two parallel chains of length 10^5, dropping the first 10^4 steps and then recording every tenth draw. We performed some sensitivity analyses, varying a_τ from 1 (uniform prior) up to 3, setting $a_\rho = b_\rho = 1, 2$ and using flatter versions of the prior for λ by setting $a_\lambda = 1, 0.1$, without finding noteworthy differences. Convergence was assessed by comparing both chains graphically and by formal tests using the CODA package (Plummer *et al.*, 2006).

4.1 DISCRETE TIME SYNTHETIC NETWORKS

In order to assess the network topology recovery power of our model, we produced two synthetic, first order autoregressive processes. One has only linear and the second a number of non-linear (S-shaped) relations. In the

non-linear case, all the functional relations were produced using Hill functions, except for the self-interactions which are linear. In both cases we set $G = 16$, $T = 40$, and $\rho \approx 0.1$.

The models with gene-wise and overall connectivity produced almost indistinguishable estimations for the network topology and thus we report the results for the simpler model only. In Figure 2a we plot the marginal posterior and prior distributions of the model precisions, λ_g (for a selection of the genes only, to avoid clutter). We also performed a sensitivity analysis on ρ, fixing the prior parameters $a_\rho = b_\rho = 2$. As shown in Figure 2b, the posterior was practically unaffected by this change.

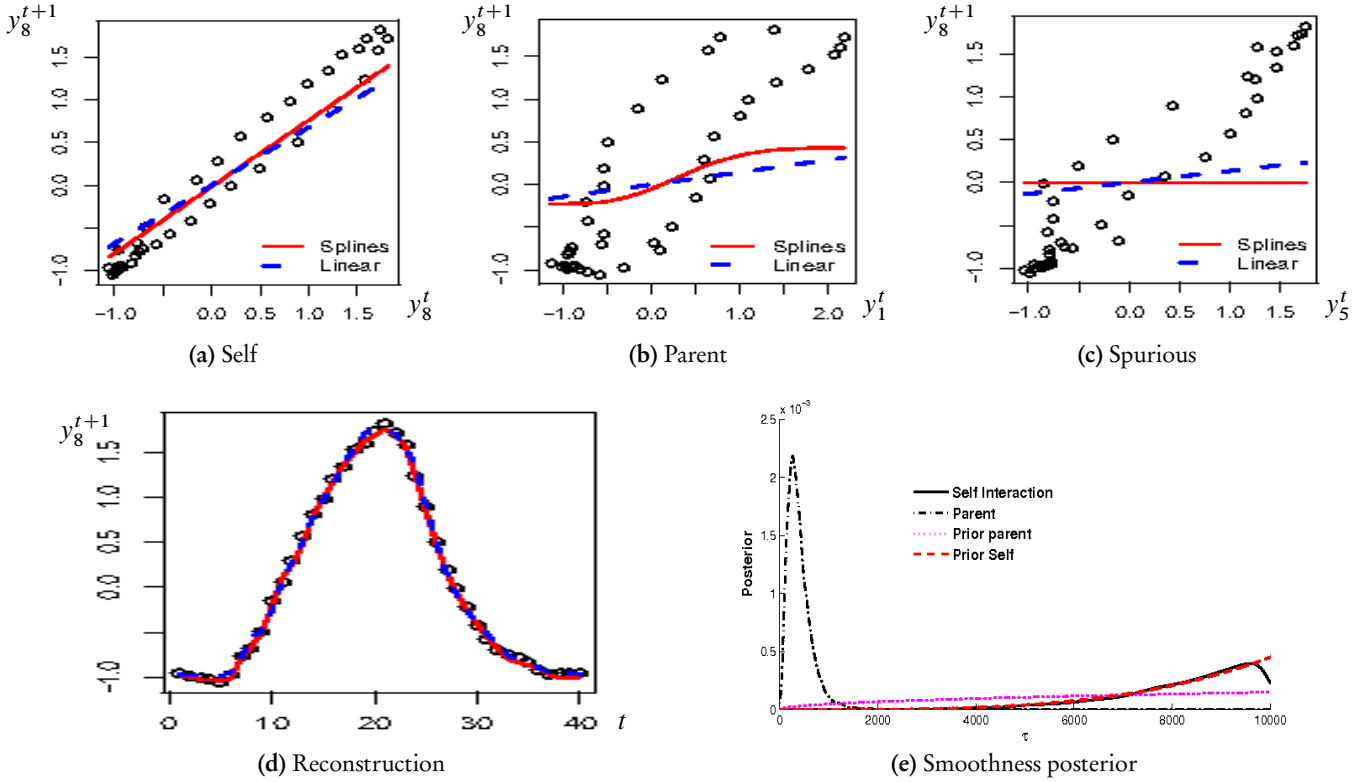


(a) Precision posteriors. **(b)** Connectivity posterior.

Figure 2. *Marginal posterior distributions calculated when fitting the splines model to the non-linear synthetic data set. (a) The posterior for a selection of the gene precisions, λ_g and the corresponding prior. (b) The posterior of the overall-connectivity, ρ, from two different priors, $a_\rho = b_\rho = 1/2$ and 2. In both panels priors are depicted by the thick (red) dashed lines.*

When the topology of the network is known, we can use the Receiver Operator Characteristic (ROC) curve to assess graphically the retrieval performance of a model (Pepe, 2000; Sing *et al.*, 2005). A more formal comparison can be carried out by calculating the area under the ROC curve (AUC): the closer the AUC to one, the better the retrieval. For the linear data set these were 0.999 for the fully parametric model and 0.998 with the splines; and when fitting the non-linear data set we obtained 0.728 and 0.912, respectively. An alternative measure of fit is the so-called mean cross entropy (MxE), calculated as the Kullback-Leibler divergence from the known network topology to that estimated by $\widehat{\Gamma}$, averaged over all possible links. MxE is bounded from below at zero, when the predicted topology is identical to the real one. Its value for a network topology predicted totally at random, *i.e.* $\widehat{\gamma}_{ij} = 1/2$, for all $i, j = 1, \ldots, G$, is $-\log 1/2 \approx 0.7$. 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.

To further understand the differences between the inferred networks under either model, we plot in Figure 3 the partial and full reconstructions for gene 8's trace in the non-linear data set, along with the posterior of the corresponding smoothing parameter. Both models provide similar predictions, as illustrated by the full reconstructions which are practically undistinguishable (Figure 3d). However, the way this fit is achieved varies significantly. As expected, both models have a very similar fit for the self-regulation (Figure 3a). As the self-interaction is linear, the splines model fits it by allocating most of the posterior mass of the corresponding smoothness parameter towards high values, depicted as the solid line in Figure 3e. Gene 8 has one parent with a non-linear interaction and the splines model is capable of reproducing the Hill functional relationship quite precisely (Figure 3b), by allocating almost all posterior mass towards small values of the corresponding smoothness parameter, illustrated in Figure 3e with a dot-dashed line. Obviously, the linear model cannot accommodate such behaviour and may need to include spurious parents in order to compensate for the lack of fit, as in this case, illustrated in Figure 3c by the dashed line with a slight positive gradient. In contrast, the splines model does not predict gene 5 as a parent (solid line in Figure 3c). Notice the mass allocation of the self-regulation link (solid) in Figure 3e: it is basically drawn from the prior (dashed), illustrating that our specification is adequate for linear relations to be reproduced accurately.



(a) Self (b) Parent (c) Spurious

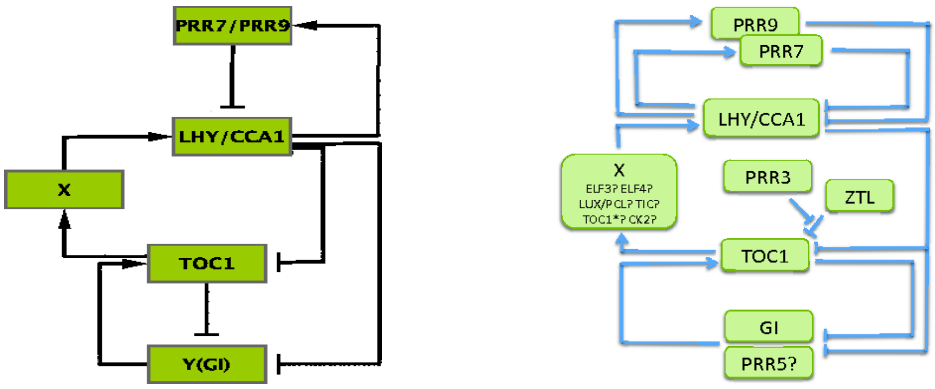
(d) Reconstruction (e) Smoothness posterior

Figure 3. *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 three 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 non-linear parent (dot-dashed, prior dotted)*

4.2 BIOLOGICAL GRN: THE PLANT CIRCADIAN CLOCK

In the following sections we focus on a partially known GRN, specifically the plant *Arabidopsis thaliana* circadian clock. Most organisms have the ability to track time even in the absence of external input (*e.g.* light). This ability allows the organism to anticipate and prepare for future events, thus enabling it to optimise the interaction with the environment. In some cases, such as in *Arabidopsis*, diurnal period tracking is achieved via a regulatory network that oscillates with a period close to 24 hrs. This period then propagates through one or more of the core genes of the clock to target genes responsible for other biological processes (reviews can be found in Harmer, 2009; Más, 2008; McClung, 2006). The circadian clock is of central importance and has been extensively studied both experimentally and through mathematical modelling. It has recently been reported to regulate up to 90% of the *Arabidopsis* genome under some environmental conditions (Michael *et al.*, 2008). While the circadian clock is able to maintain oscillations without the need of light, it is known that the period is modified by light exposure, allowing it to adapt to shorter and longer daylight hours.

Locke *et al.* (2006) developed an ODE model of the clock, involving around 80 parameters, which we use below for generating synthetic observations. The current working biological model (McClung, 2008) involves almost twice as many genes as the mathematical one. Both models include nodes X and Y, representing genes which are thought to be involved in the circadian clock, but whose identity remains unknown. These networks are schematically outlined in Figure 4. As usual, genes are represented by nodes and regulation by directed edges ending either in arrows (positive regulation) or bars (inhibition).



(a) ODE model **(b)** Functional model

Figure 4. *Models of the Circadian Clock in* Arabidopsis thaliana*. (a) The ODE model of Locke* et al. *(2006). (b) The current working model of the clock (redrawn from McClung, 2009). Nodes represent genes known (or suspected) to be part of the clock. Positive regulation is represented by an arrow and suppression by an edge with a bar end.*

4.2.1 DIFFERENTIAL EQUATION DATA

Using chemical reaction kinetics to represent the interaction between an activating protein and its target gene, the dynamics of gene expression (mRNA) can be represented by an ordinary differential equation. The ODE describes the rate of change of the concentration of mRNA as a function of the concentration of an activating protein, Y_{protein}, and its own concentration, X_{mRNA}, as:

$$\frac{\partial X_{\text{mRNA}}}{\partial t} = \beta \frac{Y^a_{\text{protein}}}{\theta + Y^a_{\text{protein}}} - \alpha \frac{X_{\text{mRNA}}}{\kappa + X_{\text{mRNA}}} ,$$

where the parameters $\{a, \theta, \alpha, \kappa\}$ govern the dynamics of the process. For a more detailed description see for example Alon (2007).

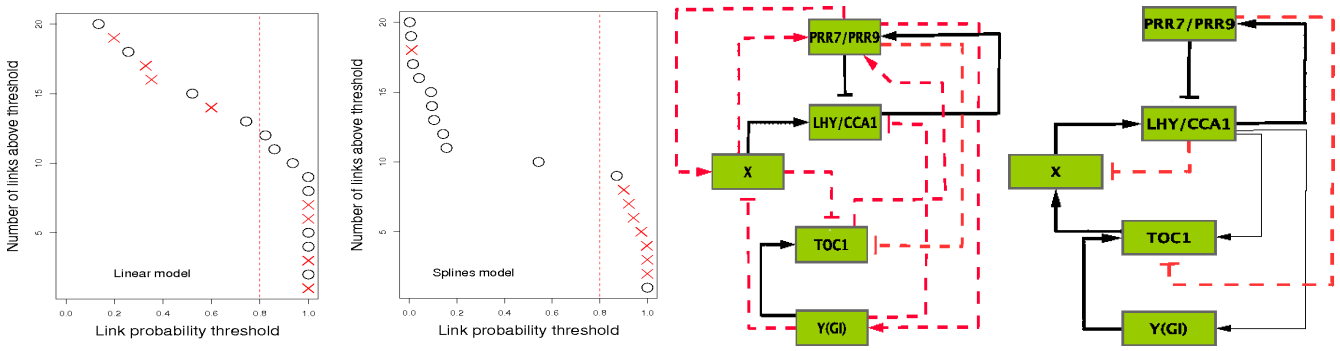
Using coupled ODEs, Locke *et al.* (2006) developed a mathematical representation of the *A. thaliana*'s circadian clock that is able to explain the period of the clock under various experimental conditions (see also Zeilinger *et al.*, 2006). The model accounts not only for the production of mRNA given the concentration of the regulatory proteins, but also for the translation of mRNA into protein and its transport in and out of the cell nucleus. The network is shown schematically in Figure 4a and consists of 5 genes and 8 connections. Two of the nodes (PRR7/PRR9 and LHY/CCA1) represent pairs of genes that have the same dynamics and play the same role in the network. Y is an unknown gene originally introduced to explain the behaviour of the network under certain experiments, and currently thought to be GI (or possibly PRR5). X was introduced to account for the fact that TOC1 has been shown experimentally to affect LHY/CCA1, though not through direct interaction.

We generated data from this model using COPASI (Hoops *et al.*, 2006) fixing the light source to be permanently on. The data was then subsampled, logged and standardised. The resulting data set has 50 time points with a time spacing of 1 hr and is depicted in the left panel of Figure 6. Given that simultaneous measurement of multiple proteins is currently very hard, usually only mRNA is available for network inference. For this reason, although the model outputs protein concentration and location, we used only the mRNA data. We expect the data generated from this model to be a reasonable reflection of experimental data, not only because it is a continuous time model with nonlinear interactions, but it also reflects realistic sampling regimes and interaction intensities.

We present the results obtained using the parent-wise connectivity structure (9)–(10). In order to interpret the output, rather than examining the ROC curves we analyse 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 5a, and when using the splines model, Figure 5b. We use a cross (circle) for a

correctly (incorrectly) predicted link; for instance, the predicted network with the splines model using a threshold of 0.85 would have 9 links (circles and crossed with link probability above 0.85 in Figure 5a), seven 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 two populations: a low probability (below 0.2) and a high probability (above 0.8) group. This contrasts with the linear model Figure 5a 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 the two rightmost panels of Figure 5. The inferred network for splines (Figure 5d) contains all links from the correct network (see Figure 4a), except for the TOC1-Y link. There are two spurious links (LHY/CCA1-X and PRR7/PRR9-TOC1) and two links with incorrect signs (LHY/CCA1-TOC1 and LHY/CCA1-Y). In the cases of links with incorrectly predicted signs, there is either a missing parent (TOC1-Y) or an additional incorrectly predicted parent (PRR7/PRR9-TOC1). While having a comparable amount of correct links, the inferred network for the linear model adds a large number of spurious links to the network. Again, we found cases where the splines model correctly predicts a single parent using a non-linear interaction, whereas the linear model predicts that link but adds extra spurious links to adjust the fit (not shown).

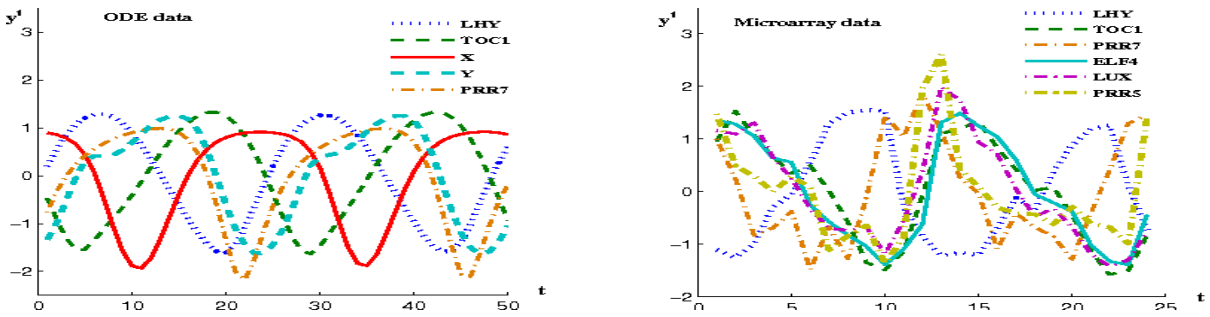


(a) LM Links (b) SM Links (c) LM Network (d) SM Network

Figure 5. *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 linear model (LM) and (d) when using the splines model (SM). Solid lines represent correct predictions, dashed lines incorrect predictions and thin lines correct predictions, but with the wrong sign.*

4.2.2 EXPERIMENTAL DATA

The data is obtained from a gene expression time series for *Arabidopsis* leaves generated using microarrays. Whole leaves were harvested every 2 hrs for 48 hrs, with four biological replicates at each time point. To reduce variability, the same leaf (the 7th leaf to emerge) was harvested for each sample. This means that the same plant was not monitored over the entire time series but leaves of 96 distinct plants grown in identical conditions were sampled (four at each of the 24 time points). Full genome expression profiles of these leaves were generated using CATMA arrays (Sclep *et al.*, 2007). Data processing and normalisation of the time series was carried out using a pipeline based on the R package MAANOVA (Wu *et al.*, 2003). Given that the replicates showed some outliers we use the median of the four bioreplicates as the observed series. We use the same genes as those that appear in the ODE model, leaving some freedom to choose which genes to use for the ambiguous nodes. For the two genes that represent pairs, we selected those that showed least variability across replicates (LHY, PRR7). To represent X we chose the genes amongst the candidates in Figure 4b that showed the strongest signal-to-noise ratio: LUX and ELF4. For Y we chose PRR5 as it had a stronger signal than GI. The data set used can be seen in Figure 6b.

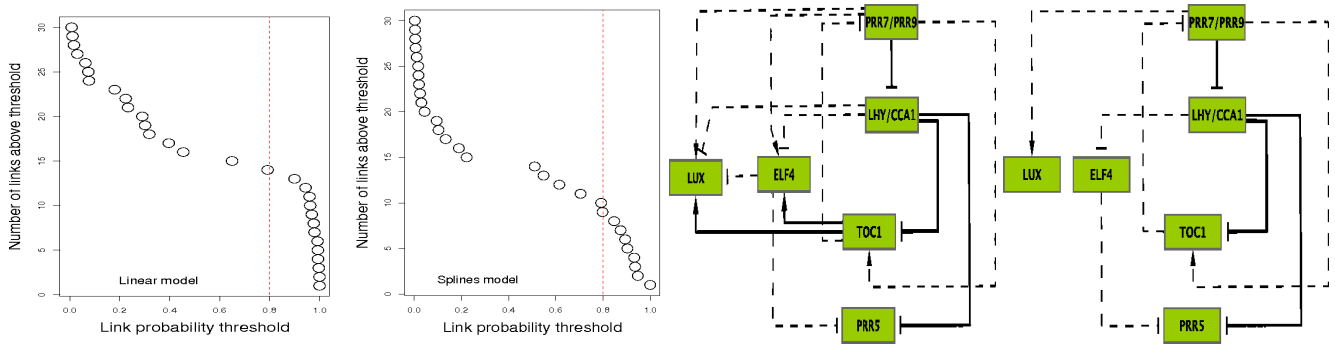


(a) ODE data (b) Microarray data

Figure 6. *Time traces of the ODE model and experimental data for the Circadian Clock in* Arabidopsis thaliana. *(a) Data simulated with the ODE clock of Locke* et al. *(2006). (b) Gene expression profiles of* Arabidopsis *leaves. Data sets are standardised.*

We analyse the output using the parent-wise connectivity structure (9)–(10). The separation between link probabilities is no longer as pronounced as in the synthetic data (see Figure 7a and 7b). This may be due to the combination of a high level of noise and fewer time points. The networks inferred for both models, using a threshold of 0.8, are shown in Figure 7c and 7d. Both inferred networks are very similar, in fact all links predicted by the splines model appear in the linear model reconstruction. However, the linear model predicts an additional two parents for ELF4 and an additional three parents for LUX. Amongst those additional links are TOC1-ELF4 and TOC1-LUX, which while we have marked as correct on the plot (for consistency with Figure 4b), 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.

Noteworthy is that for this data set only very mild nonlinearities are found, while in the previous examples the splines model found strong nonlinear relations. This could be due to the large time spacing (we have found from our simulations that hourly sampling is nearly optimal) along with the high level of noise. The mild nonlinearities help to explain why both models predict very similar networks.



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

Figure 7. *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 linear model (LM) and (b) with the splines model (SM). (c) The network obtained with a threshold of 0.8 using the linear model (LM) and (d) when using the splines model (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).*

One way to assess the accuracy of the splines inferred network topology is to restrict our analysis to the most extensively studied genes: PRR7, LHY, TOC1 and (less so) PRR5. We can reasonably assume that the connections between these links are known. As can be seen in Figure 4b, there are 6 connections amongst these genes. The

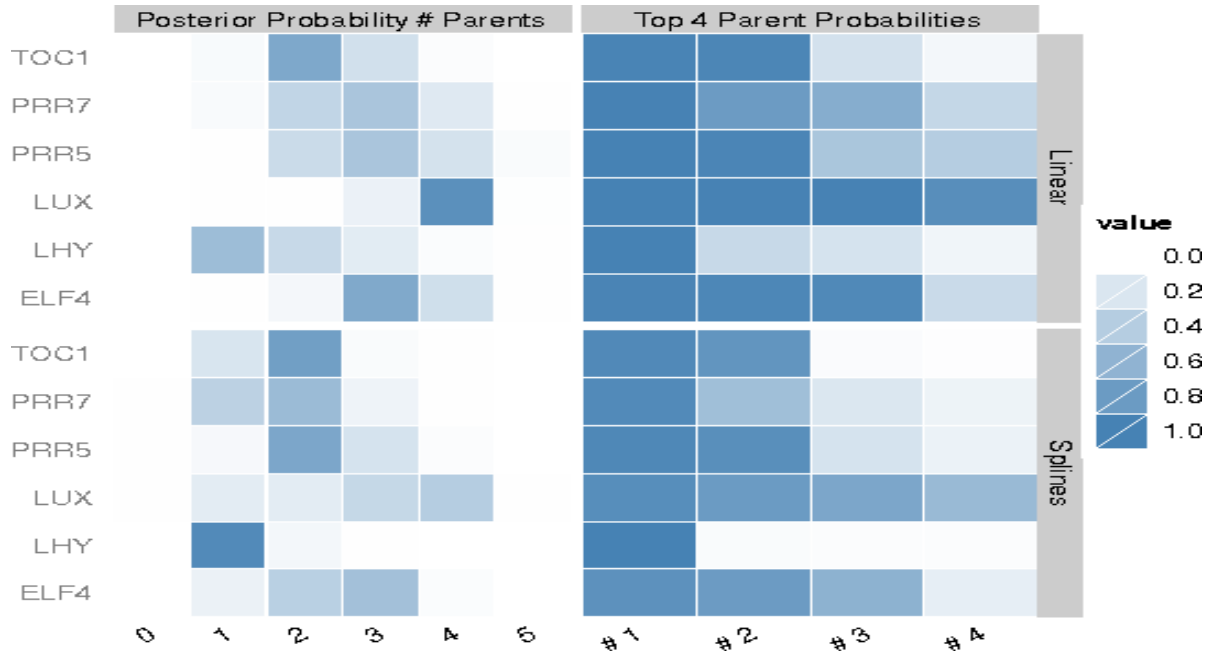


Figure 8. *Uncertainty in* Arabidopsis *circadian clock gene network reconstruction. On the left, a heatmap of the distribution of the number of parents for each gene in the clock, estimated using microarray data with the linear (top) and splines (bottom) models. On the right, a heatmap with the marginal probabilities of the top four potential parents.*

splines model predicts 5 connections, of which 3 are correctly predicted. Both of the incorrect predictions appear for genes that are missing a link in the inferred networks, indicating that the model has found the wrong parent rather than overfitting with more parents than necessary. Neither candidate gene for X (LUX and $ELF4$) regulates LHY, which would be evidence supporting the hypothesis that one of them is the unknown gene. On the other hand, these genes were proposed to be X as they are known to be involved in the clock and have some effect on the system, so the predictions can serve as a working hypotheses for determining the role they play within the network.

For network reconstruction, we have used a threshold in the examples above, fixed rather arbitrarily. 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 8 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 four 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 one or two parents with a mild probability (and three with a very slight probability), while one of the potential parents has a high marginal probability, there are two more with intermediate marginal probabilities, suggesting an ambiguity in the identity of the second potential parent. Overall, when comparing the linear with the splines predictions, there is shift to the left of the distribution for the number of parents from the splines model, strengthening the evidence of overfitting with the linear model. Likewise, marginal link probabilities from the splines model seem to be higher within a smaller number of potential parents, thus suggesting a decrease in the uncertainty in topology retrieval, compared to the linear model.

5 Discussion

In this paper we have presented a fully Bayesian implementation of P-spline based inference of a dynamic Bayesian network within a sparse connectivity context. Our application is the inference of a GRN from longitudinal data,

for instance from microarray time series data. Despite being capable of measuring up to tens of thousands of genes simultaneously, microarray time series are typically shorter than 20 time points, a consequence of their high cost and the experimental difficulties in obtaining high time resolution. This introduces significant problems for analysis and modelling, 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, by penalisation implicit in model complexity, limits the connectivity of the GRN. Within this context we are able to increase regression model complexity, providing methods for exploring whether nonlinear regulatory mechanisms are present in time series data. Our approach is to exploit the flexibility of splines to model arbitrary nonlinear relationships within a first order autoregressive process. We developed a fully Bayesian approach implemented in a (parallelised) MCMC algorithm, and provide appropriate priors such that posterior propriety holds. We found that nonlinear interactions could be successfully identified on simulated data (both discrete time and ODE models), the corresponding inferred GRN under a linear model typically acquiring additional parents, (Figure 3), 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 penalisation than a parent with a linear functional dependence model given its 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 penalisation, for instance through an arbitrary weighting $\exp(-n)$ for n parents, as in Kim *et al.* (2004).

Use of splines in inference requires handling of their functional flexibility. Firstly, there are two fundamental parameters that define the spline basis —the degree of the spline and the number of knots. The choice of a third degree spline is relatively natural and given that we have second order penalties on the coefficients, in the limit as $\tau \to \infty$, a straight line is obtained. Selecting the number of knots is not as straightforward, but depends on the number of time points and the relationship of the time resolution to the time scales in the dynamics. As illustrated by Theorem 1, a large number of knots can affect the stability of the sampler by rendering posteriors close to impropriety; on the other hand, a small number of knots can seriously affect the flexibility of the spline. 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. Secondly, the functional flexibility within the spline basis needs to be controlled; specifically spline degrees of freedom must be constrained since otherwise an interpolating spline will be fitted. We use a second order penalisation method that effectively controls the spline curvature. This entails choice of the value of the smoothing parameter τ; previous authors have optimised and fixed it before estimating the regression. We propose a fully Bayesian approach, inferring it concomitantly with the regression. Choice of the prior for τ is vital since it must allow for cases of both linearity and the levels of nonlinearity implied by the data. The commonly used conditionally conjugate gamma prior specification is not able to meet both these requirements; we propose the use of an inverted Pareto, which only requires truncating the conditional posterior obtained in the conjugate case. The sole remaining problem is then to fix the value of the cut-off value of the inverted Pareto prior, which can be interpreted as that value after which the fit of the spline is linear. We used a sensitivity analysis to confirm our prior is sufficiently weak, whilst the presence of predicted linear regulatory links in the inferred network was confirmation that our cut-off was in fact sufficiently high. Network connectivity and spline smoothness were regression/gene specific; this allowed for heterogeneity in the nonlinearity and parent number across the network.

In biological applications it is known that many regulatory mechanisms are nonlinear (Alon, 2007). However, many GRN inference methods use linear models of regulation or implement a restrictive type of nonlinearity, *e.g.* via data discretisation and logic gates (Bulashevska and Eils, 2005). Our semi-parametric model thus enables the key question of whether nonlinearity is an important factor in the GRN to be addressed. Available gene expression longitudinal data for network inference typically comprises 10–100 genes with 10–40 time points, with or without

replicates. By standardising the data, the issues facing prior choice are reduced and a generic set of spline parameters will probably work in most cases. Specifically cubic splines with 10–15 knots and an inverted Pareto prior with shape parameter in (1,2) and a cut-off in $(10^3, 10^4)$. A sensitivity analysis in the inverted Pareto prior parameter is essential, possibly performed on a subset of the data for increased speed, whilst sensitivity to the number of knots is recommended.

Our algorithm took 2.7 hrs to run 10^5 iterations with the nonlinear synthetic data ($G = 16$, $T = 40$, $\rho \approx 0.1$) and scaling is likely to be quadratic in the number of genes and number of potential parents, and linear in the number of time points; for instance, fitting a microarray gene expression data set —not shown— with $G = 30$ and $T = 37$ ($\hat{\rho} \approx 0.15$) took 20 hrs for the same run length. Thus, for data sets with a large number of genes a parallel algorithm is available which reduces computation time approximately linearly in the number of cpu-nodes; for instance, using 31 cpus the former data set took 28.6 mins and the later 3 hrs.

Our P-splines model can be extended and modified for specific purposes. Firstly, we model only direct, first-order filiation, *i.e.* single-parent–child relations. It is well known that for some processes two or more genes need to bind together in order to affect a set of target genes. One can extend the present model for allowing higher degree interactions, *e.g.* 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. Secondly, spline coefficient shrinkage can be performed in a number of ways. We argue that a second order process is natural, given that it constrains function curvature, and thus incorporates linear relations as a special case. However, additional constraints can be used, including a further term on the prior for the spline coefficients, $\mathrm{N}(\boldsymbol{\beta} \mid \mathbf{0}, \omega H)$, with H derived from the first order differences of adjacent coefficients. This effectively penalises large first order differences and favours 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 *et al.*, 1993). Finally, the splines model can be utilised to infer the functional form of the regulation, and coupled with current biological knowledge, serve as a basis of a tailor-made parametric model.

Acknowledgements

ER Morrissey was supported by the Warwick Systems Biology Doctoral Training Centre. MA Juárez was funded by BBSRC grant BB/FF003498/1. The experimental data was provided by KJ Denby through the PRESTA Project, grant number BB/F005806/1.

References

Ahn, S., Richard, W.T., Park, C.C., Lin, A., Leahy, R.M., Lange, K. and Smith, D.J. (2009) Directed mammalian gene regulatory networks using expression and comparative genomic hybridization microarray data from radiation hybrids. *PLoS Computational Biology*, 5, e1000407. doi:10.1371/journal.pcbi.1000407.

Alon, U. (2007) *An Introduction to Systems Biology: design principles of biological circuits*. Boca Raton: Chapman & Hall/CRC.

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

Bang-Jensen, J. and Gutin, G. (2009) *Digraphs : theory, algorithms, and applications*. London: Springer, second edn.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A. and di Bernardo, D. (2007) How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3, 78.

Belitz, C. and Lang, S. (2008) Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis*, 53, 61–81.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E. and Collins, J.J. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology*, 23, 377–383. doi:10.1038/nbt1075.

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

Biller, C. (2000) Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9, 122–140.

Bonneau, R., Reiss, D., Shannon, P., Facciotti, M., Hood, L., Baliga, N. and Thorsson, V. (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, 7, R36+. doi:10.1186/gb-_2006-_7-_5-_r36.

Brezger, A. and Lang, S. (2008) Simultaneous probability statements for Bayesian P-splines. *Statistical Modelling*, 8, 141–168.

Bulashevska, S. and Eils, R. (2005) Inferring genetic regulatory logic from expression data. *Bioinformatics*, 21, 2706–2713. doi:10.1093/bioinformatics/bti388.

Cao, J. and Zhao, H. (2008) Estimating dynamic models for gene regulation networks. *Bioinformatics*, 24, 1619–1624. doi:10.1093/bioinformatics/btn246.

Chen, L. and Zheng, S. (2009) Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biology*, 10, R3. doi:10.1186/gb-_2009-_10-_1-_r3.

Courcelle, J., Khodursky, A., Peter, B., Brown, P.O. and Hanawalt, P.C. (2001) Comparative gene expression profiles following UV exposure in Wild-Type and SOS-Deficient *Escherichia coli*. *Genetics*, 158, 41–64.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L. and Spiegelhalter, D.J. (1999) *Probabilistic Networks and Expert Systems*. London: Springer-Verlag.

Damien, P. and Walker, S.G. (2001) Sampling truncated Normal, Beta and Gamma densities. *Journal of Computational and Graphical Statistics*, 10, 206–215.

Dellaportas, P., Foster, J.J. and Ntzoufras, I. (2000) Bayesian variable selection using the Gibbs sampling. *Generalized linear models: a Bayesian perspective* (D.K. Dey, S.K. Ghosh and B.K. Mallick, eds.). New York: Marcel Dekker, 273–286.

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.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998) Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society B*, 60, 333–350.

Dierckx, P. (1993) *Curve and Surface Fitting with Splines*. Oxford: Clarendon Press.

DiMatteo, I., Genovese, C.R. and Kass, R.E. (2001) Bayesian curve-fitting with free-knot splines. *Biometrika*, **99**, 1055–1071.

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

Fahrmeir, L. and Kneib, T. (2009) Property of posteriors in structured additive regression models: Theory and empirical evidence. *Journal of Statistical Planning and Inference*, 139, 843–859.

Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalised additive mixed models based on Markov random field priors. *Applied Statistics*, **50**, 201–220.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. and Gardner, T.S. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5, e8+. doi:10.1371/journal.pbio.0050008.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and its applications*. London: Chapman & Hall.

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. *The Annals of Statistics*, **19**, 1–141. (with discussion).

Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805. doi: 10.1126/science.1094068.

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.

Gardner, T.S., di Bernardo, D., Lorenz, D. and Collins, J.J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105. doi:10.1126/science.1081900.

Gardner, T.S., Cantor, C.R. and Collins, J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**, 339–342.

Gentle, J.E. (2003) *Random number generation and Monte Carlo methods*. New York: Springer-Verlag, second edn.

Gilchrist, M., Thorsson, V., Li, B., Rust, A.G., Korb, M., Roach, J.C., Kennedy, K., Hai, T., Bolouri, H., Aderem, A. and Roach, J.C. (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature*, **441**, 173–178. doi:10.1038/nature04768.

Green, P.J. and Silverman, B.W. (1994) *Nonparametric regression and generalized linear models: A roughness penalty approach*. No. 58 in Monographs on Statistics and Applied Probability. Boca Raton: CRC.

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.

Hache, H., Lehrach, H. and Herwing, R. (2009) Reverse engineering of gene regulatory networks: A comparative study. *EURASIP Journal of Bioinformatics and Systems Biology*, **2009**, 1–12. doi:10.1155/2009/617281.

Harmer, S.L. (2009) The circadian system in higher plants. *Annual Review of Plant Biology*, **60**, 357–377.

Hartemik, A.J. (2005) Reverse engineering gene regulatory networks. *Nature Biotechnology*, 23, 554–555.

Hongqiang, L., Lu, L., Manly, K.F., Chesler, E.J., Bao, L., Wang, J., Zhou, M., Williams, R.W. and Cui, Y. (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Human Molecular Genetics*, 14, 1119–1125. doi:10.1093/hmg/ddi124.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. and Kummer, U. (2006) COPASI — a COmplex PAthway SImulator. *Bioinformatics*, 22, 3067–3074.

Imoto, S., Goto, T. and Miyano, S. (2002) Estimation of gene networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pacific Symposium on Biocomputing*, 7, 175–186.

Imoto, S. and Konishi, S. (2003) Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics*, 55, 671–687.

Ishwaran, H. and Rao, J.S. (2005) Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33, 730–773.

Jaffrezic, F. and Tosser-Klopp, G. (2009) Gene network reconstruction from microarray data. *BMC Proceedings*, 3, S12. doi:10.1186/1753-_6561-_3-_S4-_S12.

Jensen, F.V. and Nielsen, T.D. (2007) *Bayesian networks and decision graphs*. New York: Springer, second edn.

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. (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in Bioinformatics*, 4, 228–235.

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.

Kohanski, M.A., Dwyer, D.J., Wierzbowski, J., Cottarel, G. and Collins, J.J. (2008) Mistranslation of membrane proteins and two-component system activation trigger antibiotic-mediated cell death. *Cell*, 135, 679–690. doi:10.1016/j.cell.2008.09.038.

Kohn, R., Ansley, C.F. and Tharm, D. (1991) The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, 86, 1042–1050.

Lambert, P.C., Sutton, A.J., Burton, P.R., Abrams, K.R. and Jones, D.R. (2005) How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24, 2401–2428.

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

Lauritzen, S. (1996) *Graphical Models*. Oxford: University Press.

Lèbre, S. (2009) Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*, 8, Article 9. doi:10.2202/1544-_6115.1294.

Locke, J.C.W., Kozma-Bognár, 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. doi:10.1038/msb4100102.

Marx, B.D. and Eilers, P.H.C. (1998) Direct generalised additive modelling with penalised likelihood. *Computational Statistics & Data Analysis*, **28**, 193–209.

Más, P. (2008) Circadian clock function in *Arabidopsis thaliana*: time beyond transcription. *Trends in Cell Biology*, **18**, 273–281.

McClung, C.R. (2006) Plant circadian rhythms. *The plant Cell*, **18**, 792–803.

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

Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M., Givan, S.A., Yanovsky, M., Hong, F., Kay, S.A. and Chory, J. (2008) Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genetics*, **4**, e14. doi:10.1371/journal.pgen.0040014.

Mitchell, T.J. and Beauchamp, J.J. (1988) Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83**, 1023–1036. (with discussion).

Murphy, K. and Mian, S. (1999) Modelling gene expression data using dynamic Bayesian networks. *Tech. rep.*, Computer Science Division, University of California, Berkeley.

Ntzoufras, I. (2002) Gibbs variable selection using BUGS. *Journal of Statistical Software*, **7**, 1–19.

O'Hara, R.B. and Sillanpää, M.J. (2009) A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, **4**, 85–118.

Opgen-Rhein, R. and Strimmer, K. (2006) Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, **4**, 53–65.

Opgen-Rhein, R. and Strimmer, K. (2007) Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, **8**, S3. doi:10.1186/1471-_2105-_8-_S2-_S3.

Pepe, M.S. (2000) Receiver operating characteristic methodology. *Journal of the American Statistical Association*, **95**, 308–311.

Perrin, B., Ralaviola, L., Mazurie, A., Bottani, S., Mallet, J. and d'Alché Buc, F. (2003) Gene network inference using dynamic Bayesian networks. *Bioinformatics*, **19**, ii138–ii148.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11. URL `http://CRAN.R-_project.org/doc/Rnews/Rnews_2006-_1.pdf`.

Ronen, M., Rosenberg, R., Shraiman, B.I. and Alon, U. (2002) Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **99**, 10555–10560. doi:10.1073/pnas.152046799.

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

Schäfer, J. and Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **6**, 754–764. doi:10.1093/bioinformatics/bti062.

Sclep, G., Allemeersch, J., Liechti, R., DeMeyer, B., Beynon, J., Bhalerao, R., Moreau, Y., Nietfeld, W., Renou, J.P., Reymond, P., Kuiper, M.T.R. and Hilson, P. (2007) CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of *Arabidopsis* genes. *BMC Bioinformatics*, **8**, 400. doi:10.1186/1471-_2105-_8-_400.

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.

Shimony, S.E. (1994) Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, **68**, 399–410.

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941. doi:10.1093/bioinformatics/bti623.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–343.

Speckman, P. and Sun, D. (2003) Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, **90**, 289–302.

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.

Toyoshiba, H., Yamanaka, T., Sone, H., Parham, F.M., Walker, N.J., Martínez, J. and Portier, C.J. (2004) Gene interaction network suggests Dioxin induces a significant linkage between Aryl Hydrocarbon receptor and Retionic Acid receptor Beta. *Environmental Health Perspectives*, **112**, 1217–1224.

Wahba, G. (1990) *Spline Models for Observational Data*, *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 59. Philadelphia: SIAM.

Wand, M.P. (1999) On the optimal amount of smoothing in penalised spline regression. *Biometrika*, **86**, 936–940.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Wu, H. and Zhang, J.T. (2006) *Nonparametric Regression Methods for Longitudinal Analysis*. New Jersey: Wiley.

Wu, H., Kerr, M.K., Cui, X. and Churchill, G.A. (2003) MAANOVA: A software package for the analysis of spotted cDNA microarray experiments. *The Analysis of Gene Expression Data: Methods and Software* (G. Parmigiani, E.S. Garrett, R.A. Irizarry and S.L. Zeger, eds.). New York: Springer, 313–341.

Yin, X., Struik, P.C. and Kropff, M.J. (2004) Role of crop physiology in predicting gene-to-phenotype relationships. *Trends in Plant Science*, **9**, 426 – 432. doi:10.1016/j.tplants.2004.07.007.

Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J. and Jarvis, E.D. (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.

Zeilinger, M.N., Farré, E.M., Taylor, S.R., Kay, S.A. and Doyle, F.J. (2006) A novel computational model of the circadian clock in *Arabidopsis* that incorporates PRR7 and PRR9. *Molecular Systems Biology*, **2**, 58. doi:10.1038/msb4100101.

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.