On reverse engineering of gene interaction networks using time course data with repeated measurements

SUPPLEMENTARY MATERIAL

Morrissey, E.R., Juárez, M.A.
Systems Biology Centre
University of Warwick, Coventry CV4 7AL, UK

K. J. Denby
Warwick HRI, Systems Biology Centre
University of Warwick, Coventry CV4 7AL, UK

Burroughs, N.J.
Systems Biology Centre,
University of Warwick, Coventry CV4 7AL, UK

These supplementary materials are presented in order of appearance within the main paper. For completeness, we briefly recall the model specification.

1S THE MODEL

Denote by y_g^t the expression level of gene $g = 1, \ldots, G$, measured at time $t = 1, \ldots, T$. We model the interaction network as a linear AR(1) process,

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G} \widetilde{\beta}_{jg} y_j^t + \varepsilon_g^t \,, \tag{1}$$

where μ_g is the basal expression level of gene g; $\widetilde{\beta}_{jg} = \gamma_{jg}\beta_{jg}$ measures the influence of gene j on gene g, with $\beta_{jg} \in \mathbb{R}$ and $\gamma_{jg} = 1$ if j regulates g and $\gamma_{jg} = 0$ otherwise; finally, ε_g^t is an idiosyncratic error term, centred at zero and with precision parameter λ_g, typically assumed to be Gaussian. We augmented the model with the parenthood (link) indicator variables $\Gamma = \{\gamma_{jg}\}$ which will be the basis for estimating the network topology.

In order to account for the additional uncertainty when having repeated measurements we assume that the regulation process can be captured by (1), but instead of actually observing y_g^t, we have noisy measurements, x_{gr}^t, such that

$$x_{gr}^t = y_g^t + \eta_{gr}^t \,, \qquad r = 1 \ldots, R \,, \tag{2}$$

with η_g^t a zero mean error measurement term, with precision parameter τ_g, independent for all t, g, r. This error term is frequently assumed Gaussian; however, given that the measurement process can potentially produce outliers, we will use a Student-t specification, $\mathrm{St}\big(\eta_g^t \mid 0, \tau_g, \nu\big)$, such that $\mathrm{Var}\left[\eta_g^t\right] = \nu\, \tau_g^{-1}/(\nu - 2)$ provided the degrees of freedom, $\nu > 2$.

2S THE PRIOR

The likelihood for the Student-t measurement AR(1) model is,

$$\ell(\Theta; X) = \prod_{g=1}^{G} \prod_{t=1}^{T} \prod_{r=1}^{R} \mathrm{N}\Big(y_g^{t+1} \mid \mu_g + \widetilde{\boldsymbol{\beta}}_g \boldsymbol{y}^t, \lambda_g\Big)\, \mathrm{N}\big(x_{gr}^t \mid y_g^t, \omega_{gr}^t \tau_g\big)\, \mathrm{Ga}\big(\omega_{gr}^t \mid \nu/2, \nu/2\big)\,. \tag{3}$$

Where $Y = \{y_g^t\}$ are the unobserved expression levels, $X = \{x_{gr}^t\}$ denote their surrogate measurements and $\Theta = \{\boldsymbol{\mu}, B, \Gamma, \boldsymbol{\lambda}, \boldsymbol{\tau}, \nu\}$ collects all the parameters involved, with $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_G\}$; $B = \{\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_G'\} \in \mathbb{R}^{G\times G}$ and $\boldsymbol{\beta}_g = \{\beta_{1g}, \ldots, \beta_{Gg}\}$; $\Gamma = \{\gamma_{ij}\}$; $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_G\}$; and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_G\}$.

We specify a product form (independent) prior,

$$\pi(\Theta) = \pi(\rho)\, \pi(\nu) \left[\prod_{g=1}^{G} \pi(\mu_g)\, \pi(\boldsymbol{\beta}_g)\, \pi(\lambda_g)\, \pi(\tau_g)\, \pi(\boldsymbol{\gamma}_g)\right], \tag{4}$$

where,

$$\pi(\mu_g) = \mathrm{N}(\mu_g \mid 0, k)\,, \tag{5}$$

$$\pi(\boldsymbol{\beta}_g) = \mathrm{N}_G\big(\boldsymbol{\beta}_g \mid \mathbf{0}, k_\beta I\big),\ g = 1, \ldots, G\,, \tag{6}$$

$$\pi(\lambda_g) = \mathrm{Ga}(\lambda_g \mid a_\lambda, b_\lambda)\,, \tag{7}$$

$$\pi(\tau_g) = \mathrm{Ga}(\tau_g \mid a_\tau, b_\tau)\,, \tag{8}$$

$$\pi(\boldsymbol{\gamma}_g \mid \rho) = \prod_{j=1}^{g} \mathrm{Ber}(\gamma_{jg} \mid \rho),\ g = 1, \ldots, G\,, \tag{9}$$

$$\pi(\rho) = \mathrm{Be}(\rho \mid a_\rho, b_\rho)\,, \tag{10}$$

$$\pi(\nu) = \mathrm{Ga}(\nu \mid a_\nu, b_\nu)\,. \tag{11}$$

Given that the data is standardised before performing the estimation (zero mean and unitary standard deviation for each time series), we set $k_\mu = k_\beta = 1/4$; i.e. the prior variance of any component of $\boldsymbol{\mu}$ and B is four. In our experience, this is typically not over-informative for microarray data.

As mentioned in the paper, when repeated measurements are available it is easier to estimate $\boldsymbol{\tau}$ than $\boldsymbol{\lambda}$. Thus, we set $\{a_\tau, b_\tau\} = \{2, 1/100\}$ which is renders a rather flat prior with mode at 100 and variance of 20000.

For the autoregressive precision $\boldsymbol{\lambda}$, we used $\{a_\lambda, b_\lambda\} = \{1/10, 1/10\}$. Thus setting the prior mean at one and the variance at 10. The mode now does not exist.

Derived from the conditions given in the paper: $\mathrm{P}[\nu \leq 30] \approx 0.6$ and $\mathrm{Mode}[\nu] = 15$, it is straightforward to verify that $\{a_\nu, b_\nu\} = \{3.5, 0.15\}$.

In the absence of any prior information, we treat ρ as the probability of any given link to be present and thus use the corresponding reference prior, $\mathrm{Be}(\rho \mid 1/2, 1/2)$ (Bernardo and Smith, 1994, p. 315) .

3S THE SAMPLER

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

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

and thus can be sampled from a gamma distribution.

Constant term μ_g is conditionally Gaussian, with mean and precision

$$m_g = \frac{\overline{y}_g^{t+1} - \overline{\boldsymbol{y}}^t\widetilde{\boldsymbol{\beta}}_g}{\lambda_g + k_\mu/T} \qquad \text{and} \qquad \tau'_\mu = k_\mu + T\,\lambda_g\ ,$$

respectively, where $\overline{y}_g^{t+1} = T^{-1}\sum_t y_g^{t+1}$ and $\overline{\boldsymbol{y}} = T^{-1}\sum_t \boldsymbol{y}^t$.

Connectivity The overall connectivity, ρ, is sampled from a $\text{Be}\big(\rho \mid S + a_\rho, G^2 + b_\rho - S\big)$, with $S = \sum_{i,j=1}^{G}\gamma_{ji}$.

Measurement precision For each gene $g = 1, \ldots, G$, the measurement precision, τ_g are updated from a gamma distribution $\text{Ga}\big(\tau_g \mid a'_\tau, b'_\tau\big)$ with

$$a'_\tau = R\,T/2 + a_\tau \qquad \text{and} \qquad b'_\tau = b_\tau + \frac{1}{2}\sum_{t=1}^{T}\sum_{r=1}^{R}\omega_{gr}^t\left(x_{gr}^t - y_{gr}^t\right)^2\ .$$

Degrees of freedom We use a Metropolis-within-Gibbs strategy to draw a new value, $\nu^{(m)}$, with a gamma proposal with its mean fixed at the previous draw, $\nu^{(m-1)}$. We control for the acceptance rate to lie around 1/3 by tuning the proposal's coefficient of variation, cv. Thus, we propose a new $\nu^{(m)}$ from

$$\text{Ga}\big(\nu^{(m)} \mid cv^{-2}, cv^{-2}/\nu^{(m-1)}\big).$$

Coefficients and link probabilities The update of each indicator variable γ_{jg} is performed jointly with all the corresponding coefficients

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

with acceptance probability

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

where the subscripts have been removed for clarity. Given that we propose γ symmetrically, $q(\gamma^a)/q(\gamma^b) = 1$. The Hastings ratio is then

$$\frac{q(\boldsymbol{\beta}^a \mid \gamma^a)}{q(\boldsymbol{\beta}^b \mid \gamma^b)} = \frac{\rho}{1-\rho}\,k_\beta^{1/2}\,\frac{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^b\,\Sigma_{\boldsymbol{\beta}}^{-1b}\mu_{\boldsymbol{\beta}}^b\right]}{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^a\,\Sigma_{\boldsymbol{\beta}}^{-1a}\mu_{\boldsymbol{\beta}}^a\right]}\frac{\left|\Sigma_{\boldsymbol{\beta}}^b\right|^{1/2}}{\left|\Sigma_{\boldsymbol{\beta}}^a\right|^{1/2}}\ .$$

with $\Sigma_\cdot$ the covariance matrix obtained by considering only the relevant gene expression vectors. For the opposite move *i.e.* switching a link off, we use the reciprocal of the ratio above.

Non-observables These are drawn from a Gaussian distribution, $\mathrm{N}\left(y_g^t \mid m_g^t, p_g^t\right)$, with location

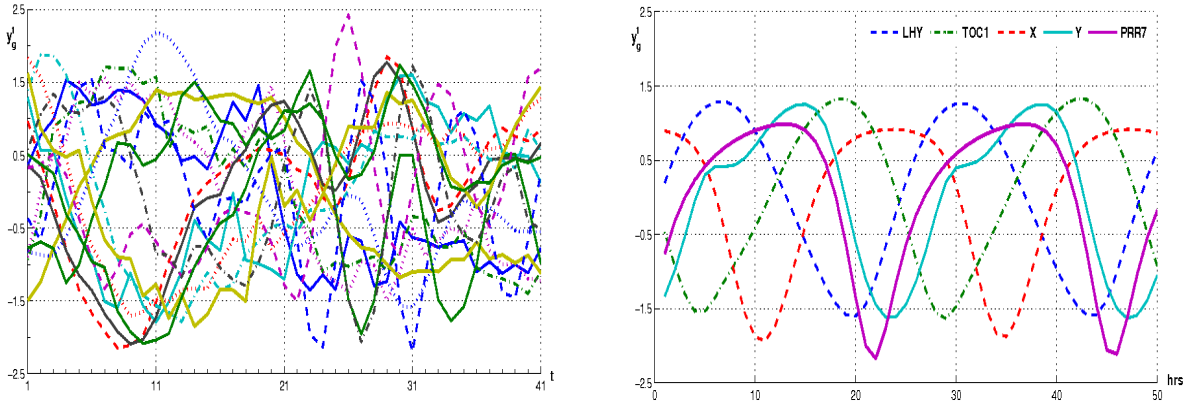
$$m_g^t = \frac{\lambda_g m_{\mathrm{AR}} + \tau_g m_{\mathrm{meas}}}{p_g^t} \qquad \text{and precision} \qquad p_g^t = \lambda_g\left(1 + \widetilde{\beta}_{gg}^2\right) + \tau_g \sum_{r=1}^{R} \omega_{gr}^t \,,$$

where

$$m_{\mathrm{AR}} = \sum_{i \neq g} \widetilde{\beta}_{ig}\left(y_i^{t-1} - \widetilde{\beta}_{gg}\, y_i^t\right) + \widetilde{\beta}_{gg}\left(y_g^{t-1} + y_g^{t+1}\right) \qquad \text{and} \qquad m_{\mathrm{meas}} = \sum_{r=1}^{R} \omega_{gr}^t x_{gr}^t \,.$$

4S DATA SETS

Time traces of the data sets used in Section 4 of the paper. The linear *in silico* data, Figure-S 1a, comprises 16 genes measured at 41 time points. The ODE data has five genes and 50 measurements in time, Figure-S 1b.



(a) *Linear interactions* **(b)** *ODE data*

Figure-S 1. In silico *data sets. Traces of the noiseless synthetic linear and ODE, non-linear data sets.*

The *Arabidopsis* data set has 5 genes with 24 time points and four repetitions.

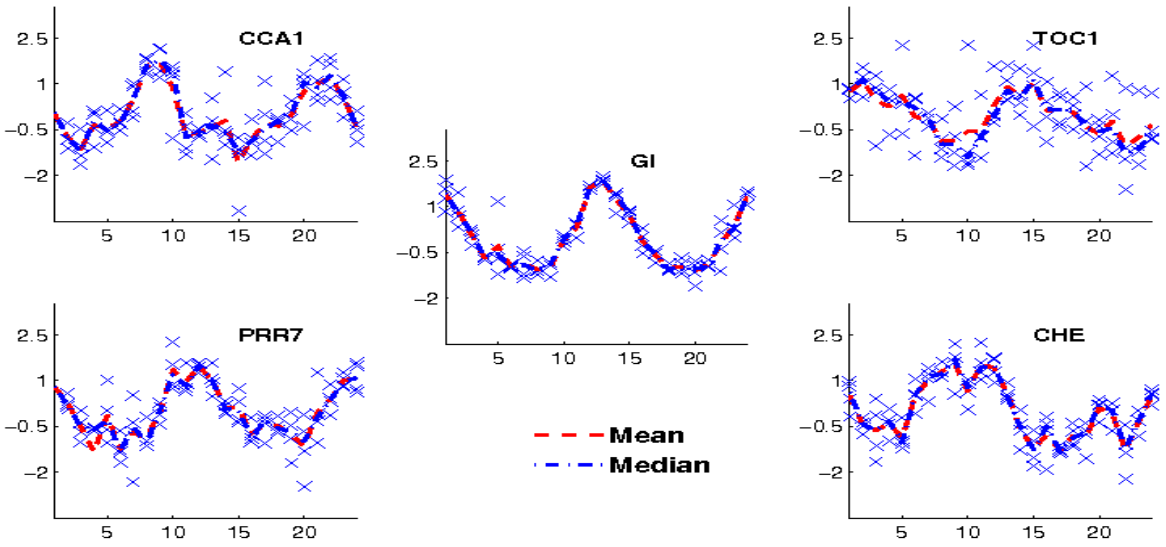


Figure-S 2. *Circadian clock related genes in* Arabidopsis thaliana. *Gene expression repeated measurements. The mean (dashed) and median (dot-dashed) of each time point are plotted as time series.*

4.1S MEAN CROSS ENTROPY

The score used in the paper, MxE, is simply the Kullback-Liebler divergency from any specific link to the true network configuration, $KL(\hat{\gamma}_{ij} \mid p_{ij})$,

$$KL(\hat{\gamma}_{ij} \mid p_{ij}) = p_{ij} \log \frac{p_{ij}}{\hat{\gamma}_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - \hat{\gamma}_{ij}}$$

averaged over all possible links, $i, j, = 1, \ldots, G$, where $p_{ij} = 1$ if the link is present and zero otherwise; with the convention of $0 \log 0 = 0$.

4.2S *In silico* DATA

As expected, when a large number of replicates are available the three models yield similar networks, for a given threshold. However, the effect of overestimation in the regression coefficients posterior precisions is apparent when looking at the probabilities predicted by MM: again, they are less disperse than those predicted by either GM or SM.

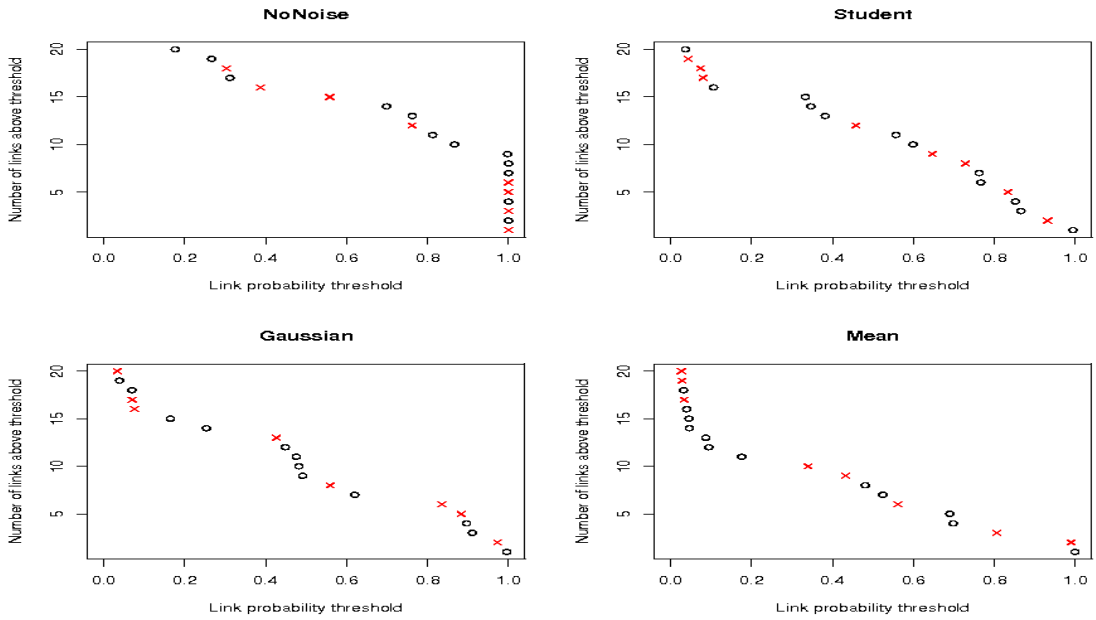


Figure-S 3. *ODE circadian clock* in silico *data set with* $R = 20$ *replicates. Number of links predicted present in the network versus posterior link probabilities for each model considered. A link present in the ODE model is highlighted with a cross.*

4.3S *In vivo* DATA

To have a more or less representative sample, we calculated all the possible subsamples with three replicates, $\mathcal{P}_3$, and then measured the Euclidean distance between the standard deviations of the original data and $\mathcal{P}_3$. These were classified into large, medium and small, based on their empirical distribution and ten series were selected from each region. For the 1-replicate case we used the Euclidean distance between the mean of the four replicate data set and the single data set. Interestingly we found no apparent effect of the Euclidean distance in the retrieved topologies and therefore we joined them when calculating the counts tables.

REFERENCES

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