Two-way sparsity for time-varying networks, with applications in genomics

Thomas E. Bartlett1*, Ioannis Kosmidis2,3, and Ricardo Silva1,3.

1. Department of Statistical Science, University College London, London, United Kingdom.
2. Department of Statistics, University of Warwick, Coventry, United Kingdom.
3. The Alan Turing Institute, London, United Kingdom.
* E-mail: thomas.bartlett.10@ucl.ac.uk

Abstract

We propose a novel way of modelling time-varying networks, by inducing two-way sparsity on local models of node connectivity. This two-way sparsity separately promotes sparsity across time and sparsity across variables (i.e., within time). Separation of these two types of sparsity is achieved with the introduction of a novel prior structure, which draws on ideas from the Bayesian lasso and from copula modelling. We provide an efficient implementation of the proposed model via a Gibbs sampler, and we apply the model to data from neural development. In doing so, we demonstrate that the model we propose is able to infer changes in genomic network structure which match current biological knowledge. The novel network structures which are inferred by the proposed model identify potential targets for further experimental investigation by neurobiologists.

1 Introduction

Network models have become an important topic in modern statistics, and the evolution of network structure over time (illustrated in Figure 1) is an important new area of study. Network structures which evolve over time naturally occur in a range of applications. Examples of recent applications include evolving patterns of human interaction [Durante et al., 2016] such as in social networks [Sekara et al., 2016], time-varying patterns of interaction between genes and their products in biological networks [Alexander et al., 2009, Lebre et al., 2010], and time-varying patterns of connectivity in the brain [Schaefer et al., 2014]. However, network models with temporal structure have only recently begun to be studied in detail by statistical scientists.

An important application area of statistical network models is genomics. Network models are a natural way to describe and analyse patterns of interactions (represented by network edges) between genes and their products (represented by network nodes). An important interaction of this type is gene regulation, in which the product of one gene influences the output level of the product of a different gene. Much

![Figure 1: Model of time-varying network structure. Each \( x_{t,i} \) represents a class label or continuous variable for node \( i \) (e.g., the expression-level of gene \( i \)) at time \( t \). The links represent network interactions or dependencies between \( x_{t,1}, x_{t,2}, \ldots \) (e.g. due to gene regulation), which may be different to those between \( x_{t,1}, x_{t,2}, \ldots \) and \( x_{t,1}, x_{t,2}, \ldots \). Hence, these network interactions may vary with time.]
gene regulation is characteristic of a particular cell type, so that a cell knows its role within the organism: these specific regulatory network structures are established during embryonic development. Changes in normal gene regulation are also inherent to cancer progression, so that cells ‘forget’ how they should act, taking on pathological roles (regulatory network re-wiring) [Suvà et al., 2014]. However, whilst network models are well established in genomics, historically these models have typically been static, ignoring the fact that genomic processes are inherently time-varying.

Recently, there has been much strong work on models of time-varying networks, reflecting the current interest in this area. In statistics, this work covers methods based on Markov processes [Crane et al., 2016], on dynamic Erdős-Rényi graphs [Rosengren and Trapman, 2016], and on sparse regression methods [Kolar et al., 2010]. It also includes work on time-varying community structure [Zhang et al., 2012] and on methods extending the stochastic block model [Xu and Hero III, 2013, Matias and Miele, 2016] and related non-parametric graphon-based methods [Pensky, 2016], as well as non-parametric methods for dynamic link prediction [Sarkar and Chakrabarti, 2014] and methods from Bayesian nonparametrics [Palla et al., 2016]. Other related work includes sparse graphical models which can account for samples/observations taken at different time-points [Kalaitzis et al., 2013].

Motivated by genomics applications, we propose a novel way of modelling time-varying networks, by inducing two-way sparsity on local models of the connectivity of each node to all the others. This is achieved as follows. We start with a regression likelihood function that assumes that observations are mutually independent over time. Importantly, we induce dependence through a novel prior structure that promotes sparsity in a two ways: across time, and within time. This decoupling of the induced sparsity is achieved through a copula specification for the parameters in the likelihood function. Specifically, the regression coefficients for one node across different time-points are jointly distributed according to a Gaussian copula with Laplace marginal distributions. The correlation matrix of the Gaussian copula is formed by assuming that the correlation between time-points decays with time in a structured, parsimonious way that also ensures its positive definiteness. In this correlation matrix, the only free parameter is the correlation between consecutive time-points, which is given a reverse-exponential prior distribution with support [0, 1]. This prior on the correlation across time discourages large differences in the regression coefficients between consecutive time-points and, as a consequence, also discourages large changes in the inferred edge structure of the network.

The decoupling of the marginal and dependence structure that is allowed by the copula specification and the particular form of the correlation matrix allows for precise control of marginal priors. This decoupling also makes the adoption of generalisations of the Bayesian lasso, such as the horseshoe [Carvalho et al., 2010], easy to implement in place of the marginal Bayesian lasso prior which we use. The prior dependence among parameters across time can also be viewed as a Bayesian version of the fused lasso [Tibshirani et al., 2005], while within each time-slice we directly utilise existing work on the Bayesian lasso [Park and Casella, 2008]. In fact, the proposed modelling framework has the Bayesian lasso as a special case, when the correlation between time-points is set to zero. From a frequentist point of view, the proposed sparsity structure would fall within the remit of the generalised lasso [Tibshirani et al., 2011], which has the fused lasso as a special case [Tibshirani et al., 2005]. Bayesian versions of the fused lasso have also been proposed [Kyung et al., 2010, Shimamura et al., 2016], however, a key difference with the modelling framework we propose is the formal decoupling of sparsity across time, which the fused lasso induces, from sparsity within time.

The modelling approach which we propose differs in important ways from earlier work by Kolar et al. [2010], which also uses sparse regression for modelling time-varying networks, and which has naturally influenced this work. In particular, in this work we show how we can estimate posterior distributions that are useful to simultaneously infer network edges and quantify edge uncertainty. In this paper, the problem is also decomposed into independent local fits for each node with sparsity both across and within time. In practice, this means that we are able to work with large networks with over 20000 nodes, at the same time allowing the data analyst to quantify parameter uncertainty.
from each local model. The novel prior structure proposed, which enables the time-varying network inference, will also be of interest more generally beyond the context of network inference. This novel prior structure will be relevant to any context where sparse regression with time-varying regression parameters is desirable.

The rest of the paper is structured as follows. In Section 2, we set up notation, and specify the model. Then, in Section 3 we present the results of fitting the model to simulated data, and in Section 4 we present the results of fitting the model to single-cell gene-expression data. Finally, in Section 5 we summarise our findings and discuss their broader context. All proofs and derivations appear in the supplement.

2 Proposed methodology

2.1 Data description

The two-way sparsity that is induced by the proposed modelling framework is motivated by the problem of inferring time-varying structure in genomic networks, in which network nodes represent genes. For each node in such a network, there are observations or measurements of the activity level of the corresponding gene (the ‘gene-expression level’): these observations constitute the data-set. The expression-level of a gene is generally influenced by the expression-level of several other genes (a process called ‘gene regulation’). Hence, a natural application of models for time-varying networks is to understand dynamic patterns of gene-regulation in biological processes, such as neural development. These time-varying patterns of gene regulation can be studied in detail with single-cell gene-expression data.

Single-cell gene-expression data are ideal for this application, because data from a study of this type will typically be obtained from a heterogeneous mixture of cells, each of which may be at a different point on a trajectory through the biological process under investigation. For example, in the context of neural development, some of these cells may be stem-cells, whereas some may be fully differentiated cells (e.g., neurons), with a whole spectrum of cells in between. Each cell can be thought of as an independent sample from an underlying latent biological process; in this example, that process is neural development. Thus, we can think of the progression of a cell through this process of neural development in terms of a ‘developmental trajectory’. The progression through such a developmental trajectory can be quantified in terms of ‘developmental time’, in the sense that it is a measure of a temporally-ordered progression through the process of cellular development.

For each of the cell-samples in the data-set, no information is available other than its high-dimensional gene-expression measurements. Hence, in data-sets such as this, it is necessary to first infer the ‘developmental time’ of the cell-samples before fitting the time-varying network model. Such inference is more generally referred to as ‘pseudo-time’ inference, and several methods exist to do this inference, e.g. the work by Qiu et al. [2011] and Trapnell et al. [2014]. Recently, ‘pseudo-time’ for the developmental time of cells has also been referred to as ‘pseudo-development’ [Nowakowski et al., 2017]. We will refer to this notion as ‘pseudo developmental time’, abbreviated as PD-time. The inferred PD-time orderings thus infer a time-stamp for each cell-sample, for use in the time-varying network modelling.

2.2 Model overview

The proposed modelling framework starts by assuming that the network structure can be decomposed locally. This assumption has been used previously by Kolar et al. [2010], and it allows the network structure to be inferred independently around each node $i$. Inference is carried out with a sparse linear model, taking the observations for node $i$ at time $t$ as the response, and the observations for all
nodes \( j \neq i \) at time \( t \) as potential predictors. From these potential predictors, the set of predictors ‘chosen’ by the sparse model fit are then used to infer the network structure, as follows.

We are concerned here with the problem of inferring network structure on a fixed set of nodes, with a set of edges that varies with time. In this scenario, only the patterns of interconnectivity change as the network evolves (Figure 1), which is the scenario most relevant to genomics applications. Such a network can be represented with a time-varying adjacency matrix \( A^{(t)} \), where \( A^{(t)}_{ij} \) denotes the absence \( (A^{(t)}_{ij} = 0) \) or presence \( (A^{(t)}_{ij} > 0) \) of an edge between nodes \( i \) and \( j \) at time \( t \). We are modelling undirected networks here, and so \( A^{(t)} \) should be symmetric, i.e., \( A^{(t)}_{ij} = A^{(t)}_{ji} \). However, because we infer the local network structure separately around nodes \( i \) and \( j \), we could end up inferring two different values for \( A^{(t)}_{ij} \) and \( A^{(t)}_{ji} \). Thus, some form of ‘symmetrisation’ is required. We follow the conservative ‘min_symmetrisation’ scheme presented by Kolar et al. [2010]. Following this scheme, we infer \( \hat{A}^{(t)}_{ij} = \hat{A}^{(t)}_{ji} > 0 \) if and only if a network edge is inferred at time \( t \) between nodes \( i \) and \( j \) according to the local model fit around node \( i \), and also according to the independent local model fit around node \( j \). Otherwise, we infer \( \hat{A}^{(t)}_{ij} = \hat{A}^{(t)}_{ji} = 0 \). We present the results of inferring networks from real data using this ‘min_symmetrisation’ scheme in Section 4.

### 2.3 Likelihood definition

We assume a likelihood function where observations are mutually independent over time. This is an assumption that is compatible with high-dimensional gene-expression data, where no single cell can be measured at more than one time point. Let \( y_t \) denote the value for some node in the system at time \( t \in \{1, ..., T\} \), and let the row-vector \( x_t \) denote the values for the other \( p - 1 \) nodes at time \( t \). Then, the dependence of \( y_t \) on \( x_t \) is modelled as

\[
y_t = a + b_{t:} x_t^\top + \epsilon_t, \tag{1}
\]

where \( b_{t:} \) is a row-vector of linear model coefficients with prior structure specified in Section 2.4, \( x_t \) is a row-vector, and \( \epsilon_t \sim \mathcal{N}(0, \tau^{-1}) \).

The response variable \( y_t \) corresponds to the observations for a ‘target’ node around which we are modelling the local network structure, whereas the variable \( x_t \) corresponds to the observations for all the other nodes of the network. To model the whole network, we must fit model (1) around each target node in turn. We note that this is an assumption about the existence of a global undirected Markov network [Lauritzen, 1996] that explains the independence constraints in the model, as used previously by other authors in an equivalent context [Kolar et al., 2010].

Using \( b_{:,j} \) to denote the column-vector of model coefficients for predictor \( j \) for \( t \in \{1, ..., T\} \), we collect parameters in matrix \( B = [b_{:,1}, b_{:,2}, ..., b_{:,p-1}] \). In the next section, we postulate a prior for dependencies within each column \( j \) of \( B \), whereas the columns of \( B \) (each corresponding to a different node as predictor) are independent of each other. We also introduce here the notation \( y_t^{(k)} \) and \( x_t^{(k)} \) to represent observations of \( y_t \) and \( x_t \) for sample \( k \), and we define \( \zeta(t) \) as the set of samples which correspond to time \( t \). This formulation is quite general, but it also makes it straightforward to model time-varying network structure when the times of the samples must be inferred [Qiu et al., 2011, Trapnell et al., 2014] before the network structure is inferred. We also denote \( y_t^{(k)} \) and \( x_t^{(k)} \) over all \( k \in \{1, ..., n\} \) as \( y = [y_1^{(1)}, y_1^{(2)}, ..., y_T^{(k)}, ..., y_T^{(n)}] \) and \( X = [(x_1^{(1)})^\top, (x_1^{(2)})^\top, ..., (x_1^{(k)})^\top, ..., x_T^{(n)})^\top] \). Hence, we can write the model likelihood for the data-set \( \{y, X\} \) for the local network structure around the target node as:

\[
P(y|X, B, a, \tau) = \prod_{t=1}^{T} \prod_{k \in \zeta(t)} \sqrt{\frac{\tau}{2\pi}} e^{-\tau (y_t^{(k)} - b_{:,1}x_t^{(k)})^2/2}. \tag{2}
\]
2.4 Priors with decoupled two-way sparsity

The elements of $b_t, (t = 1, ..., T)$ are marginally distributed as $b_{t,j} \sim \text{Laplace}(1/\lambda)$, with probability density function $rac{\lambda}{2} e^{-\lambda |t|}$ and cumulative distribution function $F_L(b_{t,j})$, for $t \in \{1, ..., T\}, j \in \{1, ..., p-1\}$. Let $u_{t,j} = F^{-1}_L(b_{t,j})$ and define $\theta_{t,j} = \Phi^{-1}(u_{t,j})$, where $\Phi$ is the standard normal cumulative distribution function. The dependencies between $b_{t-1,j} = F^{-1}_L(u_{t-1,j})$ and $b_{t,j} = F^{-1}_L(u_{t,j}), t \in \{2, ..., T\}$, are modelled through the joint distribution of these $u_{t,j}$, defined as

$$
\begin{bmatrix}
\theta_{1,j} \\
\theta_{2,j} \\
\vdots \\
\theta_{T,j}
\end{bmatrix}
\sim \mathcal{N}(0, \Sigma_j), \quad \text{with} \quad \Sigma_j =
\begin{bmatrix}
1 & \rho_j & \cdots & \rho_j^T \\
\rho_j & 1 & \cdots & \rho_j^{T-1} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_j^T & \rho_j^{T-1} & \cdots & 1
\end{bmatrix}.
$$

(3)

The correlation parameter $\rho_j$ is assumed to have a reverse-exponential distribution with support $\in [0, 1]$ and density

$$f_{\text{rexp}}(\rho_j) \sim ke^{k\rho_j}/(e^k - 1).$$

(4)

The structure of $\Sigma_j$ is such that model coefficients at adjacent points in time, such as $\theta_{t,j}$ and $\theta_{t+1,j}$, have correlation $\rho_j$ (Figure 2). Then, the coefficients separated by two time-points have correlation $\rho_j^2$, etc. Thus, the sequence of model coefficients $\theta_{1,j}, \theta_{2,j}, ..., \theta_{T,j}$ forms a Markov chain, meaning that $\Sigma_j$ is guaranteed to be positive-definite for $\rho_j \in [0, 1]$, and by construction

$$\theta_{t+1,j} \perp \theta_{t-1,j}, \theta_{t-2,j}, ..., |\theta_{t,j}. $$

(5)

The construction of $\Sigma_j$ in this way achieves sparsity across time, by discouraging differences in the regression coefficients (and hence also the inferred network structure) between adjacent time-points. Then transforming the $\theta_{t,j}$ to $b_{t,j}$, where the $b_{t,j}$ are marginally Laplace distributed, achieves sparsity within time by discouraging non-zero regression coefficients, and hence encouraging discovery of sparse network structures.

Recent work which takes a Bayesian approach to generalising the fused lasso [Shimamura et al., 2016] could be used similarly to the approach we propose, by modelling the same set of predictors at multiple time-points whilst enforcing smooth changes across time as well as sparsity overall. However, Shimamura et al. [2016] achieve their result by simply multiplying together separate frequentist-inspired priors for smoothness across time and for sparsity. Specifically, they multiply together
a Laplace prior to penalise individual non-zero model coefficients, with the ultra-sparse negative-exponential-gamma (NEG) prior to penalise non-zero differences in coefficients. The Laplace-NEG prior is defined in this context as:

\[ P(b_{t,j}) \propto \prod_{t=1}^{T} \text{Laplace}(b_{t,j}|\lambda) \prod_{t=2}^{T} \text{NEG}(b_{t,j} - b_{t-1,j}|\lambda^\dagger, \gamma), \]

where the Laplace density is defined as \( \frac{\lambda}{2} e^{-\lambda|\cdot|} \), and

\[ \text{NEG}(\cdot|\lambda^\dagger, \gamma) = \int_0^\infty \int_0^\infty f_N(\cdot|0, \tau^2) f_\gamma(\tau^2|1, 1/\psi)f_\gamma(\psi|\lambda^\dagger, 1/\gamma^2) d\tau^2 d\psi, \]

where \( f_N \) and \( f_\gamma \) are the Normal and Gamma densities, respectively. Sampling from the distribution of equation (6) is via a multivariate Normal mixture, with exponential and gamma priors on the variance components, as specified by Shimamura et al. [2016].

In contrast to the Laplace-NEG prior, the model we propose retains the property that, marginally, each coefficient still follows the Bayesian lasso prior [Park and Casella, 2008]. In particular, if we set \( \rho_j = 0 \) (for \( j = 1, 2, ..., p - 1 \)), then the model we propose is exactly the same as the Bayesian lasso. This is important because it makes it easier to set priors, including variants of the Bayesian lasso that avoid the well-known shortcomings of this prior (see for example the work by Castillo et al. [2015] and van der Pas et al. [2016]). Although we choose not to discuss such variants here to simplify the presentation, the construction and discussion based on the Bayesian lasso is immediately applicable to those variants too.

We achieve decoupled sparsity across time by specifying a prior on \( \rho_j \), and construct \( \Sigma_j \) from this \( \rho_j \) (equation (3)) to fully specify the decoupled-sparsity prior. The novel prior we use on \( \rho_j \) is a ‘reverse exponential prior’ (equation (4)). Figure 3 shows the probability density function of the reverse-exponential prior for different values of hyper-parameter \( k \). Figure 4 then shows heatmaps of the bivariate density distributions of samples from the novel decoupled-sparsity prior for a coefficient \( j \) over two time-points, i.e., \( b_{t,j} = [b_{1,j}, b_{2,j}]^\top \), for a range of values of \( \lambda \) and \( k \) (the corresponding marginal densities are shown in Figures S6 and S7).

For comparison, Figure S4 in supplement C shows samples from the Laplace-NEG prior as defined in equation (6), for various values of \( \lambda \) (which acts equivalently to \( \lambda \) in our model, controlling
2.5 Posterior inference

The order-1 Markovian relations specified by equation (5) are also computationally attractive, as it is well-known from the graphical modelling literature that such assumptions result in models with banded precision matrices. From equation (5), it follows that the partial correlation of \( \theta_{t+m,j} \) with \( \theta_{t+l,j} \) will be zero for all \( |m-l| > 1 \). Hence, all entries of the precision matrix \( \Sigma^{-1}_j \) will be zero except the diagonal and the elements immediately adjacent to it (i.e., the sub- and super-diagonals). These relationships allow all the entries of this precision matrix to be found easily in terms of \( \rho_j \) by solving \( \Sigma^{-1}_j \Sigma_j = I \). The entries of the precision matrix \( \Sigma_j^{-1} \) can be found in this way as:

\[
(\Sigma_j^{-1})_{t,t'} = \begin{cases} 
1/(1 - \rho_j^2), & \text{if } t' = t = 1 \text{ or } t' = t = T \\
(1 + \rho_j^2)/(1 - \rho_j^2), & \text{if } t' = t > 1 \text{ and } t' = t < T \\
-\rho_j/(1 - \rho_j^2), & \text{if } t' = t + 1 \text{ or } t' = t - 1 \\
0, & \text{otherwise}
\end{cases}
\]

(7)

where \( (\Sigma_j^{-1})_{t,t'} \) represents the \((t,t')\) element of the precision matrix \( \Sigma_j^{-1} \). For completeness, a full derivation of equation (7) appears in Supplement A.

The model coefficients \( \mathbf{B} \) can be sampled directly from multivariate Normal distributions, without needing the intermediate transformation to the marginally Laplace-distributed variables described in Section 2.4. This can be achieved with an algebraic manipulation which is an extension from the Bayesian lasso, as follows. The Laplace distribution can be written as an uncountable mixture of zero-mean Normal distributions with standard-deviations \( s_j \), with these \( s_j \) sampled from the Gamma(1, \( \frac{2}{\lambda^2} \)) [Andrews and Mallows, 1974, Park and Casella, 2008]. Specifically,

\[
P(b_{t,j} | \lambda) = \frac{\lambda}{2} e^{-|b_{t,j}|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-b_{t,j}^2/(2s)} \frac{\lambda^2}{2} e^{-\lambda^2 s^2/2} ds,
\]

and therefore we can write

\[
P(b_{t,j}, s_j | \lambda) = \frac{1}{\sqrt{2\pi s_j}} e^{-b_{t,j}^2/(2s_j)} \frac{\lambda^2}{2} e^{-\lambda^2 s_j/2},
\]

for \( s_j \sim \text{Gamma}(1, \frac{2}{\lambda^2}) \). This says that we will achieve \( b_{t,j} \) being marginally Laplace distributed by sampling the \( b_{t,j} \) from zero-mean Normal distributions with standard-deviations \( s_j \), with these \( s_j \) sampled from the Gamma(1, \( \frac{2}{\lambda^2} \)) prior. Hence

\[
P(b_{t,j} | s_j) = \frac{1}{\sqrt{2\pi s_j}} e^{-b_{t,j}^2/(2s_j)} = P(\sqrt{s_j} \theta_{t,j}),
\]

and so \( b_{t,j} \) follows the same Normal distribution as \( \theta_{t,j} \) with the variances and covariances scaled up by \( s_j \), with \( s_j \sim \text{Gamma}(1, \frac{2}{\lambda^2}) \). Therefore, also referring back to equation (3), it follows that

\[
P(b_{t,j} | s_j, \rho_j, \lambda) = \frac{\lambda^2}{2} e^{-\lambda^2 s_j^2/2} \frac{1}{(2\pi)^{T/2} s_j^{1/2}} e^{-b_{t,j}^T \Sigma_j^{-1} b_{t,j}/2}.
\]
To make sampling easier, at this stage we choose to make the substitution
\( s_j = \nu_j^{-1} \), leading to the density
\[
P(b_{\cdot j}, \nu_j | \rho_j, \lambda) = \frac{1}{\nu_j^2} \frac{\lambda^2}{2} e^{-\lambda^2/(2\nu_j)} \frac{\nu_j^{1/2}}{(2\pi)^{T/2} |\Sigma_j|^{1/2}} e^{-b_{\cdot j}^\top \nu_j \Sigma_j^{-1} b_{\cdot j}/2} \\
= \frac{\lambda^2}{2} e^{-\lambda^2/(2\nu_j)} \frac{\nu_j^{-3/2}}{(2\pi)^{T/2} |\Sigma_j|^{1/2}} e^{-b_{\cdot j}^\top \Sigma_j^{-1} b_{\cdot j}/2},
\] (8)
where the extra factor of \( 1/\nu_j^2 \) comes from \( |d\{\nu_j^{-1}\}/d\nu_j| \) as required for the change of variable in this probability density function. Assuming that the model will be fit to standardised data, we set the prior on the intercept as \( a \sim N(0, 1) \), and we set the prior on the model precision as \( \tau \sim \text{Gamma}(1,1) \) (which has prior mean 1, with 95% of the prior mass between 0.025 and 3.7, which we believe is reasonable for these data). Now combining equation (8) with these prior specifications, and \( P(\rho_j | k) = \frac{k}{k-1} e^{k \rho_j} \) (for \( 0 \leq \rho_j \leq 1 \)), as well as with the model likelihood (equation (2)), we get:
\[
P(y, B, \rho, \nu, a, \tau | X, \lambda, k) = \left\{ \prod_{t=1}^T \prod_{k \in \xi(t)} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}} (y_t^{(k)} - b_{\cdot t} \cdot [x_t^{(k)}]')^2/2 \right\} \\
\times \frac{1}{\sqrt{2\pi}} e^{-\frac{(\tau+a)/2}{2}} \prod_{j=1}^{p-1} \left\{ \frac{k}{e^k-1} e^{k \rho_j} \frac{\lambda^2}{2} e^{-\lambda^2/(2\nu_j)} \frac{\nu_j^{-3/2}}{(2\pi)^{T/2} |\Sigma_j|^{1/2}} e^{-b_{\cdot j}^\top \Sigma_j^{-1} b_{\cdot j}/2} \right\}.
\] (9)

Following equation (9), the model described in Sections 2.3 and 2.4 can be implemented as a Gibbs sampler with the the steps given in Algorithm 1. We note that Algorithm 1 has a relatively low computational cost, because each of the steps (with the exception of step 4) involves sampling from a known distribution for which the parameters can be easily calculated. Then for step 4, we can simply use a slice-sampler to sample \( \rho_j \), which has finite support \( \rho_j \in [0, 1] \). The full derivations of equations (10)-(14) appear in supplement B.

**Algorithm 1.** A Gibbs sampler with the following steps:

1. **Sample \( a \) from:**
   \[ P(a | y, X, ...) \propto f_N(a | \mu_a, \sigma_a) = g_a(a), \] (10)
   where \( \sigma_a^{-2} = 1 + n \tau \) and \( \mu_a = \sigma_a^2 \tau \sum_{t=1}^T \sum_{k \in \xi(t)} (y_t^{(k)} - b_{\cdot t} \cdot [x_t^{(k)}]') \).

2. **Sample \( \tau \) from:**
   \[ P(\tau | y, X, ...) \propto f_{\gamma}(\tau | k_\tau, \theta_\tau) = g_\tau(\tau), \] (11)
   where \( f_\gamma \) is the density of the gamma distribution with \( k_\tau = 1 + \frac{n}{2} \) and
   \[ \theta_\tau = 1/\left(1 + \sum_{t=1}^T \sum_{k \in \xi(t)} (y_t^{(k)} - b_{\cdot t} \cdot [x_t^{(k)}]')^2/2\right). \]

3. **Sample \( \nu_j \) from:**
   \[ P(\nu_j | y, X, ...) \propto f_{IG}(\nu_j | \mu_\nu, \lambda_\nu) = g_{\nu_j}(\nu_j), \] (12)
   where \( f_{IG} \) is the density of the inverse Gaussian distribution with parameters \( \lambda_\nu = \lambda^2 \) and \( \mu_\nu = \lambda \sqrt{b_{\cdot j}^\top \Sigma_j^{-1} b_{\cdot j}} \).

4. **Sample \( \rho_j \) from:**
   \[ P(\rho_j | y, X, ...) \propto e^{k \rho_j} \frac{1}{|\Sigma_j|^{1/2}} e^{-b_{\cdot j}^\top \nu_j \Sigma_j^{-1} b_{\cdot j}/2} = g_{\rho_j}(\rho_j). \] (13)
5. Sample $b_{i,j}$ from:

$$P(b_{i,j}|y, X, ...) \propto f_{N}(b_{i,j}|\tilde{m}_j, \tilde{\Sigma}_j) = \tilde{g}_{b_j}(b_{i,j}),$$

where $f_{N}(\cdot | \mu, \Sigma)$ is the multivariate Normal density, $\tilde{\Sigma}_j^{-1} = \nu_j \Sigma_j^{-1} + V_j^{-1}$, and $\tilde{m}_j = \Sigma_j V_j^{-1} m_j$, where the $t$th element of the vector $m_j$ is

$$m_{t,j} = \sum_{k \in \xi(t)} x_{i,j}^{(k)} \left\{ y_t^{(k)} - b_{t,\lambda_j} \cdot [x_{i,j}^{(k)}]^\top - a \right\} / \sum_{k \in \xi(t)} [x_{i,j}^{(k)}]^2,$$

where $b_{t,\lambda_j}$ and $x_{i,j}^{(k)}$ represent $b_{t,i}$ and $x_{i,j}^{(k)}$ without the $j$th elements, respectively, and $V_j$ is a diagonal matrix, with the $t$th diagonal element equal to $1/\left\{ \tau \sum_{k \in \xi(t)} [x_{i,j}^{(k)}]^2 \right\}$.

3 Simulation study

In this section, we present the results of a simulation study. We generate simulated data with structure that we expect to be typical of real data, and then fit the proposed model to the simulated data. To generate the data, the observations for each node $j$ (corresponding to a potential predictor variable, see Section 2.2) are generated such that they follow a time-series of one of four types, as follows (these types a-d are also illustrated in Figure 5):

(a) Monotonic; decreasing to no signal.
(b) Monotonic; increasing from no signal.
(c) Maximum: increasing from and decreasing to no signal.
(d) Null: random noise.

Types a and b represent node-types of interest. Type a corresponds to genes which are activated (i.e., $x_{i,j} > 0$) early in the time-series before becoming de-activated (as we would expect for genes which are important for stem-like cell identity). Type b corresponds to genes which only become activated later in the time-series (as we would expect for genes which are important for the identity of mature cells, such as neurons). Types c and d make the simulated data more realistic, by mixing in nodes with other sorts of signals. Type c corresponds to genes which are active in the middle of the time-series only, and type d are null nodes (with random activation). We then generate the observations for node $i$ (corresponding to the response variable, see Section 2.2), according to the main model equation (equation (1)). Hence, the observations for node $i$ are a mixture of the observations for nodes $j \neq i$ (where the nodes $j$ are of the same type as node $i$), as follows. First, we select the type of node $i$ by choosing randomly between type a and type b. Then, we time-stretch the time-series for each node $j$ by a random amount so that each node is distinct from the others of the same type (described in more detail below). Next, we take the average of the values for nodes of the randomly-chosen type of interest (a or b), by setting $b_{t,j} = 1/p'$ (for $j$ the same type as $i$ and for $t$ such that $x_{i,j} > 0$) or $b_{t,j} = 0$ (otherwise), where $p'$ is the number of nodes of the same type as $i$. Finally, we add Normal noise. The procedure is summarized in Figure S10 in Supplement C.

We generate each time-series with 8 time-points, and for each of the four characteristic types of nodes a-d, we generate 10 time-series (i.e., $p' = 10$). For each node $j$ of type a, we time-stretch the characteristic time-series so that we always maintain $x_{i,j} = 1$ for $t = 0$, with $x_{i,t}$ then decreasing to 0 at a time $t'$ randomly sampled from $t' \sim \mathcal{U}(3, 8)$. For each node $j$ of type b, we time-stretch the characteristic time-series so that $x_{i,j}$ increases from 0 starting at a time $t'$ randomly sampled from $t' \sim \mathcal{U}(0, 5)$, with $x_{i,t}$ always ending at 1 at $t = 8$. After combining the 10 time-series of either type a or type b, we add noise with standard deviations $\in \{0.1, 0.2, 0.3\}$ to generate the observations for
node \( i \). We then fit the model described in Section 2 to this simulated data-set, and estimate each \( b_{t,j} \) from the median of the corresponding posterior.

We expect the inferred network structure to be comprised of edges between nodes with observations from the same types of time-series, and so we infer an edge between nodes \( i \) and \( j \) if \( \hat{b}_{t,j} \neq 0 \), after thresholding the \( \hat{b}_{t,j} \) to remove trivally small values. We generate ROC (receiver-operator characteristic) curves as this threshold \( \phi \) is decreased to 0 from \( \max|\hat{b}_{t,j}| \) (for \( t \in \{1, \ldots, 8\} \) and all \( j \)), finding true positives when \( |\hat{b}_{t,j}| \geq \phi \) for \( b_{t,j} \neq 0 \), and finding false positives when \( |\hat{b}_{t,j}| \geq \phi \) for \( b_{t,j} = 0 \). We generate an average ROC curve over 100 repetitions of this procedure, and then calculate an AUC (area under curve) statistic for this average ROC curve.

We repeat this data-generation and model-fitting scheme for various values of sparsity parameter \( \lambda \), and with various amounts of noise added. Figure 6 shows the results of data generation and model fitting as described, with hyperparameter \( k = 20 \). Equivalent results with \( k = 10 \) and \( k = 50 \) are shown in Figures S11 and S12 in supplement C. A very high value of AUC=0.96 is obtained with the optimal choices of \( \lambda = 50 \) and \( k = 20 \): this represents near-perfect detection of network edges, with respect to the ground-truth in these simulated data.

### 4 Single-cell gene-expression data

We now move on to describe applying the model to real data. Times for each cell-sample were inferred according to the procedure described in supplement C. These inferred times were then used to fit the network model of Section 2. Figure 7 shows time-series of observed node values (gene-expression levels), for a selection of nodes/genes which are characteristic of stem cells, and of neurons (i.e.,
Figure 6: Accuracy of network inference, in the simulation study, with \( k = 20 \). Abbreviations: TP, true positives; FP, false positives.

mature cells). We expect stem cells to predominate at earlier times, and hence we expect decreasing time-series for genes which are characteristic of this type of cell. On the other hand, we expect mature cells such as neurons to predominate at later times, and so we expect to see increasing time-series for genes characteristic of this type of cell. The type of time-series that we would expect to observe for these cells also correspond to the simulated data presented in Figure 5a-b respectively. Equivalent results to Figure 7a-b, for larger sets of genes characteristic of these same cell types, appear in supplement C in Figures S13 and S14 respectively.

We fit the model of Section 2 to the main neuro-developmental data-set, with \( n = 1557 \) cell samples, and \( p = 22988 \) genes/nodes, reduced to \( p = 212 \) for each individual model fit by variable screening (for full details see supplement C). We fit the model using values of \( \lambda = 20 \) and \( k = 1 \): these values are chosen by grid-search stochastic EM (Figure S15 in supplement C). In fitting the model, we obtain posterior distributions for each model coefficient \( b_{t,j}(i) \), and then we use the posterior medians as posterior summary. Because we find that many of these posterior medians are close to but not exactly zero, we set \( \hat{b}_{t,j}(i) \) to zero in such cases by thresholding, thus inferring ‘no edge’ between nodes \( i \) and \( j \) when the posterior median is close to zero. Hence, if \( |\hat{b}_{t,j}(i)| > \phi \), where \( \phi \) is the threshold parameter, we would infer a network edge between nodes \( i \) and \( j \) at time \( t \) (for the
model fit around node $i$). However, if also $|\hat{b}_{t,j}(i)| > \phi$ (for the independent model fit around node $j$), we would independently infer an edge between nodes $i$ and $j$ at time $t$. Thus, some inconsistency may arise, due to these independent model fits around nodes $i$ and $j$. To deal with this, we use the ‘min_symmetrisation’ scheme of Kolar et al. [2010], as described in Section 2.2. Thus, we infer an edge between nodes $i$ and $j$ at time $t$, i.e., \( \hat{A}_{t,i,j} \neq 0 \), if and only if $|\hat{b}_{t,j}(i)| > \phi$ and $|\hat{b}_{t,i}(j)| > \phi$.

Figure 7: Time-series of mean expression for genes characteristic of: (a) stem-cells, (b) mature cells (neurons). We expect these time-series to correspond to the simulated data presented in Figure 5a-b respectively.

Figure 8: Inferred model coefficients $\hat{b}_{t,j}(i)$, for genes characteristic of: (a) stem-cells; (b) mature cells (neurons), for the same genes shown in Figure 7. Non-zero coefficients $\hat{b}_{t,j}(i)$ infer the local network structure around gene/node $i$. Coefficients which are zero for every time-point are not plotted.
Figure 9: Time-varying network structure inferred around the gene SATB2. This gene is characteristic of certain types of neuron, and hence we would expect network structure to appear at later times, when the cell type-specific gene regulatory program becomes activated.
Figure 10: Time-varying network structure of the fully-connected component (11133 nodes) of the inferred genomic network.
Figure 8 shows the values of the inferred model coefficients $\hat{b}_{t,j}(i)$, for the same genes as shown in Figure 7. These genes were chosen in an unbiased way by searching the biological sciences literature for relevant genes, and then analysing those which were present in this data-set after quality control (full details of the data-set and bioinformatic data pre-processing, including quality control, are given in Supplement C). Figure 8a shows, as would be expected for stem-cell genes, that important model coefficients $b_{t,j}$ tend to decrease in magnitude during the developmental trajectory as cells go from stem-cell to mature cell types (e.g., gene transcript MOXD1). Figure 8b then shows, as would be expected, that important model coefficients $\hat{b}_{t,j}(i)$ become non-zero (corresponding to network edges appearing) late in the developmental trajectory, when the cells become neurons and hence their characteristic gene regulatory program is activated (e.g., for SATB2). Equivalent results to Figure 8a-b, for the full sets of genes we found and analysed which are characteristic of these cell types, appear in Figures S16 and S17 respectively in supplement C. In these figures, we again see similar results: for genes which tend to be active in stem-cells, model coefficients $b_{t,j}$ tend to decrease in magnitude during the developmental trajectory as cells go from stem-cell to mature cell types (Figure S16), and vice-versa for genes which are important to mature cells such as neurons (Figure S17). The genes which do not follow this pattern in Figures S16 and S17 tend not to have the expected patterns of activation for genes relevant to stem cells and mature cells (e.g., gene transcripts FABP7 and SOX5 in S13 and S14). Plots of the inferred network structure around an example of a gene shown in Figure 8b, namely SATB2, are shown in Figure 9, after ‘min_symmetrisation’ (Section 2.2).

Finally, Figure 10 shows the fully connected component (11133 nodes) of the inferred network (again after min_symmetrisation; full inferred network is 22989 nodes).

5 Discussion

In this paper, we have proposed a new model to infer time-varying network structure. This model makes use of a novel prior structure we introduce here, which extends the Bayesian lasso to the time-varying case. The novel structure of this prior allows for effective modelling of time-varying network structure even in situations where there are very few time-points, as is typical in cell-biological (i.e., ‘omics) data. We also found that the model fitting and inference procedure we have proposed works well even in with large networks of over 20000 nodes.

We used simulated data to assess the ability of the proposed model to accurately infer time-varying network structure, and we showed that the model is effective in inferring time-varying genomic network structure, using single-cell gene-expression data. However, we note that genomic network structure which is inferred from only gene-expression data as we do here is not guaranteed to correspond to true gene regulatory patterns. To strengthen any belief that the inferred genomic network structure corresponds to true gene regulatory patterns rather than simply gene co-expression patterns, evidence from, for example, chromatin binding and epigenomic data could also be incorporated into the model [Novershtern et al., 2011]. We intend to incorporate such data as the next stage of the development of this model. Specifically, we will do this by allowing the sparsity parameter $\lambda$ to vary for each pair of nodes $i$ and $j$, depending on any prior evidence of a physical interaction between the product of gene $j$ with the DNA or surrounding chromatin of gene $i$.

Another characteristic of the single-cell gene-expression data analysed here is that the data are zero-inflated. This is data missing-not-at-random, because the dropout events which lead to the extra zeros in the data are more likely to occur when the true gene-expression level is low [Kharchenko et al., 2014]. As part of the next stage of the development of this model, we intend to explicitly include the dropout events in the model likelihood as other authors have done [Pierson and Yau, 2015]. We also note that existing time-inference methods for data such as those presented here are algorithmic, rather than model-based. Hence it is not easy to obtain uncertainties on the inferred times when using these methods. Thus, we would like to develop a model-based time-inference method which will provide such uncertainties, and then feed these uncertainties directly into the
time-varying network model we have proposed.

Understanding interactions between genes and their transcriptional regulators is a fundamental question in genomics, and network models are a natural way to represent and analyse groups of interactions between genes and their regulators. Biomedical science in the high-throughput genomic age has, for over a decade, been developing ever more innovative ways to collect increasingly vast quantities of data. However, the statistical techniques to represent, analyse and interpret such data still lag behind the means to generate them. In particular, there is currently a lack of good computational statistical methodology to represent and analyse changes in gene-regulatory interactions as cells are specified and change state - an issue we address with the model proposed here. Thus, the computational-statistical tools we are developing allow novel characterisation of genomic interactions in important settings, adding to knowledge of fundamental biological principles, and motivating further investigation by targeted experiments.

References

Roger P Alexander, Philip M Kim, Thierry Emonet, and Mark B Gerstein. Understanding modularity in molecular networks requires dynamics. Science signaling, 2(81):pe44, 2009.

David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological), 36(1):99–102, 1974.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. Biometrika, 97(2):465–480, 2010.

Ismaël Castillo, Johannes Schmidt-Hieber, Aad Van der Vaart, et al. Bayesian linear regression with sparse priors. The Annals of Statistics, 43(5):1986–2018, 2015.

Harry Crane et al. Dynamic random networks and their graph limits. The Annals of Applied Probability, 26(2):691–721, 2016.

Daniele Durante, David B Dunson, et al. Locally adaptive dynamic networks. The Annals of Applied Statistics, 10(4):2203–2232, 2016.

Alfredo Kalaitzis, John Lafferty, Neil Lawrence, and Shuheng Zhou. The bigraphical lasso. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 1229–1237, 2013.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. Nature methods, 11(7):740–742, 2014.

Mladen Kolar, Le Song, Amr Ahmed, and Eric P Xing. Estimating time-varying networks. The Annals of Applied Statistics, 4(1):94–123, 2010.

Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized regression, standard errors, and bayesian lassos. Bayesian Analysis, 5(2):369–411, 2010.

Steffen L Lauritzen. Graphical models, volume 17. Clarendon Press, 1996.

Sophie Lebre, Jennifer Becq, Frederic Devaux, Michael PH Stumpf, and Gaëlle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. BMC systems biology, 4(1):130, 2010.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(4):1119–1141, 2016.

Noa Novershtern, Aviv Regev, and Nir Friedman. Physical module networks: an integrative approach for reconstructing transcription regulation. Bioinformatics, 27(13):i177–i185, 2011.
Tomasz J Nowakowski, Aparna Bhaduri, Alex A Pollen, Beatriz Alvarado, Mohammed A Mostajo-Radji, Elizabeth Di Lullo, Maximilian Haeussler, Carmen Sandoval-Espinosa, Siyuan John Liu, Dmitry Velmeshev, Johain R Ounadjela, Joe Shuga, Xiaohui Wang, Daniel A Lim, Jay A West, Anne A Leyrat, W James Kent, and Arnold R Kriegstein. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*, 358(6368):1318–1323, 2017.

Konstantina Palla, Francois Caron, and Yee Whye Teh. Bayesian nonparametrics for sparse dynamic networks. *arXiv preprint arXiv:1607.01624*, 2016.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Marianna Pensky. Dynamic network models and graphon estimation. *arXiv preprint arXiv:1607.00673*, 2016.

Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(1):1–10, 2015.

Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10):886–891, 2011.

Sebastian Rosengren and Pieter Trapman. A dynamic erdos renyi graph model. *arXiv preprint arXiv:1604.05127*, 2016.

Purnamrita Sarkar and Deepayan Chakrabarti. Nonparametric link prediction in large scale dynamic networks. *Electronic Journal of Statistics*, 8(2):2022–2065, 2014.

Alexander Schaefer, Daniel S Margulies, Gabriele Lohmann, Krzysztof J Gorgolewski, Jonathan Smallwood, Stefan J Kiebel, and Arno Villringer. Dynamic network participation of functional connectivity hubs assessed by resting-state fMRI. *Frontiers in human neuroscience*, 8:195, 2014.

Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the national academy of sciences*, 113(36):9977–9982, 2016.

Kaito Shimamura, Masao Ueki, Shuichi Kawano, and Sadanori Konishi. Bayesian generalized fused lasso modeling via neg distribution. *arXiv preprint arXiv:1602.04910*, 2016.

Mario L Suva, Esther Rheinbay, Shawn M Gillespie, Anoop P Patel, Hiroaki Wakimoto, Samuel D Rabkin, Nicolo Riggi, Andrew S Chi, Daniel P Cahill, Brian V Nahed, et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell*, 157(3):580–594, 2014.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ryan Joseph Tibshirani, Jonathan E Taylor, Emmanuel Jean Candes, and Trevor Hastie. *The solution path of the generalized lasso*. Stanford University, 2011.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lemon, Kenneth J Livak, Tarjei S Mikkelson, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

SL van der Pas, J-B Salomond, Johannes Schmidt-Hieber, et al. Conditions for posterior contraction in the sparse normal means problem. *Electronic journal of statistics*, 10(1):976–1000, 2016.

Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 2015.

Kevin S Xu and Alfred O Hero III. Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 201–210. Springer, 2013.

Shihua Zhang, Junfei Zhao, and Xiang-Sun Zhang. Common community structure in time-varying networks. *Physical Review E*, 85(5):056110, 2012.
Availability of code and data

Code can be downloaded from www.ucl.ac.uk/statistics/people/thomas-bartlett as an R package, containing an efficient implementation of the model proposed in this paper. This package contains an R function which calls a C++ implementation of the Gibbs sampler described in Algorithm 1. The data used in this study were previously published by Nowakowski et al. [2017], and are publicly available from the NCBI database of genotypes and phenotypes (dbGaP), under accession number phs000989.v3

Acknowledgements

We are grateful to Tom Nowakowski, Alex Pollen, Aparna Bhaduri, and Aaron Diaz, for helpful discussions, insightful comments and useful advice throughout this project, and for providing early access to the data. The work of the first author was supported by the MRC grant MR/P014070/1. The work of the second and third author was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1 (Turing award numbers TU/B/000082 and TU/B/000081 respectively), and part of it was completed when the second author was a Senior Lecturer at University College London.
Supplement

Supplement A: derivation of equation (7)

Because $\theta_{t+1,j} \perp \theta_{t-1,j}, \theta_{t-2,j}, \ldots | \theta_{t,j}$ (equation (5)), the partial correlation of $\theta_{t+m,j}$ with $\theta_{t+1,j}$ will be zero for all $|m-l| > 1$. Hence, all entries of the precision matrix $\Sigma_j^{-1}$ will be zero except the diagonal and the elements immediately adjacent to it (i.e., the sub- and super-diagonals). Therefore,

$$\Sigma_j^{-1} \Sigma_j = I \implies$$

$$\begin{bmatrix}
1 & \rho_j & \rho_j^2 & \cdots & \rho_j^T \\
\rho_j & 1 & \rho_j & \cdots & \rho_j^{T-1} \\
\rho_j^2 & \rho_j & 1 & \cdots & \rho_j^{T-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_j^T & \rho_j^{T-1} & \rho_j^{T-2} & \cdots & 1
\end{bmatrix}
\begin{bmatrix}
\eta_{1,1} & \eta_{1,2} & 0 & \cdots & 0 \\
\eta_{2,1} & \eta_{2,2} & \eta_{2,3} & \cdots & 0 \\
0 & \eta_{3,2} & \eta_{3,3} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \eta_{T,T}
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1
\end{bmatrix}$$

and hence,

$$\eta_{1,1} + \rho_j \eta_{2,1} = 1,$$

$$\rho_j \eta_{1,1} + \eta_{2,1} = 0,$$

$$\implies \eta_{1,1} = \frac{1}{1 - \rho_j^2}$$

and $\eta_{2,1} = -\frac{\rho_j}{1 - \rho_j^2} = \eta_{1,2},$ \hspace{1cm} (15)

and by symmetry (or equivalent argument), also

$$\eta_{1,1} = \frac{1}{1 - \rho_j^2}$$

and $\eta_{T,T-1} = -\frac{\rho_j}{1 - \rho_j^2} = \eta_{T-1,T}.$

Then,

$$\eta_{1,2} + \rho_j \eta_{2,2} + \rho_j^2 \eta_{3,2} = 0 \implies \rho_j \eta_{2,2} + \rho_j^2 \eta_{3,2} = \frac{\rho_j}{1 - \rho_j^2}$$

and $\rho_j^2 \eta_{1,2} + \rho_j \eta_{2,2} + \eta_{3,2} = 0 \implies \rho_j \eta_{2,2} + \eta_{3,2} = \frac{\rho_j^3}{1 - \rho_j^2},$

and so subtracting the second of these equations from the first leads to

$$\left(\rho_j^2 - 1\right) \eta_{3,2} = \frac{\rho_j \left(1 - \rho_j^2\right)}{1 - \rho_j^2}.$$
and so

$$\eta_{3,2} = -\frac{\rho_j}{1 - \rho_j^2} = \eta_{2,3}, \quad (16)$$

and therefore also

$$\rho_j \eta_{2,2} - \frac{\rho_j}{1 - \rho_j^2} = \frac{[\rho_j]^3}{1 - \rho_j^2}$$

and hence

$$\eta_{2,2} = \frac{1 + \rho_j^2}{1 - \rho_j^2}.$$  

Because the sub- and super-diagonal terms found in equation (16) and (15) are the same, the derivations for the other terms $\eta_{k,t+1} = \eta_{k+1,t}$ and $\eta_{k,t}, \ t = 3, ..., T - 1$ will be identical and therefore we have

$$\left( \Sigma_j^{-1} \right)_{t,t'} = \begin{cases} 1/(1 - \rho_j^2), & \text{if } t' = t = 1 \text{ or } t' = t = T \\ (1 + \rho_j^2)/(1 - \rho_j^2), & \text{if } t' = t > 1 \text{ and } t' = t < T \\ -\rho_j/(1 - \rho_j^2), & \text{if } t' = t + 1 \text{ or } t' = t - 1 \\ 0, & \text{otherwise.} \end{cases}$$

**Supplement B: derivations of the steps in Algorithm 1**

Starting with equation (9),

$$P(y, B, \rho, s, a, \tau | X, \lambda, k) = \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} \sqrt{\frac{\tau}{2\pi}} e^{-\tau(y_t^{(k)} - b_{c_2} [x_t^{(k)}]^{\top} - a)^2 / 2} \right\} \frac{1}{\sqrt{2\pi}} e^{-\{\tau + a^2 / 2\}} \prod_{j=1}^{p-1} \left\{ \frac{k}{e^k - 1} e^{k\rho_j} \lambda^2 / 2 e^{-\lambda^2 / (2\nu_j)} \frac{\nu_j^{-3/2}}{(2\pi)^{T/2}} |\Sigma_j|^{1/2} e^{-b_{\top} j \Sigma_j^{-1} b_j} / 2 \right\},$$

we can write down the following expressions for conditional posteriors, for a Gibbs sampler:

$$P(b_{\cdot, j} | y, X, ...) \propto \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau(y_t^{(k)} - b_{c_2} [x_t^{(k)}]^{\top} - a)^2 / 2} \right\} e^{-b_{\top} j \Sigma_j^{-1} b_j} / 2 = g_{b_j}(b_{\cdot, j}), \quad (17)$$

$$P(a | y, X, ...) \propto \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau(y_t^{(k)} - b_{c_2} [x_t^{(k)}]^{\top} - a)^2 / 2} \right\} e^{-a^2 / 2}, \quad (18)$$

$$P(\tau | y, X, ...) \propto \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} \sqrt{\tau} e^{-\tau(y_t^{(k)} - b_{c_2} [x_t^{(k)}]^{\top} - a)^2 / 2} \right\} e^{-\tau}, \quad (19)$$

$$P(\nu_j | y, X, ...) \propto e^{-\lambda^2 / (2\nu_j)} \nu_j^{-3/2} e^{-b_{\top} j \Sigma_j^{-1} b_j} / 2,$$

and
\[ P(\rho_j | y, X, \ldots) \propto e^{k \rho_j} \frac{\nu_j}{|\Sigma_j|^{1/2}} e^{-b_{r,j} \nu_j \Sigma_j^{-1} b_{r,j} / 2} = g_{\rho_j}(\rho_j). \]  

Equation (17) can be written as:

\[
g_{b_j}(b_{r,j}) = \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau \left( y_t^{(k)} - b_{r,j} - x_t^{(k)} \right)^\top / 2} \right\} e^{-b_{r,j} \nu_j \Sigma_j^{-1} b_{r,j} / 2}
\]

\[
= \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau \left( b_{r,j} x_t^{(k)} - y_t^{(k)} + b_{r,j} y_t^{(k)} \right)^\top / 2} \right\} e^{-b_{r,j} \nu_j \Sigma_j^{-1} b_{r,j} / 2},
\]

and equation (21) is recognized as the product of several Normal density functions. It is well known that the product of Normal density functions (of the same variable) is another Normal density function (e.g., a Normal likelihood with a Normal prior gives a Normal posterior). Specifically, if we combine \( n \) univariate Normal density functions with means \( \mu_1, \mu_2, \ldots, \mu_n \) and variances \( \sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2 \) then we get a univariate Normal with mean and variance specified according to:

\[
\frac{1}{\sigma_{\text{combined}}^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}
\]

and

\[
\frac{\mu_{\text{combined}}}{\sigma_{\text{combined}}^2} = \sum_{i=1}^{n} \frac{\mu_i}{\sigma_i^2},
\]

and more generally if we multiply \( n \) multivariate Normal density functions with mean vectors \( \mu_1, \mu_2, \ldots, \mu_n \) and covariance matrices \( \Sigma_1, \Sigma_2, \ldots, \Sigma_n \), then we get a Normal density function with mean vector and covariance matrix given by

\[
\Sigma_{\text{combined}}^{-1} = \sum_{i=1}^{n} \Sigma_i^{-1}
\]

and

\[
\Sigma_{\text{combined}}^{-1} \mu_{\text{combined}} = \sum_{i=1}^{n} \Sigma_i^{-1} \mu_i.
\]

The inner-most product in equation (21) can be written as

\[
\prod_{k \in \zeta(t)} e^{-\tau \left( y_t^{(k)} - b_{r,j} - x_t^{(k)} \right)^\top / 2} = \prod_{k \in \zeta(t)} e^{-\tau \left( x_t^{(k)} - \frac{\{ y_t^{(k)} - b_{r,j} y_t^{(k)} \}^\top}{\sum_{k \in \zeta(t)} |x_t^{(k)}|^2} \right)^2 / 2},
\]

and so using the logic of equations (22) and (23) to combine Normal distributions of \( b_{r,j} \),

\[
\prod_{k \in \zeta(t)} e^{-\tau \left( x_t^{(k)} \right)^2 \left( b_{r,j} - \frac{\{ y_t^{(k)} - b_{r,j} y_t^{(k)} \}^\top}{\sum_{k \in \zeta(t)} |x_t^{(k)}|^2} \right)^2 / 2}
\]

\[
\propto e^{-\tau \left( \sum_{k \in \zeta(t)} |x_t^{(k)}|^2 \right) \left( b_{r,j} - \frac{\{ y_t^{(k)} - b_{r,j} y_t^{(k)} \}^\top}{\sum_{k \in \zeta(t)} |x_t^{(k)}|^2} \right)^2 / 2},
\]

where ‘proportional to’ is with respect to finding an un-normalised distribution for \( b_{r,j} \). Hence (also referring back to equation (21)),

\[
\prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau \left( y_t^{(k)} - b_{r,j} - x_t^{(k)} \right)^\top / 2}
\]

\[
\propto \prod_{t=1}^{T} e^{-\tau \left( \sum_{k \in \zeta(t)} |x_t^{(k)}|^2 \right) \left( b_{r,j} - \frac{\{ y_t^{(k)} - b_{r,j} y_t^{(k)} \}^\top}{\sum_{k \in \zeta(t)} |x_t^{(k)}|^2} \right)^2 / 2}
\]

\[
\propto e^{-(b_{r,j} - m_j)^\top V_j^{-1} (b_{r,j} - m_j) / 2}
\]
(because the product of independent univariate Normal density function of different variables is proportional to a multivariate Normal density function), where the \( t \)th element of \( \mathbf{m}_j \) is

\[
m_{t,j} = \sum_{k \in \zeta(t)} x_{t,j}^{(k)} \left\{ y_t^{(k)} - \mathbf{b}_{t,\zeta} \cdot [x_{t,j}^{(k)}]^{\top} - a \right\} / \sum_{k \in \zeta(t)} [x_{t,j}^{(k)}]^2,
\]

where \( \mathbf{b}_{t,\zeta} \) and \( x_{t,j}^{(k)} \) represent \( \mathbf{b}_t \) and \( x_{t,j}^{(k)} \) without the \( j \)th elements, respectively, and \( \mathbf{V}_j \) is a diagonal matrix, with the \( t \)th diagonal element equal to \( 1 / \left\{ \tau \sum_{k \in \zeta(t)} [x_{t,j}^{(k)}]^2 \right\} \). Hence, using the logic of equations (24) and (25), and referring also to equation (17):

\[
P(\mathbf{b}_{i,j} | \mathbf{y}, \mathbf{X}, ...) \propto g_{\mathbf{b}_i}(\mathbf{b}_{i,j}) = \prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau \left( y_t^{(k)} - \mathbf{b}_{t,\zeta} \cdot [x_{t,j}^{(k)}]^{\top} - a \right)^2 / 2} \}
\]

\[
\propto e^{-\left( \mathbf{b}_{i,j} - \mathbf{m}_j \right)^{\top} \mathbf{\Sigma}_j^{-1} \left( \mathbf{b}_{i,j} - \mathbf{m}_j \right) / 2} \propto f_{\mathbf{N}}(\mathbf{b}_{i,j} | \tilde{\mathbf{m}}_j, \mathbf{\Sigma}_j) = \tilde{g}_{\mathbf{b}_i}(\mathbf{b}_{i,j}),
\]

where \( \mathbf{\Sigma}_j^{-1} = \nu_j \mathbf{\Sigma}_j^{-1} + \mathbf{V}_j^{-1} \), and \( \tilde{\mathbf{m}}_j = \mathbf{\Sigma}_j \mathbf{V}_j^{-1} \mathbf{m}_j \), and \( f_{\mathbf{N}}(\cdot | \mathbf{\mu}, \mathbf{\Sigma}) \) is the multivariate Normal density.

Referring again to equations (22) and (23), equation (18) can be re-written as

\[
P(\mathbf{a} | \mathbf{y}, \mathbf{X}, ...) \propto e^{-\mathbf{a}^{\top} / 2} \left\{ \prod_{t=1}^{T} \prod_{k \in \zeta(t)} e^{-\tau \left( a - \left( y_t^{(k)} - \mathbf{b}_{t,\zeta} \cdot [x_{t,j}^{(k)}]^{\top} \right) \right)^2 / 2} \right\}
\]

\[
\propto f_{\mathbf{N}}(\mathbf{a} | \mu_a, \sigma_a) = g_{\mathbf{a}}(\mathbf{a}),
\]

where \( \sigma_a^{-2} = 1 + n \tau \) and \( \mu_a = \frac{\sigma_a^2 \sum_{t=1}^{T} y_t^{(k)} - \mathbf{b}_{t,\zeta} \cdot [x_{t,j}^{(k)}]^{\top}}{\sum_{k \in \zeta(t)} [x_{t,j}^{(k)}]^2} \).

Equation (19) can be written as:

\[
P(\tau | \mathbf{y}, \mathbf{X}, ...) \propto \tau^{n/2} e^{-\tau \left( 1 + \sum_{t=1}^{T} \sum_{k \in \zeta(t)} \left( y_t^{(k)} - \mathbf{b}_{t,\zeta} \cdot [x_{t,j}^{(k)}]^{\top} \right) \right)^2 / 2} \}
\]

\[
\propto f_{\gamma}(\tau | \kappa, \theta) = g_{\gamma}(\tau),
\]

where \( f_{\gamma} \) is the density of the gamma distribution with \( \kappa = 1 + \frac{n}{2} \) and

\[
\theta = 1 / \left\{ 1 + \sum_{t=1}^{T} \sum_{k \in \zeta(t)} \left( y_t^{(k)} - \mathbf{b}_{t,\zeta} \cdot [x_{t,j}^{(k)}]^{\top} - a \right)^2 / 2 \right\}.
\]

Recalling equation (20),

\[
P(\nu_j | \mathbf{y}, \mathbf{X}, ...) \propto e^{-\lambda^2 / (2 \nu_j)} \nu_j^{-3/2} e^{-\mathbf{b}_{t,\zeta}^{\top} \mathbf{\Sigma}_j^{-1} \mathbf{b}_{t,\zeta} / \nu_j / 2},
\]

and also recalling the inverse Normal density

\[
f_{IG}(x) = \frac{\lambda}{2 \pi} x^{-3/2} e^{-\lambda (x - \mu)^2 / (2 \mu^2 x)}, \quad x > 0,
\]

\[
\propto x^{-3/2} e^{-\lambda x / (2 \mu^2)} e^{-\lambda / (2 \mu x)}.
\]

4
we can write

\[ P(\nu_j | y, X, ...) \propto f_{IG}(\nu_j | \mu, \lambda) = g_{ij}(\nu_j), \]  

where \( f_{IG} \) is the density of the inverse Normal distribution (equation (26)), with parameters \( \lambda = \lambda^2 \), and \( \mu = \lambda \sqrt{\sum_{j} b_{ij}^{-1}} \).

**Supplement C: data pre-processing and time-inference**

The data used in this study were published previously [Nowakowski et al., 2017], and are publicly available from the NCBI database of genotypes and phenotypes (dbGaP), under accession number phs000989.v3. The downloaded data were normalised to give transcript read counts per million reads (CPM), hereafter referred to simply as ‘read counts’. For quality control, cells with non-zero read counts for fewer than 1000 transcripts were removed, and transcripts with non-zero read counts for fewer than 30 cells were removed. All subsequent analyses were carried out on the log(read counts + 1) for the 22989 transcripts and 4691 cells which passed quality control.

We also obtained classifications for the cells from the lab that generated the data. We visualised these classifications as follows. First, we carried out a sparse singular value decomposition: we projected the data for the 4691 cells into a reduced dimensional space corresponding to the top 42 left singular vectors. The top 42 left singular vectors were used, because the top 42 singular values were deemed to be significant, under comparison with randomised versions of the same data. Then, we used t-SNE (t-distributed stochastic neighbour embedding) [Maaten and Hinton, 2008] to further reduce the dimension of the data to two dimensions. The cells are plotted in this two dimensional space in Figures S1 and S2. The cells are clearly partitioned in this visualisation according to the classifications provided by the lab which generated the data.

As cells transition from stem-like cells (called radial glia in Figures S1 and S2) to mature cell types such as neurons, they pass through various intermediate cell types, such as intermediate progenitor cells (IPCs). Cells with similar phenotypes (i.e., physical characteristics) are expected to have similar gene-expression profiles. Therefore, cells of similar types are expected to be close together in the lower dimensional projection of Figures S1 and S2. Hence, as cells transition from stem cells to mature cells, we can expect them to pass through adjacent regions in the lower dimensional projection in Figures S1 and S2, as part of their ‘developmental trajectory’. Progression along this developmental trajectory can be quantified in terms of ‘pseudo developmental time’ (PD-time). We define 5 points in PD-time, corresponding to: \( t = 1 \), radial glia; \( t = 2 \), dividing radial glia; \( t = 3 \), IPCs (intermediate progenitor cells), \( t = 4 \), newborn neurons, \( t = 5 \), upper layer PFC (pre-frontal cortex) neurons.

We use these 5 inferred PD-time points as the times of the samples to feed into the proposed time-varying network model, with 1557 corresponding cell samples. To fit the model locally around each node whilst allowing all other 22988 other nodes to be potential predictors would lead to an unnecessarily high computational cost. Instead, we identify the ‘important’ set of genes with a lower computational burden, as follows. We adapt the variable screening method of Wang and Leng [2015], by finding the mean of their high-dimensional ordinary least-squares projection (HOLP) across each of the time-points. Then, for each gene we rank the 22988 other genes according to this mean HOLP, and select the \( n/\log(n) = 212 \) top genes according to this ranking. These 212 genes are then used as the set of possible predictors which we fit the model to. Hence, the local network structure around each node/gene is inferred from this choice of 212 other nodes/genes.
Figure S1: Low-dimensional projection of the data, with previously-obtained classifications.
Figure S2: Low-dimensional projection of the data, with previously-obtained classifications.
Figure S3: Heatmaps of the bivariate log-densities of prior samples for \( b_{i,j} = [b_{1,j}, b_{2,j}]^\top \), using the Laplace-NEG prior of Shimamura et al. [2016], for various values of \( \lambda \) and \( \lambda^\dagger \), with \( \gamma = 0.2 \).

Figure S4: Heatmaps of the bivariate log-densities of prior samples for \( b_{i,j} = [b_{1,j}, b_{2,j}]^\top \), using the Laplace-NEG prior of Shimamura et al. [2016], for various values of \( \lambda \) and \( \lambda^\dagger \), with \( \gamma = 0.5 \).

Figure S5: Heatmaps of the bivariate log-densities of prior samples for \( b_{i,j} = [b_{1,j}, b_{2,j}]^\top \), using the Laplace-NEG prior of Shimamura et al. [2016], for various values of \( \lambda \) and \( \lambda^\dagger \), with \( \gamma = 1 \).
Figure S6: Marginal densities of prior samples for $b_{1,j}$, from the proposed novel decoupled-sparsity prior, for various values of $\lambda$ and $k$.

Figure S7: Marginal densities of prior samples for $b_{2,j}$, from the proposed novel decoupled-sparsity prior, for various values of $\lambda$ and $k$.

Figure S8: Marginal densities of prior samples for $b_{1,j}$, using the Laplace-NEG prior of Shimamura et al. [2016], for various values of $\lambda$ and $\lambda^\dagger$, with $\gamma = 0.5$.

Figure S9: Marginal densities of prior samples for $b_{2,j}$, using the Laplace-NEG prior of Shimamura et al. [2016], for various values of $\lambda$ and $\lambda^\dagger$, with $\gamma = 0.5$. 

9
1) Select type a or b as 'response' node
Type a: decreasing
Type b: increasing
Type c: maximum
Type d: null
2) Time-stretch each instance of type a-c by a random amount
3) Combine instances of chosen type a or b according to equation 1 and add noise

Response variable $y_t$
Predictor variables $x_t$

Figure S10: Overview of procedure for generating the simulated data.
Figure S11: Accuracy of network inference, in the simulation study, with $k = 10$. Abbreviations: TP, true positives; FP, false positives.
Figure S12: Accuracy of network inference, in the simulation study, with $k = 50$. Abbreviations: TP, true positives; FP, false positives.
Figure S13: Time-series of mean expression for genes characteristic of stem-cells. This is as Figure 7a, with an expanded set of genes.

Figure S14: Time-series of mean expression for genes characteristic of neurons. This is as Figure 7b, with an expanded set of genes.
Figure S15: Model log-likelihood values for various values of $\lambda$ and $k$, for grid-search stochastic expectation-maximization (EM) over all model fits, for the single-cell gene-expression data.
Figure S16: Inferred model coefficients $\hat{b}_{t,j}(i)$, for genes characteristic of stem-cells. Non-zero coefficients $\hat{b}_{t,j}(i)$ infer the local network structure around gene/node $i$. Coefficients which are zero for every time-point are not plotted. This is as Figure 8a, with an expanded set of genes.

Figure S17: Inferred model coefficients $\hat{b}_{t,j}(i)$, for genes characteristic of neurons. Non-zero coefficients $\hat{b}_{t,j}(i)$ infer the local network structure around gene/node $i$. Coefficients which are zero for every time-point are not plotted. This is as Figure 8b, with an expanded set of genes.