

Sparse time series chain graphical models for reconstructing genetic networks

FENTAW ABEGAZ*, ERNST WIT

*Johann Bernoulli Institute of Mathematics and Computer Science, University of Groningen,
Nijenborgh 9, The Netherlands
f.abegaz.yazew@rug.nl*

SUMMARY

We propose a sparse high-dimensional time series chain graphical model for reconstructing genetic networks from gene expression data parametrized by a precision matrix and autoregressive coefficient matrix. We consider the time steps as blocks or chains. The proposed approach explores patterns of contemporaneous and dynamic interactions by efficiently combining Gaussian graphical models and Bayesian dynamic networks. We use penalized likelihood inference with a smoothly clipped absolute deviation penalty to explore the relationships among the observed time course gene expressions. The method is illustrated on simulated data and on real data examples from *Arabidopsis thaliana* and mammary gland time course microarray gene expressions.

Keywords: Chain graphical mode; Dynamic network; Gene expression; High-dimensional data; L_1 penalty; Model selection; Penalized likelihood; SCAD penalty; Vector autoregressive model.

1. INTRODUCTION

Recent advances in DNA microarray technology allow experimental measurement of expression levels of many genes simultaneously over time. Measuring changes over time gives a deeper understanding of the various mechanisms by which a cell controls and regulates the transcription of its genes. Reverse engineering genetic networks from time series data is an important tool to aid answering such biological questions.

Various statistical modelling approaches have been proposed for the analysis of time course gene expression data. An overview of methods for modelling and inference on gene regulatory network from time series data is given in [Sima and others \(2009\)](#). These methods are categorized into those that focus on the structure of a network, such as relevance networks and Bayesian networks, and those that focus on the structure and dynamics of a network, such as Boolean networks, probabilistic Boolean networks, Markov models, state space models, dynamic Bayesian networks and ordinary differential equations. On the other hand, [Aburatani and others \(2006\)](#) considered graphical chain models for inferring regulatory system networks from gene expression profiles of a cell cycle in yeast. Recently, chain graphical models are considered in time series setting by [Dahlhaus and Eichler \(2003\)](#) and [Gao and Tian \(2010\)](#).

*To whom correspondence should be addressed.

Time series chain graphical models (TSCGM) are modeling tools to analyze repeated multivariate time series observations. These models can be applied to explore patterns of interactions in time course gene expression data. The TSCGM provides two types of interactions among genes: dynamic or delayed interactions and contemporaneous interactions. As a result a TSCGM explores patterns of interactions resulting from both Gaussian graphical models and dynamic Bayesian networks. Recently, interest has focussed on sparse estimation of Gaussian graphical models (Meinshausen and Bühlmann, 2006; Friedman and others, 2008). This approach is essential in sparse data problems whereby it is known that the underlying network is sparse.

In this paper, we consider time series chain graphical modeling to identify genetic networks on *Arabidopsis thaliana* and mammary glands time course microarray gene expression datasets. Microarray experiment results in 10 000s of genes; however, with small samples, it may not be meaningful to infer networks for such a large number of genes. Simulation studies suggest that even with a moderate amount of noise the number of false positives and false negatives are huge and the inferred network is meaningless. We focus on a reduced number of genes extracted biologically (i.e. a subset of genes that are known or thought to be related a priori) or computationally (i.e. a subset of genes selected to have significant effect or as representatives, for example, from cluster analysis). *Arabidopsis thaliana* is a model plant used to study circadian regulation in plants. Nine genes are known to be involved to circadian regulatory network in *A. thaliana*. Based on microarray gene expression time series, Grzegorzczuk and others (2008), Grzegorzczuk and Husmeier (2011b), and Jia and Huan (2009) applied nonhomogeneous dynamic Bayesian network to construct the regulatory network of the nine circadian genes. On the other hand, data on the mammary gland experiment obtained from Stein and others (2004) consist of 12 488 probe sets representing ~8600 genes. We identified 30 genes that yield the best separation between the developmental stages of mice using cluster analysis on the gene profiles. Each gene can be thought of as an equivalence class of genes. Further discussion of these data is given in Section 3.

In this paper, our aim is to develop an efficient method that explores lag zero and lag one patterns of gene–gene interactions based on time course gene expressions. To account for highly organized structure of genes expression profiles, the proposed method aims at providing sparse estimates.

2. METHODS

2.1 Time series chain graphical models

Suppose that we have n replications of a T time point longitudinal microarray study across p genes. The data, then, can be summarized as an $n \times p \times T$ array $X = (X_1, \dots, X_n)'$ whose i th submatrix X_i has columns such that $X_{i,t} = (X_{i1t}, \dots, X_{ipt})'$ which correspond to the expression levels of p genes measured at time t . That is, X_{ijt} is the j th gene expression level at time t for the i th replicate.

The time series chain graph $G = (V, E)$ for gene networks consists of a set of vertices V , representing the gene expression profiles, and a set of edges or links E , representing contemporaneous (undirected) or dynamic (directed) interactions between pairs of genes. E is a set of ordered pairs (A, B) , $A, B \subseteq V$. The time series chain graph is based on the partitioning of V into a number of blocks $V = V_1 \cup \dots \cup V_T$. The blocks are the ordered time steps and each block consists of all genes under consideration at that time point.

In the chain graphical model, interpretation of links differs between within and across time steps. Links within a time step, V_t , are undirected as in an ordinary graphical models and reflect the instantaneous interactions among genes. Links across time steps are directed as in a Bayesian network that points from the previous, V_{t-1} , to the current time step, V_t , and represents dynamic or delayed interactions between genes across consecutive time steps.

Within a chain graphical model, the time course gene–gene interactions evolve according to Markovian dynamics. Specifically, the vector of gene expressions at time t relates to only that of at time $t - 1$, although

extension to a Markov property of order $d \geq 2$ is straightforward. Let X_t be a random variable associated to the nodes V_t . According to the first-order Markov property, the joint probability density of X_1, \dots, X_T can be decomposed as

$$f(X_1, \dots, X_T) = f(X_1)f(X_2 | X_1) \times \dots \times f(X_T | X_{T-1}). \quad (2.1)$$

We focus only on the conditional distributions in (2.1) and ignore the initial term $f(X_1)$. The conditional likelihood for the TSCGM is given in the form

$$f_c(X_1, \dots, X_T) = \prod_{t=2}^T f(X_t | X_{t-1}).$$

Furthermore, we assume a stable dynamic network structure for the conditional distribution $f(X_t | X_{t-1})$ which can be approximated via a multivariate normal

$$X_t | X_{t-1} \sim N(\Gamma X_{t-1}, \Omega). \quad (2.2)$$

This can be expressed similar to VAR(1), vector autoregressive process of order 1, and given by

$$X_t = \Gamma X_{t-1} + \epsilon_t, \quad (2.3)$$

where $\epsilon_t \sim N(0, \Omega)$.

Following [Dahlhaus and Eichler \(2003\)](#) and [Gao and Tian \(2010\)](#), the parameter elements in the matrices Γ and in the inverse of Ω represent directed and undirected links in terms of graphical modeling, respectively. Given the time series chain graph $G = (V, E)$, where E includes directed and undirected edges, a directed edge between gene a and b at consecutive time steps relate to an element in the matrix Γ as follows:

$$(a, b) \in V_{t-1} \times V_t \Leftrightarrow \Gamma_{ab} \neq 0,$$

i.e. nonzero element in Γ correspond to a directed edge in the graph and this in turn reflects the dynamic pattern of interaction. Similarly, undirected edges between any nodes a and b at time t relate to the Gaussian graphical model for the errors ϵ_t , implying contemporaneous interaction among the genes after adjusting for the effects of past and present gene expression profiles. Given Ω and the corresponding precision matrix $\Theta = \Omega^{-1}$ undirected edges relate to nonzero elements in the precision matrix:

$$(a, b) \in V_t \times V_t \Leftrightarrow \Theta_{ab} \neq 0.$$

2.2 Relationship with other models

The above modeling approach is related to state space models. [Rangel and others \(2004\)](#) considered linear state space models for the analysis of time course gene expression data. The linear state space model takes the form:

$$Z_t = AZ_{t-1} + BX_{t-1} + w_t \quad \text{and} \quad X_t = CZ_t + DX_{t-1} + v_t, \quad (2.4)$$

where, X_t denotes the observed gene expression profiles at time t and Z_t the unobserved hidden factors and w_t and v_t independent noise terms. Further simplification to express the observed data at time t as a

function of only the observed data at time $t - 1$ results similar expression as in (2.3); that is,

$$X_t = (CB + D) X_{t-1} + r_t, \quad (2.5)$$

where $r_t = v_t + Cw_t + CAZ_{t-1}$ includes all contributions from noise and previous states. Thus, the first-order interaction between gene $X_{a,t-1}$ and gene $X_{b,t}$ can be characterized by the element $[CB + D]_{ab}$ of the matrix. Inference on the chain graph models integrate out the latent component. In the chain graphical model, the matrix of autoregression coefficients Γ replaces the role of $CB + D$ and ϵ_t replaces the term r_t and introduce instantaneous gene–gene interactions through the nonzero elements of the precision matrix Θ .

The proposed approach has many advantages. First, it provides a way to simultaneously estimate Γ and Ω^{-1} . Secondly, it avoids computational burden involved due to the use of EM algorithm or Kalman filter for the estimation of hidden factors. Finally, it can be conveniently extended to handle high-dimensional gene expression data to obtain sparse solutions for Γ and Θ , which is the main focus of this paper presented in the next section.

2.3 Penalized likelihood inference

Time course genomic data typically consist of hundreds or thousands of genes measured on a comparatively small number of replications of microarray experiments across a few time steps. The model formulation in (2.3) is in a standard vector autoregressive form with correlated errors and estimation approach for high-dimensional time course genomic data is challenging. In this paper, we propose penalized maximum likelihood estimation methods for the analysis of the high-dimensional time course gene expression data. The proposed approach provides sparse estimates of the autoregressive coefficient matrix Γ and the precision matrix Θ in (2.3), which are used to reconstruct the genetic network.

Under the Gaussian assumption described in (2.2), the conditional density of the t th observation is given by

$$f_c(X_t | X_{t-1}; \Gamma, \Theta) = (2\pi)^{p/2} \det(\Theta)^{1/2} \exp[-\frac{1}{2}(X_t - \Gamma X_{t-1})' \Theta (X_t - \Gamma X_{t-1})].$$

Then the conditional log-likelihood for n replicates each at T time steps becomes

$$\begin{aligned} \ell(\Gamma, \Theta) &= \sum_{i=1}^n \sum_{t=1}^T \log f_c(X_{it} | X_{i,t-1}; \Gamma, \Theta) \\ &= -\frac{npT}{2} \log(2\pi) + \frac{nT}{2} \log \det(\Theta) - \frac{nT}{2} \text{tr}(S_\Gamma \Theta), \end{aligned} \quad (2.6)$$

where

$$S_\Gamma = (1/nT) \sum_{i=1}^n \sum_{t=1}^T (X_{it} - \Gamma X_{i,t-1})(X_{it} - \Gamma X_{i,t-1})' = C_x - C_{xx-} \Gamma' - \Gamma C'_{xx-} + \Gamma C_{x-} \Gamma',$$

with

$$C_x = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T X_{it} X'_{it}, \quad C_{xx-} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T X_{it} X'_{i,t-1}, \quad \text{and} \quad C_{x-} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T X_{i,t-1} X'_{i,t-1}.$$

Under the penalized likelihood framework, we define two penalty functions $P_\lambda(\cdot)$ and $P_\rho(\cdot)$ corresponding to Θ and Γ . The objective function for optimization based on (2.6) is defined as

$$\ell_{\text{pen}}(\Gamma, \Theta) = \log \det(\Theta) - \text{tr}(S_\Gamma \Theta) - \sum_{i \neq j}^p P_\lambda(|\theta_{ij}|) - \sum_{i,j}^p P_\rho(|\gamma_{ij}|), \quad (2.7)$$

where θ_{ij} and γ_{ij} are the (i, j) -elements of the matrix Θ and Γ and λ and ρ are the corresponding tuning parameters. Various penalty functions have been proposed in the literature. We consider L_1 and smoothly clipped absolute deviation (SCAD) penalty functions.

The L_1 penalty is convex and given by

$$P_\lambda(\theta) = \lambda|\theta|. \quad (2.8)$$

This leads to a desirable convex optimization problem when the log-likelihood is convex and computation is done efficiently using the LARS algorithm (Efron and others, 2004).

The SCAD penalty proposed in Fan and Li (2001) is given by

$$P_{\lambda,a}(\theta) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda, \\ -\frac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\theta| \leq a\lambda, \\ \frac{(a+1)^2\lambda^2}{2} & \text{if } |\theta| > a\lambda. \end{cases}$$

This is symmetric and a quadratic spline on $[0, \infty]$ but nonconvex. The first derivative is given by

$$P'_{\lambda,a}(\theta) = \lambda \left\{ \mathcal{I}(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} \mathcal{I}(|\theta| > \lambda) \right\}, \quad (2.9)$$

for $\theta \geq 0$, where $\lambda > 0$ and $a > 2$ are two tuning parameters. In the numerical studies in this paper, we use $a = 3.7$ as recommended by Fan and Li (2001). In the case of SCAD penalty, Zou and Li (2008) and Fan and others (2009) discussed the estimation procedure by using a sequence of L_1 penalized likelihoods via a local linear approximation. Later in this section we present a reformulation of this procedure with the fast and efficient graphical LASSO algorithm.

2.3.1 L_1 penalized inference. Under L_1 penalized likelihood, for the estimation of sparse precision matrix, Friedman and others (2008) proposed the graphical LASSO algorithm. This algorithm is fast and allows the re-use of the estimate under one value of the tuning parameter as a “warm” start for the next value.

To obtain the L_1 penalized likelihood, we substitute the penalty function in (2.8) into the objective function (2.7). Then, the optimization problem that gives sparse estimates of Γ and Θ is the solution of

$$\max_{\Theta, \Gamma} \left\{ \log \det(\Theta) - \text{tr}(S_\Gamma \Theta) - \lambda \sum_{i \neq j}^p |\Theta_{ij}| - \rho \sum_{i,j}^p |\gamma_{ij}| \right\}, \quad (2.10)$$

where λ and ρ are the two tuning parameters that control the sparsity of estimates of Γ and Θ . For the optimization problem defined in (2.10), we adopt the algorithm used in Rothman and others (2010) for multivariate regression setting and Yin and Li (2011) for conditional Gaussian graphical models. It implements an efficient coordinate descent algorithm similar to the idea of Friedman and others (2008) for the

LASSO penalty function. One can adopt the optimization algorithm in [Cai and others \(2011\)](#) that is based on linear programming using the primal dual and interior point algorithm, but this is less efficient.

However, [Fan and Li \(2001\)](#) and [Lam and Fan \(2009\)](#) have shown that the L_1 penalty results in substantial biases for large values of the estimates. Other nonconvex penalties such as SCAD have been shown to produce unbiased estimates for large coefficients in addition to resulting sparse solutions and ensuring consistency in model selection ([Lam and Fan, 2009](#)).

2.3.2 SCAD-based penalized inference. Under the SCAD penalty likelihood framework, estimates of Θ and Γ are obtained using the strategy followed in [Fan and others \(2009\)](#). It reformulates the local linear approximation iterative algorithm of [Zou and Li \(2008\)](#) that can be solved by the graphical LASSO algorithm. In each step, the SCAD penalties are approximated by a symmetric piecewise linear functions. Using the Taylor expansion, we approximate $P_\lambda(|\theta|)$ and $P_\rho(|\gamma|)$ in the neighborhood of $|\theta|$ and $|\gamma|$, respectively, as follows:

$$\begin{aligned} P_\lambda(|\theta|) &\approx P_\lambda(|\theta_0|) + P'_\lambda(|\theta_0|)(|\theta| - |\theta_0|), \\ P_\rho(|\gamma|) &\approx P_\rho(|\gamma_0|) + P'_\rho(|\gamma_0|)(|\gamma| - |\gamma_0|), \end{aligned}$$

where $P'_\lambda(\cdot)$ and $P'_\rho(\cdot)$ are derivatives with respect to θ and γ , respectively. These derivatives are nonnegative for $\theta, \gamma \in [0, \infty)$ due to the monotonicity of $P_\lambda(\cdot)$ and $P_\rho(\cdot)$ over $[0, \infty)$. Denote the k -step solution by $\Theta^{(k)}$ and $\Gamma^{(k)}$. Consequently, at step k , we optimize

$$\max_{\Theta, \Gamma} \left\{ \log \det(\Theta) - \text{tr}(S_\Gamma \Theta) - \sum_{i \neq j}^p w_{ij} |\theta_{ij}| - \sum_{i,k}^p \omega_{ij} |\gamma_{ij}| \right\}, \quad (2.11)$$

where $w_{ij} = P'_\lambda(\hat{\theta}_{ij}^{(k)})$, $\omega_{ij} = P'_\rho(\hat{\gamma}_{ij}^{(k)})$ and $\hat{\theta}_{ij}^{(k)}$ and $\hat{\gamma}_{ij}^{(k)}$ are the (i, j) -elements of $\Theta^{(k)}$ and $\Gamma^{(k)}$, respectively.

We then proceed a two-stage optimization of (2.11). In the first stage, for fixed Γ , we optimize

$$\arg \max_{\Theta} \left\{ \log \det(\Theta) - \text{tr}(S_\Gamma \Theta) - \sum_{i \neq j}^p w_{ij} |\theta_{ij}| \right\},$$

using the graphic LASSO algorithm. In our numerical studies, we consider a one-step iteration ($k = 1$) by using the L_1 penalty graphical LASSO estimates as initial values and for calculating the weights w_{ij} and ω_{ij} for the SCAD penalty-based estimation; see [Fan and others \(2009\)](#) and [Zou and Li \(2008\)](#) for details.

In the second-stage estimation using the SCAD penalty, given the updated Θ from the first stage, we optimize

$$\begin{aligned} &\arg \max_{\Gamma} \left\{ -\text{tr}(S_\Gamma \Theta) - \sum_{i,j}^p \omega_{ij} |\gamma_{ij}| \right\} \\ &= \arg \max_{\Gamma} \left\{ -\text{tr}(C_x \Theta - C_{xx-} \Gamma' \Theta - \Gamma C'_{xx-} \Theta + \Gamma C_{x-} \Gamma' \Theta) - \sum_{i,j}^p \omega_{ij} |\gamma_{ij}| \right\}. \end{aligned} \quad (2.12)$$

This is a quadratic function in Γ for fixed Θ and a direct coordinate decent algorithm can be used to get the penalized estimate of Γ . For the (i, j) th entry of Γ and an arbitrary $p \times p$ matrix A , $\partial \text{tr}(\Gamma \Theta) / \partial \gamma_{ij} = a_{ji} = e'_j A e_i$, where e_j and e_i are the corresponding base vectors with p dimensions each. The derivative

of the penalized negative log-likelihood from (2.12) with respect to γ_{ij} is given by

$$-\frac{\partial \ell_{\text{pen}}}{\partial \gamma_{ij}} = 2e'_j(C_{xx-}\Gamma'\Theta)e_i - 2e'_j(C'_{x-}\Theta)e_i + \text{sgn}(\gamma_{jk})\omega_{ij}, \quad (2.13)$$

where $\text{sgn}(\cdot)$ is the sign function. By setting the score function (2.13) to zero and solving for γ_{ij} , we obtain the updating formula for γ_{ij} at the $(k+1)$ th iteration,

$$\hat{\gamma}_{ij}^{(k+1)} = \frac{\mathcal{S}(g_{ij}, \omega_{ij})}{2(e'_j C_{xx-} e_j)(e'_i \Theta e_i)}, \quad (2.14)$$

where \mathcal{S} denotes the soft-thresholding operator

$$\mathcal{S}(g_{ij}, \omega_{ij}) = \text{sgn}(g_{ij})(|g_{ij}| - \omega_{ij})_+,$$

with $g_{ij} = 2\{e'_j(C'_{x-}\Theta)e_i + (e'_j C_{xx-} e_j)(e'_i \Theta e_i)\hat{\gamma}_{ij}^{(k)} - e'_j(C_{xx-}\hat{\Gamma}^{(k)'}\Theta)e_i\}$ and $\hat{\Gamma}^{(k)}, \hat{\gamma}_{ij}^{(k)}$ are the estimates in the k th step of the iteration.

To obtain the final sparse estimates of Θ and Γ using the SCAD penalty, these two updating stages are taken together iteratively until convergence using optimally chosen tuning parameters λ and ρ .

2.4 Model selection

Under the penalized maximum likelihood setting for TSCGMs, the sparsity of the estimated precision matrix Θ and the autoregressive coefficient matrix Γ are controlled by the tuning parameters λ and ρ in (2.10). We use the Bayesian information criterion (BIC) which has successfully been applied in selecting the tuning parameters in Yin and Li (2011) and references therein. However, cross-validation can also be used, similar to Cai and others (2011) and Fan and others (2009). The BIC is defined as

$$\text{BIC}(\lambda, \rho) = -nT\{\log \det(\hat{\Theta}_\lambda) - \text{tr}(S_{\hat{\Gamma}_\rho} \hat{\Theta}_\lambda)\} + \log(nT)(a_n/2 + b_n + p), \quad (2.15)$$

where p is the number of variables, a_n is the number of nonzero off-diagonal elements of $\hat{\Theta}_\lambda$ and b_n is the number of nonzero elements of $\hat{\Gamma}_\rho$. Thus, we select the values of λ and ρ that minimizes the criterion in (2.15). Here the minimization of $\text{BIC}(\lambda, \rho)$ with respect to λ and ρ is achieved by a grid search.

2.5 Simulation strategies

To evaluate and explore the performance of the proposed time series chain graphical modeling approach, we set up a simulation to generate sparse matrices Θ and Γ , similar to Yin and Li (2011). Moreover, we set up a simulation to generate sparse matrices Θ and Γ under commonly encountered network structures: random, scale-free, and cluster networks to compare the finite sample performance of the proposed method with existing approaches in the literature. Simulation from the different graphs was carried out using the R package `bigmatrix`, keeping in mind the autoregressive coefficient matrix to represent these graph structures. That is for the autoregressive coefficient matrix we took independently generated upper diagonal precision matrix along with a 2% nonzero diagonal elements sampled from uniform (0,1).

Then, we simulate initial gene expression values X_1 at time $t = 1$ from $N(0, \Theta^{-1})$ and for subsequent time steps $t = 2, \dots, T$ we simulate the gene expression values following VAR(1) model such that $X_t \sim N(\Gamma X_{t-1}, \Theta^{-1})$. This gives a time series gene expression data for p genes. Then we obtained n i.i.d. replicates of the time series gene expression data. The simulations are repeated 100 times for the various combinations of p , n and T .

Table 1. *Performance measure results of the simulation study for TSCGMs using SCAD penalized likelihood estimation for the precision and autoregressive coefficient matrices*

p	n	T	λ	Performance measures for Ω			ρ	Performance measures for Γ		
				Sensitivity	Specificity	MCC		Sensitivity	Specificity	MCC
10	10	5	0.40	0.963	0.852	0.604	0.50	0.500	0.969	0.560
		15	0.20	0.858	0.909	0.740	0.25	0.720	0.998	0.825
	50	5	0.15	0.821	0.939	0.761	0.20	0.833	0.999	0.836
		15	0.06	0.934	0.996	0.947	0.15	0.864	1.000	0.922
50	10	5	0.80	0.505	0.964	0.540	0.70	0.524	0.978	0.450
		15	0.30	0.945	0.979	0.753	0.40	0.698	0.994	0.709
	50	5	0.35	0.979	0.989	0.853	0.30	0.784	0.996	0.824
		15	0.15	0.960	0.998	0.958	0.20	0.850	1.000	0.918
	100	5	0.20	0.952	0.985	0.922	0.25	0.899	0.998	0.927
		15	0.10	0.901	0.997	0.921	0.20	0.887	1.000	0.941

We measure the performance of our proposed approach using specificity ($\text{SPE} = \text{TN}/(\text{TN} + \text{FP})$), sensitivity ($\text{SEN} = \text{TP}/(\text{TP} + \text{FN})$), and Matthews correlation coefficient ($\text{MCC} = (\text{TP} \times \text{TN} - \text{FP} \times \text{FN})/\sqrt{((\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN}))}$), where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives in identifying the nonzero elements in the matrices Θ and Γ . Here we consider the nonzero entry in the sparse Θ or Γ matrix as positive. We note that high values of specificity, sensitivity, and Matthews correlation coefficient are indicators of good performance of the proposed approach for the given combination of number of variables, number of time points, and number of replications.

3. RESULTS

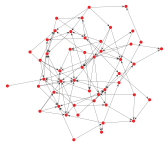
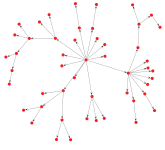
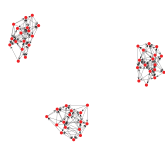
3.1 Simulation study application

We present the simulation results of sparse precision and autoregressive coefficient matrices in Table 1 based on optimal tuning parameters chosen by the minimum BICs from a sample data generated independently. Here we report only the results of sparse estimation of TSCGMs based on SCAD penalized maximum likelihood. In each simulation setting, we have very sparse matrices with only $(1/p) \times 100\%$ nonzero entries.

From the table, we can see that our method on many of the measures of performance scores high. These results suggest that, though recovering sparse network structure is a challenging task, the proposed approach has a good performance on model-based simulations. As expected, more accurate results are obtained for a large number of replications n and longer time lengths T . Importantly, the method's performance is relatively good even for a large number of variables p with very small number of time points and replications. We note here that improved model performance can be gained by allowing the tuning parameters ρ and λ to vary with each simulation. In general, the high-specificity values reveal that the zero entries in the matrices are likely to be accurately identified. The relatively high-sensitivity values are suggestive of recovering the few nonzeros entries of the matrices or true links in the network.

Furthermore, we compare the finite sample performance of the proposed approach using SCAD and GLASSO penalties and with two recently proposed approaches implemented in R: empirical Bayes dynamic Bayesian network (EBDBN) approach in [Rau and others \(2010\)](#) and vector autoregressive using dynamic correlation modeling (GeneNet) in [Opgen-Rhein and Strimmer \(2007\)](#). The two methods

Table 2. Performance measure results of the simulation study that compares various approaches in estimating sparse precision matrix (Ω) and autoregressive coefficient matrix (Γ) when $n = 50$ and $t = 10$

		 Random		 Scale-free		 Cluster	
		$p = 50$	$p = 100$	$p = 50$	$p = 100$	$p = 50$	$p = 100$
<i>Precision matrix</i>							
SCADGLASSO	SEN	0.974	1.000	0.987	0.957	0.992	0.971
	SPE	0.989	0.987	0.990	0.989	0.945	0.946
GLASSO	L_2 distance	7.403	12.342	6.289	7.484	9.576	30.192
	SEN	0.971	0.997	0.930	0.946	0.975	0.923
	SPE	0.984	0.984	0.989	0.967	0.944	0.942
	L_2 distance	7.591	12.408	6.821	13.67	9.641	30.517
<i>Autoregressive coefficients</i>							
SCADGLASSO	SEN	0.997	0.992	0.998	0.948	0.989	0.993
	SPE	0.967	0.985	0.985	0.899	0.977	0.959
	L_2 distance	0.892	1.595	0.602	1.665	0.452	3.254
GLASSO	SEN	0.997	0.991	0.939	0.935	0.977	0.989
	SPE	0.964	0.980	0.985	0.800	0.954	0.960
	L_2 distance	1.402	2.811	2.073	13.67	1.0152	4.608
EBDBN	SEN	0.392	0.201	0.343	0.195	0.394	0.226
	SPE	0.607	0.791	0.615	0.793	0.599	0.787
	L_2 distance	52.610	70.507	37.910	52.048	59.730	80.722
VAR	SEN	0.119	0.026	0.000	0.000	0.000	0.000
	SPE	0.899	0.968	0.969	0.971	0.997	0.999
	L_2 distance	16.761	26.946	10.607	15.092	19.414	31.197

estimate only the autoregressive coefficient matrix. Results of the finite sample performance simulation studies for random, scale-free, and cluster network structures are provided in Table 2.

From the simulation results, we observed that SCAD penalty performs slightly better than GLASSO penalty in terms of sensitivity and specificity and considerably better performance in the distance measure. In the EBDBN approach, the authors implemented only a test procedure for matrix D in (2.4) instead of $CB + D$ in (2.5) to describe dynamic relationship. It could be due this we found to lower performance of the EBDBN approach on these measures. The GeneNet showed good performance in terms of specificity and very low performance in terms of sensitivity, which is consistent with the comparative result reported in Opgen-Rhein and Strimmer (2007). These results are consistent across the different graph structures.

In terms of computational time, our approach provides sparse solution, for example, for 50 variables on a standard 2012 workstation based on a grid points of 10×10 for selecting the best penalty values, it requires around 3 min. A gain in computational time can be achieved for model selection using the BIC criterion by a careful selection of the path of the grid points, for instance, by running the algorithm without model selection at a few reasonably selected points. We note that this issue of model selection is a topic of further investigation. GeneNet is very fast and requires less than a second, whereas EBDBN requires 1 s without hidden variables. It takes longer time as the number of hidden variables involved increases; for example, if two hidden variables are involved, it takes around 3 min.

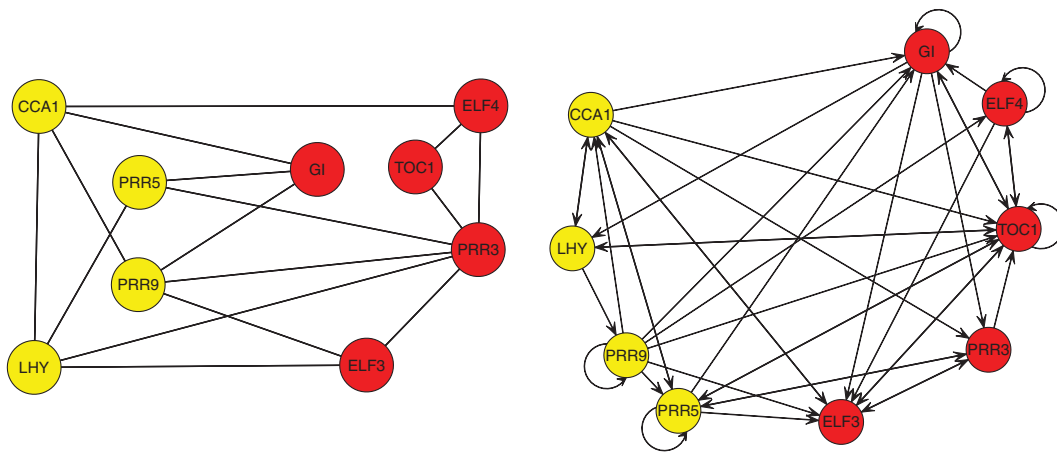


Fig. 1. Contemporaneous circadian gene interaction network in *A. thaliana* (left) and circadian gene regulatory network during 4 h of time intervals in *A. thaliana* (right).

3.2 Data analysis

In this section, we illustrate our proposed approach to time course gene expression datasets related to the study of circadian regulation in plants and regulatory gene network in mammary gland development in mice.

3.2.1 Application to *Arabidopsis thaliana* dataset. From *A. thaliana*, a model plant, circadian gene expression data were sampled to understand its internal clock-signalling network. Data were collected with the interval of 4 h during two light–dark cycles: 12 h:12 h (Reference Series: GSE3416 from GEO database). The six time points are 0 h, 4 h, 8 h, 12 h, 16 h, and 20 h. The GSE3416 dataset includes three individual samples. We consider the same group of nine genes, which from previous studies are known to be involved in circadian regulation, as in Grzegorzczuk and Husmeier (2011a, 2011b), Grzegorzczuk and others (2008), and Jia and Huan (2009). They consist of two groups of genes: “Morning genes”, which including LHY, CCA1, PRR9, and PRR5 whose expression peaks in the morning. “Evening genes”, include TOC1, ELF4, ELF3, GI, and PRR3 whose expression peaks in the evening. In this work, we analyzed the time series without combining or averaging to a single series as was sometimes done in the other approaches and applied our proposed chain graph autoregressive model of order 1.

The extended BIC criterion selects the penatly values $\lambda = 0.149$ and $\rho = 0.282$. The resulting instantaneous and delayed interaction network among the genes are displayed in Figure 1, left and right panels, respectively. In these figures morning genes are represented by lightly shaded circles and evening genes by shaded circles. In the right panel of Figure 1, there are several directed genes pointing from morning genes to evening genes and vice-versa. Some of the genes play important roles in the circadian clock network. The morning gene CCA1 found to repress the evening genes TOC1 and GI. Among the evening genes, for example, TOC1 and GI genes form a feedback loop. The evening gene ELF3 found to inhibit the morning gene LHY. Moreover, ELF3 and GI are involved in the regulatory interactions between the morning genes CCA1 and LHY and the evening genes ELF4 and PRR3. In particular ELF3 interacts negatively with LHY in Figure 1 (left panel) and positively with CCA1 in Figure 1 (right panel). Many of the results are consistent with the findings in Grzegorzczuk and Husmeier (2011a, 2011b) and references therein and the biological network referred to in Jia and Huan (2009) which is based on the work of Más (2008) and Salomé and McClung (2004).

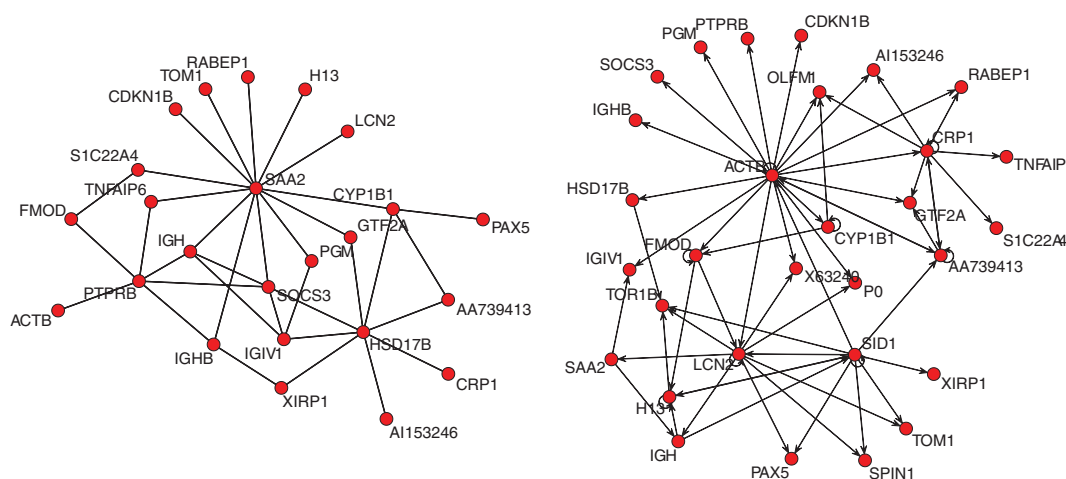


Fig. 2. Undirected (left) and directed (right) TSCGM network inferred from the mammary gland time course expression data using SCAD penalty.

3.2.2 Application to mammary gland gene expression dataset. This second example demonstrates the proposed approach on the analysis of mammary gland gene expression time course data from [Stein and others \(2004\)](#). In the mammary gland experiment, there are 12 488 probe sets representing ~ 8600 genes. These probe sets are measured over 54 arrays of 3 replicates on each of 18 time points in the following developmental stages: virgin, 6, 10, and 12 weeks; pregnancy days 1, 2, 3, 8.5, 12.5, 14.5, and 17.5; lactation days 1, 3, and 7; involution days 1, 2, 3, 4, and 20. We identified 30 genes that yield the best separation between the developmental stages using cluster analysis. We apply the proposed approach to study the interaction between these crucial genes that trigger the transitions to the main developmental events in the mammary gland of mice. We note that the VAR model assumes that the time points are equidistant which is not the case for the mammary gland data. However, one can argue that assuming equidistant measurements is justifiable at least in terms of equal relative reaction rate ([Ongen-Rhein and Strimmer, 2007](#)).

The BIC criterion selects the optimal tuning parameter values $\lambda = 0.15$ and $\rho = 0.93$ for the mammary data. With these tuning parameters, the resulting networks with coefficients greater than 0.05 are displayed in Figure 2. Figure 2 (left panel) shows the undirected links that suggest contemporaneous interactions among the genes and Figure 2 (right panel) displays the directed links that indicate delayed interactions among the genes.

It is observed that genes such as SAA2 and HSD17B from Figure 2 (left panel) and ACTB, LCN2, SID1, and CRP1 from Figure 2 (right panel) are more likely to be hubs, suggesting that they may play a fundamental role in modulating gene activities in the developmental stages of mammary gland. [Blanchard \(2007\)](#) also found SAA2 (serum amyloid protein A2), a member of a multigene family associated with high-density lipoproteins, is elevated in response to inflammation and abnormal proliferation of cells during the involution developmental stage of mice. The central role played by the HSD17B could be in support of [Labrie and others \(1997\)](#) who suggested that the HSD17B (17- β hydroxysteroid dehydrogenase) gene family provide each cell with the necessary mechanisms to control the level of intracellular androgens and/or estrogens. Moreover, past studies suggested that LCN2 (lipocalin 2) is important in the innate immune response to bacterial infection and also functions as a growth factor ([Schmidt-Ott and others, 2007](#)). It is also known that ACTB or the Actin family of genes are highly conserved genes that are involved in cell

motility, structure, and integrity. In general, these findings suggest that these genes seem to play crucial roles in mammary gland gene regulatory network that will be of interest for future work.

4. DISCUSSION AND CONCLUSION

We have presented a sparse TSCGM to infer genetic networks from time course gene expression profiles. The proposed approach combines the main features of Gaussian graphical models and dynamic Bayesian networks to infer instantaneous conditional dependencies among the genes and dynamic or delayed interactions possibly potentially “causal” relationship among genes at consecutive time steps.

To obtain sparse estimates for the high-dimensional gene expression network, we considered penalized likelihood estimation using the SCAD penalty. The use of SCAD penalty with the proposed TSCGM resulted in a much sparser and interpretable network than other methods. Simulation studies showed that the proposed sparse TSCGM can estimate the underlying gene expression networks accurately. This is demonstrated by high values of the sensitivity, specificity, and Matthews correlation coefficient.

We have illustrated the application of time series chain graphical modeling on small replicated gene expression microarray time course datasets from mammary gland and nine circadian genes in *A. thaliana*. Interesting hypotheses have emerge from the inferred networks on the biological properties of genes in the mammary gland and circadian regulation in *A. thaliana*. Some of the results are consistent with the existing literature. Here we note that the model assumptions such as Gaussian noise, linear dynamics, first-order Markovian dynamics, stable network, etc. need to be carefully investigated for time course gene expression data. We also note that developing model selection criteria other than BIC for the chain graphical model is a topic of further research. We are working on extension of the methodology used in this paper based on copulas and local kernel weights to account for non-Gaussianity, unequal time spacing and unstable and nonlinear dynamic networks for autoregressive process of order d ($d \geq 1$) including the determination of time lag.

5. SOFTWARE

The computer code that can be used for the methodology developed in this paper is available online (<http://www.math.rug.nl/stat/Main/Software>) as an R statistical computing environment package, `SparseTSCGM_1.0.tar.gz`.

ACKNOWLEDGMENTS

The authors are grateful to the Editor, an Associate Editor, and the anonymous referee for their very valuable comments. *Conflict of Interest*: None declared.

REFERENCES

- ABURATANI, S., SAITO, S., TOH, H. AND HORIMOTO, K. (2006). A graphical chain model for inferring regulatory system networks from gene expression profiles. *Statistical Methodology* **3**, 17–28.
- BLANCHARD, A. (2007). Gene expression profiling of early involuting mammary gland reveals novel genes potentially relevant to human breast cancer anne blanchard, robert shiu 2, stephanie booth 4, garrett sorensen 4, nicole decorby1, andreea nistor1, paul wong 5, etienne levygue3 and yvonne myal. *Frontiers in Bioscience* **12**, 2221–2232.
- CAI, T., LIU, W. AND LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

- DAHLHAUS, R. AND EICHLER, M. (2003). Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, 115–137.
- EFRON, B., HASTIE, T., JOHNSTONE, I. AND TIBSHIRANI, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.
- FAN, J., FENG, Y. AND WU, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics* **3**, 521.
- FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.
- GAO, W. AND TIAN, Z. (2010). Latent ancestral graph of structure vector autoregressive models. *Journal of Systems Engineering and Electronics* **21**, 233–238.
- GRZEGORCZYK, M. AND HUSMEIER, D. (2011a). Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. *Bioinformatics* **27**, 693–699.
- GRZEGORCZYK, M. AND HUSMEIER, D. (2011b). Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning* **83**, 355–419.
- GRZEGORCZYK, M., HUSMEIER, D., EDWARDS, K. D., GHAZAL, P. AND MILLAR, A. J. (2008). Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. *Bioinformatics* **24**, 2071–2078.
- JIA, Y. AND HUAN, J. (2009). The analysis of *Arabidopsis thaliana* circadian network based on non-stationary DBNS approach with flexible time lag choosing mechanism. In: *Bioinformatics and Biomedicine, 2009. BIBM'09. IEEE International Conference on IEEE*, Washington DC. pp. 178–181.
- LABRIE, F., LIN, S. X., CLAUDE, L., SIMARD, J., BRETON, R., BÉLANGER, A. and others. (1997). The key role of 17 β -hydroxysteroid dehydrogenases in sex steroid biology. *Steroids* **62**, 148–158.
- LAM, C. AND FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37**, 4254.
- MÁS, P. (2008). Circadian clock function in *Arabidopsis thaliana*: time beyond transcription. *Trends in Cell Biology* **18**, 273–281.
- MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.
- 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**(Suppl 2), S3.
- RANGEL, C., ANGUS, J., GHAHRAMANI, Z., LIOUMI, M., SOTHERAN, E., GAIBA, A., WILD, D. L. AND FALCIANI, F. (2004). Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics* **20**, 1361–1372.
- RAU, A., JAFFRÉZIC, F., FOULLEY, J. L. AND DOERGE, R. W. (2010). An empirical Bayesian method for estimating biological networks from temporal microarray data. *Statistical Applications in Genetics and Molecular Biology* **9**, DOI:10.2202/1544-6115.1513.
- ROTHMAN, A. J., LEVINA, E. AND ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962.
- SALOMÉ, P. A. AND MCCLUNG, C.R. (2004). The *Arabidopsis thaliana* clock. *Journal of Biological Rhythms* **19**, 425–435.

- SCHMIDT-OTT, K. M., MORI, K., LI, J. Y., KALANDADZE, A., COHEN, D. J., DEVARAJAN, P. AND BARASCH, J. (2007). Dual action of neutrophil gelatinase-associated lipocalin. *Journal of the American Society of Nephrology* **18**, 407–413.
- SIMA, C., HUA, J. AND JUNG, S. (2009). Inference of gene regulatory networks using time-series data: a survey. *Current Genomics* **10**, 416.
- STEIN, T., MORRIS, J. S., DAVIES, C. R., WEBER-HALL, S. J., DUFFY, M. A., HEATH, V. J., BELL, A. K., FERRIER, R. K., SANDILANDS, G. P., GUSTERSON, B. A. *and others*. (2004). Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving lbp, cd14 and stat3. *Breast Cancer Research* **6**, R75–91.
- YIN, J. AND LI, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5**, 2630–2650.
- ZOU, H. AND LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 1509.

[Received May 31, 2012; revised December 28, 2012; accepted for publication February 5, 2012]