Noname manuscript No.
(will be inserted by the editor)

Adaptive Thresholding for Reconstructing Regulatory Networks from Time Course Gene Expression Data

Ali Shojaie · Sumanta Basu · George Michailidis

Received: 2 June 2011 / Accepted: 4 December 2011

Abstract Discovering regulatory interactions from time course gene expression data constitutes a canonical problem in functional genomics and systems biology. The framework of graphical Granger causality allows one to estimate such causal relationships from these data. In this study, we propose an adaptively thresholding estimates of Granger causal effects obtained from the lasso penalization method. We establish the asymptotic properties of the proposed technique, and discuss the advantages it offers over competing methods, such as the truncating lasso. Its performance and that of its competitors is assessed on a number of simulated settings and it is applied on a data set that captures the activation of T-cells.

Keywords Regulatory networks · time-course gene expression data · graphical Granger causality · thresholding · lasso

Ali Shojaie
Department of Biostatistics, University of Washington
Tel.: +1206-615-5323
Fax: +1206-543-3286
E-mail: ashojaie@u.washington.edu

Sumanta Basu
Department of Statistics, University of Michigan
Tel.: +1734-358-7953
Fax: +1734-763-4674
E-mail: sumbose@umich.edu

George Michailidis
Department of Statistics, University of Michigan
Tel.: +1734-763-3498
Fax: +1734-763-4674
E-mail: gmichail@umich.edu

1 Introduction

Reconstructing gene regulatory networks is a critical problem in systems biology. Gene regulation is carried out by binding of protein products of transcription factors (TF) to *cis*-regulatory elements of genes, which results in change of expression levels of the regulated genes. Such relationships are often represented in the form of directed graphs with transcription factors (TF) regulating target genes. This interpretation of effects of transcription factors on regulated genes, as a physical intervention mechanism therefore implies that regulatory interactions among genes are by definition *causal*.

In the theory of graphical models, causal relationships among random variables are modeled using directed (acyclic) graphs, where an edge among two random variables indicates a direct causal effect. Statistical methods based on observational data can only determine associations among random variables and causal discovery requires additional assumptions and/or information about the underlying system. This implies that, reconstructing gene regulatory networks may be only feasible through carefully designed perturbation experiments. Such experiments are often expensive and only possible in case of model organisms and cell lines. However, regulatory mechanisms become evident if the expression level of gene Y is affected by changes in expression levels of gene X. Time course gene expression data provide a dynamic view of expression levels of all the genes under study, and therefore, can provide cues to the causal relationships among genes, which can be used to reconstruct the gene regulatory network.

Two of the most popular approaches for inferring gene regulatory networks using time course gene expression data are dynamic Bayesian Networks, Murphy (2002) and Granger causality, Granger (1969). Dynamic Bayesian Networks (DBNs), generalize the notion of Bayesian networks to allow for cycles in the graph, through expanding the state space of the model by replicating the variables in the network over time points. Cyclic networks are then transformed to directed acyclic graphs (DAGs) by breaking down cycles into interactions between variables at two different time points. Ong et al (2002) and Perrin et al (2003) discuss applications of DBNs for inferring regulatory networks from time course gene expression data.

On the other hand, Granger causality is motivated by a practical interpretation of predictability among random variables. In particular, given two random variables X and Y, if the autoregressive model of Y based on past values of both variables significantly outperforms the model based on Y alone, X is said to be Granger-causal for Y. In the context of gene expression analysis, this definition implies that changes in expression levels of Y could be explained by expression levels of X from previous time points. Exploring Granger causal relationships is closely related to analysis of vector autoregressive (VAR) models. Therefore, while applying DBNs to high-dimensional applications may be computationally prohibitive, statistical methods can be used to derive Granger causal relationships among genes from time-course gene expression data using standard techniques for analysis of VAR models (see Yamaguchi et al (2007); Opgen-Rhein and Strimmer (2007) for examples of such approaches).

Unlike the original application area of Granger causality in econometrics, in gene regulatory network applications, the number of available samples is often small compared to the number of genes in the study. As a result, sparse VAR models have been explored by a number of researchers, including Fujita et al (2007) and Mukhopadhyay and Chatterjee (2007), to obtain reliable estimates of gene regulatory networks when the number of genes, p is large compared to the sample size, n.

Penalized estimation methods provide sparse estimates of high dimensional statistical models. Arnold et al (2007) use the lasso (or ℓ_1) penalty to discover the structure of graphical models based on the concept of Granger causality in a financial setting. More recently, a similar framework, using the group lasso penalty was used by Lozano et al (2009) to group the effect of observations of each variable over past time points.

A main challenge in applying both DBN and Granger causality models to discover gene regulatory networks is that as the number of time points increases, the number of variables used in the replicated representation of the network also increases. As a result, many available methodologies simply ignore possible effects of genes on each other from time points far in the past, resulting in possible loss of information. To overcome this challenge, Shojaie and Michailidis (2010a) proposed to simultaneously estimate the order of the vector auto-regressive model, as well as the interactions among variables using a non-convex penalty, called the truncating lasso penalty, and showed that when the effects of variables on each other decay over time, the proposed penalty consistently estimates the order of the time series, as well as the structure of the regulatory network in high dimensional sparse settings.

The decay condition in Shojaie and Michailidis (2010a) (referred to as S-M henceforth) is a natural assumption in many time series models. However, when this condition is not satisfied, the truncating lasso penalty may fail to correctly estimate the order of the time series. In this study, we discuss examples where the decay assumption of S-M may fail to hold, and propose a new estimator, based on adaptive thresholding of lasso estimates, which can be used to simultaneously estimates the order of the VAR model and the structure of the network. The new estimator is based on the assumption that if the true VAR model includes non-ignorable effects at any given time point, the number of edges in the network should exceed a certain threshold. We formally state this assumption in Section 2.2.2, where we also investigate the effect of violations of this assumption on false positive and false negative errors.

The remainder of the paper is organized as follows. In Section 2, we review some background material and present the new methodology and discuss its asymptotic properties. Section 3 includes a comparative analysis of the performance of the proposed estimator over a set of simulation studies, whereas applications to time-course gene expression data from T-cell activation are presented in Section 4. Section 5 discusses some final remarks on the choice of appropriate penalty, and methods for evaluating the validity of underlying structural assumption.

2 Estimation of Regulatory Networks from Time Course Gene Expression Data

We start this section by a brief introduction of two classes of statistical models for analysis of genetic networks using time series observations, namely dynamic Bayesian Networks (DBN) and graphical Granger causality. We then discuss penalized methods for estimation of gene regulatory networks and introduce our new estimator based on an adaptively thresholded lasso penalty. Computational issues and asymptotic properties of the proposed estimator are discussed at the end of the section.

2.1 Estimation of Gene Regulatory Networks from Time Course Gene Expression Data

Bayesian networks models (BN) correspond to probability distributions over a directed acyclic graph (DAG). More specifically, let $\mathcal{G} = (V, E)$, denote a DAG with the node set V and the edge set $E \subset V \times V$. Denote the random variables on the nodes of the graph by $X_1, \ldots, X_p$, where $p = |V|$ is the cardinality of the set V. For a DAG $\mathcal{G}$, it is clear that if $(i, j) \in E \Rightarrow (j, i) \notin E$. We represent E through the adjacency matrix A of the graph, a $p \times p$ matrix whose $(j, i)-$th entry indicates whether there is an edge (and its weight) from node j to node i. We represent an edge from j to i by $j \to i$, and denote by $\mathrm{pa_i}$ the set of *parents* of node i.

A probability distribution $\mathcal{P}$ is said to be (Markov) compatible with $\mathcal{G}$ if it admits the following decomposition based on the set of parents of each node in the graph (Pearl (2000)):

$$\mathcal{P}(X_1, \ldots, X_p) = \Pi_{i \in V} \mathcal{P}(X_i | \mathrm{pa_i}). \tag{1}$$

Pearl (2000) shows that if $\mathcal{P}$ is *strictly positive*, the Bayesian network $\mathcal{G}$ associated with $\mathcal{P}$ is unique and $\mathcal{P}$ and $\mathcal{G}$ are compatible. This implies that the joint Gaussian distributions defined according to (1) on nodes of $\mathcal{G}$ are uniquely defined and Markov compatible with $\mathcal{G}$. Markov compatible probability distributions on DAGs can be defined using *structural equation models*, where each variable is modeled as a (nonlinear) function of its parents. Given latent variables $Z_i, i = 1, \ldots, p$ for each node i, the general form of these models is given by:

$$X_i = f_i(\mathrm{pa_i}, \mathrm{Z_i}), \qquad \mathrm{i = 1, \ldots, p} \tag{2}$$

In (2), the latent variables represent the unexplained variation in each node, which is independent of the effect of its parents. For Gaussian random variables, the function f_i is linear, in the sense that it corresponds to the linear regression of X_i on the set of its parents $\mathrm{pa_i}$. In other words, for Gaussian random variables (2) takes the form:

$$X_i = \sum_{j \in \mathrm{pa_i}} \rho_{ij} X_j + Z_i, \qquad i = 1, \ldots, p \tag{3}$$

where ρ_{ij} represent the *effect* of gene j on i for $j \in \mathrm{pa_i}$ and ρ_{ij} are the coefficients of the linear regression model of X_i on $X_j, j \in \mathrm{pa_i}$. Note that in this case $\rho_{ij} = 0$ whenever $j \notin \mathrm{pa_i}$.

The main limitation of Bayesian networks is the requirement that the underlying graph needs to be a DAG. However, gene regulatory networks often include cycles (e.g. the cell cycle) or feedback loops that control the expression levels of genes. Thus, a more general class of probability distributions on graphs is needed that allows for the presence of directed cycles. To overcome this shortcoming, Murphy (2002) introduced a generalization of Bayesian networks for analysis of time series data, called dynamic Bayesian networks (DBN). In DBNs, random variables in the study are replicated over time, and directed edges are only allowed from variables in each time point to those in the future time points. In its simplest form, edges in DBN are limited to those from variables in t to variables in $t+1$. Such a model corresponds to a Markov model. More generally, for variables $X_1, \ldots X_p$ observed over time points $t = 1, \ldots, T$, edges are allowed from any time point t to future time points $t' > t$.

A closely related model for analysis of time series, which we adapt in this work, was developed in the econometrics literature based on the work of Granger (1969). In this framework, called Granger causality, interactions among variables are defined if past observations of one variable result in improved prediction of other variable. More specifically, let $X^{1:T} \equiv \{X\}_{t=1}^{T}$ and $Y^{1:T} \equiv \{Y\}_{t=1}^{T}$, be trajectories of two stochastic processes X and Y up to time T. Then, X is said to be Granger-causal for Y if the joint prediction model in (4) significantly outperforms the model in (5).

$$Y^T = AY^{1:T-1} + BX^{1:T-1} + \varepsilon^T \tag{4}$$

$$Y^T = AY^{1:T-1} + \varepsilon^T \tag{5}$$

Graphical Granger causal models (GGC) extend the notion of Granger causality among two variables to p variables. In general, define a vector time series $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^{\mathsf{T}}$ and consider the corresponding vector auto-regressive (VAR) model (Lütkepohl (2005), Chapter 2):

$$\mathbf{X}^T = A^1\mathbf{X}^{T-1} + \ldots A^d\mathbf{X}^{T-d} + \varepsilon^T. \tag{6}$$

Here, d denotes the order of the time series and $A^t, t = 1, \ldots, d$ are $p \times p$ matrices whose coefficients represent the magnitude of interaction effects among variables at different time points.

In this model, X_j^{T-t} is considered Granger-causal for X_i^T if the corresponding coefficient, $A_{i,j}^t$ is statistically significant. It is then easy to see that, the

GGC corresponds to a DAG with $p \times (d+1)$ variables, in which the ordering of the set of p-variate vectors $\mathbf{X}^{T-d}, \ldots, \mathbf{X}^{T}$ is determined by the temporal index and the ordering among the elements of each vector is arbitrary. As with DBNs, the interactions in GGCs are only allowed to be forward in time, i.e. of the form $X_j^{T-t} \to X_i^T, t = 1, \ldots, d$.

2.2 Penalized Likelihood Estimation Methods for Gene Regulatory Networks

2.2.1 Background

In the analysis of gene regulatory networks, the number of genes often exceeds the available samples of the gene expression data. As a result, an estimate of the gene regulatory network based on graphical Granger causality may include spurious edges that do not correspond to interactions among the genes. In such situations, penalized estimation methods can improve the accuracy of the model, especially for reconstructing the true regulatory network. Shojaie and Michailidis (2010b) show that for Gaussian random variables, when the variables inherit a natural ordering, the likelihood function can be written as a function of the adjacency matrix of the corresponding DAG. They also show that the penalized estimate of the adjacency matrix can be obtained by solving $p-1$ penalized regression problems. Using this connection, general weighted lasso estimates of gene regulatory networks can be found by solving the following p *distinct* ℓ_1-regularized least squares problems for $i = 1, \ldots, p$:

$$\operatorname*{argmin}_{\theta^t \in \mathbb{R}^p} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t} \theta^t\|_2^2 + \lambda \sum_{t=1}^{d} \sum_{j=1}^{p} |\theta_j^t| w_j^t \tag{7}$$

where $\mathcal{X}^t$ denotes the $n \times p$ matrix of observations at time t, and $\mathcal{X}_i^t$ denotes the i^{th} column of $\mathcal{X}^t$. In this formulation, $w_j^t = 1$ corresponds to lasso estimates, and adaptive lasso estimates are obtained by setting $w_j^t = |\hat{A}_{ij}^t|^{-\gamma}$, where $\hat{A}_{ij}^t$ is a consistent estimate of A_{ij}^t. Shojaie and Michailidis (2010b) consider a modification of the adaptive lasso, which they call 2-stage lasso in which $w_j^t = 1 \vee |\hat{A}_{ij}^t|^{-\gamma}$, and $\hat{A}_{ij}^t$ is obtained using an initial lasso estimate and $\gamma = 1$.

As pointed out in S-M, the order of the VAR model d is often unknown. Therefore, to estimate the GGC, one either has to include all the previous time points by setting $d = T-1$, or set d to an arbitrary value. While the latter choice may result in ignoring some of the edges from the true network, the former results in a model with $p(T-1)$ covariates, which in turn exhibits inferior performance. To overcome this shortcoming, the authors propose to estimate the GGC using the *truncating lasso penalty*, which is given as the solution of the following non-convex optimization problem, for $i = 1, \ldots, p$:

$$\operatorname*{argmin}_{\theta^t \in \mathbb{R}^p} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t} \theta^t\|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} |\theta_j^t| w_j^t \tag{8}$$

$$\varPsi^1 = 1, \qquad \varPsi^t = M^{I\{\|A^{(t-1)}\|_0 < p^2\beta/(T-1)\}},\ t \geq 2$$

where M is a large constant, and β is the allowed false negative rate. S-M propose an efficient algorithm for solving the optimization problem in (8), and show that the proposed penalty gives a consistent estimate of the order of the underlying VAR model, as well as the structure of the network if the model satisfies a decay assumption.

2.2.2 Adaptively Thresholded Lasso Estimate

The decay assumption for the truncating lasso estimate considered in S-M is a natural assumption in many applications. However, there are examples of VAR models that do not satisfy this assumption. As an example, consider the VAR model whose adjacency matrix is depicted in the top panel of Figure 2. In this case, observations at time T are affected by those in time $T-1$ and $T-3$, whereas no significant effects exists from observations in time $T-2$. In Section 4, we show that the time series model of T-cell regulation shows a similar pattern of influence. In such cases when the decay assumption fails to hold, the truncating lasso penalty of S-M may not give a correct estimate of the order of the time series, which results in an incorrect estimate of the regulatory network. Examples of such cases are given in Sections 3.2 and 4.

To address this shortcoming, here we propose to consider the use of adaptive thresholding to provide a consistent estimate of the regulatory networks from time course gene expression data. The main idea for the proposed penalty (which replaces the decay assumption of S-M) is that a given time point includes true effects in the VAR model only if the number of edges in the network should exceed a certain threshold (we formalize this assumption in the following discussion).

Thresholding of lasso estimates has been also considered as a tool to improve the accuracy of lasso estimates in Wasserman and Roeder (2009); Meinshausen and Yu (2009). More recently, Zhou (2010) considered iterative thresholding of both lasso and Dantzig selector estimates for estimation of high dimensional sparse regression models with random design matrix. The author studied asymptotic properties of the thresholded estimator and shows that it results in accurate model selection, as well as nearly optimal ℓ_2 loss.

To obtain consistent estimates of the order d, as well as edges of the regulatory network, we modify the thresholding framework of Zhou (2010) so that only adjacency matrices with significant number of edges are included in the estimate of the regulatory network. Consider, as before, random variables $\mathbf{X}^1, \ldots \mathbf{X}^T$ from a VAR model of order d with Gaussian noise, i.e.

$$\mathbf{X}^T = A^1\mathbf{X}^{T-1} + \ldots A^d\mathbf{X}^{T-d} + \varepsilon^T, \qquad \varepsilon^T \sim N(0, \sigma^2 I_p) \tag{9}$$

where I_p denotes the $p \times p$ identity matrix. The adaptively thresholded lasso estimate of GGC is found through the following three-step procedure:

(i) Obtain the regular lasso estimate of the adjacency matrices of GGC $\tilde{A}^t_{\lambda_n}$ by solving (7) with tuning parameter $\lambda = \lambda_n$

(ii) Define $\Psi^t = \exp\left(M\mathbf{1}_{\{\|\tilde{A}^t\|_0 < p^2\beta/(T-1)\}}\right), t = 1, \ldots T$, and find the thresholded estimator by setting:

$$\hat{A}^t_{ij} = \tilde{A}^t_{ij}\mathbf{1}_{\{|\tilde{A}^t_{ij}| \geq \tau\Psi^t\}} \tag{10}$$

Here M is a large constant and τ is the tuning parameter for the thresholding step.

(iii) Estimate the order of the time series by setting

$$\hat{d} = \max_t \left\{t : \|\hat{A}^t\|_0 \geq p^2\beta/(T-1)\right\}$$

Before discussing the asymptotic properties of the proposed adaptively thresholded lasso estimator, we compare some features of the new estimator with the truncating lasso estimator of S-M, and discuss the appropriate choice of tuning parameters λ_n and τ.

The proposed adaptively thresholded estimate is found by first obtaining an estimate of the adjacency matrices using regular lasso. Then, in the thresholding step, simultaneous sparsity and order selection in VAR models is achieved by setting small values of the estimated adjacency matrix to zero, while controlling for the total number of nonzero elements of the adjacency matrix. Finally, the index of the last time point in which a significant number of nonzero elements exist in the estimated adjacency matrix is defined as the estimate of the order of VAR model.

As pointed out earlier, the thresholded estimator requires less stringent assumptions about the structure of the time series model, and as shown in Theorem 1, the consistency of the estimates of the adjacency matrix and the order of the time series are achieved under the usual sparsity and restricted eigenvalue (RE) assumptions. In addition, since the thresholded estimator is found by adaptive thresholding of the regular lasso estimates, the resulting optimization problem is convex. In contrast, although the algorithm for finding the truncating lasso estimate of S-M is shown to be convergent, the resulting estimate may correspond to a local optimum. On the other hand, the thresholded estimator requires appropriate values of two tuning parameters λ_n and τ, and hence the truncating lasso estimate may be obtained more directly. In particular, S-M propose the following error-based choice of tuning parameter, which controls a version of false positive probability:

$$\lambda_e = 2n^{-1/2}Z^*_{\frac{\alpha}{2(T-1)p^2}} \tag{11}$$

where α is the probability of false positive determined by the user, and Z^*_q denotes the upper qth quantile of the standard normal distribution. This alleviates the need for searching over the parameter space for appropriate values of λ and provides an intuitive connection to the original definition of Granger causality between two time series given earlier.

Based on the asymptotic properties of the thresholded lasso estimator, and given $\lambda_0 = \sqrt{2\log\left((T-1)p\right)/n}$, Zhou (2010) suggests the following choices for tuning parameters λ_n and τ:

$$\lambda_n = c_1 \sigma \lambda_0$$

$$\tau = c_2 \sigma \lambda_0$$

for positive constants c_1 and c_2. Considering the fact that, the choice of the thresholding parameter β is determined by the acceptable degree of false negative error, for $\lambda_0 = \sqrt{2 \log((T-1)p)/n}$, and an estimate σ, tuning parameters for the proposed adaptively thresholded estimator amount to appropriate choices of constants c_1 and c_2. A common strategy is to use cross validation (C.V.) over a grid of possible values of c_1 and c_2. We refer the interested reader to Zhou (2010) for additional details on connections between c_1 and c_2 and constants that are defined based on the conditions of the problem. For selection consistency of the estimate, we require $c_1 \geq 2\sqrt{1+\theta}$ for some constant $\theta > 0$ and $c_2 = 4c_1$. The quantity θ controls the rate at which the estimator performs consistent variable selection as reflected in Theorem 1. In Sections 3 and 4, we provide additional guidelines on practical choices of tuning parameters for the data examples considered.

We begin the discussion of asymptotic properties by providing additional notations and statements of the main assumptions.

Denote by $\mathcal{X} = [\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^{T-1}]$ the $n \times p(T-1)$ matrix of "past" observations, and define:

$$\Lambda_{\min}(m) := \min_{\nu \neq 0, \|\nu\|_0 \leq m} \frac{\|\mathcal{X}\nu\|_2^2}{n\|\nu\|_2^2} > 0$$

Denote by $E^t = \{(i,j) : A^t_{ij} \neq 0\}$ the edge set of the adjacency matrix at time lag $t = 1, \ldots, d$ and let $E = \{(i,j) : \exists 1 \leq t \leq d : A^t_{ij} \neq 0\}$ be the set of all edges in the GGC model.

Let $s = \max_i |\mathrm{pa_i}|$ be the maximum number of parents of each node in the GGC model, and define

$$a_0 = \min_{1 \leq t \leq d} \min_{1 \leq i,j \leq p, A_{ij} \neq 0} |A^t_{ij}|$$

The asymptotic analysis for the thresholded lasso in Zhou (2010) incorporates the framework of Bickel et al (2009), based on the restricted eigenvalue condition $RE(\mathcal{X})$, which states that for some integer $1 \leq s \leq (T-1)p$ and a number k, and for all $\nu \neq 0$ we have

$$\frac{1}{K(s,k)} := \min_{J \subset V, |J| \leq s} \min_{\|\nu_{J^c}\|_1 \leq k \|\nu_J\|_1} \frac{\|\mathcal{X}\nu\|_2}{n^{1/2}\|\nu_J\|_2} > 0$$

In this case, we say that $RE(\mathcal{X})$ holds with $K(s,k)$. Based on these assumptions, we have the following result on the consistency of network estimation and order selection.

Theorem 1 (Consistency of Adaptively Thresholded Lasso) *In VAR(d) model of* (6) *with independent Gaussian noise with variance σ^2, suppose $RE(\mathcal{X})$ holds with $K(s,3)$, and that $\lambda_n \geq 2\sigma\sqrt{1+\theta}\lambda_0$ for some $\theta > 0$. Also, assume $a_0 > c\lambda_n\sqrt{s}$, for some constant c depending on $\Lambda_{\min}(2s)$ and $K(s,3)$. Finally, assume $|E| = \zeta\, p^2(T-1)^1$ for some $0 < \zeta < 1$.*

Then for $b = 3K^2(s,3)/4$ and for any $\beta > \frac{(T-1)\, b\, s}{p}$, with probability at least $1 - p(\sqrt{\pi\, log(T-1)p}[(T-1)p]^{\theta})^{-1}$, the following hold for the adaptively thresholded lasso estimator with thresholding parameter β:

(i) Control of Type-I error: $FPR \leq \frac{b\, s}{(T-1)\, p\, (1-\zeta)}$

(ii) Control of Type-II error: if there exists $\delta > 0$ such that $\min_{A^t \neq 0} \|A^t\|_0 > \gamma p^2$ and β is chosen such that $\beta < \delta/(T-1)$, then $FNR = 0$, otherwise, $FNR \leq \frac{\beta}{(T-1)\, \zeta}$

(iii) Order selection consistency: under the condition in (ii), $\hat{d} = d$

Proof The proof here builds on the results in Zhou (2010) (in particular Theorems 1.1 and 3.1), with modifications to account for adaptive thresholding, control of FPR and FNR, and the time series structure. For simplicity, denote by FP and FN, the total number of false positives and false negatives. Also, let $P \equiv |E| = \zeta\,(T-1)p^2$ be total number of positives (i.e. total number of edges) and $N \equiv (T-1)p^2 - |E| = (T-1)p^2\,(1-\zeta)$ denote the number of zeros in the true adjacency matrix.

First, note that from the decomposition of likelihood in Shojaie and Michailidis (2010b) it follows that the adaptively thresholded estimator is found by solving p regular lasso regression problems according to (7), followed by the thresholding step according to (10).

Next note that, by definition of s and the RE condition, each of the p regressions satisfies the $RE(\mathcal{X})$ holds with $K(s,3)$. Therefore, for $\beta = 0$ results of Zhou (2010) apply to each individual regression.

Following Zhou (2010) we consider, for each $\theta \geq 0$, the set

$$\mathcal{T}_{\theta,i} = \left\{ \epsilon_i^T : \left\| \frac{1}{n}\mathcal{X}^T \epsilon_i^T \right\|_\infty \leq \lambda_{\sigma,\theta,p}, \text{ where } \lambda_{\sigma,\theta,p} = \sigma\sqrt{1+\theta}\lambda_0 \right\}$$

for which $\mathbb{P}(\mathcal{T}_{\theta,i}) \geq 1 - (\sqrt{\pi log(T-1)p}((T-1)p)^{\theta})^{-1}$. It then follows from Theorem 1.1 of Zhou (2010) that for $\beta = 0$, on the set $\mathcal{T}_\theta = \prod_{i=1}^{p} \mathcal{T}_{\theta,i}$, we have, for all $i = 1, \ldots p$, $\mathrm{pa_i} \subseteq \hat{\mathrm{pa}}_\mathrm{i}$. This implies that for all $t = 1, \ldots, d$, on the set $\mathcal{T}_\theta$, we have

$$E^t \subseteq \hat{E}^t$$

To obtain the upper bound on FPR, we follow the proof of theorem 3.1 in Zhou (2010) for each of the p regressions separately. First note that from the results of Bickel et al (2009) it follows that on the set $\mathcal{T}_{\theta,i}$, for $\tilde{v}_i = vec(\tilde{A}_{i:}^{1:T} - A_{i:}^{1:T})$,

$$\|\tilde{v}_{i,\mathrm{pa_i}}\|_2 \leq B_0\lambda_n\sqrt{s} \text{ and } \|\tilde{v}_{i,\mathrm{pa_i^c}}\|_1 \leq B_1\lambda_n\, s \qquad (12)$$

[1] This assumption is made for simplicity of representation. The proof can be written in terms of $|E|$, without making any explicit assumptions on the number of true edges.

where $B_0 = 4K^2(s,3)$ and $B_1 = 3K^2(s,3)$. If we threshold the lasso estimate by $4\lambda_n$, then it readily follows from (12) that (see Zhou (2010) for more details) on $\mathcal{T}_{\theta,i}$

$$|\hat{\mathrm{pa}}_\mathrm{i}\backslash\mathrm{pa}_\mathrm{i}| \le \frac{\|\tilde{\mathrm{v}}_{\mathrm{i},\mathrm{pa}_\mathrm{i}^\mathrm{c}}\|_1}{4\lambda_\mathrm{n}} \le \frac{\mathrm{B}_1\,\mathrm{s}}{4} \tag{13}$$

Hence $|\hat{\mathrm{pa}}_i\backslash\mathrm{pa}_\mathrm{i}| \le \mathrm{B}_1\,\mathrm{s}/4$, for all $i = 1,\dots p$, on $\mathcal{T}_\theta$. This implies $FP \le pbs$ where $b = 3K^2(s,3)/4$.It then follows that on $\mathcal{T}_\theta$ for $\beta = 0$, we have $FNR = 0$ and

$$FPR = FP/N \le \frac{b\,s}{(T-1)\,p\,(1-\zeta)}.$$

To complete the proof, it suffices to show that for $\beta > \frac{(T-1)\,b\,s}{p}$, FPR does not increase (or is improved) and $FNR \le \beta/(T-1)\,\zeta$. The fact that adaptive thresholding does not increase FPR follows immediately from the definition of the estimator, as the thresholding coefficient for the adaptively thresholded procedure is at least as large as the procedure of Zhou (2010).

Now suppose $A^t \ne 0$ for some $1 \le t \le T-1$. It follows from $E \subset \hat{E}$ that $\|\hat{A}^t\|_0 \ge \|A^t\|_0$ and hence, if $\|\hat{A}^t\|_0 < \frac{\beta\,p^2}{T-1}$, A^t must satisfy the same inequality. Now, if there exists $\delta > 0$ such that $\min_{A^t \ne 0} \|A^t\|_0 > \gamma p^2$ and β is chosen such that $\beta < \delta/(T-1)$, then $\|A^t\|_0 < \delta\,p^2$, which implies that $A^t \equiv 0$, and hence $FNR = 0$. On the other hand, if the condition in (ii) is not satisfied, FN could be at most βp^2, which implies that

$$FNR \le (\beta p^2)/|E| = \frac{\beta}{(T-1)\zeta}.$$

Finally, to show that $\hat{d} = d$, note that when $A^t \ne 0$, the condition in (ii) guarantees that $\hat{A}^t \ne 0$. On the other hand, if $A^t = 0$, $\|\hat{A}^t\|_0/p^2 \le \frac{b\,s}{p}$ and hence when $\beta \ge \frac{(T-1)\,b\,s}{p}$, $\hat{A}^t \equiv 0$, which completes the proof. □

Before investigating the small sample performance of the proposed estimator in Section 3, we offer some remarks regarding asymptotic properties of the estimator.

1. Consider the asymptotic regime with $n \to \infty$, $p = O(n^a)$, for some $a > 0$, and $s = o(p)$. Assume the constant $K(s,3)$ is uniformly bounded above (see the remark below on the validity of this assumption). Then theorem 1 says that with probability tending to 1, $FPR \to 0$ as long as ζ stays away from 1, i.e., the network is truly sparse. On the other hand, even if no constant δ exists to satisfy the condition in part (ii) of the Theorem, the lower bound on β, given by $\frac{(T-1)\,b\,s}{p}$, converges to zero, indicating that we can make FNR arbitrarily small as long as ζ stays away from zero, i.e., the network is not extremely sparse. The conditions on β are set to achieve a tradeoff between FPR and FNR.

2. The false positive rate in the above theorem can be improved by considering a multi-step thresholding procedure where at the second step the estimate of d is used to restrict the number of time points considered in the estimation. It can be shown that the numerator of the upper bound of FPR can be improved from $b\,s$ to $b\,\sqrt{s}$ (refer to Zhou (2010) for more details on the multi-step thresholding). However, this requires an additional assumption on the number of parents of each node in the graph, and is hence not pursued here.
3. The RE condition has been shown to hold for many non-trivial classes of Gaussian design matrices (see for example van de Geer and Bühlmann (2009), Raskutti et al (2010)). In particular Raskutti et al (2010) shows that $RE(\mathcal{X})$ holds with high probability if the sample size n is sufficiently large ($\sim O(klogp)$) and $RE(\Sigma^{1/2})$ holds, where the rows of $\mathcal{X} \sim N(0, \Sigma)$. Hence it is sufficient to ensure that $\lambda_{min}(\Sigma)$ is bounded away from zero as $n, p \to \infty$, which is not very restrictive since every node of the GGC network is a noisy observation with i.i.d innovation of variance σ^2. For the special case of stationary vector autoregressive processes, Basu et al (2011) use spectral density representation of time series to show a stationary VAR(d) process satisfies this condition if the spectral matrix operator has continuous eigenvalues and eigenvectors and the adjacency matrices for $t = 1, \ldots T$ are bounded above in spectral norm.
4. The results in Theorem 1 are non-asymptotic and are derived in the regime $n, p \to \infty$ and $p \gg n$, without any restrictions on the length of the time series T. However, it can be seen that if $T \to \infty$, then FPR and FNR converge to 0. In addition, the increase in T also improves the probability of the events under study.

3 Numerical Studies

In this section, we evaluate the performance of the proposed thresholded lasso penalty in reconstructing temporal Granger causal effects, and compare it with the performances of (adaptive) lasso and truncating (adaptive) lasso penalties. To this end, we first present the estimated adjacency matrices of two small networks with $p = 20$ and different sparsity patterns to better understand the properties of the thresholded lasso penalty. We then evaluate the phase transition behavior of the competing estimators as the sample size n and the signal to noise ratio (SNR) is varied. To compare the performances of different estimators, we consider three different criteria: (1) the False Positive Rate (FPR), (2) the True Positive Rate (TPR) and (3) the F_1 measure. The F_1 measure is the harmonic mean of $precision(P)$ and $recall(R)$ (i.e. $F_1 = 2PR/(P + R)$) for the estimated graphs. The value of this summary measure ranges between 0 and 1, with higher values corresponding to better estimates.

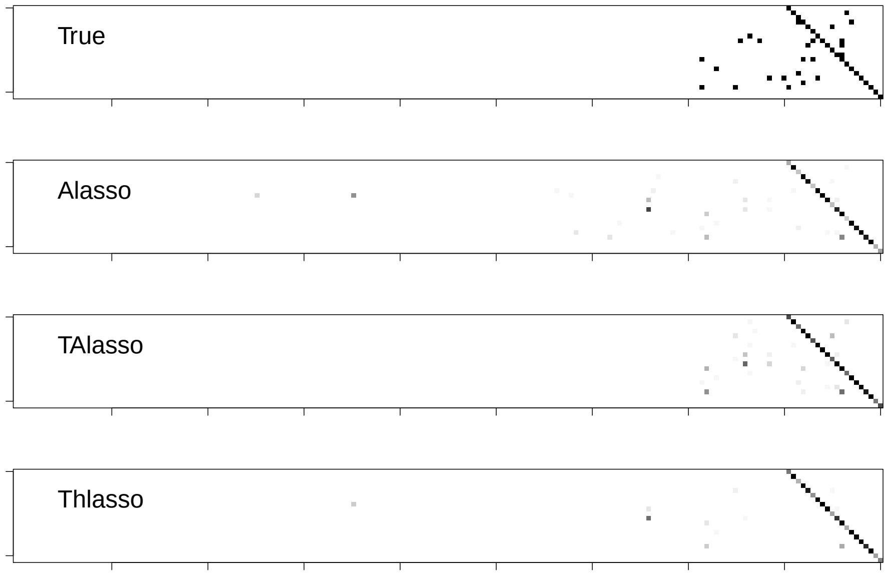


Fig. 1: True and estimated adjacency matrices of graphical Granger model (a) with T=10, d=2, p=20, n=30, SNR=2.4, the gray-scale images of the estimates represent the percentage of times an edge has been detected in the 50 iterations.

3.1 Illustrative Examples

To illustrate the effect of the proposed estimator, we begin with a simple VAR model that satisfies the decay assumption of S-M. Here $T = 20$, $d = 2$, $p = 20$ and $s \simeq min\{0.025p^2, n\}$, and every edge has an effect of $\rho = \pm 0.6$. We simulate $n = 30$ independent and identically distributed observations according to the VAR(d) model in (6), with $\sigma = 0.3$. The values of α and β are set to 0.1 each.

To obtain comparable results, we set the tuning parameter λ for all estimators to $\lambda = 0.6\lambda_e$, where λ_e is defined in (11). The thresholding parameter τ in the second stage of the thresholded lasso penalty is chosen to be $0.7\lambda\sigma$. The results over 50 replications of the above simulation and estimation procedure are presented in Figure 1 and Table 1[2].

As expected, the truncating lasso estimator outperforms the lasso and thresholded lasso estimators, and provides a consistent estimate of the order d. On the other hand, the thresholded lasso estimator offers additional improvements over its non-thresholded counterpart.

[2] Here we present the results of simulation for adaptive versions of lasso and truncating lasso estimators; the behavior of the regular versions of these estimators were similar and were excluded to save space

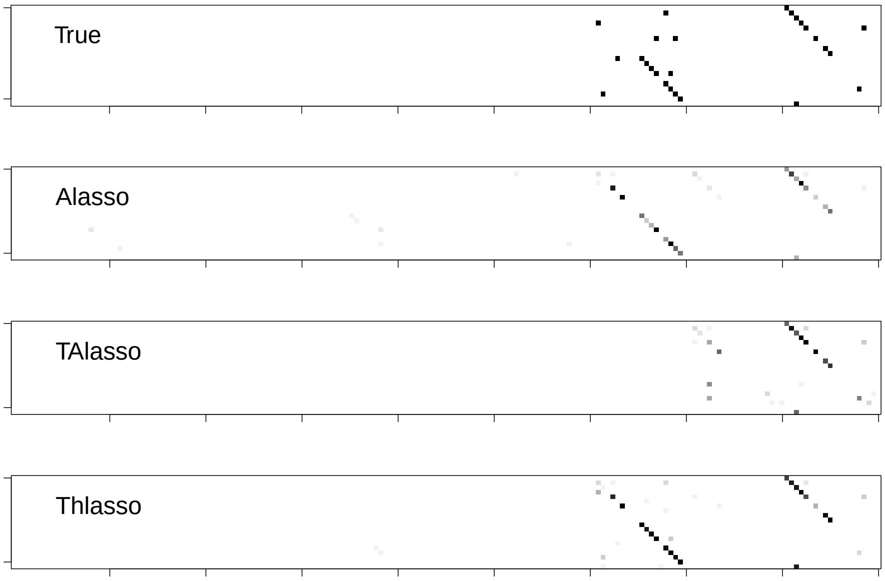


Fig. 2: True and estimated adjacency matrices of graphical Granger model (b) with T=10, d=3, p=20, n=30, SNR=2.4, the gray-scale images of the estimates represent the percentage of times an edge has been detected in the 50 iterations.

Next, we consider a more complicated structure, where the decay assumption is not satisfied. In particular, we construct a network with the same parameters as before except with $d = 3$ in such a way that there is no edge in the adjacency matrix from lag 2 (i.e., $A^2 = 0$). True and estimated adjacency matrices for this simulation setting are shown in Figure 2. The performances of the estimators in terms of TPR, FPR, and F_1 are given in Table 2.

It can be seen that the truncating lasso penalty incorrectly estimates the order of VAR as $\hat{d} = 1$, resulting in increased false positive and false negative errors. On the other hand, the (adaptive) lasso estimate includes many edges in later time lags, while failing to include some of the edges in the first time lag. This simulation illustrates the logic and advantages of the proposed thresholded lasso estimator.

	Alasso	TAlasso	Thlasso
TPR	0.3341 (0.0311)	0.4083 (0.0375)	0.3485 (0.0339)
FPR (×1000)	0.9843 (0.494)	0.8155 (0.4068)	0.4593 (0.2712)
F_1	0.4725 (0.0405)	0.5534 (0.0433)	0.5024 (0.0405)

Table 1: F_1, FPR and TPR for (adaptive) lasso, truncating (adaptive) lasso and thresholded lasso. Numbers in the table show mean and standard deviations (in parentheses) over 50 replication.

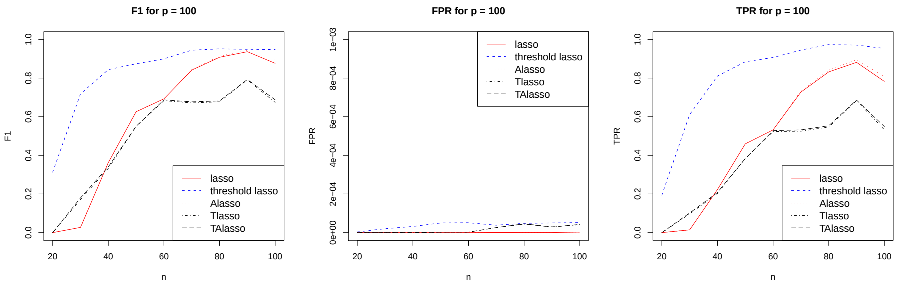


Fig. 3: Phase transition of F_1, FPR and TPR with increase in sample size

3.2 Study of Phase Transition Behavior

In this section, we study the phase transition of three performance metrics as the values of (a) sample size (n) and (b) signal-to-noise ratio ($SNR = \rho/\sigma$) is varied for different combinations of $n,\ p,\ \rho$ and σ. The results showing phase transitions for sample size are based on $p = 100,\ \rho = 0.9,\ \sigma = 0.3$, while those for phase transitions for SNR use $p = 150,\ n = 120, \sigma = 0.3$. Similar results were obtained for other choices of these parameters.

Figure 3 summarizes the phase transition results for sample size n. It can be seen that the phase transition occurs at a much smaller sample size for thresholded lasso compared to (adaptive) lasso and truncating (adaptive) lasso. However, the performances of thresholded lasso and regular lasso are almost similar when n is almost as large as p. For smaller sample sizes, thresholded lasso slightly affects the number of false positives, but greatly improves on the false negatives, resulting in a better F_1 than regular lasso.

Results of phase transition for SNR presented in Figure 4 also indicate that phase transition occurs at a smaller SNR for thresholded lasso compared to (adaptive) lasso and truncating (adaptive) lasso. As in the previous case, the performance of thresholded lasso and regular lasso become more similar as SNR increases. Also, it can be seen that for smaller SNR, thresholded lasso slightly affects the number of false positives while greatly improves the false negatives, which results in significant gain in the overall performance of the proposed estimator in terms of the F_1 measure.

	Alasso	TAlasso	Thlasso
TPR	0.3462 (0.0529)	0.3077 (0.0558)	0.6288 (0.0698)
FPR (×1000)	0.8254 (0.3454)	0.7694 (0.3729)	0.7415 (0.2611)
F_1	0.4729 (0.0591)	0.4338 (0.0654)	0.7251 (0.0581)

Table 2: F_1, FPR and TPR for (adaptive) lasso, truncating (adaptive) lasso and thresholded lasso. Numbers in the table show mean and standard deviations (in parentheses) over 50 replication.

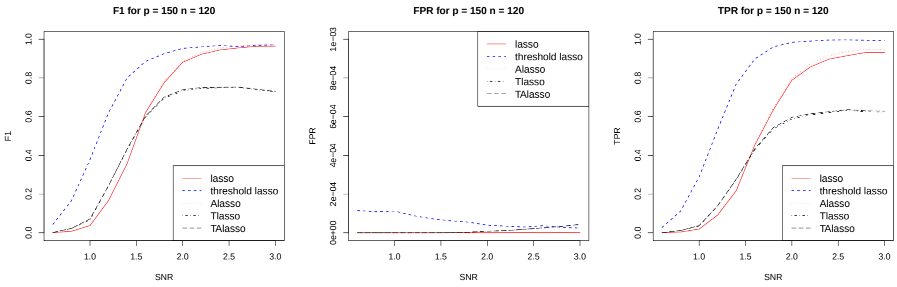


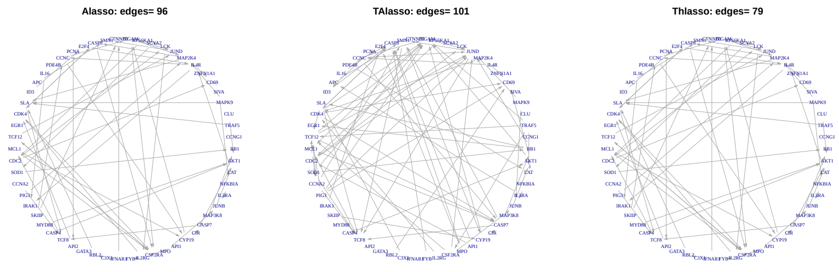
Fig. 4: Phase transition of F_1, FPR and TPR with increase in SNR

Comparison of phase transition behaviors of lasso, truncating lasso and the adaptively thresholded lasso procedures indicates that the proposed estimator provides a better estimate of Granger causal effects over the range of values of n and SNR. In addition, this advantage becomes more significant in problems with smaller sample size and/or signal to noise ratio.

4 Analysis of T-Cell Activation

We illustrate the application of GGC models in reconstructing gene regulatory networks using the time course gene expression data of Rangel et al (2004) on T-cell activation. Activated T-cells are involved in regulation of effector cells (e.g. B-cells) and play a central role in mediating immune response. The data set comprises of $n = 44$ gene expression samples of $p = 58$ genes involved in activation of T-cells, measured over 10 time points. In this study, the activity levels of genes are measured at $t = 0, 2, 4, 6, 8, 18, 24, 32, 48, 72$ hours after stimulation of cells using a T-cell receptor independent activation mechanism. Since changes in regulations often occur at early stages of activation, and to simplify the analysis from the unbalanced experiments, we consider only the earliest 5 time points.

Estimated networks of T-cell activation using the adaptive lasso, the truncating adaptive lasso and the thresholded lasso estimators are shown in Figure 5. The tuning parameters for different estimators are determined as in Section 3, where the value of σ is estimated using the standard pooled estimate. Lasso and truncating lasso estimates provided similar estimates to their adaptive counterparts and considering the advantages of the adaptive estimators over the regular estimators are not presented. The networks in Figure 5 are obtained by drawing an edge between gene i and gene j whenever there is an nonzero element in one of the adjacency matrices $\hat{A}^t_{ij}, T - \hat{d} \leq t \leq T - 1$. Comparison of the estimated networks reveals a significant overlap between the adaptive lasso and thresholded lasso estimates, whereas the truncating adaptive lasso estimate seems to give a different estimate. This is highlighted by the summary measures in Table 3, where the total number of edges in each



(a) Adaptive Lasso (b) Truncating Lasso (c) Thresholded Lasso

Fig. 5: Estimated Gene Regulatory Networks of B-cell activation. Edges indicate nonzero entries in the estimated adjacency matrix in at least one time lag.

network, along with the structural Hamming distance (SHD) between pairs of two networks, defined as the number of edges different between each two networks, are given.

The striking difference between the estimated regulatory networks using the truncating lasso estimate raises the question of whether the decay condition necessary for the performance of the truncating lasso estimator is satisfied. Although the true regulatory mechanism in this biological system is unknown, the gray-scale images of the estimated adjacency matrices in Figure 6 suggest that in this case the decay condition may be indeed violated. This example underscores the advantage of our newly proposed estimator in cases where the conditions required for the truncating lasso estimate of S-M are not met.

5 Discussion

Time course gene expression data provide a valuable source of information for the study of biological systems. Simultaneous analysis of changes in expressions of thousands of genes over time reveals important cues to the dynamic behavior of the organism and provides a unique window for discovering regulatory interactions among genes. A main challenge in applying statistical models

	Alasso	TAlasso	Thlasso
Alasso	(96)	–	–
TAlasso	99	(101)	–
Thlasso	35	102	(79)

Table 3: Structural Hamming Distance between different estimates of the T-cell regulatory network. Diagonal numbers in parentheses show the total number of edges in each network.

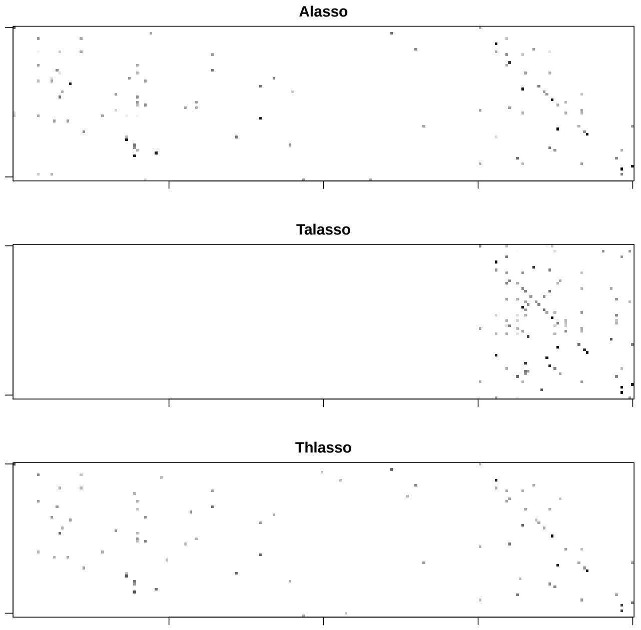


Fig. 6: Adjacency Matrices of Estimated B-Cells Networks.

for inferring regulatory networks from time course gene expression data stems from the unknown order of the time series. Simplified methods that ignore effects of genes from time points farther in the past may suffer from loss of information, and could fail to include significant regulatory interactions that are manifested after a long time lag. In contrast, methods that incorporate all of the past information may suffer from an unnecessary curse of dimensionality, and could result in inferior inference especially when the sample size is small.

To overcome this challenge, we proposed a new penalized estimation method for inferring gene regulatory networks from time series observations, based on adaptive thresholding of lasso estimates. The proposed estimator builds upon the previously proposed truncating lasso estimator Shojaie and Michailidis (2010a). Both of these estimators attempt to simultaneously estimate the order of the VAR model and the structure of the network, under two different structural assumptions. While the truncating lasso estimate is based on the assumption that the effects of genes on each other decay over time, the newly proposed adaptively thresholded lasso estimator relies on a less stringent structural assumption that sets a lower bound on the number of edges in the adjacency matrix of the GGC at each time point (see Section 2.2.2 for a formal statement of this assumption). The relaxation of the decay assumption allows the new estimator to correctly estimate the order of the time series in a broader class of models. However, while the truncating lasso penalty may fail in situations where the decay assumption is violated, it offers advantages in favorable settings.

A natural question therefore arises on the choice of the appropriate penalty for simultaneous estimation of the order of the time series and the structure of the GGC model. The truncating lasso penalty can be advantageous if its underlying assumption is satisfied, but its performance degrades markedly if it does not hold. In absence a formal methodology for determining which of the two assumptions may be more appropriate, the regular (adaptive) lasso estimate can guide the user: if the estimate from the (adaptive) lasso clearly supports the decay assumption, then one could apply the truncating lasso penalty, otherwise, the thresholded lasso penalty provides a more reliable estimate of the GGC.

Acknowledgements We would like to thank two anonymous reviewers whose comments lead to significant improvements in the presentation of the paper. AS and GM would like to acknowledge the support from the Statistical and Applied Mathematical Sciences Institute (SAMSI) as participants in the program on Analysis of Complex Networks. This work was partially supported by NIH grant 1RC1CA145444-0110.

References

Arnold A, Liu Y, Abe N (2007) Temporal causal modeling with graphical granger methods. In: Proceedings of the 13th ACM SIGKDD, pp 66–75

Basu S, Shojaie A, Michailidis G (2011) Incorporating group structure in estimation of graphical Granger causality. Tech. rep., Department of Statistics, University of Michigan

Bickel P, Ritov Y, Tsybakov A (2009) Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37(4):1705–1732

Fujita A, Sato J, Garay-Malpartida H, Yamaguchi R, Miyano S, Sogayar M, Ferreira C (2007) Modeling gene expression regulatory networks with the sparse vector autoregressive model. BMC Systems Biology 1(1):39

van de Geer SA, Bühlmann P (2009) On the conditions used to prove oracle results for the Lasso. Electron J Stat 3:1360–1392

Granger C (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica pp 424–438

Lozano A, Abe N, Liu Y, Rosset S (2009) Grouped graphical Granger modeling for gene expression regulatory networks discovery. Bioinformatics 25(12):i110

Lütkepohl H (2005) New introduction to multiple time series analysis. Springer

Meinshausen N, Yu B (2009) Lasso-type recovery of sparse representations for high-dimensional data. The Annals of Statistics 37(1):246–270

Mukhopadhyay N, Chatterjee S (2007) Causality and pathway search in microarray time series experiment. Bioinformatics 23(4):442

Murphy K (2002) Dynamic Bayesian networks: representation, inference and learning. PhD thesis, University Of California

Ong I, Glasner J, Page D, et al (2002) Modelling regulatory pathways in E. coli from time series expression profiles. Bioinformatics 18(Suppl 1):S241–S248

Opgen-Rhein R, 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

Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge Univ Press

Perrin B, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alche Buc F (2003) Gene networks inference using dynamic Bayesian networks. Bioinformatics 19(suppl 2):138–148

Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild D, Falciani F (2004) Modeling t-cell activation using gene expression profiling and state-space models. Bioinformatics 20(9):1361

Raskutti G, Wainwright MJ, Yu B (2010) Restricted eigenvalue properties for correlated Gaussian designs. J Mach Learn Res 11:2241–2259

Shojaie A, Michailidis G (2010a) Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics 26(18):i517–i523

Shojaie A, Michailidis G (2010b) Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika 97(3):519–538

Wasserman L, Roeder K (2009) High dimensional variable selection. Annals of statistics 37(5A):2178

Yamaguchi R, Yoshida R, Imoto S, Higuchi T, Miyano S (2007) Finding module-based gene networks with state-space models-Mining high-dimensional and short time-course gene expression data. IEEE Signal Processing Magazine 24(1):37–46

Zhou S (2010) Thresholded lasso for high dimensional variable selection and statistical estimation. Arxiv preprint arXiv:10021583