Network Granger Causality with Inherent Grouping Structure

George Michailidis

Department of Statistics and EECS, The University of Michigan

Joint work with Sumanta Basu and Ali Shojaie

BIRS Workshop
Current Challenges in Statistical Learning
December 2011

Motivation

Objective: Want to discover regulatory interactions from time-course data.

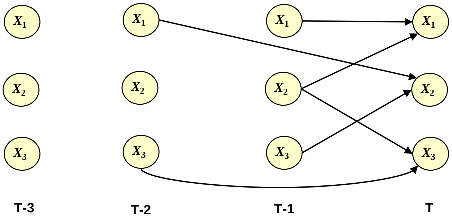
A suitable framework for infering such mechanisms is that of Granger causality.

Granger Causality

- A time series X is said to Granger-cause Y if it can be shown, usually through a series of F-tests on lagged values of X (and with lagged values of Y also known), that those X values provide statistically significant information about future values of Y.
- Granger-causality does not imply true causality; it is built on correlations.
- Recent work extends the framework beyond Gaussian rv's.

Network Granger Causality: Illustration

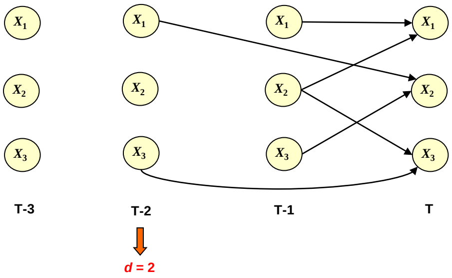
p variables observed over T time points



Network Granger Causality: Illustration

p variables observed over T time points

n_t iid observations at each time point



Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\mathsf{T}$
- Graphical Granger Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\mathsf{T}$
- Graphical Granger Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- Directed Acyclic Graph (DAG) with $(d+1) \times p$ variables, corresponding to a VAR model of order d with p variables.

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\mathsf{T}$
- Graphical Granger Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- Directed Acyclic Graph (DAG) with $(d+1) \times p$ variables, corresponding to a VAR model of order d with p variables.
- Often $d \ll T$, but not known, so $d = T - 1$ is used, many variables for large T.

Previous work on GC in a high dimensional setting

- The concept of Granger causality has been used in discovering regulatory mechanisms by Fujita et al (2007) and Mukhopadhyay and Chatterjee (2007)
- Penalized model used in Lozano et al. (2009) for grouping effects over time
- Penalized model used in Arnold et al. (2007) in a financial application

NGC and The Truncating Lasso Penalty

To avoid increasing the number of variables, need to estimate the order of the time series.

NGC and The Truncating Lasso Penalty

To avoid increasing the number of variables, need to estimate the order of the time series.

$\mathscr{X}^t$: data at time t

$$\underset{\theta^t \in \mathbb{R}^p}{\operatorname{argmin}}\, n^{-1} \| \mathscr{X}_i^T - \sum_{t=1}^{d} \mathscr{X}^{T-t} \theta^t \|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} |\theta_j^t| w_j^t$$

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

where M is a large constant, and β is the allowed false negative rate (FNR).

NGC and The Truncating Lasso Penalty

To avoid increasing the number of variables, need to estimate the order of the time series.

$\mathscr{X}^t$: data at time t

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

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

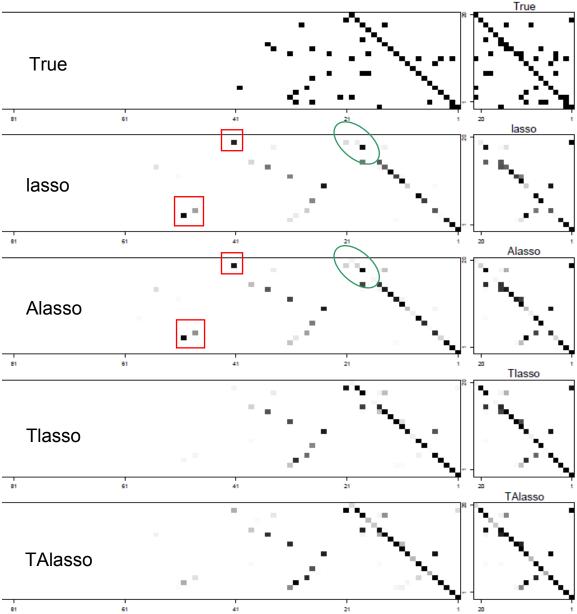
where M is a large constant, and β is the allowed false negative rate (FNR).

We propose the following value of λ that controls a version of the false positive rate (FPR):

$$\lambda(\alpha) = 2n^{-1/2} Z^*_{\frac{\alpha}{2dp^2}}$$

where Z_q^* is the $(1-q)$-th quantile of the standard normal distribution.

Illustrative Example



Properties of the estimator

- Under certain regularity conditions, if the Granger-causal effects decay over time and vanish, then in high-dimensional sparse settings
    (i) the probability of false positives is exponentially small,
    (ii) the probability of false negatives converges to the user-defined value β.
    (iii) the order of the time series is correctly estimated with probability converging to $1 - \beta$.

Asymptotics for the Truncating Lasso Estimator

Theorem

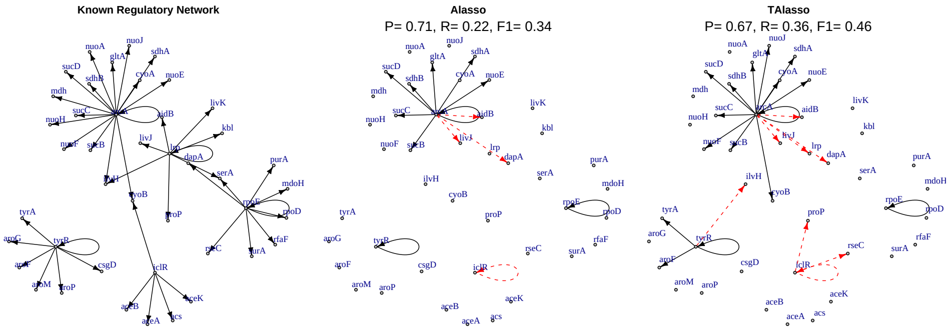
Let s be the total number of true edges in the graphical Granger model and suppose that for some $a > 0$, $p = p(n) = O(n^a)$ and $|\mathrm{pa_i}| = \mathrm{O(n^b)}$, where $sn^{2b-1}\log n = o(1)$ as $n \to \infty$. Moreover, suppose that there exists $\nu > 0$ such that for all $n \in \mathbb{N}$ and all $i \in V$, $\mathrm{Var}\left(X_i^T | X_{1:p}^{T-d:T-1}\right) \geq \nu$ and there exists $\delta > 0$ and some $\xi > b$ such that for every $i \in V$ and for every $j \in \mathrm{pa_i}$, $|\pi_{ij}| \geq \delta n^{-(1-\xi)/2}$, where π_{ij} is the partial correlation between X_i and X_j after removing the effect of the remaining variables. Assume that $\lambda \asymp dn^{-(1-\zeta)/2}$ for some $b < \zeta < \xi$ and $d > 0$, and the initial weights are found using lasso estimates with a penalty parameter λ^0 that satisfies $\lambda^0 = O(\sqrt{\log p/n})$. Also, for some large positive number g, let $\Psi^t = g\exp\left(nI\{\|A^{(t-1)}\|_0 < p^2\beta/(T-t)\}\right)$ (i.e. $M = ge^n$). Then if true causal effects diminish over time,

(i) *With probability asymptotically larger than $1-\beta$, true Granger-causal effects and the order of the VAR model are correctly determined.*

(ii) *With probability converging to 1, no additional causal effects are included in the model and the signs of causal effects are correctly estimated.*

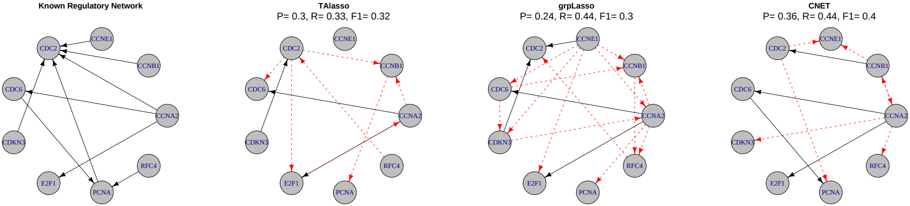
Example I: Gene Regulatory Networks of Yeast

5 Transcription Factors, 37 genes ($p = 42$), 8 time points
$d = 2$



Example II: Gene Network of HeLa Cells

9 genes, 47 time points
$d = 3$



An Adaptive Thresholding Estimation Strategy

The decay assumption for the truncating lasso plays a crucial role.

What if it is violated?

An alternative strategy is based on adaptive thresholding.

Adaptive Thresholding Algorithm

1. Obtain through the regular lasso, estimates of the adjacency matrices $\tilde{A}_t(\lambda_n)$.
2. Define $\Psi^t = \exp(MI(||\tilde{A}^t||_0 < p^2\beta/(T-1))$.
3. Set $\tilde{A}^t_{ij} = \tilde{A}^t_{ij} I(|\tilde{A}^t_{ij}| \geq \tau\Psi^t)$.
4. Estimate $\hat{d} = \max_t\{||\tilde{A}^t||_0 \geq p^2\beta/(T-1)\}$.

Guidelines for tuning parameters:

1. $\lambda_n = c_1\sigma\lambda_0$
2. $\tau = c_2\sigma\lambda_0$

where $\lambda_0 = \sqrt{2log(p)/n}$.

Asymptotic Properties: Preliminaries

1. Let $\tilde{X}$ be the $n \times p(T-1)$ matrix of past observations
2. $\Lambda_{\min}(m) = \min_{v \neq 0, ||v||_0 \leq m} \frac{||\tilde{X}v||_2^2}{n||v||_2^2} > 0$
3. $s = \max_i |pa_i|$ maximum number of parents for any node
4. $a_0 = \min_{1 \leq t \leq d} \min_{1 \leq i,j \leq p, A_{ij} \neq 0} |A_{ij}^t|$
5. Restricted Eigenvalue Condition: Define $K(s,k)^{-1} = \min_{J \subset V, |J| \leq s} \min_{||v_{J^c}||_1 \leq k||v_J||_1} \frac{||\tilde{X}v||_2}{\sqrt{n}||v_J||_2} > 0$.

Asymptotic Properties: Main Result

Theorem

In a VAR(d) *with independent Gaussian noise with variance* σ^2*, suppose RE(*$\tilde{X}$*) 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* $K(s,3)$ *and further for* $0 < \xi < 1$*, we have*

$$|E| < \xi p^2/(T-1)$$

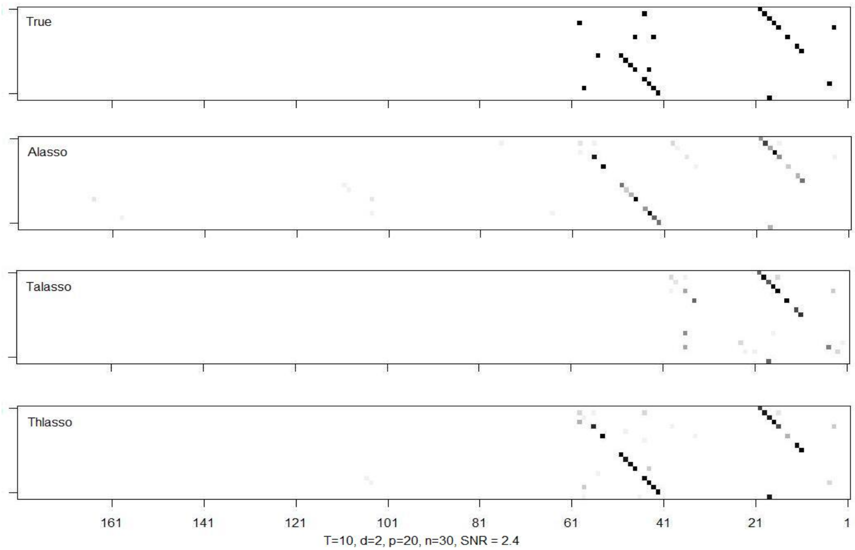
then with prob at least $1-(\sqrt{\pi \log p}p^{\theta})^{-1}$ *the following hold with thresholding parameter* $\beta \leq \xi$*:*

(i) *False positive rate* $\leq (bs)/(p-s)$ *for some constant* b *(control of Type-I error)*

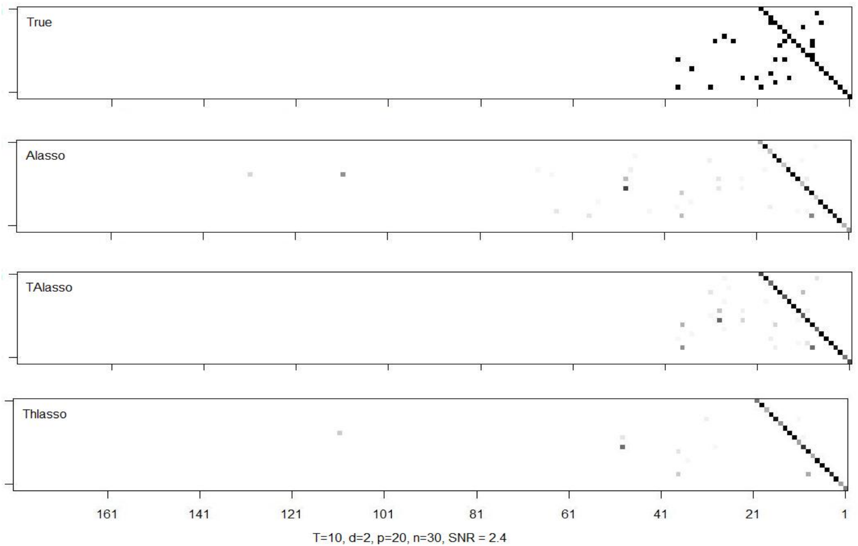
(ii) *For any* $\varepsilon > 0$*, False negative rate*$< \varepsilon$ *(control of Type-II error)*

(iii) *Order consistency:* $\hat{d} \to d$*.*

Numerical Illustration I

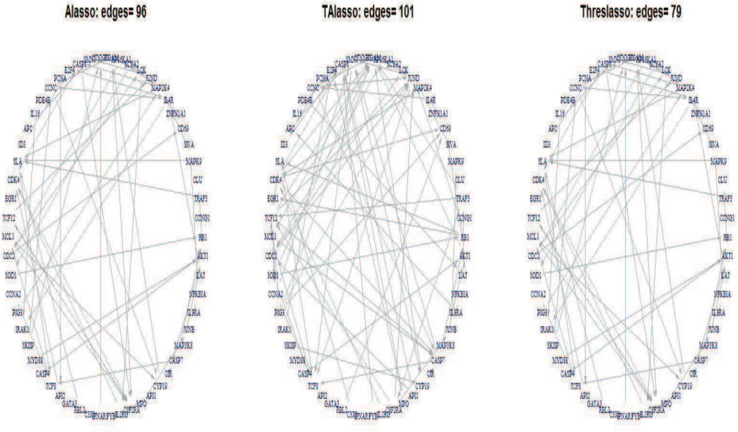


Numerical Illustration II



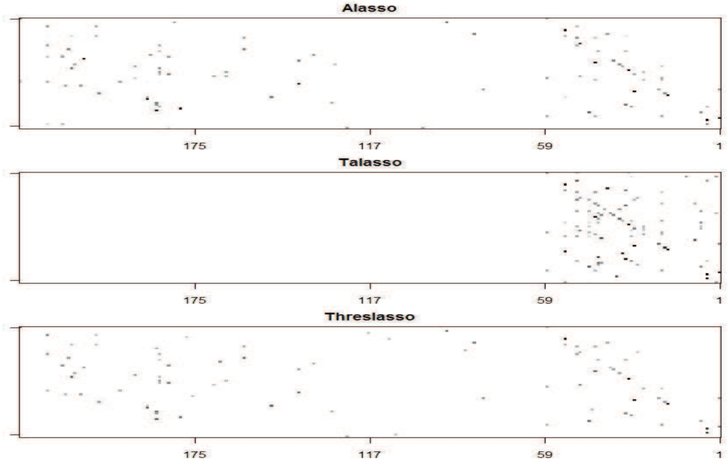
An Application to T-cell Activation

58 genes, 5 time points, n=44, $d \approx 4-5$



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

An Application to T-cell Activation



NGC with Group Sparsity

- Incorporate grouping structure into the NGC problem
  e.g. pathway information
- The node set N_G is partitioned into G non-overlapping groups $\mathscr{G}_1, \ldots, \mathscr{G}_G$ with $\left|\mathscr{G}_g\right| = k_g$ and $k_0 = \max_{1 \leq g \leq G} k_g$.
- Nodes from same group have either all zero or all non-zero effect on other nodes (signs of effects may vary)
- Last condition can be relaxed with the application of a thresholding step (allows for small misspecifications at the group level)

Group NGC estimates

- For $i = 1, \ldots, p$,

$$\begin{aligned}
\hat{A}_{i:}^{1:T-1} &= arg \min_{\theta^1, \theta^2, \ldots, \theta^{T-1} \in \mathbb{R}^p} \frac{1}{2} \| \mathscr{X}_{:i}^{T} - \sum_{t=1}^{T-1} \mathscr{X}^{T-t} \theta^t \|_2^2 \\
&\quad + \lambda_n \sum_{t=1}^{T-1} \Psi^t \sum_{g=1}^{G} \sqrt{k_g} w_{i,g}^t \| A_{i:g}^t \|_2 \qquad (1) \\
\hat{d} &= \max_{1 \le t \le T-1} \left\{ t : \hat{A}^t \neq \mathbf{0}_{p \times p} \right\} \qquad (2)
\end{aligned}$$

- $\mathscr{X}^t$: $n \times p$ design matrix corresponding to t-th time point
- $w_{i:g}^t$: weigths for adaptive version
- Ψ^t: truncating/thresholding factors

Variants of NGC estimates

- Regular: $\Psi^t = 1$, $w^t_{i,g} = 1$
- Truncating: $\Psi^1 = 1$, $w^t_{i,g} = 1$, for some very large Δ,
  $\Psi^t = exp[\Delta n I\{\sum_{g=1}^{G} I_{\{\|A^{t-1}_{:g}\|_0 > 0\}} < G^2\beta/(T-t)\}]$, $t \geq 2$
- Adaptive: $w^t_{i,g} = min\{1, \|\tilde{A}^t_{i:g}\|_2^{-1}\}$ where $\tilde{A}^t$ are the estimates from Regular GGC.
- Thresholded: For every $t = 1, \ldots, T-1$, if $j \in \mathscr{G}_g$,
  $\hat{A}^t_{ij} = \tilde{A}_{ij} I\left\{\left|\tilde{A}^t_{ij}\right| \geq \delta_1 \left\|\tilde{A}^t_{i:g}\right\|_2\right\} I\left\{\left\|\tilde{A}^t_{i:g}\right\|_2 \geq \delta_2\right\}$

Group NGC estimation as a convex optimization problem

- For every $i = 1, \ldots, p$, regular GGC estimate solves a group lasso problem

$$
\begin{aligned}
& \mathbf{Y}^n_{n\times 1} = \mathbf{X}^n_{n\times p}\beta^n_{p\times 1} + \varepsilon^n, \qquad \varepsilon^n \sim n(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n}) \\
& \{1, \ldots, p\} = \cup_{g=1}^{G} \mathscr{G}_g, \qquad |\mathscr{G}_g| = k_g \\
\hat{\beta}^n \quad = \quad & arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{Y}^n - \mathbf{X}^n \beta\|_2^2 + \lambda_n \sum_{g=1}^{G} \sqrt{k_g} \|\beta_g\|_2
\end{aligned} \tag{3}
$$

with $\mathbf{Y}^n = \mathscr{X}_i^T$, $\mathbf{X}^n = [\mathscr{X}^1 : \cdots : \mathscr{X}^{T-1}]$, $\beta^n = vec(A_{i:}^{1:(T-1)})$, $p \leftarrow (T-1)p$, $G \leftarrow (T-1)G$.

Main Results

1. Norm consistency of regression estimates β_t
2. Directional consistency of the group lasso estimates

Restricted Eigenvalue Condition for Group Lasso Estimates

RE condition for Group Lasso (Lounici et al., 2011)

In the regression framework of (3), *RE(q, L) is satisfied if there exists a postitive number* $\phi_{RE} = \phi_{RE}(q) > 0$ *which equals*

$$\min_{\substack{J \subset \mathbb{N}_G \\ |J| \leq q \\ \Delta \in \mathbb{R}^p \setminus \{\mathbf{0}\}}} \left\{ \frac{\|\mathbf{X}^n \Delta\|_2}{\sqrt{n}\|\Delta^J\|_2} : \sum_{g \in J^c} \sqrt{k_g}\|\Delta^g\|_2 \leq L \sum_{g \in J} \sqrt{k_g}\|\Delta^g\|_2 \right\}$$

Norm consistency

ℓ_2 consistency for Group Lasso

In the regression framework of (3)*, suppose* (β^t) *is contained in a set of groups* $J(\beta^n)$ *with at most* q *groups and* $RE(2q,3)$ *holds. Then for any solution* $\hat{\beta}^n$ *of* (3) *with suitably chosen* λ *the following holds with high probability:*

$$\left\|\hat{\beta}^n - \beta^n\right\|_2 \leq \frac{4\sqrt{10}}{\phi_{RE}^2(2q)} \frac{\lambda \sum_{g \in J(\beta^n)} k_g}{\sqrt{q}\sqrt{k_{min}}} \tag{4}$$

A Sufficient Condition for RE in Group NGC

Raskutti et al. (2010) show that if the sample size is “large enough" and $\Lambda_{\min}(\Sigma) > 0$ then RE holds.

Consider a stationary VAR(d) model with spectral matrix operator $f(\theta),\ \theta \in [-\pi, \pi]$. Let $\Sigma = cov(\mathbf{X}^{1:T})$. If the minimum eigenvalue $\mu(\theta)$ and a corresponding eigenvector $v(\theta)$ of $f(\theta)$ are continuous functions of θ, then the minimum eigenvalue of Σ satisfies

$$\Lambda_{min}(\Sigma) > \left(1 + \tfrac{1}{2}\mathbf{v}_{in} + \tfrac{1}{2}\mathbf{v}_{out}\right)^{-1} > 0$$

where $\mathbf{v}_{in} = \max\limits_{1 \le i \le p} \sum_{t=1}^{d} \sum_{j=1}^{p} \left|A_{ij}^{t}\right|$, $\mathbf{v}_{out} = \max\limits_{1 \le j \le p} \sum_{t=1}^{d} \sum_{i=1}^{p} \left|A_{ij}^{t}\right|$

Direction Consistency for Group Lasso Solutions

- Consider a generic group lasso estimate as in (3). Let $S = \{1, \ldots, q\}$, without loss of generality, denote the group indices in $support(\beta^n)$, i.e.,

$$\beta^n = [\beta_1^n, \ldots, \beta_q^n, \mathbf{0}, \ldots, \mathbf{0}],\ \beta_g^n \neq \mathbf{0}\ \forall\ g \in S = \{1, \ldots, q\}$$

- For a vector $\tau \in \mathbb{R}^m \backslash \{\mathbf{0}\}$ define $D(\tau) = \frac{\tau}{\|\tau\|_2}$ and $D(\mathbf{0}) = \mathbf{0}$
- $D(\beta_g^n)$ indicates the direction of influence of β_g^n at a group level as it reflects the relative importance of the influential group members
- Generalizes the notion of sign consistency

Direction Consistency for Group Lasso Solutions

- An estimate $\hat{\beta}^n$ is direction consistent at a rate r_n if there exists a sequence of positive real numbers $\delta_n \to 0$ such that $\delta_n \asymp r_n$ and

$$\mathbb{P}\left(\|D(\hat{\beta}_g^n) - D(\beta_g^n)\|_2 < \delta_n, \ \forall g \in S \right) \to 1 \text{ as } n, p \to \infty$$

- Define $\tilde{S}_g^n = \{ j \in \mathscr{G}_g : \frac{|\hat{\beta}_j^n|}{\|\hat{\beta}_g^n\|_2} > \delta_n \}$ - collection of influential group members within a group $\mathscr{G}_g$ which are detectable with a sample size of n
- If $\hat{\beta}^n$ is direction consistent then

$$\mathbb{P}(D(\hat{\beta}_j^n) = D(\beta_j^n), \ \forall j \in \tilde{S}_g^n, \forall g \in S) \to 1 \text{ as } n, p \to \infty$$

Directional Consistency in Group NGC

Under a group irrepresentable condition and some other regularity ones, we have

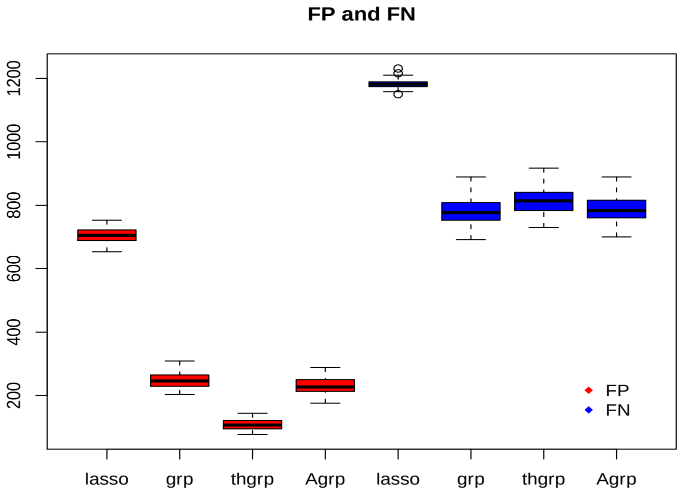
- The index set of the groups for which $\hat{\beta}_g^n \neq 0$ is correctly specified with high probability
- Directional consistency holds with high probability

Selected Numerical Results

Setup:

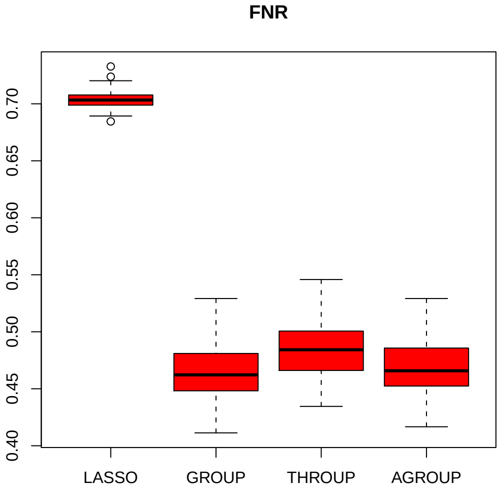
- **Nodes:** $p = 120$ nodes partitioned into $G = 15$ groups of size 8 each
- **Structure of VAR:** $d = 2$, $T = 10$
- **Sample Size:** $n = 150$
- **Network strength:** $TP = 1680$ edges from first two lag
- **Signal Strength:** SNR = 1
- **Performance Criteria:** FPR, FNR, MCC

False Positives and False Negatives



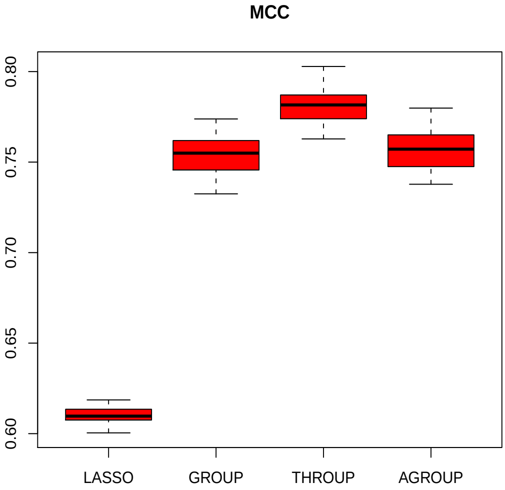
p = 120, G = 15, n = 150, d = 2, TP = 1680, TN = 27120, SNR = 1

False Negative Rate



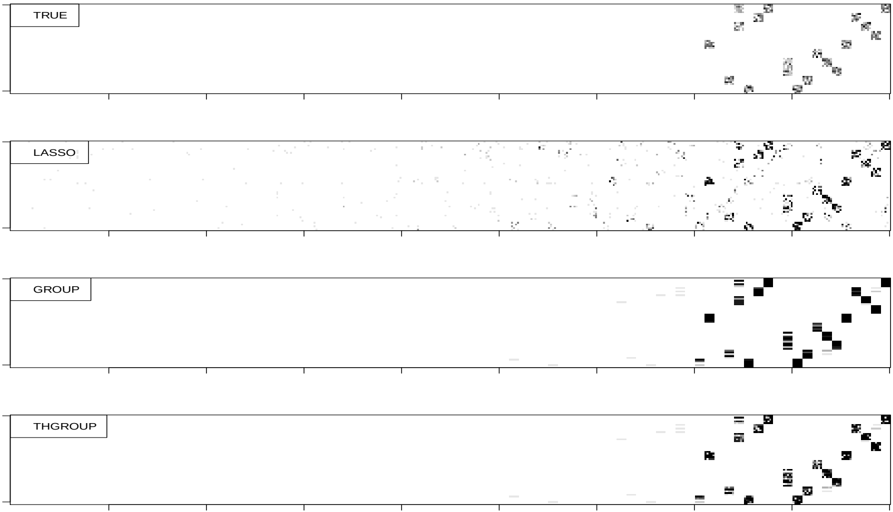
p = 120, G = 15, n = 150, d = 2, TP = 1680, TN = 27120, SNR = 1

Matthews Correlation Coefficient



p = 120, G = 15, n = 150, d = 2, TP = 1680, TN = 27120, SNR = 1

Sample Output: Adjacency Matrices

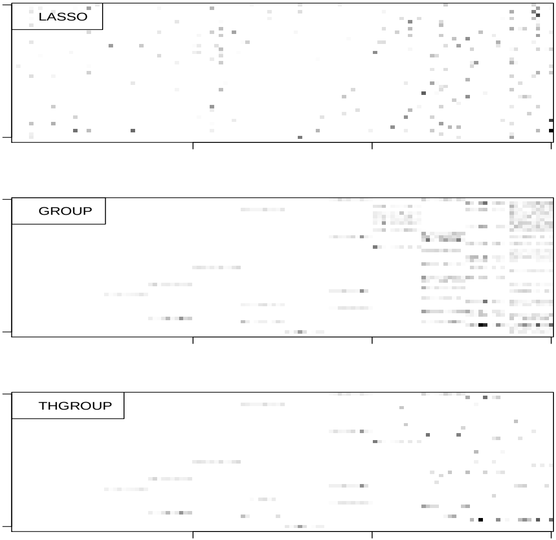


Application to Stock Returns

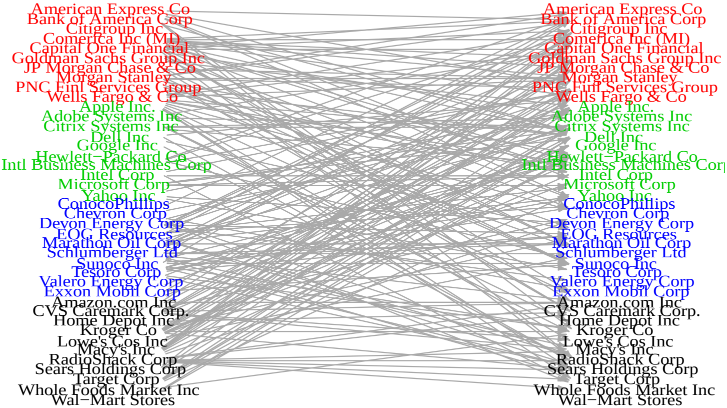
- Daily stock prices (P_t) of $p = 41$ firms from $G = 4$ different categories (Banking, IT, Energy, Retail)
  observed for $T = 4$ days (Sep 21 - Sep 24, 2010) every 5 minutes from 11 am to 3 pm
- Daily log returns $log(R_t) = log(P_t/P_{t-1})$ are calculated to reduce non-stationarity issues
- Stocks at different times of the day ($n = 48$) treated as replicates for that day
- Lasso and group lasso based NGC estimators are used to estimate the network structure of graphical Granger model
- λ chosen by ten-fold cross-validation

Data from *http://wrds-web.wharton.upenn.edu/wrds/*

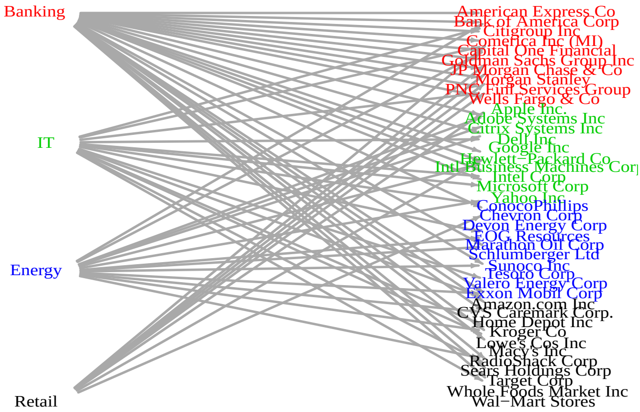
Adjacency Matrices



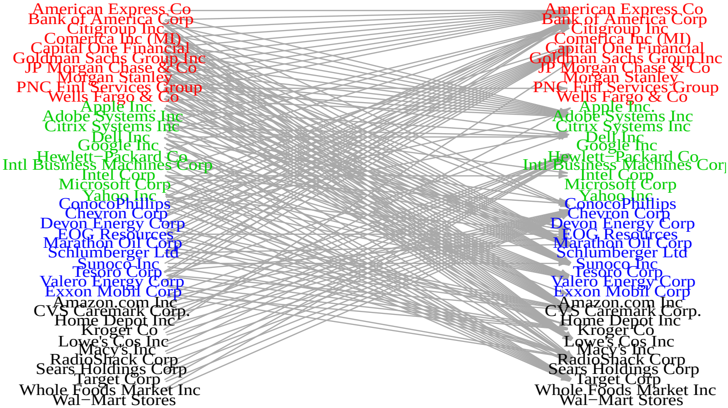
Estimated Network: Lasso



Estimated Network: Group Lasso



Estimated Network: Thresholded Group Lasso



Concluding Remarks

- Network Granger Causality can be useful for discovering temporal regulatory mechanisms
- Grouping structure of variables can be beneficial, especially if coupled with a thresholding step
- Need for correctly estimating the lag of the model
- Truncating (group) lasso performs well, when Granger causal effects decay over time, at the cost of solving a non-convex problem
- Thresholding (group) lasso a worthy alternative
- Asymptotics of pure time series model (no replicates) challenging

- Acknowledgments:
  1. National Institutes of Health

- Acknowledgments:

1. National Institutes of Health

- References:

1. Shojaie & Michailidis (2010a) Penalized Likelihood Methods for Estimation of Sparse High Dimensional DAGs, *Biometrika* 97(3): 519-538
2. Shojaie & Michailidis (2010b) Discovering Graphical Granger Causality using the Truncating Lasso Penalty, *Bioinformatics*, 26(18): i517-i523
3. Shojaie, Basu & Michailidis (2011) Adaptive Thresholding for Reconstructing Regulatory Networks from Time Course Gene Expression Data, to appear in *Statistics in Biosciences*
4. Basu, Shojaie & Michailidis (2011) Discovering Network Granger Causality in Sparse High-dimensional Networks with Inherent Grouping Structure (in preparation)