Dynamic Causal Networks with Multi-scale Temporal Structure

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

Joint work with Xinyu Kang & Apratim Ganguly

Supported in part by AFOSR award 12RSL042.

BOSTON UNIVERSITY

At 10K Feet . . .

Two simple observations:

1. The automated, simultaneous monitoring of each unit in a large complex system has become commonplace.
2. Frequently the resulting data are a high-dimensional multivariate time series.

Implication: The combination of

- *systems* and
- *time series*

perspectives suggests the use of *dynamic network modeling*, a highly active frontier in the field of network analysis.

BOSTON UNIVERSITY

Illustration: Neuroscience

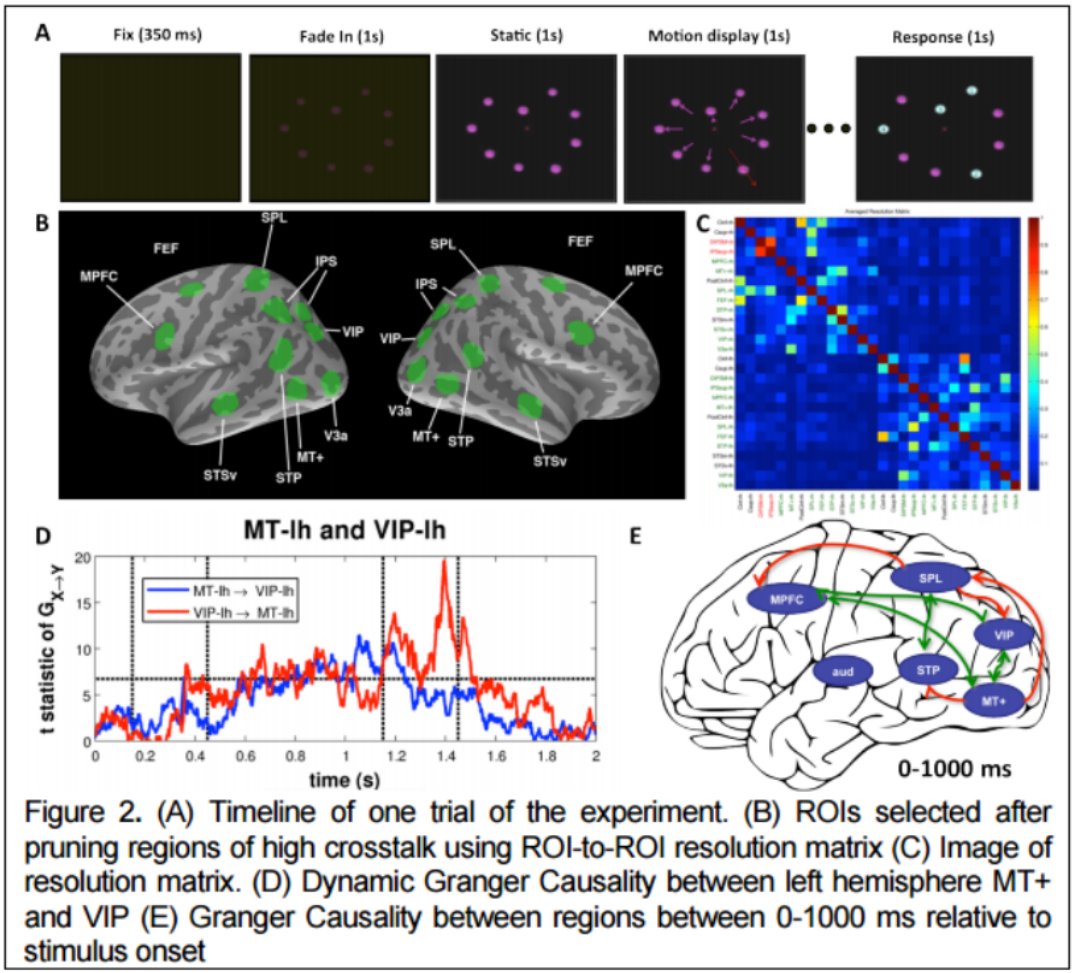
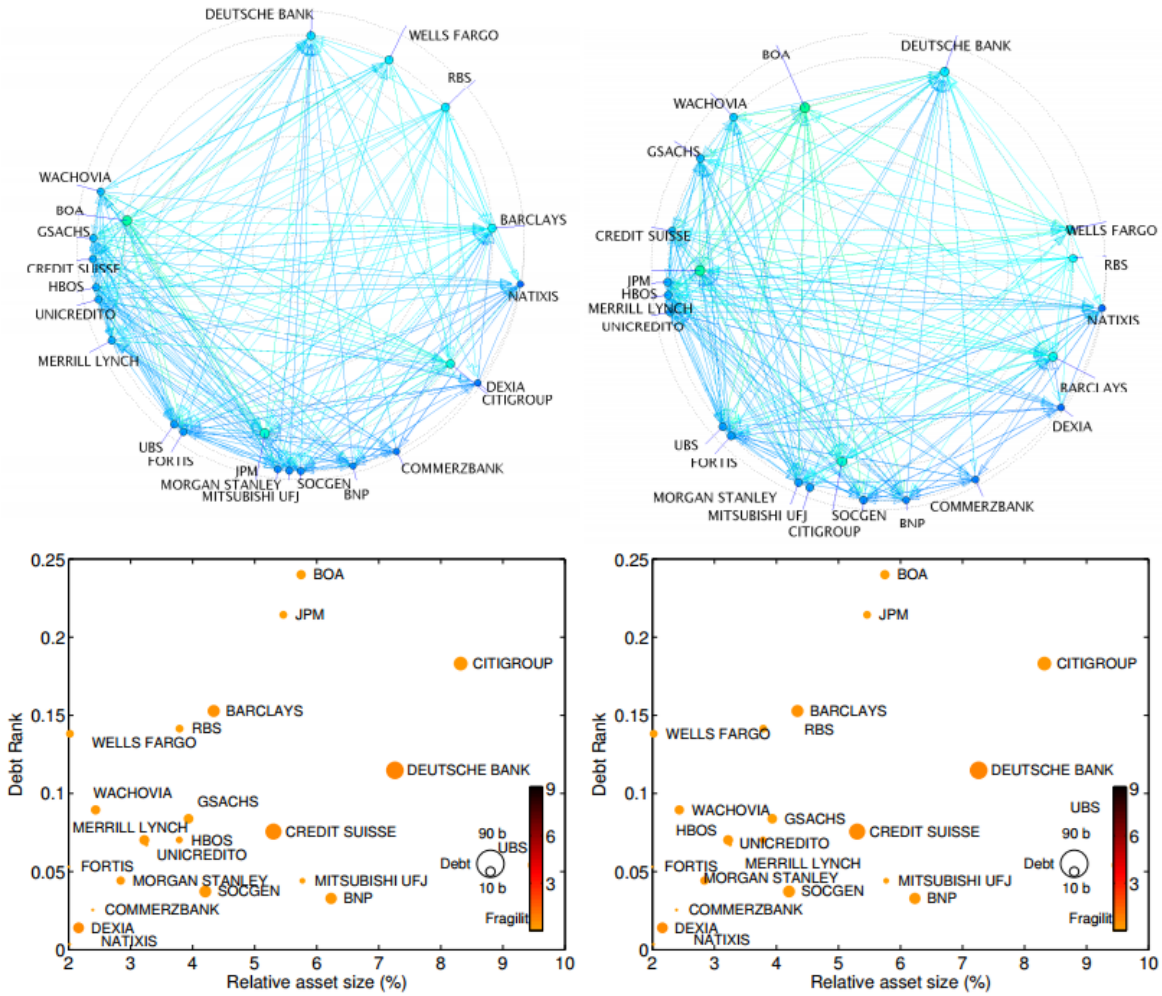


Figure 2. (A) Timeline of one trial of the experiment. (B) ROIs selected after pruning regions of high crosstalk using ROI-to-ROI resolution matrix (C) Image of resolution matrix. (D) Dynamic Granger Causality between left hemisphere MT+ and VIP (E) Granger Causality between regions between 0-1000 ms relative to stimulus onset

Source: Lucia Vaina, BU/BME

Illustration: Economics/Finance



Source: Battiston et al. (2012) *Nature Reports*

Dynamic Network Topology Inference

In these and similar application contexts, there is a basic paradigm at work:

Systems-level question of interest;

$\Rightarrow$ Collection of multivariate time series data;

$\Rightarrow$ Construction of network-based representation of system from data;

$\Rightarrow$ Network-centric answer to systems-level question.

The third step is known as *(association) network topology inference*.

Aside on Network Inference

There are a variety of methods for inference of a network[1]

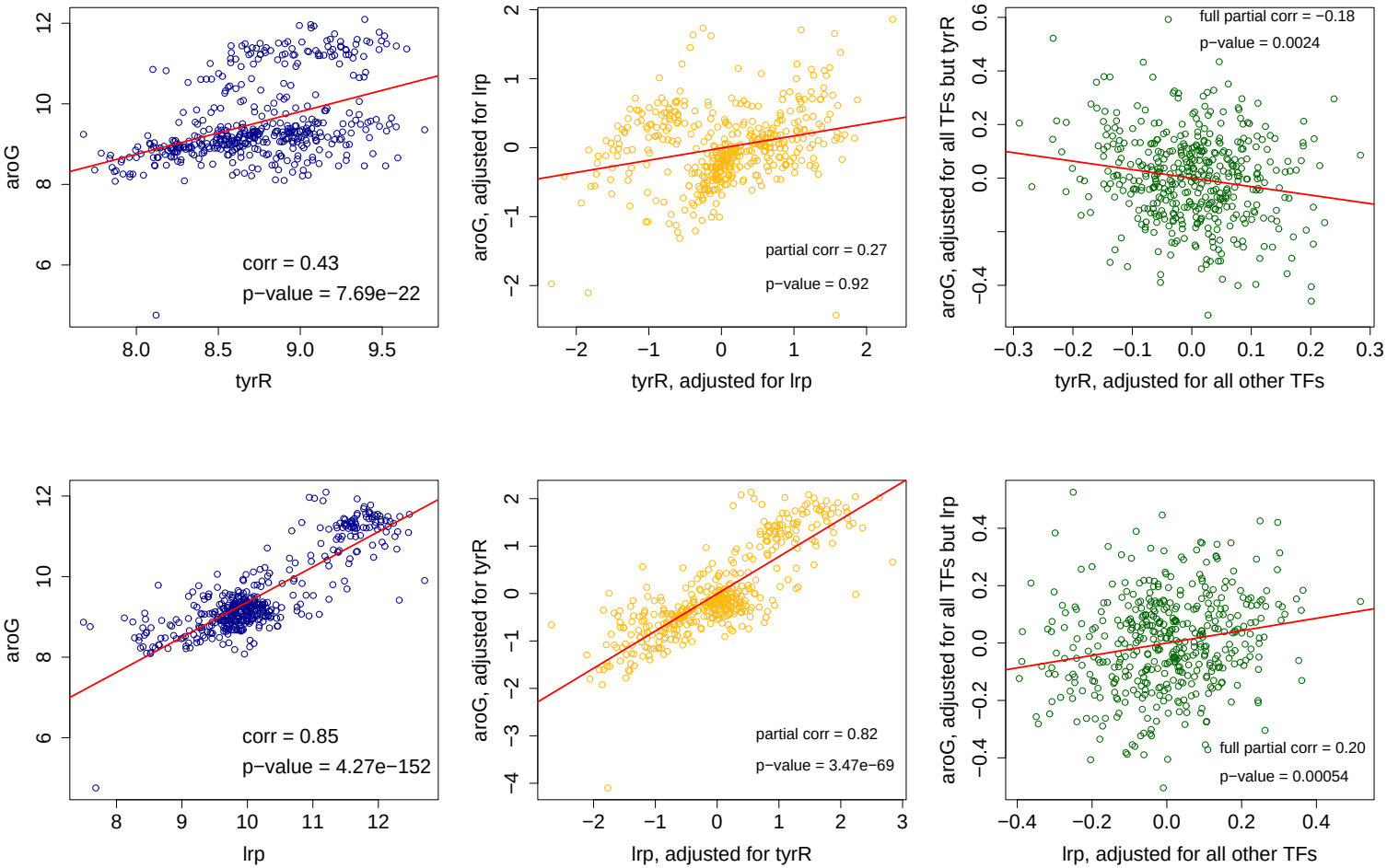
These can be largely categorized by choices in

1. notion of association (e.g., correlation, mutual information, etc.)
2. method of inference (e.g, testing, regression, etc.)
3. working parameters (e.g., significance level, smoothing parameter, etc.)

Note: Most formal methods seek to exploit sparseness typically encountered in empirical networks.

[1] See, for example, Ch7 in either of Kolaczyk (2009). *Statistical Analysis of Network Data*, or Kolaczyk & Csardi (2014). *Statistical Analysis of Network Data with R*.

Illustration with Gene Regulation: Choices Matter!



Correlation | Partial Corr. w/ 1 | Parial Corr. w/ All

BOSTON UNIVERSITY

An Additional Observation: Nature of Temporal Changes

In many contexts, we can expect changes in the system (local or global) across *multiple time scales.*

Examples:

- Neuroscience:
- Finance:

Suggests the need for *multi-scale analysis* . . . a concept well-established in time series analysis, but which has not yet emerged in dynamic network analysis.

BOSTON UNIVERSITY

Our Focus

Motivated by these various observations, we focus on the problem of detecting dynamic connectivity changes across multiple time scales in a network-centric representation of a system, based on high-dimensional multivariate time series observations.

Our approach combines

- Granger causal modeling, with
- partition-based multi-scale modeling

Granger Causal Models

Let $\mathbf{X}_t$ be an N-dim time series, at time t, corresponding to a network of N nodes $v \in V$.

We will adopt a $VAR(p)$ model, i.e.,

$$\mathbf{X}_t = \sum_{\tau=1}^{p} \Theta_\tau \mathbf{X}_{t-\tau} + \epsilon_t \ ,$$

where Θ_τ collects the influence of the nodes on each other at lag τ, and the ϵ_t are independent white noise.

Say that X_v *Granger causes* X_u if $\Theta_\tau(u, v) \neq 0$ for some $\tau = 1, \ldots, p$.

BOSTON UNIVERSITY

From Time Series to Network Graphs

In this setting, the notion of 'network' is made precise through graphs defined as a function of the underlying graphical model. That is, through *conditional independence* relations, which are coded in one-to-one correspondence with the non-zero elements of

$$\Theta = (\Theta_1, \ldots, \Theta_p) \ .$$

Two options:

1. $G = (V, E)$ is a directed graph with an edge from v to u iff $||\theta(u, v)|| \neq 0$, where $\theta(u, v) = (\Theta_1(u, v), \ldots, \Theta_p(u, v))^T$.
2. $G = (V^{(p+1)}, E)$ is a directed multi-graph, with an edge from v in layer τ to u in layer 0 iff $\theta_\tau(u, v) \neq 0$.

We will focus on the first of these options.

BOSTON UNIVERSITY

Time-indexed Directed Graphical Model

Our interest will be in *non-stationary* multivariate time series, and the corresponding representations using dynamic networks.

We adopt a changepoint perspective, so that our model class consists of concatenations of Granger causal $VAR(p)$ models, each with its own Θ constant over a given interval of time.

The result is a time-indexed directed graphical model, from which we define a dynamic network $G_t = (V, E_t)$ in analogy to the stationary case.

Our goal is to infer (i) the changepoints distinguishing the stationary intervals, and (ii) the corresponding edge sets E_t.

Granger-causal Dynamic Network Modeling

A great deal of work has been done recently in modeling multiple time series using causal network types of models.

Extending seminal work of Meinshausen and Buhlmann[2], a selection of representative examples include

- Neighborhood-selection under stationarity, using group-lasso principles[3]
- Network Granger causality with panel data[4]
- Inference of networks defined through long-run correlation[5]
- Adaptive selection of (fixed) window size[6].

Our work takes these ideas in a multi-scale direction.

[2] Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with lasso. *AoS*

[3] Bolstad, A., Van Veen, B. D., & Nowak, R. (2011). Causal network inference via group sparse regularization. *IEEE TSP*.

[4] Basu, S., Shojaie A., & Michailidis, G. (2015). Network granger causality with inherent grouping structure. *JMLR*.

[5] Barigozzi, M., Brownlees, C. (2013). Nets: network estimation for time series. *Available at SSRN*.

[6] Long, CJ., Brown, EN., Triantafyllou, C., Aharon, I., Wald, LL., Solo, V. (2005). Nonstationary noise estimation in functional MRI. *NeuroImage*.

Multi-scale Modeling

Starting in the late 1980's (building on threads of work going back at least to the early 1900's), there was an explosion of development on methods of multi-scale modeling that were

- mathematically principled;
- computationally efficient; and, often,
- domain/problem-specific.

The quintissential example is that of methods utilizing transformations with respect to bases of *wavelets*.

Wavelets & Recursive Partitioning: A connection

While the development of wavelet-based methods for standard signal and image analysis applications proceeded apace, the development of extensions for less-traditional settings like

- non-normal noise, and
- structured data (e.g., manifolds, graphs, etc.)

proved decidedly more challenging[7].

Helpful in many contexts was a fundamental result of Donoho[8] relating

1. methods of recursive (dyadic) partitioning, and
2. selection of a best-orthonormal basis,

where the basis is selected from a class of (unbalanced) Haar bases.

[7] But, nevertheless, was/is being thoroughly explored as well!

[8] Donoho, D.L. (1997). CART and Best-Ortho Basis: A Connection. *AoS*.

Illustration: Multiscale GLMs

Hence, partition-based methods – which can be more amenable than wavelets to adaptation in nontraditional settings – are nevertheless effective structures upon which to build multi-scale models and methods.

A useful illustration of this principle is in the context of generalized linear models (GLMs)[9]

Given independent observations $y_1, \ldots, y_n$, where

$$p_\theta(y_i|t_i) = \exp\left\{\frac{y_i \cdot \theta(t_i) - b[\theta(t_i)]}{\tau} + c(y,\tau)\right\} \quad ,$$

estimate $\theta(\cdot)$, a member of some inhomogeneous function class (e.g., Besov ball).

[9] Kolaczyk, E.D, and Nowak, R.D (2005). Multiscale generalized linear models for nonparametric function estimation. *Biometrika*.

Multiscale GLMs (cont)

Estimate $\theta = (\theta_1, \ldots, \theta_n)$ as

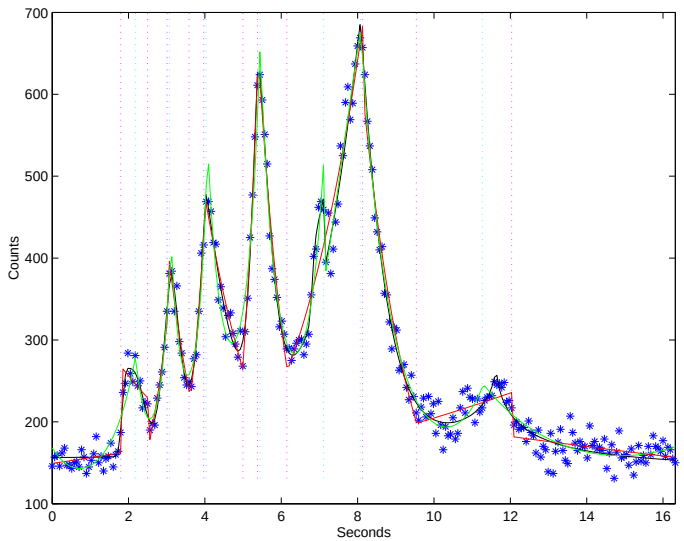
$$\hat{\theta} \equiv \arg \max_{\mathcal{P} \preceq \mathcal{P}^*_{Dy}} \max_{\theta' \in PP(\mathcal{P}; D)} \left\{ \ell(\theta') - 2\lambda \#(\mathcal{P}) \right\} \quad ,$$

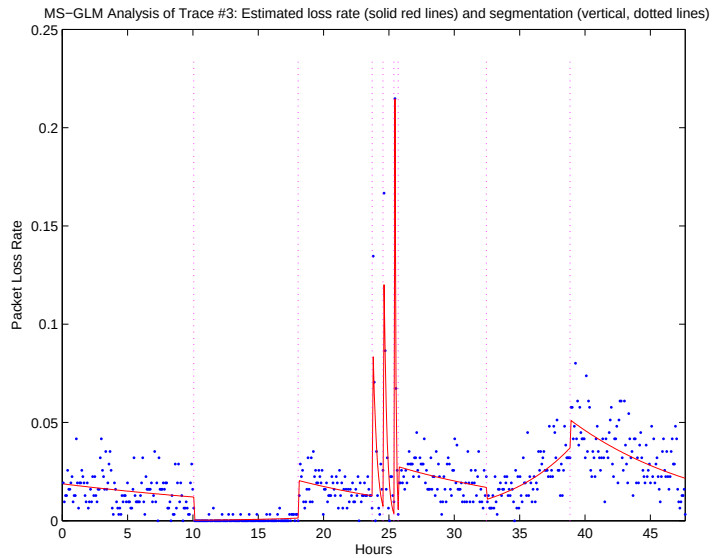
where

- $\mathcal{P}^*_{Dy}$ is a complete recursive dyadic partition;
- $\mathcal{P} \preceq \mathcal{P}^*_{Dy}$ denotes a sub-partition;
- $PP(\mathcal{P}; D) \equiv \{$ Piece-wise polynomials, of order D on $\mathcal{P}\}$; and
- $\ell(\theta)$ is the log-likelihood

Multiscale GLMs: Results Summary

- $O(n)$ alg[10] and near-optimal risk behavior
- Extension to recursive partitioning, with $O(n^3)$ alg
- Application to gamma-ray bursts & Internet packet loss.





[10] Formally, $O(n)$ model comparison steps, where complexity of comparison step depends on models being fit.

Our Contribution

We propose

1. a multi-scale Granger-based dynamic network model, and
2. a corresponding method of network topology inference

that captures the dynamics of a system in a manner sensitive to changes at multiple time scales, while encouraging sparsity of network connectivity.

Key elements of our framework:

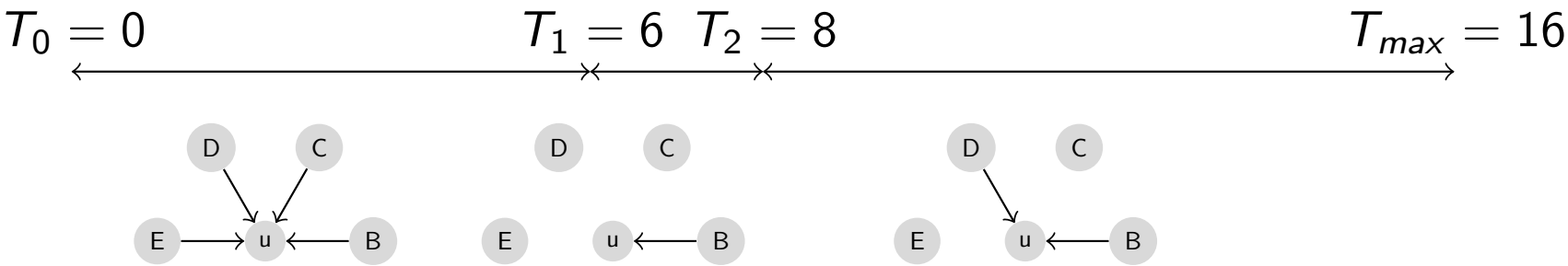
- We partition the non-stationary time space into blocks at various scales, with stationary assumed within each block;
- We do neighborhood selection with the group lasso, within each block.

BOSTON UNIVERSITY

A Cartoon Version

We estimate network topology *neighborhood by neighborhood.*

WLOG, consider the local neighborhood of node/series u.



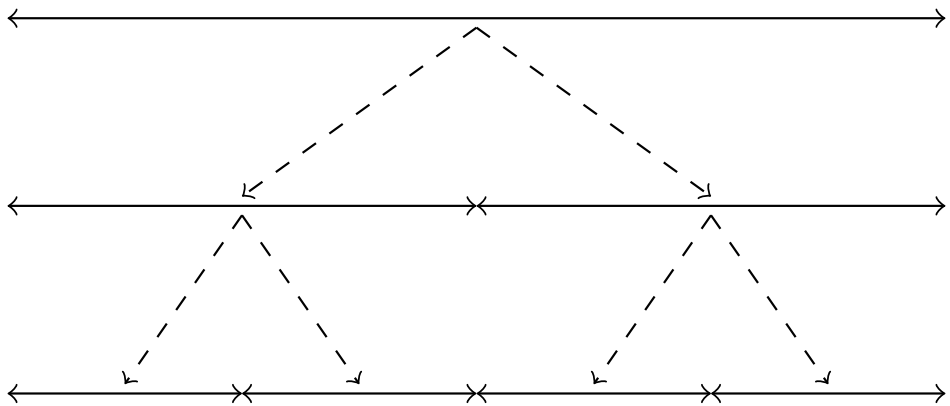
What we do:

- Changepoint Detection: Estimate the times T that the changes happened.
- Neighborhood Selection: Given the estimated change points, infer the neighborhood structure from time 0 to $\hat{T}_1$ and from time $\hat{T}_1 + 1$ to $\hat{T}_2$ and so on.

BOSTON UNIVERSITY

Parameterization of Changepoints

To capture potential change points, we partition time into candidate stationary intervals using RDP[11].



For appropriate choice of cost function, our problem can be solved in a dynamic programming fashion, in strict analogy to, e.g., MS-GLMs.

[11] Extension to RPs straightforward, with increased computational cost.

Recovery of Neighborhood Structure

Within a candidate stationary interval, WLOG with n time points, we do neighborhood selection with group lasso, solving for

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \|\mathbf{X}_u - \sum_{v \in V \setminus \{u\}} \mathbf{X}_{(v)} \theta_{(v)}\|_2^2 + \lambda \sum_{v \in V \setminus \{u\}} \|\theta_{(v)}\|_2.$$

Here

- $\mathbf{X}_u$ is the $n \times 1$ time series observed at node u;
- $\mathbf{X}_{(v)}$ is the $n \times p$ matrix formed from $\mathbf{X}_v$ at lags $\tau = 1, \ldots p$;
- $\theta_{(v)} = \theta(u, v) = (\Theta_1(u, v), \ldots, \Theta_p(u, v))^T$;
- λ is a smoothing parameter (TBD).

Write the *group lasso penalized likelihood (glPL)* as

$$\hat{glPL} = \frac{1}{n} \left\| \mathbf{X}_v - \mathbf{X}\hat{\theta} \right\|_2^2 + \lambda \sum_{v \in V \setminus \{u\}} \left\| \hat{\theta}_{(v)} \right\|_2 .$$

BOSTON UNIVERSITY

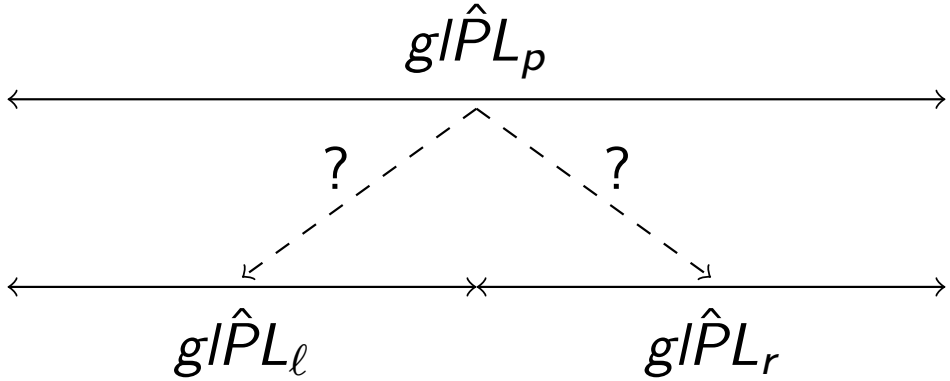
Selecting a Multiscale Dynamic Granger Causal Model

Our overall problem can be set up as solving for

$$\hat{\theta} \equiv \arg \max_{\mathcal{P} \preceq \mathcal{P}^*_{Dy}} \max_{\theta' \in PGC(\mathcal{P};p)} \left\{ \ell(\theta') - \text{Pen}(\theta') \right\} \quad ,$$

where $PGC(\mathcal{P}; p)$ is the collection of 'piecewise Granger Causal' models on $\mathcal{P}$ of lag p.

Key Algorithmic Insight: This may be solved using dyanamic programming, where at each potential break point, we solve the following problem:



Split if:

$$gl\hat{P}L_\ell + gl\hat{P}L_r < gl\hat{P}L_p$$

Consistency of Splitting

Define the truth to be: $\mathbf{X}_u = \begin{pmatrix} \mathbf{X}_u^{(\ell)} \\ \mathbf{X}_u^{(r)} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{(\ell)}\theta^{(\ell)} + \epsilon^{(\ell)} \\ \mathbf{X}^{(r)}\theta^{(r)} + \epsilon^{(r)} \end{pmatrix}$, where

$\mathbf{X}_u^{(\ell)}$ are the first $n/2$ observations and $\mathbf{X}_u^{(r)}$ the last $n/2$ observations.

Theorem

Under appropriate assumptions, for any sequence λ_n such that $\lambda_n \to 0$ and $\lambda_n n^{1/2} \to +\infty$,

$$\mathbb{P}_{\theta^{(\ell)}=\theta^{(r)}} \left(gl\hat{P}L_\ell + gl\hat{P}L_r \geq gl\hat{P}L_p \right) \longrightarrow 1 \quad (1)$$

$$\mathbb{P}_{\theta^{(\ell)}\neq\theta^{(r)}} \left(gl\hat{P}L_\ell + gl\hat{P}L_r \leq gl\hat{P}L_p \right) \longrightarrow 1 \quad (2)$$

Proof exploits consistency, within stationary intervals, established for group-lasso based neighborhood selection[12].

[12] Bolstad, A., Van Veen, B. D., & Nowak, R. (2011). Basu, S., Shojaie A., & Michailidis, G. (2015).

Finite Sample Control of Type I Error Rate

Theorem

If the penalty parameter $\lambda(\alpha)$ is chosen such that

$$\lambda(\alpha) = 2\hat{\sigma}_u \sqrt{pQ\left(1 - \frac{\alpha}{N^2}\right)}$$

where $\hat{\sigma}_u^2 = \frac{\|\mathbf{X}_u\|_2^2}{n}$, p the order of AR lags, $N = |V|$, and $Q(\cdot)$ the quantile function of $\chi^2(p)$, then

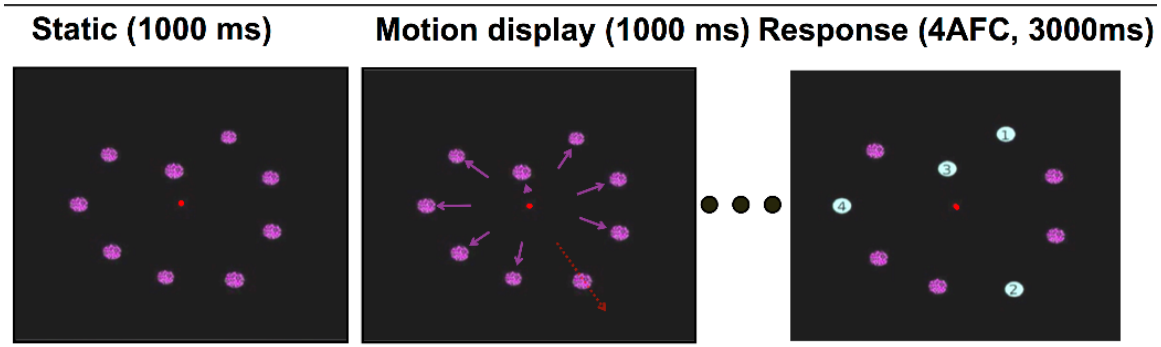
$$\mathbb{P}(\exists\, u \in V : \hat{C}_u^\lambda \nsubseteq C_u) \leq \alpha.$$

Here C_u is the connected component in G to which u belongs, and $\hat{C}_u^\lambda$, its estimate.

Proof in the spirit of result of Meinshausen & Buhlmann (2006).

Brain Image Data

We have trial-by-trial *visual search* data[13].



MEG data of 8 subjects who were asked to watch moving objects displayed on a black screen.

Each subject received 160 trials and the MEG of their brain was taken accordingly.

Each resulting time series truncated to 1190 sample points, with intervention centered at time $t = 1/2$.

[13] Data courtesy of Lucia M. Vaina

1st Attempt @ Modeling: High-level Results

Optimal piecewise Granger causal model was fit (Using VAR(7); and Recursive Dyadic Partition), yielding:

- a 3-node dynamic network.
- Changepoints detected only at dyadic positions $\{1/4, 1/2, 3/4\}$.
- Causal relations detected between pairs of nodes

BOSTON UNIVERSITY

Aggregate Rates for Changepoints and Edges

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Rate of split	(0.32)	(0.22)	(0.23)	(0.19)	(0.19)	(0.34)	(0.46)	(0.44)

Subject 1

First Half	IPS	FEF	V3a
IPS	0	0.10	0.09
FEF	0.11	0	0.08
V3a	0.11	0.09	0

Second Half	IPS	FEF	V3a
IPS	0	0.09	0.10
FEF	0.1	0	0.07
V3a	0.1	0.09	0

Subject 2

First Half	IPS	FEF	V3a
IPS	0	0.08	0.11
FEF	0.08	0	0.03
V3a	0.11	0.04	0

Second Half	IPS	FEF	V3a
IPS	0	0.05	0.09
FEF	0.04	0	0.03
V3a	0.09	0.02	0

Subject 3

First Half	IPS	FEF	V3a
IPS	0	0.01	0.18
FEF	0.01	0	0.01
V3a	0.18	0.04	0

Second Half	IPS	FEF	V3a
IPS	0	0.03	0.33
FEF	0.03	0	0.03
V3a	0.32	0.04	0

Subject 4

First Half	IPS	FEF	V3a
IPS	0	0.09	0.01
FEF	0.10	0	0.04
V3a	0.01	0.04	0

Second Half	IPS	FEF	V3a
IPS	0	0.08	0.03
FEF	0.09	0	0.06
V3a	0.02	0.03	0

Subject 5

First Half	IPS	FEF	V3a
IPS	0	0.08	0.04
FEF	0.09	0	0.01
V3a	0.03	0.06	0

Second Half	IPS	FEF	V3a
IPS	0	0.06	0.05
FEF	0.08	0	0.02
V3a	0.06	0.03	0

Subject 6

First Half	IPS	FEF	V3a
IPS	0	0.09	0.08
FEF	0.08	0	0.12
V3a	0.07	0.12	0

Second Half	IPS	FEF	V3a
IPS	0	0.09	0.06
FEF	0.09	0	0.11
V3a	0.08	0.10	0

Subject 7

First Half	IPS	FEF	V3a
IPS	0	0.19	0.21
FEF	0.20	0	0.08
V3a	0.19	0.08	0

Second Half	IPS	FEF	V3a
IPS	0	0.19	0.25
FEF	0.19	0	0.04
V3a	0.23	0.01	0

Subject 8

First Half	IPS	FEF	V3a
IPS	0	0.25	0.18
FEF	0.25	0	0.09
V3a	0.17	0.09	0

Second Half	IPS	FEF	V3a
IPS	0	0.14	0.21
FEF	0.15	0	0.09
V3a	0.21	0.10	0

BOSTON UNIVERSITY

Closing Thoughts

Dynamic network analysis is arguably one of the most active frontiers in 'network science'.

Statisticians have (uncharacteristically!) jumped in early on the problem of network topology inference.

Neuroscience and economics/finance has a great deal of interest in and activity on this topic.

Still to do for piecewise Granger causal models:

- implement RP variant (more refined than RDP)
- theoretical characaterization of overall model selection
- a more careful empirical study with MEG data

BOSTON UNIVERSITY

Thank you