Causality Analysis in Large-scale Time Series Data

CIKM 2013 Tutorial

Yan Liu

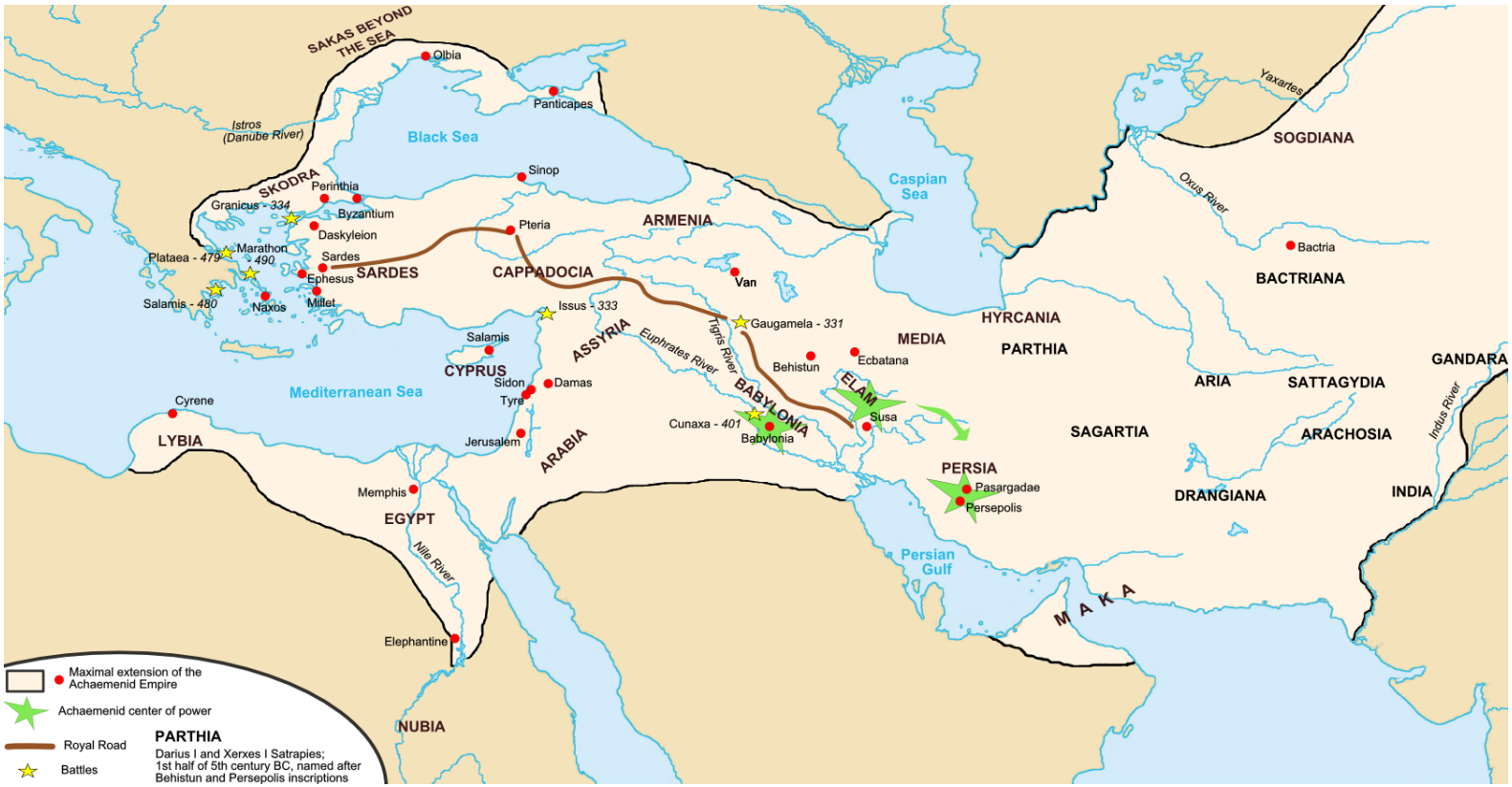
University of Southern California

October 29, 2013

History I

I would rather discover one causal law than be the king of Persia.

Democritus, (460-370 BC)



History II

Figure : The Illustrated Sutra of Cause and Effect. 8th century AD, Japan

Outline

Lecture 1: Introduction to Granger Causality (90 mins)

- Overview for Causality Analysis from Time Series Data (20 mins)
- Granger causality (40 mins)
  - Definition
  - Identification and learning
  - Examples in practical applications
- Known Issues of Granger causality (30 mins)
  - Non-linear extensions
  - Latent factors
  - Instantaneous causation

Break (30min)

Outline

Lecture 2: Alternative Approaches and New Trends (80 mins)

- Practical Issues in Granger causality (30 mins)
  - Time lag
  - Group effect
  - Non-stationary
  - Collinearity
  - Scalability
- Alternative Approaches (30 mins)
  - Randomization test
  - Auto-correlation and cross-correlation
  - Transfer entropy
- Illustration Examples (20 mins)

Disclaimer

- This is not meant for a comprehensive tutorial for causality
- Most techniques discussed in the tutorials are meant for facilitating causal discovery instead of automatic causal discovery
- This is mostly a computational view of causal discovery, i.e., how to make causal discovery practical

Outline

Lecture 1: Introduction to Granger Causality (90 mins)

- Overview for Causality Analysis from Time Series Data (20 mins)
- Granger causality (40 mins)
  - Definition
  - Identification and learning
  - Examples in practical applications
- Known Issues of Granger causality (30 mins)
  - Non-linear extensions
  - Latent factors
  - Instantaneous causation

Break (30min)

Motivation

Finding the answer to "why?" questions.

All physical laws are causal laws; for ex. The Gravity Law:

$$F \leftarrow G\frac{m_1 m_2}{r^2}$$

It is not just an algebraic equation; it also implies cause and effect relationship.

Treatment and *Symptoms* are correlated in the observations; but only changing *Treatment* has an impact on *Symptoms*.

Challenges

- Identifying True causal relationship.
- Designing Scalable algorithms for large datasets.

Philosophical View: Causality and Counter-factual Analysis

Causal Processes: Mark Transmission Theory by Wesley Salmon (1984):

> *Let P be a process that, in the absence of interactions with other processes would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both of the space-time points A and B. Then, a mark (consisting of a modification of Q into $Q^{\star}$), which has been introduced into process P by means of a single local interaction at a point A, is transmitted to point B if [and only if] P manifests the modification $Q^{\star}$ at B and at all stages of the process between A and B without additional interactions.*

Counter-factual Analysis: David Hume (1711 - 1776):

> *where if the first object had not been, the second had never existed.*

$$p \rightarrow q$$

$$\neg p \rightarrow \neg q$$

Philosophical View: Counter-factual Analysis

Philosophical Definition: Usually in *counter-factual language*:

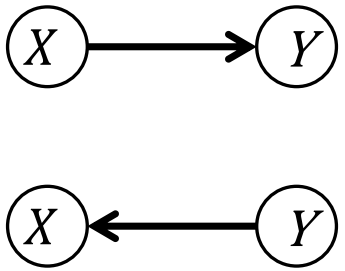
$$A \to B \quad \text{if} \quad \neg A \to \neg B.$$

- It requires *intervention* in the world and change of A to $\neg A$, while keeping the rest of the confounders unchanged. (*aka "Controlled Experiment"*)
- In many cases intervention is impossible: the car accident example.

Observational Studies: Randomization Test

Analyzing $X \to Y$:

- Fix all other factors.
- Change X and record the response of Y.
- Perform a statistical significance test.



Statistical Identifiability Problem: The joint distribution only encodes the association. Causality *(in general)* cannot be inferred from the observational data: we need *causal priors* to resolve the directions.

Advantage

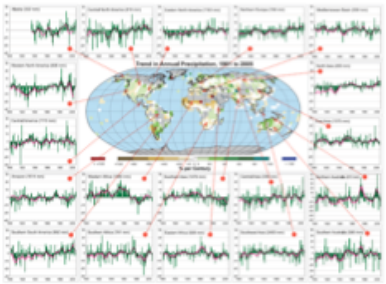
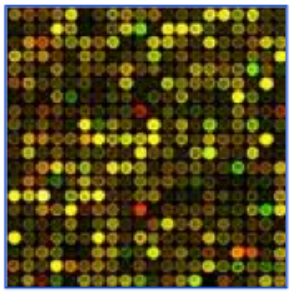
$\Rightarrow$ By far, the most accurate method
Not comparable with non-experimental (Data centric) methods.

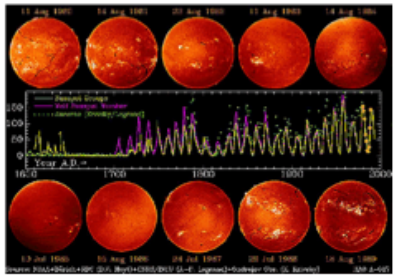
Limitations Sometimes randomized experiments are: *Expensive, Immoral or Impossible*.

Time Series Data are Everywhere

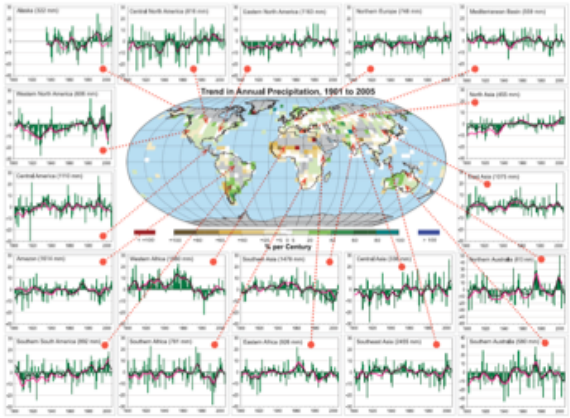
One important task in "From Data to Knowledge":

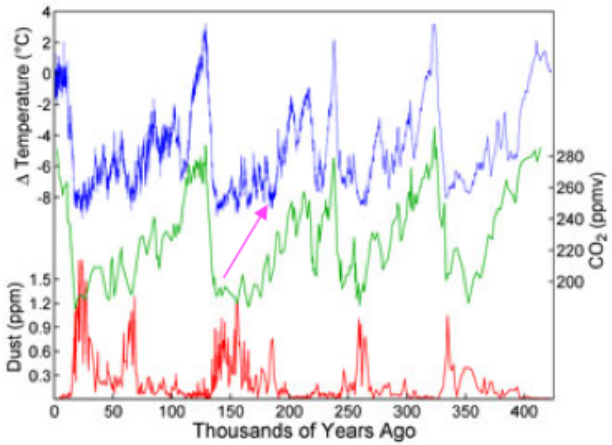
Discovery of Temporal Causal Relationships

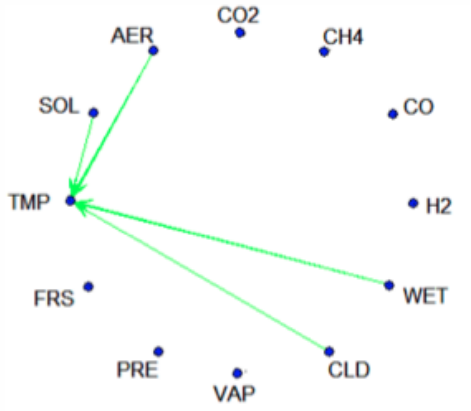



Understanding Climate Change Data

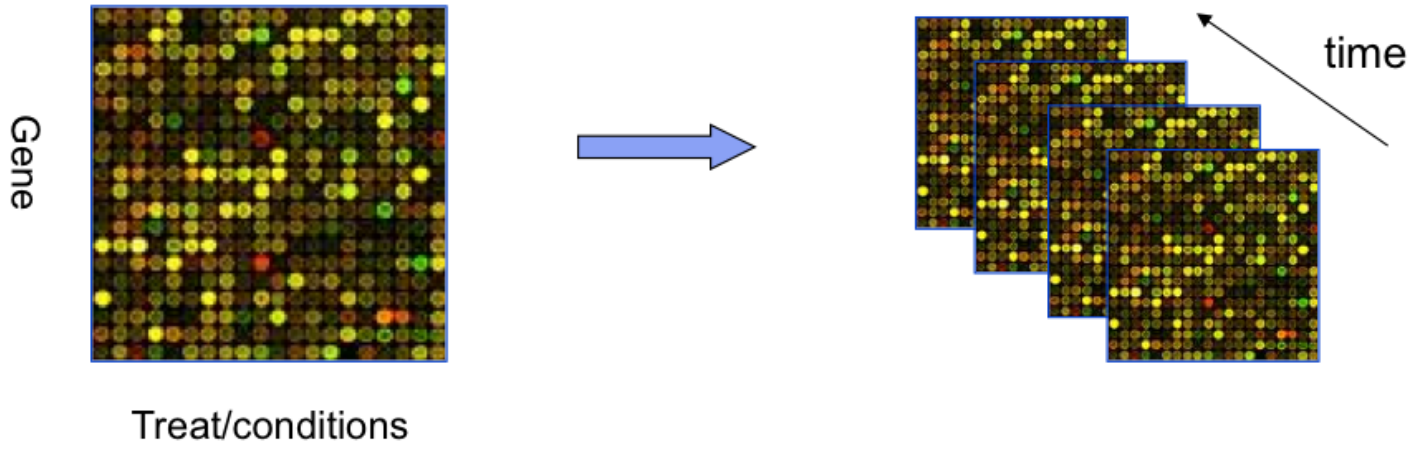




Output: temporal causal graph of climate forcing agents



Gene Regulatory Network Discovery



Output: gene regulatory networks

Social Influence Analysis

Output: social influence networks

Major Challenges

Large-scale high-dimensional time series data with complex relations

Outline

Lecture 1: Introduction to Granger Causality (90 mins)

- Overview for Causality Analysis from Time Series Data (20 mins)
- Granger causality (40 mins)
  - Definition
  - Identification and learning
  - Examples in practical applications
- Known Issues of Granger causality (30 mins)
  - Non-linear extensions
  - Latent factors
  - Instantaneous causation

Break (30min)

Granger Causality

- A cause is prior to its effect.
- If $X \xrightarrow{Granger} Y$, then X^{past} should significantly help predicting Y^{future} via Y^{past} alone.
- Based on linear regression/prediction of time series (Granger 1969)
- "Applied economists found the definition understandable and useable"
- Several writers stated that "of course, this is not real causality, it is only Granger Causality."
- "It is generally agreed that it does not capture all aspects of causality, but enough to be worth considering in an empirical test."

Clive Granger, recipient of the 2003 Nobel Prize in Economics.

Granger Causality: Formal Definition

Two Principles

1. The cause happens prior to the effect.
2. The cause makes unique changes in the effect. In other words, the causal series contains unique information about the effect series that is not available otherwise.

Define the following information sets:

- $\mathcal{I}^\star(t)$ - the set of all information in the universe up to time t;
- $\mathcal{I}^\star_{-X}(t)$ - the set of all information in the universe excluding X up to time t.

Granger's definition of causality (1969, 1980). Given two time series X and Y, we say X causes Y if

$$\mathbb{P}[Y(t+1) \in A | \mathcal{I}^\star(t)] \neq \mathbb{P}[Y(t+1) \in A | \mathcal{I}^\star_{-X}(t)],$$

Granger Causality: Practical Definition

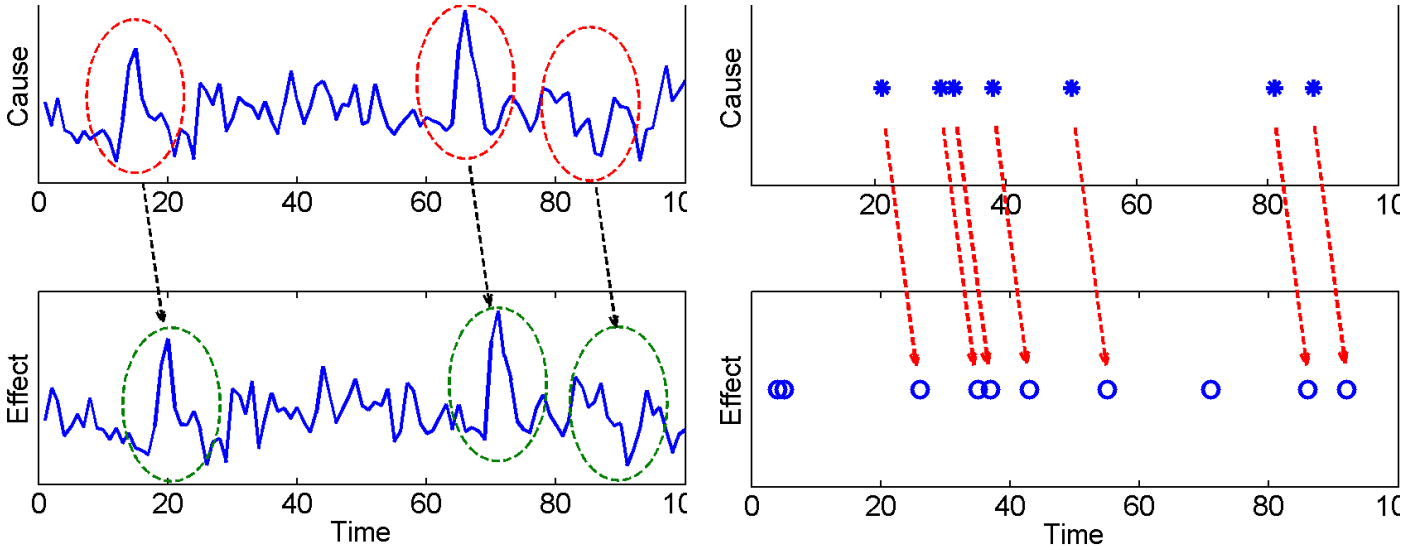
We perform two vector auto-regressions as follows:

$$Y(t) = \sum_{l=1}^{L} a_l Y(t-l) + \epsilon_1 \quad (1)$$

$$Y(t) = \sum_{l=1}^{L} a_l' Y(t-l) + \sum_{l=1}^{L} b_l' X(t-l) + \epsilon_2, \quad (2)$$

where L is the maximal time lag. We say X causes Y if eq (2) is statistically significantly better than eq (1).

Examples



(a) (b)

Figure : Illustration of the main principle behind Granger causality: in both examples the cause happens prior to its effect and its past values help predicting future values of the effect. (a) Plot of the values of two time series. (b) Plot of two point processes in which each event is shown with a mark at its happening time.

Granger Causality: Extension to Multivariate Time Series

Give p number of time series, $X_1, \ldots, X_p$, we are interested in identifying which time series Granger causes X_i.

Exhaustive-Granger Examine the pairwise relationships between pairs of time series

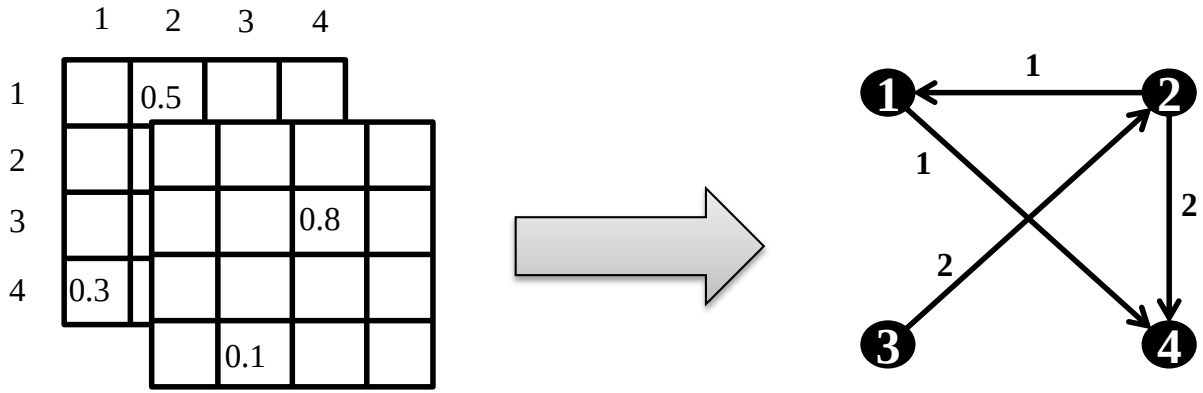
Multivariate Regression Perform vector auto-regression as follows:

$$X_i(t) = \sum_{j=1}^{p} \mathbf{a}_{i,j}^{\top} \mathbf{X}_j^{t,Lagged} + \epsilon,$$

$\mathbf{X}_j^{t,Lagged} = [X_j(t-L), \ldots, X_j(t-1)]$ is the lagged time series, and $\mathbf{a}_{i,j} = [a_{i,j,1}, \ldots, a_{i,j,L}]$ is the coefficient vector. Then we test the zeroness of $\mathbf{a}_{i,j}$ by statistical significant tests.

Granger Causality: Creating the Causal Graph

Evolution Matrices $\Rightarrow$ Granger Causality Graph



$$\text{If } \boldsymbol{\alpha}_{i,j} \neq \mathbf{0} \quad \Rightarrow \quad \mathbf{X}_j \rightarrow \mathbf{X}_i$$

Granger Causality in High Dimensions

Granger Causality with significance test

$T/L \geq p+1$ $\mathbb{P}[\text{Error}] \leq cL\sqrt{T-L}\exp\left(-\frac{c^2}{2}(T-L)\right)$

$T/L < p+1$ Inconsistent

Lasso-Granger the variable selection can be efficiently done in high dimensions: (Valdes-Sosa *et al* 2005 and Arnold *et al* 2007)

$$\min_{\{\boldsymbol{\beta}\}} \sum_{t=L+1}^{T} \sum_{i=1}^{p} \left\| X_i(t) - \sum_{i=1}^{p} \boldsymbol{\beta}_{i,j}^{\top} \mathbf{X}_j^{t,Lagged} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1,$$

Lasso-Granger

$\mathbb{P}[\text{Error}] = o(c'L\exp(-T^{\nu}))$ for some $0 \leq \nu < 1$.

Granger Causality in Practice

Advantages The simplicity, robustness, interpretability and scalability makes Granger causality widely adopted

Philosophical *Causal order* defines *time order*; the reverse is not true in general. Examples: *Instant Causation*, *The forward looking behavior of human beings*

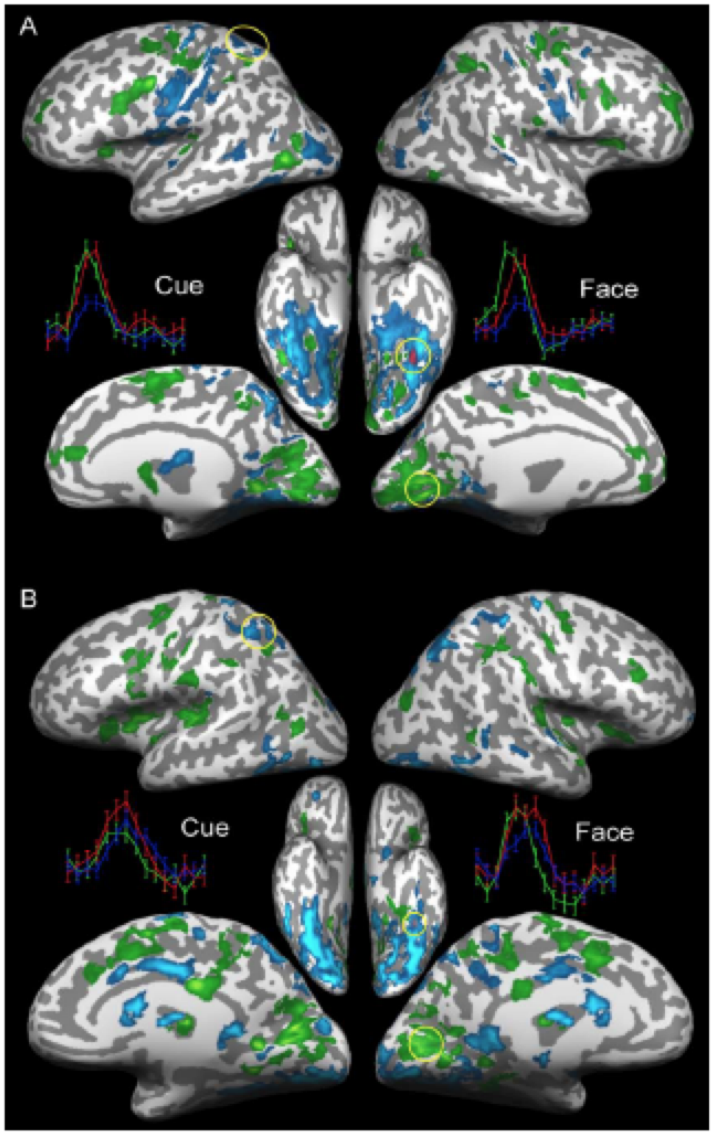
Spurious Causality Granger causality is sensitive to the absence of possible causes.

Direct Causes Granger causality only detects the *direct causality*. For example consider

$$X \xrightarrow{Granger} Z \xrightarrow{Granger} Y$$

Given Z, Granger causality test will not detect the transferred causal effects of X on Y.

Application 1: Brain Image Analysis

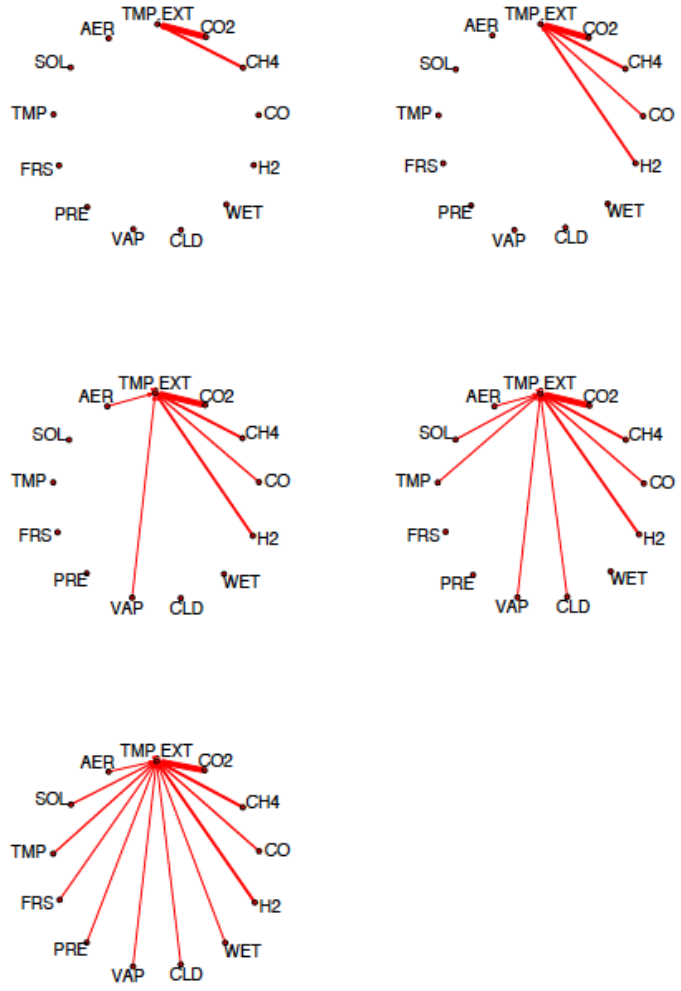


Threshold difference GCMs for a face-selective region in the left fusiform gyrus for subject 1 in (A) and subject 2 in (B). Event-related BOLD responses are shown for the circled areas in the calcarine sulcus (in green), the fusiform gyrus (the reference area, in red) and the intra-parital sulcus (in blue) for both the Cue stimulus and the Face stimulus.

Mapping directed influence over the brain using Granger causality and fMRI. Roebroeck et al, 2005.

Application 2: Climate Change Attribution

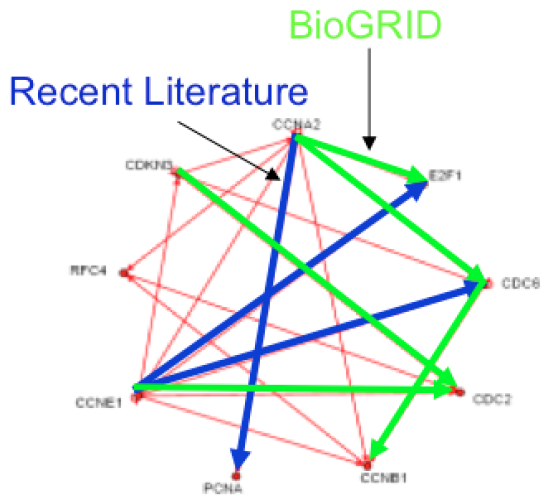
Variables (Variable group)	Type	Source
Methane (CH_4) Carbon-Dioxide (CO_2) Hydrogen (H_2) Carbon-Monoxide (CO)	Greenhouse Gases	NOAA
UV (AER)	Aerosol Index	NASA
Temperature (TMP) Temp Range (TMP) Temp Min (TMP) Temp Max (TMP) Precipitation (PRE) Vapor (VAP) Cloud Cover (CLD) Wet Days (WET) Frost Days (FRS)	Climate	CRU
Global Horizontal (SOL) Direct Normal (SOL) Global Extraterrestrial (SOL) Direct Extraterrestrial (SOL)	Solar Radiation	NCDC
1-year return level for temperature extreme (TMP.EXT)	Climate	Estimated using temp from CDIAC



Application 3: Regulatory Network Discovery

Evaluation against BioGRID

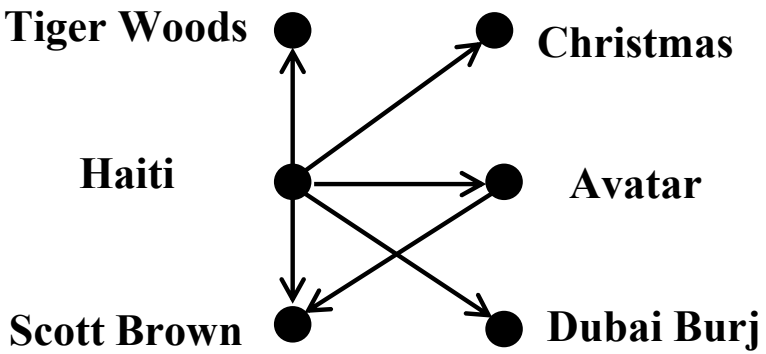
	Precision	Recall	F1
Our method	0.50	0.72	0.59
Sambo *et al*. (2008)	0.36	0.44	0.40



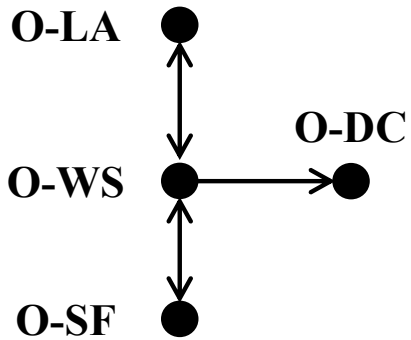
Causal graphs discovered by our method

Application 4: Twitter Influence Network Discovery

Twitter Meme:



Occupy WS:



Outline

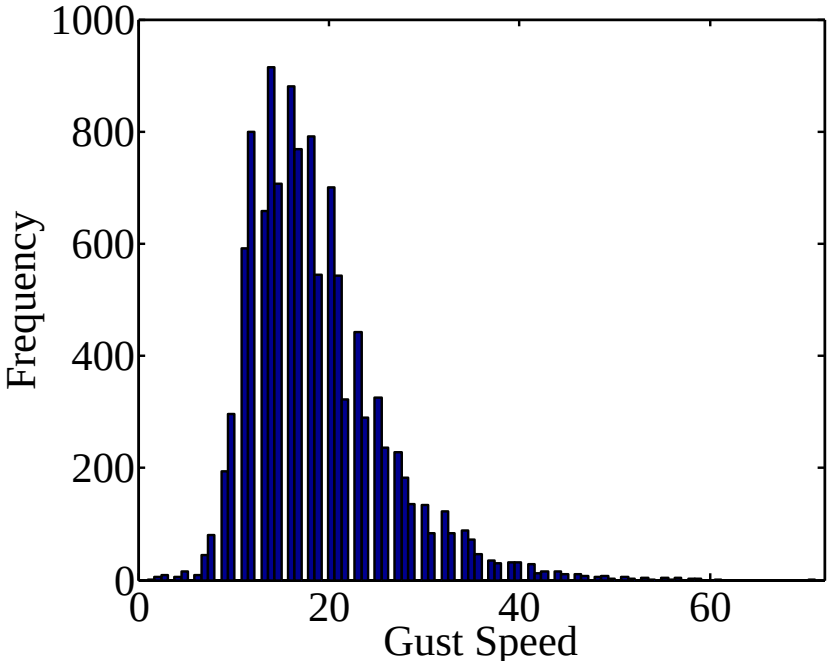
Lecture 1: Introduction to Granger Causality (90 mins)

- Overview for Causality Analysis from Time Series Data (20 mins)
- Granger causality (40 mins)
  - Definition
  - Identification and learning
  - Examples in practical applications
- Known Issues of Granger causality (30 mins)
  - Non-linear extensions
  - Latent factors
  - Instantaneous causation

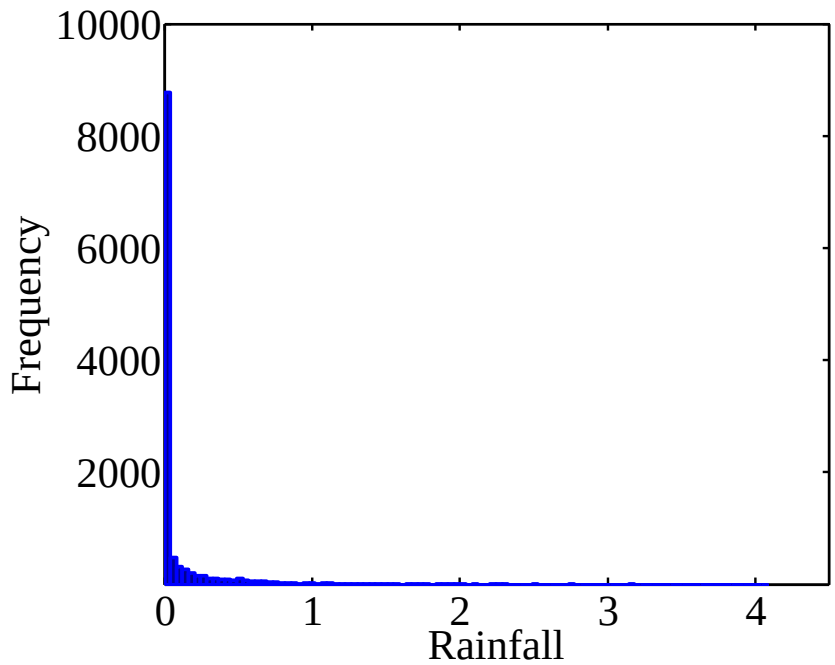
Break (30min)

Issue 1: Nonlinear Extensions

The assumption of linearity in VAR-based Granger causality test significantly simplifies the task of testing conditional independence, but it can be easily violated in the real applications.



Distribution of Wind Gust Speed



Distribution of Rainfall

Solutions to Nonlinearity

Many different approaches have been proposed to address the problem:

- ARCH, GARCH
- Non-parametric approach: Fourier or Wavelet transformation, Kernel methods
- Semi-parametric approach: copula-based approach
- heuristic additive models

Nonlinear Extensions: Kernel Methods

Given the following predictors,

$$\mathbf{U}(t) \triangleq [X_1(t-L), \ldots, X_1(t-1), \ldots, X_n(t-L), \ldots, X_n(t-1)] \tag{3}$$

$$\mathbf{V}(t) \triangleq [X_1(t-L), \ldots, X_1(t-1), \ldots, X_{j-1}(t-L), \ldots, X_{j-1}(t-1), \ldots, X_{j+1}(t-L), \ldots, X_{j+1}(t-1), \ldots, X_n(t-L), \ldots, X_n(t-1)] \tag{4}$$

where $\mathbf{V}(t)$ is $\mathbf{U}(t)$ excluding the observations of X_j, we can perform two kernel regressions [Marinazzo et al., 2008]:

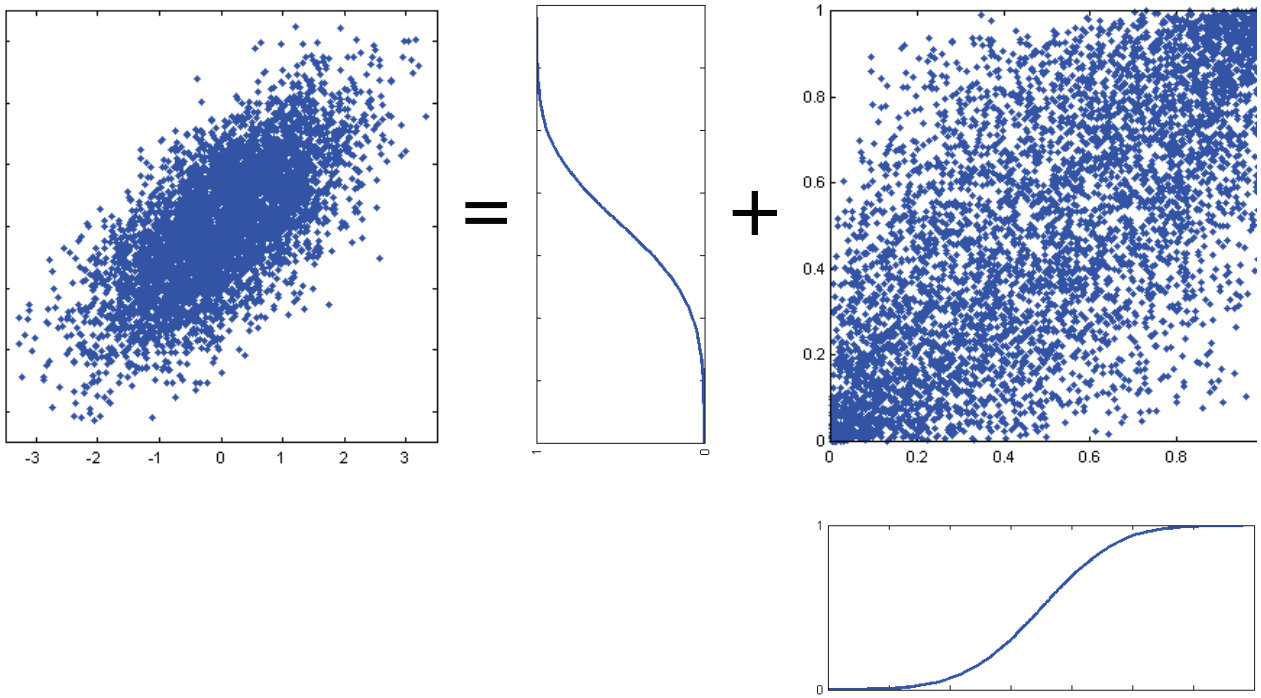
- Predicting $X_i(t)$ using $\mathbf{U}(t)$,
- Predicting $X_i(t)$ using $\mathbf{V}(t)$.

Let $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ denote the variance of noise obtained by the regression algorithms. Significance test can be applied on $1 - \widehat{\sigma}_1/\widehat{\sigma}_2$ to decide the edge $X_j \to X_i$.
Kernel functions: Gaussian kernel is a common choice.

Nonlinear Extensions: Semi-parametric Solutions

Modeling dependency requires more samples than modeling the marginals.

- Separate learning for marginals and dependency: (Liu *et al* 2009)
  - A non-parametric estimator for modeling the marginals
  - The data-efficient VAR model for modeling dependency.



[Krishner'11]

Granger Non-paranormal (G-NPN) Model

We say a set of time series $X = (X_1, \ldots, X_p)$ has Granger-Nonparanormal distribution $G - NPN(X, \boldsymbol{A}, F)$ if:

1. There exist distribution functions $\{F_j\}_{j=1}^p$ such that $Z_j \triangleq F_j(X_j)$ for $j = 1, \ldots, p$ are jointly Gaussian.
2. Z_j for $j = 1, \ldots, p$ are factorized according to the VAR model with coefficients $\boldsymbol{A}$.

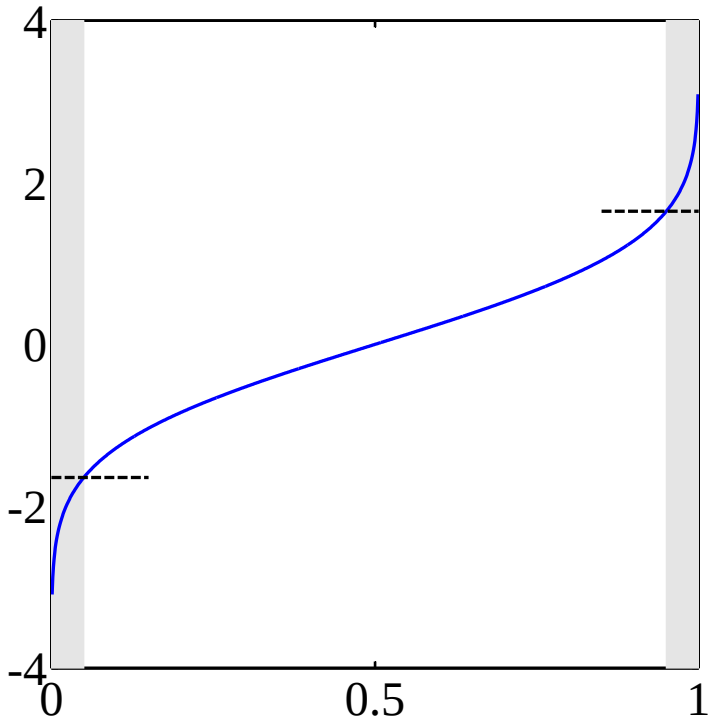
Copula-Granger Algorithm

Learning Copula-Granger:

1. Find the empirical marginal distribution for each time series $\widehat{F}_i$.
2. Map the observations into the copula space as $\widehat{f}_i(X_i(t)) = \Phi^{-1}\left(\widehat{F}_i(X_i(t))\right)$.
3. Find the Granger causality among $\widehat{f}_i(X_i(t))$ using Lasso-Granger.

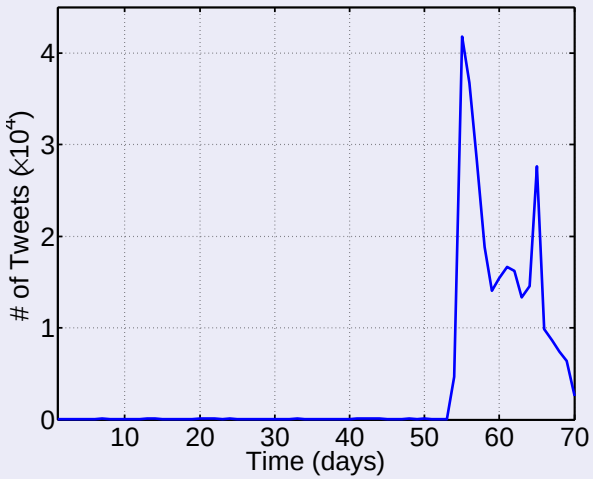
Winsorized Estimation of CDF:

$$\tilde{F}_j = \begin{cases} \delta_T & \text{if } \widehat{F}_j(X_j) < \delta_T \\ \widehat{F}_j(X_j) & \text{if } \delta_T \le \widehat{F}_j(X_j) \le 1-\delta_T \\ (1-\delta_T) & \text{if } \widehat{F}_j(X_j) > 1-\delta_T \end{cases}$$



Issue 2: Latent Factors

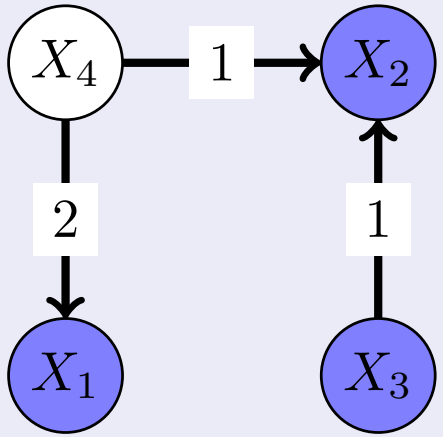
Latent Factors with Global Impact



Twitter Examples: Unobserved external factors impact the entire network activities.

Question: How can we discover causal relationships in presence of *latent factors with global impacts*?

Latent Factors with Local Impact



The complex nature of the latent factors makes the analysis challenging.

Question: How can we accurately model in presence of *latent factors with minor impacts*?

Latent Factors with Local Impact

Common Approach

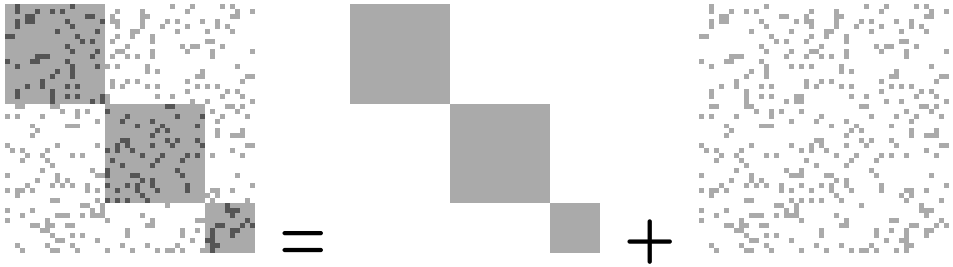
1. Explicitly model the hidden variables in the graphical model.
2. Apply the Expectation Maximization algorithm to learn the structure.

Drawbacks

1. EM algorithm is slow.
2. EM algorithm is trapped in the local minima.

Latent Factors with Global Impact

Sparse plus Low-rank Decomposition



There are convex optimization algorithms to decompose a matrix into a sparse matrix and a low-rank matrix.

Generalized Linear Auto-regressive Processes

Generalized Linear Model (GLM)

$$g(\mathbb{E}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}]) = A\mathbf{x} + \mathbf{b},$$

where the strictly monotone function $g(.)$ is called the *link function*.

Generalized Linear Auto-Regressive Processes (GLARP)

$$g(\mathbb{E}_{\mathcal{H}(t)}[\mathbf{x}(t)]) = \sum_{\ell=1}^{K} A^{(\ell)}\mathbf{x}(t-\ell) + \mathbf{b}.$$

where the matrices $A^{(\ell)}$ for $\ell = 1, \ldots, K$ are called the *Evolution Matrices*.

GLARP with Latent Factors

$$g\left(\mathbb{E}_{\mathcal{H}(t)}\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{z}(t)\end{bmatrix}\right) = \sum_{\ell=1}^{K}\begin{bmatrix}A^{(\ell)} & B^{(\ell)}\\ C^{(\ell)} & D^{(\ell)}\end{bmatrix}\begin{bmatrix}\mathbf{x}(t-\ell)\\ \mathbf{z}(t-\ell)\end{bmatrix} + \mathbf{b}$$

for $t = K+1, \ldots, T$. $\mathbf{x}(t)$, a $p \times 1$ vector, represents the observed variables, $\mathbf{z}(t)$, a $r \times 1$ vector, denotes the unobserved variables.

Key Models: Count Data

Commonly assumed Poisson distribution:

$$\log \boldsymbol{\lambda}(t) = \log(\mathbb{E}_{\mathcal{H}(t)}[\mathbf{x}(t)]) = \sum_{\ell=1}^{K} A^{(\ell)}\mathbf{x}(t-\ell) + \mathbf{b},$$

Conway-Maxwell Poisson distribution: The adjustable rate of decay:

$$\frac{\mathbb{P}[X = k-1]}{\mathbb{P}[X = k]} = \left(\frac{k}{\mu}\right)^{\nu},$$

The model:

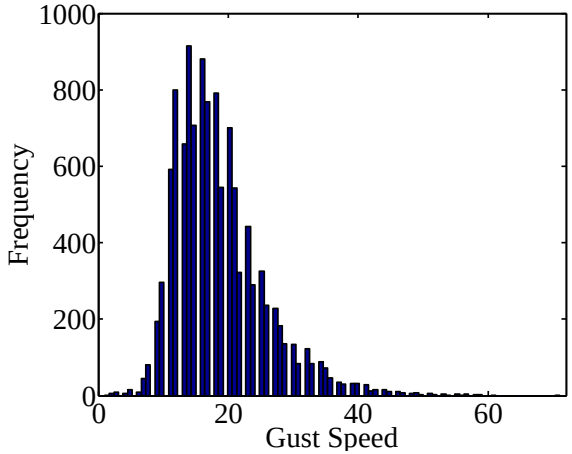
$$\mathbb{P}[x_i(t)|\mu_i(t), \nu] = \frac{1}{S(\mu_i(t), \nu)}\left(\frac{\mu_i(t)^{x_i(t)}}{x_i(t)!}\right)^{\nu}$$

$$\log\left(\boldsymbol{\mu}(t) + \frac{1}{2\nu} - \frac{1}{2}\right) = \log\left(\mathbb{E}_{\mathcal{H}(t)}[\mathbf{x}(t)]\right) = \sum_{\ell=1}^{K} A^{(\ell)}\mathbf{x}(t-\ell) + \mathbf{b}.$$

Key Models: Climate Extreme Value Data

Gumbel Distribution

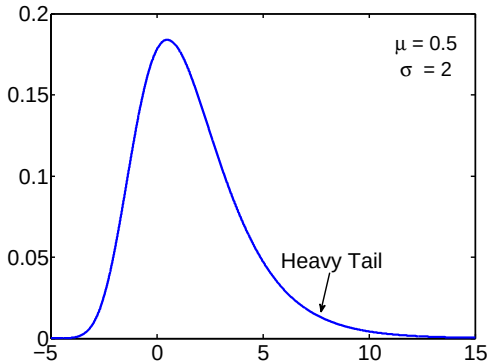
$$F(x_i(t)|\mu_i(t),\sigma) = \exp\left(-\exp\left(-\frac{x_i(t)-\mu_i(t)}{\sigma}\right)\right)$$



Marginal distribution of Wind Gust Speed

The model

$$\boldsymbol{\mu}(t) + \sigma\gamma_E = \mathbb{E}_{\mathcal{H}(t)}[\mathbf{x}(t)] = \sum_{\ell=1}^{K} A^{(\ell)}\mathbf{x}(t-\ell) + \mathbf{b},$$



Issue 3: Instantaneous causation

In some cases, there is a possibility that a time series has an instantaneous causal effect on another time series.
Solutions: *Structural Vector Authoregressive* (SVAR) formulated as follows:

$$X_i(t) = \sum_{j=1}^{n} \mathbf{a}_{i,j}^{\top} \mathbf{X}_i(t-L,\ldots,t) + \varepsilon_i(t), \tag{5}$$

where $\varepsilon_i, i = 1,\ldots,n$ are independent white noise processes.
Similar to Independent Component Analysis (ICA), SVAR suffers from identifiability when the noise processes are Gaussian processes.
Several attempts have been made to estimate the model above, such as [Hyvarinen et al, 2010] based on non-Gaussian assumption of the noise processes.

Outline

Lecture 2: Alternative Approaches and New Trends (80 mins)

- Practical Issues in Granger causality (30 mins)
  - Time lag
  - Group effect
  - Non-stationary
  - Collinearity
  - Scalability
- Alternative Approaches (30 mins)
  - Randomization test
  - Auto-correlation and cross-correlation
  - Transfer entropy
- Illustration Examples (20 mins)

Practical Issues: Time Lag

Observation: Granger causality requires prior knowledge about the maximum lag L.

Solutions:

- Cross-validation
- Modeling the distribution of maximum lag length
- Autocorrelation function (ACF) and partical autocorrelations (PACF)
- Other model selection techniques: AIC and BIC

Practical Issues: Group Effect

Observation: The existence of any non-zero element in the coefficient vector $\mathbf{a}_{i,j}$ is interpreted as X_j is a Granger cause of X_i, the penalization term should shrink the group of coefficient $\mathbf{a}_i$ instead of individually sparsifying its elements.

Solutions: We can reformulates the Lasso-Granger via the Grouped-Lasso penalization as follows:

$$\min_{\{\mathbf{a}\}} \sum_{t=L+1}^{T} \left\| X_1(t) - \sum_{i=1}^{n} \mathbf{a}_i^{\top} \mathbf{X}_i^{t,Lagged} \right\|_2^2 + \lambda \sum_{i=1}^{n} \|\mathbf{a}_i\|_2 , \tag{6}$$

Practical Issues: Non-stationary

Observation: Stationary assumptions can be easily violated in practical applications.

Solutions:

- Window-based approach: divide the time series into sub-intervals, assume that in each sub-interval the time series are stationary and apply Granger causality
- Markov model-based approach for recurrent patterns: associate each temporal structure with a hidden state and automatically infer the possible state assignment of each time stamp and its temporal structures via EM algorithm

Practical Issues: Collinearity

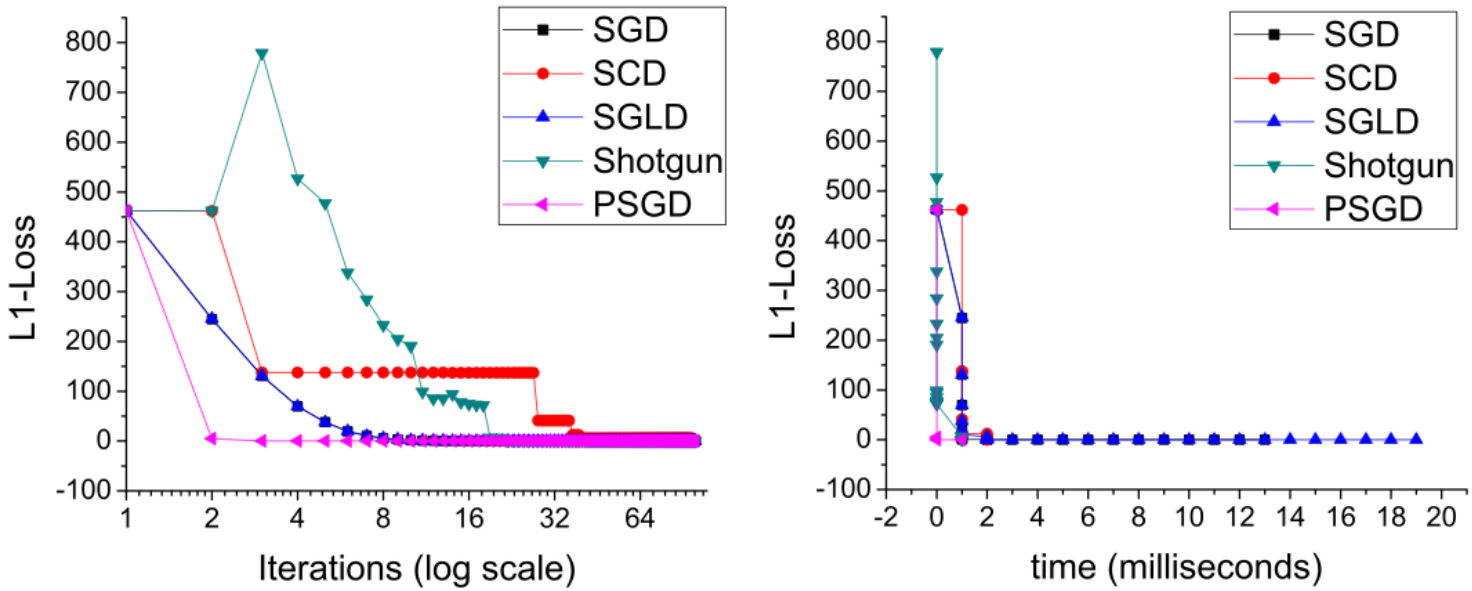
Observation: Some time series could be linearly correlated to each other, which raises identifiability issues for regression-based approaches

Solutions: feature selection methods, such as SCAD [Fan et al, 2001] and adaptive Lasso [Yuan & Zou, 2006], can be applied.

Practical Issues: Scalability

Parallel Stochastic Optimization Algorithms

- Stochastic Subgradient Langevin Dynamics (*SGLD*) [Welling & Teh, 2011]: combining mini-batch stochastic subgradient descent and Langevin Dynamics.
- Parallel Stochastic Coordinate Descent (*Shotgun*) [Bradley et al, 2011]: a parallel implementation of Stochastic Coordinate Descent.
- Parallel Stochastic Gradient Descent (*PSGD*) [Zinkevich et al, 2010]: randomly partitioning the data, giving one partition to each processor, which sequentially uses each data point of its own partition to update β using a constant step size η.



Outline

Lecture 2: Alternative Approaches and New Trends (80 mins)

- Practical Issues in Granger causality (30 mins)
  - Time lag
  - Group effect
  - Non-stationary
  - Collinearity
  - Scalability
- Alternative Approaches (30 mins)
  - Randomization test
  - Auto-correlation and cross-correlation
  - Transfer entropy
- Illustration Examples (20 mins)

Randomization Tests

Also knowns as Exact Test or Permutation Test; is a computational approach for statistical significance test.

$$\mathcal{H}_0 : X \to Y \qquad vs. \qquad \mathcal{H}_1 : X \not\to Y$$

Species Richness	Area
32	2.0
29	0.9
35	3.1
26	3.0
41	1.0
62	2.0
88	4.0
77	3.5

In 9.5% of permuted cases the computed correlation is higher than the original case. s

$\Rightarrow$ "How likely is it that if the null hypothesis were true, I would observe a value this extreme just due to chance?"

Significance Tests

Suppose We are testing $\mathcal{H}_0 : \theta \in \Theta_0$ vs $\mathcal{H}_1 : \theta \notin \Theta_0$. We can define the following statistics:

$$\lambda_n = \frac{\sup_{\theta \in \Theta} \mathcal{L}_n(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}_n(\theta)}$$

The Likelihood Ratio Test rejects $\mathcal{H}_0$ if $\lambda_n \geq c$ for some c.

Theorem *Asymptotic Distribution of Generalized Likelihood Ratio Statistics*: Subject to many assumptions (mostly smoothness) for hypotheses like $\mathcal{H}_0 : \theta_1 = \theta_2 = \ldots = \theta_r = 0$ for some $1 \leq r \leq k$, then

$$2 \log \lambda_n \to \chi^2_r \quad as\ n \to \infty$$

Practically suppose we want to guarantee with $1 - \alpha$ probability that the test is correct. Use the $\alpha_{\chi^2_r}$ quantile to find the appropriate level for c.

Structural Equation Modeling (SEM): Concepts

Goals

- Formulate a framework for counter-factual analysis
- Design statistical tools for testing *proposed* causal structure.

Terminology

Confounder background (usually unobserved) factors which can affect the cause and effect under study.

Spurious Effects The paths between the proposed cause and effect through other factors (variables).

Key claims

- Causality is associated with change.
- Causality cannot be inferred from observational data without extra assumptions.
- Given a graph, I can tell you whether you can judge about a certain causal relationship.

SEM: The Model

Structural Equations A SEM is a set of equations describing the value of each variable in $\mathcal{V}$ as a function f_X of its parents $\mathbf{pa}_X$ and a random disturbance term u_X:

$$x = f_X(\mathbf{pa}_X, u_X).$$

Counter-Factual Analysis The Counter-Factual $Y_x(u)$, Y would be y (in situation u) had X been x, is given by:

$$Y_x(u) \triangleq Y_{M_x}(u),$$

M_x is a modified version of M with equations of X replaced by $X = x$.

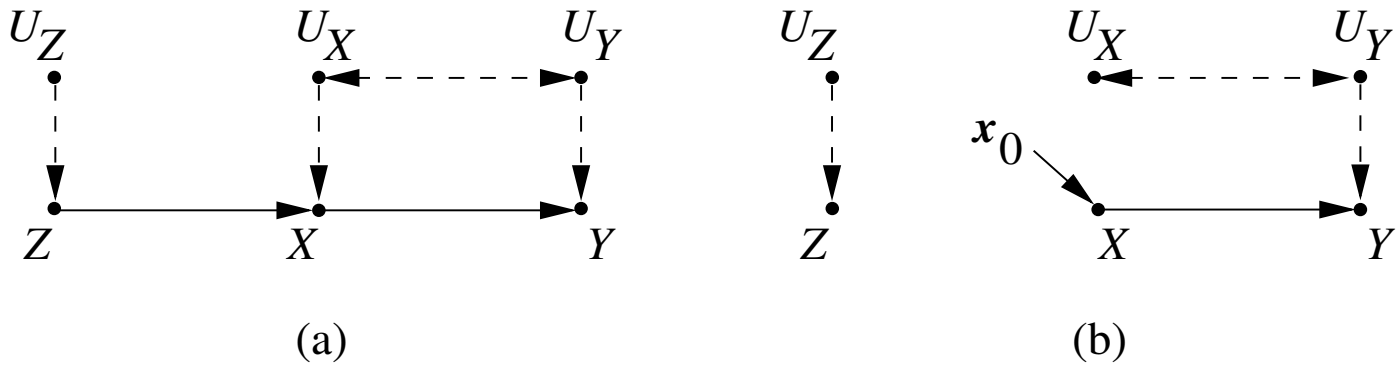


Figure : Pearl's Graph Surgery

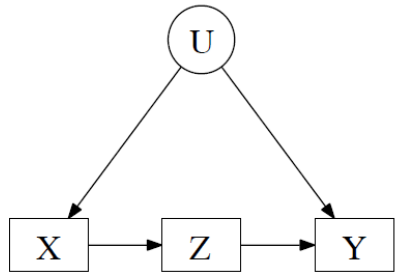
SEM: Graph Surgery with Conditional Independence Tests

CI Test Principle Condition on adequate control variables, which will block paths linking X and Y other than those which would exist in the surgically-altered graph where all paths into X have been removed.
Back-door criterion: (i) $\mathcal{S}$ blocks every path from X to Y that has an arrow into X , and (ii) no node in $\mathcal{S}$ is a descendant of X. Then

$$\mathbb{P}[Y|do(X=x)] = \sum_s \mathbb{P}[Y|X=x,\mathcal{S}=s]\mathbb{P}(\mathcal{S}=s)$$

Front-door criterion:
(i) $\mathcal{S}$ blocks all directed paths from X to Y , (ii) there are no unblocked back-door paths from X to $\mathcal{S}$, and (iii) X blocks all back-door paths from $\mathcal{S}$ to Y . For example:

$$\begin{aligned}\mathbb{P}[Y|do(X=x)] &= \mathbb{P}[Y|do(Z=z)]\mathbb{P}[Z|do(X=x)] \\ &= \mathbb{P}[Y|Z=z]\mathbb{P}[Z|X=x]\end{aligned}$$



SEM: Theoretical Results

- The back-door and front-door criteria are *Sufficient* for estimating the causal effects from probabilistic distributions,
- but they are not *Necessary*.
- The necessary conditions don't have nice forms.
- When controlling all the variables is impossible we can still estimate or bound the $\mathbb{P}[Y|do(X = x)]$. This is called *Partial Identification*.

SEM: A Practical Example

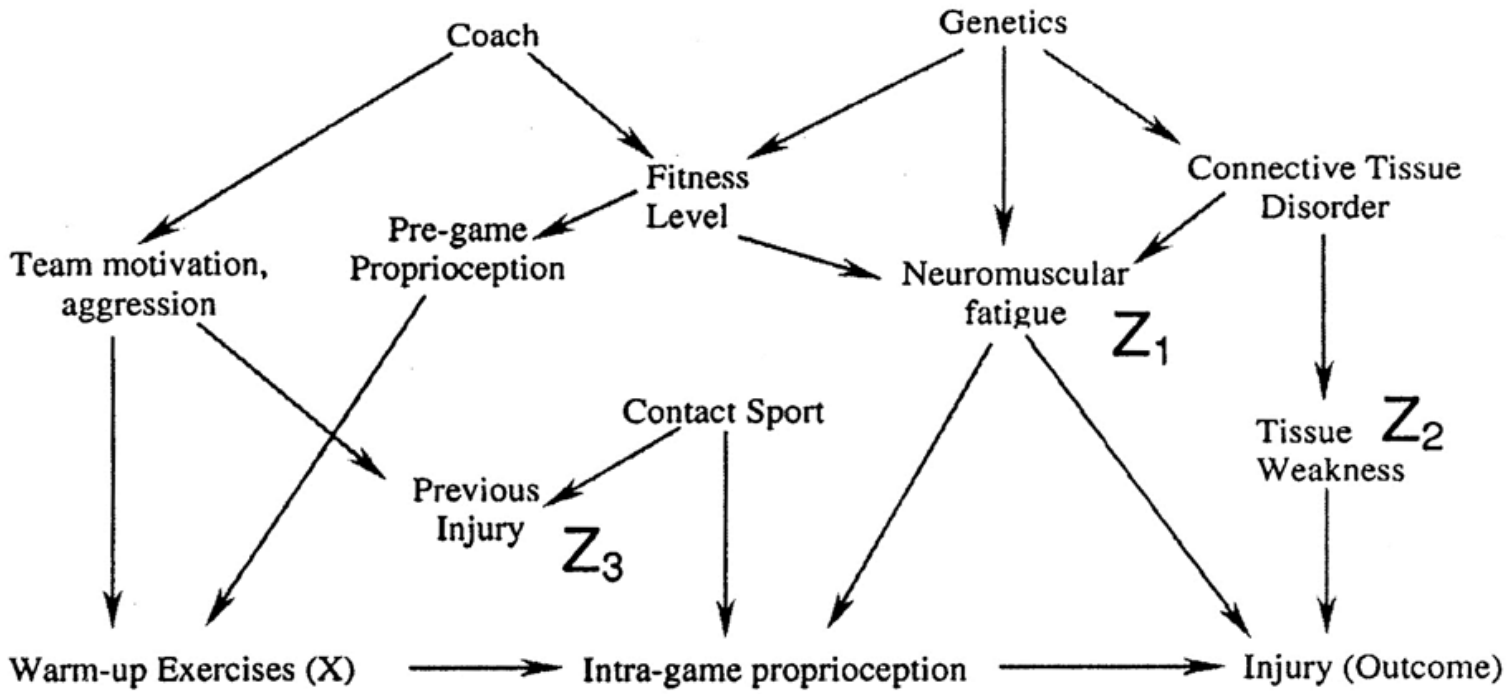


Figure : Effect of Warm-up on Injury (After Shrier & Platt, 2008)

Partial Correlation Based Algorithms List

TC Algorithm (Pellet et al., 2007) Finds correlation of (X, Y) given the rest of the variables. Then identifies the colliders and finds the direction of the edges using them. Complexity $\mathcal{O}(n^3)$.

Many BN Learning algorithms:

L1MB (Schmidt et al., 2007) Maximizing L_1 penalized likelihood function to get an undirected graph. Prune the graph using greedy hill climbing.

IAMB (Tsamardinos et al., 2003) Uses Mutual Information instead of correlation and thresholds for deciding which edges to keep. Not necessarily returns DAG. Complexity $\mathcal{O}(n^2)$.

Use of Asymmetry

Define higher order covariance as:

$$cov_{jk}[x_1, x_2] = \mathbb{E}[(x_1 - \mathbb{E}(x_1))^j (x_2 - \mathbb{E}(x_2))^k],$$

and higher order correlation as:

$$\rho_{jk}(x_1, x_2) = \frac{cov_{jk}(x_1, x_2)}{\sigma_{x_1}^j \sigma_{x_2}^k}.$$

Consider two models:

Model #1: $x_2 = b_{12} x_1 + \epsilon_1,$

Model #2: $x_1 = b_{21} x_2 + \epsilon_2,$

Prefer Model #1 if (as long as $\rho_{x_1 x_2} \neq 0$, $\mu_3(x_1) \neq 0$ and $\mu_3(x_2) \neq 0$.):

$$\hat{\rho}_{12}^2 > \hat{\rho}_{21}^2$$

and Model #2 otherwise.

If $\mu_3(\epsilon) = 0$ then the test becomes:

$$\hat{\gamma}_{x_1}^2 > \hat{\gamma}_{x_2}^2$$

The response variable should have lower skewness!

Transfer Entropy

Definition (Schreiber, 2000)

$$T_{X\to Y} = \mathbb{H}\left(Y^t|Y^{\mathbf{t-1}}, X^{\mathbf{t-1}}\right) - \mathbb{H}\left(Y^t|Y^{\mathbf{t-1}}\right)$$

The amount of resolved uncertainty in future of Y by past values of X given past values of Y.

Properties

- Non-linear causation
- Non-stationary time series
- Conceptually more meaningful;
  since uses independence instead of correlation.

Limitations

- Only well-defined for two time series
- Computation of entropy from observations is challenging.

Relationship with Granger Causality (Barnett, 2009)

Transfer Entropy and Granger Causality are equivalent if the data is distributed according to multivariate Gaussian distribution.

Outline

Lecture 2: Alternative Approaches and New Trends (80 mins)

- Practical Issues in Granger causality (30 mins)
  - Time lag
  - Group effect
  - Non-stationary
  - Collinearity
  - Scalability
- Alternative Approaches (30 mins)
  - Randomization test
  - Auto-correlation and cross-correlation
  - Transfer entropy
- Illustration Examples (20 mins)

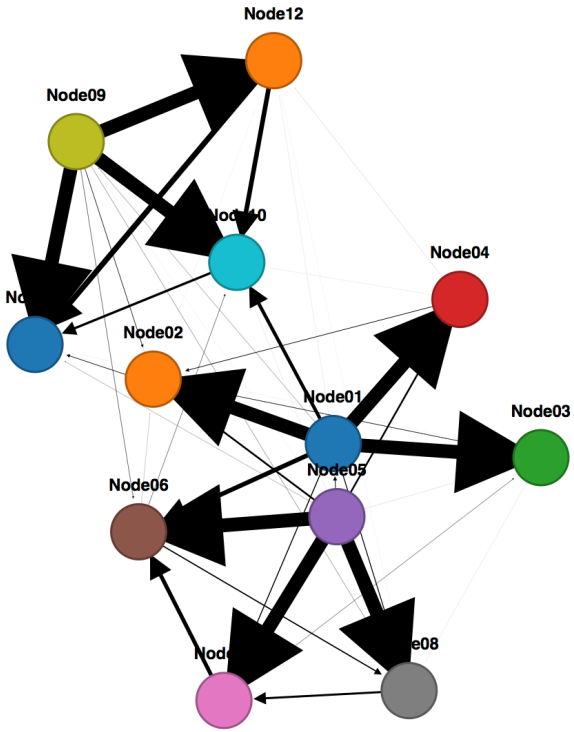
Demo

Code is available at:
`http://www-bcf.usc.edu/~liu32/code.htm`

Visualization is available at:
`http://beijing.usc.edu:22222/granger/granger.php`

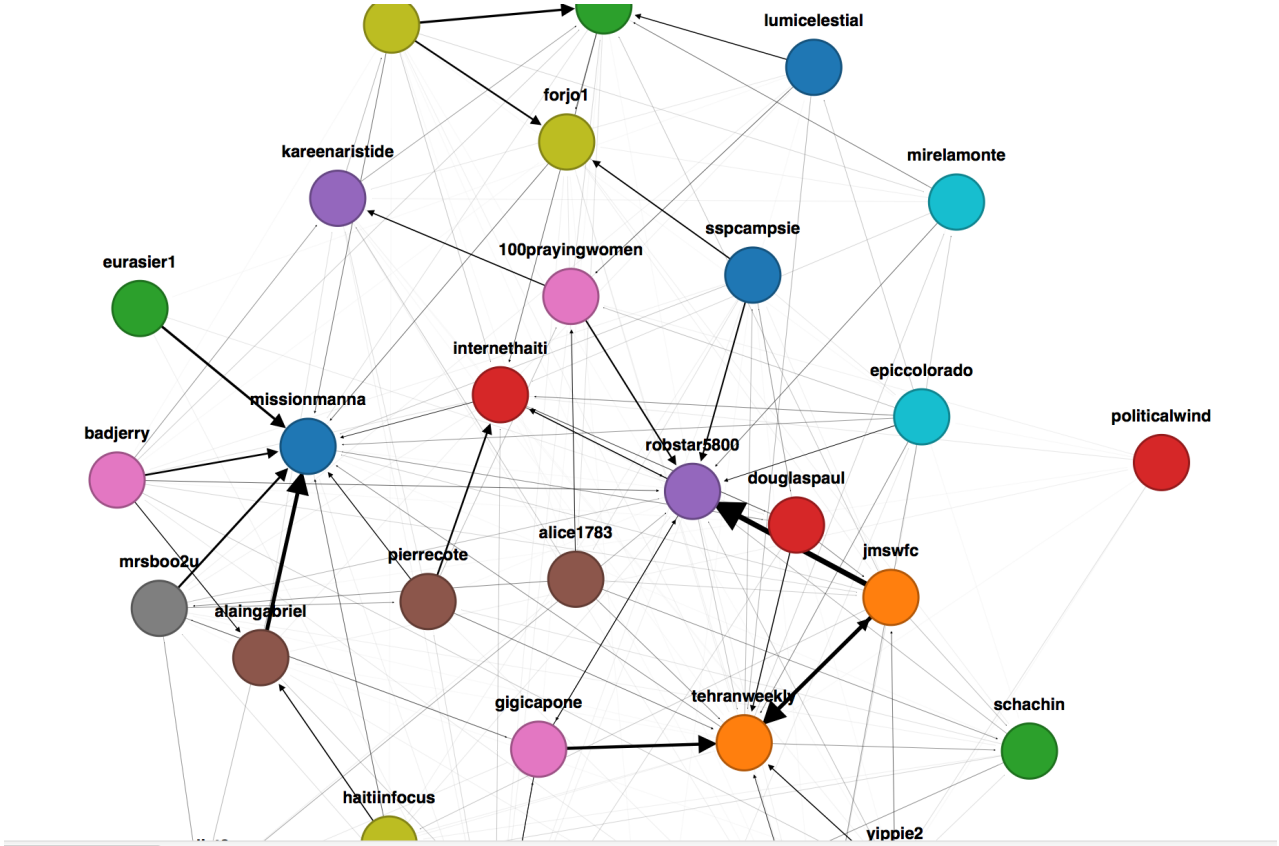
Demo 1: Synthetic Dataset

Synthetic Dataset generated from VAR models with 12 time series



Demo 2: Twitter Dataset

Twitter dataset generated from Tweet behaviors by top 20 Twitter users on Haiti-earthquake



Bibliography I

Philosophical Interpretations
— Stanford Encyc. of Philosophy. Causal Processes; Counter-factual Theories of Causation
Randomization Tests
— Sheehan, N. A., Didelez, V., Burton, P. R., & Tobin, M. D. (2008). Mendelian randomisation and causal inference in observational epidemiology. PLoS medicine.
— La Fond, T., & Neville, J. (2010). Randomization tests for distinguishing social influence and homophily effects. WWW'10
SEM and Partial Correlation Algorithms
— Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys.
— Pellet;, J.-P., & Elisseeff, A. (2007). Partial Correlation- and Regression-Based Approaches to Causal Structure Learning. IBM Technical Report
— Schmidt, M., Niculescu-Mizil, A., & Murphy, K. (2007). Learning graphical model structure using L1-regularization paths.
Theory
— Robins, J. M., Scheines, R., Spirtes, P., & Wasserman, L. (2003). Uniform consistency in causal inference. Biometrika, 90(3), 491-515. doi:10.1093/biomet/90.3.491
Non-Gaussian SEM
— Shimizu, S., & Kano, Y. (2008). Use of non-normality in structural equation modeling: Application to direction of causation. Journal of Statistical Planning and Inference.
— Shimizu, S., Hoyer, P. O., Hyvrinen, A., & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research.

— Dodge, Y., & Rousson, V. (2001). On Asymmetric Properties of the Correlation Coefficient in the Regression Setting. The Journal of American Statistical Association.

Bibliography II

Non-linear Transformations
— Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., & Schlkopf, B. (2008). Nonlinear causal discovery with additive noise models. NIPS
— Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., & Schoelkopf, B. (2010). Inferring deterministic causal relations. UAI
Granger Causality
— Anil Seth (2007) Granger causality. Scholarpedia, 2(7):1667.
— Michael Eichler. Causal inference from time series: What can be learned from granger causal- ity? Technical report, 2008.
— Valds-Sosa, P. A., Snchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernndez, M., Bosch-Bayard, J., Melie-Garca, L., & Canales-Rodrguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. Philosophical transactions of the Royal Society of London. Series B, Biological sciences.
— Look into publications of Prof. Liu
Transfer Entropy
— Schreiber, T. (2000). Measuring Information Transfer. Physical Review Letters.
— Barnett, L. (2009). Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. Physical Review Letters.
General Reading
— Cosma Shalizi, Causality Inference lecture note:
`http://www.stat.cmu.edu/~cshalizi/350/lectures/31/lecture-31.pdf`.
Causality Reference List. `http://cscs.umich.edu/~crshalizi/notabene/causality.html`