SPARSE AUTOREGRESSIVE MODEL ESTIMATION FOR LEARNING GRANGER CAUSALITY IN TIME SERIES

Chulalongkorn University
จุฬาลงกรณ์มหาวิทยาลัย
Pillar of the Kingdom

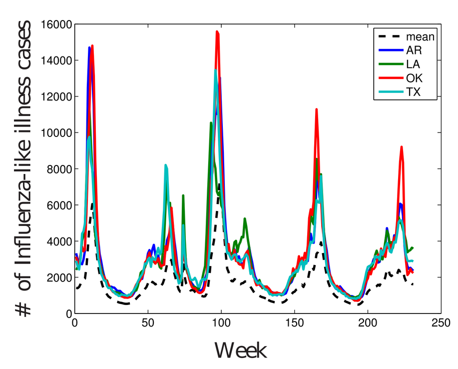
JITKOMUT SONGSIRI
email: jitkomut.s@chula.ac.th

Autoregressive Models

explain a multivariate time series by a vector AR process of order p

$$y(t) = A_1 y(t-1) + A_2 y(t-2) + \cdots + A_p y(t-p) + u(t)$$

$y \in \mathbf{R}^n$, $A_k \in \mathbf{R}^{n \times n}$, $k = 1, 2, \ldots, p$ u is noise



$n = 51$ (51 states in the U.S.)

y_1 the number of patients in AK

y_2 the number of patients in LA

$\vdots$

y_{51} the number of patients in WA

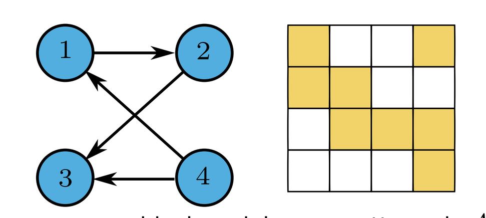
Granger Graphical Models (Granger1969)

sparsity in coefficients

$$(A_k)_{ij} = 0, \quad k = 1, 2, \ldots, p$$

is the characterization of Granger causality of AR models

- y_i is not Granger-caused by y_j
- knowing y_j does not help improve the prediction of y_i



granger graphical model zero patterns in A_k

for example, 4-dimensional AR

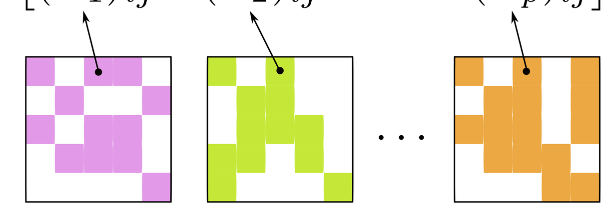
y_2 is Granger caused by y_1

y_4 is NOT Granger caused by y_2

Group Sparsity

stack the (i,j) entries of all A_k's in vector $B_{ij} \in \mathbf{R}^p$

$$[(A_1)_{ij} \quad (A_2)_{ij} \quad \cdots \quad (A_p)_{ij}] \triangleq B_{ij}$$



obtain a group sparsity in A_k's if we can enforce

$$\|B_{ij}\|_2 = 0, \quad \text{or} \quad \|[(A_1)_{ij} \quad (A_2)_{ij} \quad \cdots \quad (A_p)_{ij}]\|_2 = 0$$

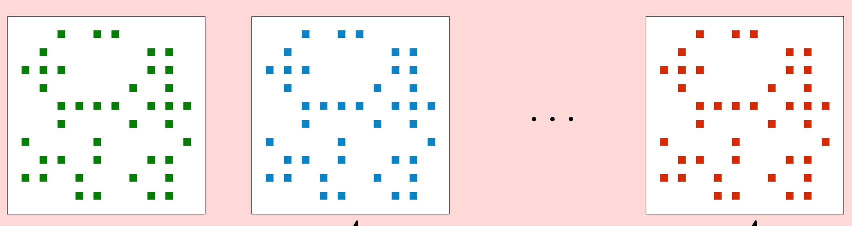
for some (i,j)

Sparse Autoregressive (AR) Models

Problem: find A_k's that minimize the sum-square error

$$\sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|_2^2$$

- A_k's contain many zeros (to infer Granger causality among variables)
- $A_1, A_2, \ldots, A_p$ have a common zero pattern



this formulation finds many applications in neuroscience and system biology

(Salvador et al. 2005, Valdes-Sosa et al. 2005, Fujita et al. 2007, ...)

Constrained AR Estimation

given the measurements $y(1), y(2), \ldots, y(N)$

$$\begin{array}{ll} \text{minimize} & \sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|^2 \\ \text{subject to} & (A_1)_{ij} = (A_2)_{ij} = \cdots = (A_p)_{ij} = 0, \quad (i,j) \notin \mathcal{V} \end{array}$$

with variables $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, 2, \ldots, p$

- $\mathcal{V}$ is the index set of a given Granger causality constraint
- the equality constraints can be eliminated, resulting in a reduced least-squares
- the solution is then analytically obtained

Sparse AR Estimation

given the measurements $y(1), y(2), \ldots, y(N)$

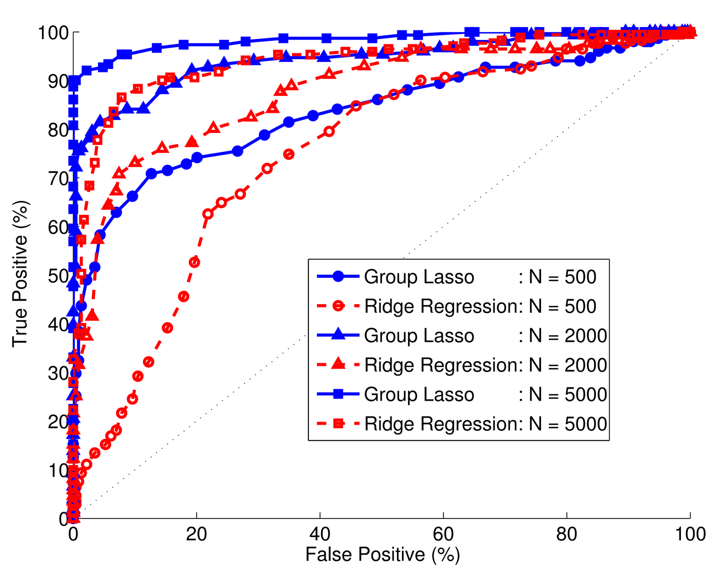
$$\text{minimize} \quad \sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|^2 + \lambda \sum_{i \neq j} \|[(A_1)_{ij} \cdots (A_p)_{ij}]\|_2$$

with variables $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, 2, \ldots, p$

- regarded as an ℓ_1-regularized least-squares problem
- summation over (i,j) plays a role of ℓ_1-type norm
- using the ℓ_2 norm of p-tuple of $(A_k)_{ij}$ yields a group sparsity
- λ is called a regularization parameter ($\lambda > 0$)

a heuristic convex approach to obtain sparse AR coefficients

ROC Curve



Receiver Operating Characteristic (ROC) curves of our approach (blue solid) and ridge regrssion (red dashed)

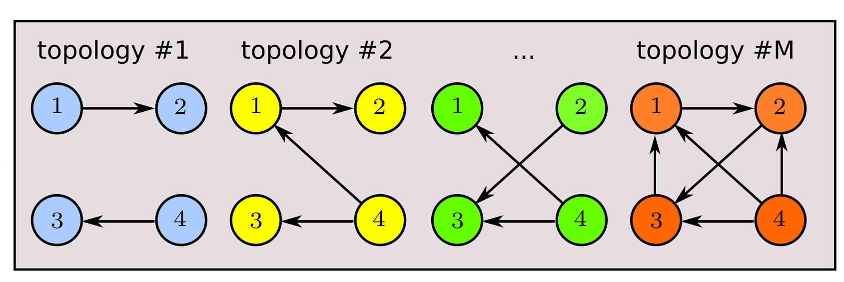
Sparse AR estimation performs better than Ridge regression even when N (number of samples) is small

Model Selection

sparse AR estimation

$$\begin{array}{ll} \text{minimize} & (1/2)\|Y - AH\|_2^2 + \lambda g(A) \\ \text{where} & g(A) = \sum_{i \neq j} \|[(A_1)_{ij} \ (A_2)_{ij} \ \cdots \ (A_p)_{ij}]\|_2 \end{array}$$

vary λ



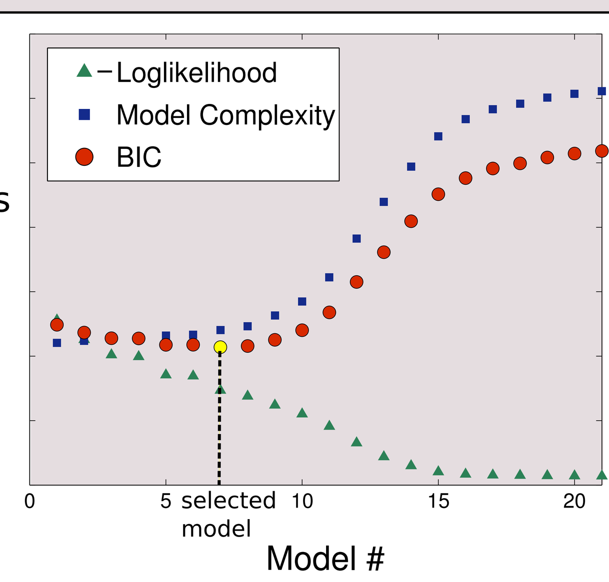
Granger constraints

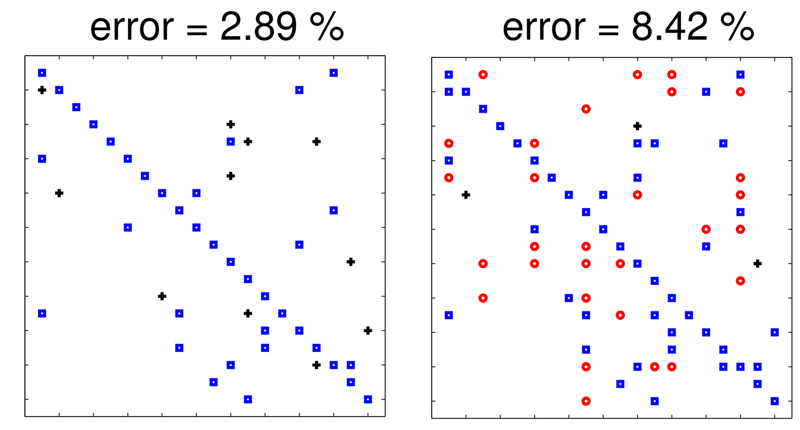
constrained AR estimation

$$\begin{array}{ll} \text{minimize} & (1/2)\|Y - AH\|_2^2 \\ \text{subject to} & (A_1)_{ij} = (A_2)_{ij} = \cdots = (A_p)_{ij} = 0 \end{array}$$

Model Candidates

BIC = -2 Loglikelihood + Model Complexity





BIC Crossvalidation

Comparison of the true and estimated sparsity patterns

- □ correctly identified nonzero entries.
- ○ misclassified entries as nonzero.
- + misclassified entries as zeros.

BIC gives a smaller error when the true model is sparse

Alternating Direction Method of Multiplier (ADMM)

Initialize $A^{(0)}, Z^{(0)}, U^{(0)}$ and set an ADMM parameter $\rho > 0$

$$A^{(k+1)} = \operatorname{argmin} \ \tfrac{1}{2}\|Y - AH\|_2^2 + \tfrac{\rho}{2}\|A - Z^{(k)} + U^{(k)}\|_F^2$$

$$Z^{(k+1)} = \operatorname{argmin} \ \left\{ (\rho/2)\|A^{(k+1)} - Z + U^{(k)}\|_F^2 + \lambda \sum_{i \neq j} \|[(Z_1)_{ij} \quad (Z_2)_{ij} \quad \cdots \quad (Z_p)_{ij}]\|_2 \right\}$$

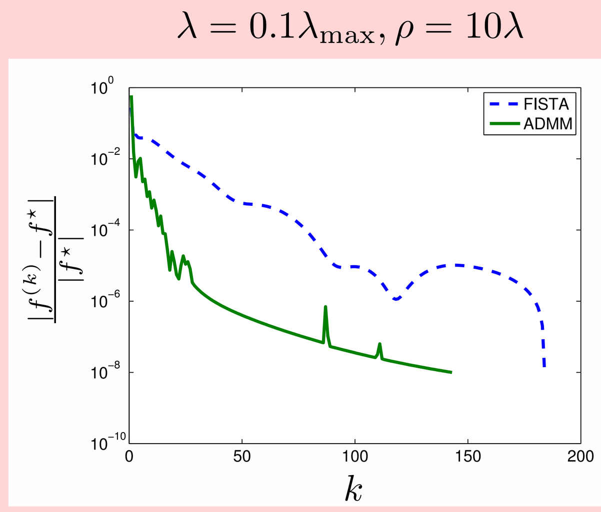
$$U^{(k+1)} = U^{(k)} + A^{(k+1)} - Z^{(k+1)}$$

until a stopping criterion is satisfied (Boyd et. al. 2010)

- $A-$ update takes the form of ridge regression
- $Z-$ update has a soft thresholding formulation
- each step can be computed efficiently

$n = 100$, $p = 3$

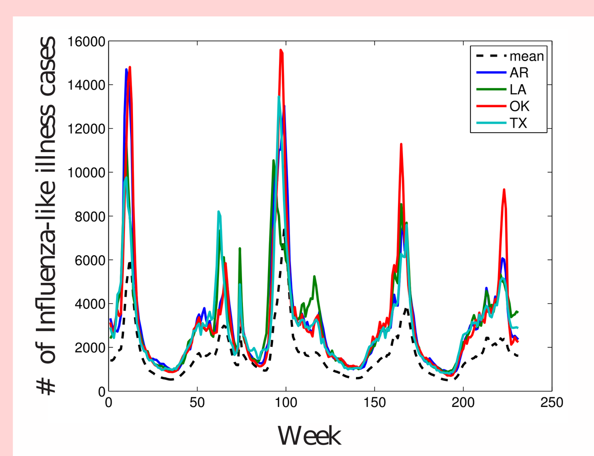
$\lambda = 0.1\lambda_{\max}$, $\rho = 10\lambda$



total 30,000 variables

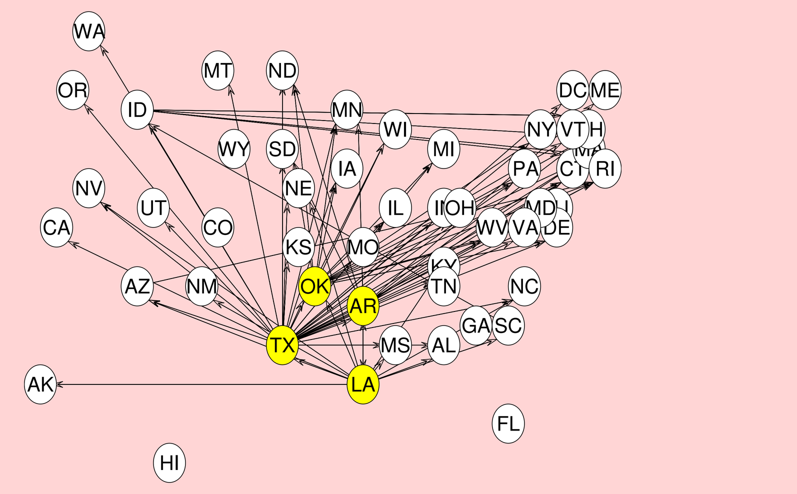
solved in 15-30 seconds

Google Flu Trend



- show the number of influenza-like illness (ILI) cases per 100,000 population (estimated by Google)
- Arkansas, Texas, Oklahoma and Louisiana are among the states that have higher numbers of ILI cases than the mean value

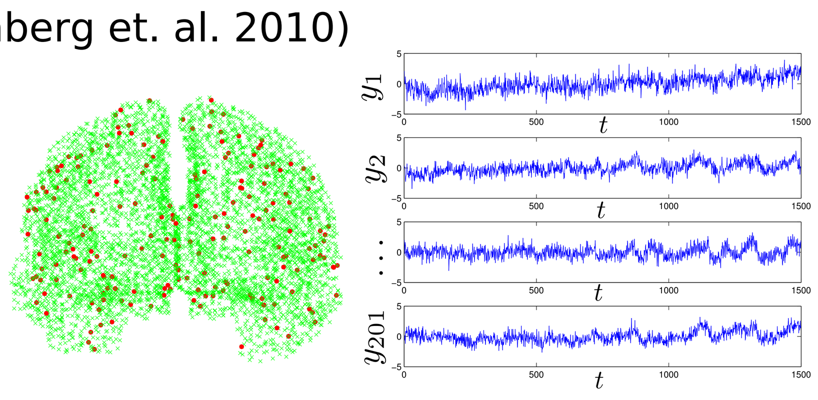
http://www.google.org/flutrends



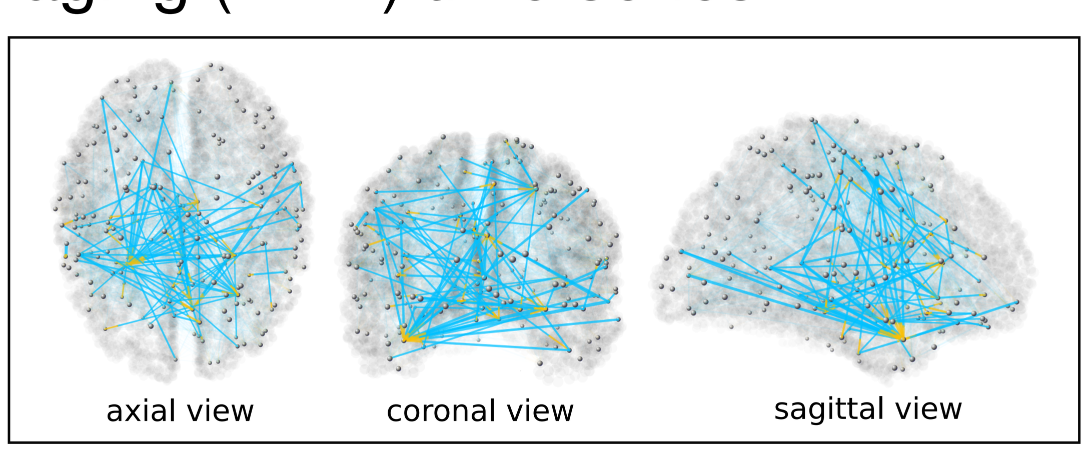
- TX, OK, LA, and AR have significant influences on many states
- factors such as climate, geography and public health policies can be taken into account to verify this result

Functional Magnetic Resonance Imaging (fMRI) time series

(Feinberg et. al. 2010)



- the data were obtained while a subject was in the resting state
- BOLD signals recorded at 6004 voxels with 1499 time samples
- reduce the number of voxels to 201 (red dots)



axial view coronal view sagittal view

- BIC selects the AR model of order 1 and the graph density is 7%
- orange color painted at the link end towards node j represents that the node j is Granger-caused by other nodes.
- temporal lobes, and the prefrontal cortex are the main elements of brain functional in the resting state

Conclusions

We have presented a convex framework for learning a topology in Granger graphical models, which is equivalent to estimating autoregressive models and promoting a joint sparsity in the AR coefficients simultaneously. The formulation is a least-squares problem with an L1-type regularization. We have investigated the ADMM algorithm which is very simple to implement numerically and has a desirable rate of convergence in practice. Moreover, we have described a model selection method for learning the most suitable sparsity pattern (or graph topology) for the given data. Using BIC score tends to pick a sparse model, which result in a low estimation error if the true model is also sparse, while the cross validation technique favorably selects a denser model. Experiment with randomly generated data sets, time series of Google flu trends and fMRI were included to confirm the effectiveness of our approach.