Estimation of Granger Causality of State-Space Models using Clustering with Gaussian Mixture Model

Jitkomut Songsiri

Nattaporn Plub-in

Department of Electrical Engineering

CHULA ENGINEERING

Foundation toward Innovation

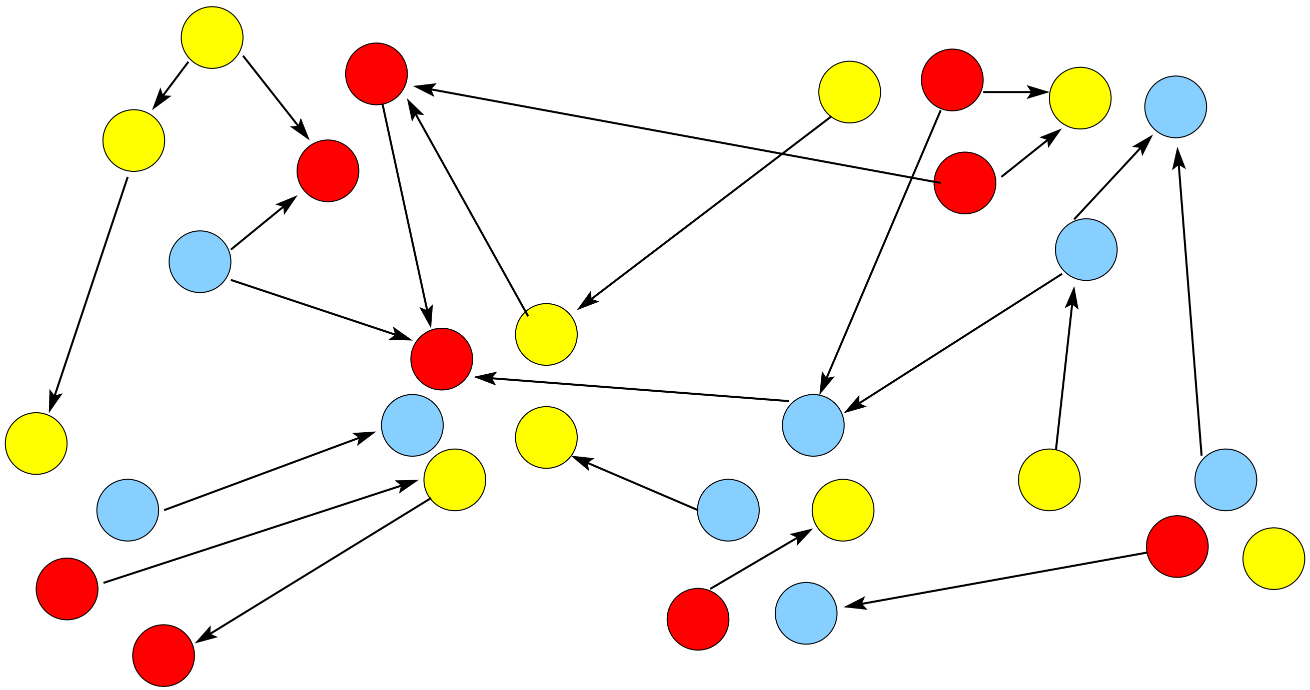
Outline

- Granger causality (GC) of state-space models
- Gaussian mixture model
- learning GC pattern
- experiment and evaluation
- results

Granger causality (Granger 1969)

let $x(t) = (x_1(t), \ldots, x_n(t))$ be multivariate time series

- x_i is not **Granger-caused** by x_j
- knowing x_j does not help to improve the prediction of x_i

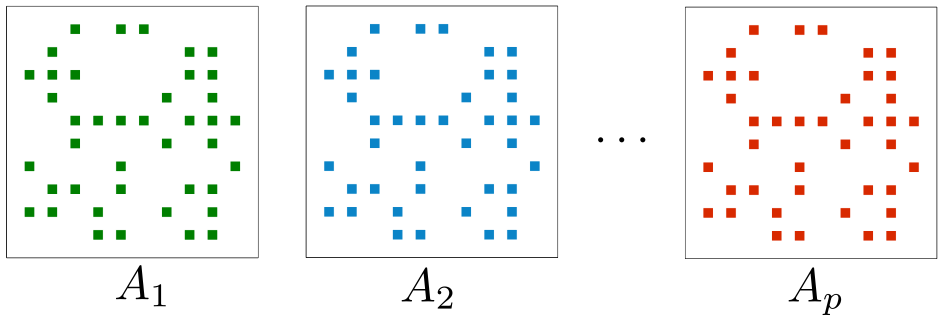


leads to a problem of learning inter-dependence relationships among variables

Characterization of GC for vector autoregressive (VAR) processes

$$x(t) = A_1 x(t-1) + A_2 x(t-2) + \cdots + A_p x(t-p) + \nu(t)$$

can be explained from a sparsity pattern of coefficients A_k



if (i, j) entries of A_k's are all zero

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

then x_j is NOT a Granger cause for x_i

problems of estimating VAR models with sparse A_k's have been proposed

Granger causality of state-space models (Barnett and Seth 2015)

state-space equations:

$$z(t+1) = Az(t) + w(t) \tag{1a}$$

$$x(t) = Cz(t) + \eta(t) \tag{1b}$$

goal: find a Granger causality characterization in terms of model parameters

- an extension of GC characterization from the typically used VAR process
- x_j is not a Granger cause for x_i if

$$F_{ij} \equiv F_{x_j \to x_i | \text{all other } x} = \log\left(\frac{\Sigma^R_{ii}}{\Sigma_{ii}}\right) = 0$$

where Σ and Σ^R are the prediction (in x) error covariances of the **full** and **reduced** models respectively

Full and reduced models

when testing if x_j has a Granger cause to x_i where x is output of state equation

$$z(t+1) = Az(t) + w(t)$$

a reduced model has all variables except x_j

$$\begin{array}{rrl} \text{full model} & x(t) & = Cz(t) + \eta(t) \\ \text{reduced model} & x^R(t) & = C^R z(t) + \eta^R(t) \end{array}$$

where x^R has all entries of x except x_j

$$x^R(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ \boxed{x_j(t)} \\ \vdots \\ x_n(t) \end{bmatrix} \rightarrow \text{removed} \quad C^R = \begin{bmatrix} C_1^T \\ \vdots \\ \boxed{C_j^T} \\ \vdots \\ C_n^T(t) \end{bmatrix} \rightarrow \text{removed}$$

Prediction error covariance via Kalman filter

if state-space parameters are known, the best estimator of z (in MMSE sense) is

$$\hat{z}(t|t-1) = \mathbf{E}[z(t)|x(t-1), \dots, x(0)]$$

whose steady-state covariance, defined as

$$P(t|t-1) = \mathbf{cov}(z(t) - \hat{z}(t|t-1))$$

can be characterized by **Kalman filter** and asymptotically solved from DARE

$$P = APA^T - (APC^T + S)(CPC^T + N)^{-1}(CPA^T + S^T) + W$$

asymptotically, covariance of output estimation error is

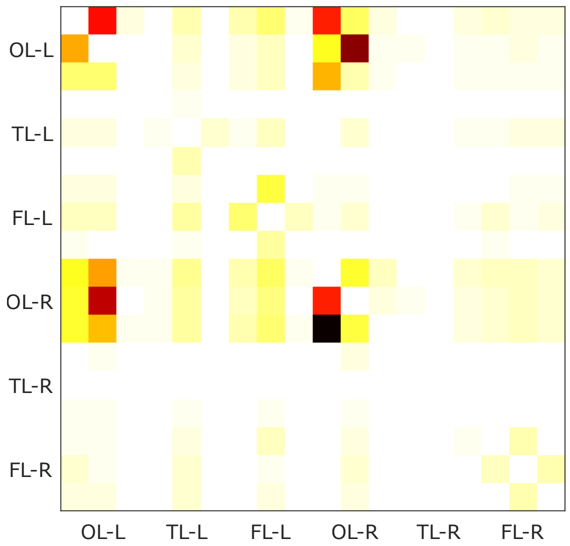
$$\Sigma = \mathbf{cov}(x(t) - \hat{x}(t|t-1) = CPC^T + N$$

GC matrix estimation

in order to estimate GC pattern, we need to estimate

- system matrices (A, B, C, D)
- noise covariances

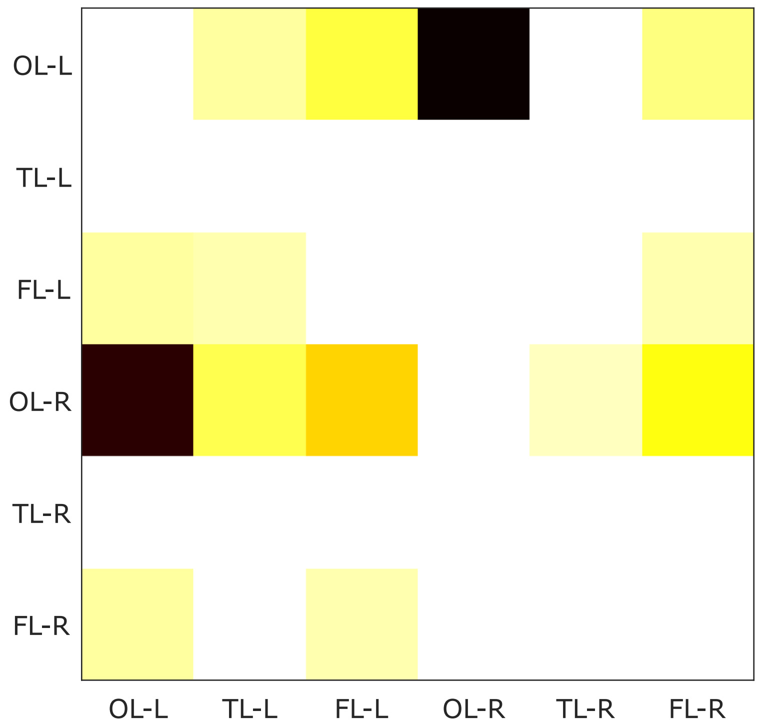
which can be done by several methods (here, using subspace identification)



- GC pattern of n variables is represented by the matrix F of size $n \times n$
- a significance test of entries estimated F_{ij} is needed
- statistical distribution of F_{ij} is still unknown
- if we have many samples of estimated F_{ij} then its sample mean can be approximated by Gaussian

Application in learning brain connectivity

learning a brain network from EEG time series



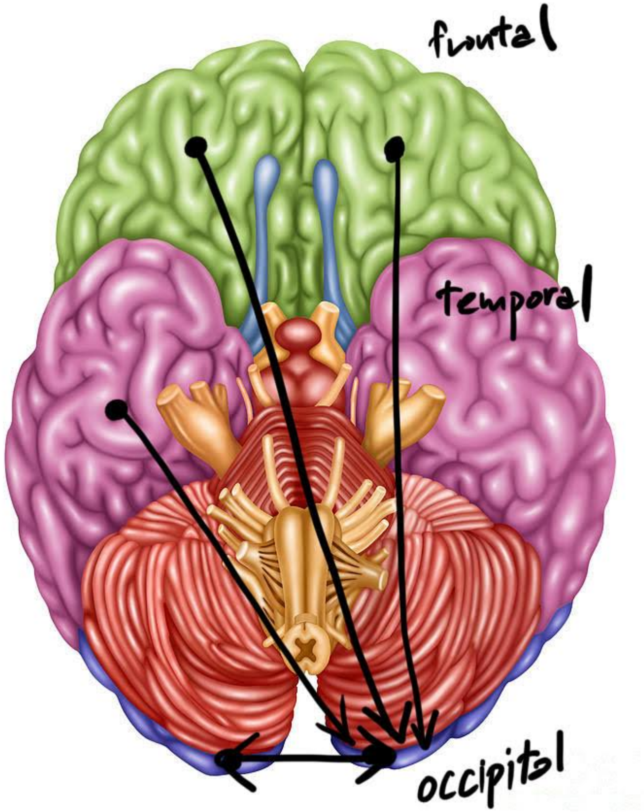


photo credit: Gwen Shockey, https://pixels.com

Gaussian mixture model (GMM)

model setting:

- let $Y_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ for $k = 1, 2, \ldots, K$
- let $(Z_1, Z_2, \ldots, Z_K)$ be latent variable with multinomial distribution

a GMM model takes the form of a linear sum of K Gaussian components:

$$Y = Z_1 Y_1 + Z_2 Y_2 + \cdots + Z_K Y_K$$

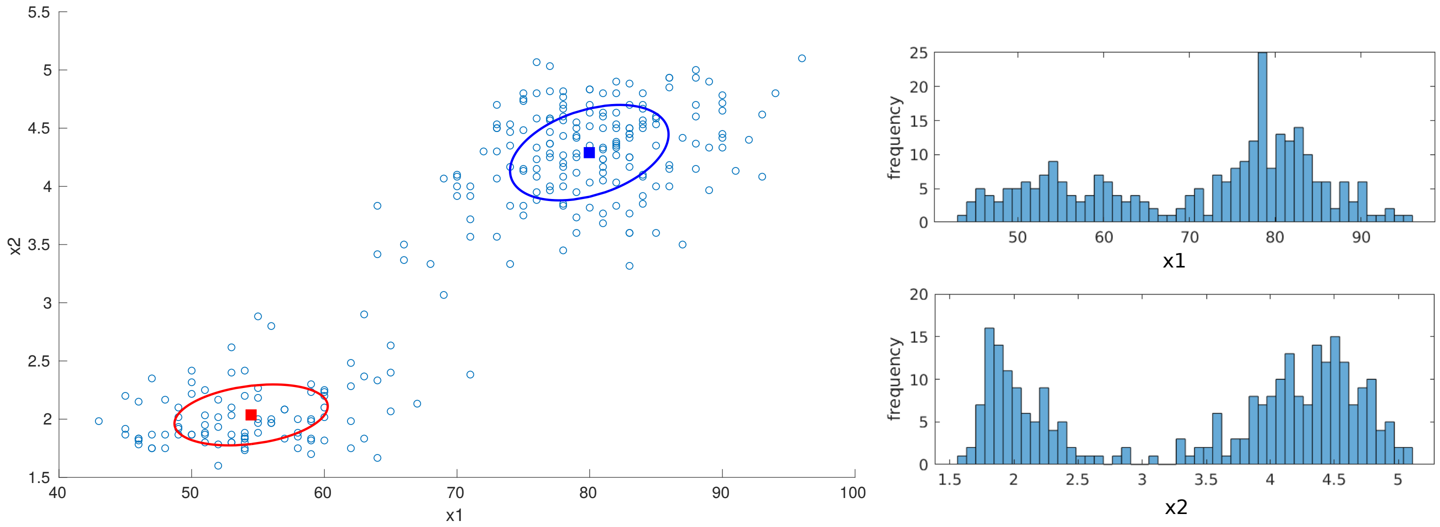
- given that $Z = \mathbf{e}_k$ (standard unit vector), Y is distributed as the kth Gaussian
- the pdf of Y is given by

$$f(y) = \pi_1 f_1(y; \mu_1, \sigma_1^2) + \cdots + \pi_K f_K(y; \mu_K, \sigma_K^2)$$

where $(\pi_1, \ldots, \pi_K)$ is pmf of Z and f_k's are Guassian density

Clustering using GMM

when samples of Y appear to be clustered as multimodal Gaussians

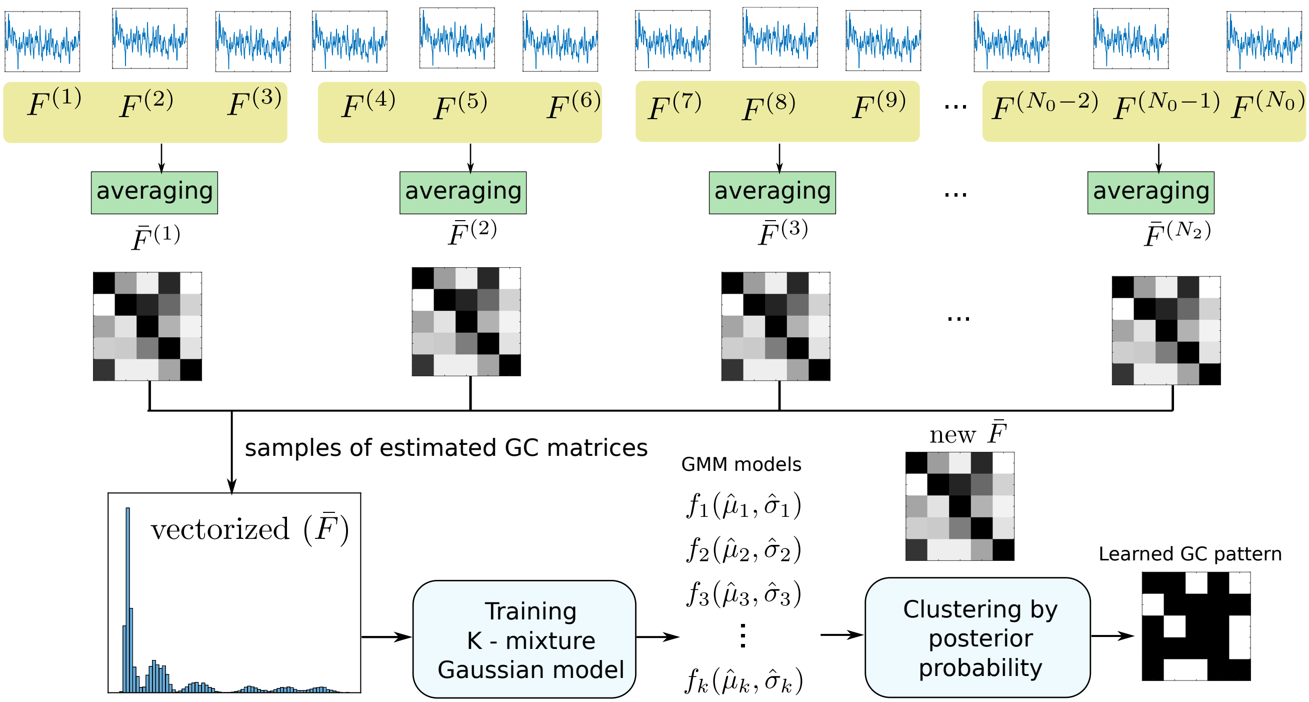


- trainning: estimate parameters of GMM $(\pi_k, \mu_k, \sigma_k^2)$ by EM algorithm
- clustering: if unseen sample of Y is given, we compute posterior pdfs

$$f_k(y \mid Z = \mathbf{e}_k; \mu_k, \sigma_k^2) P(Z = \mathbf{e}_k; \mu_k, \sigma_k^2), \quad k = 1, 2, \ldots, K$$

the kth cluster with highest posterior pdf is assigned to be the label of Y

Learning GC pattern



- assume we have many trials of estimated F
- take an average of those trials to have many samples of $\bar{F}$; each of which can be approximated by a Gaussian
- pool all F_{ij}'s and use GMM to cluster entries to each of Gaussian modes

Number of GMM components

the number of components is chosen from

- Bayesian informatic criterion score (BIC)

$$\mathrm{BIC} = -2\mathcal{L} + d \log N$$

- relative change in BIC: $\mathrm{rBIC}(k) = \mathrm{BIC}(k) - \mathrm{BIC}(k-1)$

- sillouhette score: a measure to determine how well the data are clustered
  - s is close to 1 if data are well clustered
  - s is close to 0 if data are on the the border of clusters
  - s is close to -1 if the data could have been clustered to its neighbour instead

silhouette score: consider two average distances

- from a point x_i to all points in the same cluster

$$a(x_i) = \frac{1}{\text{size of cluster}} \sum_{j \neq i} \mathbf{dist}(x_i, x_j)$$

- from a point x_i to all points points $x_j^{(k)}$ in other kth clusters

$$b(x_i) = \underset{k}{\text{minimize}} \frac{1}{\text{size of } k\text{th cluster}} \sum_{j} \mathbf{dist}(x_i, x_j^{(k)})$$

- sillouhette score of a point x_i is defined as

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}, \quad -1 \leq s(x_i) \leq 1$$

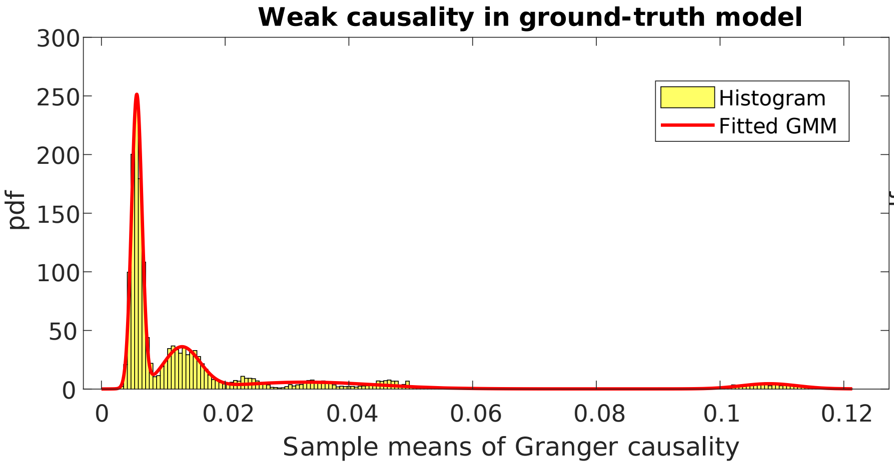
and the sillouhette score is the average over all points x_i in a cluster

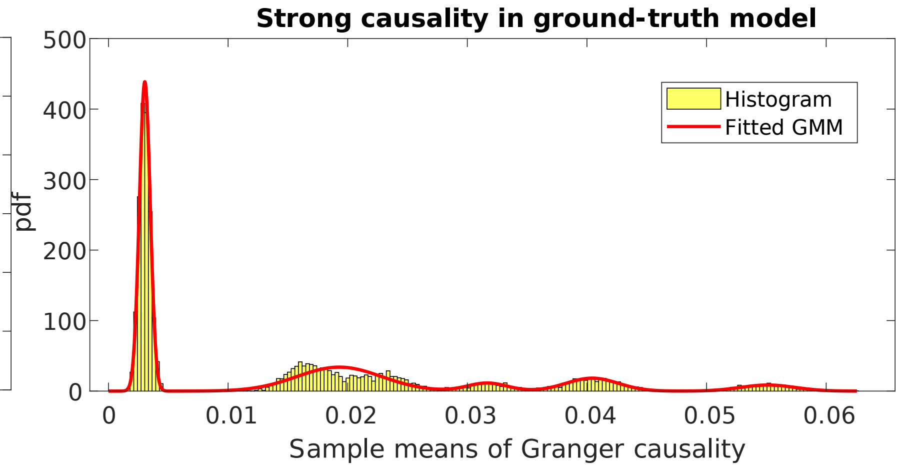
Experiment settings

our scheme of learning GC pattern is tested on simulated data sets

1. ground-truth state-space models with known GC patterns are generated
2. the ground-truth models consist of two types: strong and weak causality
3. 20,000 trials of time series are generated and state-space models are estimated using subspace identification
4. 1000 samples of $\bar{F}$ are split into training and test sets using 10-fold cross validation
5. number of GMM components are in the range of 1 to 10 and chosen by BIC, relative change of BIC (rBIC) or Silhouette score
6. classification metrics (FP,FN, and accuracy) are evaluated on test sets

Results of clustering entries in F





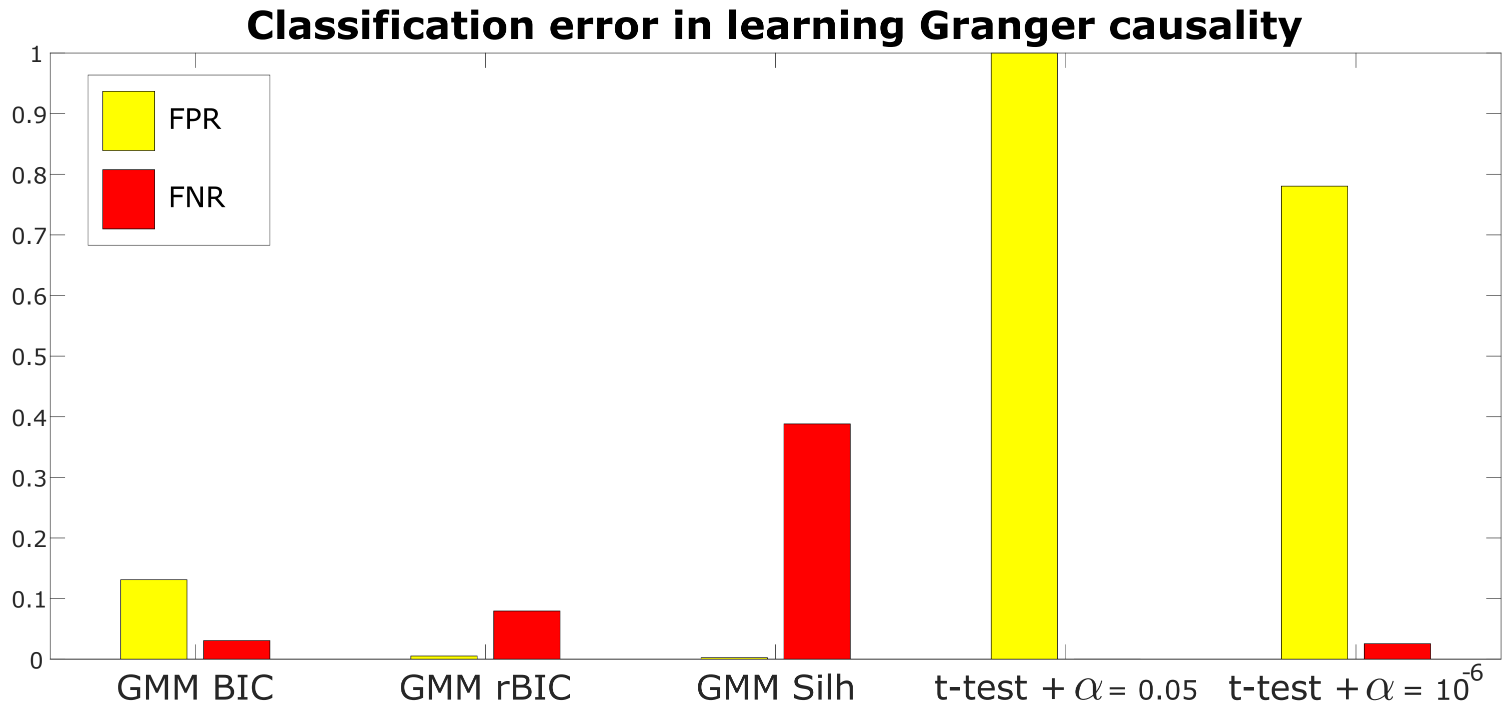
- rBIC gives a moderate number of Gaussian components
- when the ground-truth model has a strong causality, Gaussian components are well separated
- for weak causality, the fitted density functions of the first two components could be overlapped; leads to misclassifying between null and causal entries

Selected number of Gaussian components

	$N_0 = 2000$			$N_0 = 10000$		
Ground-truth model	BIC	rBIC	Silh	BIC	rBIC	Silh
Weak causality	6-9	4-7	2	7-10	4-7	2
Strong causality	6-8	3-5	2-6	6-10	3-6	2-7

- BIC tends to choose highest number of GMM components, while Silhouette score chooses lowest number
- rBIC selects a moderate number of GMM components
- GMM with too many modes tends to overly capture small entries of $\bar{F}$ (lead to high FP)
- GMM with a few modes lacks of flexibility to explain detailed characteristics of multi-modal shape of $\bar{F}$ (lead to high FN)

Errors in Granger causality learning



- the performance is best when number of GMM mode is chosen by rBIC
- t-test reject $H_0 : F_{ij} = 0$ most of the times

Conclusion

- we proposed a scheme of inferring Granger causality from estimated parameters of state-space models
- this finds applications to learning brain connectivity from EEG time series
- Granger causality measure (referred to as GC matrix) can be characterized via the concept of Kalman filter and computed from solving Riccati equation
- significant entries GC matrix can be clustered using Gaussian mixture models by an assumption that the sample mean of estimated GC matrices approaches a Gaussian distribution
- GMM performance is best when using relative change in BIC to choose the number of components
- the method requires multi-trial data for Gaussian assumption; this can be feasible for EEG application as the recordings are typically collected in a long period

Acknowledgment

Chula Engineering research grant

CHULA ENGINEERING
Foundation toward Innovation

Bangkok, Thailand