Directed Cyclic Graph for Causal Discovery from Multivariate Functional Data

Saptarshi Roy

Stat Café
November 8, 2023

Joint work with

Dr. Yang Ni

Dr. Raymond K. W. Wong

Table of contents

Background
- Causality - Statistics and beyond
- Causal Nomenclatures

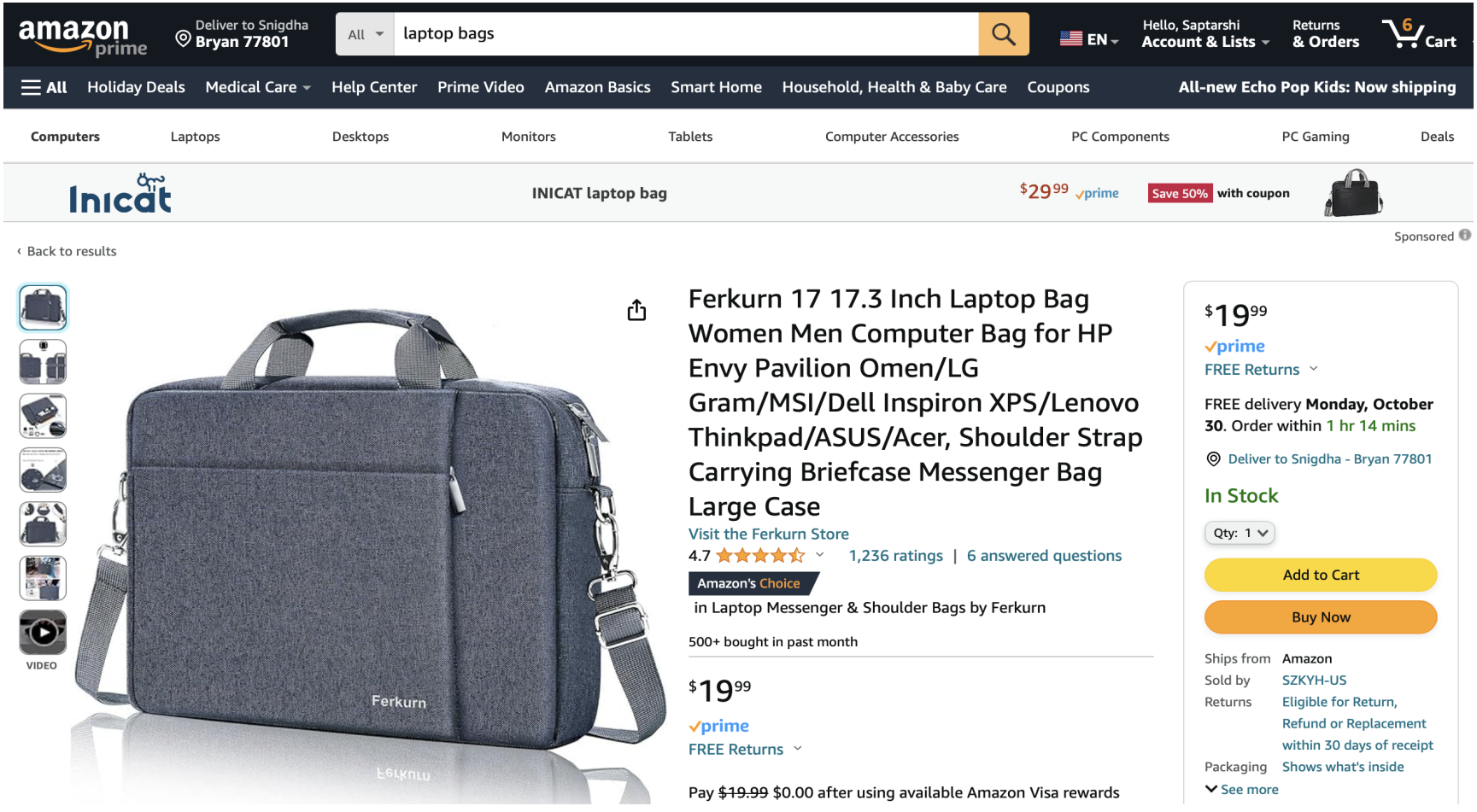
Multivariate Functional Data
- Definition & Example
- Challenges in Causal Discovery
- Motivating Example

Proposed Method
- Model Definition
- Causal Identifiability
- Bayesian Model Formulation
- Simulation Study
- Real Data Analysis

Conclusions

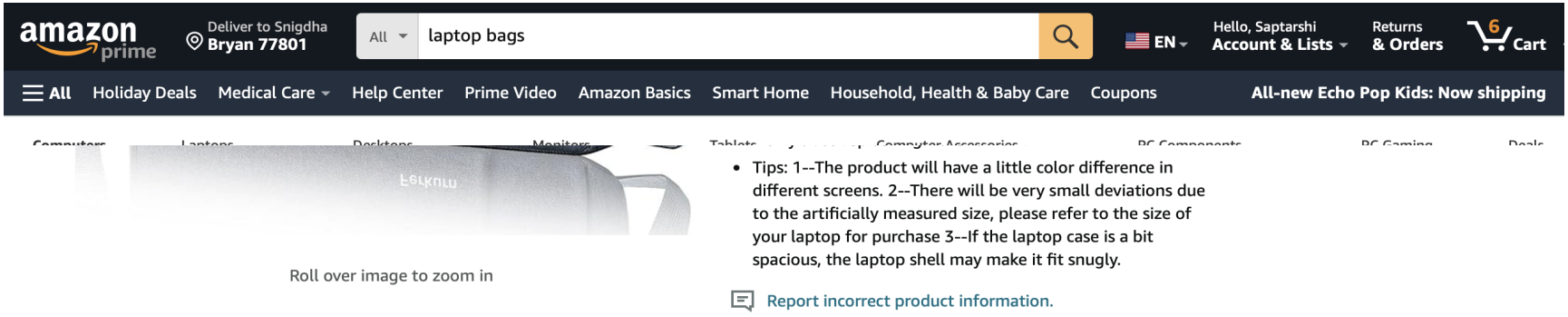
Why causality?

Example 1



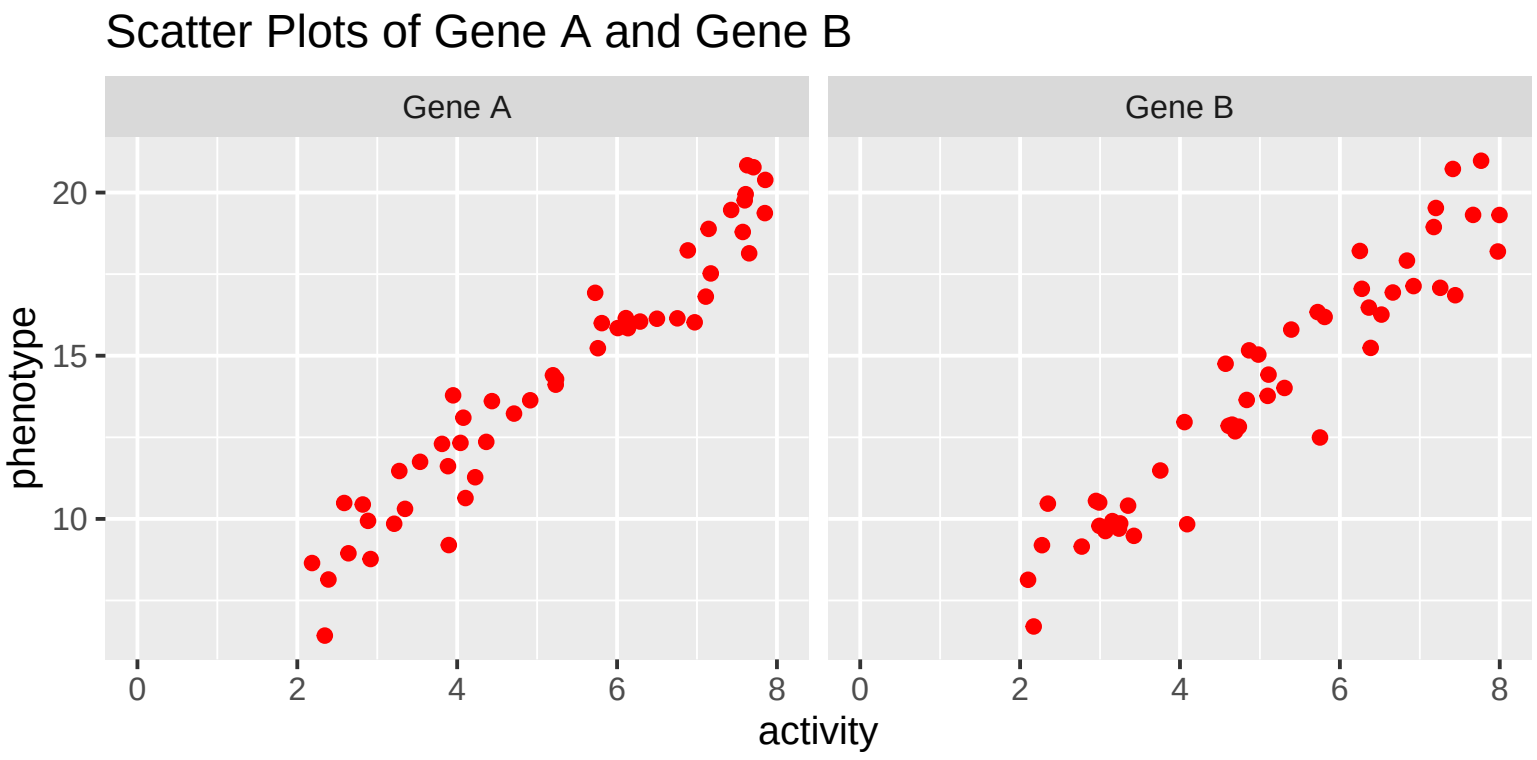
Why causality?

Example 1



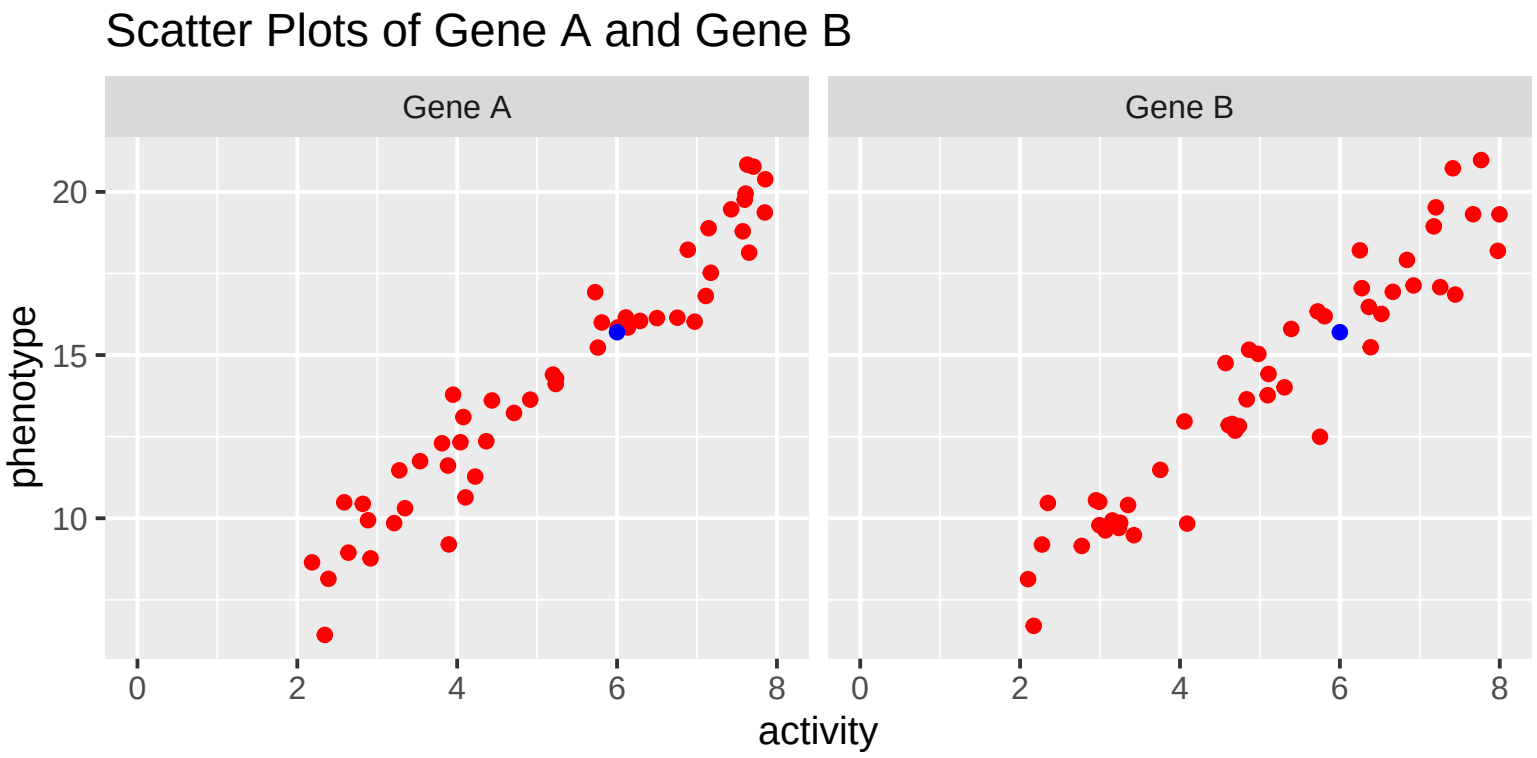
Why causality?

Example 2



Why causality?

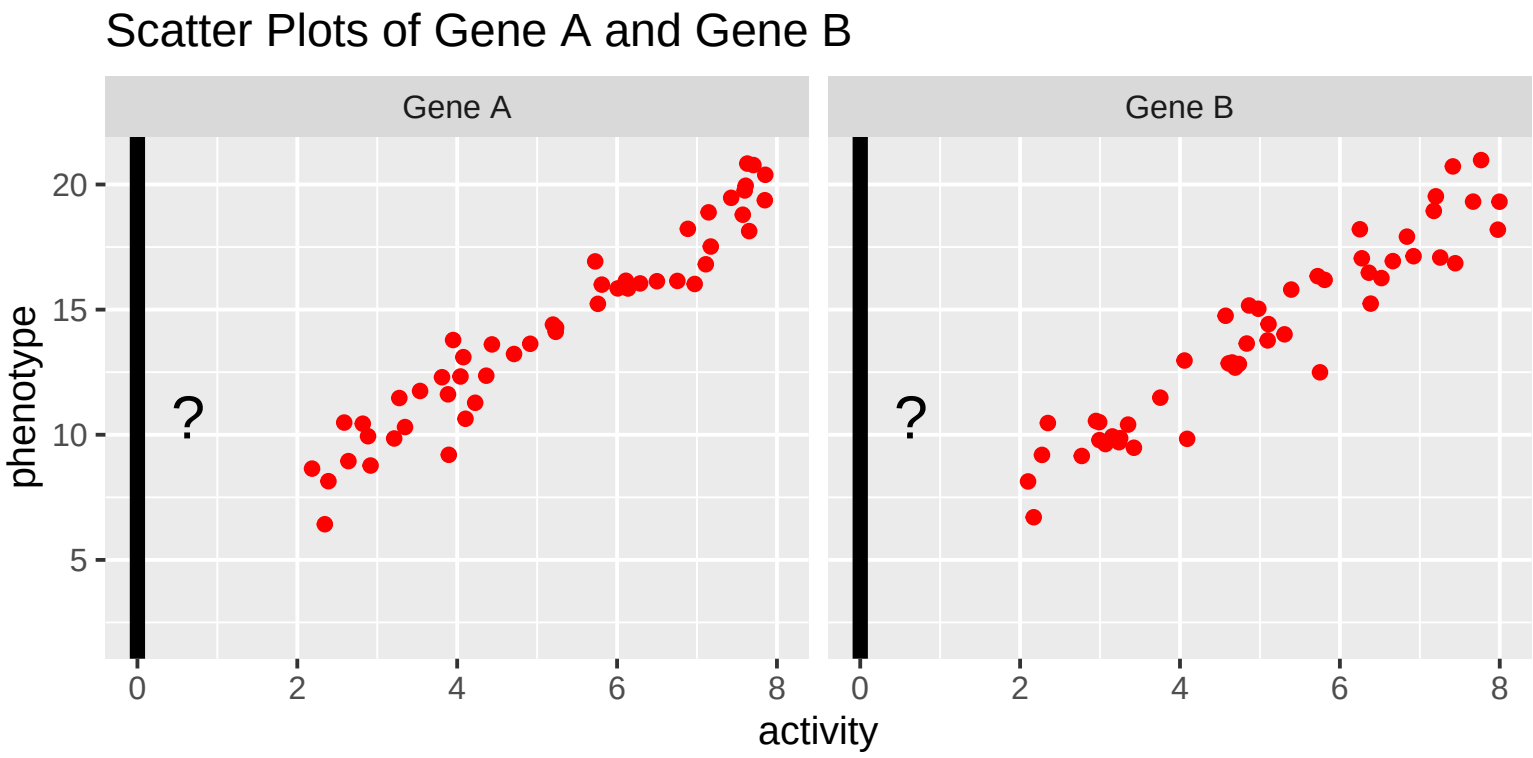
Example 2



Statistical Question!!

Why causality?

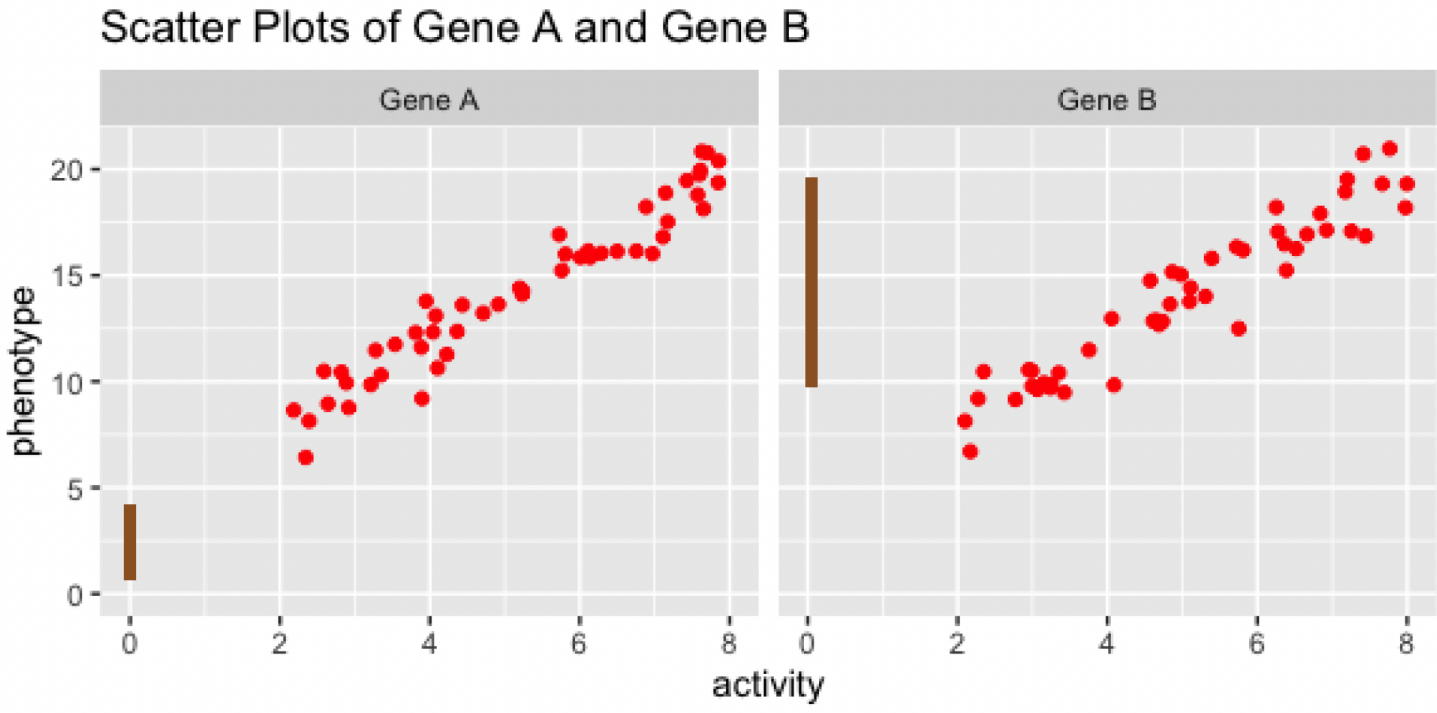
Example 2



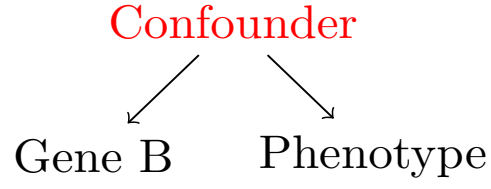
Causal Question!!

Causality matters

Example 2



Gene A $\longrightarrow$ Phenotype



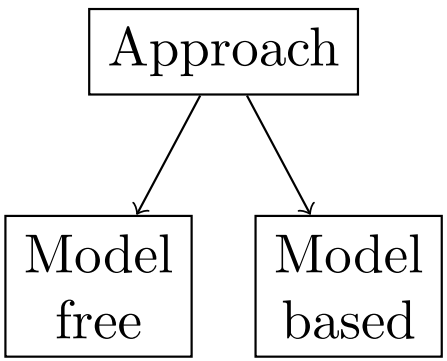
Causal discovery

- Causal discovery - infer causal structure from data.
- Gold standard is experimental data, however mostly we have observational data at our disposal.
- Fairly a difficult task to do.
- Suppose $X_1, \ldots, X_p$ forms a graph $\mathcal{G}$. Goal is to learn $\mathcal{G}$.
- Some usual assumptions:-
  1. Causal Markov condition – $X_i \perp\!\!\!\perp X_{\text{NonDesc}(i)} | X_{\text{pa}(i)}$.

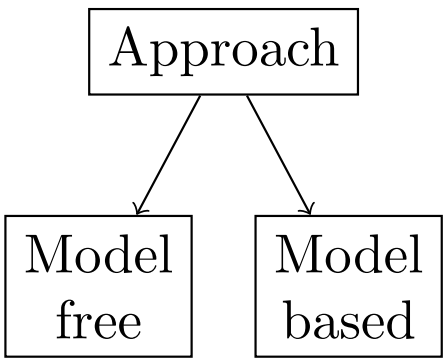
     smoking $\rightarrow$ tissue damage $\rightarrow$ lung cancer

  2. Causal faithfulness – $X_i \perp\!\!\!\perp_p X_j | X_k \implies X_i \perp\!\!\!\perp_{\mathcal{G}} X_j | X_k$.
  3. Causal sufficiency – Absence of any hidden confounders.
  4. Acyclicity – Use of Directed Acyclic Graphs (DAGs).

Current approaches

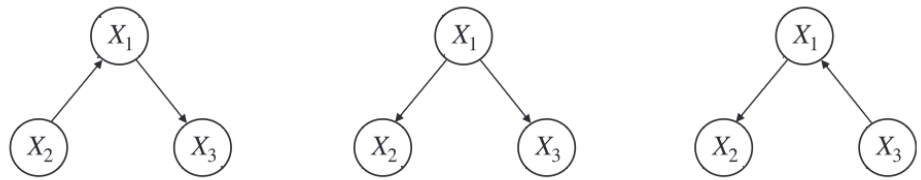


Current approaches



Model free approach:

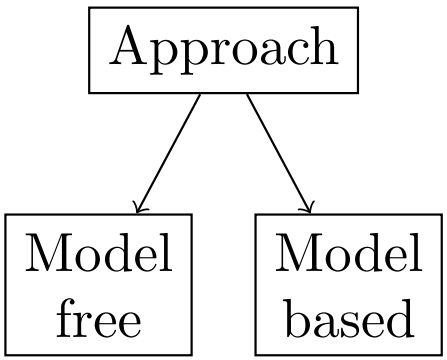
- Uses conditional independence tests – $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$.
- Non-parametric approach.
- Assumptions: causal Markov condition, causal faithfulness.
- determines the orientations of the edges up to the Markov equivalence class.
- Examples: PC, FCI, CCD.



Markov equivalence class of trivariate graphs - $X_2 \perp\!\!\!\perp X_3 | X_1$

Question: Can we do better than this?

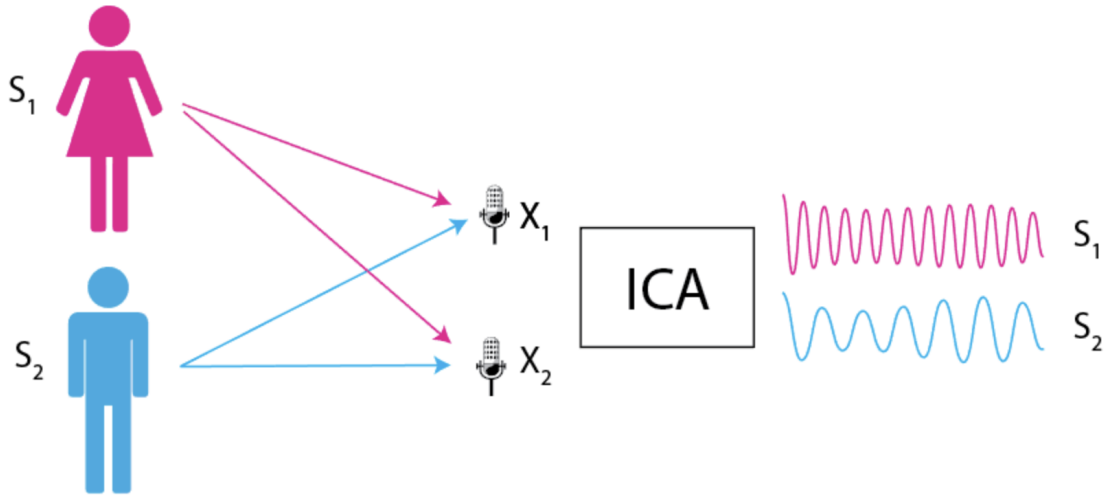
Current approaches



Model based approach:

- Structural equation model (SEM) $X_i = f_i(\boldsymbol{X}_{\mathrm{pa}(i)}, \epsilon_i)$.
- Parametric approach.
- Restricts the function class.
- The causal model is fully identifiable.
- Examples - LiNGAM, LiNG, PNL, Non-linear Additive Noise models etc.
- $f(\boldsymbol{X}_{\mathrm{pa}(i)}, \epsilon_i) = \boldsymbol{b}_i^\top \boldsymbol{X}_{\mathrm{pa}(i)} + \epsilon_i$

Independent component analysis



Goal is to recover the independent source signals S_i.

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$$

ICA & LiNGAM

ICA

- Model is $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$.
- Goal: to estimate $\boldsymbol{W} = \boldsymbol{A}^{-1}$.

LiNGAM

- Model is $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{X} = \boldsymbol{E}$.
- Underlying graph structure is acyclic.
- Goal: to estimate $\boldsymbol{B}$.

ICA & LiNGAM

ICA

- Model is $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$.
- Goal: to estimate $\boldsymbol{W} = \boldsymbol{A}^{-1}$.
- S_i are non-gaussian and mutually independent.
- Identifiable upto scaling and permutation indeterminacy of indep. components $\boldsymbol{S}$.

$$\boldsymbol{X} = \underbrace{(\boldsymbol{A}\boldsymbol{P}\boldsymbol{D})}_{\boldsymbol{A}^*}\underbrace{(\boldsymbol{D}^{-1}\boldsymbol{P}^\top\boldsymbol{S})}_{\boldsymbol{S}^*}$$

- several algorithms leads to efficient estimation of $\boldsymbol{W}$.

LiNGAM

- Model is $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{X} = \boldsymbol{E}$.
- Underlying graph structure is acyclic.
- Goal: to estimate $\boldsymbol{B}$.

ICA & LiNGAM

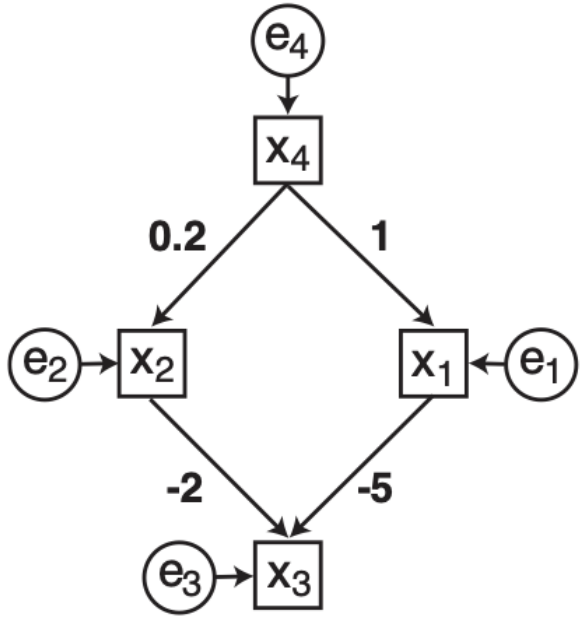
ICA

- Model is $\boldsymbol{X} = \boldsymbol{AS}$.
- Goal: to estimate $\boldsymbol{W} = \boldsymbol{A}^{-1}$.
- S_i are non-gaussian and mutually independent.
- Identifiable upto scaling and permutation indeterminacy of indep. components $\boldsymbol{S}$.

$$\boldsymbol{X} = \underbrace{(\boldsymbol{APD})}_{\boldsymbol{A}^*}\underbrace{(\boldsymbol{D}^{-1}\boldsymbol{P}^\top\boldsymbol{S})}_{\boldsymbol{S}^*}$$

- several algorithms leads to efficient estimation of $\boldsymbol{W}$.

LiNGAM



Unlike ICA, the correct correspondence between e_i and x_i is important in LiNGAM.

ICA & LiNGAM

ICA

- Model is $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$.
- Goal: to estimate $\boldsymbol{W} = \boldsymbol{A}^{-1}$.
- S_i are non-gaussian and mutually independent.
- Identifiable upto scaling and permutation indeterminacy of indep. components $\boldsymbol{S}$.

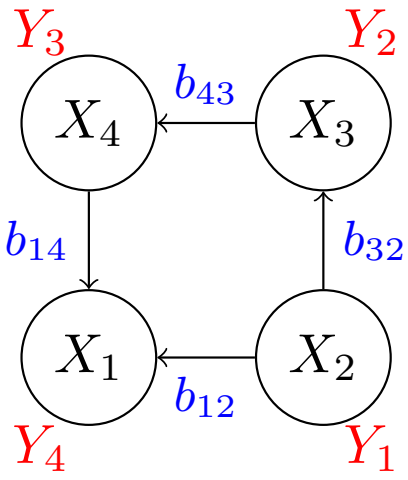
$$\boldsymbol{X} = \underbrace{(\boldsymbol{A}\boldsymbol{P}\boldsymbol{D})}_{\boldsymbol{A}^*}\underbrace{(\boldsymbol{D}^{-1}\boldsymbol{P}^\top\boldsymbol{S})}_{\boldsymbol{S}^*}$$

- several algorithms leads to efficient estimation of $\boldsymbol{W}$.

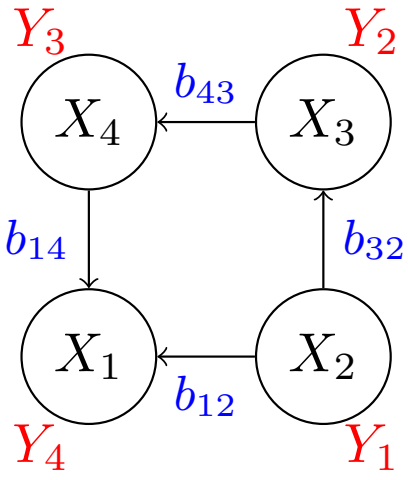
LiNGAM

- Model is $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{X} = \boldsymbol{E}$.
- Underlying graph structure is acyclic.
- Goal: to estimate $\boldsymbol{B}$.
- E_i are non-gaussian and mutually independent.
- Some post-processing needed on the $\widehat{\boldsymbol{W}}$.
- Efficient algorithm is provided in Shimizu et al. (2006).

Presence of cycles - Difficulty?



Presence of cycles - Difficulty?

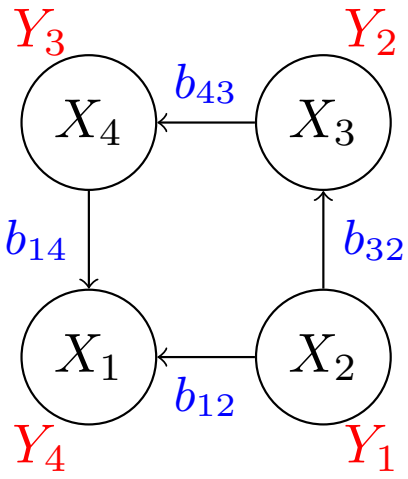


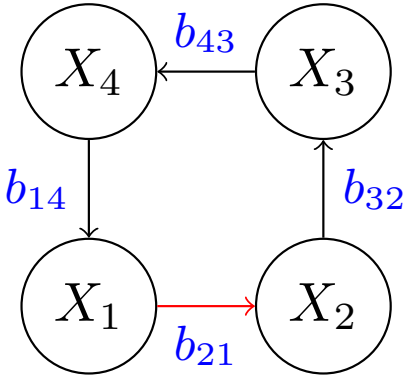
- DAGs iff topological order.
- Topological sort/order: $Y_i \to Y_j \implies j > i$.
- Example: $Y_1(X_2)$ $Y_2(X_3)$ $Y_3(X_4)$ $Y_4(X_1)$

$$Y_1 \longrightarrow Y_2 \longrightarrow Y_3 \longrightarrow Y_4$$

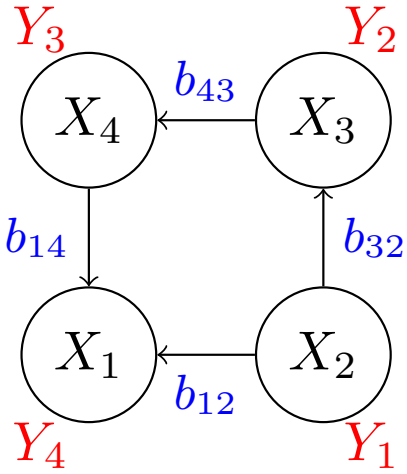
- $$\begin{pmatrix} Y_1 \\ \vdots \\ Y_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ b_{32} & 0 & 0 & 0 \\ 0 & b_{43} & 0 & 0 \\ b_{12} & 0 & b_{14} & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_4 \end{pmatrix} + \begin{pmatrix} e_2 \\ e_3 \\ e_4 \\ e_1 \end{pmatrix}$$
- $\boldsymbol{I} - \boldsymbol{B}$ is always invertible.

Presence of cycles - Difficulty?

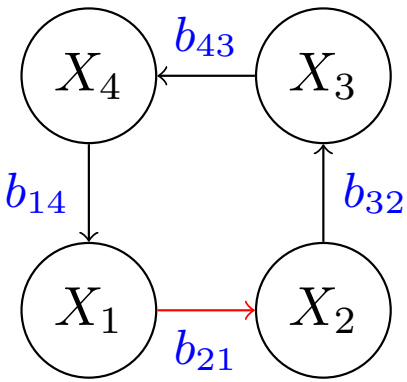




Presence of cycles - Difficulty?



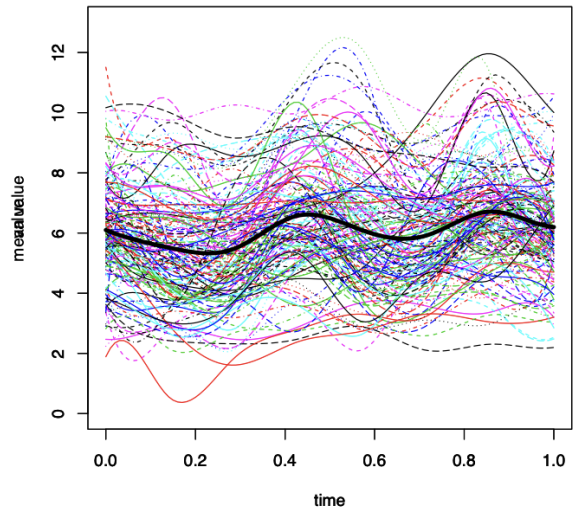
$$(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{X} = \boldsymbol{E}$$



- $\boldsymbol{B} = \begin{pmatrix} 0 & 0 & 0 & b_{14} \\ b_{21} & 0 & 0 & 0 \\ 0 & b_{32} & 0 & 0 \\ 0 & 0 & b_{43} & 0 \end{pmatrix}$
- $|\boldsymbol{I} - \boldsymbol{B}| = 1 - b_{14}b_{21}b_{32}b_{43}$.
- Some extra conditions are neccessary on $\boldsymbol{B}$ to ensure invertibility.
- The moduli of the eigenvalues of $\boldsymbol{B}$ are less than 1 and none equal to 1.

Definition

- Data where the data points are itself functions or curves.
- A realization of a (typically smooth) random object that takes values in an abstract function space.
- $\mathcal{H} := \{X | X : \mathcal{T} \to [a, b]\}$. Example: Hilbert space.
- Denoted by $X = (X_1, \ldots, X_p)^\top$.
- Examples: 1) daily PM10 concentration curves recorded in Graz, Austria in winter season 2) electrical activity measured across different regions of the brain (EEG data).



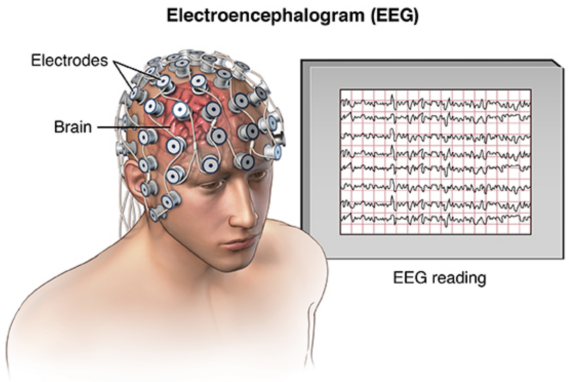
Source: Internet

Challenges

1. Infinite-dimensionality of functional data
   - Low-freq spectrum of Y_j might causally affect the high-freq. spectrum of Y_ℓ.
   - Demands identification of pertinent features.
   - The challenge is that we may not know *a priori* what these relevant features are.
2. Causal identifiability theory in presence of cycles/feedback loops.
3. Noisy functional data adds another layer of difficulty in probing the causal relationships of interest.

Motivating example

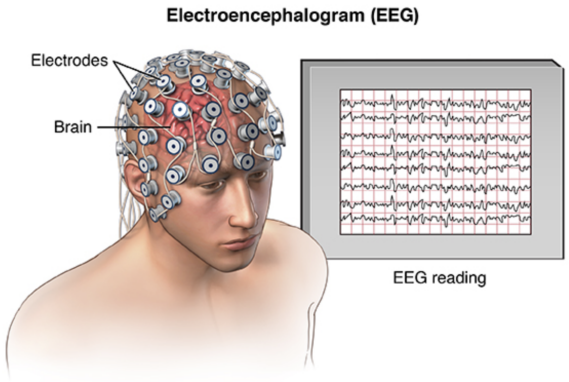
- Electroencephalography (EEG) data.



Source: Internet

Motivating example

- Electroencephalography (EEG) data.
- Continuous and the short time separation between the adjacent measuring points.
- **Goal**: To estimate brain effective connectivity among different regions.
- Strong biological evidence behind presence of feedback loops/cycles.



Source: Internet

Model definition

- Consider a multivariate stochastic process $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ where each Y_j is defined on a compact domain $\mathcal{T}_j \subset \mathbb{R}$.
- Let $Y_j \in \mathcal{H}_j$ where $\mathcal{H}_j$ is a Hilbert space of functions defined on $\mathcal{T}_j$.
- Consider an operator-based non-recursive linear SEM on $\boldsymbol{Y}$ as

$$Y_j(\cdot) = \sum_{\ell \in \mathrm{pa}(j)} (\mathcal{B}_{j\ell} Y_\ell)(\cdot) + f_j(\cdot), \quad \forall j \in [p], \tag{1}$$

- $Y_\ell \to Y_j \implies \mathcal{B}_{j\ell} \neq 0$. Assume $\mathcal{B}_{jj}$ is a null operator (no self loops).
- Suppose if $\boldsymbol{Y}_j \in \mathbb{R}^m$ $\forall j$,, (1) is just,

$$\boldsymbol{Y}_j = \sum_{\ell \in \mathrm{pa}(j)} \boldsymbol{B}_{j\ell} \boldsymbol{Y}_\ell + \boldsymbol{f}_j, \quad \forall j \in [p],$$

- However, model (1) is infinite-dimensional and hence challenging to estimate and interpret.

Causal embedded space

- Assume that the causal relationships are preserved in an unknown low-dimensional subspace $\mathcal{D}_j$ of $\mathcal{H}_j$ of dim K_j.
- $\mathcal{P}_j$ and $\mathcal{Q}_j$ are the projections onto $\mathcal{D}_j$ and its orthogonal complement resp.
- Also assume $\mathcal{B}_{j\ell} = \mathcal{P}_j \mathcal{B}_{j\ell} \mathcal{P}_\ell$.
- As such, (1) can be split into

$$\mathcal{P}_j Y_j = \sum_{\ell \in \mathrm{pa}(j)} \mathcal{B}_{j\ell}(\mathcal{P}_\ell Y_\ell) + \mathcal{P}_j f_j, \tag{2}$$

$$\mathcal{Q}_j Y_j = \mathcal{Q}_j f_j.$$

- We assume that $\mathcal{P}_j f_j$ and $\mathcal{Q}_j f_j$ are independent of each other.

Model definition

- In practice, we do not directly observe Y_j.
- For each Y_j, we observe $\{(t_{ju}, X_{ju})\}_{u=1}^{m_j}$, where $X_{ju} \in \mathbb{R}$ is the measurement of Y_j at location $t_{ju} \in \mathcal{T}_j$.
- Therefore, we consider the following measurement model:

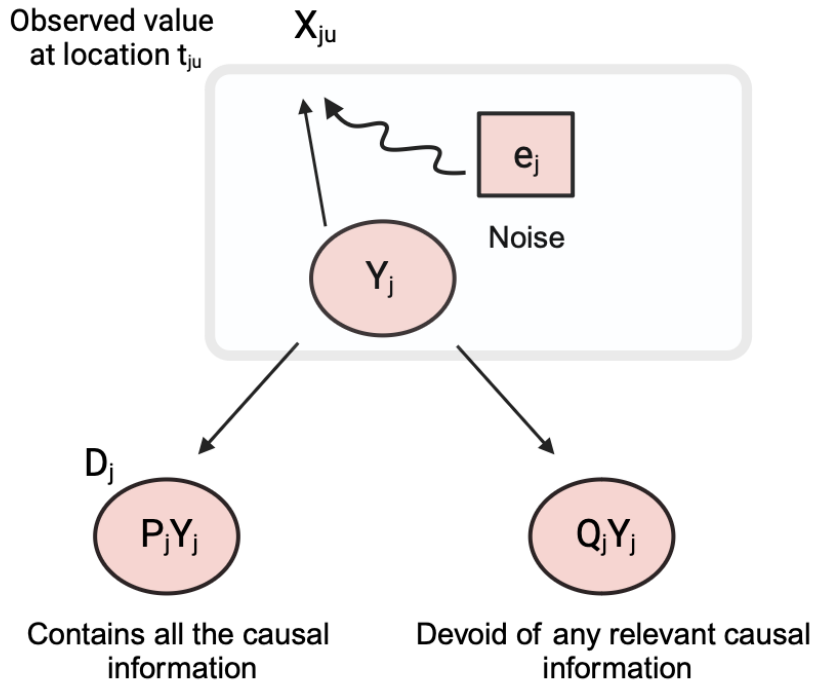
$$\begin{aligned} X_{ju} &= Y_j(t_{ju}) + e_{ju} \\ &= (\mathcal{P}_j Y_j)(t_{ju}) + (\mathcal{Q}_j Y_j)(t_{ju}) + e_{ju}, \quad \forall u \in [m_j], j \in [p], \end{aligned} \tag{3}$$

with independent noises $e_{ju} \sim N(0, \sigma_j), \forall u \in [m_j]$.

Model definition

Recall that,

$$X_{ju} = (\mathcal{P}_j Y_j)(t_{ju}) + (\mathcal{Q}_j Y_j)(t_{ju}) + e_{ju}$$



Model definition

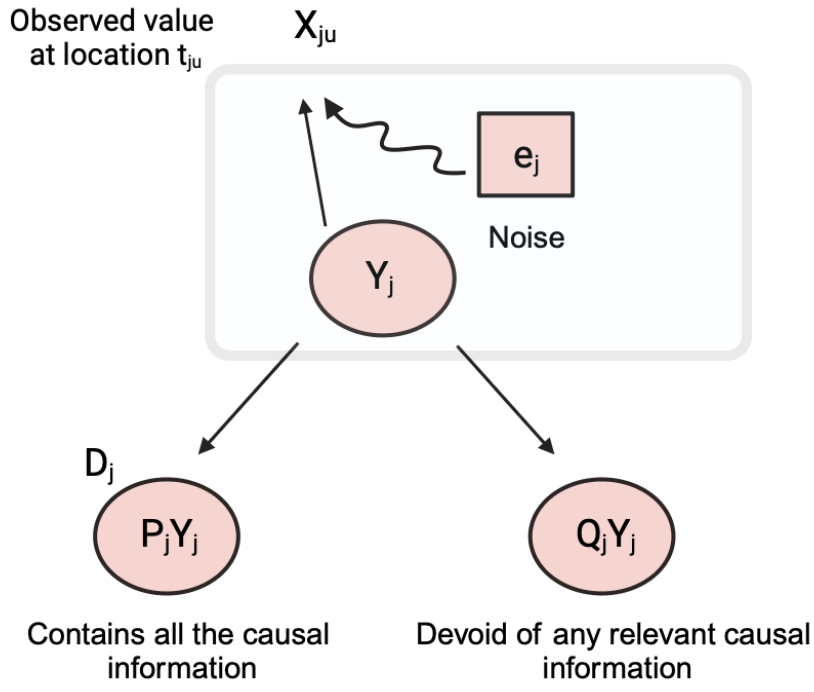
Recall that,

$$X_{ju} = (\mathcal{P}_j Y_j)(t_{ju}) + (\mathcal{Q}_j Y_j)(t_{ju}) + e_{ju}$$

- Define $\alpha_j = \mathcal{P}_j Y_j$, $\beta_j = \mathcal{Q}_j Y_j$ and $\epsilon_j = \mathcal{P}_j f_j, \forall j \in [p]$.
- More compactly,

$$\boldsymbol{X} = \boldsymbol{\alpha}(\boldsymbol{t}) + \boldsymbol{\beta}(\boldsymbol{t}) + \boldsymbol{e} \quad (4)$$

- **Question**: Is the proposed model causally identifiable? Yes, under certain assump.



Assumptions

Assumption 1 (Causal Sufficiency)

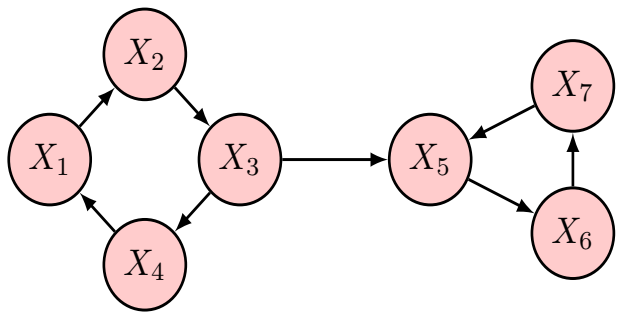
The model $\mathcal{S} = (\mathcal{G}, \mathbb{P})$ is causally sufficient, i.e., there are no unmeasured confounders.

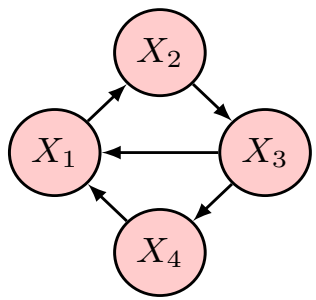
- Keeps the causal discovery task more manageable especially for cyclic graphs with purely observational data.

Assumptions contd...

Assumption 2 (Disjoint Cycles)

The cycles in $\mathcal{G}$ are disjoint, i.e., no two cycles in the graph have two nodes that are common to both.





Assumptions contd...

Assumption 3 (Stability)

For the model $\mathcal{S}$, the moduli of the eigenvalues of the finite rank operator $\mathcal{B}$ are less than or equal to 1, and none of the real eigenvalues are equal to 1.

- Similar kind of assumptions that were discussed for the univariate case have been extended for the random functions.

Assumptions contd...

Assumption 4 (Non-Gaussianity)

The exogenous variables have independent mixture of Gaussian distributions. i.e., $\epsilon_{jk} \overset{\text{ind}}{\sim} \sum_{m=1}^{M_{jk}} \pi_{jkm} \mathrm{N}(\mu_{jkm}, \tau_{jkm})$ with $M_{jk} \geq 2$.

- Can approx. any cont. distribution arbitrarily well [1]
- Useful as it induces model identifiability in the linear SEM framework.[2]

[1] Titterington, Smith, & Makov (1985). Statistical analysis of finite mixture distributions.

[2] Shimizu, Hoyer, Hyvärinen, & Kerminen. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery.

Assumptions contd...

Recall that,

$$X_{ju} = (\mathcal{P}_j Y_j)(t_{ju}) + (\mathcal{Q}_j Y_j)(t_{ju}) + e_{ju}$$

$$\beta_j = \mathcal{Q}_j Y_j$$

Assumption 5 (Non-causal dependency)

Let $\boldsymbol{\beta}(\boldsymbol{t}) = \boldsymbol{C}(\boldsymbol{t})\boldsymbol{\gamma}$, where $\boldsymbol{C}(\boldsymbol{t}) = \text{diag}(\boldsymbol{C}_{11}(\boldsymbol{t}_1), \ldots, \boldsymbol{C}_{pp}(\boldsymbol{t}_p))$ and $\boldsymbol{\gamma}$ be another exogenous component. We assume $\gamma_{jk} \overset{\text{ind}}{\sim} \sum_{m=1}^{M_{jk}} \pi'_{jkm} \text{N}(\mu'_{jkm}, \tau'_{jkm})$ with $M_{jk} \geq 1$.

- $\boldsymbol{\beta}_j(\boldsymbol{t}_j) \perp\!\!\!\perp \boldsymbol{\beta}_\ell(\boldsymbol{t}_\ell)$ for $j \neq \ell$ and $j, \ell \in [p]$.
- $\boldsymbol{C}_{jj}(\boldsymbol{t}_j)$ mixes the independent entries in $\boldsymbol{\gamma}$ to generate temporal dependence within $\boldsymbol{\beta}_j(\boldsymbol{t}_j)$.

Assumptions contd...

Recall, $\boldsymbol{X} = \boldsymbol{\alpha}(\boldsymbol{t}) + \boldsymbol{\beta}(\boldsymbol{t}) + \boldsymbol{e}$. Then if $\alpha_j(t_{ju}) = \sum_{k=1}^{K_j} \tilde{\alpha}_{jk}\phi_{jk}(t_{ju})$, then (4) can be written as,

$$\boldsymbol{X} = \boldsymbol{\Phi}(\boldsymbol{t})\tilde{\boldsymbol{\alpha}} + \boldsymbol{\beta}(\boldsymbol{t}) + \boldsymbol{e},$$

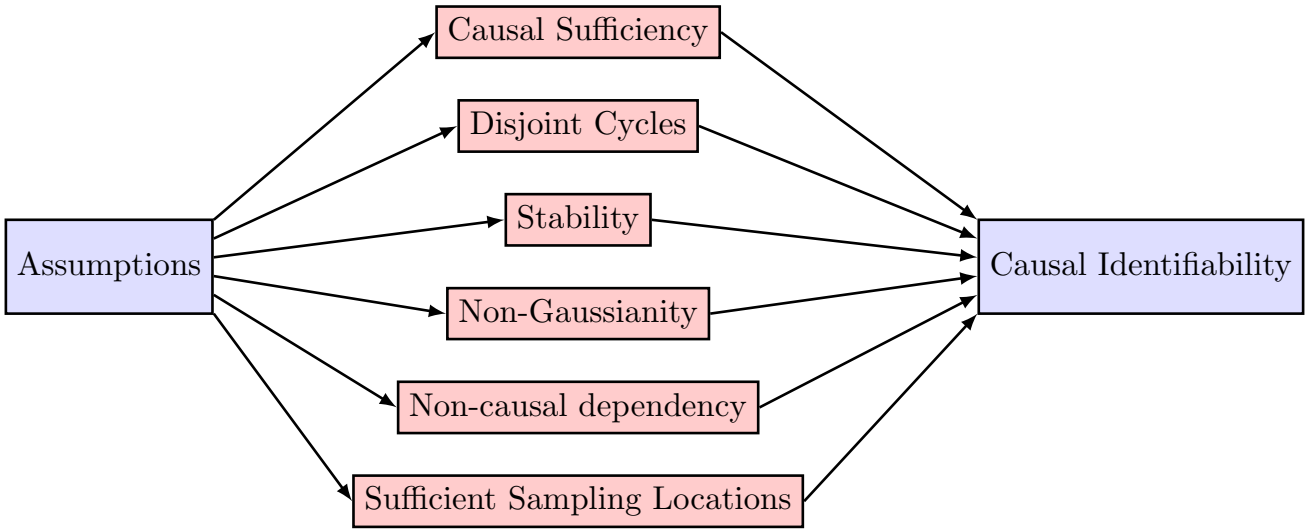
where $\boldsymbol{\Phi}(\boldsymbol{t}) = \text{diag}(\boldsymbol{\Phi}_1(\boldsymbol{t}_1), \ldots, \boldsymbol{\Phi}_p(\boldsymbol{t}_p))$ with $\boldsymbol{\Phi}_j(\boldsymbol{t}_j) = (\phi_{jv}(t_{ju}))_{u,v=1}^{m_j,K_j}$.

Assumption 6 (Sufficient sampling locations)

The basis matrix $\boldsymbol{\Phi}(\boldsymbol{t})$ of size $\sum_{j=1}^{p} m_j \times \sum_{j=1}^{p} K_j$ has a full column rank.

- Implies enough sampling locations over which each random function Y_j is observed.
- Captures all the pertinent causal information that Y_j contains.

Main thorem



Theorem (Causal Identifiability)

Under Assumptions 1-6, $\mathcal{S} = (\mathcal{G}, \mathbb{P})$ is causally identifiable.

Inference - Model parameters

- $\boldsymbol{E} = (E_{j\ell})_{j,\ell=1}^{p}$ denote the adjacency matrix.
- $X_{ju} = Y_j(t_{ju}) + e_{ju}$, $e_{ju} \sim \mathrm{N}(0, \sigma_j)$
- Suppose $Y_j = \sum_{k=1}^{S} \tilde{\alpha}_{jk}\phi_k$.
- We define SEM on first K_j components.

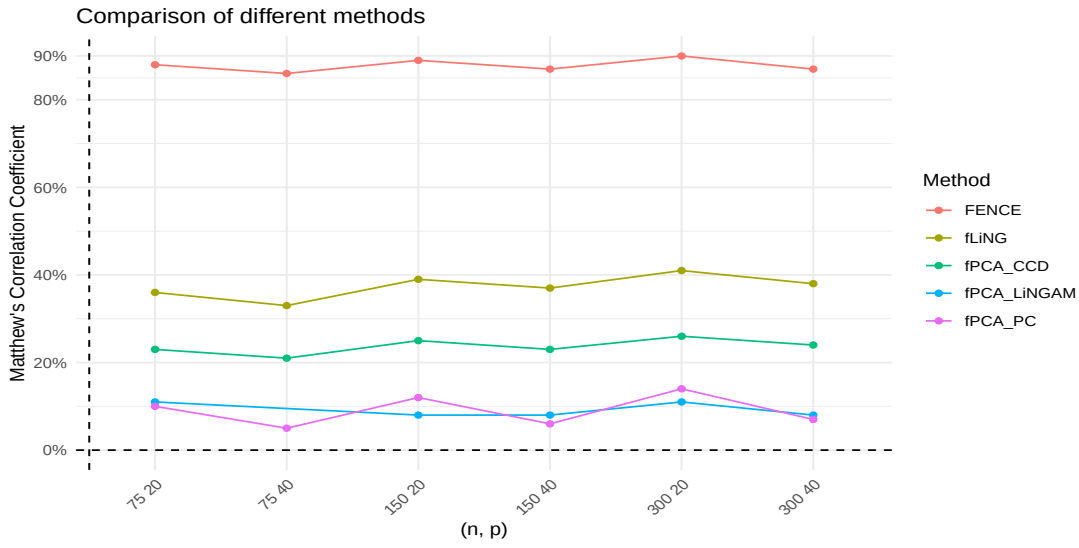
$$\tilde{\boldsymbol{\alpha}} = \boldsymbol{B}\tilde{\boldsymbol{\alpha}} + \tilde{\boldsymbol{\epsilon}}$$

  where $\tilde{\boldsymbol{\alpha}}_j = (\tilde{\alpha}_{j1}, \cdots, \tilde{\alpha}_{jK_j})^\top$ and $\tilde{\boldsymbol{\alpha}} = (\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_p$.
- ϕ_k's are useful for restricting each Y_j to D_j. Do not fix them.
- Further expand them with known cubic b-spline basis functions, $\phi_k(\cdot) = \sum_{r=1}^{R} A_{kr} b_r(\cdot)$.
- $\epsilon_{jk} \overset{\text{ind}}{\sim} \sum_{m=1}^{M_{jk}} \pi_{jkm} \mathrm{N}(\mu_{jkm}, \tau_{jkm})$.
- Some very std. priors are taken to formulate the Bayesian procedure.

Simulation study setup

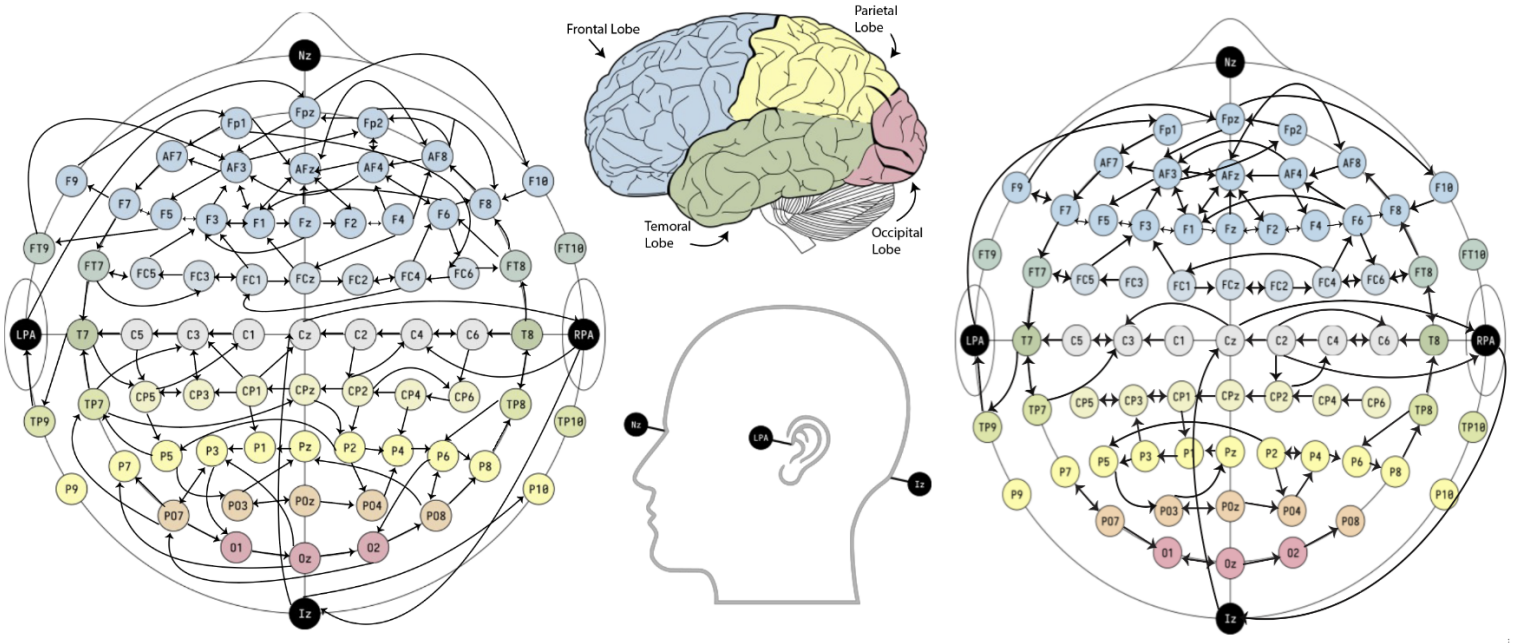
- Sample size $(n) = \{75, 150, 300\}$, number of nodes $(p) = \{20, 40\}$, time grid size $(d) = 125$.
- The causal graph G is estimated by thresholding the posterior probability of inclusion at 0.5.
- Methods compared against:
  1. Functional Bayesian Network (fLiNG).
  2. fPCA-CCD
  3. fPCA-PC
  4. fPCA-LiNGAM
- 2, 3 and 4 are two step procedures:
  – First step involves obtaining the basis coefficients by carrying out fPCA.
  – Second step involves estimating the causal graphs using existing causal discovery methods.

Results



- We used Matthew's correlation coefficient (MCC) to assess the graph recovery performance.

Application - EEG Data



- For both groups (alcoholic and control), brain regions that are spatially closer to each other tend to be more connected.
- Dense connectivity is observed in the frontal region of the brain in both groups, with multiple cycles being formed.
- Alcoholic group has more connectivity across the left parietal and occipital lobes.

Conclusions

- Proposed an operator-based non-recursive linear SEM based framework for functional data in the presence of cycles.
- Introduced the assumption of existence of a lower dimensional causal embedded space that captures all the causal information.
- Proved causal identifiability of our model under certain assumptions.
- Showed applications over an EEG dataset.

Future directions:

- Relaxing the assumption for causal sufficiency.
- Consider non linear SEM framework.

Reference

Roy, S., Wong, R. K. W., & Ni, Y. (2023). Directed Cyclic Graph for Causal Discovery from Multivariate Functional Data. Accepted in NeurIPS, 2023.

Thank you!

Backup slides

Why stability assumption?

- We assume our data are drawn from an equilibrium distribution of a dynamic system involving multivariate functions.
- $\mathbf{Y}(\cdot)[t] = (Y_1(\cdot)[t], \ldots, Y_p(\cdot)[t])^\top$ denote a vector of functions at time point t where the domain of each function $Y_j(\cdot)[t]$ is not necessarily time.
- we have used () to denote the function input/domain and [] to denote the time index of a dynamic system.
- Consider an AR1-type dynamic system of those functions,

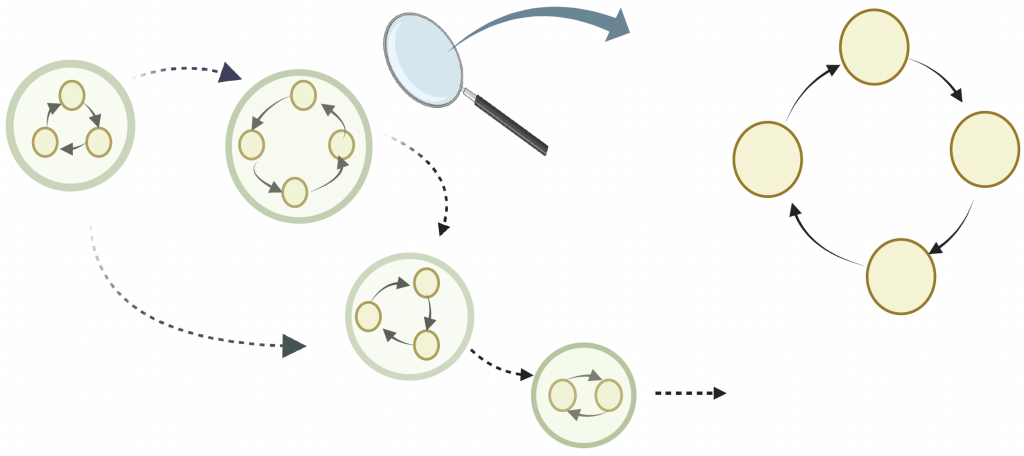
$$\mathbf{Y}(\cdot)[t] = \mathfrak{B}\mathbf{Y}(\cdot)[t-1] + \mathbf{f}(\cdot)$$

- Recursively leads to,

$$\mathbf{Y}(\cdot)[t] = \mathfrak{B}^t\mathbf{Y}(\cdot)[0] + \sum_{s=0}^{t-1} \mathfrak{B}^s\mathbf{f}(\cdot)$$

- $\mathfrak{B}^t$ and $\sum_{s=0}^{t-1} \mathfrak{B}^s$ converge as t approaches infinity.

Proof – A brief sketch pictorially



- In LHS, a hypergraph-like structure emerges when we assume the existence of disjoint cycles.
- While the true graph contains cycles, this hypergraph-like structure is essentially a DAG, thus easier to work with.
- LHS to RHS.

Prior specifications

Prior on spline coefficients $\boldsymbol{A}_k$

$$\boldsymbol{A}_k|\lambda_k \sim \mathrm{N}(\mathbf{0}, \lambda_k^{-1}\boldsymbol{\Omega}^{-})$$

where $\boldsymbol{\Omega}^{-}$ is the pseudoinverse of $\boldsymbol{\Omega} = \int \boldsymbol{b}''(t)[\boldsymbol{b}''(t)]^\top \, dt$. We constrain the regularization parameters $\lambda_1 > \cdots > \lambda_S > 0$ by putting a uniform prior:

$$\begin{aligned}
&\lambda_k \sim \mathrm{Uniform}(L_k, U_k), \ \forall\ k \in [S], \\
&U_1 = 10^8, L_k = \lambda_{k+1} \ \forall\ k \in [S-1], \\
&U_k = \lambda_{k-1} \ \forall\ k \in \{2, \ldots, S\}, L_S = 10^{-8},
\end{aligned}$$

which implies that the smoothness of $\phi_k(\cdot)$ decreases as k gets larger.

Prior specifications

Prior on adjacency matrix $\boldsymbol{E}$

$$E_{j\ell}|\rho \overset{\text{ind}}{\sim} \text{Bernoulli}(\rho)$$
$$\rho \sim \text{Uniform}(0,1)$$

- The marginal distribution of $\boldsymbol{E}$ with ρ integrated out is $\text{Beta}\left(\sum_{j\neq\ell} E_{j\ell} + 1, \sum_{j\neq\ell}(1 - E_{j\ell}) + 1\right)$.
- If $\boldsymbol{E}_0$ denotes the null adj. matrix and $\boldsymbol{E}_1$ denotes the adj. matrix with only one edge, then we can see that $p(\boldsymbol{E}_0)/p(\boldsymbol{E}_1) = p^2 - p$.
- prevents false discoveries and leads to a sparse network by increasing the penalty against additional edges as the dimension p grows.

Prior specifications

Prior on the causal effect matrix $\boldsymbol{B}$

$$\boldsymbol{B}_{j\ell}|E_{j\ell} \sim (1-E_{j\ell})MVN(\boldsymbol{B}_{j\ell};\mathbf{0}, s\gamma\boldsymbol{I}_{K_j}, \boldsymbol{I}_{K_\ell}) + E_{j\ell}MVN(\boldsymbol{B}_{j\ell};\mathbf{0}, \gamma\boldsymbol{I}_{K_j}, \boldsymbol{I}_{K_\ell})$$
$$\gamma \sim \mathrm{IG}(a_\gamma, b_\gamma)$$

with $s = 0.02, a_\gamma = b_\gamma = 1$.

- continuous spike and slab prior.
- When $E_{j\ell} = 0$, $\boldsymbol{B}_{j\ell}$ is negligibly small.

Prior on the noise variances σ_j

$$\sigma_j \sim \mathrm{IG}(a_\sigma, b_\sigma)$$

with $a_\sigma = b_\sigma = 0.01$.

Prior specifications

Prior on the mixture distribution parameters

$$(\pi_{jk1}, \ldots, \pi_{jkM_{jk}}) \sim \text{Dirichlet}(\beta, \ldots, \beta), \qquad \forall\ j \in [p], k \in [S]$$
$$\mu_{jkm} \sim N(a_\mu, b_\mu),\ \tau_{jkm} \sim \text{IG}(a_\tau, b_\tau),\ \forall\ j \in [p], k \in [S], m \in [M_{jk}]$$

We have fixed values for the hyperparameters, $\beta = 1$,
$a_\mu = 0, b_\mu = 100$, $a_\tau = b_\tau = 1$.

Choice of K_j

- Possible to use a prior to learn the number of basis functions jointly with other parameters.
- The functional observations are imputed and arranged into a $(n \times p) \times d$ matrix, where $d = |\cup_{i,j} \mathcal{T}_j^{(i)}|$ represents the size of the union of the measurement grid over all realized random functions.
- Singular value decomposition is performed, and the minimum value of K is selected such that its proportion of variance explained is at least 90%.
- We set $K_j = K$. Note that although K is fixed, the basis functions are adaptively inferred.
- Fix a grid encompassing values $\{1, 2, 3, 4, 5, 6, 7\}$ for K and subsequently selecting the K associated with the lowest WAIC.
- The graph recovery performance remains almost the same.