# Non-negative DAG Learning from Time-Series Data

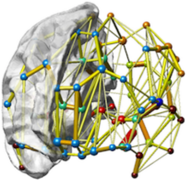
Samuel Rey† and **Gonzalo Mateos**⋆

†King Juan Carlos University - *University of Rochester

2025 Asilomar Conference on Signals, Systems, and Computers
Pacific Grove, USA – October 26-29, 2025

# Introduction

- Contemporary data exhibit temporal dynamics and irregular structure
  - $\Rightarrow$ Graphical models help explain and learn from such data [Kolaczyk09]
  - $\Rightarrow$ Graph topology is often unknown or unavailable



Brain network

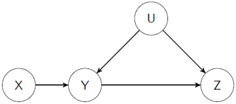
Social network


River flow data

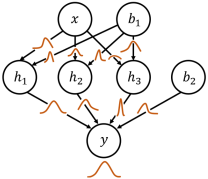
- Graph learning aims to infer the graph structure from nodal observations
  - $\Rightarrow$ **As.**: signal properties and temporal variations depend on the graph
- **This work:** learning directed acyclic graphs (DAGs) from time series

# Why DAGs?

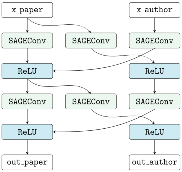
- DAGs have become prominent models in various ML applications
  - ⇒ Conditional independencies among variables in Bayesian networks
  - ⇒ DAG edges may have causal interpretations [Peters17]
  - ⇒ Applications: biology [Sachs05], genetics [Zhang13], finance [Sanford12]




Causal inference




Bayesian networks

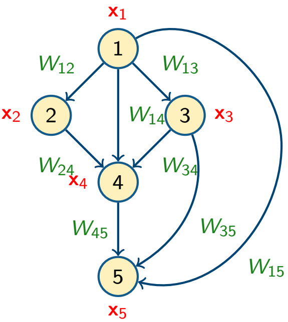



Neural networks

- Learning DAGs from observational data comes with **challenges**
  - ⇒ Imposing acyclicity is a combinatorial constraint
  - ⇒ Multiple DAGs may generate the same data distribution

# DAGs and linear SEM

- DAG $\mathcal{D}(\mathcal{V},\mathcal{E},\mathcal{W}) \in \mathbb{D}$ with $|\mathcal{V}| = N$ nodes
  - $\Rightarrow$ Adjacency matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$
  - $\Rightarrow$ Entry $W_{ij} \neq 0$ indicates a directed link $i \to j$
- Random vector $\mathbf{x} = [x_1, \ldots, x_N] \in \mathbb{R}^N$
  - $\Rightarrow$ Joint $p(\mathbf{x})$ Markov w.r.t. $\mathcal{D} \in \mathbb{D}$
  - $\Rightarrow$ $\mathcal{D}$ encodes conditional independence on $\mathbf{x}$
  - $\Rightarrow$ $x_i$ depends on parents $PA_i = \{j \in \mathcal{V} : W_{ij} \neq 0\}$




- Linear structural equation model (SEM) to generate $\mathbf{X} \in \mathbb{R}^{N\times T}$ consists of

$$\mathbf{X} = \mathbf{W}^\top \mathbf{X} + \mathbf{Z}$$

$\Rightarrow$ Exogenous input $\mathbf{Z}$ with diagonal covariance matrix

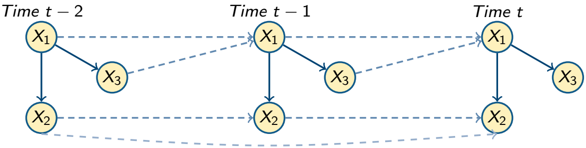
# Structural vector autoregressive model (SVARM)

- SVARM for time series $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T] \in \mathbb{R}^{N \times T}$ [Demiralp03]

$$\mathbf{x}_t = \mathbf{W}^\top \mathbf{x}_t + \sum_{p=1}^{P} \mathbf{A}_p^\top \mathbf{x}_{t-p} + \mathbf{z}_t \quad \Longrightarrow \quad \mathbf{X} = \mathbf{W}^\top \mathbf{X} + \sum_{p=1}^{P} \mathbf{A}_p^\top \mathbf{Y}_p + \mathbf{Z}$$

$\Rightarrow$ DAG $\mathbf{W}$ and $\{\mathbf{A}_p\}$ capture instantaneous and lagged dependencies

$\Rightarrow$ Matrices $\mathbf{Y}_p$ collect time-lagged versions of $\mathbf{X}$




- SVARM in matrix form as $\mathbf{X} = \mathbf{W}^\top \mathbf{X} + \mathbf{A}^\top \mathbf{Y} + \mathbf{Z}$

$\Rightarrow$ With $\mathbf{A} = [\mathbf{A}_1^\top, \ldots, \mathbf{A}_P^\top]^\top$ and $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_P]^\top$

$\Rightarrow$ SEM is a particular case when $\mathbf{A} \equiv \mathbf{0}$

# Problem statement

- Given matrices $\mathbf{X} \in \mathbb{R}^{N \times T}$ and $\mathbf{Y} \in \mathbb{R}^{NP \times T}$ adhering to a SVARM, learn matrices $\mathbf{W}$ and $\mathbf{A}$ solving a score-minimization problem

$$\min_{\mathbf{W},\mathbf{A}} \ \mathcal{S}(\mathbf{W}, \mathbf{A}; \mathbf{X}, \mathbf{Y}) \ \text{subject to} \ \mathcal{D}(\mathbf{W}) \in \mathbb{D}$$

- Learning a DAG solely from observational data is NP-hard [Chickering96]
  - $\Rightarrow$ Combinatorial acyclicity constraint $\mathcal{D} \in \mathbb{D}$ difficult to enforce
  - $\Rightarrow$ Model may not be identifiable from samples of $p(\mathbf{x})$ alone
- To address this challenging scenario we
  - $\Rightarrow$ Discuss key advances in learning **W** from SEM (iid data)
  - $\Rightarrow$ Propose a solution accounting for the **time-series structure**

# Context and contributions

**Context**

- Methods based on order search [Charpentier22,Deng23]
  - ⇒ Recovering the causal ordering is challenging with limited data
- Methods based on continuous acyclicity functions [Zheng18,Bello22]
  - ⇒ From combinatorial search to non-convex continuous optimization
  - ⇒ Focus on **iid data** modeled via SEMs
- Methods leveraging acyclicity constraint from [Zheng18] in time-series data
  - ⇒ New score functions for SVARMs [Pamfil20,Misiakos25]
  - ⇒ Resulting optimization problems are **non-convex**

**Contribution**

- Learning DAG structure based on a **convex acyclicity function**
  - ⇒ Key simplifying assumption of non-negative weights

# Non-convex acyclicity constraint

- In the iid case ($P = 0$) the score $\mathcal{S}$ only involves $\mathbf{W}$ and $\mathbf{X}$
- Characterize acyclicity via a **smooth function** $h(\mathbf{W}) : \mathbb{R}^{N \times N} \mapsto \mathbb{R}$ [Zheng18]

  $\Rightarrow$ The zero-level set corresponds to DAGs: $h(\mathbf{W}) = 0 \iff \mathcal{D}(\mathbf{W}) \in \mathbb{D}$

- From combinatorial search to non-convex continuous optimization

$$\min_{\mathbf{W}} \mathcal{S}(\mathbf{W}; \mathbf{X}) \text{ s. to } \mathcal{D}(\mathbf{W}) \in \mathbb{D} \quad \iff \quad \min_{\mathbf{W}} \mathcal{S}(\mathbf{W}; \mathbf{X}) \text{ s. to } h(\mathbf{W}) = 0$$

- Continuous acyclicity functions include:

  $\Rightarrow$ NOTEARS: $h_{\text{notears}}(\mathbf{W}) = \operatorname{Tr}\left(e^{\mathbf{W} \circ \mathbf{W}}\right) - N$ [Zheng18]

  $\Rightarrow$ DAGMA: $h^{s}_{\text{dagma}}(\mathbf{W}) = N \log(s) - \log\det(s\mathbf{I} - \mathbf{W} \circ \mathbf{W})$ [Bello22]

- **Observation**: Product $\mathbf{W} \circ \mathbf{W}$ renders the acyclicity functions non-convex

# Learning non-negative DAGs from a SEM

- Learning DAGs with non-negative weights by solving [Rey25]

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \frac{1}{2T}\|\mathbf{X} - \mathbf{W}^\top\mathbf{X}\|_F^2 + \lambda \sum_{i,j=1}^{N} W_{ij} \;\; \text{s. to} \;\; \mathbf{W} \geq 0, \;\; h(\mathbf{W}) = 0$$

$\Rightarrow$ Least squares score function with sparsity regularization

**Proposition**

For any $\mathbf{W} \in \mathbb{R}_+^{N\times N}$ with spectral radius $\rho(\mathbf{W}) < s$, $\mathcal{D}(\mathbf{W}) \in \mathbb{D}$ iff

$$h_{ldet}(\mathbf{W}) := N\log(s) - \log\det(s\mathbf{I} - \mathbf{W}) = 0$$

- Convex acyclicity $h_{ldet}(\mathbf{W})$ leads to an **abstract convex optimization**
  - $\Rightarrow$ Enables finding the global minimum due to additional structure
  - $\Rightarrow$ Guaranteed recovery of the true DAG in the infinite sample regime

# Learning non-negative DAGs from a SVARM

- When data follow a SVARM with $P > 0$ we estimate $\mathbf{W}$ and $\mathbf{A}$ solving

$$\min_{\mathbf{W},\mathbf{A}} \frac{1}{2t}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X} - \mathbf{A}^\top \mathbf{Y}\|_F^2 + \lambda_{\mathbf{W}} \sum_{i,j=1} W_{ij} + \lambda_{\mathbf{A}} \sum_{i,j=1} A_{ij}$$

$$\text{s. to } \mathbf{W} \geq 0, \quad \mathbf{A} \geq 0, \quad h_{ldet}(\mathbf{W}) = 0$$

$\Rightarrow$ Least squares term accounts for time-lagged dependencies

$\Rightarrow$ **W** and **A** are assumed to be non-negative

$\Rightarrow$ Only **W** is required to be a DAG

**Key properties**

- Additional structure leads to an **abstract convex optimization**
- Convexity enables finding the global minimizer

# DAG learning algorithm

- Learn the DAG and time-lagged dependencies via method of multipliers
  - $\Rightarrow$ Iterative method with well-known **convergence guarantees**
- Denote the augmented Lagrangian as

$$L_c(\mathbf{W}, \mathbf{A}, \alpha) = F(\mathbf{W}, \mathbf{A}) + \alpha h(\mathbf{W}) + \frac{c}{2} h(\mathbf{W})^2$$

  - $\Rightarrow$ With Lagrange multiplier $\alpha$ and penalty parameter $c$
- Sequentially performs the following steps until convergence

**Step 1:** Estimate $\mathbf{W}^{(k+1)}$ and $\mathbf{A}^{(k+1)}$ solving

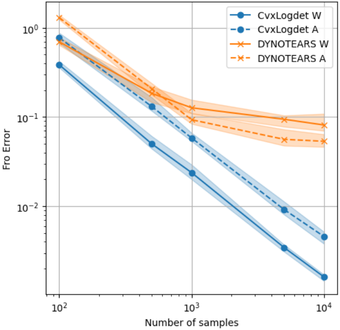
$$\{\mathbf{W}^{(k+1)}, \mathbf{A}^{(k+1)}\} = \arg \min_{\mathbf{W} \geq 0, \mathbf{A} \geq 0} L_{c^{(k)}}(\mathbf{W}, \mathbf{A}, \alpha^{(k)})$$

**Step 2:** Update Lagrange multiplier as $\alpha^{(k+1)} = \alpha^{(k)} + c^{(k)} h(\mathbf{W}^{(k+1)})$

**Step 3:** Update the penalty parameter $c^{(k+1)}$

# Numerical evaluation (I)

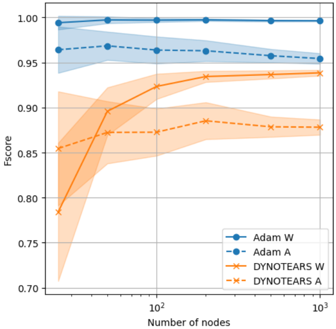
- Non-negative ER graphs with $d = 50$ nodes and average degree 4
  - $\Rightarrow$ Signals sampled from SVARM with $P = 2$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$




- Convex acyclicity constraints outperform alternatives
  - $\Rightarrow$ The proposed method achieves better error both for $\hat{\mathbf{W}}$ and $\hat{\mathbf{A}}$
  - $\Rightarrow$ Error of the convex method goes to 0 as number of samples grows

# Numerical evaluation (II)

- Non-negative ER graphs with a time series of length $T = 5000$
  - $\Rightarrow$ Represent the F-score as the number of nodes increases




- F-score of estimated $\hat{\mathbf{W}}$ consistently close to 1 with the proposed method
  - $\Rightarrow$ Illustrates how convexity helps in recovering the true DAG structure

# Concluding remarks

University of Rochester

- We address the problem of learning the DAG structure from time-series data
  - ⇒ Imposing acyclicity is challenging
  - ⇒ Combine non-convex cont. acyclicity functions with SVARM
- Leveraging the **non-negative weights** assumption allows us to
  - ⇒ Employ a convex acyclicity function
  - ⇒ Recover the global optimum using the method of multipliers
  - ⇒ Provide intuition about the recoverability of the true DAG
- Outperform state-of-the-art alternatives in synthetic experiments

# Questions?


Questions at: **samuel.rey.escudero@urjc.es**