

# Projeto de Pesquisa

Unsupervised Learning of Graph Structures: Inference and Model Selection for High-Dimensional Stochastic Processes

Área: 1.a. Aprendizagem Estatística e Ciência de Dados  
Área correlata: 1.c. Estatística Computacional

**Magno T. F. Severino**

# Agenda

- Introduction
- Theoretical Background
- Research Proposal
- Viability

# Motivation

Modern dependence analysis, including multivariate time series, sensor networks, neurodata, and large-scale, involves:

- high dimensionality,
- temporal dependence,
- continuous data,
- complex structures.

Probabilistic graphical models (MRFs) are a natural tool for representing such dependencies.

**Challenge:** existing methods mostly address:

- discrete data
- finite number of vertices,
- independence or weak dependence structures.

# Background and Definitions

# Vector Processes and the Underlying Graph

Consider the vector-valued process:

$$\mathbf{X}^{(i)} = (X_1^{(i)}, \dots, X_d^{(i)}), \quad i \in \mathbb{N}, \quad X_i \in A.$$

Assume:

- stationarity,
- existence of an invariant distribution  $\pi$ , probability space  $((A^d)^{\mathbb{N}}, \mathcal{F}, \mathbb{P})$ ,
- conditional dependencies described by a graph  $G^* = (V, E^*)$ .

This graph encodes:

$$X_v \perp X_{V \setminus (\{v\} \cup G^*(v))} \mid X_{G^*(v)}.$$

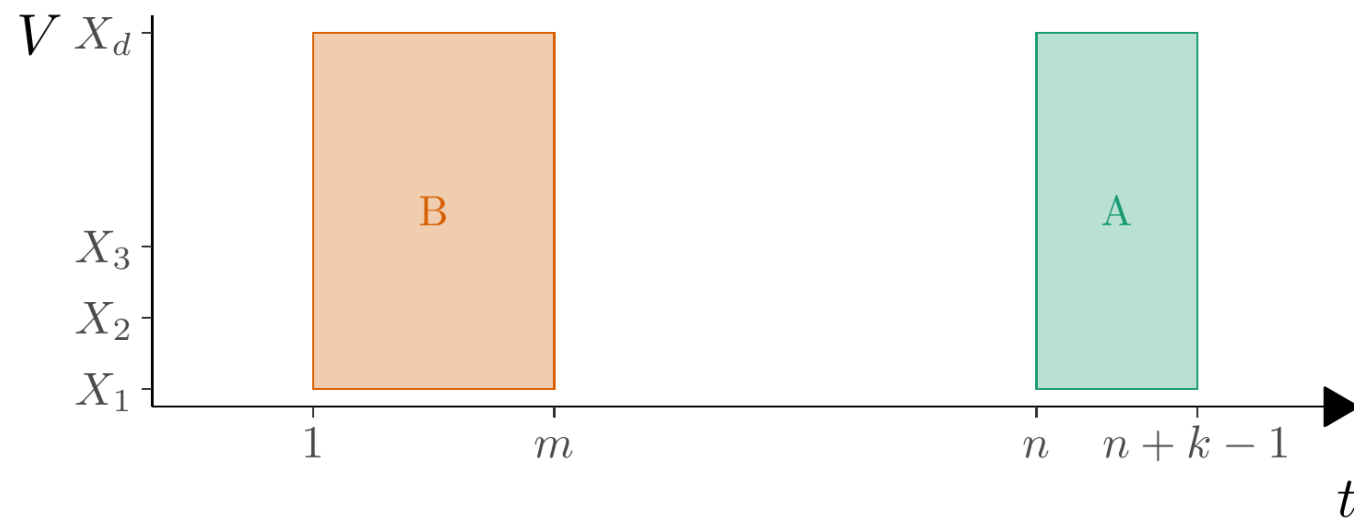
$$G^*(v) = \{u \in V : (u, v) \in E^*\}$$

# Mixing Condition

- $X^{(i:j)}$  denote the sequence of vectors  $X^{(i)}, X^{(i+1)}, \dots, X^{(j)}$ .
- $\mathbf{X} = \{\mathbf{X}^{(\mathbf{i})} : -\infty < \mathbf{i} < \infty\}$  satisfies a *mixing condition* with rate  $\{\psi(\ell)\}_{\ell \in \mathbb{R}}$  if

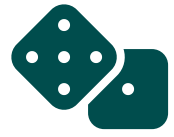
$$\left| \mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)} \mid X^{(1:m)} = x^{(1:m)}\right) - \mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)}\right) \right| \leq \psi(n-m) \mathbb{P}\left(X^{(n:(n+k-1))} = x^{(1:k)}\right),$$

for  $n \geq m + \ell$  and for each  $k, m \in \mathbb{N}$  and each  $x^{(1:k)} \in (A^d)^k, x^{(1:m)} \in (A^d)^m$  with  $\mathbb{P}(X^{(1:m)} = x^{(1:m)}) > 0$ .



$$|P(A|B) - P(A)|$$

$$\leq \psi(n-m)P(A)$$



# Empirical Probabilities

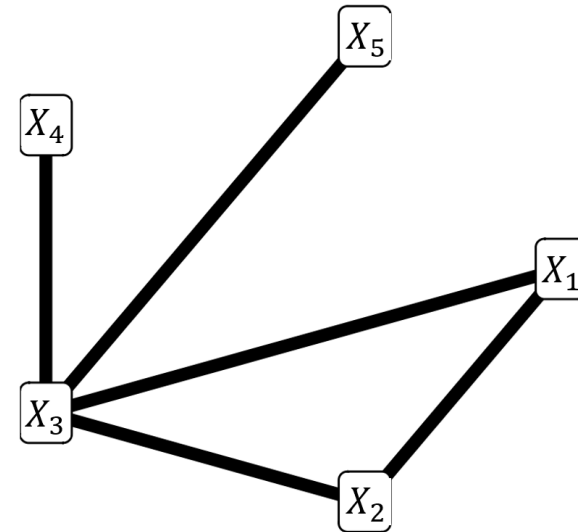
Given a sample of size  $n$  and graph  $G = (V, E)$  and  $v \in V$ , define

$$G(v) = \{u \in V : (u, v) \in E\},$$

the set of neighbors of  $v$  in graph  $G$ .

$$\hat{\pi}(X_1 = a_1) = \frac{N(X_1 = a_1)}{n}.$$

For  $v = X_1$ , we have  $G(v) = \{X_2, X_3\}$ .



# Penalized Pseudo-Likelihood

For a candidate graph  $G$ :

$$L(G) = \prod_{i=1}^n \prod_{v \in V} \pi(x_v^{(i)} | x_{G(v)}^{(i)}),$$

We estimate  $\pi$  empirically:

$$\hat{L}(G) = \prod_{i=1}^n \prod_{v \in V} \hat{\pi}(x_v^{(i)} | x_{G(v)}^{(i)}),$$

Selection criterion (Severino & Leonardi, 2025):  $\hat{G} = \arg \max$

# Consistency Theorem

**Theorem (Severino & Leonardi, 2025):**

Let  $\{X^{(i)}: i \in \mathbb{N}\}$  be a stationary process that satisfies the mixing condition presented before with rate  $\psi(\ell) = O(1/\ell^{1+\epsilon})$  for some  $\epsilon > 0$ .

Then, by taking  $\lambda_n = c \log n$ , for  $c > 0$ , we have that 
$$\widehat{G} = \underset{G}{\arg\max} \Big\{ \log \widehat{L}(G) - \lambda_n \sum_{v \in V} |A|^{(|G(v)|)} \Big\}$$
 satisfies  $\widehat{G} = G^*$  eventually almost surely as  $n \rightarrow \infty$ .

# Recent Work



## Model selection for Markov random fields on graphs under a mixing condition

*Stochastic Processes and their Applications*, 2025.

Severino, M. T. F., & Leonardi, F.

### Advances:

- global criterion based on penalized pseudo-likelihood;
- consistency theorem under a mixing condition;
- applications to discrete multivariate processes.

### Limitations:

- restriction to the **finite** vertex case,
- **discrete** variables.

This project aims to overcome both limitations in **two proposals**.

# Proposal 1

# Model Selection for MRFs with Countably Infinite Vertex Sets under Mixing Condition

## Motivation

- Massive networks (social, biological, IoT),
- Structures where  $|V| = \infty$  and grows with the sample,
- Finite-vertex methods do not generalize automatically.

## Existing methods

- **Leonardi et al. (2023)**: Penalized pseudo-likelihood for discrete MRFs. Graph estimated based on local neighborhood estimation.
- **Severino & Leonardi (2025)**: Developed theoretical results for global estimation of discrete MRFs over finite graphs.

# Model Selection for MRFs with Countably Infinite Vertex Sets under Mixing Condition

## Research goal

- Generalize the results from finite to countably infinite graphs.
- Improve global estimation, possibly reducing errors from local neighborhood estimation.

## Proposed estimation framework

- Let  $V$  be infinite and  $\{V_n\}, \{n \in \mathbb{N}\}$  be a sequence of finite subsets of  $V$ .
- Assume  $V_n \uparrow V$  as  $n \rightarrow \infty$ .
- Sample:  $\{\mathbf{X} = \{X_v: v \in V_n\}\}$ , assuming that  $\{\mathrm{ne}(v)\}$  is finite.
- Adaptation of key theorems to handle countably infinite vertex sets.

# Proposal 1 – Expected Advances

## Algorithms

- implementation in R or Python,
- simulations on large synthetic networks.

## Applications

- social networks,
- neuroscience (large neural connectivity),
- sensor systems.

# Proposal 2

# Model Selection for Continuous MRFs under Mixing

## Current limitations

- classical pseudo-likelihood is defined for finite alphabets,
- discretization leads to information loss.

## Objective

- develop a consistent estimator without discretization.

## Challenges

- replacing summations with integrals,
- defining neighborhood structure in conditional densities,
- adapting consistency proofs to the continuous setting.

# Proposal 2 – Expected Advances

## Benefits

- higher inferential precision,
- no discretization required,
- applicability to finance, hydrology, neuroscience, and bioinformatics.

## Expected results

- consistency theorems analogous to the discrete case,
- scalable algorithm,
- R or Python package.

# Final Integration

Medium-term goal:  $\text{Continuous MRFs}$  +  $\text{Infinite Vertex Sets}$  +  $\text{Mixing}$ .

Deliverables:

- unified theoretical framework,
- consistent algorithms for genuinely large-scale systems,
- general framework for complex real-world data.

# Viability

# Viability

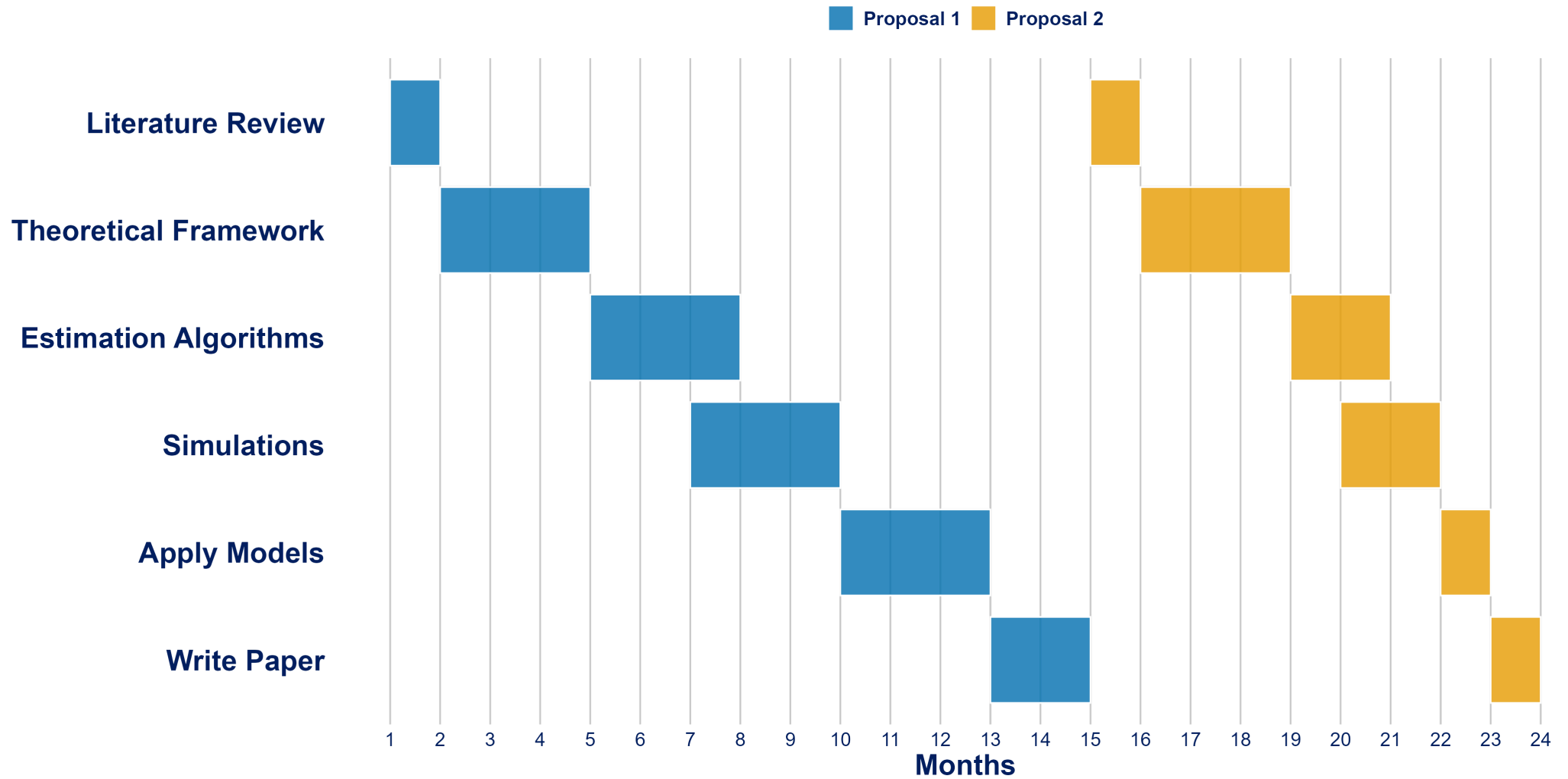
## Resources:

- consolidated background in MRFs, mixing processes, and computational expertise,
- access to computational infrastructure at IME-USP (low budget project),
- a collaborative research environment (Neuromat, UFRJ, UFRN, UBA).

## Expected output:

- two international journal articles,
- two R or Python packages,
- presentations at scientific conferences.

# Timeline



# References

- **Severino, M. T. F., & Leonardi, F. (2025).** *Model selection for Markov random fields on graphs under a mixing condition.* Stochastic Processes and their Applications.
- Leonardi, F., Lopez-Rosenfeld, M., Rodriguez, D., **Severino, M. T. F. S., & Sued, M. (2021).** *Independent block identification in multivariate time series.* Journal of Time Series Analysis.
- Leonardi, F., Carvalho, R., & Frondana, I. (2023). *Structure recovery for partially observed discrete Markov random fields on graphs under not necessarily positive distributions.* Scandinavian Journal of Statistics.
- Lauritzen, S. L. (1996). *Graphical models.* Claredon Press.
- Oodaira, H., & Yoshihara, K. I. (1971). *The law of the iterated logarithm for stationary processes satisfying mixing conditions.* Kodai Mathematical Seminar Reports.

# Obrigado

# Rate of convergence of the empirical probabilities

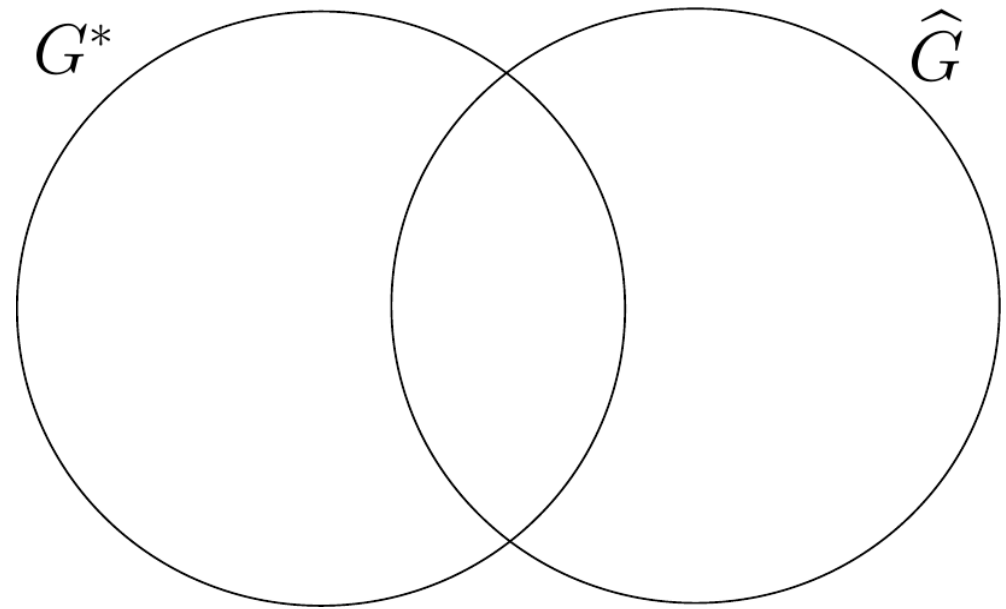
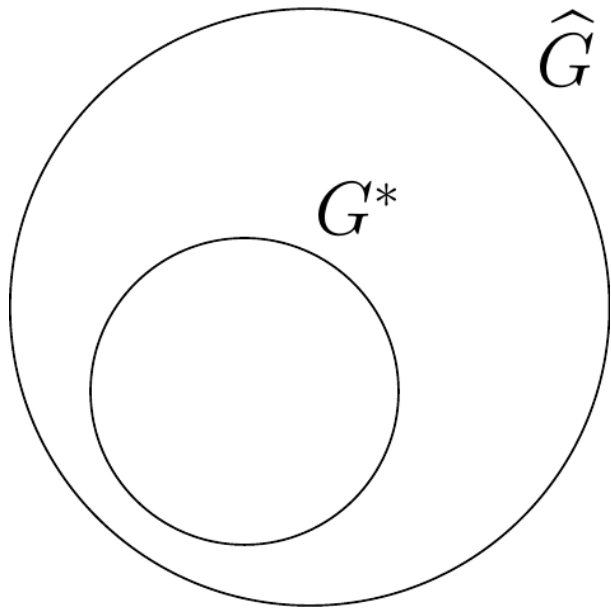
Based on the Law of the Iterated Logarithm for stationary polynomial mixing processes proved in Oodaira and Yoshihara (1971), we can derive the rate of convergence of the empirical probabilities to the true probabilities of the process.

**Proposition 1 (Typicality):** Assume the process  $\{X^{(i)} : i \in \mathbb{N}\}$  satisfies the mixing condition with rate  $\psi(\ell) = O(1/\ell^{1+\epsilon})$ , for some  $\epsilon > 0$ . Then, for any  $W \subset V$  and  $\delta > 0$ , 
$$\left| \widehat{\pi}(a_W) - \pi(a_W) \right| < \sqrt{\frac{\delta \log n}{n}}$$
 eventually almost surely as  $n \rightarrow \infty$ .

**Proposition 2 (Conditional typicality):** Then for any  $\delta > 0$ , any disjoint sets  $W, W' \subset V$  and any  $a_W \in A^W$  and  $a_{W'} \in A^{W'}$  we have that 
$$\left| \widehat{\pi}(a_W | a_{W'}) - \pi(a_W | a_{W'}) \right| < \sqrt{\frac{\delta \log n}{N(a_W)}}$$
 eventually almost surely as  $n \rightarrow \infty$ .

# Intuitive Overview of Theorem Proof

Consider  $\{\widehat{G} \neq G^*\} = \big\{G^* \subseteq \widehat{G}\big\} \cup \big\{G^* \not\subseteq \widehat{G}\big\}$ .



We prove that, eventually almost surely as  $n \rightarrow \infty$ , neither of the cases above can happen, which implies that  $\widehat{G} = G^*$ .

# Algorithms for Estimation

Define

$$\begin{equation*}\label{eq:defH} H(G) = \log \widehat{L}(G) - \lambda_n \sum_{v \in V} |A|^{G(v)}. \end{equation*}$$

**Focus:** determine the maximal value of  $H(\cdot)$  and identify the argument at which this maximum occurs.

# Exact Algorithm

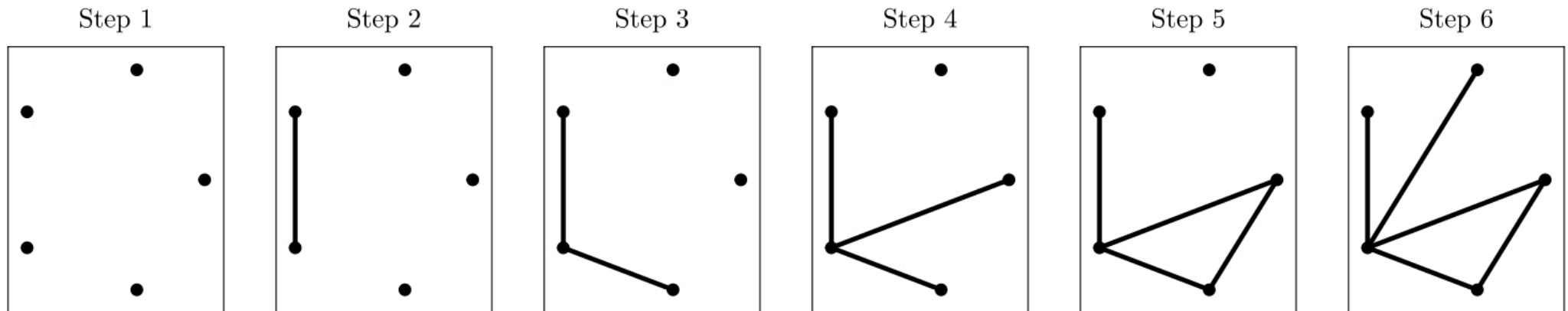
- Let  $H(G) = \log \widehat{L}(G) - \lambda_n \sum_{v \in V} |A|^{|G(v)|}$ .
- Define  $\mathcal{G} = \{ G = (V, E) : E \subseteq V \times V \setminus \{(v,v) : v \in V\} \}$ .
- Set  $\widehat{G} = \arg\max_{G \in \mathcal{G}} H(G)$ .
- Drawback: computational complexity

$$|\mathcal{G}| = 2^{\frac{d(d-1)}{2}},$$

$$|V|=d.$$

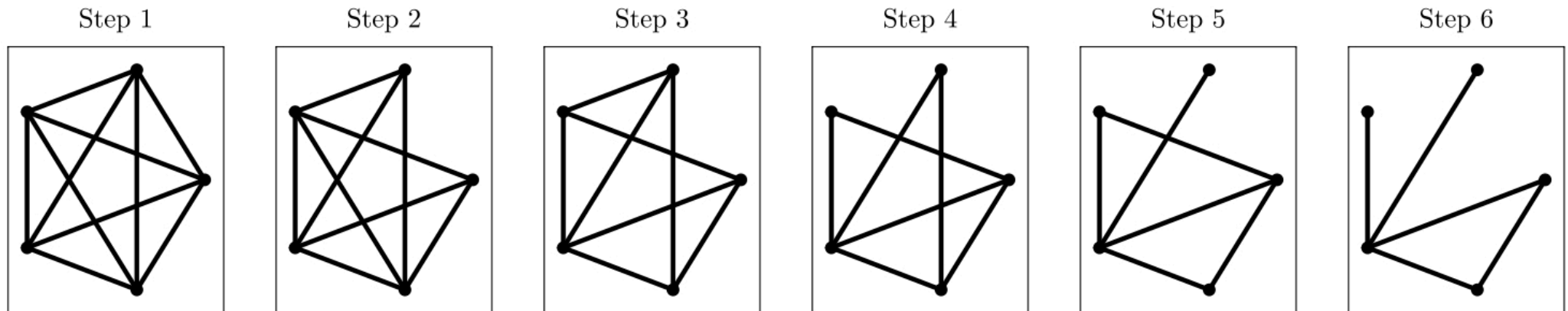
# Forward Stepwise Algorithm

- Starts with an empty graph.
- Adds edges one at a time, selecting the edge that maximizes improvement in fit.
- Stops when no further enhancement in fit is achieved.



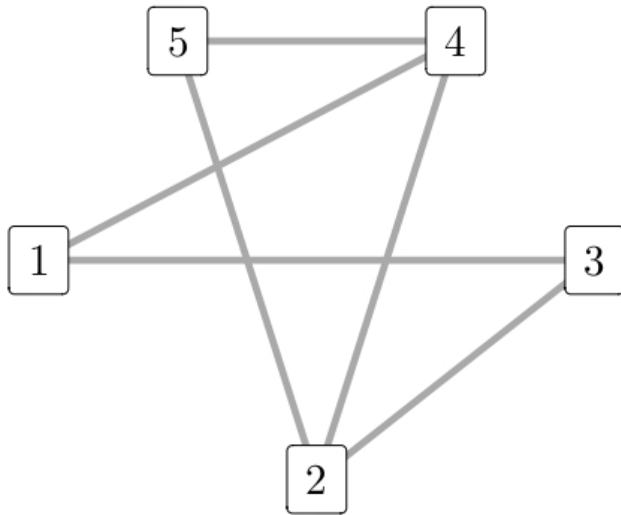
# Backward Stepwise Algorithm

- Begins with a complete graph.
- Removes edges one at a time, selecting the edge that maximizes improvement in fit.
- Stops when no further enhancement in fit is achieved.

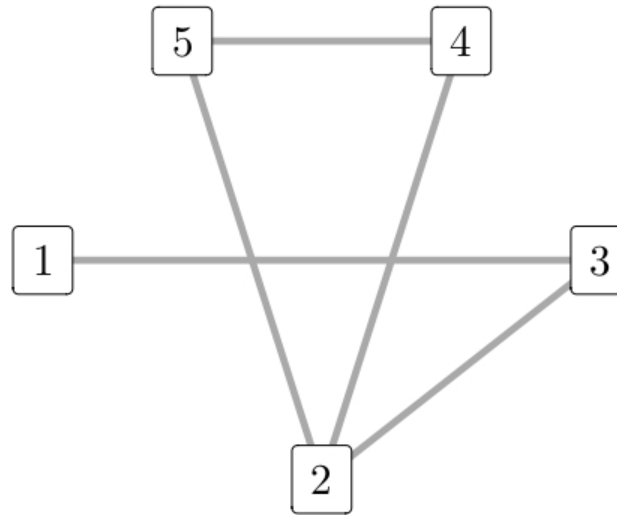


# Simulated Annealing Algorithm

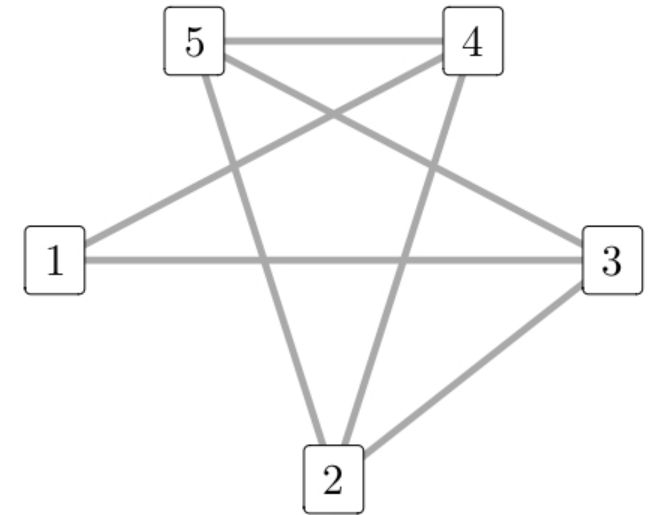
- $H(G) = \log \widehat{L}(G) - \lambda_n \sum_{v \in V} |A|^{|G(v)|}$ .
- Definition of a neighbor of a graph.



\quad\quad\quad\quad\quad G\_1



\quad\quad\quad\quad\quad G\_2



\quad\quad\quad\quad\quad G\_3

# Simulation Study

- **Objective:** Assess the performance of the proposed algorithms via simulation studies.
- **Scenario 1: Fixed True Graph**
  - Generate synthetic data based on fixed true graph.
  - Assess estimator's performance under stability conditions.
- **Scenario 2: Random Graphs with Varying Edge Number**
  - Generate random graphs with varying edge numbers.
  - Evaluate estimator's performance across graphs with different edge configuration.

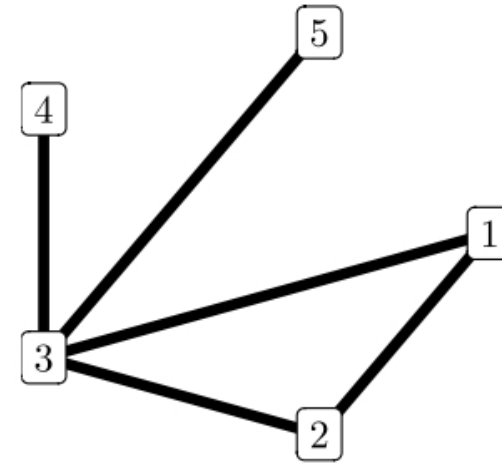
# Example 11 - Scenario 1

- Consider  $\mathbf{X} = (X_1, \dots, X_5)$ .
- Joint probability function of these variables:  $p(x_1, x_2, x_3, x_4, x_5) = p(x_3) p(x_1|x_3) p(x_2|x_1, x_3) p(x_4|x_3) p(x_5|x_3)$ .
- Conditional probabilities:

```

\begin{align*}
p(x_1|x_2, x_3, x_4, x_5) &= \\
p(x_1|x_2, x_3), \quad & \\
p(x_2|x_1, x_3, x_4, x_5) &= \\
p(x_2|x_1, x_3), \quad & \\
p(x_3|x_1, x_2, x_4, x_5) &= \\
p(x_1, x_2, x_3, x_4, x_5) / \sum_{\{x_3\}} p(x_1, x_2, x_3, x_4, x_5), \quad & \\
p(x_4|x_1, x_2, x_3, x_5) &= \\
p(x_4|x_3), \quad & \\
p(x_5|x_1, x_2, x_3, x_4) &= \\
p(x_5|x_3). \quad &
\end{align*}

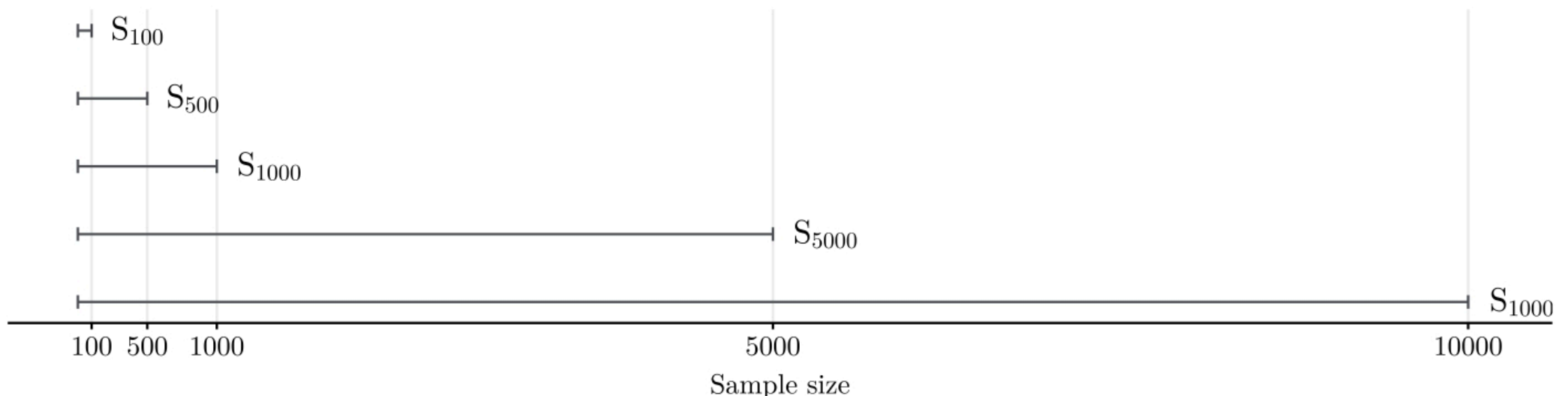
```



- **Aim:** assess the performance of the algorithms for estimation.
- **Data generation:** Gibbs Sampler algorithm.

# Example 11 - Sampling Scheme

- The Gibbs sampler (Geman and Geman, 1984 and Gelfand and Smith, 1990):
  - iterations: 15{,}000,
  - burn-in period: 5{,}000,
  - final sample: 10{,}000.
- Smaller samples were extracted from the initial sample  $s_1, \dots, s_{10{,}000}$ , with sizes  $N \in \{100, 500, 1{,}000, 5{,}000, 10{,}000\}$ .



# Example 11 - Exact Algorithm

# Example 11 - Forward Stepwise

# Example 11 - Backward Stepwise

# Example 11 - Simulated Annealing

# Example 11 - Several Replications

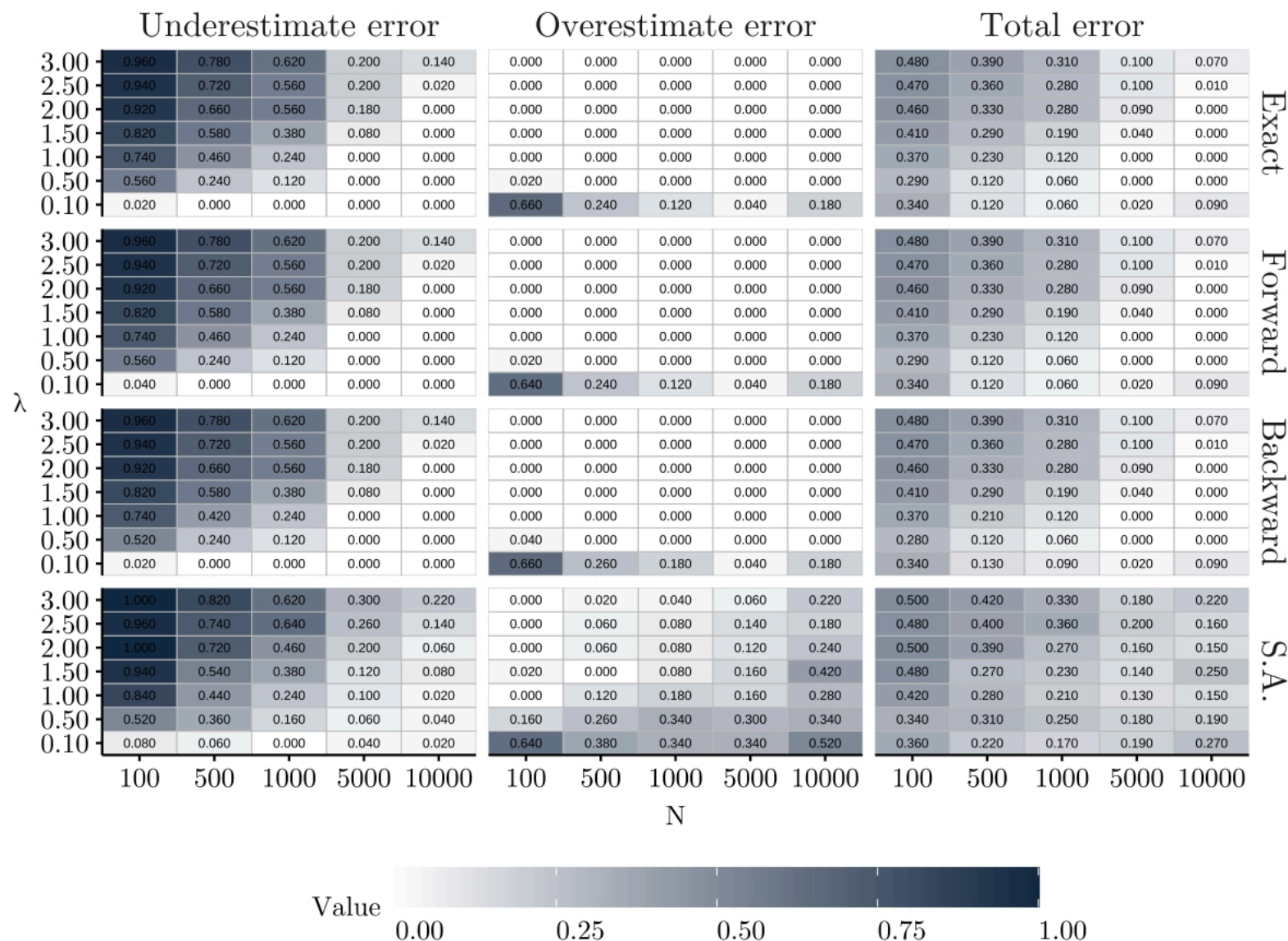
- **Objective:** generate several samples and assess the performance of the algorithms.
- **Metrics:** underestimation error (ue), overestimation error (oe), and total error (te).

$$ue(G, \hat{G}) = \frac{\sum_{(v, w)} \mathbf{1}_{\{(v, w) \in E \text{ and } (v, w) \notin \hat{E}\}}}{\sum_{(v, w)} \mathbf{1}_{\{(v, w) \in E\}}},$$

$$oe(G, \hat{G}) = \frac{\sum_{(v, w)} \mathbf{1}_{\{(v, w) \notin E \text{ and } (v, w) \in \hat{E}\}}}{\sum_{(v, w)} \mathbf{1}_{\{(v, w) \notin E\}}},$$

$$\quad \quad \quad te(G, \hat{G}) = \frac{ue \sum_{(v, w)} \mathbf{1}_{\{(v, w) \notin E\}} + oe \sum_{(v, w)} \mathbf{1}_{\{(v, w) \in E\}}}{|V|(|V|-1)/2}.$$

# Example 11 - Several Replications



# The Choice of Penalizing Constant $c$

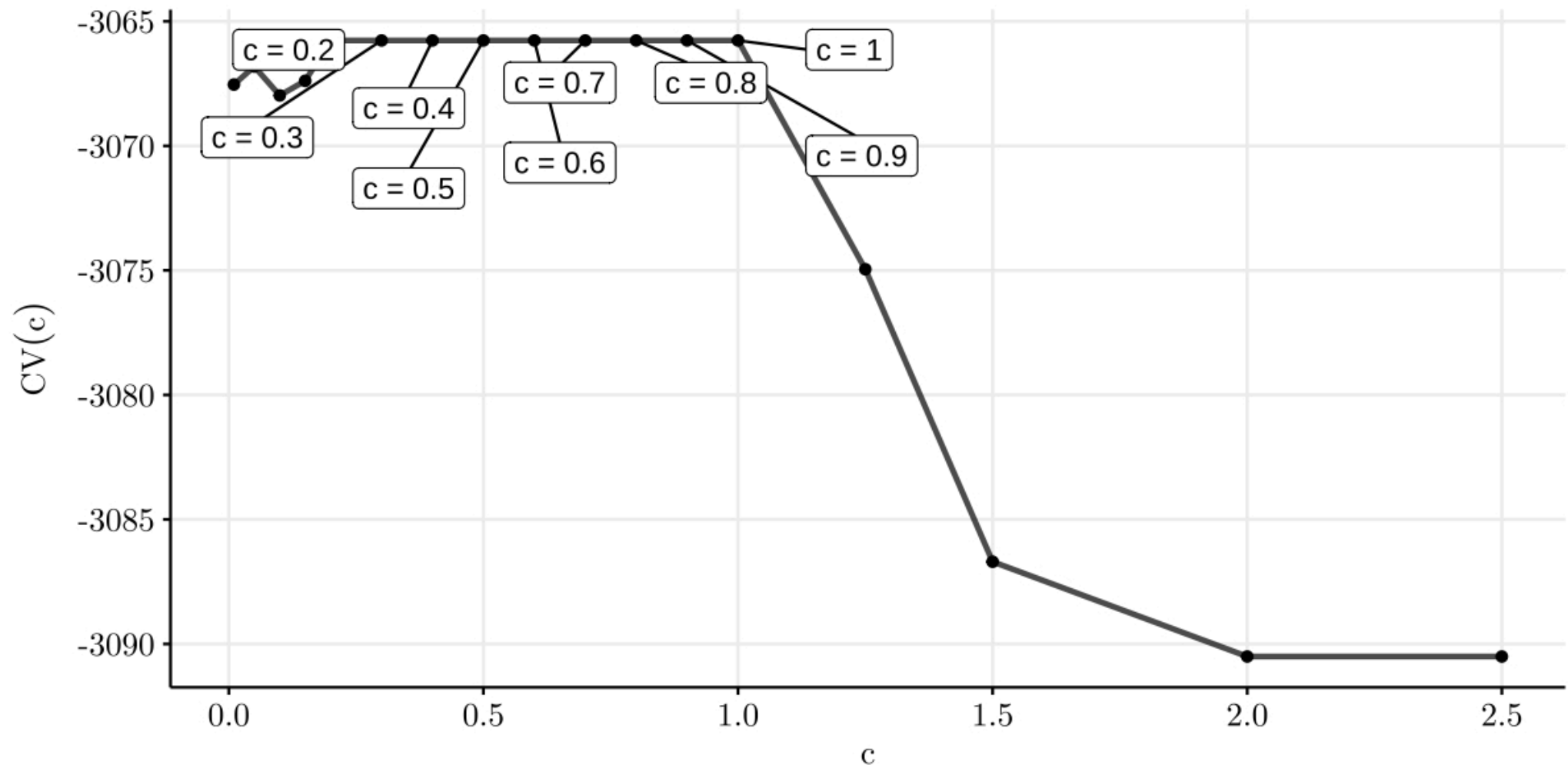
Utilize  $k$ -fold cross-validation to assess model performance and select optimal penalizing constant values.

Iteration 1	Val.	Train	Train	Train	Train
Iteration 2	Train	Val.	Train	Train	Train
Iteration 3	Train	Train	Val.	Train	Train
Iteration 4	Train	Train	Train	Val.	Train
Iteration 5	Train	Train	Train	Train	Val.

Compute pseudo-log-likelihood over validation sets for different constant values.

$$\begin{aligned} \mathrm{CV}_k^{(i)}(c) = & \sum_{v \in V} \sum_{(a_v \in A)} \\ & \sum_{a_{\hat{G}(v)} \in A^{(|\hat{G}(v)|)}} N_i(a_v, a_{\hat{G}(v)}) \log \\ & \widehat{\pi}(a_v | a_{\hat{G}(v)}), \end{aligned}$$

# The Choice of Penalizing Constant $c$

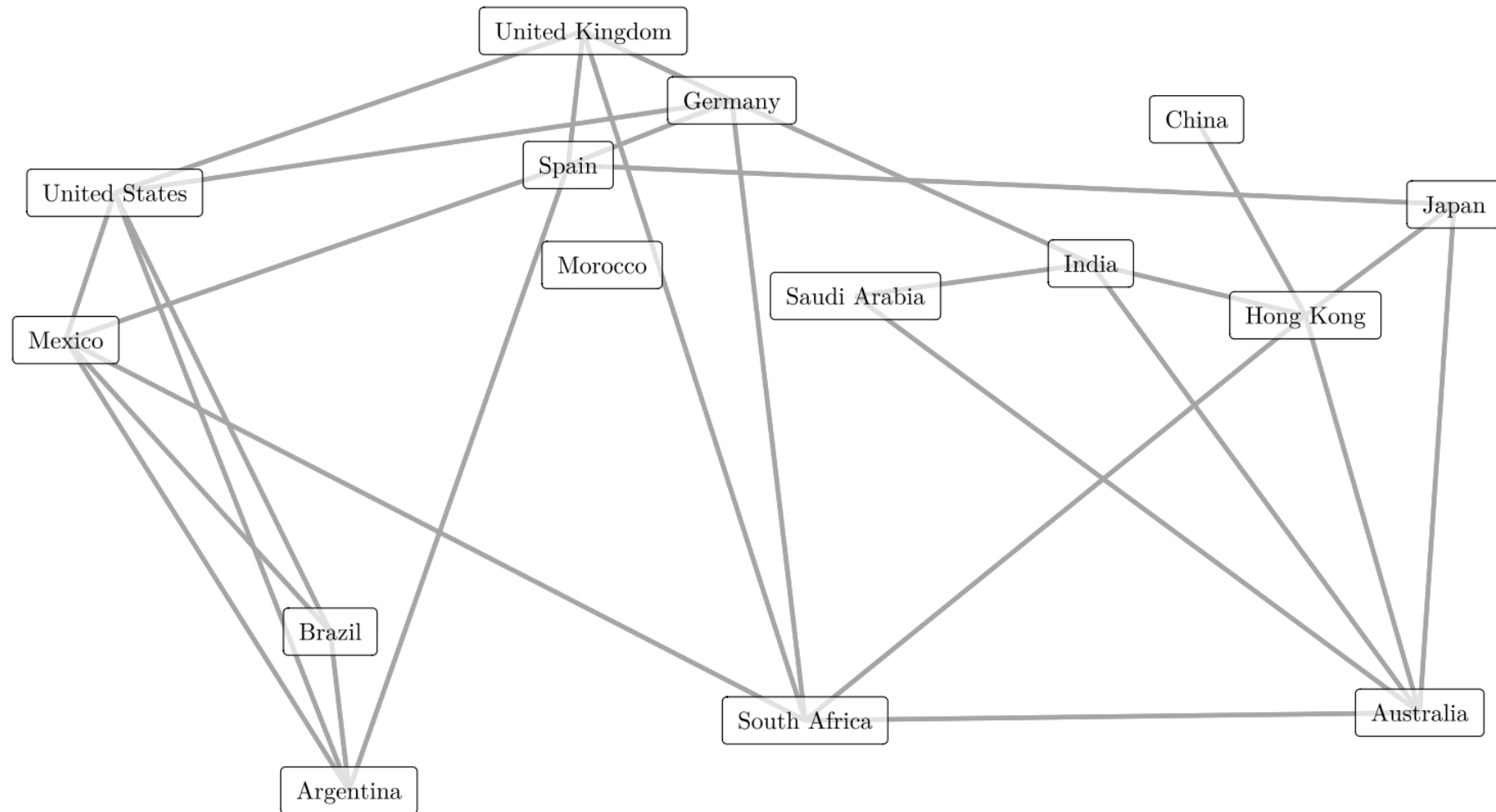


5-fold cross-validation error for Example 11, considering a sample of size 5,000 and set of values for penalizing constant.

# São Francisco River Data

- Water volume measured at  $d$  stations along the river's course, denoted as  $X_{\{u\}}$ , where  $u = 1, \dots, d$ .
- Vector  $\mathbf{X}$  observed at discrete time intervals (10-day mean), from January 1977 to December 2016.
- Process  $\mathbf{X}^n = \{\mathbf{X}^{(i)}: 1 \leq i \leq n\}$ ,  $\mathbf{X}^{(i)} = (X_{\{1\}}^{(i)}, \dots, X_{\{d\}}^{(i)})$ .
- Forward stepwise algorithm and a 5-fold cross-validation approach below.

# Stock Exchange Data



Estimated graph, considering the penalizing constant value chosen by cross-validation.