Magno Severino

Projeto de Pesquisa

Inference and Model Selection for Continuous and Infinite Markov Random Fields on Graphs

Área: 1.a. Inferência para Processos Estocásticos.

Área correlata: 3.a. Aprendizagem Estatística e Ciência de Dados.

Agenda

- Introduction
- Theorethical Background
- Research Proposal
- Viability

Introduction

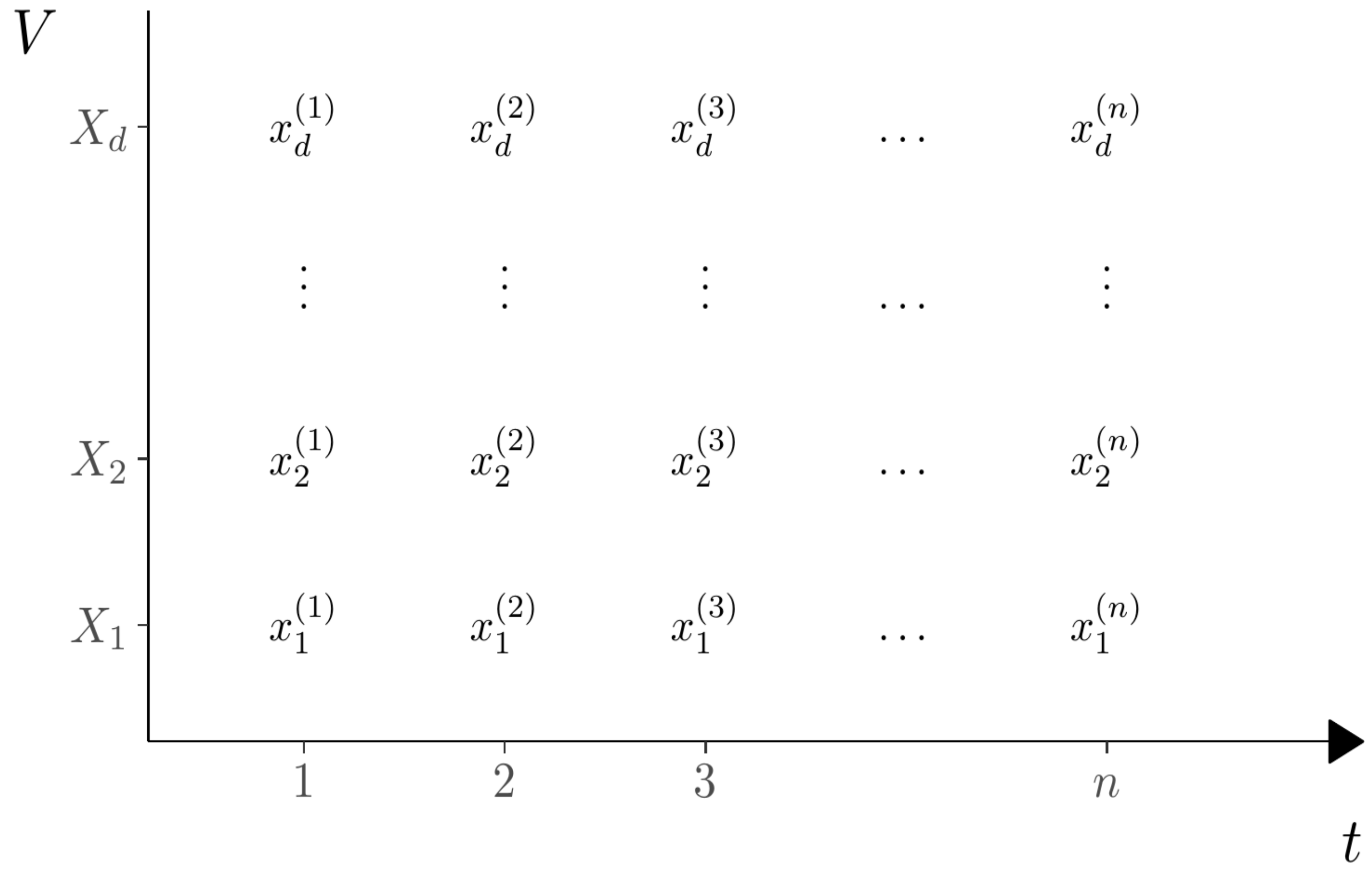
- Previous research (Severino, 2025):
  - Focus on **discrete** vector-valued stochastic processes,
  - Model selection criteria for **graphs** under **mixing conditions**,
  - Graphs with **fixed** number of nodes.

- Research proposals:
  - **Proposal 1**: Extending models to **countably infinite vertex sets**,
  - **Proposal 2**: Adapting the methodology for **continuous data**.

Background and Definitions

Vector-Valued Stochastic Processes

- $\mathbf{X^{(i)}} = (\mathbf{X_1^{(i)}}, \mathbf{X_2^{(i)}}, \dots, \mathbf{X_d^{(i)}})$.
- $X_v^{(i)} \in A$, A a finite alphabet.
- Process $\mathbf{X} = \{\mathbf{X^{(i)}} : -\infty < \mathbf{i} < \infty\}$.
- We assume the process $\mathbf{X}$ has an underlying graph $G^* = (V, E^*)$.

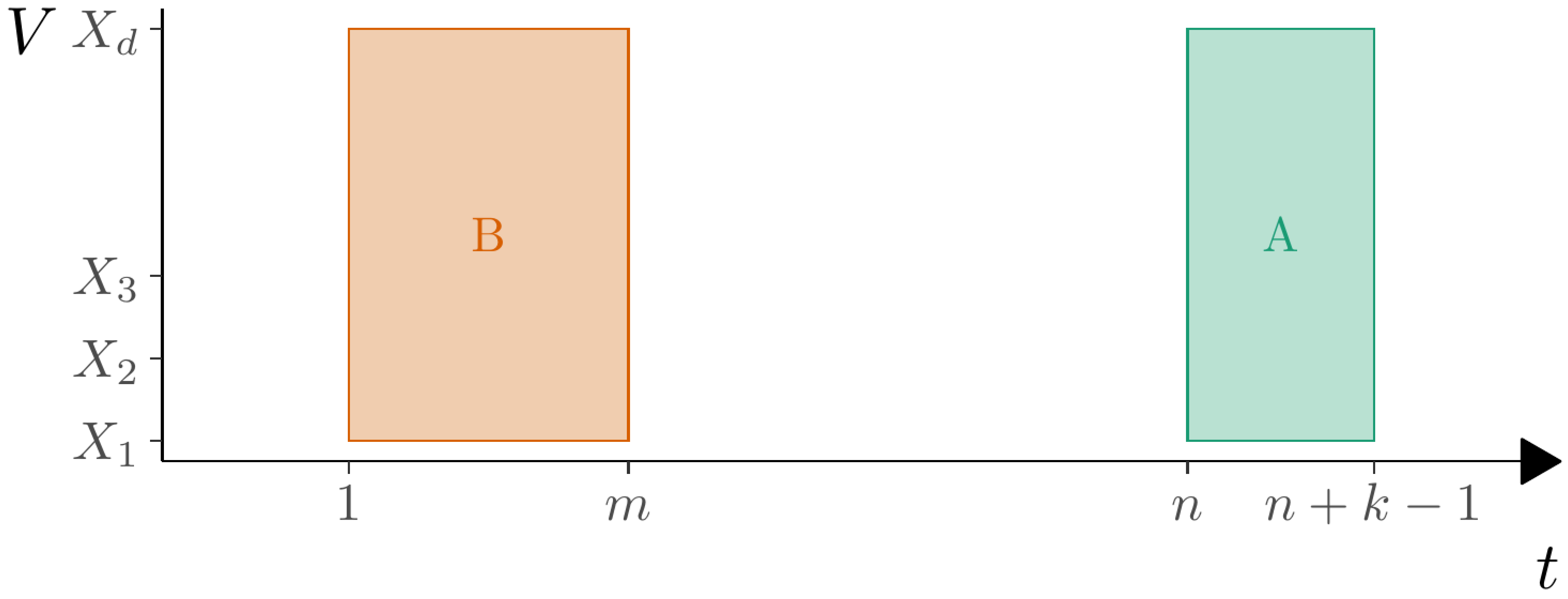


Mixing Condition

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

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

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 $\mathbb{P}(X^{(1:m)} = x^{(1:m)}) > 0$.



Empirical Probabilities

Given a process $\mathbf{X} \in \{a, b\}^3$ and the sample of size 5 below. Then

X_1	X_2	X_3
b	a	a
a	b	b
b	b	a
a	a	a
b	a	a

$$\widehat{\pi}\big(\{X_1 = a\}\big) = \frac{2}{5}, \qquad \widehat{\pi}\big(\{X_1 = a, X_3 = a\}\big) = \frac{1}{5},$$

$$\widehat{\pi}\big(\{X_1 = b\}\big|\{X_2 = a, X_3 = a\}\big) = \frac{2}{3}.$$

as $\hat{\pi}\big(\{X_2 = a, X_3 = a\}\big) > 0.$

Formally, assume we observe a sample of size n of the process, denoted by $\{x^{(i)} : i = 1, \ldots, n\}$. Then, for any $W \subset V$ and any $a_W \in A^W$,

$$\widehat{\pi}(a_W) = \frac{N(a_W)}{n}; \qquad \widehat{\pi}(a_{W'}|a_W) = \frac{\widehat{\pi}(a_{W' \cup W})}{\widehat{\pi}(a_W)},$$

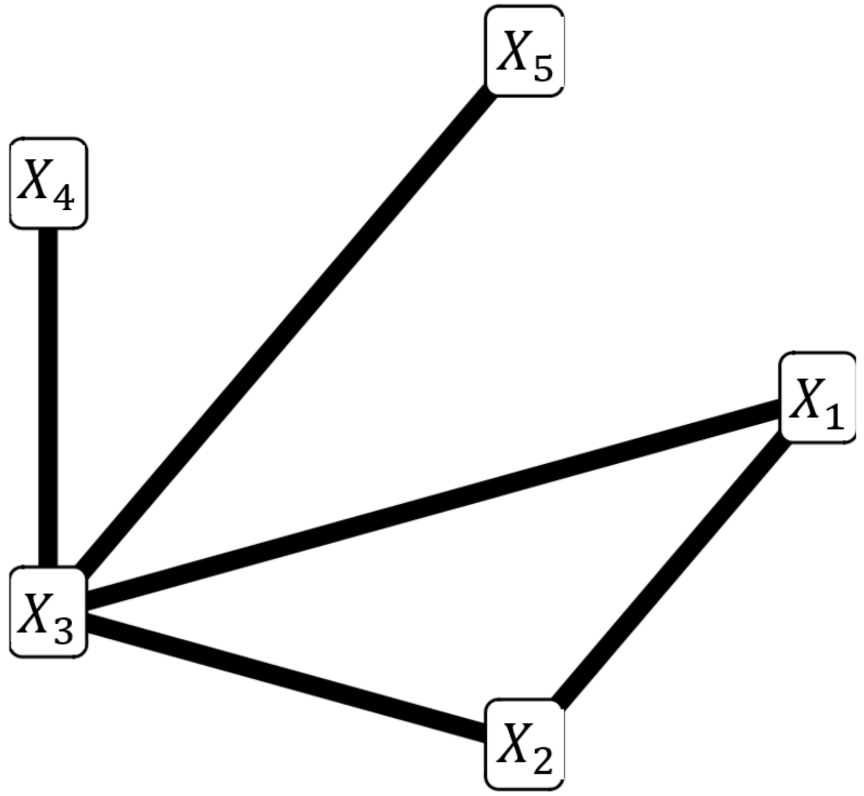
for $\widehat{\pi}(a_W) > 0$, two disjoint subsets $W, W' \subset V$, and configurations $a_W \in A^W, a_{W'} \in A^{W'}$.

Empirical Probabilities

Given a 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.



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

Then

$$\widehat{\pi}(a_v | a_{G(v)}) = \frac{\widehat{\pi}(a_{\{v\} \cup G(v)})}{\widehat{\pi}(a_{G(v)})}.$$

Graph Estimator

Given any graph $G = (V, E)$ and a sample of the process, we define the log-pseudo-likelihood function by

$$\log \widehat{L}(G) = \sum_{v \in V} \sum_{(a_v \in A)} \sum_{a_{G(v)} \in A^{|G(v)|}} N(a_v, a_{G(v)}) \log \widehat{\pi}(a_v | a_{G(v)}).$$

Theorem (Severino, 2025): Let $\{X^{(i)} : i \in \mathbb{Z}\}$ be a process that satisfies the mixing condition presented before with $\psi(\ell) = O(1/\ell^{1+\epsilon})$ for some $\epsilon > 0$. Then, by taking $\lambda_n = c \log n$, we have that

$$\widehat{G} = \arg\max_{G} \Big\{ \log \widehat{L}(G) - \lambda_n \sum_{v \in V} |A|^{|G(v)|} \Big\}$$

satisfies $\widehat{G} = G^*$ eventually almost surely as $n \to \infty$.

Proposal 1: model selection for Markov random fields with countable infinite set of vertices on graphs under a mixing condition

Proposal 1

- **Extending model selection in Markov Random Fields (MRFs)**
  - Focus on graphs with **countably infinite** vertices.
  - Applications in **neural networks** and **social networks**.
- **Motivation and existing methods**
  - **Leonardi et al. (2023)**: Penalized pseudo-likelihood for discrete MRFs.
  - Graph estimated based on local neighborhood estimation.
  - **Severino (2025)**: Developed theoretical results for global estimation of discrete MRFs over finite graphs.
- **Research goal**
  - Generalize the results from **finite** to **countably infinite graphs**.
  - Improve global estimation, possibly reducing errors from local neighborhood estimation.

Proposal 1: Estimation Approach and Applications

- **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 $n \to \infty$.
  - 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.
- **Algorithm development and evaluation**
  - Implement graph estimator in **R package**.
  - Performance assessed through **extensive simulation studies**.
- **Real-world applications**
  - **Social interaction networks**: Capturing dependencies in large-scale social systems.
  - **Online social networks**: Understanding **information diffusion & social influence**.

Proposal 2: Model selection for <u>continuous</u> Markov random fields on graphs under a mixing condition

Proposal 2

- **Extending previous work on MRFs**
  - **Severino (2025)**: Applied MRF model to **São Francisco River** water flow data.
  - Discretization was required, introducing potential limitations.
  - Goal: **Generalize the approach to continuous stochastic processes**.
- **Theoretical adaptations needed**
  - Modify key results from discrete to **continuous random variables**.
  - Replace summations with integrals in **penalized pseudo-likelihood function**.
  - Adapt **consistency and convergence proofs** for continuous measurements.
- **Key benefits**
  - Expands applicability to **environmental monitoring, finance, and signal processing**.
  - Enables more accurate inference from **continuous data sources**.

Proposal 2: Implementation & Applications

- **Algorithm development and valuation**
  - Implement adapted algorithms in **R package**.
  - Assess performance through **extensive simulation studies**.
  - Validate robustness and accuracy using real-world datasets.
- **Broader applications**
  - **Bioinformatics**: Model **gene regulatory networks** and **protein interactions**.
  - **Economics**: Capture dependencies in **financial markets** and **economic indicators**.
  - Expands to **various fields** beyond traditional MRF applications.

Future Research Direction

- Combine **Proposal 1 (countable infinite graphs)** and **Proposal 2 (continuous data)**.
- Develop a **unified method** for estimating MRFs with **continuous variables** and **countable infinite vertices**.
- Provide a **versatile tool** for modeling complex stochastic processes.

Viability

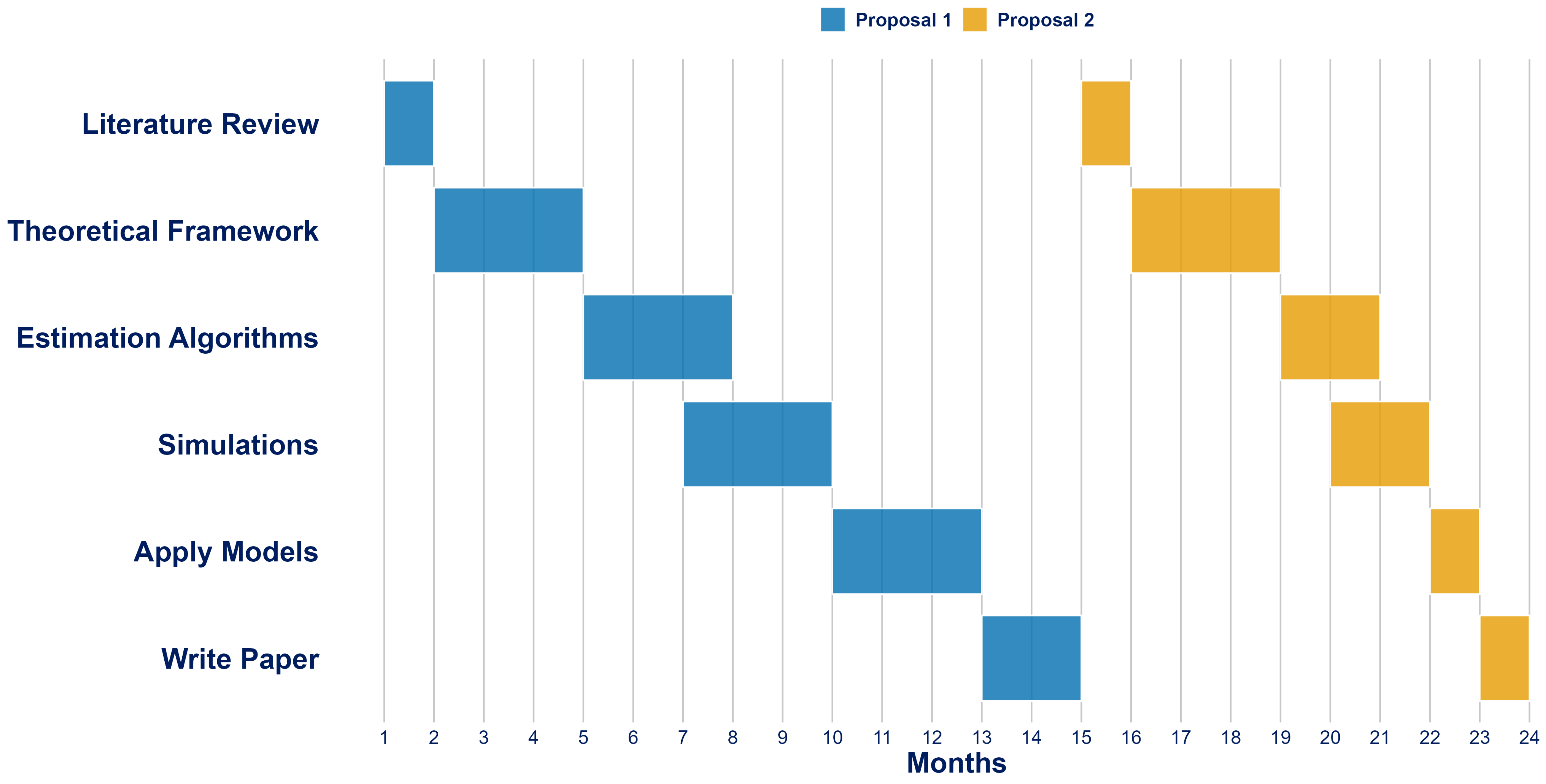
Relevance to Data Science and Statistical Learning

- **Growing Importance of Data Science**
  - Essential for innovation and decision-making across industries.
  - Need for advanced statistical methodologies to handle complex datasets.
- **Unsupervised Learning & Graphical Models**
  - Graphical model estimation enables the discovery of hidden structures in data.
  - Adaptations of methods for **continuous and dependent data processes** and **large-scale networks**.

Expected Outcomes

- **Theoretical Contributions**
  - Extension of Markov Random Fields (MRFs) theory under mixing conditions.
  - Application to graphs with countably infinite vertices and continuous data.
  - Submission to leading international journals.
- **Algorithm Development**
  - Robust, scalable algorithms for complex graphical models.
  - R packages with open-source access.
  - Comprehensive documentation with practical examples for ease of adoption.
- **Broader Scientific Impact**
  - Increased applicability in domains like **neuromathematics, finance, and ecology**.
  - Potential for interdisciplinary collaborations (UFRJ, UFRN, UBA, Neuromat researches).

Timeline


Proposal 1
Proposal 2
Literature Review
Theoretical Framework
Estimation Algorithms
Simulations
Apply Models
Write Paper
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Months

$ Budget

- **Minimal** financial costs.
- Theoretical development **does not require** any additional resources.
- Simulation and real data application of the developed algorithms will be conducted using existing computer server **resources at the department**.
- Anticipated expenses: presenting the research at international scientific conferences.

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.**, & 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 (Vol. 17)*. Clarendon Press.
- Oodaira, H., & Yoshihara, K. I. (1971). *The law of the iterated logarithm for stationary processes satisfying mixing conditions*. Kodai Mathematical Seminar Reports.