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

Projeto de Pesquisa

Área: *1.a. Inferência para Processos Estocásticos*

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

Magno Tairone de Freitas Severino

Projeto de pesquisa apresentado à
Comissão Julgadora do concurso público
para o cargo de Professor Doutor junto ao
Departamento de Estatística do
Instituto de Matemática e Estatística da
Universidade de São Paulo

São Paulo, outubro de 2024

Abstract

This research project aims to advance the field of unsupervised statistical learning by developing novel methodologies for statistical inference and model selection in stochastic processes, with a specific focus on model selection for Markov random fields under a mixing condition. Building upon the limitations of conventional methods, which primarily address finite and discrete graphical models, this study proposes two key extensions.

First, the project seeks to generalize the existing framework to accommodate graphs with a countably infinite set of vertices. This extension is important for applications in complex systems such as social networks, neural networks, and statistical physics, where infinite vertex sets reflect more realistic, large-scale structures. In addition to its relevance to these fields, the proposed methodologies can be applied to a broad range of other domain, such as economics (financial time series data), bioinformatics (gene regulatory networks) and ecology (species interactions). This flexibility underscores the potential for this research to make impactful contributions across various sectors where complex dependencies and continuous data are prevalent.

Second, the project aims to adapt the methodology originally developed for discrete data to handle continuous data processes. This adaptation will significantly broaden the applicability of the model, enabling it to address a wider array of real-world challenges, including those found in environmental sciences and finance.

Expected outcomes include the development of a robust theoretical foundation for these modelos, the creation of efficient algorithms leveraging statistical learning techniques, and the implementation of these methods in an R software package. This work will provide robust tools for researchers in data science and related fields, enabling more accurate modeling and inference in complex stochastic systems.

1 Introduction

This research project focuses on inference for stochastic processes, with applications in statistical learning and data science. It extends the work developed in my PhD thesis (Severino 2024), which addressed model selection in Markov random fields with a finite number of vertices. In the thesis, we considered a stationary vector-valued stochastic process taking values in a finite alphabet. The primary goal was to estimate the underlying graph G^* that encodes the conditional dependencies in the invariant distribution π, as well as the associated conditional probability distributions. Unlike classical models that assume independent and identically distributed observations (see, for example, Lauritzen 1996; Koller e Friedman 2009; Lerasle e Takahashi 2016; Pensar, Nyman e Corander 2017; Divino, Frigessi e Green 2000; Leonardi, Carvalho e Frondana 2023), our work considered scenarios with dependence, specifically mixing processes.

Conventional model selection techniques for graphical models often involve estimating the neighborhoods of individual nodes and constructing the graph based on these neighborhoods, as exemplified by Ravikumar, Wainwright e Lafferty (2010). But depending on the rule to combine the neighborhoods, the final estimated graph can drastically underestimate or overestimate the set of edges in the graph (Leonardi, Carvalho e Frondana 2023). For this reason, in the thesis, we adopted a global estimation perspective, optimizing a penalized pseudo-likelihood function over the set of all possible simple and undirected graphs. We proved the convergence of the estimator to the true underlying graph in finite graphical models, under a mixing condition. This result is useful in practical applications, such as EEG time series (Cerqueira et al. 2017), river stream flow data (Leonardi, Lopez-Rosenfeld et al. 2021), and stock market indices (Leonardi, Carvalho e Frondana 2023), where the assumption of independence is often too restrictive.

The current research project aims to extend this analysis in two directions: first, to situations where the set of vertices in the graph is countably infinite, which is important due to its wide-ranging applications in complex systems such as social networks, neural networks, and statistical physics. Infinite vertex sets introduce additional challenges but reflect more realistic scenarios encountered in large-scale networks and spatial processes. Second, to extend the work originally developed for discrete data to accommodate processes with continuous data, thereby broadening the applicability of the methodology to a wider range of real-world problems.

These two proposals can be pursued independently, each targeting a specific aspect of generalizing the work of Severino (2024). However, a significant advantage of this project is the potential to combine both approaches, resulting in a robust solution for globally estimating the graph of conditional dependencies in a continuous Markov random field that satisfies a mixing condition and with a countably infinite set of vertices.

This research will be carried out during the probationary period, should I be appointed as a *Doctoral Professor*, with the goal of advancing our understanding and ability to predict the behavior of complex systems through accurate estimation of underlying graph structures. Details of this proposal are outlined in Section 2. Viability, expected results, and the timetable are discussed in Section 3.

2 Proposals

This section presents two novel research proposals in the field of *inference for stochastic processes*, intersecting with *statistical learning and data science.* Both proposals build upon and extend the work developed in my PhD thesis Severino (2024), which focused on model selection criteria for graphs under mixing conditions. Our previous research concentrated on discrete multivariate stochastic processes and graphs with a fixed number of nodes. The current proposals aim to generalize this work in two distinct directions. The first proposal extends the model to cases where the set of vertices is countably infinite, while the second proposal intend to adapt our approach for continuous data.

The section is structured as follows. We begin by providing the necessary theoretical background and key definitions in Section 2.1. This foundation is necessary for understanding the subsequent proposals. Section 2.2 then details the first research proposal, exploring the extension to countably infinite vertex sets. Following this, Section 2.3 outlines the second research proposal, which focuses on adapting our method for continuous data. Through these proposals, we aim to broaden the applicability of our previous work and contribute to the advancement of inference methods for complex stochastic processes.

2.1 Background and definitions

The purpose of this section is to establish the theoretical background for this work, beginning with the concepts regarding Markov random fields on graphs in Section 2.1.1. Subsequently in Section 2.1.2, we present the definition of vector-valued stochastic processes that satisfy a mixing condition and the main contribution of Severino (2024).

2.1.1 Markov random fields on graphs

A *graph* is defined as an ordered pair $G = (V, E)$, where V represents the set of vertices (or nodes), and $E \subseteq V \times V$ is the set of edges connecting pairs of vertices. We refer to a graph as *undirected* if $(v_i, v_j) \in E$ implies that $(v_j, v_i) \in E$ for all $(v_i, v_j) \in E$, where $v_i, v_j \in V$. Furthermore, a graph is considered *simple* if $(v, v) \notin E$ for all $v \in V$. For the purposes of this work, we concentrate exclusively on undirected simple graphs, which we will henceforth simply call a *graph*.

Consider a graph $G = (V, E)$, with $V = \{1, \ldots, d\}$, for $d \in \mathbb{N}$, and assume we observe at each vertex $v \in V$ a random variable X_v, which is discrete and takes values in A, a finite alphabet. Moreover, let $X = (X_1, \ldots, X_d)$ be the vector of all variables observed on the vertices of the graph. Denote by $\mathbb{P}$ the joint probability distribution of the vector X. For any $W \subset V$ and any configuration $a_w \in A^{|W|}$ we write

$$\pi(a_W) = \mathbb{P}(X_W = a_w)\,.$$

Moreover, if $\pi(a_W) > 0$ then we denote by

$$\pi(a_U|a_W) = \mathbb{P}(X_U = a_U|X_W = a_w),$$

for $a_U \in A^{|U|}$ and $a_W \in A^{|W|}$, the corresponding conditional probability distributions.

For a given vertex $v \in V$, any set $W \subset V$, with $v \notin W$, is called a *neighborhood* of v. Furthermore, W is called *Markov neighborhood* of v if

$$\pi(a_v|a_U) = \pi(a_v|a_W)$$

for all $U \supset W$, $v \notin U$ and all $a_U \in A^{|U|}$, with $\pi(a_U) > 0$. The definition of a Markov neighborhood W of v is equivalent to request that for all $U' \subset V \setminus \{v\}$ such that $U' \cap W = \emptyset$, $X_{U'}$ is conditionally independent of X_v, given X_W. That is,

$$X_v \perp X_{U'} | X_W$$

for all U' with $U' \cap W = \emptyset$, where $\perp\!\!\!\perp$ is the usual symbol denoting independence of random variables.

As discussed in Leonardi, Carvalho e Frondana (2023), if W is a Markov neighborhood of $v \in V$, then any finite set $U \supset W$ also is a Markov neighborhood of v. In contrast, W_1 and W_2 being Markov neighborhoods of v does not imply in general that $W_1 \cap W_2$ is a Markov neighborhood of v, however this property is satisfied by some probability measures. This fact leads to the following definition.

Definition 1 (Markov intersection property). *We say that Markov intersection property holds if for all $v \in V$ and all W_1 and W_2 Markov neighborhoods of v, the set $W_1 \cap W_2$ is also a Markov neighborhood of v.*

The Markov intersection property is desirable in this context to define the smallest Markov neighborhood of a node. This property is guaranteed under the usually assumed positivity condition (see Lauritzen 1996). But the positivity assumption is not necessary to obtain consistent estimators, and then it is enough to assume the Markov intersection property, see Leonardi, Carvalho e Frondana (2023) for details.

Definition 2 (Basic neighborhood). *For $v \in V$, let $\mathcal{W}(v)$ be the set of all subsets of V that are Markov neighborhoods of v. The* basic neighborhood *of v is defined as*

$$\mathrm{ne}(v) = \bigcap_{W \in \mathcal{W}(v)} W. \tag{2.1}$$

By the Markov intersection property, $\mathrm{ne}(v)$ is the smallest Markov neighborhood of $v \in V$. Based on these basic neighborhoods, define the graph $G^* = (V, E^*)$ as

$$(v, w) \in E^* \ \text{ if and only if } w \in \mathrm{ne}(v), \tag{2.2}$$

where $E^* \subseteq V \times V$. The graph G^* with edges defined in (2.2) is *undirected*, as proved by Leonardi, Carvalho e Frondana (2023). Figure 2.1 shows two examples of graphs for Markov random fields under different settings: in (a) the finite non-homogeneous graphical model case, and in (b) the interaction graph in a classical Markov random field, or Gibbs distribution, on a regular lattice (Comets e Gidas 1992; Csiszár e Talata 2006).

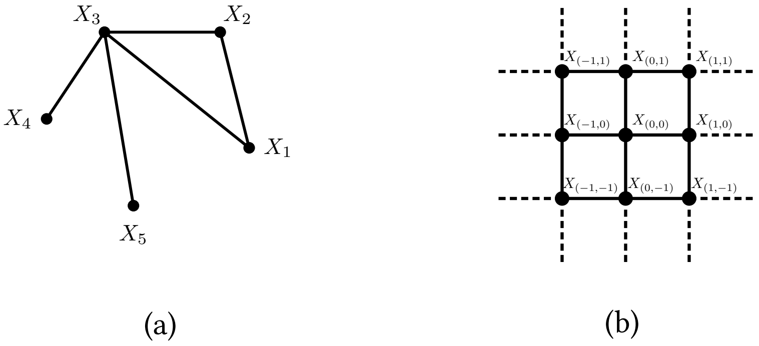

Figura 2.1: Different graph structures for Markov random fields under different settings. The figure on the left is a finite graphical model or Markov random field on a general graph, and the figure on the right represents the interaction graph in a classical Markov random field or Gibbs distribution on a regular lattice.

2.1.2 Vector-valued mixing processes

Consider a vector-valued stochastic process $X^{(1)}, X^{(2)}, \ldots$, where each variable $X^{(i)}$ is a vector of d components, belonging to the set A^d, with A a finite alphabet. We denote by $\big((A^d)^{\mathbb{N}}, \mathcal{F}, \mathbb{P}\big)$ the probability space for the process $\{X^{(i)} \colon i \in \mathbb{N}\}$. Sometimes we need to consider "slices" of the entire realization $X^{(1)}, \ldots, X^{(n)}$ on both dimensions. To avoid misleading notations we use superscripts to denote the indexes in "time" (ranging from 1 to n) and subscripts to denote indexes on "space" (a subset of $V = \{1, \ldots, d\}$). For any set $U \subset V$ and any integer interval $i : j$ we denote by $X_U^{(i:j)}$ the sequence $X_U^{(i)}, \ldots, X_U^{(j)}$ with $X_U^{(k)} = (X_u^{(k)} \colon u \in U)$, $k = i, \ldots, j$. When $U = V$ we avoid the subscript and simply write $X^{(i:j)}$. The same notation is used for "realizations" of the process, denoted in lower case $x_U^{(i:j)}$ instead of the notation for the random variables $X_U^{(i:j)}$.

We say the process $X^{(1)}, X^{(2)}, \ldots$ satisfies a mixing condition with rate $\{\psi(\ell)\}_{\ell \in \mathbb{N}}$ if for each k, m and each $x^{(1:k)}$, $x^{(1:m)}$ with $\mathbb{P}(X^{(1:m)} = x^{(1:m)}) > 0$ we have that

$$\begin{aligned} \big|\mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)} \,|\, X^{(1:m)} = x^{(1:m)}) - \mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)})\big| \\ \leq\; \psi(\ell)\, \mathbb{P}(X^{(n:(n+k-1))} = x^{(1:k)}) \end{aligned} \tag{2.3}$$

for $n \geq m + \ell$.

Assume we observe a sample of size n of the process, denoted by $\{x^{(i)} \colon i = 1, \ldots, n\}$. Since the stationary distribution of the process π is not known, we must estimate it from the data. For any $W \subset V$ and any $a_W \in A^W$ denote by

$$\widehat{\pi}(a_W) = \frac{N(a_W)}{n}\,, \tag{2.4}$$

where $N(a_W)$ denotes the number of times the configuration a_W appears in the sample $x^{(1)}, \ldots, x^{(n)}$.

If $\widehat{\pi}(a_W) > 0$, we can also define the conditional probabilities

$$\widehat{\pi}(a_W|a_{W'}) = \frac{\widehat{\pi}(a_{W\cup W'})}{\widehat{\pi}(a_{W'})}\,, \tag{2.5}$$

for two disjoint subsets $W, W' \subset V$ and configurations $a_W \in A^W, a_{W'} \in A^{W'}$.

Propositions 3 and 4 stated below show upper bounds for the rate of convergence of $\widehat{\pi}(a_W)$ into $\pi(a_W)$ and $\widehat{\pi}(a_v|a_W)$ into $\pi(a_v|a_W)$, respectively. These auxiliary results are needed for proving Theorem 5 and are proved in Severino (2024) using the Law of the Iterated Logarithm for mixing processes, as derived from Oodaira e Yoshihara 1971.

Proposition 3 (Typicality). *Assume the process $\{X^{(i)} \colon i \in \mathbb{Z}\}$ satisfies the mixing condition* (2.3) *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 $n \to \infty$.

Proposition 4 (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 $n \to \infty$.

Moreover, Severino (2024) takes a regularized pseudo maximum likelihood approach to estimate the graph G^*, given a sample $x^{(1)}, \dots, x^{(n)}$ of the stochastic process. Instead of estimating each neighborhood and then combining the results, as is proposed in several works, they globally estimate the graph G^* by optimizing a function over the set of all simple graphs over V.

Given any graph G defined on the same set of vertices V, the pseudo-likelihood function is defined by

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

where $G(v)$ denotes the neighborhood of node v in the graph G, that is $G(v) = \{u \in V \colon (u,v) \in E\}$. As the conditional probabilities of π are not known, they can be estimated them from the data, obtaining the maximum pseudo-likelihood given by

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

with $\widehat{\pi}(x_v^{(i)}|x_{G(v)}^{(i)})$ defined as in (2.5), taking $W = \{v\}$ and $W' = G(v)$.

Applying the logarithm and taking into account the number of occurrences of each configuration in the sample, we can write the log pseudo likelihood function as

$$\log \widehat{L}(G) \;=\; \sum_{v\in V}\sum_{a_v,a_{G(v)}} \;\widehat{\pi}(a_v|a_{G(v)})^{N(a_v,a_{G(v)})}\,, \tag{2.6}$$

where the sum is taken over all $v \in V$ and all configurations $a_v \in A$, $a_{G(v)} \in A^{G(v)}$ such that $N(a_v, a_{G(v)}) > 0$.

The graph estimador and its consistency are stated in Theorem 5 below and the proof is presented in Severino (2024) and is the main theoretical contribution of my PhD thesis.

Theorem 5. *Assume the process $\{X^{(i)} : i \in \mathbb{N}\}$ satisfies the mixing condition* (2.3) *with rate $\psi(\ell) = O(1/\ell^{1+\epsilon})$ for some $\epsilon > 0$. Then, taking $\lambda_n = c \log n$, with $c > 0$, we have that $\widehat{G}$*

$$\widehat{G} \;=\; \arg\max_{G}\Big\{\, \log \widehat{L}(G) - \lambda_n \sum_{v\in V} |A|^{|G(v)|} \,\Big\}\,, \tag{2.7}$$

satisfies $\widehat{G} = G^*$ eventually almost surely when $n \to \infty$. In Equation (2.7)*, $|G(v)|$ is the cardinal of the set $G(v)$.*

The work in Severino (2024) was mainly motivated by the works of Leonardi, Lopez-Rosenfeld et al. (2021) and Leonardi, Carvalho e Frondana (2023) and can be viewed as a combination and generalization of both. In the work by Leonardi, Carvalho e Frondana (2023), penalized pseudo-likelihood criterion for estimating the graph of conditional dependencies within partially observed IID discrete Markov random fields. The graph is constructed by merging the estimated neighborhood of each vertex. On the other hand, Leonardi, Lopez-Rosenfeld et al. (2021) presents a different perspective by proposing a model selection criterion for estimating points of independence within a random vector that satisfies a mixing condition. This results in decomposing the vector's distribution function into distinct independent blocks. The method, based on a general estimator of the distribution function, can be applied to both discrete and continuous random vectors, as well as IID data or dependent time series. The authors have proved the consistency of the approach under general conditions on the estimator of the distribution function and show that the consistency holds for IID data and discrete time series with mixing conditions.

In Severino (2024), overcame the limitations of the previously mentioned works. While the estimator introduced by Leonardi, Carvalho e Frondana (2023) is only applicable to IID data and the estimation is done for each vertex, the method proposed by Leonardi, Lopez-Rosenfeld et al. (2021) assumes that the random vector can only be decomposed into subvectors. In response to these limitations, we proposed a penalized pseudo-likelihood criterion for estimating the entire graph G, which consists of the set of edges E connecting the nodes V, particularly for multivariate stochastic processes satisfying a mixing condition. The primary advantage of our approach is its ability to handle non-IID data and its global estimation approach. This means that the entire set of edges E is estimated as a whole, eliminating the need to estimate the neighborhood of each node separately and then combine them to obtain the estimated graph. Theorem 5 provides rate of convergence, showing that the estimator almost surely converges to the actual underlying graph in cases of finite graphical models, provided a mixing condition holds for the generating process.

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

Markov random fields (MRFs) provide a powerful framework for modeling complex dependencies among random variables, particularly in high-dimensional data contexts. This proposal seeks to extend existing methodologies for model selection in MRFs to scenarios where the underlying graph has a countably infinite set of vertices, a situation relevant to numerous applications in neural networks, social networks (Frank e Strauss 1986; Taskar, Abbeel e Koller 2012), where interactions often go beyond finite structures.

Leonardi, Carvalho e Frondana (2023) introduced a penalized pseudo-likelihood criterion for estimating the graph of conditional dependencies in partially observed discrete Markov random fields. This technique focuses on estimating the neighborhood of each node in the graph, and the authors have demonstrated almost sure convergence of the estimator in cases where the number of variables is either finite or countably infinite, assuming the process is independent and identically distributed. Moreover, their method imposes minimal assumptions on the probability distribution, eliminating the need for the positivity condition often required by other approaches. However, as the authors discuss, the final estimated graph can significantly underestimate or overestimate the set of edges depending on the rule used to combine the individual neighborhoods.

Building on the work of Leonardi, Carvalho e Frondana (2023) and the developments proposed by Severino (2024), this proposal aims to adapt these methodologies to the context of countably infinite variables. The theoretical results presented in Severino (2024) were initially intended for discrete Markov random fields over A^V with a finite vertex set V. This proposal seeks to generalize these results to cases involving a countably infinite set of variables. The proposed methodology will take a global estimation approach, optimizing a penalized pseudo-likelihood function across the entire set of possible graphs. This approach mitigates the risks associated with estimating individual neighborhoods, which can lead to significant errors in the estimation of the graph's edges.

In Section 2.1.2, we considered a sample of size n from $\{\mathbf{X} = \{X_v : v \in V\}\}$, that can be represented as a matrix with dimensions $n \times d$, where $|V| = d$. Here in the proposed scenario, for $i = 1, \ldots, n$, we will assume the observation of a sample $\{V_v : v \in V_n\}$, where V_n is a finite subset of V', with $V_n \uparrow V'$ as $n \to \infty$, i.e., $V_1 \subseteq V_2 \subseteq V_3 \subseteq \cdots$.

The estimation of conditional probabilities for the sample will remain consistent with (2.4) and (2.5), as we should focus on finite subsets of V. However, the graph estimation must be adjusted to account for the countably infinite nature of G^*. In this context, we can only reconstruct a finite subgraph of G^* with a finite sample. Therefore, we must adapt Propositions 3 and 4 to accommodate this setting and propose a modified version of Theorem 5 to reflect these changes. The proof will draw inspiration from Theorem 1 in Leonardi, Carvalho e Frondana (2023).

Additionally, after proving the theoretical results, we will develop algorithms to implement the new graph estimator. The performance of these estimators will be thoroughly evaluated through extensive simulation studies, and the resulting code will be packaged into an R package R Core Team (2022) for distribution to the scientific community.

We plan to apply this algorithm to social interaction networks to capture the complex dependencies that arise in large-scale social systems. In such networks, each individual is represented as a vertex, while the connections between them, reflecting social ties or interactions, form the edges of the graph Lazakidou (2012). The countably infinite nature of the vertex set allows for the representation of extensive networks, such as online social networks. This approach is particularly beneficial for understanding phenomena such as information diffusion, social influence, and community detection.

2.3 Proposal 2: Model selection for continuous Markov random fields on graphs under a mixing condition

The work developed by Severino (2024) was originally designed for discrete random vectors satisfying a mixing condition and taking values in a finite alphabet A. The methodology proved effective when applied to the water flow measurements along the São Francisco River, where the vector of measurements forming the multivariate stochastic process consisted of continuous observations. However, to apply the proposed method, it was necessary to discretize the continuous data, which introduced potential limitations in the analysis. Thus, a natural extension of this work involves generalizing the approach to accommodate continuous multivariate stochastic processes. This extension would broaden the range of data types that can be effectively analyzed using this class of estimators, enhancing its versatility and applicability in various fields, including environmental monitoring, finance, and signal processing.

Several theoretical results from my previous work, originally developed for discrete data, will need to be adapted to accommodate continuous data. Key results, such as Propositions 3 and 4 and Theorem 5, must be revised to ensure their validity in the context of continuous random variables. This involves redefining the state space, replacing summations with integrals, and adapting the penalized pseudo-likelihood function to handle continuous data. Additionally, the consistency and convergence proofs of the graph estimator will need to be restructured to account for the nuances of continuous measurements, including the application of appropriate regularity conditions and measures.

Once the theoretical framework is established, we will proceed with the proposition and mplementation of the adapted algorithms in R R Core Team (2022). The performance of these algorithms will be evaluated through extensive simulation studies, similar to those conducted by Severino (2024), to ensure their robustness and accuracy. Finally, the developed code will be packaged into an R package and made available to the scientific community.

This research will extend the existing methodology to continuous data, offering a robust tool for analyzing complex systems across various domains where continuous measurements are common. In addition to applications in environmental monitoring, such as the São Francisco River data used in Severino (2024), these methods can also be extended to other important fields:

- **Bioinformatics**: In genomics, the interactions between genes and proteins often form complex networks, where continuous data (e.g., expression levels, protein concentrations) are measured. By applying continuous Markov random fields to these networks, we can more accurately model gene regulatory networks or protein-protein interactions, improving our understanding of diseases like cancer or genetic disorders.

- **Economics**: Financial markets generate continuous time series data, and dependencies between different economic indicators or stock prices are often highly complex. The proposed methodology could be used to model dependencies across different markets or economic factors, enhancing our ability to predict market behavior or identify systemic risks.

These extensions highlight the versatility and applicability of the proposed methodologies in real-world scenarios, significantly broadening their impact beyond traditional fields such as social and neural networks.

The proposal to generalize Severino (2024) to cope with continuous data, along with the previous proposal (Section 2.2) to extend it to graphs with a countably infinite number of vertices, could be combined into a third project. This new research would focus on developing a method for estimating graphs in multivariate stochastic processes that take continuous values, satisfy a mixing condition, and involve a countably infinite set of vertices. Such a comprehensive approach would offer a powerful and versatile tool for modeling complex systems, addressing challenges posed by the mixing condition, the continuous nature of the data, and the infinite structure of the graph.

3 Research Viability and Planning

This section outlines the key aspects that demonstrate the feasibility of the proposed research project (Section 3.1), along with the expected outcomes (Section 3.3), detailed timeline (Section 3.4), and budget considerations (Section 3.5). The project is designed to be both ambitious and realistic, leveraging existing resources at the institute and building on my extensive experience in the field.

This project has strong potential to be carried out in collaboration with researchers from other universities in Brazil as well as international institutions. These partnerships would enhance the project's scope and impact, contributing to a broader exchange of expertise and knowledge. Through a carefully planned approach, the project aims to produce significant theoretical and practical contributions within a 24-month period, culminating in publications in leading international journals and presentations at scientific conferences.

3.1 Viability

I have been involved in research projects since my undergraduate studies. During my PhD, I gained substantial experience in inference for stochastic processes through the development of a focused research project, which has provided me with a strong foundation in this area. I have already published one article and have another submited and under minor revision in journals with selected editorial policy. Furthermore, the Department of Statistics at IME-USP, where this project will be conducted, is well-equipped with the necessary resources to support the research.

Additionally, this project has strong potential for collaboration with researchers from other universities, further strengthening the interinstitutional relationships of the Department of Statistics at IME. Proposal 1 can be developed in partnership with Dr. Daniel Takahashi (Federal University of Rio Grande do Norte, UFRN) and Dr. Guilherme Ost (Federal University of Rio de Janeiro, UFRJ), both of whom were members of the committee during my PhD defense. Following the defense, I was invited to collaborate with them on a project in the same area as Proposal 1, which provides a strong basis for continuing this partnership.

Proposal 2 could benefit from international collaboration with researchers such as Dr. Mariela Sued and Dr. Daniela Rodriguez from the University of Buenos Aires (UBA). I had the opportunity to work with both of them during the early stages of my PhD on a research project related to multivariate stochastic processes. Their expertise in the field would be invaluable in advancing the theoretical and practical aspects of this work. Furthermore, this project could form part of a postdoctoral fellowship at UBA, where the collaboration with these researchers would enhance the quality of the research and foster greater international cooperation.

One additional factor that strengthens the viability of this project is the potential for it to be connected with the *Research, Innovation, and Dissemination Center for Neuromathematics* (NeuroMat). The project is closely aligned with the objectives of NeuroMat, which focuses on advancing mathematical and computational tools for neuroscience. During my PhD, I was affiliated with the center and actively engaged in its research activities. This existing relationship with NeuroMat provides additional resources and support, further enhancing the research environment and increasing the potential impact of this project.

Considering these factors – my prior experience, the resources at IME-USP, and the potential for collaboration with leading researchers both in Brazil and abroad – I am confident that this project can be successfully executed, including the submission of scientific articles within the planned 24-month timeline.

Despite the strong foundation and resources available for this project, certain challenges are expected in both the theoretical and computational aspects. The following obstacles have been identified, along with proposed strategies to address them:

- **Theoretical Convergence**: An important challenge is ensuring the theoretical convergence of the estimators for infinite graphs and continuous data. While existing theoretical results are promising, generalizing them to new contexts can be complex and may require reformulating certain assumptions or proving new regularity conditions. To overcome this challenge, the project will include a careful study of the theoretical conditions necessary to guarantee the consistency and efficiency of the estimators, building upon the work of Severino (2024) and extending the results of Leonardi, Carvalho e Frondana (2023). International collaboration with specialists in graph modeling and inference for stochastic processes will be essential to refine the theoretical aspects.

- **Computational Complexity**: One of the main anticipated challenges is the high computational complexity involved in generalizing the proposed methods to graphs with a countably infinite set of vertices. Execution time and memory usage can increase significantly as the number of vertices and the dimensionality of the data grow. To address this problem, I intend to employ advanced optimization techniques and parallelized algorithms to reduce execution time, leveraging the computational resources available at the Department of Statistics at IME-USP.

3.2 Relevance to Data Science and Contributions to Unsupervised Statistical Learning

In recent years, Data Science has become a critical field, driving innovation and decision-making across a wide range of industries. The rapid growth of complex datasets and the increasing demand for sophisticated models to process, analyze, and derive insights from these data have elevated the importance of developing advanced statistical methodologies. This project directly aligns with these contemporary needs, as highlighted in the justification for the position in the Department of Statistics, with a specific focus on expanding the capabilities of unsupervised learning methods.

The estimation of graphical models can be viewed as a field of unsupervised statistical learning, as it allows for the identification of hidden structures and patterns in data without predefined labels. The advancements proposed here, particularly in the adaptation of methodologies for continuous data processes and large-scale networks, represent a significant step forward in making these models more robust and applicable to diverse domains.

By contributing new methodologies and practical tools, such as the planned implementation of algorithms in an R package, this project will provide the Data Science community with powerful tools to advance the analysis and understanding of complex systems. Ultimately, the proposed research addresses critical gaps in current unsupervised learning models, aligning with the strategic goals of the department and the broader field of Data Science.

3.3 Expected Outcomes

Given the strong relevance of this project to the current needs of Data Science, particularly in the area of unsupervised statistical learning, the research is expected to yield several significant outcomes

First, it will contribute new theoretical results that extend the current understanding of Markov random fields (MRFs) under a mixing condition, especially in the context of graphs with countably infinite vertices and continuous data processes. These theoretical advances will be submitted to leading international journals, ensuring their dissemination to the broader scientific community.

Additionally, the project will develop robust, scalable algorithms designed to handle these complex graphical models. These algorithms will be implemented and released in R software packages, making them freely accessible to the research community. The software will be accompanied by comprehensive documentation, including detailed use cases and practical examples to ensure ease of adoption by other researchers. This accessibility will enhance the impact of the project by enabling its application across various domains, including bioinformatics, finance, and ecology.

Moreover, the project's outcomes have the potential to foster interdisciplinary collaborations. For example, the models developed for continuous data processes could lead to partnerships with researchers in bioinformatics to model gene regulatory networks or with economists working on financial risk analysis. These collaborations could result in joint publications, further extending the reach and impact of the research.

3.4 Timeline

The schedule for each proposal are outlined below.

Proposal 1

Month 1: Conduct a literature review focused on Markov random fields on graphs with a countable number of vertices, specifically in the context of data with conditional dependence (mixing condition).

Months 2 – 4: Develop the theoretical framework necessary for the project.

Months 5 – 6: Propose and implement efficient algorithms for estimation.

Months 7 – 9: Perform simulation studies to assess the effectiveness of the proposed methods and the efficiency of the algorithms.

Months 10 – 12: Apply the developed models to real-world data sets.

Months 13 – 14: Write and submit a research article summarizing the findings. Additionally, release the developed computational package in R.

Proposal 2

Month 15: Initiate a literature review on graph estimation in Markov random fields for continuous data.

Months 16 – 18: Focus on advancing the theoretical aspects of the project.

Month 19: Adapt the algorithms proposed in Severino 2024 for use in graph estimation with continuous data.

Months 20 – 21: Conduct simulation studies to evaluate both the performance of the methods and the efficiency of the adapted algorithms.

Month 22: Implement the models using real data sets.

Months 23 – 24: Write and submit a research article detailing the results, and release the corresponding R package.

3.5 Budget

This research project is expected to incur minimal financial costs. The theoretical development phase does not require any additional resources, and the simulation and real data application of the developed algorithms will be conducted using existing computer server resources at the institute. The only anticipated expenses are related to presenting the research findings at international scientific conferences, for which the department has allocated funds.

Bibliography

Cerqueira, Andressa et al. (2017). «A Test of Hypotheses for Random Graph Distributions Built from EEG Data». Em: *IEEE Transactions on Network Science and Engineering* 4.2, pp. 75–82.

Comets, Francis e Basilis Gidas (1992). «Parameter Estimation for Gibbs Distributions from Partially Observed Data». Em: *The Annals of Applied Probability* 2.1, pp. 142–170.

Csiszár, I. e Z. Talata (2006). «Consistent estimation of the basic neighborhood of Markov random fields». Em: *The Annals of Statistics* 34.1, pp. 123–145.

Divino, Fabio et al. (2000). «Penalized Pseudolikelihood Inference in Spatial Interaction Models with Covariates». Em: *Scandinavian Journal of Statistics* 27.3, pp. 445–458.

Frank, Ove e David Strauss (1986). «Markov graphs». Em: *Journal of the american Statistical association* 81.395, pp. 832–842.

Koller, Daphne e Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Lazakidou, Athina A (2012). *Virtual communities, social networks and collaboration*. Vol. 15. Springer Science & Business Media.

Leonardi, Florencia, Matías Lopez-Rosenfeld et al. (2021). «Independent block identification in multivariate time series». Em: *Journal of Time Series Analysis* 42.1, pp. 19–33.

Leonardi, Florencia et al. (2023). «Structure recovery for partially observed discrete Markov random fields on graphs under not necessarily positive distributions». Em: *Scandinavian Journal of Statistics (accepted)*.

Lerasle, Matthieu e Daniel Y. Takahashi (2016). «Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields». Em: *Bernoulli* 22.1, pp. 325–344.

Oodaira, Hiroshi e Ken-ichi Yoshihara (1971). «The law of the iterated logarithm for stationary processes satisfying mixing conditions». Em: *Kodai Mathematical Seminar Reports*. Vol. 23. 3. Department of Mathematics, Tokyo Institute of Technology, pp. 311–334.

Pensar, Johan et al. (2017). «Structure Learning of Contextual Markov Networks using Marginal Pseudo-likelihood». Em: *Scandinavian Journal of Statistics* 44.2, pp. 455–479.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ravikumar, Pradeep et al. (2010). «High-dimensional Ising model selection using l_1-regularized logistic regression». Em: *Ann. Statist.* 38.3, pp. 3022–1319.

Severino, Magno Tairone de Freitas (2024). «Estimation and model selection for graphical models under mixing conditions». The Digital Library of Theses and Dissertations of the University of São Paulo. Tese de doutoramento. São Paulo, Brasil: Universidade de São Paulo.

Taskar, Ben et al. (2012). «Discriminative probabilistic models for relational data». Em: *arXiv preprint arXiv:1301.0604*.