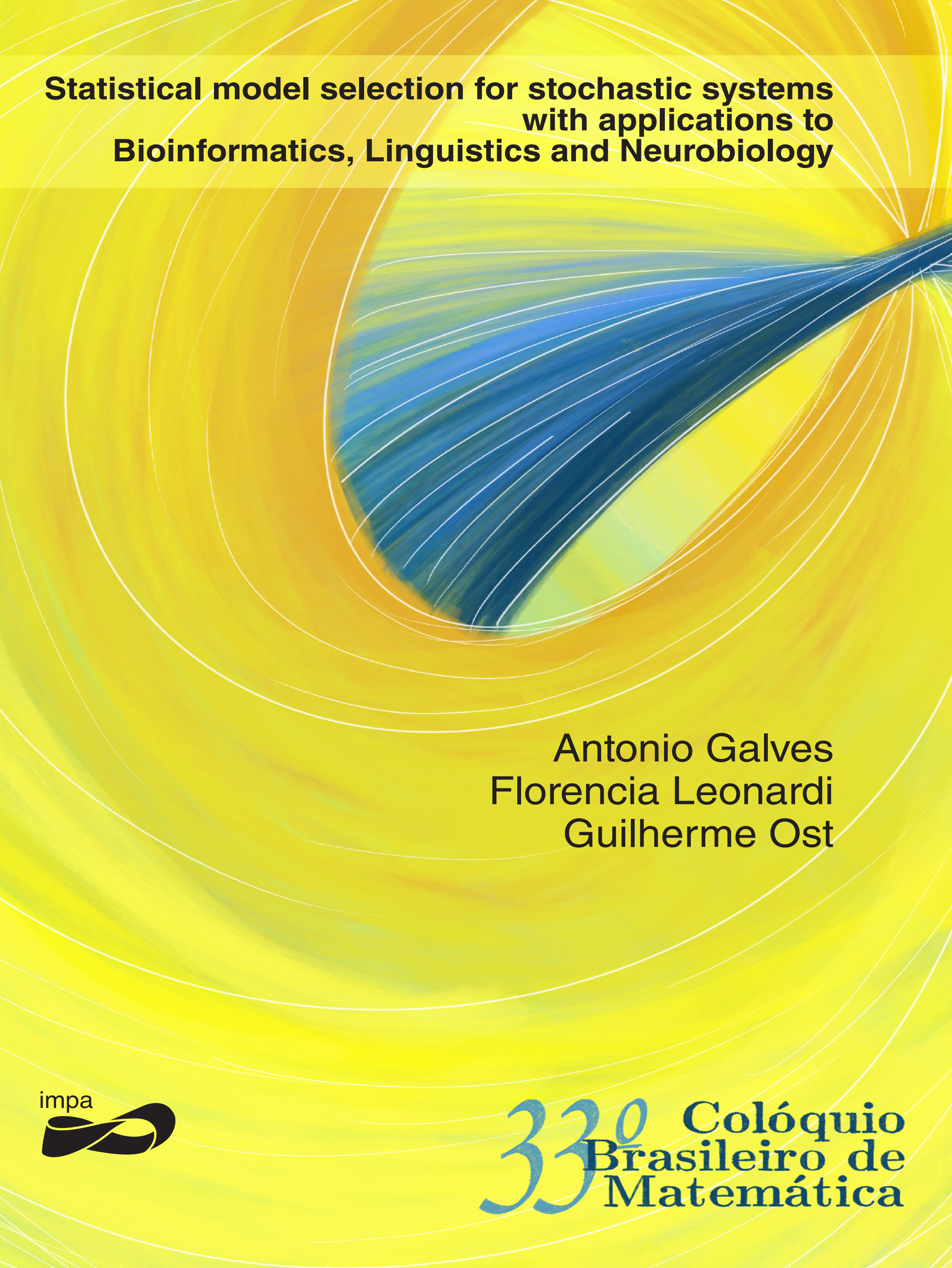

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology

Antonio Galves
Florencia Leonardi
Guilherme Ost

impa

33º Colóquio Brasileiro de Matemática

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology

Statistical model selection for stochastic systems with applications to Bioinformatics, Linguistics and Neurobiology
Primeira impressão, julho de 2021

Publicado no Brasil / Published in Brazil.

ISBN 978-65-89124-29-0
MSC (2020) Primary: 62M09, Secondary: 62M20, 60G17, 60J05, 62M05, 60J20

Coordenação Geral Carolina Araujo

Produção Books in Bytes **Capa** Izabella Freitas & Jack Salvador

Realização da Editora do IMPA
IMPA
Estrada Dona Castorina, 110
Jardim Botânico
22460-320 Rio de Janeiro RJ

`www.impa.br`
`editora@impa.br`

Contents

1 Introduction

This series of lectures present new results on statistical model selection for stochastic systems. The majority of the results to be discussed are original and first appeared in several recent papers co-authored by the authors of these notes. They all share a common feature, namely, they propose a new conceptual framework to assign appropriate models to specific samples of scientific data, displaying non trivial interactions in time and space.

The papers in which these notes are based found their original motivation in problems and data coming from linguistics or biology. In spite of their original specific motivation, we do believe that the models and statistical procedures presented here can be applied to a large variety of data sets, produced by different types of scientific data sets, representing time evolutions with structural time and space dependencies. This belief justifies the existence of the present notes.

The models and statistical procedures discussed here are interesting from an applied point of view, but not only. Actually they are also interesting from a purely theoretical point of view, as mathematical objects. In fact, all the results presented here have been rigorously proved and these proofs are presented in the notes. However, this rigour should not scare applied researchers. These notes are written in such a way the models, statistical procedures and results are presented in an intuitive way. Proofs appear only in a separated section at the end of the chapters. They are there to be read by those interested in the technical details which are necessary to prove the theorems associated to the properties of the models and procedures. Those who are only interested in the application of the models presented in the notes can skip the proofs.

Let us now summarise the goal content of these notes. Let us start by discussing the meaning of the title: statistical model selection for stochastic systems.

What is statistical model selection?
Statistical model selection is a domain of Statistics. It refers to a most important issue, namely, how to assign models to samples of experimental data.

What is a model?
A model is a description of a procedure which is able to generate samples with the same statistical features displayed by the sample of experimental data we are analysing. By procedure we mean, for instance, a computational algorithm able to generate a string of symbols.

Example: if the sample is a string of symbols, a possible model is a computational algorithm, producing sequentially the symbols, one by one, by taking into account, at each step, the last symbols already generated.

A naive model could, for instance, assume that each next symbol is produced independently of the string of past symbols. Or it could assume that each next symbol depends only on the last symbol already generated. This class of models was introduced by the Russian mathematician Andrey Andreyvich Markov in 1913 to model the occurrence of consonants and vowels in Pushkin's poem Eugene Oneguin (Markov 2006).

We could generalise Markov's original assumption and assume that each next symbol depends on the last k symbols, where k is a fixed integer greater or equal 1. More recently, in 1983, the Finish information computer scientist Jorma Rissanen observed that typically string of symbols produced by scientific experiments have a dependence from the past which is not fixed, but has a length which is a function of the past itself. This leads Rissanen to introduce what was latter called the class of *chains with memory of variable length*.

What statistics has to do with this?
The intrinsic randomness of typical samples of scientific data makes it unavoidable to use statistical criteria to select a model. In other words, we do not look for a procedure that generates a sample which is identical to the original sample of scientific data; instead we look for a procedure that is able to generate samples with

the same statistical features of those displayed by the sample of experimental data.

What is the motivation of this quest for models?
In 1867, the physicist von Helmholtz observed that the human brain does statistical model selection all the time, by making hypothesis and assigning models to sequences of stimuli, in order to be able to make predictions about what will occur in the near future. This neurobiological ability was called *unconscious inference* by von Helmholtz.

Assigning models to stimuli, in order to be able to make predictions about the future stimuli is crucial to be able to make good choices in real life, in all kind of situations, from driving a car without touching or being touched by other vehicles, to simply surviving in a hostile environment.

Less dramatically, other examples of the need of statistical model selection in real life, include being able to make reliable predictions about the time evolution of the stock market, of the weather, of the options of a set of voters, etc.

In computer science assigning models to string of bits is necessary to compress data. Medical diagnostic imaging is essentially a matter of statistical model selection. More generally, in all branches of science assigning models to samples of data is necessary to understand to structure and typical features of samples of scientific data.

What are the classes of models that will considered in this series of lectures?
In the lectures two classes of stochastic systems will be considered. First of all, the class of stochastic chains with memory of variable length, introduced by Jorma Rissanen in his 1983 seminal paper: *A universal system for data compression* (Rissanen 1983).

The title of the paper refers to the fact that models in this class are dense in the class of chains with memory of infinite order. This means that any stationary finite sample of symbols can be well approximated by models in this class.

From an applied point of view these models are interesting because they are able to maximise the likelihood of a sample, and simultaneously minimise the number of degrees of freedom of the model. This means that they are good candidates to approximate real life samples of data, in a greedy way.

We will also work with interacting systems of point process with memory of variable length and in particular systems of interacting chains with memory of infinite length. They extend Rissanen's' ideas to systems with space-time interactions, which are required to deal with medical imagery, multiunit records of neuronal activity and samples representing systems with many components interacting in time and space. From a mathematical point of view, this class of systems extends in a non trivial way class of interacting Markov systems introduced by Spitzer (Spitzer 1970).

Is this a course in Probability Theory or is this a course in Statistics? In this series of lectures, we introduce probabilistic models which are interesting mathematical objects by themselves. We also discuss how these mathematical models can be used to model sets of scientific data.

To apply the models to data analysis, is necessary to study in a rigorous way the properties of the algorithms used to select the model which best fits the data. This requires proving theorems which are mathematically challenging and technically difficult.

Besides discussing the mathematical rigorous framework required to make statistical analysis with these models, we also face the challenge of analysing real scientific data, with samples and scientific questions coming from linguistics, proteomics and neurobiology.

Is this course related to Data Science?
The answer is clearly: yes! Data Science's goal is to assign models to huge sets of data, in order to make predictions, or to classify data, putting together data with same features. It turns out that identifying essential features in the data is rarely a task which can be solved by naive "visual inspection". Real classification necessarily requires the identification of a model able to generate samples with the same statistical features as those displayed by the original data set.

A naive point of view which considers that Data science requires only computational power will only be able to produce superficial and non interesting results. To be successful Data Science requires the development of new classes of stochastic systems and new statistical selection procedures. This is precisely the goal of this series of lectures.

By the way, one of the articles that will be discussed in the lectures, A. Galves, C. Galves, et al. 2012, has received in 2020 the Johannes Kepler award discerned for the first time by the SBMAC, the Brazilian Society for Applied and Computational Mathematics. The name of the award comes from the fact Johannes Kepler can be considered the first data scientist in history.

So the answer is yes. This course is clearly related to Data Science. We hope that they will be useful for young researchers interested in the stochastic modeling of very large samples of complex data.

Acknowledgements

Each chapter of this notes starts by mentioning the original papers co-authored by us, where the models, procedures and results discussed here where first presented. So it is just fair to conclude this introduction by naming and thanking all of our co-authors. They are Ludmila Brochini, Aline Duarte, Charlotte Galves, Jesus Enrique Garcia, Pierre Hodara, Aurélien Garivier, Eva Löcherbach, Nancy Lopes Garcia, Christophe Pouzat and Patricia Reynaud-Bouret.

This work is part of USP project *Mathematics, computation, language and the brain* and FAPESP projects *Research, Innovation and Dissemination Center for Neuromathematics* (grant 2013/07699-0) and *Model selection in high dimensions: theoretical properties and applications* (grant 2019/17734-3). AG is partially supported by CNPq fellowship (grant 309501/2011-3). FL is partially supported by CNPq fellowship (grant 311763/2020-0). GO is partially supported by FAPERJ fellowship (grant E-26/211.343/2019).

2 Stochastic chains with memory of variable length

In this chapter we introduce the main definitions concerning stochastic chains with memory of variable length. We also describe the main algorithms in the literature to estimate the parameters and the structure of the context tree associated to the model. The material in this chapter is based mainly on the articles A. Galves, C. Galves, et al. (2012), A. Galves and Leonardi (2008), Garivier and Leonardi (2011), and Leonardi (2010).

2.1 Model definition

The idea behind the notion of stochastic chains with memory of variable length is that the probabilistic definition of each symbol only depends on a finite part of the past and the length of this relevant portion is a function of the past itself. The minimal relevant part of each past is called *context*. The set of all contexts satisfies the suffix property which means that no context is a proper suffix of another context. This property allows to represent the set of all contexts as a rooted labeled tree. With this representation the process is described by the tree of all contexts and an associated family of probability measures, indexed by the tree of contexts. Given a context, its associated probability measure gives the probability of the next symbol for any past having this context as a suffix. In the sequel we put these ideas in formal terms.

2.1.1 Irreducible trees

We write $\mathbb{N}$ to denote the set of natural numbers $\{0, 1, 2, \dots\}$. The set of integers $\{\dots, -1, 0, 1, \dots\}$ is denoted by $\mathbb{Z}$. The set of strictly negative and positive integers are denoted by $\mathbb{Z}_-$ and $\mathbb{Z}_+$, respectively.

Let A be a finite alphabet. We denote by $|A|$ the cardinal of the set A. For integers $m, n \in \mathbb{Z}$ with $m \leqslant n$, we will use the shorthand notation $w_{m:n}$ to denote the string $(w_m, \dots, w_n)$ of symbols in the alphabet A. The length of this string will be denoted by $\ell(w_{m:n}) = n - m + 1$. If $m > n$ we let $w_{m:n}$ denote the empty string λ. For any $j \in \mathbb{N}$, we let A^j denote the set of strings in A having length j, in particular $A^0 = \{\lambda\}$. We also let $A^\star = \cup_{j \geqslant 0} A^j$ denote the set of all finite strings on A and we denote by A^∞ the set of all left-infinite sequences $w_{-\infty:n}$ with symbols in A.

We say that a sequence $s_{j:k}$ is a *suffix* of a sequence $w_{m:n}$ if $\ell(s_{j:k}) \leqslant \ell(w_{m:n})$ and $s_{k-i} = w_{n-i}$ for all $i = 0, \dots, k - j$. This will be denoted as $s_{j:k} \preceq w_{m:n}$. If $\ell(s_{j:k}) < \ell(w_{m:n})$ then we say that s is a proper suffix of w and denote this relation by $s \prec w$. Given a sequence w, the maximal proper suffix of w (obtained bu removing the leftmost symbol) will be denoted by suf(w).

Definition 2.1. A subset $\tau \subset A^\star \cup A^\infty$ is a *tree* if it satisfies the *suffix property*, what means that no $w \in \tau$ is a proper suffix of another $s \in \tau$. If in addition, a tree τ satisfies the *irreducibility property*, which states that no string belonging to τ can be replaced by a proper suffix without violating the suffix property, then it is called *irreducible tree*.

It is easy to see that the set τ can be identified with the set of leaves of a rooted tree with a finite set of labeled branches. Elements of τ will be denoted either as w or as $w_{-k:-1}$ if we want to stress the number of symbols in the string.

Example 2.2. Suppose $A = \{0, 1\}$. Consider the following sets of sequences with

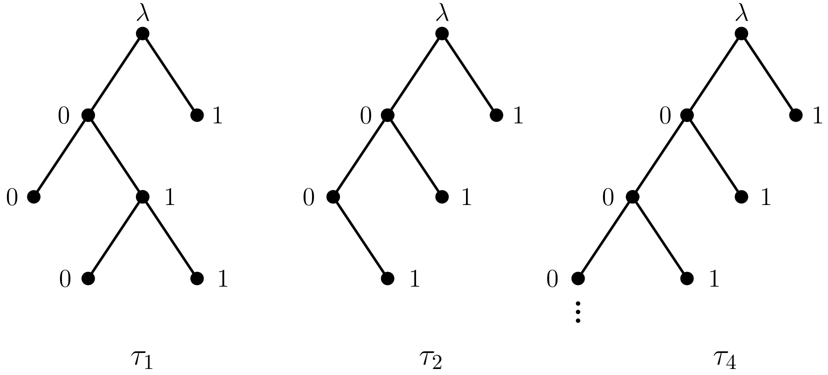

Figure 2.1: Examples of the tree representation of the sets τ_1, τ_2 and τ_4 that satisfy the suffix property in Definition 2.1. The sequences in the set (read from left to right) are read in the tree bottom-up (from leaves to root). The set τ_2 is not irreducible, because we can substitute the sequence 100 by the sequence 00 without violating the suffix property. The set τ_4 is infinite then we represent a truncated version with sequences of length up to three.

symbols in A.

$$\begin{aligned}
\tau_1 &= \{00, 010, 110, 1\} \\
\tau_2 &= \{100, 10, 1\} \\
\tau_3 &= \{000, 00, 100, 10, 1\} \\
\tau_4 &= \{10_{-k:-1} \colon k = 0, \dots\} \cup \{0_{-\infty:-1}\}\,.
\end{aligned}$$

Here, $10_{-k:-1}$ represents the sequence obtained by concatenating a 1 with k 0's. Similarly, the sequence $0_{-\infty:-1}$ is a semi-infinite sequence with all 0's. It can be seen that τ_1 and τ_4 correspond to irreducible trees over A, satisfying all the conditions in Definition 2.1. On the other hand, τ_2 does not satisfy the irreducibility property and τ_3 does not satisfy the suffix property. As τ_1, τ_2 and τ_4 satisfy the suffix property, they can be represented graphically as an (inverted) tree where each sequence is represented by a leaf in the tree, see Figure 2.1.

In the set of all trees over the alphabet A we can define a partial ordering.

Definition 2.3. We will say that $\tau \preceq \tau'$ if for every $v \in \tau'$ there exists $w \in \tau$ such that $w \preceq v$. As usual, whenever $\tau \preceq \tau'$ with $\tau \neq \tau'$ we will write $\tau \prec \tau'$.

Títulos Publicados — 33º Colóquio Brasileiro de Matemática

Geometria Lipschitz das singularidades – *Lev Birbrair e Edvalter Sena*

Combinatória – *Fábio Botler, Maurício Collares, Taísa Martins, Walner Mendonça, Rob Morris e Guilherme Mota*

Códigos Geométricos – *Gilberto Brito de Almeida Filho e Saeed Tafazolian*

Topologia e geometria de 3-variedades – *André Salles de Carvalho e Rafał Marian Siejakowski*

Ciência de Dados: Algoritmos e Aplicações – *Luerbio Faria, Fabiano de Souza Oliveira, Paulo Eustáquio Duarte Pinto e Jayme Luiz Szwarcfiter*

Discovering Euclidean Phenomena in Poncelet Families – *Ronaldo A. Garcia e Dan S. Reznik*

Introdução à geometria e topologia dos sistemas dinâmicos em superfícies e além – *Víctor León e Bruno Scárdua*

Equações diferenciais e modelos epidemiológicos – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

Differential Equation Models in Epidemiology – *Marlon M. López-Flores, Dan Marchesin, Vítor Matos e Stephen Schecter*

A friendly invitation to Fourier analysis on polytopes – *Sinai Robins*

PI-álgebras: uma introdução à PI-teoria – *Rafael Bezerra dos Santos e Ana Cristina Vieira*

First steps into Model Order Reduction – *Alessandro Alla*

The Einstein Constraint Equations – *Rodrigo Avalos e Jorge H. Lira*

Dynamics of Circle Mappings – *Edson de Faria e Pablo Guarino*

Statistical model selection for stochastic systems – *Antonio Galves, Florencia Leonardi e Guilherme Ost*

Transfer Operators in Hyperbolic Dynamics – *Mark F. Demers, Niloofar Kiamari e Carlangelo Liverani*

A Course in Hodge Theory Periods of Algebraic Cycles – *Hossein Movasati e Roberto Villaflor Loyola*

A dynamical system approach for Lane–Emden type problems – *Liliane Maia, Gabrielle Nornberg e Filomena Pacella*

Visualizing Thurston's Geometries – *Tiago Novello, Vinícius da Silva e Luiz Velho*

Scaling Problems, Algorithms and Applications to Computer Science and Statistics – *Rafael Oliveira e Akshay Ramachandran*

An Introduction to Characteristic Classes – *Jean-Paul Brasselet*

impa

Instituto de
Matemática
Pura e Aplicada

ISBN 978-65-89124-29-0

9 786589 124290