UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Haroldo Valentin Ribeiro

Identificação e Modelagem de Padrões em Sistemas Complexos

Maringá, Dezembro de 2012.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

HAROLDO VALENTIN RIBEIRO

IDENTIFICAÇÃO E MODELAGEM DE PADRÕES EM SISTEMAS COMPLEXOS

Tese de Doutorado apresentada ao Departamento de Física da Universidade Estadual de Maringá.

Orientador:
Prof. Dr. Renio dos Santos Mendes - Universidade Estadual de Maringá

Banca Examinadora:
Prof. Dr. Constantino Tsallis - Centro Brasileiro de Pesquisas Físicas
Prof. Dr. Ervin Kaminski Lenzi - Universidade Estadual de Maringá
Prof. Dr. José Soares de Andrade Jr. - Universidade Federal do Ceará
Prof. Dr. Luciano Rodrigues da Silva - Universidade Federal do Rio Grande do Norte
Prof. Dr. Luis Carlos Malacarne - Universidade Estadual de Maringá
Prof. Dr. Mauro Luciano Baesso - Universidade Estadual de Maringá

Maringá, Dezembro de 2012.

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

Ribeiro, Haroldo Valentin
R484i Identificação e modelagem de padrões em sistemas complexos/ Haroldo Valentin Ribeiro. – Maringá, 2012.
153 f., il., tabs.

Orientador: Prof. Dr. Renio dos Santos Mendes.
Tese (doutorado) – Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Física, Programa de Pós-graduação em Física, 2012.

1. Sistemas complexos. 2. Séries temporais. 3. Física estatística. 4. Difusão anômala. 5. Sistemas sociais. I. Mendes, Renio dos Santos, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Física. Programa de Pós-graduação em Física. III. Título.

CDD 22. ED. 530.13
JLM-000733

Resumo

Esta tese está focada na investigação e modelagem de diferentes sistemas complexos. Os problemas aqui apresentados foram analisados através da "lente" da Física e estão amplamente amparados por dados observacionais, experimentais ou simulados. No capítulo 1, investigamos a dinâmica sonora de aglomerações humanas e os resultados mostraram que esses sons não são ruídos triviais, pelo contrário, foram encontradas distribuições não gaussianas, não exponenciais, leis de potência e correlações. Mostramos que um processo auto-regressivo pode descrever a maioria dos resultados empíricos e que é possível distinguir entre sons pacíficos e sons de protesto usando essa análise. No capítulo 2, estudamos os sons musicais. Verificamos que a distribuição das amplitudes pode ser ajustada por uma generalização da gaussiana e que o parâmetro da distribuição fornece informações sobre a qualidade da música. Vimos que existe um acoplamento entre a forma da distribuição e propriedades correlacionais. Analisamos os padrões ordinais nesses sons via entropia e complexidade de permutação e verificamos que esses índices podem ser usados no processo de identificação automática de gêneros musicais. Analisamos, quantitativamente, a evolução de canções populares e encontramos uma tendência de "empobrecimento estatístico" dessas músicas ao longo dos anos. No capítulo 3, estudamos a evolução da vantagem em jogos de xadrez usando uma abordagem difusiva, a qual revelou vários aspectos anômalos e também um processo de aprendizagem populacional dos jogadores. Verificamos, também, que os erros dos jogadores seguem uma distribuição log-normal e que perceber esses erros é fator determinante para vencer a partida. No capítulo 4, analisamos as pontuações dos jogos de críquete como um processo difusivo. Verificamos que o processo é superdifusivo, correlacionado e autossimilar. Vimos também que uma equação de Langevin generalizada reproduz todos esses resultados empíricos. No capítulo 5, investigamos a dinâmica de bolhas em água fervente pela análise de um feixe laser que atravessa o fluido em ebulição. Observamos que existem correlações nessa dinâmica e que a distribuição dos intervalos de retorno é não exponencial. Um modelo minimalista sugere que os principais ingredientes para produzir essa dinâmica (no sinal do laser) são as correlações e a distribuição do tipo lei de potência relacionada ao tempo no qual as bolhas passam através do caminho óptico do laser. Finalmente, no capítulo 6, propusemos uma extensão da técnica de entropia e complexidade de permutação para medir a complexidade de imagens. Aplicamos esse procedimento em superfícies fractais, texturas de cristais líquidos e superfícies de Ising para comprovar a sua utilidade.

Palavras-chave: sistemas complexos, séries temporais, distribuições de probabilidade, física estatística, difusão anômala, sistemas sociais, sons musicais, aprendizado populacional, xadrez, críquete, dinâmica de bolhas, complexidade de imagens.

Abstract

This thesis is focused on the study and modeling of different complex systems. The systems investigated here were analyzed by using the "physics lens" and are all based on observational, experimental or simulated data. In chapter 1, we investigated the soundscape dynamics of human agglomeration where we showed that these noises have a non trivial dynamics with non-Gaussian, non-exponential, power-law distributions and long-range correlations. We also showed that an autoregressive model can reproduce most of the empirical findings and that is possible to distinguish between pacific and violent soundscapes. In chapter 2, we reported studies on musical sounds where the distribution of the sound amplitudes was fitted by a stretched gaussian and the parameter of the distribution gives information about the quality of the music. We also saw that there is a kind of coupling between the shape of the distribution and the long-range correlations. We analyzed the ordinal patterns in these sounds using the complexity-entropy causality plane and we employed a supported vector machine to identify the music genres of our dataset. We further investigate the evolution of these patterns over the years for a set of popular songs where we suggest that the songs are becoming more statistically poor. In chapter 3, we studied the dynamics of the advantage in chess matches by using a diffusive approach which revealed several anomalous features and a population-level learning of the game. We have also verified that the error distribution of players follows a log-normal distribution and that to note the mistakes is very important for wining the match. In chapter 4, we analyzed the scores of the game of cricket through a diffusive approach. We verified that the process is super-diffusive, long-range correlated and self-similar; all these features were modeled using a generalized Langevin equation. In chapter 5, we investigated the bubble dynamics in boiling water thought an experiment in which a laser beam was scattered by bubbles in the boiling fluid. We found that there are long-range correlations in the laser intensity and that the return intervals are exponentially distributed. A simple model suggests that the main ingredients for this non-trivial dynamics are the correlation and power-law distribution related to the time interval in which bubbles passes through the optical path. Finally, in chapter 6, we have proposed an extension of the complexity-entropy causality plane for measuring the complexity of two-dimensional patterns such as images. Our extension was worked out for fractal surfaces, textures of liquid crystals, and Ising surfaces where we proved the usefulness of our extension.

Key words: complex systems, time series, statistical physics, probability distributions, anomalous diffusion, social systems, music sounds, population-level learning, chess, cricket, bubble dynamics, image complexity.

Sumário

Introdução

Esta tese está focada no estudo empírico e teórico de vários sistemas complexos. Como o leitor irá notar, os capítulos que seguem podem ser lidos em qualquer ordem, pois são praticamente independentes. Nos apêndices, apresentamos os métodos estatísticos ou de análise de séries temporais usados aqui, na tentativa de tornar a leitura mais agradável e objetiva.

Ainda que certamente segmentados, os estudos que serão apresentados podem ser unificados dentro do que vem se chamando de Física dos Sistemas Complexos. Mas, como veremos, esses sistemas não se limitam, de modo algum, aos problemas tradicionais da Física. Trata-se de um campo recente e emergente que vem ganhando espaço nas revistas mais tradicionais de Física e que preenche considerável parte do índice remissivo de revistas multidisciplinares.

É fato curioso que, apesar de toda essa popularidade, não exista uma definição precisa ou razoável do que venha a ser um sistema complexo. Como podemos observar nas palavras de Gell-Mann [1], até mesmo o conceito de complexidade é bastante vago:

> *Probably no single concept of complexity can adequately capture our intuitive notions of what the word ought to mean. Several different kinds of complexity may have to be defined, some of which may not yet have been conceived.*

Alguns autores chegam ao ponto de dizer que [2]:

> *Complex Systems are like beauty: You know it when you see it.*

Outros enumeram várias propriedades comuns a esses sistemas, tais como [3]:

- são constituídos de um grande número de agentes interagentes;
- as interações são, em geral, não-locais e/ou não-lineares;
- existe um comportamento coletivo, auto-organizado, o qual é difícil de antecipar a partir do conhecimento da dinâmica individual dos agentes;
- esse comportamento coletivo não resulta da existência de um controle central.

Devemos notar que algumas dessas ideias já apareciam nos trabalhos de Aristóteles [4], 350 anos antes de Cristo:

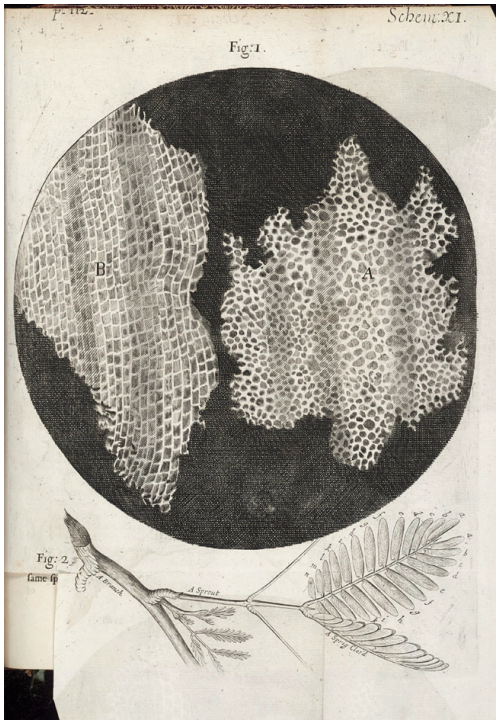
> *In the case of all things which have several parts and in which the totality is not, as it were, a mere heap, but the whole is something beside the parts...*

O que de fato parece ser consenso é a onipresença desses estudos na literatura e sua importância para a ciência e para a sociedade em geral. Fato que, no final do século passado, foi previsto por Hawking em uma entrevista para um jornal [5]:

I think the next century will be the century of complexity.

Além do que assinalamos anteriormente, existem outros autores que acreditam que as ferramentas usadas e criadas nas investigações de sistemas complexos são (e serão) as principais armas para dar algum sentido às enormes quantidades de informações que a ciência e a iniciativa pública ou privada vem acumulando.

A importância de se obter novas ferramentas para investigar a natureza é clara. Tomemos como exemplo o olho humano. Trata-se de uma ferramenta altamente sofisticada, composta por $\sim$2 milhões de partes, $\sim$130 milhões de fotorreceptores e que gera mais de $3{,}5$ gigabytes de informação por hora. A olho nu, podemos experimentar grande parte da Mecânica Clássica, como possivelmente fez Galileu Galilei [6], por exemplo, ao mostrar na Torre di Pisa que a velocidade de queda não depende da massa. Entretanto, o olho tem suas limitações. Por exemplo, só enxergamos um pequeno intervalo do espectro eletromagnético ($\sim$400 a $\sim$700 nanometros), não conseguimos observar nenhum fenômeno que ocorra em um intervalo de tempo menor que $\sim$17 milisegundos, e não vemos nenhum objeto de tamanho menor que $\sim$100 micrometros. A remoção dessas limitações pelo uso de equipamentos gerou uma infinidade de novas possibilidades para a investigação da natureza. O microscópio é um bom exemplo. Ele permite que vejamos objetos bem menores, como o fez Robert Hooke (o mesmo da Lei de Hooke). Em seu livro de 1667, *Micrographia* [7], ele relata suas observações a respeito de um pedaço de cortiça:



I Took a good clear piece of Cork, and with a Pen-knife sharpen'd as keen as a Razor, I cut a piece of it off, and thereby left the surface of it exceeding smooth, then examining it very diligently with a Microscope, me thought I could perceive it to appear little porous; but I could not so plainly distinguish then, as to be sure that they were pores, much less what Figure they were of: But judging from lightness and yielding quality Cork, that certainly the texture could not be so curious, but that possibly, if I could use some further diligence, I might find it to be discernible with a Microscope, I with the same sharp Pen-knife, cut off from the former smooth surface an exceeding thin piece of it, and placing it on a black object Plate, because it was it self a white body, and casting the light on it with a deep plano-convex Glass, I could exceeding plainly perceive it to be all perforated and porous, much like a

Honey-comb, but that the pores of it were not regular; yes it was not unlike a Honey-comb in these particulars.

O texto acima e também a figura desenhada a mão por Hooke mostram a descrição meticulosa do que ele mesmo chamou de célula, nome que está no título da seção da qual extraímos o fragmento de texto: *Of the schematisme or texture of cork, and of the cells and pores of some other such frothy bodies*. Possivelmente, Hooke não imaginou que a sua descrição do que hoje chamamos de célula eucariótica geraria uma revolução na biologia, que somente foi possível por meio da superação de uma limitação do olho humano, criando uma nova ferramenta para ver o mundo.

Algo similar também ocorre nos estudos de sistemas complexos. Hoje, muitas vezes, temos acesso a tanta informação que se torna impraticável identificar padrões por meio da examinação bruta dos dados. Uma consequência direta dessa dificuldade ocorre, por exemplo, quando decisões são tomadas sem levar em conta fatos empíricos bem estabelecidos, o que certamente prejudica a qualidade das decisões. Nessa direção, é interessante notar que a abordagem de um sistema complexo não é (ao menos em geral) única. Não existe um conjunto de variáveis análogo às posições e momentos da Mecânica. Pelo contrário, é tarefa do pesquisador determinar quais variáveis investigar e, naturalmente, quais dessas investigações são praticáveis.

Talvez a melhor maneira de compreendermos uma investigação sobre sistemas complexos seja via um exemplo. Para isso, imaginemos que queiramos compreender melhor as relações científicas do Departamento de Física da Universidade Estadual de Maringá. Não podemos pontuar que elas sejam simples. Existem vários pesquisadores e grupos de pesquisa que se relacionam por meio de colaborações em artigos e também competem por recursos financeiros. Além disso, as relações entre os pesquisadores não são triviais. Pode haver certas preferências por colaboração oriundas de uma relação de maior proximidade científica ou mesmo de afinidade pessoal.

Pois bem, como investigar esse cenário? Uma das repostas é olhar para a rede de colaboração científica dos pesquisadores do departamento. Nessa rede, consideramos que cada vértice é um pesquisador e que as ligações entre os vértices ocorrem sempre que os pesquisadores forem autores em um mesmo artigo. Aqui, por simplicidade, consideramos todos os artigos independentes do ano de publicação; entretanto, essa rede poderia ser construída ano a ano de modo a elucidar aspectos dinâmicos das colaborações no departamento. A Figura 1 mostra a rede obtida, na qual podemos observar a existência de vários aglomerados pequenos e isolados e de um grande aglomerado. Os aglomerados isolados são pequenos grupos de pesquisadores que estiveram no departamento logo após a sua criação e também contêm pesquisadores do grupo de Ensino de Física que não são co-autores em artigos com os demais pesquisadores do departamento. Em geral, a parte mais interessante é o grande aglomerado no qual estão quase todos os pesquisadores do departamento. A Figura 2 mostra essa parte da rede, muitas vezes denominada componente gigante ou principal.

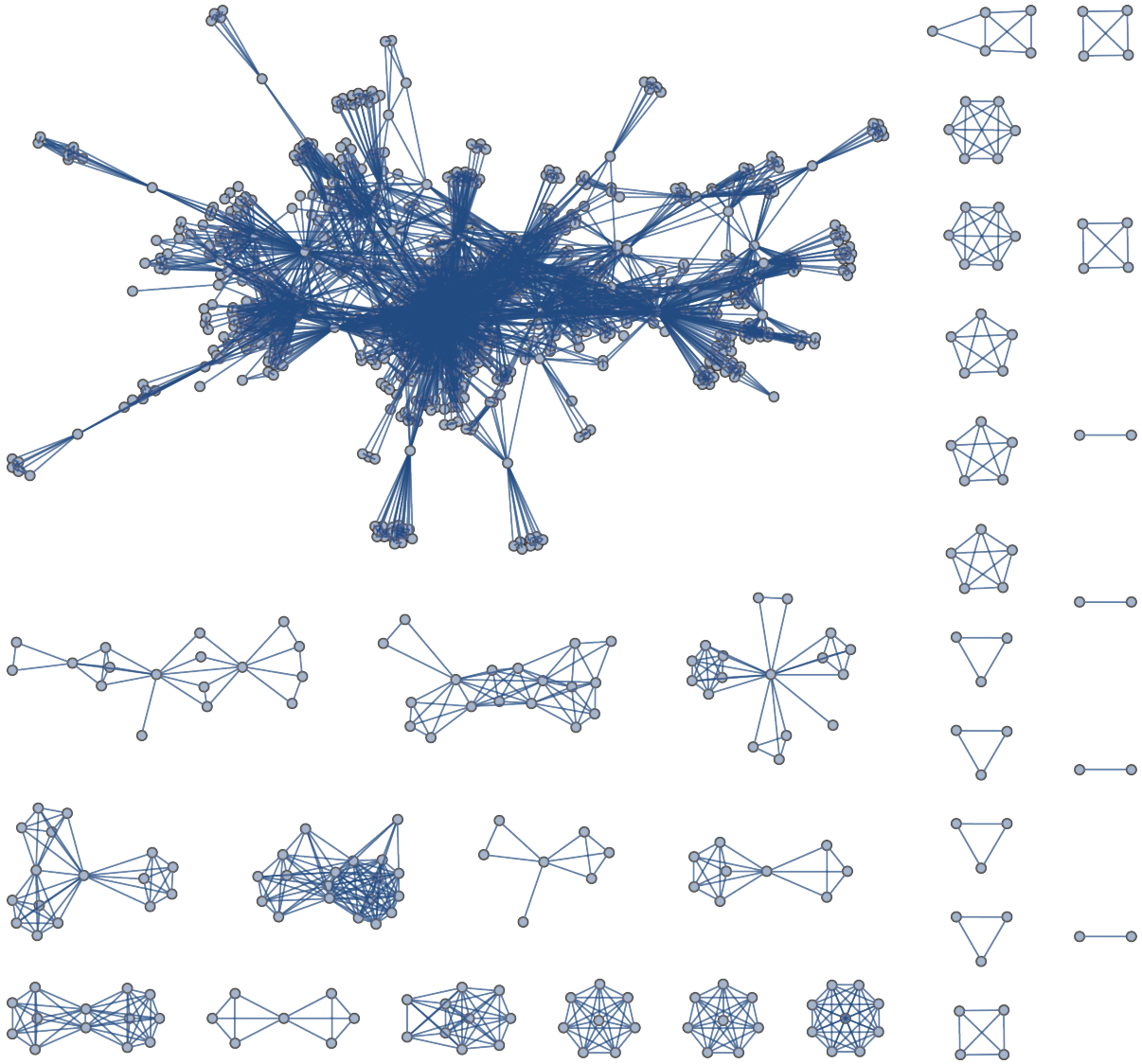


Figura 1: Rede das colaborações científicas do Departamento Física da Universidade Estadual de Maringá. Aqui, cada vértice representa um pesquisador e as ligações entre eles ocorrem sempre que os pesquisadores participam juntos em um mesmo artigo. Os dados foram obtidos no Scopus™ (`www.scopus.com`) buscando-se por todos os artigos que continham no endereço as palavras "Física" e "Maringá" e não continham a palavra "Educação". Além disso, todos os nomes foram verificados manualmente de modo a agrupar autores idênticos e a evitar ambiguidades. Observe a presença de um grande aglomerado no qual encontram-se a maioria dos pesquisadores e também a presença de pequenos aglomerados, os quais representam pesquisadores que colaboram entre si mas não colaboram/colaboraram com os demais.

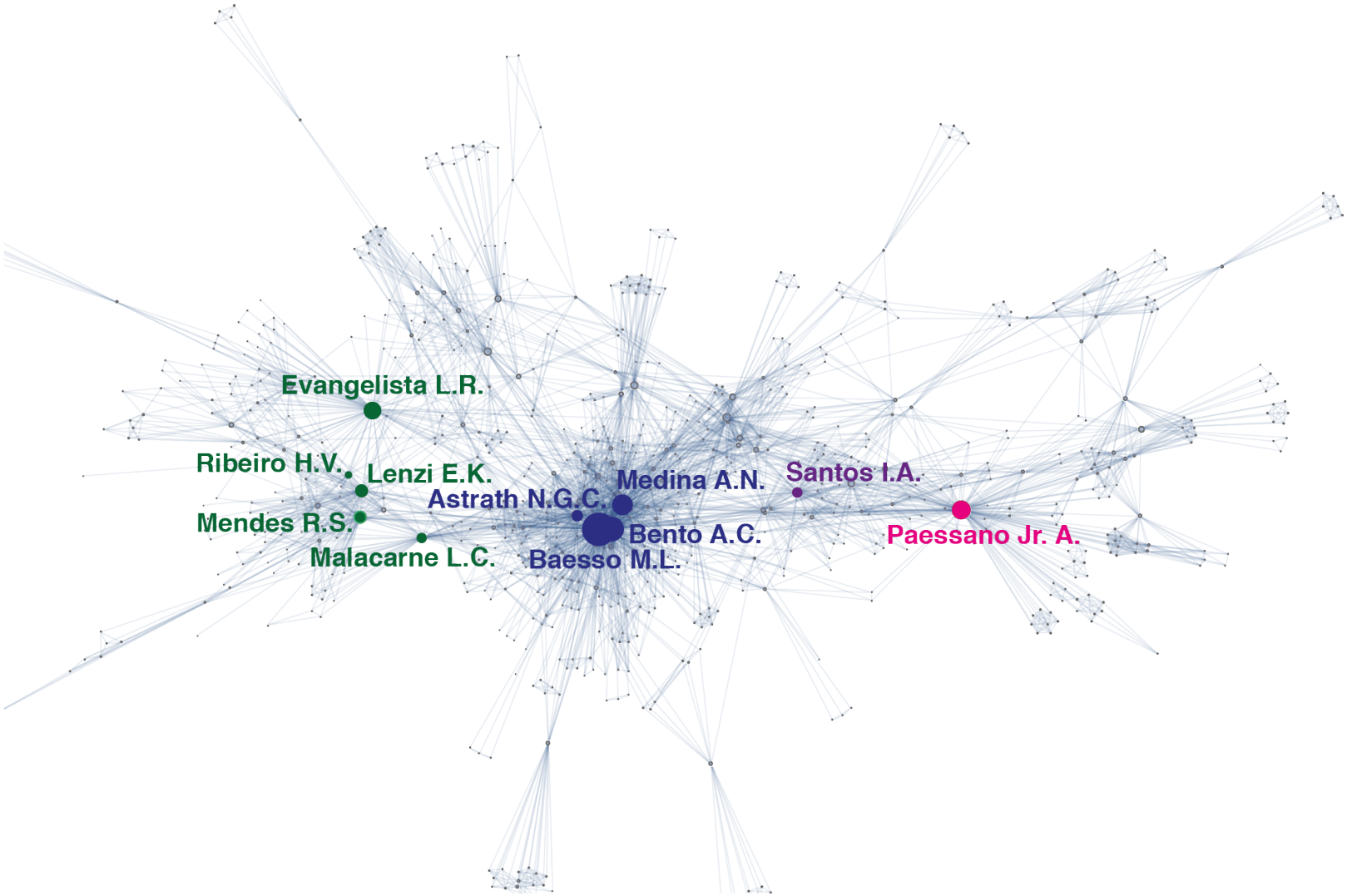


Figura 2: Componente gigante da rede da Figura 1. O tamanho dos vértices é proporcional ao valor do *pagerank*, que por sua vez mede a probabilidade de um caminhante aleatório passar por aquele vértice, saindo de uma posição aleatória da rede. Os nomes mostrados (com exceção do meu) representam os *top-10* pesquisadores ordenados pelo *pagerank* e as cores representam diferentes grupos de pesquisa.

A maneira como os vértices foram arranjados nessa figura também pode definir uma espécie de "distância para colaboração". O arranjo foi feito seguindo os chamados algoritmos baseados em forças. Nesse caso em particular, consideramos que as ligações representam molas como na Lei de Hooke e que os vértices são como partículas carregadas. Esse sistema todo é simulado integrando as equações de movimento, de modo que as forças aplicadas movem os vértices para perto ou longe até que o sistema como um todo entre em equilíbrio mecânico. Assim, podemos imaginar que a distância espacial dos vértices está de fato relacionada à probabilidade de haver colaboração entre dois autores (aqueles que conhecem o departamento devem concordar com essa análise).

Além dessa análise semiquantitativa, é possível extrair muitas outras informações estudando-se as propriedades topológicas da rede. Podemos, por exemplo, investigar a influência dos pesquisadores na rede de colaboração usando uma análise parecida com a que o Google™ faz para ordenar itens mostrados em suas pesquisas. O procedimento que foi proposto por L. Page [8], co-fundador do Google, consiste em calcular a probabilidade de passar por um determinado vértice em uma caminhada aleatória que começa e termina em outros vértices escolhidos aleatoriamente. Essa probabilidade define o índice chamado de *pagerank*, e na Figura 2, o tamanho dos vértices foi escolhido ser proporcional a esse índice. A Figura 3 mostra os pesquisadores com os dez valores mais altos do *pagerank*. Todos esses pesquisadores participam do programa de pós-graduação e são, atualmente, bolsistas produtividade do CNPq.

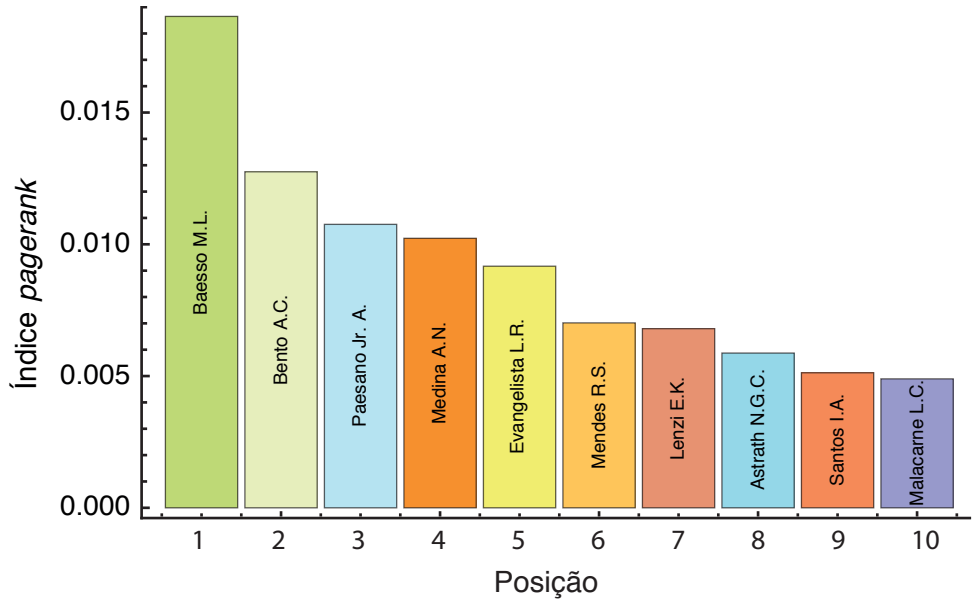


Figura 3: Medindo a influência dos pesquisadores do departamento na rede dos colaboradores via algoritmo *pagerank*. Esse gráfico de barra mostra o valor do *pagerank* para os pesquisadores com os 10 maiores valores do índice.

Naturalmente, essa análise simplificada pode ser refinada de muitas maneiras. Uma delas seria usar uma "rede ponderada", na qual cada ligação passaria a ter um peso proporcional ao número de trabalhos publicados entre aqueles dois pesquisadores. Outras possibilidades incluem estudar aspectos individuais dos pesquisadores, como as séries temporais do número de artigos

por ano e suas distribuições de probabilidade. De fato, como veremos ao longo deste texto, nossas análises estarão focadas nessas outras possibilidades e não no estudo de redes complexas.

Para ilustrar alguns desses outros aspectos, analisamos os padrões ordinais nas séries temporais do número de artigos por ano para cada um dos *top*-10 pesquisadores da Figura 3. Nesta análise, usamos as técnicas de entropia e complexidade de permutação, as quais discutiremos em detalhes nos capítulos 2 e 6. Aqui, nos limitaremos a mencionar que valores de entropia próximos de um (1) indicam padrões ordinais aleatórios, enquanto valores próximos de zero indicam padrões mais regulares; o inverso ocorre para a complexidade da série temporal. A Figura 4 mostra o diagrama da complexidade versus entropia. Notamos que os pesquisadores ocupam diferentes localizações nesse plano, o que indica que eles possuem diferentes graus de regularidade no número de publicações ano a ano. Por exemplo, aqueles pesquisadores (Medina, Evangelista, Bento) mais próximos do vértice $(1, 0)$ possuem uma aleatoriedade maior do que aqueles na região mais central do plano complexidade-entropia (Astrath e Malacarne, por exemplo).

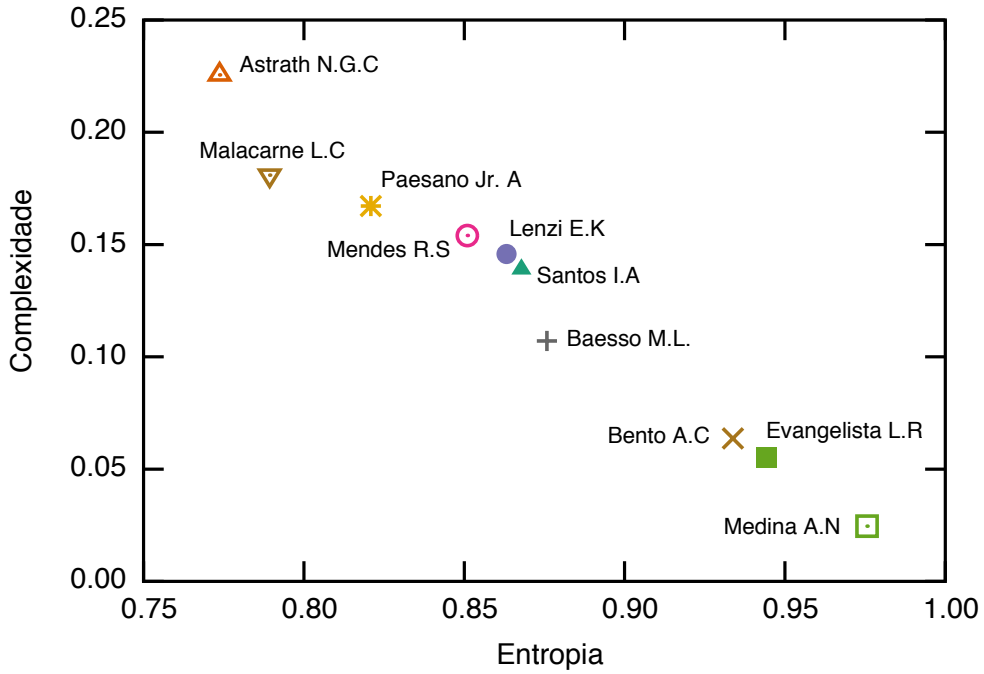


Figura 4: Plano complexidade-entropia para as séries temporais do número de artigos por ano dos *top*-10 pesquisadores ordenados pelo *pagerank*.

Outro tipo de análise que será comum ao longo desta tese é a de distribuições de probabilidade. Por exemplo, na Figura 5 mostramos a distribuição de probabilidade acumulada do número de artigos publicados por ano para cada um dos 10 pesquisadores da Figura 3. Ainda que os dados sejam poucos (desde 1996), é possível encontrar algumas regularidades nos perfis dessas distribuições. Em particular, verificamos que elas podem ser aproximadas pela distribuição gamma, $p(x) \sim x^{\alpha-1}\exp(-x/\beta)$. A distribuição gamma é uma generalização da distribuição exponencial ($\alpha = 1$) e representa a distribuição da soma de α variáveis aleatórias exponenciais de média $1/\beta$ (quando α é inteiro). Assim, podemos imaginar que os pesquisadores possuem vários mecanismos para produzir seus trabalhos, os quais podem representar, por exemplo, diferentes

colaborações ou linhas de pesquisa. Vale notar que a forma da distribuição do número de artigos e também das citações ainda representa um problema em aberto para a comunidade que estuda dados bibliográficos (*science of science*) [9, 10, 11] e que nossos dados não são suficientes para responder tais questões de maneira objetiva.

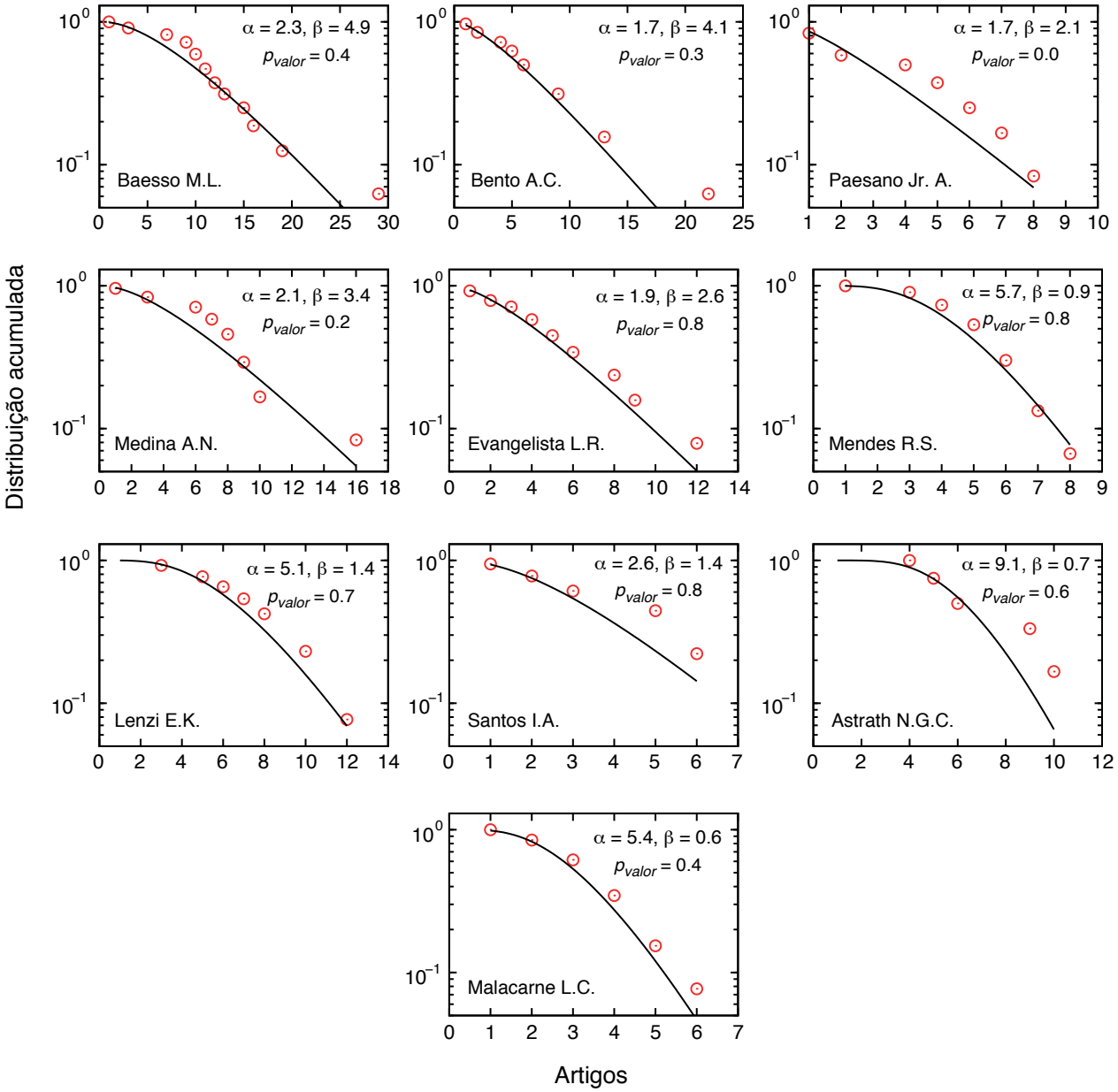


Figura 5: Distribuição acumulada do número de artigos publicados por ano para os *top*-10 pesquisadores ordenados pelo *pagerank*. As linhas contínuas são distribuições gamma, $p(x) \sim x^{\alpha-1}\exp(-x/\beta)$, e os valores dos parâmetros estão indicados nos gráficos. O p_{valor} é o valor p para o teste de hipótese de Kolmogorov-Smirnov (Apendice B.1). Notamos que, com exceção do pesquisador Paesano, não podemos rejeitar a hipótese de que as distribuições sejam distribuições gamma.

Poderíamos prosseguir com outras investigações relacionadas à produção do Departamento de Física ou mesmo estender nossas análises para toda a Universidade Estadual de Maringá. Porém, aqui, só estamos interessados em dar um exemplo de sistema complexo e mostrar o quão instigante pode ser a sua investigação.

O restante desta tese está estruturado de maneira semelhante ao problema anterior: iniciaremos pela apresentação do sistema; definiremos as variáveis empíricas que estudaremos (geralmente séries temporais), faremos as análises desses dados visando a encontrar padrões ou leis de formação e, quando possível, proporemos modelos que ajudem a compreender os mecanismos do sistema. De maneira mais específica, no capítulo 1, estudaremos os sons de aglomerações humanas [12] e veremos que é possível diferenciar sons pacíficos de sons de protesto [13]. No capítulo 2, investigaremos padrões nos sons musicais. Veremos que as distribuições das amplitudes sonoras seguem uma forma universal e que existe um acoplamento entre propriedades de correlação e a forma da distribuição [14]. Investigaremos também os padrões ordinais desses sons visando a classificar as músicas e os gêneros musicais [15] e estudaremos, quantitativamente, a evolução musical das canções mais populares, sugerindo um empobrecimento estatístico das músicas [16]. No capítulo 3, investigaremos a dinâmica da vantagem nos jogos de xadrez, revelando um processo de aprendizado, em nível populacional, dos jogadores [17]. Discutiremos também a dinâmica dos erros dos jogadores, apontando a importância de percebermos os erros e suas influências sobre o resultado das partidas [18]. No capítulo 4, a dinâmica da pontuação dos jogos de críquete será analisada do ponto de vista de um processo difusivo [19], no qual superdifusão, correlações de longo alcance e autossimilaridade serão encontradas e modeladas por uma equação de Langevin generalizada. No capítulo 5, investigaremos a dinâmica de bolhas em água fervente usando um experimento no qual um feixe laser atravessa o fluido em ebulição. Veremos que essa dinâmica possui correlações e empregaremos um modelo simples para descrever os resultados experimentais [20]. No capítulo 6, discutiremos um procedimento para medir a complexidade de imagens baseado na entropia e complexidade de permutação [21]. Finalmente, apresentaremos uma visão geral dos problemas estudados aqui.

Além dos trabalhos que serão aqui apresentados, nos quais a participação do autor foi central, gostaríamos de mencionar outros trabalhos nos quais o autor foi (em geral) apenas coadjuvante, mas que muito contribuíram para sua formação como pesquisador. São eles principalmente sobre difusão anômala [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32] e também relacionados a sistemas complexos [33, 34, 35].

Capítulo 1

Dinâmica sonora de aglomerações humanas e um sensor social

Neste capítulo, apresentaremos um estudo empírico sobre a dinâmica sonora de aglomerações humanas [12, 13], uma situação bastante comum na vida em sociedade. Contra-intuitivamente, veremos que essa dinâmica não é trivial e que esse "ruído" está longe de ser branco. Em particular, mostraremos que a distribuição das amplitudes não é gaussiana, que existem correlações temporais de longo alcance nas intensidades sonoras, que a distribuição dos intervalos de retorno não é exponencial e que a distribuição normalizada da variância local das intensidades apresenta um comportamento assintótico tipo lei de potência. Todas essas características mostraram-se em bom acordo com um modelo auto-regressivo GARGH. Veremos, ainda, que é possível distinguir sons de aglomerações humanas em situações pacíficas e violentas viabilizando a construção de um sensor social baseado em padrões sonoros.

1.1 Apresentação dos dados

É cada vez mais comum investigações de sistemas sociais por parte dos físicos, até mesmo o termo *sociophysics* [36] surgiu nesse contexto. Naturalmente, os constituintes básicos dos sistemas sociais são seres humanos, os quais são bem conhecidos por apresentarem uma dinâmica muito mais complicada que a de sistemas físicos interagentes. Além disso, aspectos individuais relacionados aos agentes sociais podem não estar disponíveis e as interações entre eles, em geral, não são bem definidas ou são muito complexas. Ainda assim, vários padrões em atividades humanas têm sido relatados em, por exemplo, eleições [35, 37, 38], colaboração entre atores [39] e também entre cientistas [40], mensagens de celulares [41], cartas [42, 43] ou e-mails [43, 44], viagens [45, 46] e competições esportivas [47] .

Aqui, investigaremos uma situação muito comum relacionada às atividades coletivas humanas: as aglomerações de pessoas. Esse tipo de situação aparece em vários lugares e por várias razões. Pode ocorrer, por exemplo, durante refeições em um restaurante e festas ou reuniões de trabalho. Em muitas dessas situações, umas das características mais notórias é o som produzido por essas

aglomerações.

Com a finalidade de estudar esses sons, fizemos uma série de gravações aqui mesmo em nossa universidade. O lugar escolhido foi uma cantina próxima ao departamento, que é ponto de encontro de estudantes nos intervalos entre as aulas. O equipamento de gravação utilizado foi um microfone de alta qualidade usado para gravação de corais (*Shure Microflex MX202W/N*) e uma mesa de som (*Yamaha MG102c*). Uma taxa de amostragem de 44,1 kHz foi usada para cobrir todo o espectro audível humano (aproximadamente entre 20 Hz e 20 kHz). O microfone foi posicionado no alto da parte central da aglomeração em uma posição bastante discreta, de modo a evitar que as pessoas soubessem que seus diálogos estavam sendo gravados. Realizamos um total de 16 gravações durante um período de 9 dias, sendo que o número de indivíduos no lugar variou aproximadamente de 100 a 200 pessoas. Cada gravação durou aproximadamente 10 minutos e durante esse tempo o número de pessoas foi aproximadamente constante. Nas análises que fizemos, essas diferentes gravações não alteram os resultados estatísticos que passaremos a apresentar, de modo que, por simplicidade, mostraremos explicitamente os resultados referentes a 3 gravações e reportaremos o desvio padrão de nossas quantidades estatísticas.

As variáveis utilizadas para investigar esse sistema são as amplitudes sonoras em função do tempo $A(t)$ e as intensidades sonoras $I(t) \propto A(t)^2$. Sendo a variável t discreta, essas duas variáveis são, portanto, séries temporais. Além disso, normalizaremos essas duas quantidades para que tenham desvio padrão unitário. A Figura 1.1 mostra um exemplo típico para essas séries temporais. Notamos nesses gráficos a existência de valores extremos que desviam em mais de 8 desvios padrões no caso das amplitudes e em mais de 20 desvios padrões nas intensidades.

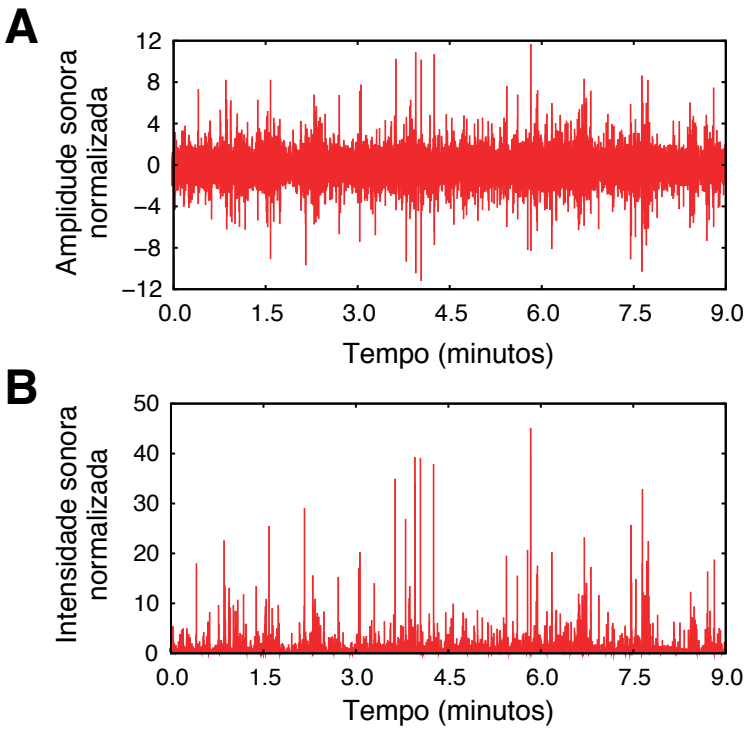


Figura 1.1: Um sinal sonoro típico das gravações que realizamos. Em **(A)**, mostramos a amplitude sonora normalizada e em **(B)** a intensidade sonora normalizada. Aqui, consideramos que a intensidade sonora é proporcional ao quadrado da amplitude.

1.2 Análise estatística

Uma das maneiras mais diretas para iniciar a caracterização das amplitudes sonoras $A(t)$ é pelo cálculo da sua distribuição de probabilidade, como mostra a Figura 1.2 para três gravações típicas. As demais gravações mostram um perfil bastante similar, e o teste de Kolmogorov-Smirnov para duas amostras (Apêndice B.1) não pode rejeitar a hipótese de que todas as distribuições sejam idênticas. Além disso, essa distribuição claramente difere da distribuição gaussiana, particularmente por apresentar um decaimento muito mais lento em suas caudas.

Esse comportamento de caudas longas reflete a existência dos eventos extremos que observamos na Figura 1.1. Mais do que isso, o fato da distribuição não ser gaussiana faz surgir a hipótese de que hajam correlações de longo alcance nessa dinâmica. Podemos imaginar que a amplitude sonora gravada é a composição das amplitudes sonoras dos indivíduos, e se essas amplitudes fossem como números aleatórios não correlacionados (e com segundo momento finito), o Teorema do Limite Central conduziria a uma distribuição gaussiana para as amplitudes. Entretanto, verificamos empiricamente que a distribuição mostrada na Figura 1.2, claramente, não é uma gaussiana.

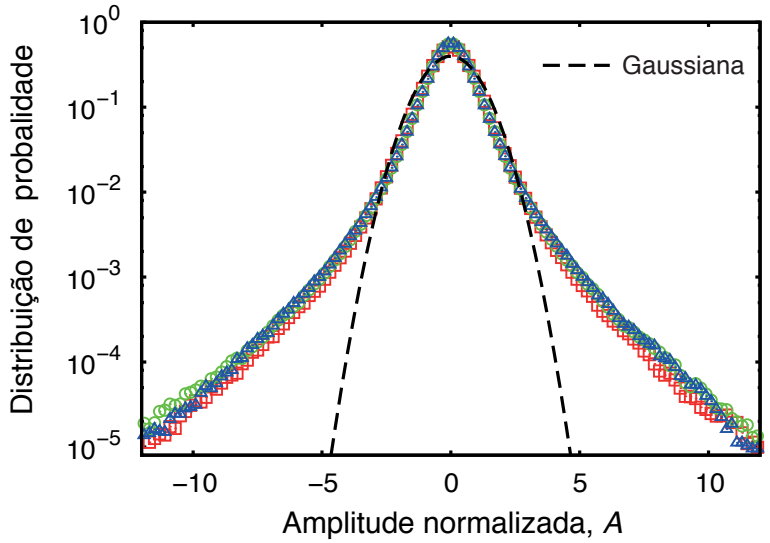


Figura 1.2: Distribuição de probabilidade das amplitudes sonoras normalizadas $A(t)$ para três gravações típicas (círculos, quadrados e triângulos) em comparação com a distribuição gaussiana de média zero e variância unitária (linha tracejada). Observemos que as três gravações apresentam praticamente o mesmo comportamento, o qual é claramente não gaussiano (veja também o Apêndice D).

Para investigar se existem ou não correlações de longo alcance nessa dinâmica, empregamos uma técnica chamada *detrended fluctuation analysis* (DFA, Apêndice A.3). DFA consiste no cálculo da função flutuação $F(n)$, que mede as flutuações da série integrada após a remoção de uma tendência polinomial para uma dada escala n. A função flutuação $F(n)$ segue uma lei de potência, *i.e.*, $F(n) \sim n^h$ com $h \neq 0{,}5$ quando a série possui correlação de longo alcance e o expoente h é o chamado expoente de Hurst. Na Figura 1.3, mostramos os resultados do DFA aplicado para a série das intensidades sonoras $I(t)$. Notamos que $h \approx 0{,}88$, o que mostra

a existência de correlações temporais de longo alcance nessa dinâmica sonora. Ademais, o valor de $h > 0{,}5$ revela a presença de um comportamento persistente, *i.e.*, a probabilidade de que grandes valores de $I(t)$ sejam seguidos por valores ainda maiores é muito maior do que em um processo aleatório. Qualitativamente, esse comportamento pode estar relacionado ao fato de que as pessoas querem ser ouvidas, e se seus vizinhos estão falando alto, é preciso aumentar ainda mais a intensidade da conversa.

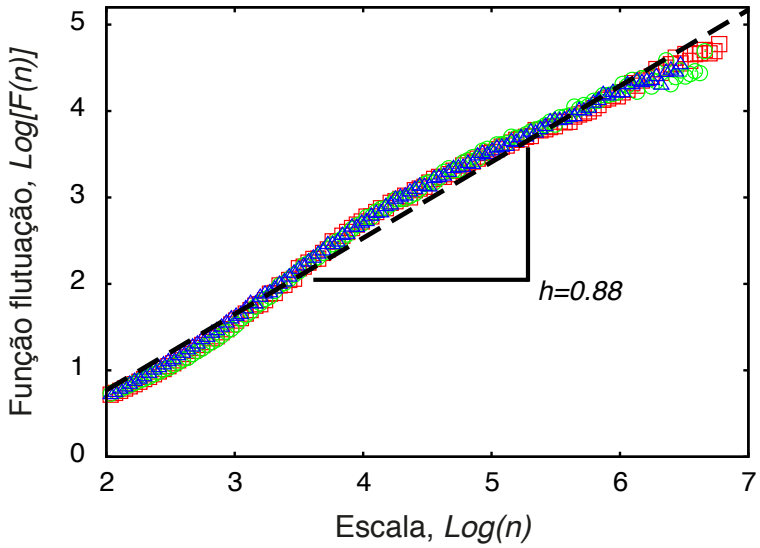


Figura 1.3: Análise DFA para as séries das intensidades sonoras $I(t)$. Nesse gráfico, mostramos a função flutuação $F(n)$ em função da escala temporal n, e n é medido em $1/44{,}1$ k segundos. Notamos que as três amostras (círculos, quadrados e triângulos) apresentam um comportamento muito semelhante, o qual é aproximado por uma lei de potência $F(n) \sim n^h$, com $h = 0{,}88 \pm 0{,}01$.

Em adição ao DFA, podemos investigar também a dinâmica dos eventos extremos nas séries das intensidades (Figura 1.1B) por meio do cálculo dos chamados intervalos de retorno. Esses intervalos temporais são obtidos considerando-se um valor limiar q e armazenando-se todos os tempos iniciais t_i para os quais a intensidade normalizada excede o valor q. Os intervalos de retorno τ_i são as diferenças entre dois valores de tempo consecutivos, *i.e.*, $\tau_i = t_{i+1} - t_i$. Para o caso de variáveis aleatórias gaussianas e sem correlação (ou fracamente correlacionadas), τ_i deve seguir a distribuição exponencial $p(\tau) \sim e^{-\tau/\bar{\tau}_q}$, na qual $\bar{\tau}_q$ é o valor médio de τ_i considerando o limiar q. Dessa maneira, desvios da distribuição exponencial são esperados, pois nosso processo apresenta correlações de longo alcance. De fato, alguns resultados empíricos têm mostrado que na presença de correlações de longo alcance as distribuições de τ_i são bem ajustadas por exponenciais *stretched* [48, 49, 50]

$$p(\tau) \sim e^{-(\tau/\bar{\tau}_q)^\gamma} \tag{1.1}$$

ou pelas distribuições de Weilbull [51]

$$p(\tau) \sim (\tau/\bar{\tau}_q)^{\gamma-1}\, e^{-(\tau/\bar{\tau}_q)^\gamma}\,, \tag{1.2}$$

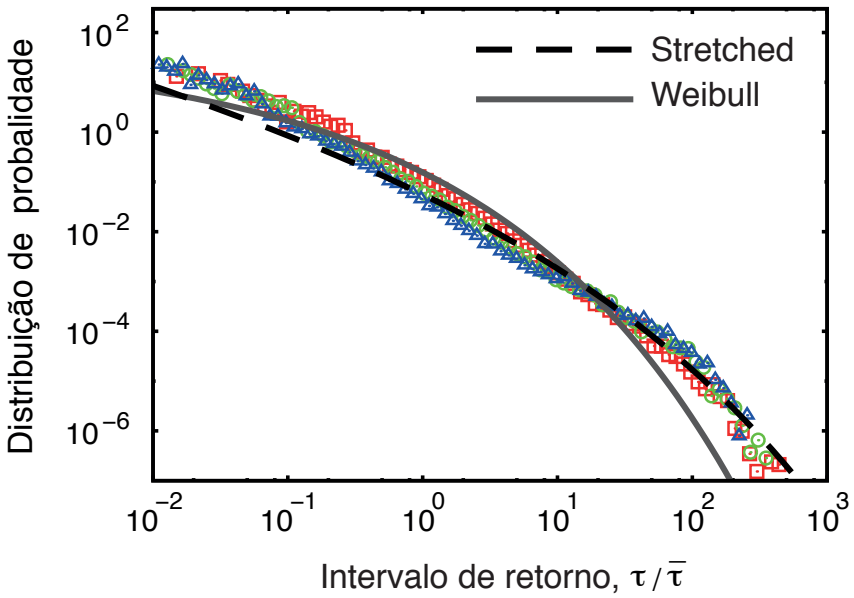


Figura 1.4: Distribuição dos intervalos de retorno escalados $\tau_i/\bar{\tau}$ considerando as intensidades sonoras da Figura 1.1B e três valores limiares: $q = 1$ (quadrados), $q = 2$ (círculos) e $q = 3$ (triângulos). A linha tracejada é a distribuição exponencial *stretched* da equação 1.1 e a linha contínua é a distribuição de Weilbull da equação 1.2, ambas com $\gamma = 2(1-h) \approx 0{,}24$. As demais amostras apresentam um comportamento muito semelhante (veja também o Apêndice D).

em que γ é o expoente da função de autocorrelação do processo que gera os τ_i. Em nosso caso, $\langle I(t)I(t+\tau)\rangle \sim \tau^{\gamma}$. Vale notar que, tanto no caso não correlacionado quanto nos casos com correlação, a distribuição é dependente do valor limiar q. Entretanto, é fácil mostrar que podemos eliminar essa dependência se empregarmos a variável escala $\tau_i/\bar{\tau}_q$. Além disso, o expoente de Hurst h e o expoente de correlação γ estão relacionados via $\gamma = 2(1-h) \approx 0{,}24$ (Apêndice A.2). Essas condições, em conjunto com a normalização e a média unitária da distribuição de τ_i, nos deixam sem nenhum parâmetro livre.

A Figura 1.4 mostra a distribuição dos intervalos de retorno para as intensidades da Figura 1.1B, considerando três valores de q, e a variável escalada em comparação com as distribuições das equações 1.1 e 1.2. Notamos um bom colapso das distribuições com o uso da variável escalada e um acordo não muito bom com as distribuições teóricas. Essas distribuições ajustam bem os intervalos de retorno obtidos de um processo correlacionado e com distribuições gaussianas (apresentamos uma análise mais detalhada de intervalos de retorno na seção 5.2). Como $I(t)$ não é gaussiano, esses desvios são esperados.

Outra possibilidade é investigar os "estouros" sonoros que observamos na Figura 1.1A. Para isso, podemos calcular a volatilidade das amplitudes normalizadas. Essa nova série pode ser definida como o desvio padrão local de $A(t)$ estimado sobre uma janela temporal móvel de tamanho $w = n\Delta t$, *i.e.*,

$$v_w^2(t) = \frac{1}{n-1}\sum_{t'=t}^{t+n-1}\left(A(t') - \langle A(t)\rangle_w\right)^2, \tag{1.3}$$

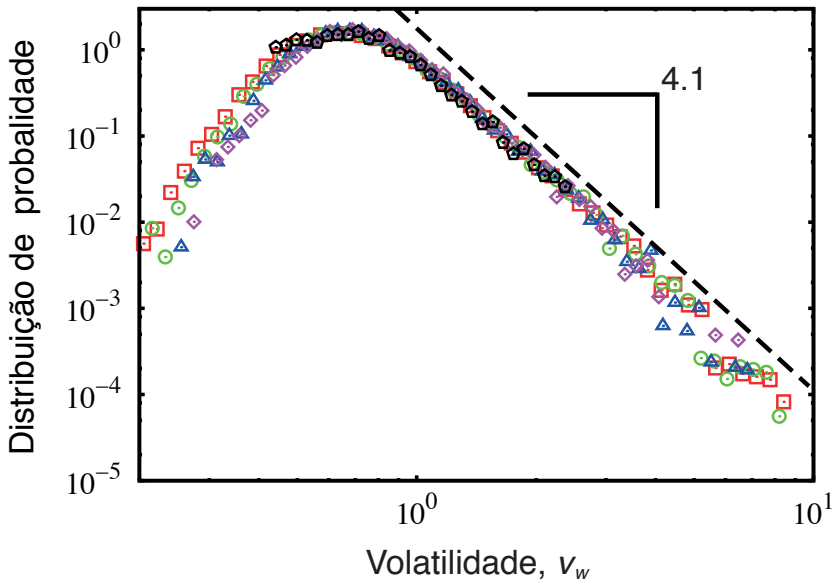


Figura 1.5: Distribuição da volatilidade v_w definida na equação 1.3 e calculada para as amplitudes da Figura 1.1A. Aqui, consideramos cinco janelas: $w = 1$ (quadrados), $w = 2$ (círculos), $w = 5$ (triângulos), $w = 10$ (losangos) e $w = 100$ (pentágonos), sendo w em unidades de centésimos de segundo. A linha tracejada é uma lei de potência com $p(v) \propto v^{-4,1}$ (veja também o Apêndice D).

em que

$$\langle A(t)\rangle_w = \frac{1}{n}\sum_{t'=t}^{t+n-1} A(t')$$

é o valor médio das amplitudes dentro da janela, n é um inteiro e Δt é o inverso da taxa de amostragem (aqui igual a $44{,}1$ kHz). Essa nova série temporal contém informação sobre a dinâmica coletiva em uma dada escala de tempo w. Se a série das amplitudes sonoras fosse estacionária, os valores da variância $v_w^2(t)$ seriam aproximadamente iguais e a distribuição de v_w deveria ser uma delta de Dirac (ou uma distribuição de cauda muito curta). Entretanto, a distribuição de v_w apresenta uma cauda aproximadamente descrita por uma lei de potência $p(v) \propto v^{-\eta}$, com expoente $\eta = 4{,}1$, como mostra a Figura 1.5. Notamos também que as distribuições são aproximadamente independentes do tamanho da janela w no intervalo de $1/100$ a 1 segundo. Como veremos na próxima seção, esse comportamento não estacionário será um aspecto-chave para a modelagem desse sistema.

1.3 Modelando via processos auto-regressivos

Nosso ponto de partida, na tentativa de propor um modelo simples para os comportamento dos dados empíricos, é o aspecto não estacionário das amplitudes sonoras, como evidenciado pela análise de volatilidade da seção anterior. De fato, a Figura 1.5 suporta a conclusão de que amplitudes sonoras são descritas por um processo estocástico com variância dependente do tempo e a Figura 1.3 indica a existência de memória de longo alcance na série das intensidades sonoras.

Características similares podem ser encontradas em dados de mercado financeiro, nos quais a volatilidade ou risco é um dos ingredientes essenciais para a dinâmica dos preços de ações. Esse cenário econômico foi e tem sido muito estudado devido, principalmente, aos interesses práticos em predizer o comportamento do mercado. Consequentemente, existe um grande número de modelos para descrever a dinâmica de preços de ações [52]. Podemos imaginar também que, de um ponto de vista qualitativo, as interações ou competições existentes em nosso sistema social podem ser parecidas com aquelas existentes entre os agentes financeiros que participam da dinâmica de preço.

Essa discussão nos motivou a empregar um modelo típico de mercado financeiro para tentar reproduzir o comportamento do som de aglomerações humanas. O modelo escolhido foi o processo $GARCH$ (*generalized autoregressive conditional heteroskedastic*). Esse modelo autoregressivo foi proposto para levar em conta a memória de longo alcance que é comum em dados financeiros [53, 54, 55]. Em sua forma mais geral, o $GARCH(p,q)$ é definido como

$$\begin{aligned} x_t &= \sigma_t \, \xi_t \, , \\ \sigma_t^2 &= \alpha_0 + \alpha_1 x_{t-1}^2 + \cdots + \alpha_p x_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \, , \end{aligned} \tag{1.4}$$

em que α_i e β_i são parâmetros de controle e ξ_t é um número aleatório não correlacionado de média zero e variância unitária. Desse modo, o processo $GARCH$ é não correlacionado em x_t e correlacionado na variância.

Aqui, por simplicidade e satisfatoriedade, consideramos o processo $GARCH(1,1)$

$$\begin{aligned} x_t &= \sigma_t \, \xi_t \, , \\ \sigma_t^2 &= \alpha_0 + \alpha_1 x_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \, . \end{aligned} \tag{1.5}$$

Consideramos, também, que ξ_t segue uma distribuição gaussiana. Nessa versão do modelo, contamos com três parâmetros: α_0, α_1 e β_1. Entretanto, devido ao fato de termos escalado as amplitudes sonoras para terem variância unitária, podemos eliminar um dos três parâmetros usando o valor esperado para a variância do processo $GARCH(1,1)$ [53]

$$\sigma_x^2 = \frac{\alpha_0}{1 - \alpha_1 - \beta_1} \, . \tag{1.6}$$

Assim, ficamos efetivamente com dois parâmetros: α_1 e β_1. Esses dois parâmetros foram variados incrementalmente e a soma das diferenças ao quadrado entre os valores simulados e os dados empíricos das amplitudes sonoras foi estimada. Os valores $\alpha_1 = 0{,}011$ e $\beta_1 = 0{,}9889$ (consequentemente $\alpha_0 = 0{,}001$) foram aqueles que minimizaram essa soma. A Figura 1.6 mostra a comparação entre os dados simulados e empíricos. Notamos que a concordância entre o $GARCH(1,1)$ e os dados empíricos é muito boa para as amplitudes sonoras (Figura 1.6A), para os intervalos de retorno (Figura 1.6C) e para as volatilidades (Figura 1.6D). Para as correlações (Figura 1.6B), não temos uma concordância tão boa quanto nos outros casos. O motivo para

isso é que as correlações geradas na variável x_t^2 não são do tipo lei de potência.

De fato, a função de correlação para x_t^2 tem uma cauda que decai exponencialmente [53], *i.e.*,

$$\langle x_t^2 \, x_{t+\tau}^2 \rangle \sim \exp(-t/\tau_c) \, , \tag{1.7}$$

sendo $\tau_c = |\ln(\alpha_1 + \beta_1)|^{-1}$. Entretanto, o processo $GARCH(1,1)$ pode imitar uma memória de longo alcance no limite $\tau_c \to \infty$. Para isso, bastaria escolher a soma $\alpha_1 + \beta_1$ próxima de 1. Em nosso caso, $\alpha_1 + \beta_1 = 0{,}9999$, o que leva a um tempo característico $\tau_c \sim 10^4$ segundos. Esse tempo é muito longo quando comparado com a escala de tempo dos sons das aglomerações humanas e, desse modo, o processo $GARCH(1,1)$ está produzindo um efeito parecido ao de uma correlação de longo alcance. Naturalmente, existem desvios maiores na comparação com os dados empíricos, manifestados principalmente na curva mais acentuada da função de flutuação. Vale ressaltar que os dados empíricos também apresentam um curvatura menor, o que pode ser um indicativo da existência de uma espécie de *cutoff* exponencial na função de autocorrelação das intensidades sonoras.

De modo geral, a semelhança entre os dados e o modelo é bem razoável, principalmente quando levamos em conta a simplicidade do processo $GARCH(1,1)$ em contraste com o cenário complexo da dinâmica humana manifestada nos sons que estamos estudando.

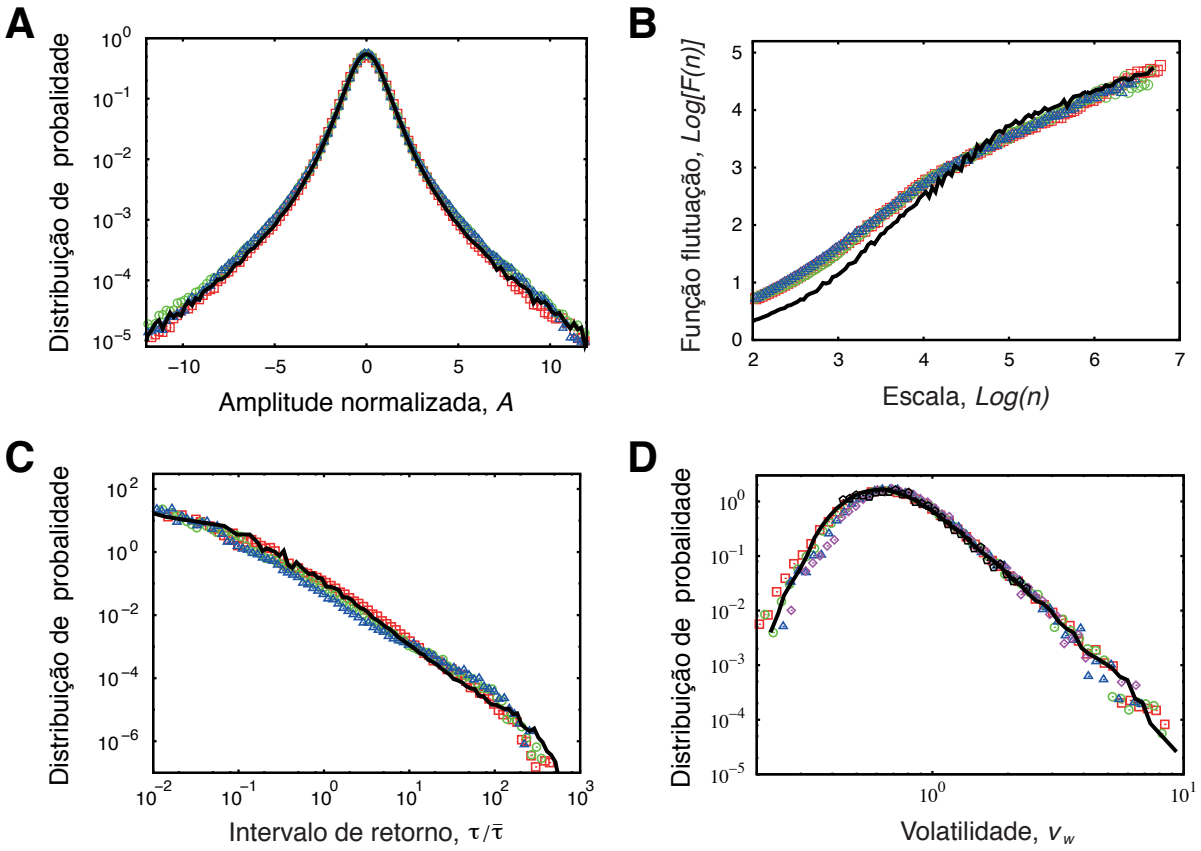


Figura 1.6: Comparação entre os dados empíricos (símbolos) e as predições do modelo $GARCH(1,1)$ (linha contínua). (**A**) Distribuição das amplitudes sonoras normalizadas. (**B**) Análise de correlação DFA. (**C**) Distribuição dos intervalos de retorno considerando três valores limiares: $q = 1$ (quadrados), $q = 2$ (círculos) e $q = 3$ (triângulos). Os dados simulados são uma média sobre esses três valores. (**D**) Distribuição da volatilidade considerando as janelas $w = 1$ (quadrados), $w = 2$ (círculos), $w = 5$ (triângulos), $w = 10$ (losangos) e $w = 100$ (pentágonos), todas em centésimos de segundo. Os dados simulados são uma média sobre esses cinco valores.

1.4 Na direção de um sensor social sonoro

Uma questão interessante sobre os sons das aglomerações humanas é verificar até que ponto os padrões que reportamos até agora são robustos perante as diferentes situações em que esses sons podem surgir. Nessa direção, um caso notório são as aglomerações pacíficas como a que estamos estudando e as aglomerações "violentas", como as que ocorrem em protestos ou brigas. Será que o padrão anterior é robusto o bastante para ser mantido nessas duas situações? Podem as nossas técnicas diferenciar essas duas situações?

Para tentar responder a essas questões procuramos, na internet, por dados relacionados a sons de protestos. Poucos dados foram encontrados, entretanto, na página *Freesound: collaborative database of creative-commons licensed sound for musicians and sound lovers* (`http://www.freesound.org`) foi possível obter alguns sons de protestos ocorridos em Barcelona (Espanha) em agosto de 2006. A Figura 1.7 mostra uma dessas gravações em comparação com um fragmento de nossas gravações anteriores. Nessa figura, é possível notar, nos sons do protesto, a existência de "explosões sonoras" muito mais bem definidas.

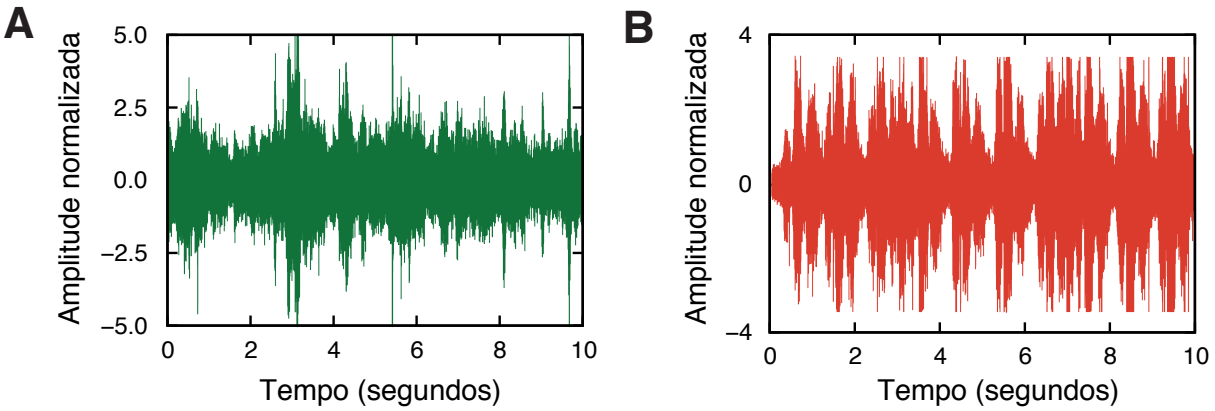


Figura 1.7: Amplitudes sonoras normalizadas (**A**) de uma aglomeração humana pacífica e (**B**) de um protesto ocorrido na Espanha, em agosto de 2006. A figura (**A**) é um segmento da gravação mostrada na Figura 1.1A.

De maneira semelhante à análise anterior, iniciamos pelo cálculo da distribuição das amplitudes sonoras normalizadas nos dois casos. Na Figura 1.8A, podemos observar pequenas diferenças na forma dessas duas das distribuições. Em particular, notamos que, no caso dos sons de protesto, existe um *cut off* nas caudas da distribuição em aproximadamente quatro desvios padrões. Esse comportamento pode ser resultado da possível baixa qualidade do microfone usado para capturar esses sons. Ainda assim, podemos considerar o perfil das distribuições muito semelhantes. Em seguida, calculamos a distribuição dos intervalos de retorno, como mostra a Figura 1.8B. Observamos que a distribuição de $\tau/\bar{\tau}_q$ é praticamente idêntica para ambos os casos. Calculamos também a distribuição das volatilidades das amplitudes sonoras, como mostra a Figura 1.8C. Como no caso das amplitudes, a distribuição das volatilidades possui um *cutoff* muito mais acentuado para os sons de protesto. Notamos que praticamente não existem valores de volatilidade maiores do que duas unidades. Entretanto, na região em que é possível comparar as duas distribuições, elas

diferem muito pouco.

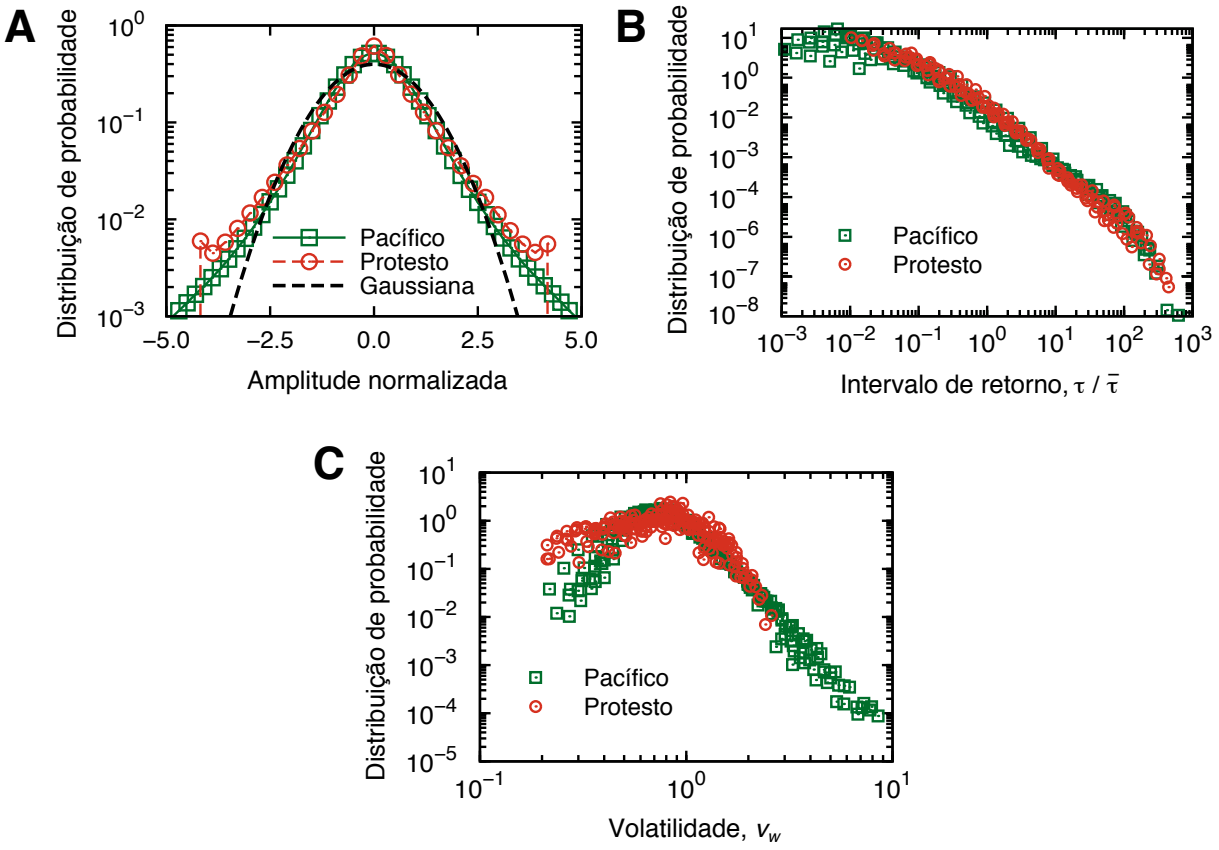


Figura 1.8: (**A**) Distribuição de probabilidade das amplitudes sonoras normalizadas. A linha tracejada é uma distribuição gaussiana de média zero e variância unitária. (**B**) Distribuição de probabilidade dos intervalos de retorno quando $q = 1$, $q = 2$, $q = 3$, $q = 4$ e $q = 5$. (**C**) Distribuição de probabilidade da volatilidade v_w para $w = 0{,}01$, $w = 0{,}05$, $w = 0{,}10$, $w = 0{,}20$, $w = 0{,}40$, $w = 0{,}60$, $w = 0{,}80$ e $w = 1{,}00$, em unidade de segundos. Todos os gráficos foram construídos usando-se os dados da Figura 1.7, sendo que os quadrados representam uma aglomeração pacífica e os círculos representam os sons de um protesto.

Os resultados anteriores sugerem que os padrões nas distribuições calculadas, usando os sons pacíficos, são aproximadamente mantidos para caso dos sons de protesto. Aqui, devemos fazer algumas ressalvas sobre a nossa falta de controle nas gravações dos sons de protesto. Resta-nos, ainda, investigar os aspectos relacionados às correlações nessas séries. Seguiremos uma rota um pouco diferente: ao invés de calcular diretamente o expoente de Hurst para as séries das intensidades sonoras $I(t)$, como fizemos no caso dos sons pacíficos (Figura 1.3), investigaremos as correlações para a série das volatilidades das intensidades sonoras, *i.e.*,

$$v_{\text{int}_w}{}^2(t) = \frac{1}{n-1}\sum_{t'=t}^{t+n-1}\left(I(t') - \langle I(t)\rangle_w\right)^2, \tag{1.8}$$

na qual

$$\langle I(t)\rangle_w = \frac{1}{n}\sum_{t'=t}^{t+n-1} I(t')$$

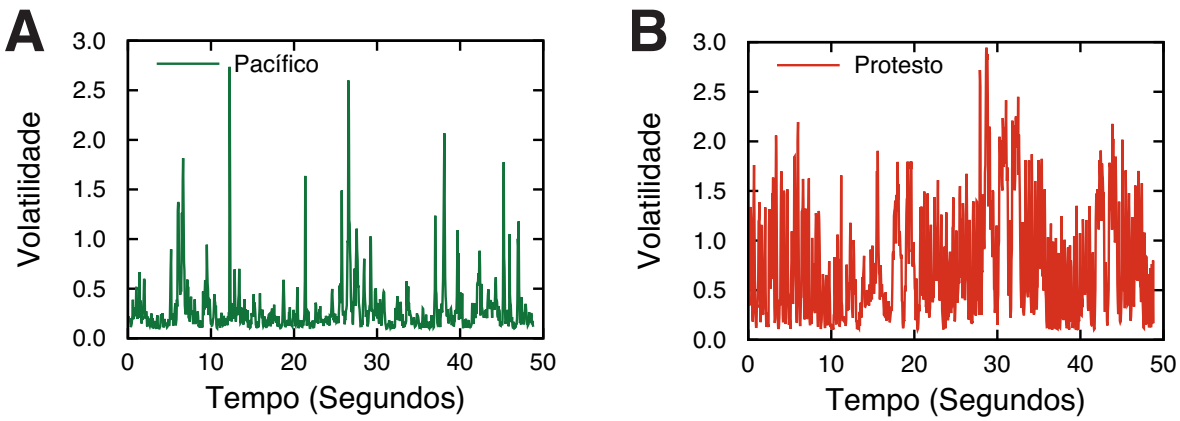


Figura 1.9: Volatilidades das intesidades sonoras (**A**) de uma aglomeração humana pacífica e (**B**) de um protesto ocorrido ao consideremos $w = 0.05$ segundos.

é o valor médio da intensidade dentro da janela $w = n\Delta t$ (Δt é o inverso da taxa de amostragem da gravação). A Figura 1.9 mostra um exemplo dessa nova série temporal.

Calculamos para essa nova série o expoente de Hurst h em função do tamanho da janela temporal w, como mostra a Figura 1.10. Notamos que para janelas de tempo pequenas o expoente de Hurst é praticamente o mesmo para ambos os casos. Observamos ainda que, para valores pequenos de w, a série das volatilidades e as séries das intensidades devem apresentar um comportamento semelhante, de modo que o valor de h, calculado a partir das intensidades é praticamente o mesmo para ambos os sons. Entretanto, à medida em que o tamanho das janelas aumenta, o expoente de Hurst diminui e se aproxima do platô $h \approx 0{,}76$ para os sons pacíficos. Por outro lado, para os sons de protesto, observamos uma diminuição de h até aproximadamente $w = 0{,}2$ e, a partir desse valor, o expoente h começa a crescer e se aproximar do valor $h \approx 1{,}04$. Do ponto de vista qualitativo, ao aumentarmos o tamanho da janela estamos capturando uma dinâmica mais "macroscópica", incluindo mais eventos sonoros dessas aglomerações. Qualitativamente, no caso das aglomerações pacíficas, esse agrupamento de eventos gera uma diminuição nas correlações, o que indica que esses eventos "macroscópicos" são menos correlacionados. O inverso ocorre nos sons de protesto, ou seja, o aumento no exponente de Hurst indica que os eventos sonoros mais "macroscópicos" são ainda mais correlacionados. De fato, protestos costumam ter coros ou "gritos de guerra" mais organizados e em uma escala temporal maior, o que pode dar origem a esse comportamento.

1.5 Conclusões e perspectivas

Estudamos neste capítulo vários aspectos estatísticos dos sons produzidos por aglomerações de pessoas. Na primeira parte, caracterizamos os sons de uma aglomeração ordinária: sons de estudantes durante os intervalos das aulas. Vimos $i)$ que as amplitudes sonoras apresentam uma distribuição não gaussiana de caudas longas, $ii)$ que existem correlações de longo alcance nas séries das intensidades sonoras, $iii)$ que os intervalos de retorno calculados a partir das

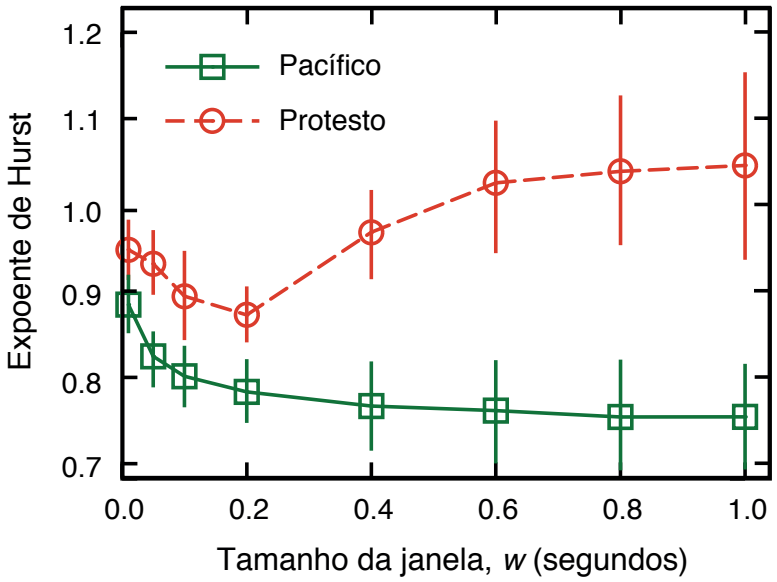


Figura 1.10: Expoente de Hurst para as séries das volatilidades, em função do tamanho da janela w. Todos os gráficos foram construídos usando-se os dados da Figura 1.7, sendo que os quadrados representam uma aglomeração pacífica e os círculos representam os sons de um protesto.

intensidades sonoras não são distribuídos exponencialmente e iv) que a volatilidade das amplitudes sonoras é não estacionária, com uma distribuição assintótica do tipo lei de potência. Todos esses comportamentos observados nos dados foram bem descritos pelo modelo auto-regressivo $GARCH(1,1)$.

Na segunda parte, especulamos sobre a robustez dos padrões estatísticos encontrados. Especificamente, comparamos os padrões encontrados nos sons dos estudantes (sons pacíficos) com aqueles oriundos de um protesto. Apesar das limitações existentes nas gravações dos sons de protesto, observamos que os padrões nas distribuições das amplitudes, dos intervalos de retorno e das volatilidades são bem robustos, de modo que os dois tipos de sons são aproximadamente indistinguíveis desse ponto de vista. Na tentativa de distingui-los de uma maneira efetiva, calculamos algumas propriedades ligadas à existência de correlações de longo alcance em nossos dados. Em particular, calculamos o expoente de Hurst para as séries das volatilidades das intensidades sonoras. Para pequenas escalas de tempo, os expoentes são praticamente idênticos; entretanto, para escalas maiores de tempo os expoentes de Hurst podem diferenciar os dois tipos de sons. Esse comportamento pode ser relacionado, qualitativamente, ao fato dos sons de protesto apresentarem uma estrutura mais organizada, tais como "gritos de guerra". No caso das volatilidades, as correlações se intensificam com o aumento da escala temporal para os protestos e diminuem para os sons pacíficos.

Assim, nossa investigação revelou um padrão não trivial nesses sons bastante comuns na vida em sociedade. Além disso, vimos que parte desse padrão é mantido quando investigamos sons de protesto, mas que aspectos de correlação podem perfeitamente diferenciar sons pacíficos de sons de protestos. É importante notar que essa segunda parte de nossos estudos ainda é preliminar,

de modo que estamos em busca de mais gravações de situações relacionadas às aglomerações humanas para reforçar nossos achados empíricos. Entretanto, apesar dessas ressalvas, estamos bastante entusiasmados pelo fato de que nossos estudos possam gerar uma aplicação prática, uma espécie de sensor social sonoro. Um mecanismo como esse poderia ajudar a monitorar cidades de um modo simples e também mais barato quando comparado com câmeras de vídeo. Outra possibilidade seria o uso em conjunto com as imagens, de modo que a análise sonora das aglomerações poderia indicar aos operadores das imagens qual região deve receber mais atenção.

Capítulo 2

Características universais e correlações nos sons musicais

Neste capítulo, também estudaremos sons, mais especificamente os provenientes de canções musicais [14, 15, 16]. Veremos que, para um grande número de músicas, é possível descrever a distribuição das amplitudes normalizadas usando uma distribuição de probabilidade com apenas um parâmetro, que possui como casos particulares a gaussiana e a distribuição de Laplace. Esse parâmetro pode classificar, de um modo bastante simples, as músicas e os gêneros musicais em uma espécie de ordem de complexidade. Além disso, veremos que esse parâmetro, que é obtido diretamente das distribuição das amplitudes, possui relações estatísticas com os aspectos de correlação medidos pelo expoente de Hurst. Em uma segunda parte deste capítulo, investigaremos um pouco mais a complexidade de músicas usando medidas entrópicas baseadas em ordens locais existentes nos sons dessas músicas. Mostraremos que esses índices entrópicos de complexidade podem ser úteis no processo de identificação automática de gêneros musicais e que eles podem também ajudar pessoas a escolherem canções parecidas. Finalmente, na última parte, especularemos sobre a evolução das canções musicais mais populares de cada época, sugerindo que as canções musicais mais populares têm ficado cada vez mais próximas de ruídos do ponto de vista estatístico.

2.1 Introdução e apresentação dos dados

Assim como no capítulo 1, estudaremos agora outro "fenômeno" produzido pelo homem: as músicas. É grande o interesse dos pesquisadores por esse tema amplo e, por conta disso, existem inúmeros trabalhos em muitas direções [56, 57, 58, 59, 60]. Por exemplo, na direção do papel social das músicas podemos citar os trabalhos de Lambiotte e Ausloos sobre o tamanho [61] e as correlações [62] em comunidades musicais de grupos de compartilhamento *on-line* de músicas e também sobre as vendas de música [63]. Além desses aspectos sociais, as músicas, ou melhor dizendo, os sons produzidos por músicas formam um sistema altamente organizado, possuindo estruturas complexas e correlações de longo alcance. Essas características têm atraído a atenção

dos físicos estatísticos desde 1975, como indica o trabalho de Voss e Clarke [64], no qual eles observaram que o espectro de potência de estações de rádio é do tipo $1/f$ e que, dependendo do gênero musical, as correlações podem se estender para escalas maiores ou menores de tempo. Outro trabalho seminal é o de Hsü e Hsü [65, 66] sobre as mudanças nas frequências acústicas em composições de Bach, Mozart e músicas modernas, no qual eles reportaram a existência de autossimilaridade e estruturas fractais. Além destes, existem muitos outros trabalhos [67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81], e a maioria está baseada em dimensões fractais, espectro de potência ou análise de correlações. Existe também uma comunidade bastante ativa trabalhando em problemas mais aplicados relacionados à classificação de músicas e muitos resultados importantes são publicados na conferência ISMIR (*The International Society for Music Information Retrieval*) [82].

Tabela 2.1: Detalhes das bases de dados das músicas.

Base 1 - 10124 músicas distribuídas em 10 gêneros musicais.

Gênero	Número de músicas
Blues	1020
Classical	997
Flamenco	679
Hiphop	1000
Jazz	700
Metal	1638
Mpb	580
Pop	1000
Tango	1016
Techno	1494

Base 2 - Músicas mais vendidas/reproduzidas nos EUA, de acordo com a *Billboard*.

Período	Número de músicas
1946 a 1948	Top 50 de cada ano
1949 a 1955	Top 30 de cada ano
1956 a 1958	Top 50 de cada ano
1959 a 2007	Top 100 de cada ano

Nosso objetivo aqui é investigar os sons provenientes de músicas de uma maneira mais direta e simples, buscando por outros padrões estatísticos e também possíveis aplicações. Para isso, construímos duas bases de dados de arquivos MP3 (um formato de arquivo de áudio): a primeira base possui $\sim$10 mil músicas distribuídas em 10 gêneros musicais. A segunda base contêm as músicas mais vendidas/reproduzidas, ano a ano, nos Estados Unidos, de acordo a revista *Billboard* [83]. A Tabela 2.1 mostra os detalhes dessas duas bases. Todos os arquivos MP3 usados em nossas análises possuem uma taxa de gravação de $44{,}1$ kHz.

Analogamente ao capítulo 1, focamos nossas análises nas séries normalizadas (média zero e variância unitária) das amplitudes sonoras (A) e nas séries das intensidades sonoras ($I \propto A^2$). A Figura 2.1 mostra exemplos dessas séries temporais para quatro músicas de gêneros musicais distintos. Essa visualização já é suficiente para revelar diferenças qualitativas entre as músicas. Para a música clássica e para o jazz (painéis superiores), podemos observar a existência de algumas "explosões" sonoras e também de regiões onde a amplitude ou a intensidade sonora possuem valores pequenos. No caso das músicas do gênero metal e techno, observamos flutuações muito mais "homogêneas" cobrindo praticamente todo o intervalo de tempo. Nas próximas seções, investigaremos mais detalhadamente essas séries temporais

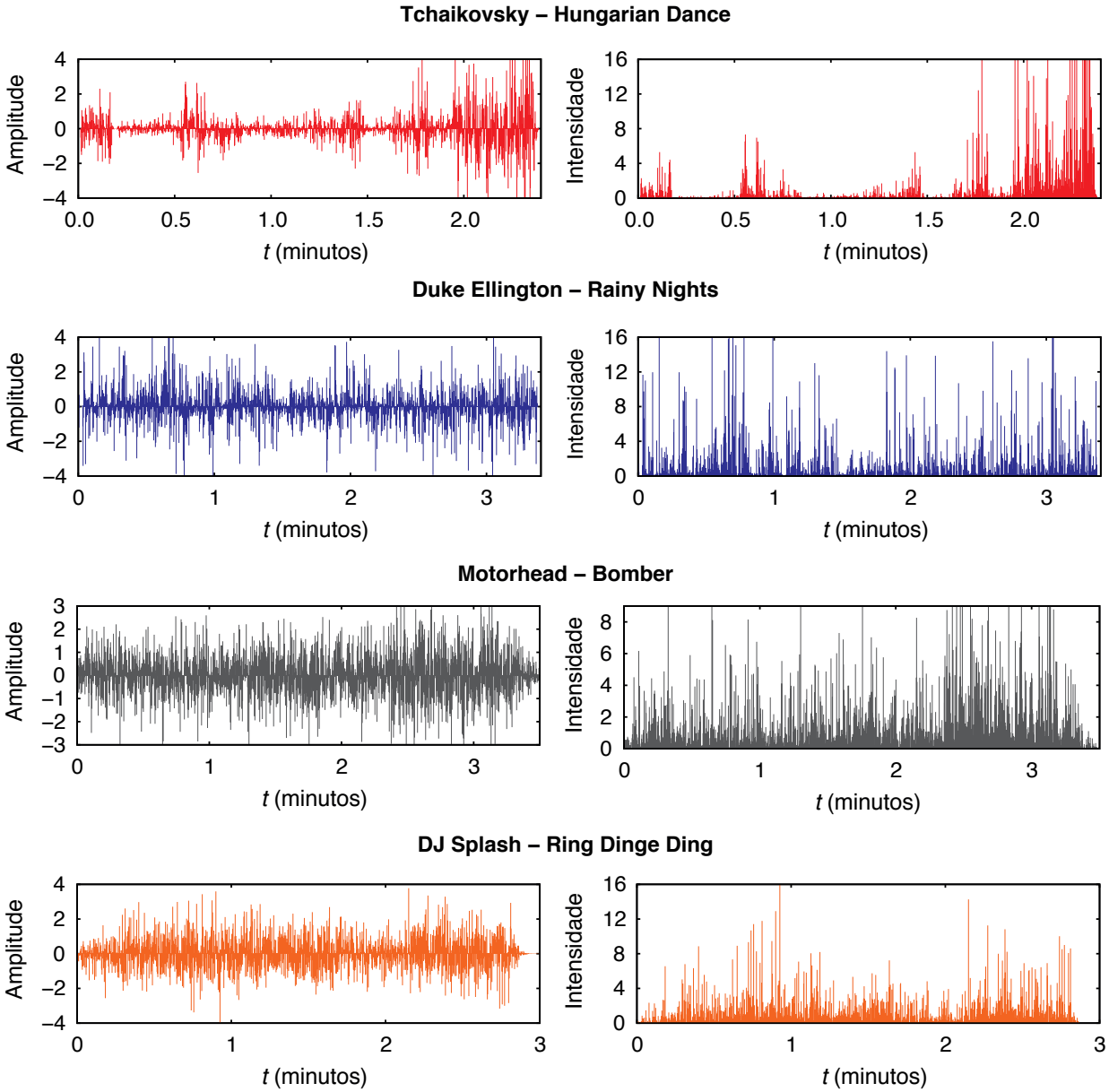


Figura 2.1: Amplitudes sonoras normalizadas (figuras à esquerda) e intensidades sonoras (figuras à direita) para quatro músicas de quatro gêneros diferentes. De cima para baixo: classical (aqui, usamos o termo em inglês para nos referirmos a música erudita ou clássica), jazz, metal e techno.

2.2 Distribuição das amplitudes sonoras e sua relação com as correlações

A análise da Figura 2.1 apontou diferenças qualitativas nas séries temporais das amplitudes sonoras normalizadas a depender da música e do gênero músical. Nesse contexto surge uma pergunta natural: as distribuições de probabilidade das amplitudes sonoras possuem uma forma bem definida?

Para responder a essa questão, calculamos as distribuições de probabilidade das amplitudes sonoras normalizadas (após subtrair o valor médio e dividir pelo desvio padrão) para todas as músicas da base de dados 1 (Tabela 2.1), com exceção das músicas do gênero blues, que foram incluídas posteriormente em nossos estudos. A Figura 2.2 mostra algumas dessas distribuições quando consideramos o canal direito ou o canal esquerdo dos arquivos de audio *stereo*. Notamos que a distribuição é praticamente independente da escolha do canal. De fato, após aplicar o teste de Kolmogorov-Smirnov (Apêndice B.1) em todas as músicas, observamos que em apenas $\sim 1\%$ delas podemos rejeitar a hipótese de igualdade das distribuições. Assim, por simplicidade, consideramos apenas os dados do canal direito em nossas próximas análises.

Mais importante do que a invariância das distribuições perante a escolha do canal é o fato que de todas as distribuições apresentam uma forma bem regular. A inspeção visual da Figura 2.2 e também de todas as outras distribuições revela que as distribuições de probabilidade das amplitudes podem ter um perfil de caudas bem longas, mais próximas de uma distribuição de Laplace ou mais próximas de uma gaussiana. Uma família de distribuições que apresenta todos esses perfis é a gaussiana *stretched* [84]

$$p(A) = \mathcal{N} \exp(-\mathcal{B}|A|^c)\,, \tag{2.1}$$

em que $\mathcal{N}$ é o fator de normalização, $\mathcal{B}$ é um parâmetro relacionado ao desvio padrão e c é o parâmetro que governa a forma da distribuição das amplitudes $p(A)$. Lembramos que essa distribuição tem como casos particulares a distribuição de Laplace ($c = 1$) e a distribuição gaussiana ($c = 2$). Além disso, visto que as amplitudes A possuem média nula e desvio padrão unitário, $\mathcal{N}$ e $\mathcal{B}$ devem ser uma função de c. De fato, escolhendo

$$\mathcal{N} = \frac{c}{2}\left(\frac{\Gamma(3/c)}{\Gamma(1/c)^3}\right)^{1/2} \quad \text{e} \quad \mathcal{B} = \frac{\Gamma(3/c)}{\Gamma(1/c)}^{c/2}\,, \tag{2.2}$$

com $\Gamma(...)$ sendo função gama de Euler, garantimos a normalização da distribuição e também o desvio padrão unitário para todo valor de c.

Com as escolhas anteriores, a distribuição 2.1 fica com apenas um parâmetro livre (c), o qual foi ajustado via método dos mínimos quadrados para cada uma das músicas. A Figura 2.2 mostra alguns dos ajustes obtidos em comparação com as distribuições empíricas. Observamos que há uma boa concordância entre a distribuição 2.1 (linha contínua) e as distribuições empíricas das amplitudes sonoras. A qualidade de todos os outros ajustes é semelhante à mostrada na Figura 2.2, sendo que o teste de Kolmogorov-Smirnov (Apêndice B.1) rejeita a hipótese de que

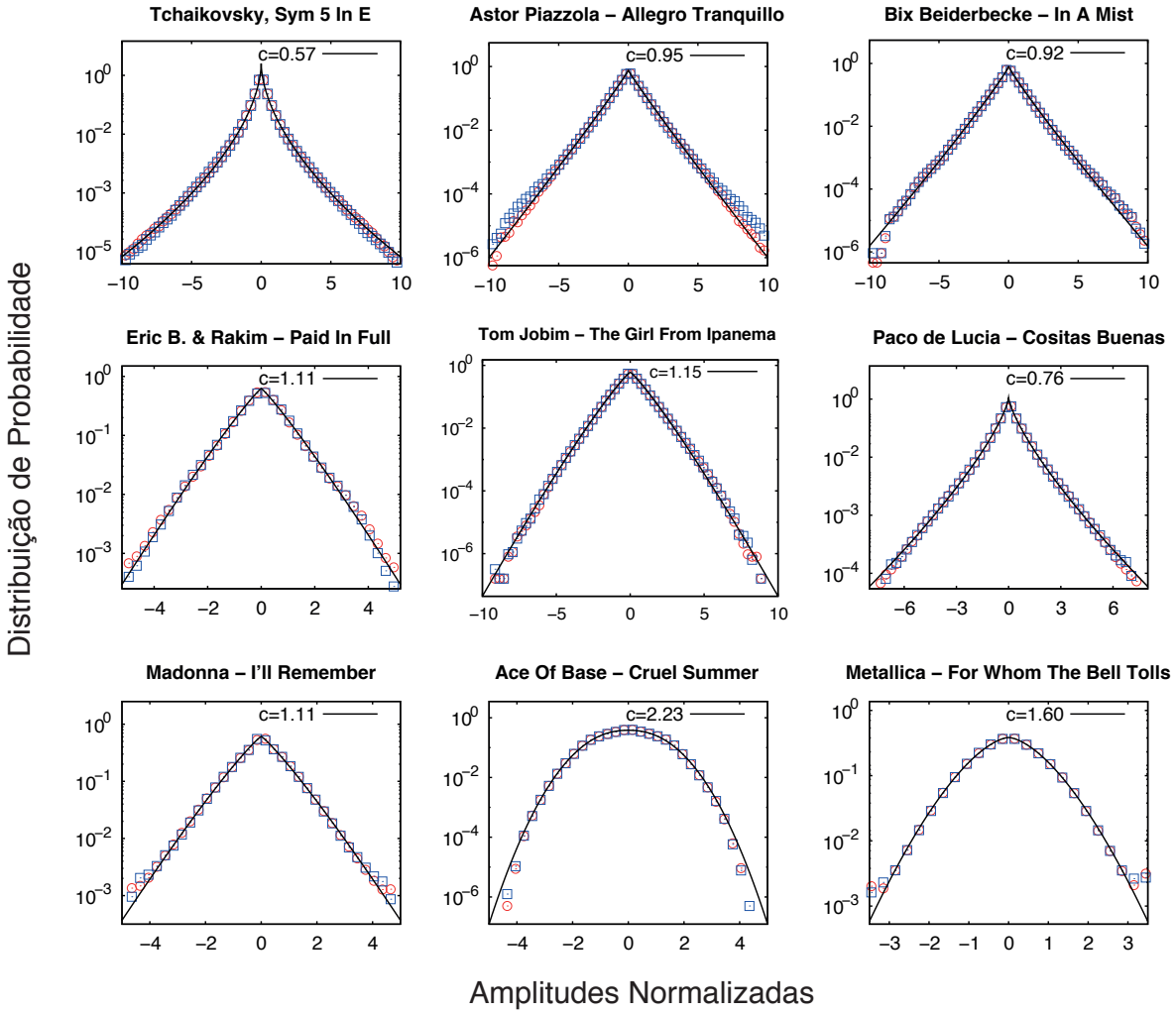


Figura 2.2: Distribuição de probabilidade das amplitudes sonoras normalizadas para nove músicas de gêneros musicais distintos. Os círculos representam os dados do canal direito e os quadrados os dados do canal esquerdo das gravações *stereo* (notamos que praticamente não existem diferenças entre os canais). A curva contínua é um ajuste da função gaussiana *stretched* (equação 2.1), da qual obtivemos os valores do parâmetro c, para cada música, mostrados nas figuras.

a distribuição 2.1 descreve os dados em apenas 10% dos casos. Ainda assim, nos casos em que a hipótese foi rejeitada, há muita semelhança na parte central da distribuição (aproximadamente cinco desvios padrões).

Podemos, então, responder à questão que propusemos no início desta seção de maneira positiva, ou seja, as distribuições das amplitudes sonoras possuem uma forma bem definida, a qual pode ser descrita por uma função gaussiana *stretched* de apenas um parâmetro (c). Surge, agora, outra questão natural: que tipo de informação o parâmetro c pode dar sobre a música? Do ponto vista estatístico, podemos considerar que c define uma espécie de distância para um processo gaussiano, o qual pode ser considerado como "simples" ou trivial no contexto do Teorema do Limite Central. Desse modo, músicas com valores de c próximos a dois podem ser consideradas "simples", enquanto músicas com parâmetros c distantes de dois podem ser

consideradas "complexas", pois um processo estocástico que dê origem a essa distribuição deve possuir uma estrutura mais intrincada do que aquele que gera uma gaussiana. Devemos enfatizar que nossa definição de complexidade para músicas é bastante simples e não leva em conta muitos outros aspectos que especialistas em música poderiam considerar. Entretanto, nosso reducionismo pode ser pensado como uma espécie de medida global para a complexidade das músicas. Na Figura 2.2 já é possível perceber o tipo de informação que o parâmetro c fornece. Notamos que a Sinfonia Número 5 em mi menor, de Tchaikovsky, possui uma distribuição de amplitudes com caudas bem longas, bem ajustada por $c = 0{,}57$. Por outro lado, a banda pop *Ace of Base* tem uma distribuição de caudas bem mais curta, com $c = 2{,}23$, para sua música *Cruel Summer*.

Para entender melhor o papel do parâmetro c, calculamos o seu valor para cada gênero musical como mostra a Figura 2.3. Nesse gráfico de barras, mostramos o valor médio de c em ordem crescente para todos os gêneros musicais. Ainda que possa ser considerada "grosseira", essa classificação faz bastante sentido, ao menos para separar musicas de alto padrão (aquela música que requer um nível significativo de treinamento e proficiência nos instrumentos, como a música clássica) das músicas dançantes (como a música eletrônica).

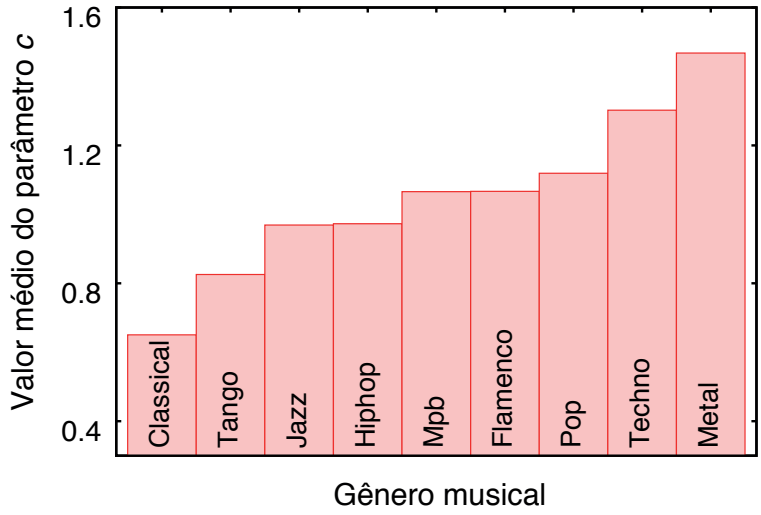


Figura 2.3: Classificação dos gêneros musicais de acordo com o valor médio parâmetro do c obtido por meio do ajuste da gaussiana $stretched$ para todas as músicas de um dado gênero. Aqui, quanto mais próximo de dois estiver o parâmetro c, maior é proximidade das distribuições das amplitudes sonoras como uma gaussiana.

Além do valor médio, calculamos também a distribuição de probabilidade dos parâmetros c para cada gênero musical, como mostra a Figura 2.4. Notamos que, por um lado, praticamente não existe superposição entre as distribuições para a música clássica e metal. Por outro lado, na região próxima de $c = 1$ existe muita superposição entre as distribuições dos gêneros jazz, hiphop, mpb, flamenco e pop. Essa superposição pode ser um reflexo da má definição e das fronteiras confusas desses gêneros musicas [72] e também da simplicidade da nossa medida de complexidade. Como já dissemos, existe uma comunidade muito ativa buscando por métodos

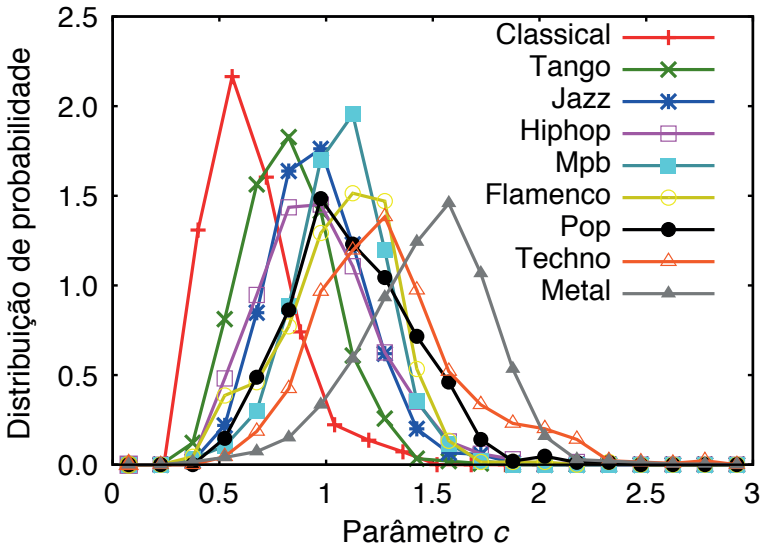


Figura 2.4: Distribuição de probabilidade dos parâmetros c para cada gênero musical. Notamos que existe muita superposição entre essas distribuições, refletindo (ao menos em parte) os limiares confusos das definições dos gêneros musicais.

para classificar músicas (ISMIR [82]) e uma resposta convincente para essa classificação ainda é um problema em aberto [72]. Na próxima seção, apresentaremos um procedimento mais promissor para classificar músicas. Por enquanto, nosso interesse maior é identificar padrões nesses sons.

Outra possibilidade é investigar a existência de correlações nas séries das intensidades sonoras. Já é bem conhecido que músicas possuem correlações desde os trabalhos de Voss e Clarke [64] e mais recentemente no trabalho de Jennings *et al.* [81]. Entretanto, até onde sabemos, aspectos de correlação diretamente nas séries das intensidades não tinham sido estudados. Nossa abordagem é análoga a que empregamos nos sons de aglomerações de pessoas, *i.e.*, empregamos a *detrended fluctuation analysis* (DFA) nas séries das amplitudes (Apêndice A.3). Essa técnica está baseada no cálculo da função de flutuação $F(n)$ para diferentes escalas temporais n: quando a série possui correlações de longo alcance, $F(n)$ apresenta um comportamento lei de potência ($F(n) \sim n^h$, com $h \neq 0{,}5$) com expoente numericamente igual ao expoente de Hurst, h. A Figura 2.5 mostra a análise DFA para as mesmas músicas da Figura 2.2. Notamos que os valores de h são todos maiores do que $0{,}5$, confirmando, por um outro procedimento, que as intensidades sonoras das músicas possuem correlações de longo alcance.

Analogamente à Figura 2.3, podemos calcular o valor médio de h para cada gênero musical. A Figura 2.6 mostra esses valores médios, mantendo a ordem crescente de c. Novamente observamos que gêneros musicais de alto padrão apresentam um valor de h mais próximo de um (1), enquanto gêneros dançantes são caracterizados por valores menores de h. Ressaltamos ainda, que valores de h próximos de um (1) estão relacionados ao ruído $1/f$ reportado por Voss e Clarke [64].

A Figura 2.6 mostra, também, que existe uma boa concordância entre o ordenamento dos gêneros musicais usando valores de c e os valores de h. É verdade que existem inversões, como no caso do gênero hiphop. Entretanto, o cálculo do coeficiente de Pearson (equação A.1,

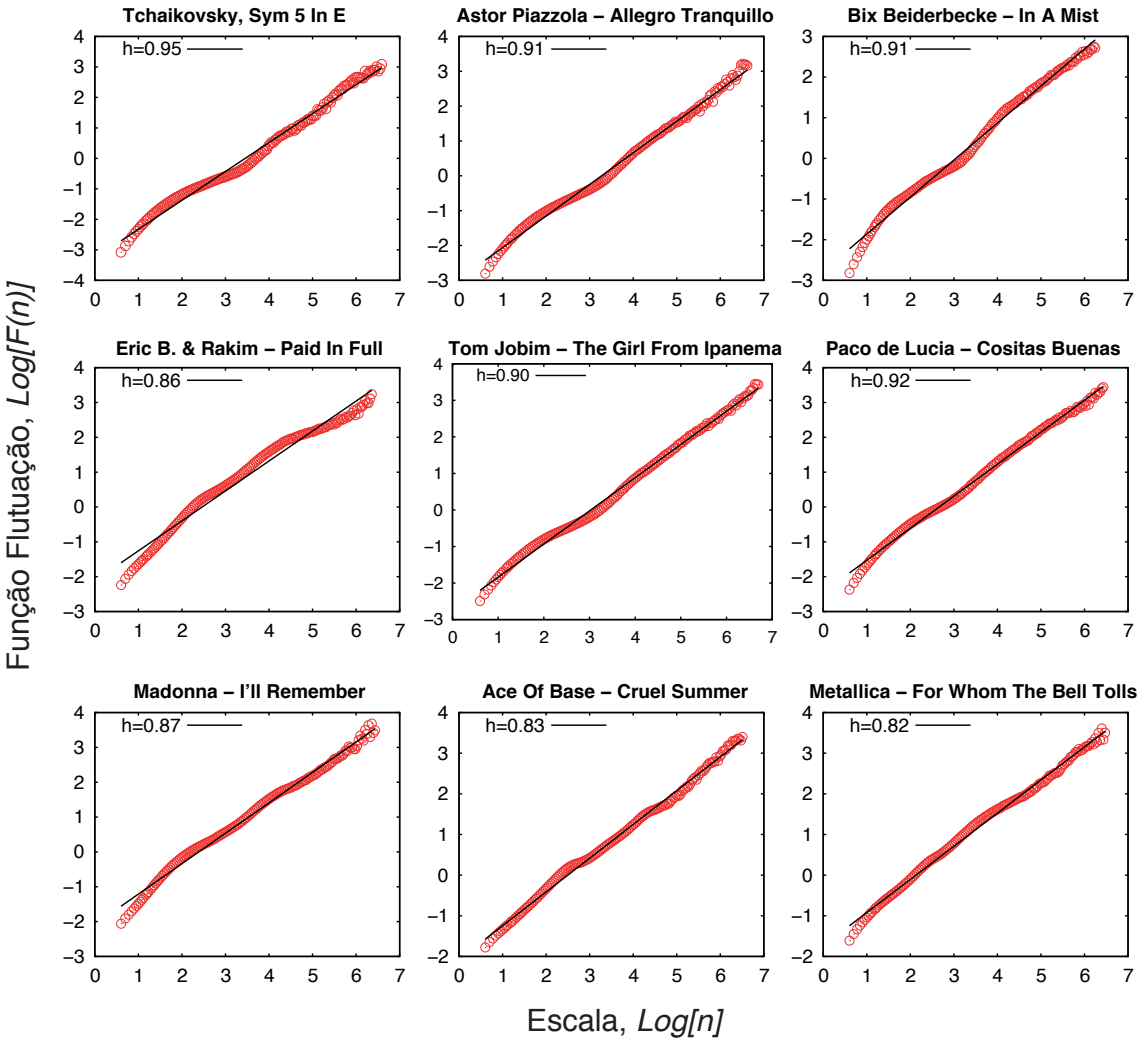


Figura 2.5: Análise DFA para as séries das intensidades sonoras de nove músicas de nove gêneros musicais. A linha contínua é um ajuste linear aos dados, sendo o coeficiente angular numericamente igual ao valor do expoente de Hurst (h). A escala n está em unidades de $1/44{,}1$ k segundos.

Apêndice A.1) entre os dois ordenamentos (também conhecido como coeficiente de Spearman) leva a um valor de $-0{,}7$, o que indica uma forte tendência de h decrescer à medida que c cresce.

Essa relação entre os ordenamentos pode ser interpretada como uma espécie de acoplamento entre as distribuições de probabilidade das amplitudes e as correlações nas intensidades sonoras. Assim, à medida que as caudas das distribuições ficam mais próximas de uma gaussiana, as intensidades também perdem um pouco de correlação. Será que esse comportamento ocorre em um nível individual? Para responder a essa pergunta, mostramos na Figura 2.7 um gráfico de dispersão do parâmetro c versus o expoente de Hurst, h. Observamos que não há uma relação perfeita entre a duas variáveis: o que existe é uma tendência estatística considerável de encontrar, por exemplo, músicas com expoente de Hurst próximos de um (1) que também possuam distribuições de amplitude de cauda bastante longa. Esse resultado é confirmado pelo valor de $0{,}74 \pm 0{,}02$ da correlação de Pearson entre as duas variáveis.

Em resumo, as investigações anteriores revelaram que a distribuição das amplitudes sonoras

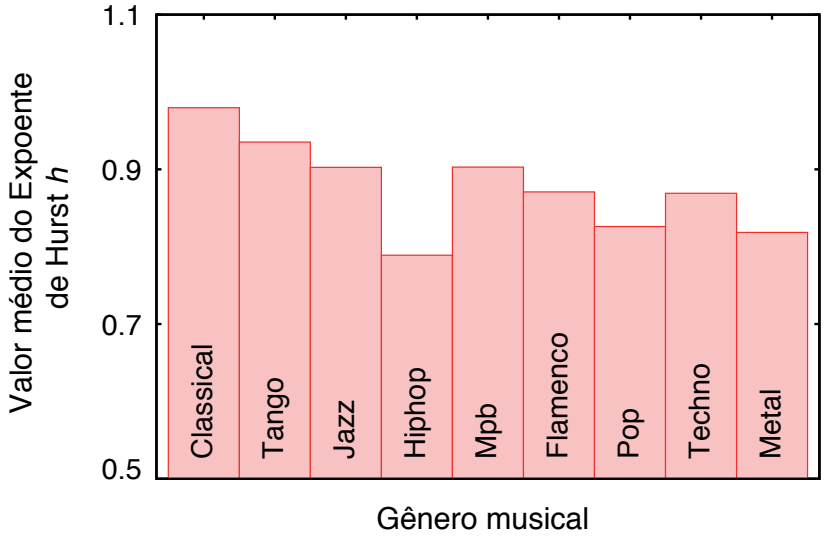


Figura 2.6: Valor médio do expoente de Hurst, h, calculado a partir das intensidades sonoras para cada gênero musical.

de várias músicas possui uma forma bem definida e em bom acordo com uma gaussiana *stretched*. Devido às condições da série, a distribuição pode ser escrita em função de apenas um parâmetro, o qual foi usado para definir uma espécie de distância para um processo gaussiano e, consequentemente, uma medida de complexidade para as músicas. Comprovamos a existência de correlações de longo alcance nas séries das intensidades sonoras e também a existência de um acoplamento entre as correlações e as distribuições das amplitudes. Na seção seguinte, usamos outra técnica para analisar as músicas em uma vertente um pouco mais aplicada.

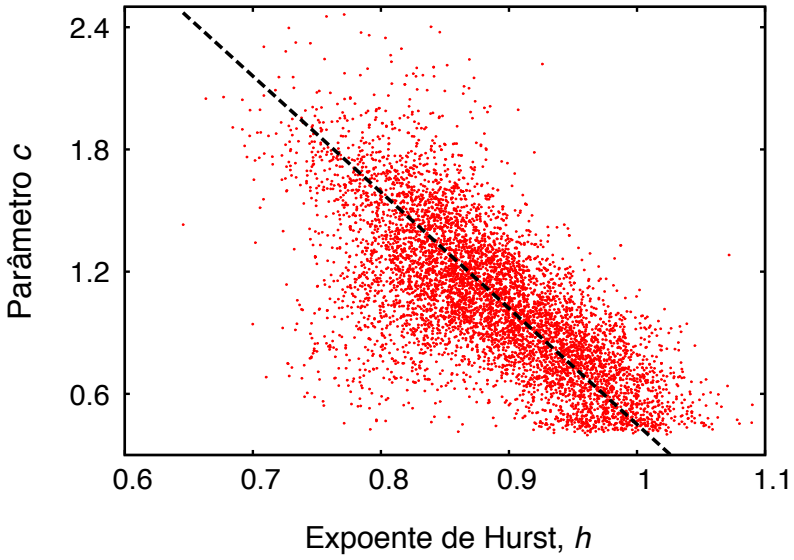


Figura 2.7: Relação existente entre o perfil da distribuição de probabilidade das amplitudes sonoras (medidos pelos parâmetros c) e aspectos relacionados às correlações existentes nas intensidades sonoras (medidos pelos expoentes de Hurst, h). A linha tracejada é apenas uma guia para os olhos.

2.3 Entropia e complexidade de permutação para classificar músicas e gêneros

Os parâmetros c e h, extraídos das séries das amplitudes e das intensidades, fornecem informações sobre a complexidade e permitem uma classificação preliminar de músicas e de gêneros musicais. Contudo, de um ponto de vista mais aplicado, surgem vários problemas práticos para sua utilização visando classificar e agrupar músicas em grandes bases de dados. Uma dessas dificuldades está relacionada à maneira numérica com a qual c e h são extraídos dos dados. Por exemplo, no caso do parâmetro c é preciso calcular inicialmente o histograma das amplitudes e em seguida realizar o ajuste da distribuição. Além de ser um procedimento lento e computacionalmente exigente, o processo envolve várias técnicas que dependem de parâmetros que devem ser bem escolhidos para o bom desempenho da classificação (o número de janelas do histograma ou o algoritmo de ajuste, por exemplo).

As características anteriores não são desejadas para um processo automatizado e não supervisionado de classificação de músicas. Na tentativa de contornar essas dificuldades, empregamos o recente método proposto por Bandt e Pompe [85] em conjunto com os desenvolvimentos de Rosso *et al.* [86], visando definir uma medida de complexidade mais natural para séries temporais. Como veremos, esse procedimento tem muitas vantagens técnicas em relação aos índices c e h, sendo um processo extremamente rápido, simples e robusto.

A essência do método é associar uma sequência de símbolos à série temporal em análise. Para sermos mais específicos, consideremos uma série temporal $\{x_t\}_{t=1,\ldots,N}$. Em nossas análises, x_t pode representar as amplitudes ou as intensidades sonoras. Consideremos também a partição representada pelo vetor d-dimensional ($d > 1, d \in \mathbb{N}$)

$$(\vec{s}) \mapsto (x_{s-(d-1)}, x_{s-(d-2)}, \ldots, x_{s-1}, x_s) \,, \tag{2.3}$$

com $s = d, d+1, \ldots, N$. Para cada um desses $(N-d+1)$ vetores, investigamos as permutações $\pi = (r_0, r_1, \ldots, r_{d-1})$ dos símbolos $(0, 1, \ldots, d-1)$ e definidas pelo ordenamento

$$x_{s-r_{d-1}} \leq x_{s-r_{d-2}} \leq \cdots \leq x_{s-r_1} \leq x_{s-r_0} \,. \tag{2.4}$$

Em seguida, calculamos para todas as possíveis $d!$ permutações de π o conjunto das probabilidades $P = \{p(\pi)\}$, dadas por

$$p(\pi) = \frac{\#\{s | s \leq N-d+1;\ (\vec{s}) \text{ do tipo } \pi\}}{N-d+1} \,, \tag{2.5}$$

no qual o símbolo $\#$ representa o número de ocorrências da permutação π. Usando esse conjunto de probabilidades $P = \{p(\pi)\}$ definiremos os índices de complexidade.

Antes de avançarmos, consideramos um exemplo específico a fim de deixar mais claro o

procedimento anterior. Suponhamos $N = 6$, $d = 3$ e

$$\{x_t\} = \{8, 7, 4, 9, 8, 5\}\,.$$

Para essa série, teremos quatro vetores $\vec{s}$, dados por:

$$\begin{aligned}
(\vec{3}) &= (8, 7, 4)\,; \\
(\vec{4}) &= (7, 4, 9)\,; \\
(\vec{5}) &= (4, 9, 8)\,; \\
(\vec{6}) &= (9, 8, 5)\,.
\end{aligned}$$

Cada vetor $\vec{s}$ deve ter seus elementos ordenados para definirmos a permutação π que será aplicada na sequência de símbolos $(0, 1, 2)$. A tabela abaixo ilustra o processo de ordenamento. Notamos que, em nosso exemplo, das 6 possíveis permutações, apenas três aparecem em nossa análise.

$\vec{s}$	$\vec{s}$ ordenado	$\pi \to (0, 1, 2)$
$(8, 7, 4)$	$(4, 7, 8)$	$(2, 1, 0)$
$(7, 4, 9)$	$(4, 7, 9)$	$(1, 0, 2)$
$(4, 9, 8)$	$(4, 8, 9)$	$(0, 2, 1)$
$(9, 8, 5)$	$(5, 8, 9)$	$(2, 1, 0)$

Usando a tabela acima, podemos obter as probabilidades $p(\pi)$, ou seja,

$$\begin{aligned}
p(\text{“012”}) &= 0\,; \\
p(\text{“021”}) &= 1/4\,; \\
p(\text{“102”}) &= 1/4\,; \\
p(\text{“120”}) &= 0\,; \\
p(\text{“201”}) &= 0\,; \\
p(\text{“210”}) &= 2/4\,.
\end{aligned}$$

Em $p(\pi)$, os números entre aspas representam as possíveis permutações: “012” significa manter os elementos na ordem em que estão; “021” significa manter o primeiro elemento em sua posição, colocar o terceiro elemento na segunda posição e colocar o segundo elemento da terceira posição; e assim por diante. Como deve estar claro, o procedimento para o cálculo dessas probabilidades é um simples processo de ordenamento local nas séries temporais e, por conta disso, é executado muito rápido do ponto de vista computacional. Além disso, uma vez escolhido o tamanho dos vetores d (*embedding dimension*), o processo não possui nenhum parâmetro extra para ser ajustado ou procedimentos adicionais de ajuste.

Uma vez definido o processo para obtenção do conjunto de probabilidades $P = \{p(\pi)\}$,

podemos calcular a entropia de permutação normalizada

$$H[P] = \frac{S[P]}{\log d!}\,, \tag{2.6}$$

em que $S[P] = -\sum p(\pi)\log p(\pi)$ representa a entropia de Shannon [87]. Naturalmente, $0 \leq H \leq 1$, sendo que $H = 1$ ocorre para uma série completamente aleatória, *i.e.*, uma série na qual todas as possíveis $d!$ permutações são equiprováveis. Caso a série apresente uma dinâmica de ordenamento mais complexa, H será geralmente menor do que 1.

Além da entropia, calculamos também uma outra medida de complexidade estatística que foi proposta inicialmente por Lamberti *et al.* [88] e definida como

$$C[P] = Q[P, P_e]\, H[P]\,, \tag{2.7}$$

na qual $P_e = \{1/d!\}$ representa a condição de equiprobabilidade das permulações π e

$$Q[P, P_e] = \frac{S[(P + P_e)/2] - S[P]/2 - S[P_e]/2}{Q_{\mathsf{max}}}\,, \tag{2.8}$$

sendo

$$Q_{\mathsf{max}} = -\frac{1}{2}\left[\frac{d!+1}{d!}\log(d!+1) - 2\log(2d!) + \log(d!)\right] \tag{2.9}$$

uma constante de normalização determinada pela condição de máximo do numerador de $Q[P, P_e]$, ou seja, quando todas as componentes de P são iguais a zero, exceto uma.

Essa nova medida C é então o produto de uma distância métrica entre as probabilidades P e a distribuição equiprovável P_e e a entropia normalizada H. A quantidade Q é normalmente denominada *desequilíbrio* e tem relações com as divergências de Jensen-Shannon [89] e Kullback-Leibler [90]. Essa nova quantidade identifica a existência de correlações [91, 92, 88] na série temporal e fornece outras informações que podem não ser obtidas somente pela entropia H. Normalmente, C vai ser diferente de zero sempre que houverem permutações π que ocorram com maior probabilidade do que as demais.

Outro aspecto interessante de C é que essa quantidade não é uma função trivial da entropia H, *i.e.*, para um dado H existe um intervalo de valores permitidos para C [93]. Devido principalmente a essa característica, Rosso *et al.* [86] propuseram o emprego do diagrama C versus H para diferenciar ruído de caos. Esse diagrama conhecido como *complexity-entropy causality plane* [86, 94, 95] será a nossa abordagem para diferenciar as músicas.

Como deve estar claro, o único parâmetro a ser escolhido para determinar H e C é a *embedding dimension* d. De fato, d tem um papel importante para a estimativa do conjunto de probabilidades $P = \{p(\pi)\}$, visto que ele define o número de permutações π acessíveis ao sistema. Contudo, a escolha de d está intimamente ligada ao tamanho N da série, de tal modo que para termos uma boa estatística a condição $d! \ll N$ deve ser satisfeita. Na maioria dos casos práticos, empregar $d = 3, \ldots, 7$, como recomendado por Bandt e Pompe [85], é suficiente. Aqui, em nossa análise

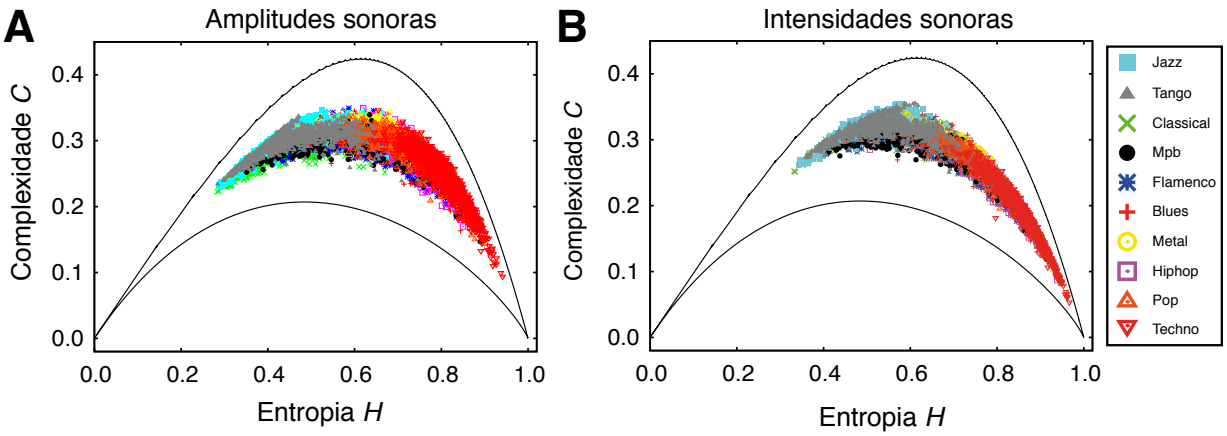


Figura 2.8: Plano complexidade-entropia, C versus H, para todas as músicas agrupadas por gênero musical, quando consideramos as séries das (**A**) amplitudes sonoras e (**B**) intensidades sonoras. Aqui consideramos $d = 5$ para definir o conjunto das probabilidades P associado às possíveis permutações nas séries de cada música.

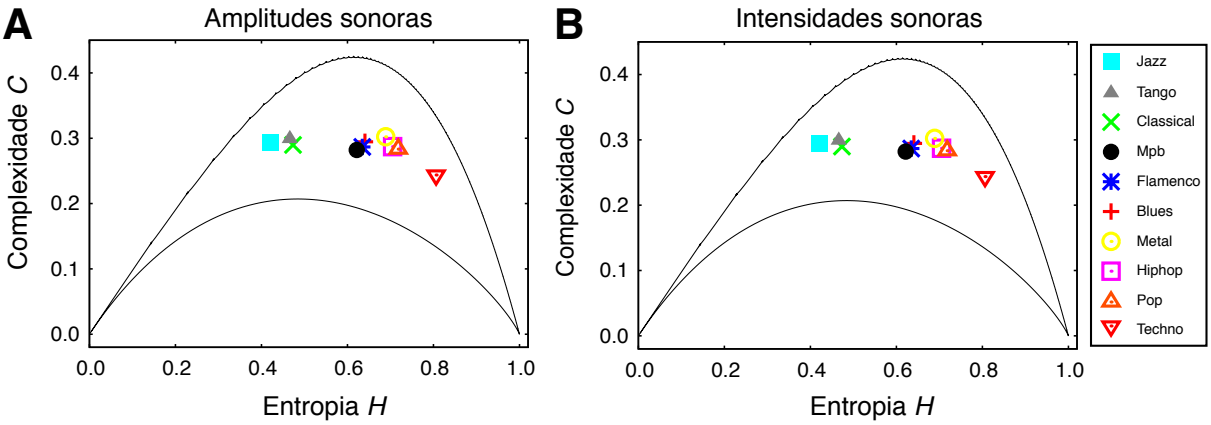


Figura 2.9: Coordenada média de cada gênero musical mostrado na Figura 2.8 para o caso das (**A**) amplitudes sonoras e (**B**) intensidades sonoras. Notamos a existência de quatro grupos de gêneros musicais bem definidos: $G1$ - jazz, tango e classical; $G2$ - mpb, flamenco e blues; $G3$ - metal, hiphop e pop; $G4$ - techno.

das músicas, fixamos $d = 5$, uma vez que as séries temporais sonoras possuem mais de 10^6 elementos.

A Figura 2.8 mostra o diagrama C versus H para todas as músicas e gêneros da tabela 2.1, quando consideramos as amplitudes e as intensidades sonoras. Notamos que ambas as séries levam a resultados similares e que as músicas se distribuem em um longo intervalo de H e C. Essa variação permite uma comparação relativa entre músicas. Por exemplo, podemos escolher qual música ouvir limitando a um intervalo de valores de C e H.

Calculamos também o valor médio de C e H para cada gênero musical, como mostrado na Figura 2.9. Esses valores médios sugerem a existência de quatro grupos de gêneros musicais: $G1$ - jazz, tango e classical; $G2$ - mpb, flamenco e blues; $G3$ - metal, hiphop e pop; $G4$ - techno. Ademais, notamos que gêneros de alto padrão como o classical e o jazz possuem valores menores de entropia H e valores maiores de complexidade C, o que confirma a existência de padrões ordinais mais complexos nas músicas desses gêneros. Outros gêneros, como o techno e o pop

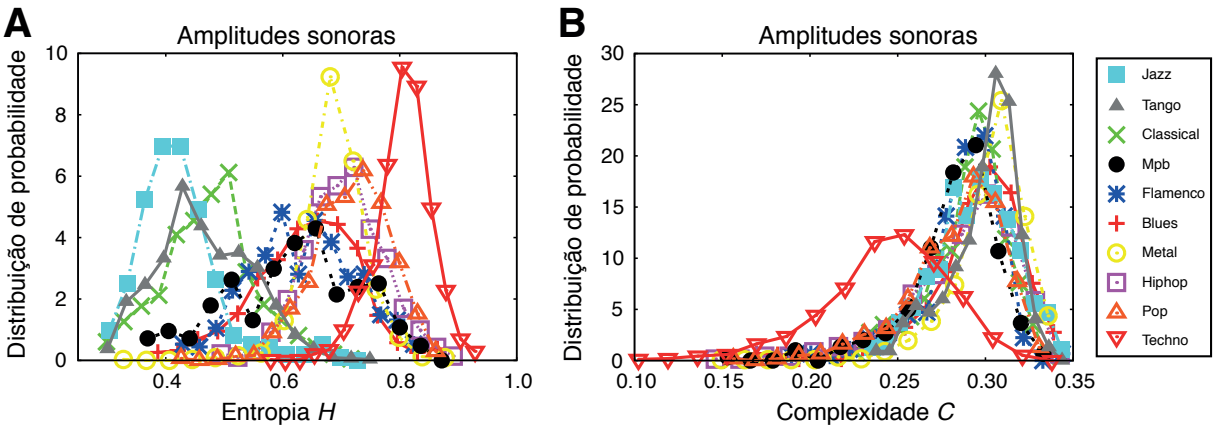


Figura 2.10: Distribuição de probabilidade para os valores (**A**) da entropia de permutação H e para (**B**) a complexidade C para cada gênero musical, quando usamos as séries das amplitudes sonoras.

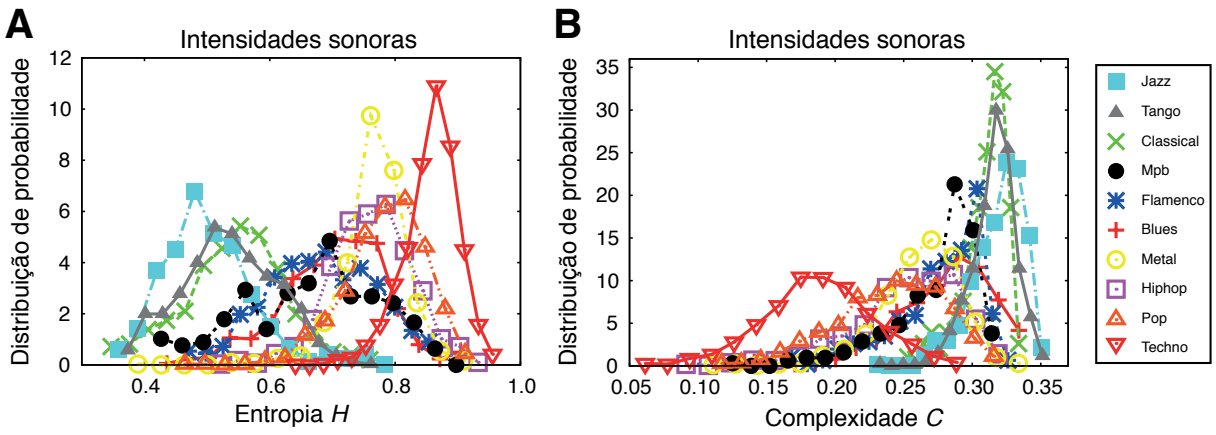


Figura 2.11: Distribuição de probabilidade para os valores (**A**) da entropia de permutação H e para (**B**) a complexidade C para cada gênero musical, quando usamos as séries das intensidades sonoras.

estão localizados mais próximos do limite de permutações aleatórias ($H = 1$ e $C = 0$), indicando que o padrão ordinal desses sons é mais "pobre" no sentido de ser mais próximo de aleatório.

Esses resultados concordam com aqueles obtidos na seção anterior, entretanto, o procedimento atual mostra-se mais promissor para diferenciar as músicas. Em particular, pudemos identificar a existência de alguns grupamentos de gêneros que passaram despercebidos na outra abordagem. Além disso, como mostram as Figuras 2.10 e 2.11, as distribuições de probabilidade dos valores de H e C apresentam uma menor superposição entre os diferentes gêneros quando comparada com a distribuição do parâmetro c (Figura 2.4). Essa melhor distinção dos gêneros fica mais evidente ao calcularmos a distribuição de probabilidade conjunta de H e C, como mostra a Figura 2.12. Vemos que praticamente não existe superposição entre os grupamentos de gêneros $G1$, $G2$, $G3$ e $G4$.

Para ir além das análises anteriores, tentamos quantificar a eficiência dos índices H e C em um cenário mais prático relacionado ao processo de detecção automática de gêneros musicais. Para isso, empregamos um algoritmo que aprende padrões a partir de exemplos, conhecido como

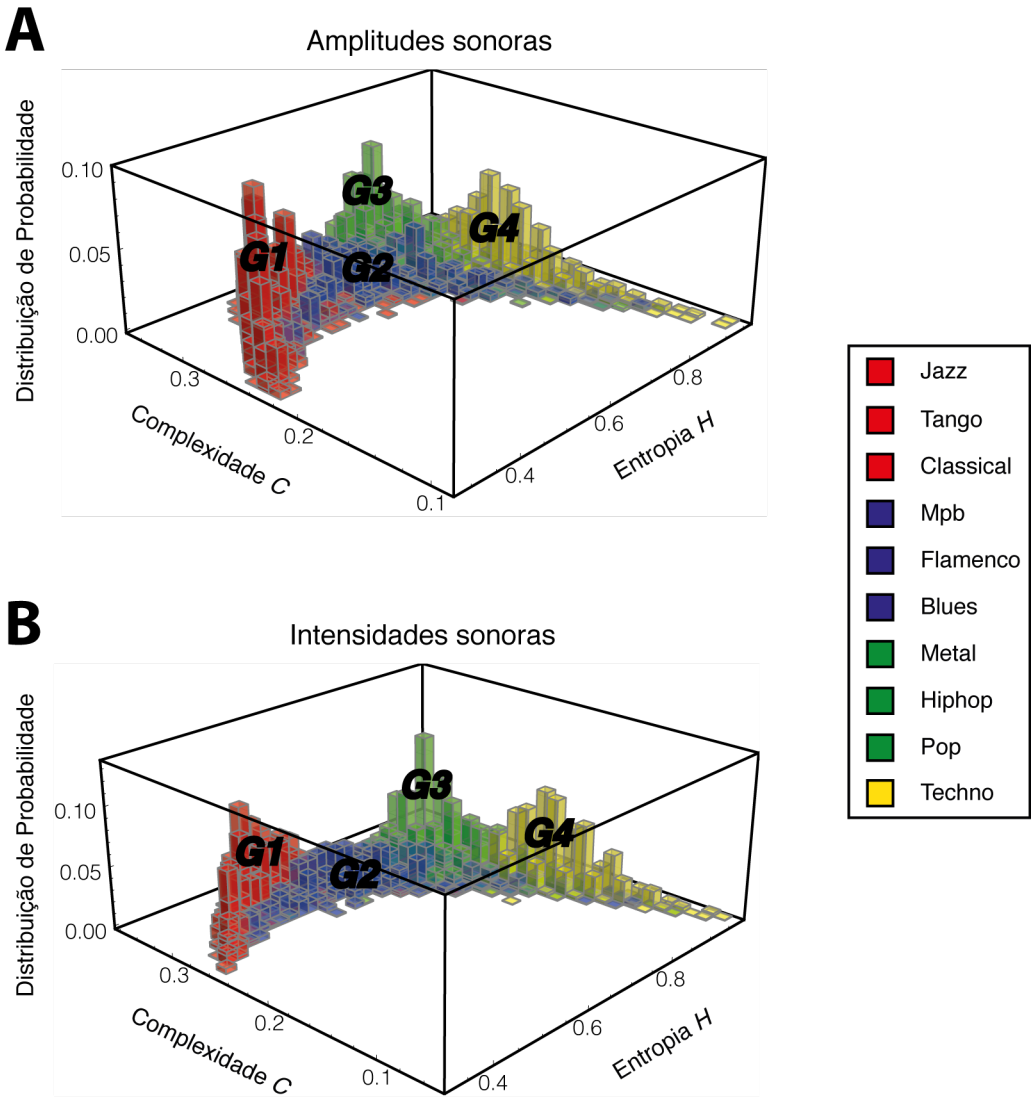


Figura 2.12: Distribuição de probabilidade conjunta da entropia de permutação H e da complexidade C considerando as (**A**) amplitudes e as (**B**) intensidades sonoras. Os diferentes gêneros musicais foram coloridos usando quatro cores, de acordo com os grupamentos sugeridos pela Figura 2.9.

support vector machine (SVM) [96]. Existem vários códigos que implementam esse algoritmo e em nosso caso, usamos uma versão simplificada que aprende padrões binários [97]. Desse modo, a pergunta que o SVM vai responder é se uma música pertence ou não a um dado gênero e não uma diferenciação completa entre todos os gêneros.

A ideia básica desse procedimento é encontrar um hiperplano que separe em duas regiões o espaço de parâmetros que caracterizam os objetos em questão. Para nossa análise, consideramos quatro parâmetros por música, *i.e.*, H e C calculados a partir das amplitudes e intensidades. O primeiro passo para aplicação da SVM é o treinamento. Nessa etapa, selecionamos aleatoriamente 90% das músicas e executamos o algoritmo SVM para aprender o padrão (ou seja, determinar o hiperplano) de separação de cada gênero com os demais. Nos 10% restantes das músicas, executamos a detecção automática de gênero. A Tabela 2.2 mostra o percentual de respostas corretas em cada gênero musical. Notamos que a precisão das escolhas é de aproximadamente

Tabela 2.2: Porcentagem de escolhas corretas da SVM para cada gênero musical.

Gênero	Precisão	Gênero	Precisão
Blues	87,87%	Metal	89,89%
Classical	92,03%	MPB	97,15%
Flamenco	95,12%	Pop	88,11%
Hiphop	88,11%	Tango	87,87%
Jazz	91,68%	Techno	87,14%

90% para todos os gêneros. Vale lembrar que estamos usando um versão binária da SVM, sendo que para escolhas múltiplas esse percentual certamente seria menor. Por outro lado, essa abordagem mostra, ao menos, a viabilidade do uso dos índices H e C em cenários mais práticos.

2.4 Uma abordagem quantitativa para a evolução das músicas populares

Vamos, agora, deixar de lado o caráter mais aplicado da seção anterior e focar em uma questão mais fundamental: a evolução temporal das músicas. Talvez a grande maioria das pessoas tenha a impressão de que a qualidade das músicas ao longo dos anos não é constante. Muitos outros devem acreditar que, para o caso das músicas populares, há um empobrecimento sistemático da músicas (letra e/ou melodia).

Para tentar investigar essa questão de uma maneira quantitativa, criamos uma base de dados com as músicas mais populares nos Estados Unidos segundo a revista *Billboard* [83] no período de 1946 a 2007 (Tabela 2.1). Usando essa base de dados composta por mais de 5 mil músicas, calculamos os valores dos parâmetros c e dos expoentes de Hurst h para cada música e, posteriormente, calculamos o valor médio ano a ano dessas quantidades. A Figura 2.13 mostra os resultados obtidos. Para o parâmetro c, observamos uma tendência de crescimento acentuada a partir dos anos 80 e para o expoente h verificamos uma tendência de decrescimento. Esse resultado indica que as amplitudes das músicas populares mais recentes então distribuídas muito mais próximas de uma gaussiana do que a anos atrás. De uma maneira menos clara, a redução no expoente de Hurst mostra uma redução nas correlações de longo alcance dessas músicas populares.

Calculamos também a entropia H e a complexidade C para cada música, bem como o seus valores médios em função dos anos, como mostram as Figuras 2.14 e 2.15. Independente de usarmos as amplitudes ou intensidades sonoras, notamos um considerável aumento da entropia H e uma considerável diminuição da complexidade C. Mais ainda, a evolução conjunta dos dois índices revela uma aproximação ao limiar de permutações aleatórias ($H = 1$ e $C = 1$). Esses resultados mostram, então, que os padrões ordinais das músicas populares tem se tornado cada vez mais aleatórios ou em outras palavras, menos complexos.

Essa nossa abordagem simplificada não deixa dúvidas que as músicas estão em evolução, o

que do ponto de vista qualitativo, pode ser até considerado trivial. Entretanto, nossos resultados são uma demonstração desse fato. Mais do que isso, nossas medidas apontam para um "empobrecimento estatístico" das músicas mais populares. Embora não tenhamos dados sobre os gêneros musicais dessas músicas populares, é possível que esses resultados sejam um reflexo da popularização de gêneros mais dançantes como o pop e techno. Por exemplo, se compararmos os resultados das Figuras 2.14 e 2.15 como os grupamentos de gêneros da Figura 2.12, observamos que a localização no plano C versus H das músicas populares estava dentro da região $G1$ na década de 50 (região dos gêneros de alto padrão) e atualmente encontra-se nas vizinhanças dos grupamentos $G3$ e $G4$ (região dos gêneros dançantes).

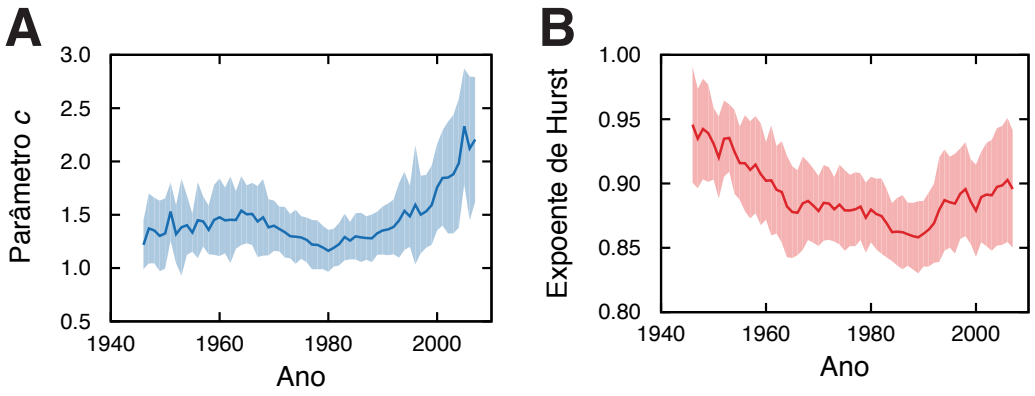


Figura 2.13: Mudanças no padrão das músicas mais populares nos Estados Unidos (*Billboard Top Charts*) medidas via evolução, ano a ano, do (**A**) parâmetro c e (**B**) do expoente de Hurst. A linhas contínuas representam o valor médio dos parâmetros em função do ano e as regiões sombreadas representam intervalos de confiança a 95%. Notamos a tendência de crescimento do parâmetro c (principalmente após os anos 80) e de decrescimento do expoente de Hurst, h. Esses resultados indicam que as músicas mais populares tem suas distribuições de probabilidade das amplitudes sonoras cada vez mais próximas de uma gaussiana. Os resultados indicam, ainda, que tem havido uma tendência de redução das correlações de longo alcance na série das intensidades sonoras.

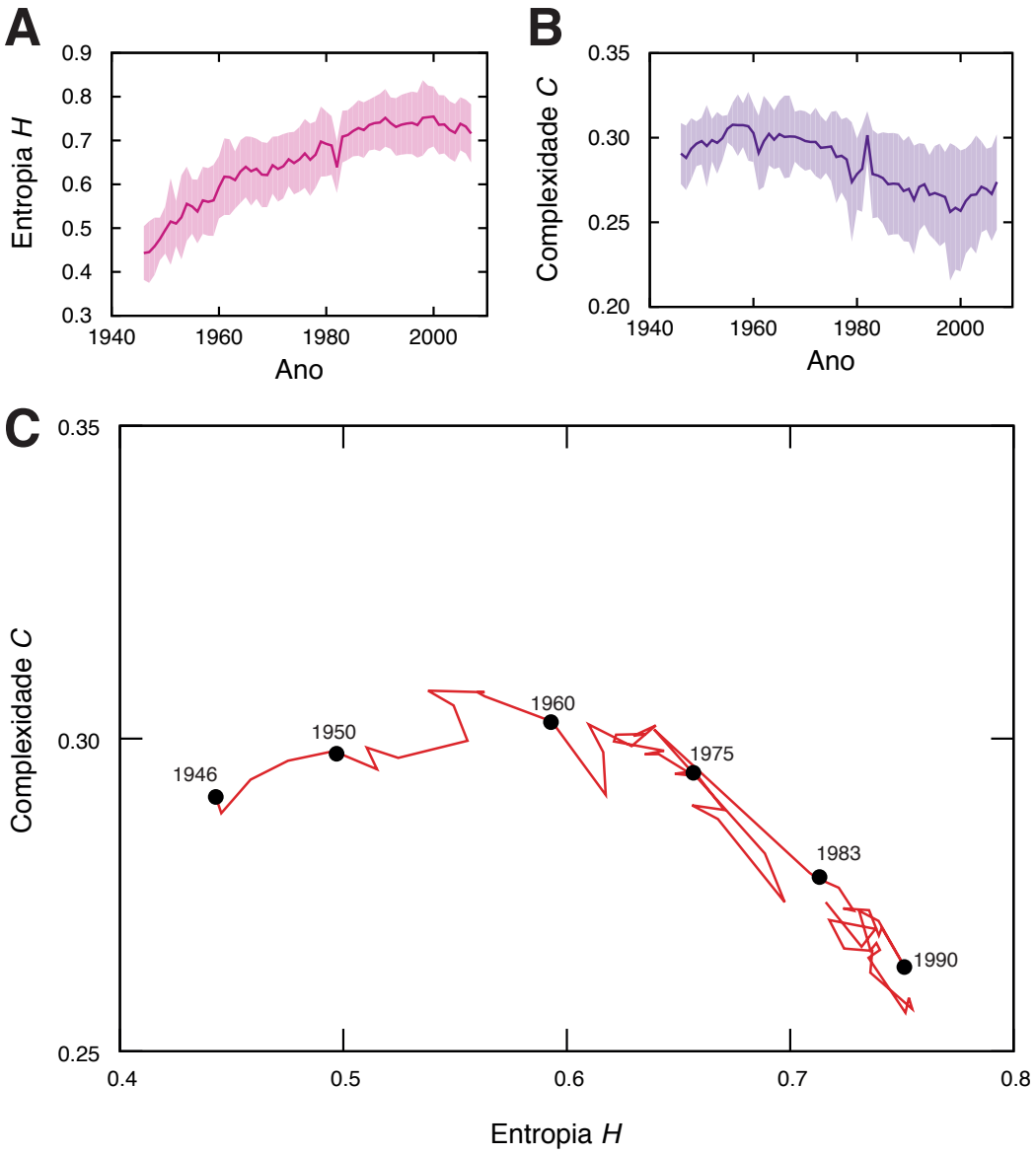


Figura 2.14: Mudanças no padrão das músicas mais populares nos Estados Unidos (*Billboard Top Charts*) medida via evolução, ano a ano, (**A**) da entropia de permutação H e (**B**) da complexidade C, calculadas para as séries das amplitudes sonoras. A linhas contínuas representam o valor médio dos índices entrópicos, em função do ano, e as regiões sombreadas representam intervalos de confiança a 95% (Apêndice B.2). Notamos que a entropia H tem aumentado com o passar dos anos e que a complexidade C tem diminuído. Isso indica que os valores de H e C das músicas mais populares têm se aproximado do limiar de permutações aleatórias ($H = 1$ e $C = 0$). Em (**C**) mostramos a evolução conjunta dos índices entrópicos H e C, na qual podemos observar mais claramente a aproximação ao limiar de permutações aleatórias.

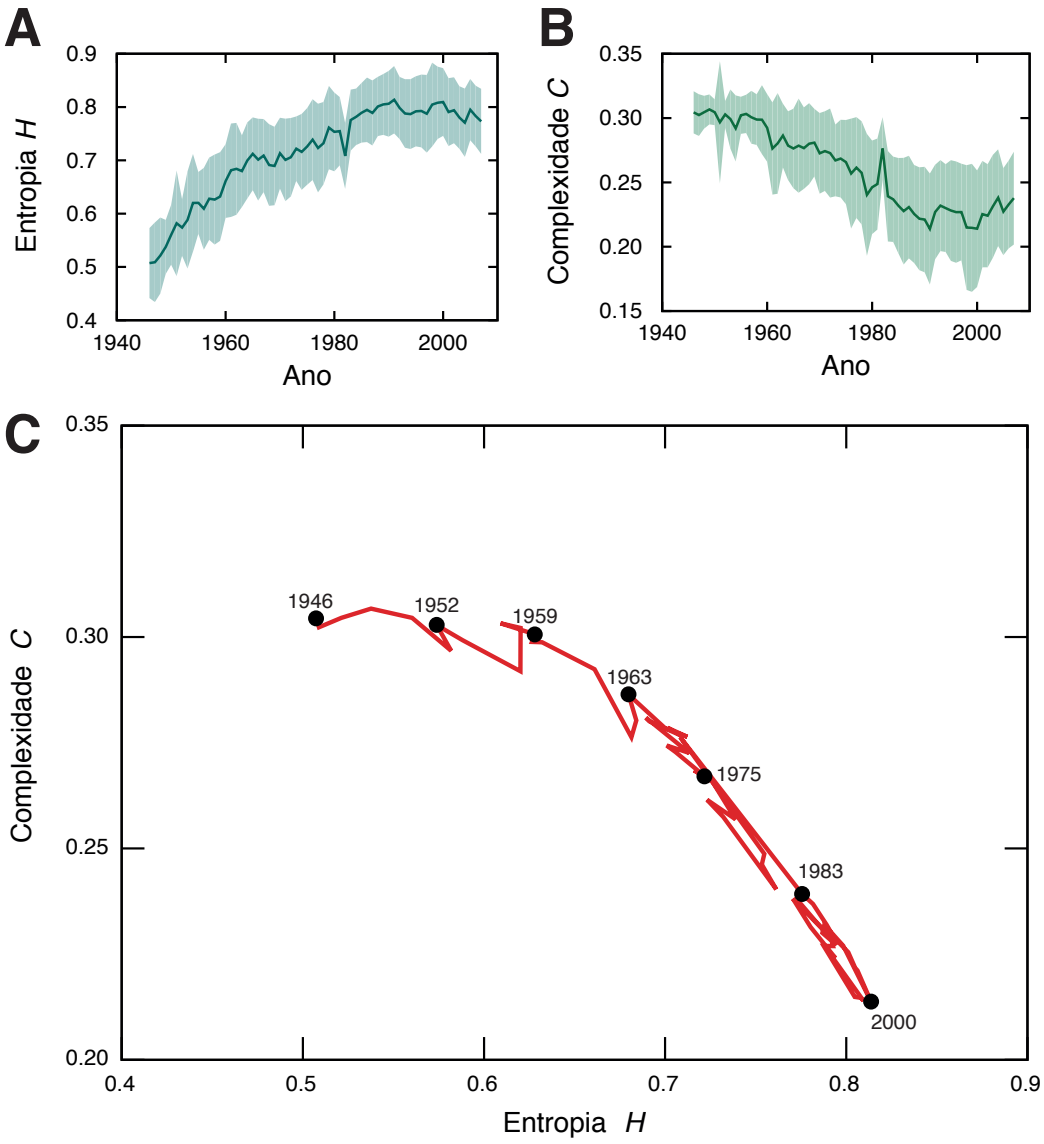


Figura 2.15: Mudanças no padrão das músicas mais populares nos Estados Unidos (*Billboard Top Charts*) medidas via evolução, ano a ano, (**A**) da entropia de permutação H e (**B**) da complexidade C, calculadas para as séries das intensidades sonoras. A linhas contínuas representam o valor médio dos índices entrópicos, em função do ano, e as regiões sombreadas representam intervalos de confiança a 95% (Apêndice B.2). Analogamente ao caso da Figura 2.14, a entropia H tem aumentado com o passar dos anos e que a complexidade C tem diminuído. Isso indica que os valores de H e C das músicas mais populares têm se aproximado do limiar de permutações aleatórias ($H = 1$ e $C = 0$). Em (**C**) mostramos a evolução conjunta dos índices entrópicos H e C, na qual podemos observar mais claramente a aproximação ao limiar de permutações aleatórias.

2.5 Conclusões e perspectivas

Neste capítulo, investigamos diversas propriedades dos sons musicais. Vimos que há um padrão bem definido para as distribuições das amplitudes sonoras, o qual é bem descrito por uma distribuição que, devido às condições dos dados, possui apenas um parâmetro. Imaginamos que esse parâmetro define uma espécie de distância para um processo gaussiano e, baseado nessa ideia, propusemos uma classificação de complexidade dos gêneros musicais. Confirmamos a existência de correlações de longo alcance nesses sons musicais e mostramos a existência de um acoplamento entre a forma da distribuição de probabilidade e os aspectos de correlação. Especificamente, vimos que à medida que a distribuição se aproxima de uma gaussiana, há uma diminuição da correlação nos sons.

Na tentativa de obter um procedimento para classificação de músicas passível de ser utilizado em uma situação mais prática, relacionada à detecção automática de gêneros, investigamos a dinâmica de ordenamento dos sons musicais. Empregamos dois índices entrópicos, baseados nas probabilidades relacionadas às possíveis permutações existentes nos sons. Vimos que essa abordagem é aplicável em um contexto mais prático. Em particular, os índices entrópicos agruparam os 10 gêneros musicais em 4 grupos de complexidade e também proporcionaram um bom percentual de acertos em um cenário simplificado de detecção automática de gêneros musicais.

Finalmente, estudamos de uma maneira quantitativa a evolução das músicas mais populares nos Estados Unidos. Verificamos que, de fato, há uma evolução nos padrões das músicas por meio dos parâmetros discutidos anteriormente. Além disso, os resultados sugerem um "empobrecimento estatístico", com o passar dos anos, das músicas mais populares.

Naturalmente, muitas outras análises podem ser feitas. Abordagens que acreditamos serem interessantes incluem análises de outras séries temporais mais ligadas à estrutura musical, como ritmo e harmonia. Outras análises mais parecidas com as que realizamos aqui (algumas já em andamento) incluem a investigação de multifractalidade, séries de intervalo de retorno e volatilidades.

Capítulo 3

Dinâmica difusiva da vantagem, aprendizado e erros em partidas de xadrez

Neste capítulo, estudaremos a dinâmica de partidas de xadrez sob dois pontos de vista. O primeiro deles será focado na evolução da vantagem dos jogadores movimento a movimento [17]. Veremos que essas trajetórias podem ser tratadas como as de partículas que executam um movimento errático. Essa "difusão" apresenta várias propriedades anômalas como distribuição não gaussiana, correlações de longo alcance e uma dependência não linear da variância com os movimentos. Devido a nossos dados cobrirem quase dois séculos de partidas, poderemos investigar também a evolução dessas características ao longo dos anos. Essa análise vai revelar um processo de aprendizagem nas aberturas e também um aumento nas diferenças técnicas entre os jogadores. O outro ponto de vista versa sobre os erros cometidos pelos jogadores [18]. Mostraremos que a distribuição da magnitude dos erros é bem descrita por uma log-normal e que a maioria, e os maiores erros, ocorrem durante o meio jogo. Classificaremos os erros em três tipos: notados, não notados e perdas de oportunidade. Essa classificação nos permitirá apontar que perceber os erros dos adversários é um fator dominante para a decisão do jogo.

3.1 Introdução e apresentação dos dados

Como já mencionamos, entender a dinâmica de sistemas complexos, sejam eles físicos, biológicos ou sociais, tem sido o foco de intensa pesquisa por mais de três décadas [98]. Entretanto, um aspecto que tem sido pouco explorado diz respeito à maneira pela qual os agentes aprendem a tomar decisões para lidar com a complexidade do sistema. Possivelmente, a carência de estudos nessa direção reflete a dificuldade em se quantificar tais situações e também em encontrar dados que cubram um período de tempo suficientemente longo para que esse comportamento seja observado de maneira mais conclusiva.

Na tentativa de compreender melhor como esses aspectos ocorrem no mundo real, vamos

estudá-los em um sistema modelo: o jogo de xadrez. O xadrez é um jogo de tabuleiro que fascina a humanidade desde sua invenção, possivelmente na Índia do século VI [99]. Trata-se de um jogo complexo com aproximadamente 10^{43} posições possíveis no tabuleiro e 10^{120} diferentes jogos [100]. Essa complexidade imita, ao menos em parte, as decisões da vida real. Além disso, é marcante o fato de que toda essa complexidade do xadrez surja de um conjunto pequeno de regras bem definidas. Estas características tornam o xadrez um excelente ponto de partida para estudar o processo de aprendizado [101, 102].

Outra característica importante é a facilidade de se encontrar bases bem documentadas de partidas de xadrez. Em particular, estudamos uma base com 73444 partidas de alto nível, que cobrem os dois últimos séculos da história do xadrez e também os torneios mais importantes do mundo (campeonatos mundiais, torneio de candidatos e famosos torneios abertos como o de Linares). Essa base é mantida pelo site PGN Mentor™ (`http://www.pgnmentor.com`) e está disponível gratuitamente. Ela consiste de arquivos PGN (*portable game notation*) que são arquivos de texto plano (sem formatação) contendo o registro de todos os movimentos e outros detalhes da partida, como os nomes dos jogadores, data da partida e nome do torneio.

Cada uma das partidas foi analisada por um programa de computador chamado Crafty™ [103], que calcula a vantagem relativa das brancas movimento a movimento. Esse procedimento de avaliação posicional é, em geral, feito por meio de uma função definida por regras heurísticas que levam em conta a diferença ponderada das peças (vantagem material) e outras características mais teóricas do jogo de xadrez, como mobilidade, segurança do Rei, controle do centro do tabuleiro (vantagem posicional). A vantagem $A(m)$, em função do movimento m, é normalmente medida em unidades de peão (peça de menor valor do jogo), o que significa que na ausência de fatores posicionais, $A(m)$ varia em uma unidade sempre que um peão é capturado. Além disso, $A(m)$ é positiva quando o jogador com as peças brancas está em vantagem na partida e negativa quando o jogador com as peças negras é quem está em vantagem. Essa é a função que os programas de xadrez procuram maximizar/minimizar na busca do melhor movimento. A Figura 3.1 mostra a evolução da vantagem das brancas $A(m)$ para 50 partidas selecionadas aleatoriamente a partir da nossa base de dados. Observamos que $A(m)$ assemelha-se ao movimento errático de partículas difusivas. Essa analogia visual será a base de nossas investigações na próxima seção.

3.2 A difusão da vantagem nas partidas de xadrez

Iniciamos nossa análise investigando o valor médio da vantagem $A(m)$, ao longo dos movimentos m, para os três possíveis resultados de uma partida, *i.e.*, vitória das brancas (33% das partidas), vitória da negras (24% das partidas) e empates (43% das partidas). Esse valor médio é definido como

$$\langle A(m) \rangle = \frac{1}{N} \sum_{i}^{N} A_i(m) \,, \tag{3.1}$$

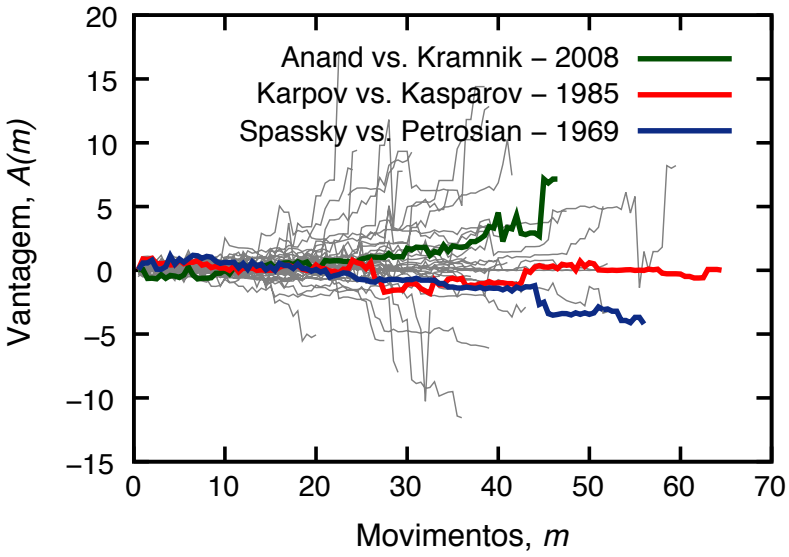


Figura 3.1: Evolução da vantagem das brancas $A(m)$ em função dos movimentos m. Conforme a notação padrão dos jogos de xadrez, um "movimento completo" m corresponde a um movimento das brancas e um das negras. Aqui, usamos valores fracionados para representar o movimento das brancas e valores inteiros para representar o movimento da negras; assim, $m = 1/2$ corresponde ao primeiro movimento das brancas, $m = 1$ ao primeiro movimento das negras (completando um "movimento completo"), e assim por diante.

sendo que o índice i rotula as partidas com um dado resultado e N é a quantidade dessas partidas. A Figura 3.2 mostra os resultados dessa análise. Naturalmente, para as partidas em que as brancas venceram, o valor médio de $A(m)$ é positivo e aumenta com o passar dos movimentos. Similarmente, para as partidas em que as negras venceram, $A(m)$ é negativo e decresce com o passar dos movimentos. Mais interessante é o comportamento das partidas que terminaram em empate. Nesse caso, o valor médio de $A(m)$ é sistematicamente positivo e, embora pequeno, o valor é estatisticamente significativo (notamos as pequenas barras de erro que representam intervalos de confiança a 95%). Observamos também que, mesmo em partidas em que as negras venceram, o valor inicial de $A(m)$ é positivo.

Essa figura pode ser a resposta para um debate histórico entre jogadores e especialistas em xadrez: existe vantagem em jogar de brancas?[1] Alguns jogadores e teóricos do jogo dizem que devido o primeiro movimento ser das brancas, elas possuem a "iniciativa" e as negras devem lutar para igualar a situação. Outros argumentam que as negras são beneficiadas pelo primeiro movimento das brancas, uma vez que a revelação desse movimento fornece informações para as negras de como será a partida.

Na tentativa de responder a essa pergunta, especialistas em xadrez normalmente reportam que as brancas vencem cerca de 10% mais partidas do que as negras. Embora essa percentagem seja uma evidência da vantagem das brancas, não é possível quantificar a magnitude dessa vantagem apenas com os resultados das partidas. Nossa análise, além de indicar a existência de vantagem

[1] Para evitar a expressão longa "jogar com as peças brancas", usamos a expressão comum entre os enxadristas "jogar de brancas"

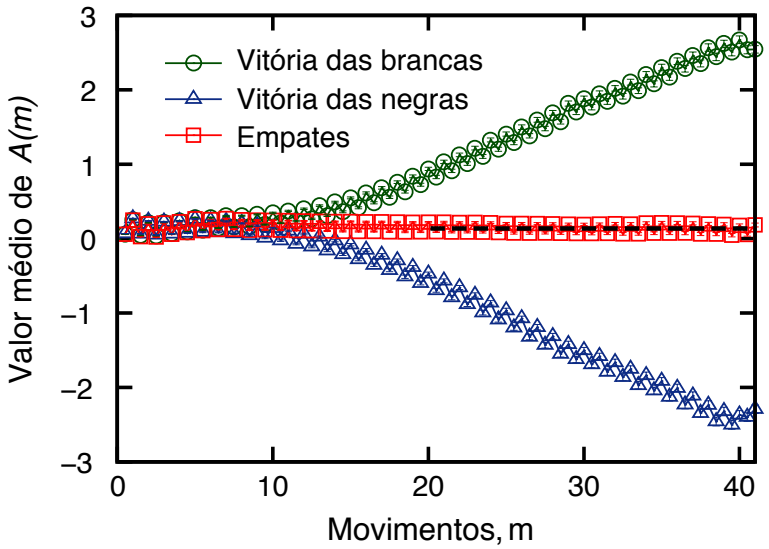


Figura 3.2: Valor médio da vantagem das brancas $A(m)$ (equação 3.1) em função dos movimentos m, para os três possíveis resultados dos jogos. A linha tracejada representa o valor estimado da vantagem inicial das brancas ($0{,}14 \pm 0{,}01$ peões).

em jogar de brancas devido ao valor positivo de $A(m)$, permite estimar numericamente essa vantagem tomando, por exemplo, o valor médio da média de $A(m)$, ao longo dos movimentos, para as partidas que terminaram em empate. Esse valor é de $0{,}14 \pm 0{,}01$ peões. Valores muito próximos são encontrados ao se calcular essa média para partidas que terminaram em vitórias das brancas ou das negras.

Em adição ao valor médio da vantagem, calculamos a evolução da variância de $A(m)$ para caracterizar o processo de difusão. A variância é definida como

$$\langle (A(m) - \langle A(m) \rangle)^2 \rangle = \frac{1}{N-1} \sum_{i}^{N} \left(A_i(m) - \frac{1}{N} \sum_{j}^{N} A_i(m) \right)^2 , \tag{3.2}$$

sendo que os índices i e j rotulam as partidas com um dado resultado e N é a quantidade dessas partidas.

Se considerarmos que m é o análogo do tempo, podemos comparar esses resultados com aqueles relacionados à difusão "física". Em particular, sabemos que para o movimento browniano usual a variância é uma função linear do tempo, *i.e.*, $\langle (A(m) - \langle A(m) \rangle)^2 \rangle \sim m$ — um reflexo da ausência de memória e da existência de uma escala característica para as flutuações. Embora essa situação seja muito comum na natureza, também existem outros comportamentos possíveis. Um caso que também é comum é uma dependência do tipo lei de potência :

$$\langle (A(m) - \langle A(m) \rangle)^2 \rangle \sim m^{\alpha} . \tag{3.3}$$

Esses casos são normalmente chamados de "difusão anômala" [104]. O caso $0 < \alpha < 1$, que é comum em sistemas porosos ou desordenados, é chamado de sub-difusão. A situação descrita por $\alpha > 1$ é denominada superdifusão e ocorre, por exemplo, em sistemas turbulentos. Para

$\alpha = 2$ temos a chamada difusão balística, que pode representar partículas difusivas a velocidades constantes. Além disso, valores de $\alpha > 2$ são também conhecidos por hiperdifusão [105] e representam, em geral, sistemas muito distantes do equilíbrio.

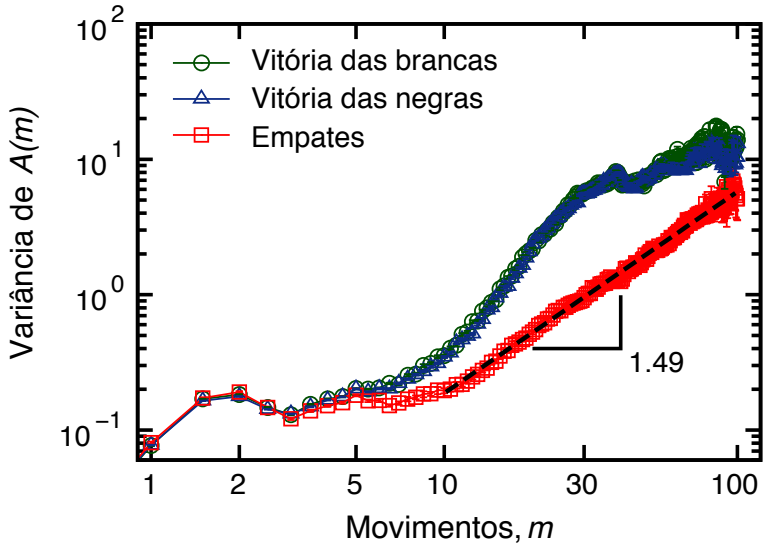


Figura 3.3: Variância da vantagem das brancas $A(m)$ (equação 3.2) em função dos movimentos m, para os três possíveis resultados dos jogos. A linha tracejada representa o regime superdifusivo observado para as partidas que terminaram em empate, ou seja, $\langle (A(m) - \langle A(m) \rangle)^2 \rangle \sim m^\alpha$ com $\alpha = 1{,}49 \pm 0{,}01$.

A Figura 3.3 mostra os resultados para os três possíveis resultados do jogo de xadrez. Notamos um cenário difusivo rico e caracterizado por mais de um regime difusivo, a depender do número de movimentos. Para os primeiros movimentos, observamos que praticamente não há difusão ou espalhamento da vantagem para todos os resultados. Esse período do jogo corresponde à *abertura*, um estágio muito estudado pelos enxadristas, no qual existem sequências de movimentos bem definidas que, quando bem jogadas, levam a posições equilibradas. Após esse estágio inicial, a variância exibe um regime difusivo mais rápido que o usual. No caso das partidas que terminaram em empate, a variância é bem descrita por uma lei de potência $\langle (A(m) - \langle A(m) \rangle)^2 \rangle \sim m^\alpha$ para $m \gtrsim 10$ e com expoente $\alpha = 1{,}49 \pm 0{,}01$. Embora não existam diferenças no perfil da variância quando consideramos vitórias das brancas ou das negras, é notório que o comportamento é mais complexo nesses casos. Em particular, para $10 \lesssim m \lesssim 30$ o processo é aproximadamente hiperdifusivo e para $m \gtrsim 40$ temos um comportamento próximo ao caso dos empates.

Para investigar melhor o comportamento difusivo das partidas e a diferença entre vitórias e empates, calculamos a variância de $A(m)$ após agrupar as partidas por tamanho, como mostra a Figura 3.4. Notamos que, se por um lado, o comportamento da variância não é dependente do tamanho da partida para os empates, por outro, temos um padrão bastante interessante para o caso das vitórias. À medida que o tamanho da partida aumenta, o perfil da variância para vitórias se torna parecido com o perfil dos empates. A principal diferença entre vitória e empate parece estar nos últimos movimentos, nos quais pode haver uma espécie de efeito avalanche que faz a vantagem passar por grandes flutuações. Esse efeito pode estar ligado ao fato de que ao atingir

vantagem suficiente para ganhar, o jogador com a vantagem tende a trocar as peças de modo a simplificar o jogo e também adquirir mais vantagem.

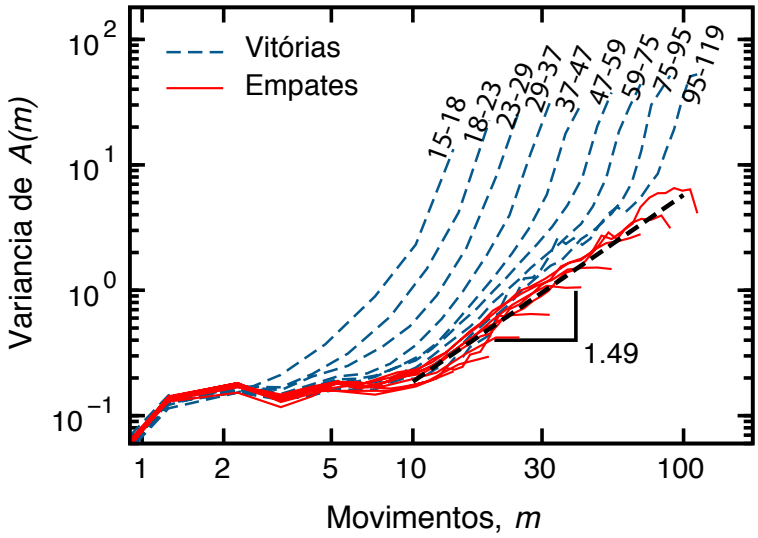


Figura 3.4: Variância da vantagem das brancas $A(m)$ (equação 3.2) em função dos movimentos m, após agrupar as partidas por tamanho. O valores usados para construir essa partição são mostrados na figura e a linha tracejada representa o mesmo regime superdifusivo da Figura 3.3.

Além do valor médio e da variância, podemos calcular a distribuição de probabilidade de $A(m)$. Para isso, consideramos a variável normalizada

$$\xi(m) = \frac{A(m) - \langle A(m) \rangle}{\langle (A(m) - \langle A(m) \rangle)^2 \rangle} \,, \tag{3.4}$$

e calculamos a distribuição de probabilidade acumulada (a probabilidade de se encontrar um valor maior ou igual a ξ). Para esse cálculo, agrupamos as partidas em vitórias e empates e consideramos ora os valores positivos ora os valores negativos de $\xi(m)$ para construir as caudas positivas e negativas das distribuições acumuladas. A Figura 3.5 mostra essas distribuições ao considerar valores de m entre 10 e 70 movimentos. Podemos ver que há um bom colapso das distribuições, o que indica que as vantagens $A(m)$ são estatisticamente autossimilares, visto que, após normalizadas, elas seguem a mesma forma universal de distribuição. Claramente, essas distribuições não são gaussianas (linhas tracejadas da Figura 3.5); em particular, observamos que as caudas das distribuições decaem muito mais lentamente do que uma gaussiana. A forma das distribuições também é um pouco diferente quando consideramos vitórias ou empates; entretanto, até onde tentamos, não foi possível descrever essas distribuições por nenhuma forma funcional.

Outra questão interessante, que investigaremos agora, é se existe ou não memória de longo alcance na série das vantagens. Para verificar essa possibilidade, consideramos as séries dos incrementos da vantagem, *i.e.*, $\Delta A(m) = A(m{+}0{,}5) - A(m)$, para todas as 5154 partidas terminadas em empate e que duraram mais de 50 movimentos. Aplicamos DFA (Apêndice A.3) nessas séries a fim de obter o expoente de Hurst, h. A Figura 3.6A mostra um exemplo de análise

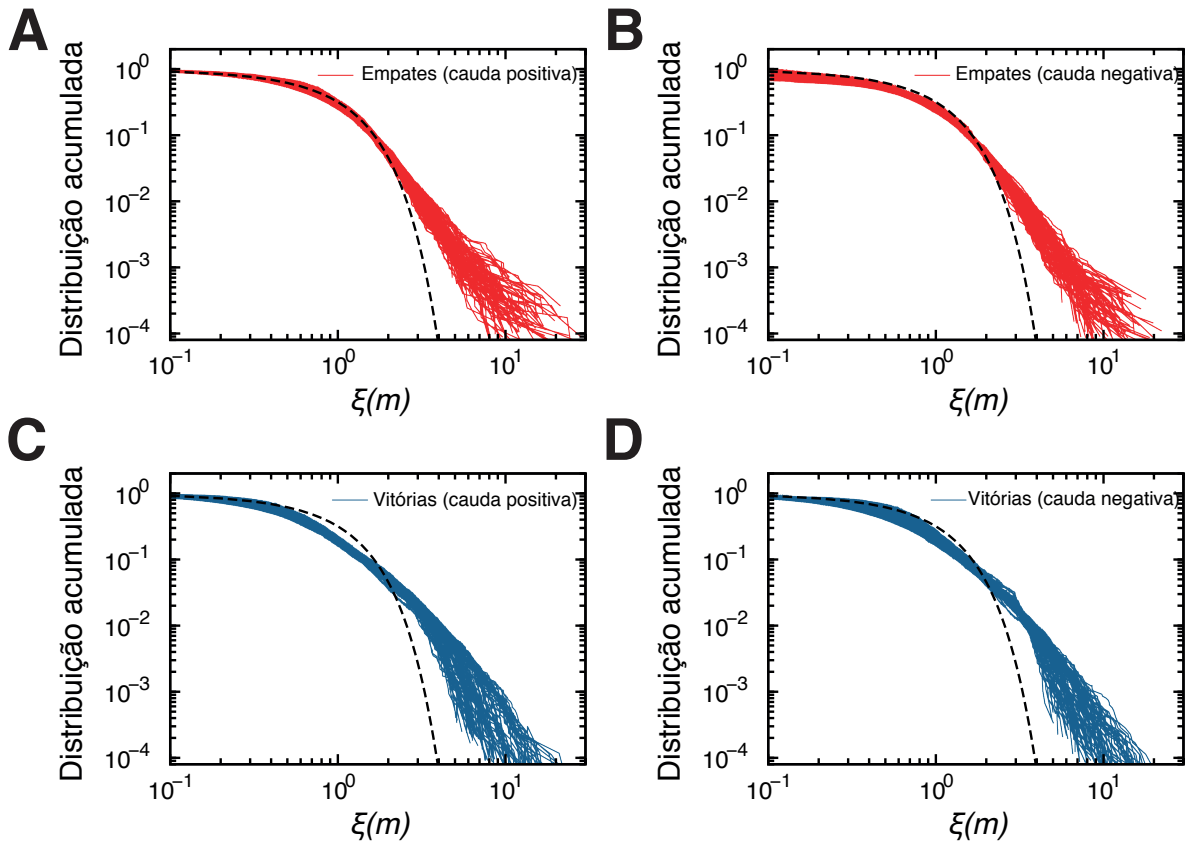


Figura 3.5: Distribuições de probabilidade acumuladas da variável normalizada $\xi(m)$ (equação 3.4) ao considerar os valores (**A**) positivos e (**B**) negativos de $\xi(m)$ para partidas que terminaram em empate e os valores (**C**) positivos e (**D**) negativos de $\xi(m)$ para partidas que terminaram em vitória. Em todos os gráficos, a linha tracejada representa a distribuição acumulada gaussiana de média zero e variância unitária (veja também o Apêndice D).

DFA para uma partida escolhida aleatoriamente da nossa base de dados. Notamos que, para essa partida, encontramos $h = 0{,}354$. O resultado para todas as demais partidas é mostrado na forma de um histograma dos valores de h na Figura 3.6B. Essa figura mostra que os expoentes h estão distribuídos gaussianamente em torno de $h = 0{,}35$ com desvio padrão igual a $0{,}1$. Portanto, a grande maioria das partidas apresenta um expoente de Hurst menor do $0{,}5$, o que indica a existência de memória de longo alcance nas séries dos incrementos. Mais do que isso, o valor menor do que $0{,}5$ também aponta a presença de antipersistência nas séries dos incrementos, ou seja, valores alternados de $\Delta A(m)$ aparecem com probabilidade muito maior do que em um processo aleatório, o que reflete a maneira alternada dos movimentos do jogadores. No caso de partidas terminadas em vitórias, o valor médio de h é um pouco menor, como mostra a Figura 3.6C. Entretanto, ao remover os últimos 5 movimentos dessas partidas, observamos um comportamento muito próximo ao das partidas terminadas em empate. Esse resultado mostra que, do ponto de vista da correlação, uma vitória difere de um empate apenas nos últimos lances, similarmente ao que verificamos para a variância de $A(m)$ (Figura 3.4).

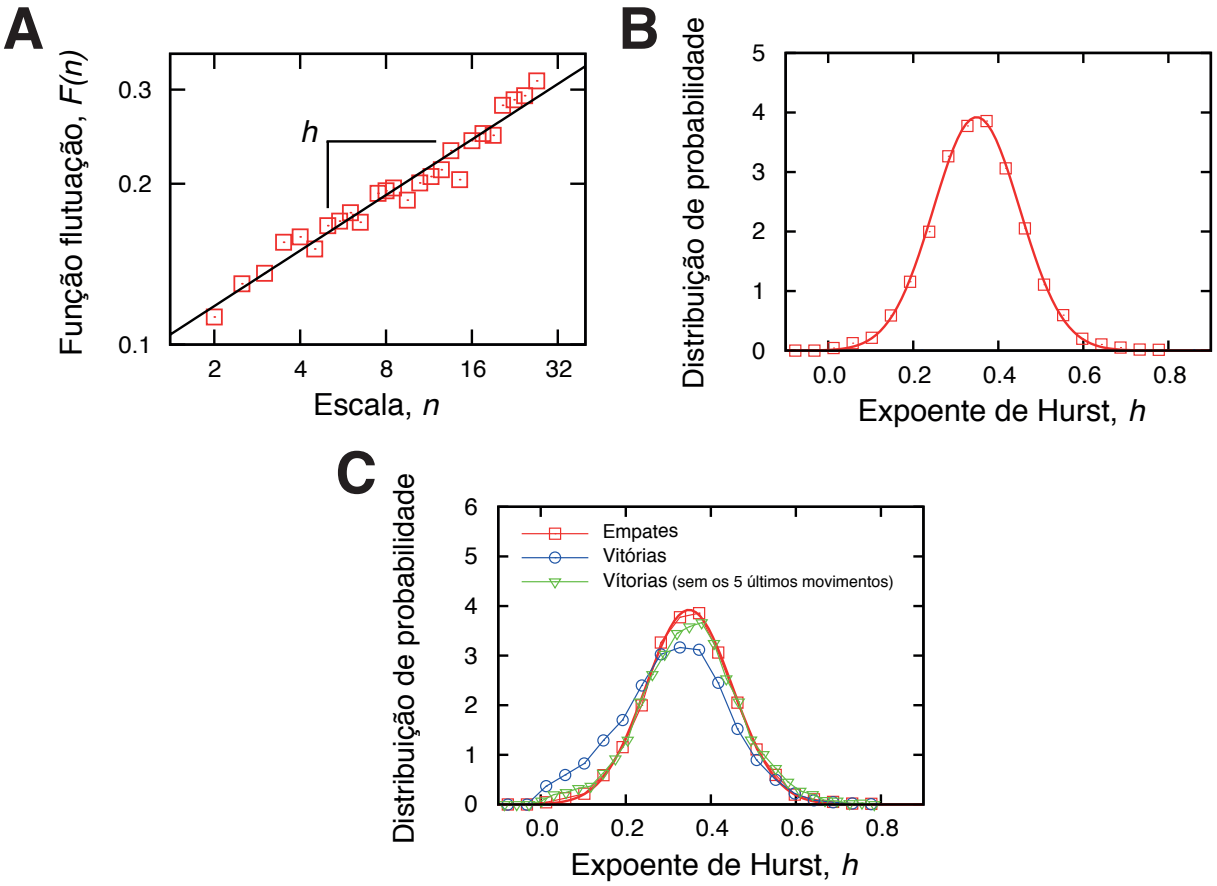


Figura 3.6: (**A**) Exemplo de uma análise DFA dos incrementos da vantagem, *i.e.*, $\Delta A(m) = A(m + 0{,}5) - A(m)$, para uma partida escolhida aleatoriamente da nossa base de dados. Notamos que a função de flutuação segue uma lei de potência ($F(m) \sim m^h$) e, portanto, nesse gráfico log-log, a relação entre F e m é aproximada por uma reta de coeficiente angular numericamente igual ao expoente de Hurst, h (neste caso $h = 0{,}345$). (**B**) Distribuição de probabilidade do expoente de Hurst (h) para todas as partidas com mais de 50 movimentos e que terminaram em empate. A linha contínua é um ajuste gaussiano aos dados, com média $0{,}35$ e desvio padrão $0{,}1$. (**C**) Comparação da distribuição de probabilidade do expoente de Hurst para empates (quadrados), vitórias (círculos) e vitórias após remover os últimos 5 movimentos (triângulos). O valores médios de h são $0{,}35$ para empates, $0{,}31$ para vitórias e $0{,}35$ para vitórias removendo-se os últimos 5 movimentos.

3.3 Tendências históricas no xadrez de alto nível

As regras do jogo de xadrez não mudam desde o século XIX. Essa estabilidade pode contribuir para o aumento da popularidade do jogo, que pode ser medida pelo número de jogadores olímpicos e títulos de Grande Mestre (a maior honraria que um enxadrista pode ganhar), como mostra Figura 3.7A. Além disso, o perfil dos jogadores tem mudado. É cada vez mais comum o aparecimento de jogadores muito jovens, diminuindo até mesmo a idade média para se o obter o título de Grande Mestre, como mostra a Figura 3.7B. Por outro lado, a habilidade dos jogadores olímpicos (medidas pelo Elo *rating* [106]) tem se mantido aproximadamente constante, enquanto o desvio padrão da habilidade dos jogadores apresenta uma forte tendência de crescimento, como mostram as Figuras 3.7C e 3.7D. Todas essas mudanças na "demografia" dos jogadores de xadrez sugere uma pergunta: até que ponto o padrão difusivo reportado na seção anterior é estável ou é possível verificar alguma tendência histórica?

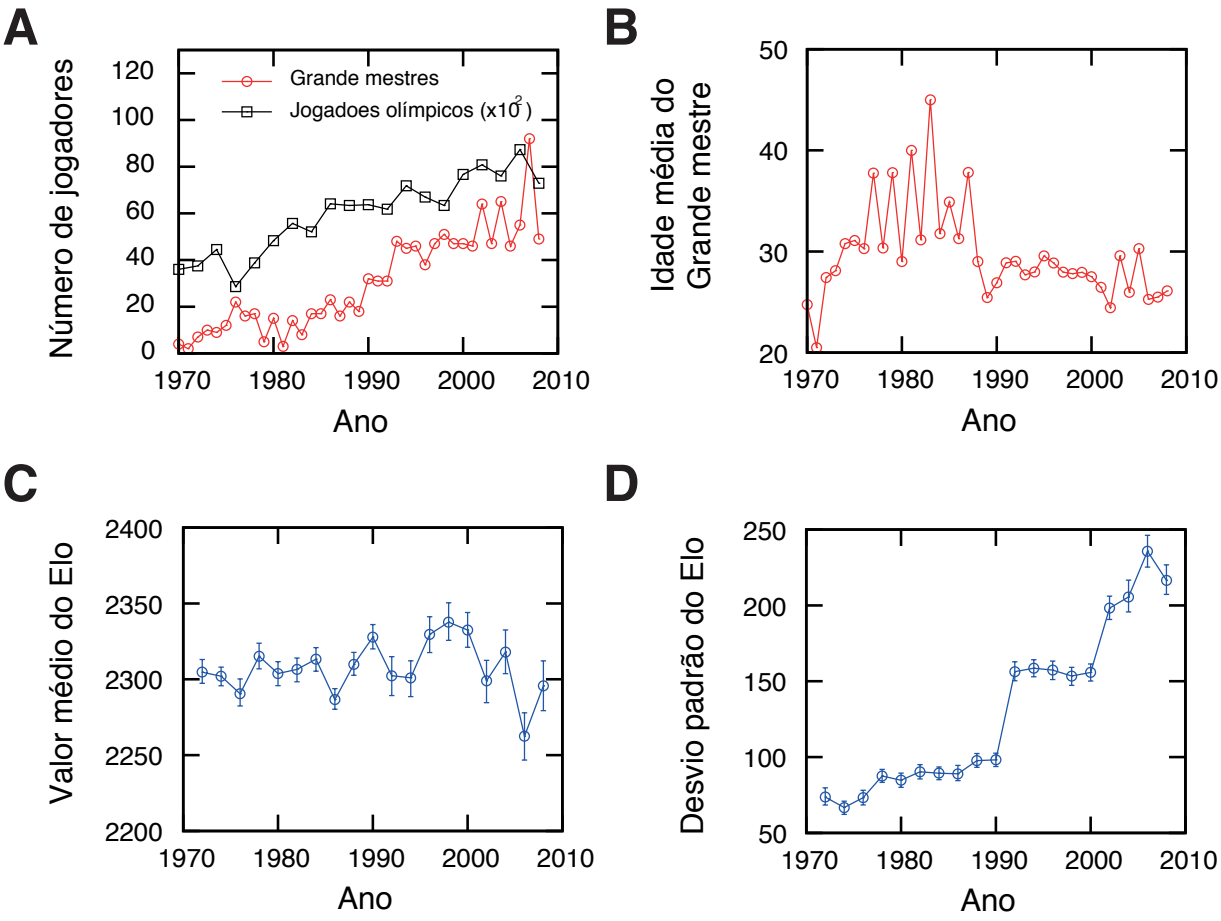


Figura 3.7: (**A**) Número de novos títulos de Grande Mestre concedidos pela Federação Mundial de Xadrez (`http://fide.com`) e número de jogadores que participaram das Olimpíadas de Xadrez (`http://www.olimpbase.org`) desde 1970. (**B**) Idade média dos jogadores ao receberem o título de Grande Mestre. (**C**) Valor médio e (**D**) desvio padrão do Elo *rating* dos jogadores que participaram das Olimpíadas de Xadrez.

Para responder a essa questão, agrupamos as partidas da nossa base em quatro períodos de tempo: 1857 a 1918, 1919 a 1949, 1950 a 1980 e 1981 a 2011. Usando os dados desses

subconjuntos, calculamos o perfil do valor médio da vantagem $A(m)$, como mostra a Figura 3.8A. Aqui, para melhor visualização, mostramos apenas três períodos e aplicamos um filtro de média móvel, usando uma janela que contém um movimento das brancas e um movimento das negras. Nesta figura, as linhas horizontais representam o valor médio do valor médio da vantagem $A(m)$ para $20 < m < 40$ e as regiões sombreadas são intervalos de confiança a 95%. Notamos que os valores apresentam diferenças estatisticamente significativas, mostrando que a vantagem inicial das brancas tem aumentado ao longo dos anos. De fato, como mostra a Figura 3.8B, a vantagem inicial das brancas cresce não linearmente ao longo dos anos. Esse perfil de crescimento pode ser bem ajustado por uma aproximação exponencial à assíntota de $0{,}23 \pm 0{,}01$ peões, com um tempo característico de $67{,}0 \pm 0{,}1$ anos. Os resultados da Figura 3.8B sugerem então que os jogadores estão aprendendo maneiras de maximizar essa vantagem inicial das brancas que também parece possuir uma limitação assintótica. Além disso, o longo tempo característico para a evolução dessa vantagem indica o quão difícil é o processo de disseminar conhecimento em nível populacional.

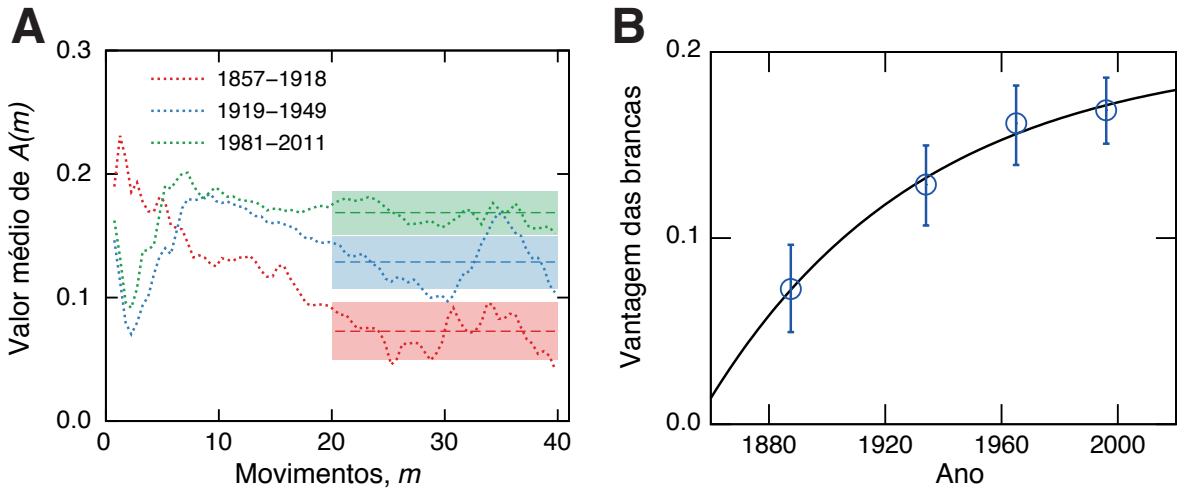


Figura 3.8: (A) Valor médio da vantagem $A(m)$ para partidas que terminaram em empate e em três períodos de tempo distintos. Essas curvas foram suavizadas usando-se um filtro de média móvel, com janela de tamanho 2. A linha horizontal representa o valor médio dos pontos para $20 < m < 40$ e as regiões sombreadas são intervalos de confiança a 95%. **(B)** Evolução temporal da vantagem das brancas em partidas que terminaram em empate. A linha sólida representa uma aproximação exponencial a um valor assintótico. Os valores do platô e do tempo característico são $0,23 \pm 0,01$ peões e $67,0 \pm 0,1$ anos, respectivamente.

Em adição aos valores médios de $A(m)$, investigamos também a evolução temporal do perfil da variância de $A(m)$ para partidas que terminaram em empate. A Figura 3.9A mostra que esse perfil tem mudado ao longo dos anos. Em particular, o expoente superdifusivo α parece estar se aproximando de dois, *i.e.*, uma difusão balística. De fato, como mostra a Figura 3.9B, o expoente α pode ser bem descrito por uma aproximação exponencial à assíntota $\alpha = 1{,}9 \pm 0{,}1$ com um tempo característico de 128 ± 9 anos. Suspeitamos que essa tendência possui relações estreitas com o aumento na diferença típica entre as qualidades dos jogadores (vale lembrar que a Figura 3.7A mostra que a variância do Elo tem aumentado ao longo do anos). Na verdade, a simples presença de diferentes qualidades ou *fitness* em um processo difusivo pode dar origem

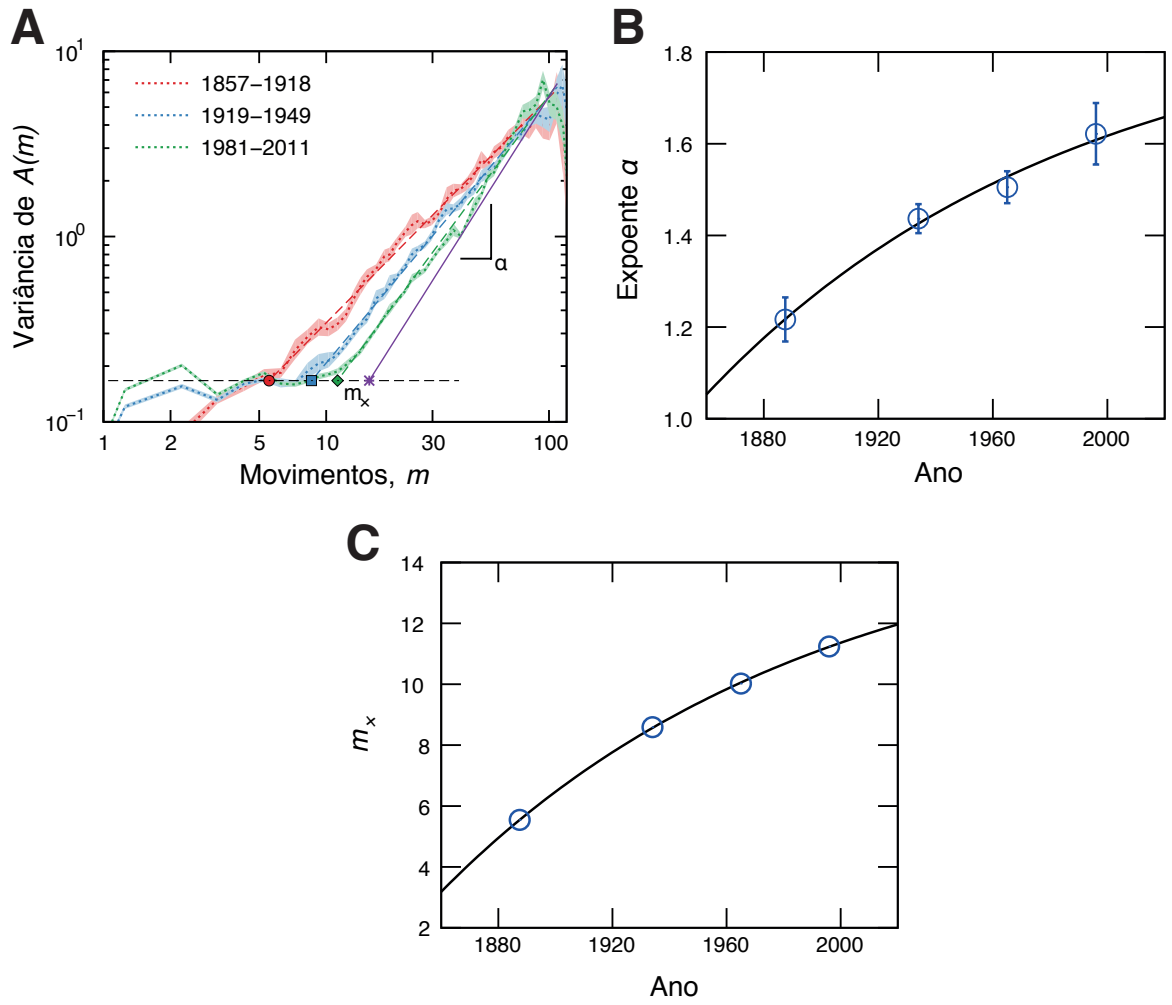


Figura 3.9: **(A)** Variância da vantagem $A(m)$ em três períodos de tempo. As regiões sombreadas são intervalos de confiança a 95% e a linhas tracejadas representam os ajustes da lei de potência em cada período. A linha horizontal tracejada representa o valor médio da variância considerando o conjunto mais recente de dados e $1 < m < 10$. Os símbolos sobre a linha horizontal indicam os valores de $m_\times$, o número de movimentos a partir do qual a difusão muda de regime. Evolução temporal do **(B)** expoente α e do **(C)** movimento característico $m_\times$. Nesses dois últimos gráficos, as linhas sólidas são ajustes de uma aproximação exponencial a um valor assintótico. Para α o valor assintótico é $\alpha = 1{,}9 \pm 0{,}1$ e o tempo característico é 128 ± 9 anos. Para $m_\times$ o valor assintótico é $m_\times = 15{,}6 \pm 0{,}6$ e o tempo característico é 130 ± 12 anos. Baseado na conjectura de que α e $m_\times$ estão aproximando seus valores assintóticos, adicionamos a figura **(A)** uma linha contínua para representar o regime difusivo limitante.

ao regime de difusão balístico [107]. Para uma ilustração de como esse comportamento pode ocorrer, assumiremos que a equação

$$A_i(m + 0{,}5) = A_i(m) + \Phi_i + \eta(m) \tag{3.5}$$

descreve a vantagem do jogador de brancas na partida i, sendo que a diferença entre as qualidades dos dois jogadores é Φ_i e $\eta(m)$ é uma variável aleatória gaussiana, não correlacionada e de média nula. Valores de $\Phi_i > 0$ levam a um *drift* positivo em $A_i(m)$, descrevendo uma situação em que

o jogador de brancas é melhor. Por outro lado, $\Phi_i < 0$ adiciona um *drift* negativo em $A_i(m)$, representando o caso em que o jogador de negras é melhor. Nessas condições, e assumindo ainda que Φ_i é sorteado de uma distribuição com variância igual a σ_Φ^2, podemos mostrar que (Apêndice C)

$$\langle (A(m) - \langle A(m) \rangle)^2 \rangle \sim \sigma_\Phi^2 \, m^2 \tag{3.6}$$

e, portanto, $\alpha = 2$.

No caso do xadrez, o cenário difusivo não é totalmente definido pelo *fitness* dos jogadores. Contudo, esse certamente é um ingrediente essencial e, assim, esse modelo fornece ao menos uma explicação parcial para os dados da Figura 3.9A, sugerindo que o aumento na diferença da qualidade entre os jogadores tem influência no perfil da variância de $A(m)$.

Outra característica marcante da Figura 3.9A é o deslocamento para valores positivos do movimento característico $m_\times$ no qual o regime superdifusivo começa. Para estimar essa evolução, consideramos que $m_\times$ é a intersecção (símbolos da Figura 3.9A) entre os regimes de lei de potência e o valor médio da variância para $1 < m < 10$ do período mais recente dos dados (linha horizontal da Figura 3.9A). Os valores de $m_\times$ em função do período considerado são mostrados na Figura 3.9A. Observamos que a evolução de $m_\times$ é bem descrita por um aproximação exponencial ao valor assintótico de $15{,}6 \pm 0{,}6$ movimentos, a um tempo característico de 130 ± 12 anos. Essa evolução também pode estar relacionada a mecanismos de aprendizagem, visto que o aumento de $m_\times$ indica um aumento da fase inicial do jogo, *i.e.*, o tamanho das sequências de aberturas do jogos de xadrez têm ficado mais longas, o que pode ser um reflexo da popularização de programas de xadrez que incentivam, principalmente, o estudo dessa fase inicial do jogo.

Investigamos também a estabilidade do perfil das distribuições da Figura 3.5. Para isso, calculamos as distribuições acumuladas para $10 < m < 70$ em cada período de tempo e, em seguida, calculamos o valor médio dessas distribuições. Os resultados são mostrados na Figura 3.10, na qual observamos que o perfil das distribuições praticamente não mudou ao longo dos anos. Verificamos também que os valores médios de h têm evoluído ao longo dos anos, como mostra a Figura 3.11. Observamos que, embora não haja uma evolução tão clara quanto as mostradas nas Figuras 3.8 e 3.9, os valores médios de h são significativamente (do ponto de vista estatístico) menores para períodos mais recentes. Esse resultado sugere que o comportamento antipersistente se intensificou com o passar dos anos.

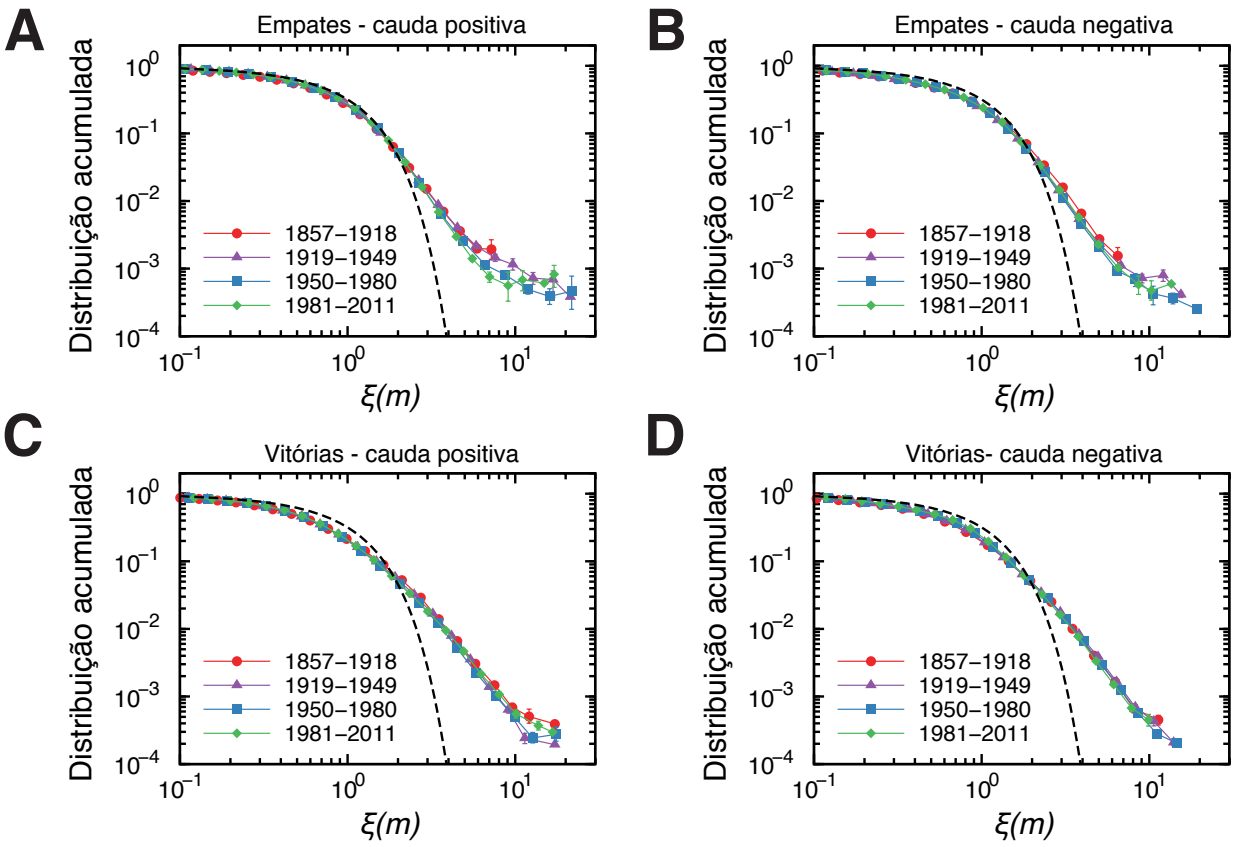


Figura 3.10: Valor médio das distribuições de probabilidade acumuladas da variável normalizada $\xi(m)$ em cada período de tempo. Em (**A**) consideramos os valores positivos e em (**B**) os negativos de $\xi(m)$ para partidas que terminaram em empate. Em (**C**) e (**D**) mostramos o análogo das figuras anteriores para partidas que terminaram em vitória. As linhas tracejadas representam a distribuição acumulada gaussiana de média zero e variância unitária.

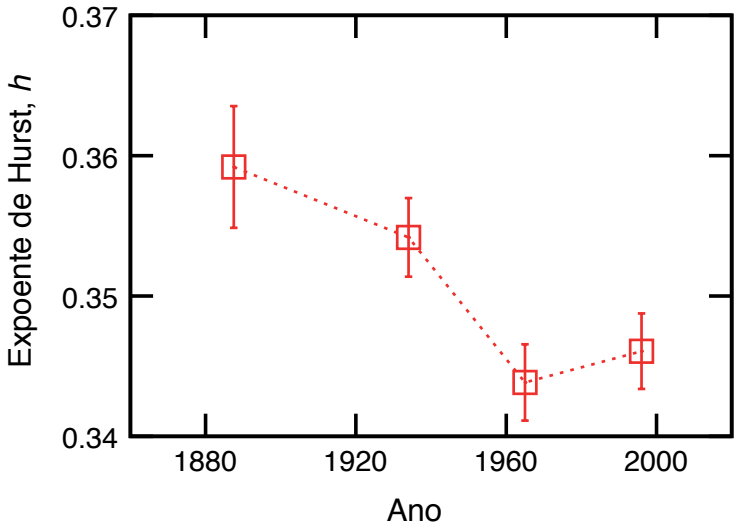


Figura 3.11: Mudanças históricas no valor médio do expoente h. Notamos que os valores médios de h são significativamente menores para os períodos mais recentes.

3.4 Dinâmica dos erros dos jogadores de xadrez

Em geral, além da vantagem das brancas $A(m)$, os programas de xadrez também permitem a identificação de possíveis erros cometidos durante uma partida de xadrez. Sempre que o programa encontrar um movimento para o qual $A^*(m) > A(m)$ durante o movimento das brancas ou $A^*(m) < A(m)$ durante o movimento das negras, consideramos que o jogador cometeu um erro. Vamos ainda definir a magnitude do erro como sendo $\varepsilon = |A^*(m) - A(m)|$, ou seja, o módulo da diferença entre a vantagem obtida pelo programa e a obtida pelo jogador. Para essa análise, por uma questão de tempo computacional, limitamos nossa base de dados aos torneios do Campeonato Mundial de Xadrez, torneio de candidatos e torneios interzonais — os mais importantes torneios do xadrez mundial. Além disso, usamos o programa de xadrez Critter™ [108] para determinar os erros. Esse programa encontra-se melhor posicionado na classificação *Computer Chess Rating Lists* (CCRL - `http://www.computerchess.org.uk/ccrl`), ocupando atualmente o quarto lugar, enquanto o Craft™ ocupa a vigésima nona colocação. Atualmente, mesmo programas mais simplificados de xadrez (como Pocket Fritz™ - `http://chessbase-shop.com/en/products/pocket_fritz_4`) possuem um nível de jogo muito alto, o que torna a avaliação das posições de jogo bastante precisas. A Figura 3.12 mostra os erros encontrados em duas partidas do Campeonato Mundial de Xadrez. Nesse gráfico, os pontos vermelhos mostram os movimentos em que o programa encontrou uma posição melhor e a distância desse ponto à curva da vantagem define a magnitude do erro.

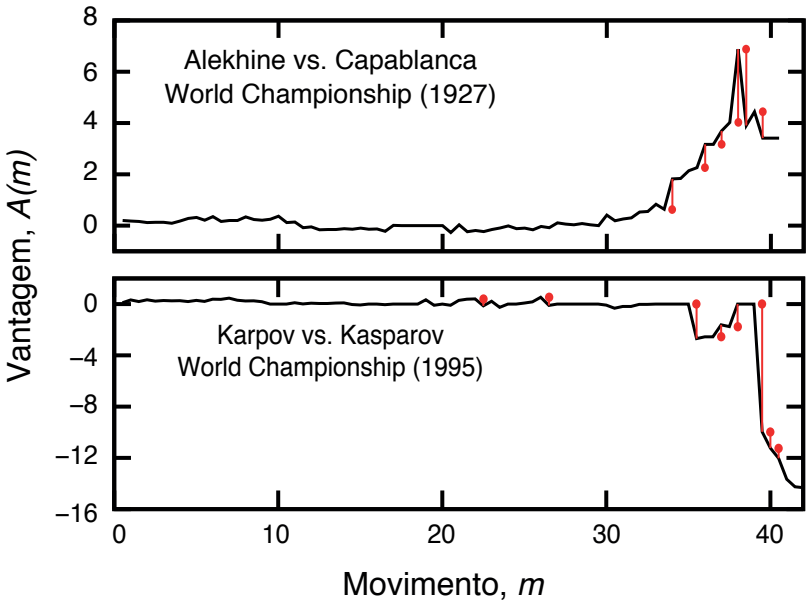


Figura 3.12: Vantagem das brancas $A(m)$ em função dos movimentos m para duas partidas do Campeonato Mundial de Xadrez. Os pontos vermelhos indicam os movimentos em que o programa de xadrez propôs um lance melhor. O valor absoluto entre a vantagem obtida pelo programa a obtida pelo jogador define a magnitude ε do erro.

Iniciamos a análise dos erros calculando a distribuição de probabilidade da variável ε para cada jogador em nossa base dados. A Figura 3.13 mostra as distribuições acumuladas para o

jogador M. Carlsen nas partidas em que ele perdeu, ganhou ou empatou. Em todos os três casos, a distribuição de ε foi bem descrita por uma log-normal, da forma

$$p(\varepsilon) = \frac{1}{\varepsilon\,\sigma\,\sqrt{2\pi}} \exp\left[\frac{1}{2}\left(\frac{\log(\varepsilon)-\mu}{\sigma}\right)^2\right], \tag{3.7}$$

sendo que μ e σ são parâmetros da distribuição. Não surpreendentemente, a distribuição dos erros no caso das derrotas encontra-se descolada mais para valores positivos, ou seja, Carlsen cometeu erros maiores nas partida em que foi derrotado. Entretanto, é inusitado o fato de que a distribuição dos erros cometidos nas partidas vencidas esteja mais descolado para valores positivos do que os erros cometido em partidas terminadas em empate. Esse fato intrigante será esclarecido mais adiante.

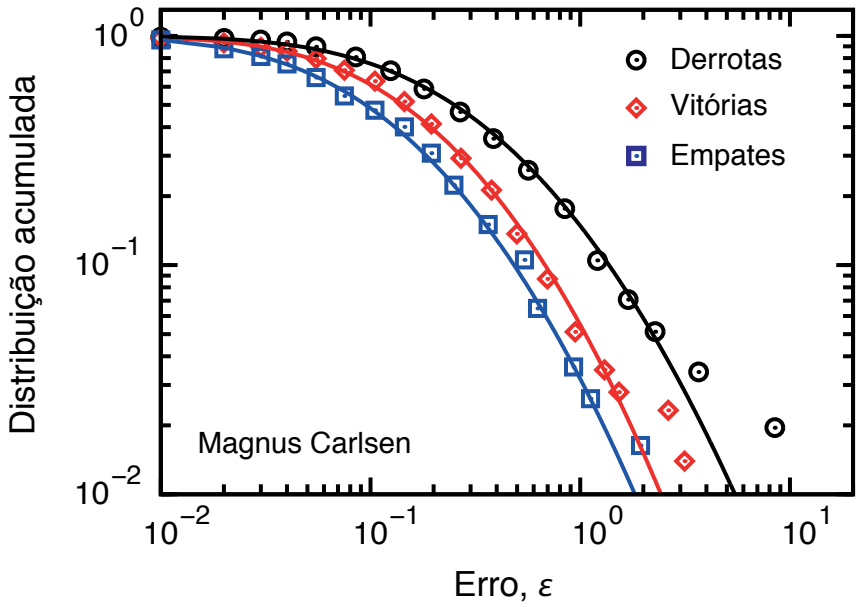


Figura 3.13: Distribuição de probabilidade acumulada da magnitude dos erros ε cometidos por Magnus Carlsen (atual melhor jogador do mundo) em partidas em que ele perdeu (círculos), ganhou (losangos) ou empatou (quadrados). As linhas contínuas são distribuições de probabilidade log-normais (equação 3.7) com $\mu = -1{,}38$ e $\sigma = 1{,}32$ para derrotas, $\mu = -1{,}95$ e $\sigma = 1{,}22$ para vitórias e $\mu = -2{,}35$ e $\sigma = 1{,}.27$ para empates. Os valores p do teste de Kolmogorov-Smirnov são $0{,}82$ para derrotas, $0{,}54$ para vitórias, e $0{,}51$ para empates; portanto, a hipótese log-normal não pode ser rejeitada (veja também o Apêndice D).

Os demais 324 jogadores nesse subconjunto de nossa base de dados mostram um comportamento muito semelhante ao obtido para M. Carlsen, ou seja, a distribuição dos erros é muito próxima de uma log-normal com parâmetros μ e σ ajustados para cada jogador. De fato, quando aplicamos o teste de Kolmogorov-Smirnov para testar a hipótese de que os erros ε sejam distribuídos de acordo com um log-normal, não podemos rejeitar a hipótese em 76% dos jogadores nas partidas terminadas em empate, 92% para vitórias e 96% para derrotas. Esse crescimento no percentual está possivelmente relacionado a efeitos de tamanho da amostragem, visto que o

número de erros é menor nas partidas empatadas do que nas partidas terminadas em vitórias ou derrotas.

Para visualizar todas as distribuições de probabilidade dos erros em uma única curva, podemos fazer uma operação de escala em todas as distribuições de probabilidade. Notemos que, se considerarmos a variável

$$\xi = \exp\left(\frac{\log(\varepsilon) - \mu}{\sigma}\right), \tag{3.8}$$

a distribuição da equação 3.7 torna-se

$$p(\xi) = \frac{1}{\xi\sqrt{2\pi}} \exp\left[\frac{1}{2}\log(\xi)^2\right], \tag{3.9}$$

a qual não possui nenhum parâmetro. A Figura 3.14 mostra as distribuições para todos os jogadores quando consideramos o erro normalizado ξ. Observamos que há um bom colapso das distribuições de probabilidade e também um bom acordo com a distribuição da equação 3.9. Esse resultado e os altos percentuais de não rejeição da hipótese log-normal mostram que os erros dos jogadores podem ser considerados como um processo aleatório multiplicativo, *i.e.*,

$$\varepsilon(t+1) = \varepsilon(t)\eta(t), \tag{3.10}$$

sendo $\eta(t)$ uma variável aleatória não correlacionada de variância finita e t os movimentos em que ocorrem o erro. Essa dependência multiplicativa dos tamanhos dos erros pode também ser a origem do crescimento rápido da variância da vantagem que observamos na Figura 3.4, pois, escolhendo-se valores convenientes para a média e a variância do ruído $\eta(t)$ é possível obter grandes flutuações em ε e, consequentemente, em $A(m)$.

Retornando aos dados da Figura 3.13, é muito curioso o fato de que existe uma probabilidade maior de se encontrar grandes erros em partidas terminadas em vitórias do que em empates. Esse comportamento nos motivou a investigar os erros mais detalhadamente, como mostra a Figura 3.15. Essa figura ilustra a evolução da vantagem e os erros na partida entre G. Kasparov (jogando de brancas) e A. Karpov, disputada pelo Campeonato Mundial de Xadrez de 1987. Nela é possível verificar a existência de diferentes "tipos" de erros. Por exemplo, no movimento $m = 22{,}5$ o programa Critter indica que Kasparov poderia ter mantido a vantagem próxima de zero ao invés de tê-la feito diminuir para aproximadamente -4 peões. Entretanto, o programa Critter também indica que Karpov cometeu um erro no próximo movimento ($m = 23$) o que fez a vantagem retornar para aproximadamente zero. O erro de Karpov foi não notar o erro de Kasparov. Assim, definimos o erro de Kasparov em $m = 22,5$ como sendo do tipo "não notado" (*UE, unnoticed error*) e o erro de Karpov em $m = 23$ como sendo do tipo "perda de oportunidade" (*LO, lost of opportunity*). Notemos que em $m = 26{,}5$ e $m = 27$ também temos um *UE* e um *LO*, respectivamente. Embora sejam bastante raras, em caso de sequências de erros com mais de dois movimentos, consideramos a sequência *UE-LO-UE-*... para definir os tipos de erros. O outro tipo de erro que definimos é o erro notado (*NE, noticed error*), que

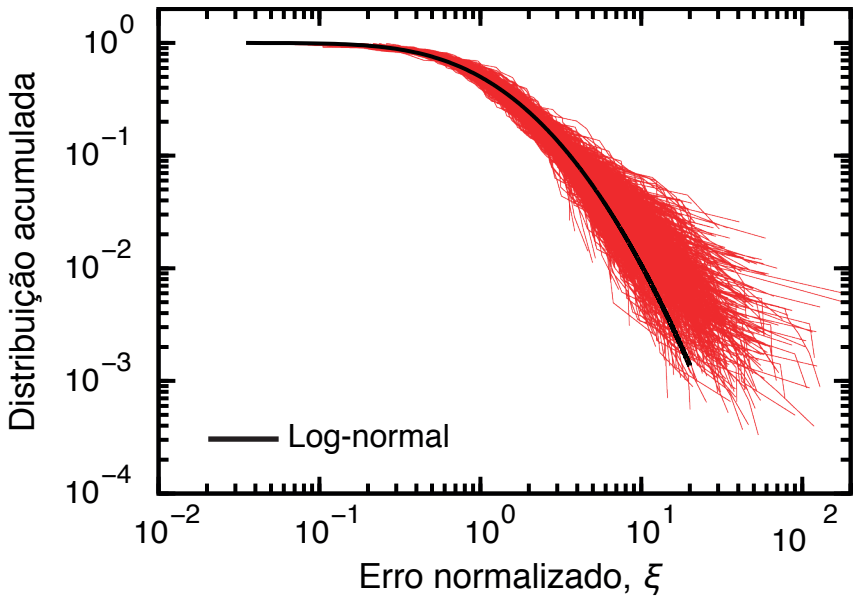


Figura 3.14: Distribuição de probabilidade acumulada da magnitude dos erros normalizados ξ (equação 3.8) para todos os jogadores e resultados. A linha contínua é uma distribuição log-normal com $\mu = 0$ e $\sigma = 1$ (equação 3.9).

vai ocorrer sempre isolado como os que ocorreram em $m = 28{,}5$ e $m = 29{,}5$. Observamos que nesses dois movimentos, Kasparov poderia ter mantido a vantagem mais próxima de zero e não o fez. Karpov, por sua vez, tomou vantagem nesses dois erros, o que torna o erro de Kasparov do tipo "notado".

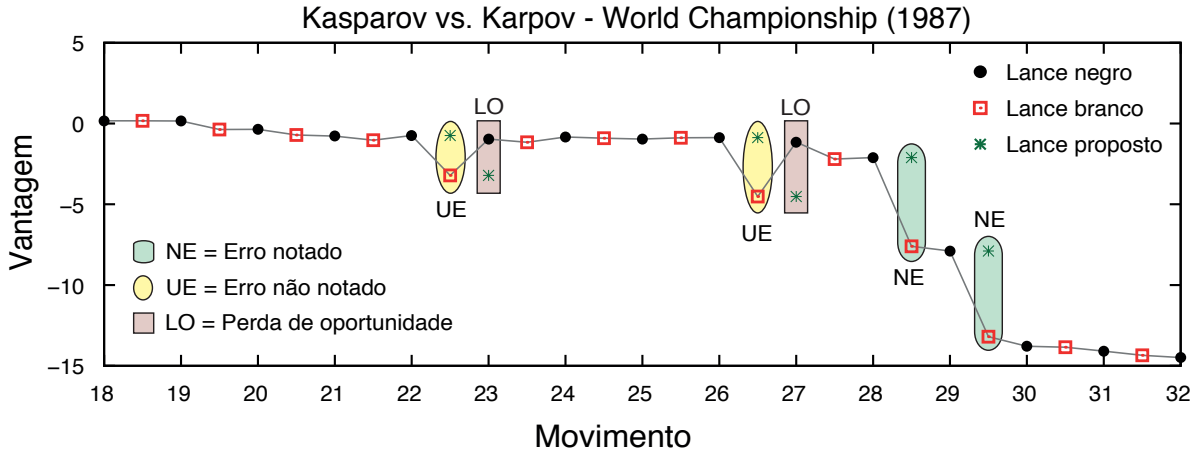


Figura 3.15: Evolução da vantagem e definição dos três tipos de erros

Analisamos, então, como esses três tipos de erros estão localizados ao longo dos movimentos para os três possíveis resultados das partidas. A Figura 3.16 mostra os resultados obtidos. Nessa figura, temos um gráfico das magnitudes dos erros em função do movimento, sendo que consideramos os erros das brancas como valores negativos e os erros das negras como valores positivos. Além disso, o código de cores indica a ocorrência de mais erros em uma dada região. Vemos que, independentemente do resultado ou do tipo de erro, são poucos e em menor magnitude

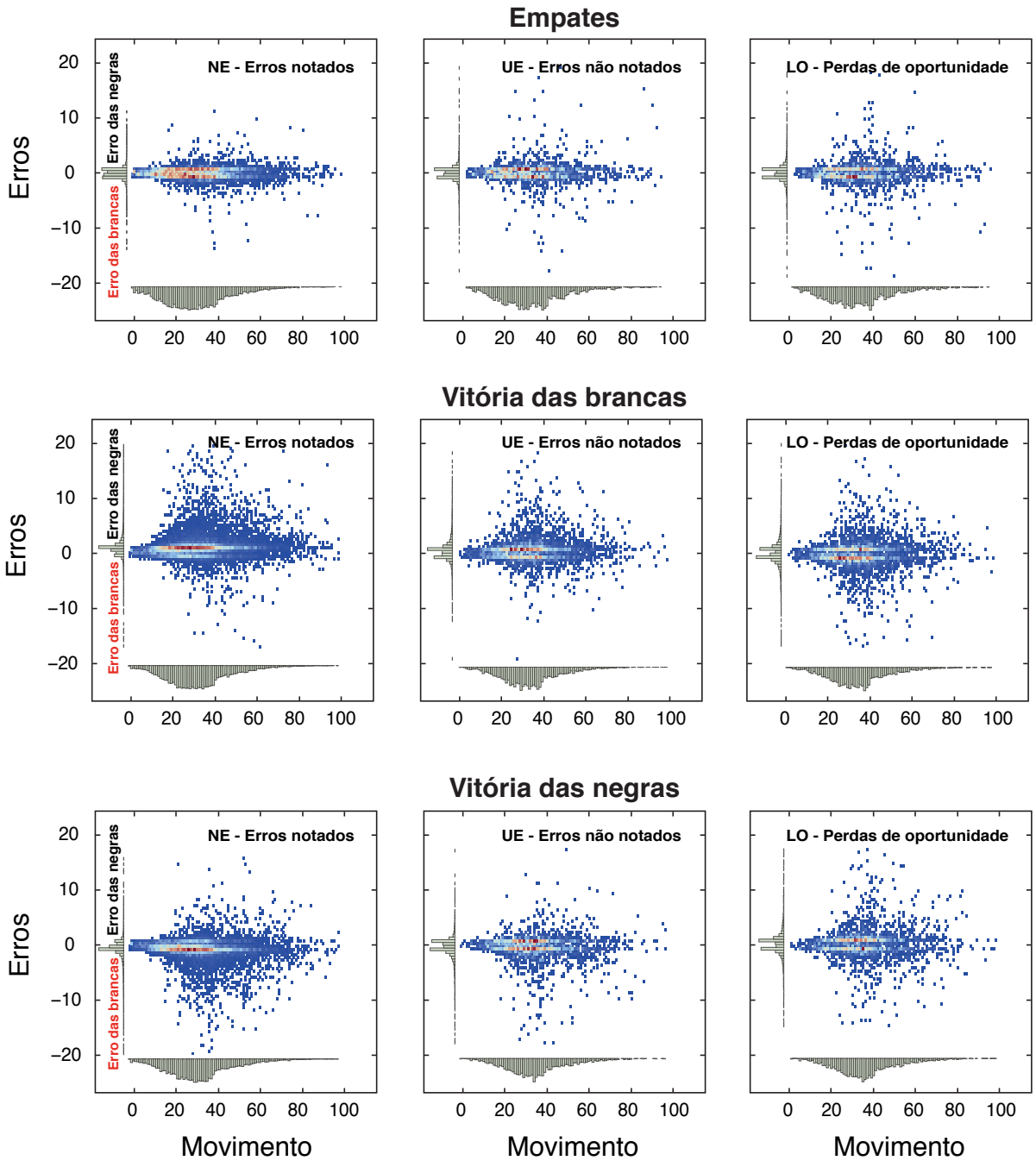


Figura 3.16: Magnitude dos erros versus o movimento em que erros ocorreram. Valores negativos indicam os erros cometidos pelas brancas e valores positivos indicam os erros cometidos pelas negras. Os tipos de erros e os possíveis resultados são mostrados nos gráficos. O código de cores indica as regiões em que há uma concentração maior de erros (regiões avermelhadas).

os erros para $m \lesssim 10$; essa região corresponde à abertura do jogo na qual, como já dissemos, existem sequências bem estabelecidas de movimentos que levam a uma situação equilibrada. Por outro lado, a grande maioria dos erros está localizada entre $10 \lesssim m \lesssim 50$ — região do meio de jogo, na qual ainda existem muitas peças no tabuleiro e uma infinidade de possibilidades a serem analisadas pelos jogadores. Para $m \gtrsim 50$, o número e a magnitude dos erros diminui,

refletindo o fato de que nesse estágio final do jogo normalmente há um número menor de peças e, consequentemente, uma simplificação do jogo.

Observamos também que as magnitudes (e também a quantidade) dos erros são consideravelmente menores nas partidas terminadas em empates. No caso das vitórias, percebemos que quando as brancas vencem os erros das negras foram, em maioria, do tipo *NE*. O análogo ocorre quando as negras vencem (neste caso a maioria dos erros das brancas também é do tipo *NE*). Isso indica que notar os erros é um fator fundamental para vencer a partida. Além disso, é possível perceber que existem mais erros do tipo *LO* das brancas quando elas venceram e mais erros do tipo *LO* das negras quando elas venceram. Esse resultado explica porque a distribuição dos erros nas partidas terminadas em vitórias está mais deslocada para valores positivos do que a distribuição dos erros em partidas terminadas em empates.

Para investigar melhor o papel dos erros nos resultados das partidas, calculamos as frações de cada tipo de erro maiores do que um certo valor ε' para cada resultado. Por exemplo, para as partidas terminadas em vitórias das brancas calculamos as frações

$$f_{NE} = \frac{\#_{\varepsilon'}\{NE\}}{\#_{\varepsilon'}\{NE + UE + LO\}}, \quad f_{UE} = \frac{\#_{\varepsilon'}\{UE\}}{\#_{\varepsilon'}\{NE + UE + LO\}} \quad \text{e} \quad f_{LO} = \frac{\#_{\varepsilon'}\{LO\}}{\#_{\varepsilon'}\{NE + UE + LO\}}.$$

Nas quais $\#_{\varepsilon'}\{\dots\}$ representa o número de erros de um dado tipo com magnitude maior do ε'. A Figura 3.17 compara essas frações entre os jogadores de brancas e de negras.

Na Figura 3.17A, mostramos as frações para os erros notados (*NE*). Para as vitórias das brancas observamos que o percentual de *NE* das negras ($\sim 60\%$) é consideravelmente maior do que o das brancas ($\sim 40\%$). Basicamente o oposto ocorre nas partidas vencidas pelas negras. Nas partidas terminadas em empates, os percentuais são praticamente os mesmos para as brancas e negras. Esse resultado confirma a hipótese de que notar os erros do adversário é um fator determinante para vencer a partida.

As Figuras 3.17B e 3.17C mostram os resultados para os erros não notados (*UE*) e perdas de oportunidade (*LO*). Notemos que, mesmo nas partidas vencidas pelas brancas, o percentual de *UE* é maior para as negras e o percentual de *LO* é maior para brancas. O oposto ocorre em caso de vitórias das negras. Esse resultado indica que mesmo nas partidas vencidas, brancas ou negras falharam em aproveitar oportunidades que poderiam ter levado a vitórias mais fáceis ou mais expressivas. Isso também reflete o alto nível dos programas de xadrez, uma vez que eles encontraram posições melhores mesmo nas partidas com vitórias.

Para os empates, não observamos diferenças significativas nos percentuais de *NE*, *UE* e *LO* para as brancas e negras.

3.5 Conclusões e perspectivas

Estudamos neste capítulo vários aspectos relacionados à vantagem nas partidas de xadrez. Em uma primeira parte, focamos no estudo da evolução temporal da vantagem relativa das brancas. Tratamos essa vantagem como um processo difusivo e essa abordagem revelou características

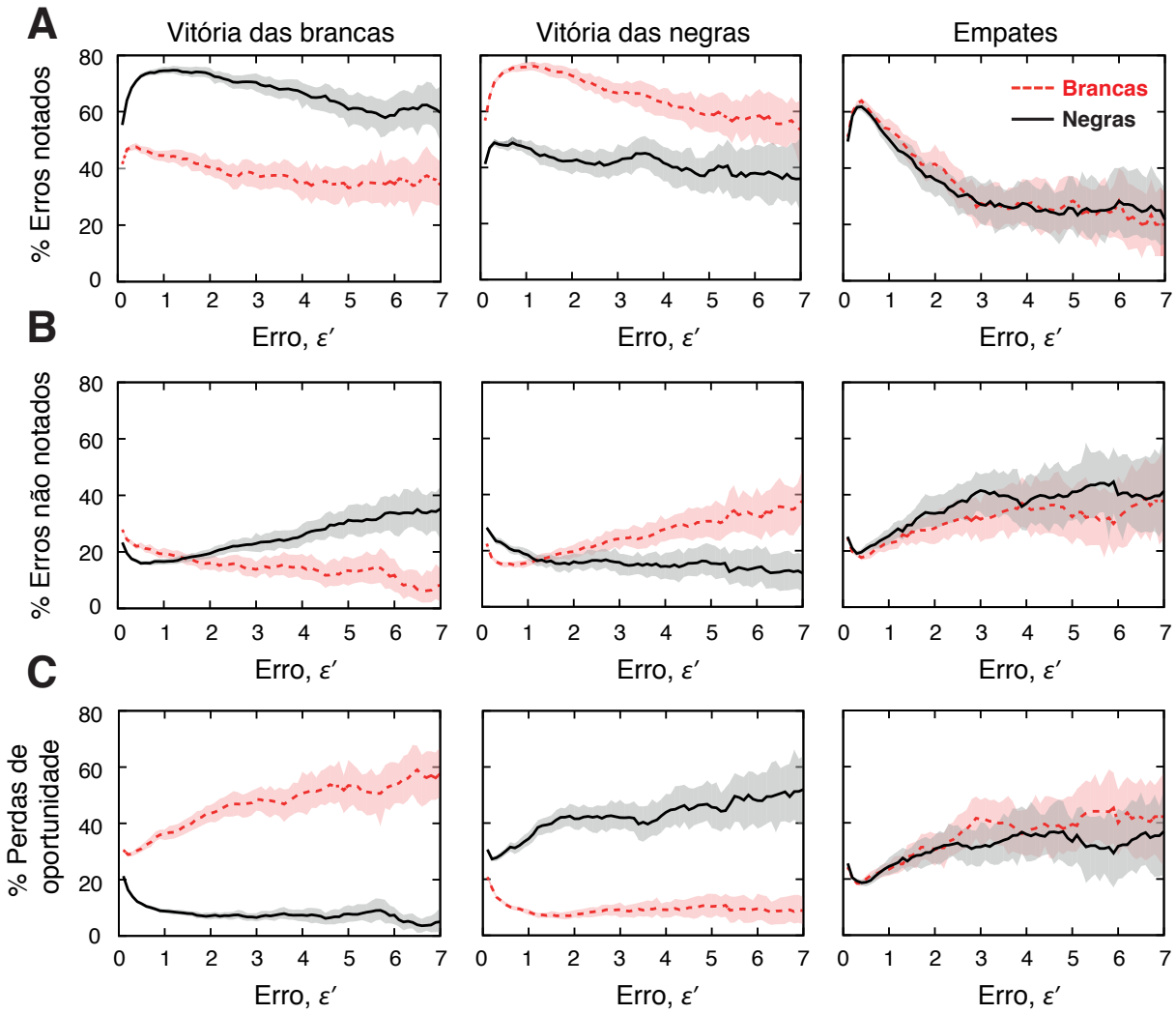


Figura 3.17: **(A)** Fração dos erros notados (*NE*) das brancas (linhas vermelhas) e das negras (linhas pretas) com magnitude maior que ε' para cada possível resultado do jogo. Em **(B)** e **(C)** mostramos o mesmo para os erros não notados (*UE*) e do tipo perdas de oportunidade (*LO*), respectivamente. Em todos os casos, as regiões sombreadas são intervalos de confiança a 95%.

bastante interessantes do jogo de xadrez. Por meio do cálculo do valor médio da vantagem, apontamos a existência de uma pequena vantagem das brancas por iniciarem a partida — a chamada iniciativa do jogo. Caracterizamos o processo difusivo da vantagem com sendo anômalo e dependente dos resultados das partidas. Além disso, observamos a existência de mais de um regime difusivo, sendo que para lances iniciais praticamente não há difusão da vantagem, fato que relacionamos com as sequências iniciais bem definidas do jogo, conhecidas como *aberturas*. Para o meio do jogo, observamos a presença de um regime superdifusivo para as partidas terminadas em empates e de um regime intermediário, aproximadamente hiperdifusivo, para partidas terminadas em vitórias. Estudamos essas diferenças nos regimes difusivos entre partidas terminadas em vitórias e empates, e observamos que o comportamento das vitórias é dependente do tamanho da partida e que apresenta maiores diferenças em relação aos empates apenas nos últimos movimentos. Para esses movimentos, uma espécie de efeito cascata ou avalanche provoca uma

variação muito rápida da variância da vantagem, produzindo o regime hiperdifusivo aproximado. Verificamos também que a distribuição de probabilidade da vantagem é invariante por escala e apresenta uma forma não gaussiana de caudas longas. Além disso, apontamos a existência de correlações de longo alcance nos incrementos da vantagem.

Em uma segunda parte, investigamos a estabilidade dos comportamentos anteriores ao longo da história do xadrez nos últimos dois séculos. Vimos que a vantagem inicial das brancas tem aumentado durante esse período como uma função exponencial que se aproxima da assíntota $\approx 0{,}23$ peões, com um tempo característico de ≈ 67 anos. Observamos que o expoente superdifusivo α e o movimento $m_\times$, no qual tem início o regime superdifusivo, também cresceram ao longo do período considerado. Em particular, α aproxima-se exponencialmente do valor ≈ 2 a um tempo característico de ≈ 128 anos, enquanto $m_\times$ aproxima-se de ≈ 16 movimentos a um tempo característico de ≈ 130 anos. Acreditamos que o comportamento da vantagem inicial e de $m_\times$ esteja relacionado a um mecanismo de aprendizagem em nível populacional, no qual os jogadores têm aprendido melhores maneiras de aproveitar o jogo com as brancas e também tornado o estágio de abertura dos jogos mais longo. A aproximação de α ao regime balístico pode ser relacionada ao crescimento da diferença entre os jogadores, sendo que, por meio de um modelo simplificado, mostramos que a simples presença dessas diferenças dá origem ao regime difusivo balístico.

A terceira e última parte deste capítulo foi destinada a investigar os erros dos jogadores. Vimos que as distribuições das magnitudes dos erros dos jogadores são bem descritas por distribuições log-normais. O processo estocástico que dá origem aos erros é então multiplicativo, o que explica (em parte) os comportamentos super ou hiperdifusivos observados para a variância da vantagem. Vimos que existe uma probabilidade maior de encontrar grandes erros em partidas terminadas em vitórias do que em partidas terminadas em empate. Esse fato nos levou a uma investigação mais detalhada sobre os erros, na qual propusemos a classificação dos erros em três tipos: os erros notados, os erros não notados e as perdas de oportunidade. Nessa análise, observamos que a maioria e os maiores erros ocorrem na região do meio jogo — região mais complexa do jogo de xadrez. Ao calcular as frações dos erros maiores do que um certo limiar, verificamos que perceber os erros dos adversários é um fator determinante para vencer as partidas. Vimos também que, mesmo nas partidas vencidas (por brancas ou negras), os jogadores deixam de notar muitos erros e perdem muitas oportunidades que poderiam ter levado a uma vitória mais fácil e convincente. No caso das partidas terminadas em empates, não observamos diferenças significativas nos percentuais dos erros notados, não notados e perdas de oportunidade.

Muitas outras possibilidades ainda podem ser exploradas (algumas, de fato, já estão sendo) nesse contexto dos jogos de xadrez. Entre elas, podemos investigar como se relacionam os parâmetros das distribuições dos erros ou o percentual de cada tipo de erro com a classificação de jogadores, proposta pela Federação Mundial de Xadrez (FIDE). Uma vez que o processo de classificação deles é baseado apenas nos resultados dos jogos, informações relacionadas à estrutura interna das partidas (como as que estudamos aqui) podem fornecer uma classificação melhor e remover possíveis ambiguidades. Outra possibilidade que pretendemos explorar está

relacionada a estudos de modelos difusivos que possam reproduzir as características presentes na primeira seção deste capítulo.

Capítulo 4

Difusão anômala e correlações na pontuação dos jogos de críquete

Neste capítulo, estudaremos a dinâmica das pontuações em uma competição esportiva: os jogos de críquete [19] — o segundo esporte mais popular do mundo. Veremos que a evolução da pontuação dos times ao longo de um jogo pode ser considerada como um processo difusivo anômalo, em que existe superdifusão, autossimilaridade e correlações de longo alcance. Verificaremos também que uma equação de Langevin generalizada, com termo de ruído correlacionado, pode descrever todos os nossos achados empíricos. Finalmente, discutiremos a relação entre as correlações de longo alcance e o fenômeno denominado "mão quente" nos esportes.

4.1 Introdução e apresentação dos dados

Os movimentos difusivos são, de fato, onipresentes na natureza. Eles podem representar desde o espalhamento de uma gota de tinta em água, descrever como organismos vivos, tais como peixes [107] ou bactérias [109], se movem ou até mesmo como informações se propagam em redes complexas [110]. Uma das análises mais comuns ao se estudar processos difusivos é a de verificar como as partículas ou objetos em questão se espalham. Essa medida de espalhamento é, geralmente, a variância das posições das partículas após um certo período de tempo. A chamada *difusão usual* ocorre quando a variância cresce linearmente com o passar do tempo. Duas hipóteses estão subjacentes a esse comportamento linear. A primeira é a não existência de memória ao longo da trajectória da partícula, o que permite escrever a sua próxima posição como uma função dependente apenas da sua posição imediatamente anterior (hipótese markoviana). A segunda hipótese é que existe uma escala característica para os incrementos das posições.

Naturalmente, em muitas situações da natureza, essas hipóteses são violadas. Consequentemente, desvios desse comportamento usual aparecem e são, normalmente, denominados *difusão anômala*. Uma situação bem estabelecida ocorre quando os incrementos das posições não possuem um comprimento característico. Nesse caso, a variância das posições não é finita e as distribuições do processo são as distribuições de Lévy. Exemplos de processos de Lévy incluem

animais durante forrageamento [111], difusão de átomos frios [112] e sistemas térmicos fora do equilíbrio [113]. A situação é mais complexa quando o processo difusivo apresenta memória, pois existem muitas maneiras de se correlacionar as posições das partículas. Essas escolhas mudam drasticamente as propriedades difusivas como o perfil da variância em função do tempo. Entretanto, um comportamento típico da variância nessas situações é uma dependência temporal do tipo lei de potência com um expoente característico α [104], sendo que $\alpha < 1$ representa sub-difusão e $\alpha > 1$ superdifusão.

Existe uma comunidade bastante interessada em estudar propriedades difusivas anômalas de um ponto de vista mais formal. Geralmente, esses formalismos envolvem equações de difusão fracionárias [104], equações de Fokker-Planck [114] e equações de Langevin generalizadas [115]. Por outro lado, ainda existe uma carência de resultados empíricos que visem testar esses formalismos ou que busquem um melhor entendimento dos mecanismos que levam os processos as serem anômalos e correlacionados. Nesse sentido, o estudo de processos difusivos em sistemas complexos pode representar um ponto de vista ideal para a aplicação desses formalismos, pois, como se tratam de sistemas consideravelmente diferentes da matéria inanimada, comportamentos mais complexos são esperados.

Estudaremos, agora, o processo difusivo relacionado à evolução das pontuações em um esporte chamado críquete. Ainda que não seja muito popular no Brasil, o jogo de críquete é o segundo esporte mais popular do mundo, atrás apenas do futebol. Trata-se de um jogo de "bastão e bola", similar ao beisebol, que é disputado entre dois times de 11 jogadores. Atualmente, existem três formatos de críquete que diferem apenas em tamanho (duração) da partida. O "Twenty20" críquete (T20) é o formato mais curto e dura aproximadamente 3 horas, o "One Day International" críquete (ODI) dura aproximadamente 8 horas e o "Test" críquete é o mais longo, levando até 5 dias para terminar. O jogo consiste em um dos times rebatendo as bolas (os *innings*) e buscando marcar o máximo de pontos possíveis (os *runs*) para definir um limite de pontos para o time adversário. Em seguida, o time adversário é quem rebate, visando exceder o limite de pontos do time da rodada anterior. Os *innings* de um time terminam sempre que excede-se a cota dos chamados *overs* (seis bolas lançadas consecutivamente) ou quando o time perde 10 *wickets* (estacas de madeira que são usadas como alvo pelo lançador de bolas do outro time). O limite de *overs* é 20 para o T20 críquete, 50 para o ODI críquete e 200 para o Test críquete. Uma versão simplificada desse jogo é jogada no Brasil com o nome de "bets" ou "bétia".

Surpreendentemente, os registros de um jogo de críquete (*score cards*) incluem não somente os resultados das partidas, mas também incluem a evolução evento por evento das pontuações. Em particular, coletamos informações das pontuações por *overs* para o T20 nos anos de 2005 a 2011 e para o ODI e Test críquete nos anos de 2002 a 2011 na página de internet *cricinfo* [116]. Usando esses dados, foram construídas 2144 séries temporais, nas quais o tempo t representa um *over* completo. Nas Figuras 4.1A, 4.1B e 4.1C, mostramos a evolução temporal das pontuações $S(t)$ para 100 partidas, dos três diferentes formatos do jogo, selecionadas aleatoriamente de nossa base de dados. Notamos que há uma tendência de crescimento nas pontuações e também um movimento errático ao longo dessa tendência média. Para melhor visualização das flutuações,

mostramos nas Figuras 4.1D, 4.1E e 4.1F a evolução de $S(t)$ após subtrair a tendência média de crescimento $\langle S(t)\rangle$.

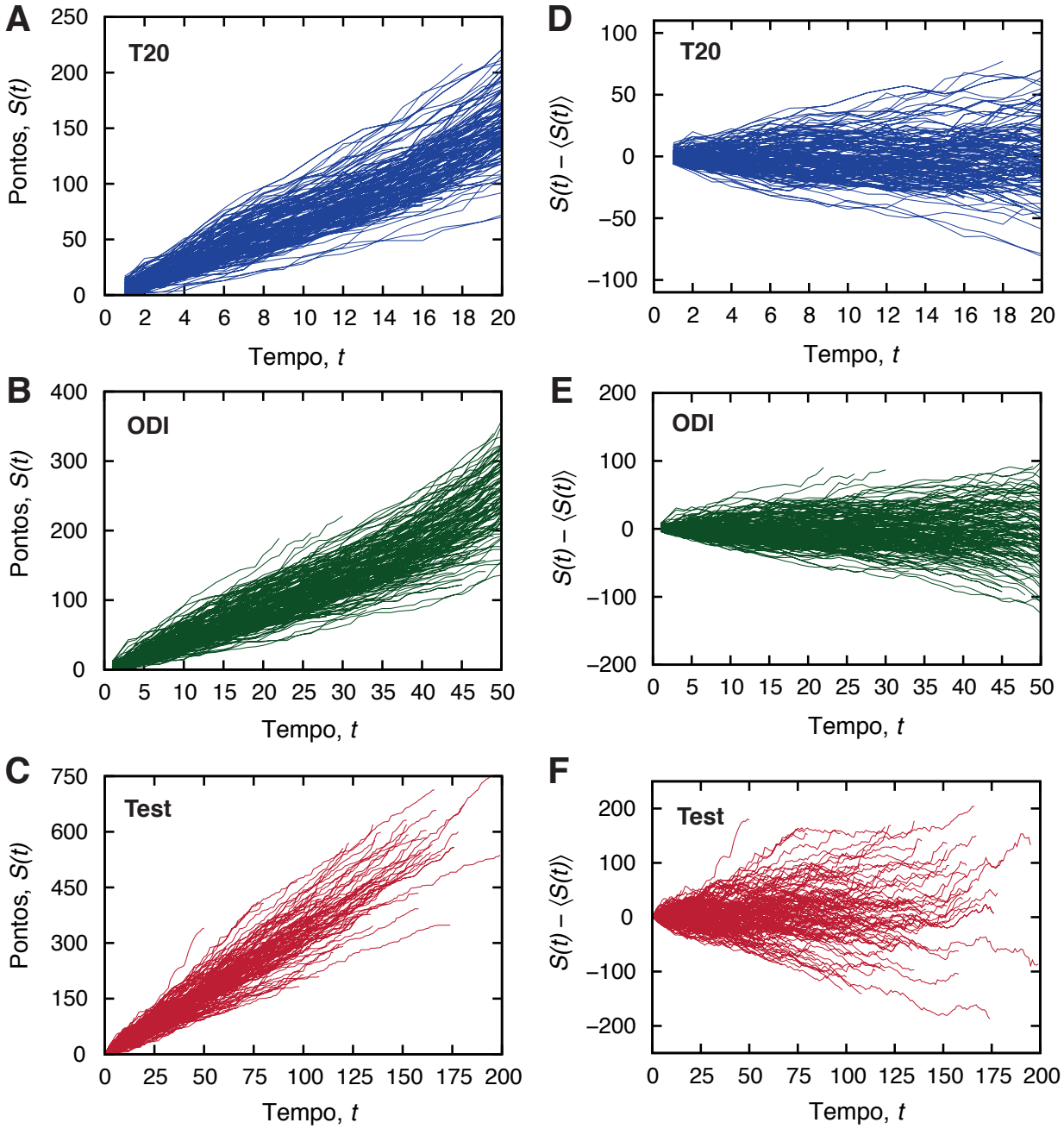


Figura 4.1: Evolução da pontuação $S(t)$ para os três tipos de críquete: (**A**) "Twenty20" ou T20, (**B**) "One Day International" ou ODI e (**C**) Test críquete. Aqui, selecionamos aleatoriamente as pontuações de 100 jogos de cada tipo de críquete. Em (**D**), (**E**) e (**F**) mostramos $S(t)$ após subtrair a tendência média de crescimento $\langle S(t)\rangle$.

4.2 Análise estatística da pontuação

Iniciamos nossa investigação da evolução das pontuações $S(t)$, calculando os seus valores médios em função do tempo

$$\langle S(t)\rangle = \frac{1}{N}\sum_{i=1}^{N} S_i(t)\,, \tag{4.1}$$

em que N é o número de evoluções das pontuações para cada formato de críquete e $S_i(t)$ é a pontuação de um time em dado jogo i. A Figura 4.2 mostra os valores médios $\langle S(t)\rangle$ para os três formatos do críquete. Notemos que $\langle S(t)\rangle$ cresce linearmente com t, *i.e.*, $\langle S(t)\rangle = A\,t$ para os três formatos. A única diferença é a taxa de crescimento A, com $A = 6{,}4 \pm 1{,}0$ para T20, $A = 3{,}9 \pm 1{,}0$ para ODI e $A = 3{,}3 \pm 1{,}0$ para Test críquete. Essas diferenças mostram que o desempenho dos times está relacionado à duração do jogo. Para os formatos T20 (duração de ~ 3 horas) e ODI (duração de ~ 8 horas) as taxas de crescimento da pontuação são maiores, o que indica que os jogadores trabalham duro para marcar o máximo de pontos possíveis. Por outro lado, para o Test críquete a taxa A é menor que nos outros dois formatos, sugerindo que no Test críquete os jogadores preferem poupar esforços, visto que a partida é muito longa.

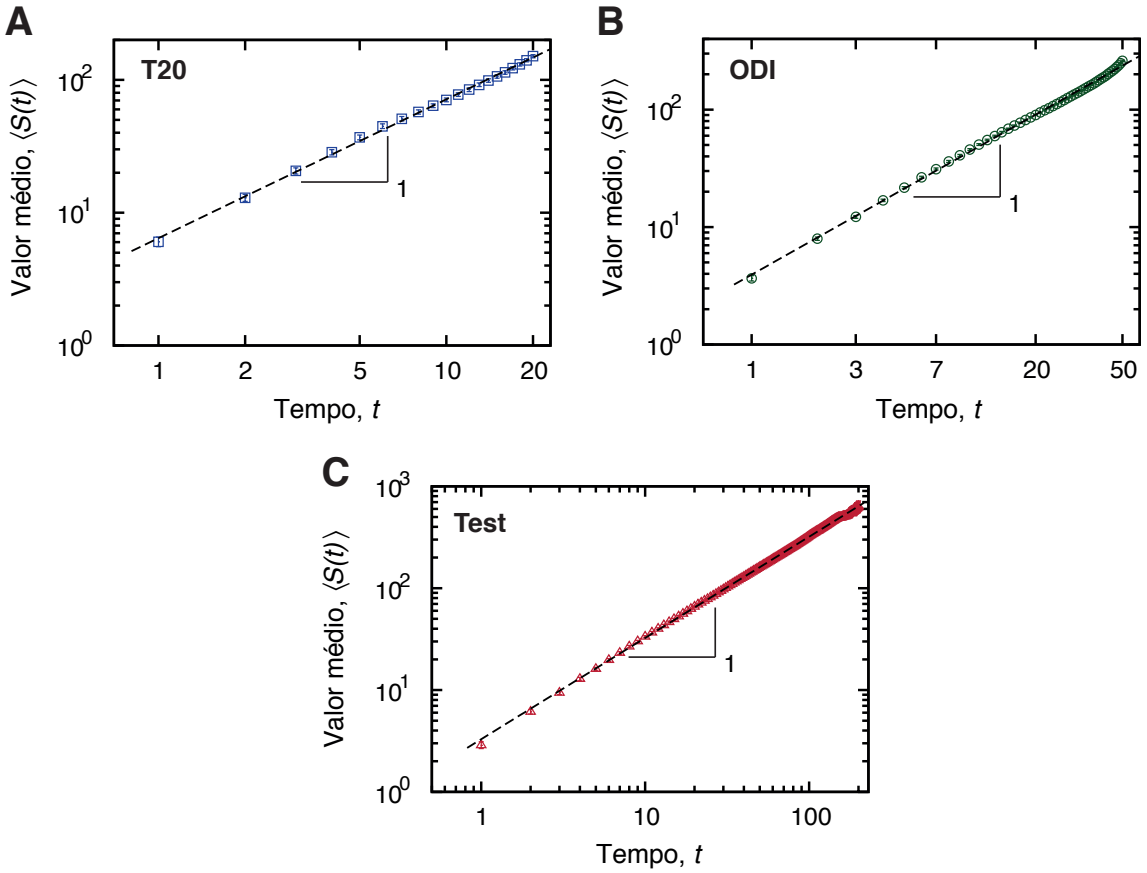


Figura 4.2: Valor médio das pontuações em função do tempo, $\langle S(t)\rangle$, para os três tipos de críquete: (**A**) "Twenty20" ou T20, (**B**) "One Day International" ou ODI e (**C**) Test críquete. As linhas tracejadas representam ajustes lineares a cada conjunto de dados em log-log. As barras de erros (ainda que menores que o tamanho do ponto) representam intervalos de confiança a 95% obtidos via bootstrapping (Apêndice B.2).

Como estamos interessados no processo difusivo das pontuações, calculamos também a variância

$$\sigma^2(t) = \langle [S(t) - \langle S(t) \rangle]^2 \rangle = \frac{1}{N-1} \sum_{i=1}^{N} (S_i(t) - \langle S(t) \rangle)^2 \,. \tag{4.2}$$

A Figura 4.3 mostra essa análise para os três formatos do críquete, na qual observamos um crescimento não linear de $\sigma^2(t)$, *i.e.*, ao ajustar esses dados encontramos $\sigma^2(t) \propto t^{\alpha}$, com $\alpha \approx 1{,}3$ para todos os três formatos. Esse valor de $\alpha > 1$ mostra que o processo difusivo relacionado à evolução de $S(t)$ é superdifusivo. Essa característica intrigante sugere que o competição natural do jogo induz as pontuações a se espalharem mais rapidamente do que em um movimento browniano usual.

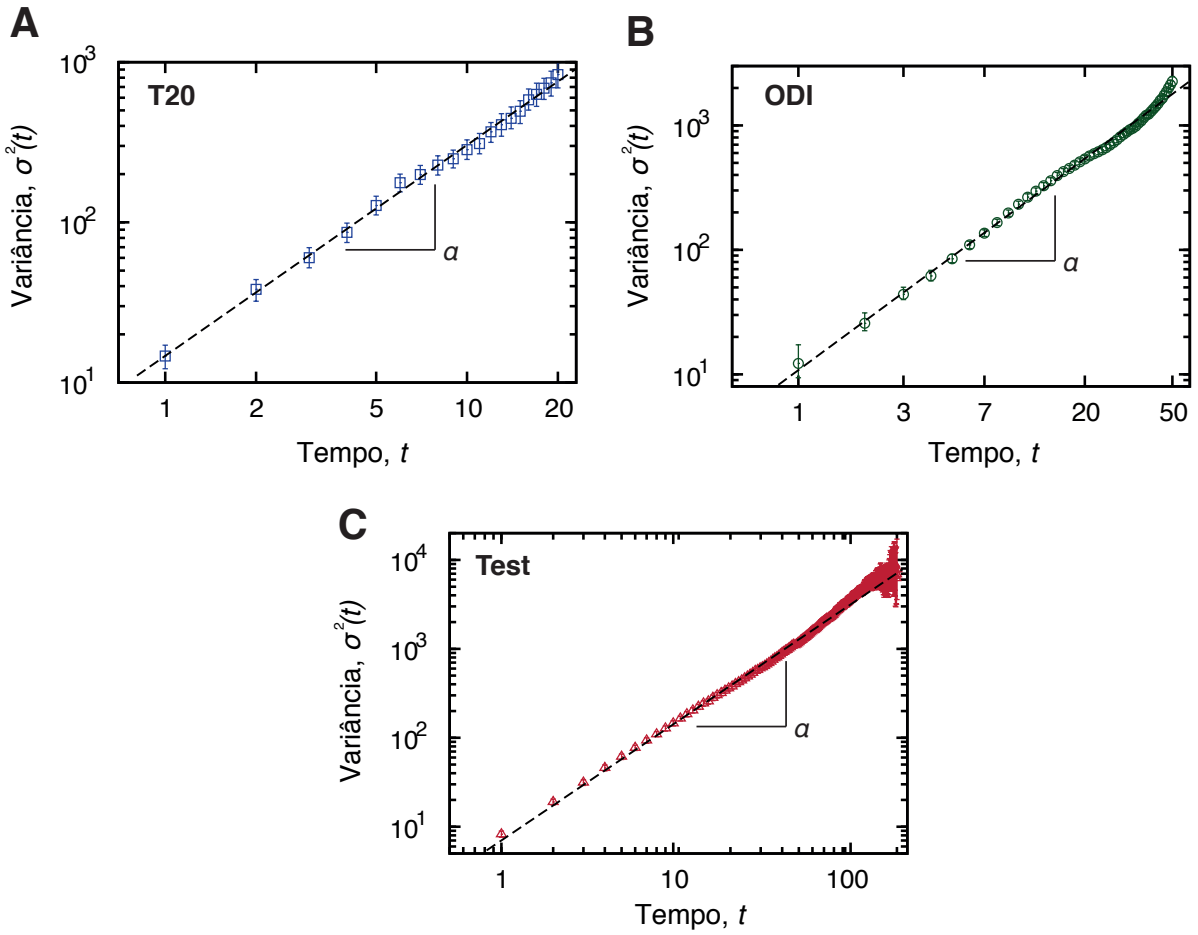


Figura 4.3: A difusão anômala das pontuações. Mostramos os espalhamentos das trajetórias medidos por meio da variância $\sigma^2(t) = \langle [S(t) - \langle S(t) \rangle]^2 \rangle$ para os três tipos de críquete: (**A**) "Twenty20" ou T20, (**B**) "One Day International" ou ODI e (**C**) Test críquete. As linhas tracejadas são ajustes de funções tipo lei de potência, em que encontramos $\sigma^2(t) \sim t^{\alpha}$ com $\alpha = 1{,}32 \pm 0{,}02$ para T20, $\alpha = 1{,}31 \pm 0{,}02$ para ODI, e $\alpha = 1{,}30 \pm 0{,}02$ para Test. As barras de erros representam intervalos de confiança a 95% obtidos via bootstrapping (Apêndice B.2).

Outra pergunta que podemos fazer sobre as pontuações é se a distribuição de probabilidade de $S(t)$ é autossimilar e se ela segue uma forma funcional particular. Para responder a essas questões, calculamos as distribuições das pontuações para cada tempo t, como mostrado na

Figura 4.4. Essa figura ilustra bem o processo de difusão das pontuações, na qual observamos o deslocamento das distribuições para valores positivos e também o aumento da largura das distribuições com o passar do tempo. Além disso, esses gráficos em mono-log também indicam que as distribuições são bem parecidas com distribuições gaussianas.

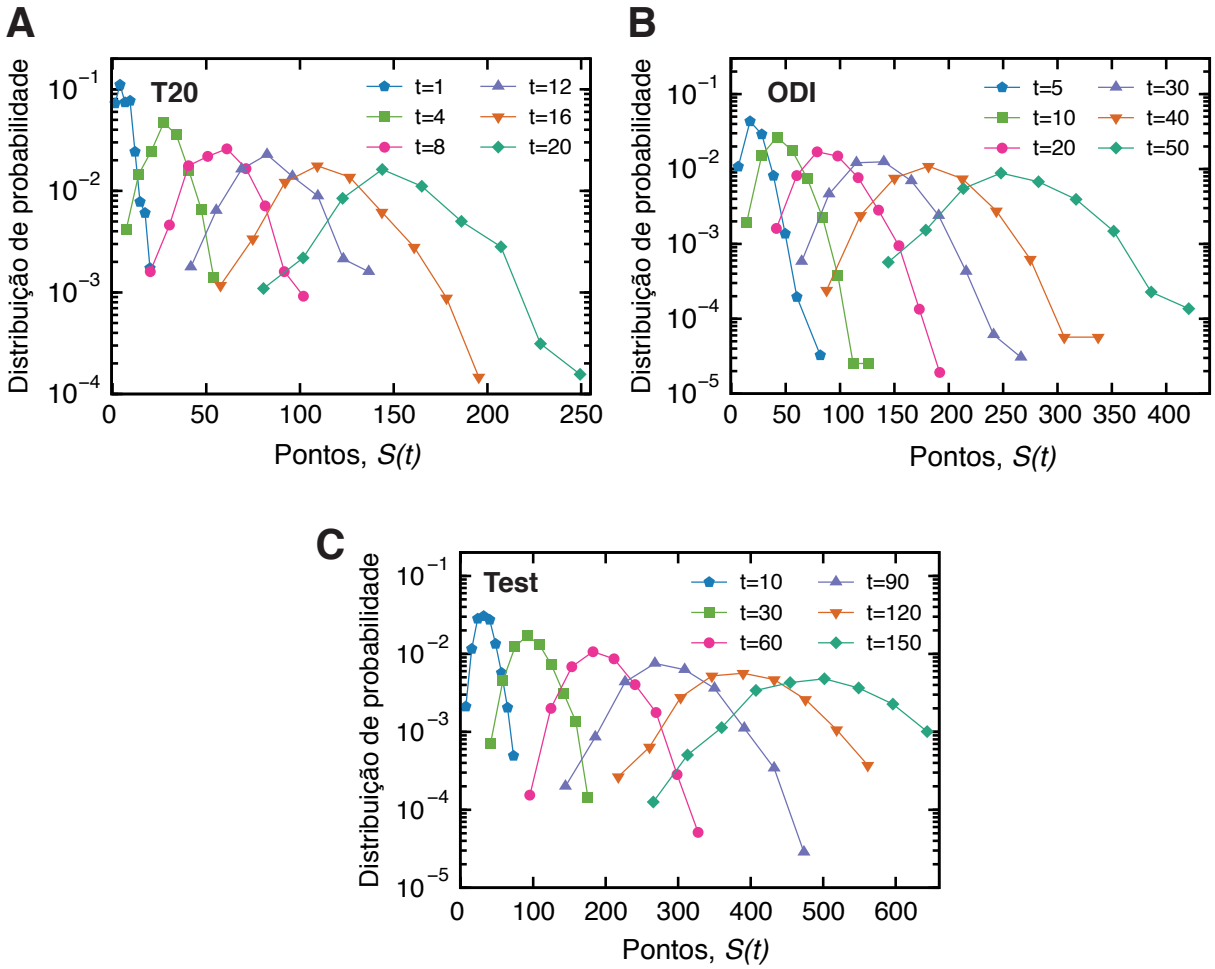


Figura 4.4: Evolução das distribuições de probabilidade das pontuações $S(t)$ para alguns valores de t (mostrados na figura) e para os três tipos de críquete: (**A**) "Twenty20" ou T20, (**B**) "One Day International" ou ODI e (**C**) Test críquete. Notemos que o centro das distribuições se descola para valores positivos e que a largura das distribuições aumenta com passar do tempo. Observamos também que as distribuições assemelham-se a distribuições gaussianas.

Para verificar a normalidade e a autossimilaridade das distribuições, calculamos as distribuições das pontuações normalizadas usando $\xi(t) = \frac{S(t)-\langle S(t)\rangle}{\sigma(t)}$. A Figura 4.5 mostra essas distribuições, na qual podemos notar o bom colapso das distribuições em uma única curva muito próxima de uma distribuição gaussiana. Observemos que os valores médios dessas distribuições são bem ajustados pela gaussiana de média nula e desvio padrão unitário (linha tracejada).

Podemos, ainda, aplicar o teste de Kolmogorov-Smirnov para verificar de uma maneira mais precisa a normalidade das distribuições das pontuações. Para isso, calculamos o valor p (Apêndice B.1) em função do tempo t, como mostra a Figura 4.6. Aqui, observamos que a hipótese de que as distribuições sejam gaussianas é rejeitada para todas as distribuições das pontuações do

T20 críquete e também do ODI críquete. Somente para valores de $t \gtrsim 90$ e para o Test críquete é que não podemos descartar a hipótese de que as distribuições de $S(t)$ sejam gaussianas. Esses resultados estão em contraste com os ajustes mostrados na Figura 4.5. Contudo, o fato do teste falhar para tempos curtos é um reflexo da condição inicial das pontuações, *i.e.*, como os pontos iniciam-se do valor zero, leva um certo tempo para que as distribuições fiquem simétricas e gaussianas, fato que podemos observar na Figura 4.4.

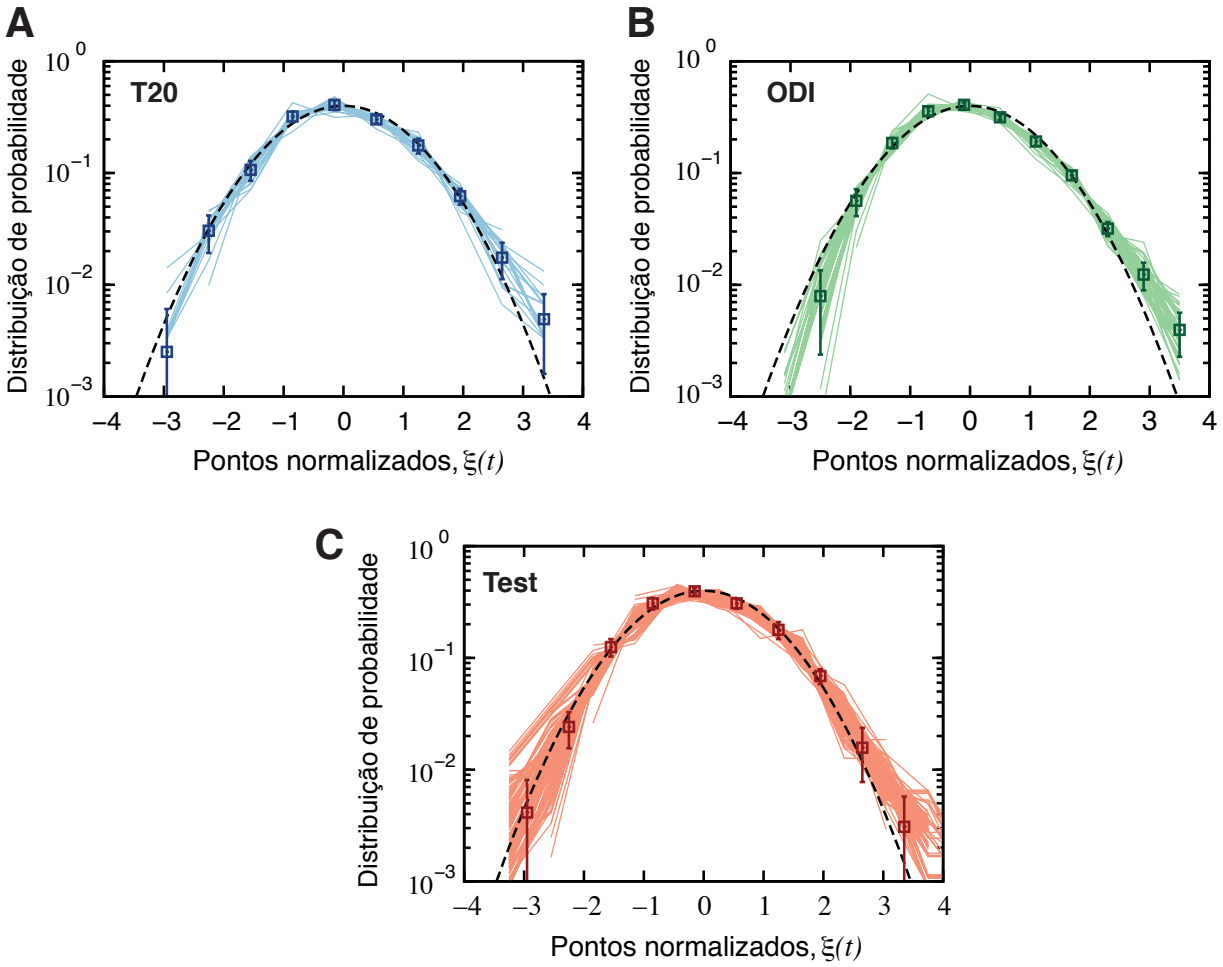


Figura 4.5: Invariância de escala das pontuações. As figuras mostram as distribuições usando a variável escalada $\xi(t) = \frac{S(t)-\langle S(t)\rangle}{\sigma(t)}$ (linhas contínuas) para todos os valores de t e para os três tipos de críquete: (**A**) "Twenty20" ou T20, (**B**) "One Day International" ou ODI e (**C**) Test críquete. Nessas figuras, os símbolos representam os valores médios das distribuições para diferentes valores de t e as barras de erro são intervalos de confiança a 95% obtidos via bootstrapping (Apêndice B.2). A linha tracejada é uma distribuição gaussiana de média zero e variância unitária.

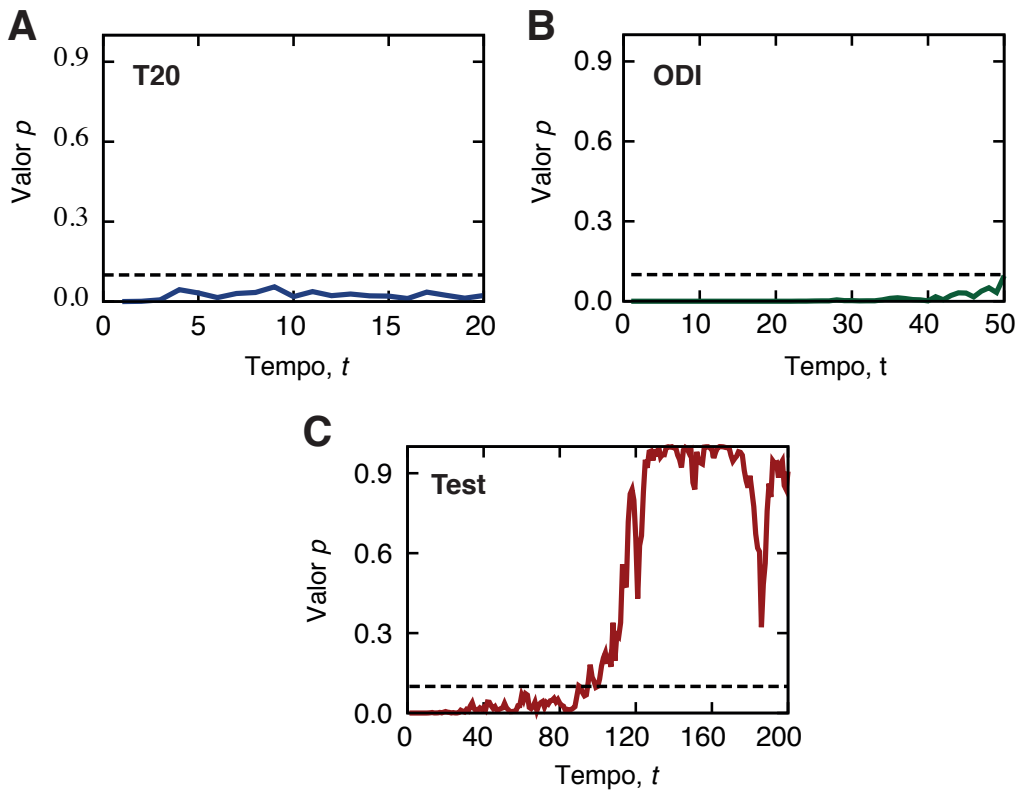


Figura 4.6: Testes de hipótese da normalidade das distribuições de probabilidade das pontuações $S(t)$. Mostramos o valor p do teste de Kolmogov-Smirnov (Apêndice B.1) em função do tempo t, para todos os valores de t e para os três tipos de críquete: (**A**) "Twenty20" ou T20, (**B**) "One Day International" ou ODI e (**C**) Test críquete. A linha tracejada representa o limiar a partir do qual não podemos rejeitar a hipótese de que as distribuições sejam gaussianas.

Vamos, agora, fazer uma análise de correlações na evolução da pontuação para tentar verificar se o processo é markoviano. Para isso, selecionamos todas as partidas do formato Test críquete que duraram mais de 100 unidades de tempo, totalizando 431 jogos. Com esse subconjunto de dados, definimos a série dos incrementos da pontuação dada por

$$\Delta S(t) = S(t+1) - S(t)\,. \tag{4.3}$$

Aplicamos análise DFA (Apêndice A.3) nessas novas séries temporais para obter o expoente de Hurst, h. Sabemos que, para séries com correlações de longo alcance, a função de flutuação segue uma lei potência, da forma $F(n) \sim n^h$. Em nossos dados, verificamos que h é praticamente independente do jogo e possui um valor médio igual a $h = 0{,}63 \pm 0{,}01$. Na Figura 4.7, mostramos as análises DFA. Esses resultados indicam que existe memória de longo alcance na evolução das pontuações, e o valor de $h > 0{,}5$ aponta um comportamento persistente nos incrementos $\Delta S(t)$, ou seja, valores positivos são seguidos de valores positivos e valores negativos são seguidos de valores negativos muito mais frequentemente do que em um processo completamente aleatório.

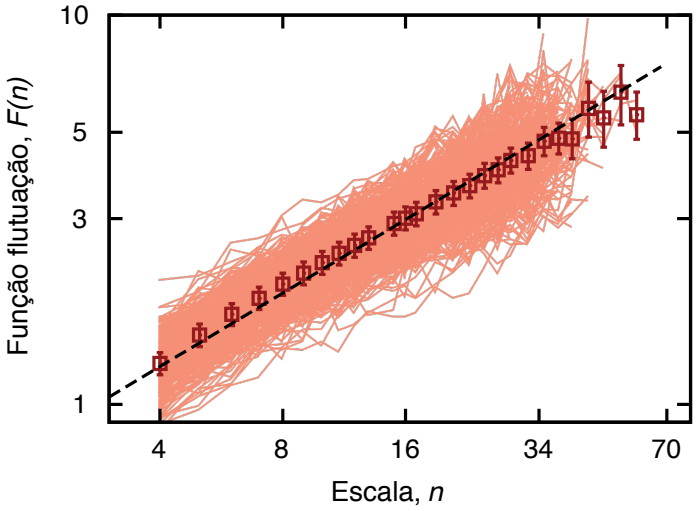


Figura 4.7: As linhas contínuas mostram a análise DFA dos incrementos da pontuação, $\Delta S(t) = S(t+1) - S(t)$, para todas as partidas de Test críquete que duraram mais de 100 unidades de tempo (431 jogos). Os símbolos representam os valores médios da função de flutuação e as barras de erros são intervalos de confiança a 95% (Apêndice B.2). A linha tracejada é uma lei de potência em que $F(n) \sim n^h$ com $h = 0{,}63 \pm 0{,}01$ sendo o valor médio dos expoentes de Hurst obtidos para cada partida.

4.3 Modelando com uma equação de Langevin generalizada

Na tentativa de modelar os achados empíricos da seção anterior, consideramos uma equação de Langevin generalizada que tem sido bastante estudada do ponto de vista mais formal. Em particular, suponhamos que pontuação $S(t)$ possa ser descrita pela equação

$$\frac{d^2S(t)}{dt^2} + \int_0^t \lambda(t-\tau)\frac{dS(\tau)}{d\tau}d\tau + K = \xi(t)\,, \tag{4.4}$$

em que $\lambda(t-\tau)$ é uma função relacionada ao efeito retardado da força de dissipação, $K < 0$ é uma constante e $\xi(t)$ é um ruído gaussiano de média nula. Notamos que para $\lambda(t-\tau) = \gamma\delta(t-\tau)$ recuperamos a equação de Langevin usual na presença de uma força constante:

$$\frac{d^2S(t)}{dt^2} + \gamma\frac{dS(t)}{dt} + K = \xi(t)\,. \tag{4.5}$$

Visando a ganhar um pouco de intuição sobre a equação 4.4, resolvemos inicialmente a equação 4.5. Para isso, consideramos a variável $v(t) = dS(t)/dt$ e aplicamos a transformada de Laplace ($\mathcal{L}\{\dots\} \equiv \int_0^\infty \exp(-st)\dots dt$) à equação 4.5, obtendo

$$v(s) = \frac{v_0}{s+\gamma} - \frac{K}{s(s+\gamma)} + \frac{\xi(s)}{s+\gamma}\,, \tag{4.6}$$

na qual $v_0 = v(t=0)$ é a condição inicial do sistema. Após aplicarmos a transformada inversa

de Laplace na equação 4.6, teremos

$$v(t) = v_0 \exp(-\gamma\, t) - \frac{K}{\gamma}[1 - \exp(-\gamma\, t)] + \int_0^t \exp[-\gamma\,(t - t')]\xi(t')dt'\,, \tag{4.7}$$

que pode ser integrada, resultando em

$$\begin{aligned} S(t) &= \frac{\gamma\, v_0 + K}{\gamma^2}[1 - \exp(-\gamma\, t)] - \frac{K\, t}{\gamma} + \int_0^t dt' \int_0^{t'} dt'' \exp[-\gamma(t' - t'')]\xi(t'') \\ &= \frac{\gamma\, v_0 + K}{\gamma^2}[1 - \exp(-\gamma\, t)] - \frac{K\, t}{\gamma} + \frac{1}{\gamma}\int_0^t dt''\xi(t'')\{1 - \exp[-\gamma(t - t'')]\}\,. \end{aligned} \tag{4.8}$$

Usando a expressão 4.8 e lembrando que $\langle \xi(t') \rangle = 0$, podemos calcular o valor médio da pontuação

$$\langle S(t) \rangle = \frac{\gamma\, v_0 + K}{\gamma^2}[1 - \exp(-\gamma\, t)] - \frac{K\, t}{\gamma}\,, \tag{4.9}$$

o qual, para tempos longos, pode ser aproximado por

$$\langle S(t) \rangle \approx -\frac{K\, t}{\gamma}\,. \tag{4.10}$$

Assim, a equação de Langevin usual com o termo de força constante (K) é capaz de reproduzir o crescimento linear da pontuação no tempo (Figura 4.2). Entretanto, ao calcularmos a variância

$$\langle (S(t) - \langle S(t) \rangle)^2 \rangle = \left\langle \left(\frac{1}{\gamma} \int_0^t dt''\xi(t'')\{1 - \exp[-\gamma(t - t'')]\} \right)^2 \right\rangle\,, \tag{4.11}$$

obtemos

$$\begin{aligned} \langle (S(t) - \langle S(t) \rangle)^2 \rangle &= \frac{\Lambda}{\gamma^2} \int_0^t dt' \{1 - \exp[-\gamma(t - t')]\}^2 \\ &= \frac{\Lambda}{\gamma^2} \left(t - \frac{3}{2\gamma} - \frac{\exp(-2\gamma\, t)}{2\gamma} + \frac{\exp(-\gamma\, t)}{\gamma} \right)\,. \end{aligned} \tag{4.12}$$

Nesse cálculo, assumimos que o ruído é não correlacionado, *i.e.*, $\langle \xi(t')\xi(t'') \rangle = \Lambda\delta(t' - t'')$, sendo Λ uma constante. A expressão 4.12 pode ser aproximada para tempos longos, como

$$\langle (S(t) - \langle S(t) \rangle)^2 \rangle \approx \frac{\Lambda}{\gamma^2}\, t\,, \tag{4.13}$$

ou seja, a variância apresenta uma dependência linear no tempo (difusão usual), a qual não reproduz os regimes superdifusivos encontrados na Figura 4.3. Além disso, usando a expressão 4.7, podemos calcular a autocorrelação das velocidades,

$$\langle v(t)v(t + \bar{t}) \rangle \sim \exp(-\gamma\, \bar{t})\,, \tag{4.14}$$

para verificar que a equação de Langevin 4.5 não reproduz as correlações de longo alcance nos

incrementos das pontuações (Figura 4.7).

Podemos, ainda, calcular a distribuição de probabilidade de $S(t)$ da equação 4.5. Para isso, desprezamos o termo de derivada segunda e lembramos que a equação de Langevin 4.5 é equivalente à seguinte equação de Fokker-Planck [117]

$$\frac{\partial P}{\partial t} = \frac{K}{\gamma^2}\frac{\partial P}{\partial S} + \frac{\Lambda}{2\gamma}\frac{\partial^2 P}{\partial S^2}\,, \tag{4.15}$$

em que $P = P(S,t)$ é a distribuição das pontuações. A solução para a equação 4.15 pode ser encontrada usando a transformada de Laplace para a variável t e a transformada de Fourier na variável x, que após invertidas levam a

$$P(S,t) = \frac{1}{\sqrt{2\pi\Lambda t}}\exp\left[-\frac{(S+\frac{K}{\gamma^2}t)^2}{2\Lambda t}\right]\,. \tag{4.16}$$

Notemos que usamos como condição inicial $P(S,0) = \delta(S)$. Desse modo, temos uma distribuição gaussiana centrada em $-\frac{K}{\gamma^2}t$ que reproduz os resultados na Figura 4.4.

Assim, verificamos que embora a equação Langevin usual com termo de força constante reproduza o crescimento linear do valor médio da pontuação e a distribuição gaussiana dos pontos, ela não descreve o regime superdifusivo da variância nem as correlações de longo alcance dos incrementos da pontuação. Para tentar descrever esses dois últimos resultados empíricos, consideramos a equação de Langevin generalizada 4.4 com termo de ruído com correlação de longo alcance, dado por $\langle\xi(t')\xi(t'')\rangle = \mathcal{A}\,|t'-t''|^{-\alpha}$, sendo $\mathcal{A}$ e α constantes. Assumimos também que $\lambda(t) = \langle\xi(0)\xi(t)\rangle$ para satisfazer o Teorema da Flutuação-Dissipação [118]. Nessas condições, com $K = 0$, a equação 4.4 foi estudada nas referências [115, 119, 120]. Em nosso caso, para $K \neq 0$, podemos proceder como fizemos para a equação 4.5, ou seja, calculamos a transformada de Laplace para obter

$$v(s) = \frac{v_0}{s+\mathcal{A}\,\Gamma(1+\alpha)s^{-1-\alpha}} - \frac{K}{s(s+\mathcal{A}\,\Gamma(1+\alpha)s^{-1-\alpha})} + \frac{\xi(s)}{s+\mathcal{A}\,\Gamma(1+\alpha)s^{-1-\alpha}}\,, \tag{4.17}$$

a qual, após ser invertida, conduz a

$$\begin{aligned} v(t) &= v_0 E_{2+\alpha,1}[-\mathcal{A}\,\Gamma(1+\alpha)t^{2+\alpha}] - Kt\,E_{2+\alpha,2}[-\mathcal{A}\,\Gamma(1+\alpha)t^{2+\alpha}] \\ &+ \int_0^t E_{2+\alpha,1}[-\mathcal{A}\,\Gamma(1+\alpha)t'^{2+\alpha}]\xi(t')dt' \end{aligned} \tag{4.18}$$

e que pode ser integrada para obtermos

$$\begin{aligned} S(t) &= v_0(t^{2+\alpha}E_{2+\alpha,2}[-\mathcal{A}\,\Gamma(1+\alpha)t^{2+\alpha}] - 1) + Kt^{2+\alpha}E_{2+\alpha,3}[-\mathcal{A}\,\Gamma(1+\alpha)t^{2+\alpha}] \\ &+ \int_0^t dt' \int_0^{t'} dt'' E_{2+\alpha,2}[-\mathcal{A}\,\Gamma(1+\alpha)t^{2+\alpha}]\xi(t'')\,. \end{aligned} \tag{4.19}$$

Nas expressões 4.19, a função

$$E_{a,b}(z) = \sum_{k=0}^{\infty} \frac{z}{\Gamma(ak+b)}\,, \tag{4.20}$$

com $\Gamma(y) = \int_0^\infty \exp(y) t^{y-1} dt$ (função gamma de Euler) é a função de Mittag-Leffler [121], uma generalização da função exponencial ($E_{1,1}(z) = \exp(z)$). Essa função aparece ao calcularmos a transformada de Laplace de

$$G(s) = \frac{1}{s + \mathcal{B} s^{-1-\alpha}} \tag{4.21}$$

presentes nos termos da equação 4.17.

De posse das soluções para $S(t)$, podemos calcular o seu valor médio e variância e verificar que, para tempos longos,

$$\langle S(t) \rangle \sim t \tag{4.22}$$

e

$$\langle (S(t) - \langle S(t) \rangle)^2 \rangle \sim t^{\alpha}\,, \tag{4.23}$$

ou seja, a equação de Langevin generaliza 4.4 reproduz o regime superdifusivo de espalhamento das pontuações. Além disso, o cálculo da função de correlação de $v(t)$ leva a

$$\langle v(t) v(t+\bar{t}) \rangle \sim \bar{t}^{\alpha-2}\,, \tag{4.24}$$

descrevendo adequadamente as correlações de longo alcance que encontramos nos incrementos das pontuações. Além disso, usando a relação existente entre o expoente de correlação $\gamma' = \alpha - 2$ e o expoente de Hurst h, $\gamma' = 2(1-h)$ (Apêndice A.2), obtemos uma relação entre o expoente α da superdifusão e o expoente h de Hurst:

$$\alpha = 2h\,. \tag{4.25}$$

A Figura 4.8 mostra um gráfico de barras em que comparamos os dois membros da relação 4.25. Observamos que o valor de $2h = 1{,}26$ está próximo ao valor de $\alpha = 1{,}30$ e que existe superposição entre os intervalos de confiança de cada um. Portanto, a previsão do modelo de que $\alpha = 2h$ é válida.

Para finalizar, precisamos obter a distribuição de probabilidade das pontuações. Entretanto, não existe uma equivalência bem definida entre a equação de Langevin generalizada e uma equação do tipo Fokker-Planck [30]. Aqui, utilizamos a aproximação proposta na referência [119]

$$\frac{\partial P}{\partial t} = \frac{K}{\gamma^2} \frac{\partial P}{\partial S} + D_e(t) \frac{\partial^2 P}{\partial S^2}\,, \tag{4.26}$$

em que $D_e(t) = 1/\int_0^t \lambda(t')dt' \sim \frac{\alpha+1}{\mathcal{A}} t^{\alpha-1}$. Após resolver essa equação, usando a condição inicial

$P(S,0) = \delta(S)$, devemos obter

$$P(S,t) = \frac{1}{\sqrt{2\pi\Lambda t^{\alpha}}} \exp\left[-\frac{(S+\frac{K}{\gamma^2}t)^2}{2\Lambda t^{\alpha}}\right], \tag{4.27}$$

ou seja, uma distribuição gaussiana centrada em $-\frac{K}{\gamma^2}t$.

Assim, verificamos que as expressões 4.22, 4.23, 4.24 e 4.27 reproduzem todos os comportamentos observados nos dados empíricos e, portanto, a equação de Langevin generalizada 4.4 descreve muito bem a evolução das pontuações.

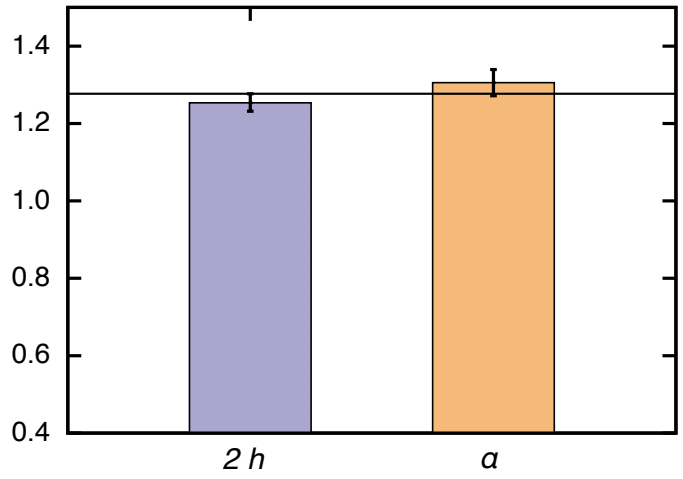


Figura 4.8: Comparação da previsão do modelo, $\alpha = 2h$, para o jogos de Test críquete. A barra à esquerda mostra o valor empírico de $2h$ e a barra da direita o valor de α. As barras de erro são intervalos de confiança a 95% (Apêndice B.2) e a linha horizontal representa o limite superior do intervalo de confiança para $2h$. Notemos que existe superposição entre os intervalos de confiança e, portanto, podemos aplicar a relação $\alpha = 2h$ para os jogos de Test críquete.

4.4 Conclusões e perspectivas

Estudamos neste capítulo a evolução temporal das pontuações dos jogos de críquete como um processo difusivo. Nossa análise revelou que o valor médio da pontuação cresce linearmente com o passar do tempo, enquanto a variância das pontuações apresenta uma dependência temporal do tipo lei de potência, com um expoente $\alpha > 1$, ou seja, um regime superdifusivo. Mostramos, também, que as pontuações são estatisticamente autossimilares e seguem uma distribuição universal gaussiana. Pelo uso do DFA, verificamos que o processo difusivo é não markoviano, visto que os incrementos das pontuações possuem correlações de longo alcance. Verificamos que todos esses aspectos empíricos podem ser descritos por uma equação de Langevin generalizada, na qual o termo de ruído possui correlações do tipo lei de potência.

Do ponto de vista esportivo, o comportamento persistente manifestado pelo valor do expoente de Hurst $h > 0{,}5$ pode ter relações com o fenômeno esportivo de "mão quente" (*hot hand*) [122].

Esse fenômeno é uma crença entre esportistas e estudiosos de esportes de que jogadores que fizeram pontos anteriormente no jogo têm uma probabilidade maior de marcar novamente no futuro. Talvez o exemplo mais famoso seja no basquete, em que os fãs do esporte acreditam que o jogador tem uma probabilidade maior de acertar a cesta se ele já acertou os dois ou três último arremessos. Nesse caso, eles dizem que o jogador está com a "mão quente". Entretanto, a existência ou não desse fenômeno representa um debate histórico desde o trabalho seminal de Gilovich *et al.* [123], no qual os autores atribuem o fenômeno de "mão quente" a uma interpretação errônea de sequências de números aleatórios, algo semelhante à falsa crença de que, depois uma sequência de caras, há uma probabilidade maior de ocorrer coroa em lançamentos de moedas (a chamada falácia de Monte Carlo). Entretanto, evidências recentes parecem não deixar dúvidas de que o fenômeno realmente existe no basquete [124] e no boliche [125]. Em nosso estudo, as correlações persistentes de longo alcance que encontramos, além de indicar a existência desse fenômeno nos jogos de críquete, também sugerem que esse fenômeno pode se manifestar em escalas longas de tempo.

Nesse sentido, nossa análise poderia ser aplicada a outros esportes, nos quais novos resultados podem revelar outros tipos de cenários difusivos para serem comparados com o crescente número de resultados teóricos em difusão anômala, além de poder investigar questões mais ligadas ao esporte, como a existência ou não de fenômenos do tipo "mão quente".

Capítulo 5

Dinâmica de bolhas em água fervente via transmitância de um feixe laser

Neste capítulo, estudaremos a dinâmica das bolhas em uma amostra de água em ebulição [20]. Para extrair informações sobre essa dinâmica, construímos um experimento simples no qual um feixe laser atravessa uma amostra de água fervente e tem sua intensidade monitorada. Usando essa série temporal da intensidade, veremos que a dinâmica apresenta alguns aspectos não triviais, em particular, distribuições não exponenciais e correlações de longo alcance. Um modelo minimalista também será proposto para tentar descrever essa dinâmica de intensidade do sinal.

5.1 Introdução e descrição do experimento

As bolhas estão presentes em vários fenômenos da natureza e podem, muitas vezes, apresentar uma dinâmica bastante complexa [126, 127]. Por essa razão e também por aplicações tecnológicas, o estudo da dinâmica de bolhas tem sido fonte de muita pesquisa. Relacionado às tecnologias, estudos de bolhas aparecem, por exemplo, no contexto de geração de energia [128, 129] e nos processos de sonoluminescência [130], nos quais o colapso de bolhas produz pulsos luz. Do ponto de vista mais básico, pesquisadores têm estudado o processo de convecção térmica em bolhas de sabão, efeitos cooperativos na dinâmica de espumas [131, 132], bolhas de ar em água [133], entre outros [126].

Um caso muito comum, em que temos a presença de bolhas, é o processo de fervura de água [134]. Sabemos que, sob pressão atmosférica de 1 atm, quando a temperatura de uma amostra de água atinge 100°C, podemos observar o processo de formação expontânea de bolhas [135]. Apesar de ser ordinário, trata-se um de um processo bastante complexo que envolve interações térmicas entres as bolhas e a superfície aquecida e também entre as regiões onde as bolhas são formadas, além das interações hidrodinâmicas [136]. Nesse sentido, uma "abordagem de sistemas complexos" ao problema pode ser útil para extrair padrões e esclarecer aspectos dessa dinâmica aparentemente simples. De fato, alguns estudos já têm feito isso. É o caso, por exemplo, dos processos de evaporação em pequenos canais [137] ou em tubos capilares [138].

Contudo, verificamos que uma atenção menor foi dada para sistemas com várias bolhas [139]. Este será o ponto que abordaremos nesse capítulo.

Para investigar essa situação, construímos um experimento simples para extrair dados sobre a dinâmica de um conjunto de bolhas durante o processo de ebulição da água. Uma representação esquemática do experimento é mostrada na Figura 5.1. Basicamente, temos uma amostra de água fervente atravessada por um feixe laser. Monitoramos a intensidade desse feixe após o processo de ebulição se estabilizar e a Figura 5.2 mostra um sinal típico dessa dinâmica. Notamos que o sinal é altamente intermitente, caracterizado por picos e vales agudos. Apesar do sinal apresentar certa complexidade devido ao grande número de bolhas no sistema, o processo que gera as mudanças na intensidade do laser é qualitativamente simples. As bolhas são produzidas na base do recipiente que está em contato com o aquecedor. Nessa região, pequenas quantidades de água evaporam ao entrar em contato com as altas temperaturas, o que produz as bolhas. Essas bolhas (menos densas) sobem para a superfície da amostra e, durante esse movimento de ascensão, interagem com outras bolhas e com as paredes do recipiente. Quando uma ou mais bolhas atravessam o caminho óptico do laser, o sinal é espalhado produzindo um decaimento na intensidade monitorada.

O sinal da intensidade do laser, em função do tempo, será a nossa fonte de informação sobre a dinâmica do processo de ebulição da água.

5.2 Análise estatística dos dados

Devido às características do sinal de intensidade do laser, uma variável natural para investigar a dinâmica do sistema é o intervalo de tempo entre dois eventos caracterizados por grandes variações na intensidade do laser. Essa é uma análise comum na literatura de física e econofísica [140, 141, 142, 143] e tem se mostrado útil para revelar os processos que estão governando o sistema [48, 50, 49, 51].

Uma maneira eficiente de se obter esses eventos extremos é considerar um valor limiar q e registrar todos os tempos iniciais t_i para os quais o sinal do laser é menor ou igual a q. A diferença entre dois tempos consecutivos, $\tau_i = t_{i+1} - t_i$, é o chamado intervalo de retorno (esse procedimento também foi usado na seção 1.2). As linhas horizontais da Figura 5.2 representam esses intervalos de retorno para $q = 0{,}5$ e a distribuição de probabilidade de τ_i para três valores de q é mostrada na Figura 5.3.

Claramente, essas distribuições são dependentes do valor q. Além disso, sabemos que para processos estocásticos gaussianos e não correlacionados, a distribuição de τ_i segue uma distribuição exponencial [144], na forma

$$\rho(\tau) = \frac{1}{\bar{\tau}_q} e^{-\tau/\bar{\tau}_q}\,, \tag{5.1}$$

em que $\bar{\tau}_q$ é o valor médio dos intervalos de retorno ao considerarmos o valor limiar q. Na Figura 5.3 mostramos também uma comparação entre essas distribuições e as distribuições empíricas. Notamos que a concordância entre os dados empíricos e o modelo exponencial é bastante

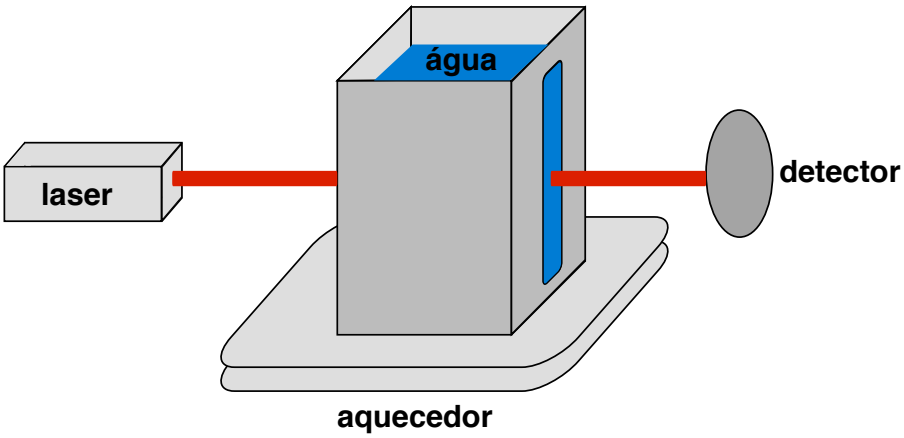


Figura 5.1: Representação esquemática da configuração experimental usada para aquisição dos dados. O porta amostra é feito de alumínio com janelas laterais de vidro, por onde passa um feixe laser de hélio-neônio com potência de 10 mW. O porta amostra, contendo aproximadamente 300 ml de água destilada, está em contado com aquecedor elétrico com potência de aproximadamente 300 W. A temperatura na região de contato entre a amostra e o aquecedor é de aproximadamente 300 °C. A intensidade do feixe laser é monitorada por um fotodiodo (Thorlabs DET100A) acoplado a um osciloscópio (Tektronix TDS5032B) com taxa de amostragem de 1 kHz. O processo de gravação se inicia após o processo de ebulição tornar-se estável e dura aproximadamente 10 minutos. Em nossos experimentos, verificamos que, evitando-se a interface ar-água, a altura do feixe laser não influencia os resultados apresentados aqui. No total, realizamos mais de 10 medidas experimentais e verificamos que os resultados obtidos praticamente não mudam quando consideramos qualquer uma dessas medidas.

ruim, em particular, as caudas das distribuições decaem muito mais lentamente do que uma função exponencial.

A grande diferença entre os resultados experimentais e o modelo da equação 5.1 é um indicativo da existência de memória na dinâmica das bolhas estudadas aqui. Para verificar ou não a existência de memória no sinal da intensidade do laser, aplicamos a análise de flutuação DFA. Esse procedimento consiste, basicamente, no cálculo da função de flutuação $F(n)$ para a série integrada e destendenciada do sinal do laser, para diferentes valores de uma escala temporal n (Apêndice A.3). Quando a série é invariante por escala (fractal), $F(n)$ segue a lei de potência $F(n) \sim n^h$, em que h (expoente de Hurst) mede o grau de correlação na série temporal: se $h = 0{,}5$ a série é não correlacionada e valores de $h \neq 0{,}5$ indicam a existência de correlações de longo alcance na série. A Figura 5.4 mostra a análise DFA para o sinal de intensidade do laser, em que encontramos $h \approx 0{,}65$ e, portanto, confirmamos a existência de memória nessa série temporal.

Alguns resultados empíricos têm mostrado que, na presença de correlações de longo alcance, a distribuição dos intervalos de retorno τ_i é bem ajustada por uma exponencial *stretched* [48, 50, 119] ou por uma distribuição de Weilbull [51], dadas por

$$\rho(\tau) \sim e^{-A(\tau/\bar{\tau}_q)^{\gamma}} \quad \text{ou} \quad \rho(\tau) \sim (\tau/\bar{\tau}_q)^{\gamma-1}\, e^{-B(\tau/\bar{\tau}_q)^{\gamma}}\,, \tag{5.2}$$

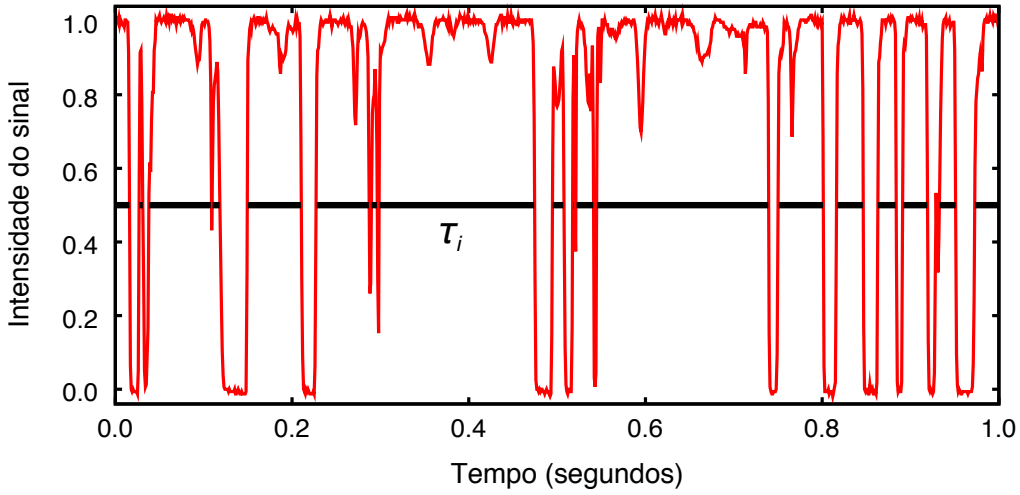


Figura 5.2: Exemplo de uma gravação típica da intensidade do feixe laser que atravessa a amostra de água em ebulição. O sinal foi escalado para que ficasse entre 0 e 1, sendo 1 correspondente a situação em que o feixe passa sem ser defletido pelas bolhas e 0 é o caso em que o feixe é completamente defletido. As linhas horizontais representam os intervalos de retorno para $q = 0{,}5$.

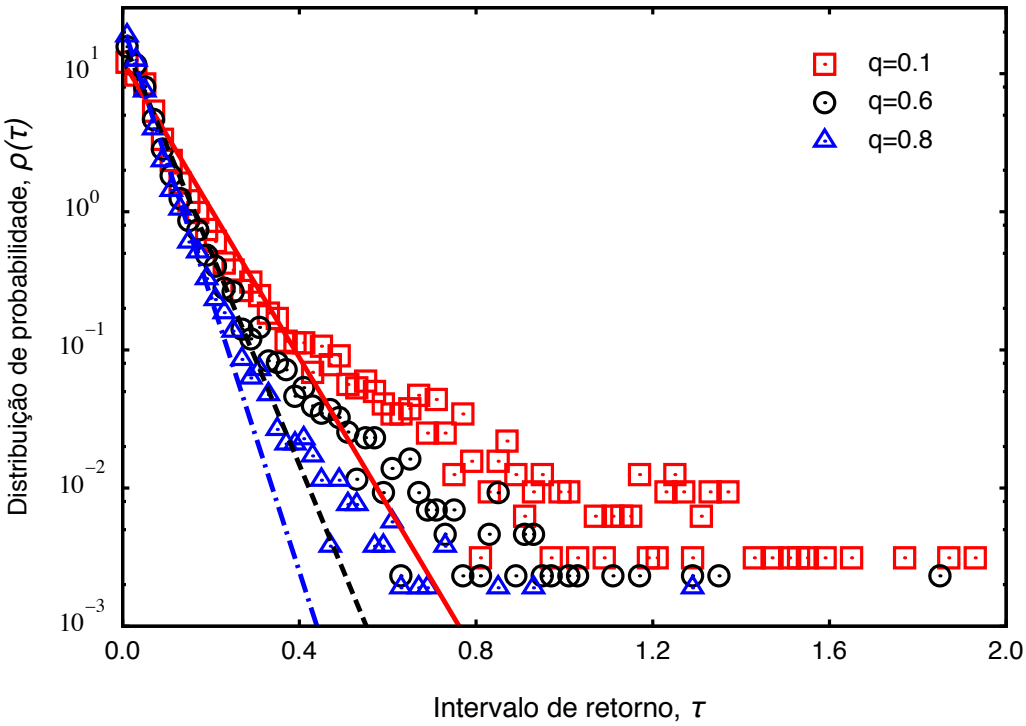


Figura 5.3: Distribuição de probabilidade dos intervalos de retorno τ para três valores de q em comparação com a distribuição exponencial da equação 5.1 para $\bar{\tau}_{0,1} = 0{,}0808$ s (linha contínua), $\bar{\tau}_{0,6} = 0{,}0564$ s (linha tracejada) e $\bar{\tau}_{0,8} = 0{,}0438$ s (linha pontilhada e tracejada).

sendo A e B constantes e γ o expoente da função de autocorrelação do processo que produz os intervalos de retorno (em nosso caso, o sinal da intensidade do laser). Além disso, Santhanam e Kantz [144] mostraram que, para um processo x_t descrito pelo movimento browniano fracionário

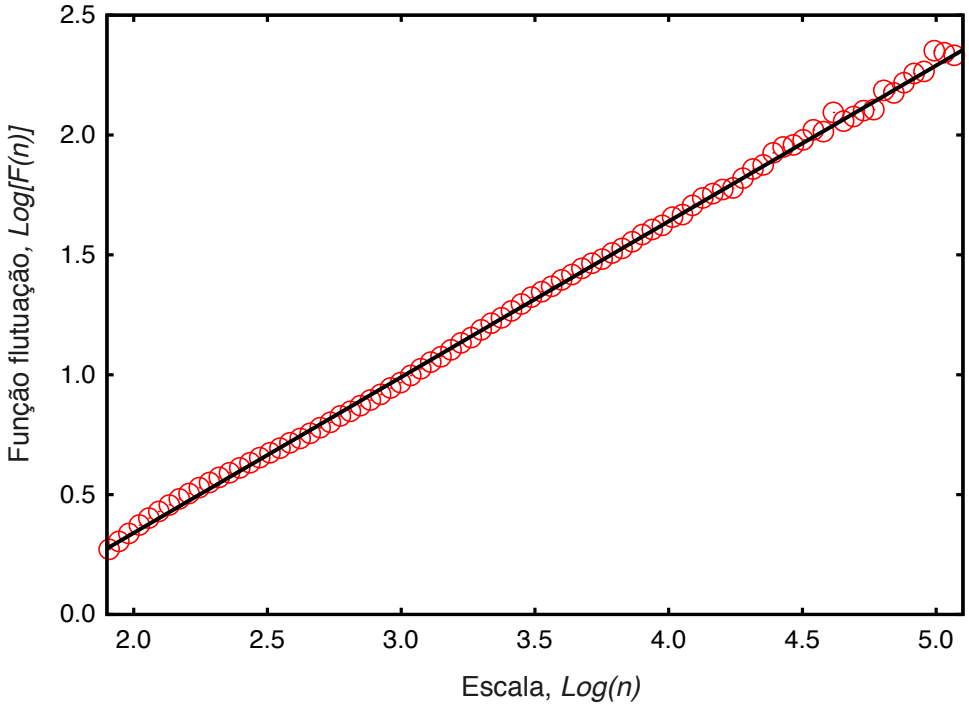


Figura 5.4: Análise DFA para as séries das intensidades do laser. Nesse gráfico, temos $\log_{10}[F(n)]$ versus $\log_{10}(n)$ em comparação com um ajuste linear, no qual encontramos $F(n) \propto n^h$, com $h \approx 0{,}65$. Aqui, n é medido em unidades de 10^{-3} segundos.

(Apêndice A.2), a distribuição dos intervalos de retorno é, de fato, uma Weilbull. Além disso, o parâmetro γ é o expoente da função de autocorrelação, $\langle x_t x_{t+\bar{t}} \rangle \sim \bar{t}^{-\gamma}$, relacionado ao expoente de Hurst (h) do movimento browniano fracionário via $\gamma = 2(1-h)$.

Sendo assim, tentamos ajustar os intervalos de retorno considerando as distribuições anteriores, utilizando a variável de escala $\xi = \tau/\bar{\tau}_q$. Ao realizarmos essa substituição, as distribuições da equação 5.2 tornam-se independentes do valor limiar q. Além disso, o valor médio de ξ torna-se unitário. Com essa condição e a normalização, as distribuições da equação 5.2 ficam dependentes apenas de γ que, por sua vez, é determinado a partir de h. Na Figura 5.5A mostramos a comparação dos dados empíricos com as distribuições da equação 5.2 ao considerarmos $\xi = \tau/\bar{\tau}_q$. Notemos que, embora tenhamos um bom colapso dos dados empíricos em uma única curva, as distribuições *stretched* ou Weilbull não são capazes de descrever os dados satisfatoriamente, por exemplo, as distribuições subestimam os dados para $\xi \in (0, 2)$ e superestimam para $\xi \in (3, 7)$. Verificamos também que a qualidade do ajuste não melhora quando obtemos γ via método dos mínimos quadrados.

Outro aspecto que investigamos foi a série dos incrementos consecutivos dos intervalos de retorno, ou seja, $\Delta\xi = \xi_{i+1} - \xi_i$. Essa análise é mostrada na Figura 5.5B, na qual observamos que a distribuição de $\Delta\xi$ não é uma gaussiana e apresenta caudas longas. De fato, é bem conhecido em teoria de probabilidade que a distribuição das diferenças entre duas variáveis aleatórias não

correlacionadas, $X - Y$, é dada pela correlação cruzada (*cross-correlation*) [145]

$$f_{X-Y}(\tau) = \int_{-\infty}^{\infty} f_X(x) f_Y(x+\tau) dx \,, \tag{5.3}$$

em que $f_X(x)$ representa a distribuição de probabilidade de X e $f_Y(x)$ a distribuição de Y. Na Figura 5.5B mostramos também essas distribuições, ao considerarmos que $f_X(x)$ e $f_Y(x)$ são distribuições exponenciais *stretched* ou de Weilbull. Novamente, o acordo entre essas distribuições da equação 5.3 (obtidas ao realizar a integração numericamente) e os dados empíricos não é bom. Em nosso caso, essa expressão deve ser vista apenas como uma aproximação, visto que as séries dos ξ_i também são correlacionadas de longo alcance (encontramos $h \approx 0{,}55$ para essas séries).

5.3 Um modelo simples para descrever os dados experimentais

Em princípio, deveríamos ser capazes de descrever completamente a dinâmica da água em ebulição. Entretanto, dificuldades técnicas relacionadas à estabilidade numérica das soluções e também à grande complexidade das condições de contorno das duas fases (líquido e vapor) torna essa tarefa muito difícil. Além disso, nosso objetivo aqui é tentar entender essa dinâmica complexa do ponto vista de um modelo mínimo, que retenha apenas os ingredientes mais relevantes para a descrição do sinal da intensidade do laser.

Qualitativamente, as bolhas se originam em posições não estáticas localizadas na base do recipiente. Entretanto, essas posições não são tão importantes (ao menos numa primeira abordagem) para o espalhamento do feixe laser. Mais do que isso, o número de bolhas e também as trajetórias que elas realizam antes de atravessar o caminho óptico do laser também podem ser desprezados nesse modelo mínimo. Desse modo, consideramos uma espécie de modelo de dois estados, no qual o que importa é se existem ou não bolhas no caminho óptico do laser. Caso existam bolhas, o feixe laser será completamente espalhado e a intensidade será nula. Por outro lado, caso não existam bolhas, o feixe laser atravessa a amostra sem ser espalhado e a intensidade será unitária.

Além dessa aproximação de dois estados para o sinal do laser, um ingrediente muito importante presente nos dados empíricos é a correlação de longo alcance revelada pela análise DFA (Figura 5.4). Nesse contexto de correlações de longo alcance, Buiatti *et al.* [146] (e também [147]) mostraram que é possível produzir sequências de símbolos (aqui, zeros e uns) com correlações utilizando dois números aleatórios não correlacionados. Para a descrição desse procedimento, consideramos uma rede unidimensional de tamanho n, na qual cada sítio representa a intensidade do feixe laser em um dado tempo. Essa rede será preenchida, a partir do primeiro

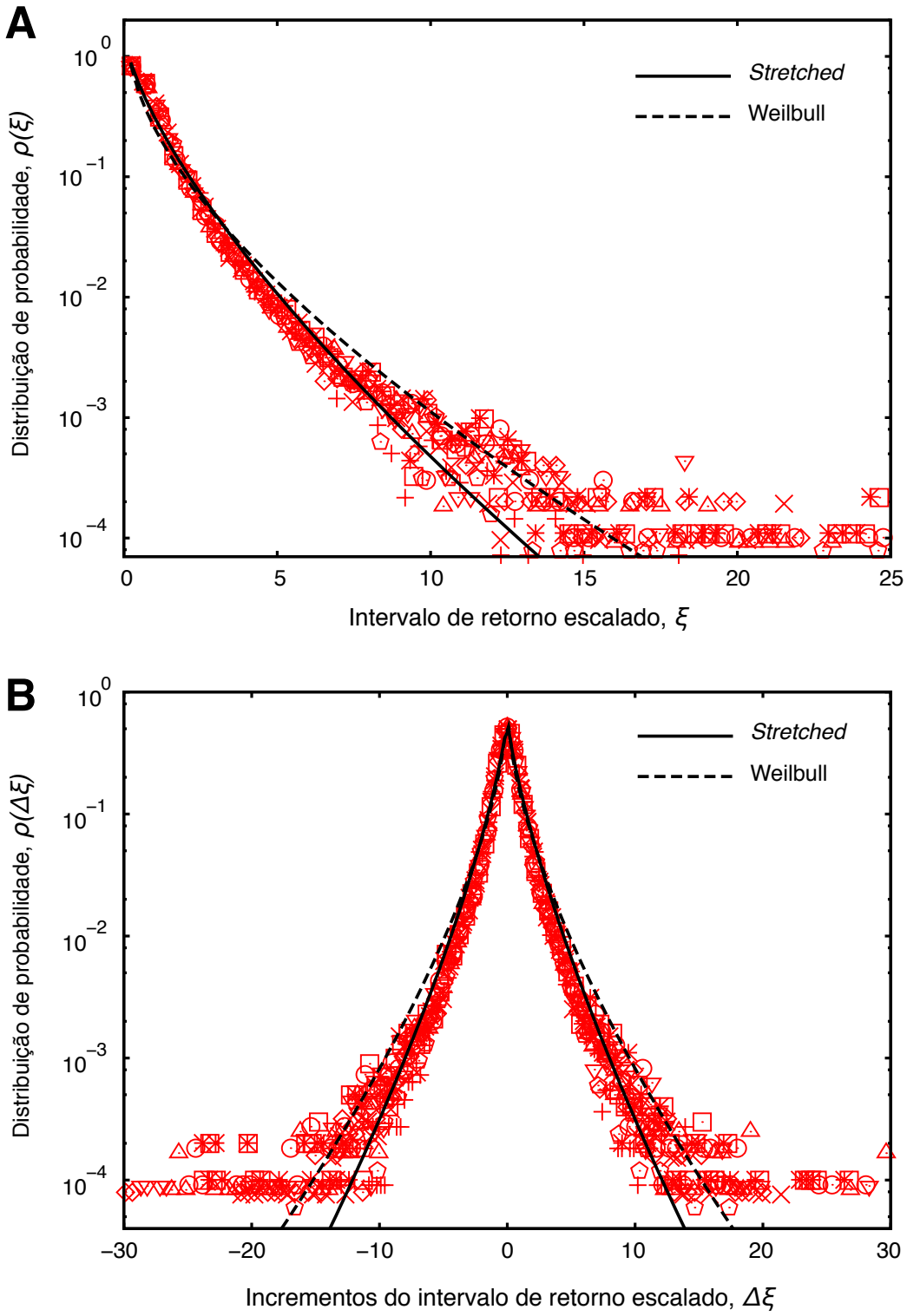


Figura 5.5: Distribuição de probabilidade (**A**) dos intervalos de retorno escalados ($\xi = \tau/\bar{\tau}_q$) e de (**B**) seus incrementos, $\Delta\xi = \xi_{i+1} - \xi_i$, para oito valores de q igualmente espaçados entre $0{,}1$ e $0{,}9$. A linha contínua é uma distribuição exponencial *stretched* e a linha tracejada uma distribuição de Weilbull, ambas com $\gamma = 2(1-h) = 0{,}7$ (equação 5.2). Veja também o Apêndice D.

sítio, sorteando-se um número aleatório discreto σ com distribuição de Bernoulli dada por

$$\rho(\sigma) = \begin{cases} p & \text{se } \sigma = 0, \\ 1-p & \text{se } \sigma = 1, \end{cases} \tag{5.4}$$

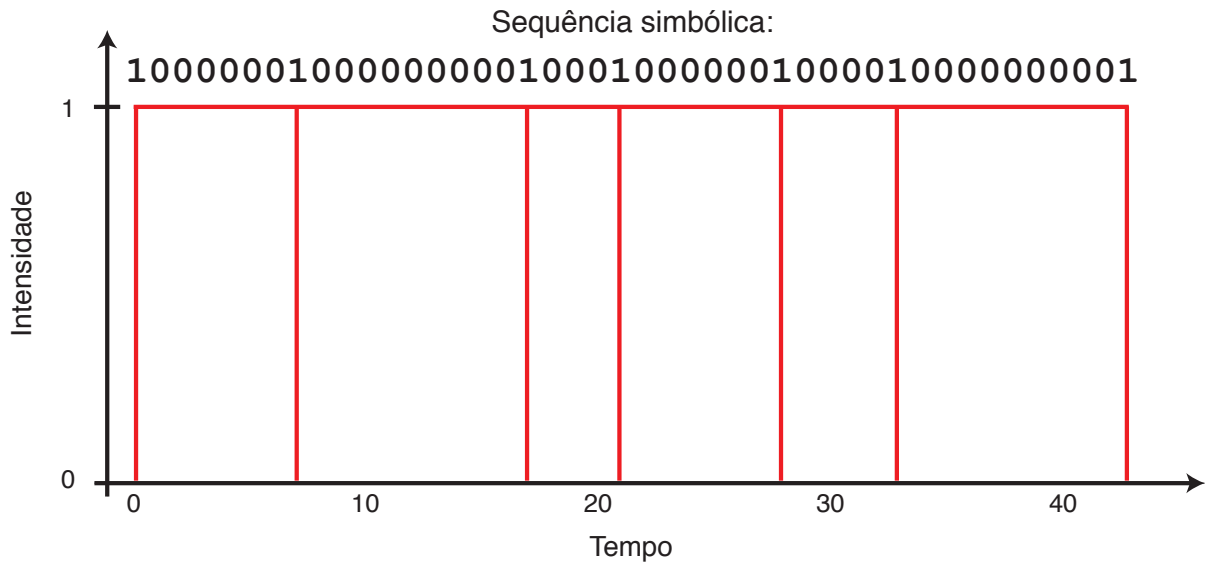


Figura 5.6: Ilustração do procedimento para produzir as intensidades do feixe laser a partir da sequências simbólicas. Neste exemplo, o primeiro sítio/símbolo da sequência é 1, o que produz uma intensidade nula. Após isso, temos seis zeros consecutivos, que representam regiões em que o feixe laser não foi espalhado. A restante da sequência é convertido em intensidade de maneira análoga.

sendo p um parâmetro. Se o número sorteado for $\sigma = 1$, existem bolhas no caminho óptico do laser e a intensidade será nula. Caso o número seja $\sigma = 0$, não existem bolhas no caminho óptico do laser e o sinal será unitário. Além disso, quando $\sigma = 0$ vamos preencher os próximos $[x]$ sítios vizinhos com zeros, sendo $[x]$ a parte inteira de um número aleatório x cuja distribuição é $P(x)$. A Figura 5.6 ilustra o procedimento para obter as intensidades do feixe laser a partir das sequências simbólicas. Notemos que nesse modelo simplificado, os intervalos de retorno serão independentes do valor limiar q. De fato, os intervalos de retorno são exatamente os comprimentos dos sítios preenchidos consecutivamente com zeros. Por conta disso, focamos nossa comparação na variável escalada ξ.

Apesar do modelo parecer bastante *ad hoc*, visto que temos uma função $P(x)$ e um parâmetro p para escolher, ao realizar as simulações verificamos que distribuições de cauda curta, tais como a exponencial, a gaussiana, a log-normal e a gamma não são capazes de produzir um ajuste melhor do que as expressões analíticas da equação 5.2. Por outro lado, ao utilizar a distribuição de Pareto (a distribuição de caudas longas mais simples), dada por

$$P(x) = \begin{cases} \alpha k^{\alpha} x^{-\alpha-1} & \text{if } x > k, \\ 0 & \text{if } x < k, \end{cases} \tag{5.5}$$

em que $k > 0$ e $\alpha > 0$ são parâmetros, a concordância com os dados empíricos foi consideravelmente melhor. A Figura 5.7 permite uma comparação visual entre os intervalos de retorno empíricos e os simulados. Para as simulações, fixamos $k = 1$ e o tamanho da rede de modo a obtermos $\sim 10^5$ intervalos de retorno (número típico encontrado nos dados experimentais).

Desse modo, o modelo possui apenas dois parâmetros a serem ajustados. Esses dois parâmetros, α e p, foram variados de maneira incremental e para cada um de seus valores, a distribuição dos intervalos de retorno foi calculada e comparada com a distribuição experimental via método dos mínimos quadrados. Os valores que minimizam as diferenças entre a distribuição empírica e a simulada são $\alpha = 1{,}8$ e $p = 0{,}2$. A Figura 5.8 mostra a comparação das distribuições empíricas de ξ e de $\Delta\xi$ com as simuladas, em que confirmamos o bom ajuste que esse modelo produz.

A simplicidade desse modelo sugere que uma solução analítica é possível. De fato, os intervalos de retorno podem ser pensados como uma soma de números aleatórios, na qual o número de termos somado é também um número aleatório. Assim, podemos escrever

$$\rho(\tau) = N \sum_{n=1}^{\infty} (1-p)^n S_n(\alpha_n, k_n)\,, \tag{5.6}$$

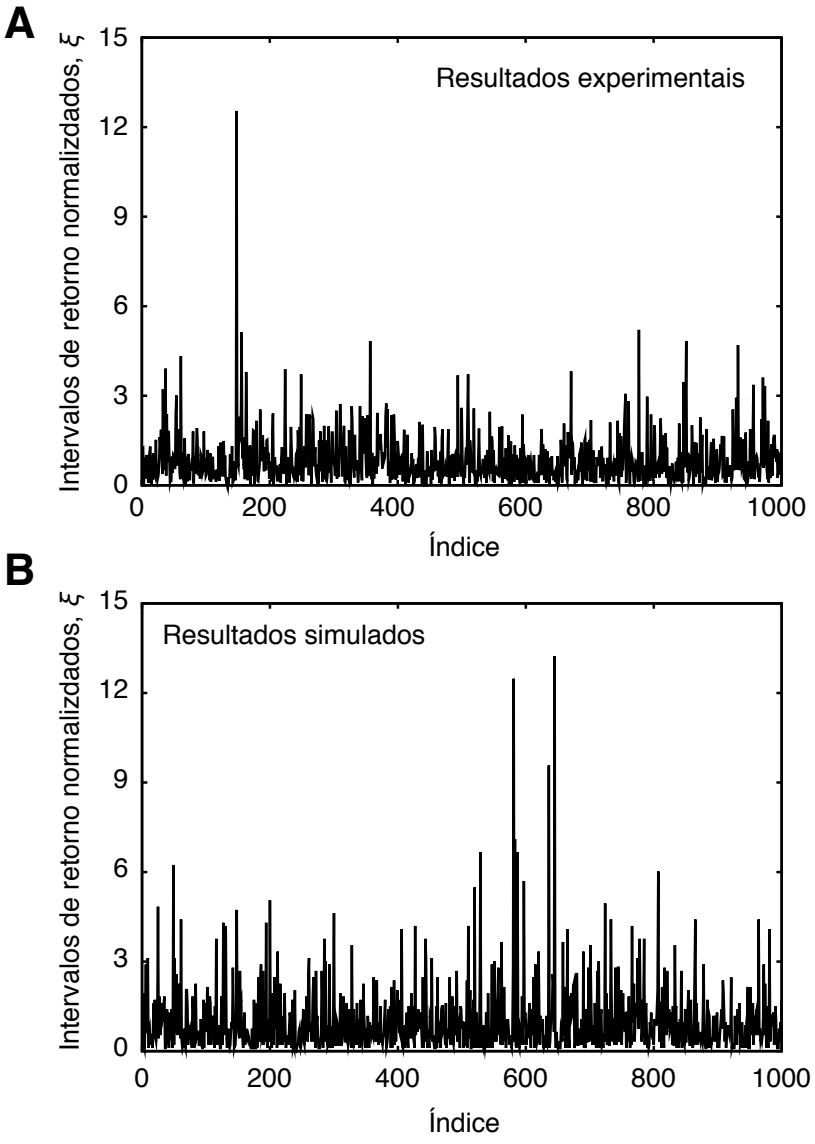


Figura 5.7: Exemplos de intervalos de retorno normalizados (**A**) experimentais e (**B**) simulados usando o modelo descrito no texto.

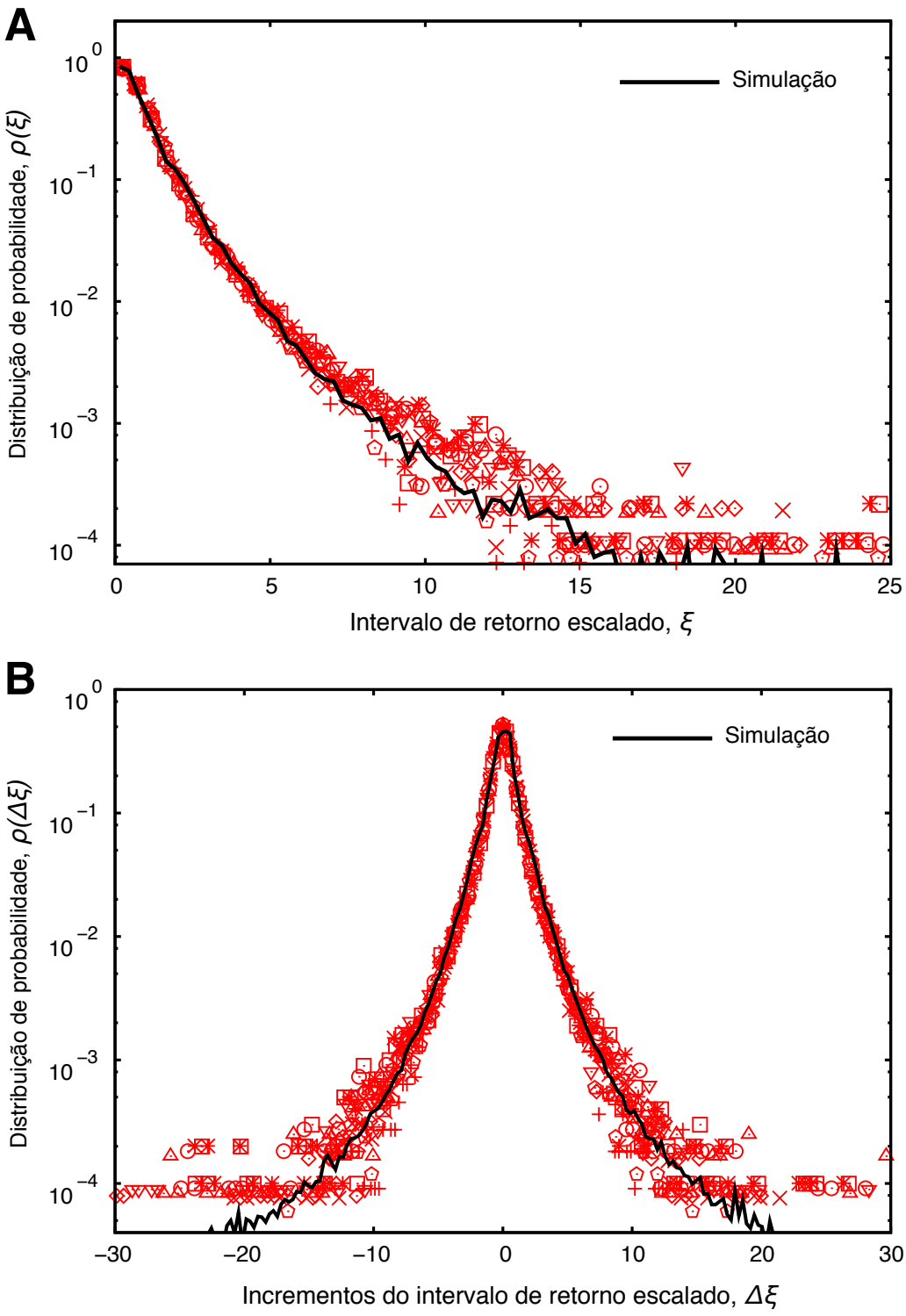


Figura 5.8: Distribuições dos (**A**) intervalos retorno normalizados ξ e dos seus (**B**) incrementos $\Delta\xi = \xi_{i+1} - \xi_i$ quando consideramos os dados experimentais (pontos) e simulados (linha contínua). Cada conjunto de pontos representa um valor de q igualmente espaçados entre $0{,}1$ e $0{,}9$. Os resultados da simulação foram obtidos considerando a distribuição de Pareto com $k = 1$ e $\alpha = 1{,}8$, $p = 0{,}2$ e uma rede com 5×10^4 sítios. Além disso, uma média sobre 200 realizações da simulação também foi considerada.

na qual N é um fator de normalização, $(1-p)^n$ é a probabilidade de somar n números consecutivos e $S_n(\alpha_n, k_n)$ é a distribuição da soma de n variáveis aleatórias do tipo Pareto. Entretanto, obter

uma expressão geral para $S_n(\alpha_n, k_n)$ não é uma tarefa fácil e representa um problema em aberto em teoria de probabilidades [148, 149, 150]. Por conta disso, limitamos nossa discussão aos resultados numéricos obtidos com o modelo descrito neste capítulo.

5.4 Conclusões e perspectivas

Apresentamos neste capítulo uma análise estatística sobre a dinâmica de bolhas em uma amostra de água em ebulição. Para extrair informações sobre essa dinâmica, construímos um experimento simples baseado no espalhamento de um feixe laser que atravessa a amostra em ebulição e tem sua intensidade gravada. Essa série temporal das intensidades do feixe laser foi a nossa fonte de informação sobre essa dinâmica complexa. Devido às características dessa série, verificamos que uma boa variável para investigar a dinâmica é o intervalo de retorno. Após calcular as distribuições dos intervalos de retorno, verificamos que essas distribuições não são exponenciais e que a intensidade do laser apresenta correlações de longo alcance. Esse resultado indica fortemente que o processo que produz as bolhas pode ser também correlacionado. Verificamos ainda que um modelo minimalista foi capaz de reproduzir alguns aspectos presentes nos dados empíricos. Além disso, o modelo também sugere que os principais ingredientes que produzem essa dinâmica no sinal do laser são as correlações e a distribuição do tipo lei de potência relacionada ao tempo no qual as bolhas passam através do caminho óptico do laser.

Naturalmente, outras análises podem ainda ser feitas. É o caso, por exemplo, de se estudar os estágios anteriores à estabilização do processo de ebulição. Durante o início da transição de fase líquido-vapor, podem haver outros comportamentos diferentes dos que foram obtidos aqui. Além disso, também seria interessante verificar se a potência do aquecedor tem alguma influência sobre essa dinâmica; por exemplo, talvez para grandes potências a ebulição do fluído torne-se mais aleatória, produzindo distribuições de intervalo de retorno exponenciais. Do ponto vista mais aplicado, poderíamos também investigar fluídos com diferentes viscosidades e tentar relacionar esse parâmetro do material com aqueles obtidos da série temporal das intensidades do feixe laser.

Capítulo 6

Entropia e complexidade de permutação de estruturas bidimensionais

Neste capítulo, apresentaremos um método para medir a complexidade de estruturas bidimensionais [21], tais como imagens. O método é, na verdade, uma extensão de uma medida de complexidade para séries temporais que foi proposta recentemente. Trata-se da entropia e complexidade de permutação que já utilizamos no capítulo 2 para medir a complexidade de músicas; entretanto, não faremos referência àqueles resultados para manter os capítulos independentes. Após apresentarmos uma extensão do procedimento unidimensional, aplicaremos o método em $i)$ superfícies fractais, em que será possível distinguir entre diferentes dimensão fractais; $ii)$ texturas de cristais líquidos, em que o método permitirá identificar transições de fases e também diferenciar texturas características desses materiais; $iii)$ superfícies construídas somando-se os valores dos spins do modelo Ising, em que o método irá identificar a temperatura crítica do modelo além de mostrar-se estável.

6.1 Introdução e apresentação do problema

Assim como os estudos apresentados nos capítulos anteriores, uma grande parte dos trabalhos relacionados a sistemas complexos está focada em investigar dados empíricos com o objetivo de extrair padrões, regularidades ou leis que estejam governando a dinâmica do sistema. Nessas investigações, o conceito de medidas de complexidade aparece com frequência. Medidas de complexidade podem comparar séries temporais e classificá-las, por exemplo, entre regular, caótica ou aleatória [86], enquanto outras medidas podem diferenciar entre tipos de correlações [151]. Alguns exemplos dessas medidas de complexidade incluem a complexidade algorítmica [152], entropias [87], entropias relativas [153], dimensões fractais [154] e expoentes de Lyapunov [155]. Esses trabalhos seminais motivaram e ainda motivam novas definições, de modo que temos atualmente um grande número de medidas de complexidade, as quais têm se mostrado úteis em

aplicações na medicina [156, 157], ecologia [158, 159, 160, 161], astrofísica [162, 163, 164] e música [67, 165].

Entretanto, foi com surpresa que percebemos que a maioria dessas medidas aplica-se somente a dados unidimensionais, tais como séries temporais. Uma atenção consideravelmente menor foi dada para estruturas de maior dimensão, como imagens. Exceções incluem os trabalhos de Grassberger [166] e mais recentemente os trabalhos das referências [167, 168, 169]. Contudo, como foi argumentado por Bandt e Pompe [85], a maioria das medidas de complexidade é dependente de algoritmos especializados ou receitas para processar os dados que, por sua vez, dependem de parâmetros ajustáveis. Por conta disso, muitas vezes, é bastante complicado reproduzir resultados de outros trabalhos sem o conhecimento desses detalhes metodológicos.

Além de apontar o problema anterior, Bandt e Pompe também propuseram uma solução, introduzindo o que eles chamaram de uma medida "natural" de complexidade para séries temporais: a entropia de permutação [85]. Existem muitas aplicações recentes que fazem uso dessa nova técnica em dados unidimensionais e confirmam a sua utilidade [170, 171, 172, 173, 174, 175, 176]. Uma dessas aplicações foi feita por Rosso *et al.* [86], na qual as ideias de Bandt e Pompe foram aplicadas em conjunto com um índice entrópico relativo [88] para diferenciar séries temporais caóticas de séries temporais estocásticas. Rosso *et al.* construíram um diagrama chamado de *complexity-entropy causality plane* (o qual foi primeiramente proposto por López-Ruiz *et al.* [177]). Esse diagrama é composto pelos valores do índice entrópico relativo em função da entropia de permutação. Intrigantemente, séries temporais de origem caótica e estocástica se localizam em regiões diferentes nesse diagrama.

Neste capítulo, mostraremos uma maneira conveniente para se estender o método anterior de Rosso *et al.* para estruturas de maior dimensão, como imagens.

6.2 Entropia e complexidade de permutação em duas dimensões

A ideia de Bandt e Pompe [85] foi definir uma medida que pode ser facilmente aplicada a qualquer tipo de série temporal. O método baseia-se em associar uma sequência simbólica a segmentos da série temporal por meio de um ordenamento local. Esses "símbolos" ou estados definirão probabilidades que serão usadas no cálculo dos índices entrópicos. Para propósito de definição, consideramos uma série temporal $\{x_t\}_{t=1,\ldots,n}$ composta por n elementos e também os vetores de dimensão $d > 1$, definidos por

$$(\vec{s}) \mapsto (x_{s-(d-1)}, x_{s-(d-2)}, \ldots, x_{s-1}, x_s) \, , \tag{6.1}$$

em que $s = d, d+1, \ldots, n$. Em seguida, para todos os $(n - d + 1)$ vetores, calculamos as permutações $\pi = (r_0, r_1, \ldots, r_{d-1})$ de $(0, 1, \ldots, d-1)$ definidas pelo ordenamento

$$x_{s-r_{d-1}} \leq x_{s-r_{d-2}} \leq \cdots \leq x_{s-r_1} \leq x_{s-r_0}. \tag{6.2}$$

As $d!$ permutações possíveis definem os estados acessíveis ao sistema. Cada permutação π possui probabilidade

$$p(\pi) = \frac{\#\{s | s \leq N - d + 1;\ (\vec{s}) \text{ do tipo } \pi\}}{N - d + 1}\,, \tag{6.3}$$

de ser encontrada, em que o símbolo $\#$ representa o número de ocorrências da permutação π. Essas probabilidades formam o conjunto de probabilidades $P = \{p(\pi)\}$ que será usado no cálculo dos índices entrópicos (um exemplo do caso unidimensional foi dado na seção 2.3).

Aqui, nosso objetivo é estender esse procedimento para estruturas de dados com dimensão maior do que um, tais como imagens. Para isso, consideremos agora que, no lugar da série temporal, tenhamos uma matriz bidimensional $\{y_i^j\}_{i=1,\ldots,n_x}^{j=1,\ldots,n_y}$ de tamanho $n_x \times n_y$. Essa matriz pode representar, por exemplo, uma imagem na qual cada elemento y_i^j é um pixel dessa imagem. Em analogia ao vetor $(\vec{s})$, definimos as matrizes de tamanho $d_x \times d_y$ $(d_x, d_y > 1)$ dadas por

$$(s_x, s_y) \mapsto \begin{pmatrix} y_{s_x-(d_x-1)}^{s_y-(d_y-1)} & y_{s_x-(d_x-2)}^{s_y-(d_y-1)} & \cdots & y_{s_x-1}^{s_y-(d_y-1)} & y_{s_x}^{s_y-(d_y-1)} \\ y_{s_x-(d_x-1)}^{s_y-(d_y-2)} & y_{s_x-(d_x-2)}^{s_y-(d_y-2)} & \cdots & y_{s_x-1}^{s_y-(d_y-2)} & y_{s_x}^{s_y-(d_y-2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{s_x-(d_x-1)}^{s_y-1} & y_{s_x-(d_x-2)}^{s_y-1} & \cdots & y_{s_x-1}^{s_y-1} & y_{s_x}^{s_y-1} \\ y_{s_x-(d_x-1)}^{s_y} & y_{s_x-(d_x-2)}^{s_y} & \cdots & y_{s_x-1}^{s_y} & y_{s_x}^{s_y} \end{pmatrix}\,, \tag{6.4}$$

nas quais $s_x = d_x, d_x + 1, \ldots, n_x$ e $s_y = d_y, d_y + 1, \ldots, n_y$. Em seguida, para cada uma dessas $(n_x - d_x + 1)(n_y - d_y + 1)$ matrizes, calculamos as permutações

$$\pi = [(r_0, u_0), (r_1, u_0), \ldots, (r_{d_x-1}, u_0), \ldots, (r_0, u_{d_y-1}), (r_1, u_{d_y-1}), \ldots, (r_{d_x-1}, u_{d_y-1})]$$

de $(0, 1, \ldots, d_x\, d_y - 1)$, definidas por

$$\begin{aligned} y_{s_x-r_{d_x-1}}^{s_y-u_{d_y-1}} \leq y_{s_x-r_{d-2}}^{s_y-u_{d_y-1}} \leq \cdots \leq y_{s_x-r_1}^{s_y-u_{d_y-1}} \leq y_{s_x-r_0}^{s_y-u_{d_y-1}} \leq \ldots \\ \leq y_{s_x-r_{d_x-1}}^{s_y-u_0} \leq y_{s_x-r_{d-2}}^{s_y-u_0} \leq \cdots \leq y_{s_x-r_1}^{s_y-u_0} \leq y_{s_x-r_0}^{s_y-u_0}\,. \end{aligned} \tag{6.5}$$

O sistema agora pode acessar $(d_x\, d_y)!$ estados, para os quais calculamos o conjunto de probabilidades $P = \{p(\pi)\}$ usando a frequência relativa dada por

$$p(\pi) = \frac{\#\{(s_x, s_y) | s_x \leq n_x - d_x + 1 \text{ e } s_y \leq n_y - d_y + 1;\ (s_x, s_y) \text{ é do tipo } \pi\}}{(n_x - d_x + 1)(n_y - d_y + 1)}\,. \tag{6.6}$$

Para melhor entendimento, a Figura 6.1 ilustra esse procedimento para uma matriz pequena.

Naturalmente, o procedimento que escolhemos para ordenar os elementos da matriz (s_x, s_y) e definir as permutações π não é mais único como no caso unidimensional. Por exemplo, ao invés de ordenar os elementos linha por linha, poderíamos ter ordenado coluna por coluna. Contudo, essas diferentes definições mudariam apenas os "nomes" dos estados, de modo que o conjunto $P = \{p(\pi)\}$ permaneceria idêntico. Assim, não há perda de generalidade em assumir uma dada ordem para definir π. Além disso, vale notar que esse procedimento pode ser facilmente

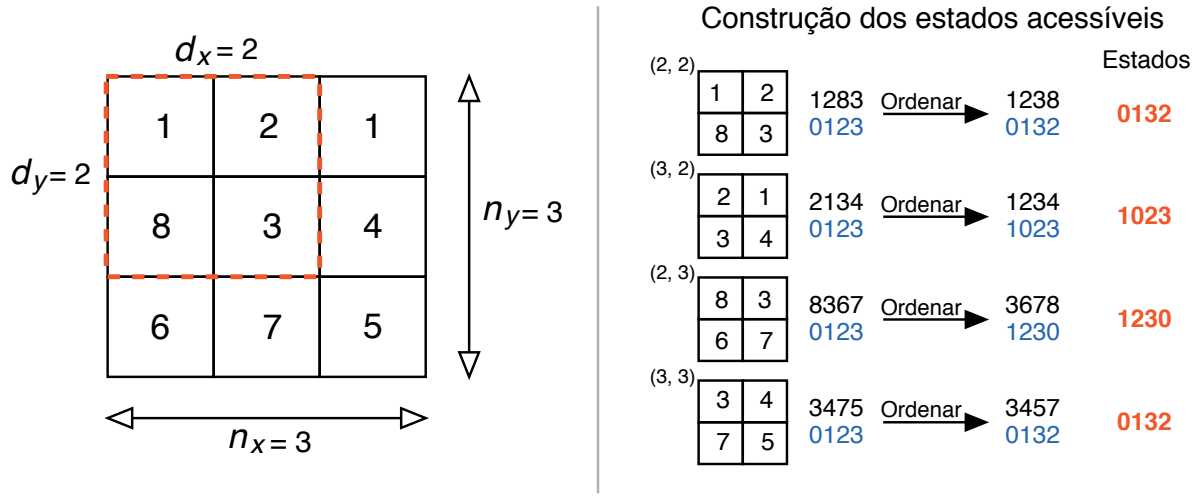


Figura 6.1: Representação esquemática da construção dos estados acessíveis. Neste exemplo, temos uma uma matriz 3×3 (painel da esquerda) e escolheremos as dimensões *embedding* $d_x = 2$ e $d_y = 2$. No painel da direita ilustramos a construção dos estados. O primeiro passo é obter a sub-matriz correspondente a $s_x = 2$ e $s_y = 2$ que tem como elementos $(1, 2, 8, 3)$. Após ordenar os seus elementos, obtemos o estado "0132". Em seguida, movemos para a próxima sub-matriz $s_x = 3$ e $s_y = 2$ que tem como elementos $(2, 1, 3, 4)$ e que, após ordenados, levam ao estado "1023". As outras duas sub-matrizes geram os estados "1230" e "0132". De posse dos estados, podemos estimar as probabilidades $p(\pi)$, as quais são: $p(\text{"0132"}) = 2/4 = 0{,}5$, $p(\text{"1023"}) = 1/4 = 0{,}25$, $p(\text{"1230"}) = 1/4 = 0{,}25$ e $p = 0$ para as demais permutações de π. Finalmente, usamos essas probabilidades nas equações 6.7 e 6.9 para obter $H \approx 0{,}33$ e $C \approx 0{,}27$.

estendido para estruturas de dimensão maior (por exemplo, as imagens volumétricas do cérebro obtidas via ressonância magnética) e que esse procedimento recupera o caso unidimensional ao fazermos $n_y = 1$ e $d_y = 1$. Aqui, por simplicidade, focaremos nossas análises em estruturas bidimensionais.

Os parâmetros d_x e d_y (chamadas de dimensão *embedding*) têm um papel importante para definir o conjunto de probabilidades P, pois eles determinam o número de estados acessíveis. No caso unidimensional, é comum escolher $d! \ll n$ para obter uma boa estatística. De fato, Bandt e Pompe recomendam o uso de $d = 3, \ldots, 7$ [85]. Para o caso bidimensional, uma relação similar deve valer, especificamente: $(d_x d_y)! \ll n_x n_y$.

Para prosseguir, devemos reescrever as medidas entrópicas usadas por Bandt e Pompe [85] e por Rosso *et al.* [86]. A primeira delas é chamada entropia de permutação normalizada [85] e é obtida aplicando-se a entropia de Shannon ao conjunto das probabilidades $P = \{p(\pi)\}$, *i.e.*,

$$H[P] = \frac{S[P]}{S_{\max}}\,, \tag{6.7}$$

em que

$$S[P] = -\sum p(\pi) \log p(\pi) \tag{6.8}$$

e $S_{\max} = \log[(d_x d_y)!]$. O valor de $S_{\max}$ é obtido quando todos os $(d_x d_y)!$ estados são equipro-

váveis, ou seja, $P = P_e = 1/(d_x d_y)!$. Devido a essa normalização, $0 \leq H[P] \leq 1$, sendo que o limite superior ocorrerá para matrizes complemente aleatórias. Por outro lado, esperamos valores de $H[P] < 1$ para matrizes que possuam algum tipo de padrão mais complexo.

A outra medida [86] é definida por

$$C[P] = Q[P, P_e]\, H[P]\,, \tag{6.9}$$

no qual $Q[P, P_e]$ é uma métrica entrópica entre o conjunto de probabilidades $P = \{p(\pi)\}$ e o estado equiprovável $P_e = 1/(d_x d_y)!$. A quantidade $Q[P, P_e]$ é muitas vezes denominada "desequilíbrio" e é definida em termos da divergência de Jensen-Shannon [89] (ou ainda em termos da divergência Kullback-Leibler simetrizada [90]) e pode ser escrita como

$$Q[P, P_e] = \frac{S[(P + P_e)/2] - S[P]/2 - S[P_e]/2}{Q_{\mathsf{max}}}\,, \tag{6.10}$$

em que

$$Q_{\mathsf{max}} = -\frac{1}{2}\left\{\frac{(d_x d_y)! + 1}{(d_x d_y)!}\log[(d_x d_y)! + 1] - 2\log[2(d_x d_y)!] + \log[(d_x d_y)!]\right\} \tag{6.11}$$

é o valor máximo do numerador de $Q[P, P_e]$, obtido quando um dos componentes de P é igual a um e todos os outros são nulos.

A medida C pode quantificar a existência de estruturas correlacionadas nas matrizes, fornecendo informações adicionais que podem não estar sendo levadas em conta pela entropia de permutação. Além disso, para um dado valor de $H[P]$ existe um intervalo de possíveis valores para $C[P]$ [93]. Esse comportamento foi a principal razão pela qual Rosso *et al.* [86] propuseram empregar um diagrama de $C[P]$ versus $H[P]$ como uma ferramenta de diagnóstico, construindo o que vem sendo chamado de *complexity-entropy causality plane*. Essa também será a nossa abordagem para medir a complexidade e distinguir diferentes estruturas bidimensionais.

Nas próximas três seções aplicaremos esse método em diferentes cenários visando a verificar sua eficiência e utilidade.

6.3 Aplicação I: superfícies fractais

A primeira aplicação que trabalhamos está relacionada a superfícies fractais. Verificaremos se o nosso método é capaz de distinguir diferentes superfícies com distintas dimensões fractais.

Para isso, geramos superfícies fractais usando o algoritmo *random midpoint displacement* introduzido por Fournier *et al.* [178]. Para apresentar esse algoritmo, imagine um quadrado no qual atribuímos um valor aleatório a cada vértice para representar a altura da superfície. Em seguida, adicionamos um ponto localizado no centro do quadrado. A esse novo ponto atribuímos um valor igual ao valor médio dos quatro vértices do quadrado inicial, somado a um número aleatório gaussiano de média nula e desvio padrão δ_1. Adicionamos ainda outros quatro pontos

localizados no meio dos segmentos que conectam os vértices do quadrado inicial. Cada um desse novos pontos terá valor igual ao valor médio entre os vértices mais próximos e o ponto do meio, somando-se ainda um número aleatório de média nula e desvio padrão δ_1. Assim, nesse estágio do algoritmo, temos 9 pontos que podem ser imaginados como quatro novos quadrados. Repetimos o procedimento anterior para cada um desses 4 novos quadrados e usando δ_2. A repetição recursiva desse processo por k vezes e usando $\delta_k = \delta_0\, 2^{-\frac{kh}{2}}$ irá produzir uma superfície quadrada de tamanho $(2^k + 1) \times (2^k + 1)$. Aqui, h é o expoente de Hurst e $D = 3 - h$ é a dimensão fractal da superfície. Os parâmetros δ_0 e k não têm influência nas propriedades fractais da superfície, produzindo apenas um efeito multiplicativo global.

A Figura 6.2 mostra algumas superfícies geradas usando esse procedimento para diferentes valores de h. Notamos que, para valores pequenos de h, a superfície exibe uma alternância entre picos e vales muito mais frequente do que para valores maiores de h. Para valores maiores de h as superfícies mostram-se mais suaves, refletindo o comportamento persistente induzido por $h > 0{,}5$.

Aplicamos o nosso método nessas superfícies para verificar se os índices H e C são sensíveis às mudanças no expoente de Hurst h, como mostra a Figura 6.3. Nesses gráficos tridimensionais, mostramos a localização no plano complexidade-entropia em função de h para superfícies de tamanho 1025×1025 ($k = 10$). Na Figura 6.3A, usamos $d_x = 2$ e $d_y = 3$ (círculos), e $d_x = 3$ e $d_y = 2$ (quadrados). Notamos que os valores de H e C são praticamente invariantes perante a rotação $d_x \to d_y$ e $d_y \to d_x$. Essa invariância está relacionada ao fato de que essas superfícies fractais não possuem nenhuma direção preferencial. Na Figura 6.3B, empregamos $d_x = 3$ e $d_y = 3$, obtendo basicamente o mesmo perfil da Figura 6.3A; entretanto, mudanças nos valores de H e C são observadas devido ao crescimento no número de estados acessíveis ao sistema. Esses resultados indicam que nosso método identifica e permite distinguir diferentes superfícies fractais. Mais do que isso, verificamos que o método é bastante robusto perante a várias realizações do algoritmo *random midpoint displacement*, visto que os valores de H e C praticamente não mudam. Por exemplo, o desvio padrão dessas quantidades é menor do que 10^{-4} quando considerando $k = 10$ e 50 realizações do algoritmo.

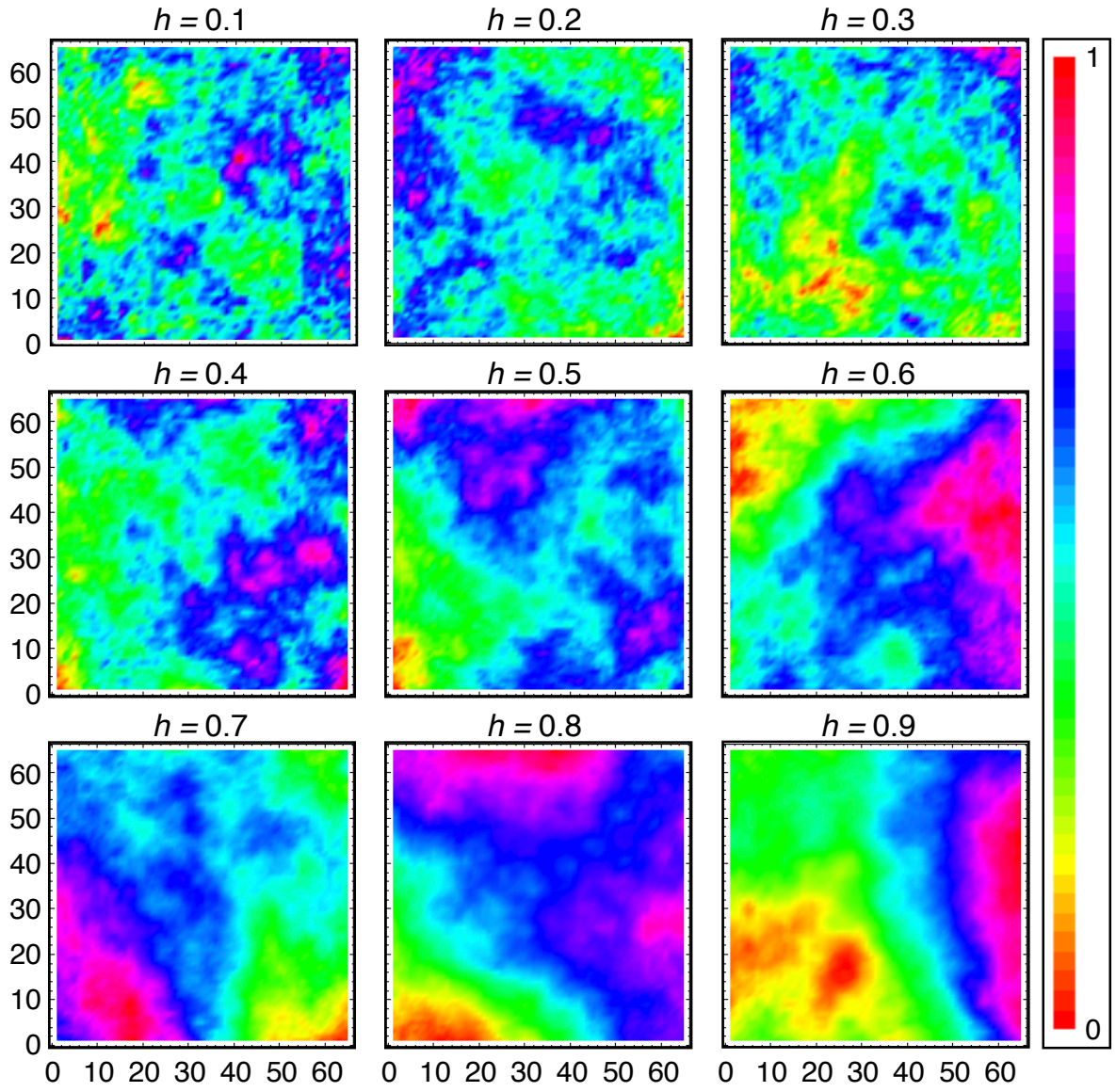


Figura 6.2: Exemplos de superfícies fractais geradas pelo uso do método *random midpoint displacement*. Essas são superfícies de tamanho 65×65 ($k = 6$) para diferentes valores do expoente de Hurst h. Aqui, para melhor visualização, os valores das superfícies foram escalados para ficarem entre 0 e 1.

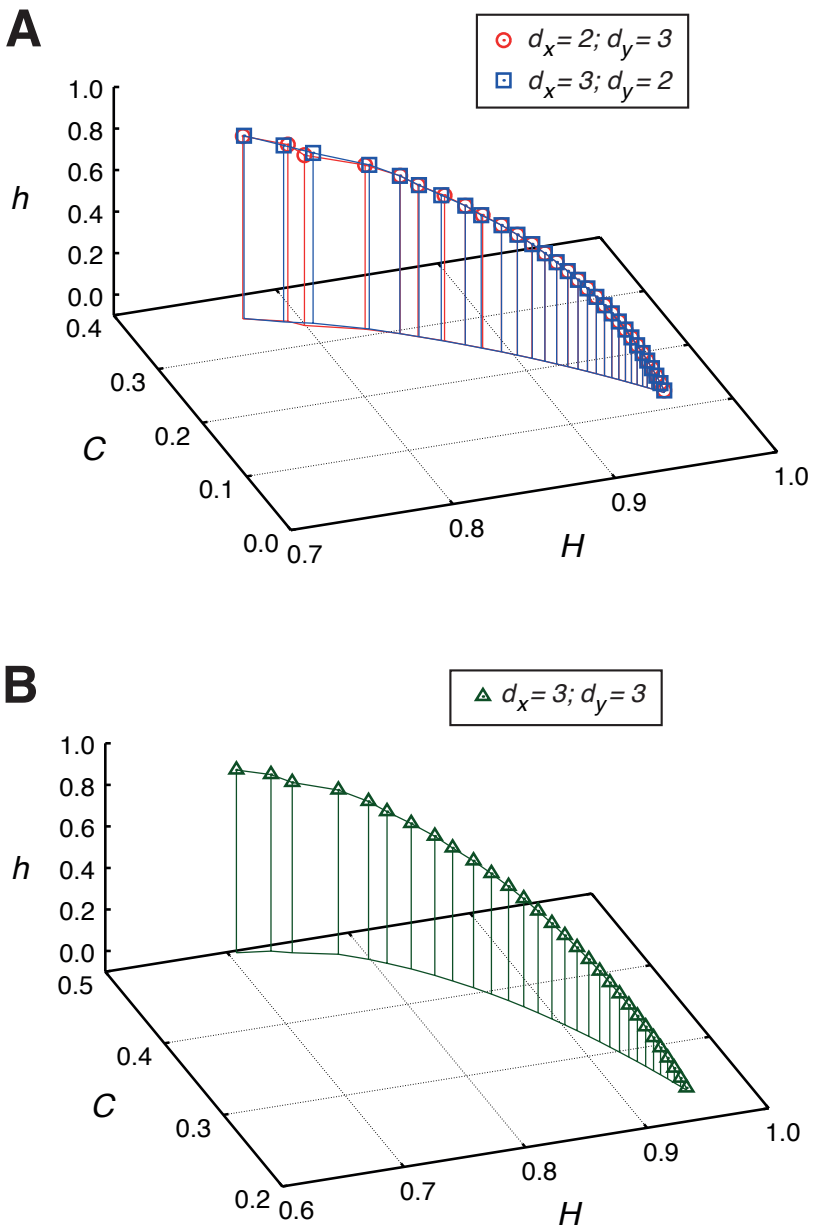


Figura 6.3: Dependência dos índices entrópicos H e C em relação ao expoente de Hurst h. Esses valores foram calculados usando-se superfícies de tamanho 1025×1025 ($k = 10$). Em (**A**) mostramos C e H versus h para $d_x = 2$ e $d_y = 3$ (círculos) e também para $d_x = 3$ e $d_y = 2$ (quadrados). Em (**B**) mostramos a mesma análise para $d_x = d_y = 3$.

6.4 Aplicação II: texturas de cristais líquidos

Outra aplicação interessante que efetuamos está relacionada aos padrões exibidos por cristais líquidos, as chamadas texturas. Essas texturas (imagens) são geralmente obtidas pela observação, em microscópio de luz polarizada, de uma amostra de cristal líquido posicionada entre dois polarizadores cruzados. Essas texturas são conhecidas por fornecerem informações úteis a respeito da estrutura macroscópica do cristal líquido. Por exemplo, diferentes "fases" líquido-cristalinas possuem texturas típicas e diferentes. Além disso, estudando a evolução dessas imagens é possível identificar transições de fase do material.

Primeiramente, estudamos um sistema simples que possui apenas duas fases líquido-cristalinas. Trata-se de um cristal líquido liotrópico que, do ponto de vista químico é, geralmente, formado por uma mistura homogênea de água, sal e sabão. O sabão ou *surfatante* é composto por moléculas que possuem uma parte polar ligada a uma cadeia carbônica longa apolar. A parte polar é altamente solúvel em água, enquanto a parte apolar não. Em concentrações e temperaturas específicas [179], aparecem nessa mistura aglomerados de moléculas (as micelas) que podem assumir várias formas. Em nosso caso, em particular, elas assumem formas cilíndricas (a chamada fase nemática cilíndrica). Essas estruturas apresentam uma tendência a se alinharem, gerando uma anisotropia no sistema, a qual, por sua vez, pode ser observada nas texturas. A Figura 6.4 mostra três dessas texturas em diferentes temperaturas. Essas imagens foram obtidas na página de internet do *Liquid Crystal Institute* da *Kent State University* [180]. Nas imagens à esquerda e à direita, o cristal líquido encontra-se na fase isotrópica e na imagem do centro está na fase nemática cilíndrica. Aqui, já podemos notar que o padrão é muito mais complexo na fase nemática,

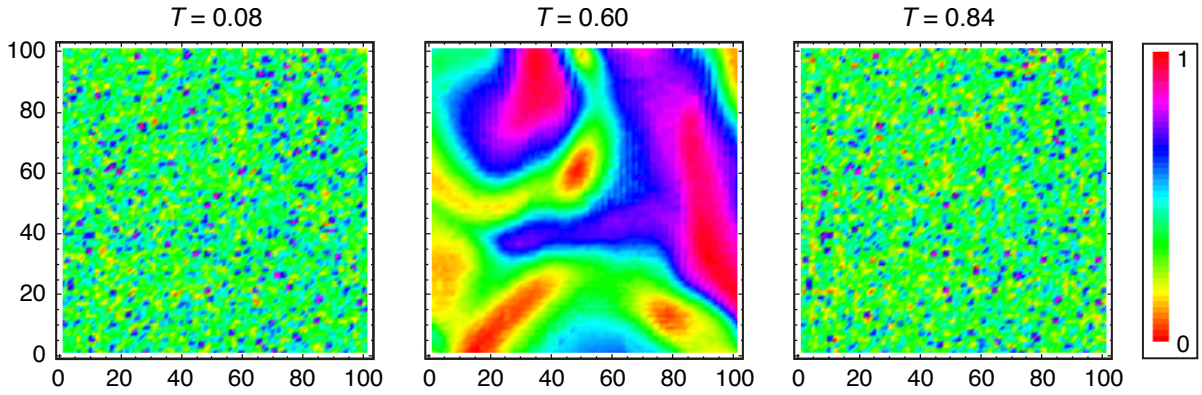


Figura 6.4: Texturas características de um cristal líquido liotrópico para diferentes temperaturas e fases. O sistema liotrópico usado aqui é uma mistura de laurato de potássio ($\approx 27{,}00\%$), decanol ($\approx 6{,}24\%$) e óxido de deutério ($\approx 66{,}76\%$). Nessas concentrações, ao aumentar a temperatura, obtemos a sequência de transição de fase isotrópica $\rightarrow$ nemática $\rightarrow$ isotrópica [179]. Essas imagens foram obtidas pela observação em microscópico óptico de um capilar plano contendo a mistura em várias temperaturas. Os gráficos dessa figura representam os valores médios das camadas *RGB* da imagem original. Além disso, a temperatura está em unidades arbitrárias, uma vez que não possuímos dados sobre a verdadeira temperatura.

enquanto para a fase isotrópica temos basicamente um padrão aleatório.

Calculamos H e C em função da temperatura para diferentes valores das dimensões *embedding* d_x e d_y, como mostra a Figura 6.5. Nesses gráficos, as diferentes regiões sombreadas representam as diferentes fases do cristal líquido (conhecidas *a priori*). Observamos que as transições são propriamente identificadas independentemente dos valores de d_x e d_y. Contudo, as Figuras 6.5C e 6.5D mostram pequenas diferenças na dependência de H e C com a temperatura ao considerarmos $d_x = 2$ e $d_y = 3$ ou $d_x = 3$ e $d_y = 2$. Essas diferenças, provavelmente, estão relacionadas aos efeitos de superfície, uma vez que esse material é normalmente colocado em tubos capitares alongados, e as paredes do tubo podem induzir a uma direção preferencial para as moléculas.

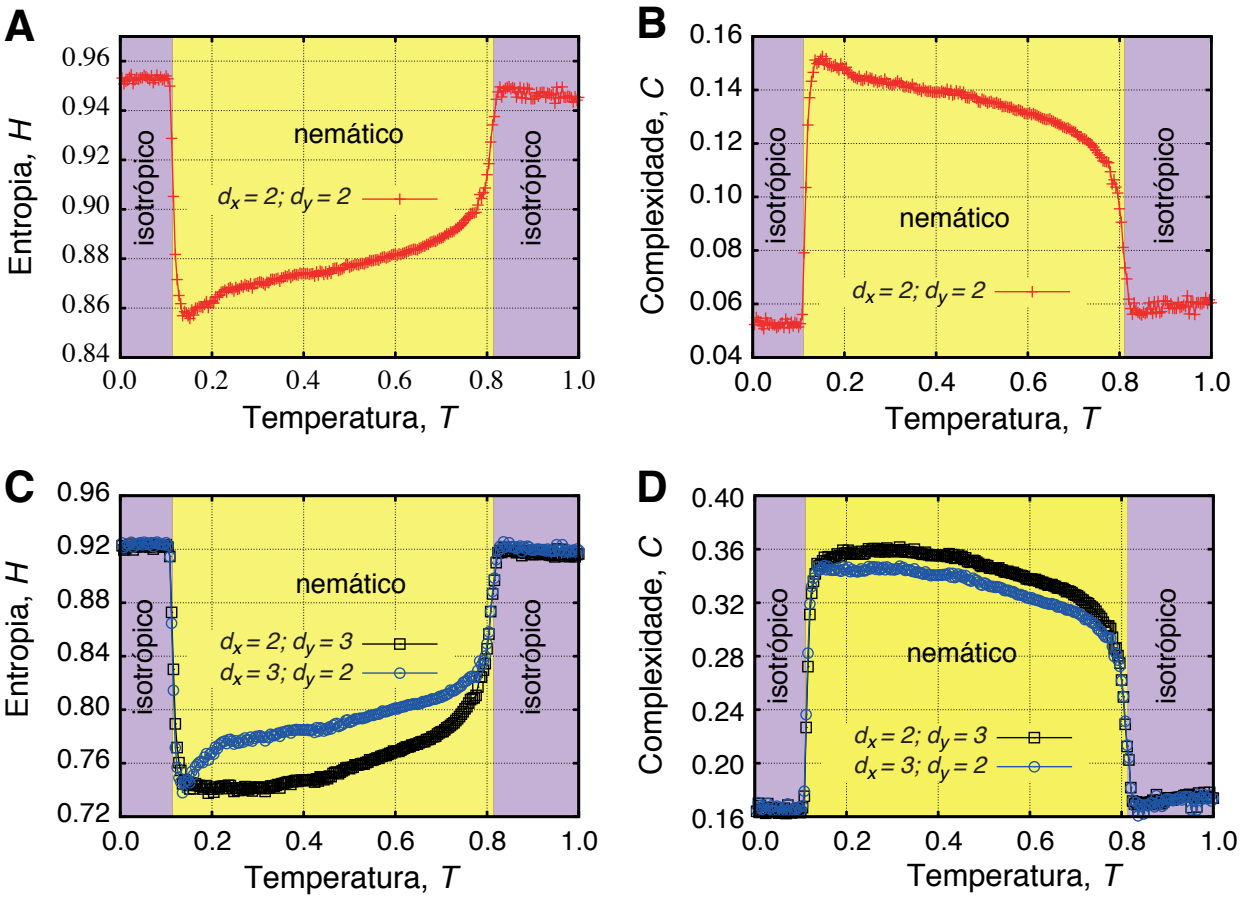


Figura 6.5: Dependência dos índices entrópicos com a temperatura do cristal líquido liotrópico. Em (**A**) mostramos H versus a temperatura T e em (**B**) mostramos C versus a temperatura T para $d_x = d_y = 2$. As figuras (**C**) e (**D**) apresentam os resultados quando consideramos $d_x = 2$ e $d_y = 3$ e também para $d_x = 3$ e $d_y = 2$. As regiões sombreadas indicam os intervalos de temperatura de cada fase.

A transição estudada anteriormente é, de fato, bastante simples. As diferenças entre as texturas são suficientemente grandes a ponto de podermos identificar as transições visualmente. Contudo, esse não é sempre o caso e muitas outras transições são difíceis de se identificar. Nesse contexto, uma questão interessante é se o nosso método pode distinguir outras fases. Para tentar a responder essa questão, calculamos H e C para 12 texturas características de cristais líquidos

sob diferentes condições. Essas texturas também foram obtidas na página de internet do *Liquid Crystal Institute* da *Kent State University* [180]. A Figura 6.6 mostra essas texturas apenas como uma ilustração, visto que não possuímos outras informações além dos nomes dessas texturas. A Figura 6.7 mostra os valores de H e C para cada uma dessas texturas no plano complexidade-entropia. Esses resultados indicam que nosso método classifica as texturas em uma espécie de ordem de complexidade em que cada textura ocupa uma localização distinta nesse plano. Além disso, esses valores diferentes de H e C sugerem que os índices entrópicos poderiam identificar possíveis transições entre as fases mostradas na Figura 6.6.

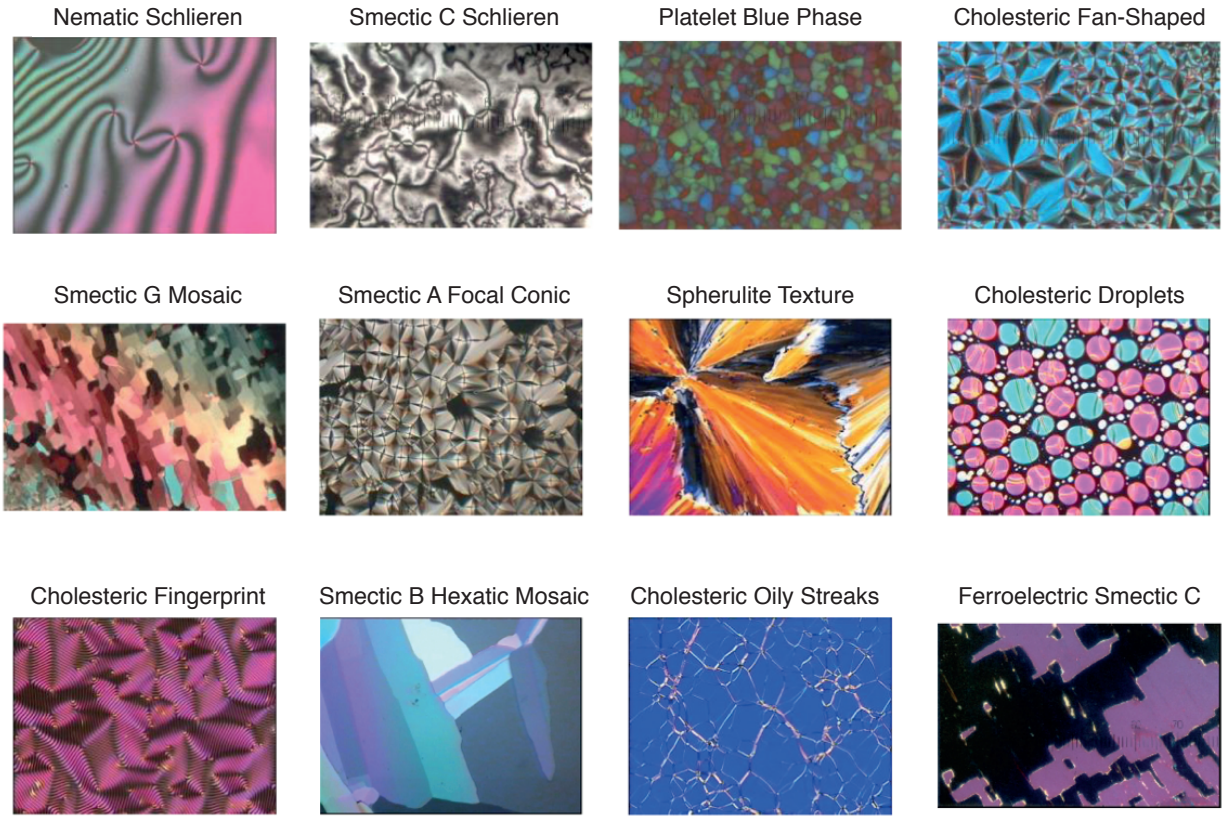


Figura 6.6: Diferentes texturas de cristais liquídos de diferentes tipos e em diferentes fases, obtidas de uma página de internet do *Liquid Crystal Institute* da *Kent State University* [180].

Naturalmente, a localização de cada textura no plano complexidade-entropia deve estar relacionada às propriedades físicas do cristal líquido. Nesse sentido, um melhor entendimento dessa relação talvez mereça uma investigação mais cuidadosa, pois medir propriedades físicas desses materiais pode ser, muitas vezes, uma tarefa complicada. Assim, por exemplo, uma relação clara entre o parâmetro de ordem do cristal líquido e H ou C seria muito útil do ponto vista experimental. Aqui, devido a termos disponíveis somente as imagens das texturas (e também devido à inexperiência do autor nesse tema), torna-se muito complicado sugerir tais relações, de modo que deixaremos para futuros trabalhos essa investigação mais detalhada.

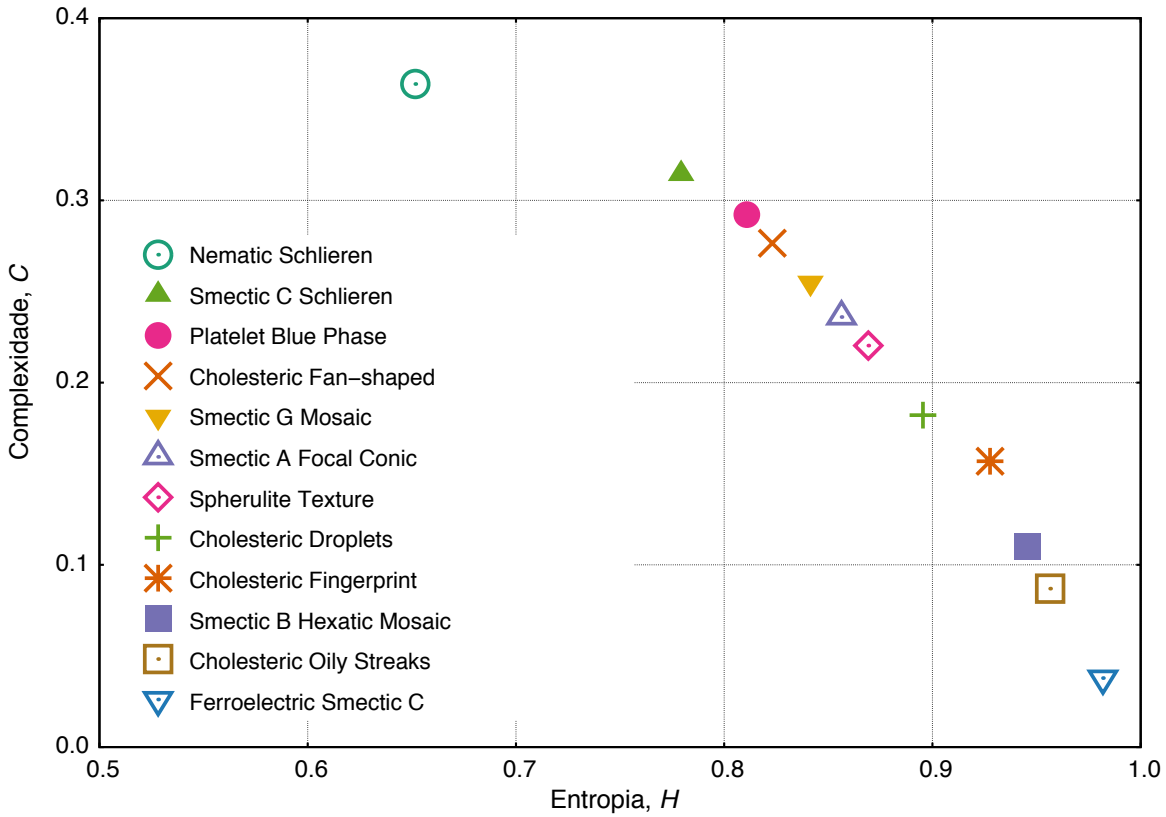


Figura 6.7: Plano complexidade-entropia, C versus H, aplicado para as texturas da Figura 6.6. Para esse cálculo, usamos os valores médios dos pixels nas três camadas (*RGB*) da imagem original e $d_x = 2$ e $d_y = 3$. O tamanho médio das imagens é de aproximadamente 270×200 pixels.

6.5 Aplicação III: superfícies de Ising

Nessa última aplicação de nosso método, estudamos as medidas entrópicas H e S aplicadas a superfícies de Ising [181, 182]. Essas superfícies são obtidas pela acumulação dos valores dos *spins* $\sigma_i \in \{-1, 1\}$ do modelo de Ising definido pela Hamiltoniana

$$\mathcal{H} = -\sum_{\langle i,j \rangle} \sigma_i \sigma_j \,, \tag{6.12}$$

na qual a soma é realizada sobre todos os primeiros vizinhos da rede. Aqui, resolvemos esse modelo numericamente usando uma rede quadrada de tamanho $L \times L$ via método de Monte Carlo com condições periódicas de contorno e condição inicial $\sigma_i = 1 \; \forall i$. O algoritmo de Monte Carlo consiste em varrer todos os sítios da rede e calcular a energia

$$\Delta\mathcal{H} = 2\,\sigma_i(\sigma_{\text{direita}} + \sigma_{\text{esquerda}} + \sigma_{\text{acima}} + \sigma_{\text{abaixo}}) \tag{6.13}$$

relacionada à inversão do valor do *spin* do sítio i ($\sigma_i \to -\sigma_i$). Aqui σ_{direita}, σ_{esquerda}, σ_{acima} e σ_{abaixo} representam os primeiros vizinhos ao sítio i. Caso $\Delta\mathcal{H} < 0$, a inversão é efetivada na rede. Caso $\Delta\mathcal{H} \geq 0$, efetivamos a mudança de valor com probabilidade

$$P = \exp\left(-\frac{\Delta\mathcal{H}}{T}\right) , \tag{6.14}$$

sendo T a temperatura do modelo Ising. Ao percorrer todos os *spins* da rede, completamos um passo de Monte Carlo.

Após isso, definimos a altura da superfície relacionada ao sítio i da rede como

$$\mathcal{S}_i = \sum_t \sigma_i(t) \,, \tag{6.15}$$

em que t representa o número de passos do algoritmo Monte Carlo.

A Figura 6.8 mostra três superfícies obtidas pelo procedimento anterior para diferentes valores da temperatura reduzida T/T_c, em que $T_c = 2/\ln(1+\sqrt{2})$ é a temperatura crítica do modelo de Ising usado aqui. Observamos que há um padrão complexo quando $T/T_c = 1$ e um padrão praticamente aleatório para $T/T_c > 1$ ou $T/T_c < 1$.

Investigamos, primeiramente, a dependência dos índices entrópicos H e C com a temperatura reduzida T/T_C, após um grande número de passos do algoritmo Monte Carlo (10^5) e para $L = 500$. Essa análise é mostrada nas Figuras 6.9A para $d_x = 2$ e $d_y = 3$ e na Figura 6.9B para $d_x = 3$ e $d_y = 2$. Podemos ver nesta figura que ambos os índices apresentam um pico/vale agudo quando o sistema está na temperatura crítica. Ademais, observamos que os valores de H e C são praticamente invariantes perante a rotação $d_x \to d_y$ e $d_y \to d_x$. Na Figura 6.9C, apresentamos a variação conjunta de H e C em função de T/T_C para $d_x = d_y = 3$. Essa representação em três dimensões pode ser útil na investigação de transições de fase mais complicadas, visto que

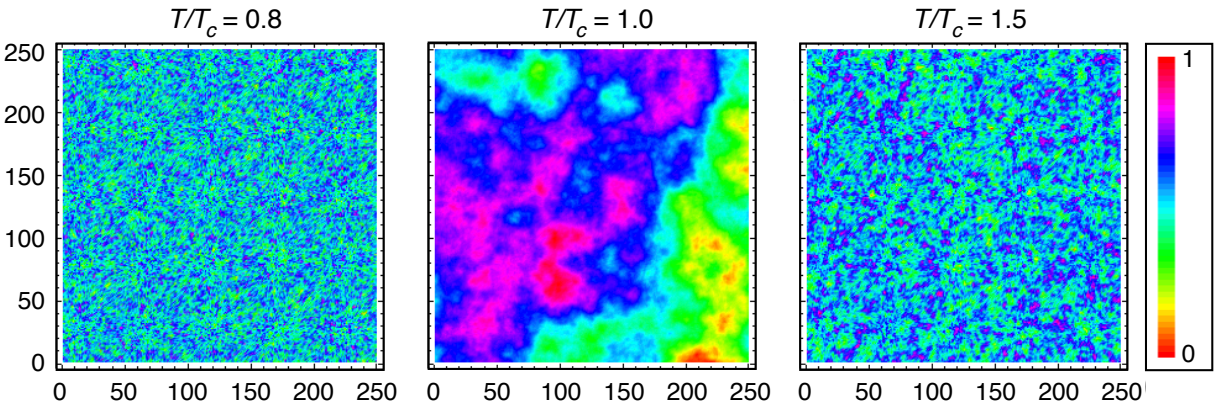


Figura 6.8: Exemplos de superfícies de Ising para três temperaturas diferentes: abaixo de T_c, igual a T_c e acima de T_c. Essas superfícies foram obtidas após somar os valores dos *spins* em 10^5 passos do algoritmo Monte Carlo. Para melhor visualização, as alturas das superfícies foram escaladas para ficarem ente 0 e 1.

nesse gráfico o ponto crítico pode tornar-se mais visível.

Investigamos também a evolução temporal (com o número de passos Monte Carlo) de H e C para diferentes valores da temperatura reduzida T/T_C, como apontamos na Figura 6.10. Os valores iniciais dos *spins* foram escolhidos todos iguais a 1; entretanto, os valores de H e C começam a se diferenciar logo após o primeiro passo Monte Carlo. Para $T \neq T_c$, o valor de H cresce com o número de passos Monte Carlo até atingir um platô em $t \sim 10^2$. Para $T = T_c$, o valor cresce até um máximo localizado em $t \sim 10^2$ e, em seguida, começa a diminuir e aproximar-se de um platô. No caso do índice C, para todas as temperaturas, a complexidade apresenta um máximo antes de começar a aproximar-se de um platô. Podemos observar que ambos os índices entrópicos C e H são bem estáveis após $\sim 10^4$ passos Monte Carlo.

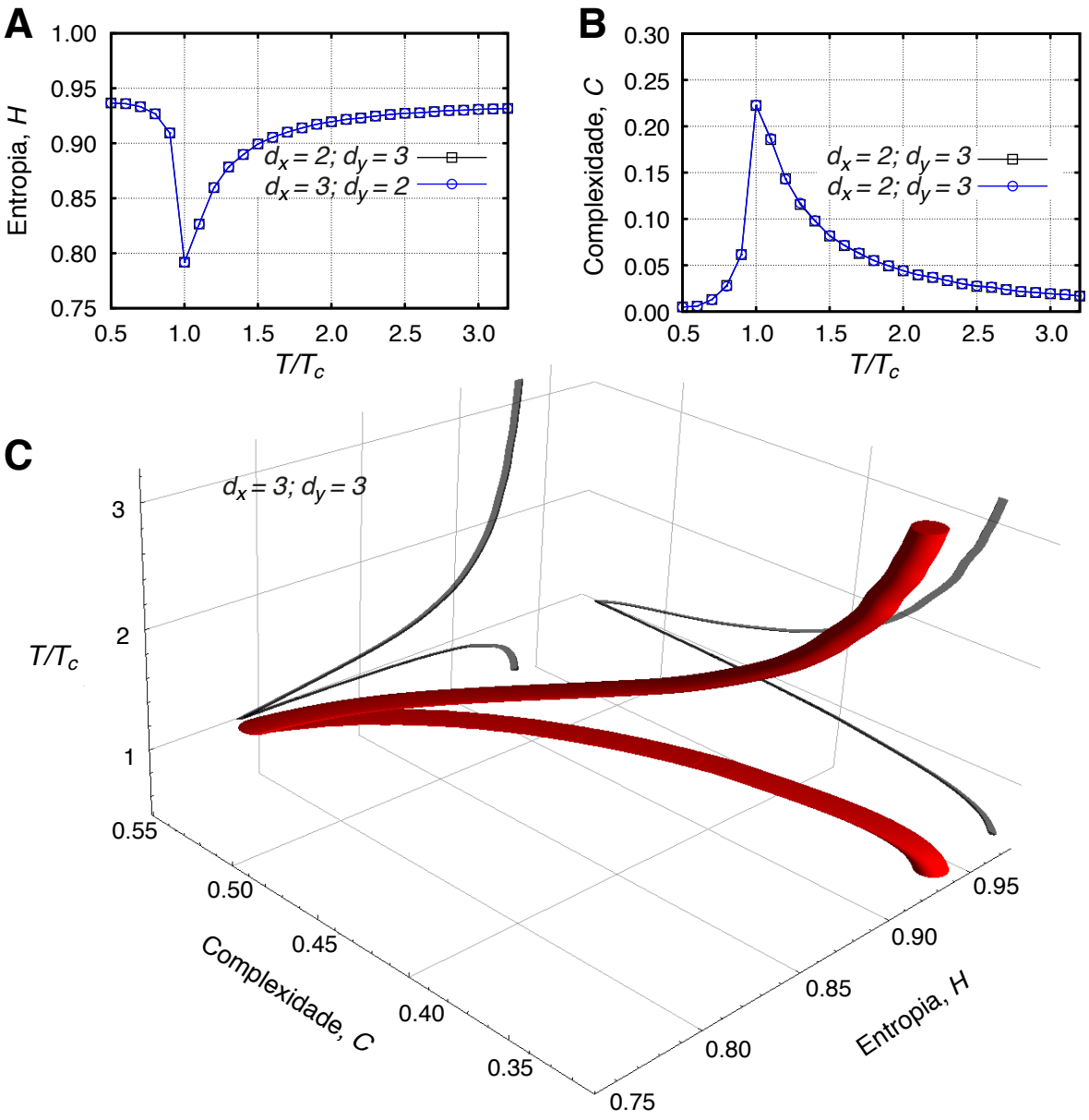


Figura 6.9: Dependência dos índices entrópicos (C e H) com a temperatura reduzida (T/T_c). Em (**A**) mostramos H versus T/T_c e em (**B**) C versus T/T_c quando consideramos $d_x = 2$ e $d_y = 3$ e também $d_x = 3$ e $d_y = 2$. Em (**C**) mostramos a evolução conjunta de C e H em função de T/T_c quando usamos $d_x = d_y = 3$.

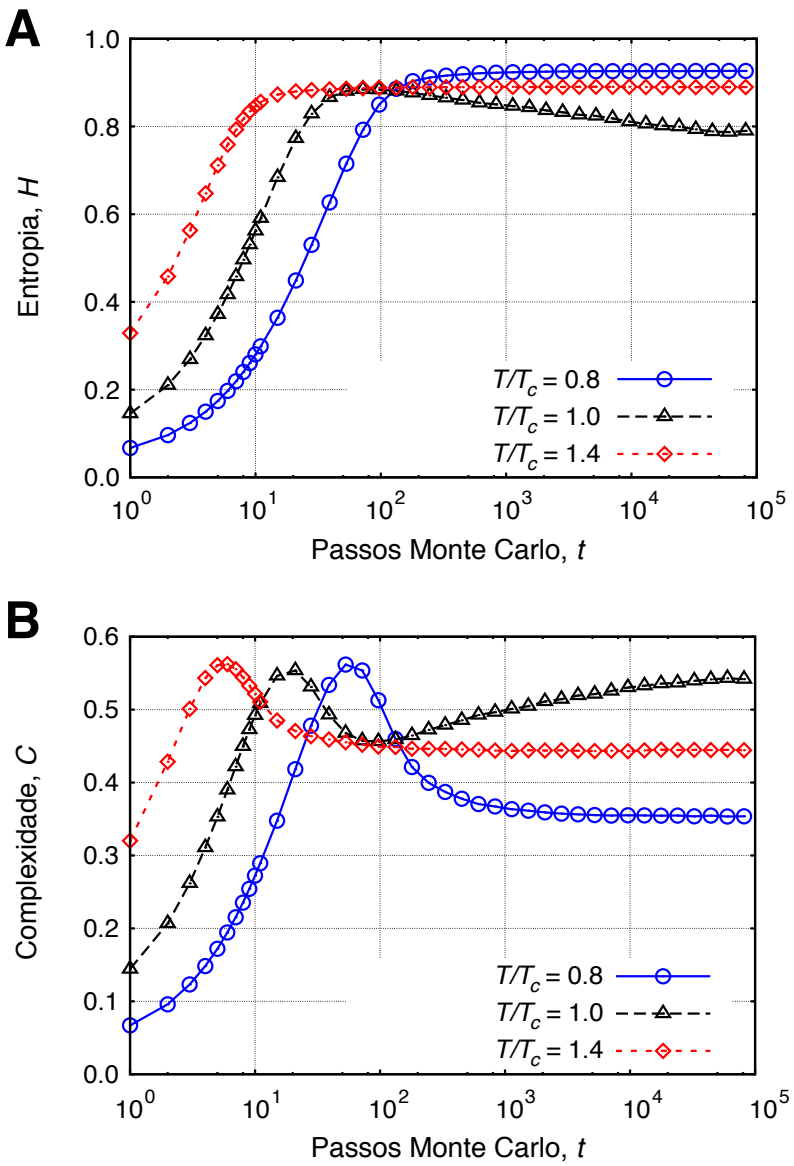


Figura 6.10: Dependência dos índices entrópicos (C e H) com o número de passos do algoritmo Monte Carlo (tempo t) para três valores da temperatura reduzida T/T_c. Em (**A**) mostramos H versus t e em (**B**) C versus t, ambos calculados com $d_x = d_y = 3$.

6.6 Conclusões e perspectivas

Neste capítulo, apresentamos uma proposta de generalização do plano complexidade-entropia para investigar padrões em duas dimensões. Aplicamos o procedimento para estudar superfícies fractais, texturas de cristal líquido e superfícies de Ising. Nessas aplicações, observamos que os índices H e C são capazes de distinguir as diferentes rugosidades das superfícies fractais. Eles também identificaram transições de fases em cristais líquidos liotrópicos e classificaram diferentes texturas características em uma espécie de ordem de complexidade. Finalmente, sobre as superfícies de Ising, o método, além de identificar a temperatura crítica do modelo, mostrou-se bastante estável após $\sim 10^4$ passos Monte Carlo. Considerando esses resultados e adicionando-se o fato de que o procedimento é simples e muito rápido do ponto de vista computacional, estamos muito otimistas com o fato de que nosso método possa ajudar a reduzir as discrepâncias entre medidas de complexidade unidimensional e de dimensões maiores.

Além de testar o método, alguns dos estudos que realizamos mostraram-se interessantes por si sós, como é o caso das diferentes texturas em cristais líquidos. Para tentar entender melhor se é possível relacionar os parâmetros C e H com propriedades físicas dos cristais líquidos, buscamos colaborações com especialistas no tema. Além disso, existem vários padrões nessas texturas que talvez mereçam ser estudados usando o plano complexidade-entropia. No caso das superfícies de Ising, é bastante curiosa a evolução dos parâmetros C e H com o número de passos Monte Carlo. Em particular, os valores máximos de C mostrados na Figura 6.10 parecem ser aproximadamente iguais e parece haver uma dependência com a temperatura reduzida na localização desses máximos. Notemos que para temperaturas mais altas o máximo ocorre para um número menor de passos Monte Carlo. Além disso, quando o sistema está na temperatura crítica, o perfil de C versus t apresenta ainda um valor mínimo. Acreditamos que esses comportamentos curiosos e extensões do modelo Ising (por exemplo, o modelo de Pots) também mereçam uma investigação mais cuidadosa. Entretanto, nosso objetivo aqui foi apenas introduzir a nossa generalização e verificar que o plano complexidade-entropia pode ser útil para medir a complexidade de estruturas com dimensão maior do que um.

Visão geral dos problemas apresentados

Nesta tese, analisamos vários sistemas complexos com características completamente diferentes e que podem ser considerados, em sua maioria, como multidisciplinares. A multidisciplinariedade é, talvez, um dos mais interessantes e estimulantes aspectos desses estudos, visto que o pesquisador é constantemente desafiado a interpretar resultados de diferentes áreas do conhecimento e, consequentemente, acaba por assimilar novos conhecimentos. Por outro lado, toda essa variabilidade nos temas estudados, muitas vezes, não se reflete nas técnicas utilizadas nas investigações. De fato, como o leitor deve ter notado, as técnicas utilizadas e a abordagem que empregamos apresentam muitas características semelhantes. Na maioria do casos, a pergunta central que conduziu nossas análises é quase sempre a mesma: qual o tipo de dado que estamos analisando? A depender dessa reposta, certo conjunto de ferramentas torna-se disponível para investigar e extrair padrões do sistema em questão. A Figura 1 ilustra esse processo para os sistemas que aqui estudamos. A primeira distinção que fizemos diz respeito à quantidade de séries temporais relacionadas ao sistema. Por exemplo, nos capítulos 1 (primeira parte) e 5, as informações sobre o sistema se concentravam em uma única série temporal: as amplitudes (intensidades) sonoras e a intensidade do laser. Desse modo, os padrões foram obtidos diretamente da série temporal ou de séries derivadas da série original (intervalos de retorno, por exemplo). Por outro lado, nos capítulos 3 e 4, o sistema era representado por um conjunto de dados: as $\sim$70 mil partidas de xadrez e os $\sim$2 mil jogos de críquete. Nesse caso, nossa abordagem esteve focada em estudar propriedades difusivas ligadas a essas séries temporais. No capítulo 2, também estudamos várias séries temporais associadas às músicas, porém, nesse caso, não há sentido em aplicar a abordagem difusiva. Por isso, focamos nos padrões individuais das músicas buscando uma classificação de complexidade. No capítulo 6, a abordagem foi um pouco diferente, visto que a questão central desse capítulo não era analisar dados e sim propor e aplicar um procedimento para medir complexidade de estruturas bidimensionais.

Outra pergunta (talvez mais importante que a do parágrafo anterior) que buscamos responder ao longo de nossas análises diz respeito às consequências, implicações e aplicações dos padrões encontrados. Para essa pergunta não temos uma resposta tão direta como a que mostramos na Figura 1. Pelo contrário, uma resposta satisfatória pode muitas vezes consumir muito mais esforço e tempo do que propriamente a análise dos dados. Ainda assim, acreditamos que fomos bem sucedidos nesse aspecto na maioria dos problemas que abordamos. Por exemplo, no capítulo 1,

propusemos a construção de um "sensor social"; no capítulo 2, implementamos uma detecção automática de gêneros musicais e revelamos o "empobrecimento" das músicas populares; no capítulo 3, conectamos os aspectos difusivos da vantagem nos jogos de xadrez com o processo de aprendizagem coletivo e mostramos a importância de perceber os erros dos oponentes; no capítulo 4, relacionamos a memória de longo-alcance com o fenômeno esportivo de "mão quente"; no capítulo 5, nosso modelo simples identificou os principais ingredientes que produzem a dinâmica da intensidade do feixe laser que atravessa a amostra de água fervente; finalmente, no capítulo 6, a aplicação do nosso método para medir a complexidade de imagens foi posta em prática e mostrou-se bastante promissora.

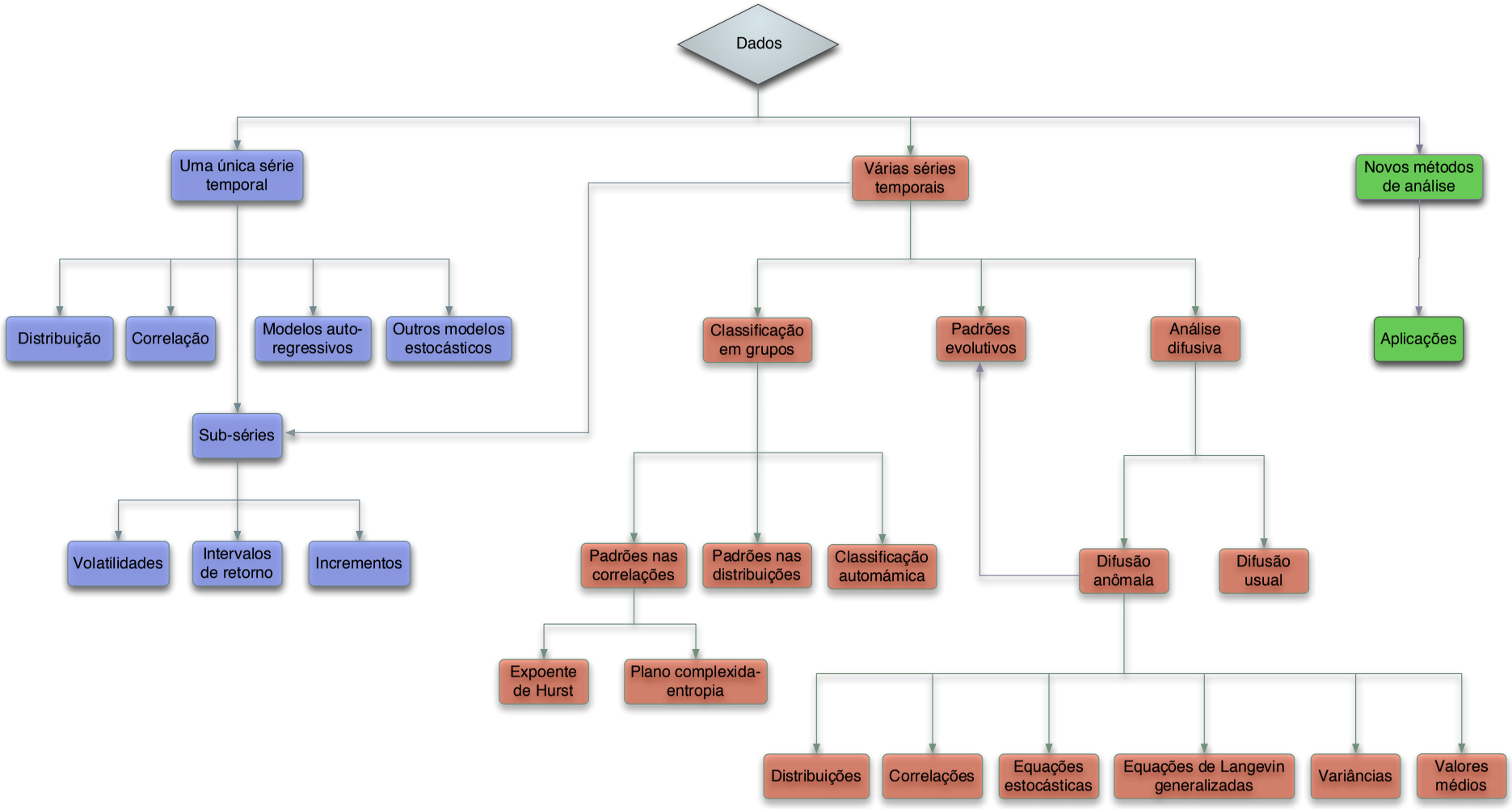


Figura 1: Representação esquemática da maneira como os problemas foram abordados nesta tese.

Apêndice A

Correlações em séries temporais

A.1 Função de autocorrelação

Medidas de correlação são bastante comuns em estatística e uma das mais populares é o coeficiente de correlação de Pearson [183]. Este diz o quão linear é a relação entre dois conjuntos de variáveis x_i e y_i ($i \in \{1, 2, \ldots, n\}$), sendo definido como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\,, \tag{A.1}$$

em que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Esse coeficiente varia de -1 a 1, e $r = 1$ implica em uma relação linear crescente perfeita entre os conjuntos x_i e y_i e $r = -1$ implica em uma relação linear decrescente perfeita. Para $r = 0$ não há uma relação linear entre esses dois conjuntos.

A ideia de Pearson conduz naturalmente à função de autocorrelação[1]. Basta imaginarmos que, ao invés de investigar a correlação entre dois conjuntos, desejamos saber como os elementos de uma série temporal estão relacionados com seus próprios elementos defasados por uma ou mais unidades de tempo. Suponhamos que z_i ($i \in \{1, 2, \ldots, n\}$) representa uma série temporal, para a qual desejamos saber como é a correlação entre os elementos imediatamente espaçados. Para isso, podemos usar o coeficiente de Pearson aplicado aos conjuntos $x_i = \{z_1, z_2, \ldots, z_{n-1}\}$ e $y_i = \{z_2, z_3, \ldots, z_n\}$, ou seja,

$$r_1 = \frac{\sum_{i=1}^{n-1}(z_i - \bar{x})(z_{i+1} - \bar{y})}{\sqrt{\sum_{i=1}^{n-1}(z_i - \bar{x})^2 \sum_{i=1}^{n-1}(z_{i+1} - \bar{y})^2}}\,. \tag{A.2}$$

Se a série temporal z_i é estacionária[2] e n é razoavelmente grande, podemos considerar que

[1] Muitas vezes referida apenas por função de correlação. Trata-se de um abuso de linguagem muito comum na literatura em geral e este texto não é uma exceção.

[2] Um processo estocástico $x(t)$ é dito ser estacionário quando a sua distribuição de probalidade é invariante por translação temporal. Essa definição é muitas vezes considerada bastante restritiva, dando origem a outras definições mais amplas. Vejamos, por exemplo, a referência [52].

$\bar{x} = \bar{y} = \bar{z}$ e $\sum_{i=1}^{n-1}(z_i - \bar{x})^2 = \sum_{i=1}^{n-1}(z_{i+1} - \bar{y})^2$, conduzindo a

$$r_1 = \frac{\sum_{i=1}^{n-1}(z_i - \bar{z})(z_{i+1} - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}\,, \tag{A.3}$$

sendo $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Naturalmente, essa ideia pode ser facilmente generalizada para elementos espaçados por k unidades temporais. Matematicamente, temos:

$$r_k = \frac{\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}\,. \tag{A.4}$$

Essa última quantidade é conhecida como função de autocorrelação defasada em k unidades de tempo (*lag-k auto-correlation*).

É interessante observar que o numerador da expressão A.4 é uma generalização da variância. De fato, a expressão

$$c_k = \frac{1}{n-k}\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z}) \tag{A.5}$$

é conhecida como covariância, sendo que

$$c_0 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \tag{A.6}$$

é precisamente[3] a variância da série temporal z_t. Assim, podemos reescrever a função de auto-correlação (A.4) da seguinte maneira:

$$r_k = \frac{c_k}{c_0}\,. \tag{A.7}$$

Obviamente, esses resultados podem ser facilmente estendidos para o caso contínuo. Com relação a isso, é interessante notar que alguns autores usam diferentes denominações para o caso discreto e contínuo, sendo a função de autocorrelação no caso discreto também chamada de coeficiente de autocorrelação. Para o caso contínuo, a série temporal z_i deve ser substituída pela variável estocástica $x(t)$ com $t \in \mathbb{R}$. Assim, a covariância fica determinada por

$$C(\tau) = \langle x(t)x(t+\tau)\rangle\,, \tag{A.8}$$

em que $\langle \dots \rangle$ representa o valor médio sobre um conjunto de realizações do processo estocástico

[3] Alguns autores usam uma definição ligeiramente diferente, sendo $\frac{1}{n-k}$ substituído por $\frac{1}{n-k-1}$ na equação (A.5), ou seja,

$$c_k = \frac{1}{n-k-1}\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z})\,.$$

Acredita-se que essa definição tenha um erro estatístico menor para amostras pequenas. Contudo, as duas definições são praticamente indistinguíveis para séries de tamanho razoável [183].

que gera $x(t)$. Uma vez definida a covariância, podemos definir também a autocorrelação

$$R(\tau) = \frac{C(\tau)}{C(0)}. \tag{A.9}$$

A definição na forma contínua é bastante conveniente para descrever o tipo de memória existente na variável estocástica $x(t)$. Uma questão fundamental diz respeito à existência ou não de uma escala típica para a função de autocorrelação. Para investigar essa questão, consideremos a integral [52]

$$\int_0^\infty R(\tau)d\tau \tag{A.10}$$

e os três cenários descritos a seguir.

- Se a integral $\int_0^\infty R(\tau)d\tau$ é finita, então existe uma escala de tempo de memória típica, a qual é chamada de tempo de correlação do processo $x(t)$. Por exemplo:
  - $R(\tau) \sim \exp(-\frac{\tau}{\tau_c})$ implica $\int_0^\infty R(\tau)d\tau \sim \tau_c$;
  - $R(\tau) \sim \exp(-\frac{\tau^\nu}{\tau_c})$ implica $\int_0^\infty R(\tau)d\tau \sim \tau_c^{1/\nu}$.
- Se a integral $\int_0^\infty R(\tau)d\tau$ é indeterminada, então é possível que a função de correlação seja oscilante. Por exemplo:
  - $R(\tau) \sim \sin(a\tau)$ conduz à uma integral oscilante que não pode ser determinada. Nestes casos, dizemos que a correlação é harmônica.
- Se a integral $\int_0^\infty R(\tau)d\tau$ não é finita, então não existe uma escala de tempo de memória típica. Por exemplo:
  - $R(\tau) \sim \tau^{\eta-1}$ com $0 < \eta \leq 1$ implica $\int_0^\infty \tau^{\eta-1}d\tau \to \infty$.

Variáveis aleatórias caracterizadas por uma função de autocorrelação do tipo

$$R(\tau) \sim \tau^{\eta-1} \qquad \text{ou} \qquad R(\tau) \sim \tau^{-\gamma}, \tag{A.11}$$

com $\gamma = 1 - \eta$, são ditas serem correlacionadas de longo alcance (*long-range correlated*). Por outro lado, variáveis aleatórias caracterizadas por funções de autocorrelação do tipo $R(\tau) \sim \exp(-\frac{\tau}{\tau_c})$ são ditas serem correlacionadas de curto alcance (*short-range correlated*).

A.2 Invariância de escala e movimento browniano fracionário

Outra propriedade diretamente relacionada aos aspectos de correlação descritos na seção anterior é a invariância de escala. De uma maneira matemática, dizemos que uma função

$\Phi(x_1, x_2, \dots)$ é invariante por escala (ou também autossimilar) sempre que

$$\Phi(x_1, x_2, \dots) = \gamma^a \Phi(\gamma^b x_1, \gamma^c x_2, \dots)\,. \tag{A.12}$$

No caso de uma série temporal estacionária, z_i $(i \in \{1, 2, \dots, n\})$, um modo de investigar a invariância de escala é definir a série das flutuações

$$\Delta z_i(k) = x_{i+k} - x_i\,. \tag{A.13}$$

Essa nova série é dita ser invariante por escala, autossimilar ou fractal se a distribuição de probabilidade de $\Delta z_i(k)$ possuir a mesma forma funcional para diferentes valores de k, isto é,

$$p(\Delta z_i, k) = \frac{1}{k^\delta}\,\Psi\left(\frac{\Delta z_i}{k^\delta}\right)\,, \tag{A.14}$$

sendo δ o expoente de escala e $\Psi(x)$ a função escala. Por exemplo, se z_i representar uma série temporal oriunda de um movimento browniano usual, $\Psi(x)$ será uma gaussiana e $\delta = 1/2$. Outra propriedade interessante das flutuações de uma série fractal é a dependência em k do segundo momento

$$\langle \Delta z_i(k)^2 \rangle = \frac{1}{n-k}\sum_{i=1}^{n-k} \Delta z_i(k)^2 \sim k^{2\delta}\,, \tag{A.15}$$

em que δ é novamente o expoente de escala.

Podemos fazer a conexão entre invariância de escala e correlações de longo alcance por meio do movimento browniano fracionário. Assim como o movimento browniano usual pode ser visto como a integração de um ruído branco, o movimento browniano fracionário é usualmente definido por meio da seguinte integração estocástica [184]:

$$\begin{aligned} B_h(t) &= \frac{1}{\Gamma(h+1/2)}\left(\int_{-\infty}^{0}[(t-s)^{h-1/2} - (-s)^{h-1/2}]dB(s)\right. \\ &+ \left.\int_0^t (t-s)^{h-1/2} dB(s)\right)\,, \end{aligned} \tag{A.16}$$

na qual $\Gamma(y) = \int_0^\infty x^{\alpha-1}\exp(-x)dx$, $B(s)$ representa um movimento browniano usual e $0 < h < 1$ é um parâmetro chamado de expoente de Hurst. Notamos que para $h = 1/2$ recuperamos o movimento browniano usual,

$$B_{1/2}(t) = \int_0^t dB(s)\,. \tag{A.17}$$

Esse processo possui algumas propriedades, as quais nos limitaremos apenas a enunciá-las[4]. Trata-se de um processo estocástico estacionário e de valor médio nulo. Além disso, sua variância é dada por $\langle B_h(t)^2 \rangle \sim t^{2h}$ e a covariância é $\langle B_h(t) B_h(\tau) \rangle \sim t^{2h} + \tau^{2h} - |t-\tau|^{2h}$. Para investigarmos a invariância de escala no processo browniano fracionário, devemos considerar sua série das

[4]Uma descrição detalhada pode ser encontrada nas referências [184, 185].

flutuações ou dos incrementos,

$$\begin{aligned}\Delta B_h(\tau) &= B_h(t+\tau) - B_h(t) = \frac{1}{\Gamma(h+1/2)} \\ &\times \left[\int_{-\infty}^{t+\tau} (t+\tau-s)^{h-1/2} dB(s) - \int_{-\infty}^{t} (t-s)^{h-1/2} dB(s)\right]. \end{aligned} \tag{A.18}$$

Usando essa definição, é possível mostrar que $\Delta B_h(\tau)$ é invariante por escala, caracterizado por uma função escala $\Psi(x)$ gaussiana e

$$\langle \Delta B_h(\tau)^2 \rangle \sim \tau^{2h}\,. \tag{A.19}$$

Além disso, sua covariância é dada por

$$\langle \Delta B_h(\tau) \Delta B_h(\tau+k) \rangle \sim h(2h-1)k^{2h-2}\,. \tag{A.20}$$

Por essas duas expressões, podemos observar que, no caso dos incrementos do movimento browniano fracionário (também chamado de ruído gaussiano fracionário), o expoente de escala δ e o expoente da função de correlação de longo-alcance γ estão diretamente relacionados com o expoente h de Hurst. De fato, por comparação direta das equações (A.18) e (A.20) com as equações (A.11) e (A.15), obtemos $\delta = h$ e $\gamma = 2(1-h)$. Essa relação entre invariância de escala e correlações de longo alcance é a base para o método DFA que apresentaremos na próxima seção.

Sabemos, ainda, que esses resultados valem em um cenário mais geral. Por exemplo, uma característica necessária (mas não suficiente) para a validade da igualdade $\delta = h$ é que função de escala $\Psi(x)$ tenha o segundo momento finito.

A.3 Análise de flutuações DFA

Um dos métodos mais utilizados para detectar correlações em séries temporais é o chamado DFA, sigla inglesa para *detrended fluctuation analysis*. Ele foi proposto por Peng *et al.* [186] (vejamos também as referências [187, 188, 189, 190]) para o estudo de correlações em séries temporais construídas a partir do DNA. Como indica o nome, o método DFA analisa as flutuações de séries temporais removendo uma possível tendência local. Trata-se de um método de fácil implementação e que produz excelentes resultados mesmo para séries temporais moderadamente pequenas (empiricamente, da ordem de mil termos).

Para apresentarmos o método, consideremos novamente uma série temporal x_i ($i \in \{1, 2, \ldots, l\}$). O primeiro passo do DFA consiste em obter a série integral de x_i, *i.e.*,

$$y_i = \sum_{j=1}^{i} x_j\,. \tag{A.21}$$

Logo após, dividimos a série integrada em $s = l/n$ partições não superpostas, de maneira que cada partição tenha n elementos, ou seja,

$$\underbrace{\{y_1, y_2, \ldots, y_n\}}_{w_j^{(1,n)}}, \underbrace{\{y_{n+1}, y_{n+2}, \ldots, y_{2n}\}}_{w_j^{(2,n)}}, \ldots, \underbrace{\{y_{(s-1)n+1}, y_{(s-1)n+2}, \ldots, y_{sn}\}}_{w_j^{(s,n)}}, \tag{A.22}$$

em que $w_j^{(i,n)}$ $(j \in \{1, 2, \ldots, n\})$ representa o conjunto dos elementos contidos na i-ésima partição. Para cada conjunto $w_j^{(i,n)}$ ajustamos um polinômio de grau v e obtemos a flutuação em torno desse ajuste

$$\chi_2^{(i,n)} = \frac{1}{n-1} \sum_{j=1}^{n} [w_j^{(i,n)} - f_v(j)]^2 \,, \tag{A.23}$$

em que $f_v(j)$ representa o polinômio ajustado ao conjunto $w_j^{(i,n)}$. Em seguida, calculamos o valor médio dessa flutuação sobre todas as s partições:

$$F(n) = \left(\frac{1}{s} \sum_{i=1}^{s} \chi_2^{(i,n)} \right)^{1/2} . \tag{A.24}$$

Naturalmente, essa flutuação média será uma função de n e está diretamente relacionada com o expoente de escala δ da seguinte maneira:

$$F(n) \sim n^{\delta} \,. \tag{A.25}$$

Para o caso em que a função escala $\Psi(x)$ possui o segundo momento finito, teremos a igualdade $\delta = h$. Nos casos em que $\Psi(x)$ não possui o segundo momento, o DFA pode conduzir a falsas correlações. Na prática, costuma-se aplicar o método na série embaralhada de maneira aleatória para verificar a validade da igualdade $\delta = h$. Se $h \neq 0{,}5$ for obtido para a série embaralhada, nada podemos dizer sobre o expoente h da série não embaralhada. Por outro lado, se $h \approx 0{,}5$ para a série embaralhada, o expoente h da série original será verdadeiramente o expoente de Hurst para aquela série. Nos resultados apresentados neste texto, todas as análises DFA conduzem à $h \approx 0{,}5$ para as versões embaralhadas da série temporal em questão. Além disso, utilizou-se um polinómio de grau 1 em todas as análises DFA apresentadas neste trabalho.

A derivação do resultado (A.25) pode ser encontrada na referência [191] para o caso em que remove-se uma tendência linear ($v = 1$). Para evitar as longas (embora simples) manipulações algébricas relacionadas à remoção das tendências, apresentaremos uma dedução considerando o caso em que a série x_i não apresenta tendências. Sob essa hipótese, calcular a flutuação $\chi_2^{(i,l)}$ é

o mesmo que calcular $\langle y_i^2 \rangle$, ou seja,

$$\begin{aligned}
\langle y_i^2 \rangle &= \left\langle \left(\sum_{j=1}^{i} x_j \right)^2 \right\rangle = \left\langle \sum_{j=1}^{i} x_j^2 \right\rangle + \left\langle \sum_{j=1}^{i} \sum_{k \neq j}^{i} x_j x_k \right\rangle \\
&= \sum_{j=1}^{i} \langle x_j^2 \rangle + \sum_{j=1}^{i} \sum_{k \neq j}^{i} \langle x_j x_k \rangle \\
&= i\langle x_j^2 \rangle + \sum_{j=1}^{i} \sum_{k \neq j}^{i} C(|k-j|) \\
&= i\langle x_j^2 \rangle + 2 \sum_{j=1}^{i-1} (i-j) C(j) \,, \qquad \text{(A.26)}
\end{aligned}$$

em que $C(|k-j|) = \langle x_j x_k \rangle$. Agora, supondo $i \gg 1$, podemos aproximar a soma do segundo termo de (A.26) pela integral

$$2 \sum_{j=1}^{i-1} (i-j) C(j) \sim \int_1^i (i-j) C(j) dj \sim \int_1^i (i-j) j^{-\gamma} dj \,, \qquad \text{(A.27)}$$

considerando que a função correlação seja uma lei de potência, $C(j) \sim j^{-\gamma}$, com $0 < \gamma < 1$. Assim, obtemos1

$$\langle y_i^2 \rangle \sim i^{-\gamma+2} + i[\langle x_j^2 \rangle + (\gamma - 1)^{-1}] - i^{-\gamma} \,, \qquad \text{(A.28)}$$

na qual o termo dominante é

$$\langle y_i^2 \rangle \sim i^{-\gamma+2} \,, \qquad \text{(A.29)}$$

visto que estamos considerando i muito grande. Notemos que esse deslocamento quadrático médio cresce mais rápido que uma função linear, correspondendo a um processo superdifusivo. Finalmente, substituindo esse resultado em (A.24) e imaginando que i faça o papel de l, encontramos

$$F(l) \sim l^{1-\gamma/2} \,, \qquad F(l) \sim l^h \qquad \text{ou} \qquad F(l) \sim l^\delta \,, \qquad \text{(A.30)}$$

em que usamos $h = 1 - \gamma/2$ e $\delta = h$. Se a série for correlacionada de curto alcance, $C(j)$ decai exponencialmente e o primeiro termo em (A.26) será dominante, conduzindo a $\langle y_i^2 \rangle \sim i$ (difusão usual) e $F(l) \sim l^{1/2}$, de modo que $h = 0{,}5$ indica ausência de correlações de longo alcance, como já havíamos discutido. Observemos que a validade desses resultados está condicionada à existência do segundo momento $\langle x_j^2 \rangle$.

A aplicação da equação (A.30) é bastante simples: basta calcular essa função de flutuação para um conjunto de valores de l e construir um gráfico log-log ou um gráfico linear de $\log(F)$ versus $\log(n)$. A inclinação dessa reta será numericamente igual ao expoente δ. Observemos que uma das engenhosidades desse método está no fato dele construir diversas trajetórias, aproximadamente independentes, a partir de uma única série temporal. Isto faz com que, mesmo para séries moderadamente pequenas, os erros envolvendo a determinação do expoente h sejam

pequenos.

Apêndice B

Teste de hipótese de Kolmogorov-Smirnov e o método bootstrapping

B.1 Teste de Kolmogorov-Smirnov

Suponhamos que tenhamos uma amostra de números aleatórios $\mathcal{A} = \{X_1, X_2, \ldots, X_n\}$ cuja distribuição de probabilidade $P(x)$ não é conhecida, mas pode ser estimada dos dados como $P_n(x)$. O teste Kolmogorov-Smirnov [192, 193] verifica estatisticamente a hipótese de que $P(x)$ seja uma dada distribuição $P_\star(x)$. O problema posto, em uma linguagem mais estatística, consiste em decidir entre duas hipóteses:

$$\begin{aligned} H_0 : P &= P_\star \\ H_1 : P &\neq P_\star \,, \end{aligned} \tag{B.1}$$

sendo H_0 denominada hipótese nula e H_1 é a hipótese alternativa. Para apresentarmos o método devemos, primeiro, definir a distribuição acumulada empírica dos dados

$$F_n(x) = P_n(X \leq x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \,, \tag{B.2}$$

em que a função I retorna 1 se $X_i \leq x$ e 0 caso contrário, desse modo, o somatório representa o número de elementos em $\mathcal{A}$ menores do que x. Além disso, precisamos também da distribuição acumulada de $P_\star(x)$,

$$F_\star(x) = \int_{-\infty}^{x} P_\star(y) dy \,. \tag{B.3}$$

O teste de Kolmogorov-Smirnov consiste em calcular a distância estatística

$$D_n = \sqrt{n}\, \sup_{x \,\in\, \mathbb{R}} |F_n(x) - F_\star(x)| \tag{B.4}$$

na qual $\sup_{x \in \mathbb{R}}$ é o supremo do conjunto das distâncias. A decisão do teste baseia-se na seguinte regra:

$$\begin{aligned} &\text{Não podemos rejeitar } H_0 \quad \text{se} \quad D_n \leq c_\alpha \,. \\ &\text{Não podemos rejeitar } H_1 \quad \text{se} \quad D_n > c_\alpha \,. \end{aligned}$$

O valor limiar c_α depende do nível de significância α e pode ser encontrado em livros de tabelas estatística [194]. Essas tabelas são determinadas, em geral, ao resolver-se a equação

$$\alpha = 1 - K(c_\alpha) \,, \tag{B.5}$$

sendo K a distribuição acumulada de Kolmogorov-Smirnov, definida por [195]

$$K(c_\alpha) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2i^2 c_\alpha) \,. \tag{B.6}$$

Por exemplo, para termos 95% ($\alpha = 0{,}05$) de probabilidade do teste de hipótese estar correto, teremos

$$0{,}05 = 1 - K(c_{0{,}05}) \implies c_{0{,}05} \approx 1{,}36 \,. \tag{B.7}$$

Além disso, é muito comum o cálculo do chamado p_{valor} (valor p ou p-value), que é a probabilidade de se obter uma distância estatística menor do que a D_n que foi observada, supondo que a hipótese H_0 seja válida. Esse valor pode ser calculado usando a distribuição das distâncias D_n que é dada pela distribuição de Kolmogorov-Smirnov (equação B.6) no limite de $n \to \infty$. Assim, o valor p pode ser aproximado por $p_{\text{valor}} \approx K(D_n)$. O teste de Kolmogorov-Smirnov é considerado como estatisticamente significante quando $p_{\text{valor}} > \alpha$.

O teste de Kolmogorov-Smirnov também permite comparar duas distribuições empíricas e dizer se elas são idênticas. Esse é o chamado teste de duas amostras de Kolmogorov-Smirnov. Nesse caso, suponhamos que tenhamos o conjunto de dados $\mathcal{A} = \{X_1, X_2, \ldots, X_n\}$ e o conjunto de dados $\mathcal{B} = \{Y_1, Y_2, \ldots, Y_m\}$. A partir desses dados, podemos calcular as distribuições acumuladas

$$F_n(x) = P_n(X \leq x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{B.8}$$

e

$$G_m(x) = P_m(Y \leq x) = \frac{1}{m} \sum_{i=1}^{m} I(Y_i \leq x) \,. \tag{B.9}$$

Desejamos, agora, testar as hipóteses

$$\begin{aligned} H_0 : F &= G \\ H_1 : F &\neq G \,. \end{aligned} \tag{B.10}$$

Para isso, basta calcular a distância estatística

$$D_{m\,n} = \sqrt{\frac{mn}{m+n}}\ \sup_{x\,\in\,\mathbb{R}} |F_n(x) - G_m(x)| \tag{B.11}$$

e usar as mesmas regras do teste para uma amostra.

B.2 Método bootstrapping

Suponhamos que tenhamos uma amostra aleatória $\mathcal{A} = \{X_1, X_2, \ldots, X_n\}$ e que desejamos estimar alguma medida estatística $\hat{\theta}$ dessa amostra e também os intervalos de confiança dessa estimativa (por exemplo, $\hat{\theta}$ pode ser o valor médio ou a variância de $\mathcal{A}$). Um modo de se obter esses intervalos de confiança é usando o procedimento conhecido como *bootstrapping* [196]. O método consiste em obter o valor de $\hat{\theta}$ para os subconjuntos

$$\mathcal{B}_i = \{Y_1, Y_2, \ldots, Y_n\} \tag{B.12}$$

em que Y_j é um elemento de $\mathcal{A}$ escolhido aleatoriamente. Assim, teremos um novo conjunto

$$\Theta = \{\hat{\theta}[B_1], \hat{\theta}[B_2], \ldots, \hat{\theta}[B_N]\}\,, \tag{B.13}$$

a partir do qual podemos determinar os intervalos de confiança ($(\hat{\theta}[\mathcal{A}])_{\text{inferior}}$ e $(\hat{\theta}[\mathcal{A}])_{\text{superior}}$) de $\hat{\theta}[\mathcal{A}]$ com nível de significância α, definidos como

$$(\hat{\theta}[\mathcal{A}])_{\text{inferior}} = \hat{Q}_{\alpha/2}[\Theta] \tag{B.14}$$

$$(\hat{\theta}[\mathcal{A}])_{\text{superior}} = \hat{Q}_{1-\alpha/2}[\Theta]\,, \tag{B.15}$$

em que $\hat{Q}_\beta$ é o β-quantil da distribuição de probabilidade dos elementos de Θ. $\hat{Q}_\beta$ retorna o valor de x tal que a probabilidade de encontrar um evento menor do que x seja β; por exemplo, $\hat{Q}_{1/2}[\mathcal{A}]$ retorna a mediana do conjunto $\mathcal{A}$.

Apêndice C

Cálculo da variância da expressão 3.6

Para obter a variância 3.6, consideramos a equação 3.5

$$A_i(m+0{,}5) = A_i(m) + \Phi_i + \eta(m)$$

e assumimos que a variável discreta m torna-se uma variável contínua t. Nessa aproximação, podemos rescrever 3.5 como

$$\frac{dA(t)}{dt} = \Phi + \eta(t)\,. \tag{C.1}$$

Desse modo, temos uma equação de Langevin, porém na presença de dois ruídos de naturezas diferentes. O termo $\eta(t)$ está associado à evolução temporal de $A(t)$ e irá produzir um ensemble de trajetórias para cada possível valor de Φ. Por outro lado, os diferentes valores de Φ produzem outro ensamble. Denotaremos por $\langle \ldots \rangle_\eta$ as médias associadas ao primeiro ensamble e $\langle \ldots \rangle_\Phi$ as médias associadas ao segundo ensamble. Sabemos, ainda, que o ruído $\eta(t)$ possui média nula, $\langle \eta(t) \rangle_\eta = 0$, e não é correlacionado, $\langle \eta(t')\eta(t'') \rangle_\eta = b\delta(t'-t'')$. Aqui, b é uma constante e δ é a função delta de Dirac.

O primeiro passo para obter a variância 3.6 é integrar a equação C.1 com relação a t, obtendo

$$A(t) = A_0 + \Phi\, t + \int_0^t \eta(t')dt'\,, \tag{C.2}$$

sendo A_0 a condição inicial para a equação C.1. Calculamos, em seguida, os valores médios

$$\begin{aligned} \langle A(t) \rangle_\eta &= A_0 + \Phi\, t + \int_0^t \langle \eta(t') \rangle_\eta \, dt' \\ &= A_0 + \Phi\, t \end{aligned} \tag{C.3}$$

e

$$\begin{aligned}\langle A(t)^2\rangle_\eta &= A_0^2 + \Phi^2\, t^2 + 2A_0\,\Phi\, t + 2A_0 \int_0^t \langle \eta(t')\rangle_\eta\, dt' \\ &+ 2\,\Phi\, t \int_0^t \langle \eta(t')\rangle_\eta dt' + \int_0^t \int_0^{t'} \langle \eta(t')\eta(t'')\rangle_\eta\, dt' dt'' \\ &= A_0^2 + 2A_0\,\Phi\, t + bt + \Phi^2 t^2\,. \end{aligned} \tag{C.4}$$

Resta-nos calcular os valores médios no ensemble de Φ, *i.e.*,

$$\langle\langle A(t)\rangle_\eta\rangle_\Phi = A_0 + \langle\Phi\rangle_\Phi\, t \tag{C.5}$$

e

$$\langle\langle A(t)^2\rangle_\eta\rangle_\Phi = A_0^2 + 2A_0\,\langle\Phi\rangle_\Phi\, t + bt + \langle\Phi^2\rangle_\Phi t^2\,. \tag{C.6}$$

Finalmente, podemos escrever a variância da vantagem como

$$\begin{aligned}\sigma^2 &= \langle\langle A(t)^2\rangle_\eta\rangle_\Phi - (\langle\langle A(t)\rangle_\eta\rangle_\Phi)^2 \\ &= bt + t^2(\langle\Phi^2\rangle_\Phi - \langle\Phi\rangle_\Phi^2) \\ &= bt + \sigma_\Phi^2\, t^2\,. \end{aligned} \tag{C.7}$$

Assim, para escolhas convenientes de b ou t suficientemente grande, é possível obter $\sigma^2 \sim t^2$.

Apêndice D

Possíveis conexões com a Mecânica Estatística Não Extensiva

Este apêndice foi adicionado após a defesa da tese e visa a responder algumas indagações dos membros da banca relacionadas a possíveis conexões entre alguns resultados desse trabalho e a chamada Mecânica Estatística Não Extensiva. Iniciaremos com uma pequena introdução ao tema e, posteriormente, pontuaremos sobre alguns resultados que podem ser ajustados por funções presentes nesse formalismo.

D.1 Breve introdução à Mecânica Estatística Não Extensiva

A Mecânica Estatística Não Extensiva foi inicialmente proposta por Tsallis [197] como uma extensão da Mecânica Estatística Boltzmann-Gibbs. A ideia inicial foi substituir a forma funcional da entropia de Shannon pela expressão

$$S_q = k\frac{1-\sum_{i=1}^{W} p_i^q}{q-1}\,, \tag{D.1}$$

sendo k uma constante, W o número de estados acessíveis ao sistema, p_i a probabilidade de encontrar o sistema ocupando o i-ésimo estado e $q \in \mathbb{R}$ um parâmetro. Observamos que essa forma funcional tem como caso particular a entropia Shannon no limite de $q \to 1$ e, portanto, a Mecânica Estatística Não Extensiva contém a de Boltzmann-Gibbs.

A entropia S_q compartilha várias propriedades importantes do ponto de vista da Termodinâmica com entropia de Shannon. Entre elas: não negatividade, valor extremo na equiprobilidade dos estados e concavidade. Entretanto, S_q é intrinsecamente não aditiva, i.e, para dois sistemas independentes (A e B) a relação

$$S_q(A+B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B) \tag{D.2}$$

é válida para todo valor de q. Além disso, se compusermos um sistema de um número n de subsistemas independentes S_q é não extensiva para $q \neq 1$ (desse fato segue a denominação de Mecânica Estatística Não Extensiva). Por outro lado, se compusermos um sistema de um número n de subsistemas *dependentes* em alguma maneira não trivial, a entropia de Shannon será não extensiva, enquanto a entropia S_q pode ser extensiva para algum valor de q. As referências [146, 147, 198, 199, 200] reportam sobre alguns sistemas nos quais essa propriedade se manifesta.

Um outro resultado importante relacionado à Mecânica Estatística Não Extensiva é que a distribuição de um sistema em equilíbrio, que seria gaussiana em uma situação usual, não é gaussiana. De fato, ao usarmos vínculos convenientes [201], o processo de maximização da entropia S_q conduz à função

$$P_q(x) = \mathcal{A}_q \left[1 + (1-q)\mathcal{B}_q(x-\mu_q)^2\right]^{\frac{1}{1-q}} \qquad (\forall q \in \mathbb{R}|q<3), \tag{D.3}$$

na qual A_q, B_q e μ_q são constantes relacionadas à normalização de $P_q(x)$ e ao valor médio e desvio padrão de x. A distribuição $P_q(x)$ é conhecida como q-gaussiana e é uma extensão da distribuição gaussiana (recuperada quando $q \to 1$). As q-gaussianas possuem uma cauda longa (i.e., leis de potência como comportamento assintótico) quando $q > 1$ e apresentam um suporte finito quando $q < 1$.

Dentro do formalismo da Mecânica Estatística Não Extensiva é comum a utilização de algumas "funções especiais" que tornam a notação e as expressões do formalismo mais elegantes. Além disso, existem relações mais elementares do formalismo que se tornam mais evidentes ao utilizarmos essas funções. Do mesmo modo que a exponencial é a solução da equação diferencial $(y(0) = 1)$

$$\frac{dy(x)}{dx} = y(x)\,, \tag{D.4}$$

a chamada q-exponencial é a solução da seguinte equação diferencial não linear $(y(0) = 1)$

$$\frac{dy(x)}{dx} = y(x)^q\,, \tag{D.5}$$

ou seja,

$$y(x) = (1 + (1-q)x)^{\frac{1}{1-q}} \equiv e_q(x)\,. \tag{D.6}$$

Assim, podemos também definir a função inversa de $e_q(x)$ como

$$g(x) = \frac{x^{1-q} - 1}{1-q} \equiv \ln_q(x)\,, \tag{D.7}$$

a qual é denominada q-logaritmo.

Com as definições anteriores é possível re-escrever a entropia de Tsallis como

$$S_q = -k \sum_{i=1}^{W} p_i^q \ln_q(p_i) \tag{D.8}$$

e as distribuições q-gaussianas como

$$P_q(x) = \mathcal{A}_q \, e_q^{-\mathcal{B}_q(x-\mu_q)^2} \, . \tag{D.9}$$

Muitos trabalhos têm sido feitos dentro desse formalismo e consequentemente uma vasta lista de referências relacionadas ao tema está disponível (veja `http://tsallis.cat.cbpf.br/biblio.htm`). Vários desses trabalhos buscam evidências de que o formalismo não extensivo possa ser útil em situações distantes do equilíbrio onde a hipótese de caos e ergodicidade deixam de ser válidas e, consequentemente, a Mecânica Estatística Boltzmann-Gibbs deixa de ser o formalismo ideal para tratar desses sistemas. Em geral, ao tratar de sistemas complexos, a hipótese de equilíbrio termodinâmico, muitas vezes, sequer faz sentido e, portanto, é possível que funções relacionadas ao formalismo não extensivo possam ser encontradas na análise de dados oriundos de sistemas complexos. Nas seções que seguem, tentaremos fazer tais conexões com alguns dos resultados apresentados nessa tese.

D.2 Dinâmica sonora de aglomerações humanas

No Capítulo 1, no qual estudamos os sons de aglomerações humanas, vários desvios do que podemos chamar de comportamento trivial (distribuição gaussianas, ausência de correlações e distribuições exponenciais para intervalos de retorno) foram reportados. Em particular, a distribuição das amplitudes sonoras normalizadas difere muito de uma gaussiana e visualmente se parece com uma q-gaussiana. Desse modo, a primeira conexão que buscamos verificar foi com essa distribuição. Essa análise é mostrada na Figura D.1, na qual as distribuições empíricas são confrontadas com a distribuição q-gaussiana. Nesse ajuste, devido as amplitudes possuírem média nula e desvio padrão unitário, a distribuição q-gaussiana pode ser escrita como

$$P_{q\text{-gauss}}(x) = \frac{\sqrt{q-1}\,\Gamma\left(\frac{1}{q-1}\right)}{\sqrt{\pi}\sqrt{5-3q}\,\Gamma\left(\frac{1}{q-1}-\frac{1}{2}\right)}\left(1-(1-q)\frac{x^2}{5-3q}\right)^{\frac{1}{1-q}} \, , \tag{D.10}$$

ou seja, o único parâmetro a ser ajustado é q. Observamos na Figura D.1 que a q-gaussiana é certamente um descrição muito melhor do que a gaussiana que mostramos na Figura 1.2. Entretanto, é possível observar que enquanto a região central da distribuição (entre 4 desvios padrões) é muito bem descrita, a q-gaussiana subestima a calda da distribuição empírica. Durante os ajustes, também consideramos um maior peso para as caldas, o que torna o ajuste melhor naquela região; contudo, o ajuste na parte central deixa de ser tão bom.

Outra possível conexão com a Mecânica Estatística Não Extensiva aparece na análise dos intervalos de retorno da Figura 1.4. Naquela análise, consideramos como possíveis candidatas a ajustar os dados empíricos as distribuições exponenciais *stretched* [48, 49, 50] e Weilbull [51] (equações 1.1 e 1.2). Vamos agora considerar também a distribuição q-exponencial que aqui

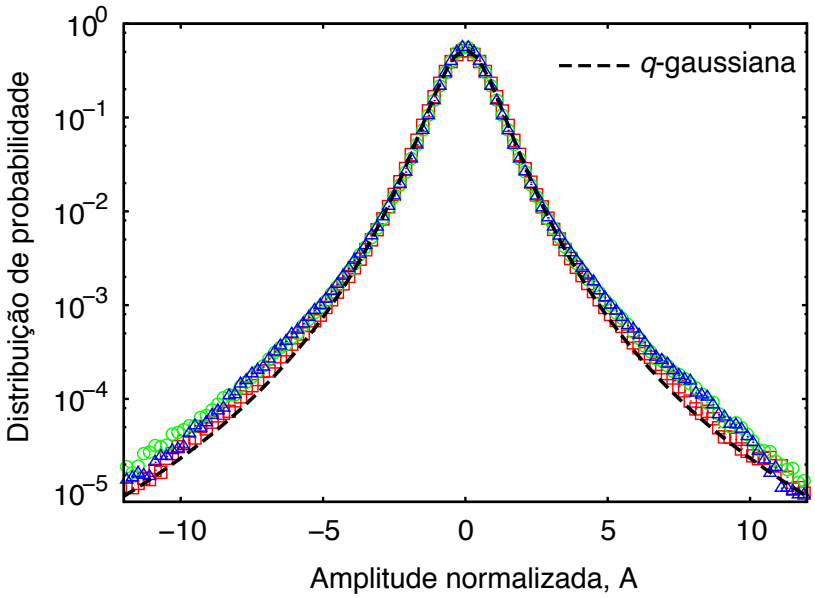


Figura D.1: Distribuição de probabilidade das amplitudes sonoras normalizadas $A(t)$ para três gravações típicas (círculos, quadrados e triângulos) em comparação com a distribuição q-gaussiana dada pela equação D.10 para $q = 1{,}38$ (linha tracejada).

pode ser escrita como

$$P_{q\text{-exp}}(x) = \frac{2-q}{3-2q}\left(1-(1-q)\frac{x}{3-2q}\right)^{\frac{1}{1-q}}, \tag{D.11}$$

visto que os intervalos de retorno escalados possuem média unitária. A Figura D.2 mostra a comparação entre a distribuição q-exponencial e a distribuição empírica dos intervalos de retorno. Ainda que o acordo como os dados empíricos não seja muito bom, é possível observar que a distribuição q-exponencial fornece um acordo superior as distribuições *stretched* e de Weilbull (Figura 1.4).

Ainda sobre os dados do Capítulo 1, investigamos se a distribuição das volatilidades das amplitudes sonoras pode ser descrita por funções ligadas ao formalismo não extensivo. Nesse caso, existem alguns trabalhos na literatura que descrevem com sucesso dados relacionados ao volume de movimentação financeira em bolsas de valores [202, 203]. A distribuição que esses autores utilizam pode ser definida como

$$P_{q\text{-gamma}}(x) = \frac{(q-1)^{\alpha}\,\Gamma\left(\frac{1}{q-1}\right)}{\theta\,\Gamma\left(\frac{1}{q-1}-\alpha-1\right)\Gamma(\alpha+1)}\left(\frac{v}{\theta}\right)^{\alpha}\left(1-(1-q)\frac{v}{\theta}\right)^{1/(1-q)}, \tag{D.12}$$

a qual pode ser considerada como uma generalização da distribuição gamma (no limite de $q \to 1$). Na Figura D.3, confrontamos essa distribuição com a distribuição empírica das volatilidades. Diferentemente dos dois casos anteriores, a distribuição definida pela expressão D.12 está mais distante de um bom acordo com os dados empíricos. Note que dois ajustes foram realizados, um utilizando método do mínimos com χ^2 obtido a partir do logaritmo dos valores das probabilidades

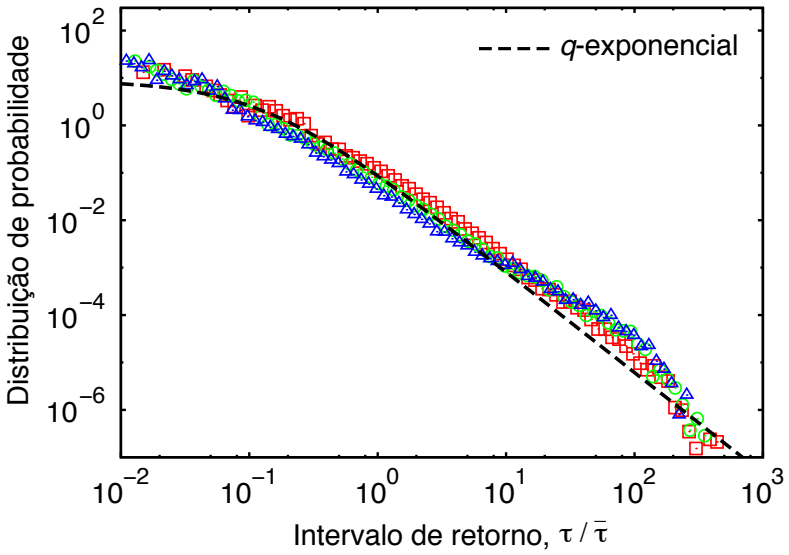


Figura D.2: Distribuição dos intervalos de retorno escalados $\tau_i/\bar{\tau}$ considerando as intensidades sonoras da Figura 1.1B e três valores limiares: 1 (quadrados), 2 (círculos) e 3 (triângulos). A linha tracejada é a distribuição $q-$exponencial com $q = 1{,}47$.

(linha contínua) e outro utilizando χ^2 usualmente (linha tracejada).

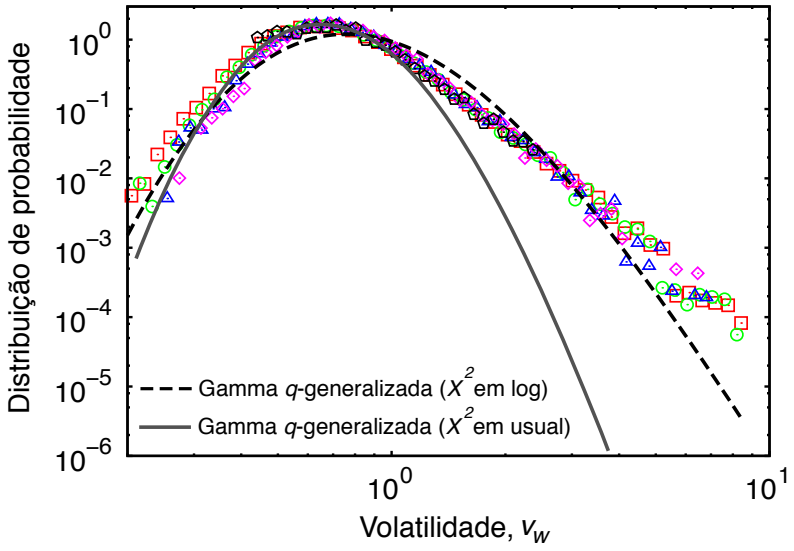


Figura D.3: Distribuição da volatilidade v_w definida na equação 1.3 e calculada para as amplitudes da Figura 1.1A. Aqui, consideramos cinco janelas: $w = 1$ (quadrados), $w = 2$ (círculos), $w = 5$ (triângulos), $w = 10$ (losangos) e $w = 100$ (pentágonos), sendo w em unidades de centésimos de segundo. A linha tracejada é a distribuição gamma q-generalizada com $\alpha = 22{,}52$, $\theta = 0{,}01$ e $q = 1{,}03$, e a linha contínua é a mesma distribuição para $\alpha = 26{,}69$, $\theta = 0{,}01$ e $q = 1{,}02$.

D.3 Dinâmica da vantagem e dos erros no xadrez

Nesta seção, investigamos conexões entre o formalismo não extensivo e as distribuições da vantagem e magnitude dos erros em partidas de xadrez. Na Figura 3.5, mostramos a distribuição da vantagem normalizada (após subtrair a médio e dividir pelo desvio padrão) em comparação com a distribuição gaussiana de valor médio nulo e desvio padrão unitário. Claramente, a gaussiana não é uma boa descrição para esses dados empíricos e, por isso, realizamos um ajuste a esses dados utilizando a q-gaussiana da equação D.10. A Figura D.4 mostra esses ajustes para as distribuições acumulas da vantagem para todas as partidas agrupadas por resultado. Notamos que q-gaussiana representa uma descrição bastante superior (ainda que desvios sistemáticos possam ser observados).

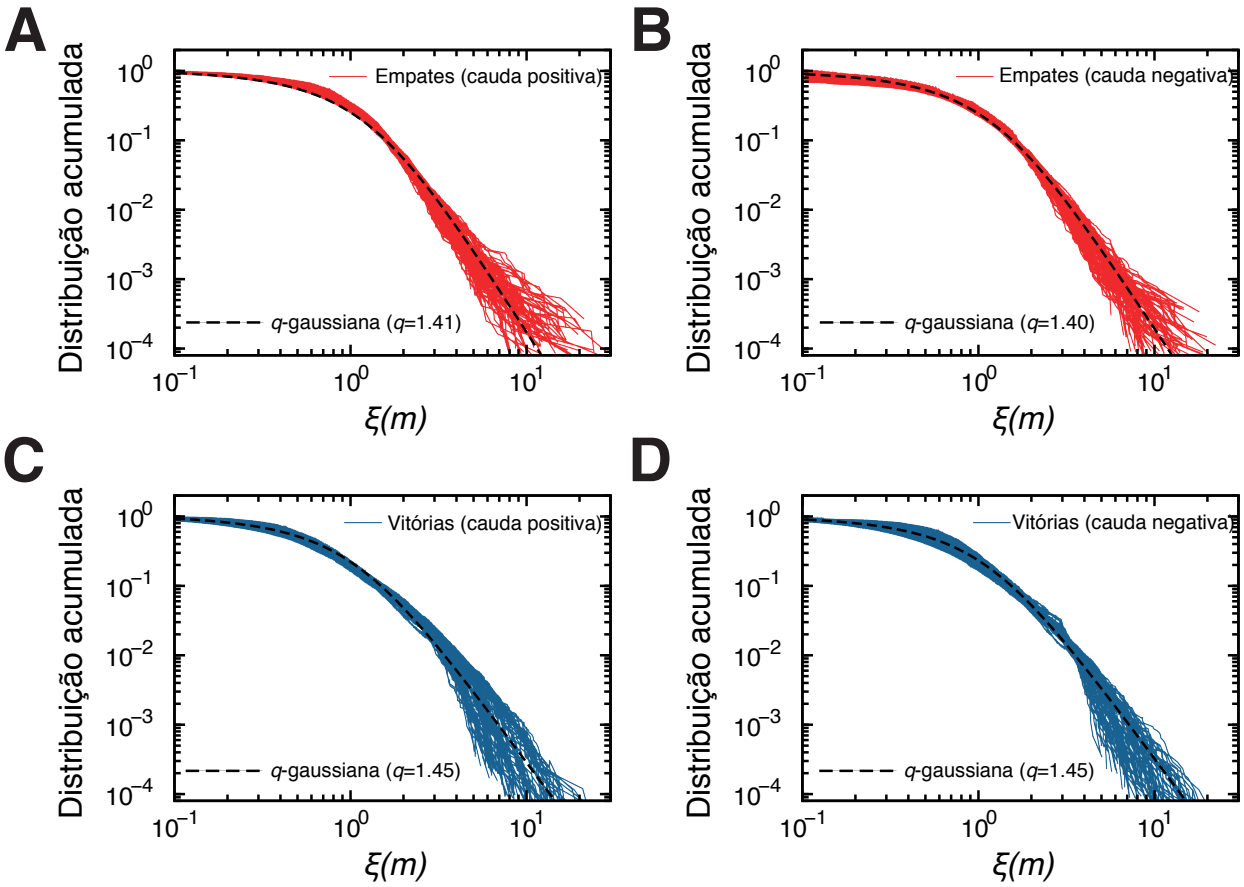


Figura D.4: Distribuições de probabilidade acumuladas da variável normalizada $\xi(m)$ (equação 3.4) ao considerar os valores (**A**) positivos e (**B**) negativos de $\xi(m)$ para partidas que terminaram em empate e os valores (**C**) positivos e (**D**) negativos de $\xi(m)$ para partidas que terminaram em vitória. Em todos os gráficos, a linha tracejada representa a distribuição acumulada $q-$gaussiana de média zero, variância unitária e parâmetro q indicado no gráfico.

No caso da distribuição das magnitudes dos erros, a distribuição do formalismo não extensivo a ser testada é a q-exponencial

$$P_{q\text{-exp}}(x,\lambda) = \lambda\,(2-q)(1-(1-q)\lambda\,x)^{\frac{1}{1-q}}\,, \tag{D.13}$$

visto que a magnitude do erro é sempre positiva. Para esse ajuste, temos dois parâmetros (λ e q) pois a distribuição empírica dos erros não apresenta nenhuma restrição quando aos momentos dos erros. Na Figura D.5, mostramos os ajustes da q-exponencial para os erros do jogador Magnus Carlsen. Note que, as q-exponenciais não fornecem um ajuste superior ao da log-normal e, para esse ajuste, o teste de Kolmogorov-Smirnov rejeita a hipótese q-exponencial. Para os demais jogadores, não podemos rejeitar a hipótese q-exponencial em 0.5% dos jogadores nas partidas terminadas em empate, 1.3% para vitórias e 2% para derrotas.

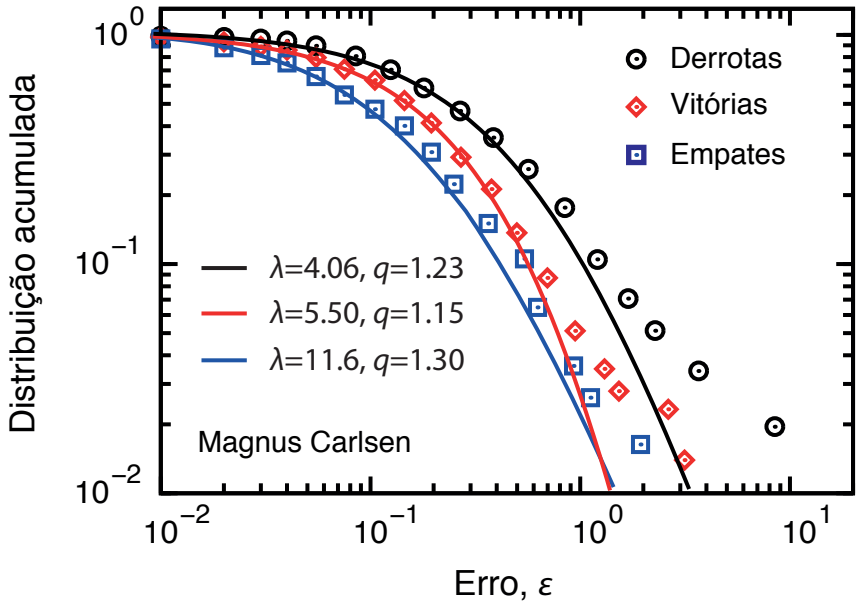


Figura D.5: Distribuição de probabilidade acumulada da magnitude dos erros ε cometidos por Magnus Carlsen (atual melhor jogador do mundo) em partidas em que ele perdeu (círculos), ganhou (losangos) ou empatou (quadrados). As linhas contínuas são distribuições de probabilidade q-exponenciais (equação D.13) com valores de λ e q indicados nos gráficos. Os valores p do teste de Kolmogorov-Smirnov são todos menores do que 0.01 e, portanto, a hipótese de que a q-exponencial descreva esses dados é rejeitada.

D.4 Dinâmica de bolhas em água fervente

Nesta última seção, investigamos conexões do formalismo não extensivo e os intervalos de retorno relacionados à dinâmica de bolhas em água fervente. Como já dissemos, processos estocásticos gaussianos e não correlacionados apresentam uma distribuição exponencial [144] para os intervalos de retorno. Se mantivermos a distribuição desses processos gaussiana e os correlacionarmos como no movimento browniano fracionário, a distribuição dos intervalos de retorno passa a ser a distribuição de Weilbull [144]. Para o melhor do conhecimento do autor, não existe uma teoria que considere os dois cenários, i.e., que descreva o comportamento dos intervalos de retorno quando o processo estocástico é não gaussiano e correlacionado. Nesse sentido, empregar outras distribuições para descrever os intervalos de retorno nessas condições parece ser um problema em aberto.

Retomando aos dados da dinâmica de bolhas em água fervente, mostramos na Figuras 5.3 e 5.5 que ambas as distribuições exponencial e de Weilbull (e também a exponencial *stretched*) não podem descrever esses dados. De fato, como discutimos anteriormente, a intensidade do laser é correlacionada de longo alcance e não é normalmente distribuída; portanto, estas distribuição devem ser vistas apenas como aproximações. Aqui, na tentativa de descrever melhor os intervalos de retorno, vamos utilizar duas distribuições relacionadas ao formalismos não extensivo, a distribuição q-exponencial [204, 205]

$$P_{q\text{-exp}}(x) = \frac{2-q}{3-2q}\left(1-(1-q)\frac{x}{3-2q}\right)^{\frac{1}{1-q}}\,, \tag{D.14}$$

e também a distribuição de Weilbull q-generalizada (q-Weilbull)

$$P_{q\text{-Weibull}}(x) = \gamma\,(2-q)\,\Lambda\,x^{\gamma-1}\,(1-(1-q)\Lambda\,x^{\gamma})^{\frac{1}{1-q}}\,, \tag{D.15}$$

em que

$$\Lambda = \frac{\left(\frac{\gamma\,\Gamma\left(\frac{1}{q-1}-1\right)}{\Gamma\left(\frac{1}{\gamma}\right)\Gamma\left(-\frac{1}{\gamma}-1+\frac{1}{q-1}\right)}\right)^{-\gamma}}{q-1}\,.$$

Na forma como estão escritas, essas distribuições apresentam média unitária e, por isso, ajustamos os seus parâmetros utilizando os intervalos de retorno escalados ($\xi = \tau/\bar{\tau}_q$). A Figura D.6 mostra esses ajustes. Observamos que ambas as distribuições representam descrições superiores as distribuições exponencial e de Weilbull. Em particular, para esses dados, a distribuição q-Weibull (que possui um parâmetro a mais) mostra-se ligeiramente superior a q-exponencial. Desse modo, podemos sugerir que o uso que tais distribuições para ajustar intervalos de retorno fornece uma melhor descrição quando a serie original (da qual obtemos os intervalos de retorno) é não gaussiana e correlacionada. Devemos ressaltar que esse tópico merece uma investigação mais sistemática, na qual poderia se avaliar melhor o efeito conjunto da série original ser não gaussiana e correlacionada.

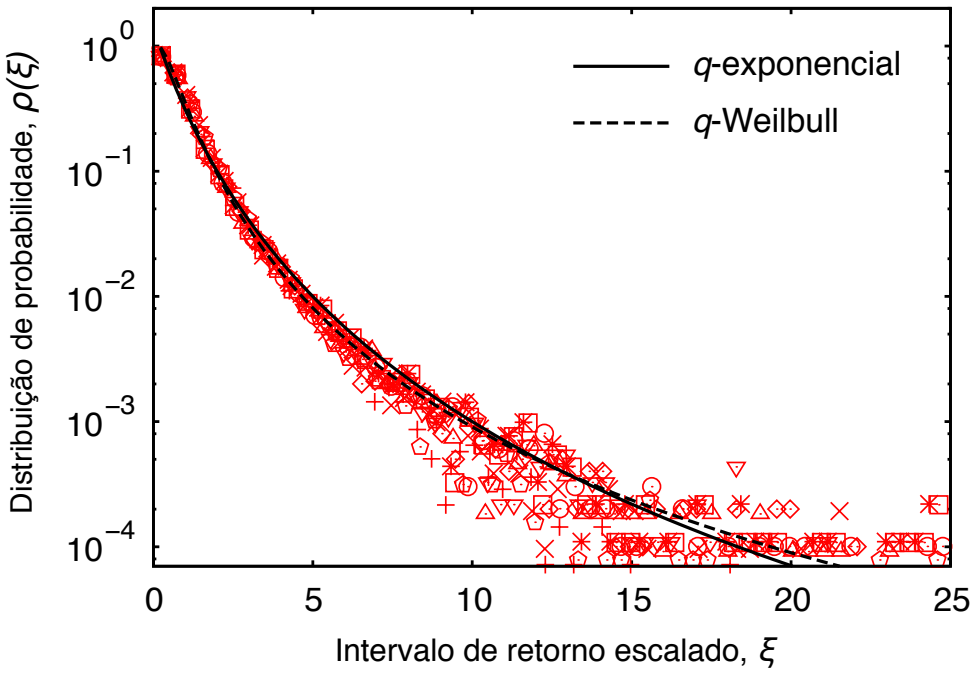


Figura D.6: Distribuição de probabilidade dos intervalos de retorno escalados ($\xi = \tau/\bar{\tau}_q$) para oito valores do limiar igualmente espaçados entre $0{,}1$ e $0{,}9$. A linha contínua é uma distribuição q-exponencial (equação D.14) com $q = 1{,}22$ e a linha tracejada uma distribuição q-Weilbull (equação D.15) com $q = 1{,}35$ e $\gamma = 1{,}31$.

Referências Bibliográficas

[1] M. Gell-Mann, *The Quark and the Jaguar - Adventures in the simple and the complex* (W. H. Freeman and Company, New York, 1994).

[2] R. Lopez-Pena, R. Capovilla, R. Garcia-Pelayo, H. Waelbroeck, *Complex System and Binary Networks* (Springer, Berlin, 1995).

[3] N. Boccara, *Modeling complex systems* (Springer-Verlag, New York, 2004).

[4] Aristotle, Metaphysics 1045a 8-10 (`http://classics.mit.edu/Aristotle/metaphysics.html`, traduzido por W. D. Ross).

[5] *"Unified Theory" is getting closer, Hawking Predicts*, San Jose Mercury News, Jan 23, 2000.

[6] Embora existam dúvidas se Galileu teria ou não realizado esse experimento (`http://www.if.ufrgs.br/historia/galileu.html`).

[7] R. Hooke, *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses with observations and Inquiries thereupon* (Royal Society, 1667).

[8] L. Page, *Method for node ranking in a linked database*, U.S. Patent 6285999, filed Jan 9, 1998, and issued Sep 4, 2001.

[9] W. T. Chiu, Y. S. Ho, *Bibliometric analysis of tsunami research*, Scientometrics **73**, 3 (2007).

[10] F. Radicchi, S. Fortunato, C. Castellano, *Universality of citation distributions: Toward an objective measure of scientific impact*, Proc. Natl. Acad. Sci. U. S. A. **105**, 17268 (2008).

[11] A. M. Petersen, H. E. Stanley, S. Succi, *Statistical regularities in the rank-citation profile of scientists*, Scientific Reports **1**, 181 (2011).

[12] H. V. Ribeiro, R. T. de Souza, E. K. Lenzi, R. S. Mendes, L .R. Evangelista, *The soundscape dynamics of human agglomeration*, New J. Phys. **13**, 023028 (2011).

[13] H. V. Ribeiro, E. K. Lenzi, R. S. Mendes, *The soundscape dynamics of human agglomeration: towards a social sound sensor* (em preparação).

[14] R. S. Mendes, H. V. Ribeiro, F. C. M. Freire, A. A. Tateishi, E. K. Lenzi, *Universal patterns in sound amplitudes of songs and music genres*, Phys. Rev. E **83**, 017101 (2011).

[15] H. V. Ribeiro, L. Zunino, R. S. Mendes, E. K. Lenzi, *Complexity-entropy causality plane: a useful approach for distinguishing songs*, Physica A **391**, 2421 (2012).

[16] H. V. Ribeiro, L. Zunino, R. S. Mendes, E. K. Lenzi, *A quantitative approach to quantify music evolution: Is popular music becoming poor?* (em preparação).

[17] H. V. Ribeiro, R. S. Mendes, E. K. Lenzi, M. del Castillo-Mussot, L. A. N. Amaral, *Move-by-move dynamics of the advantage in chess matches reveals population-level learning of the game* (PloS One, 2013).

[18] H. V. Ribeiro, L. A. N. Amaral, *Large-scale analysis of under pressure decision making in complex environments* (em preparação).

[19] H. V. Ribeiro, S. Mukherjee, X. H. T. Zeng, *Anomalous diffusion and long-range correlations in the score evolution of the game of cricket*, Phys. Rev. E **86**, 022102 (2012).

[20] H. V. Ribeiro, R. S Mendes, E. K. Lenzi, M. P. Belancon, L. C. Malacarne, *On the dynamics of bubbles in boiling water*, Chaos Solitons and Fractals *44*, 178 (2011).

[21] H. V. Ribeiro, L. Zunino, E. K. Lenzi, P. A. Santoro, R. S. Mendes, *Complexity-entropy causality plane as a complexity measure for two-dimensional patterns*, PLoS One **7**, e40689 (2012).

[22] E. K. Lenzi, H. V. Ribeiro, H. Mukai, R. S. Mendes, *Continuous-time random walk as a guide to fractional Schrödinger equation*, J. Math. Phys. **51**, 092102 (2010).

[23] E. K. Lenzi, L. R. Evangelista, M. K. Lenzi, H. V. Ribeiro, E. C. Oliveira, *Solutions for a non-Markovian diffusion equation*, Phys. Lett. A **374**, 4193 (2010).

[24] E. K. Lenzi, P. R. G. Fernandes, T. Petrucci, H. Mukai, H. V. Ribeiro, *Anomalous-diffusion approach applied to the electrical response of water*, Phys. Rev. E **84**, 041128 (2011).

[25] H. V. Ribeiro, E. K. Lenzi, R. S. Mendes, P. A. Santoro, *Anomalous diffusion in a symbolic model*, Physica Scripta **83**, 045007 (2011).

[26] A. T. Silva, E. K. Lenzi, L. R. Evangelista, M. K. Lenzi, H. V. Ribeiro, A. A. Tateishi, *Exact propagator for a Fokker-Planck equation, first passage time distribution, and anomalous diffusion*, J. Math. Phys., **52**, 083301 (2011).

[27] E. K. Lenzi, H. V. Ribeiro, J. Martins, M. K. Lenzi, G. G. Lenzi, S. Spheccia, *Non-Markovian diffusion equation and diffusion in a porous catalyst*, Chem. Eng. J. **172**, 1083 (2011).

[28] A. A. Tateishi, E. K. Lenzi, H. V. Ribeiro, L. R. Evangelista, R. S. Mendes, L. R. da Silva, *Solutions for a diffusion equation with a backbone term*, J. Stat. Mech. Theor. Exp., P02022 (2011).

[29] J. Martins, H. V. Ribeiro, L. R. Evangelista, L. R. da Silva, E. K. Lenzi, *Fractional Schrödinger equation with noninteger dimensions*, Appl. Math. Comput. **219**, 2313 (2012).

[30] A. A. Tateishi, E. K. Lenzi, L. R. da Silva, H. V. Ribeiro, S. Picoli, R. S. Mendes, *Different diffusive regimes, generalized Langevin and diffusion equations*, Phys. Rev. E **85**, 011147 (2012).

[31] L. S. Lucena, L. R. da Silva, A. A. Tateishi, H. V. Ribeiro, E. K. Lenzi, *Solution for fractional diffusion equation with noninteger dimensions*, Nonlinear Anal. Real World Appl. **13**, 1955 (2012).

[32] H. V. Ribeiro, R. Rossato, A. A. Tateishi, E. K. Lenzi, R. S. Mendes, *Continuous Time Random Walk and different diffusive regimes*, Acta Sci-Technol. **34**, 201 (2012).

[33] R. S.Mendes, L. C. Malacarne, R. P. B. Santos, H. V. Ribeiro, S. Picoli, *Earthquake-like patterns of acoustic emission in crumpled plastic sheets*, EPL **92**, 29001 (2010).

[34] S. Picoli, J. J. V. Teixeira, H. V. Ribeiro, L. C. Malacarne, R. P. B. Santos, R. S. Mendes, *Spreading patterns of the influenza A (H1N1) pandemic*, Plos One **6**, e17823 (2011).

[35] M. C. Mantovani, H. V. Ribeiro, M. V. Moro, S. Picoli Jr., R.S. Mendes, *Scaling laws and universality in the choice of election candidates*, EPL **96**, 48001 (2011).

[36] C. Castellano, S. Fortunato, V. Severo, *Statistical physics of social dynamics*, Rev. Mod. Phys. **81**, 591 (2009).

[37] S. Fortunato, C. Castellano, *Scaling and universality in proportional elections*, Phys. Rev. Lett. **99**, 138701 (2007).

[38] R. N. Costa Filho, M. P. Almeida, J. S. Andrade, J. E. Moreira, *Scaling behavior in a proportional voting process*, Phys. Rev. E **60**, 1067 (1999).

[39] D. J. Watts, S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature **393**, 440 (1998).

[40] M. E. J. Newman, *Scientific collaboration networks. Network construction and fundamental results*, Phys. Rev. E **64**, 016131 (2001).

[41] T. Zhou, H. A. T. Kiet, B. J. Kim, B. H. Wang, P. Holme, *Role of activity in human dynamics*, EPL **82**, 28002 (2008).

[42] J. G. Oliveira, A.L. Barabási, *Human dynamics: Darwin and Einstein correspondence patterns*, Nature **437**, 1251 (2005).

[43] A. Vazquez, *Impact of memory on human dynamics*, Physica A **373**, 747 (2007).

[44] A. L. Barabási, *The origin of bursts and heavy tails in human dynamics*, Nature **435**, 207 (2005).

[45] D. Brockmann, L. Hufnagel, T. Geisel, *The scaling laws of human travel*, Nature **439**, 462 (2006).

[46] M. C. González, C. A. Hidalgo, A. L. Barabási, *Understanding individual human mobility patterns*, Nature **453**, 779 (2008).

[47] H. V. Ribeiro, R. S. Mendes, L. C. Malacarne, S. Picoli Jr., P. A. Santoro, *Dynamics of tournaments: the soccer case*, Eur. Phys. J. B **75**, 327 (2010).

[48] A. Bunde, J. F. Eichner, J. W. Kantelhardt, S. Havlin, *Long-term memory: a natural mechanism for the clustering of extreme events and anomalous residual times in climate records*, Phys. Rev. Lett. **94**, 048701 (2005).

[49] F. Wang, K. Yamasaki, A. Havlin, H. E. Stanley, *Scaling and memory of intraday volatility return intervals in stock markets*, Phys. Rev. E **73**, 026117 (2006).

[50] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, H. E. Stanley, *Scaling and memory in volatility return intervals in financial markets*, Proc. Natl. Acad. Sci. U.S.A. **102**, 9424 (2005).

[51] R. Blender, K. Fraedrich, F. Sienz, *Extreme event return times in long-term memory processes near* $1/f$, Nonlinear Processes Geophys. **15**, 557 (2008).

[52] R. N. Mantegna, H. E. Stanley, *An introduction to econophysics* (Cambridge University Press, Cambridge, 1999).

[53] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, Econometrics **31**, 307 (1986).

[54] S. M. D. Queirós, *On a generalised model for time-dependent variance with long-term memory*, EPL **80**, 30005 (2007).

[55] S. M. D. Queirós, *On discrete stochastic processes with long-lasting time dependence in the variance*, Eur. Phys. J. B **66**, 137 (2008).

[56] T. DeNora, *The music of everyday life* (Cambridge University Press, Cambridge, 2000).

[57] D. L. Silva, M. M. Soares, M. V. C. Henriques, M. T. S. Alves, S. G. de Aguiar, T. P. de Carvalho, G. Corso, L. S. Lucena, *The complex network of the brazilian popular music*, Physica A **332**, 559 (2004).

[58] J. A. Davies, *The individual success of musicians, like that of physicists, follows a stretched exponential distribution*, Eur. Phys. J. B **27**, 445 (2002).

[59] E. P. Borges, *Comment on The individual success of musicians, like that of physicists, follows a stretched exponential distribution by J.A. Davies*, Eur. Phys. J. B **30**, 593 (2002).

[60] H. B. Hu, D. Y. Han, *Empirical analysis of individual popularity and activity on an online music service system*, Physica A **387**, 5916 (2008).

[61] R. Lambiotte, M. Ausloos, *Uncovering collective listening habits and music genres in bipartite networks*, Phys. Rev. E **72**, 066107 (2005).

[62] R. Lambiotte, M. Ausloos, *On the genre-fication of music: a percolation approach*, Eur. Phys. J. B **50**, 183 (2006).

[63] R. Lambiotte, M. Ausloos, *Endo- vs. exogenous shocks and relaxation rates in book and music sales*, Physica A **362**, 485 (2006).

[64] R. F. Voss, J. Clarke, $1/f$ *noise in music and speech*, Nature **258**, 317 (1975).

[65] K. J. Hsü, A. Hsü, *Self-similarity of the* $1/f$ *noise called music*, Proc. Natl. Acad. Sci. U.S.A. **88**, 3507 (1991).

[66] K. J. Hsü, A. Hsü, *Fractal geometry of music*, Proc. Natl. Acad. Sci. U.S.A. **87**, 938 (1990).

[67] J. P. Boon, O. Decroly, *Dynamical systems theory for music dynamics*, Chaos **5**, 501 (1995).

[68] M. Bigerelle, A. Iost, *Fractal dimension and classification of music*, Chaos Solitons Fractals **11**, 2179 (2000).

[69] P. Diodati, S. Piazza, *Different amplitude and time distribution of the sound of light and classical music*, Eur. Phys. J. B **17**, 143 (2000).

[70] G. Gündüz, U. Gündüs, *The mathematical analysis of the structure of some songs*, Physica A **357**, 565 (2005).

[71] Z. Y. Su, T Wu, *Multifractal analyses of music sequences*, Physica D **221**, 188 (2006).

[72] N. Scaringella, G. Zoia, D. Mlynek, *Automatic genre classification of music content: a survey*, IEEE Signal Process. Mag. **23**, 133 (2006).

[73] G. R. Jafari, P. Pedram, L. Hedayatifar, *Long-range correlation and multifractality in Bach's Inventions pitches*, J. Stat. Mech. P04012 (2007).

[74] L. Dagdug, J. Alvarez-Ramirez, C. Lopez, R. Moreno, E. Hernandez-Lemus, *Correlations in a Mozart's music score (K-73x) with palindromic and upside-down structure*, Physica A **383**, 570 (2007).

[75] Z. Y. Su, T. Wu, *Music walk, fractal geometry in music*, Physica A **380**, 418 (2007).

[76] M. Beltrán del Río, G. Cocho, G. G. Naumis, *Universality in the tail of musical note rank distribution*, Physica A **387**, 5552 (2008).

[77] W. Ro, Y. Kwon, $1/f$ *noise analysis of songs in various genre of music*, Chaos Solitons Fractals **42**, 2305 (2009).

[78] J. Serrà, X. Serra, R. G. Andrzejak, *Cross recurrence quantification for cover song identification*, New J. Phys. **11**, 093017 (2009).

[79] M. M. Mostafa, N. Billor, *Recognition of Western style musical genres using machine learning techniques*, Expert Syst. Appl. **36**, 11378 (2009).

[80] J. P. Boon, *Complexity, time and music*, Adv. Complex. Syst. **13**, 155 (2010).

[81] H. D. Jennings, P. Ch. Ivanov, A. M. Martins, P. C. da Silva, G. M. Viswanathan, *Variance fluctuations in nonstationary time series: a comparative study of music genres*, Physica A **336**, 585 (2004).

[82] ISMIR - The International Society for Music Information Retrieval (`http://www.ismir.net`).

[83] `www.billboard.com` (acessado em Agosto de 2012).

[84] L. F. Richardson, *Atmospheric diffusion shown on a distance-neighbour graph*, Proc. R. Soc. London Ser. A **110**, 709 (1926).

[85] C. Bandt, B. Pompe, *Permutation entropy: a natural complexity measure for time series*, Phys. Rev. Lett. **88**, 174102 (2002).

[86] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, M. A. Fuentes, *Distinguishing noise from chaos*, Phys. Rev. Lett. **99**, 154102 (2007).

[87] C. E. Shannon, *A mathematical theory of communication*, Bell. Syst. Tech. J. **27**, 623 (1948).

[88] P. W. Lamberti, M. T. Martin, A. Plastino, O. A. Rosso, *Intensive entropic non-triviality measure*, Physica A **334**, 119 (2004).

[89] I. Grosse I, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, H. E. Stanley, *Analysis of symbolic sequences using the Jensen-Shannon divergence*, Phys. Rev. E **65**, 041905 (2002).

[90] J. Lin, *Divergence measures based on the Shannon entropy*, IEEE Trans. Inf. Theory **37**, 145 (1991).

[91] R. López-Ruiz, H. L. Mancini, X. Calbet, *A statistical measure of complexity*, Phys. Lett. A **209**, 321 (1995).

[92] M. T. Martin, A. Plastino, O. A. Rosso, *Statistical complexity and disequilibrium*, Phys. Lett. A **311**, 126 (2003).

[93] M. T. Martin, A. Plastino, O. A. Rosso, *Generalized statistical complexity measures: Geometrical and analytical properties*, Physica A **369**, 439 (2006).

[94] L. Zunino, M. Zanin, B. M. Tabak, D. G. Pérez, O. A. Rosso, *Complexity-entropy causality plane: A useful approach to quantify the stock market inefficiency*, Physica A **389**, 1891 (2010).

[95] L. Zunino, B. M. Tabak, F. Serinaldi, M. Zanin, D. G. Pérez, O. A. Rosso, *Commodity predictability analysis with a permutation information theory approach*, Physica A **390**, 876 (2011).

[96] V. N. Vapnik, *The nature of statistical learning theory* (Springer, New York, 1995).

[97] T. Joachims, `http://svmlight.joachims.org` (acessado em Novembro de 2011).

[98] L. A. N. Amaral, J. M. Ottino, *Augmenting the framework for the study of complex systems*, Eur. Phys. J. B **38**, 147 (2004).

[99] C. O'Brien, *Checkmate for chess historians*, Science **265**, 1168 (1994).

[100] C. E. Shannon, *Programming a computer for playing chess*, Philosophical Magazine **41**, 314 (1950).

[101] F. Gobet, A. de Voogt, J. Retschitzki, *Moves in mind: the psychology of board games* (Psychology Press, Hove and New York, 2004).

[102] P. Saariluoma, *Chess players' thinking: A cognitive psychological approach* (Routledge, London, 1995).

[103] R. M. Hyatt, Crafty Chess v23.3, `www.craftychess.com` (acessado em julho de 2011).

[104] R. Metzler, J. Klafter, *The random walk's guide to anomalous diffusion: a fractional dynamics approach*, Phys. Rep. **339**, 1 (2000).

[105] P. Siegle, I. Goychuk, P. Hänggi, *Origin of hyperdiffusion in generalized brownian motion*, Phys. Rev. Lett. **105**, 100602 (2010).

[106] A. Elo, *The tating of chess players, past and present* (Batsford, London, 1978).

[107] G. T. Skalski, J. F. Gilliam, *Modeling diffusive spread in a heterogeneous population: A movement study with stream fish*, Ecology **81**, 1685 (2000).

[108] R. Vida, Critter Chess v1.6a, `http://www.vlasak.biz/critter/` (acessado em Julho de 2011).

[109] T. Ishikawa, N. Yoshida, H. Ueno, M. Wiedeman, Y. Imai, T. Yamaguchi, *Energy transport in a concentrated suspension of bacteria*, Phys. Rev. Lett. **107**, 028102 (2011).

[110] J. L. Iribarren, E. Moro, *Impact of human activity patterns on the dynamics of information diffusion*, Phys. Rev. Lett. **103**, 038702 (2009).

[111] N. E. Humphries, N. Queiroz, J. R. M. Dyer, N. G. Pade, M. K. Musyl, K. M. Schaefer, D. W. Fuller, J. M. Brunnschweiler, T. K. Doyle, J. D. R. Houghton, G. C. Hays, C. S. Jones, L. R. Noble, V. J. Wearmouth, E. J. Southall , D. W. Sims, *Environmental context explains Lévy and Brownian movement patterns of marine predators*, Nature **465**, 1066 (2010).

[112] Y. Sagi, M. Brook, I. Almog, N. Davidson, *Observation of anomalous diffusion and fractional self-similarity in one dimension*, Phys. Rev. Lett. **108**, 093002 (2012).

[113] F. Bardou, J. Bouchaud, A. Aspect, C. Cohen-Tannoudji, *Lévy statistics and laser cooling* (Cambridge University Press, Cambridge, England, 2002).

[114] S. A. Trigger, *Anomalous transport in velocity space: from Fokker-Planck to the general equation*, J. Phys. A: Math. Theor. **43**, 285005 (2010).

[115] S. C. Lim, S. V. Muniandy, *Self-similar Gaussian processes for modeling anomalous diffusion*, Phys. Rev E **66**, 021114 (2002).

[116] `www.espncricinfo.com` (acessado em Fevereiro de 2012).

[117] H. Hisken, *The Fokker-Planck equation: methods of solutions and applications* (Springer, New York, 1996).

[118] R. Kubo, *The fluctuation-dissipation theorem*, Rep. Prog. Phys. **29**, 255 (1966).

[119] K. G. Wang, C. W. Lung, *Long-time correlation effects and fractal Brownian motion*, Phys. Lett. A **151**, 119 (1990).

[120] K. G. Wang, *Long-time-correlation effects and biased anomalous diffusion*, Phys. Rev. A **45**, 833 (1992).

[121] I. Podlubny, *Fractional differential equations* (Academic Press, San Diego, 1999).

[122] A. Reifman, *Hot Hand: The statistics behind sports' greatest streaks* (Potomac Books Inc., Dulles, 2011).

[123] T. Gilovich, R. Vallone, A. Tversky, *The hot hand in basketball: on the misperception of random sequences*, Cognitive Psychology **17**, 295 (1985).

[124] G. Yaari, S. Eisenmann, *The hot (invisible?) hand: can time sequence patterns of success/failure in sports be modeled as repeated random independent trials?*, PLoS One **6**, e24532 (2011).

[125] G. Yaari, S. Eisenmann, *"Hot hand" on strike: bowling data indicates correlation to recent past results, not causality*, PLoS One **7**, e30112 (2012).

[126] A. Prosperetti, *Bubbles*, Phys. Fluids **16**, 1852 (2004).

[127] D. Lohse, *Bubble puzzles*, Physics Today **56**, 36 (2003).

[128] J. B. Joshi, *Computational flow modelling and design of bubble column reactors*, Chem. Eng. Sci. **56**, 5893 (2001).

[129] J. Alvarez-Ramirez, G. Espinosa-Paredes, A. Vazquez, *Detrended fluctuation analysis of the neutronic power from a nuclear reactor*, Physica A **351**, 227 (2005).

[130] M. P. Brenner, S. Hilgenfeldt, D. Lohse, *Single-bubble sonoluminescence*, **74**, 425 (2002).

[131] N. Vandewalle, J. F. Lentz, S. Dorbolo, F. Brisbois, *Avalanches of popping bubbles in collapsing foams*, Phys. Rev. Lett. **86**, 179 (2001).

[132] H. Ritacco, F. Kiefer, D. Langevin, *Lifetime of bubble rafts: cooperativity and avalanches*, Phys. Rev. Lett. **98**, 244501 (2007).

[133] L. E. Schmidt, N. C. Keim, W. W. Zhand, S. R. Nagel, *Memory-encoding vibrations in a disconnecting air bubble*, Nature Phys. **5**, 343 (2009).

[134] D. Zahn, *How does water boil?*, Phys. Rev. Lett. **93**, 227801 (2004).

[135] C. E. Brennen, *Cavitation and Bubble Dynamics* (Oxford University Press, New York, 1995).

[136] M. Shoji, *Studies of boiling chaos: a review*, Int. J. Heat Mass Transfer **47**, 1105 (2004).

[137] J. R. Thome, *Boiling in microchannels: a review of experiment and theory*, Int. J. Heat Fluid Flow **25**, 128 (2004).

[138] A. Cordonet, R. Lima, E. Ramos, *Two models for the dynamics of boiling in a short capillary tube*, Chaos **11**, 344 (2001).

[139] Y. Iida, J. Lee, T. Kozuka, K. Yasui, A. Towata, T. Tuziuti, *Optical cavitation probe using light scattering from bubble clouds*, Ultrason. Sonochem. **16**, 5519 (2009).

[140] E. J. Gumbel, *Statistics of extremes* (Dover Publications Inc., New York, 2004).

[141] J. Galambos, *The asymptotic theory of extreme order statistic* (John Wiley & Sons Inc, New York, 1978).

[142] R. D. Reiss, M. Thomas, R. D. Reiss, *Statistical analysis of extreme values: from insurance, finance, hydrology and other fields* (Birkhauser, Boston, 1997).

[143] P. Embrechts, C. Kluppelberg, T. Mikosch, *Modelling extremal events for insurance and finance* (Springer, New York, 1997).

[144] M. S. Santhanam, H. Kantz, *Return interval distribution of extreme events and long-term memory*, Phys. Rev. E **78**, 051113 (2008).

[145] V. K. Rohatgi, *Statistical inference* (John Wiley, New York, 1984).

[146] M. Buiatti, P. Grigolini, L. A. Palatella, *A non extensive approach to the entropy of symbolic sequences*, Physica A **268**, 214 (1999).

[147] H. V. Ribeiro, E. K. Lenzi, R. S. Mendes, G. A. Mendes, L. R. da Silva, *Symbolic sequences and Tsallis entropy*, Braz. J. Phys. **39**, 444 (2009).

[148] C. Kluppelberg, T. Mikosch, *Large deviations of heavy-tailed random sums with applications in insurance and finance*, J. Appl. Probab. **34**, 293 (1997).

[149] I. V. Zaliapin, Y. Y. Kagan, F. P. Schoenberg, *Approximating the distribution of Pareto sums*, Pure Appl. Geophys. **162**, 1187 (2005).

[150] S. Nadarajah, M. M. Ali, *Pareto random variables for hydrological modeling*, Water Resour. Manag. **22**, 1381 (2008).

[151] O. A. Rosso, L. Zunino, D. G. Pérez, A. Figliola, H. A. Larrondo, M. Garavaglia, M. T. Martín, A. Plastino, *Extracting features of Gaussian self-similar stochastic processes via the Bandt-Pompe approach*, Phys. Rev. E **76**, 061114 (2007).

[152] A. N. Kolmogorov, *Three approaches to the quantitative definition of information*, Probl. Inf. Transm. **1**, 3 (1965).

[153] S. Kullback, R. A. Leibler, *On information and sufficiency*, Ann. Math. Statist. **22**, 79 (1951).

[154] B. B. Mandelbrot, *The fractal geometry of nature* (Freeman, San Francisco, 1982).

[155] A. M. Lyapunov, *The general problem of the stability of motion* (Taylor-Francis, London: Translated by A. T. Fuller, 1992).

[156] F. Maes, A. Collignon, A. Vandermeulen, G. Marchal, P. Suetens, *Multimodality image registration by maximization of mutual information*, IEEE Trans. Med. Imag. **16**, 187 (1997).

[157] M. Khader, A. B. Hamza, *Nonrigid image registration using an entropic similarity*, IEEE Trans. Inf. Technol. Biomed. **15**, 681 (2011).

[158] L. Parrott, *Quantifying the complexity of simulated spatiotemporal population dynamics*, Ecological Complexity **2**, 175 (2005).

[159] L. Jost, *Entropy and diversity*, Oikos **113**, 363 (2006).

[160] R. S. Mendes, L. R. Evangelista, S. M. Thomaz, A. A. Agostinho, L. C. Gomes, *A unified index to measure ecological diversity and species rarity*, Ecography **31**, 450 (2008).

[161] L. Parrott, *Measuring ecological complexity*, Ecological Indicators **10**, 1069 (2010).

[162] U. Schwarz, A. O. Benz, J. Kurths, A. Witt, *Analysis of solar spike events by means of symbolic dynamics methods*, Astron. Astrophys. **277**, 215 (1993).

[163] G. Consolini, R. Tozzi, P. De Michelis, *Complexity in the sunspot cycle*, Astron. Astrophys. **506**, 1381 (2009).

[164] M. Lovallo, L. Telesca, *Complexity measures and information planes of x-ray astrophysical sources*, J. Stat. Mech. P03029 (2011).

[165] Z. Y. Su, T. Wu, *Multifractal analyses of music sequences*, Physica D **221**, 188 (2006).

[166] P. Grassberger, *Toward a quantitative theory of self-generated complexity*, Int. J. Theor. Phys. **25**, 907 (1986).

[167] Y. A. Andrienko, N. V. Brilliantov, J. Kurths, *Complexity of two-dimensional patterns*, Eur. Phys. J. B **15**, 539 (2000).

[168] F. D. Feldman, J. P. Crutchfield, *Structural information in two-dimensional patterns: entropy convergence and excess entropy*, Phys. Rev. E **67**, 051104 (2003).

[169] Z. Cai, E. Shen, F. Gu, Z. Xu, J. Ruan, Y. Cao, *A new two-dimensional complexity measure*, Int. J. Bifurcation Chaos **16**, 3235 (2006).

[170] G. Ouyang, D. Dang, D. A. Richards, X. Li, *Ordinal pattern based similarity analysis for EEG recordings*, Clin. Neurophysiol. **121**, 694 (2010).

[171] X. Li, G. Ouyang, *Estimating coupling direction between neuronal populations with permutation conditional mutual information*, NeuroImage **52**, 497 (2010).

[172] N. Nicolaou, J. Georgiou, *The use of permutation entropy to characterize sleep electroencephalograms*, Clin. EEG Neurosci. **42**, 24 (2011).

[173] C. Masoller, O. A. Rosso, *Quantifying the complexity of the delayed logistic map*, Phil. Trans. R. Soc. A **369**, 425 (2011).

[174] M. Barreiro, A. C. Marti, C. Masoller, *Inferring long memory processes in the climate network via ordinal pattern analysis*, Chaos **21**, 013101 (2011).

[175] J. A. Cánovas, A. Guillamón, M. del Carmen-Ruíz, *Using permutations to detect dependence between time series*, Physica D **240**, 1199 (2011).

[176] N. Nicolaou, J. Georgiou, *Detection of epileptic electroencephalogram based on permutation entropy and support vector machines*, Expert Syst. Appl. **39**, 202 (2012).

[177] R. López-Ruiz, H. L. Mancini, X. Calbet, *A statistical measure of complexity*, Phys. Lett. A **209**, 321 (1995).

[178] A. Fournier, D. Fussel, L. Carpenter, *Computer rendering of stochastic models*, Commun. ACM **25**, 371 (1982).

[179] L. J. Yu, A. Saupe, *Observation of a biaxial nematic phase in potassium laurate-1-decanol-water mixtures*, Phys. Rev. Lett. **45**, 1000 (1980).

[180] `http://dept.kent.edu/spie/liquidcrystals/` (acessado em Janeiro de 2012).

[181] A. F. Brito, J. A. Redinz, J. A. Plascak, *Dynamics of rough surfaces generated by two-dimensional lattice spin models*, Phys. Rev. E **75**, 046106 (2007).

[182] A. F. Brito, J. A. Redinz, J. A. Plascak, *Two-dimensional XY and clock models studied via the dynamics generated by rough surfaces*, Phys. Rev. E **81**, 031130 (2010).

[183] R. V. Hogg, A. Craig, *Introduction to Mathematical Statistics* (Prentice Hall, New York, 1995).

[184] B. B. Mandelbrot, J. W. V. Ness, *Fractional Brownian motions, fractional noises and applications*, SIAM Review **10**, 422 (1968).

[185] D. Koutsoyiannis, *The Hurst phenomenon and fractional Gaussian noise made easy*, Hydrological Sciences Journal **47**, 573 (2002).

[186] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, *Mosaic organization of DNA nucleotides*, Phys. Rev. E **49**, 1685 (1994).

[187] D. Vjushin, R. B. Govindan, R. A. Monetti, S. Havlin, A. Bunde, *Scaling analysis of trends using DFA*, Physica A **302**, 234 (2001).

[188] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, H. E. Stanley, *Effect of trends on detrended fluctuation analysis*, Phys. Rev. E **64**, 011114 (2001).

[189] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, A. Bunde, *Detecting long-range correlations with detrended fluctuation analysis*, Physica A **295**, 441 (2001).

[190] Z. Chen, P. C. Ivanov, K. Hu, H. E. Stanley, *Effect of nonstationarities on detrended fluctuation analysis*, Phys. Rev. E **65**, 041107 (2002).

[191] M. S. Taqqu, V. Teverovsky, W. Willinger, *Estimators for long-range dependence: an empirical study*, Fractals **3**, 785 (1995).

[192] G. W. Corder, D. I. Foreman, *Nonparametric statistics for non-statisticians: a step-by-step approach* (Wiley, New Jersey, 2009).

[193] W. T. Eadie, D. Drijard, F. E. James, M. Roos, B. Sadoulet, *Statistical methods in experimental physics* (North-Holland Publishing, Amsterdam, 1971).

[194] E. S. Pearson, H. O. Hartley, *Biometrika tables for statisticians* (Cambridge University Press, Cambridge, 1972).

[195] G. Marsaglia G, W. W. Tsang, J. Wang, *Evaluating Kolmogorov's distribution*, Journal of Statistical Software **8**, 1 (2003).

[196] B. Efron, R. J. Tibshirani, *An introduction to the bootstrap* (Chapman and Hall/CRC, London, 1993).

[197] C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics*, J. Stat. Phys. **52**, 479 (1988).

[198] C. Tsallis, M. Gell-Mann and Y. Sato, *Asymptotically scale-invariant occupancy of phase space makes the entropy S_q extensive*, Proc. Natl. Acad. Sc. USA **102**, 15377 (2005).

[199] F. Caruso, C. Tsallis, *Extensive nonadditive entropy in quantum spin chains, in Complexity, Metastability and Nonextensivity*, American Institute of Physics Conference Proceedings **965**, 51 (2007).

[200] F. Caruso, C. Tsallis, *Nonadditive entropy reconciles the area law in quantum systems with classical thermodynamics*, Phys. Rev. E **78**, 021102 (2008).

[201] C. Tsallis , R. S. Mendes, A. R. Plastino, *The role of constraints within generalized nonextensive statistics*, Physica A **261**, 534 (1998).

[202] R. Osorio, L. Borland, C. Tsallis , *Nonextensive Entropy - Interdisciplinary Applications, edited by Gell-Mann M. and Tsallis C.* (Oxford University Press, New York, 2004).

[203] S. M. D. Queirós, *On the emergence of a generalised Gamma distribution. Application to traded volume in financial markets*, EPL **71**, 339 (2005).

[204] C. J. Keylock, *Describing the recurrence interval of extreme floods using nonextensive thermodynamics and Tsallis statistics*, Advances in Water Resources **28**, 773 (2005).

[205] J. Ludescher, C. Tsallis, A. Bunde, *Universal behaviour of interoccurrence times between losses in financial markets: An analytical description*, EPL **95**, 68002 (2011).

Agradecimentos

São muitas as pessoas que contribuíram diretamente ou indiretamente para os resultados apresentados nesta tese. Abaixo cito aquelas que considero mais importantes.

Agradeço ao *Renio*, meu orientador e amigo, por ter me mostrado um caminho fértil para pesquisar e pela liberdade concedida na condução de meus estudos. Agradeço ao *Ervin*, amigo e ávido colaborador, por ter me envolvido em seus trabalhos e pelas tantas conversas e caminhadas que tivemos ao longo deste doutoramento. Sou também muito grato ao amigo *Perseu* pela confiança em mim depositada, orientação no início da graduação e pelas colaborações em vários trabalhos. Ao *Malarcarne* agradeço pelas poucas, porém efetivas, colaborações e pelo seu interesse no meu trabalho.

Agradeço também aos colegas de curso por tornarem a convivência junto à pós-graduação mais divertida. Em especial, aos amigos: *Angel* (o famoso "Angelito"), *Marcos Paulo* (o famoso "Pop", "Usain Bolt" ou "Monsieur Confusão") e *Rodolfo* (o famoso "Rodolfo"). Agradeço ainda à senhora *Akiko* pela extrema competência junto à secretaria da pós-graduação.

Agradeço aos amigos *Xiaohan*, *Satyam* e *João* por terem me acolhido e partilhado suas experiências junto ao Amaral Lab. Sou também muito grato ao *Luís Amaral* por ter me recebido em seu grupo, pelo seu interesse em meu trabalho e pelas valiosas lições que lá aprendi.

Finalmente, agradeço ao governo brasileiro que, por meio dos programas de bolsas de estudo, torna a pós-graduação e a pesquisa básica possíveis no Brasil.

"Dedico este trabalho à minha família: Vera Lúcia, Gilberto e Bruno; e à Mônica."