UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Tese de Doutorado

Métodos de Estimação de Medidas de Informação e Aplicações em Neurociências

Doutoranda
Juliana Martins de Assis

Orientador
Francisco Marcos de Assis, Dr.

Campina Grande – PB
Novembro – 2017

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Métodos de Estimação de Medidas de Informação e Aplicações em Neurociências

Juliana Martins de Assis

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Campina Grande como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências, no domínio da Engenharia Elétrica.

Área de concentração: Processamento de Informação
Linha de Pesquisa: Eletrônica e Telecomunicações

Dr. Francisco Marcos de Assis
(Orientador)

Campina Grande – PB
Novembro – 2017

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

A848m Assis, Juliana Martins de.
Métodos de estimação de medidas de informação e aplicações em Neurociências / Juliana Martins de Assis. – Campina Grande, 2017.
119 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2017.
"Orientação: Prof. Dr. Francisco Marcos de Assis".
Referências.

1. Informação Mútua. 2. Informação Direcional. 3. Entropia de Transferência. 4. Estimação – Engenharia Elétrica. I. Assis, Francisco Marcos de. II. Título.

CDU 621.391(043)

"MÉTODOS DE ESTIMAÇÃO DE MEDIDAS DE INFORMAÇÃO E APLICAÇÕES EM NEUROCIÊNCIAS"

JULIANA MARTINS DE ASSIS

TESE APROVADA EM 23/11/2017

FRANCISCO MARCOS DE ASSIS, Dr., UFCG
Orientador(a)

BENEMAR ALENCAR DE SOUZA, D.Sc., UFCG
Examinador(a)

CHARLES CASIMIRO CAVALCANTE, Dr., UFC
Examinador(a)

DANILO SILVA, Ph.D., UFSC
Examinador(a)

LEOCARLOS BEZERRA DA SILVA LIMA, D.Sc., UFCG
Examinador(a)

ALEXANDRE JEAN RENÉ SERRES, Dr., UFCG
Examinador(a)

CAMPINA GRANDE - PB

À minha avó, Terezinha de Melo Martins, in memoriam

Agradecimentos

Agradeço a minha família, Aida Maria, Juana Tereza, Joaquim Marcos, em especial a meu pai, Francisco Marcos, que me servem de exemplo e orientação. Agradeço a todos meus amigos e colegas do Iquanta. Em especial agradeço a Viviane Martins, pela ajuda com o programa de simulação de atividade neuronal, a Mikaelle Santos pelo incentivo e colaboração no nosso primeiro artigo, e a Milena Marinho pelas discussões sobre informação direcional e entropia de transferência. Agradeço a Francisco Revson pela ajuda com as referências e com os programas de computador. Agradeço a Vinícius Vieira pelo auxílio com os *slides* da qualificação e a Edmar José pelo auxílio com os *slides* da tese. Agradeço a Saulo Dornellas, pela ajuda com algumas figuras e pelo apoio sempre presente. Agradeço também a todos os membros da banca, que contribuíram com críticas e encorajamento para seguir neste caminho. Agradeço por fim a todos os meus antigos professores, amigos e colegas que contribuíram *causalmente* na minha formação até aqui.

Resumo

Muitos projetos de sistemas modernos dependem de seus componentes, os quais podem se relacionar por meio de dependências ou causalidades entre si. Os sistemas com quais lidamos neste trabalho referem-se àqueles que apresentam componentes que podem ser quantificadas ou medidas e que permitem um tratamento matemático. Podemos citar como exemplos sistemas financeiros, cujas componentes podem ser valores das ações vendidas ou compradas; sistemas biológicos/médicos, cujas componentes podem ser grandezas como pressão arterial e taxa de batimentos cardíacos; sistemas neurofisiológicos, cujas componentes podem ser registros de eletroencefalogramas ou imageamento por ressonância magnética das diversas regiões cerebrais; dentre muitos outros sistemas. A teoria da informação tem se mostrado eficaz na quantificação dessas relações, por meio de grandezas como a informação mútua, a informação direcional e a entropia de transferência. Uma questão fundamental quando se lida com sistemas reais é a dificuldade em se modelar as distribuições de probabilidade das variáveis envolvidas, distribuições estas presentes na definição das já citadas medidas de informação. É neste contexto que surge o presente trabalho, propondo-se a investigar e contribuir com as formas de estimação de medidas de informação necessárias no estudo de sistemas. Este trabalho dá um enfoque especial em aplicações que lidam com sistemas neuronais.

Palavras-chave. Informação Mútua, Informação Direcional, Entropia de Transferência, Estimação.

Abstract

Many modern system projects depend on their components, which may relate to each other by dependency or causality relations. What is meant by systems in this work are those whose components may be evaluated or measured. For example: financial systems, whose components may be stock markets; medical/biological systems, whose components may be respiration, blood pressure, and heart rate; neurophysiological system, whose components may be electroencephalogram or functional magnetic resonance imaging from different parts of the brain; among many other systems that allow mathematical treatment. Information theory has been presented as an efficient mean to quantify the relations in these systems, bringing useful concepts and evaluating measures such as mutual information, directed information, and transfer entropy. A fundamental question when dealing with real systems concerns the difficulty to model the true underlying probability densities of the involved variables. The definitions of mutual information, directed information, and transfer entropy rely on these densities. In this context, the present work evolves to investigate and contribute with estimation methods to measure the relations among variables when studying systems. This work gives a special attention to neuronal systems.

Keywords. Mutual Information, Directed Information, Transfer Entropy, Estimation.

Sumário

Lista de Figuras

Lista de Tabelas

Notação e Terminologia

Nesta tese, variáveis aleatórias são escritas com letras maiúsculas e valores particulares que elas assumem com letras minúsculas. Processos aleatórios são escritos com letras em maiúsculo e em negrito. Séries de variáveis aleatórias são indexadas com o primeiro índice no subscrito e o último índice no sobrescrito, por exemplo, a série de variáveis $\{X_1, X_2, X_3, X_4, X_5\} = X_1^5$. É possível também omitir-se o subscrito quando o primeiro índice da série da variável aleatória for 1, ou seja, no exemplo anterior, $X_1^5 = X^5$. O símbolo $\mathbb{E}X$ denota esperança de uma variável aleatória X. Uma barra sobre uma variável aleatória denota sua média amostral, isto é:

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

O símbolo $\mathrm{var}(X)$ denota a variância de X, ao passo que $\mathrm{cov}(X, Y)$ denota covariânca entre as variáveis aleatórias X e Y, logo:

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))],$$

em que a esperança é tomada na distribuição conjunta das variáveis X e Y envolvidas. Observa-se que $\mathrm{cov}(X, X) = \mathrm{var}(X)$.

A função logarítmica, quando escrita como log, está na base 2 (caso em que as medidas de informação são feitas em bits). Quando a função logarítmica for escrita ln, ela está na base neperiana (caso em que as medidas de informação são feitas em *nats*). A entropia binária de parâmetro p é denotada por $\mathcal{H}(p)$, sendo dada por:

$$\mathcal{H}(p) = -p \log p - (1 - p) \log(1 - p) \text{ bit}.$$

Capítulo 1

Introdução

Um sistema, segundo o dicionário Aurélio, é a combinação de partes reunidas para concorrerem para um resultado, ou de modo a formarem um conjunto. Os sistemas constituem objeto de estudo de diversas áreas do conhecimento. As áreas que lidam com ciências exatas, em geral, permitem um estudo matemático dessas combinações e resultados, o qual pode ser feito por medições de variáveis aleatórias, de modo a retirar conclusões relevantes para o entendimento do funcionamento do sistema considerado.

A teoria da informação, em particular, apresenta-se como uma ferramenta interessante na investigação dessas relações de dependência, seja ela dirigida (constituindo-se causalidade) ou não. A teoria da informação, estabelecida em 1948 por um artigo de Claude Shannon [74], trouxe conhecidas contribuições às comunicações, à ciência da computação, à probabilidade e à estatística [13]. Alguns de seus conceitos, como o de informação mútua, informação direcional e entropia de transferência, podem ser aplicados em diversos sistemas, não apenas aqueles de sinais naturais, tais quais os sistemas biológicos e climáticos, como também os de sinais "artificiais", tais quais os sistemas econômicos e de processamento industrial.

A teoria da informação foi criada tendo em vista a resolução de problemas oriundos da engenharia de telecomunicações e a delimitação de seu escopo foi uma das razões principais de seu sucesso. Contudo, devido ao seu poder de explanação e suas várias aplicações práticas, os conceitos introduzidos pela teoria da informação foram progressivamente estendidos em sua utilização em outras áreas. O bem conhecido modelo de sistema de comunicação ponto a ponto é composto por fonte de informação, transmissor, canal, receptor e sorvedouro de informações. A fonte de informação é modelada a partir de um conjunto com dada distribuição de probabilidade. O transmissor gera sinais codificando mensagens e as transmite através de um canal representado por uma distribuição de probabilidade condicionada. Esta distribuição de probabilidade condicionada é modelada de acordo com o tipo de canal, ou mais especificamente, de acordo com o ruído que este canal adiciona. O sinal perturbado pelo ruído é processado no receptor cuja saída é uma estimativa da mensagem transmitida. Nesse sentido, informação é incerteza. Uma fonte de informação é modelada como uma variável aleatória ou um processo aleatório e a teoria de probabilidade é fundamental no desenvolvimento da teoria da informação. Lembramos que a ideia de significado, que está relacionada no campo de linguística à área denominada semântica, é completamente descartada nesta teoria. Entretanto, há estudos recentes que buscam pelo surgimento de símbolos e significado dentro de um contexto computacional [71].

Os exemplos de sistemas biológicos nos quais já houve uma aplicação da teoria da

informação são vários. Em genética, por exemplo, a teoria da informação foi utilizada para analisar sequências de ácido desoxirribonucleico [26]. Considerando sistemas cardíacos, é possível aferir medidas de pressão sanguínea e de batimentos cardíacos de modo a estabelecer relações de dependência ou de causalidade entre essas variáveis e o diagnóstico de doenças cardiovasculares [72].

Em sistemas neuronais, são vastas as explorações da informação mútua como medida de dependência entre estímulos físicos externos ao animal e características neurofisiológicas do mesmo [68, 63, 16, 15]. Por exemplo, na referência [16], é investigado o quanto a frequência de tons sonoros altera as alturas e latências de registros invasivos no córtex auditivo de ratos, em especial registros de baixa frequência (potenciais pós-sinápticos e potenciais de campo local). Estimando-se a informação mútua entre essas grandezas, frequência do tom *versus* altura ou latência do registro, verificou-se que há mais informação sobre os estímulos sonoros, nos registros em baixas frequências, quando a informação é codificada em grupos neuronais, de que em neurônios individualmente. A referência [16] foi uma primeira contribuição deste doutorado.

Ainda em sistemas neuronais, é possível investigar, por exemplo, como a taxa de potenciais de ação, ou *spikes* (registros em altas frequências), de um neurônio afeta a taxa de *spikes* de outro neurônio através da entropia de transferência [23]. A entropia de transferência também tem sido apontada como uma maneira eficaz de quantificar conectividade efetiva (causalidade) entre registros não invasivos do cérebro, como eletroencefalogramas (EEG) e magnetoencefalogramas (MEG) [84]. Estimativas de informação mútua entre registros de EEG de diferentes áreas cerebrais, especialmente de áreas distantes, revelaram fielmente estados de consciência distintos em pacientes se recuperando de um coma [38, 17]. Além disso, estimativas de entropia de transferência entre registros de EEG permitiram a identificação confiável do hemisfério cerebral contendo o foco epiléptico sem a necessidade de observar uma crise epiléptica de fato [77]. Tanto a entropia de transferência quanto a informação direcional já foram apontadas como formas eficazes de medir conexões neuronais a partir de registros de trens de *spikes* apenas [32, 78, 67].

Ainda pensando em termos de sistemas, em outros campos de conhecimento, como a meteorologia, pode ser feita a medição de variáveis como a velocidade dos ventos e a umidade para fazer previsões climáticas. Mesmo em campos em que as grandezas observáveis sejam artificiais, isto é, tenham sido produzidas pelo homem, como é o caso em alguns processos químicos industriais, muitas vezes essas grandezas são desconhecidas e pode ser necessário investigar a causa de alguma pertubação no processo [18, 9].

Geralmente, as distribuições de probabilidade das variáveis aleatórias envolvidas em um sistema real são desconhecidas. Essas probabilidades estão presentes nas definições das já citadas medidas de informação — informação mútua, direcional e entropia de transferência. Neste caso, torna-se necessário fazer inferências e estimações dessas medidas de informação.

Há diversas maneiras de se estimar medidas de informação, algumas das quais abordadas nesta tese. A tabela 1.1 apresenta os métodos de estimação investigados para o caso discreto, ou seja, o caso em que as variáveis aleatórias são do tipo discreto. Neste caso, os métodos são chamados aqui de método *plug-in*, método de Jiao, método de Quinn e método KDE (de *kernel density estimation*). Estes métodos serão pormenorizados no capítulo 3. Já a tabela 1.2 apresenta os métodos de estimação investigados para o caso contínuo, ou seja, o caso em que as variáveis aleatórias assumem valores reais. Neste caso, os métodos são chamados aqui de método do particionamento do suporte, método do particionamento adaptativo do suporte, método KDE, método do mapeamento sim-

bólico, método KSG e método BI-KSG. Estes métodos são pormenorizados no capítulo 5. A tabela 1.3 apresenta os métodos de estimação investigados para o caso misto, ou seja, o caso em que algumas variáveis assumem valores discretos e outras valores reais. Estes métodos são pormenorizados no capítulo 8. Neste caso, os métodos são chamados aqui de método do particionamento do suporte e método de Ross.

Medidas/Métodos	*Plug in*	Jiao	Quinn	KDE
Informação Mútua	X	X		
Informação Direcional		X	X	
Entropia de Transferência				X

Tabela 1.1: Relação entre funcionais de medidas de informação e métodos utilizados para estimá-los: caso discreto.

Medidas/Métodos	Partic.	Partic. Adapt.	KDE	Map. Simb.	KSG	BI-KSG
Informação Mútua	X	X	X	X	X	X
Informação Direcional		X	X			X
Entropia de Transferência	X	X	X	X	X	

Tabela 1.2: Relação entre funcionais de medidas de informação e métodos utilizados para estimá-los: caso contínuo. Partic.: particionamento do suporte. Partic. Adapt.: particionamento adaptativo do suporte. Map. Simb.: Mapeamento simbólico.

Medidas/Métodos	Particionamento do suporte	Ross
Informação Mútua	X	X
Informação Direcional		
Entropia de Transferência		

Tabela 1.3: Relação entre funcionais de medidas de informação e métodos utilizados para estimá-los: caso misto.

No caso discreto, o método *plug-in* foi estudado na análise de ácido desoxirribonucleico em [26]. O método de Jiao foi proposto em [34] e foi utilizado para analisar influências na economia, o método de Quinn foi proposto em [67] e foi utilizado para detectar conexões sinápticas entre neurônios. O método KDE foi utilizado em [73].

No caso contínuo, o método do particionamento do suporte foi extensivamente utilizado em neurociências [15, 68], no estudo das codificações entre estímulos físicos e registros eletrofisiológicos (estimação de informação mútua). Recentemente o método do particionamento do suporte foi aplicado para estimar a entropia de transferência entre as fases de registros de EEG [51]. O método do particionamento adaptativo do suporte foi proposto em [14], inicialmente para estimar informação mútua e posteriormente estendido para informação direcional [49] e entropia de transferência [48] por Liu *et al.*. O método KDE foi utilizado em neurociências, por Malladi *et al.* [54], para detectar zonas de focos

epilépticos no cérebro através da informação direcional, mas já é bem estabelecido que o método serve para estimar informação mútua [50], entropia de transferência [73], e até mesmo outras medidas de informação não abordadas aqui (como a entropia de Rényi [66]). O método do mapeamento simbólico foi proposto inicialmente para estimar entropia [8], sendo em seguida estendido para entropia de transferência [77] e para informação mútua [38]. O método KSG foi proposto Kraskov *et al.* [44], sendo estendido para entropia de transferência em aplicações de neurociências, dentre outras [84, 88]. Recentemente, Gao *et al.* propuseram modificações no método KSG, que levaram a um método similar, o método BI-KSG [19][20], que foi estendido por Murin *et al.* para estimar informação direcional [59].

1.1 Principais Contribuições

Nesta seção, apresentamos as principais contribuições desta tese:

1. Extensão dos estimadores de informação direcional de Jiao para processos assumindo valores reais;

2. Apresentação de estimadores de entropia de transferência para o caso misto;

3. Comparação entre os métodos de estimação de informação direcional de Jiao e de Quinn aplicados a trens de *spikes*, a fim de detectar sinapses químicas neuronais;

4. Comparação entre métodos de estimação de medidas de informação em termos de acurácia e tempo de execução.

O primeiro item investiga a aplicação de um dos estimadores de informação direcional para processos de alfabeto finito a processos assumindo valores contínuos, pela discretização prévia destes processos.

O segundo item não foi abordado na literatura revisada até o presente momento. O modelo de um canal de Poisson discreto no tempo utiliza um processo aleatório contínuo em amplitude como entrada e um processo discreto em amplitude como saída [28]. Além disso, a informação direcional dá um limitante mais justo à capacidade do canal com *feedback* [55]. A informação direcional está intimamente relacionada à entropia de transferência, como será visto no capítulo 2. Portanto, a entropia de transferência para o caso misto pode ser utilizada no contexto de um canal de Poisson com *feedback*, por exemplo.

O terceiro item compara dois métodos de estimação de informação direcional entre trens de *spikes*. Valores positivos de taxa de informação direcional normalizada entre trens de *spikes* detectaram, em média, todas as sinapses químicas de uma rede de neurônios simulada, com os dois estimadores de informação direcional estudados. Como apresentado no capítulo 9, a reconstrução de circuitos neuronais não é uma tarefa trivial com a tecnologia atual.

A investigação proposta aqui compara alguns dos diferentes estimadores para os funcionais de informação mútua, informação direcional e entropia de transferência. A comparação é feita em termos do enviesamento, da acurácia dos estimadores e do tempo de estimação em uma mesma máquina. Especificamente, as simulações expostas aqui foram executadas em um computador com um processador de 2.67GHz. As simulações realizadas foram no ambiente do *Matlab*. Para algumas simulações do capítulo 6, que

são indicadas ao longo do texto, o pacote JIDT foi utilizado (que por sua vez utiliza linguagem de programação *Java*).

1.2 Organização da Tese

O capítulo 2 apresenta as definições e algumas propriedades das medidas de informação consideradas (informação mútua, informação direcional e entropia de transferência). O capítulo 3 apresenta os métodos de estimação escolhidos para os casos em que as variáveis relacionadas sejam discretas, enquanto o capítulo 4 mostra os resultados de suas estimações. O capítulo 5 apresenta os métodos de estimação estudados para os casos em que as variáveis relacionadas sejam contínuas, enquanto o capítulo 6 mostra os resultados de suas estimações. O capítulo 7 é o que investiga a aplicação do estimador de informação direcional para processos com alfabeto finito a processos contínuos em amplitude. O capítulo 8 investiga estimadores de informação mútua para o caso misto e em particular sua aplicação para detecção de causalidade (estimação de entropia de transferência) no caso misto. O capítulo 9 apresenta alguns resultados no uso de estimadores com dados neuronais simulados, em particular trens de *spikes*. Finalmente, o capítulo 10 resume as principais conclusões obtidas ao longo desta tese. Além dos capítulos citados, o presente trabalho apresenta no apêndice A algumas definições de estatística relevantes ao problema de estimação, tais quais a inferência e a amostragem. Os apêndices B e C trazem as definições da função digamma e da distribuição de Dirichlet, que serão úteis ao longo da tese. Já os apêndices D e E apresentam alguns cálculos referentes à informação direcional e à entropia de transferência, que são utilizados nos capítulos 4 e 6, respectivamente. Por fim, o apêndice F traz as publicações oriundas deste doutorado.

Capítulo 2

Medidas de Informação

O presente capítulo descreve as medidas de informação que serão estudadas e estimadas ao longo desta tese. Elas servem para medir a dependência ou causalidade entre variáveis ou processos aleatórios. Como veremos mais adiante, o conceito de dependência difere da causalidade fundamentalmente pela exigência de uma ordenação na definição do segundo conceito (causalidade), ordenação esta que usualmente é feita pelos índices temporais. A independência entre variáveis aleatórias é medida segundo a distribuição conjunta de seus eventos. Seja A um evento associado à variável aleatória X apenas e B um evento associado à variável aleatória Y apenas. As duas variáveis aleatórias X e Y são ditas independentes se e somente se $P(A, B) = P(A)P(B)$, para todo A, B, em que P denota a probabilidade [85].

Uma medida bastante conhecida e utilizada de dependência entre variáveis aleatórias X e Y é a correlação de Pearson [46]. Quando calculada sobre uma população das variáveis aleatórias X e Y, a correlação de Pearson apresenta a seguinte forma:

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{cov}(X, X)\mathrm{cov}(Y, Y)}}. \tag{2.1}$$

Quando aplicada a uma amostra de X e Y, a correlação de Pearson apresenta a seguinte forma:

$$r = \frac{\sum(X_n - \bar{X})(Y_n - \bar{Y})}{[\sum(X_n - \bar{X})^2 \sum(Y_n - \bar{Y})^2]^{1/2}}, \tag{2.2}$$

em que (X_n, Y_n) é uma realização amostral das N realizações do par (X, Y).

Essa medida pode ser estendida para processos aleatórios considerando suas realizações nos tempos $t_1, t_2 \in I$, em que I é o conjunto infinito de índices dos processos. Sabe-se, no entanto, que a correlação é uma medida que captura apenas associações lineares entre as variáveis envolvidas [46, 47].

A dependência dirigida, isto é, causalidade, considerada nesta tese é baseada na fundamentação feita por Granger [24, 25]. A ideia da causalidade de Granger remonta às ideias de Hume (1739), que afirmou que a mente humana é incapaz de reconhecer relações causais [56]. Esta afirmação tem como base a crença de que o ser humano é capaz de observar apenas um evento de cada vez. Dessa forma, o ser humano seria incapaz de reconhecer as relações causais que ocorrem em um tempo contínuo. Para definir uma relação de causa e efeito, Hume propôs que

- A causa deve preceder o efeito temporalmente;

- A causa inclui informação sobre o efeito que não está disponível em uma grande variedade de outras variáveis.

A causalidade de Granger também foi inspirada por ideias de Wiener (1956) [56, 89]. Wiener considerou que para dois sinais $\mathbf{X}$ e $\mathbf{Y}$ medidos simultaneamente, o sinal $\mathbf{X}$ causa o sinal $\mathbf{Y}$ se for melhor predizer $\mathbf{Y}$ conhecendo informação sobre o passado de $\mathbf{X}$ do que sem tal conhecimento. Esta consideração estabelece o conceito de causalidade preditiva.

Assim, a definição de causalidade de Granger é baseada em três pressupostos [56]:

1. O passado e o presente podem causar o futuro, mas a recíproca não é verdadeira;

2. Ω (todo conhecimento disponível no universo no tempo t) não possui informações redundantes. Ou seja, se existe uma variável Z funcionalmente relacionada a outras variáveis, de maneira determinística, então Z deve ser excluída de Ω;

3. Todas as relações causais se mantêm constantes em direção durante o tempo.

Granger admite ter se inspirado nas ideias de Hume para os 2 primeiros pressupostos. O terceiro pressuposto foi adicionado para evitar confusões entre as definições de dependência e causalidade [56].

Apesar da validade teórica do conceito de causalidade de Granger como descrito pelos três pressupostos mencionados acima, há um problema prático na medição de causalidade devido ao número finito de séries temporais que podem ser consideradas quando se faz um cálculo. O termo numérico que define a causalidade de Granger foi estabelecido primeiro para ser utilizado em economia, porém tem sido bastante utilizado em diversas áreas para identificar relações causais entre dois processos [67]. Para quantificar a causalidade existente, são calculadas as variâncias dos termos de correções para modelos autorregressivos:

$$Y_n = \sum_{m=1}^{p} a_m Y_{n-m} + E_n, \tag{2.3}$$

$$Y_n = \sum_{m=1}^{p} (b_m Y_{n-m} + c_m X_{n-m}) + \tilde{E}_n, \tag{2.4}$$

em que E_n e $\tilde{E}_n$ são termos de correção de erros. Na referência [67], a causalidade de Granger é definida como uma medida específica:

$$G_{\mathbf{X}\to\mathbf{Y}} = \log \frac{\mathrm{var}(\mathbf{E})}{\mathrm{var}(\tilde{\mathbf{E}})}, \tag{2.5}$$

a qual compara as variâncias dos modelos quando o passado de $\mathbf{X}$ é incluso na análise de $\mathbf{Y}$ ou não. Nesse desenvolvimento consideram-se $\mathbf{E}$ e $\tilde{\mathbf{E}}$ processos estacionários. A razão da equação (2.5) é denominada *ratio gap* na literatura [61].

Dadas as informações acima sobre a maneira com que a causalidade de Granger é comumente calculada, observa-se que uma de suas limitações é requerer que a interação entre os dois processos observados seja linear. Essa é uma limitação análoga ao caso da medição de dependência através da correlação.

Além disso, a literatura reporta uma série de problemas quando utilizando testes de Granger para detectar causalidade [56]. Ocorre que o uso dos diferentes testes, a arbitrariedade de escolha nos coeficientes de regressão, efeitos de linearização/transformação

dos dados e mesmo a escolha do tempo de amostragem dos sinais podem resultar em diferentes detecções de relações de causalidade (para os mesmos dados).

Outra questão crítica concerne aos conceitos de causa comum ou causa em cascata. A causa comum ocorre quando há um sinal $\mathbf{X}$ que influencia outros dois sinais $\mathbf{Y}$ e $\mathbf{Z}$. A causalidade preditiva, abordando apenas duplas de sinais, pode detectar uma causalidade espúria entre os sinais $\mathbf{Y}$ e $\mathbf{Z}$, especialmente se $\mathbf{X}$ causa $\mathbf{Y}$ e $\mathbf{Z}$ com diferentes relações temporais. Por outro lado, a causa em cascata ocorre quando o sinal $\mathbf{X}$ influencia o sinal $\mathbf{Y}$ que por sua vez influencia o sinal $\mathbf{Z}$. Nesta situação, a causalidade preditiva detecta uma relação causal entre $\mathbf{X}$ e $\mathbf{Z}$, muito embora a relação seja indireta.

O espaço $\mathbf{V}$ sobre o qual as observações de causalidade são feitas também tem fundamental importância na determinação da causalidade de um processo aleatório sobre outro. Um exemplo de causa em cascata pode esclarecer esse ponto. Sejam os processos aleatórios $\mathbf{X}$, $\mathbf{Y}$ e $\mathbf{Z}$ definidos como:

$$Y_{n+1} = X_n + \epsilon_{n+1} \tag{2.6}$$
$$Z_{n+1} = Y_n + \eta_{n+1}, \tag{2.7}$$

em que ϵ e η são sequências independentemente e identicamente distribuídas (i.i.d.), independentes uma da outra e $n = 1, 2, \ldots$. Considerando o espaço $\mathbf{V} = \{\mathbf{X}, \mathbf{Z}\}$, verificamos que $P(Z_{n+1}|Z^n X^n) \neq P(Z_{n+1}|Z^n)$, logo, neste contexto $\mathbf{X}$ causa $\mathbf{Z}$. Contudo, considerando o espaço $\mathbf{V} = \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$, $P(Z_{n+1}|Z^n Y^n X^n) = P(Z_{n+1}|Z^n Y^n)$, e neste novo contexto $\mathbf{X}$ não causa $\mathbf{Z}$ [5].

Um exemplo do cotidiano ilustra uma aplicação do conceito de causalidade preditiva levando a detecções errôneas de causalidade. Considere o número S de sorvetes vendidos em uma praia em um dia. Considere também o número A de mortes por afogamento nesta mesma praia em um dia. Em dias de sol, que poderia ser a causa comum do aumento de ambas variáveis S e A, poderia-se chegar à conclusão de que o aumento na quantidade S influenciaria causalmente a quantidade A. Ou seja, chegaria-se à conclusão errada de que o aumento do consumo de sorvete estaria causando mortes por afogamento, quando a causa real seria o clima ensolarado aumentando ambas as taxas de consumo de sorvete e de afogamentos.

Portanto, quando utiliza-se o conceito de causalidade preditiva, como no presente trabalho, é interessante considerar processos aleatórios que tenham alguma relação provável de causalidade. Ou seja, é importante que a pesquisa feita considere os mecanismos físicos que relacionam os sinais envolvidos a fim de detectar relações verdadeiras. O simples cálculo matemático preditivo sem ponderações teóricas pode trazer conclusões inválidas.

2.1 Informação Mútua

Já é um fato bem conhecido que a informação mútua mede a dependência geral entre variáveis aleatórias e, portanto, é mais precisa que a correlação para medir dependências [47]. A informação mútua é uma medida confiável da dependência entre variáveis aleatórias, medindo a quantidade de incerteza que é reduzida de uma variável aleatória pelo conhecimento do resultado de outra variável aleatória [13], como por exemplo entre as

variáveis aleatórias discretas X e Y, sendo dada pela equação (2.8):

$$\begin{aligned} I(X;Y) &= \sum_{\mathcal{X},\mathcal{Y}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} && (2.8)\\ &= \mathrm{D}(P(X,Y)||P(X)P(Y)) && (2.9)\\ &= \mathbb{E}\left(\log\frac{P(X,Y)}{P(X)P(Y)}\right) \\ &= H(X) - H(X|Y) && (2.10)\end{aligned}$$

em que D(.||.) denota a distância de Kullback-Leibler.

Na equação (2.10), $H(X)$ é a entropia de Shannon da variável aleatória X, medindo sua incerteza, ao passo que $H(X|Y)$ é a entropia, ou incerteza, da variável aleatória X quando a variável aleatória Y é conhecida. Ambas grandezas são definidas como:

$$\begin{aligned} H(X) &= -\sum_{x\in\mathcal{X}} P(x)\log P(x), \\ H(X|Y) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log P(x|y). \end{aligned}$$

A informação mútua também pode ser aplicada a mais de duas variáveis aleatórias através da regra da cadeia, conforme equação (2.12):

$$\begin{aligned} I(X^N;Y^N) &= H(Y^N) - H(Y^N|X^N) && (2.11)\\ &= \sum_{n=1}^{N} I(Y_n;X^N|Y^{n-1}). && (2.12)\end{aligned}$$

Os termos em (2.11) são calculados como:

$$\begin{aligned} H(Y^N) &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}), \\ H(Y^N|X^N) &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}X^N), \end{aligned}$$

em que $H(Y^N)$ é a entropia conjunta, ou incerteza, das variáveis aleatórias $Y_1,\ldots,Y_N$, e $H(Y^N|X^N)$ é a entropia condicional de Y^N em relação a X^N. De maneira análoga, $H(Y^n|Y^{n-1})$ e $H(Y_n|Y^{n-1}X^N)$ são as entropias condicionadas de Y_n em relação a Y^{n-1} e em relação a Y^{n-1} e X^N, respectivamente.

Na equação (2.12), a informação mútua condicional entre as variáveis Y_n, X^N e Y^{n-1} é calculada como:

$$I(Y_n;X^N|Y^{n-1}) = \mathbb{E}_{P(Y^n,X^N)}\left(\log\frac{P(Y_n,X^N|Y^{n-1})}{P(Y_n|Y^{n-1})P(X^N|Y^{n-1})}\right). \quad (2.13)$$

Para o caso de variáveis aleatórias contínuas, a informação mútua é definida como:

$$I(X;Y) = \int f(x,y)\ln\frac{f(x,y)}{f(x)f(y)}dxdy. \quad (2.14)$$

Analogamente, para o caso de variáveis aleatórias mistas, isto é, uma variável aleatória discreta e outra contínua, a informação mútua é definida como:

$$I(X;Y) = \sum_{x\in\mathcal{X}} \int f(x,y) \ln \frac{f(x,y)}{P(x)f(y)} dy \tag{2.15}$$

$$= \sum_{x\in\mathcal{X}} P(x) \int f(y|x) \ln \frac{f(y|x)}{f(y)} dy. \tag{2.16}$$

2.1.1 Taxas de Informação Mútua

Quando se fala em processos aleatórios, recomenda-se fazer medidas de entropia e informação mútua em relação às taxas que esses processos assumem. Isso decorre do fato de que processos aleatórios são sequências de variáveis aleatórias, sequências estas indexadas e por vezes infinitas [85]. Assim, faz mais sentido calcular a taxa do crescimento da entropia ou da informação mútua. Para uma sequência finita de tamanho N, define-se a taxa de entropia como:

$$H_N(X) = \frac{1}{N} H(X^N). \tag{2.17}$$

A entropia de um processo aleatório $\mathbf{X}$ na realidade é definida como o limite de sua taxa de entropia, se o limite existir [13]:

$$H_\infty(X) = \lim_{N\to\infty} H_N(X). \tag{2.18}$$

Defina o limite

$$H'_\infty(X) = \lim_{N\to\infty} H(X_N|X^{N-1}), \tag{2.19}$$

e admita sua existência. Note que a quantidade (2.18) representa a entropia por símbolo de uma sequência de N variáveis aleatórias e a quantidade (2.19) representa a entropia da última variável da sequência condicionada ao seu passado.

O Teorema (4.2.1), p. 75, em [13] assegura que se o processo aleatório for estacionário então

$$H_\infty(X) = H'_\infty(X). \tag{2.20}$$

É conveniente observar que a estacionariedade é *condição suficiente* mas não necessária para a igualdade (2.20), ou seja, é possível que processos aleatórios não estacionários obedeçam (2.20). Um exemplo ilustrativo interessante é apresentado na solução do exercício 4.12, p.93, em [13].

De maneira análoga, computam-se a taxa de informação mútua entre os processos aleatórios $\mathbf{X}$ e $\mathbf{Y}$ e seu limite quando $N \to \infty$:

$$I_N(X;Y) = \frac{1}{N} I(X^N;Y^N), \tag{2.21}$$

$$I_\infty(X;Y) = \lim_{N\to\infty} I_N(X;Y). \tag{2.22}$$

2.1.2 Propriedades da Informação Mútua

As propriedades da informação mútua são as mesmas, tanto para o caso discreto como para o caso contínuo [13]. Algumas delas são:

Propriedade 1. $I(X;Y) \geq 0$.

Propriedade 2. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X)$.

A primeira das propriedades mostradas é a de não-negatividade. Por ser uma distância de Kullback-Leibler, $I(X;Y)$ é sempre maior ou igual a zero, com igualdade se e somente se $P(x,y) = P(x)P(y)$, no caso discreto, ou $f(x,y) = f(x)f(y)$, no caso contínuo, isto é, se e somente se X e Y forem variáveis independentes [13]. A segunda das propriedades refere-se ao fato da informação mútua ser uma grandeza simétrica.

Para o caso discreto, pode ser mostrada ainda uma caracterização adicional acerca da informação mútua: ela sempre é limitada pela menor das entropias das variáveis envolvidas. Isto é, $I(X;Y) \leq \min\{H(X), H(Y)\}$.

2.2 Informação Direcional

Em termos de causalidade, a informação direcional é baseada no mesmo princípio de Granger — a quantidade em que o processo aleatório **X** causa o processo aleatório **Y** é medida segundo quão mais fácil é prever o valor de **Y** baseado no passado de **X** e no passado de **Y** conjuntamente [67]. Contudo, diferentemente da causalidade de Granger, essa medição não pressupõe nenhum modelo estatístico específico. Se os processos aleatórios forem discretos, é possível interpretar a informação direcional como a redução do número mínimo de bits requerido para especificar uma fonte **Y** dado o conhecimento causal de **X** [67].

Similarmente à informação mútua, e denotada por $I(X^N \to Y^N)$, a informação direcional de X^N para Y^N é definida como [67, 42]:

$$\begin{aligned} I(X^N \to Y^N) &= H(Y^N) - H(Y^N||X^N) && (2.23)\\ &= \sum_{n=1}^{N} I(X^n; Y_n|Y^{n-1}) && (2.24)\\ &= \sum_{n=1}^{N} \mathbb{E}\left(\log \frac{P(Y_n|Y^{n-1}X^n)}{P(Y_n|Y^{n-1})}\right) \end{aligned}$$

em que o termo da entropia causalmente condicionada é:

$$H(Y^N||X^N) = \sum_{n=1}^{N} H(Y_n|Y^{n-1}X^n). \quad (2.25)$$

A informação direcional $I(X^N \to Y^N||Z^N)$ fluindo de X^N para Y^N quando causalmente condicionada à sequência Z^N é definida como [42]:

$$\begin{aligned} I(X^N \to Y^N||Z^N) &= H(Y^N||Z^N) - H(Y^N||X^N Z^N) && (2.26)\\ &= \sum_{n=1}^{N} I(X^n; Y_n|Y^{n-1}Z^n). && (2.27) \end{aligned}$$

2.2.1 Taxas de Informação Direcional

Em muitas aplicações, a informação direcional aumenta linearmente com N. Um parâmetro importante é a taxa deste crescimento, i.e., a taxa de informação direcional ou informação direcional por letra. Para este propósito, assim como na seção anterior, definem-se as seguintes taxas de entropia causalmente condicionada e taxas de informação direcional [42]:

$$\begin{aligned} H_N(X||Y) &= \frac{1}{N}H(X^N||Y^N), \\ I_N(X \to Y) &= \frac{1}{N}I(X^N \to Y^N), \\ I_N(X \to Y||Z) &= \frac{1}{N}I(X^N \to Y^N||Z^N). \end{aligned} \tag{2.28}$$

Na maioria dos casos de interesse, os termos em (2.28) têm limites quando N tende ao infinito. Denotam-se tais limites por [42]:

$$\begin{aligned} H_\infty(X||Y) &= \lim_{N\to\infty}\frac{1}{N}H(X^N||Y^N), \\ I_\infty(X \to Y) &= \lim_{N\to\infty}\frac{1}{N}I(X^N \to Y^N), \\ I_\infty(X \to Y||Z) &= \lim_{N\to\infty}\frac{1}{N}I(X^N \to Y^N||Z^N). \end{aligned} \tag{2.29}$$

Quando $\mathbf{X}$ e $\mathbf{Y}$ são processos estacionários, a taxa de informação direcional pode ser escrita como [48, 6]

$$I_\infty(X \to Y) = \lim_{N\to\infty} I(X^N; Y_N|Y^{N-1}) \tag{2.30}$$

$$= \lim_{N\to\infty}(H(Y_N|Y^{N-1}) - H(Y_N|Y^{N-1}X^N)). \tag{2.31}$$

2.2.2 Propriedades da Informação Direcional

Foram derivadas algumas propriedades da informação direcional que a relacionam com a informação mútua, para o caso de variáveis assumindo valores discretos [42, 34]:

Propriedade 3. $0 \leq I(X^N \to Y^N) \leq I(X^N; Y^N)$*, com igualdade na esquerda se e somente se* $I(X^n; Y_n|Y^{n-1}) = 0$ *para todo* $n = 1, 2, \ldots, N$*, e com igualdade na direita se e somente se* $H(Y_n|Y^{n-1}X^n) = H(Y_n|Y^{n-1}X^N)$ *para todo* $n = 1, 2, \ldots, N$*.*

Propriedade 4. $I(X^N; Y^N) = I(X^N \to Y^N) + I(DY^N \to X^N)$*, em que* D *denota* $DY^N = (0, Y_1, Y_2, \ldots, Y_{N-1})$*.*

2.3 Comparação entre Informação Mútua e Informação Direcional

Agora, retomando a comparação entre informação mútua e informação direcional, reescrevendo as equações (2.11) e (2.23), percebe-se a sutil diferença entre elas: apenas

o sobrescrito N ou n na variável X, dentro do somatório.

$$\begin{aligned} I(X^N;Y^N) &= \sum_{n=1}^{N}[H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^N)], \\ I(X^N \to Y^N) &= \sum_{n=1}^{N}[H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^n)]. \end{aligned} \tag{2.32}$$

Esta sutil diferença tem grande impacto no sentido das duas medidas, significando que, no caso da informação direcional, a entropia do processo $\mathbf{Y}$ está sendo condicionada apenas a valores síncronos ou anteriores do processo $\mathbf{X}$, revelando aí a ideia de causalidade.

Um simples exemplo ilustra a diferença entre informação mútua e informação direcional e revela como esta última traz a noção de causalidade. Considere uma cadeia de Markov discreta e binária $\mathbf{X}$ como mostrada no diagrama de estados da Fig. 2.1, e o processo $\mathbf{Y}$, tal que X_n começa com $n = 1$ e $Y^N = DX^N$. Ou seja, Y^N é o atraso de X^N por uma unidade de tempo (com o descarte da última componente). $Y_n = X_{n-1}$, logo $\mathbf{Y}$ é causalmente condicionado por $\mathbf{X}$. A taxa de informação mútua é calculada como:

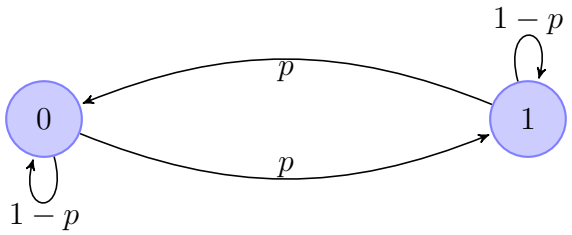


Figura 2.1: Diagrama de estados para processo $\mathbf{X}$.

$$\begin{aligned} I_\infty(X;Y) &= \lim_{N\to\infty}\frac{1}{N}\left(\sum_{n=1}^{N} H(Y^N) - H(Y^N|X^N)\right) \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=2}^{N}\Big(H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^N)\Big) && (2.33) \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=2}^{N}\Big(H(Y_n|Y_{n-1}) - H(Y_n|Y^{n-1}X^N)\Big) && (2.34) \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=2}^{N}(H(Y_n|Y_{n-1}) - H(Y_n|X_{n-1})) && (2.35) \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=2}^{N}\mathcal{H}(p) \\ &= \lim_{N\to\infty}\frac{N-1}{N}\mathcal{H}(p) \\ &= \mathcal{H}(p). \end{aligned}$$

Na equação (2.33) consideramos que as sequências $\mathbf{X}$ e $\mathbf{Y}$ começam em $n = 1$, de modo que o primeiro termo do somatório é $H(Y_1) - H(Y_1|X_1^N) = H(Y_1) - H(Y_1) = 0$. Além disso, (2.34) advém do fato de $\mathbf{Y}$ também ser uma cadeia de Markov e (2.35) advém do fato de que $Y_n = X_{n-1}$.

Por outro lado, é possível calcular a taxa de informação direcional de $\mathbf{X}$ para $\mathbf{Y}$ e de $\mathbf{Y}$ para $\mathbf{X}$ e observar a diferença:

$$\begin{aligned}
I_\infty(X \to Y) &= \lim_{N\to\infty} \frac{1}{N}\Big(H(Y^N) - H(Y^N||X^N)\Big) \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{n=2}^{N}\Big(H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^n)\Big) \qquad (2.36)\\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{n=2}^{N}\left(H(Y_n|Y_{n-1}) - H(Y_n|X_{n-1})\right) \qquad (2.37)\\
&= \mathcal{H}(p) \\
I_N(Y \to X) &= \lim_{N\to\infty} \frac{1}{N}\Big(H(X^N) - H(X^N||Y^N)\Big) \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{n=2}^{N}\Big(H(X_n|X^{n-1}) - H(X_n|X^{n-1}Y^n)\Big) \qquad (2.38)\\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{n=2}^{N}\left(H(X_n|X_{n-1}) - H(X_n|X_{n-1})\right) \qquad (2.39)\\
&= 0.
\end{aligned}$$

em que (2.36) e (2.38) advém da definição, (2.37) advém do fato de que $Y_n = X_{n-1}$ e (2.39) advém do fato de $\mathbf{X}$ formar uma cadeia de Markov.

Está claro neste exemplo que $I_N(X \to Y)$ mede a taxa de informação em bits de $\mathbf{Y}$ que é causalmente dada por $\mathbf{X}$ e que $I_N(Y \to X)$ dá o total de informação na direção reversa, neste caso, zero.

2.4 Entropia de Transferência

A entropia de transferência foi introduzida por Schreiber [73], a fim de quantificar a coerência estatística entre sistemas no tempo, capturando a dinâmica da transferência de informação. A entropia de transferência é definida como:

$$\begin{aligned}
TE_n(X \to Y) &= \sum_{y_n, y_{n-m}^{n-1}, x_{n-l}^{n-1}} P(y_n, y_{n-m}^{n-1}, x_{n-l}^{n-1}) \log \frac{P(y_n|y_{n-m}^{n-1}, x_{n-l}^{n-1})}{P(y_n|y_{n-m}^{n-1})} \qquad (2.40)\\
&= I(Y_n; X_{n-l}^{n-1}|Y_{n-m}^{n-1}) \\
&= I(Y_n, Y_{n-m}^{n-1}; X_{n-l}^{n-1}) - I(Y_{n-m}^{n-1}; X_{n-l}^{n-1}). \qquad (2.41)
\end{aligned}$$

Observa-se que a entropia de transferência é a distância de Kullback-Leibler entre duas distribuições: $P(Y_n|Y_{n-m}^{n-1}, X_{n-l}^{n-1})$ e $P(Y_n|Y_{n-m}^{n-1})$. Portanto, a entropia de transferência é uma distância entre a distribuição de Y_n conhecendo l valores passados de $\mathbf{X}$ e m valores passados de $\mathbf{Y}$ e a distribuição de Y_n conhecendo apenas m valores passados de $\mathbf{Y}$. Em outras palavras, a entropia de transferência mede o desvio da seguinte suposição da propriedade de Markov generalizada:

$$P(Y_n|Y_{n-m}^{n-1}) = P(Y_n|Y_{n-m}^{n-1}, X_{n-l}^{n-1}). \qquad (2.42)$$

Geralmente a entropia de transferência é aplicada a processos conjuntamente estacionários, de maneira que a dependência com o tempo n em que é feita a estimação não é considerada.

A entropia de transferência também pode ser escrita em função de entropias convencionais:

$$TE_n(X \to Y) \;=\; H(Y_n|Y_{n-m}^{n-1}) - H(Y_n|Y_{n-m}^{n-1}X_{n-l}^{n-1}). \tag{2.43}$$

Assim como nos casos da entropia convencional, da informação mútua e da informação direcional, pode ser interessante considerar o limite da entropia de transferência quando a duração n tende ao infinito, ou seja:

$$TE_\infty(X \to Y) = \lim_{n\to\infty} TE_n(X \to Y). \tag{2.44}$$

Por ser um distância de Kullback-Leibler, a entropia de transferência não constitui uma métrica verdadeira, no sentido de não ser simétrica nem satisfazer a desigualdade triangular [13]. Esta é uma propriedade genérica das divergências estatísticas. Todas elas são não-simétricas mas são uma distância no espaço de Hilbert (espaço de Hilbert é um espaço vetorial normado completo com um produto interno, maior detalhamento na referênca [58]).

A entropia de transferência apresenta a propriedade de ser não negativa:

Propriedade 5. $TE_n(X \to Y) \geq 0$.

2.4.1 Relação entre Entropia de Transferência e Informação Direcional

Na referência [48], a relação entre entropia de transferência e informação direcional é explorada. A informação direcional entre dois processos $\mathbf{X}$ e $\mathbf{Y}$ pode ser decomposta em [6]:

$$\begin{aligned} I(X^N \to Y^N) &= I(DX^N \to Y^N) + I(X^N \to Y^N||DX^N) && (2.45)\\ &= \sum_{n=1}^{N} I(Y_n; X^{n-1}|Y^{n-1}) + \sum_{n=1}^{N} I(Y_n; X_n|Y^{n-1}X^{n-1}). && (2.46) \end{aligned}$$

Analogamente, para processos estacionários, a taxa de informação direcional pode ser decomposta em

$$\begin{aligned} I_\infty(X \to Y) &= I_\infty(DX \to Y) + I_\infty(X \to Y||DX)\\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \Big(I(X^{N-1} \to Y^N) + I(X^N \to Y^N||X^{N-1}) \Big)\\ &= \lim_{N\to\infty} I(X^{N-1}; Y_N|Y^{N-1}) + \lim_{N\to\infty} I(X_N; Y_N|X^{N-1}Y^{N-1}). && (2.47) \end{aligned}$$

O primeiro termo da equação (2.47) é a taxa de informação direcional do processo $\mathbf{X}$ atrasado em uma unidade para o processo $\mathbf{Y}$. Já o segundo termo da equação (2.47) é a transferência instantânea de informação de $\mathbf{X}$ para $\mathbf{Y}$, condicionados nos valores passados dos dois processos.

Liu e Aviyente provaram que, se

- os processos $\mathbf{X}$ e $\mathbf{Y}$ são estacionários,
- a transferência instantânea de informação entre $\mathbf{X}$ e $\mathbf{Y}$ é nula e

- $P(Y_n|Y_1^{n-1}, X_1^{n-1}) = P(Y_n|Y_{n-m}^{n-1}, X_{n-l}^{n-1})$,

o limitante superior da taxa de informação direcional é igual à entropia de transferência [48].

Convém ressaltar aqui que as definições tanto de informação direcional como de entropia de transferência são estendidas para o caso de variáveis aleatórias assumindo valores contínuos [54, 73]. O capítulo seguinte descreve métodos para estimar medidas de informação entre variáveis assumindo valores discretos.

Capítulo 3

Métodos de Estimação de Medidas de Informação entre Variáveis Aleatórias Discretas

Conforme já visto no capítulo 2, as medidas de informação investigadas neste trabalho são definidas em função das distribuições de probabilidade das variáveis ou processos aleatórios envolvidos. Estas medidas permitem quantificar o grau de associação entre as variáveis, seja ele relativo à dependência ou à causalidade. Contudo, muitas vezes estas distribuições de probabilidade não estão disponíveis a priori, de maneira que torna-se necessário estimar as medidas através de realizações amostrais. Neste capítulo investigam-se estimadores de medidas de informação para o caso em que as variáveis envolvidas são discretas.

3.1 Estimadores *Plug-in*

Para o caso de variáveis aleatórias discretas, o método mais direto de se estimar medidas de informação, em particular a informação mútua, é chamado de *plug-in*. O método *plug-in* consiste em computar as frequências relativas dos eventos e atribuir estas frequências às distribuições de probabilidades, como pode ser visto no apêndice A. A partir daí os funcionais destas medidas podem ser calculados. Todavia, por serem baseadas em um número finito de realizações de eventos, as frequências relativas em geral não correspondem às probabilidades verdadeiras, promovendo um viés nas estimativas de medidas de informação.

Devido à existência deste viés, surgiram diversas técnicas para corrigi-lo. Dentre as primeiras correções de erro de estimação cita-se o trabalho de Miller, que encontrou aproximações numéricas para o viés b da informação mútua em um regime de amostragem assintótica, utilizando uma expansão de Taylor [29, 31]. De maneira não rigorosa, diz-se que no regime de amostragem assintótica o número de amostras é grande. Miller mostrou que, neste regime, o viés de uma estimativa *plug-in* é inversamente proporcional ao tamanho amostral N, sendo aproximado pela expressão:

$$b = \frac{1}{2N \ln 2} \left\{ \sum_x [\tilde{y}_x - 1] - (\tilde{y} - 1) \right\}. \tag{3.1}$$

Na expressão (3.1), $\tilde{y}_x$ denota o número de valores distintos que Y assume para um

determinado valor de $X = x$ tal que $P(Y|x) > 0$, ao passo que $\tilde{y}$ denota o número de valores distintos que Y assume tal que $P(Y) > 0$ [31].

Outra técnica de correção de viés que pode ser citada e que será explorada aqui é o da extrapolação quadrática (QE) [79], que também pode ser aplicada na estimação de outros funcionais (como a entropia de Shannon, por exemplo). Esta técnica divide o conjunto de amostras em partições com metade e um quarto do tamanho populacional original [31]. Para estes subgrupos, a informação mútua é estimada segundo o método *plug-in*, sendo feita em seguida a média dos dois e quatro valores estimados. Os pares $(N, \hat{I}_1(X;Y))$, $(N/2, \hat{I}_2(X;Y))$ e $(N/4, \hat{I}_4(X;Y))$ são usados de forma a ajustar os parâmetros $\hat{I}'$, a e b da curva descrita nas equações (3.2), (3.3) e (3.4) abaixo [15]:

$$\hat{I}_1(X;Y) = \hat{I}' + \frac{a}{N} + \frac{b}{N^2}, \tag{3.2}$$

$$\hat{I}_2(X;Y) = \hat{I}' + \frac{a}{(N/2)} + \frac{b}{(N/2)^2}, \tag{3.3}$$

$$\hat{I}_4(X;Y) = \hat{I}' + \frac{a}{(N/4)} + \frac{b}{(N/4)^2}, \tag{3.4}$$

que são equivalentes ao sistema de equações

$$\begin{bmatrix} N^2 & 4N & 16 \\ N^2 & 2N & 4 \\ N^2 & N & 1 \end{bmatrix} \begin{bmatrix} \hat{I}' \\ a \\ b \end{bmatrix} = \begin{bmatrix} N^2\hat{I}_4 \\ N^2\hat{I}_2 \\ N^2\hat{I}_1 \end{bmatrix}. \tag{3.5}$$

Neste caso, desenhando a curva $\hat{I}$ em função de N, quando N tende a ∞ (ou $1/N \to 0$), a curva de estimativa $\hat{I}$ tende ao seu valor analítico. A ideia da correção QE é aproximar o valor verdadeiro de $I(X;Y)$ pelo parâmetro $\hat{I}'$ da curva. A Fig. 3.1 ilustra a aplicação do método QE sobre uma estimativa *plug-in* de informação mútua, em que a estimativa *plug-in* QE está mais próxima do valor analítico que a estimativa *plug-in* sem correção QE.

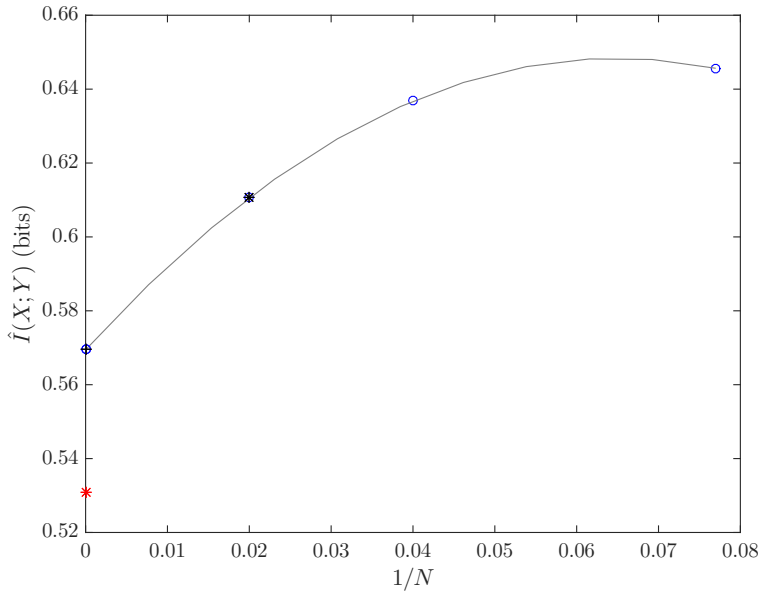


Figura 3.1: Ilustração da correção QE sobre estimativa *plug-in* de informação mútua, em que o valor analítico era $I(X,Y) = 0.5310$ bits e o tamanho amostral era $N = 50$. O asterisco preto revela a estimativa *plug-in* sem correção, o valor analítico está indicado no asterisco vermelho e a intersecção da curva com o eixo $1/N = 0$ indica a estimativa com correção QE para a amostra obtida, $\hat{I}'$.

3.2 Estimadores para Informação Direcional

Introduzem-se aqui duas maneiras encontradas na literatura de estimar informação direcional a partir dos dados. Ambas consideram que os processos aleatórios que geraram os dados são estacionários ergódicos de alfabeto finito, logo são processos assumindo valores discretos. O primeiro método de estimação introduzido aqui é baseado em métodos universais para processos arbitrários estacionários e ergódicos de alfabeto finito. Por outro lado, o segundo método de estimação foi desenvolvido para inferir causalidade em registros de trens de *spikes* neurais, pressupondo um modelo generalizado linear e paramétrico.

3.2.1 Estimador de Jiao

Jiao *et al.* propuseram quatro estimadores similares de informação direcional em [34], todos quatro utilizam as probabilidades ponderadas segundo o algoritmo CTW (*context tree weighting*). O CTW considera uma fonte cuja saída depende até certa profundidade da sequência emitida anteriormente e para compreendê-lo algumas definições são relevantes.

Fontes em Árvore

Primeiramente, pode-se afirmar que uma fonte de informação sobre um alfabeto finito $\mathcal{X}$ é definida por uma família de medidas de probabilidade:

$$P_n(x^n),\ n = 0, 1, \ldots,$$

cada P_n definido sobre o conjunto $\mathcal{X}^n$ de todas as sequências de comprimento n, tais que a condição marginal:

$$\sum_{x\in\mathcal{X}} P_{n+1}(x^n x) = P_n(x^n). \tag{3.6}$$

é mantida para todo n [87].

No que se refere às sequências, convém lembrar que sequências no alfabeto finito $\mathcal{X}$ são definidas como $s = q_{1-l}q_{2-l}\dots q_0$, $q \in \mathcal{X}$, e que a contatenação de sequências é feita como $s's \triangleq q'_{1-l'}q'_{2-l'}\dots q'_0 q_{1-l}q_{2-l}\dots q_0$. Além disso, uma sequência $s = q_{1-l}q_{2-l}\dots q_0$ é sufixo de outra sequência $s' = q'_{1-l'}q'_{2-l'}\dots q'_0$ se $l \leq l'$ e $q_{-i} = q'_{-i}$ para $i = 0, \dots, l-1$ e a sequência vazia λ é o sufixo de todas as sequências.

O comportamento estatístico de uma fonte em árvore binária de memória finita pode ser descrito por seu conjunto de sufixos $\mathcal{S}$. O conjunto $\mathcal{S}$ deve ser próprio e completo, isto é, nenhuma sequência em $\mathcal{S}$ deve ser sufixo de outra e cada sequência semi-infinita de símbolos deve ter um sufixo pertencente a $\mathcal{S}$, respectivamente. A fonte em árvore é limitada e tem memória não maior que D se cada sufixo em $\mathcal{S}$ for menor ou igual a D.

Para simplificar, considera-se o caso binário. A cada sufixo em $\mathcal{S}$ corresponde um parâmetro $\theta_{\mathcal{S}}$, que, para o caso de fonte binária, é a probabilidade do próximo símbolo ser 1, dada que sequência emitida foi $s \in \mathcal{S}$. As probabilidades de próximo símbolo para uma fonte em árvore binária de memória finita com conjunto de sufixos $\mathcal{S}$ e vetor de parâmetros $\Theta_{\mathcal{S}}$ são

$$P_a\Big(X_n = 1|x_{n-D}^{n-1}, \mathcal{S}, \Theta_{\mathcal{S}}\Big) = \theta_{\beta_{\mathcal{S}}\left(x_{n-D}^{n-1}\right)}, \quad \text{para todo } \ n. \tag{3.7}$$

Na equação (3.7), a função de sufixo $s = \beta_{\mathcal{S}}\Big(x_{n-D}^{n-1}\Big)$ mapeia sequências semi-infinitas $x_{-\infty}^{n-1}$ em um sufixo único, utilizando para tanto os últimos D símbolos, já que esta é a largura máxima dos sufixos.

As probabilidades de bloco são produtos das probabilidades de próximo símbolo:

$$P_a\Big(X_1^n = x_1^n|x_{1-D}^0, \mathcal{S}, \Theta_{\mathcal{S}}\Big) = \prod_{\tau=1}^{n} P_a\left(X_\tau = x_\tau|x_{\tau-D}^{\tau-1}, \mathcal{S}, \Theta_{\mathcal{S}}\right). \tag{3.8}$$

Um exemplo retirado da referência [90] e cuja estrutura encontra-se ilustrada na Fig. 3.2 esclarece as noções de fonte em árvore binária expostas aqui. Seja $D = 3$. Considere a fonte com $\mathcal{S} = \{00, 10, 1\}$ e parâmetros $\theta_{00} = 0.5$, $\theta_{10} = 0.3$ e $\theta_1 = 0.1$. A probabilidade (condicional) da fonte emitir 0110100 dados os símbolos passados $\dots 010$ pode ser calculada como:

$$P_a(0110100|\dots 010) = (1-\theta_{10})\theta_{00}\theta_1(1-\theta_1)\theta_{10}(1-\theta_1)(1-\theta_{10}) = 0.0059535. \tag{3.9}$$

Codificação para uma Fonte em Árvore Desconhecida

O uso do CTW pode servir para comprimir uma sequência gerada por uma fonte em árvore com parâmetros $\theta_{\mathcal{S}}$ e conjunto de sufixos $\mathcal{S}$ desconhecidos pelo codificador e decodificador. A árvore de contexto $\mathcal{T}_D$ é um conjunto de nós etiquetados s, em que s é uma sequência com comprimento $l(s)$ tal que $0 \leq l(s) \leq D$, que para o caso binário, se divide em dois nós $0s$ e $1s$. O nó s é chamado de pai dos nós $0s$ e $1s$, os quais são

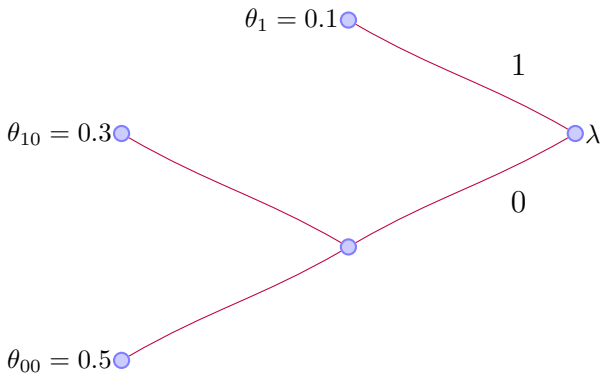


Figura 3.2: Exemplo de fonte em árvore binária genérica.

chamados de nós filhos de s. Para cada nó $s \in \mathcal{T}_D$, há contagens $a_s \geq 0$ e $b_s \geq 0$, para o número de zeros e uns que precedem a sequência s, respectivamente. Para os filhos $0s$ e $1s$ do nó pai s, as contagens devem satisfazer $a_{0s} + a_{1s} = a_s$ e $b_{0s} + b_{1s} = b_s$ [90].

Dito isto, para entender o algoritmo torna-se interessante um exemplo, como o da Fig. 3.3, para uma árvore de contexto binária de profundidade $D = 2$, de uma sequência de tamanho $n = 6$: $x_{-1}x_0x_1 \ldots x_6 = 00100110$ A árvore começa na raiz que conta o número de ocorrências de cada símbolo em uma sequência $x_1x_2 \ldots x_n$, começando pelo índice 1. A cada nó da árvore, a contagem continua, só que agora observando quantas vezes o contexto indicado seguindo os ramos da árvore precede o símbolo que está sendo contado. Por exemplo, a contagem de quantas vezes o contexto 11 precede o 0 e o 1 na sequência dada está na folha superior da árvore, resultando nos valores $a_{11} = 1$ e $b_{11} = 0$.

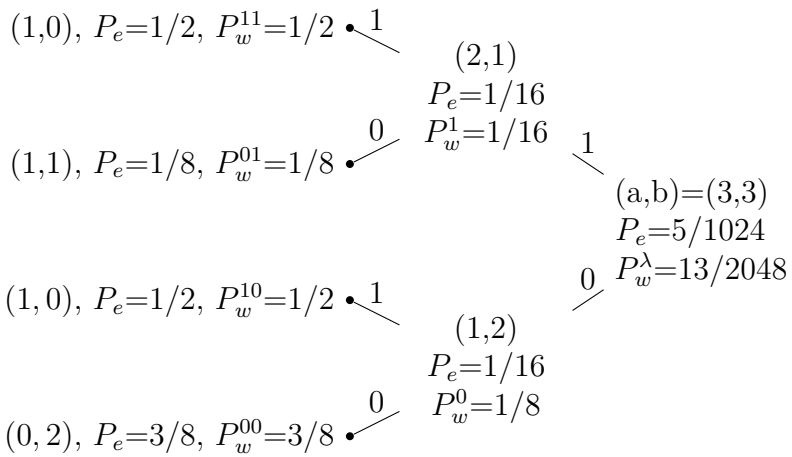


Figura 3.3: Exemplo de árvore de contexto, com o cálculo das probabilidades estimadas e ponderadas.

Após essa contagem, a probabilidade ponderada de cada nó pode ser encontrada utilizando-se a probabilidade estimada de Krichevsky-Trofimov (KT). A probabilidade de uma fonte binária, com probabilidade θ de gerar 1, gerar uma sequência com a zeros e b uns é $(1-\theta)^a\theta^b$. Quando o parâmetro θ é desconhecido, é possível ponderar esta probabilidade utilizando a distribuição $\left(\frac{1}{2}, \frac{1}{2}\right)$-Dirichlet, como a seguir:

$$P_e(a,b) = \int_0^1 \frac{1}{\pi\sqrt{(1-\theta)\theta}}(1-\theta)^a\theta^b d\theta, \tag{3.10}$$

que é a probabilidade estimada KT (maiores informações sobre a distribuição de Dirichlet encontram-se no apêndice C). Contudo, é possível escrever uma relação sequencial para encontrar as probabilidades KT para outros valores de a e b ($P_e(0,0) = 1$) [90]:

$$P_e(a+1,b) = \frac{a+1/2}{a+b+1}P_e(a,b), \tag{3.11}$$

$$P_e(a,b+1) = \frac{b+1/2}{a+b+1}P_e(a,b). \tag{3.12}$$

De modo análogo, para uma fonte emitindo M símbolos, encontra-se a generalização da relação sequencial para as probabilidades estimadas KT ($P_e(0,0,\ldots,0) = 1$), sendo b_i a contagem do símbolo i, $i \in \{0,\ldots,M-1\}$ [34]

$$P_e(b_0,b_1,\ldots b_{i-1},b_i+1,b_{i+1},\ldots,b_{M-1}) = \frac{b_i+1/2}{b_0+\cdots+b_i+\cdots+b_{M-1}+M/2} \times P_e(b_0,b_1,\ldots b_{i-1},b_i,b_{i+1},\ldots,b_{M-1}). \tag{3.13}$$

A probabilidade ponderada, a ser utilizada no CTW, é P_w^s para cada nó da árvore de contexto. Para o caso de alfabeto binário em particular, temos:

$$P_w^s = \begin{cases} \frac{1}{2}P_e(a_s,b_s) + \frac{1}{2}P_w^{0s}P_w^{1s}, & \text{para } 0 \le l(s) < D \\ P_e(a_s,b_s), & \text{para } l(s) = D. \end{cases} \tag{3.14}$$

Já para o caso geral de um alfabeto M-ário, temos as probabilidades ponderadas:

$$P_w^s = \begin{cases} \frac{1}{2}P_e^s(x^n) + \frac{1}{2}\prod_{i=0}^{M-1}P_w^{is}(x^n), & \text{para } 0 \le l(s) < D \\ P_e^s(x^n), & \text{para } l(s) = D. \end{cases} \tag{3.15}$$

Estimadores

Jiao apresenta quatro estimadores da taxa de informação direcional que utilizam as probabilidades ponderadas P_w^λ encontradas segundo o algoritmo CTW, para cada saída da fonte de comprimento N [34]. Os quatro estimadores são bastante similares. Os dois primeiros, "E1" e "E2", apresentam taxas de convergência estabelecidas como vantagem. Os outros dois, "E3" e "E4", apresentam a vantagem de sempre serem não negativos. O estimador "E2", utilizado no capítulo 9, é encontrado pelas seguintes fórmulas:

$$\hat{I}_N(X^N \to Y^N) = \hat{H}(Y^N) - \hat{H}(Y^N||X^N), \text{ em que} \tag{3.16}$$

$$\hat{H}(Y^N) = \frac{1}{N}\sum_{n=1}^{N}\sum_{y_{n+1}} Q(y_{n+1}|Y^n)\log\frac{1}{Q(y_{n+1}|Y^n)} \tag{3.17}$$

$$\hat{H}(Y^N||X^N) = \frac{1}{N}\sum_{n=1}^{N} f(Q(x_{n+1}, y_{n+1}|X^n, Y^n)), \tag{3.18}$$

$$f(P) = -\sum_{x,y} P(x,y)\log P(y|x), \tag{3.19}$$

$$Q(x_{n+1}|x^n) = \frac{P_w^\lambda(x^{n+1})}{P_w^\lambda(x^n)} \tag{3.20}$$

Este estimador $\hat{I}_N$ segue o seguinte algoritmo:

- Fixe o comprimento N da sequência e a profundidade D da árvore de contexto;
- Inicialize a estimativa $\hat{I} \leftarrow 0$;
- Para n indo de 1 até N:
  - Faça supersímbolo $z_n = (x_n, y_n)$;
- Para n indo de $D+1$ até $N+1$:
  - Capture o contexto z_{n-D}^{n-1} para o n-ésimo símbolo z_n;
  - Atualize a árvore de contexto para cada valor possível de z_n;
  - Estime $Q(z_n|Z^{n-1})$;
  - Capture o contexto y_{n-D}^{n-1} para o n-ésimo símbolo y_n;
  - Atualize a árvore de contexto para cada valor possível de y_n;
  - Estime $Q(y_n|Y^{n-1})$;
  - Atualize $\hat{I} \leftarrow \hat{I} + f(Q(z_n|Z^{n-1})) - f(Q(y_n|Y^{n-1}))$, em que $f(.)$ é dado pela equação (3.19);
- Atualize $\hat{I}_N \leftarrow \hat{I}/(N-D)$.

Se Q é a atribuição de probabilidade pelo algoritmo CTW, $\{\mathbf{X}, \mathbf{Y}\}$ formam um processo de Markov aperiódico de alfabeto finito cuja ordem não excede a profundidade na árvore no algoritmo CTW, e $\mathbf{Y}$ também é um processo de Markov aperiódico de alfabeto finito de mesma ordem que $\{\mathbf{X}, \mathbf{Y}\}$, então $\lim_{N\to\infty} \hat{I}_N(X \to Y) \to I_\infty(X \to Y)$, com probabilidade 1, o que siginifica que este estimador é consistente. A consistência dos demais estimadores "E1", "E3" e "E4" também está provada. O código para os estimadores de Jiao [34] está atualmente disponível na página: `http://web.stanford.edu/~tsachy/DIcode/`.

3.2.2 Estimador de Quinn

O estimador de Quinn [67] foi desenvolvido para o caso particular em que os processos envolvidos na estimação de informação direcional são trens de *spikes* neurais. Primeiramente, consideram-se dois processos aleatórios $\mathbf{X} = (X_\tau : 0 \leq \tau \leq T)$ e $\mathbf{Y} = (Y_\tau : 0 \leq \tau \leq T) \in \mathcal{Y}_T$. $\mathcal{Y}_T$ é o conjunto de funções $y : (0, T] \to \mathbb{Z}_+$, não-decrescentes, contínuas pela direita e com $y_0 = 0$. A Fig. 3.4 ilustra uma função amostra de $\mathbf{Y}$.

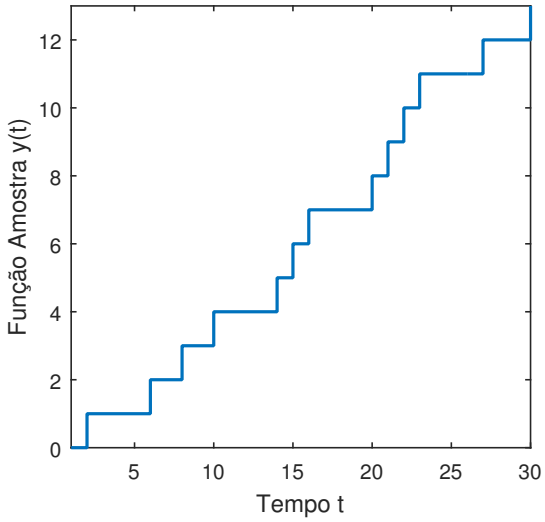


Figura 3.4: Gráfico de uma função amostra $y(t)$ do processo aleatório $\mathbf{Y}$ de contagem de eventos.

Definem-se as histórias no tempo t para os processos de contagem de eventos $\mathbf{Y}$ como a σ-álgebra gerada pelos processos aleatórios apropriados até o tempo t (a definição de σ-álgebra encontra-se no apêndice A):

$$\mathcal{F}_t = \sigma\left(X_\tau : \tau \in [0, t], Y_t : \tau \in [0, t)\right), \tag{3.21}$$

$$\mathcal{F}_t' = \sigma\left(Y_\tau : \tau \in [0, t)\right). \tag{3.22}$$

Sabe-se que a função de intensidade condicional caracteriza completamente a estrutura estatística de todos processos de contagem de eventos bem comportados usados em inferência estatística de dados neurais [12], sendo definida como:

$$\lambda(t||\mathcal{F}_t) = \lim_{\Delta \to 0} \frac{P(Y_{t+\Delta} - Y_t = 1|\mathcal{F}_t)}{\Delta}, \tag{3.23}$$

$$\lambda(t||\mathcal{F}_t') = \lim_{\Delta \to 0} \frac{P(Y_{t+\Delta} - Y_t = 1|\mathcal{F}_t')}{\Delta}, \tag{3.24}$$

A função de intensidade condicional $\lambda(t||\mathcal{F}_t)$ dá a probabilidade instantânea de *spike* por unidade de tempo, dados os *spikes* neurais prévios [67]. Outra equação interessante que está relacionada a λ é a da probabilidade de intervalo entre *spikes*:

$$p(t||\mathcal{F}_t') = \lambda(t||\mathcal{F}_t') \exp\left\{-\int_0^t \lambda(u||\mathcal{F}_u')du\right\}, \tag{3.25}$$

que dá a probabilidade de ocorrência de um *spike* em t dado o passado $\mathcal{F}_t^{'}$. Note que a equação (3.25) apresenta a mesma forma que a densidade de probabilidade de uma variável aleatória exponencial [85] (exceto pela integral). Cabe aqui lembrar que muitas vezes trens de *spikes* são modelados por processos de Poisson não homogêneos. Em processos de Poisson, a densidade de probabilidade do tempo de ocorrência do próximo evento é dada pela distribuição exponencial. A integral da função λ em (3.25) se deve justamente ao fato do processo ser de Poisson e não homogêneo, ou seja, se deve ao fato de λ variar no tempo.

Para um processo de contagem de eventos $\mathbf{Y} \in \mathcal{Y}_T$ com funções de intensidade condicional $\lambda(t||\mathcal{F}_t)$ e $\lambda(t||\mathcal{F}_t^{'})$, a densidade de Y em y dado x é dada por [12]:

$$f(y||x;\lambda) = \exp\left\{\int_0^T \log\lambda(t||\mathcal{F}_t)dy_t - \lambda(t||\mathcal{F}_t)dt\right\}, \tag{3.26}$$

e, analogamente, a densidade de Y em y é dada por:

$$f(y;\lambda) = \exp\left\{\int_0^T \log\lambda(t||\mathcal{F}_t^{'})dy_t - \lambda(t||\mathcal{F}_t^{'})dt\right\}. \tag{3.27}$$

Discretizando $(0,T)$ em $N = T/\Delta$ intervalos de comprimento $\Delta \ll 1$ de modo que $dy = (dy_1, dy_2, \ldots, dy_N)$ com $dy_n = y_{(n+1)\Delta} - y_{n\Delta} \in \{0,1\}$, é possível aproximar as equações (3.26) e (3.27) por [67]:

$$-\log f(y||x;\lambda) \approx \sum_{n=1}^{N} -\log\lambda(n||\mathcal{F}_n)dy_n + \lambda(n||\mathcal{F}_n)\Delta, \tag{3.28}$$

$$-\log f(y;\lambda) \approx \sum_{n=1}^{N} -\log\lambda(n||\mathcal{F}_n^{'})dy_n + \lambda(n||\mathcal{F}_n^{'})\Delta. \tag{3.29}$$

Agora que já definiram-se as funções de intensidade condicional, é possível estimar a taxa de informação direcional entre dois trens de *spikes* distintos, processos $\mathbf{X}$ e $\mathbf{Y}$, a partir de alguns pressupostos [67]. Além do já citado pressuposto de estacionariedade e ergodicidade dos processos $\mathbf{X}$ e $\mathbf{Y}$, pressupõe-se que eles sejam de memória finita (ou seja, formem cadeias de Markov). Além disso, pressupõe-se que para um conjunto pré-especificado de funções $\{h_k : k \geq 0\}$ a função de intensidade condicional pertença aos modelos lineares generalizados de h (ou seja, $\lambda(i||\mathcal{F}_i) \in \mathrm{GLM}(h)$) [67]. Modelos GLM têm a função de intensidade condicional atendendo a:

$$\log\lambda(n||\mathcal{F}_n) = \alpha_0 + \sum_{j=1}^{J}\alpha_j dy_{n-j} + \sum_{k=1}^{K}\beta_k h_k(x_{n-(k-1)}), \tag{3.30}$$

em que h_k é alguma função sobre as covariáveis extrínsecas $(x_{n-(k-1)})$ e $\theta = \{\alpha_0, \alpha_1, \ldots, \alpha_J, \beta_1, \ldots, \beta_K\}$ é o vetor de parâmetros. Maiores informações sobre GLM são mostradas nas referências [60, 67, 82].

Em geral, o vetor de parâmetros $\theta = \{\alpha_0, \alpha_1, \ldots, \alpha_J, \beta_1, \ldots, \beta_K\}$, pertencente ao espaço vetorial de possíveis parâmetros $\Omega(J,K)$, e os valores de J e K são desconhecidos. Neste caso, propõe-se que uma estimativa de máxima verossimilhança seja utilizada para encontrar θ (equação (3.31)) e que o princípio MDL (*minimum description length*) [27] seja utilizado para encontrar os valores de J e K (equação (3.32)):

$$\hat{\theta}(J,K) = \arg\min_{\theta\in\Omega(J,K)} -\frac{1}{N}\log f(Y_1^N||X_1^N;\theta) \tag{3.31}$$

$$(\hat{J},\hat{K}) = \arg\min_{(J,K)} \min_{\theta\in\Omega(J,K)} -\frac{1}{N}\log f(Y_1^N||X_1^N;\theta) + \frac{J+K}{2N}\log N \tag{3.32}$$

Além disso, sabe-se que [67]:

$$\begin{aligned} H_\infty(Y||X) &= \lim_{N\to\infty} -\frac{1}{N}\log f(Y_1^N||X_1^N;\theta) \\ &= \lim_{N\to\infty} -\frac{1}{N}\sum_{n=1}^{N}\log(\lambda_\theta(n|\mathcal{F}_n))dy_n - \lambda_\theta(n|\mathcal{F}_n)\Delta), \end{aligned} \tag{3.33}$$

de modo que propõe-se o seguinte método para estimar a informação direcional:

- Encontrar $\hat{J}$, $\hat{K}$ e $\hat{\theta}$ de acordo com as equações (3.31) e (3.32);
- Encontrar $\hat{H}_\infty(Y||X)$, estimativa de $H_\infty(Y||X)$, de acordo com (3.33) com os parâmetros obtidos no passo anterior;
- Utilizar estimadores $\hat{H}_\infty(Y)$ universais de taxa de entropia $H_\infty(Y)$, para processos $\mathbf{Y}$ estacionários ergódicos que formem cadeias de Markov de ordem finita;
- Computar a diferença $\hat{I}_\infty(X \to Y) = \hat{H}_\infty(Y) - \hat{H}_\infty(Y||X)$.

O código para este estimador de informação direcional foi disponibilizado pelo próprio autor de [67] sob pedido.

Diante do exposto, percebe-se que há alguma complexidade computacional nos algoritmos descritos, em particular para estimação de informação direcional (estimadores de Jiao e de Quinn). No próximo capítulo, comparamos os estimadores citados aqui de informações mútua e direcional.

Capítulo 4

Simulação de Métodos de Estimação de Medidas de Informação entre Variáveis Aleatórias Discretas

Este capítulo mostra o resultado das simulações com alguns dos estimadores vistos no capítulo 3, de medidas de informação para o caso discreto. Aqui serão apresentadas simulações para estimação de informação mútua com os estimadores *plug-in*, *plug-in* com correção de viés QE, e com o estimador de Jiao. Já para estimação de informação direcional, serão comparados os estimadores de Quinn e de Jiao.

4.1 Informação Mútua

A seguir, são apresentados os gráficos de diversos exemplos simulados em que se conhece o valor analítico de informação mútua. Os gráficos mostram o desempenho dos estimadores em termos de acurácia. Os gráficos com as médias dos estimadores, que indicam o comportamento estatístico do viés, também apresentam na linha em vermelho o valor analítico para a informação mútua.

4.1.1 1º Caso: Canal BSC

O canal BSC (*binary symmetric channel*) é um canal que apresenta entrada X e saída Y, $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, e parâmetro p de transição. O valor analítico da informação mútua entre entrada e saída, quando a entrada X é uniforme, é dado por [13]:

$$I(X;Y) = 1 - \mathcal{H}(p). \tag{4.1}$$

As Fig. 4.1 e Fig. 4.2 indicam o desempenho dos estimadores quando $p = 0.5$. Já as Fig. 4.3 e Fig. 4.4 indicam o desempenho dos estimadores quando $p = 0.1$.

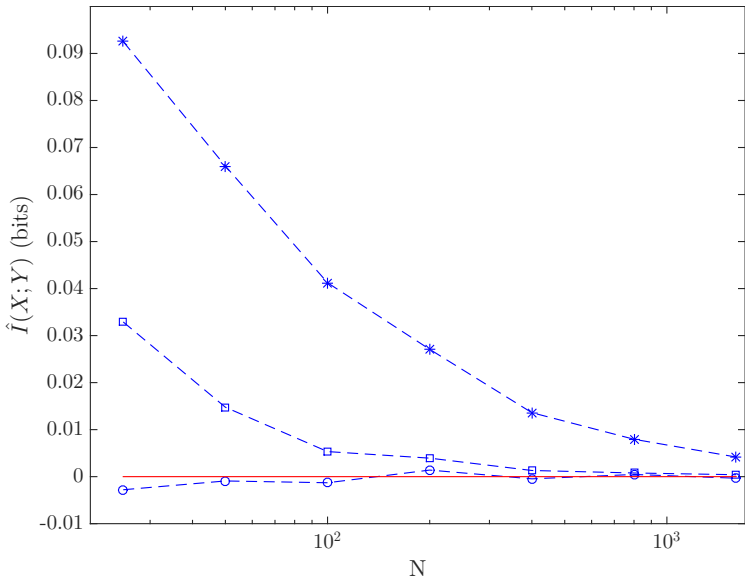


Figura 4.1: Canal BSC, $p = 0.5$, $I(X, Y) = 0$. Curva com quadrado: *plug-in*, curva com bola: *plug-in* com correção QE, curva com asterisco: Jiao. $N = \{25, 50, 100, 200, 400, 800, 1600\}$. Médias em 50 realizações para cada tamanho amostral.

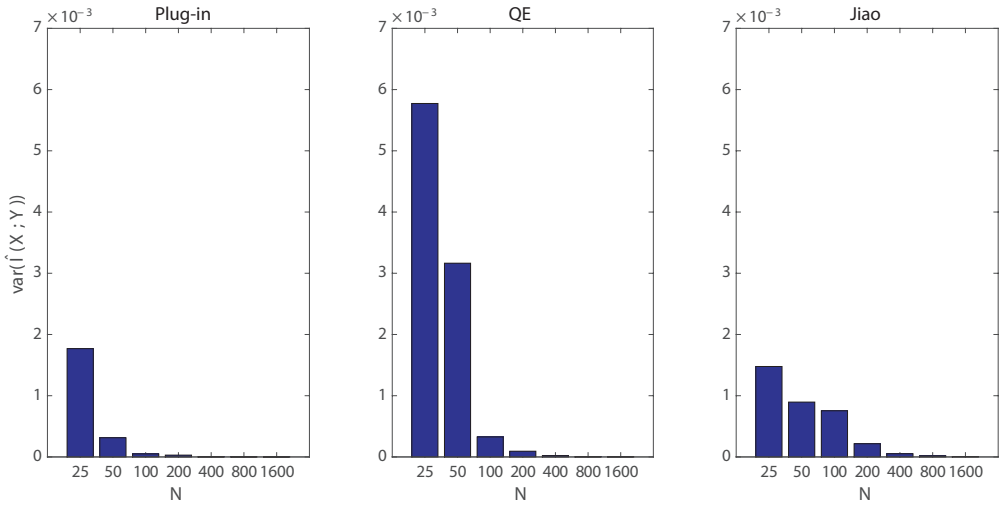


Figura 4.2: Canal BSC, $p = 0.5$, $I(X, Y) = 0$. Variâncias em 50 realizações para cada tamanho amostral. $N = \{25, 50, 100, 200, 400, 800, 1600\}$.

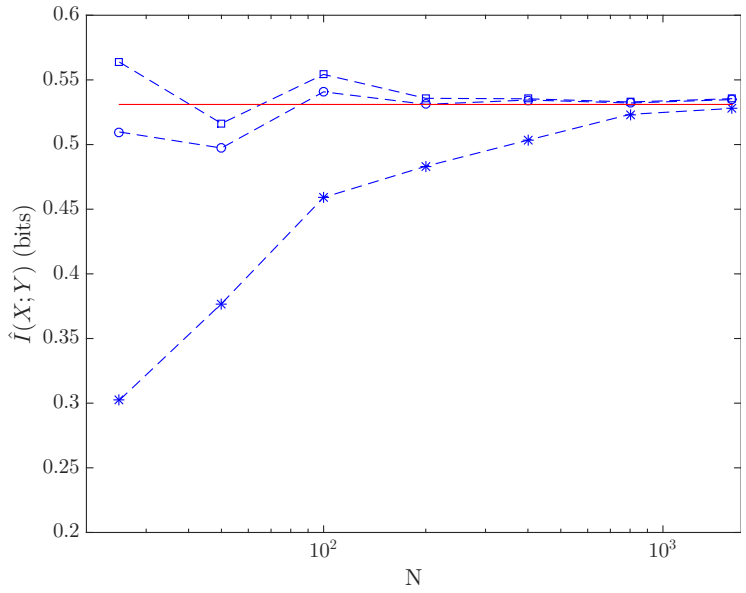


Figura 4.3: Canal BSC, $p = 0.1$, $I(X, Y) = 0.5310$. Curva com quadrado: *plug-in*, curva com bola: *plug-in* com correção QE, curva com asterisco: Jiao. $N = \{25, 50, 100, 200, 400, 800, 1600\}$. Médias em 50 realizações para cada tamanho amostral.

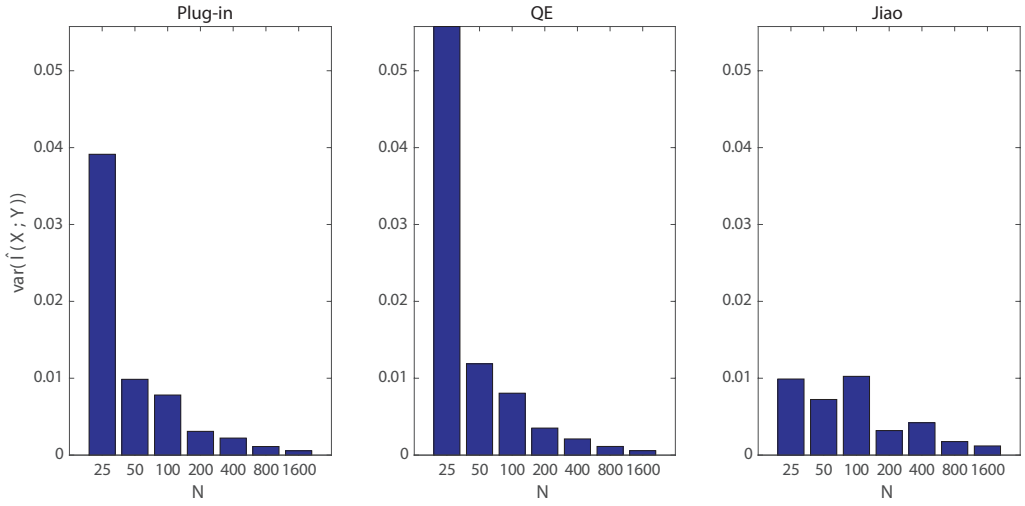


Figura 4.4: Canal BSC, $p = 0.1$, $I(X, Y) = 0.5310$. Variâncias em 50 realizações para cada tamanho amostral. $N = \{25, 50, 100, 200, 400, 800, 1600\}$.

4.1.2 2º Caso: Canal BEC

O canal BEC (*binary erasure channel*) é um canal que apresenta entrada X e saída Y, $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, e, 1\}$, e parâmetro α de apagamento. O valor analítico da informação mútua entre entrada e saída, quando a entrada X é uniforme, é dado por [13]:

$$I(X;Y) = 1 - \alpha. \tag{4.2}$$

As Fig. 4.5 e Fig. 4.6 indicam o desempenho dos estimadores quando $\alpha = 0.8$. Já as Fig. 4.7 e Fig. 4.8 indicam o desempenho dos estimadores quando $\alpha = 0.2$.

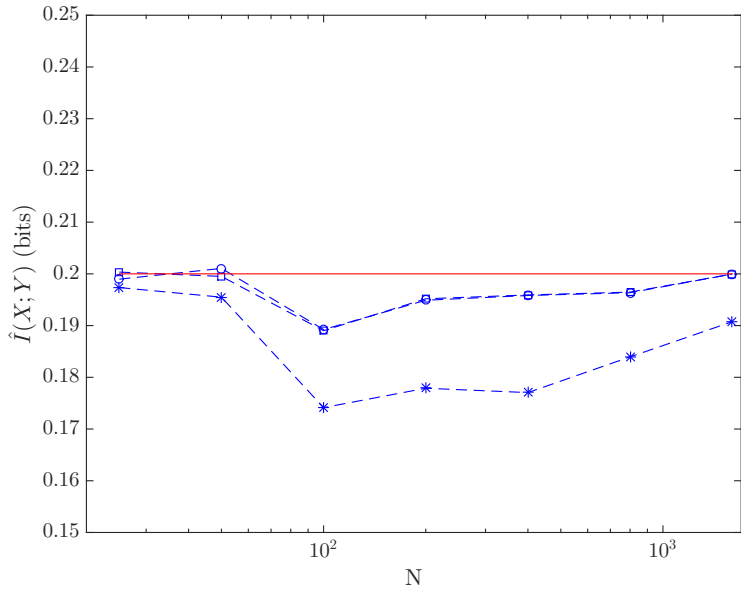


Figura 4.5: Canal BEC, $\alpha = 0.8$, $I(X,Y) = 0.2$. Curva com quadrado: *plug-in*, curva com bola: *plug-in* com correção QE, curva com asterisco: Jiao. $N = \{25, 50, 100, 200, 400, 800, 1600\}$. Médias em 50 realizações para cada tamanho amostral.

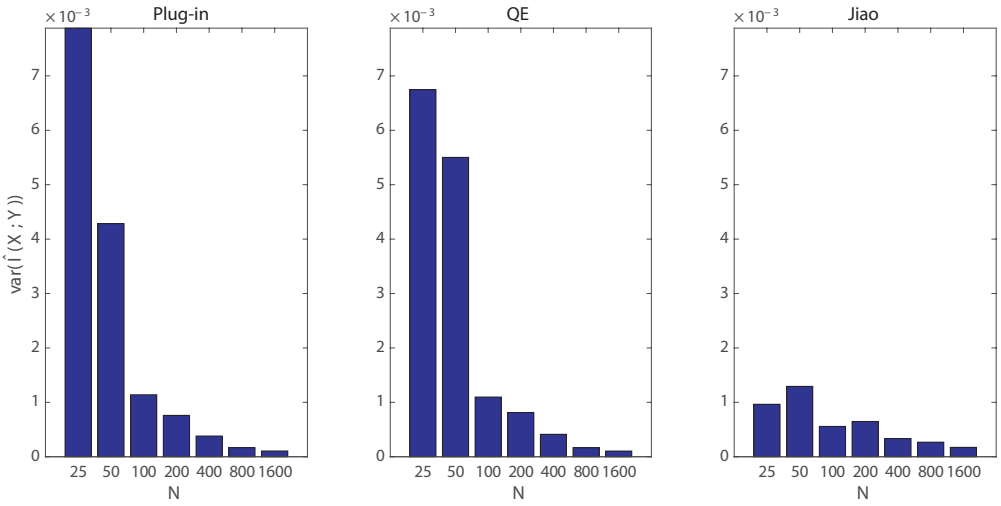


Figura 4.6: Canal BEC, $\alpha = 0.8$, $I(X,Y) = 0.2$. Variâncias em 50 realizações para cada tamanho amostral. $N = \{25, 50, 100, 200, 400, 800, 1600\}$.

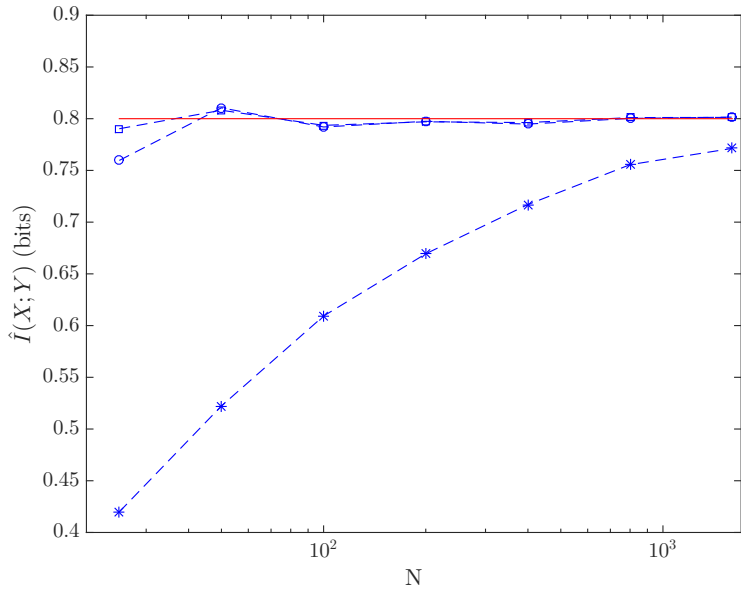


Figura 4.7: Canal BEC, $\alpha = 0.2$, $I(X,Y) = 0.8$. Curva com quadrado: *plug-in*, curva com bola: *plug-in* com correção QE, curva com asterisco: Jiao. $N = \{25, 50, 100, 200, 400, 800, 1600\}$. Médias em 50 realizações para cada tamanho amostral.

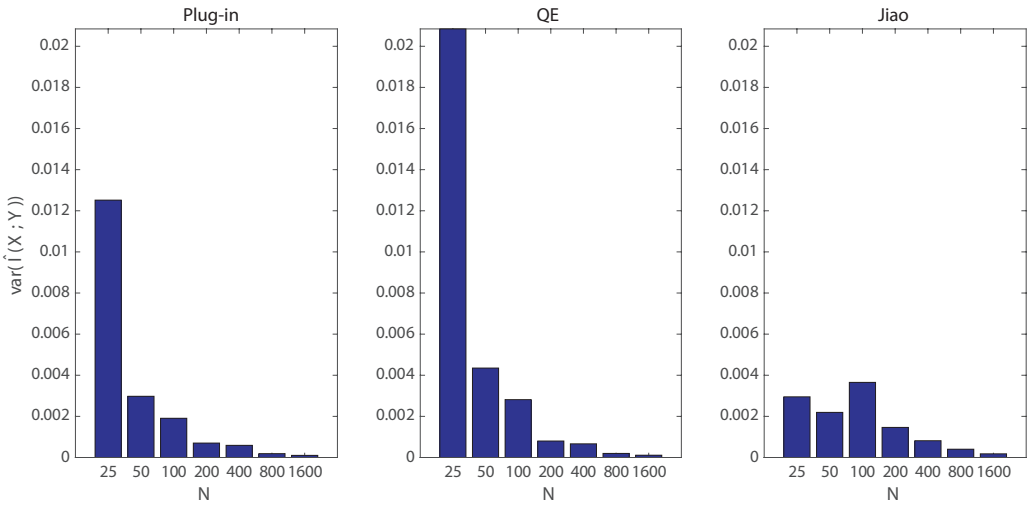


Figura 4.8: Canal BEC, $\alpha = 0.2$, $I(X,Y) = 0.8$. Variâncias em 50 realizações para cada tamanho amostral. $N = \{25, 50, 100, 200, 400, 800, 1600\}$.

4.1.3 3º Caso: Canal Simétrico

Repetimos os experimentos para um canal simétrico. Neste caso, e para uma entrada distribuída uniformente, o valor de analítico de informação mútua é dado por:

$$I(X;Y) = \log |\mathcal{Y}| - H(\mathbf{r}), \tag{4.3}$$

em que $|\mathcal{Y}|$ denota a cardinalidade do alfabeto $\mathcal{Y}$, e $\mathbf{r}$ é uma linha da matriz de transição $P(Y|X)$ (no caso do canal simétrico, vale lembrar que $|\mathcal{Y}| = |\mathcal{X}|$).

Para o canal com a seguinte matriz de transição:

$$P(Y|X) = \begin{bmatrix} 0.5 & 0.2 & 0.15 & 0.15 \\ 0.2 & 0.5 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.5 & 0.2 \\ 0.15 & 0.15 & 0.2 & 0.5 \end{bmatrix}, \tag{4.4}$$

o valor $I(X;Y) = 0.2145$bit. As Fig. 4.9 e Fig. 4.10 indicam o desempenho dos estimadores para este exemplo.

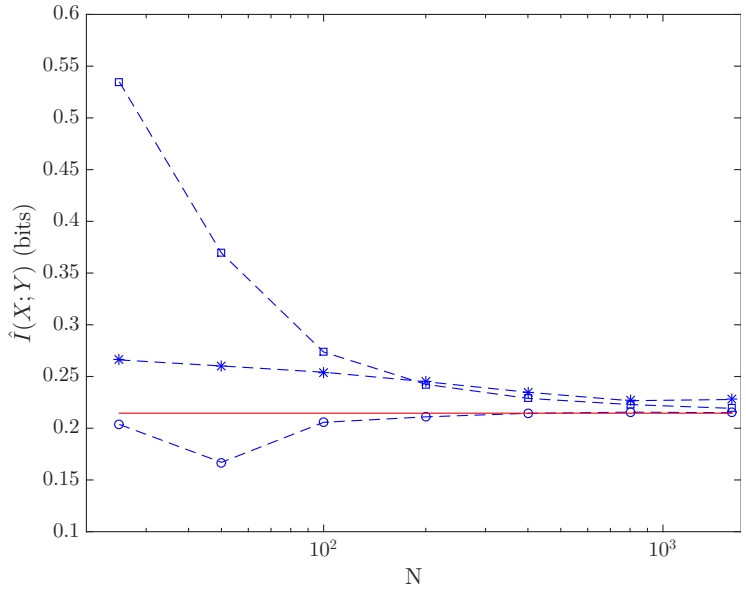


Figura 4.9: Canal simétrico, $|\mathcal{X}| = |\mathcal{Y}| = 4$. Curva com quadrado: *plug-in*, curva com bola: *plug-in* com correção QE, curva com asterisco: Jiao. $N = \{25, 50, 100, 200, 400, 800, 1600\}$. Médias em 50 realizações para cada tamanho amostral.

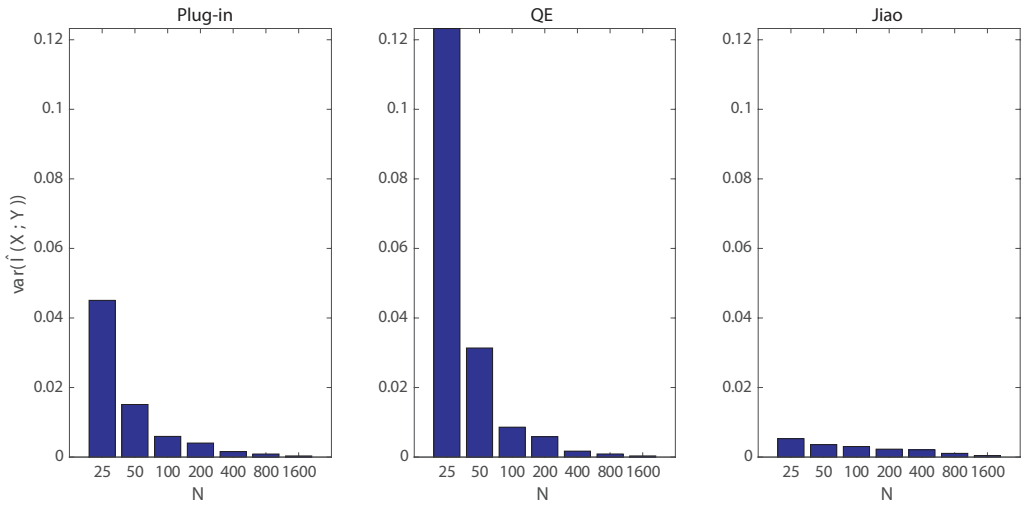


Figura 4.10: Canal simétrico, $|\mathcal{X}| = |\mathcal{Y}| = 4$. Variâncias em 50 realizações para cada tamanho amostral. $N = \{25, 50, 100, 200, 400, 800, 1600\}$.

Repetimos os experimentos para o canal simétrico com uma matriz de transição cuja primeira linha é:

$$P(Y=0|X) = \left[\begin{array}{cccccccc} 1/2 & 1/4 & 1/8 & 1/16 & 1/32 & 1/64 & 1/128 & 1/128 \end{array} \right], \qquad (4.5)$$

e as demais linhas são rotações desta primeira linha, sempre mantendo a diagonal principal da matriz com valor 1/2. As Fig. 4.11 e Fig. 4.12 indicam o desempenho dos estimadores neste exemplo.

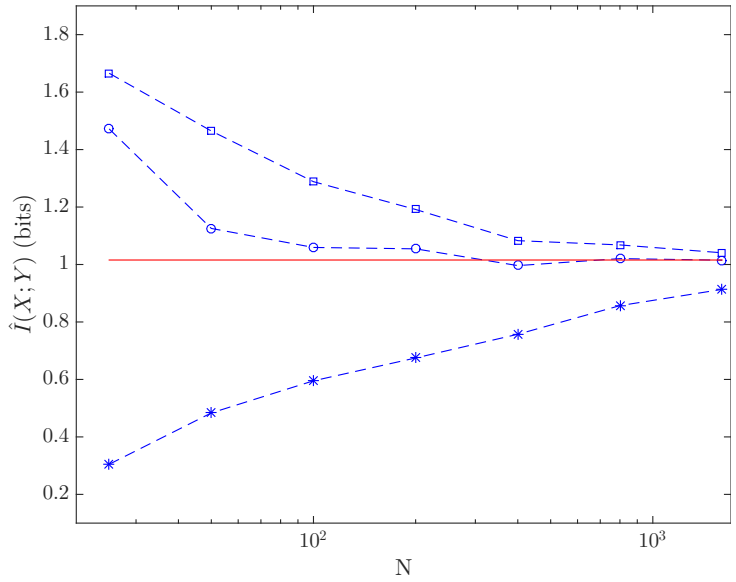


Figura 4.11: Canal simétrico, $|\mathcal{X}| = |\mathcal{Y}| = 8$. Curva com quadrado: *plug-in*, curva com bola: *plug-in* com correção QE, curva com asterisco: Jiao. $N = \{25, 50, 100, 200, 400, 800, 1600\}$. Médias em 50 realizações para cada tamanho amostral.

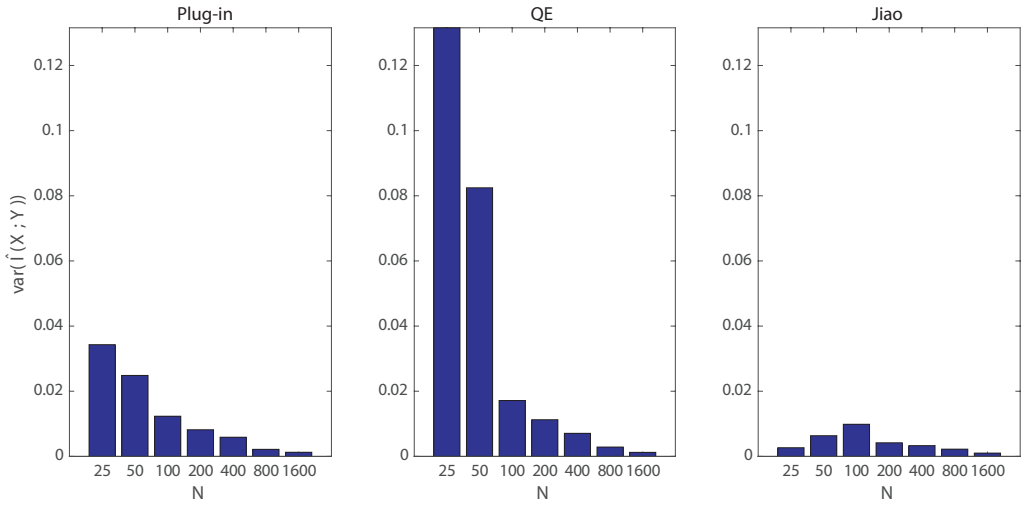


Figura 4.12: Canal simétrico, $|\mathcal{X}| = |\mathcal{Y}| = 8$. Variâncias em 50 realizações para cada tamanho amostral. $N = \{25, 50, 100, 200, 400, 800, 1600\}$.

Observando os resultados gráficos que refletem a acurácia dos estimadores, algumas observações são pertinentes. Para alfabetos muito pequenos ($|\mathcal{X}| = \{2, 3\}$) e tamanhos amostrais também pequenos ($N = 25$), o método *plug-in* sem correção QE apresenta viés, porém variância menor, sendo preferível ao método *plug-in* com correção de viés. Já para alfabetos um pouco maiores (exemplo, $|\mathcal{X}| = \{4, 8\}$) ou para tamanhos amostrais intermediários ($N \geq 50$), o método QE em geral obteve resultados mais acurados. Já o metodo de Jiao, além de mais lento computacionalmente, demora mais pra convergir conforme N aumenta, apresentando estimativas mais enviesadas, apesar da variância menor do estimador.

No que concerne o tempo de execução dos estimadores para informação mútua, observou-se que tamanhos amostrais de $N = 1600$ tomaram aproximadamente 0.001s com o método *plug-in*, para cada estimativa realizada. Com correção QE, o método tomou aproximadamente 0.004s. O método de Jiao tomou aproximadamente 0.396s, com profundidade de árvore $D = 1$, cardinalidade do alfabeto $|\mathcal{X}| = 2$. Portanto, observamos que em termos de velocidade, o melhor desempenho foi do método *plug-in*, seguido do método *plug-in* acompanhado da correção de viés QE e por último o método de Jiao.

4.2 Informação Direcional

Nesta seção montamos um exemplo em que se conhece o valor de informação direcional analítica entre dois processos $\mathbf{X}$ e $\mathbf{Y}$ e verificamos o desempenho dos estimadores de informação direcional de Jiao e de Quinn.

Seja $\mathbf{X}$ uma fonte em árvore apresentada na Fig. 3.2. O diagrama de estados correspondente para o processo $\mathbf{X}$ é dado na Fig. 4.13. Lembramos que θ_s é a probabilidade da fonte binária emitir 1 dado que o sufixo da sequência foi s.

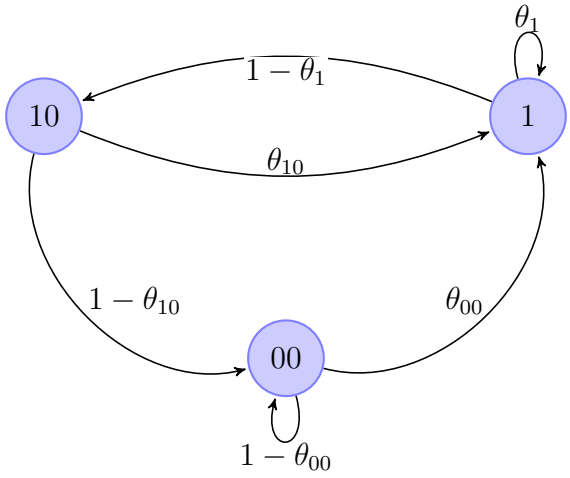


Figura 4.13: Diagrama de estados para processo $\mathbf{X}$.

O processo $\mathbf{X}$ passa por um canal BSC (*binary symmetric channel*) resultando no processo $\mathbf{Y}$.

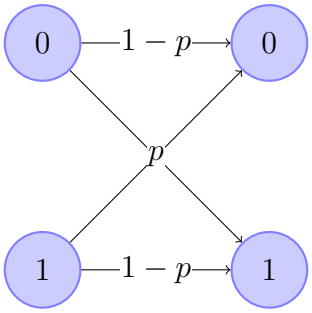


Figura 4.14: Canal BSC.

É possível encontrar $I_\infty(Y \to X)$. Primeiro, encontra-se a taxa de entropia de $\mathbf{X}$:

$$\begin{aligned} H_\infty(X) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n|X^{n-1}) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n|X_{n-2}^{n-1}) \qquad (4.6) \\ &= \lim_{N\to\infty} H(X_N|X_{N-2}^{N-1}) \qquad (4.7) \\ &= \pi_{00}\mathcal{H}(\theta_{00}) + \pi_{10}\mathcal{H}(\theta_{10}) + \pi_1\mathcal{H}(\theta_1) \end{aligned}$$

em que a equação (4.6) se deve ao fato da memória de $\mathbf{X}$ ser 2, ao passo que a equação (4.7) se deve à média de Cesáro e ao fato de, no limite, $\mathbf{X}$ estar em regime estacionário. Os termos π_1, π_{10} e π_{00} são as probabilidades da cadeia se encontrar nos estados 1, 10 e 00.

Depois, encontra-se a taxa de entropia causalmente condicionada:

$$\begin{aligned} H_\infty(X||Y) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n|X^{n-1}Y^n) \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} H(X_n|X_{n-2}^{n-1}Y_n) \qquad (4.8) \\ &= \lim_{N\to\infty} H(X_N|X_{N-2}^{N-1}Y_N) \qquad (4.9) \end{aligned}$$

em que a equação (4.8) se deve ao fato de que o canal é sem memória e a equação (4.9) se deve à média de Cesáro e ao fato de que, no limite, $\mathbf{X}$ e $\mathbf{Y}$ são conjuntamente estacionários.

A fim de simplificar a notação, consideramos que a cadeia $\mathbf{X}$ é estacionária desde o início do processo, e escrevemos:

$$\begin{aligned} \lim_{N\to\infty} H(X_N|X_{N-2}^{N-1}Y_N) &= H(X_3|X_1^2, Y_3) \\ &= -\sum_{x_1^2, y_3} P(x_1^2, y_3) \sum_{x_3} P(x_3|x_1^2, y_3) \log P(x_3|x_1^2, y_3) \end{aligned}$$

Para encontrar os valores de $P(x_3|x_1^2, y_3)$, considera-se que $X_1^2 \to X_3 \to Y_3$, ou seja, essas variáveis formam uma cadeia de Markov e, portanto,

$$P(x_1^2, y_3|x_3) = P(y_3|x_3)P(x_1^2|x_3).$$

Assim, desenvolve-se a seguinte relação:

$$\begin{aligned} P(x_3|x_1^2, y_3) &= \frac{P(x_3, x_1^2, y_3)}{P(x_1^2, y_3)} \\ &= \frac{P(y_3, x_3|x_1^2)P(x_1^2)}{P(y_3|x_1^2)P(x_1^2)} \\ &= \frac{P(y_3, x_3|x_1^2)}{\sum_{x_3} P(y_3, x_3|x_1^2)} \\ &= \frac{P(y_3|x_3, x_1^2)P(x_3|x_1^2)}{\sum_{x_3} P(y_3, x_3|x_1^2)} \\ &= \frac{P(y_3|x_3)P(x_3|x_1^2)}{\sum_{x_3} P(y_3|x_3)P(x_3|x_1^2)}, \end{aligned}$$

em que $P(y_3|x_3)$ é dado pelo gráfico da Fig. 4.14, ao passo que $P(x_3|x_1^2)$ é dado pelo gráfico da Fig. 4.13. Similarmente, as probabilidades $P(x_1^2, y_3)$ são dadas por:

$$\begin{aligned} P(x_1^2, y_3) &= \sum_{x_3} P(x_1^2, y_3|x_3)P(x_3) \\ &= \sum_{x_3} P(x_1^2|x_3)P(y_3|x_3)P(x_3) \\ &= \sum_{x_3} P(x_1^2, x_3)P(y_3|x_3) \\ &= \sum_{X_3} P(x_3|x_1^2)P(x_1^2)P(y_3|x_3), \end{aligned}$$

em que $P(X_1^2)$ assume os valores π_1, π_{10} ou π_{00}.

Ajustando os valores $\theta_1 = 0.1$, $\theta_{10} = 0.3$, $\theta_{00} = 0.5$, encontram-se $\pi_1 = 0.32$, $\pi_{10} = 0.28$ e $\pi_{00} = 0.40$. O apêndice D evidencia como proceder as contas. Fazendo a probabilidade de transição do canal BSC $p = 0.1$, encontram-se:

$$\begin{aligned} H_\infty(X) &= 0.80\text{bit} \\ H_\infty(X||Y) &= 0.39\text{bit} \\ I_\infty(Y \to X) &= 0.41\text{bit} \end{aligned}$$

As estimativas segundo o método de Jiao foram feitas com profundidade da árvore $D = 2$, escolhendo-se o estimador "E4". A realização de cada estimativa de tamanho $N = 10^5$ durou aproximadamente 30 segundos. Os gráficos das médias amostrais das estimativas e das variâncias amostrais das estimativas encontram-se nas Fig. 4.15 e 4.16, respectivamente.

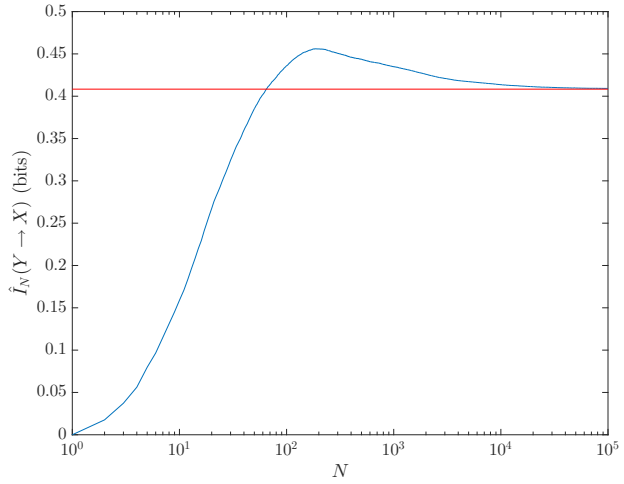


Figura 4.15: Médias das estimativas de Jiao para taxa de informação direcional em 50 realizações, variando o comprimento N das amostras (em escala logarítmica). Valor analítico de $I_N(Y \to X)$ indicado na linha vermelha.

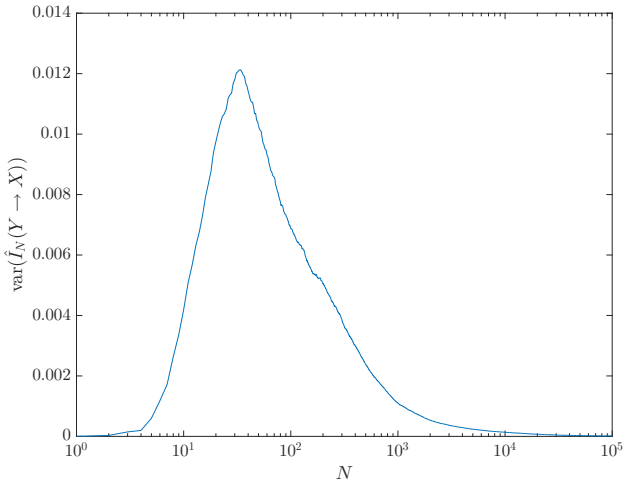


Figura 4.16: Variâncias das estimativas de Jiao para taxa de informação direcional em 50 realizações, variando o comprimento N das amostras (em escala logarítmica).

As estimativas também foram feitas segundo o método de Quinn, em uma busca com parâmetros $J = K = 7$. Cada estimativa durou aproximadamente 60 segundos. No entanto, todas as estimativas deram resultado exatamente nulo, o que significa que este método não foi capaz de detectar a causalidade existente. Tal resultado deve-se ao fato de que o método de Quinn provê um estimador paramétrico cujo modelo não se ajusta ao caso simulado em estudo. Além disso, ressalta-se aqui que o estimador de Quinn pode se tornar ainda mais lento, de acordo com a busca em um espaço maior de parâmetros do algoritmo MDL ($J > 7$, $K > 7$, no exemplo analisado).

Capítulo 5

Métodos de Estimação de Medidas de Informação entre Variáveis Aleatórias Contínuas

Seguindo a linha geral deste trabalho e estendendo o assunto do capítulo 3, este capítulo lida com o problema da estimação de medidas de informação, mas agora para o caso em que as variáveis envolvidas assumem valores contínuos (as variáveis são reais). Neste caso, a estimação é ainda mais delicada que no do capítulo 3, pois distribuições em que as variáveis aleatórias envolvidas podem assumir infinitos valores precisam ser estimadas a partir de um conjunto finito de dados.

5.1 Método do Particionamento do Suporte

Este método pode ser visto como uma extensão do método *plug-in*. Nele, as variáveis aleatórias contínuas são discretizadas, por sua realização ser observada em partições do conjunto de valores que elas podem assumir. Portanto, neste método é feita uma segmentação de intervalos que a variável pode assumir e posterior contagem de quantas vezes cada segmento (ou *bin*) ocorreu.

Uma pergunta natural ao utilizar o método do particionamento do suporte é em quantos segmentos deve ser dividido o intervalo que cada variável pode assumir e como variam as correspondentes estimativas de informação. Ingenuamente, poderia-se pensar que o uso de muitos intervalos refinaria a estimativa de informação mútua, melhorando sua estimação. Contudo, se a partição do espaço conjunto das variáveis X e Y tiver um número igual ou maior que o tamanho amostral, a entropia conjunta pode ser igual à $\ln N$, estando mais ligada ao tamanho amostral de que à estrutura do sistema em estudo de fato. Ao passo que a entropia conjunta pode saturar em um valor dado pelo tamanho amostral e/ou pelo número de valores distintos de dados, as entropias individuais não saturam com o refinamento das partições. Lembrando que informação mútua pode ser escrita como $I(X;Y) = H(X)+H(Y)-H(X,Y)$, chega-se à conclusão que um refinamento exagerado da partição acaba por sobrestimar as estimativas de informação mútua [29].

Já foi proposto o uso de intervalos equipovoados como um simples método de particionamento adaptativo, promovendo segmentos equipovoados marginalmente. Mesmo assim, o número de segmentos é fundamental neste método. No caso de estimar a informação mútua de duas variáveis X e Y com tamanho amostral N, propôs-se o número marginal de segmentos $Q \leq \sqrt[3]{N}$ [62, 29].

Apesar de apresentar a vantagem da simplicidade, há indícios de que esse método apresente viés [43, 44, 16]. Uma melhoria no seu desempenho pode ser feita através do algoritmo do particionamento adaptativo do suporte, que será explorado na seção seguinte.

5.2 Método do Particionamento Adaptativo do Suporte

Proposto por Darbellay e Vajda [14] em 1999, o método do particionamento adaptativo do suporte apresentou uma melhora significativa em relação ao simples particionamento do suporte. Para compreender este método, tornam-se relevantes algumas definições.

Primeiramente, consideremos uma partição $\mathcal{R}$ de $\mathcal{X} \times \mathcal{Y}$ em retângulos do tipo $A \times B$ e as densidades de distribuições condicionais

$$P_{X,Y|A\times B} = P_{X,Y|X\in A, Y\in B} \tag{5.1}$$

e

$$\begin{aligned} P_{X|A} &= P_{X|X\in A} \\ P_{Y|B} &= P_{Y|Y\in B} \\ f_{X,Y|A\times B} = \frac{1_{A\times B} f_{X,Y}}{\int 1_{A\times B} f_{X,Y}} &= \frac{1_{A\times B} f_{X,Y}}{P_{X,Y}(A\times B)}. \end{aligned} \tag{5.2}$$

Similarmente, definem-se

$$f_{X|A} = \frac{1_A f_X}{P_X(A)} \tag{5.3}$$

$$f_{Y|B} = \frac{1_B f_Y}{P_Y(B)}, \tag{5.4}$$

em que 1_E denota a função indicadora do conjunto E. Isto é,

$$1_E(x) = \begin{cases} 1, \text{ se } x \in E, \\ 0, \text{ se } x \in E^c. \end{cases}$$

A divergência restrita é definida como

$$D^{\mathcal{R}}(X;Y) = \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \ln \frac{P_{X,Y}(A\times B)}{P_X(A) P_Y(B)}. \tag{5.5}$$

Já a divergência residual é definida como

$$D_{\mathcal{R}}(X;Y) = \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \int f_{X,Y|A\times B} \ln \frac{f_{X,Y|A\times B}}{f_{X|A} f_{Y|B}}. \tag{5.6}$$

Darbellay e Vajda propõem que, para cada partição $\mathcal{R}$, a informação mútua é igual à soma das divergências restrita e residual, ou seja:

$$I(X;Y) = D^{\mathcal{R}}(X;Y) + D_{\mathcal{R}}(X;Y). \tag{5.7}$$

Diz-se que uma sequência de partições $\{\mathcal{R}^{(k)}, k \in \mathbb{N}\}$ é aninhada se cada célula $C \in \mathcal{R}^{(k)}$ é uma união disjunta

$$C = \sum_{\ell=1}^{L} C_\ell \tag{5.8}$$

de células $C_\ell \in \mathcal{R}^{(k+1)}$, em que L varia com C. $\mathcal{R}^{(k+1)}$ é chamada um refinamento de $\mathcal{R}^{(k)}$.

Uma sequência aninhada $\{\mathcal{R}^{(k)}, k \in \mathbb{N}\}$ é dita assintoticamente suficiente para X, Y se para cada $\epsilon > 0$ há um k_ϵ tal que para cada $C \subset \mathcal{X} \times \mathcal{Y}$ mensurável pode-se encontrar $C_0 \in \mathcal{S}(\mathcal{R}^{(k_\epsilon)})$ satisfazendo a condição

$$P_{X,Y}(C \Delta C_0) < \epsilon, \tag{5.9}$$

em que Δ denota a diferença simétrica.

Dessa maneira, propõe-se que se uma sequência de partições aninhadas $\mathcal{R}^{(k)}$ é assintoticamente suficiente para X, Y, então

$$\lim_{k\to\infty} D^{\mathcal{R}^{(k)}}(X;Y) = I(X;Y), \tag{5.10}$$

$$\lim_{k\to\infty} D_{\mathcal{R}^{(k)}}(X;Y) = 0, \tag{5.11}$$

em que ambas convergências são monótonas.

Agora, de posse das definições e proposições dadas, é possível explanar como estimar a informação mútua de uma amostra de N realizações independentes do par (X, Y). Seja

$$P_N(C) = \frac{1}{N} \sum_{n=1}^{N} 1_C(X_n, Y_n),\ C \subset \mathcal{X} \times \mathcal{Y} \tag{5.12}$$

a distribuição de probabilidade empírica em $\mathcal{X} \times \mathcal{Y}$. Lembramos que 1_C denota a função indicadora da célula C.

A estimativa de divergência restrita é dada por:

$$\begin{aligned}
\hat{D}_{N,k}(X;Y) &= \sum_{A\times B \in \mathcal{R}^{(k)}} P_N(A \times B) \ln \frac{P_N(A \times B)}{P_N(A \times \mathbb{R}) P_N(\mathbb{R} \times B)} \\
&= \frac{1}{N} \sum_{A\times B \in \mathcal{R}^{(k)}} \sum_{n=1}^{N} 1_{A\times B}(X_n, Y_n) \ln \frac{P_N(A \times B)}{P_N(A \times \mathbb{R}) P_N(\mathbb{R} \times B)} \\
&= \frac{1}{N} \sum_{n=1}^{N} \ln \frac{P_N(A_n \times B_n)}{P_N(A_n \times \mathbb{R}) P_N(\mathbb{R} \times B_n)},
\end{aligned} \tag{5.13}$$

em que $A_n \times B_n$ é a célula em que (X_n, Y_n) se encontra. Está provado que [14]:

$$\lim_{N\to\infty} \hat{D}_{n,k} = D^{\mathcal{R}^{(k)}}(X;Y) \text{ em probabilidade.} \tag{5.14}$$

E que

$$\lim_{N\to\infty} P(|\hat{D}_{n,k_\epsilon} - I(X;Y)| < \epsilon) = 1. \tag{5.15}$$

Um ϵ menor requer um k_ϵ maior, ou seja, uma partição mais refinada.

A ideia fundamental do particionamento adaptativo do suporte consiste em escolher as células, ou retângulos, apropriadas para aproximar a estimativa de divergência restrita do valor verdadeiro de informação. Para tanto, Darbellay e Vajda propõem um algoritmo baseado nos dados, em que os parâmetros $r, s \in \{2, 3, \dots\}, s \geq r$, e $\delta > 0$ (geralmente escolhe-se $r = 2$):

- **Passo 0**. Seja $\mathcal{R}_N^{(1)} = \mathcal{P}_N \times \mathcal{Q}_N$, em que $\mathcal{P}_N$ e $\mathcal{Q}_N$ são partições de $\mathbb{R}$ em r intervalos especificados pelos quantis marginais

$$a_j = F_N^{-1}(j/r),\ 1 \leq j \leq r-1$$

e

$$b_j = G_N^{-1}(j/r),\ 1 \leq j \leq r-1.$$

$F_N(x) = P_N((-\infty, x) \times \mathbb{R})$, $G_N(x) = P_N(\mathbb{R} \times (-\infty, x))$ são as distribuições marginais de P_N.

- **Passo k (repetir até critério de parada).** $\mathcal{R}_N^{(k+1)}$ contém todos os retângulos de $\mathcal{R}_N^{(k)}$ que não intersectam com a amostra. Para os demais retângulos $A \times B \in \mathcal{R}_n^{(k)}$ há duas possibilidades. Definimos primeiro as funções de distribuição marginal condicional

$$F_{A,N}(x) = \frac{P_N(((-\infty, x) \cap A) \times \mathbb{R})}{P_N(A \times \mathbb{R})},\ x \in \mathbb{R},$$

$$G_{B,N}(x) = \frac{P_N(\mathbb{R} \times ((-\infty, x) \cap B))}{P_N(\mathbb{R} \times B)},\ x \in \mathbb{R}$$

e a partição $\mathcal{R}_{A\times B} = \mathcal{P}_A \times \mathcal{P}_B$ de $A \times B$, em que $\mathcal{P}_A$ e $\mathcal{P}_B$ são partições de A e B especificadas pelos correspondentes quantis condicionais amostrais marginais definidos de maneira análoga ao passo 0, mas com r substituído por $s \geq r$. A decomposição da equação (5.8) se mantém para $A_\ell \times B_\ell \in \mathcal{R}_{A\times B}$ e $L = s^2$. Se

$$D_{A\times B}(X;Y) = \sum_{\ell=1}^{s^2} \frac{P_N(A_\ell \times B_\ell)}{P_N(A \times B)} \ln\left(\frac{P_N(A_\ell \times B_\ell)}{P_N(A \times B)} \Big/ \frac{P_N(A_\ell \times \mathbb{R})P_N(\mathbb{R} \times B_\ell)}{P_N(A \times \mathbb{R})P_N(\mathbb{R} \times B)}\right)$$

for maior que $\delta = \delta_s$ então $\mathcal{R}_N^{(k+1)}$ contém todos os retângulos da partição $\mathcal{R}_{A\times B}$ construídos r quantis marginais. Caso contrário, $\mathcal{R}_N^{(k+1)}$ contém a célula $A \times B$ em si.

- **Passo 2 ou critério de parada.** Caso $\mathcal{R}_N^{(k+1)} \neq \mathcal{R}_N^{(k)}$, o passo 1 é repetido. Caso contrário o processo é terminado e $\mathcal{R}_N$ é definido como $\mathcal{R}_N^{(k)}$.

Através do passo 2, estima-se a informação mútua pelo uso da equação (5.13).

A Fig. 5.1 ilustra o procedimento do particionamento simples do suporte em (a) *versus* o procedimento do particionamento adaptativo do suporte em (b), para o caso em que há uma variável aleatória uniforme U em $[0, 2\pi]$ determinando duas variáveis aleatórias:

$$\begin{aligned} X &= \cos U, \\ Y &= \text{sen} U. \end{aligned}$$

A área sombreada na Fig. 5.1 (b) não será mais refinada segundo o algoritmo do particionamento adaptativo do suporte, mostrando que o algoritmo faz o particionamento refinado apenas das células de interesse de realização do par de variáveis aleatórias (X, Y).

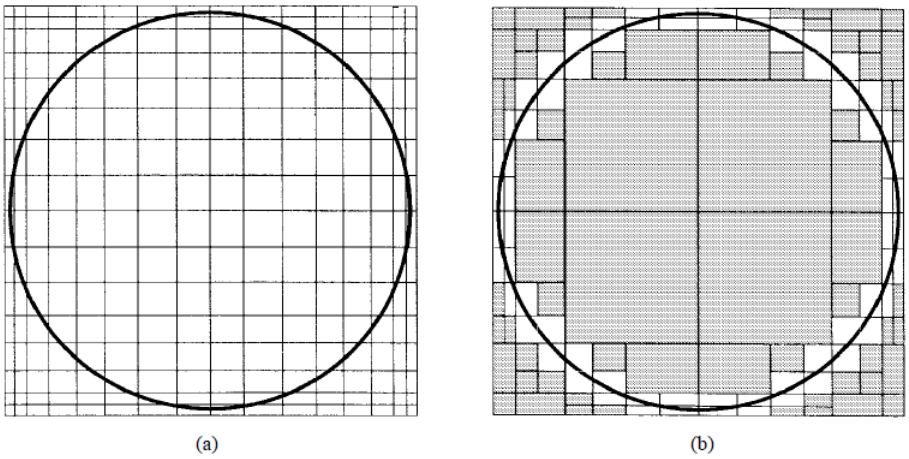


Figura 5.1: Métodos de particionamento do suporte. À esquerda, particionamento simples do suporte, à direita, particionamento adaptativo. Figura adaptada da referência [14].

O método do particionamento adaptativo do suporte de Darbellay e Vajda em [14] apresentou desempenho satisfatório segundo os autores. Possivelmente, parte do sucesso do estimador deve-se ao fato de que a informação mútua também é definida de maneira geral como um supremo entre todas as partições [13], e o método busca as melhores partições para estimar a informação mútua.

5.3 Método KDE

O método KDE, de *kernel density estimator*, é um método aplicado para estimar densidades de probabilidades proposto por Emanuel Parzen [65]. Essas densidades de probabilidade são então aplicadas nos funcionais de medidas de informação a fim de estimá-los. O método consiste na utilização de uma função *kernel*, a qual pondera pesos aos eventos ocorridos x_n e reconstrói a densidade em um ponto x como:

$$\hat{f}(x) = \frac{1}{N\epsilon} \sum_{n=1}^{N} K\left(\frac{x - x_n}{\epsilon}\right) \tag{5.16}$$

em que N é o número de pontos, a função $K(x)$ é a função *kernel* e ϵ é o seu tamanho (ou parâmetro de largura de banda). A função $K(x)$ determina o peso de cada ponto dependente de sua distância, satisfazendo aos seguintes requisitos [66]:

- $K(x) \geq 0$,
- $\int K(x)dx = 1$,
- $\lim_{x\to\infty} |xK(x)| = 0$.

É possível estabelecer condições sob as quais prova-se a consistência da estimativa $\hat{f}(x)$ na média quadrática [65], isto é,

$$\mathbb{E}|\hat{f}(x) - f(x)|^2 \to 0 \text{ quando } N \to \infty. \tag{5.17}$$

Além disso, o viés deste estimador de densidade de probabilidade é diretamente proporcional ao quadrado da largura ϵ da função *kernel*, ao passo que a variância deste estimador

de densidade de probabilidade é inversamente proporcional à largura ϵ da função *kernel* [66]. Assim como no método do particionamento do suporte, a densidade de probabilidade estimada pode ser utilizada nos funcionais das medidas de informação mútua, entropia de transferência, e de informação direcional, a fim de estimá-las. Portanto, a estimação KDE de medidas de informação também é enviesada e sensível à escolha do parâmetro ϵ [50].

Recentemente, foi publicado um artigo em que é proposto um estimador de informação direcional que utiliza o método KDE para identificar a zona de início de uma convulsão em uma ataque epiléptico [54]. Este estimador foi proposto para o caso em que os processos sob análise são estacionários e ergódicos, ou em janelas temporais que possuam tais características. Este estimador obteve um desempenho satisfatório, no sentido de encontrar muitas vezes o mesmo foco epiléptico encontrado pela análise de um especialista médico.

5.4 Método do Mapeamento Simbólico

O método do mapeamento simbólico foi proposto inicialmente por Bandt e Pompe para a estimação de entropia [8], sendo em seguida estendido para estimação de informação mútua [38] e de entropia de transferência [77]. O método consiste essencialmente em fazer uma ordenação dos valores contínuos que as variáveis envolvidas podem assumir e utiliza um parâmetro k que é chamado parâmetro de imersão. Por exemplo, na estimação de entropia, dada a sequência temporal:

$$x = (2.1, 3.4, 4.5, 5, 3.2, 5.7, 1),$$

e o valor escolhido de $k = 2$, observa-se quantas vezes $x_n < x_{n+1}$ e quantas vezes $x_n > x_{n+1}$. Nessas situações são atribuídos os símbolos 01 e 10, respectivamente. Há, na sequência dada, 4 ocorrências do símbolo 01 e 2 ocorrências do símbolo 10. Portanto, a entropia neste caso, de ordem $k = 2$, é dada por:

$$H(2) = -(4/6)\log(4/6) - (2/6)\log(2/6) \approx 0.918 \text{ bit}.$$

Analogamente, quando considera-se o parâmetro de imersão $k = 3$, observam-se 2 vezes o símbolo 012 ($(2.1, 3.4, 4.5)$, $(3.4, 4.5, 5)$), 2 vezes o símbolo 120 ($(4.5, 5, 3.2)$, $(3.2, 5.7, 1)$), 1 vez o símbolo 102 (5, 3.2, 5.7), de modo que encontra-se a entropia:

$$H(3) = -(2/5)\log(2/5) - (2/5)\log(2/5) - (1/5)\log(1/5) \approx 1.522 \text{ bit}.$$

Definindo π como uma das $k!$ permutações de ordem k, e a frequência relativa

$$p(\pi) = \frac{\text{número de}\{n | n \leq N - k, (x_{n+1}, \ldots, x_{n+k}) \text{ tem tipo } \pi\}}{N - k + 1},$$

em que N é a duração temporal da série x_1^N, define-se a entropia de permutação de ordem $k \geq 2$ como

$$H(k) = -\sum p(\pi)\log p(\pi). \tag{5.18}$$

Bandt e Pompe verificaram que a equação (5.18) cresce com k, de maneira que propuseram a entropia de permutação por letra:

$$h_k = H(k)/(k-1). \tag{5.19}$$

A entropia de transferência simbólica é definida de maneira análoga [77]. Seguindo a ideia do teorema de Takens [81], é possível reescrever as séries temporais $x_n^{n+l-1} = (x(n), x(n+\tau), \ldots, x(n+(l-1)\tau))$ e $y_n^{n+m-1} = (y(n), y(n+\tau), \ldots, y(n+(m-1)\tau))$, em que τ é o tempo de reconstrução e l e m são dimensões de imersão, dos sinais X e Y, respectivamente. É possível ordenar as séries na ordem crescente, por exemplo: $x(n+(k_{n1}-1)\tau) \leq x(n-(k_{n2}-1)\tau) \leq \cdots \leq x(n-(k_{nl}-1)\tau)$. Dessa maneira obtém-se o sinal $\hat{x}_n = (k_{n1}, k_{n2}, \ldots, k_{nl})$. Com as sequências $\hat{x}_n$ e $\hat{y}_n$, obtém-se a estimativa de transferência simbólica

$$T^S(X \to Y) = \sum p(\hat{y}_{n+\delta}, \hat{y}_n, \hat{x}_n) \log \frac{p(\hat{y}_{n+\delta}|\hat{y}_n\hat{x}_n)}{p(\hat{y}_{n+\delta}|\hat{y}_n)}, \tag{5.20}$$

em que δ é um passo temporal de predição.

5.5 Método KSG

Kraskov *et al.* [44, 43] estabelecem dois algoritmos para estimar informação mútua entre variáveis contínuas. Em geral, os dois algoritmos apresentam desempenho similar. Para o caso da estimação de informação mútua de uma distribuição bidimensional, o método KSG foi testado em [44] com a distribuição gaussiana, com a distribuição Γ-exponencial [14] e com a distribuição Weinman exponencial ordenada [14].

Para o caso bidimensional, a função de densidade de probabilidade para a distribuição gaussiana conjunta é dada por [85]:

$$f(x,y) = \frac{1}{2\pi(\det(K))^{1/2}} e^{-(1/2)[x-\mathbb{E}X \;\; y-\mathbb{E}Y]K^{-1}[x-\mathbb{E}X \;\; y-\mathbb{E}Y]^T}, \tag{5.21}$$

em que K é a matriz de covariância

$$K = \begin{pmatrix} \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{var}(Y) \end{pmatrix}.$$

Quando $(X,Y) \sim \mathcal{N}(0,K)$, isto é, quando X e Y apresentam distribuição gaussiana conjunta com média 0 e matriz de covariância atendendo a

$$K = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix},$$

a informação mútua entre X e Y é dada por:

$$I(X;Y) = -\frac{1}{2}\ln(1-\rho^2), \tag{5.22}$$

em que ρ é o coeficiente de correlação entre X e Y.

Para o caso da distribuição Γ-exponencial, a densidade de probabilidade depende de um parâmetro θ, sendo dada por [44]:

$$f(x,y) = \frac{1}{\Gamma(\theta)} x^\theta e^{-x-xy}. \tag{5.23}$$

A informação mútua para essa distribuição é:

$$I(X;Y) = \psi(\theta+1) - \ln\theta, \tag{5.24}$$

em que ψ representa a função digamma (informações sobre a função digamma no apêndice B).

Já para o caso da distribuição Weinman exponencial ordenada, existe uma dependência a 2 parâmetros θ e θ_0, cuja análise em [44] é limitada ao caso em que $\theta_0 = 1$, de modo que a densidade de probabilidade conjunta é dada por:

$$f(x,y) = \frac{2}{\theta} e^{-2x-(y-x)/\theta}, \tag{5.25}$$

para $x, y > 0$ e sendo $f(x,y) = 0$ caso contrário. Para esta distribuição, a informação mútua é dada por [44]:

$$I(X;Y) = \begin{cases} \ln \dfrac{2\theta}{1-2\theta} + \psi\left(\dfrac{1}{1-2\theta}\right) - \psi(1), & \text{se } \theta < 1/2 \\ -\psi(1), & \text{se } \theta = 1/2 \\ \ln \dfrac{2\theta-1}{\theta} + \psi\left(\dfrac{2\theta}{2\theta-1}\right) - \psi(1), & \text{se } \theta > 1/2 \end{cases}. \tag{5.26}$$

Como há em todos os casos citados fórmulas exatas para a informação mútua, é possível verificar o desempenho dos estimadores. Para o caso gaussiano, o desempenho do método KSG implementado diretamente foi considerado satisfatório em [44]. No caso da distribuição Γ-exponencial o desempenho foi satisfatório mediante a transformação das variáveis $x'_n = \ln x_n$ e $y'_n = \ln y_n$. Isso pode ter ocorrido porque a distribuição Γ-exponencial pode apresentar um ponto de máximo muito agudo. No caso da distribuição Weinman exponencial ordenada, as estimativas foram menores que os valores teóricos de informação mútua, mesmo com a melhoria decorrente da transformação logarítmica ($x'_n = \ln x_n$ e $y'_n = \ln y_n$) [44].

Além disso, testaram-se os casos em que as variáveis eram independentes, apresentando distribuição uniforme, exponencial, ou quando uma delas era gaussiana e a outra uniforme ou exponencial. Em todos os casos, as estimativas de informação mútua apresentaram desempenho satisfatório ($\hat{I} \approx 0$, com erros estatísticos da ordem de $10^{-4}/10^{-3}$) [44]. Para o caso de dimensões mais altas, só distribuições gaussianas foram testadas, com resultados satisfatórios, especialmente quando as distribuições são independentes [44].

Convém ressaltar aqui que, além dos resultados promissores de diversas simulações meticulosas em [44], recentemente em [19][20], Gao *et al.* provam a consistência do estimador KSG, diante dos seguintes pressupostos genéricos:

- $\int f(x,y) |\ln f(x,y)| dxdy < \infty$,
- Existe uma constante C' tal que as distribuições condicionais $f(x|y) < C'$ e $f(y|x) < C'$ em quase todo ponto,
- $f(x,y)$ é duplamente continuamente diferenciável e a matriz de Hess H_f satisfaz

$$||H_f(x,y)||_2 = \left|\left| \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x \partial y} \\ \dfrac{\partial^2 f}{\partial y \partial x} & \dfrac{\partial^2 f}{\partial y^2} \end{bmatrix} \right|\right|_2 < C \tag{5.27}$$

  em quase todo ponto,

- o número de vizinhos escolhido k satisfaz $k > \max\{d_x/d_y, d_y/d_x\}$, $d_x, d_y = O(1)$, em que d_x é a dimensão de X e d_y é a dimensão de Y.

Além disso, sob algumas modificações dos pressupostos, Gao *et al.* [19] encontram taxas de convergência para o viés e a variância do estimador KSG.

Os dois algoritmos dos estimadores KSG se baseiam no método de Kosachenko-Leonenko para estimar entropia [41]. Este último se baseia na distribuição média de probabilidade dos vizinhos de cada ponto da amostra, e convém esclarecê-lo a fim de compreender como a partir dele pode-se derivar também a estimação de informação mútua.

5.5.1 Estimação de Kosachenko-Leonenko para Entropia de Shannon

Seja X uma variável aleatória contínua assumindo valores em algum espaço métrico, isto é, há uma distância bem definida $||x - x'||$ entre duas de suas realizações. Lembramos brevemente aqui a definição de norma ℓ_p de um vetor de dimensão J [58]:

$$||x||_p = \left(\sum_{j=1}^{J} |x_j|^p \right)^{(1/p)}. \tag{5.28}$$

Por exemplo, seja a norma definida ℓ_2 (distância euclidiana). Considera-se o ponto $X = x$ e ordenam-se suas distâncias a todas as outras realizações de X na norma ℓ_2. Seja $f(x)$ sua densidade de probabilidade. A entropia diferencial de X, como uma extensão da entropia de Shannon, pode ser escrita como:

$$H(X) = -\int f(x) \ln f(x) dx, \tag{5.29}$$

que é a média de $-\ln f(x)$. Portanto, como pode ser visto no apêndice A, um estimador não enviesado para $H(X)$ é a média amostral de $\ln f(x)$:

$$\hat{H}(X) = -\frac{1}{N} \sum_{n=1}^{N} \widehat{\ln(f(x_n))}, \tag{5.30}$$

se por sua vez $\widehat{\ln(f(x_n))}$ for um estimador não enviesado de $\ln f(x)$. Para encontrar um estimador $\widehat{\ln(f(x_n))}$ adequado, considera-se a distribuição de probabilidade de $P_k(\epsilon)$ para as distâncias entre x_n e seu k-ésimo vizinho mais próximo (segundo a norma ℓ_2, por exemplo).

A probabilidade $P_k(\epsilon)d\epsilon$ é igual à chance de que exista um vizinho a uma distância de x_n dentro do intervalo $[\epsilon/2, \epsilon/2 + d\epsilon/2]$, $k-1$ vizinhos a distâncias menores e $N-k-1$ vizinhos a distâncias maiores. Denotando p_n como a massa da bola ϵ centrada em x_n, $p_n(\epsilon) = \int_{||\xi - x_n|| < \epsilon/2} f(\xi) d\xi$. Utilizando a fórmula trinomial, obtém-se:

$$P_k(\epsilon)d\epsilon = \frac{(N-1)!}{1!(k-1)!(N-k-1)!} \frac{dp_n(\epsilon)}{d\epsilon} d\epsilon \times p_n^{k-1} \times (1-p_n)^{(N-k-1)} \tag{5.31}$$

Utilizando a equação (5.31), é possível computar o valor esperado de $\ln p_n(\epsilon)$:

$$\begin{aligned}\mathbb{E}(\ln p_n) &= \int_0^\infty d\epsilon P_k(\epsilon) \ln p_n(\epsilon) \\ &= k\binom{N-1}{k} \int_0^1 dp\ p^{k-1}(1-p)^{N-k-1} \ln p, \\ &= \psi(k) - \psi(N), \end{aligned} \tag{5.32}$$

em que ψ é a função digamma. Pode-se obter um estimador para $\ln f(x)$ supondo que $f(x)$ é constante em toda a bola ϵ, escrevendo-se:

$$p_n(\epsilon) \approx c_d \epsilon^d f(x_n), \tag{5.33}$$

em que d é a dimensão de X e c_d é volume da bola unitária de dimensão d.

Utilizando as equações (5.32) e (5.33), desenvolve-se:

$$\begin{aligned}\ln p_n(\epsilon) &= \ln c_d \epsilon^d f(x_n) \\ \ln p_n(\epsilon) &= \ln c_d + d \ln \epsilon + \ln f(x_n) \\ -\frac{1}{N}\sum_n \ln f(x_n) &= -\frac{1}{N}\sum_n (\ln p_n(\epsilon) - \ln c_d - d \ln \epsilon) \\ \hat{H}(X) = \widehat{\ln f(x_n)} &= -\psi(k) + \psi(N) - \ln c_d + \frac{d}{N}\sum_{n=1}^{N} \ln \epsilon_n, \end{aligned} \tag{5.34}$$

em que ϵ_n é o dobro da distância de x_n ao seu k-ésimo vizinho.

Pela derivação do estimador (5.34), percebe-se que a única razão de seu viés vem da suposição de que $f(x)$ é constante em toda a bola ϵ.

5.5.2 Estimador de Informação Mútua (1)

Agora considera-se a variável aleatória contínua $Z = (X, Y)$. Analisando o caso para um ponto z_n da amostra de tamanho N, encontra-se a distância $\epsilon/2$ ao seu k-ésimo vizinho, cuja distribuição é novamente dada pela equação (5.31). A equação (5.32) também continua válida. A primeira diferença da subseção anterior é na equação (5.33), em que é necessário substituir d por $d_Z = d_X + d_Y$, c_d por $c_{d_X} c_{d_Y}$ e x_n por z_n. Com tais modificações, obtém-se:

$$\hat{H}(X,Y) = -\psi(k) + \psi(N) + \ln(c_{d_X} c_{d_Y}) + \frac{d_X + d_Y}{N}\sum_{n=1}^{N} \ln \epsilon_n. \tag{5.35}$$

Para obter $I(X,Y)$, é necessário subtrair a estimativa (5.35) do somatório das estimativas de $H(X)$ e $H(Y)$. Para essas últimas, já mostrou-se a derivação de (5.34), que pode ser utilizado com o mesmo k para os dois casos. Contudo, este procedimento significaria usar efetivamente escalas distintas de distância no espaços conjuntos e marginais. Para qualquer k fixo, a distância ao k-ésimo vizinho no espaço conjunto será maior que as distâncias dos vizinhos nos espaços marginais. O viés da equação (5.34) resulta da não-uniformidade da densidade. Como o efeito desta densidade depende das distâncias do k-ésimo vizinho, os vieses de $\hat{H}(X)$, $\hat{H}(Y)$ e $\hat{H}(X,Y)$ seriam muito diferentes e não iriam se cancelar.

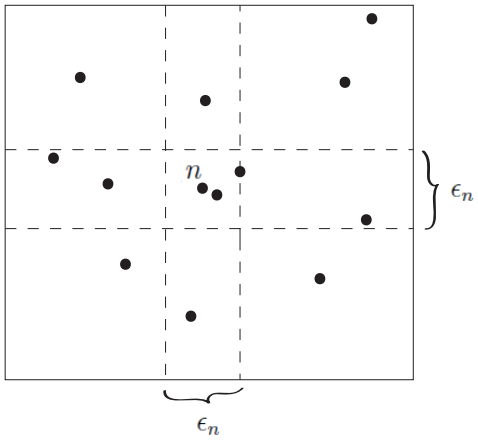


Figura 5.2: Contagem de vizinhos para o ponto z_n, com $k = 2$. Neste caso, $m_x(n) = 3$, $m_y(n) = 5$, $N = 13$

Para evitar este problema, Kraskov *et al.* [44] perceberam que a equação (5.34) é válida para qualquer valor de k e que não é necessário considerar o mesmo k na estimação de entropias marginais. Observando a Fig. 5.2, supõe-se que o k-ésimo vizinho de z_n está sobre um dos lados verticais projetados do quadrado de comprimento de lado ϵ_n. Nesta observação, $\epsilon_n/2$ está definido como o máximo da distância projetada do ponto z_n ao seu k-ésimo vizinho z, ou seja, $\epsilon_n/2 = \max\{||x_n - x||, ||y_n - y||\}$. Neste caso, há no total $m_x(n)$ pontos dentro das linhas verticais $x = x_n \pm \epsilon_n/2$, então $\epsilon_n/2$ é a distância ao $(m_x(n) + 1)$-ésimo vizinho de x_n, e

$$\hat{H}(X) = \frac{-1}{N}\sum_{n=1}^{N}\psi(m_x(n)+1) + \psi(N) + \ln c_{d_X} + \frac{d_X}{N}\sum_{n=1}^{N}\ln\epsilon_n. \tag{5.36}$$

Na outra direção, isto não é exatamente verdadeiro, ou seja, ϵ_n não é exatamente igual ao dobro da distância ao $(m_y(n) + 1)$-ésimo vizinho, quando $m_y(n)$ é definido analogamente como o número de pontos com $||y - y_n|| < \epsilon_n/2$. Contudo, é possível considerar a equação (5.36) como uma aproximação razoável para $H(Y)$, substituindo X por Y. Tal procedimento leva à seguinte equação:

$$\hat{I}(X,Y) = \psi(k) - (\overline{\psi(m_x+1) + \psi(m_y+1)}) + \psi(N), \tag{5.37}$$

em que N é o número de realizações de (X, Y).

O segundo estimador proposto por Kraskov *et al.* é obtido de forma análoga a este primeiro, mas considerando diferentemente a contagem de pontos $m_x(n)$ e $m_y(n)$ dentro das distâncias $||x_n - x||$ e $||y_n - y||$ ao k-ésimo vizinho. Neste caso, ϵ não é considerado o mesmo nas projeções em X e Y.

Vale salientar que a entropia de transferência entre variáveis aleatórias contínuas também pode ser estimada através das distâncias dos vizinhos mais próximos, similarmente ao método KSG [84, 50, 22].

5.6 Método BI-KSG

No desenvolvimento anterior, a métrica utilizada para encontrar o número de vizinhos para estimar a entropia marginal é $\epsilon_n/2 = \max\{||x_n - x||, ||y_n - y||\}$. Neste caso, a norma utilizada é ℓ_∞.

Gao *et al.* [19][20], além de provar a consistência do estimador KSG, propõem e provam a consistência de um outro estimador ligeiramente modificado, encontrando as mesmas taxas de convergência para o viés e a variância do estimador KSG. Tal estimador foi denominado BI-KSG.

A principal diferença entre o estimador KSG e o estimador BI-KSG é que, ao invés de utilizar a norma ℓ_∞, o estimador BI-KSG utiliza a norma ℓ_2 para encontrar o número de vizinhos para estimar a entropia marginal. O estimador BI-KSG para informação mútua é dado por:

$$\begin{aligned}\hat{I}_{BI-KSG}(X;Y) &= \psi(k) + \ln N + \ln \frac{c_{d_x,2}c_{d_y,2}}{c_{d_x+d_y,2}} - \\ &\quad \frac{1}{N}\sum_{n=1}^{N}[\ln(m_{x,n,2}) + \ln(m_{y,n,2})], \end{aligned} \tag{5.38}$$

em que d_x e d_y são as dimensões das variáveis X e Y, respectivamente. O parâmetro k continua sendo o número de vizinhos escolhido e N o tamanho amostral. As constantes na forma:

$$c_{d,2} = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} \tag{5.39}$$

são o volume da bola unitária ℓ_2 em d dimensões. Finalmente, os termos $m_{x,n,2}$ e $m_{y,n,2}$ são definidos como

$$\begin{aligned} m_{x,n,2} &= \sum_{j\neq n} 1_{\{||X_j - X_n||_2 \leq \rho_{k,n,2}\}} \\ m_{y,n,2} &= \sum_{j\neq n} 1_{\{||Y_j - Y_n||_2 \leq \rho_{k,n,2}\}} \end{aligned}$$

em que $\rho_{k,n,2}$ é a distância do ponto (X_n, Y_n) ao k-ésimo vizinho mais próximo na norma ℓ_2, e $1_{\{.\}}$ é a função indicadora da condição no subscrito entre chaves. Em geral, $m_{x,n,2} \geq k$ e $m_{y,n,2} \geq k$.

O estimador BI-KSG é indicado como mais eficiente, por meio de simulações, que o estimador KSG em [19][20]. Ao usar a norma ℓ_2, o vizinho utilizado para estimar uma das entropias marginais não é exatamente o mesmo que o utilizado para estimar a entropia conjunta, como no caso do estimador KSG. Contudo, no caso do estimador KSG, uma das entropias marginais acaba por ser sobrestimada. Por exemplo, na Fig. 5.2, a componente da entropia marginal $H(X)$ é estimada com o mesmo vizinho que a componente da entropia conjunta $H(X,Y)$ no ponto z_n. Contudo, a componente de entropia marginal $H(Y)$ no ponto z_n é sobrestimada, pois a distância projetada por $\epsilon_n/2 = \max\{||x_n - x||, ||y_n - y||\}$ é maior que a distância do k-ésimo (segundo) vizinho mais próximo projetada em Y. O segundo vizinho mais próximo do ponto z_n na Fig. 5.2 corresponde ao terceiro mais próximo de Y_n, na projeção em Y, e não ao $(m_y(n)+1)$-ésimo (sexto), como utilizado no método KSG.

No próximo capítulo, alguns dos métodos explanados neste capítulo serão utilizados na estimação de funcionais de medidas de informação envolvendo variáveis aleatórias ou processos aleatórios contínuos em amplitude.

Capítulo 6

Simulação de Métodos de Estimação de Medidas de Informação entre Variáveis Aleatórias Contínuas

Este capítulo mostra o resultado das simulações com alguns dos estimadores vistos no capítulo 5, de medidas de informação para o caso contínuo. Aqui serão apresentadas simulações para estimação de informação mútua com os estimadores KDE, KSG e BI-KSG e de particionamento do suporte. Já para estimação de entropia de transferência, serão comparados os estimadores KSG, KDE, e de particionamento do suporte.

Tanto para as estimativas de informação mútua como para as de entropia de transferência, para determinar o número de segmentos do particionamento (quantis), utilizou-se a recomendação de Paluš [62, 29]

$$Q \leq \sqrt[r+1]{N}, \tag{6.1}$$

em que r é o número de variáveis aleatórias na estimação de informação mútua. O resultado em (6.1) nem sempre é um número inteiro, então fixou-e o valor de Q em

$$Q = \left\lfloor \sqrt[r+1]{N} \right\rfloor. \tag{6.2}$$

6.1 Informação Mútua

A seguir, são apresentados os gráficos de diversos exemplos simulados em que se conhece o valor analítico de informação mútua. Os gráficos mostram o desempenho dos estimadores em termos de acurácia. Nas estimativas KSG e BI-KSG, utilizamos o parâmetro $k = 4$. Nas estimativas KDE, utilizamos o pacote JIDT [50], que já vem com largura do *kernel* (gaussiano) pré-ajustada.

6.1.1 Distribuição Uniforme

Nesta subseção consideramos distribuições uniformes de suporte finito, sendo que no primeiro exemplo simulado há independência entre as variáveis aleatórias, ao passo que no segundo exemplo há dependência.

Exemplo em que há independência entre X e Y

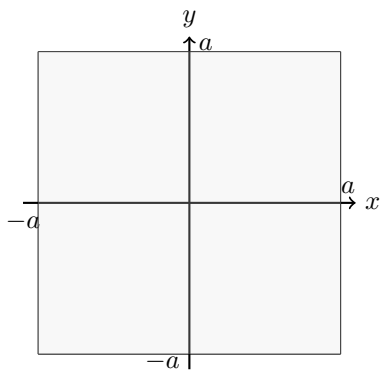


Figura 6.1: Distribuição $f(x, y)$ uniforme na região sombreada - caso de independência entre X e Y.

Neste caso, a distribuição $f(x, y)$ foi feita uniforme na região sombreada da Fig. 6.1. Neste caso, $f(x, y)$ é o produto das distribuições marginais. Logo, X e Y são independentes ($I(X;Y) = 0$). A Fig. 6.2 ilustra o desempenho dos estimadores em termos da média amostral do viés ao passo que a Fig. 6.3 ilustra o desempenho dos estimadores em termos de variância amostral das estimativas.

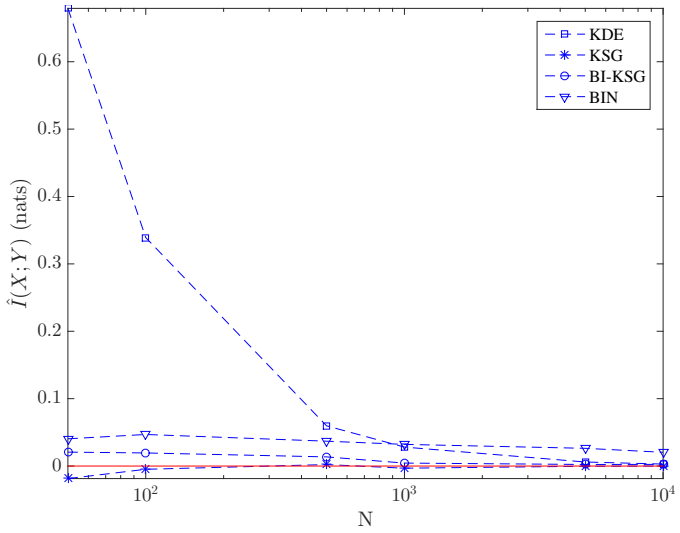


Figura 6.2: Caso uniforme com independência entre X e Y. Médias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$. Curva vermelha indica valor analítico de informação mútua. Curva com quadrado indica média amostral KDE, curva com asterisco indica média amostral das estimativas KSG, curva com círculo indica média amostral das estimativas BI-KSG e curva com triângulo indica média amostral com método do particionamento do suporte. Médias em 50 estimativas.

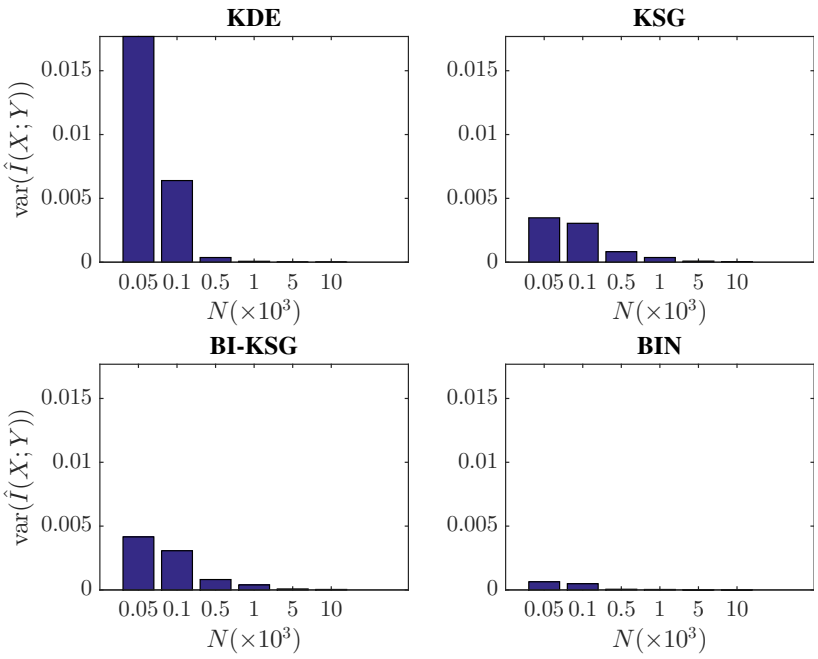


Figura 6.3: Caso uniforme com independência entre X e Y. Variâncias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$, para os métodos KDE, KSG, BI-KSG e de particionamento do suporte ("BIN").

Exemplo em que há dependência entre X e Y

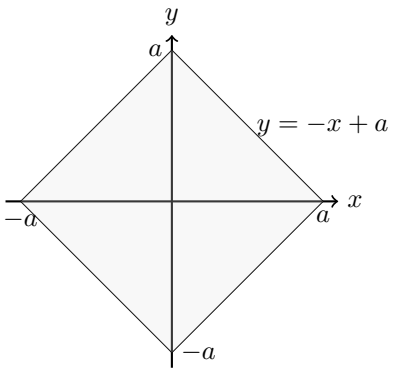


Figura 6.4: Distribuição $f(x, y)$ uniforme na região sombreada - caso de dependência entre X e Y.

Neste exemplo, a distribuição $f(x, y)$ é uniforme na região sombreada da Fig. 6.4. Assim como no caso anterior, é possível encontrar o valor analítico de informação mútua. Primeiramente escrevemos:

$$f(x,y) \quad = \quad \frac{1}{(a\sqrt{2})^2} = \frac{1}{2a^2}$$

Por simetria,

$$\begin{aligned} H(X,Y) &= -4\int_0^a \int_0^{-x+a} \frac{1}{2a^2} \ln \frac{1}{2a^2} dydx \\ &= -4\frac{1}{2a^2} \ln \frac{1}{2a^2} \int_0^a (-x+a)dx \\ &= \ln(2a^2) \end{aligned}$$

Para encontrar a entropia marginal $H(X) = H(Y)$, considera-se, para $-a \leq x \leq 0$:

$$\begin{aligned} f(x) &= \int_{-x-a}^{x+a} \frac{1}{2a^2} dy \\ &= \frac{x+a}{a^2}. \end{aligned}$$

Já para $0 \leq x \leq a$:

$$\begin{aligned} f(x) &= \int_{x-a}^{-x+a} \frac{1}{2a^2} dy \\ &= \frac{-x+a}{a^2}. \end{aligned}$$

Fazendo uma transformação de variáveis $u = \dfrac{x+a}{a^2}$, e por simetria, obtém-se:

$$\begin{aligned} H(X) &= 2\left(-\int_{-a}^{0} \frac{x+a}{a^2} \ln \frac{x+a}{a^2} dx\right) \\ &= 2\left(-\int_0^{1/a} a^2 u \ln u du\right) \\ &= -2a^2 \left[-\frac{1}{4}u^2 + \frac{1}{2}u^2 \ln u\right]_0^{1/a} \\ &= \frac{1}{2} + \ln a \end{aligned}$$

Logo,

$$\begin{aligned} I(X;Y) &= 2\left(\frac{1}{2} + \ln a\right) - \ln(2a^2) \\ &= 1 - \ln 2. \end{aligned}$$

A Fig. 6.5 ilustra o desempenho dos estimadores em termos da média amostral do viés ao passo que a Fig. 6.6 ilustra o desempenho dos estimadores em termos de variância amostral das estimativas.

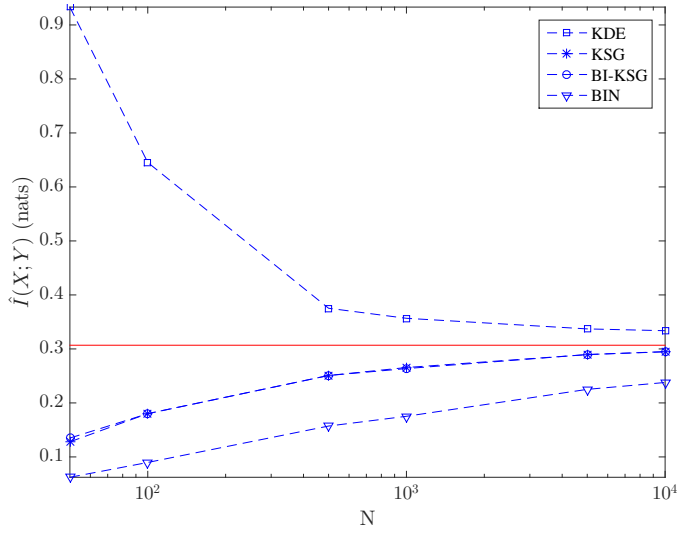


Figura 6.5: Caso uniforme com dependência entre X e Y. Médias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$. Curva vermelha indica valor analítico de informação mútua. Curva com quadrado indica média amostral KDE, curva com asterisco indica média amostral das estimativas KSG, curva com círculo indica média amostral das estimativas BI-KSG e curva com triângulo indica média amostral com método do particionamento do suporte. Médias em 50 estimativas.

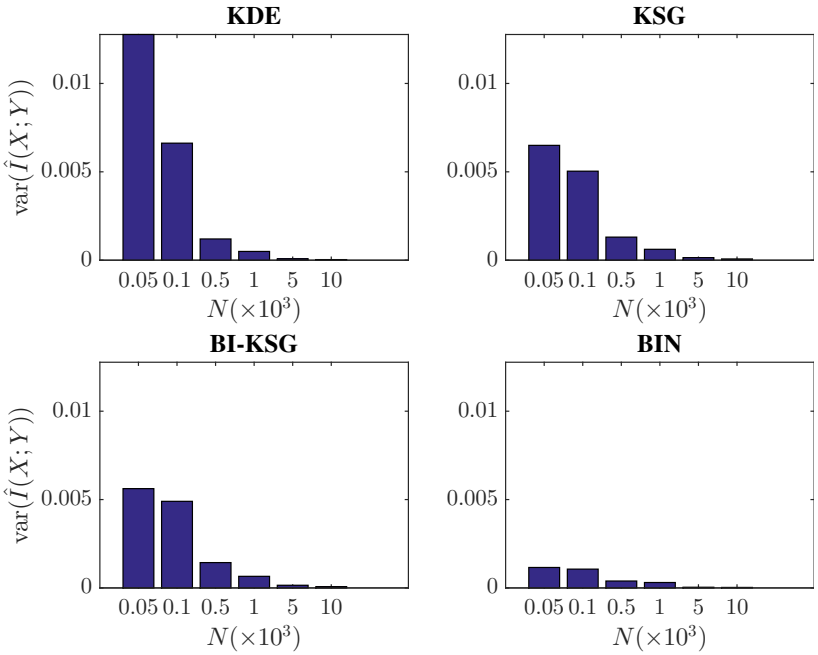


Figura 6.6: Caso uniforme com dependência entre X e Y. Variâncias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$, para os métodos KDE, KSG, BI-KSG e de particionamento do suporte ("BIN").

6.1.2 Distribuição Gaussiana

Nesta subseção consideramos distribuições gaussianas, de suporte infinito, sendo que no primeiro exemplo simulado há independência entre as variáveis aleatórias, ao passo que no segundo exemplo há dependência. A fórmula analítica para a informação mútua neste caso encontra-se na equação (5.22), do capítulo 5.

Exemplo em que há independência entre X e Y

Neste caso $I(X;Y) = 0$. A Fig. 6.7 ilustra o desempenho dos estimadores em termos da média amostral do viés ao passo que a Fig. 6.8 ilustra o desempenho dos estimadores em termos de variância amostral das estimativas.

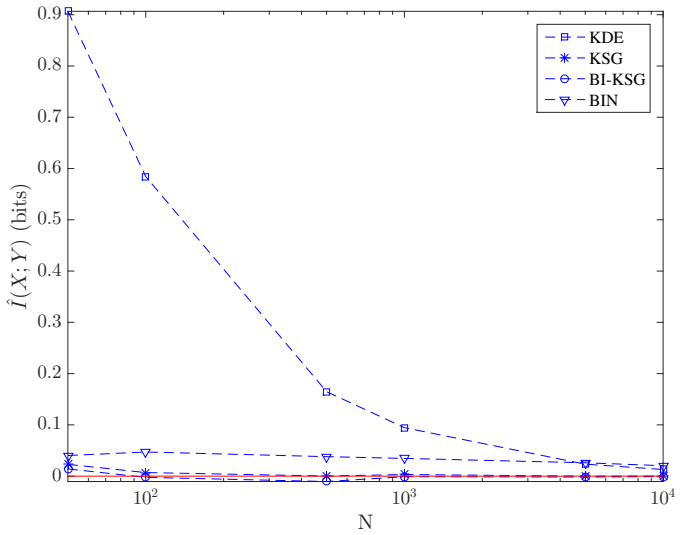


Figura 6.7: Caso gaussiano com independência entre X e Y. Médias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$. Curva vermelha indica valor analítico de informação mútua. Curva com quadrado indica média amostral KDE, curva com asterisco indica média amostral das estimativas KSG, curva com círculo indica média amostral das estimativas BI-KSG e curva com triângulo indica média amostral com método do particionamento do suporte. Médias em 50 estimativas.

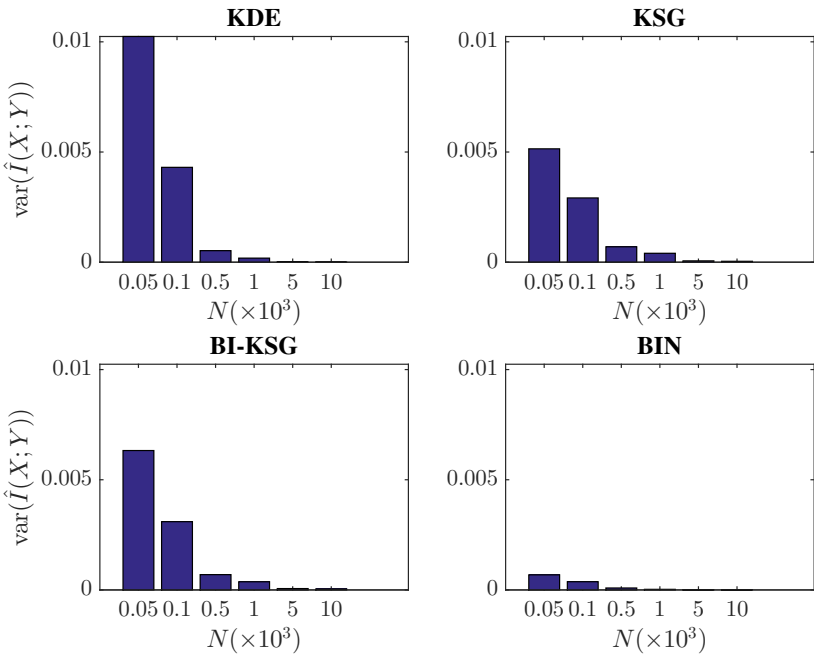


Figura 6.8: Caso gaussiano com independência entre X e Y. Variâncias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$, para os métodos KDE, KSG, BI-KSG e de particionamento do suporte ("BIN").

Exemplo em que há dependência entre X e Y

Neste caso, simularam-se gaussianas com coeficiente de correlação $\rho = 0.6$. A Fig. 6.9 ilustra o desempenho dos estimadores em termos da média amostral do viés ao passo que a Fig. 6.10 ilustra o desempenho dos estimadores em termos de variância amostral das estimativas.

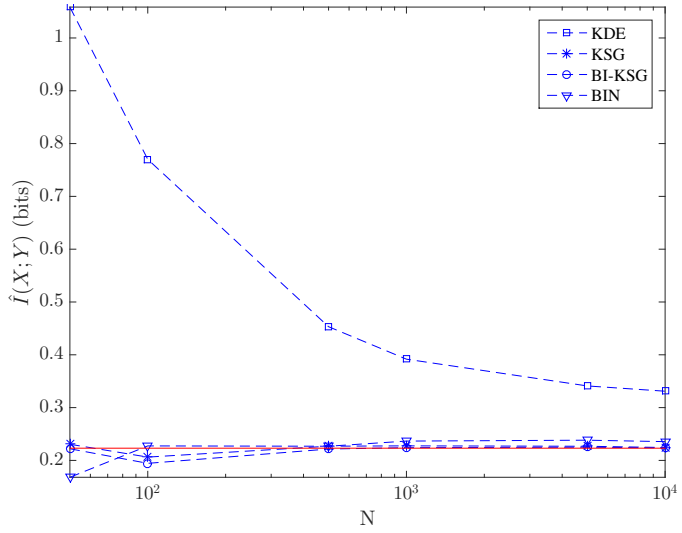


Figura 6.9: Caso gaussiano com dependência entre X e Y ($\rho = 0.6$). Médias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$. Curva vermelha indica valor analítico de informação mútua. Curva com quadrado indica média amostral KDE, curva com asterisco indica média amostral das estimativas KSG, curva com círculo indica média amostral das estimativas BI-KSG, e curva com triângulo indica média amostral com método do particionamento do suporte. Médias em 50 estimativas.

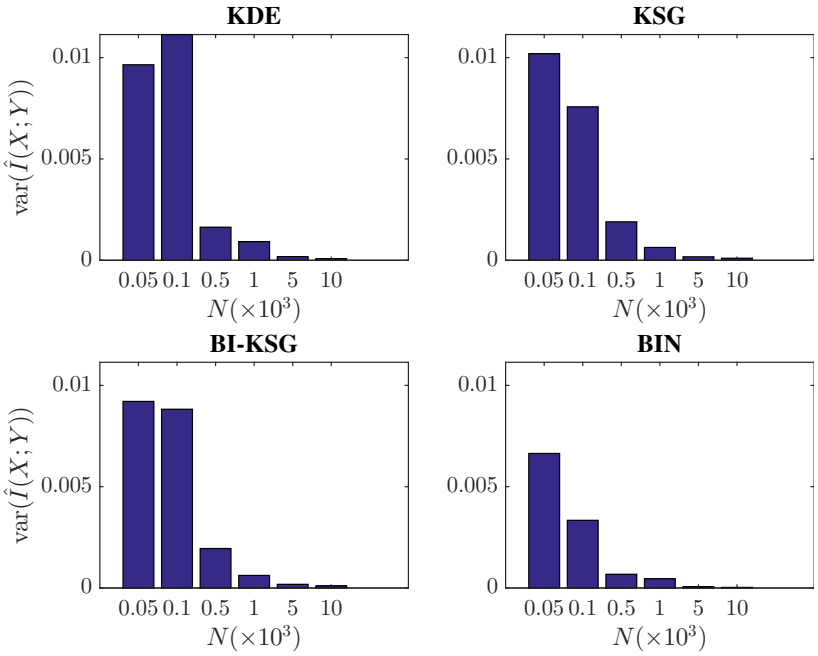


Figura 6.10: Caso gaussiano com dependência entre X e Y ($\rho = 0.6$). Variâncias amostrais das estimativas em função do tamanho amostral $N = \{50, 100, 500, 1000, 5000, 10000\}$, para os métodos KDE, KSG, BI-KSG e de particionamento do suporte ("BIN").

Observou-se que em todos os exemplos simulados, os estimadores KSG e BI-KSG estiveram próximos do valor verdadeiro de informação mútua, em especial quando N

aumenta, apresentando praticamente o mesmo desempenho. Já o estimador de particionamento do suporte mostrou um viés muito reduzido no caso de independência tanto com a distribuição uniforme como com a distribuição gaussiana, mesmo para o maior valor de N testado. O estimador de particionamento do suporte, no caso de dependência, revelou um viés ainda menor com a distribuição gaussiana. Por outro lado, com a dstribuição uniforme com dependência entre X e Y, o estimador de particionamento do suporte mostrou-se bastante enviesado, mesmo para o maior valor de $N = 10000$. O estimador KDE sempre se aproxima do valor verdadeiro conforme N aumenta, mas no panorama geral esteve mais enviesado que os demais. Todos os quatro estimadores têm suas varâncias amostrais reduzidas conforme N aumenta.

Na tabela 6.1 indicam-se os tempos de cada estimativa com os tamanhos $N = \{10000,$ 5000, 1000, 500$\}$. Para os demais tamanhos amostrais, $N = \{50, 100\}$, as estimativas foram da ordem de centésimos de segundos para os métodos KDE, KSG e BI-KSG, e de milésimos de segundo para o método do particionamento do suporte. Observamos que os métodos KSG e BI-KSG foram bem mais lentos que o método KDE, e que todos os métodos foram bem mais lentos que o método de particionamento do suporte. A referência [44] aponta que, computacionalmente, o método KSG gasta a maior parte de seu tempo na procura pelos vizinhos mais próximos. O método BI-KSG foi um pouco mais rápido que o método KSG, o que provavelmente se deve ao uso da função ln no lugar da função ψ (comparando as equações (5.38) e (5.37) do capítulo 5).

$N = 10000$	Método	KDE	KSG	BI-KSG	BIN
	Tempo	2s	9min	7min30s	0.03s
$N = 5000$	Método	KDE	KSG	BI-KSG	BIN
	Tempo	0.33s	2min	1min38s	0.01s
$N = 1000$	Método	KDE	KSG	BI-KSG	BIN
	Tempo	0.02s	4.5s	3.9s	0.004s
$N = 500$	Método	KDE	KSG	BI-KSG	BIN
	Tempo	0.04s	1.16s	1s	0.002s

Tabela 6.1: Tempo aproximado de cada estimativa de informação mútua de tamanho amostral N, caso contínuo. BIN: método de particionamento do suporte.

6.2 Entropia de Transferência

Para avaliar o desempenho dos estimadores de entropia de transferência, utilizou-se o caso analítico descrito em [35], em que:

$$X_{n+1} = \alpha X_n + \eta_n^X, \tag{6.3}$$
$$Y_{n+1} = \beta Y_n + \gamma X_n + \eta_n^Y, \tag{6.4}$$

supondo que $\eta_n^X \sim \mathcal{N}(0,1)$, $\eta_n^Y \sim \mathcal{N}(0,1)$, $\mathbf{X}$ e $\mathbf{Y}$ são conjuntamente estacionários, $\mathbb{E}(X_n) = \mathbb{E}(Y_n) = 0$ e que todos os processos envolvidos são gaussianos.

O valor analítico para a entropia de transferência de $\mathbf{X}$ para $\mathbf{Y}$, $TE(X \to Y)$, quando considerando os índices de passado $l = m = 1$, é dado pela equação:

$$TE(X \to Y) = \frac{1}{2} \ln \frac{\det(K(Y_n, X_n)) \det(K(Y_{n+1}, Y_n))}{\det(K(Y_{n+1}, Y_n, X_n))\sigma_Y^2}, \tag{6.5}$$

em que σ_Y^2 é a variância de Y_n e K denota a matriz de covariâncias. O desenvolvimento dos termos das matrizes de covariância é feito no apêndice E.

Para o caso em que $\alpha = 0.5$, $\beta = 0.6$ e $\gamma = 0.4$, encontra-se:

$$TE(X \to Y) = 0.0923 \; nats. \tag{6.6}$$

Variando o valor de acoplamento γ, encontram-se os gráficos da Fig. 6.11. Observa-se que aparentemente o valor de TE não converge aumentando o acoplamento entre $\mathbf{X}$ e $\mathbf{Y}$ (γ), crescendo indefinidamente.

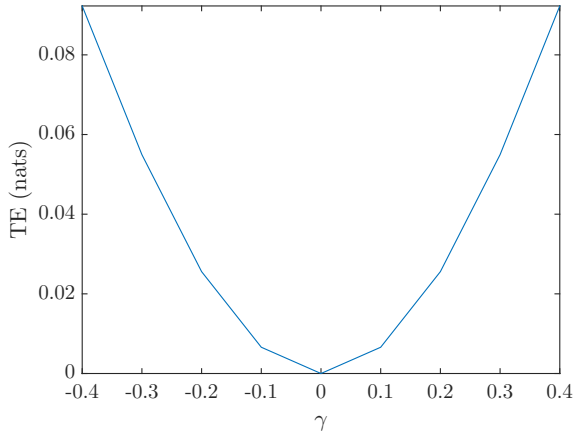


Figura 6.11: Valores analíticos de entropia de transferência, TE, variando-se o termo de acoplamento γ.

É importante ressaltar que as estimativas de entropia de transferência realizadas aqui consideram que os processos $\mathbf{X}$ e $\mathbf{Y}$ são ergódicos. Isto significa que as estimativas de entropia de transferência são feitas com médias temporais. Na implementação, tanto do método KDE quanto do método KSG, foi utilizado o pacote JIDT [50]. Já com o método do particionamento do suporte, utilizou-se o número de quantis:

$$Q = \left\lfloor \sqrt[m+l+1]{N} \right\rfloor = \left\lfloor \sqrt[3]{N} \right\rfloor, \tag{6.7}$$

visto que as dimensões dos vetores na equação

$$TE(X \to Y) \;\; = \;\; I(Y_n, Y_{n-1}; X_{n-1}) - I(Y_{n-1}; X_{n-1}), \tag{6.8}$$

são iguais aos índices de passado dos processos $\mathbf{X}$ e $\mathbf{Y}$, $l = 1$ e $m = 1$, respectivamente. Em seguida, a estimação de entropia de transferência foi implementada através da estimação de informações mútuas, valendo-se da equação (6.8). As Fig. 6.12 e Fig. 6.13 ilustram o desempenho dos estimadores com o método KDE, o método KSG e o método do particionamento do suporte.

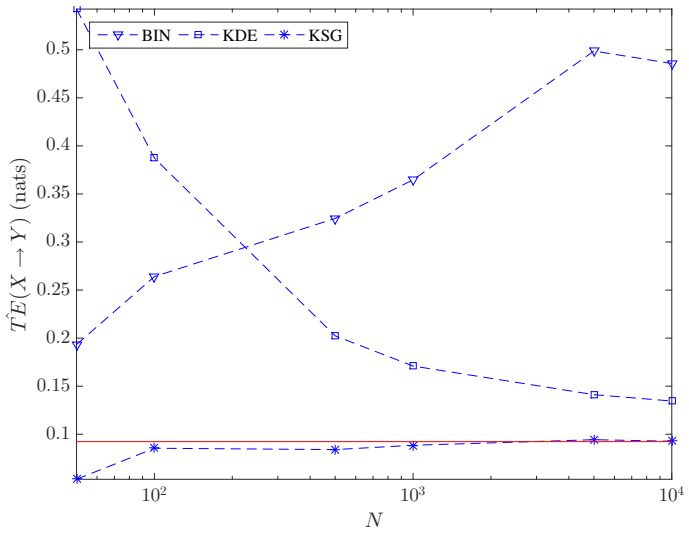


Figura 6.12: Caso simulado de entropia de transferência com valor conhecido de $TE(X \to Y) = 0.0923 nats$. Médias amostrais das estimativas em função da duração $N = \{50, 100, 500, 1000, 5000, 10000\}$ dos processos. Curva vermelha indica valor analítico de entropia de transferência. Curva com quadrado indica média amostral das estimativas KDE, curva com asterisco indica média amostral das estimativas KSG e curva com triângulo indica média amostral das estimativas de particionamento do suporte. Médias em 50 estimativas.

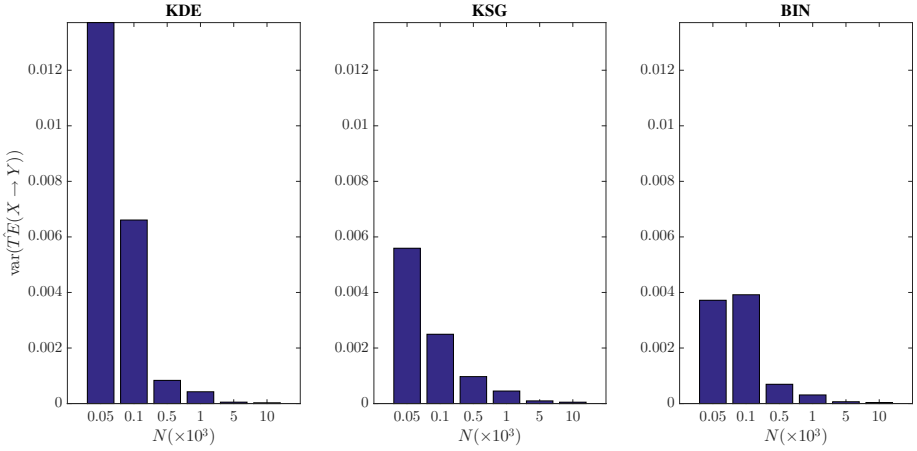


Figura 6.13: Caso simulado de entropia de transferência com valor conhecido de $TE(X \to Y) = 0.0923 nats$. Variâncias amostrais das estimativas em função da duração $N = \{50, 100, 500, 1000, 5000, 10000\}$ dos processos, com os métodos KDE, KSG e de particionamento do suporte ("BIN"). Variâncias em 50 estimativas.

Observou-se neste caso de estimação de entropia de transferência que os métodos KSG e KDE se aproximam do valor verdadeiro conforme N cresce, mas que essa convergência é mais rápida no método KSG. Por outro lado, infelizmente o estimador do particionamento do suporte não convergiu para o valor verdadeiro de entropia de transferência, mesmo com o auxílio da regra da equação (6.7). Para os três estimadores, a variância diminuiu conforme N aumentou.

Utilizando o pacote JIDT, o desempenho em termos de velocidade do método KSG melhorou bastante. Para o exemplo estudado de entropia de transferência, com amostras dos processos $\mathbf{X}$ e $\mathbf{Y}$ com duração $N = 10000$, o tempo de cada estimativa foi aproximadamente 0.26s, similar ao tempo do método do particionamento do suporte, que foi de apenas 0.22s. O método KDE, utilizando o pacote JIDT, foi o mais lento, levando 2.22s para cada estimativa. Deste modo, é possível afirmar que os estimadores baseados na busca pelos vizinhos mais próximos, KSG e BI-KSG, que foram muito lentos nas simulações de informação mútua da seção anterior, podem ser consideravelmente melhorados com o uso de código em *Java*, como o JIDT implementa.

Capítulo 7

Estimação de Informação Direcional entre Processos Contínuos em Amplitude com Estimadores de Jiao

No capítulo 5, apresentaram-se alguns estimadores de medidas de informação para processos assumindo valores contínuos, enquanto no capítulo 6 o desempenho de alguns desses estimadores foi comparado, para as medidas de informação mútua e entropia de transferência. Neste capítulo, investiga-se o uso de um dos estimadores de informação direcional de Jiao, introduzidos no capítulo 3, para estimar informação direcional entre processos estocásticos contínuos em amplitude através da discretização desses valores contínuos.

A discretização é feita e analisada de acordo com três métodos diferentes: equidistante, equipovoado e simbólico. O método equidistante consiste em segmentar o suporte dos processos em L segmentos equidistantes, de acordo com a distância euclidiana, do valor mínimo para o máximo. Os métodos equipovoado e simbólico foram introduzidos no capítulo 5. O método equipovoado segmenta o suporte, mas em L segmentos equipovoados (ou aproximadamente equipovoados). O método simbólico consiste na ordenação dos k valores consecutivos do processo, representando sua ordem de transição por um número que representa uma das $k!$ permutações possíveis, como descrito na seção 5.4. Por exemplo, na sequência:

$$X^7 = [0.5\ 0.75\ -0.1\ -0.23\ 0.05\ 0.52\ 0.49],$$

se for escolhido $k = 2$, há as seguintes transições:

$$[01\ 10\ 10\ 01\ 01\ 10].$$

Rotulando o símbolo 01 como 1 e o símbolo 10 como 2, obtém-se a seguinte sequência discretizada (começando no índice $n = 2$):

$$\tilde{X}_2^7 = [1\ 2\ 2\ 1\ 1\ 2].$$

Por outro lado, se for escolhido $k = 3$, as correspondentes transições são:

$$[120\ 210\ 102\ 012\ 021].$$

Rotulando cada transição como na tabela 7.1, obtém-se a seguinte sequência discretizada (começando no índice $n = 3$):

$$\tilde{X}_3^7 = [4\ 6\ 3\ 1\ 2].$$

Tabela 7.1: Valores de permutação

Transição	Valor Discretizado
012	1
021	2
102	3
120	4
201	5
210	6

Portanto, no método simbólico, se foi escolhido o parâmetro k, o processo contínuo é discretizado em $L = k!$ símbolos.

Para comparar o desempenho dos três métodos de discretização na estimação de informação direcional com o estimador de Jiao, utiliza-se o cálculo de um exemplo de base em que um processo **X** causa outro processo **Y** (**X** e **Y** contínuos em amplitude).

7.1 Exemplo de Base

O **X** é um processo aleatório gaussiano i.i.d., com média zero e variância 1 (σ_X^2=1), ao passo que o processo **Y** é dado de acordo com a equação:

$$Y_n = \beta X_{n-2} + Z_n \tag{7.1}$$

em que β é um parâmetro de acoplamento e Z_n também é um processo gaussiano i.i.d. com média zero e variância 1 ($\sigma_Z^2 = 1$). O valor verdadeiro de informação direcional é calculado como segue, em termos de $H(Y^N||X^N)$ e $H(Y^N)$ separadamente:

$$\begin{aligned}
\frac{1}{N}H(Y^N||X^N) &= \frac{1}{N}\sum_{n=1}^{N} H(Y_n|Y^{n-1}X^n) \\
&= \frac{1}{N}\sum_{n=1}^{N} H(\beta X_{n-2} + Z_n|Y^{n-1}X^n) \\
&= \frac{1}{N}\sum_{n=1}^{N} H(\beta X_{n-2} + Z_n|X^{n-2}) \\
&= \frac{1}{N}\sum_{n=1}^{N} H(Z_n) \\
&= \frac{1}{2}\log(2\pi e\sigma_Z^2) \\
&= \frac{1}{2}\log(2\pi e),
\end{aligned}$$

e

$$\begin{aligned}
\frac{1}{N}H(Y^N) &= \frac{1}{N}\sum_{n=1}^{N} H(Y_n|Y^{n-1}) \\
&= H(Y_n),
\end{aligned}$$

porque **Y** não depende de seu próprio passado.

Para calcular $H(Y_n)$, como Y_n é gaussiana com média 0, encontra-se primeiro sua variância:

$$\begin{aligned}\mathrm{var}(Y_n) &= \mathbb{E}(\beta X_{n-2} + Z_n)^2 \\ &= \mathbb{E}(\beta^2 X_{n-2}^2 + 2\beta X_{n-2} Z_n + Z_n^2) \\ &= \beta^2 \mathbb{E}(X_{n-2}^2) + \sigma_Z^2 \\ &= \beta^2 \sigma_X^2 + \sigma_Z^2 = \beta^2 + 1.\end{aligned}$$

Logo,

$$\begin{aligned}\frac{1}{N} H(Y^N) &= \frac{1}{2}\log(2\pi e(1+\beta^2)), \text{ e} \\ I_N(X \to Y) &= \frac{1}{2}\log(1+\beta^2).\end{aligned}$$

7.2 Simulação

Considerando o exemplo de base, simularam-se 50 amostras de $\mathbf{X}$, $\mathbf{Y}$ e $\mathbf{Z}$ com duração $N = 10^5$, parâmetro de profundidade da árvore $D = 2$, para cada parâmetro de acoplamento β entre -1 e 1, com passo 0.2. Para todos os processos de estimação, selecionaram-se os níveis de discretização $L = 2$ ou $L = 6$. Além disso, executaram-se simulações com os níveis de discretização $L = 4$ com os métodos de discretização equidistante e equipovoado. O método simbólico não foi possível com $L = 4$ porque não há inteiro k tal que $k! = 4$. O estimador de Jiao utilizado foi o "E4".

As Fig. 7.1, Fig. 7.2 e Fig. 7.3 ilustram os resultados para $L = 2$. As Fig. 7.4, Fig. 7.5 e Fig. 7.6 ilustram os resultados para $L = 6$. As Fig. 7.7 e 7.8 ilustram os demais resultados para $L = 4$. Em todas as figuras, curvas pontilhadas vermelhas indicam a taxa de informação direcional analítica, enquato curvas contínuas azuis indicam medianas das estimativas de taxa de informação direcional e curvas tracejadas azuis indicam de 10% a 90% das estimativas.

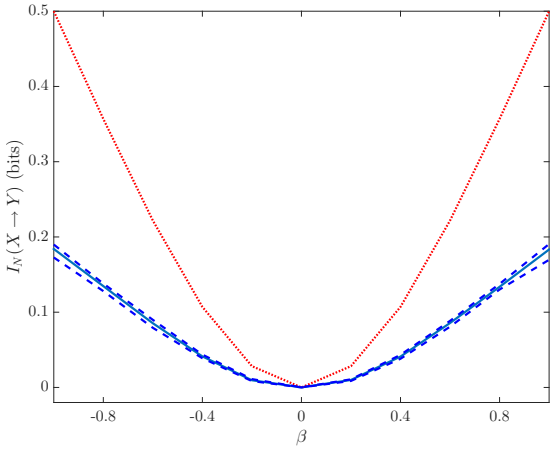


Figura 7.1: Método equidistante, $L = 2$

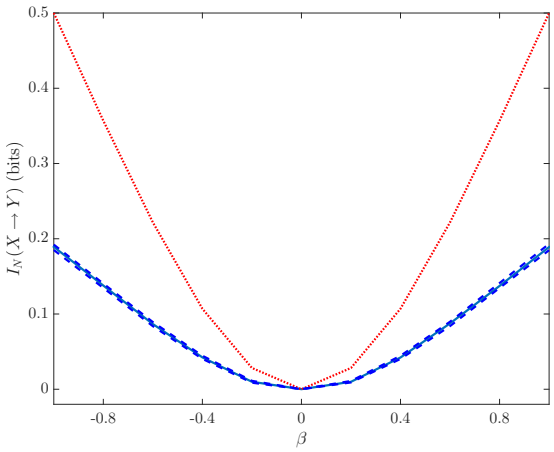


Figura 7.2: Método equipovoado, $L = 2$

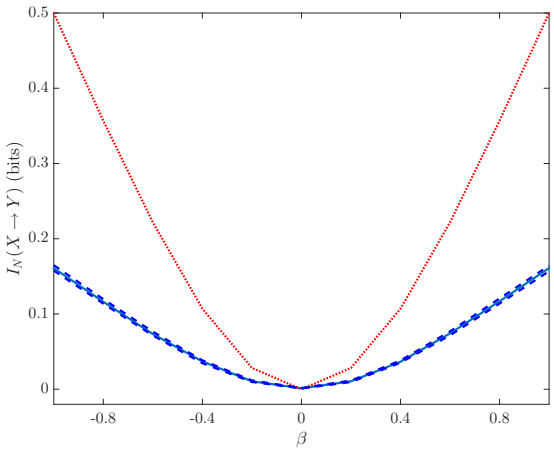


Figura 7.3: Método simbólico, $L = 2$

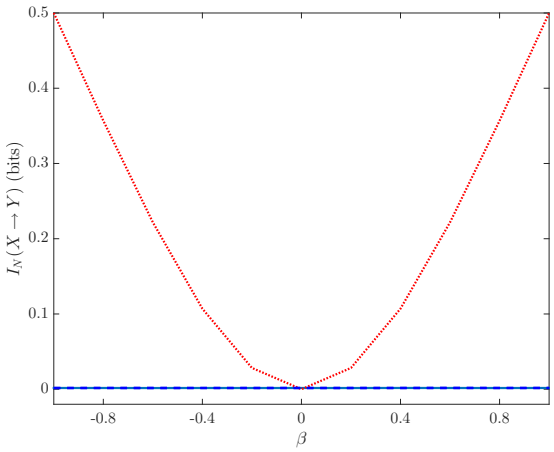


Figura 7.4: Método equidistante, $L = 6$

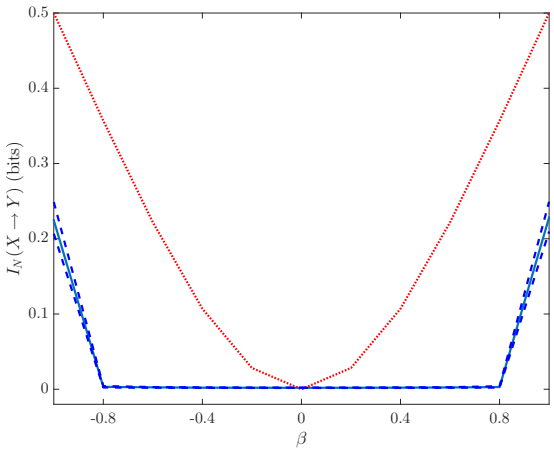


Figura 7.5: Método equipovoado, $L = 6$

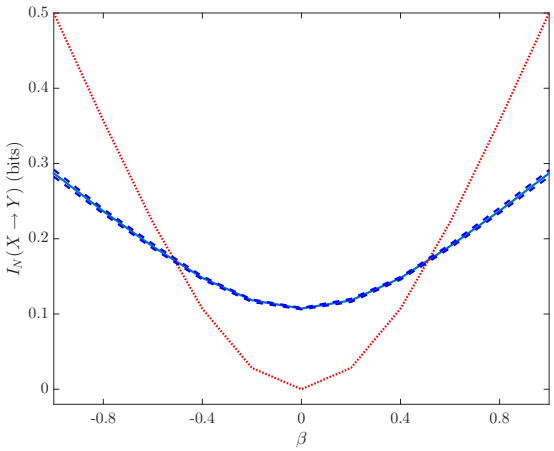


Figura 7.6: Método simbólico, $L = 6$

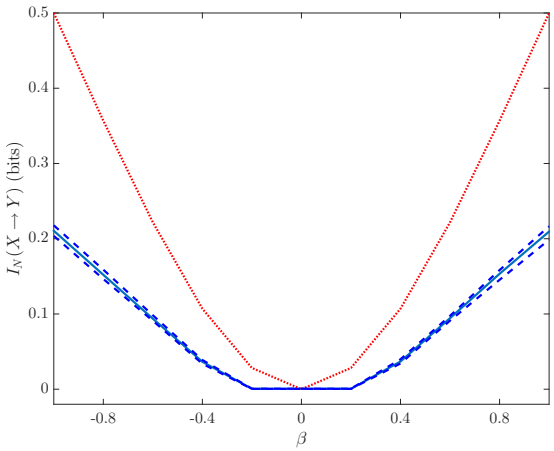


Figura 7.7: Método equidistante, $L = 4$

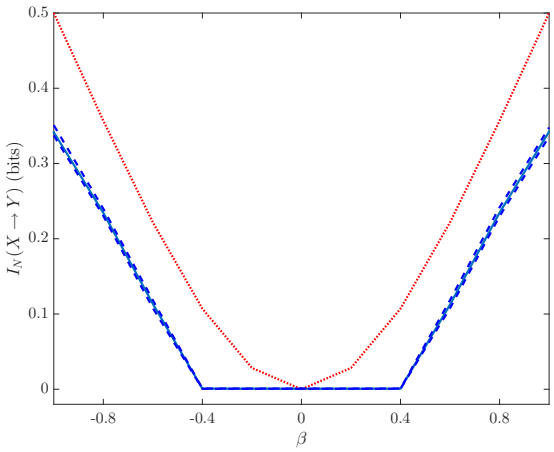


Figura 7.8: Método equipovoado, $L = 4$

Algumas características dos métodos propostos podem ser observadas. Primeiramente, no que concerne às estimativas, observa-se que para $L = 2$, todos os três métodos apresentam o mesmo desempenho em termos de acurácia. Eles sempre subestimam as taxas de informação direcional e apresentam pequena variância amostral, como indicado pelas linhas tracejadas. Por outro lado, para $L = 6$, observou-se que os métodos equidistante e equipovoado em geral não capturam a causalidade para todos os valores do parâmetro β, exceto com o método equipovoado com $\beta = 1$. Isto pode ter ocorrido porque as probabilidades da árvore de contexto diminuem bastante com um alfabeto maior (L). Infelizmente, o método simbólico sobrestima a taxa de informação direcional para pequenos valores de β e subestima a taxa de informação direcional para valores maiores de β. Contudo, as estimativas de taxa de informação direcional neste caso acompanham o comportamento dos valores analíticos. Para os casos em que $L = 4$, observamos que o método equipovoado supera o método equidistante, mas não captura causalidade para valores de β pequenos ($\beta = 0.2, 0.4$) e ainda subestima a taxa de informação direcional. Contudo, essa subestimação com o método equipovoado é menor que a subestimação com o método equidistante.

No que concerne o tempo de estimação, todos os métodos de discretização consomem pouco tempo, com um pequeno incremento no método simbólico. Entretanto, como já observou-se no capítulo 4, a estimação de informação direcional com o estimador de Jiao é demorada. Cada realização amostral dos processos levou aproximadamente 35s com $L = 2$ ou $L = 4$, enquanto cada realização amostral levou aproximadamente 50s com $L = 6$.

Com o propósito de avaliar a presença de detecção espúria de causalidade entre processos que não apresentavam qualquer causalidade, também simularam-se 50 amostras de dois processos gaussianos i.i.d. com duração $N = 10^5$ e estimaram-se taxas de informação direcional de acordo com os três métodos de discretização. Neste caso, $I_N(X^N \to Y^N) = 0$. Novamente, ajustou-se a profundidade da árvore de contexto $D = 2$, e a discretização foi feita com $L = 2, 4$ ou 6. A tabela 7.2 mostra as medianas das estimativas de taxa de informação direcional neste caso.

Aparentemente, há uma tendência de que as medianas das estimativas apresentem um pequeno aumento com o uso de alfabetos maiores de discretização. No entanto, o único caso com estimativas comparavelmente grandes, apesar da inexistência de causalidade neste caso, foi quando utilizando o método simbólico com $L = 6$.

Por fim, deve-se ressaltar aqui que existem muitas outras formas possíveis de discre-

Tabela 7.2: Medianas da taxa de informação direcional de acordo com os 3 métodos de discretização e níveis (L), quando o valor analítico $I_N(X^N \to Y^N) = 0$.

Método de discretização / Níveis	$L = 2$	$L = 4$	$L = 6$
Equidistante	$\approx 10^{-5}$	$\approx 10^{-4}$	$\approx 10^{-3}$
Equipovoado	$\approx 10^{-5}$	$\approx 10^{-4}$	$\approx 10^{-3}$
Simbólico	$\approx 10^{-3}$	-	$\approx 10^{-1}$

tização. Por exemplo, na referência [34], a causalidade em bolsas de valores é estimada pela discretização em três valores. O valor -1 indicou que a bolsa de valor reduziu em um dia mais de 0.8%, o valor 1 indicou que a bolsa de valores aumentou em um dia mais de 0.8%, enquanto o valor 0 indicou que o valor absoluto da mudança foi menos de que 0.8%. Contudo, este método apresenta a desvantagem de ter que escolher um valor apropriado para a mudança absoluta (o valor 0.8% na referência [34]). Trabalhos futuros podem verificar um modo de usar este método de uma maneira genérica.

Capítulo 8

Estimação de Medidas de Informação entre Variáveis Aleatórias Mistas

A necessidade de relacionar grandezas discretas e contínuas é bem comum em ciências. Neste capítulo investigamos o desempenho de dois estimadores de informação mútua para o caso em que as variáveis aleatórias são mistas, isto é, algumas das variáveis aleatórias assumem valores discretos e outras assumem valores contínuos. Além disso, analisamos o comportamento destes dois estimadores para o caso particular de detecção de causalidade, em estimativas de entropia de transferência.

8.1 Estimadores de Informação Mútua

8.1.1 Estimador de Particionamento do Suporte

Esse estimador foi extensivamente utilizado em neurociências [69, 7, 10, 37]. Ele consiste essencialmente em particionar o suporte da variável aleatória contínua Y em segmentos equiprováveis, método citado no capítulo 5. Já a função massa de probabilidade da variável aleatória discreta X, $\hat{P}(x)$, é estimada pelo método *plug-in*.

De maneira mais formal, sejam (X_1^N, Y_1^N) as N amostras i.i.d. geradas a partir de uma densidade de probabilidade subjacente $f(x, y) = P(x)f(y|x)$. As amostras Y_1^N são ordenadas crescentemente, e Q intervalos equipovoados são escolhidos [86]:

$$\{\tilde{Y}\}_{i=1,2,\ldots,Q} = \{(-\infty, Y_{(1)}], (Y_{(1)}, Y_{(2)}], \ldots, (Y_{(Q-1)}, \infty)\}, \tag{8.1}$$

em que $Y_{(i)}$ é o i-ésimo Q-quantil da amostra Y_1^N. A função de massa de probabibilidade estimada de $\tilde{Y}$ será

$$\hat{P}(i) = \frac{n_i}{N} \approx \frac{Q}{N}, \tag{8.2}$$

em que n_i conta quando ocorreu $Y \in \tilde{Y}_i$ na amostra Y_1^N. Já a função de massa de probabibilidade estimada de X será

$$\hat{P}(x) = \frac{n_x}{N}, \tag{8.3}$$

em que n_x conta o número de ocorrências $X = x$ na amostra X_1^N. Similarmente, a função massa de probabilidade conjunta de $(X, \tilde{Y}_i)$ será

$$\hat{P}(x, i) = \frac{n_{x,i}}{N}, \tag{8.4}$$

em que $n_{x,i}$ conta quantas vezes ocorreu $(X = x, Y \in \tilde{Y}_i)$ conjuntamente na amostra (X_1^N, Y_1^N).

Assim, a estimativa de informação mútua entre X e Y fica

$$\hat{I}(X;Y) = \sum_{\mathcal{X}} \sum_{i=1}^{Q} \hat{P}(x, i) \log \frac{\hat{P}(x, i)}{\hat{P}(x)\hat{P}(i)}. \tag{8.5}$$

Uma grande questão na utilização deste método, como já descrito no capítulo 5, é como escolher o número apropriado de quantis Q para uma estimativa acurada de informação mútua.

8.1.2 Estimador de Ross

Ross [70], baseado no trabalho de Kraskov *et al.* [44], propôs um estimador para o caso misto, isto é, entre uma variável aleatória discreta e outra contínua. Relembra-se do capítulo 2 que neste caso a informação mútua é dada pela equação (2.16). Como já existe o estimador de Kozachenko-Leonenko para a entropia marginal, agora torna-se necessário um estimador para a entropia condicional de Y dado um valor discreto de X. Isso é feito novamente escolhendo-se diferentes ks para cada estimativa de entropia. Para a entropia marginal escolhe-se o j-ésimo vizinho mais próximo e para a entropia condicionada escolhe-se, dentre os pontos que tiveram um mesmo valor de $X_n = x_n$, o k-ésimo vizinho mais próximo. Ross tenta diminuir o viés de suas estimativas através da escolha de k e j de modo que ambas estimativas de entropia utilizem o mesmo vizinho. A Fig. 8.1 especifica como é feita a escolha do vizinho.

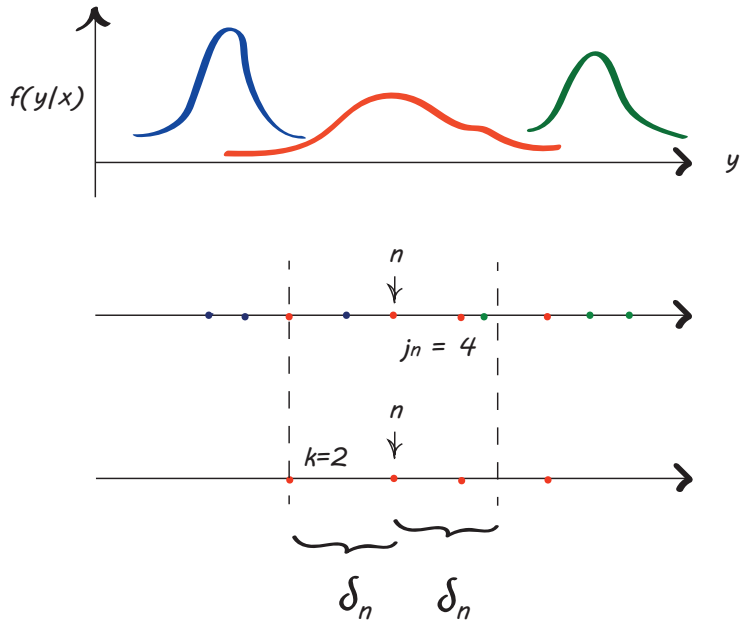


Figura 8.1: Ilustração do estimador de Ross. Na linha superior, exemplos de distribuições condicionadas $f(Y|X)$, em que X é a variável discreta assumindo 3 valores simbolizados pelas curvas azul, vermelha e verde. Na linha do meio, um conjunto de pares de dados (X, Y), em que os valores de Y são representados pela posição dos pontos no eixo y e os valores de X são representados pelas cores destes pontos. Na linha inferior, mostra-se o ponto n analisado indicado por uma seta vertical e o 2º vizinho mais próximo (dado que o valor de X é "vermelho"). Utilizando $k = 2$, percebe-se que o 2º vizinho de n na linha inferior é o 4º na linha do meio - que considera todos os valores de X. Linhas tracejadas mostram a distância δ_n do ponto n ao 2º vizinho. Neste exemplo, $N = 10$, $k = 2$, e para este ponto n, $N_{X_n} = 4$ e $j_n = 4$ (incluindo o vizinho à distância δ_n).

Portanto, o estimador usa, para cada ponto n, a distribuição das distâncias do k-ésimo vizinho mais próximo da variável contínua, para um dado valor da variável discreta (equação (8.6)):

$$\hat{I}_n = \psi(N) - \psi(N_{X_n}) + \psi(k) - \psi(j_n). \tag{8.6}$$

N é o tamanho da amostra, N_{X_n} é o número de pontos para os quais a variável discreta é $X_n = x_n$ e k é o número escolhido de vizinhos. O termo j_n corresponde ao número de pontos a uma distância dada pelo k-ésimo vizinho de n dentre os pontos N_{X_n}.

A estimativa de informação mútua é obtida pelo cálculo da média de $\hat{I}_n$ sobre toda a amostra:

$$\hat{I} = \frac{1}{N} \sum_{n=1}^{N} \hat{I}_n \tag{8.7}$$

8.1.3 Simulações

Foram realizadas simulações para verificar o funcionamento deste estimador, em particular comparando-o com o processo de estimação que usa o método do particionamento do suporte da variável aleatória contínua. Para tanto, utilizou-se o toolbox da referência [53], com a correção de viés QE (extrapolação quadrática). Foi estabelecida uma variável aleatória discreta X, assumindo valores 1 ou 2, que alterava a distribuição de uma variável aleatória contínua Y segundo uma distribuição condicionada uniforme:

$$f(Y|X) = \begin{cases} \frac{1}{2}\text{ret}_2(y-1), \text{ se } x=1, \\ \frac{1}{2}\text{ret}_2(y-2), \text{ se } x=2, \end{cases} \tag{8.8}$$

em que ret_2 é um pulso retangular de largura 2, centrado em zero.

A informação mútua verdadeira para esse exemplo pode ser calculada analiticamente como segue

$$\begin{aligned}
X &\in \{1,2\} \\
f(y|x) &= \frac{1}{2}\text{ret}_2(y-x) \\
f(y) &= \sum P(x)f(y|x) \\
f(y|x) &= \begin{cases} 1/2 & \text{se } 0 \le y \le 2 \text{ e } x = 1 \\ 1/2 & \text{se } 1 \le y \le 3 \text{ e } x = 2 \end{cases} \\
f(y) &= \begin{cases} 1/4 & \text{se } 0 \le y < 1 \\ 1/2 & \text{se } 1 \le y < 2 \\ 1/4 & \text{se } 2 \le y \le 3 \end{cases} \\
I(X;Y) &= H(Y) - H(Y|X) \\
H(Y) &= -\int_0^1 \frac{1}{4}\log\frac{1}{4}dy - \int_1^2 \frac{1}{2}\log\frac{1}{2} - \int_2^3 \frac{1}{4}\log\frac{1}{4} \\
&= \frac{3}{2} \\
H(Y|X) &= -\int P(x)f(y|x)\log f(y|x)dy \\
&= \sum P(x)\int f(y|x)\log f(y|x)dy \\
&= -\frac{1}{2}\int_0^2 \frac{1}{2}\log\frac{1}{2}dy - \frac{1}{2}\int_1^3 \frac{1}{2}\log\frac{1}{2}dy \\
&= 1 \\
I(X;Y) &= \frac{3}{2} - 1 \\
&= 0.5\text{bit}
\end{aligned}$$

Os resultados das simulações encontram-se expressos na Fig. 8.2, que revelam um desempenho mais acurado do estimador de Ross sobre o do método do particionamento do suporte. O maior problema de se utilizar o método do particionamento do suporte parece estar na dificuldade de escolher o número adequado de segmentos. A regra observada no capítulo 5, de utilizar um número de segmentos $Q \le \sqrt[3]{N} = \sqrt[3]{400} \approx 7$ leva a estimativas com valores abaixo do esperado para este caso.

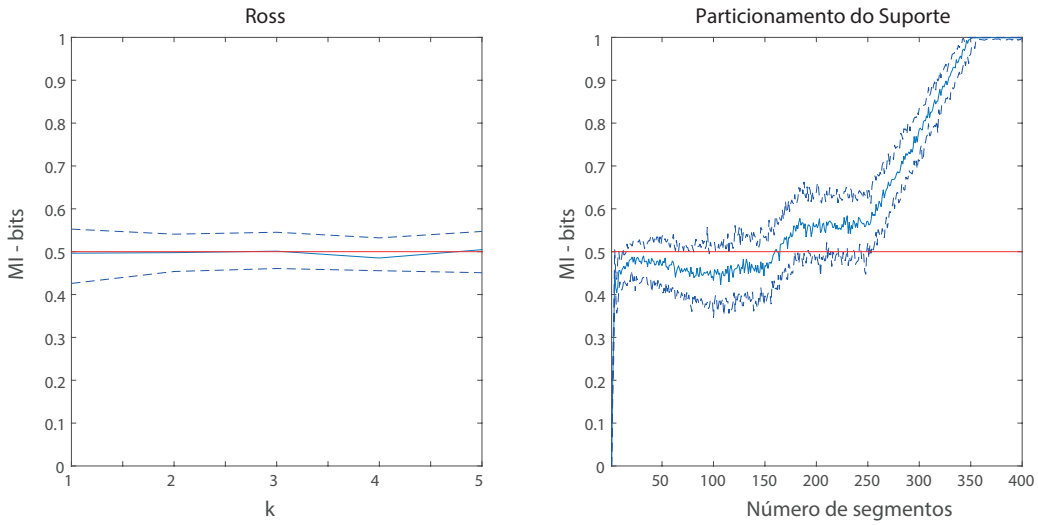


Figura 8.2: Estimativas de acordo com o método de Ross e o método do particionamento do suporte, em bits, caso uniforme-uniforme. À esquerda, estimação de informação mútua pelo método de Ross, em função do número de vizinhos k. À direita, estimação de informação mútua pelo método do particionamento do suporte com correção de viés QE, como uma função do número de segmentos usados. O tamanho amostral foi $N = 400$, linhas vermelhas indicam o verdadeiro valor de informação mútua, ao passo que linhas azuis indicam a mediana de informação mútua para 50 conjuntos de dados de tamanho 400 cada. O intervalo entre linhas azuis tracejadas indicam de 10% a 90% das estimativas.

Outro exemplo simulado foi o que X assume os valores discretos $X \in \{1, 2\}$ uniformemente, ao passo que a distribuição condicionada de Y ao valor de X é:

$$
\begin{aligned}
f(y|x) &= \begin{cases} \phi_1 = \dfrac{1}{\sqrt{2\pi}} e^{-\dfrac{(y-1)^2}{2}} & \text{se } x = 1 \\ \phi_2 = \dfrac{1}{\sqrt{2\pi}} e^{-\dfrac{(y-2)^2}{2}} & \text{se } x = 2 \end{cases} \\
f(y) &= \frac{1}{2\sqrt{2\pi}} e^{\dfrac{(y-1)^2}{2}} + \frac{1}{2\sqrt{2\pi}} e^{-\dfrac{(y-2)^2}{2}}
\end{aligned}
$$

É possível encontrar $H(Y|X)$:

$$
\begin{aligned}
H(Y|X) &= -\sum_x \int P(x) f(y|x) \ln f(y|x) dy \\
&= -\sum_x P(x) \int f(y|x) \ln f(y|x) dy \\
&= -\frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\dfrac{-(y-1)^2}{2}} \ln \left(\frac{1}{\sqrt{2\pi}} e^{\dfrac{-(y-1)^2}{2}} \right) dy - \\
&\quad \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\dfrac{-(y-2)^2}{2}} \ln \left(\frac{1}{\sqrt{2\pi}} e^{\dfrac{-(y-2)^2}{2}} \right) dy
\end{aligned}
\tag{8.9}
$$

em que

$$
\begin{aligned}
-\frac{1}{2}\int_{-\infty}^{\infty} \quad & \frac{1}{\sqrt{2\pi}}e^{-\frac{(y-1)^2}{2}} \ln\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-1)^2}{2}}\right) dy = \\
&= -\frac{1}{2}\int \left(\phi_1 \ln(1/\sqrt{2\pi}) + \phi_1(-(y-1)^2/2)\right) dy \\
&= \frac{1}{2}\ln(\sqrt{2\pi}) + \frac{1}{2}\mathbb{E}\left(\frac{(Y-1)^2}{2}\right) \\
&= \frac{1}{2}\frac{1}{2}\ln(2\pi) + \frac{1}{2}\frac{1}{2}\mathbb{E}(Y-\mathbb{E}Y)^2 \\
&= \frac{1}{4}\ln(2\pi) + \frac{1}{4}\mathrm{var}(Y) \\
&= \frac{1}{4}\ln(2\pi e),
\end{aligned}
$$

e, similarmente,

$$
-\frac{1}{2}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{\frac{-(y-2)^2}{2}} \ln\left(\frac{1}{\sqrt{2\pi}}e^{\frac{-(y-2)^2}{2}}\right) dy \quad = \quad \frac{1}{4}\ln(2\pi e),
$$

logo,

$$
H(Y|X) = \frac{1}{2}\ln(2\pi e)\ nats. \tag{8.10}
$$

O valor da entropia $H(Y)$, foi encontrado integrando numericamente a função

$$
-f(y)\ln f(y). \tag{8.11}
$$

A razão para isso foi que o gráfico da função em (8.11) mostrou valores nulos para valores menores que -5 ou maiores que 10. O valor aproximado obtido foi:

$$
H(Y) \approx 1.5304\ nats. \tag{8.12}
$$

Portanto, o valor aproximado de informação mútua para este caso foi

$$
I(X;Y) = H(Y) - H(Y|X) \approx 0.111\ nats. \tag{8.13}
$$

A Fig. 8.3 mostra o gráfico de $-f(y)\ln f(y)$ e a Fig. 8.4 mostra as estimativas de informação mútua de acordo com o método do particionamento do suporte e com o método de Ross.

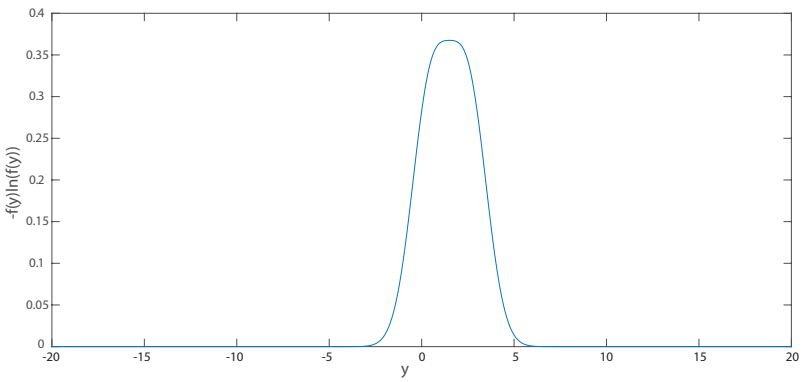


Figura 8.3: Gráfico de $-f(y)\ln f(y)$ em função dos valores y.

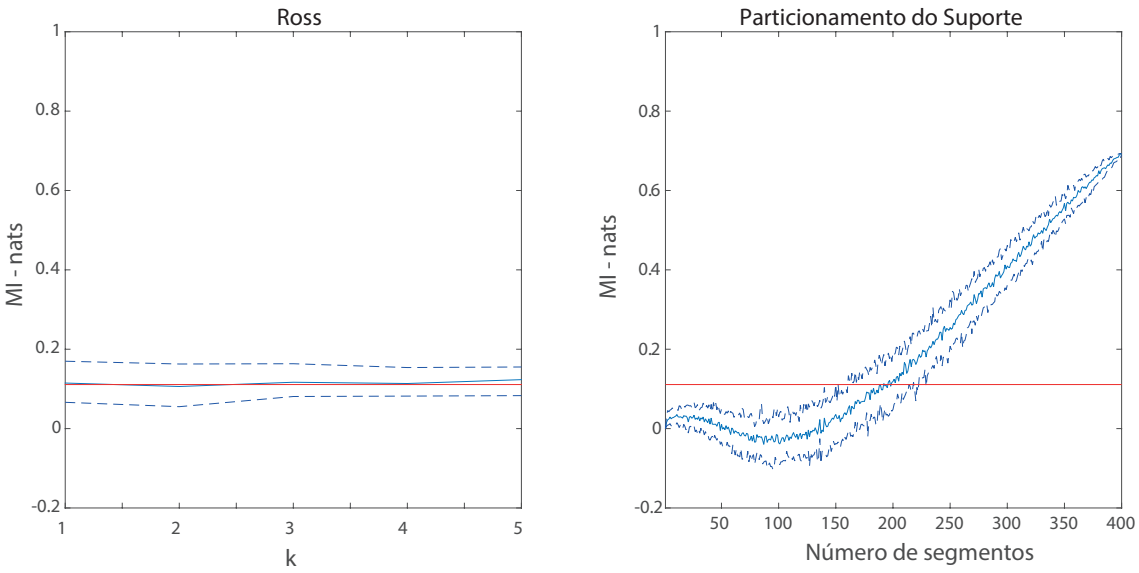


Figura 8.4: Estimativas de acordo com o método de Ross e o método do particionamento do suporte, em *nats*, caso uniforme-gaussiano. À esquerda, estimação de informação mútua pelo método de Ross, em função do número de vizinhos k. À direita, estimação de informação mútua pelo método do particionamento do suporte com correção de viés QE, como uma função do número de segmentos usados. O tamanho amostral foi $N = 400$, linhas vermelhas indicam o verdadeiro valor de informação mútua, ao passo que linhas azuis indicam a mediana de informação mútua para 50 conjuntos de dados de tamanho 400 cada. O intervalo entre linhas azuis tracejadas indicam de 10% a 90% das estimativas.

8.2 Estimação de Entropia de Transferência

Nesta seção analisamos o comportamento dos dois estimadores de informação mútua explorados neste capítulo para o caso particular de detecção de causalidade. Para tanto, estima-se a entropia de transferência através da identidade (2.41) do capítulo 2. Assim como no capítulo 6, aqui a estimação de entropia de transferência entre os processos é feita considerando médias temporais, isto é, os processos envolvidos são considerados ergódicos.

Com o objetivo de avaliar o desempenho dos estimadores propostos, alguns exemplos envolvendo causalidade em casos mistos foram elaborados. Em todas as simulações seguintes, quando usando o método de Ross (vizinhos mais próximos), o parâmetro número de vizinhos foi ajustado em $k = 3$, como recomendado em [44, 70]. Além disso, quando usando o método do particionamento do suporte, assim como no capítulo 6, ajustou-se o parâmetro número de segmentos como

$$Q = \left\lfloor \sqrt[m+l+1]{N} \right\rfloor . \tag{8.14}$$

Ajustou-se o número mínimo de $Q = 2$, quando o valor encontrado de Q na equação (8.14) era na verdade 1, porque o método do particionamento do suporte não faz sentido usando apenas um segmento (pois não captura qualquer variação dos processos aleatórios envolvidos).

8.2.1 Primeiro Exemplo: Entropia de Transferência dos Discretos para os Contínuos

No primeiro exemplo, nós temos um processo aleatório discreto $\mathbf{X}$ que está influenciando causalmente um processo aleatório contínuo $\mathbf{Y}$. $\mathbf{X}$ é a cadeia de Markov cujo diagrama de estados é dado na Fig. 8.5.

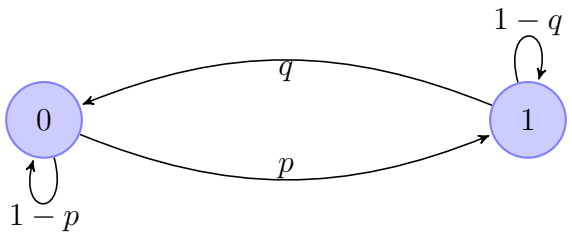


Figura 8.5: Diagrama de estados para processo $\mathbf{X}$.

Por outro lado, o processo $\mathbf{Y}$ é dado pela seguinte relação:

$$Y_n = \alpha Y_{n-1} + \gamma X_{n-1} + \epsilon \eta_n, \tag{8.15}$$

em que α, γ e ϵ são parâmetros fixos e η_n é um ruído aleatório Gaussiano padrão ($\eta_n \sim \mathcal{N}(0,1)$).

Para realizar as simulações e avaliar os estimadores, primeiramente ajustaram-se os parâmetros $p = 1/2$ e $q = 3/4$ ($P(X_n = 0) = 0.6$ e $P(X_n = 1) = 0.4$ no regime estacionário). O parâmetro γ variou no alcance de $[-0.5, 0.5]$, usando passo 0.1. Para cada valor de γ, encontraram-se limitantes para $TE_\infty(X \to Y)$ como explicado a seguir e simularam-se 50 amostras destes processos aleatórios com duração $N = 1000$. Fixaram-se $\alpha = 0.5$ e $\epsilon = 0.1$.

Neste exemplo, o processo $\mathbf{Y}$ não é estacionário para todos os valores de α, γ e ϵ. Por exemplo, quando $\alpha = \gamma = 1$ e $\epsilon = 0$, a média do processo $\mathbf{Y}$ não se mantém constante. Contudo, ainda é possível encontrar limitantes para o valor analítico de $TE_\infty(X \to Y)$, considerando que a cadeia de Markov $\mathbf{X}$ esteja em regime estacionário (o que ocorre quando $n \to \infty$). Primeiramente, escreve-se a seguinte identidade:

$$\begin{aligned} TE_n(X \to Y) &= H(Y_n|Y_{n-1}) - H(Y_n|Y_{n-1}X_{n-1}) \\ TE_\infty(X \to Y) &= \lim_{n\to\infty}[H(Y_n|Y_{n-1}) - H(Y_n|Y_{n-1}X_{n-1})], \end{aligned} \tag{8.16}$$

considerando os índices de passado, de $\mathbf{X}$ e de $\mathbf{Y}$, $l = m = 1$. O segundo termo à direita da igualdade (8.16) pode ser calculado como:

$$\begin{aligned} \lim_{n\to\infty} H(Y_n|Y_{n-1}X_{n-1}) &= \lim_{n\to\infty} H(Y_n - \alpha Y_{n-1} - \gamma X_{n-1}|Y_{n-1}X_{n-1}) \\ &= \lim_{n\to\infty} H(\epsilon\eta_n) \\ &= \frac{1}{2}\ln(2\pi e\epsilon^2). \end{aligned}$$

Já o primeiro termo apresenta o seguinte limitante superior:

$$\lim_{n\to\infty} H(Y_n|Y_{n-1}) \leq \lim_{n\to\infty} H(\gamma X_{n-1} + \epsilon\eta_n), \tag{8.17}$$

que é uma desigualdade, pois condicionamento não aumenta a entropia [13] e Y_{n-1} e X_{n-1} não são independentes (ambos dependem da variável X_{n-2}). Por outro lado, é certo considerar como limitante inferior

$$\begin{aligned}\lim_{n\to\infty} H(Y_n|Y_{n-1}) &= H(\gamma X_{n-1} + \epsilon\eta_n|Y_{n-1}) \\ &\geq \lim_{n\to\infty} H(\gamma X_{n-1} + \epsilon\eta_n|X_{n-2}),\end{aligned} \tag{8.18}$$

pois a soma $\gamma X_{n-1} + \epsilon\eta_n$ depende diretamente de X_{n-2}, ao passo que o ruído η_n é i.i.d..

Para encontrar o limitante superior, com o processo $\mathbf{X}$ em regime estacionário, considera-se a distribuição subjacente de $U = \gamma X_{n-1} + \epsilon\eta_n$ na equação (8.17):

$$f_U(u) = P(X = 0)g_0(u) + P(X = 1)g_1(u), \tag{8.19}$$

em que

$$\begin{aligned}g_0(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-u^2/(2\epsilon^2)}, \\ g_1(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-(u-\gamma)^2/(2\epsilon^2)},\end{aligned}$$

visto que η é gaussiana padrão. Como pode ser visto, $f_U(u)$ é uma mistura de duas gaussianas, uma com média 0 e outra com média γ.

A entropia diferencial de uma mistura gaussiana não apresenta uma solução analítica conhecida [30]. Para encontrar uma aproximação numérica para o limitante na inequação (8.17), utilizou-se a regra trapeizoidal:

$$\int_a^b f(x)dx \approx (b-a)\frac{f(b) - f(a)}{2}. \tag{8.20}$$

No caso particular de interesse, deseja-se uma aproximação para

$$H(U) = -\int_{-\infty}^{\infty} f_U(u) \ln f_U(u) du. \tag{8.21}$$

Com o intuito de descobrir um intervalo de integração apropriado, o gráfico de

$$-f_U(u) \ln f_U(u)$$

foi traçado (Fig. 8.6), com o valor ajustado de $\gamma = 0.5$. Observa-se claramente que a função $f_U(u)$ já é aproximadamente nula para valores de u fora do intervalo $[-1.5, 1.5]$. Portanto, para utilizar a regra trapezoidal, houve o somatório da regra na equação (8.20) em intervalos $\Delta_u = 0.001$, desde $u = -1.5$ até $u = 1.5$. O valor obtido foi $H(U) \approx -0.2275$ *nats*.

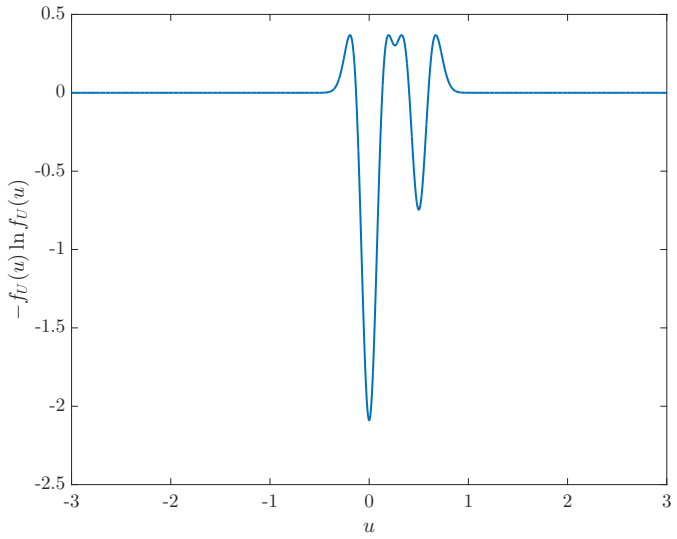


Figura 8.6: Gráfico de $-f_U(u)\ln f_U(u)$.

Já para encontrar o limitante inferior, temos:

$$\begin{aligned}
U &= \gamma X_{n-1} + \epsilon\eta_n \\
f_U(u|x_{n-2}) &= f_{U|X_{n-2}}(\gamma x_{n-1} + \epsilon\eta_n | x_{n-2}) \\
&= \sum_{x_{n-1}} f_{U,X_{n-1}|X_{n-2}}(\gamma x_{n-1} + \epsilon\eta_n, x_{n-1}|x_{n-2}) \\
&= f_{U,X_{n-1}|X_{n-2}}(\epsilon\eta_n, X_{n-1}=0|x_{n-2}) + f_{U,X_{n-1}|X_{n-2}}(\gamma+\epsilon\eta_n, X_{n-1}=1|x_{n-2}) \\
&= f_{U|X_{n-2}^{n-1}}(\epsilon\eta_n|X_{n-2}^{n-1})P(X_{n-1}=0|x_{n-2}) + \\
&\quad + f_{U|X_{n-2}^{n-1}}(\epsilon\eta_n+\gamma|x_{n-2})P(X_{n-1}=1|x_{n-2}) \\
&= g_0(u)P(X_{n-1}=0|x_{n-2}) + g_1(u)P(X_{n-1}=1|x_{n-2}). \qquad (8.22)
\end{aligned}$$

Portanto, quando $X_{n-2}=0$, encontra-se

$$f_{U|X_{n-2}}(u|X_{n-2}=0) = g_0(u)(1-p) + g_1(u)p, \qquad (8.23)$$

e, quando $X_{n-2}=1$, encontra-se

$$f_{U|X_{n-2}}(u|X_{n-2}=1) = g_0(u)q + g_1(u)(1-q). \qquad (8.24)$$

Assim, o limitante inferior pode ser calculado através de:

$$\lim_{n\to\infty} H(U|X_{n-2}) = \lim_{n\to\infty} -\sum_{x_{n-2}} P(x_{n-2}) \int_{-\infty}^{\infty} f_{U|X_{n-2}}(u|x_{n-2}) \ln f_{U|X_{n-2}}(u|x_{n-2}) du,$$

que, assim como foi feito para determinar o limitante superior, pode ser encontrado por integração numérica, utilizando a regra trapezoidal, para cada valor de γ utilizado.

A Fig. (8.7) mostra as medianas das estimativas de entropia de transferência, com ambos métodos de estimação, junto com os limitantes teóricos obtidos.

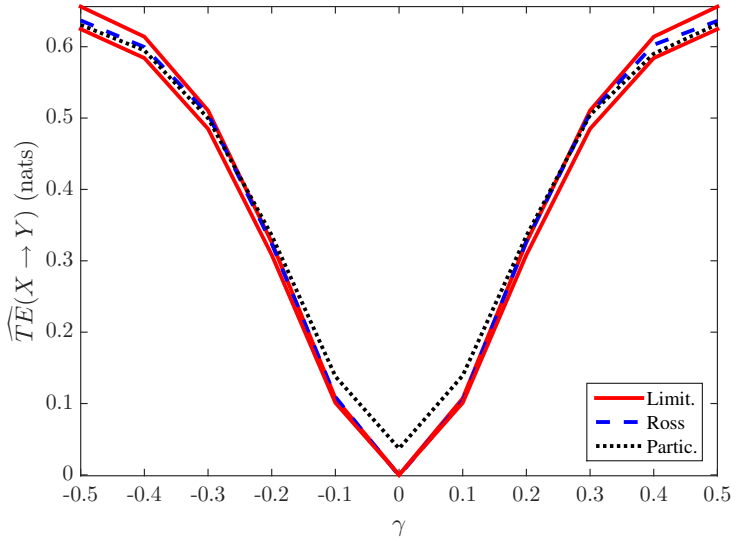


Figura 8.7: Medianas de estimativas de entropia de transferência em função do parâmetro de acoplamento de causalidade γ. Curva azul tracejada indica medianas das estimativas com método de Ross, curva pontilhada preta indica medianas das estimativas com o método do particionamento do suporte, para cada valor de γ, e curvas contínuas vermelhas indicam limitantes teóricos. Medianas em 50 funções amostras, com duração $N = 1000$.

Observa-se pela Fig. 8.7 que ambos métodos se aproximam dos valores teóricos, de acordo com o parâmetro γ utilizado. Entretanto, observa-se que o método do particionamento do suporte sobrestima um pouco as estimativas de entropia de transferência quando o parâmetro de acoplamento foi muito baixo ($|\gamma| = \{0, 0.1, 0.2\}$).

Também estimou-se a entropia de transferência com parâmetro de acoplamento fixo $\gamma = 0.5$, para diferentes durações dos processos ($N = \{50, 100, 500, 1000, 5000\}$). A Fig. 8.8 mostra que ambos métodos convergem para o mesmo valor conforme N cresce, e que este valor de convergência encontra-se dentro dos limitantes.

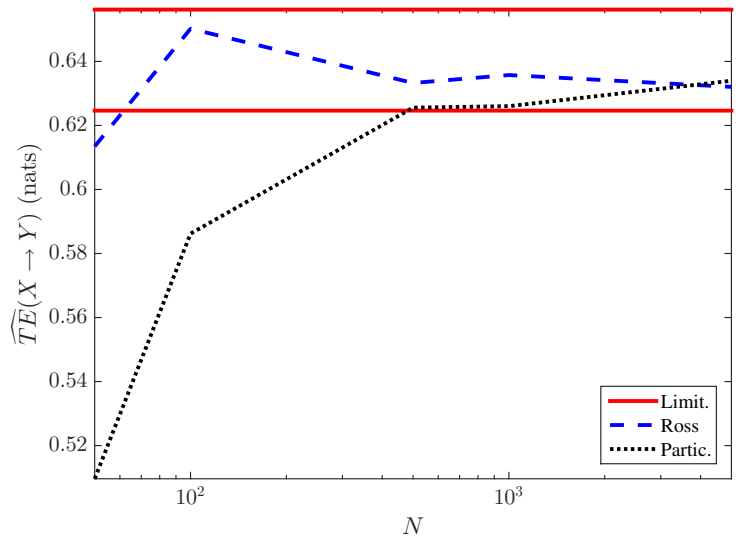


Figura 8.8: Medianas de estimativas de entropia de transferência em função da duração N dos processos, em 50 estimativas, para cada duração N. Curva azul tracejada indica mediana das estimativas com método de Ross, curva preta pontilhada indica estimativas com o método do particionamento do suporte, para valor fixo $\gamma = 0.5$, e linhas contínuas vermelhas indicam limitantes teóricos.

A Fig. 8.9 ilustra o desempenho dos estimadores de acordo com a variância amostral das estimativas, para ambos casos simulados — variando parâmetro γ ou duração N. As estimativas do particionamento do suporte apresentaram menor variância em todos os casos, revelando desempenho superior neste critério, e com ambos estimadores a variância diminui conforme N aumenta.

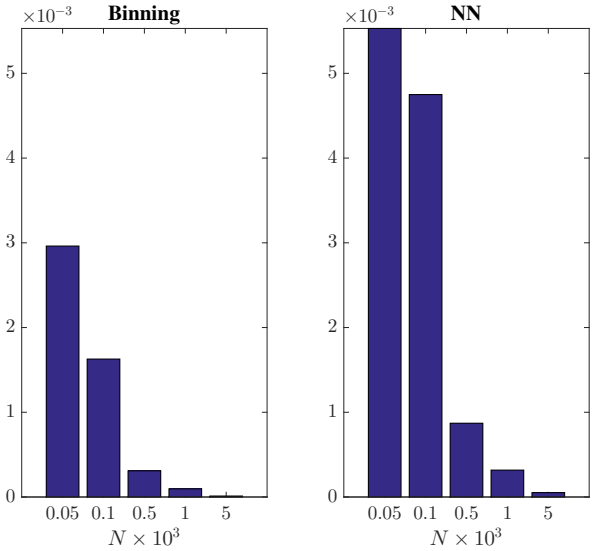


Figura 8.9: Variâncias amostrais de estimativas de TE como uma função da duração N dos processos. Estatísticas realizadas em 50 amostras, parâmetro fixo $\gamma = 0.5$. Método do particionamento do suporte: "*binning*", método de Ross: "NN".

8.2.2 Segundo Exemplo: Entropia de Transferência dos Contínuos para os Discretos

Nesta subseção, trata-se da estimação de entropia de transferência de um processo contínuo para um processo discreto. Primeiramente trata-se de um caso em que há uma aproximação do valor analítico de TE, como descrito a seguir.

Considere $\mathbf{X}$ um processo i.i.d. tal que cada X_n é distribuído uniformemente no intervalo $[\alpha, \beta]$ ($X_n \sim U(\alpha, \beta)$, $0 < \alpha < \beta$). Considere o processo $\mathbf{Y}$ "causado" por $\mathbf{X}$ da seguinte maneira:

$$P(Y_n = y | X_{n-1} = x) = \frac{(x)^y e^{-x}}{y!}. \tag{8.25}$$

Portanto, $P(Y_n = y)$ pode ser calculado como:

$$\begin{aligned} P(Y_n = y) &= \int_\alpha^\beta P(Y_n = y | X_{n-1}) f_{X_{n-1}}(x) dx \\ &= \int_\alpha^\beta \frac{x^y e^{-x}}{y!} \frac{1}{\beta - \alpha} dx \\ &= \frac{1}{y!(\beta - \alpha)} \int_\alpha^\beta x^y e^{-x} dx \end{aligned} \tag{8.26}$$

Logo,

$$P(Y_n = y) = \begin{cases} -e^{-x}|_\alpha^\beta = e^{-\alpha} - e^{-\beta}, & \text{se } y = 0, \\ (-x-1)e^{-x}|_\alpha^\beta, & \text{se } y = 1, \\ -x^y e^{-x}|_\alpha^\beta + y \int_\alpha^\beta x^{y-1} e^{-x} dx, & \text{se } y > 1. \end{cases} \tag{8.27}$$

Quando $y > 1$, é possível calcular $P(Y_n = y)$ recursivamente através da igualdade em (8.27). E através das probabilidades $P(Y_n = y)$, $y \in \mathbb{Z}$, $y \geq 0$, é possível aproximar a entropia $H(Y_n|Y_{n-1}) = H(Y_n)$, visto que Y_n é independente de Y_{n-1}, para todo n.

Já para encontrar a entropia $H(Y_n|Y_{n-1}X_{n-1}) = H(Y_n|X_{n-1})$, convém lembrar que o valor de Y_n quando condicionado ao valor de X_{n-1} apresenta uma distribuição de Poisson, com taxa $X_{n-1} = x$, $x > 0$. Há na literatura aproximações para entropia de uma distribuição de Poisson, quando $x > 10$ [39]:

$$H(Y_n | X_{n-1} = x) \approx \frac{1}{2} \ln(2\pi e x) - \frac{1}{12x} + O(x^{-2}). \tag{8.28}$$

Desconsiderando o termo $O(x^{-2})$, encontra-se a aproximação para a entropia condicionada:

$$\begin{aligned} H(Y_n|X_{n-1}) &= \int_\alpha^\beta \frac{1}{\beta - \alpha} H(Y_n | X_{n-1} = x) dx \\ &\approx \int_\alpha^\beta \frac{1}{\beta - \alpha} \left(\frac{1}{2} \ln(2\pi e x) - \frac{1}{12x} \right) dx \\ &= \frac{1}{2(\beta - \alpha)} \left(x \ln(2\pi e x) - x - \frac{1}{12} \ln(x) \right) \Big|_\alpha^\beta \\ &= \frac{1}{2(\beta - \alpha)} \left(\beta \ln(2\pi e \beta) - \alpha \ln(2\pi e \alpha) + \alpha - \beta - \frac{1}{12} \ln\left(\frac{\beta}{\alpha}\right) \right). \end{aligned} \tag{8.29}$$

Selecionando os valores de $\alpha = 25$ e $\beta = 55$, por exemplo, que garantem a condição $x > 10$ utilizada na aproximação da entropia da equação (8.29), encontra-se

$$TE_n(X \to Y) = TE_\infty(X \to Y) \quad \approx \quad 0.589 nats. \tag{8.30}$$

As estimativas para este exemplo estão na Fig. 8.10, em que observa-se que o método de Ross, neste exemplo, converge para a aproximação do valor analítico conforme N aumenta, ao passo que o método do particionamento do suporte diverge, ao menos para o maior valor de N testado.

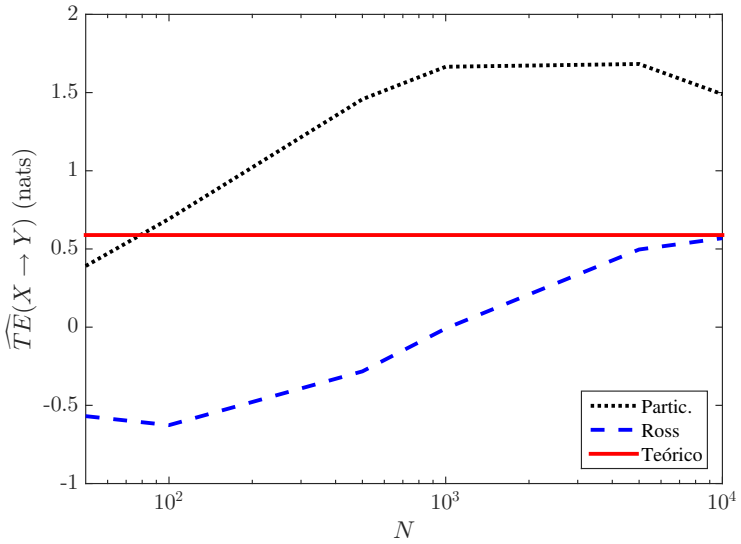


Figura 8.10: Medianas das estimativas de entropia de transferência de um processo contínuo para um processo discreto (equação 8.29) com o método do particionamento do suporte e com o método de Ross em função da duração dos processos N, em 50 realizações. Aproximação do valor analítico na linha contínua vermelha.

As variâncias amostrais obtidas neste exemplo estão indicadas na Fig. 8.11. Mais uma vez, o método do particionamento do suporte se mostrou superior no critério de apresentar menores variâncias nas estimativas.

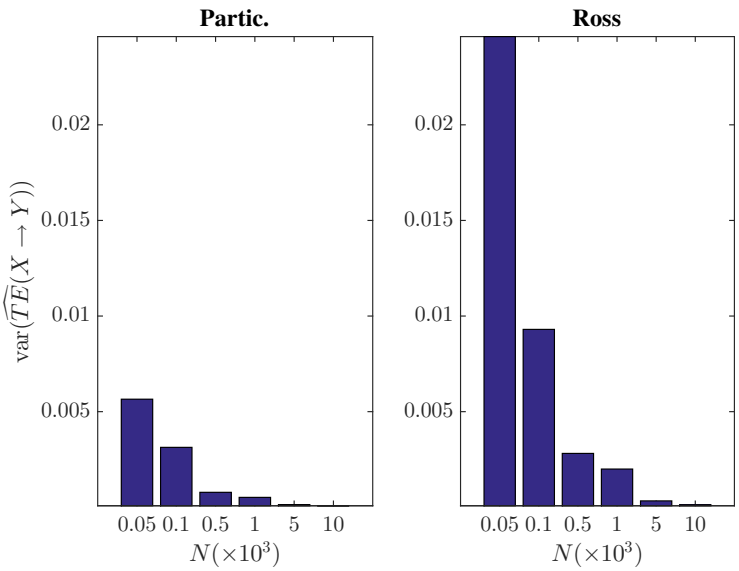


Figura 8.11: Variâncias amostrais das estimativas de entropia de transferência de um processo contínuo para um processo discreto (equação 8.29) com o método do particionamento do suporte e com o método de Ross em função da duração dos processos N, em 50 realizações.

No segundo exemplo desta subseção, simulou-se o caso em que $\mathbf{X}$ é dado por:

$$X_n = \alpha X_{n-1} + \eta_n, \tag{8.31}$$

em que η_n é um processo aleatório i.i.d. com distribuição gaussiana padrão ($\mathbf{X}$ é autorregressivo). Por outro lado, $\mathbf{Y}$ é um processo aleatório discreto. Para cada índice temporal n, a função massa de probabilidade da variável aleatória Y_n condicionada ao valor de X_{n-1} é dada por

$$P(Y_n = y | X_{n-1} = x) = \frac{|x|^y}{y!} e^{-|x|}, \tag{8.32}$$

que, assim como no exemplo anterior, é uma distribuição de Poisson cuja taxa é determinada por um valor passado do processo $\mathbf{X}$. Portanto, o processo discreto $\mathbf{Y}$ é causalmente influenciado pelo processo contínuo $\mathbf{X}$.

Nestas simulações, ajustou-se $\alpha = 0.5$. Em cada uma das 50 amostras do experimento, obtiveram-se 4 estimativas de entropia de transferência: duas foram com os métodos de Ross e do particionamento do suporte. As outras duas estimativas de entropia de transferência foram construídas sem qualquer relação de dependência ou causalidade entre $\mathbf{X}$ e $\mathbf{Y}$. Mais especificamente, $\mathbf{Y}$ foi gerado a partir de um processo i.i.d. gaussiano $\mathbf{Z}$, da mesma forma que na equação (8.32), substituindo x_{n-1} por z_{n-1}.

A Fig. 8.12 ilustra os resultados.

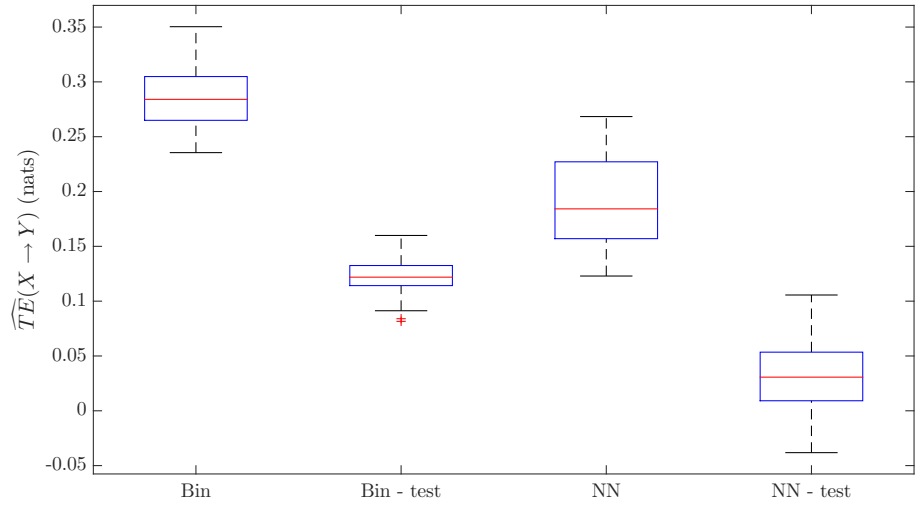


Figura 8.12: Boxplots das estimativas de entropia de transferência para a simulação do segundo exemplo com o método do particionamento do suporte ("Bin"), com o método do particionamento do suporte sobre processos sem qualquer causalidade ("Bin - test"), com o método de Ross ("NN") e com o método de Ross sobre processos sem qualquer causalidade ("NN - test"). Estatísticas realizadas em 50 amostras, duração dos processos $N = 500$.

Está claro da Fig. 8.12 que há uma diferença significativa entre as estimativas dos dados originais e as estimativas dos dados sobre processos sem qualquer causalidade, que foi verificada com um teste t em ambos os casos (nível de confiança do teste em 5%). Isto significa que, apesar de não convergirem para o mesmo valor mediano para a duração $N = 500$, ambos métodos de estimação indicam fielmente a presença de uma relação causal que não foi detectada na mesma intensidade com os dados sem causalidade.

Além disso, aumentando a duração N dos processos, observou-se que as estimativas se aproximam, como ilustrado na Fig. 8.13.

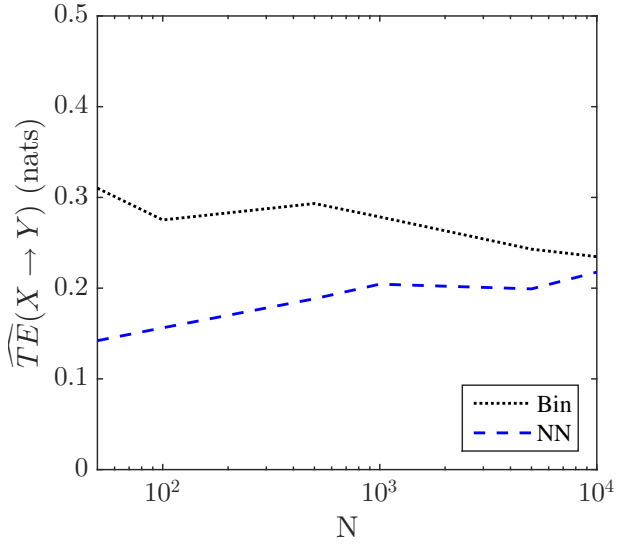


Figura 8.13: Estimativas de entropia de transferência para a simulação do segundo exemplo com o método do particionamento do suporte ("Bin") e com o método de Ross ("NN") em função da duração dos processos N.

Um fato curioso é que no primeiro exemplo desta subseção, o método do particionamento do suporte diverge da aproximação do valor analítico (enquanto o método de Ross converge), mas no segundo exemplo ambos métodos convergem para o mesmo valor conforme N aumenta. Tal fato pode ter ocorrido porque no primeiro exemplo, a taxa $X_{n-1} = x$ do processo de Poisson varia em um intervalo grande, de modo que os valores assumidos $\mathbf{Y}$ neste exemplo variam em um intervalo maior de que no segundo exemplo. Deste modo, muito embora o alfabeto de $\mathbf{Y}$ nos dois exemplos seja teoricamente infinito (contável), as realizações de $\mathbf{Y}$ no segundo exemplo apresentam alfabeto bem reduzido em relação ao do primeiro. Relembra-se aqui que o método do particionamento do suporte usa o método *plug-in*, cuja aproximação do viés depende do alfabeto de valores possíveis assumidos por $\mathbf{Y}$ (equação (3.1), capítulo 3). A Fig. 8.14 ilustra os histogramas dos valores de $\mathbf{X}$ e $\mathbf{Y}$, do primeiro e do segundo exemplo, em funções amostras destes processos, com duração $N = 10000$.

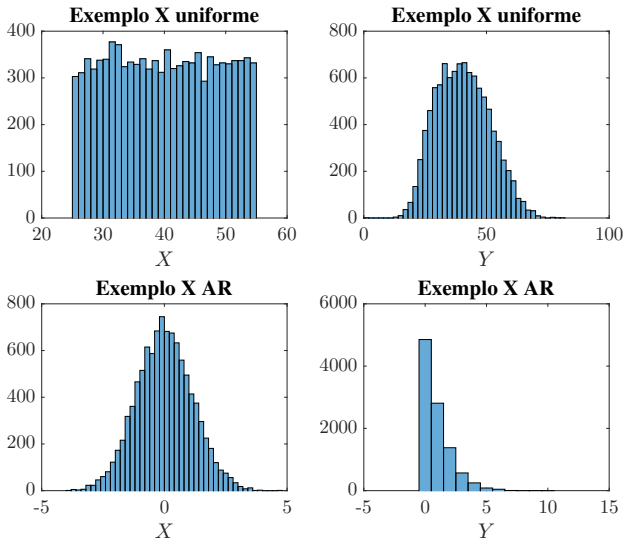


Figura 8.14: Histogramas de valores assumidos pelos processos $\mathbf{X}$ e $\mathbf{Y}$ desta subseção, segundo o exemplo em que $\mathbf{X}$ é distribuído uniformemente e segundo o exemplo em que $\mathbf{X}$ é autorregressivo (AR), em funções amostras de duração $N = 10000$.

8.2.3 Terceiro Exemplo: Maior Acoplamento Temporal

Nesta seção considera-se o efeito de um maior acoplamento temporal entre os processos $\mathbf{X}$ e $\mathbf{Y}$ nos métodos de estimação propostos (e o efeito de maiores índices de passado l e m na estimação de entropia de transferência). Para esta consideração, geraram-se processos $\mathbf{X}$ como os do exemplo do capítulo 4, cujo diagrama de estados é dado na Fig. 4.13.

Para cada estado do processo $\mathbf{X}$, que são 1, 10 ou 00, associa-se outro processo discreto $\mathbf{Z}$, tal que

$$Z_n = \begin{cases} 1, \text{ se } X_{n-1} = 1, \\ 0, \text{ se } X_{n-2}^{n-1} = 10, \\ -1, \text{ se } X_{n-2}^{n-1} = 00. \end{cases} \tag{8.33}$$

Então, define-se o processo $\mathbf{Y}$ como:

$$Y_n = \alpha Y_{n-m} + \gamma Z_n + \epsilon \eta_n. \tag{8.34}$$

Neste terceiro exemplo, ajustaram-se os parâmetros $\alpha = 0.5$, $\gamma = 0.5$, $\epsilon = 0.1$, e as probabilidades condicionadas do diagrama de estados como $\theta_1 = 0.1$, $\theta_{10} = 0.3$ e $\theta_{00} = 0.5$. Em regime estacionário, as probabilidades dos estados de $\mathbf{X}$ são $\pi_1 = 0.32$, $\pi_{10} = 0.28$ e $\pi_{00} = 0.4$ (vide apêndice D).

Assim como na subseção 8.2.1, é possível encontrar uma aproximação para um limitante superior de entropia de transferência, quando $\mathbf{X}$ está em regime estacionário (mas agora considerando os índices de passado $l = 2$, m):

$$\begin{aligned} \lim_{n\to\infty} H(Y_n | Y_{n-m}^{n-1} X_{n-2}^{n-1}) &= \lim_{n\to\infty} H(Y_n - \alpha Y_{n-m} - \gamma Z_n | Y_{n-m}^{n-1} X_{n-2}^{n-1}) \\ &= \lim_{n\to\infty} H(\epsilon \eta_n) \\ &= \frac{1}{2} \ln(2\pi e \epsilon^2). \end{aligned} \tag{8.35}$$

$$\lim_{n\to\infty} H(Y_n | Y_{n-m}^{n-1}) \leq \lim_{n\to\infty} H(\gamma Z_n + \epsilon \eta_n). \tag{8.36}$$

Com o processo $\mathbf{X}$ em regime estacionário, considera-se a distribuição subjacente de $U = \gamma Z_n + \epsilon\eta_n$ na equação (8.36):

$$f_U(u) = \pi_1 g_1(u) + \pi_{10} g_{10}(u) + \pi_{00} g_{00}(u), \tag{8.37}$$

em que

$$\begin{aligned} g_1(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-(u-\gamma)^2/(2\epsilon^2)}, \\ g_{10}(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-u^2/(2\epsilon^2)}, \text{ e} \\ g_{00}(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-(u+\gamma)^2/(2\epsilon^2)}, \end{aligned}$$

visto que η é gaussiana padrão. Como pode ser visto, assim como na seção 8.2.1, que $f_U(u)$ é uma mistura de gaussianas, mas neste caso de três gaussianas, com médias γ, 0 e $-\gamma$.

O gráfico de

$$-f_U(u) \ln f_U(u)$$

foi traçado (Fig. 8.15), com o valor ajustado de $\gamma = 0.5$. Observa-se claramente que a função $f_U(u)$ já é aproximadamente nula para valores de u fora do intervalo $[-1.5, 1.5]$. Portanto, para utilizar a regra trapezidal, houve o somatório da regra na equação (8.20) em intervalos $\Delta_u = 0.001$, desde $u = -1.5$ até $u = 1.5$. O valor obtido foi $H(U) \approx 0.1821$ *nats*. Logo

$$\begin{aligned} TE(X \to Y) &\leq 0.1821 - \frac{1}{2}\ln(2\pi e\epsilon^2) \\ &= 1.0658 nats. \end{aligned} \tag{8.38}$$

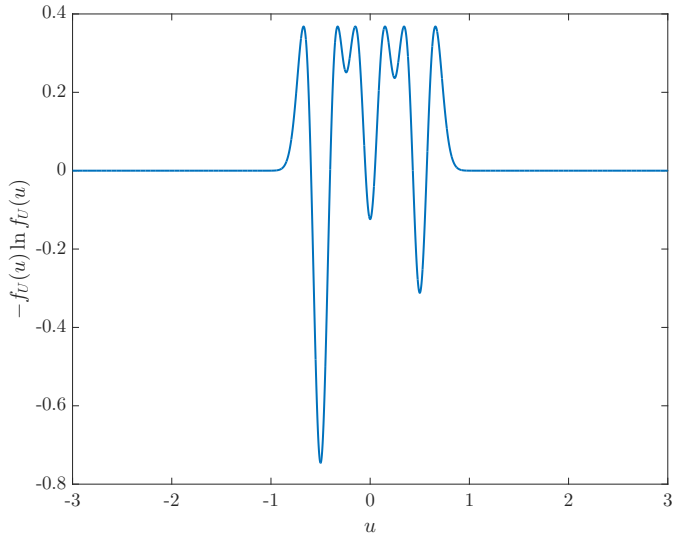


Figura 8.15: Gráfico de $-f_U(u) \ln f_U(u)$.

O índice de passado de $\mathbf{X}$ utilizado para estimar entropia de transferência foi $l = 2$. O acoplamento temporal m utilizado nas simulações para gerar o processo $\mathbf{Y}$, na equação

(8.34), foi o mesmo índice de passado m utilizado para estimar entropia de transferência. O termo m foi variado a fim de observar sua influência nas estimativas de entropia de transferência. A Fig. 8.16 ilustra os resultados através das medianas das estimativas. A Fig. 8.17 mostra os resultados através de boxplots. Assim como no segundo exemplo da subseção 8.2.2, aqui também geraram-se dados sem qualquer causalidade a fim de comparação. Mais especificamente, o processo $\mathbf{Y}_{test}$ foi gerado para teste com a mesma equação (8.34), mas sendo $\mathbf{Z}$ um processo i.i.d., discreto e uniformemente distribuído no alfabeto $\{-1, 0, 1\}$.

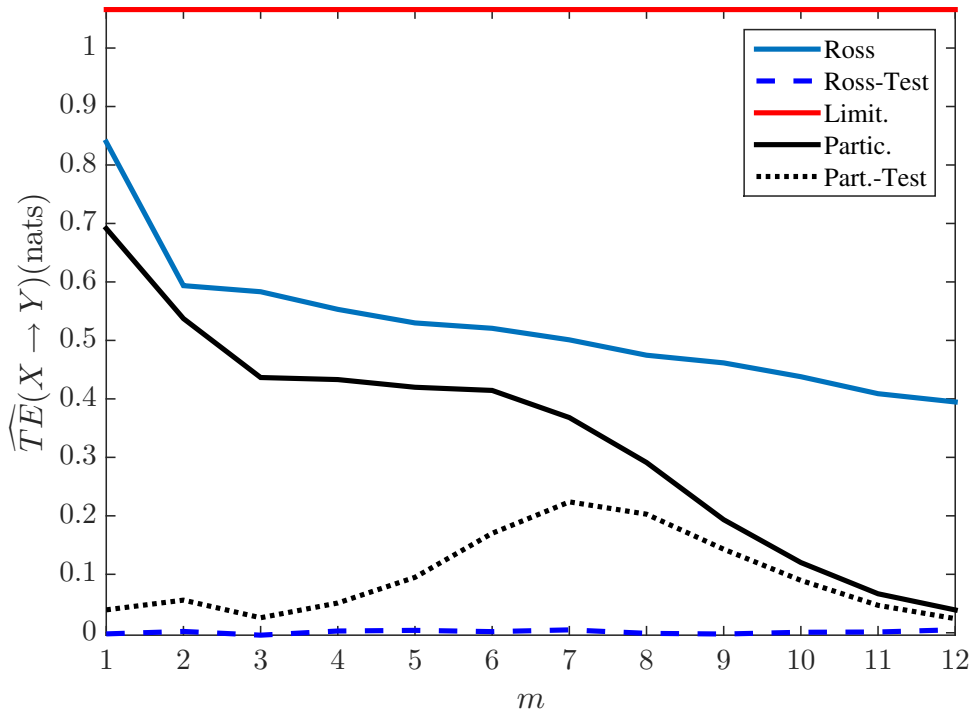


Figura 8.16: Medianas das estimativas de entropia de transferência do acoplamento temporal m do processo $\mathbf{Y}$, de acordo com a equação (8.34). Curva contínua azul indica medianas das estimativas com método de Ross, curva tracejada azul indica medianas das estimativas de Ross sobre dados sem qualquer dependência. Curva contínua preta indica medianas das estimativas com o método do particionamento do suporte e curva preta pontilhada indica medianas das estimativas com método do suporte sobre dados sem qualquer dependência. Aproximação do limitante superior teórico na linha contínua vermelha. Estatísticas sobre 50 amostras, duração dos processos $N = 500$.

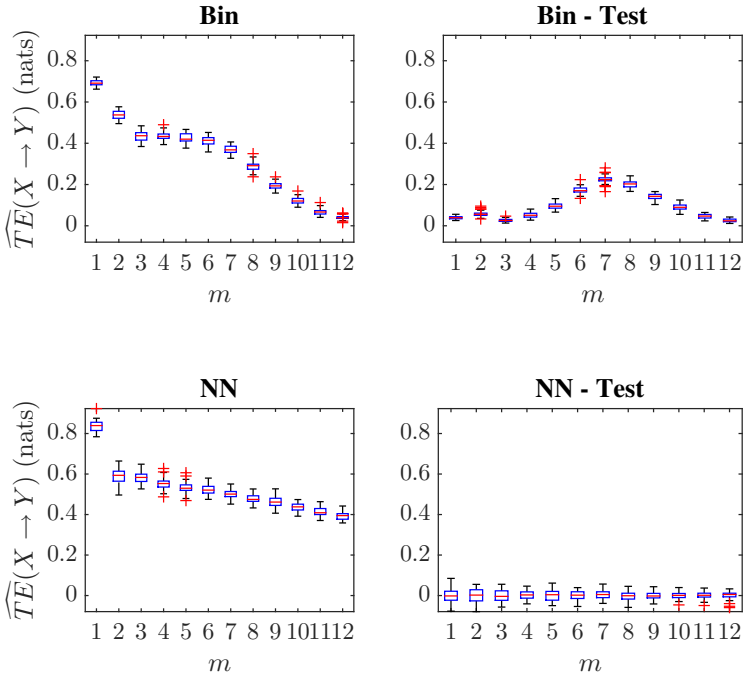


Figura 8.17: Boxplots de estimativas de entropia de transferência em função do acoplamento temporal m do processo $\mathbf{Y}$, de acordo com a equação (8.34). Método de Ross: "NN", método de Ross sobre dados sem qualquer dependência: "NN - Test", método do particionamento do suporte: "Bin" e método do particionamento do suporte sobre dados sem qualquer dependência: "Bin - Test". Estatísticas sobre 50 amostras, duração dos processos $N = 500$.

Observando as Fig. 8.16 e Fig. 8.17, percebe-se que as estimativas com ambos os métodos estão abaixo do limitante superior, conforme esperado. Também percebe-se que o método de Ross não detectou causalidade espúria. As estimativas com o método de Ross para $\widehat{TE}(X \to Y_{test})$ foram praticamente nulas. Contudo, estimativas com o método do particionamento do suporte para $\widehat{TE}(X \to Y_{test})$ foram sempre maiores que zero. Além disso, há uma diferença mais nítida entre $\widehat{TE}(X \to Y)$ e $\widehat{TE}(X \to Y_{test})$ com o método de Ross, especialmente para $m \geq 6$. Contudo, ao executar um teste t, ambos métodos apresentaram uma diferença significativa entre as estimativas de $\widehat{TE}(X \to Y)$ e $\widehat{TE}(X \to Y_{test})$, para cada m utilizado (nível do teste ajustado em 5%).

Percebe-se também que com ambos métodos, quando m cresce, a observação da influência causal de $\mathbf{X}$ sobre $\mathbf{Y}$ é diminuída (as estimativas de entropia de transferência diminuem, apesar da causalidade subjacente). Isto ocorre porque a dimensão da variável Y_{n-m}^{n-1} na estimação também aumenta, apesar de manter-se o tamanho amostral (duração dos processos, $N = 500$).

8.2.4 Desempenho dos Estimadores em Termos de Velocidade

Registraram-se os tempos de estimação de entropia de transferência entre os processos do primeiro exemplo simulado (subseção 8.2.1) em função da sua duração N. O tempo que o método do particionamento do suporte levou foi sempre inferior a 0.1s. A Fig. 8.18 ilustra o tempo mediano levado pelo método de Ross normalizado pelo tempo mediano

levado pelo método do particionamento do suporte, em 50 realizações dos processos do exemplo da subseção 8.2.1, em função da duração N dos processos. Está evidente que o método de Ross consome mais tempo em sua execução, conforme já mencionado na referência [70].

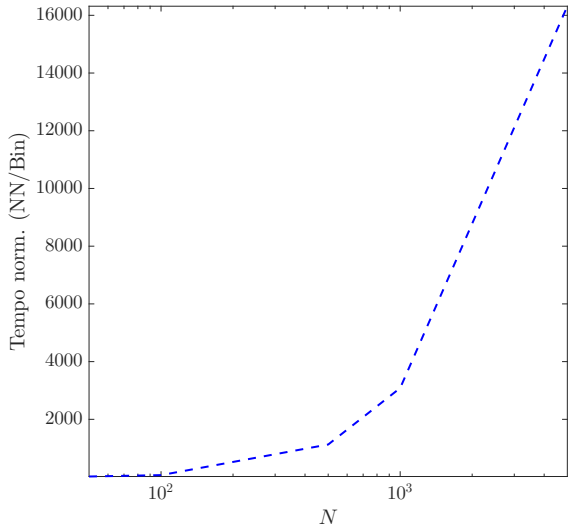


Figura 8.18: Medianas do tempo de estimação levado pelo método de Ross ("NN") normalizadas pelo tempo de estimação levado pelo método do particionamento do suporte ("Bin"), em 50 realizações dos processos do exemplo da subseção 8.2.1, em função da duração N.

Capítulo 9

Estimação de Informação Direcional para Trens de *Spikes* Neurais Simulados

Em média, cada neurônio humano forma cerca de 1000 conexões sinápticas e recebe ainda mais, talvez na ordem de 10 000 conexões [36]. Determinar a conectividade entre grupos de neurônios é importante na determinação de circuitos neurais e em como eles controlam o comportamento animal [21, 40, 11]. Embora seja possível desvendar as conexões de pequenos ciruitos neuronais *in vitro* usando microscopia eletrônica, a reconstrução na escala de uma coluna cortical é muito difícil com a tecnologia atual [76].

O presente capítulo investiga a capacidade dos estimadores da taxa de informação direcional de inferir conexões sinápticas entre neurônios a partir de registros extracelulares. Esses registros podem ser obtidos pelos neurocientistas atualmente graças a avanços nas técnicas envolvendo múltiplos eletrodos em tecidos nervosos, capazes de registrar a atividade de centenas de neurônios simultânea e individualmente [67]. Essa ideia é apresentada na referência [67], aqui a análise é feita com os dois estimadores apresentados no capítulo 3.

Antes de relatar o trabalho em si, alguns conceitos pertinentes de neurociências são relevantes neste capítulo. Conforme citado no capítulo 1, *spikes* são registros neurofisiológicos extracelulares em altas frequências [69]. Também chamados de potenciais de ação, os *spikes* são sinais que são transmitidos em um neurônio, tipicamente a partir do soma e então ao longo do axônio. A frequência com que os *spikes* podem ocorrer em um mesmo neurônio sempre respeita um período refratário de 1ms, ou seja, um mesmo neurônio não pode disparar *spikes* mais de uma vez em 1ms [36].

As conexões entre neurônios costumam se dar por sinapses químicas ou elétricas [36], neste trabalho investigamos apenas sinapses químicas. As sinapses químicas se caracterizam por uma fenda sináptica entre os neurônios pré e pós-sinápticos por onde são transmitidas vesículas sinápticas contendo neurotransmissores do neurônio pré-sináptico para o neurônio pós-sináptico. Quando um neurônio pré-sináptico dispara um *spike*, seus neurotransmissores são liberados. Estes neurotransmissores, quando em contato com um neurônio pós-sináptico, alteram a tensão superficial de membrana do neurônio pós-sináptico [36]. Quando a sinapse for excitatória, a tensão superficial do neurônio pós-sináptico será incrementada. Quando a sinapse for inibitória, a tensão superficial do neurônio pós-sináptico será reduzida. Quando a tensão superficial no neurônio pós-sináptico atinge um determinado limiar de disparo, o neurônio pós-sináptico também

dispara um *spike* [36]. Portanto, quando um neurônio pré-sináptico, em uma sinapse excitatória, dispara um *spike*, há uma chance maior do neurônio pós-sináptico também disparar um *spike*. Por outro lado, quando um neurônio pré-sináptico dispara um *spike*, em uma sinapse inibitória, há uma chance menor do neurônio pós-sináptico também disparar um *spike*. A Fig. 9.1 ilustra uma sinapse entre dois neurônios.

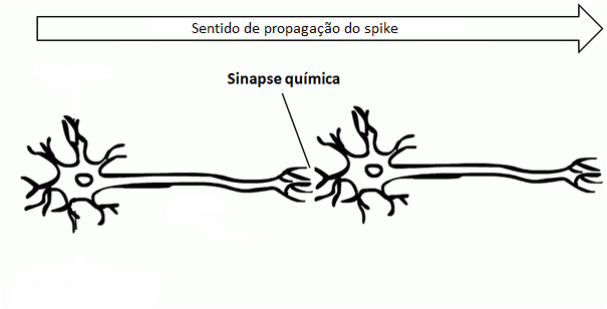


Figura 9.1: Ilustração de sinapse química entre dois neurônios. Figura adaptada do site http://maxaug.blogspot.com.br/2013/11/

Há vários modelos possíveis para descrever a atividade neuronal. O modelo de Hodgkin–Huxley é um modelo que descreve biologicamente com acurácia a atividade neuronal, contudo, é bastante complexo e de grande demanda computacional [33]. Há também o modelo LIF (*leaky-integrate-and-fire* [75]), que apresenta simplicidade computacional, mas que não reflete realisticamente a complexidade da dinâmica neuronal [33]. Em uma abordagem mais simplista, como a LIF, o limiar de disparo é constante no tempo. Contudo, observa-se experimentalmente que o limiar de disparo na realidade é dinâmico [80, 2]. Um modelo possível que articula simplicidade computacional e fidelidade à dinâmica neuronal é o modelo proposto por Izhikevich [33].

O modelo proposto por Izhikevich em [33] foi desenvolvido preferencialmente para simular a atividade de grandes redes neuronais (por exemplo, com 1000 neurônios). Para tanto utiliza um sistema de equações diferenciais de duas dimensões da forma:

$$\begin{aligned} \frac{dv}{dt} &= 0.04v^2 + 5v + 140 - u + I, \\ \frac{du}{dt} &= a(bv - u), \end{aligned}$$

com a condição auxiliar de pós-*spike*:

$$\text{se } v \geq 30\text{mV, então } v \leftarrow c, u \leftarrow u + d.$$

Utilizou-se o mesmo programa apresentado em [33] para simular a atividade neuronal deste trabalho. Para que houvesse um comportamento heterogêneo dos neurônios, isto é, cada um tivesse sua própria dinâmica, a cada neurônio excitatório atribuíram-se os parâmetros $(a_i, b_i) = (0.02, 0.2)$ e $(c_i, d_i) = (-65, 8) + (15, -6)r_i^2$, em que i indica o índice do neurônio simulado e r_i é uma variável aleatória uniforme em $[0, 1]$. Similarmente, a cada neurônio inibitório atribuíram-se os parâmetros $(a_i, b_i) = (0.02, 0.25) + (0.08, -0.05)r_i$ e $(c_i, d_i) = (-65, 2)$. Além desses parâmetros, o programa original de Izhikevich utiliza pesos sinápticos aleatórios que indicam que o *spike* do neurônio i imediatamente modifica a tensão superficial de membrana do neurônio j por w_{ij}.

A rede simulada no nosso trabalho, entretanto, apresentava apenas 5 neurônios, como ilustrada na Fig. 9.2. Para observar o padrão comum na atividade cortical, como ilustrado na Fig. 9.3, alguns parâmetros tiveram de ser modificados do programa original apresentado em [33]. Tais parâmetros foram a entrada ruidosa talâmica I e os pesos sinápticos w_{ij}. Como o número de neurônios da rede original foi drasticamente reduzido, para que houvesse alguma atividade neuronal, o parâmetro I foi aumentado, ajustado em 50 para neurônios excitatórios e 20 para neurônios inibitórios. Já os pesos sinápticos dos neurônios excitatórios variaram aleatória e uniformente no intervalo $[0, 40]$, ao passo que os pesos sinápticos dos neurônios inibitórios variaram aleatória e uniformente no intervalo $[0, -80]$. É certo que tais valores de pesos sinápticos não são observáveis na prática, mas este ajuste na simulação permitiu que o padrão observado de trens de *spikes* estivesse de acordo com o que é observado no córtex de mamíferos acordados. Além disso, mantém-se observada a regra de que as sinapses inibitórias apresentam conexões sinápticas com pesos maiores, em média e valor absoluto, que as sinapses excitatórias, como ocorre no córtex de mamíferos [33]. Mantém-se também no exemplo simulado a proporção de 4:1 entre neurônios excitatórios e inibitórios, que também é observada no córtex de mamíferos [33].

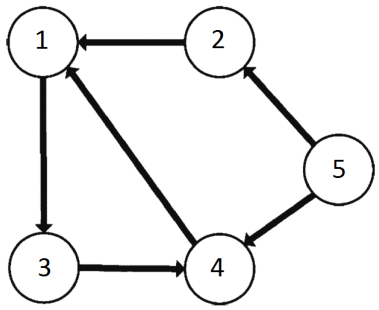


Figura 9.2: Diagrama para mostrar conexões entre neurônios simulados. Neurônio 5 apresenta conexões sinápticas inibitórias. Figura adaptada da referência [67].

As estimativas de taxa de informação direcional foram feitas com a duração de 100000 ms, cada milisegundo é uma janela temporal com a possibilidade de 0 ou 1 *spike*, respeitando o período refratário dos neurônios [36]. Os pesos das conexões sinápticas foram variados aleatoriamente a partir de uma distribuição uniforme ao longo de 50 amostras. A saída de uma possível amostra em uma janela de 1000ms é ilustrada na Fig. 9.3.

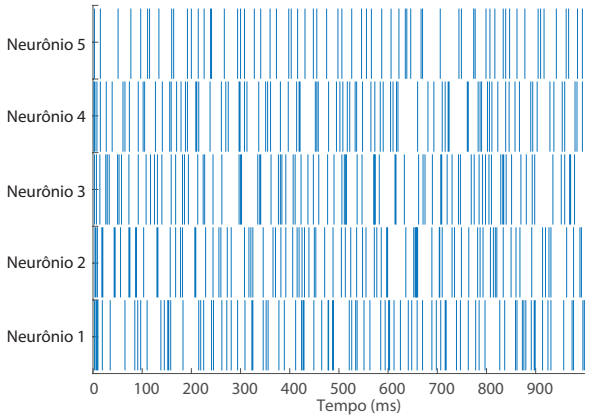


Figura 9.3: Gráfico de uma amostra de 5 neurônios disparando *spikes* em uma janela de 1000ms.

Para cada amostra estimaram-se as taxas de informação direcional entre cada par de trens de *spikes* neurais. Após esta etapa, a estimativa da taxa de informação direcional normalizada foi calculada como na equação (9.1):

$$\frac{\hat{I}_\infty(X \to Y)}{\hat{H}_\infty(Y)} = \frac{\hat{H}_\infty(Y) - \hat{H}_\infty(Y||X)}{\hat{H}_\infty(Y)} = 1 - \frac{\hat{H}_\infty(Y||X)}{\hat{H}_\infty(Y)}. \tag{9.1}$$

Quando a taxa de informação direcional normalizada é próxima de 1, há uma indicação de forte relação causal, ao passo que quando a taxa de informação direcional normalizada está próxima de 0, há uma indicação de fraca relação causal.

Estimaram-se taxas de informação direcional normalizadas entre trens de *spikes* de neurônios cujas conexões já eram conhecidas, a fim de investigar a relação entre esses valores e as conexões neurais. A seguinte matriz de adjacências serve de exemplo para uma amostra possível de pesos sinápticos, em que w_{ij} é peso sináptico do neurônio i para o neurônio j:

$$W = \begin{bmatrix} 0 & 0 & 15 & 0 & 0 \\ 20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 34 & 0 \\ 31 & 0 & 0 & 0 & 0 \\ 0 & -21 & 0 & -75 & 0 \end{bmatrix}. \tag{9.2}$$

A Fig. 9.4 ilustra a matriz de adjacências da equação 9.2 dos pesos sinápticos de conexão entre cada par de neurônios.

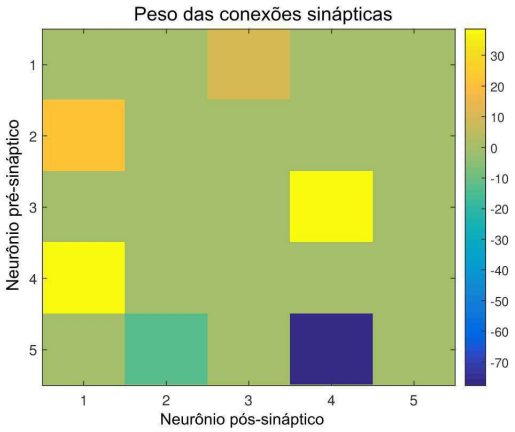
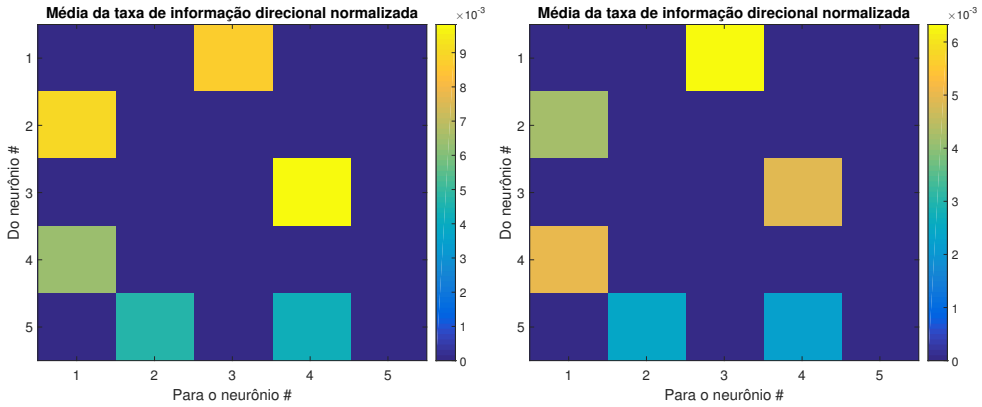


Figura 9.4: Pesos das conexões sinápticas entre neurônios em uma amostra.

A Fig. 9.5 ilustra a média de taxa de informação direcional normalizada para cada par de trens de *spikes* dos 5 neurônios, segundo o estimador de Quinn e o estimador de Jiao (algoritmo 2), explorados no capítulo 3. Na Fig. 9.5, os valores na diagonal principal foram ajustados a 0 para a percepção visual dos demais valores na escala da barra de cores, mas o valor real desses elementos é 1. Na subfigura 9.5b, alguns dos valores de taxa de informação direcional normalizada encontrados foram negativos. Esse resultado é aparentemente errôneo, haja visto que a informação direcional deve ser sempre maior ou igual a 0 (como visto no capítulo 2 e na referência [42]). Todavia, a possibilidade de ocorrência de resultados negativos para este estimador ("E2") já foi prevista [34], devido à insuficiência de dados ou à violação da suposição de sua estacionariedade. Na Fig. 9.5b, as estimativas negativas de taxa de informação direcional normalizada foram ajustadas para 0.

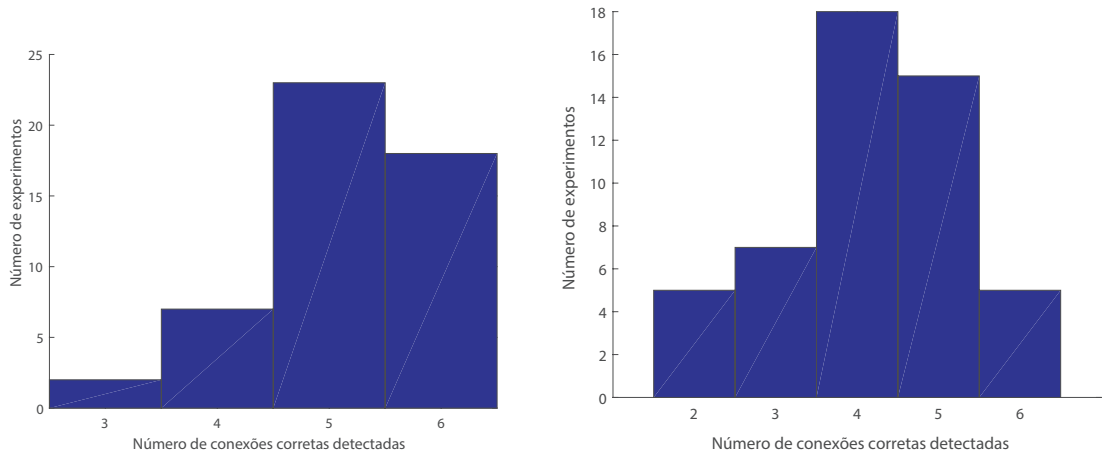


(a) Estimador de Quinn. (b) Estimador de Jiao.

Figura 9.5: Média de taxa de informação direcional normalizada entre trens de *spikes* neurais, em 50 amostras.

Observamos que valores positivos de taxa de informação direcional normalizada sempre indicaram conexões sinápticas verdadeiras entre os neurônios, nas 50 amostras realizadas para cada um dos estimadores. Também observou-se, para os dois estimadores, que valores nulos de taxa de informação direcional normalizada poderiam indicar a ausência de conexão sináptica ou não. Portanto, mesmo valores positivos e pequenos de taxa

de informação direcional normalizada, da ordem de 10^{-3}, indicando uma fraca relação causal, foram capazes nessas simulações de identificar corretamente conexões sinápticas. Por outro lado, valores nulos (ou negativos, no caso do estimador de Jiao) não fizeram uma predição confiável da ausência de conexões sinápticas. O número de predições corretas detectadas pelos dois estimadores é ilustrada na Fig. 9.6 (o número verdadeiro de conexões existentes foi 6).



(a) Estimador de Quinn. (b) Estimador de Jiao.

Figura 9.6: Histograma com o número de conexões verdadeiras detectadas pelos dois estimadores estudados. O número verdadeiro de conexões era 6.

Os resultados encontrados aqui são em geral diferentes dos encontrados em [67]. Lá foram utilizadas outras formas para simular a atividades dos neurônios e através de outra topologia, encontrando muitas vezes valores positivos de taxa de informação direcional normalizada. Estes valores positivos geralmente ocorreram por causa de influências indiretas de um neurônio sobre outro. Por exemplo, na Fig. 9.2, o neurônio 1 indiretamente causa o trem de *spikes* do neurônio 4. Contudo, o maior problema enfrentado aqui foi o de não detectar relações causais quando havia. Este problema foi observado com os dois tipos de estimadores, sendo que o estimador de Quinn esteve mais vezes predizendo todas as conexões verdadeiras (Fig. 9.6). Isto não diminui a aplicabilidade do estimador de Jiao, já que ele atua para casos e alfabetos mais gerais que o estimador de Quinn. Além disso, na simulação do estimador de Jiao, a profundidade da árvore foi feita $D = 3$, mas é possível que outros valores tragam predições mais condizentes com as conexões verdadeiras.

Finalmente, deve-se enfatizar também que os valores encontrados de taxa de informação direcional normalizada variam de acordo com os valores dos pesos sinápticos simulados. Além disso, observa-se neste estudo que apesar de em média terem pesos sinápticos maiores, as sinapses inibitórias apresentam valores médios mais baixos de informação direcional entre trens de *spikes* neurais (Fig. 9.5).

Capítulo 10

Conclusões e Perspectivas

Ao longo desta tese, investigaram-se diversos métodos de estimação de medidas de informação, em termos de acurácia e velocidade, tanto para processos discretos como para processos contínuos em amplitude. Também examinou-se o uso dos estimadores de Jiao para informação direcional entre processos contínuos através de três formas de discretização prévia dos processos. Além disso, avaliaram-se dois métodos de estimação de entropia de transferência para o caso misto, que ainda não tinham sido analisados na literatura. Finalmente, investigou-se a capacidade dos estimadores de informação direcional de inferir conexões neuronais sinápticas através de registros de *spikes* apenas, em uma contribuição específica para neurociências. Neste capítulo, resumimos as principais conclusões desta tese, além de apontar direções para novas pesquisas.

10.1 Estimadores de Medidas de Informação entre Processos Aleatórios Discretos

No capítulo 4 observou-se que para estimação de informação mútua, o método *plug-in* com correção QE obteve os melhores resultados em termos de acurácia. A exceção ocorre quando o tamanho amostral é muito reduzido (por exemplo, $N \approx 25$), caso em que a variância do estimador é elevada e a estimação *plug-in* sem correção é mais indicada. O método de Jiao para estimar informação mútua apresenta consistência, mas converge mais lentamente para o valor analítico. Em termos de velocidade, o método *plug-in* é mais eficiente, mas a diferença quando realiza-se o método *plug-in* com correção QE é muito pequena. O método de Jiao é mais lento computacionalmente também.

Por outro lado, a sofisticação do método de Jiao é bastante útil para estimação de informação direcional entre processos discretos com alguma memória finita. O método de Quinn é um método paramétrico e deve ser utilizado quando os processos envolvidos podem ser modelados como processos de contagem de eventos, com funções de intensidade condicional pertencentes aos modelos lineares generalizados. Por ser um método não paramétrico, o método de Jiao é mais abrangente. Pode-se observar que ele é aplicável e obtém resultados similares ao método de Quinn quando utilizado especificamente para o modelo de trens de *spikes*, na detecção de causalidade. Contudo, o método de Quinn obteve melhores resultados nesta aplicação específica.

Em termos de velocidade, ambos métodos são lentos, e o método de Quinn foi mais lento nos exemplos simulados. Contudo, a velocidade dos métodos pode ser aumentada pela diminuição do parâmetro de profundidade da árvore D, para o estimador de Jiao,

ou dos parâmetros de busca do algoritmo MDL, para o estimador de Quinn. Entretanto, a diminuição destes parâmetros pode atrapalhar a detecção de causalidade.

Ainda sobre a estimação de informação direcional com o estimador de Jiao, observou-se que o uso de um alfabeto muito extenso, por exemplo, $|\mathcal{X}| \geq 6$, pode prejudicar a estimação. Tal fato decorre não apenas pela questão da velocidade computacional, mas também pela própria aplicação do algoritmo CTW, que produz uma árvore de contexto com vários nós, cada qual com probabilidade muito pequena.

10.2 Estimadores de Medidas de Informação entre Processos Aleatórios Contínuos

No capítulo 6, observou-se que a estimação de informação mútua com os métodos KSG e BI-KSG apresentam desempenho similar em termos de velocidade e acurácia, sendo que o método BI-KSG foi um pouco mais rápido que o método KSG original. O método KDE foi mais rápido mediante uso de uma *toolbox* (JIDT), porém com resultados mais enviesados. Tais resultados possivelmente podem ser melhorados com um ajuste adequado do tamanho da função *kernel*. O método de particionamento do suporte foi o mais rápido, e suas estimativas apresentaram pouco viés no caso geral de distribuições gaussianas e de distribuição uniforme com independência entre as variáveis aleatórias. No caso particular de distribuição uniforme com dependência entre as variáveis aleatórias, o método de particionamento do suporte demorou mais a convergir e foi o mais enviesado para o maior tamanho amostral testado. O número de segmentos utilizado com o método do particionamento do suporte foi obtido com o auxílio da recomendação de Paluš.

Para a estimação de entropia de transferência, o método KSG se mostrou superior ao método KDE tanto em termos de velocidade como de acurácia, com o auxílio da *toolbox* JIDT. O método do particionamento do suporte não convergiu para o valor analítico de entropia de transferência, mesmo com o aumento da duração dos processos, ao menos da forma que foi implementado (como uma soma de informações mútuas e com a escolha do número de quantis fixada pela equação (6.7)).

A estimação de informação direcional com estimadores de Jiao a processos contínuos pela discretização prévia destes também foi investigada. No panorama geral, os métodos de discretização equidistante e equipovoado subestimam a taxa de informação direcional. O método de discretização simbólico em geral também produz estimativas abaixo do valor analítico, exceto quando usando alfabeto maior e quando não há causalidade de fato. Assim, não recomenda-se o uso deste método de discretização neste contexto. Para estimativas conservadoras, isto é, para detecção de causalidade quando realmente há (não detectar falsos positivos), recomenda-se o uso de alfabeto pequeno (discretização em até 4 níveis) com o método equipovado.

10.3 Estimadores de Medidas de Informação entre Processos Aleatórios Mistos

Nesta tese investigou-se também o uso de estimadores de informação mútua e entropia de transferência com o método do particionamento do suporte e com o método de Ross (que é baseado nas distâncias dos k vizinhos mais próximos). O método do particionamento do suporte se mostrou superior em termos de velocidade, ao passo que o método

de Ross se mostrou mais eficaz em termos de acurácia. A escolha do parâmetro k entre 1 e 5 não alterou muito as estimativas obtidas. Já o número de quantis Q utilizados no método do particionamento do suporte fez grande diferença nas estimativas obtidas, para um mesmo tamanho amostral.

Para o caso particular de estimação de entropia de transferência, foram considerados exemplos em que eram conhecidos limitantes justos para o valor teórico de entropia de transferência ou uma aproximação de seu valor teórico. Nestes exemplos, o método do particionamento do suporte forneceu resultados com um viés positivo quando não havia causalidade. Este método trouxe estimativas dentro dos limitantes quando havia causalidade de fato e o tamanho do alfabeto do processo discreto era pequeno (por exemplo, igual a 2), com uma duração dos processos relativamente pequena ($N = 500$). Contudo, quando o tamanho do alfabeto era maior (por exemplo, igual a 40) o método do particionamento do suporte não convergiu para o valor aproximado teórico de entropia de transferência, mesmo para a maior duração dos processos ($N = 10000$). Por outro lado, o método de Ross forneceu resultados que estavam dentro dos limitantes, ou convergiam para a aproximação teórica, nestes exemplos.

Também avaliaram-se os métodos em exemplos em que limitantes justos para a entropia de transferência, ou uma aproximação teórica da entropia de transferência, não eram conhecidos. Nestas situações, as estimativas dos dois métodos foram comparadas com estimativas obtidas a partir de dados sem qualquer causalidade (mas com estísticas similares). Um teste t foi capaz de detectar a diferença entre as estimativas em que os processos tinham uma causalidade subjacente daquelas em que não havia qualquer relação de causalidade entre os processos, com ambos os métodos. Esta diferença se manteve quando consideraram-se diferentes acoplamentos temporais (e índices de passado) nas estimativas de entropia de transferência. Contudo, a diferença foi mais evidente com o método de Ross.

Portanto, no panorama geral, o método de Ross é capaz também de detectar causalidade de maneira mais confiável, em particular trazendo menos falsos positivos. Contudo, o método do particionamento do suporte apresentou menor variâncias nas estimativas e desempenho muito melhor em termos de velocidade.

É interessante ressaltar que a recomendação de Paluš (equação (8.14) utilizada) mostrou-se mais eficaz para estimação de entropia de transferência no caso misto de que no caso contínuo. Tal fato é coerente com a diminuição da variabilidade que se encontra quando um dos processos é discreto em relação ao caso em que os dois processos são contínuos.

10.4 Perspectivas para Pesquisas Futuras

O presente trabalho abre espaço para novas pesquisas. Na aplicação do estimador de Jiao de informação direcional para processos contínuos, outros métodos de discretização podem ser avaliados. Uma maneira de formular esta avaliação seria: como estabelecer um valor percentual que indique um crescimento ou decrescimento nos processos contínuos que seja significativo na discretização?

Outra perspectiva futura de pesquisa seria como outras identidades para a entropia de transferência, como por exemplo, escrevê-la como uma soma de entropias convencionais, ao invés de uma soma de informações mútuas, poderia melhorar as estimativas utilizando o método do particionamento do suporte, para o caso de processos assumindo valores contínuos. Além disso, é possível realizar estimativas de entropia de transferência que

não considerem a suposição de ergodicidade dos processos, feitas ao longo desta tese, a fim de comparação.

Há ainda a possibilidade de desenvolver pesquisas que envolvam mais de dois processos, que foi o caso ao qual nos restringimos nesta tese. Finalmente, há espaço para a melhoria dos códigos, especialmente para aqueles que estimam informação direcional, a fim de aumentar sua velocidade.

Apêndice A

Conceitos de Probabilidade e Estatística

Este apêndice apresenta alguns conceitos de probabilidade e estatística que podem ser úteis no entendimento deste trabalho. Primeiramente, é interessante afirmar que enquanto a probabilidade é uma disciplina matemática desenvolvida como um modelo abstrato e suas deduções são baseadas em axiomas, a estatística lida com aplicações da teoria a problemas reais e suas conclusões são inferências baseadas em observações [64]. Em diversas aplicações, é mais viável retirar conclusões acerca de uma população inteira a partir de uma amostra aleatória. A inferência estatística consiste em retirar conclusões acerca da natureza de algum sistema baseado em dados sujeitos a variação aleatória [83].

Um conceito anterior necessário ao uso da teoria de probabilidade é o de σ-álgebra. Este conceito é utilizado no capítulo 3. Considere o conjunto universal S e a coleção $\mathcal{Q}$ de subconjuntos de S. $A, B \in \mathcal{Q}$ são dois elementos arbitrários dessa coleção e A_n, $n = 1, 2, \ldots$, é uma sequência de conjuntos com $A_n \in \mathcal{Q}$, $\forall n$. A coleção $\mathcal{Q}$ é chamada de campo de as seguintes propriedades são satisfeitas [85]

1. $S \in \mathcal{Q}$,
2. $A \in \mathcal{Q}$ implica $A^c \in \mathcal{Q}$,
3. $A, B \in \mathcal{Q}$ implica $A \cup B \in \mathcal{Q}$.

Um campo é chamado de σ-álgebra se a seguinte propriedade é satisfeita:

$$A_1, A_2, \cdots \in \mathcal{Q} \text{ implica } \cup_{n=1}^{\infty} A_n \in \mathcal{Q}. \tag{A.1}$$

Uma σ-álgebra portanto contém todas as uniões, finitas ou infinitas, de seus elementos, e só assim garante-se que operações arbitrárias de conjuntos sobre eventos criam conjuntos cujas probabilidades podem ser definidas [85]. Evento, em probabilidade, é um subconjunto de um espaço amostral, ao passo que espaço amostral de um determinado experimento aleatório é o conjunto de todos resultados possíveis do experimento aleatório [57].

Outro conceito importante de probabilidade é o de variável aleatória. Variável aleatória é uma função que associa a cada elemento do espaço amostral um número real. Uma amostra aleatória de uma variável aleatória X é um vetor de variáveis aleatórias $(X_1, X_2, \ldots, X_N)$, cada uma com a mesma distribuição de X [57].

Na investigação estatística, há duas classes gerais de problemas. Na primeira classe, há um modelo probabilístico conhecido e deseja-se fazer predições de futuras observações

— portanto, parte-se do modelo para as observações. Já na segunda classe, um ou mais parâmetros do modelo são desconhecidos e deseja-se estimá-los (estimação de parâmetro) ou decidir se estão em um determinado subconjunto de valores (teste de hipótese). Nesse caso, parte-se das observações para o modelo [64]. Ao longo da presente tese, o problema considerado é o da segunda classe.

O problema de estimação está bem definido para o caso paramétrico, ou seja, aquele em que se conhece a distribuição mas não se conhece um ou mais de seus parâmetros θ. Neste caso, estimador é uma função de uma amostra aleatória, i.e., $\hat{\theta} = g(X_1, X_2, \ldots, X_N)$, logo constitui outra variável aleatória. Seu valor específico para uma determinada realização amostral é chamada de estimativa $\hat{\theta} = g(x_1, x_2, \ldots, x_N)$. Qualquer função de uma amostra aleatória é chamada de estatística, logo, um estimador é uma estatística [64].

Diz-se que um estimador $\hat{\theta}$ não é enviesado do parâmetro θ se $\mathbb{E}\hat{\theta} = \theta$. Caso contrário, é enviesado com viés $b = \mathbb{E}\hat{\theta} - \theta$. A média amostral $\bar{X}$ de uma variável aleatória X é um exemplo de estimador não enviesado da esperança de X, $\mathbb{E}X$:

$$\begin{aligned}
\mathbb{E}\bar{X} &= \mathbb{E}\frac{1}{N}\sum_{n=1}^{N} X_n \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}X_n \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}X \\
&= \mathbb{E}X.
\end{aligned}$$

Se a função $g(X_1^N)$ for selecionada adequadamente, o erro de de estimação $\hat{\theta} - \theta$ diminui à medida que N aumenta [64]. Há vários critérios de convergência para uma sequência de variável aleatória, por exemplo:

- $\hat{\theta}_n$ converge em probabilidade se $\lim_{n\to\infty} P[|\hat{\theta}_n - \theta| > \epsilon] = 0$ [64];
- $\hat{\theta}_n$ converge com probabilidade 1 (ou quase certamente) para θ se $P[\lim_{n\to\infty} \hat{\theta}_n = \theta] = 1$ [13];

Um estimador é dito consistente se seu erro de estimação converge para zero em probabilidade. Além disso, diz-se que um estimador $\hat{\theta}_1$ de um parâmetro θ é mais acurado que um estimador $\hat{\theta}_2$ se [3]

$$\mathbb{E}(\hat{\theta}_1 - \theta)^2 < \mathbb{E}(\hat{\theta}_2 - \theta)^2. \tag{A.2}$$

Quanto menor o erro médio quadrático de um estimador em relação ao parâmetro que se deseja estimar, maior sua acurácia. Seja b o viés do estimador $\hat{\theta}$ do parâmetro θ de uma distribuição. É possível desenvolver a seguinte relação entre acurácia ($\mathbb{E}(\hat{\theta} - \theta)^2$), viés e variância de um estimador:

$$\begin{aligned}
b &= \mathbb{E}\hat{\theta} - \theta, \text{ viés do estimador,} \\
\mathbb{E}(\hat{\theta} - \theta)^2 &= \mathbb{E}(\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2) \\
&= \mathbb{E}\hat{\theta}^2 - 2\theta\mathbb{E}\hat{\theta} + \theta^2 \\
&= \mathbb{E}\hat{\theta}^2 - (\mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta})^2 - 2\theta\mathbb{E}\hat{\theta} + \theta^2 \\
&= \text{var}(\hat{\theta}) + (b + \theta)^2 - 2\theta(b + \theta) + \theta^2 \\
&= \text{var}(\hat{\theta}) + b^2.
\end{aligned}$$

Portanto, quanto menores a variância e o viés do estimador $\hat{\theta}$, maior será sua acurácia.

É de se esperar que o valor verdadeiro de um parâmetro da população esteja relacionado à amostra aleatória encontrada na realização de experimentos. Ou seja, é de se esperar que a função densidade de probabilidade $f(x_1, x_2, \ldots, x_N; \theta)$ dependa do valor de θ. Por exemplo, uma amostra aleatória de uma variável aleatória gaussiana apresenta valores que dependem de sua média. Dessa forma, em geral procura-se achar a melhor função dos dados para estimar um parâmetro, o que pode ser feito no sentido da máxima verossimilhança.

A estimativa de máxima verossimilhança (ML - *maximum likelihood*) do parâmetro θ associada à distribuição de probabilidade conjunta de $f(x_1, x_2, \ldots, x_n; \theta)$ de uma variável aleatória pode ser determinada a partir da equação de verossimilhança como

$$\hat{\theta}_{ML} = \sup_{\theta} f(x_1, x_2, \ldots, x_n; \theta) \tag{A.3}$$

ou através da função de log-verossimilhança

$$L(x_1, x_2, \ldots, x_n; \theta) = \ln f(x_1, x_2, \ldots, x_n; \theta). \tag{A.4}$$

Se $L(x_1, x_2, \ldots, x_n; \theta)$ for diferenciável e um supremo $\hat{\theta}_{ML}$ existir, então ele deve satisfazer à equação [64]:

$$\left.\frac{\partial \ln f(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{ML}} = 0 \tag{A.5}$$

Observa-se que a conexão entre a probabilidade e a realidade é baseada na frequência relativa de eventos, isto é, a probabilidade de um evento A costuma ser estimada pela contagem de quantas vezes ele ocorreu dividida pelo tamanho amostral [64]:

$$p = \frac{N_A}{N}. \tag{A.6}$$

Já está estabelecido que a frequência relativa é o estimador de máxima verossimilhança para um vetor de probabilidades de eventos discretos [26]. Apesar disso, o uso deste estimador para probabilidades que serão usadas no funcional de informação mútua, por exemplo, resulta em estimativas de informação mútua enviesadas. Diferentemente da estimação de entropia, em que o uso da frequência relativa para estimá-la resulta na média em estimativas menores que a verdadeira, não há uma direção para o viés da informação mútua quando a frequência relativa é utilizada [26]. O viés da estimação de informação mútua quando a frequência relativa é utilizada pode depender das distribuições conjuntas de probabilidade $p(X, Y)$ e o tamanho amostral N.

Além de inferir parâmetros, muitas vezes em estatística deseja-se retirar alguma conclusão acerca dos dados observados — verificar se há de fato algum padrão neles ou se o que foi obtido deve-se somente ao acaso. Para tanto, existem diversos testes de hipóteses, para alguns deles foi feito um breve resumo [4]:

- Testes paramétricos:
  - teste t pareado (compara a diferença entre as médias de variáveis dependentes);
  - teste t não pareado (compara a diferença entre as médias de variáveis independentes);
- Testes não paramétricos:

– Wilcoxon rank-sum (leva em conta magnitude e direção da diferença entre duas variáveis independentes);
– Wilcoxon sign-rank (leva em conta magnitude e direção da diferença entre duas variáveis dependentes);

Muitos testes paramétricos são robustos, apontando corretamente a evidência de uma hipótese ser verdadeira ou falsa ainda que a distribuição subjacente não seja exatamente aquela considerada para o modelo (normal, exponencial, binomial, etc.). Ao se testar uma mesma hipótese sob ambos tipos de teste — paramétrico e não paramétrico — há um resultado mais confiável no primeiro caso se a distribuição suposta pelo teste corresponder à verdadeira. Contudo, quando não se pode afirmar com certeza a lei de probabilidade envolvida, o teste não paramétrico é preferível [45].

Para compreender como os testes de hipóteses paramétricos funcionam, considera-se a questão a seguir. Dada uma variável X, que possua um modelo conhecido de distribuição, mas com parâmetro desconhecido θ, testa-se a suposição de que $\theta = \theta_0$ (hipótese nula, H_0) contra a suposição de que $\theta \neq \theta_0$ (hipótese alternativa, H_1). A hipótese alternativa também pode ser unilateral, isto é, testa-se se $\theta > \theta_0$ ou se $\theta < \theta_0$. As hipóteses são testadas baseadas em evidências experimentais, isto é, na amostra aleatória X_1^N observada. O propósito do teste não é determinar a veracidade de H_0 ou H_1, mas sim estabelecer se há evidências que apóiam a rejeição de H_0 [64].

Portanto, há dois tipos de erros que podem ocorrer em testes de hipótese:

- Erro do tipo I: rejeitar H_0 muito embora H_0 seja verdadeiro. A probabilidade de tal erro é denotada por α e é chamada de nível de significância do teste;
- Erro do tipo II: aceitar H_0 muito embora H_0 seja falso. A probabilidade de tal erro é denotada por β e a diferença $1 - \beta$ é chamada de potência do teste.

Apêndice B

Função Digamma

A função digamma é definida como a primeira derivada do logaritmo da função gamma ($\Gamma(x)$) [1]:

$$\begin{aligned} \psi(x) &= \frac{d}{dx}\ln\Gamma(x) \\ &= \frac{\Gamma^{'}(x)}{\Gamma(x)}, \end{aligned} \tag{B.1}$$

ao passo que a função $\Gamma(x)$ é dada por:

$$\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt, \tag{B.2}$$

para $x > 0$.

É possível derivar uma relação recursiva para a função $\psi(x)$:

$$\begin{aligned} \psi(x+1) - \psi(x) &= \frac{\Gamma^{'}(x+1)}{\Gamma(x+1)} - \frac{\Gamma^{'}(x)}{\Gamma(x)} \\ &= \frac{\Gamma^{'}(x+1)\Gamma(x) - \Gamma(x+1)\Gamma^{'}(x)}{\Gamma(x+1)\Gamma(x)} \\ &= \frac{\Gamma(x)^2\left(\Gamma(x+1)/\Gamma(x)\right)^{'}}{\Gamma(x+1)\Gamma(x)} \qquad \text{(B.3)} \\ &= \frac{\Gamma(x)^2}{\Gamma(x+1)\Gamma(x)} \qquad \text{(B.4)} \\ &= \frac{1}{x}, \end{aligned}$$

em que (B.3) utilizou-se a propriedade da derivada de uma fração de funções e em (B.4) utilizou-se a equação funcional de Euler [1]:

$$\Gamma(x+1) = x\Gamma(x). \tag{B.5}$$

Apêndice C

Distribuição de Dirichlet e Ponderação de Probabilidades

A distribuição de Dirichlet de parâmetros $\alpha_1, \alpha_2, \ldots, \alpha_k$ tem como suporte o simplex [1]

$$S_k = \{(\theta_1, \theta_2, \ldots, \theta_k) | \theta_1 \geq 0, \theta_2 \geq 0, \ldots, \theta_k \geq 0, \theta_1 + \theta_2 + \cdots + \theta_k = 1\}, \tag{C.1}$$

em que $k = 2, 3, \ldots$, e é determinada pela densidade de probabilidade

$$p(\theta_1, \theta_2, \ldots, \theta_k) = \begin{cases} C_k \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}, \text{ se } (\theta_1, \theta_2, \ldots, \theta_k) \in S_k, \\ 0, \text{ caso contrário,} \end{cases} \tag{C.2}$$

em que $\alpha_1 \geq 0, \alpha_2 \geq 0, \ldots, \alpha_k \geq 0$ e

$$C_k = \Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k) \prod_{i=1}^{k} \frac{1}{\Gamma(\alpha_i)}, \tag{C.3}$$

em que Γ é a função gamma.

Para o caso particular do capítulo 3, deseja-se estimar a probabilidade P_e de uma fonte binária de parâmetro θ desconhecido gerar a zeros e b uns. A distribuição de Dirichlet é um modelo de como as probabilidades variam, e os parâmetros α_i estão relacionados às amostras observadas (proporções dos dados) [52]. Na derivação do capítulo 3, a distribuição de Dirichlet é obtida com vetor de parâmetros $(\alpha_1, \alpha_2) = (1/2, 1/2)$. Assim tem-se:

$$p(\theta, 1-\theta) = C_2 \theta^{1/2-1} (1-\theta)^{1/2-1} = C_2 \frac{1}{\sqrt{\theta(1-\theta)}}, \tag{C.4}$$

e

$$C_2 = \frac{\Gamma(1/2 + 1/2)}{\Gamma(1/2)\Gamma(1/2)} = \frac{1}{\pi}, \tag{C.5}$$

de modo que

$$p(\theta, 1-\theta) = p(\theta) = \frac{1}{\pi\sqrt{\theta(1-\theta)}}, \tag{C.6}$$

que é a probabilidade usada para ponderar o parâmetro θ de 0 até 1 na equação (3.10).

Apêndice D

Cálculo das Probabilidades da Cadeia de Markov Estacionária (Exemplo do Capítulo 4)

Neste apêndice encontra-se a derivação das probabilidades estacionárias dos estados 1, 10 e 00 da cadeia de Markov ilustrada na Fig. 4.13, do capítulo 4.

Primeiramente, no estado estacionário observa-se a seguinte relação:

$$\begin{aligned} (\pi_1\ \pi_{10}\ \pi_{00})T &= (\pi_1\ \pi_{10}\ \pi_{00}) \\ (\pi_1\ \pi_{10}\ \pi_{00}) \begin{bmatrix} \theta_1 & (1-\theta_1) & 0 \\ \theta_{10} & 0 & 1-\theta_{10} \\ \theta_{00} & 0 & 1-\theta_{00} \end{bmatrix} &= (\pi_1\ \pi_{10}\ \pi_{00}), \end{aligned} \tag{D.1}$$

em que T é a matriz de transição do processo $\mathbf{X}$.

Acrescida à igualdade D.1, deve ser satisfeita a seguinte relação:

$$\pi_1 + \pi_{10} + \pi_{00} = 1.$$

Resolvendo este sistema de equações, obtém-se, para os valores do texto de $\theta_1 = 0.1$, $\theta_{10} = 0.3$, $\theta_{00} = 0.5$:

$$\begin{aligned} \pi_1 &= 0.32, \\ \pi_{10} &= 0.28, \\ \pi_{00} &= 0.40. \end{aligned}$$

Apêndice E

Cálculo do Exemplo Analítico de Entropia de Transferência para Caso Contínuo

Neste apêndice encontra-se a derivação da expressão analítica de entropia de transferência entre variáveis aleatórias contínuas, como descrito no capítulo 6. As variáveis envolvidas são $X_{n+1} = \alpha X_n + \eta_n^X$ e $Y_{n+1} = \beta Y_n + \gamma X_n + \eta_n^Y$, e supõe-se que $\eta_n^X \sim \mathcal{N}(0,1)$, $\eta_n^Y \sim \mathcal{N}(0,1)$, X_n, Y_n são conjuntamente estacionários, $\mathbb{E}X_n = \mathbb{E}Y_n = 0$ e todos os processos envolvidos são gaussianos. Deseja-se encontrar a entropia de transferência de $\mathbf{X}$ para $\mathbf{Y}$.

Primeiramente, sabe-se que:

$$\begin{aligned} TE(X \to Y) &= \mathrm{D}(P(Y_{n+1}|Y_n X_n)||P(Y_{n+1}|Y_n)) \\ &= \mathbb{E}_{P(Y_{n+1},Y_n,X_n)} \log \frac{P(Y_{n+1}|Y_n X_n)}{P(Y_{n+1}|Y_n)} \\ &= \mathbb{E} \log \frac{P(Y_{n+1},Y_n,X_n)}{P(Y_n,X_n)} - \mathbb{E} \log \frac{P(Y_{n+1},Y_n)}{P(Y_n)} \\ &= -H(Y_{n+1},Y_n,X_n) + H(Y_n,X_n) + H(Y_{n+1},Y_n) - H(Y_n) \end{aligned} \tag{E.1}$$

No caso de processo gaussiano multivariado (Y^n), tem-se

$$H(Y^n) = -\frac{1}{2} \ln[(2\pi e)^n \det(K(Y^n))], \tag{E.2}$$

em que

$$K(Y^n) = \begin{pmatrix} \mathrm{cov}(Y_1,Y_1) & \mathrm{cov}(Y_1,Y_2) & \cdots & \mathrm{cov}(Y_1,Y_n) \\ \mathrm{cov}(Y_2,Y_1) & \mathrm{cov}(Y_2,Y_2) & \cdots & \mathrm{cov}(Y_2,Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_n,Y_1) & \mathrm{cov}(Y_n,Y_2) & \cdots & \mathrm{cov}(Y_n,Y_n) \end{pmatrix}$$

Considerando a expressão E.1, encontra-se:

$$TE(X \to Y) = \frac{1}{2} \ln \frac{\det(K(Y_n,X_n)) \det(K(Y_{n+1},Y_n))}{\det(K(Y_{n+1},Y_n,X_n))\sigma_Y^2}, \tag{E.3}$$

em que σ_Y^2 é a variância de Y_n.

As matrizes de covariância são escritas como segue:

$$K(Y_{n+1}, Y_n, X_n) = \begin{pmatrix} \text{cov}(Y_{n+1}, Y_{n+1}) & \text{cov}(Y_{n+1}, Y_n) & \text{cov}(Y_{n+1}, X_n) \\ \text{cov}(Y_n, Y_{n+1}) & \text{cov}(Y_n, Y_n) & \text{cov}(Y_n, X_n) \\ \text{cov}(X_n, Y_{n+1}) & \text{cov}(X_n, Y_n) & \text{cov}(X_n, X_n) \end{pmatrix}$$

$$K(Y_{n+1}, Y_n) = \begin{pmatrix} \text{cov}(Y_{n+1}, Y_{n+1}) & \text{cov}(Y_{n+1}, Y_n) \\ \text{cov}(Y_n, Y_{n+1}) & \text{cov}(Y_n, Y_n) \end{pmatrix}$$

$$K(Y_n, X_n) = \begin{pmatrix} \text{cov}(Y_n, Y_n) & \text{cov}(Y_n, X_n) \\ \text{cov}(X_n, Y_n) & \text{cov}(X_n, X_n) \end{pmatrix}$$

Como $\mathbf{Y}$ é estacionário, tem-se $\text{cov}(Y_{n+1}, Y_{n+1}) = \text{cov}(Y_n, Y_n)$. Então, calcula-se:

$$\begin{aligned} \text{cov}(X_n, X_n) &= \mathbb{E}[(\alpha X_{n-1} + \eta^X_{n-1})(\alpha X_{n-1} + \eta^X_{n-1})] \\ &= \mathbb{E}[\alpha^2 X^2_{n-1} + 2\alpha X_{n-1}\eta^X_{n-1} + (\eta^X_{n-1})^2] \\ &= \alpha^2 \mathbb{E}X^2_{n-1} + 2\alpha \mathbb{E}X_{n-1}\mathbb{E}\eta^X_{n-1} + \sigma^2_{\eta^X} \\ &= \alpha^2 \text{cov}(X_{n-1}, X_{n-1}) + \sigma^2_{\eta^X} \\ (1-\alpha^2)\text{cov}(X_n, X_n) &= \sigma^2_{\eta^X} = 1 \\ \text{cov}(X_n, X_n) &= 1/(1-\alpha^2) = u. \end{aligned}$$

$$\begin{aligned} \text{cov}(X_n, Y_n) &= \mathbb{E}[X_n Y_n] \\ &= \mathbb{E}[X_n(\beta Y_{n-1} + \gamma X_{n-1} + \eta^Y_{n-1})] \\ &= \beta \mathbb{E}[X_n Y_{n-1}] + \gamma \mathbb{E}[X_n X_{n-1}] \\ \text{cov}(X_n, Y_{n-1}) &= \mathbb{E}(X_n Y_{n-1}) \\ &= \mathbb{E}[X_n(\beta Y_{n-2} + \gamma X_{n-2} + \eta^Y_{n-2})] \\ &= \mathbb{E}[(\alpha X_{n-1} + \eta^X_{n-1})(\beta Y_{n-2} + \gamma X_{n-2} + \eta^Y_{n-2})] \\ &= \mathbb{E}[\alpha\beta X_{n-1}Y_{n-2} + \alpha\gamma X_{n-1}X_{n-2} + \alpha X_{n-1}\eta^Y_{n-2} + \\ &\quad + \beta\eta^X_{n-1}Y_{n-2} + \gamma\eta^Y_{n-1}X_{n-2} + \eta^X_{n-1}\eta^Y_{n-2}] \\ &= \alpha\beta \mathbb{E}(X_{n-1}Y_{n-2}) + \alpha\gamma \mathbb{E}(X_{n-1}X_{n-2}) \end{aligned}$$

$$\begin{aligned} (1-\alpha\beta)\mathbb{E}(X_n Y_{n-1}) &= \alpha\gamma \mathbb{E}(X_n X_{n-1}) \\ \mathbb{E}[X_{n-1}X_n] &= \mathbb{E}[X_{n-1}(\alpha X_{n-1} + \eta^X_{n-1})] \\ &= \alpha \mathbb{E}(X^2_{i-1}) \\ &= \alpha \text{cov}(X_n, X_n) \\ &= \alpha u \\ \mathbb{E}(X_n Y_{n-1}) &= \frac{\alpha^2 \gamma u}{1-\alpha\beta} \\ \text{cov}(X_n, Y_n) &= \beta \mathbb{E}(X_n Y_{n-1}) + \gamma \mathbb{E}(X_n X_{n-1}) \end{aligned}$$

Chamando $w = 1/(1-\alpha\beta)$:

$$\text{cov}(X_n, Y_n) = \beta\alpha^2\gamma uw + \gamma\alpha u = \alpha\gamma uw$$

$$
\begin{aligned}
\mathrm{cov}(Y_{n+1}, Y_{n+1}) &= \mathbb{E}[(\beta Y_n + \gamma X_n + \eta_n^Y)(\beta Y_n + \gamma X_n + \eta_n^Y)] \\
&= \mathbb{E}[\beta^2 Y_n^2 + 2\beta\gamma Y_n X_n + 2\beta Y_n \eta_n^Y + 2\gamma X_n \eta_n^Y + \\
&\quad + \gamma^2 X_n^2 + (\eta_n^Y)^2] \\
&= \beta^2 \mathbb{E} Y_n^2 + 2\beta\gamma \mathbb{E} X_n Y_n + \gamma^2 \mathbb{E} X_n^2 + \sigma_{\eta^Y}^2 \\
\mathrm{cov}(Y_n, Y_n)[1 - \beta^2] &= 2\beta\gamma^2 \alpha u w + \gamma^2 u + 1
\end{aligned}
$$

Chamando $v = 1/(1-\beta^2)$, tem-se:

$$\mathrm{cov}(Y_n, Y_n) = v + \gamma^2(1 + \alpha\beta)uvw$$

.

$$
\begin{aligned}
\mathrm{cov}(Y_{n+1}, Y_n) &= \mathbb{E}[(\beta Y_n + \gamma X_n + \eta_n^Y) Y_n] \\
&= \beta \mathbb{E} Y_n^2 + \gamma \mathbb{E} X_n Y_n + \mathbb{E}(\eta_n^Y Y_n) \\
&= \beta \mathrm{cov}(Y_n, Y_n) + \gamma \mathrm{cov}(X_n, Y_n) \\
&= \beta v + \gamma^2(\alpha + \beta) uvw \\
\mathrm{cov}(Y_{n+1}, X_n) &= \mathbb{E}[(\beta Y_n + \gamma X_n + \eta_n^Y) X_n] \\
&= \beta \mathbb{E}(X_n Y_n) + \gamma \mathbb{E}(X_n X_n) \\
&= \alpha\beta\gamma uw + \gamma u
\end{aligned}
$$

Agora substituindo as expressões de covariâncias encontradas nas matrizes de covariância $K(Y_{n+1}, Y_n, X_n)$, $K(Y_{n+1}, Y_n)$ e $K(Y_n, X_n)$ é possível encontrar a entropia de transferência deste exemplo. Em particular, para os valores de $\alpha = 0.5$, $\beta = 0.6$ e $\gamma = 0.4$, encontra-se:

$$TE(X \to Y) = 0.0923 \; nats$$

.

Apêndice F

Lista de Publicações do Doutorado

- Juliana M. de Assis, Mikaelle O. Santos, Francisco M. de Assis. "Auditory Stimuli Coding by Postsynaptic Potential and Local Field Potential Features". Plos One, v. 11, p. e0160089, 2016;

- Juliana M. de Assis, Edmar C. Gurjão. "The Use of Discrete Prolate Spheroidal Sequences and Trig Prolates to Compressed Sensing". XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2016, Santarém, Pará. Anais do XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2016. v. 1. p. 294-298;

- Juliana M. de Assis, Francisco M. de Assis. "An Application of Directed Information to Infer Synaptic Connectivity". XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2016, Santarém, Pará. Anais do XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2016. v. 1. p. 528-532;

- Juliana M. de Assis, Francisco M. de Assis. "Estimation of Directed Information to Processes Assuming Continuous Values with CTW Algorithm". XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2017, São Pedro, São Paulo. Anais do XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2017. v. 1. p. 225-229.

- (Submetido, 20 de julho de 2017) Juliana M. de Assis, Francisco M. de Assis. "Estimation of Transfer Entropy between Discrete and Continuous Random Processes". Journal of Communication and Information Systems.

Referências Bibliográficas

[1] Encyclopedia of mathematics. https://www.encyclopediaofmath.org/index.php/Dirichlet_distribution. Acessado: 10 de novembro de 2016.

[2] Nonlinear integrate-and-fire models. http://neuronaldynamics.epfl.ch/online/Ch5.html. Acessado: 07 de agosto de 2017.

[3] Série estatística básica - estimação. http://www.mat.ufrgs.br/~viali/estatistica/mat2246/material/apostilas/A5_Estimacao.pdf. Acessado: 14 de fevereiro de 2017.

[4] Types of statistical tests. https://cyfar.org/types-statistical-tests. Acessado: 28 de setembro de 2016.

[5] Pierre-Olivier Amblard and Olivier J. J. Michel. The relation between granger causality and directed information theory: A review. *Entropy*, 15:113–143, 2013.

[6] Pierre-Olivier Amblard and Olivier JJ Michel. On directed information theory and granger causality graphs. *Journal of computational neuroscience*, 30(1):7–16, 2011.

[7] Ehsan Arabzadeh, Stefano Panzeri, and Mathew E Diamond. Whisker vibration information carried by rat barrel cortex neurons. *Journal of Neuroscience*, 24(26):6011–6020, 2004.

[8] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.

[9] Margret Bauer, John W Cox, Michelle H Caveness, James J Downs, and Nina F Thornhill. Finding the direction of disturbance propagation in a chemical process using transfer entropy. *IEEE transactions on control systems technology*, 15(1):12–21, 2007.

[10] Andrei Belitski, Arthur Gretton, Cesare Magri, Yusuke Murayama, Marcelo A Montemurro, Nikos K Logothetis, and Stefano Panzeri. Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *Journal of Neuroscience*, 28(22):5696–5709, 2008.

[11] JG Bernstein and ES Boyden. Optogenetic tools for analyzing the neural circuits of behavior. *Trends Cogn Sci*, 15(12):592–600, 2011.

[12] Emery N Brown, Riccardo Barbieri, Uri T Eden, and Loren M Frank. Likelihood methods for neural spike train data analysis. *Computational neuroscience: A comprehensive approach*, pages 253–286, 2003.

[13] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[14] Georges A Darbellay, Igor Vajda, et al. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

[15] Juliana M. de Assis. Medida de uma codificação de estímulos sonoros por potenciais pós-sinápticos e potenciais de campo local usando teoria da informação. Master's thesis, Brain Institute, UFRN, Natal, 2014.

[16] Juliana M. de Assis, Mikaelle O. Santos, and Francisco M. de Assis. Auditory stimuli coding by postsynaptic potential and local field potential features. *PLoS One*, 11(8), 2016.

[17] Stanislas Dehaene. *Consciousness and the Brain*. Viking Press, 2014.

[18] Ping Duan, Fan Yang, Sirish L Shah, and Tongwen Chen. Transfer zero-entropy and its application for capturing cause and effect relationship between variables. *IEEE Transactions on Control Systems Technology*, 23(3):855–867, 2015.

[19] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *arXiv preprint arXiv:1604.03006*, 2016.

[20] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1267–1271. IEEE, 2017.

[21] Felipe Gerhard, Tilman Kispersky, Gabrielle J Gutierrez, Eve Marder, Mark Kramer, and Uri Eden. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. *PLoS Comput Biol*, 9(7):e1003138, 2013.

[22] Germán Gómez-Herrero, Wei Wu, Kalle Rutanen, Miguel C Soriano, Gordon Pipa, and Raul Vicente. Assessing coupling dynamics from an ensemble of time series. *Entropy*, 17(4):1958–1970, 2015.

[23] Boris Gourévitch and Jos J Eggermont. Evaluating information transfer between auditory cortical neurons. *Journal of Neurophysiology*, 97(3):2533–2543, 2007.

[24] C. W. J. Granger. Economic processes involving feedback. *Information and Control*, 6:28–48, 1963.

[25] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[26] Ivo Grosse. *Applications of statistical physics and information theory to the analysis of DNA sequences*. PhD thesis, Boston University, 2000.

[27] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[28] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and conditional mean estimation in poisson channels. *IEEE Transactions on Information Theory*, 54(5):1837–1849, 2008.

[29] Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

[30] Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for gaussian mixture random vectors. In *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pages 181–188. IEEE, 2008.

[31] Robin AA Ince, Riccardo Senatore, Ehsan Arabzadeh, Fernando Montani, Mathew E Diamond, and Stefano Panzeri. Information-theoretic methods for studying population codes. *Neural Networks*, 23(6):713–727, 2010.

[32] Shinya Ito, Michael E Hansen, Randy Heiland, Andrew Lumsdaine, Alan M Litke, and John M Beggs. Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. *PLoS one*, 6(11):e27431, 2011.

[33] Eugene M Izhikevich et al. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

[34] Jiantao Jiao, Haim H.Permuter, Lei Zhao, Young-Han Kim, and Tsachy Weissman. Universal estimation of directed information. *IEEE Transactions on Information Theroy*, 59(10):6220–6242, 2013.

[35] A Kaiser and T Schreiber. Information transfer in continuous processes. *Physica D: Nonlinear Phenomena*, 166(1):43–62, 2002.

[36] Eric R Kandel, James H Schwartz, Thomas M Jessell, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

[37] Christoph Kayser, Marcelo A Montemurro, Nikos K Logothetis, and Stefano Panzeri. Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron*, 61(4):597–608, 2009.

[38] Jean-Rémi King, Jacobo D Sitt, Frédéric Faugeras, Benjamin Rohaut, Imen El Karoui, Laurent Cohen, Lionel Naccache, and Stanislas Dehaene. Information sharing in the brain indexes consciousness in noncommunicative patients. *Current Biology*, 23(19):1914–1919, 2013.

[39] Murray S. Klamkin, editor. *Problems in applied mathematics: selections from SIAM review*. Society for Industrial and Applied Mathematics, 1990.

[40] Bryan Kolb and Ian Q. Whishaw. Brain plasticity and behavior. *Annu Rev Psychol.*, 49(7):43–64, 1998.

[41] LF Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[42] Gerhard Kramer. *Directed Information for Channels with Feedback*. PhD thesis, Swiss Federal Institute of Technology, Zurich, 1998.

[43] Alexander Kraskov. *Synchronization and Interdependence Maesures and their Applications to the Electroencephalogram of Epilepsy Patients and Clustering of Data*. PhD thesis, Universität Wuppertal, Fakultät für Mathematik und Naturwissenschaften» Physik» Dissertationen, 2004.

[44] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review*, 69, 2004.

[45] Harold J. Larson. *Introduction to Probability Theory and Statistical Inference*. John Wiley and Sons, 1982.

[46] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[47] Wentian Li. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60(5-6):823–837, 1990.

[48] Ying Liu and Selin Aviyente. The relationship between transfer entropy and directed information. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 73–76. IEEE, 2012.

[49] Ying Liu, Selin Aviyente, and Mahmood Al-khassaweneh. A high dimensional directed information estimation using data-dependent partitioning. In *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 606–609. IEEE, 2009.

[50] Joseph T. Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014.

[51] Muriel Lobier, Felix Siebenhühner, Satu Palva, and J Matias Palva. Phase transfer entropy: a novel phase-based measure for directed connectivity in networks coupled by oscillatory interactions. *Neuroimage*, 85:853–872, 2014.

[52] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[53] Cesare Magri, Kevin Whittingstall, Vanessa Singh, Nikos K Logothetis, and Stefano Panzeri. A toolbox for the fast information analysis of multiple-site lfp, eeg and spike train recordings. *BMC neuroscience*, 10(1):1, 2009.

[54] Rakesh Malladi, Giridhar Kalamangalam, Nitin Tandon, and Behnaam Aazhang. Identifying seizure onset zone from the causal connectivity inferred using directed information. *IEEE Journal of Selected Topics in Signal Processing*, 10(7):1267–1283, 2016.

[55] James Massey. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pages 303–305. Citeseer, 1990.

[56] Mariusz Maziarz. A review of the granger-causality fallacy. *The Journal of Philosophical Economics: Reflections on Economic and social issues*, 8(2):86–105, 2015.

[57] Paul L. Meyer. *Probabilidade - Aplicações à Estatística*. Livros Técnicos e Científicos Editora, 1983.

[58] Todd K Moon and Wynn C Stirling. *Mathematical methods and algorithms for signal processing*. Prentice Hall, 2000.

[59] Yonathan Murin, Jeremy Kim, and Andrea Goldsmith. Tracking epileptic seizure activity via information theoretic graphs. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 583–587. IEEE, 2016.

[60] John A Nelder and R Jacob Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[61] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet. On hölder projective divergences. *Entropy*, 19(3):122, 2017.

[62] Milan Paluš. Testing for nonlinearity using redundancies: Quantitative and qualitative aspects. *Physica D: Nonlinear Phenomena*, 80(1):186–205, 1995.

[63] Stefano Panzeri, Riccardo Senatore, Marcelo A. Montemurro, and Rasmus S. Petersen. Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, 98:1064–1072, 2007.

[64] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. Mc Graw Hill, 2002.

[65] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[66] Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[67] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J Comput Neurosci*, 30:17–44, 2011.

[68] Rodrigo Quian Quiroga and Stefano Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10:173–185, 2009.

[69] Rodrigo Quian Quiroga and Stefano Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185, 2009.

[70] Brian C. Ross. Mutual information between discrete and continuous data sets. *PLoS One*, 9(e87357), 2014.

[71] Nicolas P. Rougier. Implicit and explicit representations. *Neural Networks*, 22(2):155–160, 2009.

[72] Jakob Runge, Maik Riedl, Andreas Müller, Holger Stepan, Jürgen Kurths, and Niels Wessel. Quantifying the causal strength of multivariate cardiovascular couplings with momentary information transfer. *Physiological measurement*, 36(4):813, 2015.

[73] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[74] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

[75] Nima Soltani and Andrea Goldsmith. Directed information between connected leaky integrate-and-fire neurons. *IEEE International Symposium on Information Theory*, 2014.

[76] Sen Song, Per Jesper Sjöström, Markus Reigl, Sacha Nelson, and Dmitri B Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS biology*, 3(3):e68, 2005.

[77] Matthäus Staniek and Klaus Lehnertz. Symbolic transfer entropy. *Physical Review Letters*, 100(15):158101, 2008.

[78] Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Comput Biol*, 8(8):e1002653, 2012.

[79] Steven P Strong, Roland Koberle, Rob R de Ruyter van Steveninck, and William Bialek. Entropy and information in neural spike trains. *Physical review letters*, 80(1):197, 1998.

[80] Prapun Suksompong and Toby Berger. Capacity analysis for integrate-and-fire neurons with descending action potential thresholds. *IEEE Transactions on Information Theory*, 56(2):838–851, 2010.

[81] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[82] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[83] Graham Upton and Ian Cook. *A dictionary of statistics 3e*. Oxford university press, 2014.

[84] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy - a model-free measure of effective connectivity for the neurosciences. *J Comput Neurosci*, 30:45–67, 2011.

[85] Yannis Viniotis. *Probability and Random Processes*. McGrall-Hill International Editions, 2000.

[86] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

[87] Marcelo J. Weinberger, Jorma J. Rissanen, and Meir Feder. A universal finite memory source. *IEEE Transactions on Information Theory*, 41(3):643–652, 1995.

[88] Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T Lizier, and Raul Vicente. Measuring information-transfer delays. *PLoS one*, 8(2):e55809, 2013.

[89] Norbert Wiener. The theory of prediction. *Modern mathematics for engineers*, 1:125–139, 1956.

[90] Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41(3):653–664, 1995.