



Agenda

- Introdução**
 - Números Aleatórios
 - Testes
 - Descritores baseados em Teoria da Informação
 - Delimitação do Problema
- Proposta**
 - Fundamentação Teórica
- Resultados**
 - Análise global das sequências
 - Histogramas suavizados
 - Análise das regiões de confiança
 - Aplicações
- Conclusão**
- Referências Bibliográficas**



Introdução

Números Aleatórios



Figura: Visão de Dilbert de um gerador de números aleatórios

Introdução

Números Aleatórios



Números aleatórios perfazem uma das partes mais importantes em aplicações computacionais nos vários campos do conhecimento, como aborda Knuth (1998):

- ▶ Simulação
- ▶ Amostragem
- ▶ Programação de Computadores
- ▶ Tomada de Decisão
- ▶ Criptografia
- ▶ Estética
- ▶ Diversão

Introdução

GNAs e GNPA



Geradores de Números Aleatórios - GNAs

A maior parte dos geradores utiliza-se de fenômenos físicos naturais como, decaimento radioativo, ruidos termais em semicondutores, amostras de som num local ruinoso, ruído no espectro eletromagnético, dentre outros que, por óbvia dedução, carecem de algum hardware específico para serem capturados, o que dificulta sua obtenção.

Geradores de Números Pseudo-Aleatórios - GNPA

Dadas as dificuldades descritas anteriormente, atualmente a maneira mais conveniente e confiável de se gerar números aleatórios para diversas aplicações é através de algoritmos com um sólido embasamento matemático.

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo (UNIFESP) | Faculdade de Informática

Introdução

GNAs e GNPA



Figura: Setup do Gerador de Números Aleatórios de www.random.org (1998)

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo (UNIFESP) | Faculdade de Informática

Introdução

Testes



Existem duas abordagens para testar-se a capacidade de geradores aleatórios ou pseudoaleatórios produzirem sequências ditas aleatórias. Segundo L'Ecuyer (1992) são elencados em teóricos e empíricos.

Testes teóricos

Os testes teóricos são bastante específicos para cada tipo de GNPA, pois analisam as propriedades das sequências a partir da definição do gerador.

Testes empíricos

Já os testes empíricos valem-se de técnicas estatísticas objetivando avaliar o quanto boas são as sequências produzidas por um determinado gerador.

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo (UNIFESP) | Faculdade de Informática

Introdução

Testes



NIST

Desde 1997, o Grupo de Trabalho Técnico em Geração de Números Aleatórios (RNG-TWG) tem trabalhado no desenvolvimento de uma bateria de testes estatísticos apropriados para a avaliação de geradores de números aleatórios e pseudoaleatórios utilizados em aplicações criptográficas.

Diehard[er]

George Marsaglia desenvolveu a bateria de testes Diehard em 1995, e os disponibilizou em CD-ROM. Robert Brown identificou limitações nessa bateria de testes, os implementou novamente na linguagem de programação C, acrescentou testes da bateria NIST e disponibilizou um conjunto ampliado de testes denominado Dieharder.

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo (UNIFESP) | Faculdade de Informática

Introdução

Testes



ENT

ENT Walker (2017) realiza uma variedade de testes no fluxo de bytes de um arquivo (ou na entrada padrão se nenhum arquivo for especificado). É mantido no Fournilab (<http://www.fournilab.ch/random/>) por John Walker.

TestU01

Considerado como o estado da arte dos testes para geradores de números aleatórios L'Ecuyer and Simard (2007), o TestU01 se apresenta como uma biblioteca de software escrita em ANSI C que oferece uma coleção de utilitários para testagem, ele provê implementações generalistas dos testes estatísticos clássicos para geradores de números aleatórios, bem como de vários outros propostos na literatura, além de propor alguns originais.

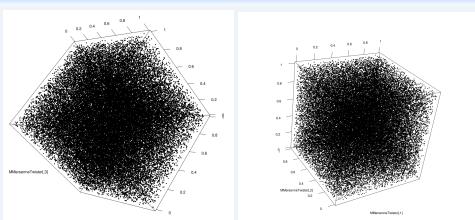
Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Informática

Introdução

Testes



Visualização



(a) Primeira perspectiva

(b) Segunda perspectiva

Figura: Visualização 3D de sequências disjuntas produzidas pelo gerador Mersenne-Twister

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Informática

Introdução

Testes



Visualização

O intuito desses testes é a verificação de que os dados produzidos por um gerador, seja ele algorítmico ou físico, não se afastam significativamente da hipótese de serem eventos de variáveis independentes e identicamente distribuídas segundo uma lei Uniforme no intervalo $(0, 1]$.

A componente mais difícil é a de verificar a independência, por se tratar de independência coletiva e não apenas aos pares.

A falta de independência de uma sequência de N pontos pode se manifestar de várias maneiras. Uma delas é quando os pontos jazem em subespaços de dimensão menor a N , ao invés de preencher o espaço por completo.

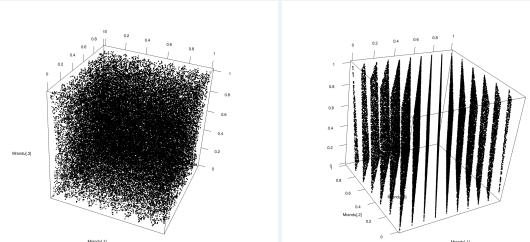
Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Informática

Introdução

Testes



Visualização



(a) Primeira perspectiva

(b) Segunda perspectiva

Figura: Visualização 3D de sequências disjuntas produzidas pelo gerador RANDU

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Informática

Introdução

Descritores baseados em Teoria da Informação



Descritores baseados em Teoria da Informação

O trabalho pioneiro de Bandt and Pompe (2002) representa uma mudança de paradigma na análise de séries temporais, e será o nosso marco referencial teórico.

Eles propõem uma técnica não-paramétrica de análise de sequências que consiste em transformar palavras de D dados não necessariamente subsequentes em símbolos ordinais.

Esses símbolos codificam a ordem que as D observações têm na sequência e, portanto, são bem menos suscetíveis a contaminação dos que os próprios valores.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - Centro de Ciências da Computação - Programa de Pós-Graduação em Ciência da Computação - Teoria da Informação

Introdução

Testes



Visualização

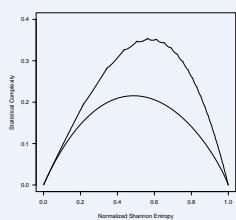


Figura: Plano Entropia-Complexidade.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - Centro de Ciências da Computação - Programa de Pós-Graduação em Ciência da Computação - Teoria da Informação

Introdução

Descritores baseados em Teoria da Informação



Descritores baseados em Teoria da Informação

Descritores baseados em Teoria da Informação

Forma-se então um histograma de proporções dos símbolos observados, e calculam-se duas quantidades: a Entropia e a Divergência de Jensen-Shannon a uma distribuição de referência (usualmente a uniforme). A série, finalmente, é representada pelo par de valores Entropia-Complexidade Estatística, sendo que esta última é o produto da Entropia e a Divergência de Jensen-Shannon.

O conjunto de valores possíveis dos pontos característicos de qualquer série não varre \mathbb{R}^2 , mas constitui-se em um subconjunto compacto do plano: o plano Entropia-Complexidade.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - Centro de Ciências da Computação - Programa de Pós-Graduação em Ciência da Computação - Teoria da Informação

Introdução

Testes



Introdução

Descritores baseados em Teoria da Informação



Algumas aplicações emblemáticas dessa abordagem:

Larrondo et al. (2006) mostram que o plano Entropia-Complexidade permite prever o resultado dos testes Diehard de qualidade de GNPA.

Martin et al. (2006) analisam o mapa caótico logístico e discutem cotas no plano Entropia-Complexidade.

Rosso et al. (2006) analisam dados de eletroencefalogramas de pacientes com epilepsia utilizando decomposição wavelet e o plano Entropia-Complexidade.

De Micco et al. (2008) avaliam melhorias da qualidade de sequências pseudoaleatórias.

Carpí et al. (2011) analisam a evolução de redes dinâmicas com descritores baseados em Teoria da Informação.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - Centro de Ciências da Computação - Programa de Pós-Graduação em Ciência da Computação - Teoria da Informação

Introdução

Descritores baseados em Teoria da Informação



Algumas aplicações emblemáticas dessa abordagem:

Cabral et al. (2013) utilizam descritores de Teoria da Informação para descrever a dinâmica de redes de sensores sem fios.

Aquino et al. (2015) analisam o comportamento de veículos em larga escala em função da topologia de diversas cidades.

Rosso et al. (2016) mostram a expressividade dos descritores de Teoria da Informação para o problema de classificação e verificação de assinaturas.

Aquino et al. (2017) conseguem caracterizar o tipo de dispositivo elétrico observando o ponto no plano Entropia-Complexidade em que o histórico do seu consumo é mapeado.

Introdução

Delimitação do Problema



Delimitação do Problema

A motivação deste trabalho é o desenvolvimento de um teste baseado em Teoria da Informação para verificar a hipótese de que uma sequência é ruído branco, isto é, formada por observações de variáveis aleatórias independentes e identicamente distribuídas (*vaiid*).

Para tanto, trabalharemos com atributos derivados da simbolização de Bandt & Pompe. Cada sequência sob análise será transformada em um ponto no plano Entropia-Complexidade, e será medida a sua distância ao ponto característico de sequências ideais.

Analisaremos, então, a distribuição empírica de uma variedade de situações de interesse para, finalmente, propor regiões de confiança da hipótese nula. Por fim, analisaremos sequências com o ferramental aqui proposto.

Proposta

Fundamentação Teórica



Simbolização

Dada uma série temporal a tempo discreto $X = x_t : 1 \leq t \leq M$ de comprimento N , uma dimensão D e um tempo de atraso (*delay*) τ , o particionamento é efetuado por meio da reorganização do sistema em conjuntos seguindo os passos:

- ▶ **Composição dos grupos:** Os conjuntos, ou palavras, de comprimento D e *delay* τ são definidas por um segmento da série:

$$(x_{t+1}, x_{t+\tau}, \dots, x_{t+D\tau}).$$

- ▶ **Formação dos padrões:** Cada palavra é então relacionada a um padrão ordinal π_j de ordem D , isto é, um elemento indexado univocamente por

$$j \in \{1, 2, \dots, D-1, D\}.$$

Proposta

Fundamentação Teórica



Simbolização

Neste trabalho utilizaremos a atribuição lexicográfica, isto é, se os valores da palavra $(x_{t+1}, x_{t+\tau}, \dots, x_{t+D\tau})$ são tais que, ordenados, eles têm índices crescentes b_1, b_2, \dots, b_D , então o padrão correspondente será

$$\pi = b_1 b_2 \dots b_D$$

Calcula-se, então, o histograma de proporções

$$\mathcal{H} = (p_1, \dots, p_D)$$

dos padrões observados:

$$p_j = \frac{1}{t - N + 1} \# \{\text{padrões } \pi_j \text{ observados}\}.$$

O seguinte passo consiste em calcular descritores a partir desse histograma.



Simbolização

Os trabalhos já citados utilizam dois descritores: a Entropia de Shannon e a Complexidade Estatística da série.

A Entropia de Shannon é definida como

$$E(\mathcal{H}) = - \sum_{j=1}^{D!} p_j \log p_j,$$

Esta é uma medida da desordem ou imprevisibilidade da lei subjacente a \mathcal{H} . A entropia apenas não consegue caracterizar de forma plena a dinâmica que produz a série. Torna-se interessante, então, o uso de um outro descritor baseado em quão diferente o histograma \mathcal{H} é de uma lei de probabilidade de referência. A nossa referência será a lei uniforme, e a medida de distância entre elas será a distância de Jensen-Shannon:

$$JS(\mathcal{H}, \mathcal{H}_R) = E((\mathcal{H} + \mathcal{H}_R)/2) - \frac{1}{2}(E(\mathcal{H}) + E(\mathcal{H}_R)).$$

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Física | Programa de Pós-Graduação em Física



Regiões de confiança no plano Entropia-Complexidade

Não conhecemos, contudo, a distribuição conjunta do par (E, JS) para uma sequência de variáveis aleatórias coletivamente independente e identicamente distribuídas segundo uma lei uniforme. Como também não conhecemos a distribuição do par (H, C) decidimos fazer a análise empírica de dados obtidos de fontes “confiáveis”.

Para isso, utilizamos três fontes de dados: duas físicas e uma algorítmica. As fontes físicas foram dados de medidas de estados quânticos Gabriel et al. (2010) (que denominaremos *sequências quânticas*) e de sinais de rádio Stamey (2016) (que chamaremos *sequências de rádio*). A fonte algorítmica é o gerador Mersenne-Twister Matsumoto and Nishimura (1998), considerado um padrão de qualidade de algoritmos de geração de números pseudoaleatórios.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Física | Programa de Pós-Graduação em Física



Regiões de confiança no plano Entropia-Complexidade

Obtivemos com os autores 54 000 000 de observações de cada gerador físico, e as submetemos à análise dos padrões de Bandt & Pompe.

Obtivemos os valores de entropia e complexidade estatística para todas as sequências possíveis de tamanho 18 000, palavras de tamanho $D \in \{3, 4, 5, 6\}$ e $lag \tau \in \{1, 10, 30, 50\}$.

Segundo a proposta de Bandt (2017), nossa medida de qualidade é a distância do ponto padrão observado (E, C) no plano Entropia-Complexidade ao ponto ideal $(1, 0)$. Empregamos, como esse autor, a distância euclidiana.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Física | Programa de Pós-Graduação em Física



Análise global das sequências

Trabalhamos com sequências de observações provindas de três geradores: dois físicos (considerados “verdadeiramente aleatórios”) e um algorítmico (o gerador Mersenne-Twister, que é reputado um dos melhores geradores pseudoaleatórios).

Para cada gerador considerado coletamos sequências disjuntas de tamanho 1.000 e 50.000, de cada tamanho de palavra D e a cada $lag \tau$. Temos, assim, quatro fatores a serem analisados.

Cada sequência passou pelo processo de simbolização, e foi calculado o histograma dos símbolos. Foram então calculados os valores de Entropia e de Complexidade de cada histograma, bem como a distância euclidiana desses valores ao ponto de referência $(1, 0)$.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Física | Programa de Pós-Graduação em Física

Resultados

Análise global das sequências

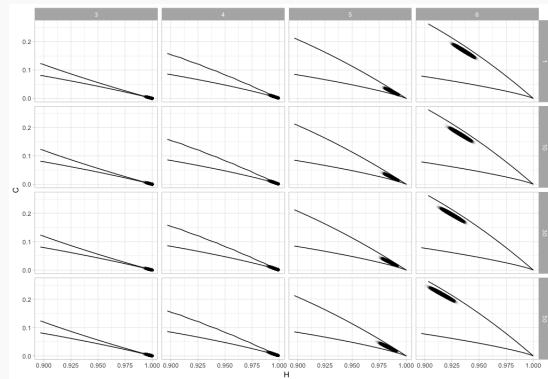


Figura: Diagramas de dispersão das sequências quânticas com 1.000 observações

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Análise global das sequências

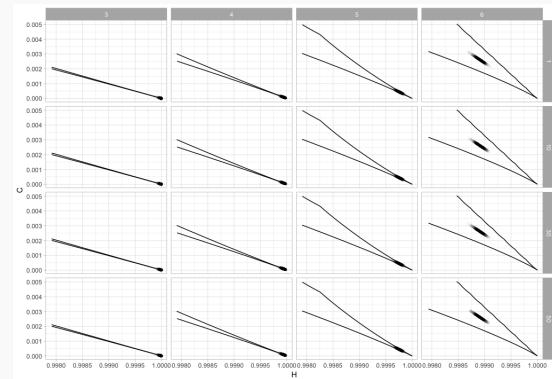


Figura: Diagramas de dispersão das sequências quânticas com 50.000 observações

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Análise global das sequências

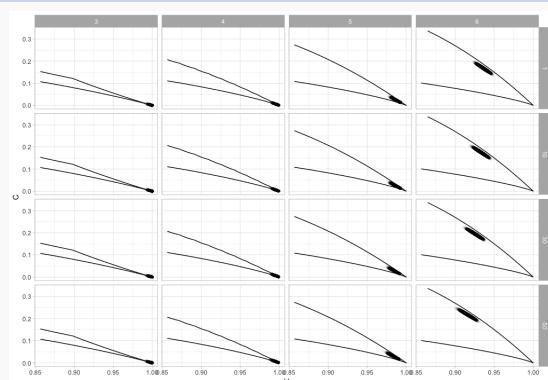


Figura: Diagramas de dispersão das sequências de rádio com 1.000 observações

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Análise global das sequências

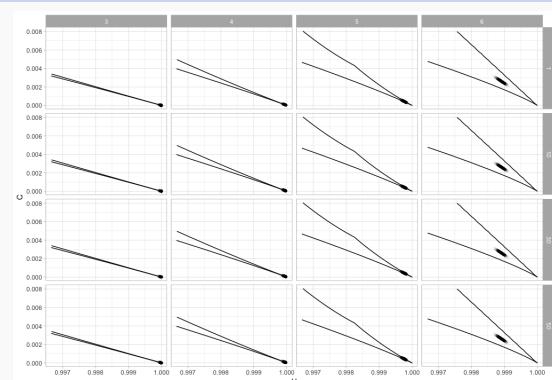


Figura: Diagramas de dispersão das sequências de rádio com 50.000 observações

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Análise global das sequências

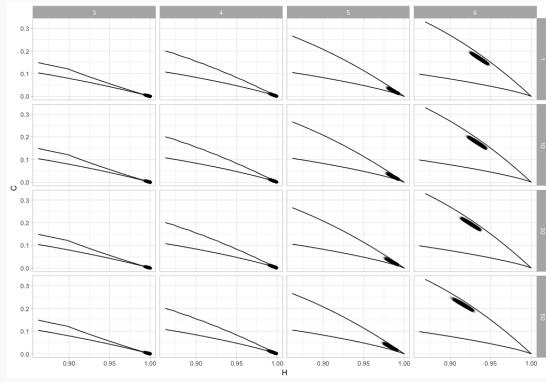


Figura: Diagramas de dispersão de Mersenne-Twister com 1.000 observações

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Análise global das sequências

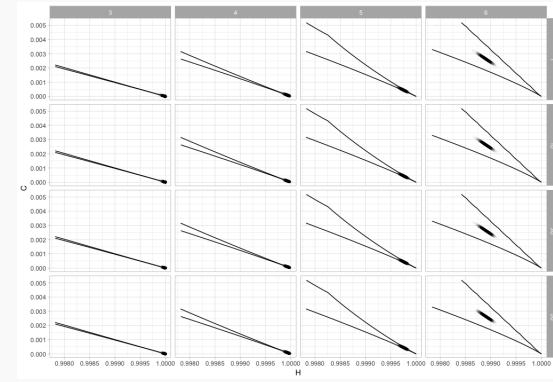


Figura: Diagramas de dispersão de Mersenne-Twister com 50.000 observações

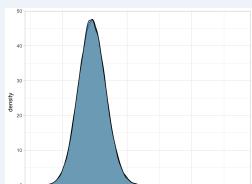
Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

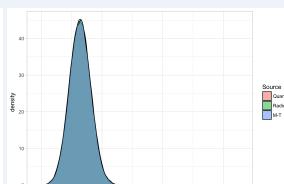
Histogramas suavizados



A olho nu, D é um fator relevante pois os diagramas de dispersão mostram comportamentos que merecem uma análise mais aprofundada, por outro lado não temos certeza de como τ e o gerador influenciam os resultados.



(a) $D = 6, \tau = 1$



(b) $D = 6, \tau = 50$

Figura: Histogramas suavizados de situações que sugerem que o gerador é um fator irrelevante para $N = 1.000$

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Histogramas suavizados

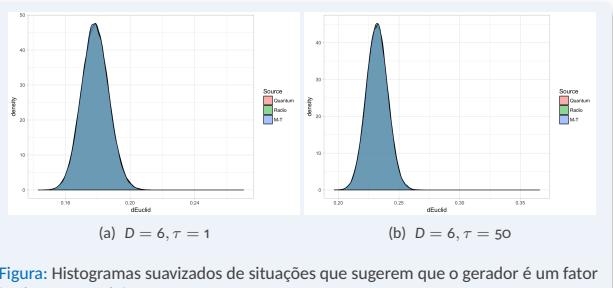


Figura: Histogramas suavizados de situações que sugerem que o gerador é um fator irrelevante também para $N = 50.000$

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Computação

Resultados

Histogramas suavizados



A figura a seguir mostra os histogramas suavizados das distâncias dos três geradores para palavras de tamanho $D = 6$ e dois valores de lag ($\tau = 1, 50$).

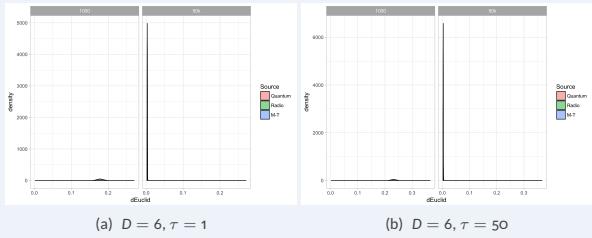


Figura: Histogramas suavizados de situações que sugerem que o N é um fator relevante

Resultados

Histogramas suavizados



A seguir mostramos os histogramas suavizados das distâncias dos três geradores para dois tamanhos de sequências ($N = 1.000, 50.000$) e $D = 6$.

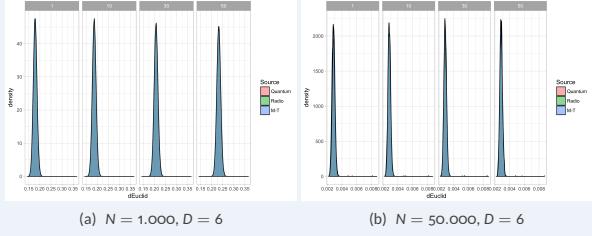


Figura: Histogramas suavizados de situações que sugerem que o τ é um fator relevante

Resultados

Histogramas suavizados



A seguir mostramos os histogramas suavizados das distâncias dos três geradores para dois tamanhos de sequências ($N = 1.000, 50.000$) e dois valores de lag ($\tau = 1, 50$).

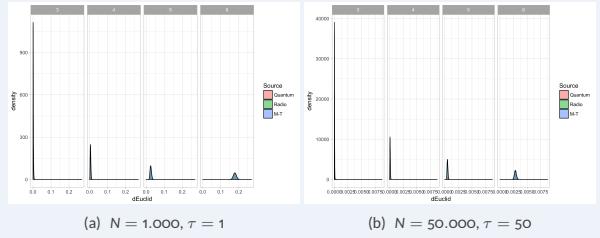


Figura: Histogramas suavizados de situações que sugerem que o D é um fator relevante

Resultados

Teste de Kolmogorov-Smirnov



Teste de Kolmogorov-Smirnov aplicado a pares de sequências para 1.000 observações

Par	D	$\tau = 1$	$\tau = 10$	$\tau = 30$	$\tau = 50$
Quântica vs. Rádio	$D = 3$	0.18034676	0.08582490	0.58096350	0.32626542
	$D = 4$	0.21708776	0.60690204	0.08116764	0.46372312
	$D = 5$	0.32388371	0.53394280	0.46970138	0.02768674
	$D = 6$	0.61501858	0.63403661	0.54795745	0.15353799
Quântica vs. M-T	$D = 3$	0.09400120	0.22995096	0.36766759	0.03706359
	$D = 4$	0.25188769	0.35844686	0.16768952	0.18237754
	$D = 5$	0.97552039	0.79878301	0.12852918	0.08764347
	$D = 6$	0.47615384	0.42420007	0.55290011	0.79669144
Rádio vs. M-T	$D = 3$	0.008560614	0.496450214	0.982419336	0.390237891
	$D = 4$	0.003804157	0.229503619	0.629158543	0.651783589
	$D = 5$	0.254216237	0.179697451	0.824743440	0.071709252
	$D = 6$	0.846441994	0.033726860	0.493286733	0.184856861

Resultados

Teste de Kolmogorov-Smirnov



43

Teste de Kolmogorov-Smirnov aplicado a pares de sequências para 50.000 observações

Par	D	$\tau = 1$	$\tau = 10$	$\tau = 30$	$\tau = 50$
Quântica vs. Rádio	D = 3	0.13862662	0.93677447	0.07714702	0.46405291
	D = 4	0.68079537	0.90035466	0.60801914	0.77908261
	D = 5	0.14371256	0.76662067	0.64456996	0.91315843
	D = 6	0.02268670	0.49307044	0.53135926	0.30074267
Quântica vs. M-T	D = 3	6.074571e-10	1.388898e-01	2.682058e-01	4.822849e-01
	D = 4	6.592620e-09	2.987721e-02	8.438134e-01	7.923220e-01
	D = 5	9.424114e-08	3.289299e-02	7.681676e-01	8.405168e-01
	D = 6	9.058821e-04	7.075225e-01	2.982731e-01	3.614763e-01
Rádio vs. M-T	D = 3	1.226271e-09	1.609665e-01	1.465970e-01	1.914937e-01
	D = 4	2.190462e-10	1.277762e-02	2.069821e-01	9.963221e-01
	D = 5	1.438372e-11	1.267622e-02	4.364935e-01	9.999998e-01
	D = 6	3.749001e-06	2.263466e-01	7.419248e-01	4.014641e-01

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Medicina de São Paulo - FMSP

Resultados



45

Os p -valores observados na Tabela de teste de Kolmogorov-Smirnov aplicado a pares de sequências para 50.000 observações nos levam a concluir que não é possível desconsiderar a fonte de dados como um fator relevante quando se trata do gerador de Mersenne-Twister. Já as distâncias das sequências produzidas pelos geradores Quântico e de Rádio são indistinguíveis e, portanto, não podemos descartar a hipótese dessas fontes serem idênticas para a medida considerada.

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Medicina de São Paulo - FMSP

Resultados

Teste de Kolmogorov-Smirnov

Resultados

Teste de Kolmogorov-Smirnov



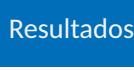
44

Teste de Kolmogorov-Smirnov aplicado a pares de sequências para 50.000 observações

Par	D	$\tau = 1$	$\tau = 10$	$\tau = 30$	$\tau = 50$
Quântica vs. Rádio	D = 3	0.13862662	0.93677447	0.07714702	0.46405291
	D = 4	0.68079537	0.90035466	0.60801914	0.77908261
	D = 5	0.14371256	0.76662067	0.64456996	0.91315843
	D = 6	0.02268670	0.49307044	0.53135926	0.30074267
Quântica vs. M-T	D = 3	6.074571e-10	1.388898e-01	2.682058e-01	4.822849e-01
	D = 4	6.592620e-09	2.987721e-02	8.438134e-01	7.923220e-01
	D = 5	9.424114e-08	3.289299e-02	7.681676e-01	8.405168e-01
	D = 6	9.058821e-04	7.075225e-01	2.982731e-01	3.614763e-01
Rádio vs. M-T	D = 3	1.226271e-09	1.609665e-01	1.465970e-01	1.914937e-01
	D = 4	2.190462e-10	1.277762e-02	2.069821e-01	9.963221e-01
	D = 5	1.438372e-11	1.267622e-02	4.364935e-01	9.999998e-01
	D = 6	3.749001e-06	2.263466e-01	7.419248e-01	4.014641e-01

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Medicina de São Paulo - FMSP

Resultados



46

A figura a seguir sugere que o comportamento das distâncias euclidianas ao ponto de referência muda conforme o tamanho do padrão D varia. Além disso, também são perceptíveis mudanças de comportamento em função do $lag \tau$ quando $D = 5, 6$.

Marcelo Queiroz de Assis Oliveira | Universidade Federal de São Paulo - UNIFESP | Faculdade de Medicina de São Paulo - FMSP

Resultados



Resultados

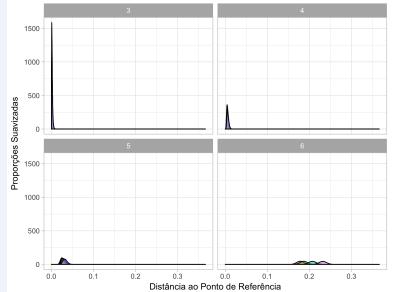


Figura: Histogramas suavizados das distâncias euclidianas dos padrões ao ponto de referência, para $D \in \{3, 4, 5, 6\}$ e $\tau \in \{1, 10, 30, 50\}$.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Informação

Resultados

Análise das regiões de confiança



Sequências “Aleatórias”

Feita a junção das distâncias dos pontos característicos ao ponto de referência das sequências produzidas pelos geradores quântico e de rádio (sequências aleatórias), para cada situação de $N=(1.000, 50.000)$, $D=(3, 4, 5, 6)$ e $\tau=(1, 10, 30, 50)$, o próximo passo consiste em calcular os quantis relevantes.

Inicialmente ilustraremos apenas duas situações. As figuras subsequentes mostram os padrões das sequências aleatórias para o caso $N=1.000$, $D=3$, $\tau=1$ e para o caso $N=50.000$, $D=6$, $\tau=50$ no plano Entropia-Complexidade, junto com os quantis de 90 %, 95 %, 99 % e 99.9 % em escala linear e em escala logarítmica.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Informação

Da análise aqui apresentada concluímos que, à luz da distância do ponto característico de uma sequência ao ponto de referência, os dois geradores físicos produzem sequências indistinguíveis. Diante disso, nos cálculos subsequentes faremos a junção desses conjuntos de dados, que denominaremos simplesmente “aleatórios” no que segue.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Informação

Resultados

Análise das regiões de confiança



Análise das regiões de confiança

Nestas figuras, os pontos característicos foram desenhados com um gradiente de cores que vai do amarelo ao preto em função da distância ao ponto de referência. Identificamos também os valores de Entropia e de Complexidade Estatística correspondentes a cada quantil de interesse, estes últimos plotados em vermelho.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná - UFPR | Departamento de Ciências da Informação

Resultados

Análise das regiões de confiança

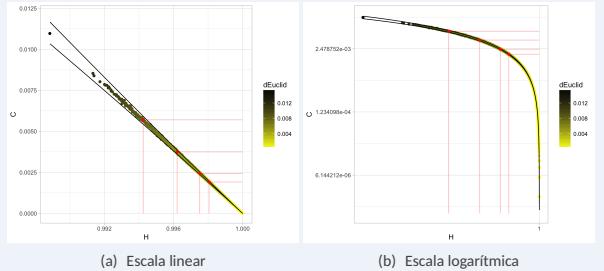


Resultados

Análise das regiões de confiança



Análise das regiões de confiança



(a) Escala linear

(b) Escala logarítmica

Figura: Diagramas de dispersão das sequências aleatórias para o caso $N = 1.000$, $D = 3$ e $\tau = 1$.

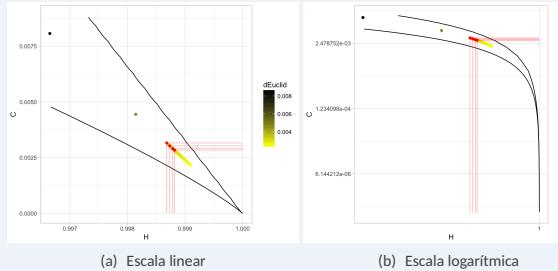
Marcelo Queiroz de Assis Oliveira | Mestrado em Ciência da Computação - Universidade Federal do Paraná - Programa de Pós-Graduação

Resultados

Análise das regiões de confiança



Análise das regiões de confiança



(a) Escala linear

(b) Escala logarítmica

Figura: Diagramas de dispersão das sequências aleatórias para o caso $N = 50.000$, $D = 6$ e $\tau = 50$.

Marcelo Queiroz de Assis Oliveira | Mestrado em Ciência da Computação - Universidade Federal do Paraná - Programa de Pós-Graduação

Resultados

Análise das regiões de confiança



Resultados

Análise das regiões de confiança



Análise das regiões de confiança

As figuras a seguir apresentam os resultados centrais dessa dissertação, nelas exibimos os quantis de interesse para amostras de tamanho 1.000 e $\tau=(1, 10, 30, 50)$, logo em seguida para amostras de tamanho 50.000 e, assim como anteriormente, $\tau=(1, 10, 30, 50)$. Cada uma destas figuras inclui as quatro situações de $D=(3, 4, 5, 6)$.

Adicionalmente aos gráficos são apresentadas tabelas, mostrando os valores da distância euclidiana dos pontos de interesse dos quantis ao ponto de referência.

Marcelo Queiroz de Assis Oliveira | Mestrado em Ciência da Computação - Universidade Federal do Paraná - Programa de Pós-Graduação

Resultados

Análise das regiões de confiança

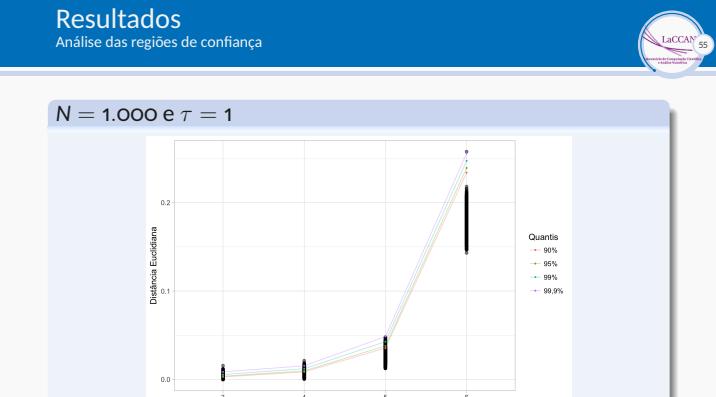


Figura: Intervalos de confiança para o caso $N = 1.000$ e $\tau = 1$.

Marcelo Queiroz de Assis Oliveira | Mestrado em Ciência da Computação - Universidade Federal do Paraná - Programa de Pós-Graduação

Resultados

Análise das regiões de confiança



$N = 1.000 \text{ e } \tau = 10$

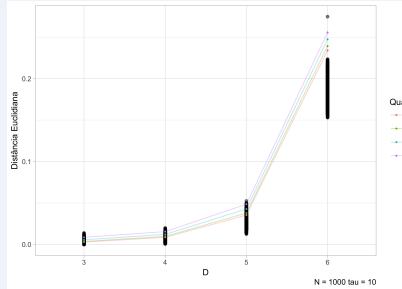


Figura: Intervalos de confiança para o caso $N = 1.000 \text{ e } \tau = 10$.

Resultados

Análise das regiões de confiança



$N = 1.000 \text{ e } \tau = 50$

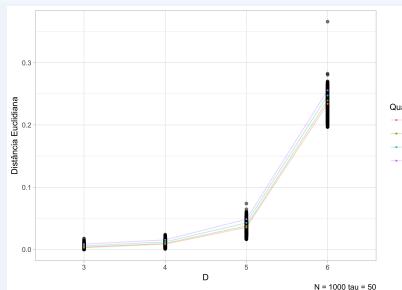


Figura: Intervalos de confiança para o caso $N = 1.000 \text{ e } \tau = 50$.

Resultados

Análise das regiões de confiança

Resultados

Análise das regiões de confiança



$N = 1.000 \text{ e } \tau = 30$

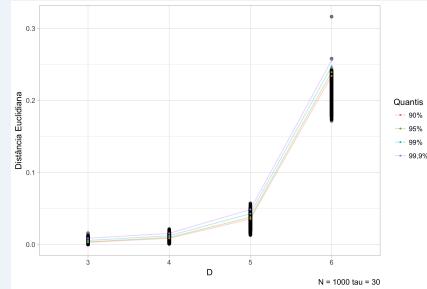


Figura: Intervalos de confiança para o caso $N = 1.000 \text{ e } \tau = 30$.

Resultados

Análise das regiões de confiança



Resultados

Análise das regiões de confiança



Quantis das distâncias, sequências de 1.000 observações.

$N = 1.000$	D	τ	90 %	95 %	99 %	99.9 %
3	1	2.728065e-03	3.478919e-03	0.0053313857	0.0081048900	
3	10	2.802528e-03	3.577539e-03	0.0054960059	0.0083091257	
3	30	2.961344e-03	3.749702e-03	0.0056871429	0.0087472165	
3	50	3.120298e-03	3.950138e-03	0.0059008109	0.0090272065	
4	1	7.9640766e-03	9.015899e-03	0.0112777244	0.0143372255	
4	10	8.199472e-03	9.295153e-03	0.0116255590	0.0147762539	
4	30	8.738506e-03	9.883617e-03	0.0123589751	0.0156235388	
4	50	9.368242e-03	1.054840e-02	0.0131188349	0.0166817556	
5	1	3.117067e-02	3.304803e-02	0.0366895915	0.0413215927	
5	10	3.235895e-02	3.425898e-02	0.0380480169	0.0427998033	
5	30	3.545788e-02	3.752600e-02	0.0417425086	0.0467667352	
5	50	3.914194e-02	4.142425e-02	0.0459567507	0.0514563584	
6	1	1.891794e-01	1.923893e-01	0.1984529319	0.2050583463	
6	10	1.975446e-01	2.007669e-01	0.2069269144	0.2139321164	
6	30	2.185870e-01	2.218804e-01	0.2280312709	0.2345272440	
6	50	2.431686e-01	2.464034e-01	0.2526103765	0.2596196960	

Resultados

Análise das regiões de confiança



$N = 50.000$ e $\tau = 1$

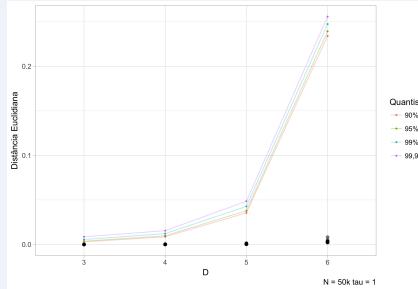


Figura: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 1$.

Resultados

Análise das regiões de confiança



$N = 50.000$ e $\tau = 30$

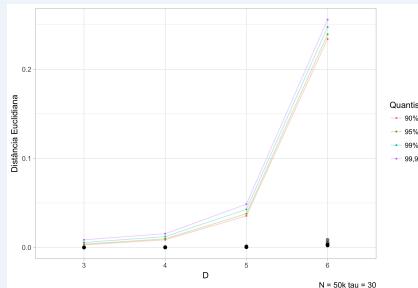


Figura: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 30$.

Resultados

Análise das regiões de confiança

Resultados

Análise das regiões de confiança



$N = 50.000$ e $\tau = 10$

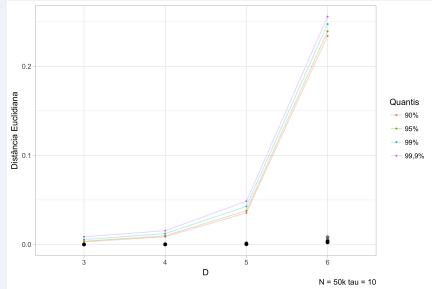


Figura: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 10$.

Resultados

Análise das regiões de confiança

Resultados

Análise das regiões de confiança



$N = 50.000$ e $\tau = 50$

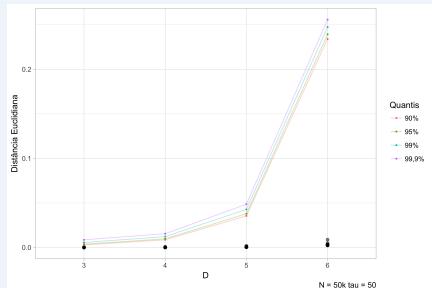


Figura: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 50$.

Resultados



Resultados Aplicações

Quantis das distâncias, sequências de 50.000 observações.

$N = 50.000$	D	τ	90 %	95 %	99 %	99.9 %
3	1	5.407679e-05	7.015910e-05	0.0001178519	0.0001917689	
3	10	5.636160e-05	7.130762e-05	0.0001000853	0.0001585019	
3	30	5.731458e-05	7.245769e-05	0.000101931	0.0001580179	
3	50	5.951595e-05	7.541980e-05	0.0001093136	0.0001983728	
4	1	1.589081e-04	1.818144e-04	0.0002250318	0.0003038889	
4	10	1.588585e-04	1.790301e-04	0.0002264259	0.0002903681	
4	30	1.631504e-04	1.863802e-04	0.0002330694	0.0002893557	
4	50	1.619028e-04	1.809200e-04	0.0002311957	0.0003017229	
5	1	6.062508e-04	6.389830e-04	0.0007179730	0.0008832568	
5	10	5.985032e-04	6.289105e-04	0.0007041024	0.0007855387	
5	30	6.040569e-04	6.400727e-04	0.0007268196	0.0008213228	
5	50	6.055769e-04	6.381391e-04	0.0007134216	0.0008042788	
6	1	3.071590e-03	3.150152e-03	0.0033104814	0.0035923810	
6	10	3.050841e-03	3.177779e-03	0.0032553996	0.0033734229	
6	30	3.066360e-03	3.129649e-03	0.0032669327	0.0034581195	
6	50	3.074748e-03	3.147889e-03	0.0032828968	0.0034178859	

Marcelo Queiroz de Assis Oliveira |

Universidade Federal do Rio Grande do Sul - UFRGS | Mestrado em Engenharia Civil - Programa de Pós-Graduação

Aplicações

Nesta seção mostramos a aplicação da nossa proposta a sequências de tamanho 1000.

Utilizando a metodologia descrita, aplicamos o teste a uma seqüência de 1000 observações produzidas pelos geradores Mersenne-Twister e Randu, além de séries não estacionárias, estacionárias e mapas logísticos, todos com a mesma dimensão.

Resultados Aplicações



Aplicações - Mersenne-Twister e Randu

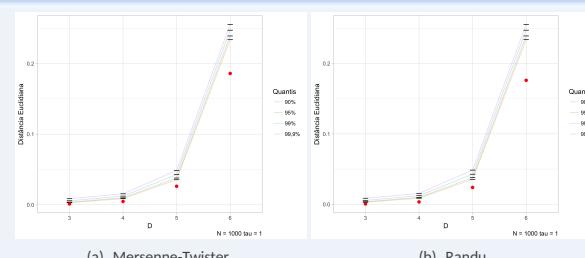


Figura: Aplicação do teste aos pontos de Mersenne-Twister e Randu

Marcelo Queiroz de Assis Oliveira |

Universidade Federal do Rio Grande do Sul - UFRGS | Mestrado em Engenharia Civil - Programa de Pós-Graduação

Resultados Aplicações



Aplicações - Estruturas autocorrelacionadas

A seguir, geramos sequências estocásticas com estrutura de autocorrelação:

- ▶ **Estacionária** - ruído gaussiano filtrado;
- ▶ **Não estacionária** - uma trajetória de movimento browniano.

Marcelo Queiroz de Assis Oliveira |

Universidade Federal do Rio Grande do Sul - UFRGS | Mestrado em Engenharia Civil - Programa de Pós-Graduação

Resultados

Aplicações



Aplicações - Séries não estacionárias

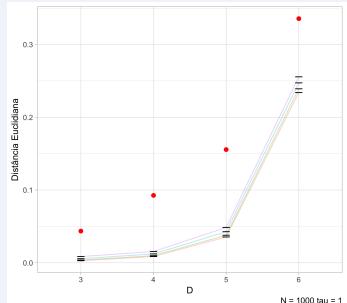


Figura: Pontos característicos das séries não estacionárias e intervalos de confiança.

Marcelo Queiroz de Assis Oliveira | Teste para Verificação da Hipótese de Ruído Branco Utilizando Teoria da Informação

Resultados

Aplicações



Aplicações - Série Estacionária

Uma série estacionária em que aplicamos nosso teste foi obtida pela convolução de uma sequência de variáveis aleatórias independentes e identicamente distribuídas segundo uma lei gaussiana padrão, convolucionada com uma máscara de tamanho 3 com valores não negativos: $(\beta, 1, \beta)$, $0 \leq \beta \leq 1$.

Quando $\beta=0$ temos a sequência original, e para valores crescentes de β temos sequências com cada vez maior estrutura de correlação. Calculamos o ponto característico da série assim obtida, e o contrastamos com os quantis empíricos obtidos anteriormente.

Marcelo Queiroz de Assis Oliveira | Teste para Verificação da Hipótese de Ruído Branco Utilizando Teoria da Informação

Resultados

Aplicações



Aplicações - Poder do teste aplicado a Séries Estacionárias

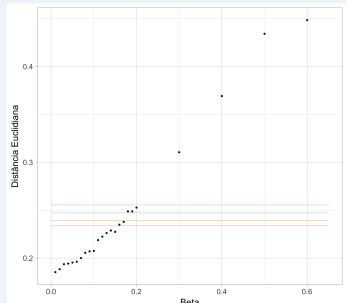


Figura: Poder do teste aplicado a uma sequência de séries estacionárias no caso particular $N = 1.000$, $\tau = 1$, variando-se a máscara de convolução $(\beta, 1, \beta)$.

Marcelo Queiroz de Assis Oliveira | Teste para Verificação da Hipótese de Ruído Branco Utilizando Teoria da Informação

Resultados

Aplicações



Aplicações - Poder do teste aplicado a Séries Estacionárias

No gráfico desenhamos também os quantis que correspondem a essa situação, e verificamos quais situações não são rejeitadas a cada nível de significância:

90 %: máscaras com $\beta \leq 0, 15$.

95 %: máscaras com $\beta \leq 0, 17$.

99 %: máscaras com $\beta \leq 0, 17$.

99.9 %: máscaras com $\beta \leq 0, 20$.

Estes resultados sugerem a necessidade de se fazer uma análise exaustiva do poder do teste em função de uma diversidade de parâmetros. Esse estudo é objeto de trabalhos futuros.

Marcelo Queiroz de Assis Oliveira | Teste para Verificação da Hipótese de Ruído Branco Utilizando Teoria da Informação

Resultados

Aplicações



Aplicações - Série Estacionária

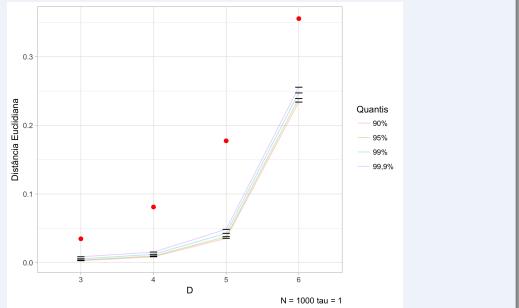


Figura: Pontos característicos das séries estacionárias e intervalos de confiança.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Ciências Exatas e da Terra | Faculdade de Informática

Resultados

Aplicações



Aplicações - Mapa logístico

A seguir analisamos uma série determinística com comportamento caótico: o mapa logístico. O mapa logístico é a sequência obtida pela recursão:

$$x_{n+1} = rx_n(1 - x_n), \quad (1)$$

com $0 < x_1 < 1$ e $0 < r \leq 4$. Utilizamos $r = 4$, $x_0 = 0.01$.

Iteramos o mapa 10 000 vezes para alcançar estabilidade, e só coletamos a partir de $n = 10001$.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Ciências Exatas e da Terra | Faculdade de Informática

Resultados

Aplicações



Aplicações - Mersenne-Twister e Randu

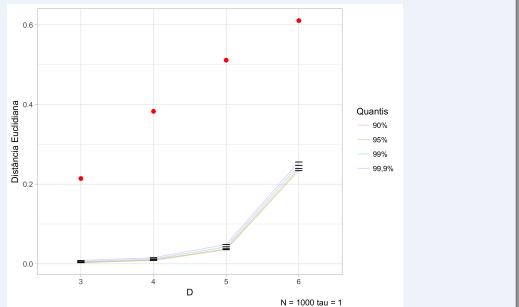


Figura: Pontos característicos do mapa logístico e intervalos de confiança.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Ciências Exatas e da Terra | Faculdade de Informática

Conclusão



Conclusão

Neste trabalho analisamos a possibilidade de a distância euclidiana de pontos no plano ($H \times C$) de sequências ao ponto $(1, 0)$, referência teórica de ruído branco, poderem ser usadas como uma estatística de teste para a hipótese de a sequência ser ruído branco. Verificamos que essa possibilidade existe, e que essa estatística é capaz de identificar, com limitações, o mapa logístico (que já foi usado como gerador de números pseudoaleatórios), movimento browniano e ruído com autocorrelação. Para este último, fizemos uma análise preliminar do poder do teste em função da intensidade da correlação.

Marcelo Queiroz de Assis Oliveira | Universidade Federal do Paraná | Departamento de Ciências Exatas e da Terra | Faculdade de Informática

Conclusão



Conclusão

Verificamos, também, que os geradores Mersenne-Twister e Randu são considerados ruído branco, mesmo sendo elas técnicas algorítmicas de geração de observações pseudoaleatórias. Uma limitação deste trabalho é que apenas verificamos a qualidade do gerador em relação a um de estrutura ideal. Com isso, limitamos a aplicabilidade do nosso trabalho à análise de séries que, potencialmente, são ocorrências de variáveis aleatórias independentes e identicamente distribuídas.

Conclusão



Conclusão

Há farta literatura que caracteriza diferentes tipos de estruturas como, por exemplo, processos estocásticos do tipo f^{-k} . A nossa metodologia pode, em princípio, ser aplicada a quaisquer processos mas, para isso, é necessário o conhecimento da distribuição dos padrões ordinários do processo de referência. No nosso caso, trata-se da lei uniforme sobre os padrões, que é característica de ruído branco. Não conhecemos resultados que caracterizem de forma teórica as leis de outros processos.

Há, contudo, uma solução para esse problema: estimar a lei característica do padrão de interesse. Isso pode ser feito através de estudos Monte Carlo, mas tal extensão foge ao objetivo deste trabalho.

Referências



- Aquino, A. L. L., Cavalcante, T. S. G., Almeida, E. S., Frery, A. C., and Rosso, O. A. (2015). Characterization of vehicle behavior with information theory. *The European Physical Journal B: Condensed Matter and Complex Systems*, 88(10):257–269.
- Aquino, A. L. L., Ramos, H. S., Frery, A. C., Viana, L. P., Cavalcante, T. S. G., and Rosso, O. A. (2017). Characterization of electric load with information theory quantifiers. *Physica A*, 465:277–284.
- Bandt, C. (2017). A new kind of permutation entropy used to classify sleep stages from invisible EEG microstructure. *Entropy*, 19(5):197.
- Bandt, C. and Pompe, B. (2002). Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88:174102-1–174102-4.
- Cabral, R. S., Aquino, A. L. L., Frery, A. C., Rosso, O. A., and Ramírez, J. A. (2013). Structural changes in data communication in wireless sensor networks. *Central European Journal of Physics*, 11(12):1645–1652.

Referências (cont.)



- Carpi, L. C., Rosso, O. A., Saco, P. M., and Gómez Ravetti, M. (2011). Analyzing complex networks evolution through Information Theory quantifiers. *Physics Letters A*, 375:801–804.
- De Micco, L., González, C. M., Larondo, H. A., Martin, M. T., Plastino, A., and Rosso, O. A. (2008). Randomizing nonlinear maps via symbolic dynamics. *Physica A: Statistical Mechanics and its Applications*, 387(14):3373–3383.
- Gabriel, C., Wittmann, C., Sych, D., Dong, R., Mauerer, W., Andersen, U. L., Marquardt, C., and Leuchs, G. (2010). A generator for unique quantum random numbers based on vacuum states. *Nature Photonics*, 4(10):711–715.
- Knuth, D. E. (1998). *The Art of Computer Programming, Volume 2: (2Nd Ed.) Seminumerical Algorithms*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.

Referências (cont.)



- Larrondo, H. A., Martín, M. T., González, C. M., Plastino, A., and Rosso, O. A. (2006). Random number generators and causality. *Physics Letters A*, 352(4–5):421–425.
- L'Ecuyer, P. (1992). Testing Random Number Generators. In *Proceedings of the 1992 Winter Simulation Conference*, pages 305–313. [IEEE] Press.
- L'Ecuyer, P. and Simard, R. (2007). Testuo1: A c library for empirical testing of random number generators. *ACM Trans. Math. Softw.*, 33(4):22:1–22:40.
- Martin, M. T., Plastino, A., and Rosso, O. A. (2006). Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A*, 369:439–462.
- Matsumoto, M. and Nishimura, T. (1998). Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions Model. Comput. Simul.*, 8(1):3–30.
- Rosso, O. A., Martin, M. T., Figliola, A., Keller, K., and Plastino, A. (2006). EEG analysis using wavelet-based information tools. *Journal of Neuroscience Methods*, 153:163–182.

Marcelo Queiroz de Assis Oliveira | marcelo@lacca.ufal.br | marcelo@lacca.ufal.br | marcelo@lacca.ufal.br

Referências (cont.)



- Rosso, O. A., Ospina, R., and Frery, A. C. (2016). Classification and verification of handwritten signatures with time causal information theory quantifiers. *PLoS ONE*, 11(12):e0166868.
- Stamey, L. (2016). How RANDOM.ORG's journey from radio static to true randomness generates reliable results for games, security, and clinical trials.
- Walker, J. (2017). Ent. a pseudorandom number sequence test program.

Marcelo Queiroz de Assis Oliveira | marcelo@lacca.ufal.br | marcelo@lacca.ufal.br | marcelo@lacca.ufal.br

