Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Sobre a Detecção de Autocorrelações em Séries Temporais: Uma Comparação Objetiva entre Análise de Flutuações, Transformações Wavelet e Análise Entrópica

Adriana Camila Braga[1]

Departamento de Matemática, UTFPR, Apucarana, PR

Ademir Alves Ribeiro

Departamento de Matemática, UFPR, Curitiba, PR

Manoel Messias Alvino de Jesus

Departamento de Física, UTFPR, Apucarana, PR

Haroldo Valentin Ribeiro

Departamento de Física, UTFPR, Apucarana, PR

Departamento de Física, UEM, Maringá, PR

Resumo. Nesse trabalho revisitou-se três dos principais métodos existentes na literatura, que conseguem identificar autocorrelações de longo alcance. Nominalmente: a análise de flutuação DFA (detrend fluctuation analysis), as transformações wavelet e a análise entrópica DEA (diffusion entropy analysis). Fez-se uma comparação entre os três métodos, quanto a sua convergência para o verdadeiro valor do expoente h de Hurst, em função do tamanho das séries geradas. Nesta comparação, observou-se algumas peculiaridades de cada método, por exemplo, que o DFA converge por valores superiores de h, enquanto wavelet e DEA o fazem por valores inferiores. Com base nesse achado empírico, propôs aplicar simultaneamente DFA e wavelet. Isso fez com que a convergência para o valor verdadeiro de h fosse alcançada para séries razoavelmente pequenas.

Palavras-chave. Séries Temporais, Expoente de Hurst, Autocorrelações de Longo Alcance, Análise de Flutuações, Transformações Wavelet, Análise Entrópica.

1 Introdução

É cada vez maior o interesse em estudar os chamados sistemas complexos. A prova disso são os recentes avanços neste campo, materializados sobre uma crescente quantidade de artigos e livros publicados. Trata-se de um campo de pesquisa vasto e, em geral, multidisciplinar. Contudo, é interessante ressaltar que não existe uma definição muito precisa de um sistema complexo. O que se tem são propriedades gerais, as quais, muitas

[1]adrianacamila@utfpr.edu.br

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC

vezes, estão presentes nos sistemas complexos. De fato, na maioria das vezes, o estado da arte das investigações em sistemas complexos baseia-se em uma pequena quantidade de informação relacionada ao sistema. Essa quantidade limitada de informação torna ainda mais impressionante as implicações dos estudos em sistemas complexos. Nessas investigações, é comum o estudo de séries temporais. Séries temporais são compostas de quantidades observáveis de um sistema que podem ser ordenadas no tempo ou espaço. Em trabalhos com séries temporais, uma análise muito importante é a de autocorrelação. Medidas de autocorrelação indicam o quão correlacionados estão os termos de uma série temporal. Com medidas como essa, pode-se intuir muitos aspectos diretamente ligados à dinâmica do sistema sob investigação, como, por exemplo, se existe invariância por escala ou alguma outra propriedade fractal. Visando detectar autocorrelação, alguns métodos foram propostos. Entre eles, destacam-se a análise de flutuação DFA [3, 5], as transformações wavelet [1, 4] e a análise entrópica DEA [2]. Com este trabalho, pretende-se revisar aspectos relacionados a esses três métodos bem como compará-los de uma maneira objetiva.

2 Análise de flutuações e o método DFA

Um dos métodos mais utilizados para detectar autocorrelações em séries temporais é o chamado DFA, sigla inglesa para *detrended fluctuation analysis*. Ele foi proposto por Peng *et al.* [5] (veja também a Ref. [3]) para o estudo de autocorrelações em séries temporais construídas a partir do DNA. Como indica o nome, o método DFA analisa as flutuações de séries temporais removendo uma possível tendência local. Trata-se de um método de fácil implementação e que produz excelentes resultados mesmo para séries temporais de comprimento da ordem de mil termos. Para apresentar o método, considera-se uma série temporal x_i ($i \in \{1, 2, \ldots, n\}$). O primeiro passo do DFA consiste em obter a série integral de x_i, isto é,

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

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

$$\underbrace{\{y_1, y_2, \ldots, y_l\}}_{w_j^{(1,l)}}, \underbrace{\{y_{l+1}, y_{l+2}, \ldots, y_{2l}\}}_{w_j^{(2,l)}}, \ldots, \underbrace{\{y_{(s-1)l+1}, y_{(s-1)l+2}, \ldots, y_{sl}\}}_{w_j^{(s,l)}}, \tag{2}$$

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

$$\chi_2^{(i,l)} = \frac{1}{l-1} \sum_{j=1}^{l} [w_j^{(i,l)} - f_v(j)]^2 \,, \tag{3}$$

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC

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

$$F(l) = \left(\frac{1}{s} \sum_{i=1}^{s} \chi_2^{(i,l)} \right)^{1/2} . \tag{4}$$

Naturalmente, esta flutuação média será uma função de l, que está diretamente relacionada com o expoente de escala δ da seguinte maneira:

$$F(l) \sim l^{\delta} . \tag{5}$$

Para o caso em que a função escala $\Psi(x)$ possui o segundo momento finito, tem-se a igualdade $\delta = h$. Nos casos em que $\Psi(x)$ não possui o segundo momento, o DFA pode conduzir a falsas autocorrelações. Na prática, costuma-se aplicar o método na série embaralhada de maneira aleatória para verificar a validade da igualdade $\delta = h$. Se $h \neq 0.5$ for obtido para a série embaralhada, nada podemos dizer sobre o expoente h da série não embaralhada. Por outro lado, se $h \approx 0.5$ para a série embaralhada, o expoente h da série original será verdadeiramente o expoente de Hurst para aquela série.

A aplicação da equação (5) é bastante simples. Basta calcular essa função de flutuação para um conjunto de valores de l e construir um gráfico log-log. A inclinação dessa reta será numericamente igual ao expoente δ. Observe que a grande engenhosidade desse método é o fato dele construir diversas trajetórias aproximadamente independentes a partir de uma única série temporal. Isto faz com que, mesmo para séries moderadamente pequenas, os erros envolvendo a determinação do expoente h sejam pequenos.

3 Transformações wavelet

A literatura de análise de séries temporais parece estar dividida (essencialmente) em duas vertentes concorrentes no que diz respeito à análise de autocorrelações. Como dito, DFA é um dos métodos mais utilizados para determinar o expoente de Hurst e talvez o seu maior "concorrente" seja as transformações wavelet [1, 4]. As transformações wavelet são parametrizadas por um parâmetro de escala $s > 0$ e por um parâmetro de translação u. Além disso, deve-se escolher uma função do tipo

$$\psi_{s,u}(x) = \psi \left(\frac{x-u}{s} \right) , \tag{6}$$

a qual é denominada de função mãe ou *analyzing wavelet*. Esta função deve ter média nula (quando $u = 0$), ser localizada no espaço usual e também no espaço de Fourier. Assim, dada uma função $g(x)$, define-se sua transformada wavelet como

$$\mathcal{W}[g(x)](s,u) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} g(x)\ \psi \left(\frac{x-u}{s} \right) dx , \tag{7}$$

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC

Para o caso de um conjunto discreto, isto é, uma série temporal x_i ($i \in \{1, 2, \ldots, n\}$), a definição acima toma a seguinte forma

$$\mathcal{W}[x_i](s,u) = \frac{1}{\sqrt{s}} \sum_{i=1}^{n} \psi^* \left(\frac{i-u}{s} \right) x_i \,. \tag{8}$$

em que $\psi^*(x)$ denota o complexo conjugado de $\psi(x)$. No caso de uma função $g(x)$ invariante por escala, isto é, $g(x) \doteq \lambda^{-h} g(\lambda x)$, em que o símbolo $\doteq$ representa uma igualdade estatística, ou seja, ambos os lados possuem a mesma distribuição (função escala), deve-se obter

$$\mathcal{W}[g(x)](\lambda s, \lambda u) \doteq \lambda^{(1/2)+h} \, \mathcal{W}[g(x)](s,u) \,. \tag{9}$$

Como deve estar claro, o domínio de uma transformação wavelet de uma função unidimensional é bidimensional: uma dimensão correspondente ao parâmetro de escala s e outra dimensão correspondente ao parâmetro de translação u. Assim, fixando-se o parâmetro s, tem-se um número infinito de amplitudes relacionadas ao parâmetro de translação u. Uma maneira de contornar essa dupla dependência é tomar o valor médio sobre o parâmetro de translação u para cada valor de escala s. Essa manipulação conduz a

$$\mathcal{W}[g(x)](s) = \langle |W[g(x)](s,u)| \rangle_u \,, \tag{10}$$

em que $\langle \ldots \rangle_u$ denota o valor médio com respeito à variável u. Outros tipos de média podem ser considerados, como por exemplo, a média harmônica. A inclusão do valor absoluto também é opcional. Porém, seu uso é bastante comum e acredita-se que resultados numéricos melhores são obtidos usando essa forma. Após considerar esse valor médio, a relação de escala (9) leva a

$$\mathcal{W}[g(x)](\lambda s) \doteq \lambda^{1/2+h} \, \mathcal{W}[g(x)](s) \,. \tag{11}$$

Assim, em um gráfico log-log do valor médio $\mathcal{W}[g(x)](s)$ versus s, tem-se uma reta cuja inclinação é numericamente igual a $1/2 + h$. Quando lida-se com séries temporais, a expressão (10) pode ser escrita como

$$\mathcal{W}(s) = \frac{1}{n} \sum_{u=1}^{n} \left[\frac{1}{\sqrt{s}} \sum_{i=1}^{n} \psi^* \left(\frac{i-u}{s} \right) x_i \right] , \tag{12}$$

sendo que omitiu-se a notação de funcional por questão de simplicidade. Resta escolher qual função mãe $\psi_{s,u}(x)$ a usar. Algumas são bastante comuns na literatura, como é o caso da derivada de ordem $\beta \in \mathbb{N}$ da gaussiana

$$\psi_{s,u}(x) = \frac{(-1)^{\beta+1}}{\sqrt{\Gamma(\beta + 1/2)}} \frac{d^\beta}{dx^\beta} \exp(-x^2/2) \tag{13}$$

Usualmente a escolha da função mãe pode influenciar na análise wavelet, entretanto, na estimativa do expoente h, observa-se que o uso das funções do tipo (13) ou também da wavelet de Paul conduzem a resultados similares. Os resultados apresentados aqui,

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC

emprega-se a expressão (13). É interessante observar que o DFA utiliza a série integrada para calcular a função de flutuação, enquanto as transformações wavelet usam a série original. Assim, para uma comparação correta entre os dois métodos é necessário o uso da série integrada do movimento browniano fracionário. Vale notar também, que se por um lado o DFA gera muitas trajetórias aproximadamente independentes, as transformações wavelet levam uma série temporal para um espaço bidimensional. Dessa maneira, bons resultados podem ser obtidos para séries razoavelmente pequenas.

4 Análise entrópica e o método DEA

Outro método para detecção de autocorrelações em séries temporais é o chamado DEA, sigla inglesa para *diffusion entropy analysis* [2]. Sua implementação também é bastante simples e, como sugere o nome, este método está baseado no cálculo da entropia de trajetórias criadas a partir da série temporal. Analogamente aos outros dois métodos, o DEA está baseado no cálculo do expoente de escala δ. Para apresentar o algoritmo, considera-se a série temporal z_i ($i \in \{1, 2, \ldots, n\}$). Usando essa série, construa as seguintes trajetórias

$$x^{(s)}(t) = \sum_{i=1}^{t} x_{i+s} \qquad \forall t \in \{1, 2, \ldots, n\}\,, \tag{14}$$

em que $s \in \{0, 1, \ldots, n-t\}$. Note que essas trajetórias nada mais são do que a série original integrada de diferentes pontos iniciais. Dessa maneira, para cada t, tem-se um conjunto de $n - t + 1$ trajetórias denotadas por $x^{(s)}(t)$. Pode-se imaginar que essas trajetórias representam partículas em um movimento difusivo, para o qual pode-se definir a probabilidade $\rho(x, t)$ de encontrar uma partícula localizada entre x e $x + dx$, em um intervalo de tempo entre t e $t + dt$. Usando essa distribuição, é possível calcular a entropia do processo difusivo

$$S(t) = -\int_{-\infty}^{\infty} \rho(x, t) \ln[\rho(x, t)] dx\,. \tag{15}$$

Nos casos em que existir invariância de escala, isto é, $\rho(x, t) = t^{-\delta}\, \Psi(x\, t^{-\delta})$, tem-se

$$S(t) \;\; = \;\; -\int_{-\infty}^{\infty} \Psi(y) \ln[\Psi(y)] dy + \delta \ln(t)\,. \tag{16}$$

Note que fez-se a mudança de variável $y = x\, t^{-\delta}$ e considerou-se $\int_{-\infty}^{\infty} \Psi(y) dy = 1$. Pode-se escrever ainda

$$S(t) = A + \delta \ln(t)\,, \tag{17}$$

visto que o primeiro termo em (16) não depende de t. Assim, em um gráfico de $S(t)$ versus $\ln(t)$, deve-se obter uma reta cuja inclinação é numericamente igual ao expoente δ, o qual coincide com o expoente h de Hurst quando a função escala $\Psi(x)$ possuir o segundo momento finito. Naturalmente, o cálculo da distribuição de probalidade $\rho(x, t)$ deve ser feito utilizando histogramas. Uma possível implementação consiste em construir M partições de tamanho ϵ no espaço das trajetórias $x^{(s)}(t)$ e contar o número de partículas

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC

existentes em cada partição para um dado t. Denotando esse número por $N_j(t)$, pode-se definir a versão discreta de $\rho(x,t)$ como

$$p_j(t) = \frac{N_j(t)}{n-t+1}\,, \tag{18}$$

sendo que j representa o índice da partição. Além disso, deve-se utilizar a versão discreta da fórmula (15), isto é,

$$S(t) = -\sum_{j=1}^{M} p_j(t)\ln[p_j(t)]\,. \tag{19}$$

Empiricamente, uma maneira eficiente de escolher o tamanho ϵ das partições, é assumí-lo independente de t e determiná-lo por uma fração conveniente do desvio padrão da série z_i. Aplicou-se esse método para a série dos incrementos do movimento browniano fracionário. Novamente, bons resultados são obtidos para séries moderadamente pequenas.

5 Comparando os métodos

Nesta seção, apresenta-se uma comparação objetiva entre os três métodos apresentados anteriormente. Mais especificamente, investiga-se como é a convergência desses métodos com relação ao tamanho (número de termos) das séries. Para isso, gerou-se séries fractais usando o movimento browniano fracionário com diferentes valores de h e diferentes tamanhos ($n \sim 10$ até $n \sim 10^5$). Aplicou-se os três métodos em cada uma das séries, obtendo os valores do expoente h para cada método. Uma média sobre 2000 realizações desse procedimento foi levada em conta, visando obter o valor médio de h fornecido por cada método, para cada tamanho de série. Esses resultados são mostrados na Figura 1. Nesta figura, observou-se que para valores pequenos de n, todos os métodos conduzem a valores médios distantes dos verdadeiros valores de h. Naturalmente, ao aumentar o tamanho das séries, todos os métodos convergem para o valor teórico de h. Contudo, pode-se ver que o DFA converge por valores superiores ao verdadeiro valor de h, enquanto as transformações wavelet e o DEA, o fazem por valores inferiores. No que diz respeito à especificidade de cada método notou-se que: *i*) Para valores de $h < 0.3$ o DEA apresenta uma convergência superior aos demais. Porém, para valores de $h > 0.5$ sua convergência é consideravelmente mais lenta; *ii*) Comparado com as transformações wavelet, o método DFA apresenta uma convergência ligeiramente mais rápida para séries persistentes ($h > 0.5$) e ligeiramente mais lenta para séries anti-persistentes ($h < 0.5$); *iii*) Para séries com mais de 10000 termos os três métodos produzem valores de h que praticamente coincidem com os verdadeiros valores. A simetria (aproximada) existente entre as transformações wavelet e DFA, em relação ao valor de h da série, sugere que os métodos podem conduzir a melhores resultados se forem aplicados simultaneamente. Para testar essa ideia, aplicou-se DFA e wavelet em uma mesma série e calculou-se o valor médio do expoente h obtido para cada método. Na Figura 1, mostrou-se também como se dá a convergência para o verdadeiro valor de h, utilizando essa combinação de métodos. Veja que há uma considerável melhora na convergência, sendo que mesmo para séries pequenas ($n \sim 100$) o valor obtido por essa combinação é muito próximo ao valor verdadeiro de h.

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC

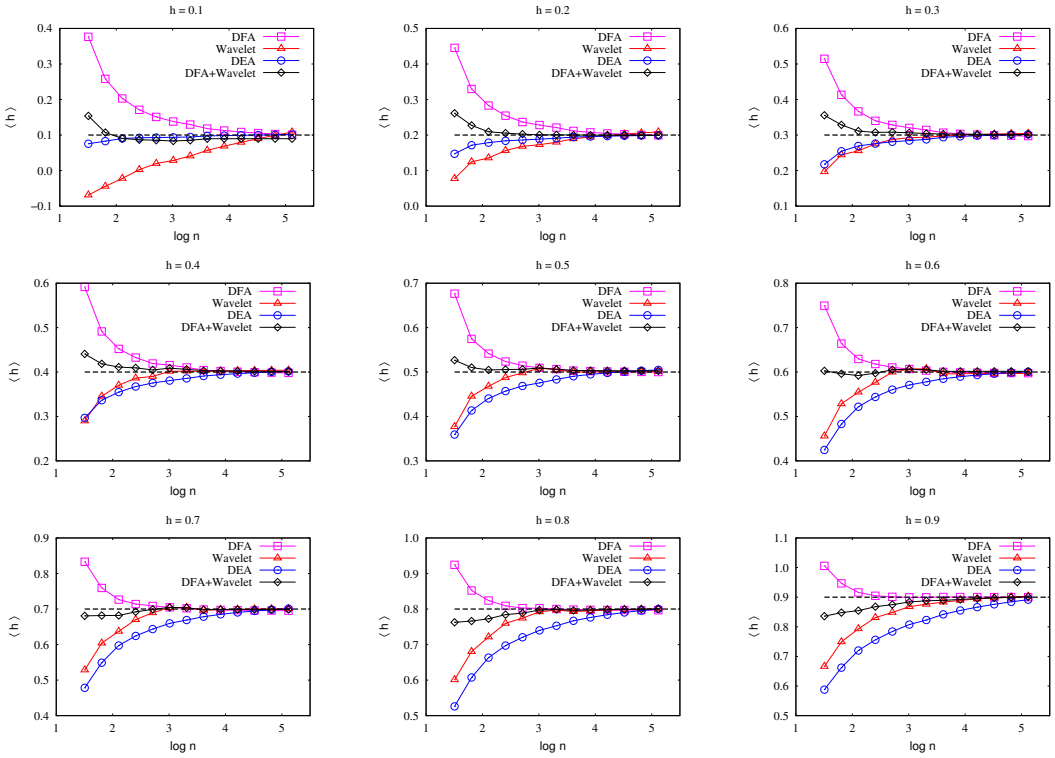

Figura 1: Valor médio do expoente h calculado para diferentes valores de n, usando DFA, wavelet, DEA e combinação média entre DFA e wavelet. A linha tracejada indica o valor verdadeiro de h, isto é, o valor de h usado para gerar as trajetórias do movimento browniano fracionário.

Referências

[1] A. Arneodo, E. Bracry, P. V. Graves and J. F. Muzy, Charactering long-range correlations in DNA sequences from wavelet analysis, Phys. Rev. Lett., vol. 74, 3293-3296, (1995).

[2] P. Grigolini, L. Palatella and G. Raffaelli, Asymmetric anomalous diffusion: An efficient way to detect memory in time series, Fractals, vol. 9, 439-449, (2001).

[3] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin and A. Bunde, Detecting long-range correlations with detrended fluctuation analysis, Physica A, vol. 295, 441-454, (2001).

[4] P. Manimaran, P. K. Panigrahi and J. C. Parikh, Wavelet analysis and scaling properties of time series, Phys. Rev. E, vol. 72, 046120, (2005).

[5] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E, vol. 49, 1685-1689, (1994).

DOI: 10.5540/03.2015.003.02.0058 © 2015 SBMAC