Lista de exercícios Matemática Computacional
Parte B – Prof. Dr. Reinaldo Rosa - 2020

Denis M. A. Eiras

Exercício 6 - Descrição

6.1. Considere as séries temporais listadas na tabela dataset_signal e obtenha, para cada série, os valores respectivos do seguintes atributos: S^2 , K, β (via PSD) e α (via DFA). Confira para todas as series se β (via PSD) está bem ajustado a partir da formula WKP: $\beta = 2\,\alpha - 1$. Construa dois espaços de parâmetros EPSB-K-means: S^2 x K x β e EDF-K-means: S^2 x K x α .

6.2. Classifique, nos espaços de parâmetros do exercício anterior, as séries temporais: (a) ST-Sol3GHz, (b) ST-surftemp504 e (c) ST-OWS_NDC_C

6.3. Aplique k-means para todas as series ST-OWS_NDC_Covid1 considerando os seguintes Espaços de atributos: S^2 x α e K x α. Obtenha os melhores agrupamentos, identifique os grupos e discuta os resultados.

Exercício 6.1 – Detalhes da implementação

Foi criada a função calcula_df_estatistico_por_familia_e_sinal, para gerar arquivos csv das tabelas estatísticas dos espaços de parâmetros requisitados do enunciado.

As seguintes funções de outros exercícios foram utilizadas para gerar os dados da tabela data_set_sinal, reaproveitando assim toda a lógica:

Exercicio1.exercicio1_1 - gerador_de_sinais_aleatorios
Exercicio2.exercicio2 - gerador_de_sinais_colored_noise
Exercicio3.exercicio3 - gerador_de_sinais_pmodel
Exercicio5.exercicio5_1 - gerador_de_sinais_logisticos, gerador_de_sinais_henon

Exercício 6.2 – Detalhes da implementação

Para a leitura de arquivos gerada no exercício anterior e a leitura das séries Covid, Sol e SurfTemp, foi reutilizado o leitor genérico implementado no exercício 4.2. Para calcular as estatísticas, foi reaproveitada uma função do exercício 6.1. Para realizar o agrupamento das séries dentro do espaço de parâmetros, foi utilizado o k-means, do exercício 1.3. Em resumo, foram reaproveitados:

Exercicio4.exercicio4_2_2 - ler_serie_generica_de_arquivo_ou_url
Exercicio6.exercicio6_1 – calcula_df_estatistico
Exercicio1.exercicio1_3 - k_means_e_metodo_do_cotovelo

Exercício 6.3 – Detalhes da implementação

O programa implementado cria séries estatísticas à partir de uma coluna agrupadora e uma coluna de valores da série, configuráveis no programa. Também é possível configurar valores da coluna agrupadora a serem removidos.

Exercício 6.1 – Análise

Para verificar se β (via PSD) está bem ajustado a partir da formula WKP: β = 2 α – 1, foram gerados gráficos contendo os pontos, em azul, no plano α x β, para cada série, como mostram os gráficos da figura 1. Em seguida, foi gerada uma reta que interpola esses pontos, em azul claro, e uma outra reta gerada calculada a partir da equação β = 2 α – 1, em rosa.

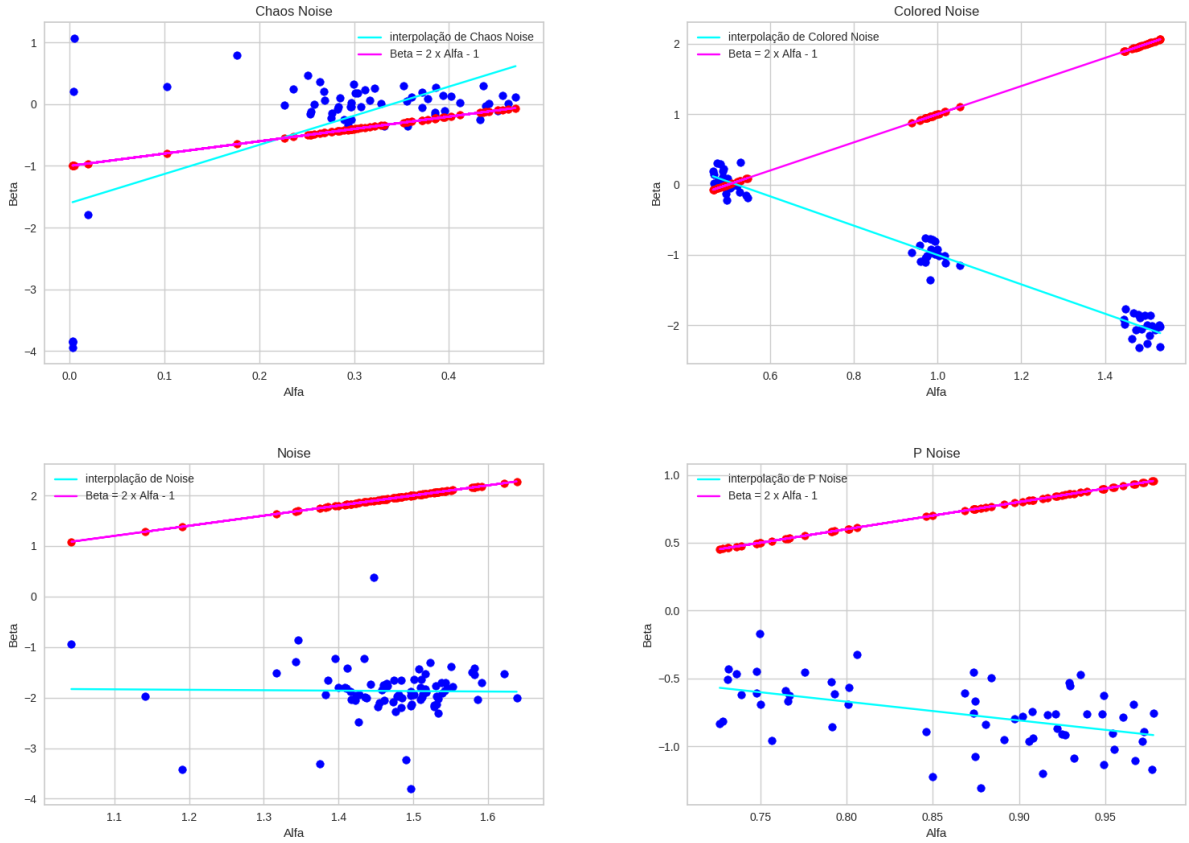

Figura 1. Retas interpoladoras dos pontos, em azul, e retas calculadas À partir da equação β = 2 α – 1, em rosa.

Observando os gráficos, a reta Chaos é a que mais se aproxima da reta calculada pela equação, o que só ocorreu devido à alguns pontos mais fora da reta fora do agrupamento.

Os pontos Colored Noise possuem uma inclinação de reta que parece ser a negativa da equação, isto é, β = -2 α + 1.

Para se ajustar ao sinal Noise, a equação precisaria de um descréscimo de aproximadamente mais 3 pontos em y, e algum ajuste na inclinação.

O sinal P Noise também requer um ajuste em y e na inclinação da reta.

Exercício 6.1 - Conclusão

Observando os gráficos, a reta Chaos é a que mais se aproxima da reta calculada pela equação.

Exercício 6.2 – Análise

Para verificar em qual espaço de parâmetros, ESPB ou EDF, onde cada uma das séries poderia estar contida, foram gerados dois datasets para cada uma das séries Covid, Sol e Surf_Temp: Um

dataset contém o espaço ESPB e espaço (um ponto) de uma série, e outro dataset contendo o espaço EDF e o espaço (um ponto) de uma série. Isto é:

- Espaço ESPB (um ponto para cada sinal) + Espaço Covid (um ponto)
- Espaço ESPB (um ponto para cada sinal) + Espaço Sol (um ponto)
- Espaço ESPB (um ponto para cada sinal) + Espaço Surf_Temp. (um ponto)
- Espaço EDF (um ponto para cada sinal) + Espaço Covid (um ponto)
- Espaço EDF (um ponto para cada sinal) + Espaço Sol (um ponto)
- Espaço EDF (um ponto para cada sinal) + Espaço Surf_Temp. (um ponto)

Para verificar se as séries poderiam estar em agrupamentos dos espaços ESPB e EDF, o K-means foi executado entre 2 e 14, onde o melhor k foi encontrada através do método do cotovelo, e então, localizadas as séries dentro do melhor k.

Série COVID-19

As estatísticas de novos casos foram geradas para o país Estados Unidos, entre 10/03/202 e 28/05/2020.

Estatísticas da série da Covid:

S^2	curtose	β	α
0,34	-0,11	-0,64	0,92

ESPB – Melhor k = 5 – Ponto da série na classe 3
EDF – Melhor k = 6 – Ponto da série na classe 1

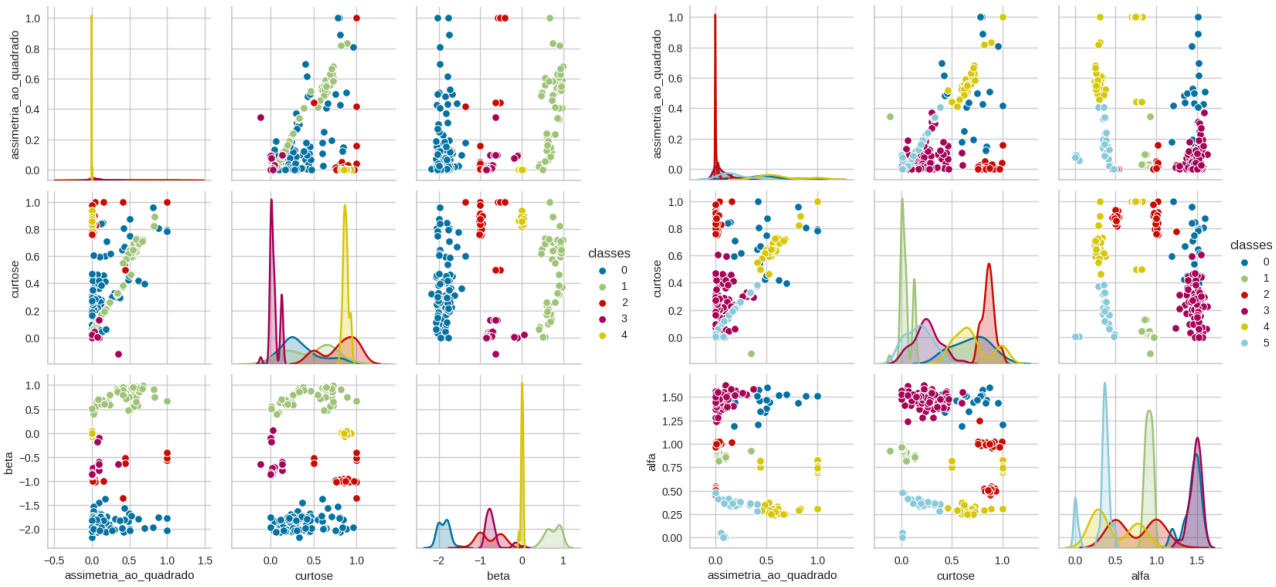

Figura 2. Ponto da série Covid no espaço ESPB, à esquerda, e no espaço EDF, à direita

Série Sol 3ghz

Estatísticas da serie sol:

S^2	curtose	β	α
2,66	3,80	-1.87	1,36

ESPB – Melhor k = 6 – Ponto da série na classe 3
EDF – Melhor k = 6 – Ponto da série na classe 4

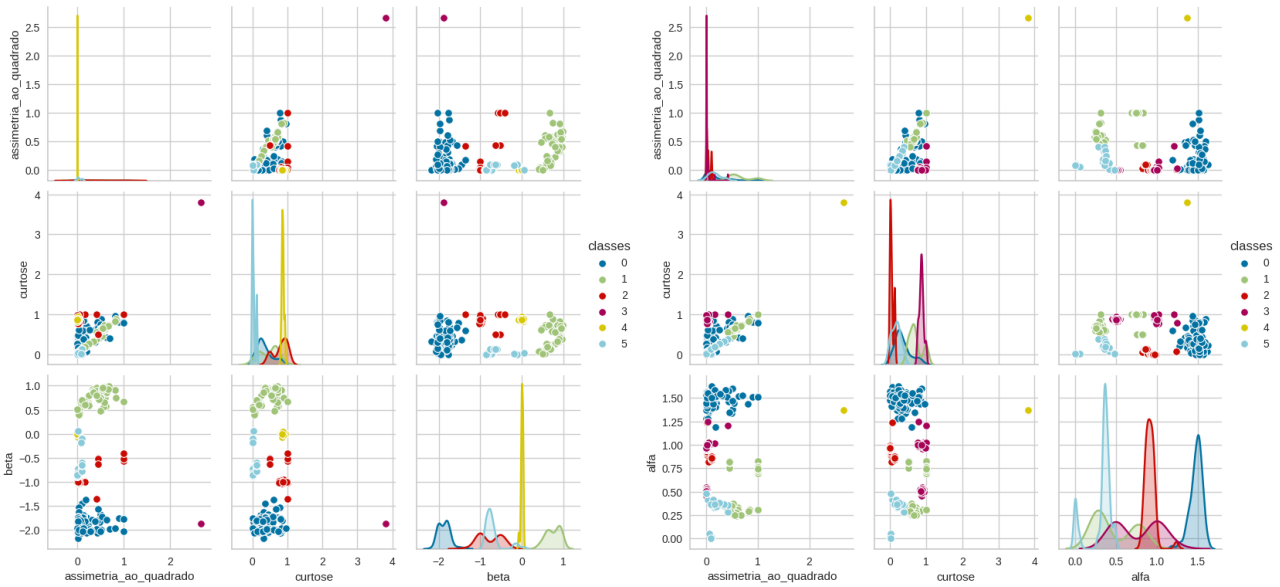

Figura 3. Ponto da série Sol no espaço ESPB, à esquerda, e no espaço EDF, à direita

<u>Série Surface Temperature</u>

Estatísticas da serie surf_temp:

S^2	curtose	β	α
0,26	0,39	-1.09	0,74

ESPB – Melhor k = 5 – Ponto da série na classe 2
EDF – Melhor k = 6 – Ponto da série na classe 1

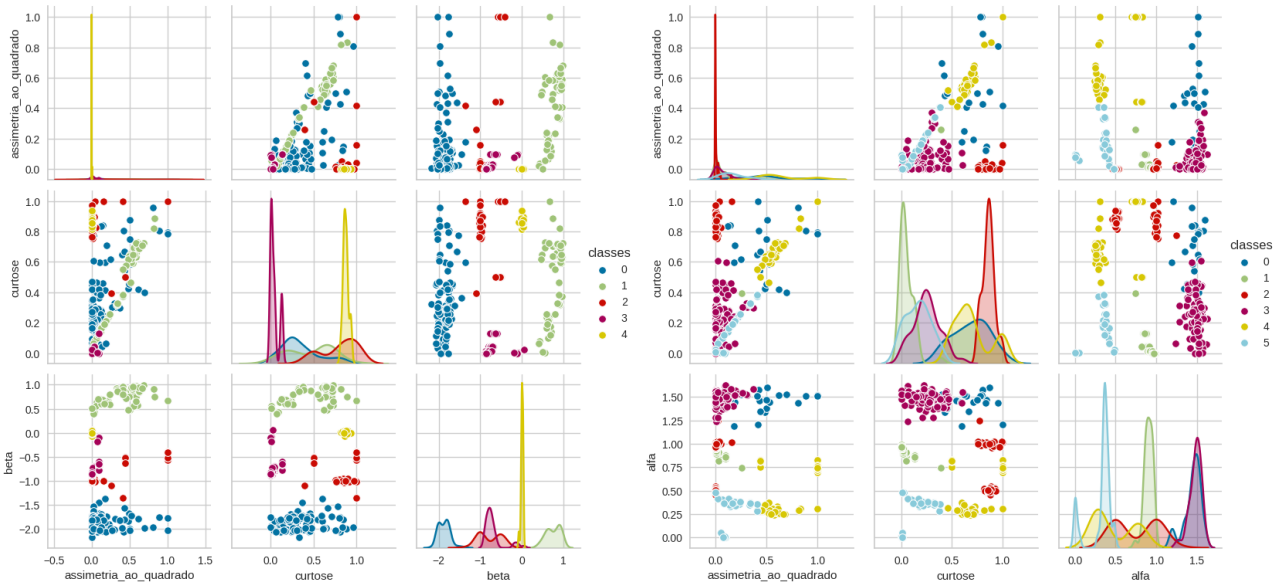

Figura 4. Ponto da série Surface Temperature no espaço ESPB, à esquerda, e no espaço EDF, à direita

Exercício 6.2 – Conclusão

As séries COVID e Surface Temperature foram classificadas dentro dos espaços ESPB e EDF. A série Sol ficou agrupada em um espaço isolado, fora dos dois espaços, devido à assimetria e à curtose, ambas elevadas.

Exercício 6.3 – Análise

As séries estatísticas sobre o número de casos diários da COVID-19 ‘new_cases’ foram criadas utilizando o agrupador ‘location’, isto é, um agrupamento por país. Foram removidos os valores perntencentes à localização ‘World’, que representa o total de todos os países, ‘International’ e valores com iso_code vazios.

Devido àlta variabilidade da assimetria, poucos grupos foram criados. Por esse motivo, a assimetria e curtose foi normalizada.

O K-means foi executado para os espaços estatístico $S^2_x\alpha$ e $K_x\alpha$, e para ambos os casos, o melhor k = 6, como pode ser observado na figura 5.

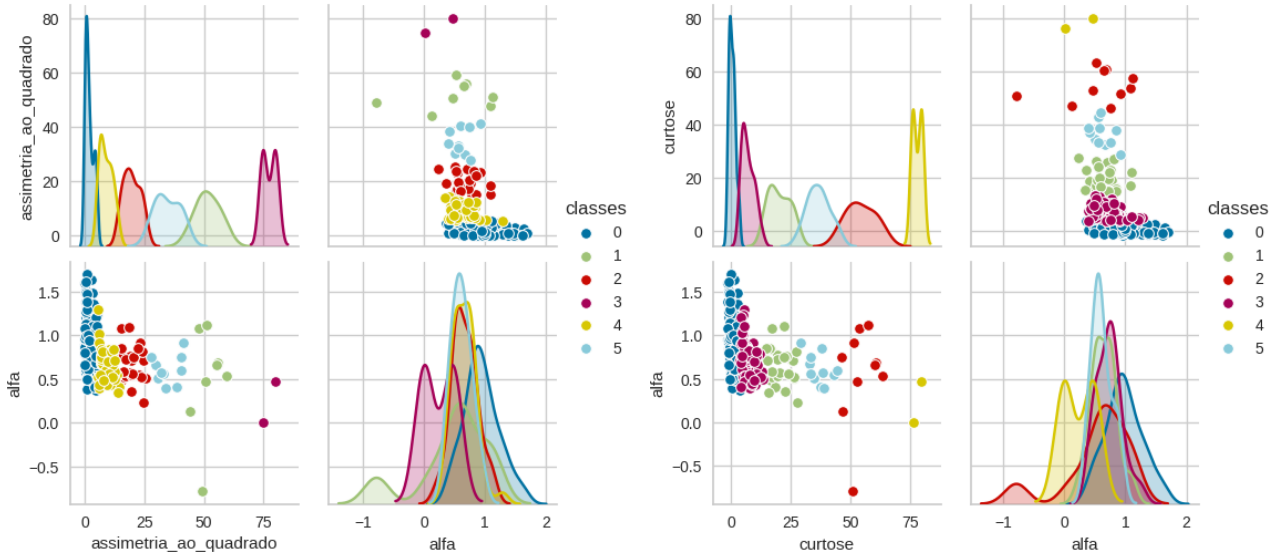

Figura 5. Agrupamentos K-means. a) $S^2_x\alpha$; b) $K_x\alpha$.

Foi implementada uma funcionalidade onde se faz a previsão da classe de cada país nos espaços de parâmetros, de onde foi construída a tabela 1 para o espaço S^2 e a tabela 2, no espaço K.

Tabela 1. Alguns países no espaço S^2.

Classe 0	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
United States	Comoros	Aruba	Zimbabwe	Australia	Cambodia
Brazil	Faeroe Islands	Belize	Western Sahara	China	Ecuador
Germany	Mongolia	FrenchPolynesia		Japan	Palestine
United Kindgon	Nicaragua	Paraguay		Syria	Sao Tome and Principe
Italy	Papua New Guinea	Uganda		Venezuela	Tanzania
Jamaica	Timor	Greenland		Haiti	Vatican

Tabela 1. Alguns países no espaço K.

Classe 0	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
United States	Aruba	Comoros	Australia	Western Sahara	Cambodia
Brazil	Belize	Faeroe Islands	Japan	Yemen	Sao Tome and Principe
Germany	FrenchPolynesia	Mongolia	Jamaica		Tanzania
United Kindgon	Paraguay	Nicaragua	Haiti		Uganda

Italy	Venezuela	Papua New Guinea	Syria		Vatican
Kenya	Greenland	Trinidad and Tobago	Niger		Benin

Exercício 6.3 - Conclusão

Um baixo valore de alfa, é um bom indicador de países emergentes, pois a curva de casos ainda está em ascenção. O contrário nem sempre é verdadeiro, conforme observado nos valores.

O agrupamento por S^2 contribui na identificação de países em ascenção do pico, como no caso do Paraguai.

A curtose pode ajudar a identificar os países que apresentam uma estabilidade no número de casos diários, quando o valor é baixo, ou picos em alguns países, quando esse número é alto. Esse número também pode estar relacionado à quantidade de testes diários, pois o número de casos tende a ser mais estável em países que fazem mais testes.

Os agrupamentos nos espaços ficaram bastante parecidos, principalmente a classe 0, que contém menores valores de assimetria e curtose, pois são países que em sua maioria já passaram pelo pico da pandemia.

Devido a alterações na metodologia de coleta dos dados, alguns países têm registros de valores negativos, o que leva à erros na classificação, como é o caso da Uganda. Devido à alterações nas metodogias e falta de consistência dos dados, um estudo mais aprofundado deveria ser feito, com maiores validações e ajustes nos dados.