Mackleyn Andrade Santos Lira de Vasconcelos

PROPRIEDADES COMPUTACIONAIS E FÍSICAS DE GRANDES CONJUNTOS DE DADOS CARDÍACOS HUMANOS

João Pessoa - Paraíba - Brasil

2023, Maio

Mackleyn Andrade Santos Lira de Vasconcelos

PROPRIEDADES COMPUTACIONAIS E FÍSICAS DE GRANDES CONJUNTOS DE DADOS CARDÍACOS HUMANOS

Trabalho de Conclusão de Curso realizado sob a orientação do Prof. Dr. Jorge Gabriel Gomes de Souza Ramos, apresentado à Coordenação de Graduação em Física da Universidade Federal da Paraíba, em complementação aos requisitos para finalização da grade curricular da Graduação em Física Bacharelado.

Universidade Federal da Paraíba – UFPB

Centro de Ciências Exatas e da Natureza

Departamento de Física

Orientador: Prof. Doutor Jorge Gabriel Gomes de Souza Ramos

João Pessoa - Paraíba - Brasil

2023, Maio

Catalogação na publicação
Seção de Catalogação e Classificação

V331p Vasconcelos, Mackleyn Andrade Santos Lira de.
Propriedades computacionais e físicas de grandes conjuntos de dados cardíacos humanos / Mackleyn Andrade Santos Lira de Vasconcelos. - João Pessoa, 2023.
74 p. : il.

Orientação: Jorge Gabriel Gomes de Souza Ramos.
TCC (Curso de Bacharelado em Física) - UFPB/CCEN.

1. Teoria do Caos. 2. Dados Cardíacos. 3. Wavelet.
I. Ramos, Jorge Gabriel Gomes de Souza. II. Título.

UFPB/CCEN CDU 53(043.2)

Elaborado por Josélia Maria Oliviera da Silva - CRB-15/113

Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Coordenação dos Cursos de Graduação em Física

Ata da Sessão Pública da Defesa do Trabalho de Conclusão de Curso de Bacharelado em Física, do discente Mackleyn Andrade Santos Lira de Vasconcelos.

Aos 26 dias do mês de maio do ano de 2023, às 14h, na sala 202 - Departamento de Física/CCEN/UFPB, realizou-se a Sessão Pública da Defesa do Trabalho de Conclusão de Curso de Bacharelado em Física, do discente Mackleyn Andrade Santos Lira de Vasconcelos, sendo a Banca Examinadora constituída pelos docentes Prof. Dr. Jorge Gabriel Gomes de Souza Ramos (UFPB), orientador e presidente da banca, Prof. Dr. Paulo Sérgio Rodrigues da Silva (UFPB), Prof. Dr. Abel Cavalcante Lima Filho (UFPB) e Prof. Dr. Herondy Francisco Santana Mota (UFPB - suplente). Dando início aos trabalhos, o professor orientador e presidente da banca examinadora comunicou aos presentes a finalidade da reunião. A seguir, concedeu a palavra ao discente para que fizesse a explanação de seu Trabalho de Conclusão de Curso, intitulado "*Propriedades Computacionais e Físicas de Grandes Conjuntos de Dados Cardíacos Humanos*". Concluída a exposição, o discente foi arguido pelos membros presentes da Banca Examinadora. Após as arguições, a Banca, de comum acordo, declarou que o Trabalho apresentado foi aprovado com nota 10,0. E para constar, encerrada a sessão, lavrou-se esta ata que será assinada pelos presentes. João Pessoa, 26 de maio de 2023.

Prof. Dr. Jorge Gabriel Gomes de Souza Ramos
UFPB – Orientador

Prof. Dr. Paulo Sérgio Rodrigues da Silva
UFPB

Prof. Dr. Abel Cavalcante Lima Filho
UFPB

Prof. Dr. Herondy Francisco Santana Mota
UFPB - suplente

Mackleyn Andrade Santos Lira de Vasconcelos
Discente

Este trabalho é dedicado à pessoas como você,
que subiram o longo e fino pelo do coelho
e agora admiram o truque do mágico com seus próprios olhos.

Agradecimentos

Primeiramente agradeço à minha mãe, Carmenlúcia Santos, e ao meu falecido pai, José Lira de Vasconcelos, por terem desde cedo incentivado minha sede por conhecimento, sempre me auxiliando a alçar meus próprios voos graças às minhas próprias perguntas, então nada mais justo de que fique registrado neste meu agradecimento a realidade mais pura que resultou desta dedicação parental plena: **Enquanto eu viver, em cada mente que as asas do meu conhecimento tocarem, eu levarei um pedaço de vocês para a eternidade.**

Aos meus amigos e colegas da turma 2017.1 do curso de bacharelado em Física, com quem compartilhamos discussões, estudos e sem sombra de dúvidas momentos alegres e tristes juntos, graças a vocês que mesmo nos meus momentos mais pessimistas este curso se tornou possível de ser finalizado, meu muito obrigado.

Ao meu orientador Prof. Dr. Jorge Gabriel Gomes de Souza Ramos, que me ensinou sua didática, incitou maravilhosas e calorosas discussões que fomentassem meu conhecimento e me guiou para que este curso fosse mais leve e proveitoso, sua compreensão foi um dos pilares que me fizeram retomar e superar a síndrome de impostor que me afligia.

Aos meus amigos mais próximos que formei por toda vida, foram graças às nossas discussões de temas de diversos gêneros (incluindo o assunto deste TCC) que me fizeram ter noção da importância de que a linguagem com a qual eu me comunicasse nesta monografia teria que ser acessível ao maior público possível.

E por último e não menos importante, agradeço à minha psicóloga, Joana D'Arc Urbano, que me ajudou a recuperar minha autoconfiança quando nem eu estava mais acreditando em mim, suas palavras foram o impulso que eu precisava para que eu mesmo construísse o principal pilar da minha mente: O da força de vontade.

Agora realmente por fim, pois não poderia deixar de agradecer a você que está lendo até aqui, espero que sua leitura que virá a seguir te ajude tanto quanto todas estas pessoas acima me ajudaram. Obrigado por me dar a oportunidade de que você me conheça num dos âmagos mais profundos de meu ser: A didática.

Que você tenha uma ótima leitura.

O importante é viver bem, não viver por muito tempo;
e muitas vezes vive bem quem não vive muito.
(Lúcio Aneu ***Sêneca****)*

Resumo

Desde o início da humanidade, quando nossos ancestrais hominínios aprenderam a criar lanças e a manipular o fogo, nós temos que lidar com o fluxo e a interpretação de dados, assim como em diversas outras áreas mais recentes, tais como a análise do clima na meteorologia, ou com diagnóstico de doenças cardíacas com as avaliações de eletrocardiogramas (ECG´s) e por conta de tamanha checagem conseguimos compreender ainda mais o funcionamento da natureza ao nosso redor e consequentemente aumentamos nossa própria longevidade, simplesmente por realizar uma checagem de dados eficiente.
Entretanto, a análise de tais ECG's só costumam considerar os dados principais: Os picos e os espaçamentos entre os valores. Desconfiando que os dados menores ignorados (antes considerados apenas como ruído) pudessem de fato dizer algo além do que víamos (um caos oculto), nós iniciamos este projeto e estudamos desde o básico da estatística (média, variância e correlação) até técnicas avançadas de filtragem como *Wavelets* e *Bézier* (este último abandonado por ser ineficiente para os nossos objetivos), então fizemos a aplicação delas em séries temporais de sinais biológicos do MIT utilizando a ferramenta MATLAB para podermos filtrar (2.8) os ruídos e analisarmos os dados outrora desprezados.
Por fim, nós obtivemos a filtragem prevista teoricamente para o filtro Wavelet (2.8) ao aplicarmos nos sinais biológicos da *MIT-BIH Arrhythmia Database*, com picos bem definidos e pouco ruidosos em todos os níveis detalhados (seção 3.4), além de obtermos resultados que mostram a proximidade entre os dados computacionalmente obtidos e a previsão de contagem de picos da seção (3.3). Sendo assim, nossa expectativa a partir daqui é continuar avançando para a criação de um método de detecção analítica para qualquer sistema biológico no qual se poderia gerar números universais de fácil diagnóstico para prevenção de doenças (não apenas cardíacas) a partir de dados outrora ignorados, que é todo o objetivo inicial.

Palavras-chave: Correlação. Dados. Picos. Sinais biológicos. *Wavelet*.

Abstract

Since the beginning of humanity, when our hominines ancestors have learned how to create spears and manipulate fire, we have to deal with data flow and interpretation of it, as well as in several other areas, such as climate analysis in Meteorology, or with heart disease diagnosis with the evaluations of eletrocardiograms (ECG's) and because of so many data checking we become able to comprehend even more the behavior of nature around us and consequently we raised our own longevity, simply for performing an efficient data checking. However, in such ECG's analysis only are considered the principal data: The peaks and the spacing between values. Suspecting that the ignored small data (only considered noise until now) could indeed mean something that we didn't see before (a hidden chaos), so we started this research and studied since the basics of statistic (mean, variance and correlation) until advanced filtration techniques, like Wavelets and Bézier (this one was left behind for being inefficient for our objectives), so we did the application of these in temporal series of biological signal from MIT Database using the MATLAB tool to filter (2.8) the noises and analyse the data once neglected.

In the end, we achieved the filtration theoretically provided by the Wavelet filter (2.8) when we applied it on the biological signals from *MIT-BIH Arrhythmia Database*, with well defined peaks and with low noise in all detailed levels (section 3.4), in addition to achieve results that shows the proximity between computationally achieved data and the predicted peaks counting given by section (3.3). So, our expectation from now on is to keep going our study until creating an analytical detection method for every biological system in which it could be created universal numbers of easy diagnosis for diseases prevention (not only cardiovascular ones) from the previously ignored data, which it is the whole initial objective.

Keywords: Biological Signal. Correlation. Data. Peaks. Wavelet.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
BIH	*Beth Israel Hospital*
ECG	Eletrocardiograma
MIT	*Massachusetts Institute of Technology*
UFPB	Universidade Federal da Paraíba
TCC	Trabalho de Conclusão de Curso

Lista de símbolos

μ	Letra grega minúscula mu
ρ	Letra grega minúscula rho
τ	Letra grega minúscula tau
ϕ	Letra grega minúscula phi
ψ	Letra grega minúscula psi

Sumário

1 Introdução

1.1 Dados ao Longo do Tempo

Há muitos anos, tantos que nem sabemos a quantidade ao certo, a humanidade lida com os inúmeros dados que inundam o mundo ao nosso redor, seja para relacionar na antiguidade que se um raio caía em uma árvore ela pegaria fogo, seja na atualidade quando entendemos que se um semáforo está com uma luz vermelha acesa então nós devemos parar e esperar, a interpretação de dados está **sempre** presente em nossas vidas. Entretanto, a complexidade dos dados que precisamos avaliar cresceu exponencialmente mais rápido do que nossa capacidade cerebral de os analisarmos e, para suprir tal limitação biológica, foram criadas máquinas e computadores capazes de ver o que nossa mente não conseguia. Com isso saltos computacionais enormes foram realizados, tais como na meteorologia, que na década de 90 tinha uma precisão de acerto de 50% (o mesmo que você mesmo tentar no cara ou coroa se vai chover ou não), enquanto que na década passada já chegou à casa dos 90% (BROCHADO, 2017), isto se deve ao estudo de uma área bem recente, a Teoria do Caos.

1.2 Teoria do Caos

O caos pode parecer um conceito distante, mas é uma propriedade da natureza tão presente em nossas vidas quanto a própria gravidade. Caos nada mais é do que quando você tem um conjunto tão grande de dados (MASCENA; RAMOS, 2021) que, apesar de você saber as Leis físicas para aplicar ainda assim seria impossível calcular com exatidão todo o comportamento deste sistema pela **complexidade** do mesmo. Um exemplo de caos é o dia em que você conheceu seu(sua) melhor amigo(a), por exemplo, você já parou para pensar como sua vida seria extremamente diferente se naquele momento inicial você nunca tivesse conhecido esta pessoa? Você viveria momentos diferentes, recordações diferentes, tudo isto porque um único dia na sua vida foi mudado. Existem enredos inteiros de filmes baseados nesta premissa, como *Shrek 4*, por exemplo. Este conhecimento já existe há muito tempo na cultura pop e em especulações filosóficas, porém ele só ganhou peso de estudo científico por volta da década de 60, quando o meteorologista estadunidense Edward Lorenz (ESTRANHO, 2011) estava testando um programa de computador de massas de ar para gerar os dados do clima do mês e resolveu diminuir em quase nada a precisão do aplicativo, pensando que iria poupar tempo e em pouco iria mudar o resultado, pelo contrário, o que antes era um dia de sol no computador se tornou um dia com presença de tornados e céu completamente nublado. Isto foi a base que ele precisou para criar a Teoria

do Caos, ou como é mais popularmente conhecida na cultura pop: Efeito Borboleta. Com o passar do tempo foi se percebendo que o Caos estava presente até na Bolsa de Valores, mas surge uma pergunta que será o guia de todo este trabalho: E no nosso coração, o Caos está presente? Antes de avançar nisto precisamos entender como ele funciona.

1.3 Coração e o Sistema Circulatório

O coração é o principal órgão do sistema circulatório, responsável por distribuir o oxigênio que todo o corpo necessita da seguinte maneira (GUYTON; HALL, 2006):

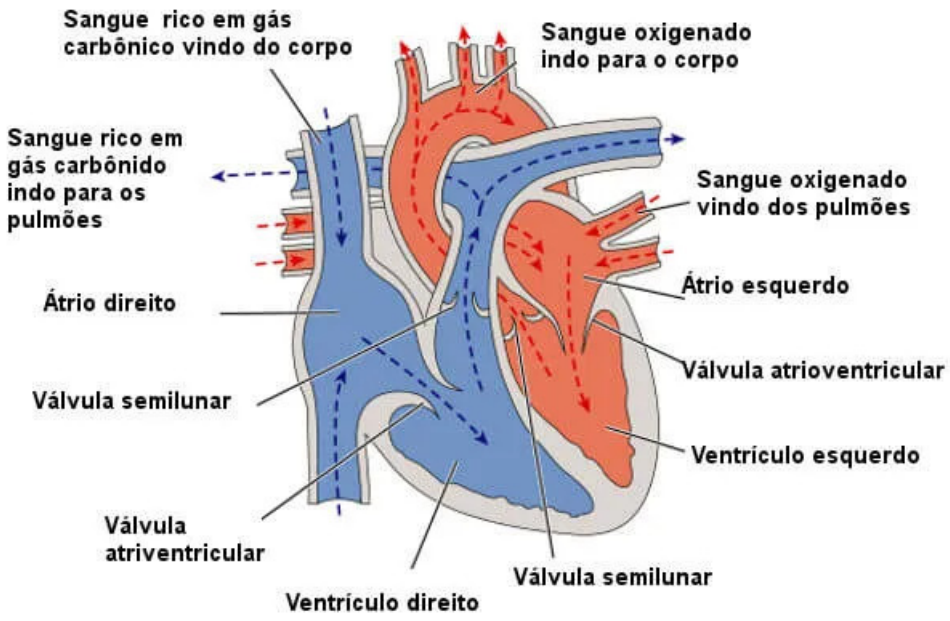

Figura 1 – Imagem de como o sangue flui no coração
Fonte:
<https://s5.static.brasilescola.uol.com.br/img/2019/06/fluxo-de-sangue-no-coracao.jpg>

Primeiramente o oxigênio chega ao corpo através dos pulmões e cai na corrente sanguínea, chegando ao coração através das veias pulmonares pelo átrio esquerdo e faz o caminho pelo ventrículo esquerdo seguindo para oxigenar o corpo saindo pela artéria aorta, a partir daqui vai ser uma competição entre duas energias que irão atuar: a cinética (de movimento) e a potencial (da gravidade, puxando para o chão), em que quanto mais descer, maior a cinética e menor a potencial e o oposto quando subir; ao longo do caminho estas energias oscilarão da seguinte forma:

- **Arterial:** No setor arterial (sangue oxigenado) o sangue é acelerado pela queda gravitacional até chegar na região em que vai ser designado, onde entra em estruturas menores, os capilares.

- **Capilar:** Aqui é onde o sangue oxigenado é espalhado nos tecidos, perdendo seu oxigênio e absorvendo o gás carbônico do local e em seguida retornando para estruturas maiores de retorno: as veias.

- **Venoso:** Em tal situação (gás carbônico no sangue) o sangue está retornando ao coração. Por conta da falta de velocidade, o movimento do sangue ocorre principalmente pela pressão que o bombeamento do coração causa, como se o sangue arterial empurrasse o venoso como um carro em uma subida de ladeira. E assim ele retorna ao coração, onde este sangue rico em gás carbônico é levado ao pulmão, que expele todo o gás carbônico que estava acumulado no sangue e absorvendo o oxigênio para reiniciar todo o ciclo.

Como se deve imaginar, cada uma destas etapas devem ter **milhares** de variáveis quando analisado, seja a relação entre os diâmetros de cada estrutura, além de energias que se dissipam por conta do atrito, seja a alimentação da pessoa, humor, sono, até às condições de voltagem da própria máquina com a qual se analisa o coração.

Tais sinais biológicos exibem diversas flutuações aleatórias diferentes em função de um parâmetro externo que podem variar, como o tempo, a voltagem, o campo magnético aplicado, etc. Então, para analisar como funciona o caos no coração, identificar médias paramétricas com médias em ensembles (conjuntos de dados) é um procedimento padrão e permite acumular estatísticas em função destes parâmetros experimentais. Isto nos convida a pensar sobre a possibilidade de extração de informação estatística útil (determinística) de uma única curva de um sinal biológico caótico. Já foi mostrado que é possível fazê-lo em muitos sistemas da matéria condensada. Mais especificamente, obteve-se sucesso em calcular a densidade de picos (máximos da função) e relacioná-la com a função de autocorrelação. Como resultado, nós tentamos achar um novo mensurável universal para sinais caóticos, a partir da análise de ensembles de dados de ECG fornecidos pelo MIT, utilizando o MATLAB para retirar tais flutuações aleatórias dos sinais biológicos. (GUYTON; HALL, 2006; RAMOS et al., 2011; MATLAB, 2018)

Parte I

Referenciais teóricos

2 Procedimentos Metodológicos

2.1 Fundamentação Teórica Estatística

Como o objetivo deste estudo é fazer uma análise de um grande banco de dados, é crucial entendermos os conceitos básicos de estatística, desde o conceito básico de conjunto às inúmeras maneiras de relacionarmos umas com as outras e/ou autorrelacionarmos estas coleções, e será toda esta base conceitual que criaremos nestas próximas seções:

2.2 Conjunto

Assim como muitos outros conceitos matemáticos, este é mais um em que o nome faz parecer ser mais difícil do que realmente é. Conjunto nada mais é do que qualquer agrupamento de elementos que possuam alguma característica em comum. Mas de qual maneira agrupamos? Ou de qual característica em comum estamos falando? A resposta para estas duas perguntas é: A que for conveniente. Da lista de compras até a relação fiscal de uma empresa, conjuntos podem ser organizados das mais variadas formas.

Eles podem ter vários elementos:

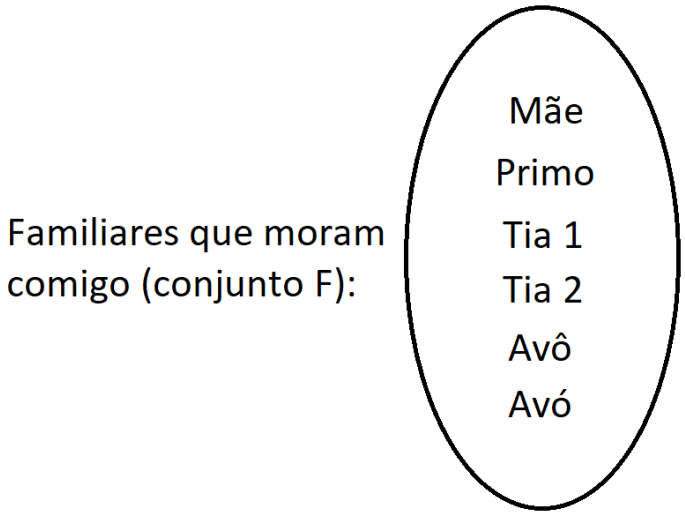

Figura 2 – Diagrama do conjunto F

Eles podem ter um único elemento:

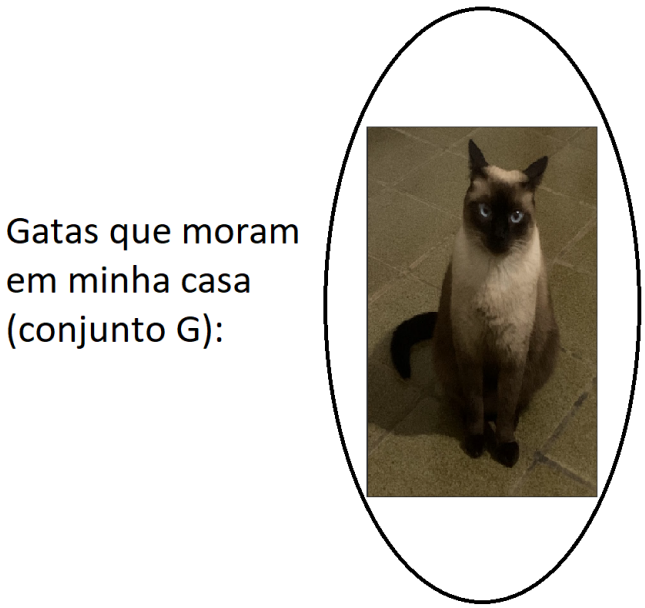

Figura 3 – Diagrama do conjunto G

Eles podem nem ter elemento:

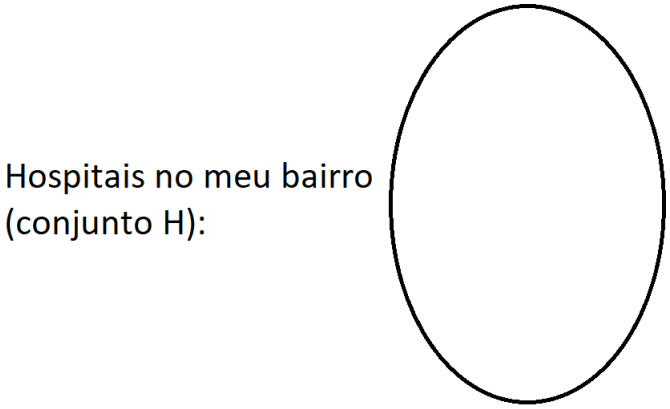

Figura 4 – Diagrama do conjunto H

E esta maneira gráfica de representarmos conjuntos com diagramas de Venn (esta elipse com os elementos dentro, presente nas 3 figuras acima) não é a única, outra forma bem comum é representar por enumeração (listagem de itens) e os 3 conjuntos anteriores utilizando esta forma de representação ficam assim:

$$F = \{\text{Mãe; Primo; Tia 1; Tia 2; Avô; Avó}\}$$
$$G = \{\text{Gatinha}\}$$
$$H = \{\} \text{ ou } \emptyset$$

E além de todas estas formas de representação, ainda falta ressaltar uma das características mais importantes dos conjuntos: **Eles podem ser comparados e operacionalizados**.

Para esta característica ser bem explicitada, analisemos este conceito com o exemplo bem cotidiano a seguir:

2.2.1 Exemplo de Comparação e Operação de Conjuntos

Imagine que você foi ao supermercado fazer uma pequena feira e para isto você fez uma lista (conjunto) das frutas que você está indo comprar, a qual representaremos assim:

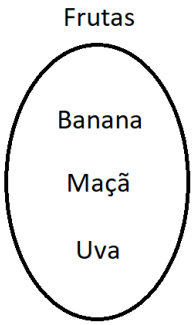

Figura 5 – Diagrama representativo da lista de Frutas (F) que estão sendo procuradas

Chegando no estabelecimento você encontra os valores para cada item que você procura:

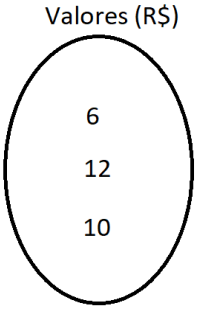

Figura 6 – Diagrama do conjunto de Valores (V) dos itens que estão sendo procurados

E a análise que subconscientemente já fazemos neste momento é o que matematicamente representamos desta maneira:

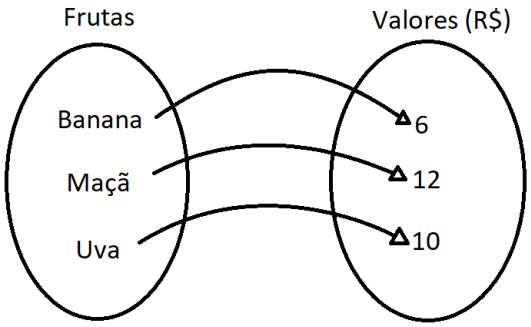

Figura 7 – Comparação de elementos de conjuntos

Nós comparamos e atribuímos valores para cada elemento que antes eram só nomes em uma lista. Por exemplo, o elemento "Banana" deixou de ser apenas "Banana" e se tornou "Banana custa R$6", nós paramos de analisar um único conjunto anterior para a partir de agora analisarmos dois conjuntos como um todo.

Pergunta: Mas de que realmente nos serve essa comparação?

Resposta: Para realizarmos a **operação** que bem entendermos. Agora que os elementos do conjunto F possuem **valores** nós podemos somar e descobrir quanto dará nossa feira, por exemplo:

$$\text{Banana} + \text{Maçã} + \text{Uva} = 6 + 12 + 10 = \text{R\$28}$$

E a partir de tal resultado nós podemos descobrir quão caro ou barato este valor está ao comparar com outros conjuntos, tais como: Valores do mês passado deste supermercado, valores do supermercado vizinho, etc.

E tudo isto foi possível graças as características dos conjuntos (vistas aqui: 2.2), em que agrupamos elementos com características em comum, comparamos estas coleções e realizamos uma operação (soma) em cima deles; entretanto, existem muitas outras operações que podemos fazer com conjuntos, das mais variadas complexidades e razões de utilização e são elas que começaremos a estudar a partir de agora.

2.3 Média

Na etimologia, a palavra média vem derivada do latim "mediu-"(PRESS, 2022) que tem o sentido de meio e na estatística o seu significado também não foge de sua origem, pois uma maneira de definir argumentativamente média é a seguinte:

Média: Dado um conjunto com "n" números, a média é um termo central com relação a eles, que diferentemente de outros termos estatísticos, como moda e mediana, o termo central da média é tal que se for multiplicado por "n" (quantidade de termos do conjunto) o resultado será exatamente o mesmo valor que ter apenas somado todos os termos do próprio conjunto inicial.

Matematicamente isto pode ser explicitado na seguinte equação:

$$\langle X \rangle = \mu = \frac{\sum x}{n} \tag{2.1}$$

Legenda: **X**: conjunto de termos do qual queremos saber a média
$\langle \mathbf{X} \rangle$ ou $\boldsymbol{\mu}$: Média do conjunto **X**
$\sum \boldsymbol{x}$: Soma de todos os termos do conjunto **X**
n: Quantidade de termos do conjunto **X**.

Mas como esta equação (2.1) se relaciona com a argumentação do que é média? Para explicitar melhor, vejamos o exemplo a seguir:

2.3.1 Exemplo Introdutório de Média

As notas de um aluno ao longo dos 3 trimestres do ano na disciplina de Geografia são mostradas a seguir:

Tabela 1 –Exemplo: Notas de um aluno na disciplina de geografia ao longo do ano.

Notas	Trimestre
4	1º
8	2º
9	3º

Pergunta: Qual é a média deste aluno na disciplina?

Resposta: Começamos este problema analisando a tabela (1) e vendo que o conjunto "Notas" (N) possui 3 elementos e é formado da seguinte maneira:

$$N = \{4; 8; 9\}$$

Agora que temos todos estes dados, podemos jogar tudo na equação (2.1) e teremos o resultado da média:

$$\langle N \rangle = \frac{4+8+9}{3} = 7$$

Temos o resultado matemático para este exemplo, porém o que podemos tirar disto? Uma interpretação que podemos ter é que a média é a criação de um novo conjunto semelhante ao inicial, de maneira que agora temos:

$$N = \{4; 8; 9\}$$
$$M = \{7; 7; 7\}$$

Como a soma dos elementos dos dois conjuntos são idênticos (total igual a 21, neste caso específico), temos que o conjunto M pode **representar** o conjunto N sem perda de sentido no quesito: Quantidade de pontos que o aluno tinha feito na disciplina de Geografia ao longo do ano. Entretanto, fica a pergunta: Por que calcular a média das notas, se eu poderia só analisar a soma de todas elas? Fazemos isto por alguns motivos, sendo eles:

1. Com a média você pode criar uma base comparativa com outros valores de média já estabelecidos, como a média mínima para passar de ano, por exemplo, e esta base não vai variar entre 0 e a soma das notas máximas, mas só entre 0 e a nota máxima de uma única prova, ou seja, você cria uma margem bem estabelecida de valores possíveis e pode trabalhar em cima deles;

2. Esta base comparativa com média não sofre variações se você acrescentar mais notas (no exemplo que vimos, se você acrescentar mais notas escolares, a média ainda vai continuar sendo um valor entre 0 e 10, diferente da soma das notas, que só vai crescer cada vez mais), o que mantém a estabilidade do item anterior.

Então na prática a grande vantagem de se usar média é que possuindo uma quantidade maior de dados, com a média você pode resumir todos eles para apenas extrair a informação de onde deve ser o termo central do seu conjunto de dados, assim como a média 7 (informação que o professor precisa saber) foi o termo do meio para as notas 4, 8 e 9.

Entretanto, quando discutimos anteriormente sobre o fato da semelhança entre os conjuntos média (M) e nota (N) ser que a soma de seus elementos são iguais (veja aqui: 2.3.1) nós passamos direto por um detalhe muito importante: **A soma dos elementos são iguais, não os conjuntos.**

Os valores do conjunto N são completamente diferentes dos valores do conjunto M, quando comparamos um com o outro até vemos que existem valores que são próximos ao valor da média, por exemplo: o valor 8 está a apenas 1 de distância da média 7, enquanto 4 já está mais longe, a 3 de distância da média 7. Mas de que maneiras nós poderíamos

quantificar a diferença entre todo um conjunto e sua média, são justamente estas medidas de **dispersão** de conjuntos que estudaremos a partir de agora.

2.4 Desvio Padrão

Como comentei brevemente na provocação do final da seção anterior, existem várias maneiras de se medir a dispersão entre os valores de um conjunto e sua média, a principal delas é o desvio padrão (σ), que analisa a diferença entre um conjunto e sua média da seguinte maneira:

Dado um conjunto **X**, com uma quantidade **N** grande dados finitos ($\boldsymbol{x_1, x_2, \ldots, x_N}$) com mesma chance de aparição (mesmo peso) $\mathbf{1/N}$, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \langle X\rangle)^2} \tag{2.2}$$

Legenda: $\boldsymbol{\sigma}$: Desvio padrão do conjunto **X**
$\frac{\mathbf{1}}{\boldsymbol{N}}$: Probabilidade de aparição de cada elemento do conjunto **X**
N: Quantidade de termos do conjunto **X**.
$\boldsymbol{x_i}$: Elementos do conjunto **X**;
$\langle \boldsymbol{X}\rangle$: Média do conjunto **X**.

Quanto mais próximo o desvio padrão estiver de zero, mais semelhante à média o conjunto será. Por consequência lógica, quanto mais alto o valor deste desvio padrão, mais afastados da média os elementos do conjunto serão.
Imagino que simplesmente apresentar a equação e já partir para a análise minuciosa dela acabe mais atrapalhando o entendimento do que facilitando, então, antes de analisarmos as nuances da equação (2.2) vamos primeiro a um exemplo da aplicação de desvio padrão para que tudo possa ser melhor assimilado.

2.4.1 Exemplo Introdutório de Desvio Padrão

Imagine que dois alunos (Alice e Bob, ou A e B, como preferir) estão concorrendo a uma única vaga de bolsa de pesquisa e, para decidir quem ficará com ela, tanto Alice quanto Bob realizarão 4 provas e, quem obtiver a maior média conseguirá a vaga. Em caso de empate, o critério de desempate favorecerá quem obtiver a menor diferença entre as notas de cada prova e a média total. Os resultados nas 4 provas, de cada um deles, estão mostrados na tabela a seguir:

Tabela 2 –Exemplo: Notas de Alice e Bob nas provas para bolsa de pesquisa

Prova	Nota de Alice	Nota de Bob
1º	10	9
2º	10	10
3º	9	8
4º	7	9

Pergunta: Qual dos dois conquistou a bolsa de pesquisa?

Resposta: Aplicando a equação (2.1) conseguimos ver que as médias de Alice e Bob são idênticas e iguais a 9.

$$\langle A \rangle = \langle B \rangle = 9$$

Visto que ambos concorrentes obtiveram a mesma média, o desempate será calculado a partir do desvio padrão de cada candidato, e então, aplicando a equação (2.2) e utilizando $\boldsymbol{\sigma_A}$ para Alice e $\boldsymbol{\sigma_B}$ para Bob, temos que:

$$\sigma_A = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \langle X \rangle)^2} = \sqrt{\frac{1}{4}\{(10-9)^2 + (10-9)^2 + (9-9)^2 + (7-9)^2\}} \approx 1,2$$

$$\sigma_B = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \langle X \rangle)^2} = \sqrt{\frac{1}{4}\{(9-9)^2 + (10-9)^2 + (8-9)^2 + (9-9)^2\}} \approx 0,7$$

Ou seja: As notas de Bob oscilaram menos de valor, já que seu desvio padrão ($\boldsymbol{\sigma_B}$) foi mais baixo, o que indica que suas notas permaneceram mais próximas do seu valor médio do que as de Alice, portanto, ele ficou com a bolsa de pesquisa.

Agora que vimos a equação e um exemplo de sua aplicação, neste momento você pode estar se perguntando:

Pergunta: Por que eu estou elevando ao quadrado as diferenças entre os elementos e a média para depois passar a raiz quadrada (operação inversa da potenciação) na equação (2.2)? Não seria muito mais simples e prático simplesmente calcular as diferenças e somar, sem elevar e passar raiz?

Resposta: De fato seria mais simples, mas estaria **errado**, pois, imagine que alguém tirou duas notas: **10** e **4**. Pelo conhecimento que temos da equação (2.1), a média desta pessoa é **7**, se apenas calculássemos a diferença (sem elevar ao quadrado) para saber o desvio padrão, a parte dentro do parêntese do somatório da equação (2.2, sem considerar a contribuição ao quadrado) ficaria assim:

$$(10-7) + (4-7) = 3 - 3 = 0$$

que somando com todos os outros termos da equação (2.2, sem considerar a contribuição ao quadrado) resultaria em $\boldsymbol{\sigma}$=0, ou seja, implicando que os dois conjuntos são idênticos ({10;4}={7;7}), **uma afirmação claramente falsa** que só ocorreu de ser feita porque a maneira que utilizamos para medir dispersão fez com que contribuições positivas (10, que é maior que 7) se anulassem com contribuições negativas (4, que é menor que 7). E é justamente para evitar isto que elevamos ao quadrado, para que números negativos, ao serem elevados ao quadrado se tornem positivos e assim todos os valores **contribuam** para o desvio padrão, ao invés de se aniquilarem.

Para além do desvio padrão existem várias outras maneiras de analisar dispersão de valores com relação à média (detalharemos por que usamos uma ao invés da outra em cada situação, mas não neste momento), antes, iremos estudar mais uma delas a seguir.

2.5 Variância

Como dito ao fim da seção anterior: A variância é mais uma maneira de medir a dispersão dos dados em relação a média de um conjunto. Mas não é qualquer maneira, é uma extremamente semelhante ao próprio desvio padrão, tão semelhante que até o símbolo de variância ($\boldsymbol{\sigma^2}$ ou **var**) é inspirado no de desvio padrão ($\boldsymbol{\sigma}$) e esta elevação ao quadrado no símbolo de variância não é a toa, pois a equação que define a variância em si é dada da seguinte maneira:

Dado um conjunto **X**, com **N** dados finitos ($\boldsymbol{x_1, x_2, \ldots, x_N}$) com mesma chance de aparição (mesmo peso) $\mathbf{1/N}$, a variância é dada por:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \langle X\rangle)^2 \tag{2.3}$$

Legenda: $\boldsymbol{\sigma^2}$: Variância do conjunto **X**
$\frac{\mathbf{1}}{\boldsymbol{N}}$: Probabilidade de aparição de cada elemento do conjunto **X**
N: Quantidade de termos do conjunto **X**.
$\boldsymbol{x_i}$: Elementos do conjunto **X**;
$\boldsymbol{\langle X\rangle}$: Média do conjunto **X**.

Sim, a equação (2.2) do desvio padrão, só que elevada ao quadrado, por conta disto acaba mantendo praticamente todas as propriedades do desvio padrão, tais como: ser **sempre** maior ou igual a zero; e quanto mais próximo de zero o valor da variância, mais próximos da média os termos do conjunto estarão e por consequência lógica quanto mais distante de zero mais espalhados ou distantes da média serão os termos do conjunto.

Uma outra maneira de analisar a variância é a partir de uma outra definição. Dado um conjunto **X**, sua variância será dada por:

$$var(X) = \langle X^2 \rangle - \langle X \rangle^2 \tag{2.4}$$

Entretanto, uma maneira de se escrever esta equação (2.4) é, ao invés de utilizar conjuntos, utilizarmos **vetores**.

Vetores: Nada mais são do que um **conjunto** de valores (assim como vistos antes), só que agora para onde eles estão apontados, ou sua **direção** e **sentido**, começam a importar, exemplo: Quando se está dirigindo um carro, a velocidade dele não é o único fator crucial, em qual sentido você dirige também é extremamente importante, se está dirigindo para fora da pista você estará com um grande problema justamente pelo seu sentido estar diferente de para onde a pista aponta.

Utilizando este conceito, a equação (2.4) com um vetor $[\overrightarrow{v} = (v_1, v_2, \ldots, v_n)]$ de **n** termos resultará na seguinte equação:

$$var(\overrightarrow{v}) = \langle \overrightarrow{v} \cdot \overrightarrow{v} \rangle - \langle |\overrightarrow{v}| \rangle^2 \tag{2.5}$$

Em que $(\cdot)$ é o produto escalar.

Uma vez entendidos os conceitos das seções 2.4 e 2.5 pode ser que surja a seguinte dúvida:

Pergunta: Se a variância é realmente tão semelhante ao desvio padrão, para que utilizar as duas quando poderíamos usar só uma?

Resposta: Pela mesma razão que usamos hectares (ha) para medir área de grandes terrenos e metros quadrados (m^2) para medir a área de um quarto, pois objetos diferentes servem a propósitos diferentes, sendo assim deixarei um motivo para a utilização de cada um dos métodos de dispersão:

- **Por que usar variância?** Por conta de uma propriedade única sua, em que, para conjuntos de variáveis aleatórias e independentes (que não tem relação) uma da outra, **a variância da soma (ou subtração) é igual a soma das variâncias**. Matematicamente, esta relação se apresenta assim:

$$var(X + Y) = var(X) + var(Y)$$

- **Por que usar desvio padrão?** Por conta que seu valor é dado na mesma unidade de medida do conjunto que se analisou, enquanto que a variância eleva ao quadrado, exemplo: Se analisarmos o desvio padrão da idade de um grupo de trabalhadores, a resposta será algo do tipo: “Desvio de x anos em relação à média”. Enquanto que a

variância seria uma resposta do tipo: "Variância de y $(anos)^2$ em relação à média". Anos ao quadrado não tem nem significado prático e é assim para boa parte das respostas de variância: **Elas não costumam ter significado físico**.

Agora que entendemos conceitualmente os principais métodos de dispersão e suas diferenças, vamos entender visualmente o que eles representam, na seção a seguir.

2.6 Função Normal

Se for necessário analisar as dispersões de dados de um conjunto **X** com valores dados pela função normal a seguir:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2.6}$$

O conhecimento que obtivemos até agora não será suficiente, pois apenas tratamos de conjuntos com uma quantidade discreta de termos. Para tratar de funções contínuas, as conhecidas equações deixarão de ser **somatórios**, como nas equações (2.1, 2.2, 2.3) e se tornarão **integrais**, mantendo as mesmas propriedades e relações, recebendo novas aparências, demonstradas a seguir, as quais usaremos na função (2.6):

Dado um conjunto de variáveis aleatórias contínuas, **X**, com função densidade de probabilidade **f(x)**, então temos que:
A **média** será:

$$\mu = \langle X \rangle = \int_{-\infty}^{+\infty} x f(x) dx \tag{2.7}$$

A **variância** será:

$$\sigma^2 = var(X) = \int_{-\infty}^{+\infty} [x-\mu]^2 f(x) dx \tag{2.8}$$

E o **desvio padrão**:

$$\sigma = DP(X) = \sqrt{var(X)} \tag{2.9}$$

Apesar destas equações poderem dar os resultados de média e dispersão da função normal dada pela equação (2.6), felizmente não precisaremos utilizá-las, pois a função normal já é boa por alguns bons motivos, dentre eles, temos:

- O gráfico de sua função descreve vários um comportamento geral de inúmeros eventos naturais, tais como movimento browniano;
- Encare com um pouco mais de atenção a equação (2.6), veja bem como velhos conhecidos já estão **explícitos na própria função**. Observe que, na equação (2.6),

sem nem precisar calcular nada a própria função já nos diz quem são a **média**, **desvio padrão** e **variância**, conforme destacado a seguir:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2.10}$$

Média

Variância

Desvio Padrão

Por conta de tais características tão únicas, quando uma variável aleatória (conjunto) **X** possui tantos termos que se aproxima de uma função normal, não se costuma escrever a função normal por extenso, mas sim escreve-se **X** como uma nova notação **N** que se refere a função normal em termos da **média** e **variância**, cuja notação matemática é: $\boldsymbol{X \sim N(\mu, \sigma^2)}$.

Entretanto, uma dúvida pode surgir...

Pergunta: Por que quando uma variável **X** se aproxima da função normal nós a escrevemos como $\boldsymbol{X \sim N(\mu, \sigma^2)}$, sem colocar o desvio padrão junto?

Resposta: Como variância e desvio padrão são relacionados diretamente um ao outro sempre do mesmo modo ($\boldsymbol{DP(X) = \sqrt{var(X)}}$), independente de como o conjunto é formado, desvio padrão e variância **sempre** se relacionam deste jeito, por conta disto, escrever a dependência da função normal com desvio padrão **e** variância seria simplesmente uma redundância (não acrescenta informação alguma), pois só se precisa de um para ter o entendimento completo, por conveniência, o escolhido foi a variância.

Compreendemos as dispersões e como elas estão presentes na função normal (2.10), entretanto, qual será a mudança gráfica que acontecerá nesta função se os valores de média e variância oscilarem? É o que será visto no exemplo a seguir.

2.6.1 Exemplo Gráfico: Função Normal

Agora que os aspectos mais cruciais da função (ou distribuição) normal (equação 2.6) foram compreendidos, veremos como ela se comporta de fato. Existem infinitas distribuições normais, cada uma com suas respectivas médias e variâncias (consequentemente desvio padrão também, como visto na pergunta final da seção 2.6), contudo, a distribuição com média $\mu = 0$ e variância $\sigma^2 = 1$ é chamada de **distribuição normal padrão** e possui notação $\boldsymbol{N(0, 1)}$ e a mesma se distribui conforme o gráfico a seguir:

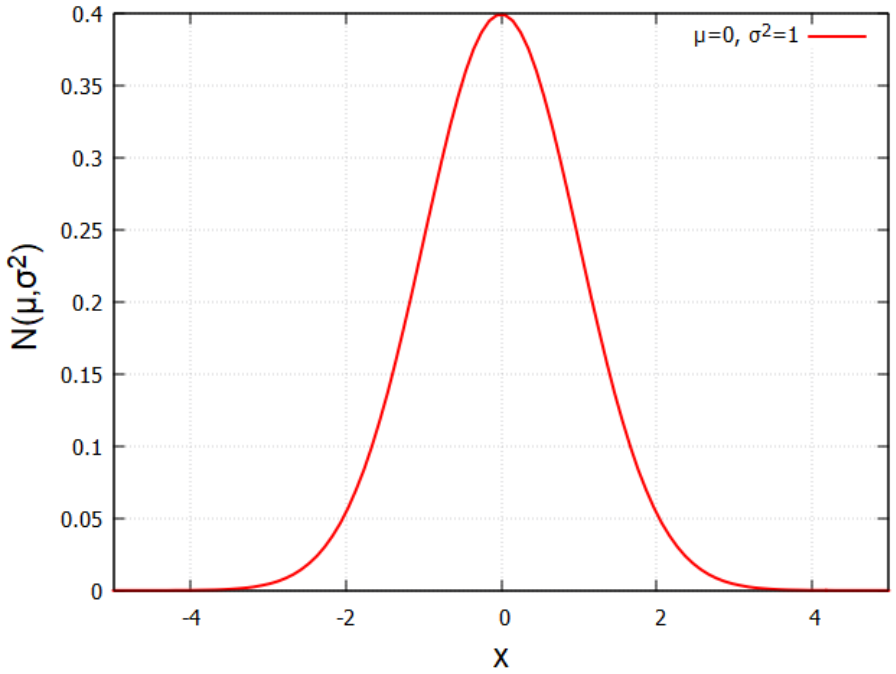

Figura 8 – Gráfico da distribuição normal padrão $\boldsymbol{N(0,1)}$

E é justamente por este formato que a distribuição normal é chamada de curva de sino, mas neste momento, mais importante do que saber seu outro nome é interessantíssimo ressaltar onde esta curva (a função em si) está centralizada, que é justamente no valor da média. Ou seja, a curva, mesmo com valores espalhados nos valores positivos e negativos de "x", possui seu máximo, seu **termo central** de dados, exatamente onde é sua média, em $\boldsymbol{x = 0}$, ou seja, a função normal mostra **explicitamente** o que estudamos argumentativamente na seção (2.3): **A média é o termo que por definição (para distribuições como esta) é a centralização dos valores do conjunto.**

Agora que uma noção básica de como a distribuição normal se comporta foi adquirida, vamos retomar ao fato anteriormente citado de que existem infinitas distribuições normais e vamos, em único gráfico, comparar a distribuição padrão que exibimos na figura (8) com vários outros exemplos:

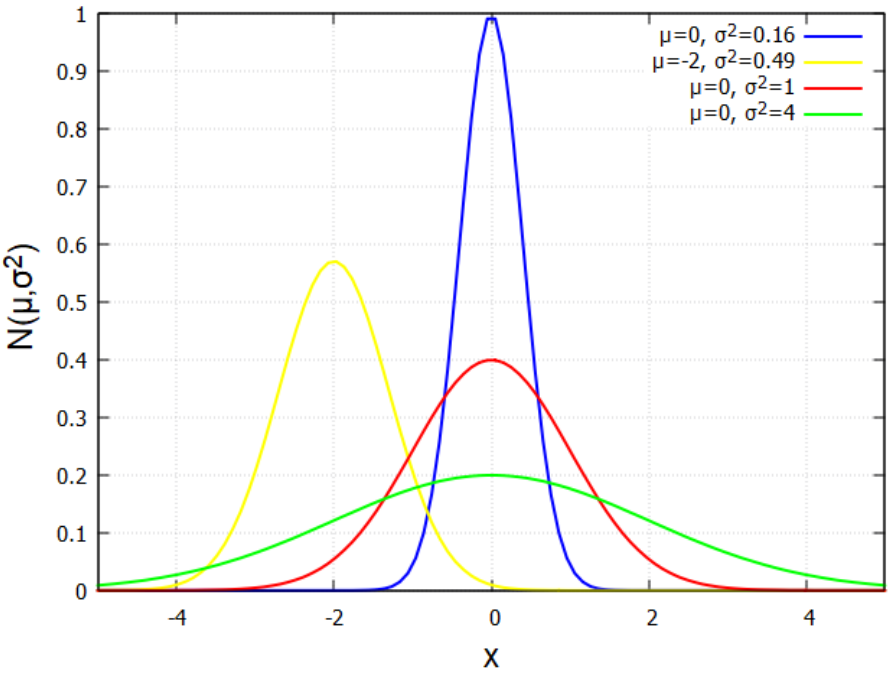

Figura 9 – Gráfico de várias distribuições normais para efeito comparativo de média e desvio padrão

Vê-se primeiramente neste gráfico a reconfirmação de como a média centraliza a função, pois mesmo distribuições normais diferentes com mesmo valor de média se centralizam em torno da mesma, e quando se desloca o valor de média na curva $\boldsymbol{N(-2, 0.49)}$ a função inteira se desloca e centraliza em conjunto no valor da média.

Porém, um outro fato excelente pode ser visto ao comparar os formatos das funções em si, pois, aparentemente quanto menor o valor da variância (σ^2) mais a função se afina, como um pico de uma montanha, em contrapartida, quanto maior o valor da variância (σ^2) mais a função "abaixa" e suaviza, como uma duna em um deserto. Isso não é a toa, pois é exatamente isto o que a variância faz em uma função, nas seções (2.4 e 2.5) eu não as nomenclaturei como **medidas de dispersão** por acaso. Como já citado nas seções (2.4 e 2.5): o que o desvio padrão e a variância fazem é justamente calcular quão perto ou afastado os termos de um conjunto estão da média dele próprio. Se o valor da medida de dispersão for cada vez mais próximo de 0, menos espalhado seus valores estarão, ou seja: Todos seus dados estarão extremamente centralizados, na figura (9) a curva que melhor representa isto é a $\boldsymbol{N(0, 0.16)}$, que, não coincidentemente, possui a menor variância. Em contrapartida, quanto maior for o valor das medidas de dispersão, mais espalhados seus valores estarão, ou seja: Todos seus dados estarão cada vez mais distantes do termo central, na figura (9) a curva que melhor representa isto é a $\boldsymbol{N(0, 4)}$ que é justamente a com a maior variância de todas.

Até o presente momento já foi explicado a noção de termo central de um conjunto e como os elementos deste podem se dispersar com relação à media dele mesmo, entretanto,

como se pode analisar relações de dependência entre dois conjuntos diferentes? Por conta disto, estudaremos a seguir a mais importante das medidas de dependência.

2.7 Correlação

A correlação é uma medida de dependência de conjuntos, e como tal ela existe com o intuito de analisar quão semelhante (parecido) um conjunto é com relação a um ou mais outros conjuntos, de tal maneira que quanto mais correlato (quão maior for o valor da correlação), mais um conjunto se assemelhará ao outro, ou seja, menos **aleatórios** todos os dados analisados serão, já que existe alguma **semelhança**.

Uma maneira simples de entender o conceito de correlação é com o exemplo a seguir:

2.7.1 Exemplo Conceitual: Correlação

Imagine que uma fábrica produz faixas e elas são feitas no seguinte padrão: Começa-se com faixas de uma única cor e a cada 3 faixas desse padrão de cor fabricadas as próximas 3 faixas receberão uma nova cor (ou seja, terão duas cores, ao invés de uma) e assim em diante. Dado este padrão, as 9 primeiras faixas produzidas nesta fábrica podem ser vistas na imagem a seguir:

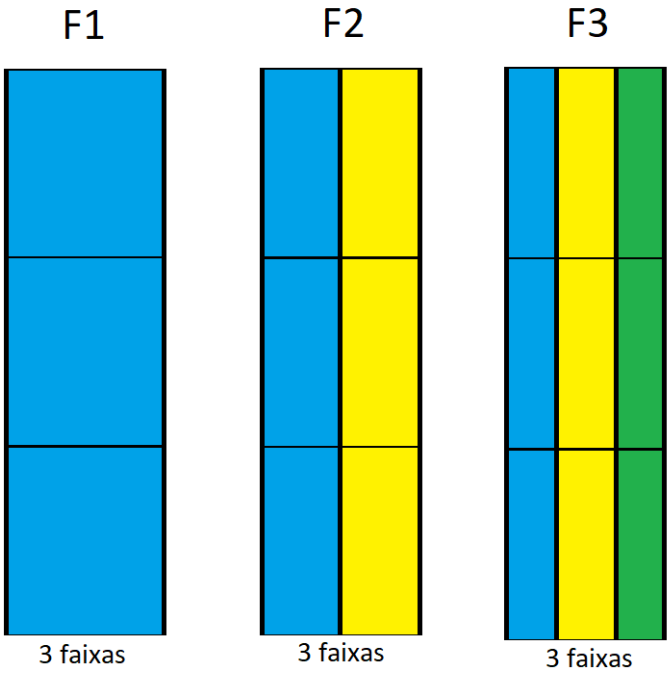

Figura 10 – Exemplo de correlação: Modelos de faixas produzidos em uma fábrica

A noção de correlação começa quando você olha a figura (10) e tem o seguinte pensamento lógico: As 3 primeiras faixas F1 se parecem bastante com as faixas F2, elas também se parecem com as de F3, **mas F1 se parece mais com F2 do que com F3.**

Pergunta: Mas por que essa frase em negrito logo acima seria tão importante?

Resposta: Porque aí mora justamente o cerne do que é correlação: Nesta mesma frase você entende que F1 possui semelhanças (ou seja, correlação) com F2 e com F3, ao mesmo tempo em que também entende que F1 se parece mais com F2 do que com F3, ou seja: você entendeu que existem correlações e que elas podem ser **maiores** que outras, isto sem fazer uma conta sequer.

Agora que foi explicado como o conceito de semelhança está relacionado à correlação, uma outra pergunta importante precisa ser feita.

Pergunta: Dentre todas as colunas possíveis da figura (10), com qual a coluna de faixas F1 se parece mais (possui maior correlação)?

Resposta: Se a sua resposta tiver sido F2, F3 ou qualquer conjunto com ainda mais cores, saiba que sua resposta está **errada**. O conjunto que mais se parece com F1 é **o próprio F1**, do mesmo jeito que a pessoa que mais se parece com você no mundo inteiro é você, o conjunto mais parecido consigo próprio é ele mesmo, ou seja: **A maior correlação que um conjunto pode ter é com o próprio conjunto.**

Convertendo as noções que tiramos da figura (10) para a linguagem matemática somos levados ao seguinte: dados os seguintes conjuntos de vetores,
$\overrightarrow{v^{(1)}} = ({v_1}^{(1)}, {v_2}^{(1)}, \ldots, {v_n}^{(1)})$, $\overrightarrow{v^{(2)}} = ({v_1}^{(2)}, {v_2}^{(2)}, \ldots, {v_n}^{(2)}), \cdots, \overrightarrow{v^{(n)}} = ({v_1}^{(n)}, {v_2}^{(n)} \ldots {v_n}^{(n)})$,
temos que a correlação entre o primeiro vetor e todos os outros é dada por:

$$\begin{aligned} C_1^{(1)} &= \langle \overrightarrow{v^{(1)}} \cdot \overrightarrow{v^{(1)}} \rangle - \langle \overrightarrow{v^{(1)}} \rangle \langle \overrightarrow{v^{(1)}} \rangle = var(\overrightarrow{v^{(1)}}) \\ C_1^{(2)} &= \langle \overrightarrow{v^{(1)}} \cdot \overrightarrow{v^{(2)}} \rangle - \langle \overrightarrow{v^{(1)}} \rangle \langle \overrightarrow{v^{(2)}} \rangle \\ &\vdots \\ C_1^{(n)} &= \langle \overrightarrow{v^{(1)}} \cdot \overrightarrow{v^{(n)}} \rangle - \langle \overrightarrow{v^{(1)}} \rangle \langle \overrightarrow{v^{(n)}} \rangle \end{aligned} \tag{2.11}$$

Legenda: $\boldsymbol{C_1^{(n)}}$: Correlação entre o vetor $\overrightarrow{v^{(1)}}$ e o vetor $\overrightarrow{v^{(n)}}$

Vemos da primeira das equações presentes na equação (2.11) que a maior correlação possível (afinal, a maior semelhança possível do vetor é com ele mesmo) é justamente a própria variância (leia seção 2.5 e a equação 2.5 para maior entendimento), por conta disto que para fins de normalização (organização) dos dados, quando formos criar gráficos de correlação dividiremos todos os valores pelo valor da variância (valor máximo). Assim quanto mais próximo de 1, mais os dados se assemelham e quanto mais próximo de 0, menos se assemelham. O que esperamos dos nossos dados é que eles comecem com correlação máxima e decaiam lentamente com o tempo (pois terão cada vez menos relação com

os dados iniciais), um esboço de como é o decaimento leve de um gráfico de correlação **normalizado** está na figura a seguir:

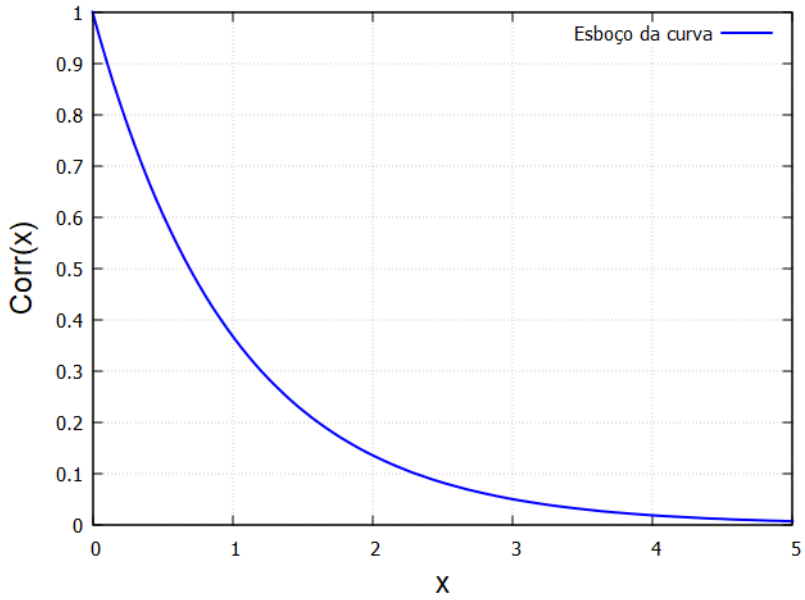

Figura 11 – Esboço de como uma série correlacionada deve se comportar ao longo do tempo.

Uma vez entendido os conceitos básicos de correlação, é de extrema importância ressaltar uma das características cruciais da correlação, é tão relevante que separei a próxima subseção só para ela.

2.7.2 Correlação Não Implica Causalidade

Uma forma mais simples de ler este título seria: "Semelhança não indica causa". Apesar do exemplo dado na subseção (2.7.1) ter uma regrinha de construção (a cada 3 faixas a faixa seguinte teria uma cor a mais), o fato de existir alguma semelhança (correlação) que seja entre os dados não nos diz **NADA** a respeito de se algum desses dados é de alguma forma a origem do outro ou não. Vamos analisar a imagem a seguir para deixar explícito o que significa este argumento.

Figura 12 – Correlação entre um gato e um telhado amassado.
Fonte: <https://pbs.twimg.com/media/FCd-PlWWUAM0L3g?format=jpg&name=small>

Olhando a imagem é clara a correlação que existe: Há um gato em cima de um amassado no telhado do tamanho do próprio gato. Entretanto, pelo conhecimento que temos da massa que um gato costuma ter, podemos afirmar que: Apesar da **correlação** existente entre o gato e o telhado amassado, não tem condição do gato ser a **causa** deste telhado amassado.

Ou seja: Independente de qual correlação se encontre, ela **NADA** estará dizendo sobre a causa e consequência dos eventos. Todo esse raciocínio é crucial para este trabalho justamente pelo fato de que não basta apenas encontrar correlação entre os dados, **também** é necessário provar que eles possuem causas e consequências, que não são advindos da própria aleatoriedade.

Agora que compreendemos a estrutura central dos conjuntos, a dispersão dentro deles próprios e como podemos analisar a semelhança entre conjuntos diferentes, nos falta mais um desafio: Remover o ruído; de nada adianta termos ferramentas para analisar dados e não conseguirmos identificar nada por tantos valores errados estarem no meio. Por conta disto, iniciaremos agora os estudos dos processos de filtragem de dados.

2.8 Wavelet

A família de filtros *Wavelet* é conhecida pela sua característica peculiar de filtrar os dados através de um processo de desmontagem e remontagem dos valores, destacando as "peças" que somam para formar o sinal original. Este processo é chamado de **análise multiresolução** e as "peças" são o sinal de aproximação mais todos os níveis de discretização. A formalização matemática pode ser vista a seguir.

2.8.1 Ideia Abstrata da Aproximação de Dados

Da Álgebra Linear, nós podemos decompor um sinal em uma combinação linear dos vetores da base se o próprio sinal está no espaço gerado por tal base, matematicamente temos:

$$f(t) = \sum_k a_k \phi_k(t) \tag{2.12}$$

Em que k é um índice inteiro da soma finita ou infinita, a_k são os coeficientes da expansão e $\phi_k(t)$ são as funções da base. Em geral, se escolhemos as bases corretamente, vai existir uma outra base formada por $\{\tilde{\phi}_k(t)\}$ tal que $\{\phi_k(t)\}$ e $\{\tilde{\phi}_k(t)\}$ são ortonormais, ou seja, seu produto interno é:

$$< \phi_i(t), \tilde{\phi}_j(t) >= \int \phi_i(t)\tilde{\phi}_j^*(t)dt = \delta_{ij} \tag{2.13}$$

em que $\{\tilde{\phi}_k(t)\}$ é chamado de função dual de $\{\phi_k(t)\}$ e utilizando tal propriedade da equação (2.13) temos como achar todos os coeficientes da expansão da seguinte maneira:

$$a_k =< f(t), \tilde{\phi}_k(t) >= \int f(t)\tilde{\phi}_k^*(t)dt \tag{2.14}$$

Tal procedimento pode ser aplicado na compressão de sinal caso preserve a semelhança com o sinal original ao mesmo tempo em que comprime.

2.8.2 Dois Exemplos sobre Multiresolução

2.8.2.1 Exemplo 1: Aproximando Sinais Discretos no Tempo Usando Função Delta

Consideremos o sinal discreto no tempo (n)

$$x[n] = \left(\frac{1}{2}\right)^{|n|} \tag{2.15}$$

Escolhemos a base $\{\phi_k(n)\} = \{\delta[n-k]\}$ e o seu dual $\{\tilde{\phi}_k(n)\} = \{\phi_k(n)\} = \{\delta[n-k]\}$

Vemos da equação 2.14 que:

$$a_k =< x[n], \delta[n-k] >= \sum_{n=-\infty}^{\infty} \left(\frac{1}{2}\right)^{|n|} \delta[n-k] = \left(\frac{1}{2}\right)^{|k|} \tag{2.16}$$

Se restringirmos os valores do índice k $\in [K_1, K_2]$ obtemos as aproximações das figuras abaixo:

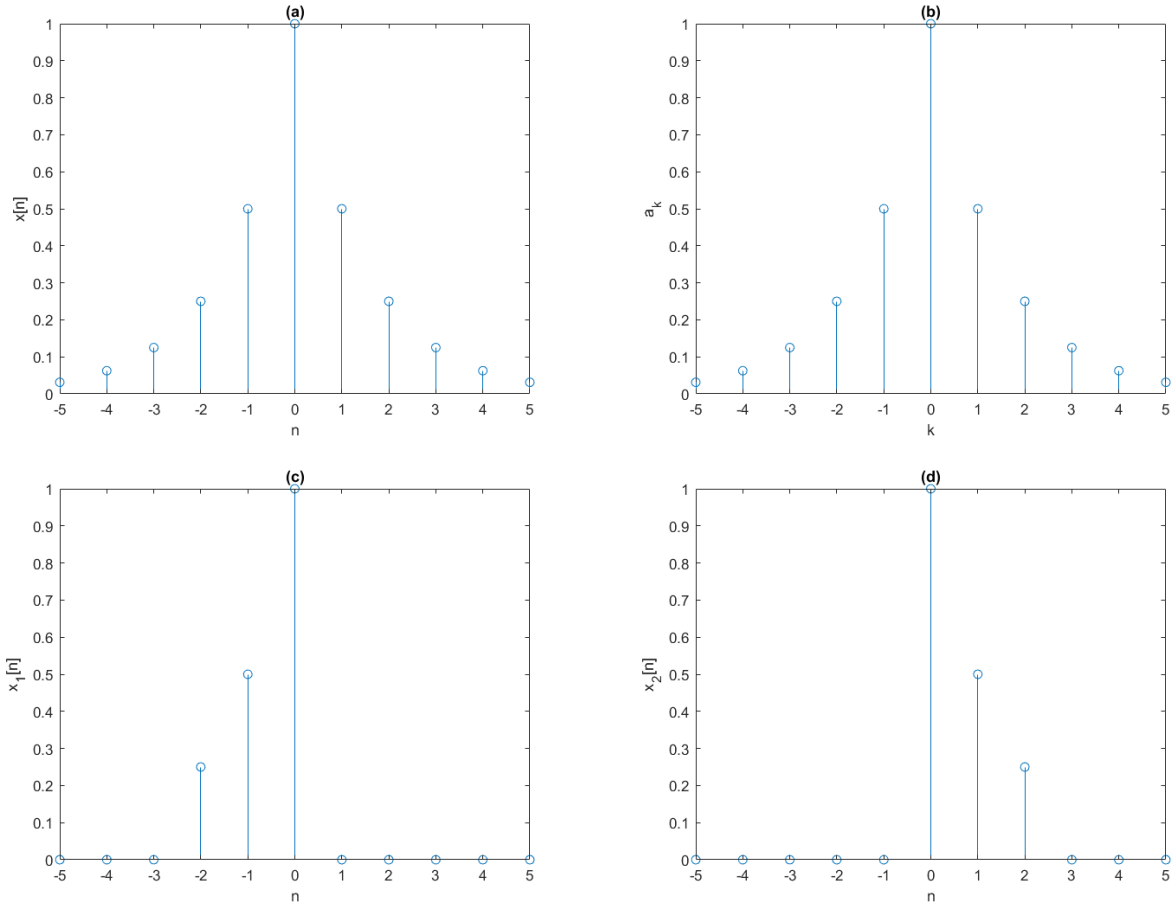

Figura 13 – Resultado da aproximação por função delta. **(a)** Sinal original x[n] **(b)** Coeficientes a_k obtidos **(c)** Sinal reconstruído $x_1[n] = \sum_{k=-2}^{0} a_k \delta[n-k]$ **(d)** Sinal reconstruído $x_2[n] = \sum_{k=0}^{2} a_k \delta[n-k]$

Tal exemplo usa a base $\delta[n-k]$ para achar os coeficientes, mas pode também ser interpretada como uma translação da delta normal $\delta[n]$, tal que $\delta[n-k]$ é a posição do impulso localizado em n=k. O sinal reconstruído usa coeficientes parciais das posições que nos interessam (para estudo da discretização no tempo), por exemplo, para analisar o sinal em n $\in$ [-2,0], podemos usar a_{-2}, a_{-1} e a_0 para achar a versão reconstruída.

2.8.2.2 Exemplo 2: Aproximando Sinais com Ajuste de Escala

Diferente do caso anterior em que fizemos nossa base transladando a função delta, agora vamos trabalhar com ajuste de escala. Considerando a função contínua abaixo:

$$\phi(t) = \begin{cases} 1, & 0 \leqslant t < 1 \\ 0, & \text{Nos outros casos} \end{cases} \tag{2.17}$$

Da sua definição temos que $\phi(t)$ é uma função retangular centrada em $t = \frac{1}{2}$ e média igual a 1. A versão re-escalada é definida como $\phi_s(t)$ tal que:

$$\phi_s(t) = \phi(st) \tag{2.18}$$

Veja que s é um fator de escala contínuo, tal que conforme s aumenta, o intervalo não nulo da função é estreitado. Para uma base discreta, como no exemplo em que s=2^n (n inteiro), temos que:

$$\phi_n(t) = \phi(2^n t) \tag{2.19}$$

Veja nas figuras (14) que, conforme a frequência (2^n) cresce, a largura da função diminui, assim os coeficientes de alta frequência podem ser aproximados pelas escalas de alta ordem.

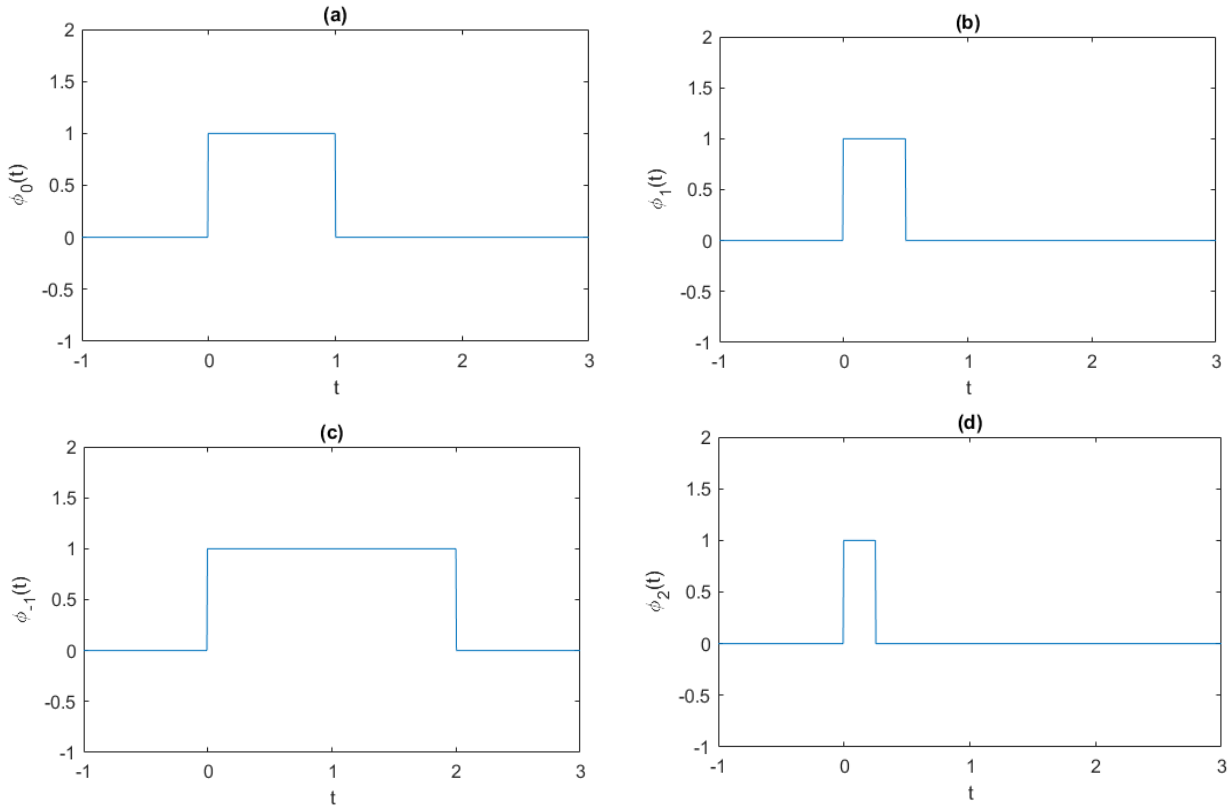

Figura 14 – Gráfico da operação por função de escala (Usei n como índice do gráfico). **(a)** $\phi_n(t) = \phi_0(t)$ **(b)** $\phi_n(t) = \phi_1(t)$ **(c)** $\phi_n(t) = \phi_{-1}(t)$ **(d)** $\phi_n(t) = \phi_2(t)$

Teste de ortogonalidade da família de funções $\{\phi_n(t)\}$, dadas as duas primeiras funções $\{\phi_0(t)\}$ e $\{\phi_1(t)\}$.

$$< \phi_0(t), \phi_1(t) >= \int_{-\infty}^{\infty} \phi_0(t)\phi_1(t)dt = \int_0^{\frac{1}{2}} dt = \frac{1}{2} \tag{2.20}$$

Como não é ortogonal nós faremos o procedimento de Gram-Schmidt para obter uma base ortonormal a $\{\phi_n(t)\}$:

$$\begin{aligned}
\phi_0^{'}(t) &= \phi_0(t) = \phi(t) \\
\phi_1^{'}(t) &= \phi_1(t) - \frac{<\phi_1(t), \phi_0(t)>}{<\phi_0(t), \phi_0(t)>}\phi_0(t) \\
&= \begin{cases} \frac{1}{2} & 0 \leqslant t < \frac{1}{2}, \\ -\frac{1}{2} & \frac{1}{2} \leqslant t < 1, \\ 0 & \text{Nos outros casos} \end{cases} \\
&= \frac{\psi(t)}{2}
\end{aligned}$$

Podemos continuar fazendo isto para estender tal base, mas por inspeção vemos que $\phi(t)$ tem média $\neq 0$, enquanto que $\psi(t)$ tem média 0. $\phi(t)$ tem dois saltos em t = 0, 1, enquanto que $\psi(t)$ pula em t = 0, $\frac{1}{2}$, 1. Contudo, a potência de $\phi(t)$ é melhor aplicada em baixas frequências, enquanto que a potência de $\psi(t)$ é melhor concentrada nas frequências mais elevadas. De maneira formal, $\phi(t)$ é chamada de função de escala para fazer aproximação e $\psi(t)$ é chamada de função wavelet para achar os detalhes, citadas no início da seção (2.8).

2.8.3 Análise de Multiresolução

Combinando as propriedades dos dois exemplos da seção (2.8.2), temos como construir uma base da função de escala e da *Wavelet* com dois parâmetros: escala e translação, definido assim:

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k) \tag{2.21}$$

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \tag{2.22}$$

em que j $\in \mathbb{Z}$ é o parâmetro de dilatação e k $\in \mathbb{Z}$ é o parâmetro de posição. No geral, queremos que todos os dados tenham resolução desejada, para alguma resolução j, por exemplo. Assim definimos os subespaços:

$$V_j = \text{Span}\{\phi_{j,k}(t)\} \tag{2.23}$$

$$W_j = \text{Span}\{\psi_{j,k}(t)\} \tag{2.24}$$

Para uma definição mais formal, estabelecemos algumas regras:

1. A função de escala é ortogonal às duas translações inteiras;
2. Os subespaços periódicos da função de escala a baixas escalas estão dentro dos gerados pelas escalas maiores. Da figura 2 temos que $\phi_{-1}(t) = \phi_0(t) + \phi_0(t-1)$ que vamos denotar agora como $V_j \subset V_{j+1}$;
3. A única função comum a qualquer V_j é $f(x) = 0$
4. Qualquer função pode ser representada com aproximação arbitrária.

Expandindo a notação de $\phi_{j,k}(t)$, temos que:

$$\phi(t) = \frac{1}{\sqrt{2}}(\sqrt{2}\phi(2t)) + \frac{1}{\sqrt{2}}(\sqrt{2}\phi(2t-1)) \tag{2.25}$$

Que é chamada de equação de refinamento, por conta da eliminação dos termos fora do alcance, em que podemos reescrever assim,

$$\phi(t) = \sum_n h_\phi[n]\sqrt{2}\phi(2t-n) \tag{2.26}$$

Em que temos o filtro passa-baixa: $h_\phi[n] = \{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\}$ para as funções de escala de *Haar*. De maneira similar, temos que:

$$\psi(t) = \sum_n h_\psi[n]\sqrt{2}\phi(2t-n) \tag{2.27}$$

Tal que o filtro passa-alta para a função Haar será: $h_\psi[n] = \{\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\}$. E os dois filtros se relacionam da seguinte maneira:

$$h_\psi[n] = (-1)^n h_\phi[1-n] \tag{2.28}$$

Em que isto serve para realizarmos as separações de todos os sinais separados que formam o sinal original. E podemos esquematizar todas estas transformações e filtros da seguinte maneira:

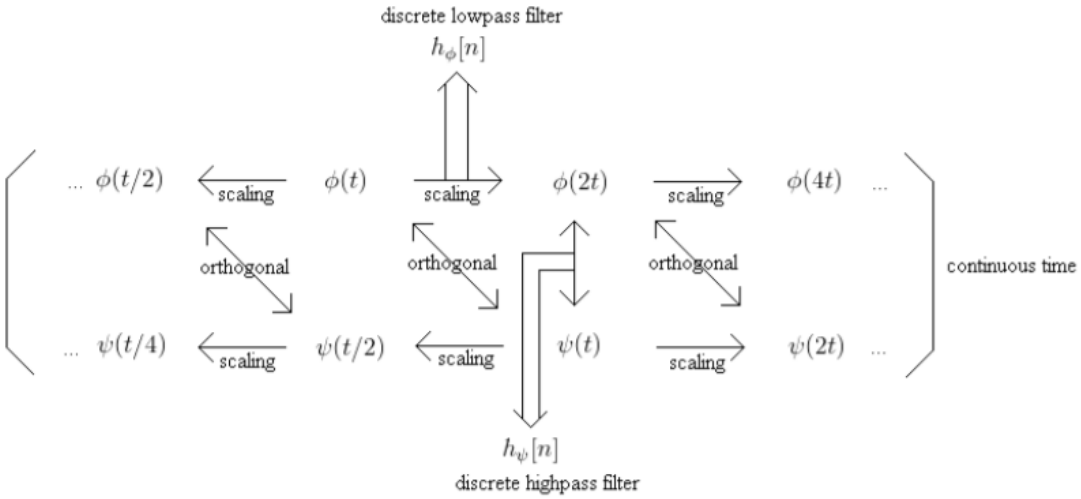

Figura 15 – Relação entre as funções Wavelet e de escala
fonte: Liu, Chun-Lin. (2010). A Tutorial of the Wavelet Transform. (CHUN-LIN, 2010)

Uma vez que definimos o conjunto infinito das Wavelets, tal que é igual ao conjunto dos quadrado integráveis $L^2(\mathrm{R}) = \{f(x)|\int |f(x)|^2 dx < \infty\}$, requisito 4 desta seção, ou formalmente,

$$L^2(\mathbf{R}) = V_0 \oplus W_0 \oplus W_1 \oplus \dots \tag{2.29}$$

E qualquer função pode ser decomposta no $L^2(\mathbf{R})$. E agora podemos relacionar todos os conjuntos e funções com o seguinte diagrama de Venn

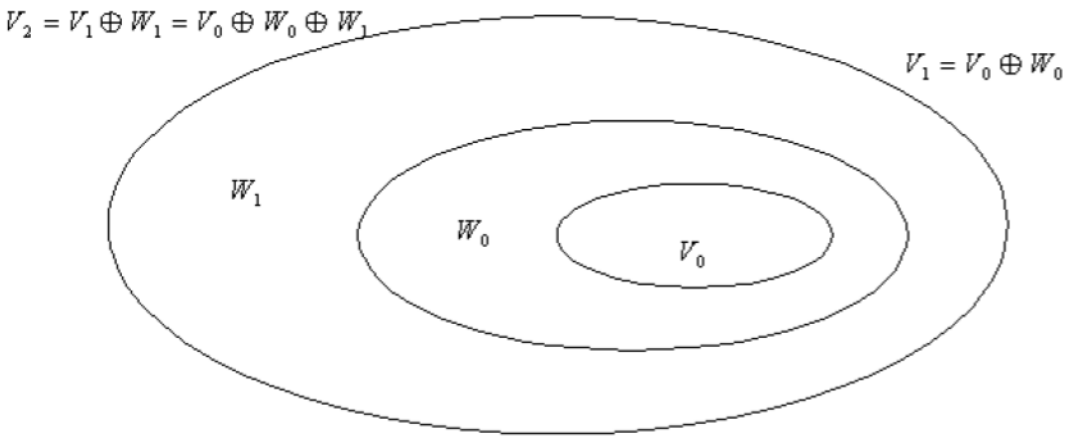

Figura 16 – Relação entre os espaços das funções Wavelet e de escala
fonte: Liu, Chun-Lin. (2010). A Tutorial of the Wavelet Transform. (CHUN-LIN, 2010)

2.8.4 Transformada Wavelet Discreta

Suponha que nós conheçamos as funções de escala e Wavelet. Então podemos aproximar um sinal discreto de $l^2(\mathbf{Z}) = \{f[n] | \sum_{n=-\infty}^{\infty} |f[n]|^2 < \infty\}$ por:

$$f[n] = \frac{1}{\sqrt{M}} \sum_k W_\phi[j_0, k]\phi_{j_0,k}[n] + \frac{1}{\sqrt{M}} \sum_{j=j_0}^{\infty} \sum_k W_\psi[j, k]\psi_{j,k}[n], \tag{2.30}$$

Em que $f[n]$, $\phi_{j_0,k}[n]$ e $\psi_{j,k}[n]$ são funções discretas definidas em [0, M-1]. Pelos conjuntos $\{\phi_{j_0,k}[n]\}_{k\in\mathbf{Z}}$ e $\{\psi_{j,k}[n]\}_{(j,k)\in\mathbf{Z}^2, j\geqslant j_0}$ serem ortogonais entre si, podemos fazer o produto interno e, finalmente, obter os coeficientes Wavelet:

$$W_\phi[j_0, k] = \frac{1}{\sqrt{M}} \sum_n f[n]\phi_{j_0,k}[n] \tag{2.31}$$

$$W_\psi[j, k] = \frac{1}{\sqrt{M}} \sum_n f[n]\psi_{j,k}[n] \qquad j \geqslant j_0, \tag{2.32}$$

Em que as equações indicam, respectivamente, o coeficiente de aproximação e o de detalhe. Assim, a transformação através do corte contínuo em diversos filtros passa-bandas podem ser sintetizadas pelo esquema abaixo (o nome dos filtros foram alternados):

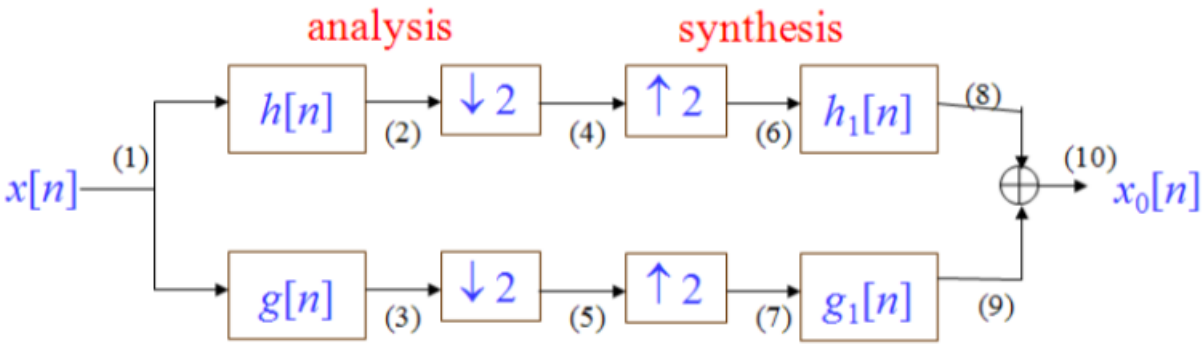

Figura 17 – Esquematização da Transformada Wavelet Discreta
fonte: Liu, Chun-Lin. (2010). A Tutorial of the Wavelet Transform. (CHUN-LIN, 2010)

2.9 Aplicação

Para este projeto, aplicaremos o formalismo apresentado acima da seguinte maneira: Tomaremos um conjunto de dados de ECG do MIT e, primeiro, utilizaremos a função *Wavelet*, encontrada no próprio MATLAB, para remover os ruídos e podermos utilizar os dados menores, outrora ignorados por todos, que é o objetivo deste projeto. Depois aplicaremos um programa de correlação criado por mim também em MATLAB para testarmos se tais dados outrora considerados ruidosos são de fato apenas ruído. (MATLAB, 2018; CHUN-LIN, 2010)

Parte II

Resultados

3 Análise dos Dados: Resultados e Discussão

3.1 MIT-BIH *Arrhythmia Database*

Esta base de dados que está sendo utilizada neste trabalho foi criada por (MOODY; MARK, 2005) em 1975 nos laboratórios BIH (*Beth Israel Hospital*), e gravados pelo MIT (*Massachusetts Institute of Technology*) utilizando de valores de pacientes externos e internos ao hospital, além de arritmias raras, para efeito de análise. Tudo isto no intuito de se criar melhores detectores de arritmia.

É importante ressaltar (para efeito de descobrir possíveis origens de ruído em nossos resultados) que as gravações destes dados foram digitalizadas em 360 amostras por segundo (360Hz) por cada canal, contando com uma precisão de 11-bit ($2^{11} = 2048$ valores possíveis, ou seja, existindo acima de 2048 possíveis valores, **haverá** margem de erro) em uma faixa de valores de 10mV de margem, em que mais de um cardiologista registrou tais valores simultaneamente, discutindo sua tomada de valores afim de diminuir a chance de incluir dados discrepantes.

3.2 Demonstração dos Conhecimentos obtidos

Agora que sabemos de onde virão nossos dados, finalmente dá para mostrar o poder da ferramenta *Wavelet* no processo de filtragem que precisaremos, vamos testá-la analisando um sinal ruidoso qualquer:

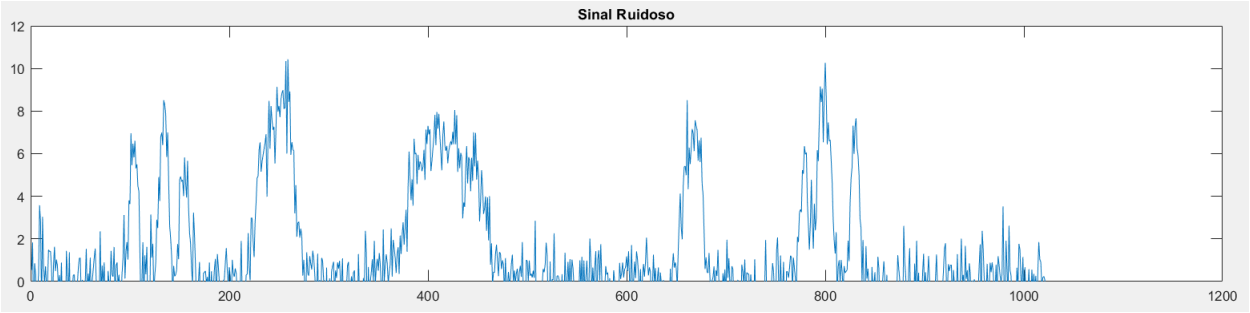

Figura 18 – Função ruidosa a qual queremos remover o ruído, intuito descrito na seção 2.9 importantíssimo para o prosseguimento desta pesquisa

E aplicando a ferramenta *Wavelet* é esta impressionante diferença que obtemos entre o sinal ruidoso que tínhamos e a aproximação que a *Wavelet* proporciona:

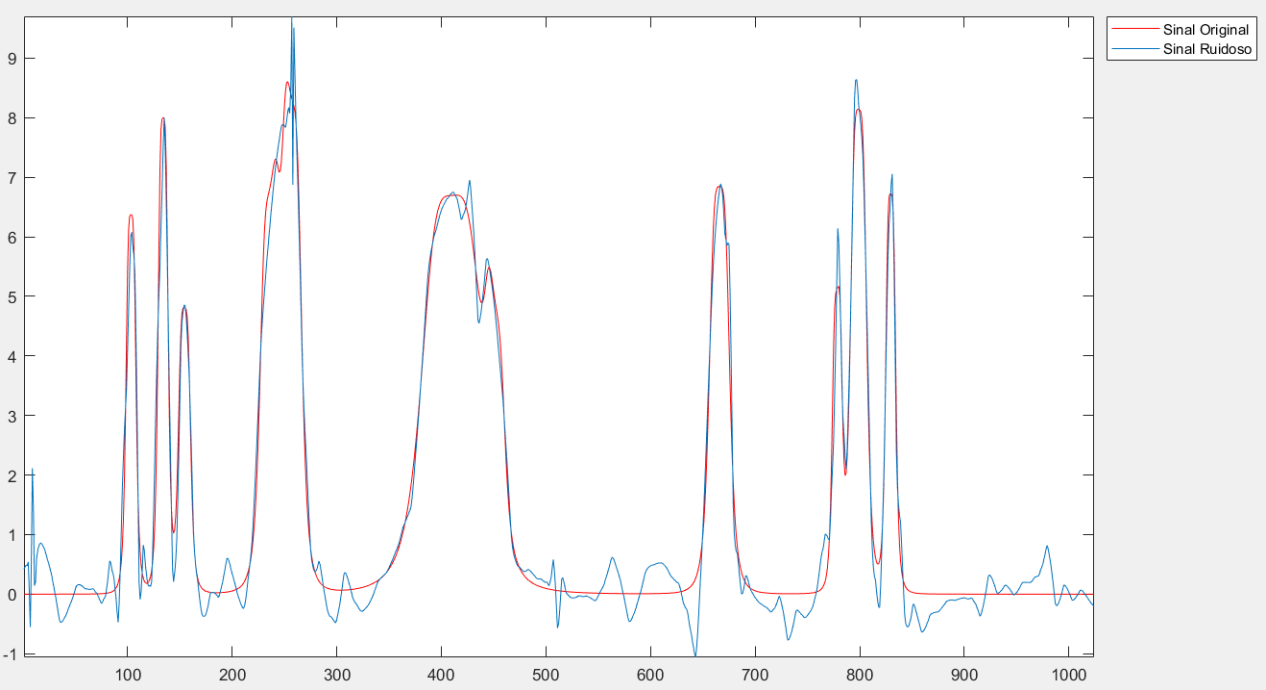

Figura 19 – Comparativo entre a função com ruído e a função com Wavelet aplicado (sem ruído)

Não só o filtro eliminou explosões pontuais dos valores mais altos, como ele simplesmente conseguiu aniquilar os ruídos que existiam em torno dos valores nulos. Restando agora apenas um único ponto a ser tratado antes de adentrar nos valores obtidos da seção (3.4).

3.3 Contagem de Picos

Após a filtragem Wavelet aniquilar os ruídos aleatórios de nosso sistema, estamos finalmente aptos a mensurar o comportamento da correlação dos dados ao longo do tempo, entretanto, como já explicado na subseção (2.7.2): **CORRELAÇÃO NÃO IMPLICA CAUSA**. Entretanto, as causas inerentes à correlação (caso tenham) podem ser indiretas e desconhecidas, então para que possamos categorizar os nossos dados correlacionados iremos precisar calcular o seu comprimento de correlação para que possamos comparar com outros valores. Infelizmente este processo é extremamente demorado e computacionalmente exaustivo, felizmente o trabalho da referência (RAMOS et al., 2011) (**que só é válido para sistemas caóticos**) descobriu uma nova maneira de saber o tempo de decaimento (tempo para que uma correlação caia para 50% do valor original), não pela análise de queda da função correlação, mas sim pela contagem de picos da função original filtrada, então o

que antes era uma computação maçante se tornou uma contagem de pontos. Portanto, nós precisaremos testar quão próxima da realidade a aproximação por contagens de pico é, fazendo o estudo da quantidade de picos de nossa base de dados filtrada (útil para nosso estudo de correlação) e uma maneira de analisar tal densidade de picos é analisando os dados filtrados diretamente e comparando a análise computacional da correlação com o que a referência (RAMOS et al., 2011) estima para tal densidade de máximos, que é dada pela seguinte equação:

$$< \rho >= \frac{1}{6\tau} \tag{3.1}$$

Legenda:
$< \rho >$: Média da densidade de máximos
τ: Tempo de decaimento da correlação

Com tal base estamos finalmente prontos para analisar os resultados da pesquisa.

3.4 Resultados Obtidos

Inicialmente nós utilizamos sinais biológicos de batimentos cardíacos advindos da MIT-BIH *Arrhythmia Database*, tais como este a seguir, para fazer a análise que será base para o nosso objetivo:

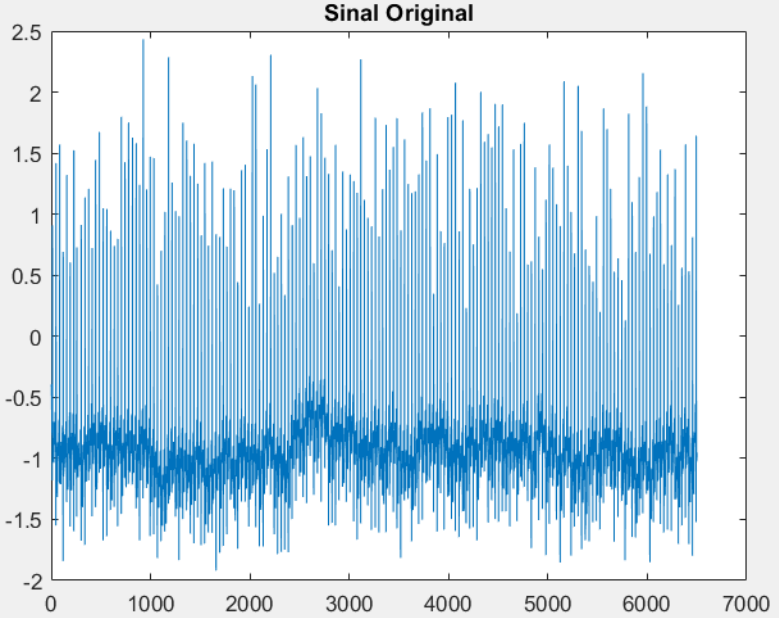

Figura 20 – Recorte do sinal biológico original do *MIT-BIH Arrhythmia Database*, eixo x é a leitura de dados a 360Hz, eixo y é a amplitude do sinal em mV

Em seguida aplicamos a Transformada Wavelet Discreta (subseção 2.8.4) de sexto nível (primeiro nível de aproximação, cinco níveis de discretização) e os resultados obtidos são apresentados abaixo:

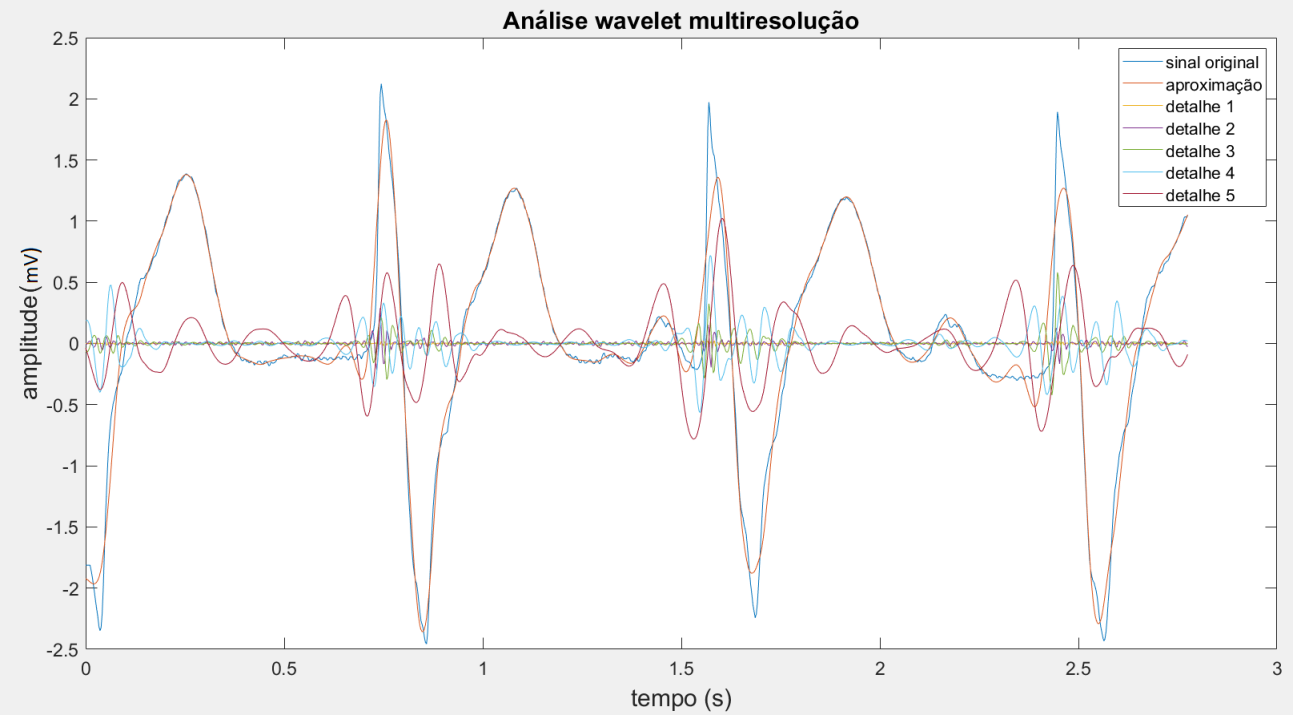

Figura 21 – Coeficientes do recorte do sinal biológico após ter sido aplicado Wavelet Daubechies 6

Vemos dos dados anteriores que a Wavelet fez a separação de ruído com um resultado de acordo com o que a Teoria (seção 2.8) prevê, em cada um dos 6 níveis analisados (1 de aproximação, 5 de discretização).

Então, a partir daqui, faremos uma análise mais profunda. Analisaremos a nossa base de dados (3.1) com algumas das poderosas *Wavelets* de análise multiescala, as dos tipos Biortogonal 8 e *Daubechies* 8 nos seus respectivos 6º, 7º e 8º níveis de discretização, a intenção é descobrir qual destes filtros é o mais eficiente no processo de filtragem dos nossos valores e utilizar suas discretizações para a análise do caos (capítulo 1). Então, é importante frisar que esta análise comparativa está acontecendo porque **mais níveis de discretização NÃO necessariamente implicam melhor leitura dos dados**. Outro ponto que também precisa ser ressaltado é que esta análise comparativa também envolverá a correlação (2.7).Para isto, separaremos várias análises do conjunto de dados em amostras diferentes (de 100 em 100 valores, de 250 em 250 valores e de 325 em 325 valores) para que saibamos qual dessas apurações (além do filtro) é a que carrega menos erros inerentes.

Os resultados da análise acima descrita para cada um dos 47 conjuntos de dados cardíacos presentes na MIT-BIH *Arrhythmia Database* (3.1) estão presentes nos apêndices (A e B), nos quais foram grifados em vermelho os valores de erro que mais se afastaram da média (2.3) observada nas outras densidades de máximos (3.3) do mesmo conjunto (2.2) de nível de discretização dos dados cardíacos.
O intervalo de frequência compreendido por essas decomposições são:

Para a *Wavelet* biortogonal 8:
6º nível: 5,625 Hz - 2,8125 Hz
7º nível: 2,8125 Hz - 1,40625 Hz
8º nível: 1,40625Hz - 0.703125Hz

Para a *Wavelet Daubechies* 8:
6º nível: 3.5117Hz - 1.8759Hz
7º nível: 1.8759Hz - 0.9380Hz
8º nível: 0.9380Hz - 0.4690Hz

Importante lembrar que cada nível de decomposição de uma filtragem *Wavelet* funciona como um filtro passa-banda, onde os detalhes compreendem um intervalo de frequência que vai da metade da frequência original, até 1/4 (um quarto) da frequência. Com isso, foi procurado um intervalo que se aproximasse do sinal que queremos recuperar, que no caso é o da frequência cardíaca, em que considerando batimentos cardíacos entre 60bpm e 100bpm, temos uma frequência entre 1Hz e 1,7 Hz. Com isto procuramos os detalhes que compreendessem a frequência cardíaca no intervalo de frequência mencionado, sendo necessário fazer diferença de escala na *Daubechies* 8, pois sua frequência central é de 0.6667Hz, enquanto que a Biortogonal não necessita de ajuste, uma vez que sua frequência central é de $1.0008 \approx 1$, imersa no intervalo cardíaco.

Olhando estes resultados dos apêndices A e B, podemos ver que dentre as duas *Wavelets* a biortogonal apresentou menos explosões pontuais nas densidades de máximos, mas um outro ponto muito importante que também se destaca é que as únicas análises que **não** apresentaram explosões fizeram correlação a cada 100 amostras (muito provavelmente se deve a não conter dados do próximo batimento cardíaco, o que não deve ocorrer com cortes maiores), tanto que seus valores mais altos (grifados em amarelo nos apêndices A e B) não chegam nem perto das explosões presentes em outras densidades de máximos. Isto pode indicar principalmente que a correlação é falha para 250 e 325 dados pois eles podem ser tão longos que já estão guardando informação do próximo batimento cardíaco, fazendo com que a correlação cresça ao invés de diminuir com o tempo.

Por fim, utilizamos estes valores dos apêndices A e B e fizemos um comparativo de erro entre a análise destes dados pela conta explícita do tempo de decaimento da

correlação por análise gráfica complexa e a análise do tempo de decaimento pela contagem de máximos (seção 3.3) para toda a base de dados em cada *Wavelet* utilizada, o erro resultante entre um método e outro está exposto nas tabelas a seguir:

Tabela 3 – Médias e desvios padrões dos erros obtidos na comparação dos métodos de tempo de decaimento de correlação para toda base de dados (MOODY; MARK, 2005) com uso da *Wavelet* Biortogonal 8 (dados do apêndice A).

Wavelet TYPE-samples	Média dos erros (%)	Desvio padrão dos erros (%)
D6-100	4.05	3.01
D6-250	8.63	18.13
D6-325	12.76	25.66
D7-100	5.25	3.53
D7-250	10.89	20.28
D7-325	13.06	25.57
D8-100	3.66	3.15
D8-250	9.67	19.37
D8-325	14.72	26.19

Tabela 4 – Médias e desvios padrões dos erros obtidos na comparação dos métodos de tempo de decaimento de correlação para toda base de dados (MOODY; MARK, 2005) com uso da *Wavelet Daubechies* 8 (dados do apêndice B).

Wavelet TYPE-samples	Média dos erros (%)	Desvio padrão dos erros (%)
D6-100	15.15	18.02
D6-250	21.73	27.71
D6-325	31.36	36.55
D7-100	8.99	5.24
D7-250	35.72	35.02
D7-325	35.02	34.80
D8-100	3.77	3.57
D8-250	17.26	27.26
D8-325	40.90	36.56

Confirmando nossa análise anterior da correlação, em que vemos que principalmente na *Wavelet* Biortogonal 8 por 100 amostras nós temos erros mínimos entre a aproximação da contagem direta da densidade de máximos e a análise computacional do tempo de correlação em cada um dos níveis de discretização analisados, além disto, são valores **realmente** próximos no geral, sem muitas explosões pontuais, visto que o desvio padrão (2.4) **também** é baixo. Assim como agora é visível como o erro cresce com o crescimento das amostras, indicando o viés de repetição destes dados anteriormente discutido.

3.5 Erros Também Ensinam

Os filtros *Wavelet* biortogonal e *Daubechies* não foram os únicos testados para este TCC, outros dois foram testados e seus resultados não foram satisfatórios pelas seguintes razões:

- ***Wavelet Haar***: A Wavelet Haar, ou Daubechies 1, possui uma filtragem muito boa (CHUN-LIN, 2010), porém peca em um detalhe: Ela é para conjuntos de sinais bem específicos. A sua aplicabilidade ideal ocorre em sinais com queda abrupta de valores, tais como funções com aspecto degrau, um bom exemplo são as figuras (14), infelizmente nosso sinal cardíaco não se enquadrou nestes limites e a filtragem deixou a desejar.

- **Filtro *Bézier***: A primeira abordagem utilizada neste TCC foi com o filtro *Bézier*, porém depois de algum tempo ele se mostrou muito impreciso, visto que ao aplicá-lo tínhamos que regular manualmente o filtro até quando nós julgássemos que o sinal estava de acordo com o que queríamos ver, e é exatamente aí que mora o problema, isto poderia acabar causando viés cognitivo, de alguma natureza, em "querer ver" um determinado resultado, e como a natureza não é enviesada, este filtro se mostrou ineficiente.

4 Conclusão

4.1 Conclusões

A análise dos dados obtidos pela MIT-BIH *Arrhythmia Database* (3.1) indicou que existem diferenças entre os processos de filtragem da correlação dos dados cardíacos dos pacientes, o filtro *Wavelet Daubechies* (mais comum em análises de dados) se mostrou eficiente, porém com mais explosões pontuais (2.4) do que o filtro *Wavelet* Biortogonal, muito provavelmente isto se deve ao fato citado em (3.4) de que foi necessário fazer ajuste de escala para a *Daubechies* se enquadrar na faixa de frequência cardíaca, enquanto que a Biortogonal já estava imersa sem nenhum tratamento estatístico necessário.

Além disto também ficou explícito nos dados como a análise por amostragem de 100 em 100 valores se mostrou a mais precisa, em contraste com as de 250 e 325, que só cresceram a correlação (2.7) e o erro (duas coisas que se esperavam diminuir em suas respectivas partes). Possivelmente, isto deve ter ocorrido por conta destes conjuntos serem tão grandes que estariam absorvendo valores de batimentos seguintes e diminuindo o comprimento de correlação, ao invés de aumentar.

Ao final, analisamos a possibilidade de fazer análise de comprimento de correlação através da contagem de máximos (3.3), (RAMOS et al., 2011) e para testar sua validade comparamos com a análise computacional direta (procedimento muito mais demorado) e obtivemos resultados (3.4), comprovando o baixíssimo erro médio (2.3) com valores sem explosões pontuais (2.4) do método da contagem de máximos, ou seja, agora não só temos um método de análise de dados comprovadamente funcional, como ele também é rápido.

E isto tudo foi possível graças ao arcabouço teórico formado até aqui, não apenas por todo o estudo estatístico, como visto no capítulo (2), mas também pela quantidade de ferramentas computacionais que tive que desenvolver familiaridade para construir esta monografia, tais como:

- **Linguagem C e Gnuplot:** Utilizadas para montagem de executáveis que foram úteis para organização dos dados antes de analisá-los em MATLAB (MATLAB, 2018); além de gerar gráficos (tais como as figuras 8, 9 e 11) que me permitiram desenvolver melhor a didática da parte estatística para que sua leitura desta monografia pudesse fluir sem maiores empecilhos.

- **MATLAB:** Aqui (3.4), (MATLAB, 2018) foi onde toda a análise de sinais aconteceu, em que pude aplicar todo o conhecimento estatístico, principalmente a parte de

Wavelet (2.8) e Correlação (2.7) e, assim, finalmente poder analisar de forma minuciosa como os dados cardíacos se comportam.

- **LaTeX**: E por último e não menos importante, a ferramenta de edição textual LaTeX, na qual **todos** estes conhecimentos adquiridos puderam ser condensados e agora estão sendo transmitidos para você que está lendo.

Por fim, como perspectiva, é importantíssimo destacar a futura possibilidade de utilização desta atual análise para a criação de um grande método de detecção analítica preventiva para qualquer sistema biológico, no qual se poderia gerar números universais de fácil diagnóstico que revelariam problemas caóticos desde sua origem, antes mesmo de atingirem condições macroscópicas, e tais análises não estariam limitadas apenas a dados cardíacos, detecção de epilepsia (OSORIO; WILKINSON, 1998; ACHARYA et al., 2013; IASEMIDIS et al., 2004) é um outro exemplo bem próximo de algo que pode ser detectado a partir da análise de dados discretos outrora somente ignorados, mas secretamente inundados em dados caóticos, por tudo isto exposto que com felicidade digo que o desdobramento desta pesquisa só está se iniciando.

Referências

ACHARYA, U. R. et al. fautomated eeg analysis of epilepsy: A review. *Knowledge-Based Systems*, v. 45, p. 147–165, 2013. Citado na página 66.

BROCHADO, S. Por que a previsão dos meteorologistas erra tantas vezes? *SUPERINTERESSANTE, São Paulo*, v. 30, Nov 2017. Disponível em: <https://encurtador.com.br/bdwW4>. Citado na página 23.

CHUN-LIN, L. [S.l.]: "A Tutorial of the Wavelet Transform.", 2010. Citado 5 vezes nas páginas 13, 52, 53, 54 e 63.

ESTRANHO, R. M. O que é a teoria do caos? *Mundo Estranho, São Paulo*, v. 18, Abril 2011. Disponível em: <https://super.abril.com.br/mundo-estranho/o-que-e-a-teoria-do-caos>. Citado na página 23.

GUYTON, A. C.; HALL, J. E. *Tratado de Fisiologia Médica.* 11ªed. ed. Rio de Janeiro: Elsevier, 2006. Citado 2 vezes nas páginas 24 e 25.

IASEMIDIS, L. D. et al. *Dynamical resetting of the human brain at epileptic seizures: application of nonlinear dynamics and global optimization techniques, IEEE Trans.* [S.l.]: Eng. 51 (3, 2004. 493-506 p. Citado na página 66.

MASCENA, J. H. A.; RAMOS, J. G. G. S. *Sistemas dinâmicos e caos: do clássico ao quântico.* Trabalho de Conclusão de Curso (Bacharelado), Nov 2021. Disponível em: <https://repositorio.ufpb.br/jspui/handle/123456789/21675>. Citado na página 23.

MATLAB. 2018. Citado 3 vezes nas páginas 25, 54 e 65.

MOODY, G.; MARK, R. Mit-bih arrhythmia database. *PhysioNet, EUA*, Fev 2005. Disponível em: <https://doi.org/10.13026/C2F305>. Citado 3 vezes nas páginas 15, 57 e 62.

OSORIO, M. G. F.; WILKINSON, S. B. Real-time automated detection and quantitative analysis of seizures and short-term prediction of clinical onset. *Epilepsia*, v. 39, n. 6, p. 615–627, 1998. Citado na página 66.

PRESS, C. U. *Cambridge Latin Course.* 5ª. ed. Reino Unido: Cambridge University Press, 2022. Citado na página 33.

RAMOS, J. G. G. S. et al. Conductance peaks in open quantum dots. *Physical Review Letters*, v. 107, 2011. Citado 4 vezes nas páginas 25, 58, 59 e 65.

Apêndices

APÊNDICE A – Valores de análise utilizando Wavelet Biortogonal

A imagem completa foi colocada na página seguinte com o intuito de garantir uma melhor visualização.

	A	B	C	D	E	F	G	H	I
1	D6 – 100	D6 – 250	D6 – 325	D7 – 100	D7 – 250	D7 – 325	D8 – 100	D8 – 250	D8 – 325
2	3,36880127	3,36698545	0,21199093	4,29499027	3,01401068	4,6042378	0,70512344	1,20107596	74,29621845
3	0,4284731	3,19772441	1,00037882	0,05657108	4,80225748	0,04173535	0,41549793	6,01000670	0,83802541
4	1,18413797	0,15001328	1,75724604	0,12780637	0,74641893	3,30367471	4,36809759	3,98108549	4,78080126
5	10,8466912	5,57545659	10,0954389	1,69904500	1,53169936	2,14319244	4,32309097	4,17527413	8,29603765
6	3,08309104	4,19347119	2,30194234	7,13033976	8,37006481	6,53785627	0,26311809	0,02562465	0,68061165
7	3,02862154	5,58241947	0,34295542	14,9094712	17,5498099	12,8102157	15,6856909	12,3928192	8,25207405
8	1,58294591	2,70425321	0,84909970	8,11247244	5,18601378	6,41542561	1,20532350	0,73164808	5,40049935
9	1,05363857	1,67541116	1,02738466	2,74821984	2,49447294	3,62763159	0,01671907	0,57720978	0,88365268
10	3,40681446	4,54470313	2,615581	1,66657730	2,22325357	0,25187926	8,91264517	5,03593466	10,8911455
11	7,48445943	91,02986801	88,5122358	0,76645090	89,4607744	89,7491464	1,22897946	74,0710850	37,5107795
12	6,37307061	8,29283471	57,3514918	2,91598467	1,69372768	3,39871267	5,51441998	4,29511685	7,90482108
13	9,07658703	27,3318700	95,2713337	6,84233452	62,9480619	68,9715911	3,56946522	3,34871946	81,5838816
14	1,35719374	1,55760881	0,68623800	5,33671328	3,11098323	3,09543647	6,77562851	7,98367960	5,48926026
15	9,55717089	6,45758270	4,15753350	8,4170009	5,65201209	2,73607404	5,1564154	6,87984072	5,16405949
16	2,12432341	2,25198666	0,52340566	1,32635461	2,65582331	3,32121961	4,26072699	3,97244719	5,97389592
17	5,42474930	2,08345912	8,72574269	11,612223	14,5281239	10,2903341	9,03469583	9,22926719	10,075557
18	5,05388818	6,41830627	5,08258192	0,49029175	0,8342093	1,46758220	2,62863077	2,23832004	3,43795168
19	2,51901263	0,81127897	1,76279806	12,2249618	8,89773555	1,51542685	3,14220018	46,617778	82,9759713
20	3,13579730	0,22377426	2,60796290	2,91927137	3,02849417	3,80274732	0,20299093	2,45720723	0,59252917
21	8,06637282	10,2449108	11,645434	5,77894517	3,09457153	4,8141433	5,86163417	5,860378	4,76427346
22	2,97167644	3,94130341	89,3927665	9,38712356	53,3796972	88,1842191	2,47866405	79,8108858	81,1523066
23	3,49272673	1,71717294	4,19106138	10,170763	8,15788492	8,94837047	10,661719	8,84874305	9,73351548
24	13,0912385	14,0233170	19,5069333	0,54423132	1,84324236	0,15273690	3,9153974	6,90017647	3,71566209
25	1,84145776	2,55628432	4,01558287	5,97166707	7,263316	4,03413335	5,34703988	6,0951272	7,362740
26	2,08403910	4,89386568	1,32915395	1,84717695	3,49380123	6,25669428	3,76593951	0,04656011	5,345131
27	3,83536091	5,88585803	7,35409496	4,81111257	4,45419226	6,17931349	3,13851108	4,4705092	4,98136037
28	3,91090148	3,42315243	1,85219132	6,35135805	7,63804015	3,42326751	3,86269218	5,50028495	0,0117650
29	1,65868003	2,42522291	4,85709803	9,77856659	8,32191339	5,69285872	2,38700072	1,78513955	1,3583368
30	2,87691967	4,40791890	0,16674470	6,57162206	2,35687780	4,38451669	2,09158387	2,52726323	0,87060066
31	8,05125479	6,84135968	8,40356259	5,19468409	2,28810601	4,17310557	3,13508274	3,41735352	6,31097855
32	3,71918657	5,06487387	3,67330858	4,2592590	6,12752586	7,00307161	4,58195085	4,21392321	3,07537683
33	5,4805797	3,9563832	5,27319334	9,53772866	8,21216936	8,30614825	2,94710193	0,36920793	0,71755216
34	3,58803995	2,87527715	3,73360914	2,85386724	3,91094942	3,39768532	0,86223918	1,27237257	0,0254652
35	3,30997708	91,4767912	91,6183908	6,72760497	4,64313695	6,15273546	8,45104316	8,25470964	8,97313479
36	2,68603536	7,32951594	1,98734515	7,06260474	8,39213346	6,08351716	0,92390649	2,55907572	81,3853567
37	5,97036346	8,17381504	2,76800818	2,40749612	0,70561465	1,28002040	0,81059233	0,93568406	1,98757628
38	0,35771508	4,41453621	6,43940189	4,47757158	3,85060077	0,35966249	0,52036054	0,33150601	0,67148841
39	1,65689550	4,49036992	5,86503157	7,63758325	8,46627721	4,50453321	0,49654173	1,06404252	1,41820376
40	4,44941647	7,48677640	2,04880111	2,29249038	2,63737658	4,83407569	0,54586604	0,7330156	0,99889168
41	0,56400078	0,18927664	3,81391743	8,29182197	5,47821770	6,65840528	4,13519450	3,26369158	3,11720435
42	0,58827828	3,197105	1,68992420	6,46458581	5,3659468	0,87284592	6,18861731	7,90013409	7,13987156
43	2,65472084	0,87978079	1,82759221	6,33041807	7,58745129	90,0564972	2,92606464	3,51922700	2,40272846
44	0,01626068	2,86071708	7,26433115	2,57025275	4,44076675	1,54069056	2,2343216	3,37463971	1,91647184
45	2,48102749	4,52758606	3,6243017	4,07016728	0,32276499	4,34549136	4,7555725	4,65422220	6,13990574
46	6,33528983	6,99627243	10,3853849	0,20751144	89,7009519	90,549253	5,44125271	9,77097739	5,67213663
47	4,84885275	5,24952655	4,17534489	3,24865243	2,07592375	1,87184122	2,05645328	1,02299574	2,19012870
48	9,64249046	8,51091106	5,82151080	8,28828552	9,07709155	11,7765241	0,08266348	80,8593556	83,4066425

Figura 22 – Erro percentual obtido na aplicação da *Wavelet* Biortogonal 8 na base de dados, comparando os dois métodos de obtenção do tempo de correlação

APÊNDICE B – Valores de análise usando *Wavelet Daubechies* de 6º, 7º e 8º níveis de discretização

A imagem completa foi colocada na página seguinte com o intuito de garantir uma melhor visualização.

	A	B	C	D	E	F	G	H	I
1	**D6 – 100**	**D6 – 250**	**D6 – 325**	**D7 – 100**	**D7 – 250**	**D7 – 325**	**D8 – 100**	**D8 – 250**	**D8 – 325**
2	6,92379176	4,87131080	5,16101187	8,87831378	7,11058106	8,28073609	1,31972932	79,83371925	82,74910063
3	7,37732995	8,13510578	4,62997065	4,15400125	84,53453132	85,17066680	1,54135208	9,35323259	0,07258535
4	9,34650169	8,02255025	11,13516452	2,43315707	61,14102418	87,78945715	2,04214823	2,74160326	84,18578675
5	11,02353706	8,98681063	96,02042747	0,59375450	85,21078122	86,69631073	0,35848427	2,15545772	81,82315712
6	11,90650156	13,75319972	11,21411661	5,52328363	4,66239086	6,13002916	0,68738201	0,40751209	82,88812273
7	6,44062357	11,27834494	94,71341346	18,4468156	14,82934695	20,25978497	16,99435425	19,65021352	20,31595479
8	6,06235092	9,23937698	6,87920692	12,76017862	13,48502919	9,31822534	2,92382773	0,63197759	2,30476905
9	9,46733844	10,4814653	95,71080669	1,03981040	83,03599538	81,97908	2,68765644	2,70414145	78,41039627
10	8,28746423	7,67839034	5,87064564	5,10718130	6,63764893	7,19395414	4,71811184	4,67963146	2,93578317
11	12,8591439	94,76042345	94,863049	5,82559693	71,103463	91,28581414	1,00386544	28,10489448	83,4964758
12	85,72285281	86,07077884	93,94071467	3,43686941	4,33596829	48,03107555	3,74571539	2,23111620	7,03327841
13	12,97520149	94,70616525	95,39844502	20,92195727	87,44773051	89,64724359	0,78344327	79,28449195	11,4363366
14	12,96767576	15,62147873	8,67289803	7,01415958	6,64769141	4,82632171	1,63376477	1,93175348	50,39162589
15	8,93301577	13,82280655	8,81246572	5,04090410	74,79981185	69,05939404	0,01967922	0,25549119	82,32874083
16	10,2547404	9,82586707	9,72368037	5,56441100	5,16977265	4,13432360	0,93834596	0,20504183	78,6747151
17	11,21087538	12,0225853	9,98607204	15,10929723	17,33741853	13,76083432	6,33395266	8,19136295	7,08816541
18	12,56551378	13,09670016	12,1178613	17,00847988	18,70175639	15,38778441	0,60407451	0,73240410	0,28740114
19	77,75032157	83,68801699	83,64333497	15,68833597	15,20920639	15,66095838	1,52627078	80,57813462	52,15574136
20	6,68238357	5,41783458	3,39919718	13,1937089	12,92298705	12,86223844	6,88438851	6,62791503	6,94621283
21	12,12535038	12,00621869	12,37484124	11,304696	11,16475733	11,93809635	0,77882655	1,76999542	81,40288775
22	9,09119715	9,64280922	83,17855999	13,52447923	51,67971759	82,58666783	2,03019906	49,57928718	81,33837157
23	12,38177381	11,95656889	12,08097011	8,52642767	8,81042807	8,50443921	6,07167042	7,49469493	4,25647588
24	8,59357408	9,43768399	9,67124485	13,02141273	14,25951416	14,14600556	0,42411458	0,94124039	1,19463186
25	12,46474624	12,39646215	17,53232427	13,68614907	10,46686058	12,71759847	9,00219441	9,51503513	10,56085482
26	10,27232731	10,80756665	6,59473825	11,68086908	12,77520234	12,93743419	6,79834547	6,06419731	7,60764074
27	10,16549104	10,8707494	10,92462948	9,20350952	8,50743929	11,67827667	5,60412372	6,49057290	5,28811072
28	16,12226325	14,83371223	16,42510248	7,17643206	11,41687249	4,55729559	7,20712674	12,36351113	4,00626651
29	16,27026395	16,94715608	11,77782947	11,30727521	9,36277721	16,30127074	5,39832704	1,32891820	58,31424322
30	17,23295834	16,01606887	18,27634798	5,17355907	5,05111508	3,97451327	1,1101314	0,97651785	82,13012059
31	12,2715486	12,07086401	11,26787826	11,09672697	10,87794399	8,24690208	5,14143570	5,68450948	7,58894794
32	7,79488986	6,75419255	8,14664172	14,60886157	12,73755295	17,25264394	9,41829289	9,53994740	10,18916176
33	10,14143139	10,21083147	11,84080127	9,58073899	11,84596819	8,64948249	4,08502881	7,27808310	5,21167815
34	10,27890897	9,48520358	10,56690917	11,30820046	83,06388465	89,17919346	3,44880557	3,13393125	83,16557577
35	83,93625572	90,81922602	91,92858348	0,60818239	84,83118891	3,47396773	10,0531437	10,4115071	8,21340982
36	13,14762627	11,87086949	15,70786385	6,06503796	7,43174393	8,68371158	1,28452707	71,05736125	84,79699603
37	9,20441931	11,81084028	8,66376822	2,47614805	85,44268532	82,68844098	1,94165022	0,67417168	1,04764954
38	8,6003526	5,17782082	5,83715613	5,13521199	5,66127716	2,24970174	1,17120440	0,00156066	78,94367782
39	7,37839074	7,68326721	8,86877000	11,75093929	12,65019912	8,90995487	0,43895787	81,11155983	83,82895879
40	10,2510952	12,05341311	95,92808783	1,60968252	85,22509764	89,61181944	0,50977603	1,15785307	47,62524086
41	9,61718214	10,92354337	10,9080302	11,85781985	12,16185153	7,19262598	9,23652019	9,98355448	9,23176859
42	3,59293704	1,30845776	5,80329967	6,83509111	2,81845440	4,23418429	9,01143217	8,60368283	9,39285333
43	16,71507552	66,22286174	92,97668178	11,70332974	90,09095669	90,44106219	6,01018535	3,95004917	7,03459210
44	8,49178636	7,47004793	11,16018145	11,00464245	9,61036846	12,96997663	0,98498324	1,62113500	84,13713002
45	9,57423256	9,60003864	13,42978919	4,98109613	81,2933233	60,41053716	2,78792927	3,04824828	82,33152576
46	15,73793402	88,08492105	95,09698529	4,28072657	90,18023786	66,38464554	1,17671647	81,30936929	82,62659276
47	12,06720166	9,98693650	12,05641494	4,7204263	85,51292306	85,51140574	7,67376404	8,91457085	7,35933666
48	9,99503782	9,54009779	11,88418249	20,80447199	89,25010064	62,89657152	1,55193097	77,08551845	77,84909491

Figura 23 – Erro percentual obtido na aplicação da *Wavelet Daubechies* 8 na base de dados, comparando os dois métodos de obtenção do tempo de correlação