VOZES DIVERSAS
DIFERENTES SABERES

Salão UFRGS 2018 — Conhecimento Formação Inovação

SALÃO DE INICIACAO CIENTIFICA XXX SIC

15 A 19 OUTUBRO CAMPUS DO VALE

UFRGS PROPESQ

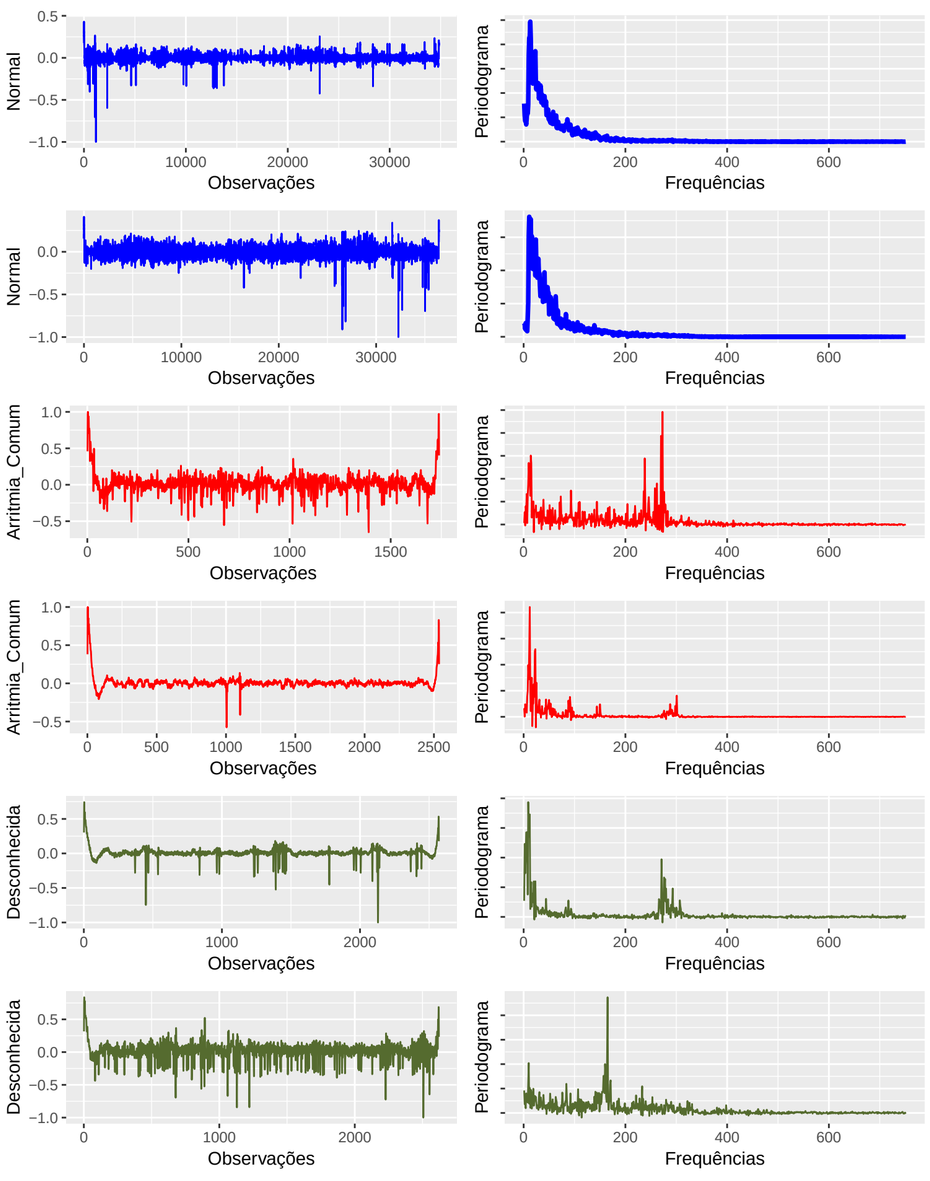
Inferência Estatística para Classificação de Doenças Cardíacas

Autora: Mikaela Baldasso
Orientador: Marcio Valk

Sinais cardíacos

No estudo de doenças cardiácas é comum a coleta de dados como Eletrocardiogramas (ECG) e Fonocardiogramas (PCG) que fornecem importantes informaões a respeito de possíveis denças patológicas.
Esses dados podem ser vistos como séries temporais, em que a técnica de classificação e agrupamento baseada em U-estatísticas pode ser aplicada, possibilitando-nos mensurar a confiabilidade de um método de diagnóstico.

Exemplo: Sinais cardíacos com arritmia e sem arritmia com seus respectivos periodogramas que são tranformações dos dados usadas na busca por padrões.



Métodos de agrupamento para identificar padrões

O método de *clustering* é um conjunto de técnicas computacionais cujo propósito consiste em separar objetos em grupos distintos de acordo com as características que eles apresentam. De forma geral, a técnica consiste em colocar elementos similares em um mesmo grupo de acordo com algum critério já estipulado.

Uhclust - Método baseado em U-estatísticas

Seja uma amostra $X = (X_1, ..., X_n)$ de n vetores L-dimensionais (sinais cardíacos, por exemplo) dividida em dois grupos G_1 e G_2 de tamanhos n_1 e n_2 respectivamente, onde $n = n_1 + n_2$. Uma medida de dissimilaridade dos grupos é dada por

$$B_n = \frac{n_1 n_2}{n(n-1)}(2U^{(1,2)}_{n_1 n_2} - U^{(1)}_{n_1} - U^{(2)}_{n_2})$$

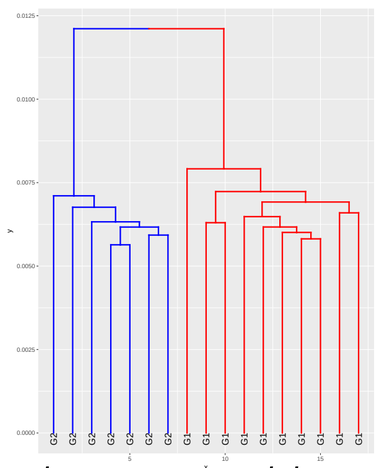
em que $U^{(1,2)}_{n_1,n_2}$ é uma U-estatística que mede a distância entre grupos e $U^{(1)}_{n_1}$ e $U^{(2)}_{n_2}$ medem as distâncias dentro dos grupos 1 e 2, respectivamente. É sabido que B_n segue uma distribuição Normal e que, sob H_0 (não existe separação entre os grupos), o valor esperado de B_n é zero.

Simulações de Monte Carlo

Nesse estudo utilizamos os processos autorregressivos de oredem 1 (AR(1)) para gerar os grupos. O processo é definido por $Y_t = \phi y_{t-1} + \varepsilon_t$, em que o parâmetro ϕ deve satisfazer $|\phi| < 1$ e ε_t é um ruído branco gaussiano.
Na Tabela 1, as $n_1 = 10$ séries temporais que compõem o grupo 1 (G_1) são geradas com $\phi = 0.3$, conforme coluna do ϕ_1, e as $n_2 = 7$ séries que compõem o grupo 2 (G2) são geradas a partir de diferentes valores para ϕ, conforme a coluna do ϕ_2.

$n_1 = 10$ e $n_2 = 7$				
ϕ_1	ϕ_2	Poder	ARI hclust	ARI uhclust
0.30	−0.20	1.00	0.99	1.00
0.30	−0.10	1.00	0.77	0.99
0.30	0.00	0.97	0.46	0.88
0.30	0.10	0.24	0.11	0.42
0.30	0.20	0.05	0.02	0.06
0.30	0.30	0.04		
0.30	0.40	0.08	0.02	0.06
0.30	0.50	0.53	0.17	0.61
0.30	0.70	1.00	0.99	1

Tabela 1: *Proporção de rejeição (poder) do uhclust e ARI do uhclust e hclust*



Dendrograma para dois grupos

Resultados

Durante a realização do presente trabalho, exploramos vários bancos de dados de diferentes fontes e caracaterísticas, e neles aplicamos diversas tranformações na busca por padrões. Simulações de Monte Carlo foram realisadas em um contexto controlado e sugerem que o método *uhclust* pode ser usado para caracterizar sinais com dinâmicas diferentes desde que a métrica correta seja utilizada. Os próximos passos serão na direção da aplicação a dados reais.

Referências

[1] Valk, Marcio, and Gabriela Bettella Cybis. ”U-statistical inference for hierarchical clustering.”arXiv preprint arXiv:1805.12179 (2018).