Juliano Capelo Nagy

Aplicação de análise multivariada em detecção de disparos de neurônios

Natal - RN

2 de setembro de 2021

Juliano Capelo Nagy

Aplicação de análise multivariada em detecção de disparos de neurônios

Monografia de Graduação apresentada ao Departamento de Estatística do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Universidade Federal do Rio Grande do Norte

Centro de Ciências Exatas e da Terra

Departamento de Estatística

Orientador: Bruno Monte de Castro

Natal - RN

2 de setembro de 2021

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Ronaldo Xavier de Arruda - CCET

Nagy, Juliano Capelo.
Aplicação de análise multivariada em detecção de disparos de neurônios / Juliano Capelo Nagy. - 2021.
35f.: il.

Monografia (bacharelado) - Universidade Federal do Rio Grande do Norte, Centro de Ciências Exatas e da Terra, Departamento de Estatística. Natal, 2021.
Orientador: Prof. Dr. Bruno Monte de Castro.

1. Estatística - Monografia. 2. Atividade neural - Monografia. 3. Desvio mediano absoluto - Monografia. 4. Análise de componentes principais - Monografia. 5. Agrupamento - Monografia. I. Castro, Bruno Monte de. II. Título.

RN/UF/CCET CDU 519.2

Elaborado por Joseneide Ferreira Dantas - CRB-15/324

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

ATA DE DEFESA DE MONOGRAFIA
BACHARELADO EM ESTATÍSTICA

Às 14h30min do dia 02 de setembro de 2021, por meio da ferramenta de transmissão na plataforma GoogleMeet, compareceu para defesa da monografia do curso de graduação em Estatística o aluno *Juliano Capelo Nagy* tendo como Título da Monografia "Aplicação de análise multivariada em detecção de disparos de neurônios". Constituiu a Banca Examinadora os professores Bruno Monte de Castro (presidente), Antonio Marcos Batista do Nascimento (examinador) e Eliardo Guimarães da Costa (examinador). Após a apresentação e as observações dos membros da banca avaliadora, ficou definido que o trabalho foi considerado aprovado.

_____
Prof. Dr. Bruno Monte de Castro (DEST/UFRN)
Orientador

_____
Prof. Dr. Antonio Marcos Batista do Nascimento (DEST/UFRN)
Examinador

_____
Prof. Dr. Eliardo Guimarães da Costa (DEST/UFRN)
Examinador

Emitido em 09/09/2021

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO Nº 1/2021 - EST/CCET (12.02)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 09/09/2021 09:56)
ANTONIO MARCOS BATISTA DO NASCIMENTO
PROFESSOR DO MAGISTERIO SUPERIOR
EST/CCET (12.02)
Matrícula: 1048587

(Assinado digitalmente em 09/09/2021 09:44)
BRUNO MONTE DE CASTRO
PROFESSOR DO MAGISTERIO SUPERIOR
EST/CCET (12.02)
Matrícula: 2354162

(Assinado digitalmente em 09/09/2021 12:00)
ELIARDO GUIMARAES DA COSTA
PROFESSOR DO MAGISTERIO SUPERIOR
EST/CCET (12.02)
Matrícula: 3010614

Para verificar a autenticidade deste documento entre em https://sipac.ufrn.br/documentos/ informando seu número: **1**, ano: **2021**, tipo: **ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO**, data de emissão: **09/09/2021** e o código de verificação: **e427dc9150**

Dedico este trabalho ao meu pai, mãe e irmã, que me deram apoio e me incentivaram a chegar até aqui.

Agradecimentos

Meus agradecimentos vão para meu pai, minha mãe e irmã, João Carlos Nagy, Elizabeth Capelo Nagy e Aline Capelo Nagy Baccarin, por me darem todo o apoio possível desde sempre.

À UFRN e ao Departamento de Estatística por oferecerem a estrutura necessária para as aulas e os estudos e também ao LEA por permitir uma primeira experiência fazendo consultorias para pessoas de fora da área da Estatística.

Ao professor Dr. Bruno Monte de Castro por me orientar durante o desenvolvimento deste trabalho, incluindo os momentos de dificuldades, sempre diposto a ajudar e tirar dúvidas.

À todos os professores do Departamento de Estatística por todo o conhecimento compartilhado durante as aulas do curso.

Resumo

A classificação por picos é um problema na neurociência contemporânea e consiste na detecção e agrupamento de atividades neurais, também chamados de picos. O problema existe porque há uma dificuldade na hora de diferenciar estas atividades, quando vários neurônios são gravados simultaneamente, fazendo com que diferentes atividades neurais sejam detectadas. Este trabalho tem o objetivo de apresentar um processo para a resolução do problema da classificação por picos.

Para isso vamos, usando o *software* Rstudio, aplicar algumas técnicas estatísticas em um conjunto de dados reais sobre atividades neurais de uma espécie de gafanhoto *Schistocerca americana*, com o desvio mediano absoluto e técnicas de análise multivariada como análise de componentes principais e K-means.

Palavras-chave: Atividade neural. Desvio mediano absoluto. Análise de componentes principais. Agrupamento.

Abstract

Spike sorting is a problem in contemporary neuroscience and consists of detecting and grouping neural activities, also called peaks. The problem exists because there is a difficulty when it comes to differentiating these activities, when several neurons are recorded simultaneously, causing different neural activities to be detected. This work aims to present a process for solving the spike sorting problem.

For this we will, using *software* Rstudio, apply some statistical techniques on a real dataset about neural activities of a species of grasshopper *Schistocerca americana*, with the absolute median deviation and multivariate analysis techniques as analysis of main components and K-means.

Keywords: Neural activity. Median absolute deviation. Principal component analysis. Clustering.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Na neurociência contemporânea, uma das principais questões em aberto é entender como o cérebro (cérebro humano, de vertebrados ou de insetos) realiza determinadas atividades. Por isso, para responder esta questão, são feitos experimentos tentando registrar os potenciais de ação ou disparos (*spike*) do maior número possível de neurônios. Nestes experimentos os registros das atividades neurais do cérebro são detectados usando dispositivos chamados eletrodos que ficam no "espaço extracelular"; o resultado é uma sequência temporal dos registros de todos os neurônios no alcance do eletrodo.

Segundo Lewicki e Michael (1998) a detecção da atividade de disparos de neurônios, conhecida como *spike sorting*, é uma técnica desafiadora e é um pré-requisito para estudar quaisquer tipos de atividade cerebrais. A maioria dos neurônios no cérebro se comunica entre si através dos disparos. Quando um certo neurônio dispara ele pode influenciar os disparos dos neurônios conectados a ele.

Entretanto, com isso temos um problema. As atividades neurais de vários neurônios são detectadas simultaneamente, o que faz com que os dados brutos fiquem confusos e difíceis de interpretar. Podemos fazer uma analogia de uma sala com várias pessoas conversando e com microfones instalados nesta sala. Com o registro dos microfones, queremos descobrir quantas pessoas falam durante a gravação, quais as características de cada pessoa (tom da voz, sonoridade, variabilidade) e qual o discurso de cada pessoa. Este é o problema da "classificação de picos".

Para tentar resolver este problema, usaremos algumas técnicas estatísticas, como desvio mediano absoluto, análise de componentes principais e análise de agrupamento: *k-means* para dados funcionais.

O desvio mediano absoluto foi popularizado por Hampel e Frank (1974) e, segundo Leys et al. (2013), pode ser usado como uma alternativa para o desvio médio absoluto, com a vantagem de ser mais resistente à *outliers*. Segundo Johnson e Wichern (2014) Análise Multivariada é um conjunto de técnicas estatísticas para dados multivariados com o objetivo de redução de dimensão de dados, agrupamento, investigação de dependência entre variáveis, predição e construção de hipóteses. Dentro da análise multivariada temos a análise de componentes principais com foco em reduzir as dimensões dos dados com p variáveis para j componentes principais onde $j < p$, enquanto o k-means é uma técnica que tem o objetivo de agrupar n observações em k grupos. Vamos aplicar estas técnicas em dados funcionais, que são dados que fornecem informações sobre curvas dentro de um espaço contínuo de tempo, como pode ser visto em Ramsay, Hooker e Graves (2009).

1.1 Objetivo

O objetivo deste trabalho é apresentar técnicas estatísticas, principalmente na área de Análise Multivariada, para realizar o *Spike Sorting*.

Estas técnicas foram inspiradas no estudo de Pouzat (2012) sobre o *Spike Sorting*.

Os objetivos específicos deste trabalho são:

- Apresentar um método para identificar a atividade de disparos de neurônios;
- Separar essas atividades do restante dos dados brutos e, assim, gerar um novo conjunto de dados somente com as atividades detectadas;
- Reduzir a dimensão do novo conjunto de dados;
- Fazer o agrupamento das atividades detectadas por grau de similaridade de amplitude.

2 Metodologia

Neste capítulo, vamos apresentar e explicar as técnicas estatísticas que serão usadas no desenvolvimento deste projeto.

2.1 Desvio mediano absoluto

Definição. Considere uma amostra $x_1, x_2, \ldots, x_n$. O desvio mediano absoluto (DMA) amostral, denotado por $\tilde{\sigma}$, é definido pela mediana dos desvios absolutos mediano dos dados, da forma

$$\tilde{\sigma} = mediana(|x_i - \tilde{X}|),\ i = 1, 2, \ldots, n$$

sendo que $\tilde{X}$ é a mediana dos dados.

Exemplo. Considere os dados (0, 1, 1, 2, 4, 7). A mediana dos dados é 1.5 então

$$x_i - \tilde{x} = (-1.5, -0.5, -0.5, 0.5, 2.5, 5.5)$$

$$|x_i - \tilde{x}| = (1.5, 0.5, 0.5, 0.5, 2.5, 5.5)$$

ordenando os módulos dos desvios absolutos medianos temos

$$|x_i - \tilde{x}| = (0.5, 0.5, 0.5, 1.5, 2.5, 5.5).$$

Portando, o DMA é obtido pela mediana dos desvios acima, ou seja, $\tilde{\sigma} = \frac{0.5+1.5}{2} = 1$.

O DMA é uma medida de dispersão para dados que possuem *outliers*, ou seja, para este conjunto específico de dados, esta medida é mais robusta que o desvio padrão. No caso do desvio padrão, as distâncias entre cada observação estão elevadas ao quadrado. Desta forma grandes desvios são fortemente influenciados por *outliers*. Em compensação, a medida de dispersão DMA não é influenciada quando o conjunto de dados possui um pequeno número de *outliers*. No caso em que uma amostra aleatória segue uma distribuição de probabilidade Cauchy, que sabemos que não possui média, nem variância finitas, talvez, o DMA seja um possível estimador para a mediana populacional.

Existe uma relação entre o estimador do DMA e o estimador do desvio padrão, da seguinte forma:

$$\hat{\sigma} = k \cdot \tilde{\sigma}$$

em que k é uma constante e o seu valor depende da distribuição da amostra, ver em Ruppert (2011), página 155. No caso em que os dados seguem uma distribuição normal, o k é obtido a partir de uma função de distribuição acumulada, da seguinte forma:

$$k = \frac{1}{\Phi^{-1}(3/4)} \approx 1.4826$$

A função quantil da função inversa da distribuição normal padrão $Z = \frac{(X-\mu)}{\sigma}$, em que $X \sim N(\mu,\sigma^2)$, cobre 50% (entre 1/4 e 3/4) da distribuição acumulada padrão quando recebe o argumento 3/4. Vamos verificar o resultado acima. Partindo de

$$P(\mid X-\mu \mid \leq \tilde{\sigma}) = P\left(\left|\frac{X-\mu}{\sigma}\right| \leq \frac{\tilde{\sigma}}{\sigma}\right) = P\left(|Z| \leq \frac{\tilde{\sigma}}{\sigma}\right) = \frac{1}{2},$$

Observação. Igualamos a $\frac{1}{2}$ porque queremos que a probabilidade acima esteja entre os quantis $\frac{1}{4}$ e $\frac{3}{4}$, ou seja, $\frac{3}{4} - \frac{1}{4} = \frac{1}{2}$

devemos ter isto :

$$P\left(|Z| \leq \frac{\tilde{\sigma}}{\sigma}\right) = \Phi\left(\frac{\tilde{\sigma}}{\sigma}\right) - \Phi\left(-\frac{\tilde{\sigma}}{\sigma}\right) = \frac{1}{2} \Rightarrow$$

$$\Phi\left(\frac{\tilde{\sigma}}{\sigma}\right) - \left(1 - \Phi\left(\frac{\tilde{\sigma}}{\sigma}\right)\right) = \frac{1}{2} \Rightarrow$$

$$\Phi\left(\frac{\tilde{\sigma}}{\sigma}\right) = \frac{3}{4}.$$

Usando agora a inversa da distribuição acumulada da normal padrão, obtemos

$$\frac{\tilde{\sigma}}{\sigma} = 0.67449,$$

Assim, a relação entre o DMA e o desvio padrão é

$$\frac{\sigma}{\tilde{\sigma}} = \frac{1}{0.67449} = 1.4826.$$

2.2 Análise de componentes principais

A análise de componentes principais (ACP) é uma técnica de Análise Multivariada que tem o objetivo de reduzir a dimensão dos dados e facilitar a interpretação das análises obtidas. Considere um vetor aleatório p-variado de variáveis, $\mathbf{X} = (X_1, \ldots, X_p)$, com vetor de médias $\boldsymbol{\mu}$ e matriz de variâncias e covariâncias Σ, com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, da forma:

$$\sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}.$$

Considere agora uma amostra aleatória de tamanho n do vetor $\mathbf{X}$, representada pela matriz abaixo

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Definição. As componentes principais (CP), denotadas por Y_1,Y_2,...,Y_p, são combinações lineares das variáveis originais X_1,X_2,...,X_p e são não correlacionadas entre si.

$$Y_1 = a_1^\top X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = a_2^\top X = a_{21}X_2 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\vdots$$
$$Y_p = a_p^\top X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

Em que $a_i = (a_{i1}$,$a_{i2}, \ldots, a_{ip})$ são os vetores de coeficientes da i-ésima combinação linear, com as restrições de $a_i^T a_i = 1$ e $Cov(a_i^T X$,$a_j^T X) = 0$ para $j < i$. Considere que a matriz de variâncias e covariâncias Σ tenha autovetores $e_1, \ldots, e_p$, então as componentes principais $Y_1, \ldots, Y_p$ também podem ser escritas como

$$Y_i = e_i^\top X = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p.$$

A análise de componentes principais consiste em criar um novo conjunto de variáveis, $\mathbf{Y} = (Y_1, \ldots, Y_p)$, usando as variáveis originais $\mathbf{X} = (X_1, \ldots, X_p)$. Assim, podemos reduzir a dimensão dos dados pela análise das variáveis de maior variabilidade de $\mathbf{Y}$, pois cada componente principal representa uma proporção da variância total. A variância total dos dados pode ser obtida da seguinte forma:

$$\sum_{i=1}^{p} Var(X_i) = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

e por consequência a proporção da variância total explicada da i-ésima componente principal, denotado por P_i, é dada por

$$P_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i},\ i = 1,\ 2,\ \ldots,\ p$$

Onde λ_i é o i-ésimo autovalor da matriz de covariância da matriz dos dados.

A proporção da variância explicada acumulada até o i-ésimo componente principal, denotado como $P^{(i)}$, é definido como,

$$P^{(i)} = \frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{p} \lambda_i},\ i = 1,\ 2,\ \ldots,\ p$$

As proporções da variância explicada são posicionadas em ordem decrescente, ou seja, $P_1 \geq \ldots \geq P_p$ e portanto, durante a análise, devemos escolher as k-ésimas primeiras componentes principais para a análise de acordo com a proporção que queremos atingir.

Exemplo. Considere um conjunto de dados com as variáveis X_1, X_2, X_3 e X_4 com a seguinte matriz de variância e covariância

$$\Sigma = \begin{pmatrix} 5 & -2 & 0 & 0 \\ -2 & 3 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

Os autovalores e autovetores são

$$\lambda_1 = 6.24,\ e_1^\top = (0.85, -0.53,0,0)$$

$$\lambda_2 = 3.62,\ e_2^\top = (0,\ 0,\ 0.53,\ 0.85)$$

$$\lambda_3 = 1.76,\ e_3^\top = (0.53,\ 0.85,\ 0,\ 0)$$

$$\lambda_4 = 1.38,\ e_4^\top = (0,0,0.85, -0.53)$$

As componentes principais podem ser escritas como

$$Y_1 = 0.85X_1 - 0.53X_2$$

$$Y_2 = 0.53X_3 + 0.85X_4$$

$$Y_3 = 0.53X_1 + 0.85X_2$$

$$Y_4 = 0.85X_3 - 0.53X_4$$

A variância total dos dados é dada como

$$\sum_{i=1}^{p} Var(X_i) = \sigma_{11} + \sigma_{22} + \sigma_{33} + \sigma_{44} = 5 + 3 + 2 + 3 = 13$$

ou

$$\sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 6.24 + 3.62 + 1.76 + 1.38 = 13.$$

Dividindo cada autovalor pela variância total dos dados podemos encontrar a proporção da variância explicada:

Tabela 2.1 – Variância acumulada exemplo

	CP 1	CP 2	CP 3	CP 4
proporção da variância explicada	0.48	0.28	0.14	0.10
proporção da variância explicada acumulada	0.48	0.76	0.90	1

Para dados funcionais, que são dados que fornecem informações sobre curvas dentro de um período de tempo, segundo Ramsay, Hooker e Graves (2009), uma componente principal representa a variação em torno da média, por isso é bom somar e subtrair as médias de cada componente pelos seus respectivos autovetores para conseguir novas interpretações sobre os dados e quais os pontos onde os dados possuem maior variância. Na Figura 2.1 temos um exemplo retirado do livro de Ramsay, Hooker e Graves (2009) que mostra visualmente o comportamento de adicionar e subtrair o autovetor relacionado ao primeiro autovalor na primeira componente principal. O eixo X representa o índice do tempo da duração da curva e o eixo Y representa a média da variável em cada ponto do período de tempo, e as três curvas do gráfico são: a do meio é a média da variável

em cada ponto do período, a de cima é a média somado pelo autovetor e a de baixo é a média subtraindo pelo autovetor, os dois últimos representados pelos sinais + e − respectivamente.

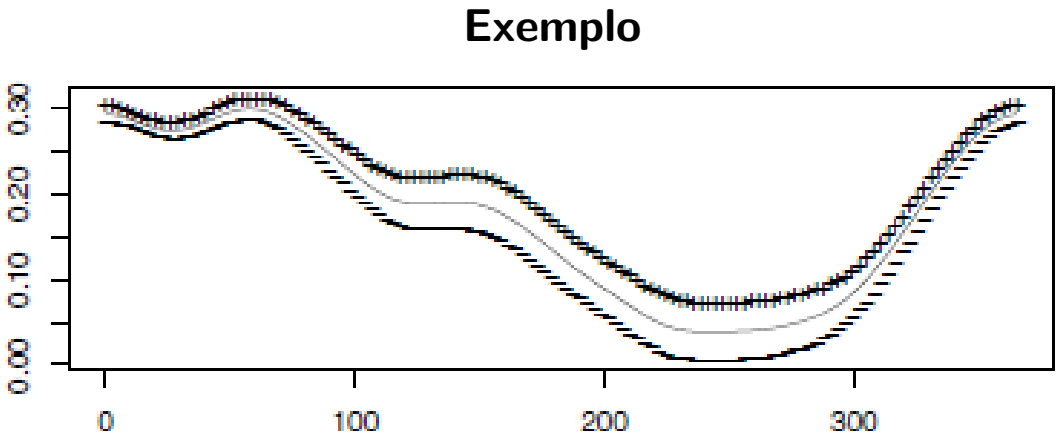

Figura 2.1 – Exemplo de um gráfico de uma componente principal somando e subtraindo pelo seu respectivo autovetor.

2.3 Análise de agrupamento: K-means

Definição. A análise de agrupamento (ou análise de *cluster*) é uma técnica de Análise Multivariada que tem por objetivo dividir os dados de uma amostra e classificar, em um mesmo grupo, os elementos que possuem características semelhantes em relação às variáveis, e ao mesmo tempo separando, em grupos diferentes, os elementos mais heterogêneos. O método K-means é uma técnica não-hierárquica que tem a função de encontrar diretamente os K agrupamentos pré-especificados pelo pesquisador. Além da especificação inicial do número de grupos, a novidade é que nas técnicas não-hierárquicas os novos grupos podem ser construídos a partir de outros grupos já formados, ou seja, se dois elementos estão juntos em um *cluster*, não necessariamente eles estarão juntos até o final do processo, como consequência, não se pode construir dendrogramas. O método mais utilizado é conhecido como K-means, que pode ser obtido da seguinte forma:

1. Fazer a separação inicial das observações em K grupos.
2. Alocar as observações para o grupo com a centróide mais próxima

   (As centróides de um grupo são as coordenadas centrais de cada um dos K grupos). A j-ésima centróide pode ser obtida pela média de todas as observações alocadas a ela, ou seja

   $$c_j = \frac{1}{|C_j|} \sum_{K:x_k \in C_j} x_k$$

   em que C_j é o número de observações alocadas no j-ésimo grupo.)

usando a distância euclidiana; se necessário trocar o grupo atual da observação para um novo grupo e recalcular as centróides.

Definição. Sejam A e B dois vetores $A = (a_1,a_2,\ldots,a_n)$ e $B = (b_1,b_2,\ldots,b_n)$ a distância euclidiana, entre A e B denotado por $d(A,B)$ em um espaço n-dimensional é dada por

$$d(A,B) = \sqrt{(a_1 - b_1)^2 + \ldots + (a_n - b_n)^2};$$

3. Repetir o passo anterior para todos os elementos da amostra até que não haja mais realocações de observações entre os grupos.

Exemplo. Vamos considerar um conjunto de dados com quatro observações com duas variáveis X_1 e X_2. Queremos dividí-los $K = 2$ grupos.

Tabela 2.2 – K-means exemplo

	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Primeiro, definimos um chute inicial: Depois, vamos calcular as coordenadas das centróides

Tabela 2.3 – Chute inicial exemplo

Primeiro grupo	A,B
Segundo grupo	C,D

.

Tabela 2.4 – Cálculo das centróides iniciais: exemplo

	X_1	X_2
AB	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
CD	$\frac{1+(-3)}{2} = -1$	$\frac{(-2)+(-2)}{2} = -2$

Agora, vamos fazer os cálculos de uma nova centróide e da distância euclidiana para o elemento A, supondo que o elemento A não mude de grupo, e depois supondo que o elemento A mude de grupo.

A fórmula para calcular as centróides são:

- caso o elemento A entre em algum outro grupo que não seja o seu do chute inicial, fazemos:

$$\tilde{x}_{i,novo} = \frac{n\tilde{x}_i + x_{ij}}{n+1};$$

- caso o elemento A saia do seu próprio grupo do chute inicial, escrevemos:

$$\tilde{x}_{i,novo} = \frac{n\tilde{x}_i - x_{ij}}{n-1}.$$

Em que x_{ij} representa o j-ésimo elemento do i-ésimo grupo, $\tilde{x}_i$ a média do i-ésimo grupo e n o número de elementos no i-ésimo grupo.

Calculando as novas centróides com base nas suposições acima:

$$Grupo(B) : \tilde{x}_{1,novo} = \frac{2(2)-5}{2-1} = -1; \tilde{x}_{2,novo} = \frac{2(2)-3}{2-1} = 1$$

$$Grupo(ACD) : \tilde{x}_{1,novo} = \frac{2(-2)+5}{2+1} = 1; \tilde{x}_{2,novo} = \frac{2(-2)+5}{2+1} = 0.33$$

Agora calcularemos a distância euclidiana entre o componente A e as duas centróides definidas acima.

- Caso o item A não mude de grupo

$$d(A,(AB))^2 = (5-2)^2 + (3-2)^2 = 10$$

$$d(A,(CD))^2 = (5+1)^2 + (3-1)^2 = 61$$

- Caso o item A mude para o grupo CD

$$d(A,(B))^2 = (5-1)^2 + (3-1)^2 = 40$$

$$d(A,(ACD))^2 = (5-1)^2 + (3+0.33)^2 = 27.07$$

A menor distância euclidiana é a distância entre o elemento A e grupo AB, portanto, o componente A não vai sair do primeiro grupo do chute inicial, e de forma análoga, podemos aplicar a mesma lógica para os elementos restantes. O restante deste exemplo pode ser visto em Johnson e Wichern (2014).

3 Aplicação em dados reais

Neste capítulo, vamos apresentar um exemplo de *spike sorting* e vamos utilizar os *softwares* livres Rstudio, usando a linguagem R, e o Ggobi para fazer as análises dos dados. O código em R utilizado para a aplicação dos dados foi obtido em Pouzat (2012).

3.1 Sobre os dados

Os dados são uma sequência temporal de ações neurais detectados por quatro tétrodos, que é um dispositivo que contém quatro eletrodos, instalados em locais diferentes no tecido extracelular no cérebro de gafanhotos (*Schistocerca americana*). Cada gravação possui 20 segundos e há registro na gravação a cada 15 KHz; no total cada gravação possui 300000 registros.

3.2 Análise descritiva

Tabela 3.1 – Análise descritiva

	gravação 1	gravação 2	gravação 3	gravação 4
Mínimo	-9.074	-8.229	-6.890	-7.350
1° quantil	-0.371	-0.450	-0.530	-0.490
Mediana	-0.029	-0.036	-0.042	-0.040
Média	0.000	0.000	0.000	0.000
3° quantil	0.326	0.396	0.469	0.430
Máximo	10.626	11.742	9.849	10.560

Observação. Além disso, o desvio padrão dos dados para cada gravação é 1, o que significa que os dados foram normalizados.

Na Figura 3.1 temos o gráfico dos registros no primeiro 0.2 segundo separado pelas quatro gravações, nomeados de g.1 até g.4. Podemos ver que na maioria do tempo não acontece nada, apenas o chamado ruído de fundo, e de vez em quando acontece o registro dos disparos.

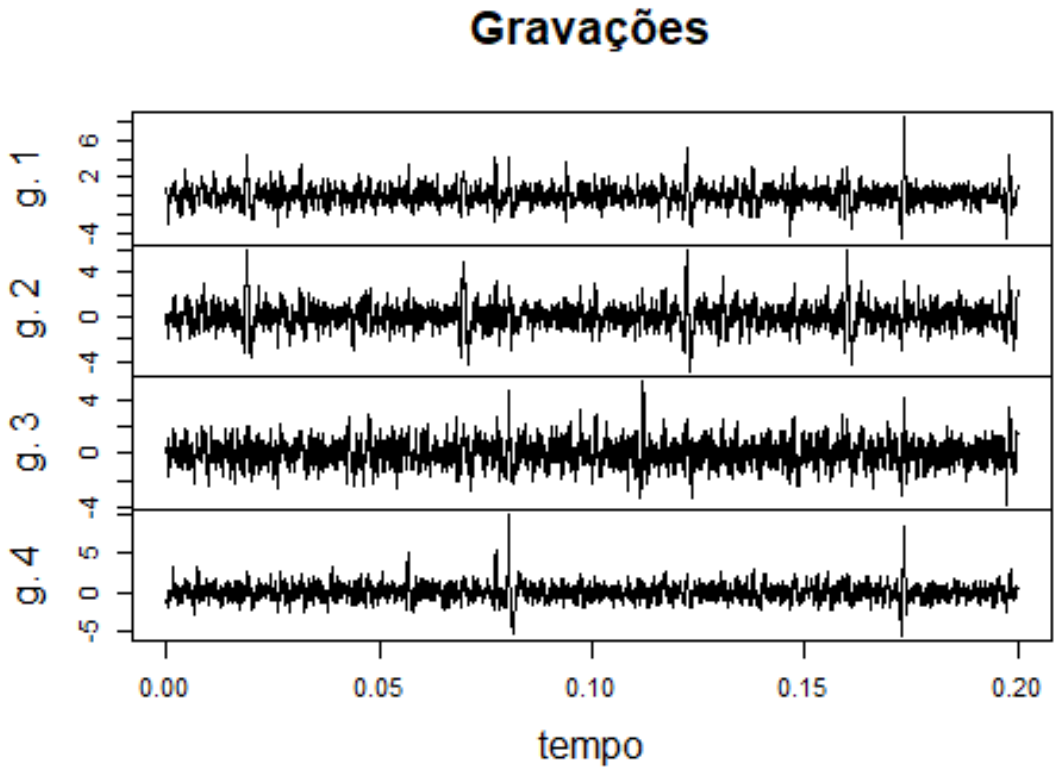

Figura 3.1 – Primeiro 0.2 segundo dos dados em cada um dos tétrodos

3.3 Renormalização dos dados

Observando os gráficos, é possível perceber que em muitos momentos não ocorrem os disparos. Enquanto isso, ocorre o chamado ruído de fundo. É importante caracterizar o ruído de fundo, pois fica mais fácil de detectar os disparos. Entretanto, com o desvio padrão atual, é possível que haja confusão, com pontos que não são disparos sendo considerados como disparos. Para evitar isso, vamos renormalizar os dados caracterizando o ruído de fundo.

Para caracterizar o ruído de fundo e diferenciar os disparos quando ocorrerem, precisamos fazer com que o desvio padrão dos ruídos de fundo seja aproximadamente 1. Para isso, podemos usar o DMA para estimar o desvio padrão dos dados e renormalizar os dados usando a média igual à mediana e o novo desvio padrão estimado. Com isso, ficará mais fácil detectar os picos. Na Figura 3.2 podemos ver graficamente a diferença entre o desvio padrão dos dados originais e o desvio padrão dos dados renormalizados para a primeira gravação que são 1 e 1.93 respectivamente.

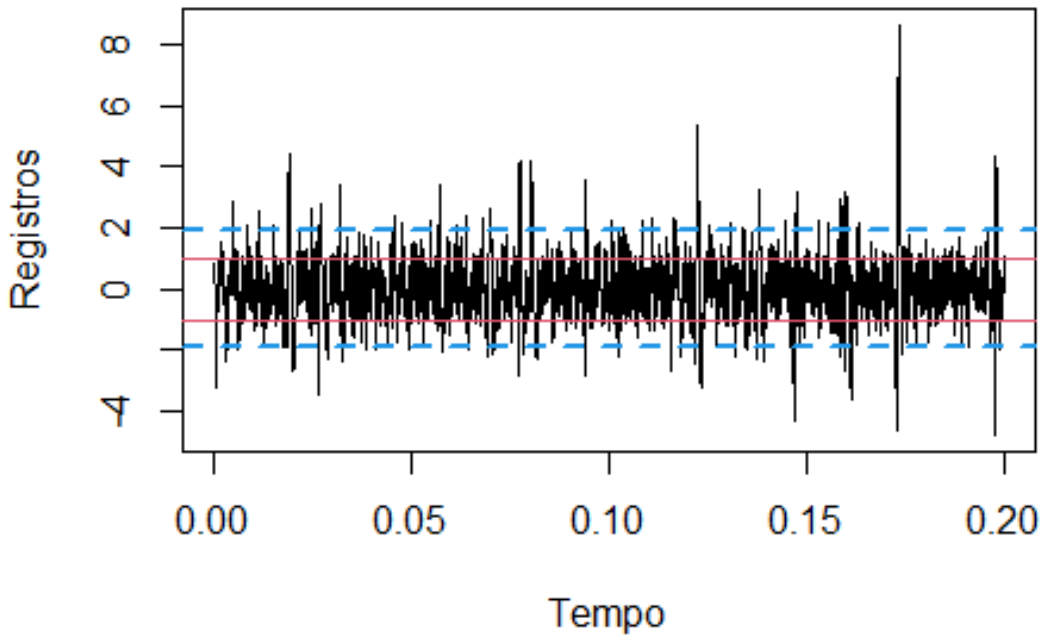

Figura 3.2 – Primeiro 0.2 segundo do primeiro tétrodo destacando o desvio padrão (vermelho) e o desvio padrão estimado pelo DMA (azul).

Fazendo um Q-Q plot, que é um gráfico que compara os quantis de duas distribuições, visto na Figura 3.3, para testar e comparar a normalidade de cada gravação antes (linhas tracejadas) e depois (linhas contínuas) da renormalização, podemos ver que vai existir mais disparos nos dados depois da renormalização, ou seja, isso quer dizer que vamos conseguir mais picos usando a normalização baseado no DMA do que usando a normalização baseada na normal padrão e, por isso, a partir de agora, vamos usar os dados com o desvio padrão estimado pelo DMA.

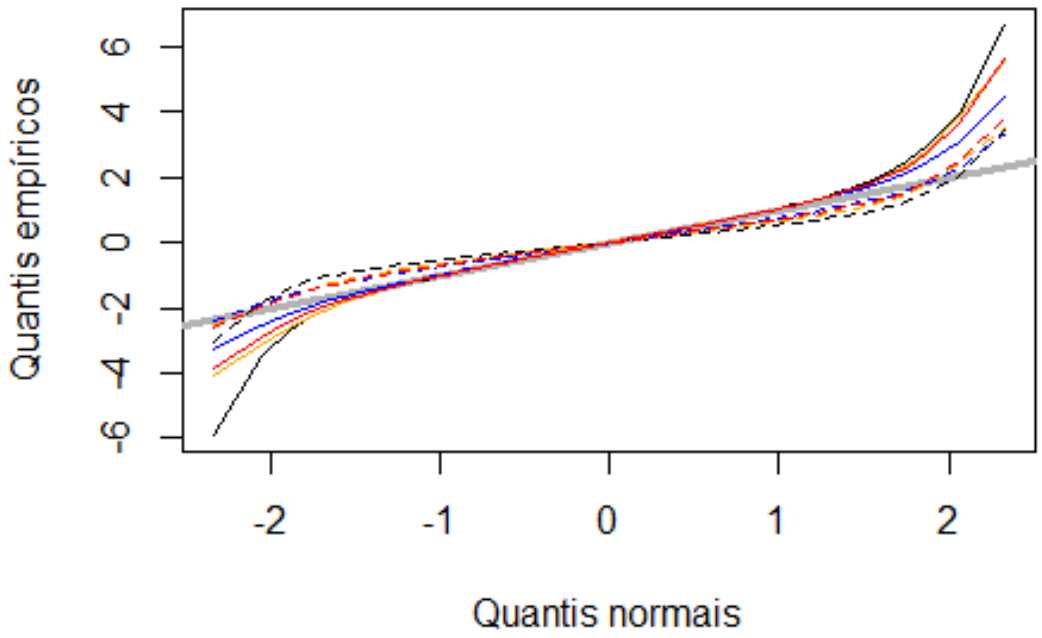

Figura 3.3 – Q-Q plot dos dados usando a normalização baseada no DMA (linhas contínuas) e a normalização padrão (linhas tracejadas); As cores representam as gravações: g.1, preto; g.2, laranja; g.3, azul; g.4, vermelho.

3.4 Detecção de picos

Para identificar os picos, calculamos a média de cada elemento com seus dois vizinhos mais próximos, tanto à esquerda quanto à direita, e se esta média for maior do que 4, será tratado como um pico, ou máximo local. Foram identificados 1795 disparos no conjunto de dados ao longo dos quatro tétrodos.

Depois, vamos dividir o conjunto de dados em duas partes: A primeira metade será para criar um modelo que será testado nas duas partes. Todo procedimento descrito a partir de agora será realizado usando apenas a primeira metade.

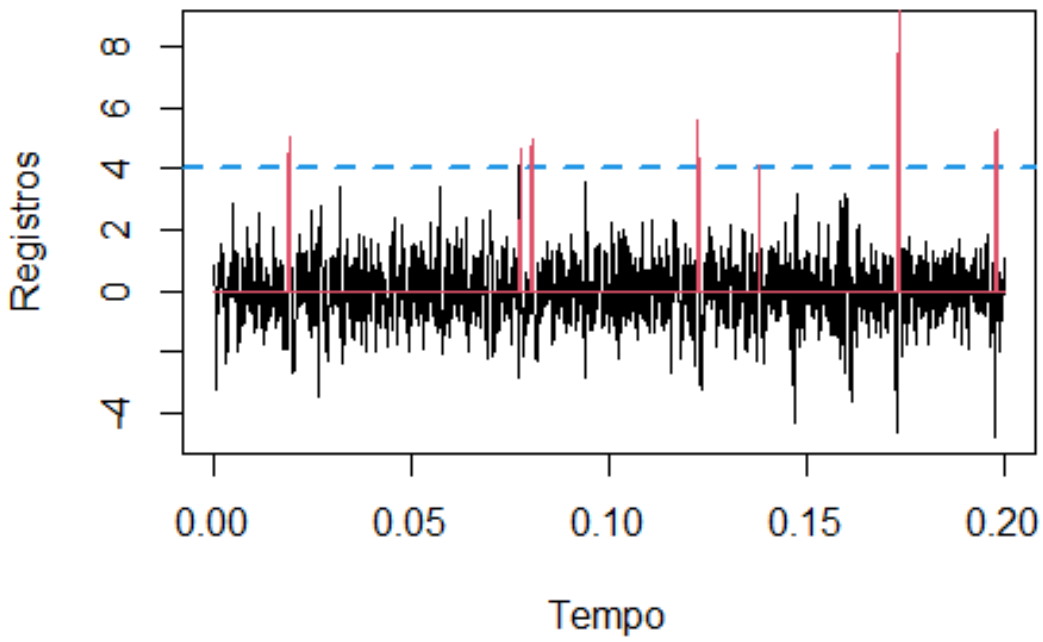

Figura 3.4 – Picos detectados durante o primeiro 0.2 segundo no primeiro tétrodo

3.5 Cortes

Com os máximos locais detectados, fazemos um corte para cada pico, separando o trecho selecionado do conjunto de dados. Mas, antes, precisamos determinar o tamanho amostral de cada corte. Para determinar o tamanho dos cortes seguimos os seguintes passos:

1. Faça o corte mais longo que o necessário (neste caso vai ser 50 para direita e 50 para esquerda);
2. Compute as estimativas robustas do evento central (mediana e DMA) dos cortes para cada um dos pontos de registro selecionados para os quatro locais de gravações;
3. Faça os gráficos dos dois parâmetros e verifique quando o traço atinge o ruído de fundo (mediana é 0 e DMA 1).

Na Figura 3.5 encontra-se o DMA (vermelho) e a mediana (preto). Podemos ver que inicialmente o valor da mediana é 0 e do DMA é 1 e ao longo dos 100 pontos selecionados anteriormente, em cada um dos quatro locais de gravação, os valores destes parâmetros

mudam aproximadamente 15 pontos antes e 30 pontos depois do disparo, portanto devemos alterar o tamanho destes cortes para ter estes tamanhos.

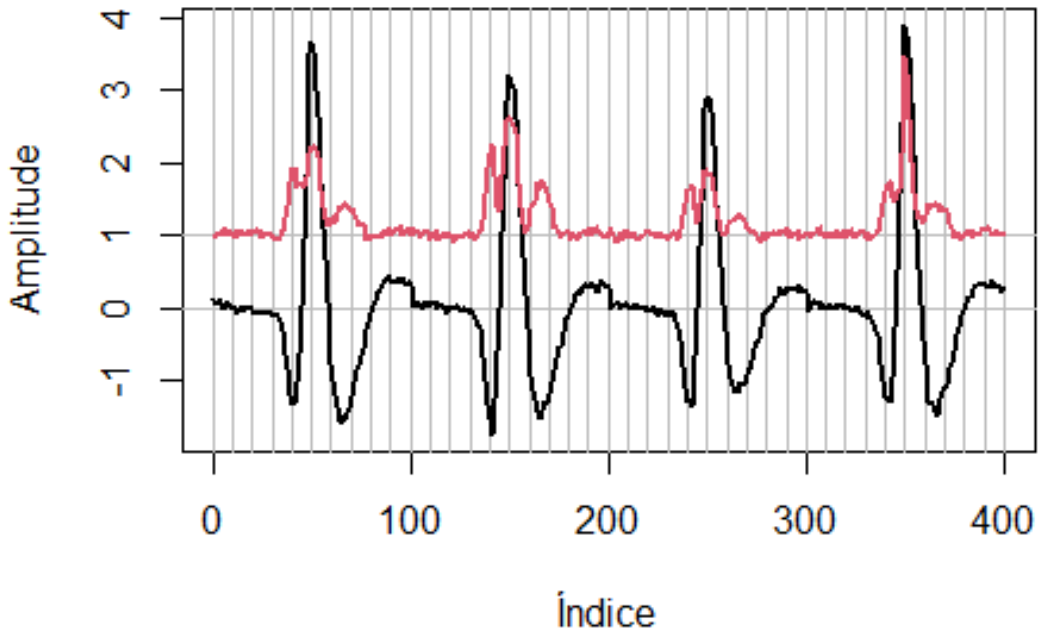

Figura 3.5 – Mediana (preto) e DMA (vermelho) de cada ponto de registro dos cortes, a cada 100 registros representa um tédroto, totalizando 400 observações.

3.6 Limpeza dos dados

O objetivo agora é "limpar" os eventos, ou seja, eliminar eventos em que acontecem dois ou mais picos quase ao mesmo tempo. Fazemos isso para facilitar as interpretações do conjunto de dados e evitar interpretações erradas envolvendo mais de um disparo em cada evento; o objetivo é que se tenha somente um pico por evento.

Depois deste procedimento o número de cortes foi reduzido de 1795 para 858.

Abaixo, as Figuras 3.6 e 3.7 mostram os cortes antes e depois da limpeza, respectivamente, com seus tamanhos amostrais ajustados destacando a mediana (vermelho) e o DMA (azul). Perceba que depois do máximo local os valores são menores depois da limpeza dos eventos.

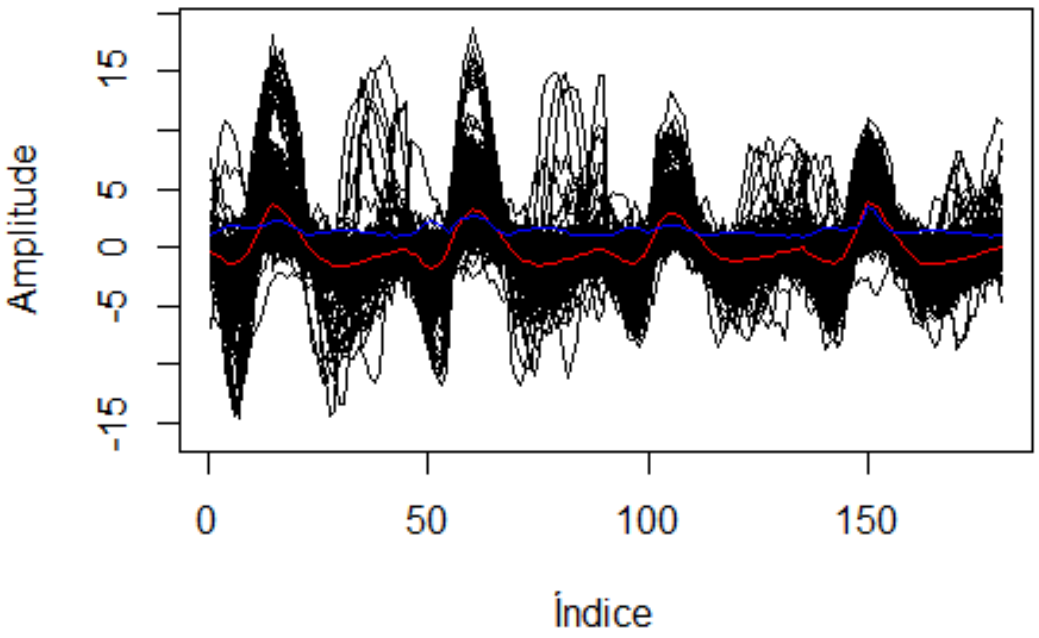

Figura 3.6 – 200 primeiros eventos sobrepostos

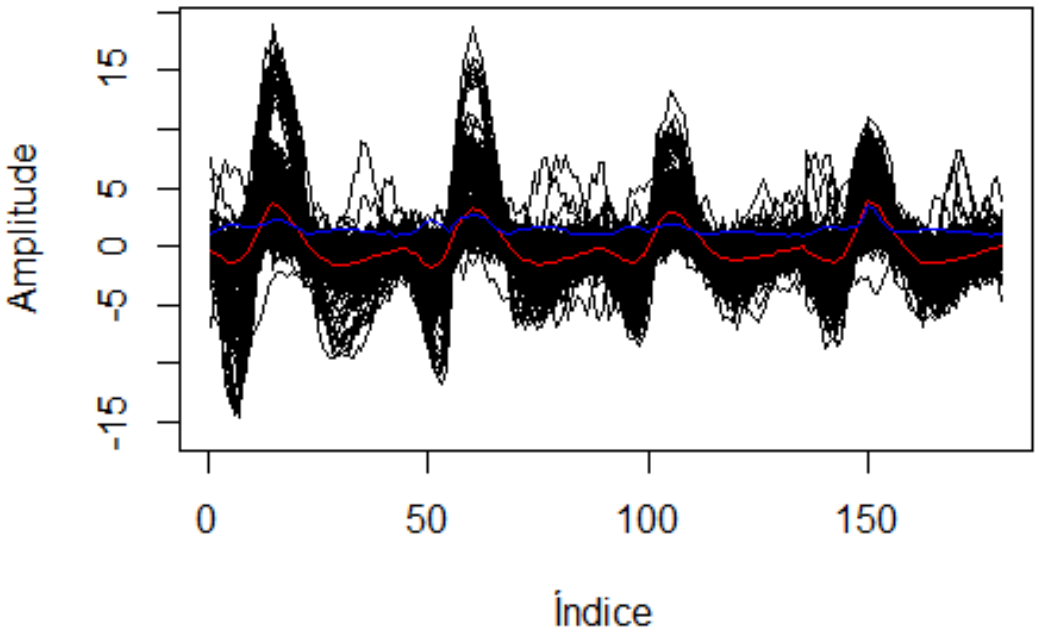

Figura 3.7 – 200 primeiros eventos sobrepostos depois da limpeza

3.7 Redução de dimensão

O objetivo agora é reduzir as dimensões dos dados, pois a dimensão dos dados atualmente é de 858 cortes, com tamanho amostral de 180 cada (cada corte com tamanho amostral de 45 para cada um dos quatro tétrodos). Aplicando a análise de componentes principais, as oito primeiras variáveis do conjunto de dados, depois de limpos, representam 80.14% da variância total, com isso vamos usar até as oito primeiras componentes principais nas análises abaixo.

As Figuras 3.8 até 3.11 mostram a média de cada uma das 180 variáveis (preto) e essa mesma média somando (vermelho) e subtraindo (azul) os seus respectivos autovetores da matriz de variâncias e covariâncias dos dados, multiplicado por 5 das primeiras quatro componentes principais. Cada pico destes gráficos representa os registros de cada tétrodo e eles representam a média dos eventos limpos de cada gravação em ordem.

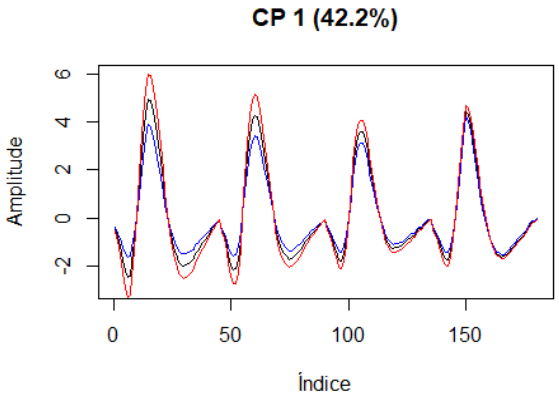

Figura 3.8 – Primeira componente principal.

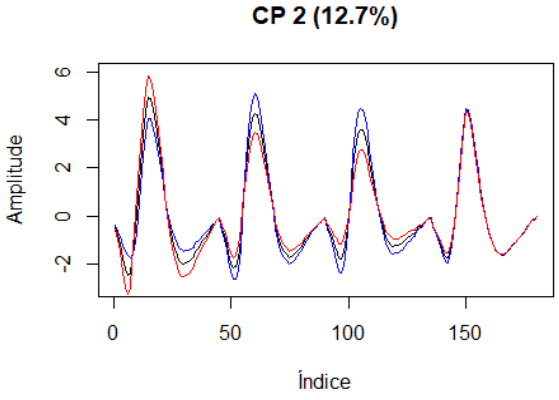

Figura 3.9 – Segunda componente principal.

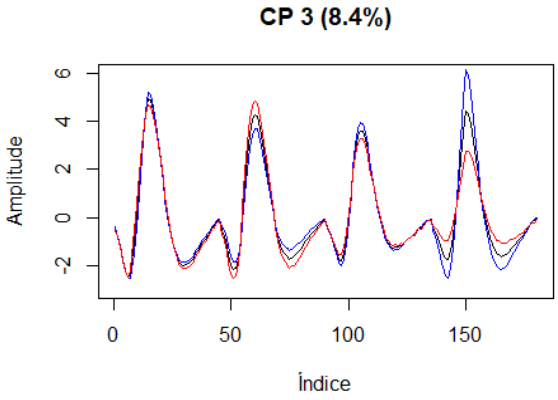

Figura 3.10 – Terceira componente principal.

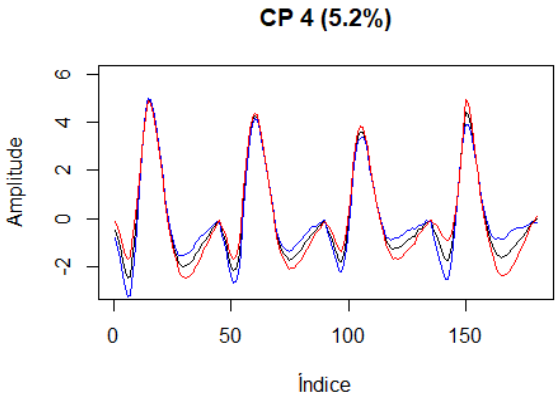

Figura 3.11 – Quarta componente principal.

A partir destes gráficos conseguimos as seguintes interpretações:

1. As três primeiras componentes principais correspondem às variações de amplitude;
2. Para a primeira componente principal, um evento com grande projeção é menor que a média nos locais de gravação 1, 2 e 3;
3. Para a segunda componente principal, um evento com grande projeção é maior que a média nos locais de gravação 1, menor nos locais 2 e 3, e igual no local 4;
4. Para a terceira componente principal, um evento com grande projeção é maior que a média no local 4;
5. A quarta componente principal é a primeira componente que revela uma oposição à amplitude.

Depois vamos fazer o mesmo para as próximas quatro componentes principais, vistas nas Figuras 3.12 até 3.15

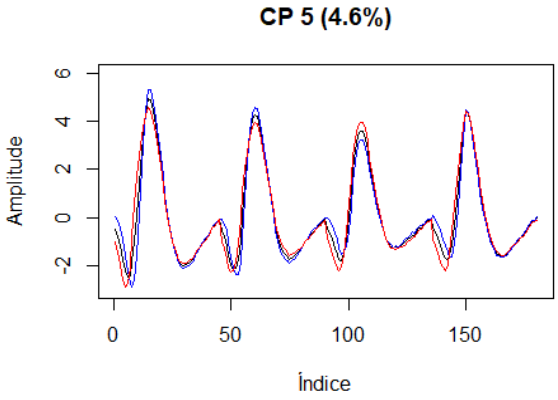

Figura 3.12 – Quinta componente principal.

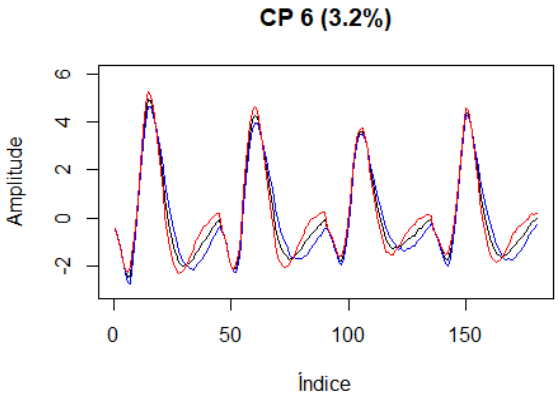

Figura 3.13 – Sexta componente principal.

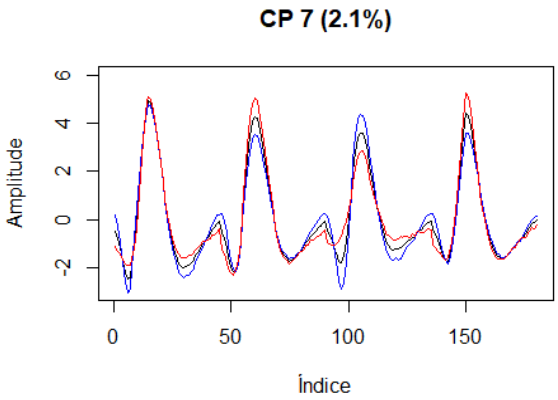

Figura 3.14 – Sétima componente principal.

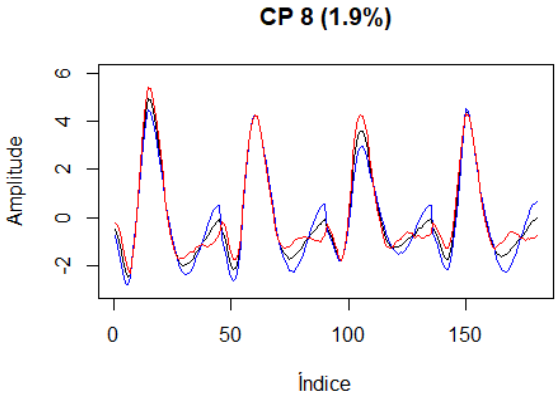

Figura 3.15 – Oitava componente principal.

1. Na quinta componente principal, um evento com alta projeção tende a ser mais lento que o evento médio;
2. As componentes 5 e 6 correspondem a efeitos compartilhados entre os locais de gravação;
3. A componente 7 apresenta efeitos da mudança de forma nos registros em todos os locais, exceto no primeiro;
4. Na oitava componente principal a projeção aumenta mais rápido e decai mais lentamente que a média em todos os locais de gravação;
5. Na componente principal 8 há evidências de que estamos chegando no limite da variedade dos eventos, indicando que as próximas componentes principais serão mais próximas do ruído de fundo.

Na Figura 3.16 podemos ver os gráficos de dispersão 2x2 das primeiras quatro componentes principais. Elas são as que possuem as maiores proporções da variância total explicada. Juntas, representam 68.5% da variância total dos dados. Entretanto, quando fazemos o gráfico de dispersão com duas dimensões perdemos informação, e por isso, usando um software chamado Ggobi, um software livre, fazemos um gráfico com três dimensões das três primeiras componentes principais, o máximo possível, representando 63.3%. Com esta nova visualização dos dados, conseguimos distinguir exatamente dez grupos diferentes.

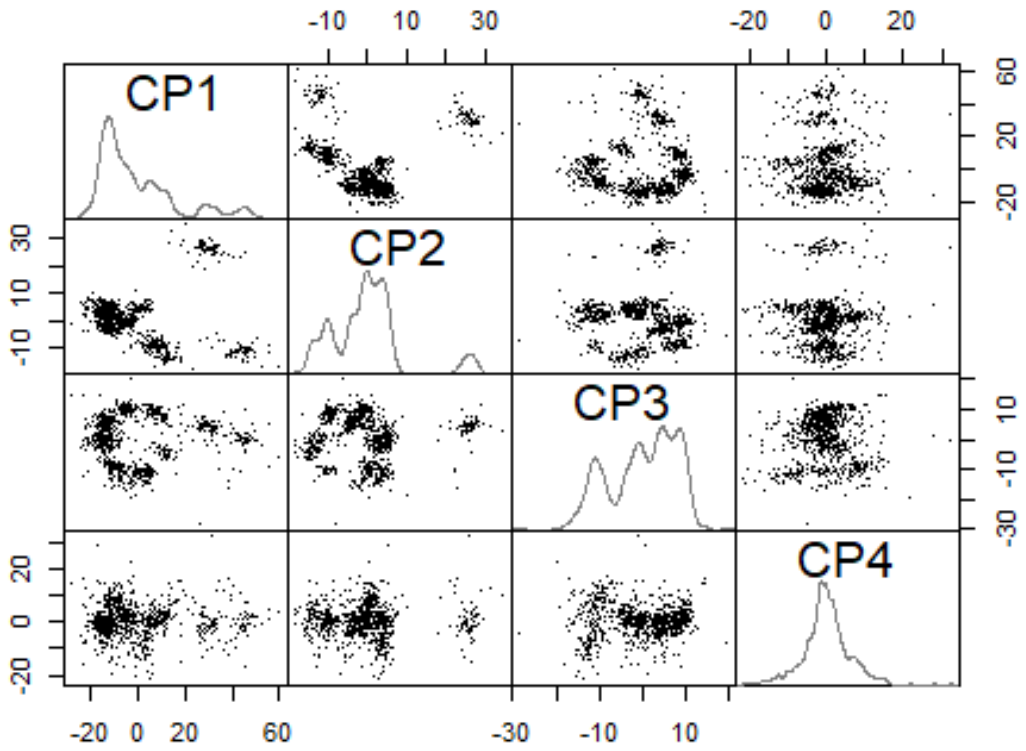

Figura 3.16 – Gráfico de dispersão das quatro primeiras componentes principais junto com suas respectivas densidades.

3.8 Agrupamento

Agora que sabemos o número de grupos que queremos para fazer a divisão dos dados, vamos fazer o agrupamento usando o K-means com $K = 10$.

Na Figura 3.17 podemos ver o gráfico das duas primeiras componentes principais feito no Ggobi, um software que cria gráficos interativos para melhor visualização dos dados, destacando os grupos em diferentes cores, sendo 10 grupos no total.

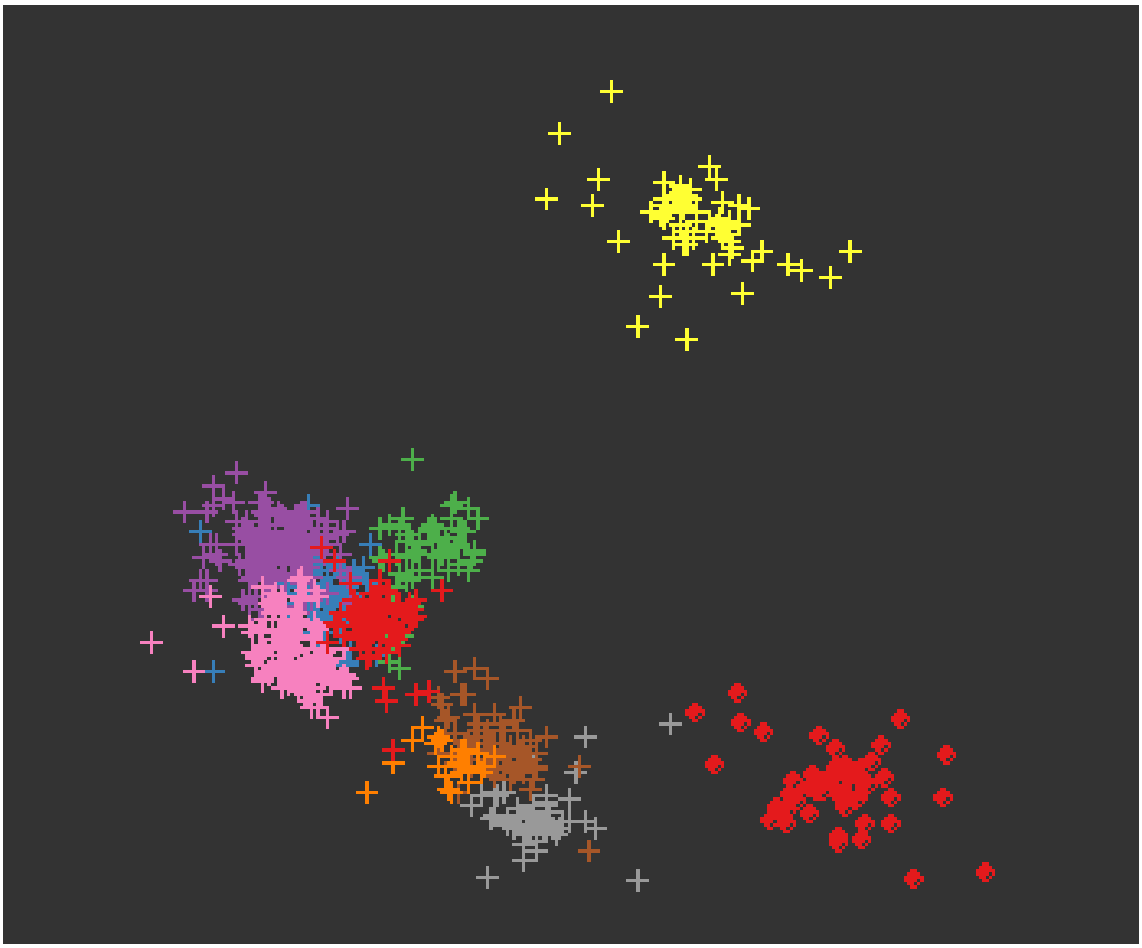

Figura 3.17 – Duas primeiras componentes principais destacando os grupos

Os próximos gráficos, Figuras 3.18 e 3.19, mostram os eventos sobrepostos, mediana (vermelho) e o DMA (azul) de cada um dos grupos. Assim podemos ver que eles possuem similaridades dentro de seus respectivos grupos.

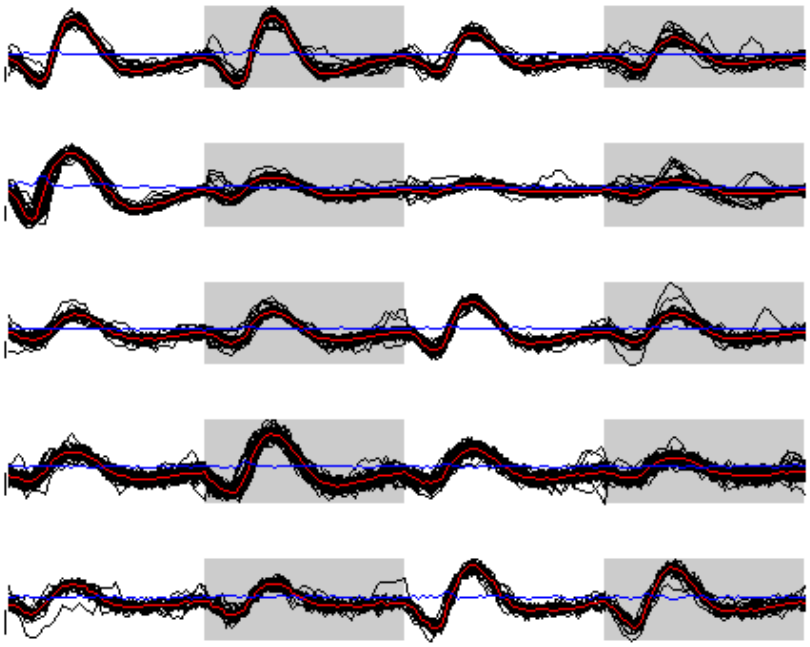

Figura 3.18 – Grupos 1 ao 5 dos eventos

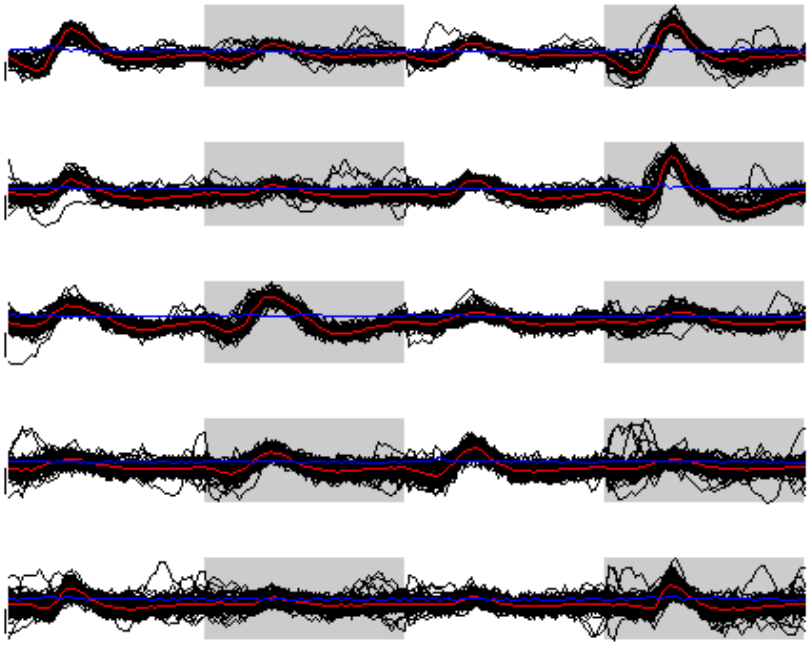

Figura 3.19 – Grupos 6 ao 10 dos eventos

4 Considerações Finais

Neste trabalho, foi mostrado o processo de agrupar as atividades neurais de acordo com a similaridade de amplitude usando técnicas estatísticas. Também foi mostrada a metodologia para chegar a esse processo. Dá para notar que nenhuma suposição sobre os dados foi necessária para a execução do processo proposto.

O objetivo deste trabalho foi desenvolver e estudar este processo. Para este estudo foi de grande importância o uso de dois softwares livres, o Ggobi e o R.

As técnicas usadas durante o desenvolvimento deste projeto foram, principalmente, de Análise Multivariada. Também aplicamos a técnica de renormalização dos dados usando o desvio mediano absoluto, o que faz com que os dados sejam transformados mudando seu desvio padrão, por isso não foi necessária nenhuma suposição sobre os dados.

Referências

HAMPEL; FRANK, R. The influence curve and its role in robust estimation. *Journal of the american statistical association*, Taylor & Francis, v. 69, n. 346, p. 383–393, 1974.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. [S.l.]: Pearson London, UK:, 2014. v. 6.

LEWICKI; MICHAEL, S. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, IOP Publishing, v. 9, n. 4, p. R53, 1998.

LEYS, C. et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, Elsevier, v. 49, n. 4, p. 764–766, 2013.

POUZAT, C. *Spike Sorting*. 2012. <http://xtof.perso.math.cnrs.fr/sorting.html>.

RAMSAY, J.; HOOKER, G.; GRAVES, S. *Functional data analysis with R and MATLAB*. New York: Springer, 2009.

RUPPERT, D. *Statistics and data analysis for financial engineering*. New York: Springer, 2011.