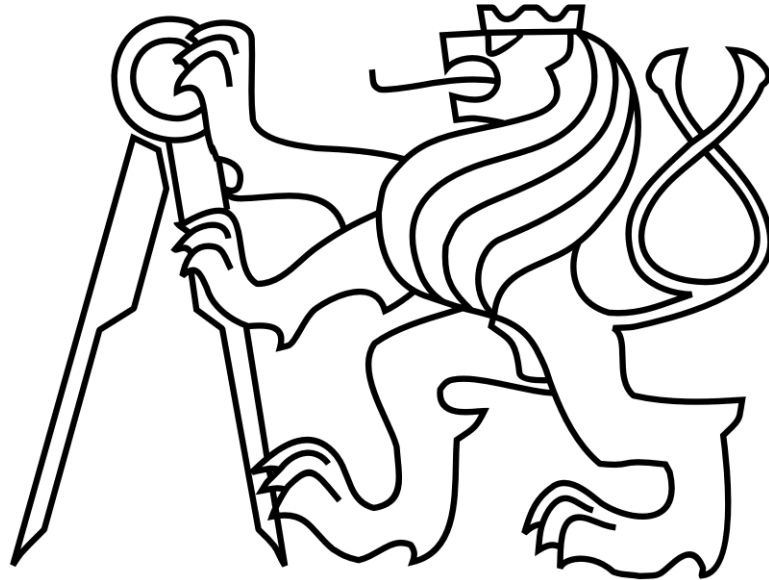


ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE  
FAKULTA ELEKTROTECHNICKÁ



## BAKALÁŘSKÁ PRÁCE

Vizualizace dat z hierarchického shlukování

*Radek Tomis*

*Květen 2014*

*Vedoucí práce: Ing. Václav Gerla, Ph.D.*



## ZADÁNÍ BAKALÁŘSKÉ PRÁCE

**Student:** Radek T o m i s

**Studijní program:** Kybernetika a robotika (bakalářský)

**Obor:** Robotika

**Název tématu:** Vizualizace dat z hierarchického shlukování

### Pokyny pro vypracování:

1. Implementujte v MATLABu vybranou metodu hierarchického shlukování a použijte ji v úloze analýzy neznámého EEG signálu (vedoucí práce dodá studentovi EEG signály, výsledky adaptivní segmentace i vypočtené parametry z nalezených segmentů).
2. Sestavte přehledné GUI umožňující zobrazit dendrogram a výsledky shlukové analýzy. Součástí navrženého řešení bude možnost dynamicky měnit počet shluků a sledovat vliv na výsledné zobrazení. Zvolte postup, který umožní přehledně zobrazit velký počet shluků.
3. V EEG záznamu barevně odlište nalezené shluky.
4. Implementujte algoritmus umožňující export výsledků shlukové analýzy do souboru.
5. Otestujte navržené řešení nad reálnými klinickými daty.

### Seznam odborné literatury:

- [1] Lopes da Silva Fernando, Niedermeyer Ernst: Electroencephalography – Basic principles, clinical applications and related field. Philadelphia. 2005.
- [2] Fung Glenn.: A Comprehensive Overview of Basic Clustering Algorithms. 2001.
- [3] The MathWorks, Inc., Creating Graphical User Interfaces - R2013b, Natick, 2013. online: [http://www.mathworks.com/help/pdf\\_doc/matlab/buildgui.pdf](http://www.mathworks.com/help/pdf_doc/matlab/buildgui.pdf)

**Vedoucí bakalářské práce:** Ing. Václav Gerla, Ph.D.

**Platnost zadání:** do konce letního semestru 2013/2014

L.S.

doc. Dr. Ing. Jan Kybic  
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.  
děkan

V Praze dne 10. 1. 2014



# Poděkování

Chtěl bych poděkovat vedoucímu své práce Ing. Václavu Gerlovi, Ph.D. za trpělivost, inspiraci a rady k dané problematice, které mi ochotně poskytoval po celou dobu psaní práce.



# Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze dne .....

.....

Podpis





# Abstrakt

Práce se zabývá implementací hierarchických shlukovacích metod v MATLABu aplikovaných na vícerozměrná data charakterizující EEG lidského těla a vizualizací jejích výsledků. Teoretická část práce se zabývá třemi základními oblastmi, a to popisem signálů lidského těla, jejich zpracováním a shlukovou analýzou. V rámci praktické části byl vytvořen program umožňující provést hierarchickou shlukovou analýzu a zobrazit její výsledky. Nad programem byla vytvořena grafická nadstavba. V závěru práce se nachází zhodnocení otestování programu na reálných datech. Výsledky práce ukazují, že metody hierarchického shlukování obsažené v MATLABu mohou být vhodným základem pro analýzu neurobiologických signálů.

## Klíčová slova

EEG, GUI, MATLAB, shluková analýza, zpracování signálu

# Abstract

The focus of this project is MATLAB implementation of hierarchical clustering methods applied on multidimensional data representing EEG of human body and visualization of their results. Theoretical part of the project deals with three basic issues that are: description of human body signals, their processing and clustering analysis. In the practical part, there is summary of creating GUI including application providing hierarchical cluster analysis and visualization of its results. At the conclusion, there is assessment of testing the application on real clinical data. The results of this project show that hierarchical clustering methods, implemented in MATLAB, can be instrumental towards neurobiologic signals analysis.

## Keywords

EEG, GUI, MATLAB, cluster analysis, signal processing

# Obsah

|          |                                       |           |
|----------|---------------------------------------|-----------|
| <b>1</b> | <b>Úvod</b>                           | <b>1</b>  |
| <b>2</b> | <b>Biomedicínské signály</b>          | <b>3</b>  |
| 2.1      | EEG . . . . .                         | 3         |
| 2.1.1    | Využití . . . . .                     | 3         |
| 2.1.2    | Historie . . . . .                    | 3         |
| 2.1.3    | Druhy rytmu EEG . . . . .             | 5         |
| 2.2      | Evokované potenciály . . . . .        | 6         |
| 2.3      | Polysomnografie . . . . .             | 7         |
| 2.3.1    | Fáze spánku . . . . .                 | 7         |
| 2.4      | Artefakty . . . . .                   | 8         |
| 2.4.1    | Fyzikální artefakty . . . . .         | 9         |
| 2.4.2    | Biologické artefakty . . . . .        | 9         |
| <b>3</b> | <b>Měření a zpracování EEG</b>        | <b>11</b> |
| 3.1      | Elektrody . . . . .                   | 11        |
| 3.2      | Zapojení elektrod . . . . .           | 11        |
| 3.3      | Předzpracování signálu . . . . .      | 12        |
| 3.4      | Segmentace . . . . .                  | 14        |
| <b>4</b> | <b>Shluková analýza</b>               | <b>15</b> |
| 4.1      | Nehierarchické metody . . . . .       | 16        |
| 4.2      | Hierarchické metody . . . . .         | 17        |
| 4.2.1    | Aglomerativní metody . . . . .        | 18        |
| 4.2.2    | Divizní metody . . . . .              | 18        |
| 4.3      | Metrika . . . . .                     | 19        |
| 4.4      | Výběr shlukovacího kriteria . . . . . | 20        |

---

|          |   |           |
|----------|---|-----------|
| 4.4.1    | K-means . . . . .   | 20        |
| 4.4.2    | Metoda průměrné vzdálenosti . . . . .                               | 21        |
| 4.4.3    | Metoda těžišť . . . . .   | 21        |
| 4.4.4    | Metoda mediánová . . . . .  | 21        |
| 4.4.5    | Metoda nejbližšího souseda . . . . .                                | 21        |
| 4.4.6    | Metoda nejbližšího souseda . . . . .                                | 22        |
| 4.4.7    | Metoda Wardova . . . . .  | 22        |
| 4.4.8    | MacNaughton–Smithova metoda . . . . .                               | 22        |
| 4.4.9    | Další metody . . . . .  | 23        |
| 4.5      | Zhodnocení metod shlukování . . . . .                               | 23        |
| <b>5</b> | <b>Program PSG-HC</b>   | <b>25</b> |
| 5.1      | Vstupní soubory . . . . .   | 25        |
| 5.1.1    | Hodnoty signálu . . . . .   | 26        |
| 5.1.2    | Hranice segmentů . . . . .  | 26        |
| 5.1.3    | Vyčíslené příznaky . . . . .  | 26        |
| 5.1.4    | Popis příznaků . . . . .  | 26        |
| 5.2      | GUI . . . . .   | 26        |
| 5.2.1    | Úvodní okno . . . . .   | 27        |
| 5.2.2    | Okno analýzy . . . . .  | 28        |
| 5.3      | Export výsledků . . . . .   | 32        |
| 5.3.1    | Formát výstupních dat . . . . .                                     | 33        |
| 5.4      | Popis funkcí využívaných programem . . . . .                        | 33        |
| 5.4.1    | Vytvoření dendrogramu . . . . .                                     | 33        |
| 5.4.2    | Nastavení prahu . . . . .   | 36        |
| 5.4.3    | Získání barev listů dendrogramu . . . . .                           | 36        |
| 5.4.4    | Logaritmické měřítko . . . . .                                      | 37        |
| 5.4.5    | Vykreslení segmentového grafu . . . . .                             | 38        |
| 5.4.6    | Vykreslení grafu signálu . . . . .                                  | 38        |
| 5.5      | Otestování na klinických datech . . . . .                           | 39        |
| 5.5.1    | Demonstrace práce programu při detekci paradoxního spánku . . . . . | 39        |
| 5.5.2    | Demonstrace práce programu při detekci artefaktů . . . . .          | 41        |
| 5.5.3    | Použití aplikace nad jinými typy dat . . . . .                      | 43        |
| <b>6</b> | <b>Závěr</b>  | <b>45</b> |

|   |    |
|---|----|
| A Seznam příznaků, graf hledání artefaktů | 51 |
| B Obsah CD                                | 55 |



# Kapitola 1

## Úvod

Program PSG-HC vytvořený v rámci této bakalářské práce by se měl stát jedním z prostředků umožňujících náhled na zpracování a analýzu EEG záznamu, nebo libovolného biomedicínského signálu obecně, za použití grafického uživatelského prostředí. Může být využíván buď jako samostatný analyticko-vizualizační nástroj nebo v budoucnu jako jeden z modulů pro MATLAB toolbox PSGLab vytvořený na Katedře Kybernetiky na ČVUT v Praze.

Účelem programu je provést shlukovou analýzu signálu předzpracovaného do formátu pro PSGLab. Umožní hierarchickou analýzu za použití několika různých matematických metod a několik způsobů zobrazení výsledků. Program byl vytvořen se záměrem nezatěžovat GUI programu zbytečnými ovládacími prvky a ponechat co nejvíce prostoru vizualizačním nástrojům.

Celá práce kromě úvodu a závěru obsahuje teoretickou a praktickou část, které jsou děleny do kapitol a podkapitol.

Ve dvou kapitolách teoretické části práce se čtenář seznamuje s vybranými biomedicínskými signály měřitelnými na lidském těle, s jejich strukturou, způsobem jejich vzniku a využití v medicínském prostředí. Dále tyto kapitoly popisují jejich měření a předzpracování do formy vhodné pro práci programu. Poslední kapitola teoretické části se zabývá vysvětlením problematiky shlukové analýzy. Popisuje různé metriky a shlukovací kritéria, přičemž klade důraz zejména na hierarchické metody.

Kapitola věnovaná praktické části představuje autorem vytvořený program PSG-HC. První část kapitoly kopíruje uživatelův průchod programem, podává vysvětlení ovládacích prvků a interpretuje vizuální výsledky analýzy. V druhé části kapitoly práce vysvětluje nejdůležitější funkce a algoritmy použité při vytváření programu. Závěr praktické části je věnován zhodnocení praktické použitelnosti shlukovacích

---

algoritmů na reálných klinických datech s přihlédnutím k přehlednosti vizualizace výsledků a využití strojového času a kapacity při analýze. Jako konkrétní příklad užití slouží demonstrace práce programu při detekci paradoxního spánku a odhalování artefaktů v EEG záznamu.



# Kapitola 2

## Biomedicínské signály

Biomedicínský signál je časově proměnná veličina schopná nést informaci, která je vyvolávána a snímána ze živého organismu.

### 2.1 EEG

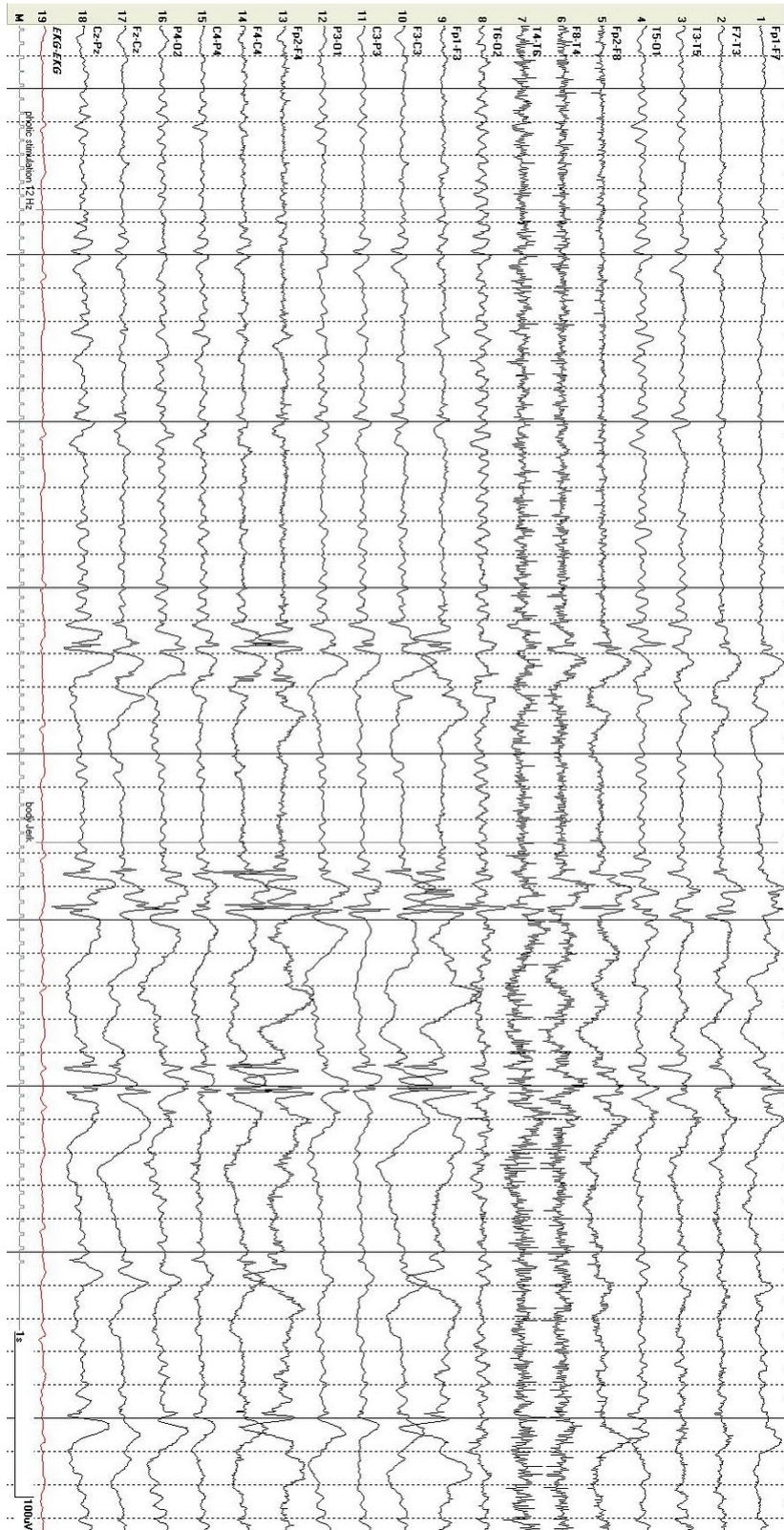
EEG, neboli elektroencefalogram, je obvykle vícekanálový časový záznam nestacionárního (amplitudově a frekvenčně proměnného) signálu reprezentujícího elektrické potenciály snímané zpravidla neinvazivně z povrchu skalpu. Amplitudy potenciálu se pohybují v řádu jednotek až desítek mikrovoltů. Například při epileptickém záchvatu však může vystoupat až na  $300 \mu\text{V}$ . Frekvence může dosahovat až 100 Hz, přičemž maximum výkonu se přenáší v pásmu 0,5 - 30 Hz [1].

#### 2.1.1 Využití

Největšího využití se EEG dostává v neurologii a psychiatrii. Slouží lékaři jako nástroj diagnózy a monitorování stavů vědomí a různých chorob a dysfunkcí, jako například epilepsie, zranění hlavy, poruchy spánku, poruchy centrální nervové soustavy, neurologické poruchy, ale i migrény, kóma nebo roztroušenou sklerózu. Jedná se však pouze o podpůrnou disciplínu, vždy je nutný expertní názor lékaře nebo odborníka.

#### 2.1.2 Historie

Počátky této techniky sahají do roku 1875, kdy britský vědec Richard Caton objevil elektrické impulsy na povrchu hlavy savců. Z jeho pozorování vychází práce



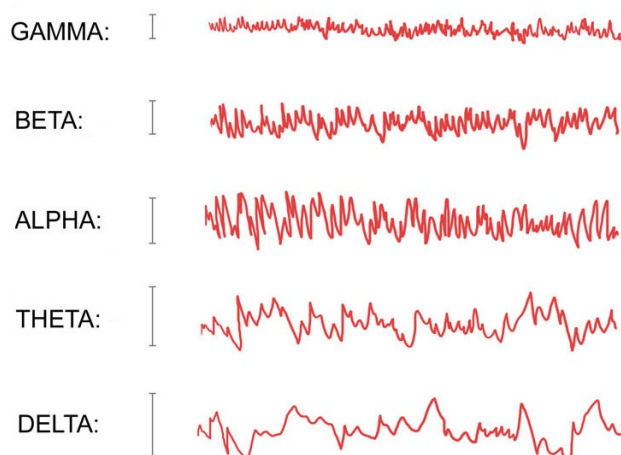
Obrázek 2.1: Příklad naměřených biosignálů [2]

německého neurologa Hanse Bergera, který roku 1929 sestrojil první přístroj na měření a zaznamenávání aktivity mozku, elektroencefalogram. Jako první popsal *alfa* vlny<sup>1</sup> a definoval je jako vlny o frekvenci 7,81-13,28 Hz. Dále popsal i jejich potlačení a nahrazení rychlejšími *beta* vlnami. Jako důsledek rozvoje elektroniky obecně v 60. a 70. letech 20. století dochází k rozvoji techniky jak snímací, tak analytické a EEG přístroje se začínají objevovat v klinické praxi.

### 2.1.3 Druhy rytmu EEG

Mozková aktivita se dá rozdělit na čtyři základní druhy<sup>2</sup>:

- Alfa aktivita
  - Frekvence 8-13 Hz, amplituda 30-80  $\mu\text{V}$ .
  - Přirozená aktivita bdícího zdravého mozku při zavřených očích.
- Beta aktivita
  - Frekvence 14-30 Hz, amplituda 10-30  $\mu\text{V}$ .
  - Bdící člověk v motorickém klidu.



Obrázek 2.2: Příklady mozkových vln [3]

<sup>1</sup>Alternativně *Bergerovy vlny*

<sup>2</sup>Meze frekvencí i amplitud se podle různých zdrojů mohou měnit, zde převzato z [1]

- Theta aktivita
  - Frekvence 4-7,5 Hz, amplituda  $<30 \mu\text{V}$ .
  - Bdící člověk ve stresu nebo vzrušení.
- Delta aktivita
  - Frekvence 0,5-4 Hz.
  - U dospělých v bdělém stavu nenormální, může indikovat nádor, u dětí hluboký spánek.
- Někdy se objevují i další typy rytmu, například
  - Spánková vřeténka 12-16 Hz, amplituda náhle narůstá a klesá. Anglický název zní *sleep spindle*.
  - K-komplexy [4].
  - Gamma aktivita,  $>30$  Hz, v EEG se často kvůli zatím pouze povrchnímu prozkoumání neuvádí.
  - Epileptické grafoelementy, zřetelně odlišitelné hroty o šířce kolem 50 ms.

## 2.2 Evokované potenciály

Na podobné bázi jako EEG vyšetření fungují i diagnostické metody jednotlivých nervových soustav. Zprůměrováním epoch EEG lze zjistit takzvané evokované potenciály (EP). Jedná se o reakce mozku nebo příslušné nervové soustavy na určitý vnější podnět [6]:

- Senzorický - SEP: Vyšetření periferní nervové soustavy pomocí stimulace senzitivních receptorů pomocí krátkého elektrického impulsu (30/60 ms), laseru nebo krátkým vzduchovým pulsem.
- Motorický - MEP: Testuje integritu motorických drah sledováním motorické odpovědi svalu na krátký vysokovoltážní stimul nebo stimul pulzním magnetickým polem.
- Vizuální - VEP: Sleduje reakce zrakového kortexu na vizuální podněty dvojího druhu, buď krátkými záblesky nebo rytmickými zvraty políček na černobílé šachovnici. Může odhalit například roztroušenou sklerózu.

- Sluchový - (B)AEP: Odezva na zvukové kliknutí nebo tón, zkoumá činnost mozkového kmene nebo poruchy sluchu.

## 2.3 Polysomnografie

Polysomnografie, zkráceně PSG, je komplexní vyšetřovací metoda spánkových fází. Může sloužit k diagnóze celé řady spánkových a dýchacích poruch jako insomnie, narkolepsie nebo chrápání.

Před vyšetřením obvykle pacient musí vykonat určité úkony jako mrknutí, pohyb nohou, zadržení dechu a další, aby systém zaregistroval tyto pohyby a později je odborník mohl při konečném zpracování odhalit.

Vyšetření obvykle trvá celou noc a probíhá ve speciální spánkové laboratoři. K rozlišení bdělosti a jednotlivých spánkových fází je třeba snímat více signálů. Podle různých zdrojů [1], [7], [8] se měří 11, 14 nebo až 55 kanálů. Přirozeně záleží na konkrétním druhu vyšetření a možnostech spánkové laboratoře poskytnout kvalitní vybavení. Kromě minimálně čtyř EEG se obvykle jedná o EKG a PNG a dále o párové EOG a EMG<sup>3</sup>, které mimo jiné pomáhají s detekcí pohybových artefaktů. Dále se mohou snímat pohyby jednotlivých částí těla a úroveň saturace kyslíku v hemoglobinu.

Po samotném vyšetření provádí odborník analýzu dat, která jsou rozdělena do většinou třicetisekundových intervalů. Z nich pak určí, za jak dlouho po vypnutí světel pacient usnul, procentuální čas na lůžku strávený spánkem a rozdělení spánku do spánkových fází.

### 2.3.1 Fáze spánku

Normální lidský spánek se skládá ze dvou typických fází. První se v angličtině nazývá Non-rapid eye movement (NREM, Non-REM), druhý Rapid eye movement (REM). Jak je z názvů patrné, nejzřetelnějším rozeznávaným znakem je pohyb očí pod zavřenými víčky. V češtině se pro Non-REM fázi vžilo pojmenování klidný a pro REM fázi aktivní nebo paradoxní spánek. U novorozenců každá z fází zabírá přibližně polovinu celkové doby spánku, ale už od věku několika let dochází k výraznému přelivu procentuálního zastoupení ve prospěch klidného spánku. Ten v dospělosti u

---

<sup>3</sup>EKG - elektrokardiogram, PNG - pneumogram, EOG - elektrookulogram, EMG - elektromyogram

člověka zabírá kolem 75% doby a tento poměr se stářím ještě stoupá [4]. Střídání fází u dospělého člověka je zobrazeno na Obr. 2.3.1.

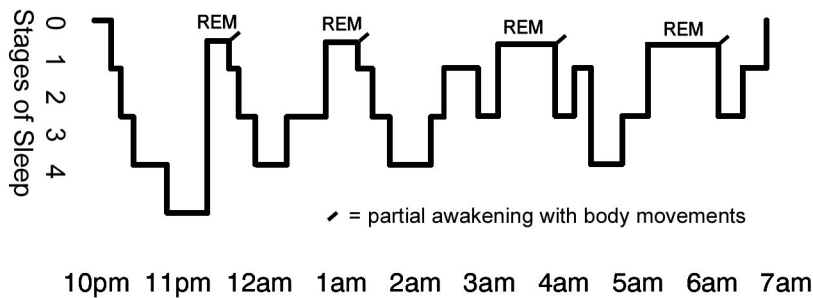
**Non-REM** Klidný spánek se skládá ze čtyř stádií (fáze 1 - fáze 4).

Fáze 1 nastává zejména při usínání nebo při krátkodobém procitnutí v průběhu spánku. Trvá jen několik minut, během kterých odeznívají alfa vlny a objevují se nízkonapěťové vlny se smíšenou frekvencí.

Fáze 2 je nejčastější prožívaná fáze spánku, může zabírat až 55% jeho celkové délky. Vyskytují se spánková vřeténka a K-komplexy.

Fáze 3 a 4 jsou fáze hlubokého spánku, kde se objevují pomalé delta vlny s vysokými amplitudami.

**REM** Spánek projevující se rychlými pohyby očí přichází v přibližně 90 minutových intervalech a objevuje se čtyřikrát až šestkrát během normálního spánku. Je charakterizován nízkonapěťovými střídajícími se frekvencemi v rytmech alfa a beta. Na konci této fáze spánku dochází k spontánním probuzením, které si člověk nemusí uvědomit ani zapamatovat, a pohybům těla, zejména šubáním končetin. Vzhledem k tomu, že se mozková aktivita REM fáze podobá aktivitě při bdění, vzil se pro ni název aktivní spánek.



Obrázek 2.3: Střídání fází spánku [5]

## 2.4 Artefakty

Jako artefakty se v biomedicíně označují jevy nebo procesy, které nemají fyziologický původ ve vyšetřovaném orgánu. Obvykle se dělí na artefakty fyzikální a biologické.

### 2.4.1 Fyzikální artefakty

Mezi fyzikální (technické) artefakty se řadí takové jevy, které nejsou důsledkem pohybové nebo jiné biologické aktivity zkoumaného subjektu. Do této skupiny patří:

**Elektrostatické potenciály.** Vznikají v důsledku nízké jakosti elektrod nebo jejich špatného kontaktu s pokožkou. Ovlivnit je může zvýšené pocení, změna obsahu kyslíku nebo prokrvení v mozkové tkáni, změny parametrů elektrod nebo pohyby z elektrostatických hmot (např. oblečení) v blízkosti elektrod nebo citlivých součástí v elektroencefalografu.

**Síťový brum.** Nežádoucí efekty vznikající zapojením do elektrické sítě. Může se jednat buď o síťový kmitočet 50 Hz, vliv harmonických složek u nedokonale uzemněných elektroinstalací nebo o vliv blízkosti výkonových zařízení jako jsou ledničky, rentgeny a další.

**Šum elektronického zařízení.** V obvodech každého elektronického zařízení může vznikat napěťový a proudový šum, zejména u aktivních prvků jako jsou tranzistory, operační zesilovače nebo diody.

**Magnetická pole.** Vliv nedostatečně odstíněných magnetických polí se projevuje zejména v biomagnetismu.

### 2.4.2 Biologické artefakty

Tyto artefakty vznikají působením procesů uvnitř nebo na povrchu těla zkoumaného subjektu.

**EOG.** Pohyby očí nebo mrkání vede ke změně EOG potenciálu, který se projevuje zejména ve frontálních svodech EEG. Jeho příspěvek se dá částečně nebo úplně eliminovat zavřením očí nebo průběžným stanovováním hodnot EOG potenciálů, jejich váhováním a odečtením od naměřeného EEG. EOG artefakty se frekvenčně projevují jako delta až alfa vlny.

**EKG.** Artefakty rytmu EKG mají stejný kmitočet jako srdeční stahy a projevují se jako hrotové vlny s nízkou amplitudou.

**Dýchací projevy.** Při dýchání dochází k nežádoucím pohybům svodových kabelů a případně i elektrod, což vede ke vzniku artefaktů. Mají pravidelný rytmus s periodou několika sekund a různé amplitudy. Odstranit se dají uložením pacienta do co nejstatičtější pozice [9].



# Kapitola 3

## Měření a zpracování EEG

### 3.1 Elektrody

K snímání elektrického potenciálu z povrchu skalpu se obvykle používá 19 snímacích elektrod a dvě referenční. Většinou jsou fixovány v elastické masce, kterou má měřený subjekt nasazenou na hlavě po celou dobu snímání. Pro lepší kvalitu snímání potenciálů a omezení nežádoucích vlivů okolí nebo pohybů pokožky hlavy, se mezi elektrodu a kůži nanáší vhodná vodivá pasta nebo tkanina namočená ve fyziologickém roztoku. Standardní rozložení elektrod se označuje jako 10-20, což vyjadřuje procentuální vzdálenosti elektrod vůči obvodu hlavy [1], viz Obr. 3.1.

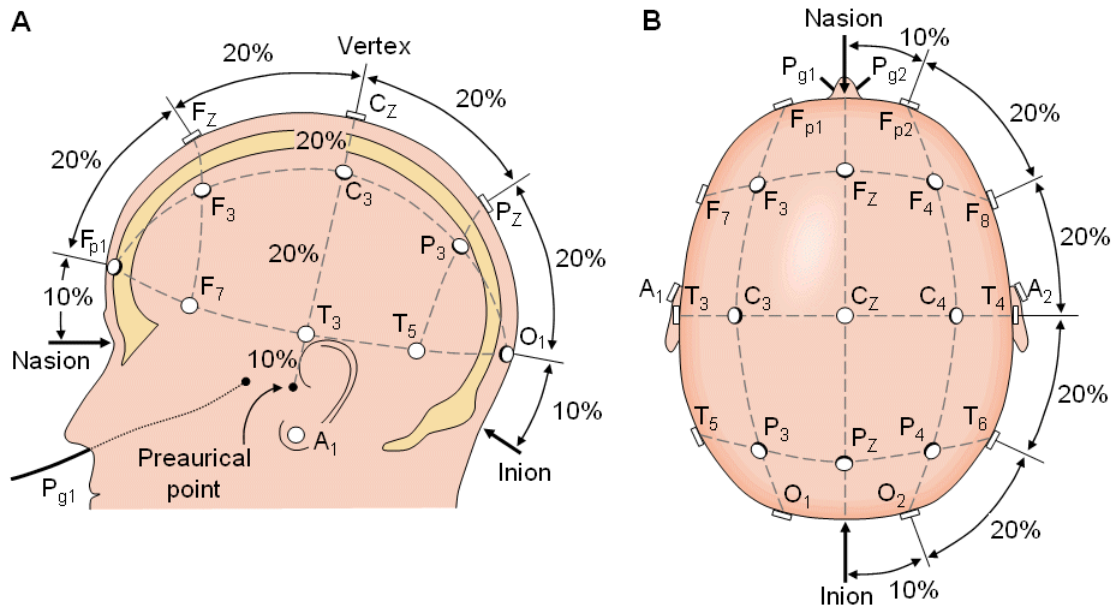
Označení elektrod na schématu vychází z následujícího klíče: První písmeno určuje oblast umístění elektrody, F - Frontální, C - Centrální, T - Temporální, P - Parietální, O - Okcipitální, číselný index pak rozlišuje mezi levou (lichá) hemisférou a pravou (sudá) hemisférou mozku, případně střed (Z - zero). Elektrody  $A_1$  a  $A_2$  jsou referenční na ušních lalůčkách.

Systém, kdy každý snímaný kanál označuje jednu elektrodu, bývá zpravidla ještě doplněn dalším kanálem snímajícím EKG a volitelně ještě dalšími pro jiné biomedicínské signály. Tyto signály pak lze odečíst od EEG a získat tím čistší výsledky [9].

### 3.2 Zapojení elektrod

Nejčastěji se používají tři režimy zapojení elektrod:

**Unipolární.** Používá společnou referenční elektrodu, která může být dvou typů:



Obrázek 3.1: Schéma rozložení elektrod systémem 10-20 [10]

- Ipsilaterální ucho - Snadné rozlišení tranzientů, ale nebezpečí kontaminace referenční elektrody mozkovou aktivitou.
- Propojené ušní elektrody - Redukuje výskyt EEG artefaktů, ale při jejich průniku do jedné z elektrod ovlivní svody obou stran.

**Bipolární.** Vzniká odečtením dvou sousedních elektrod.

**Průměrné.** Nebo také z aj. averaged, jako reference slouží průměr všech elektrod.

### 3.3 Předzpracování signálu

Před samotnou analýzou naměřeného signálu musí signál projít předzpracováním. To se skládá ze tří kroků:

- Zesílení.
- Analogově - digitální převod.
- Filtrace.

V praxi se většinou pro všechny tři úkony používá jediná elektronická součástka. Ta dokáže zesílit naměřené hodnoty EEG z původních mikrovoltů na velikost vhodnou pro další práci se signálem.

Pro počítačové aplikace je nejdůležitější část předzpracování analogově - digitální  $B$ -bitový převodník. A/D převod se skládá ze dvou zásadních kroků, vzorkování a kvantizace.

**Vzorkování.** Vzorkováním se rozumí snímání hodnot spojitého signálu v pravidelných časových úsecích. Převrácená hodnota tohoto časového úseku se nazývá *vzorkovací frekvence*  $f_{vz}$ . K jejímu správnému určení je nutné dodržet *Nyquistův teorém*<sup>1</sup>, který praví: *Aby bylo možné přesně rekonstruovat signál daných charakteristik podle jeho vzorků, je nutné, aby vzorkovací frekvence byla nejméně dvojnásobná oproti maximální frekvenci vzorkovaného signálu, tedy:*

$$f_{vz} \geq 2f_{max} \quad (3.1)$$

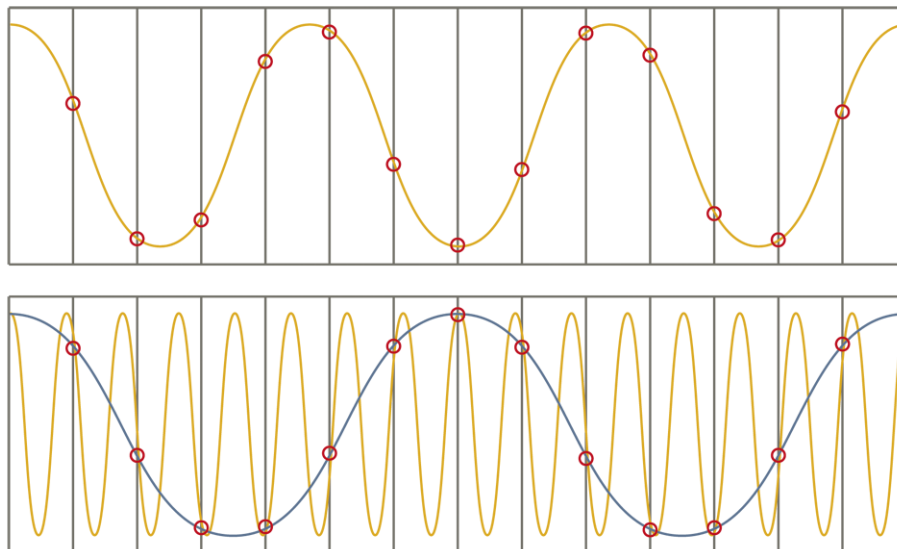
Maximální frekvence může být zajištěna například průchodem analogového signálu předřazeným filtrem s dolní propustí. Pokud není Nyquistův teorém dodržen a vzorkovací frekvence je příliš nízká, může dojít k *aliasingu*. To je jev, kdy spojení vzorkovaných bodů vytvoří jiný než původní signál. Správná a chybná volba vzorkovací frekvence je demonstrována na Obr. 3.2. Ve frekvenční oblasti dochází při aliasingu ke slévání spekter. Pokud je vzorkovací frekvence zvolena příliš vysoká, pak jsou zaznamenány i redundantní vzorky a dochází k zbytečnému zahlcování paměti počítače.

V klinické praxi se používá vzorkovací frekvence do 512 Hz, nejčastěji mezi 70-256 Hz (maximum beta aktivity se nachází kolem 30 Hz).

**Kvantování.** Kvantování je proces, kdy se spojitý rozsah vstupní proměnné rozdělí do  $B^2$  nepřekrývajících se úrovní [1]. Každému vzorku je přidělena diskrétní hodnota odpovídající dané úrovni.

**Digitální filtrace.** Filtrace je proces, kterým prochází signál před začátkem zpracování. Odstraňuje artefakty nebo jiné nežádoucí prvky obsažené v naměřeném signálu. Obecné požadavky na filtry jsou co nejvyšší rychlost a přesnost.

<sup>1</sup>Jinak také Shannonův nebo Kotělnikovův



Obrázek 3.2: Časový aliasing [11]

Pokud není signál měřen pro účely počítačového zpracování, ale pro záznam na papír, používá se za výše popsanými obvody ještě D/A převodník a další zesilovač, který dodává dostatečné napětí pro pohyb zapisovacích hrotů.

### 3.4 Segmentace

EEG je obecně nestacionární signál, při jehož zpracování se využívá lokálních stacionarit. Segmentací se rozumí rozdělení signálu do konstantně nebo různě dlouhých úseků (segmentů), u kterých je žádoucí, aby byly právě stacionární. Konstantní segmentace jako jednoduchý způsob dělení však nedokáže zaručit, že hranice lokálních stacionarit budou shodné s hranicemi segmentů. Splnění této podmínky zajistí užití segmentačních metod s adaptivní periodou, které dokáží odhalit změny stacionarity a do těchto míst lokalizovat hranice segmentů. Více informací o metodách segmentace obsahuje např. [1].

# Kapitola 4

## Shluková analýza

Shluková analýza je nástroj datové analýzy zahrnující řadu matematických metod, které popisují míru podobnosti zkoumaných objektů popsaných sadou obecně vícerozměrných dat. Na základě podobnosti dat k sobě do shluků (aj. *clusters*) přiřazuje objekty tak, aby podobnost mezi objekty ve stejném shluku byla co největší a podobnost mezi jednotlivými shluky co nejmenší. Jedná se o metodu učení bez učitele [12].

Využití se jí dostává v širokém spektru oborů. Nejčastěji tam, kde objekty projevují přirozenou tendenci se seskupovat. V biologii se shluková analýza zabývá numerickou taxonomií<sup>1</sup>, v lékařství identifikací nemocí a jejich stádia. Dokáže najít podobnosti mezi shlukovanými objekty, ale nedokáže rozlišit jejich významné a nevýznamné znaky. Volba znaků musí být provedena na základě pojmových, teoretických a praktických hledisek. Měly by být využity pouze takové znaky, které dostatečně rozlišují mezi objekty, jelikož nesprávné zařazení znaků vede k zahrnutí takových objektů, které mohou mít rušivý vliv na výsledek analýzy [13].

Z matematického hlediska se jedná o hledání takového předpisu vzdálenosti  $d$  pro objekty  $O_i$  a  $O_j$ , pro který platí následující čtyři podmínky [14]:

$$d(O_i, O_j) \geq 0 \quad (4.1)$$

$$d(O_i, O_i) = 0 \quad (4.2)$$

$$d(O_i, O_j) = d(O_j, O_i) \quad (4.3)$$

$$d(O_i, O_j) \leq (O_i, O_h) + (O_h, O_j) \quad (4.4)$$

---

<sup>1</sup>Taxon = systematická jednotka. Taxonomie = vědní kategorie zabývající se identifikací a klasifikací organismů do hierarchické vývojové struktury: **Nadříše/Říše/.../Druh/Poddruh** [25].

Algoritmy výpočtu se podle typu shlukovacího kritéria dají rozdělit na dvě základní skupiny a to nehierarchické a hierarchické.

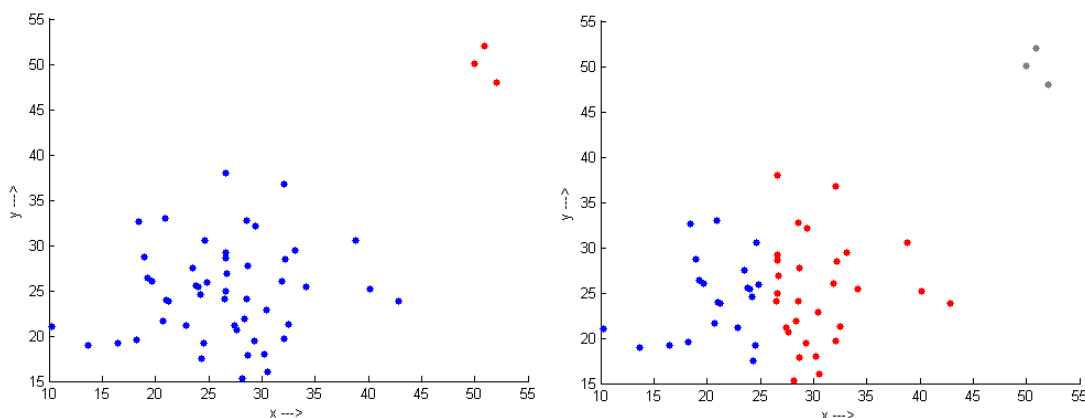
## 4.1 Nehierarchické metody

Nehierarchické metody si kladou za cíl minimalizaci kritériální funkce [1] optimálním rozkladem množiny objektů. Tento rozklad lze chápat jako nejvhodnější klasifikaci objektů [14]. Pro nehierarchické shlukování je třeba předem určit optimální počet  $K$  shluků. Ten lze provést buď kvalifikovaným odhadem podle cíle analýzy nebo lze použít jeden z následujících výpočtů:

- Calinski-Harabasz index
- C index
- Goodman-Cruskal index
- The Banfeld-Raftery index

Jejich definice přesahuje obsah této práce, podrobněji viz [15].

Výsledkem nehierarchického shlukování je  $K$  shluků, pro které platí, že průnikem dvou nebo více shluků vždy vznikne prázdná množina. Tyto metody se dají dále rozdělit na metody s konstantním nebo proměnným počtem shluků. U konstantních metod nedochází ke změně  $K$ , což je nevýhodné, pokud se v sadě dat objevuje množina dat výrazně odlišných. Vztaheno k náplni praktické části této práce se může



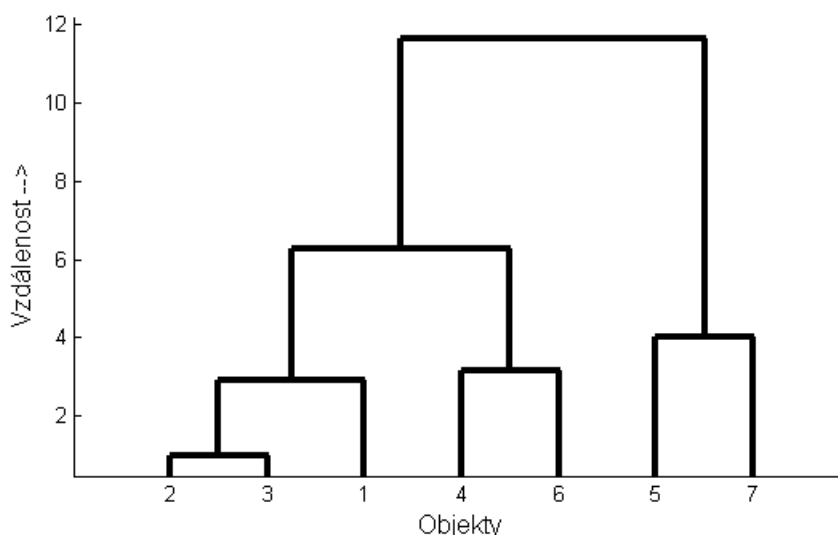
Obrázek 4.1: Nevhodná data v jednom shluku

jednat například o segmenty signálu, které obsahují artefakty. Tato data jsou pak sdružena ve vlastním shluku a odebírají zbytku dat možnost podrobnějšího členění (Obr. 4.1a). Tomu lze předejít použitím některé z metod pracujících s proměnným počtem shluků. Ty dokáží vytvořit nový shluk pro výrazně odlišná data (Obr. 4.1b).

Vybrané metody nehierarchické shlukové analýzy jsou popsány v následujících sekcích.

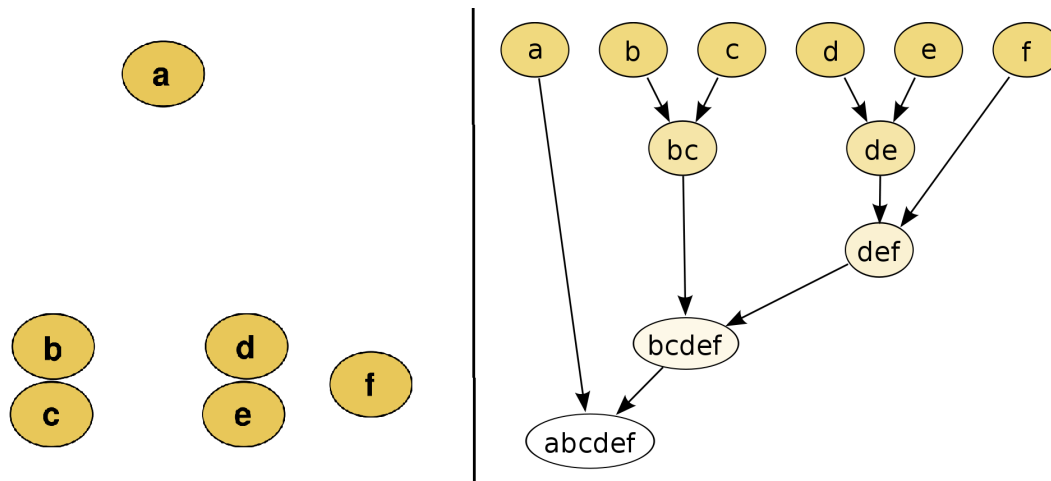
## 4.2 Hierarchické metody

Jak název napovídá, tyto metody se snaží najít určitou hierarchii v sadě dat za pomoci vyhodnocování kritériální funkce. Podobně jako nehierarchické metody, lze i hierarchické rozdělit na dvě skupiny. Podle směru průchodu množinou objektů existují metody *aglomerativní* (zdola nahoru<sup>2</sup>) a *divizní* (shora dolů). Pro  $n$  objektů v množině se jedná o posloupnost  $n-1$  rozkladů, které lze znázornit binárním stromem zvaným dendrogram [16]. Každý  $i$ -tý uzel stromu znázorňuje shluk rozkladu  $\Omega_i$  a svislá vzdálenost mezi rozklady  $\Omega_i$  a  $\Omega_{i-1}$  uvádí vzdálenost mezi shluky, respektive míru jejich vzájemné nepodobnosti. Průnikem výsledných nebo dílčích shluků/množin je buď prázdná množina nebo jeden z nich.



Obrázek 4.2: Příklad dendrogramu

<sup>2</sup>Myšleno z hlediska binárního stromu s listy dole



Obrázek 4.3: Schematický příklad aglomerativního shlukování[27]

### 4.2.1 Aglomerativní metody

Jak je ukázáno na Obr. 4.3, aglomerativní metody začínají v úrovni rozkladu  $\Omega_0$ , kde je každý objekt právě jediným členem svého shluku, tedy množiny shluků jsou:

$$\Omega_0 = \{(a), (b), (c), (d), (e), (f)\}$$

V dalším kroku se spojí podle určitého kritéria nejpodobnější dvojice shluků a tyto spojené shluky se stávají objektem. Vzniká

$$\Omega_1 = \{(a), (bc), (de), (f)\}$$

Analogicky se tvoří rozklady  $\Omega_{i+1}$  až po dosažení konečného  $\Omega_{n-1}$ , který obsahuje všechny objekty původní množiny, v tomto případě

$$\Omega_4 = \{(abcdef)\}$$

Alternativně může být shlukování ukončeno při dosažení požadovaného počtu shluků.

### 4.2.2 Divizní metody

Divizní metody (nebo také divizivní) postupují přesně opačně než aglomerativní, tedy začínají s jedním shlukem, který obsahuje všechny zkoumané objekty. Ten rozdělí podle určitého kritéria podobnosti na dva<sup>3</sup> nové shluky a tento postup opa-

<sup>3</sup>Obecně nemusí být dva, ale i více

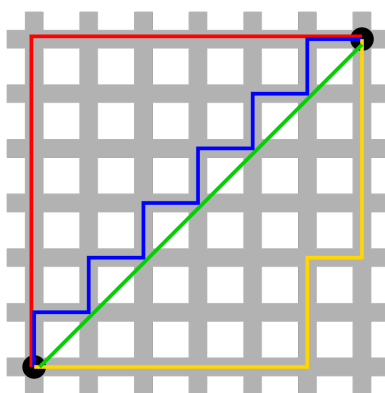


kuje až do konečného rozdělení, kde každému shluku přísluší právě jeden z počáteční množiny, alternativně mohou opět končit při dosažení daného počtu shluků.

### 4.3 Metrika

Úvodem této podkapitoly je třeba definovat metrický prostor. Metrický prostor je uspořádaná dvojice  $(\mathcal{K}, d)$ , kde množina  $\mathcal{K}$  je neprázdná množina objektů a metrika  $d$  je zobrazení  $d : \mathcal{K} \times \mathcal{K} \rightarrow \mathcal{R}$ . Všechna  $d$  musí splňovat podmínky uvedené v úvodu této kapitoly (Rovnice 4.1-4.4).

Níže následuje popis některých základních metrik.<sup>4</sup>



Obrázek 4.4: Srovnání manhattanské a eukleidovské metriky[29]

**Manhattanská** neboli také součtová, city-block nebo taxicab

$$d_T(\vec{x}, \vec{y}) = \sum_{i=1}^n |x_i - y_i| \quad (4.5)$$

**Eukleidovská** nejčastěji používaná

$$d_E(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4.6)$$

Z Obr. 4.6 je jasně patrný rozdíl mezi různými způsoby výpočtu manhattanské (červeně, modře a žlutě) a eukleidovské vzdálenosti.

<sup>4</sup>V metrikách jsou  $n$ -rozměrné vektory  $\vec{x}$  a  $\vec{y}$  vektory souřadnic zkoumaných objektů

Minkovského

$$d_M(\vec{x}, \vec{y}) = \sqrt[z]{\sum_{i=1}^n |x_i - y_i|^z} \quad (4.7)$$

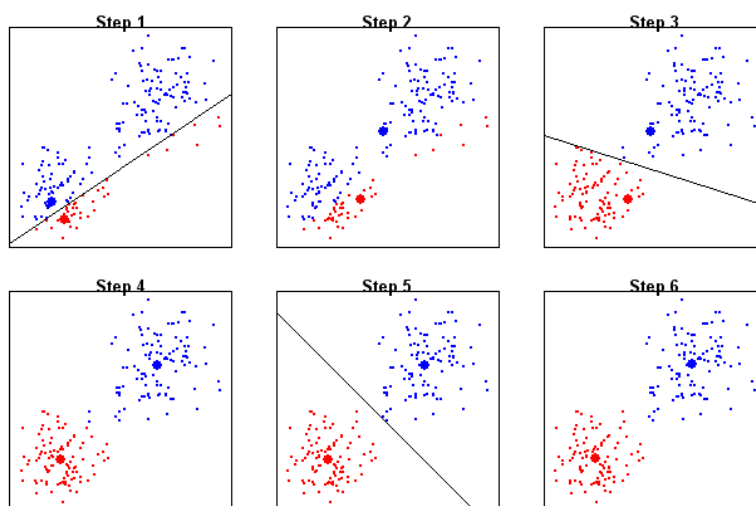
Parametr  $z$  určuje zdůraznění rozdílu mezi vzdálenými objekty. Pro  $z = 1$  se Minkovského metrika stává manhattanská a pro  $z = 2$  eukleidovská.[13]

## 4.4 Výběr shlukovacího kriteria

### 4.4.1 K-means

Nebo také MacQueenova metoda podle jejího autora, nehierarchická metoda.

K-means neboli metoda K-středů má definovanu kriteriální funkci jako eukleidovskou vzdálenost objektů od těžiště shluku. Jedná se o optimalizační iterativní metodu, která probíhá v několika opakujících se krocích, jak lze názorně vidět na Obr. 4.5 pro  $K = 2$ .



Obrázek 4.5: Algoritmus K-means[26]

Tyto kroky jsou:

1. Volba těžišť  $K$  shluků. Probíhá buď náhodně nebo vhodnou heuristikou.
2. Přiřazení objektů do shluků podle minimální vzdálenosti k těžišti.
3. Spočítání nových těžišť.

Opakováním kroků 2. a 3. se optimalizuje rozložení objektů ve shlucích. Nejlepšího rozložení je dosaženo, pokud jsou nově vypočtená těžiště shodná s předchozími. Problémem u této metody je, že vždy vyhledává pouze lokální optima, což může vést k neoptimálnímu výsledku.

#### 4.4.2 Metoda průměrné vzdálenosti

Average linkage, hierarchická metoda.

Vzdálenost dvou shluků se počítá jako průměr z možných mezishlukových vzdáleností dvou objektů, kdy se mezishlukovou vzdáleností objektů rozumí vzdálenost  $d$  mezi všemi objekty jednoho a všemi objekty jiného shluku. Nejbližší jsou shluky, které mají nejmenší průměrnou vzdálenost  $d$ .

#### 4.4.3 Metoda těžišť

Centroid linkage, hierarchická metoda.

Vzdálenost  $d$  se počítá jako eukleidovská vzdálenost shlukových těžišť. Nejbližší jsou ty shluky, jejich vzdálenost  $d$  je nejmenší. Při počítání s různě velikými shluky může u této metody docházet k jevu, že těžiště většího shluku bude mít větší „váhu“ než těžiště menšího. Toto odstraňuje její vylepšení, mediánová metoda.

#### 4.4.4 Metoda mediánová

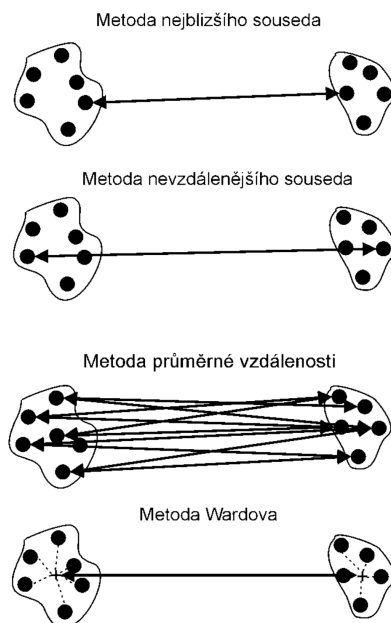
Median linkage, hierarchická metoda.

Vylepšuje centroidní metodu odstraněním rozdílné váhy různě velkých shluků tím, že jako kritérium vzdálenosti používá nejmenší vzdálenost mediánu shluku.

#### 4.4.5 Metoda nejbližšího souseda

Single linkage, hierarchická metoda.

Kritérium vzdálenosti je minimální mezishluková vzdálenost  $d$ . Vytváří nový shluk na základě  $d$  dvou nejbližších objektů, což může zejména u špatně separovaných shluků vést k řetězení, až k mylným závěrům.



Obrázek 4.6: Nejčastěji užívané metody shlukování[13]

### 4.4.6 Metoda nejvdálenějšího souseda

Complete linkage, hierarchická metoda.

Je podobná metodě nejbližšího souseda s tím rozdílem, že jako kritérium vzdálenosti nebere nejmenší vzdálenost  $d$ , ale naopak největší. Nepodobnost shluků je určována nepodobností dvou nejvdálenějších objektů z různých shluků.

### 4.4.7 Metoda Wardova

Ward linkage, hierarchická metoda.

Wardova metoda počítá v každém kroku eukleidovskou vzdálenost každého objektu shluku od jeho těžiště a shluky spojuje na základě minimální hodnoty součtu čtverců těchto vzdáleností. Metoda časově nejnáročnější[13],[17].

### 4.4.8 MacNaughton–Smithova metoda

MacNaughton–Smith method, hierarchická metoda.

Divizní metoda. Snaží se, byť za cenu možnosti neoptimálního výsledku, snížit náročnost divizního postupu. V praktické části této práce nebyla implementována, zájemci o podrobnější informace viz například [18].

#### 4.4.9 Další metody

Metod samozřejmě existuje více, než je uvedeno v této práci. Jedná se buď o odvozeniny předešlých metod jako například Fuzzy C-Means, což je fuzzy-logická varianta K-means, nebo o samostatné metody, například učící se klasifikátory. Metoda vhodná pro práci s velkodimenzionálními daty je například metoda CURE. Její podrobný popis však přesahuje rozsah této práce, více o ní v [19].

### 4.5 Zhodnocení metod shlukování

V minulosti byly hierarchické metody populárnější a zejména Wardova a průměrová metoda byly považovány za nejlepší techniky. Kvůli náročnosti  $O(n^2)$  až  $O(n^2 \log(n))$  jejich komplexita roste s množstvím příznaků shlukovaných objektů velmi rychle a tím se stávají nevhodnými pro práci s velkou sadou dat.[24]

Naopak v poslední době se do popředí dostávají nehierarchické metody. Jejich kvalita záleží na schopnosti uživatele vybrat optimální počet shluků.

Poslední stále užívanější metodou je kombinace nehierarchického i hierarchického shlukování. Na celou sadu dat se nejdříve použije vybraná nehierarchická metoda, která vytvoří shluky, na které je po jednom aplikována vhodná hierarchická metoda.



# Kapitola 5

## Program PSG-HC

Tato kapitola popisuje práci s programem PSG-HC pro vizualizaci hierarchické shlukové analýzy biomedicínského signálu. Program je kompletně vytvořen v prostředí MATLAB za použití vývojářského nástroje MATLAB GUIDE. Program může načíst datové sady více kanálů, u kterých z principu nezáleží, o jaké z biomedicínských signálů se jedná, ale najednou zvládne pracovat pouze s jedním signálem.

### 5.1 Vstupní soubory

Program pracuje se vstupními soubory ve formátu odpovídajícím PSGLab toolboxu. Pro spuštění samotné analýzy je třeba v dialogovém okně vybrat postupně tři soubory, které se všechny vztahují ke stejnému signálu. Z menu Data lze dále volitelně vybrat čtvrtý soubor, který definuje druhy příznaků segmentů. Všechny soubory určitého kanálu program spojí do jedné datové sady, se kterou nadále pracuje jako s celkem. Pro demonstraci práce programu jsou názvy souborů uvedeny v následujícím tvaru<sup>1</sup>:

- Channel-xxx.mat - Časově závislé hodnoty signálu
- Segments-xxx.txt - Hranice segmentů
- Features-xxx-data.txt - Vyčíslené příznaky
- Features-desc.txt<sup>2</sup> - Popis příznaků

---

<sup>1</sup>„xxx“ jsou tři čísla, kde první označuje sadu dat (0 pro novorozenecká data, 1 pro data s artefakty podle podkapitoly 5.5) a další dvě pořadí signálu v rámci sady

<sup>2</sup>Volitelný

### 5.1.1 Hodnoty signálu

*Channel-xxx.mat* - zde jsou uloženy číselné hodnoty naměřeného signálu ve tvaru MATLAB Cell Array.

### 5.1.2 Hranice segmentů

*Segments-xxx.txt* - tento textový soubor obsahuje dva sloupce celých čísel reprezentující začátky a konce jednotlivých segmentů signálu. Program je přizpůsoben k práci s různě dlouhými segmenty a proto nevyžaduje segmentaci s konstantní periodou, ale dokáže pracovat i s adaptivní segmentací.

Při načítání souboru s mezemi segmentů program kontroluje dvě kritéria, s jejichž pomocí dokáže odhalit nesprávný soubor. Tato kritéria jsou:

- Počet sloupců musí být dva.
- Maximální hodnota hranic segmentů musí být rovna počtu prvků v souboru s číselnými hodnotami.

Pokud jsou obě podmínky splněny, program vyzve uživatele k vybrání souboru *features-xxx-data.txt* obsahující vypočtené příznaky. V opačném případě se opakuje vybrání souboru s hranicemi segmentů.

### 5.1.3 Vyčíslené příznaky

*Features-xxx-data.txt* je textový soubor obsahující  $[m \times n]$  číselných hodnot, kde  $m$  je počet segmentů a  $n$  počet příznaků pro každý z nich. Program k výpočtu implicitně využívá všechny příznaky, pokud uživatel nerozhodne jinak (viz 5.2.2).

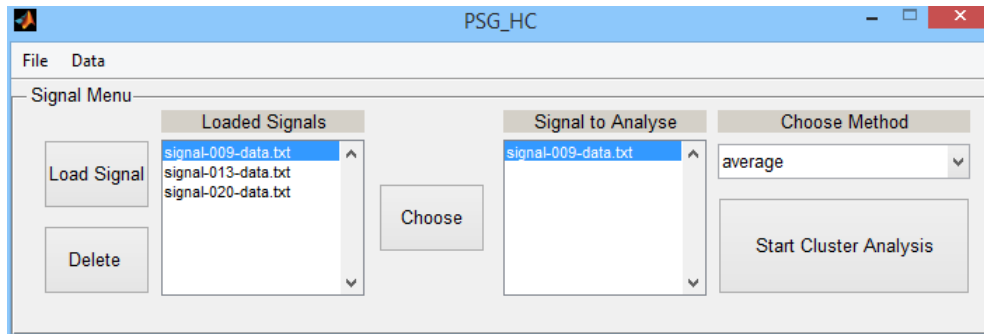
### 5.1.4 Popis příznaků

*Features-desc.txt* obsahuje názvy jednotlivých příznaků ze souboru *Features-xxx-data.txt* a jeho výběr není povinný. Výchozí soubor *features-desc.txt* je umístěn v adresáři programu.

## 5.2 GUI

Grafické uživatelské prostředí se skládá ze dvou základních a několika doplňkových oken.



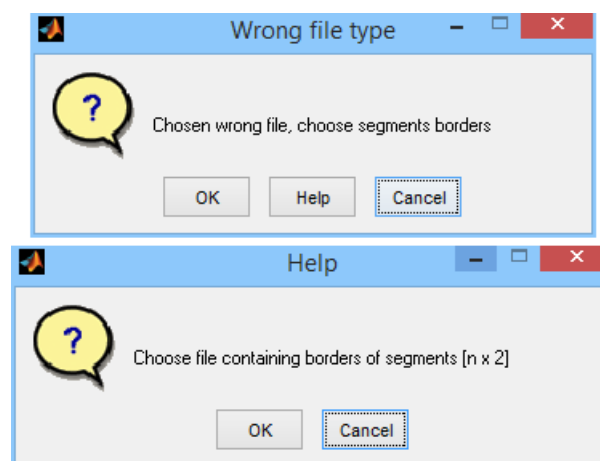


Obrázek 5.1: Úvodní okno programu po načtení signálů

### 5.2.1 Úvodní okno

Po spuštění programu se objeví okno z Obr 5.1. Toto obsahuje čtyři tlačítka, jedno pop-up menu a dvě toolbarová menu.

**Load Signal:** Vyvolá okno pro načtení zdrojových souborů. Nejdříve je třeba vybrat soubor s číselnými hodnotami signálu, poté s hranicemi segmentů a nakonec s hodnotami příznaků. Pokud dojde k výběru nesplňujícímu kritéria z bodu 5.1.2, vyvolá se chybová hláška s požadavkem na vybrání nového souboru, viz Obr. 5.2.1. Po načtení správných souborů se v levém listboxu zobrazí položka odpovídající danému signálu. Takto lze načíst téměř neomezeně mnoho signálů, jejichž jména se všechna budou zobrazovat v tomto listboxu.



Obrázek 5.2: Chybová hláška pro špatný soubor segmentů a okno s nápovědou

**Delete:** Odstraní signál označený v levém listboxu.

**Choose:** Vybere signál označený v levém listboxu a zobrazí jeho název v pravém listboxu. Tento signál se stane předmětem samotné shlukové analýzy.

**Start Cluster Analysis** Vyvolá okno samotné analýzy a spustí ji.

**Choose Method:** Rozbalovací menu, v němž lze zvolit požadovanou variantu hierarchické shlukové analýzy, zvané *linkage*. Nabídka metod odpovídá hierarchickým aglomerativním metodám popsaných v 4.4.2:

- Average
- Centroid
- Complete
- Median
- Single
- Ward

Plus ještě navíc metodu Weighted - WPGMA, její popis například v [22].

**File** - V tomto menu lze spustit program od začátku nebo jej zavřít.

**Data** - Pomocí této nabídky lze exportovat data nebo vybrat soubor s popisem příznaků z bodu 5.1.4.

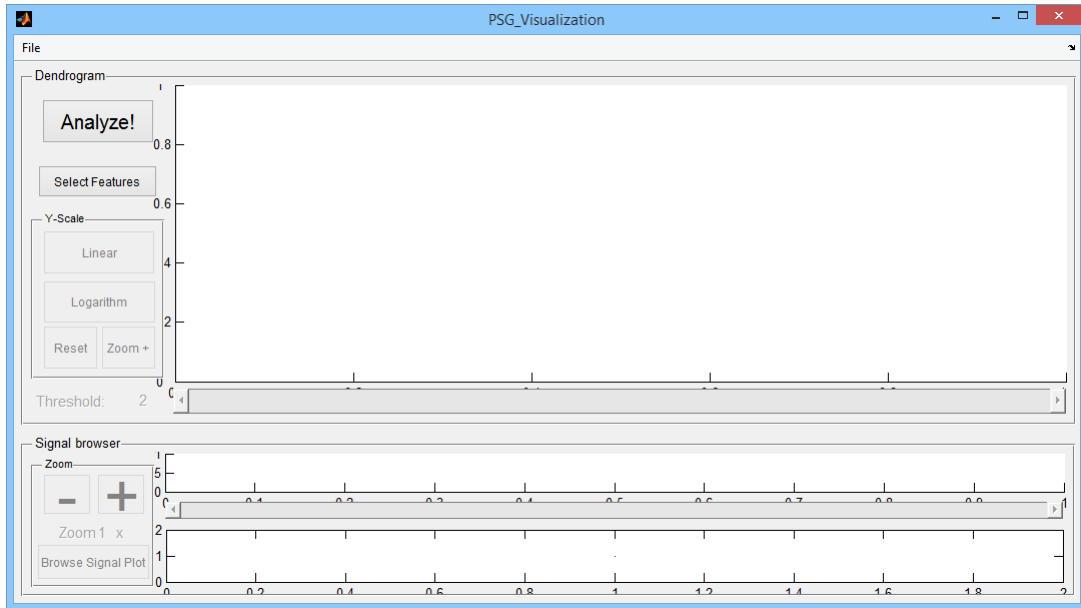
### 5.2.2 Okno analýzy

Program si klade za cíl v tomto okně (Obr. 5.3) zobrazit co nejpřehledněji výsledky shlukové analýzy. Okno je rozděleno na dvě části pojmenované *Dendrogram* a *Signal Browser*. V části *Dendrogram* se nachází následující prvky:

**Analyze!** Po stisknutí tohoto tlačítka se provede shluková analýza za použití dříve vybrané shlukovací metody. Do výpočtu vzdáleností objektů<sup>3</sup> se zahrnou implicitně všechny změřené a spočítané příznaky, pokud uživatel nevybere jinak, viz dále.

---

<sup>3</sup>Zde představující segmenty

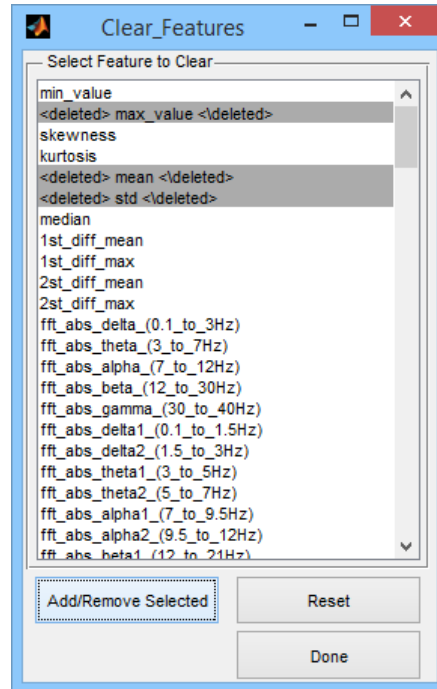


Obrázek 5.3: Okno s připravenými grafy pro vizualizaci

**Select Features:** Vyvolá okno se seznamem příznaků. Zde je možné vybrat příznaky, které se dle názoru uživatele nehodí k zařazení do analýzy. Tyto vyřazené příznaky se označí prefixem <deleted> a sufixem < \deleted>, jak lze vidět na Obr. 5.4.

**Dendrogram s posuvníkem:** Hlavním způsobem zobrazení výsledků analýzy je dendrogram. Vodorovná modrá přerušovaná čára reprezentuje hranici (práh, threshold) obarvení zobrazených shluků. Její úroveň se nastavuje posuvníkem umístěným pod grafem dendrogramu. Samotným dendrogramem jsou reprezentovány výsledky hierarchické shlukové analýzy. Zobrazena je vždy celá struktura od kořene až po listy. Pomocí změny úrovně prahu může uživatel dosáhnout uříznutí části pod prahem a zobrazit tak požadovaný počet shluků, kde každý z nich pak může obsahovat více než jeden objekt. Všechny objekty v daném shluku jsou obarveny stejnou barvou. Pokud dojde k situaci, že některý shluk obsahuje pouze jeden objekt, je tento obarven černě.

**Y-Scale:** Tento box slouží k přehlednějšímu zobrazení dendrogramu pomocí úprav měřítka a limit svlé osy. Tlačítka **Logarithm** a **Linear** se přepíná měřítko svlé osy na logaritmické, respektive lineární. Tato funkce byla implementována do programu kvůli skutečnosti, že při zobrazení více než cca 200 objektů,



Obrázek 5.4: Seznam příznaků

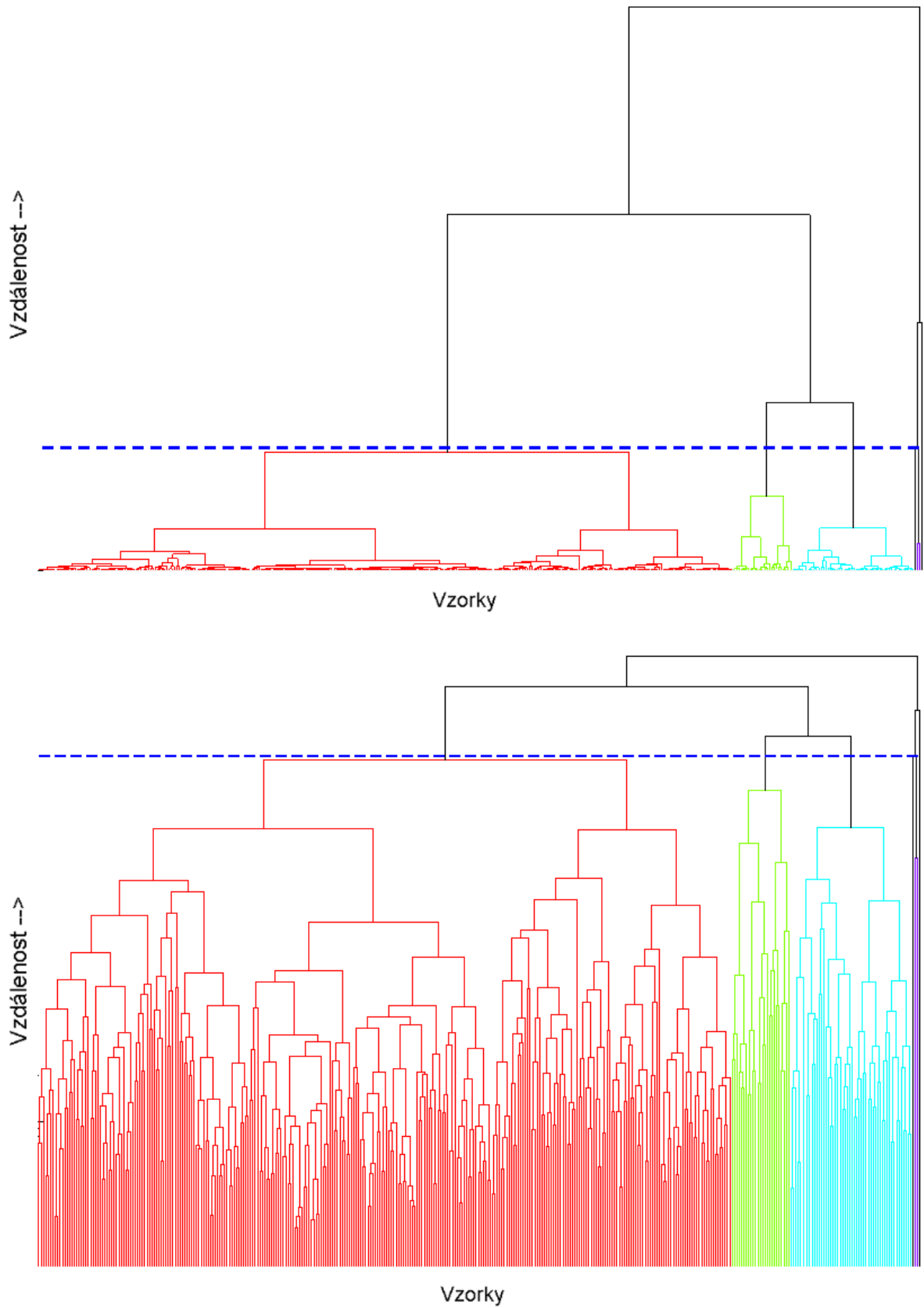
tedy při této aplikaci obvyklého počtu, dochází ke splývání jednotlivých listů dendrogramu. Při přepnutí do logaritmického režimu se listy stávají výraznějšími a tedy přehlednějšími, jak jde vidět na Obr. 5.5. Tlačítko **Zoom+** slouží k posunutí limit zobrazení grafu. V lineárním režimu snižuje horní hranici a v logaritmickém zvyšuje dolní. Tlačítko **Reset** pak vrací zobrazení do původního stavu.

V části pojmenované *Signal Browser* (Obr. 5.6) se nachází další dva grafy s časově závislými způsoby zobrazení výsledků analýzy a několik dalších ovládacích prvků.

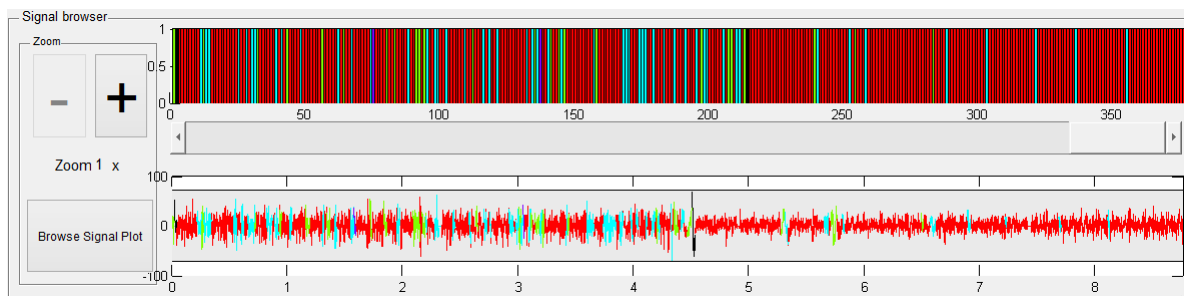
**Graf signálu:** Ve spodním grafu je zobrazen naměřený signál se segmenty obarvenými dle výsledků analýzy. Šedě orámovaná část se zobrazuje v horním segmentovém grafu.

**Segmentový graf:** V horním není pro lepší přehlednost zobrazen průběh signálu, ale segmenty jsou reprezentovány barevnými obdélníky.

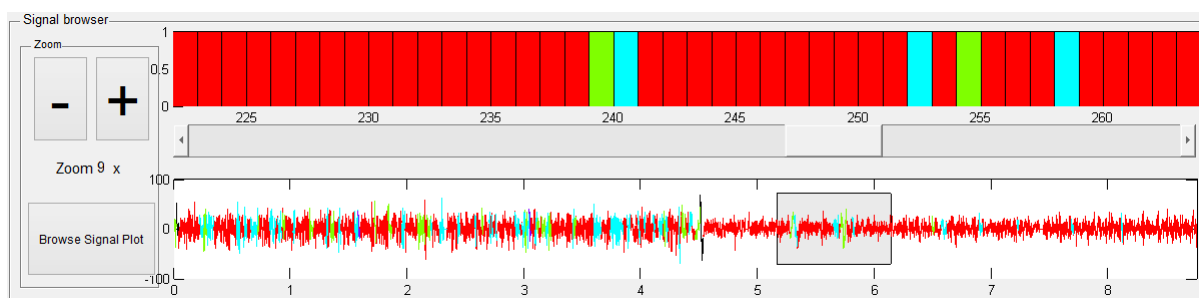
**Zoom:** Pomocí tlačítek v boxu Zoom lze v segmentovém grafu zobrazit detail na posuvníkem vybranou část signálu.



Obrázek 5.5: Dendrogram s lineárním a logaritmickým měřítkem s prahem 6



Obrázek 5.6: Barevné odlišení segmentů v časových grafech



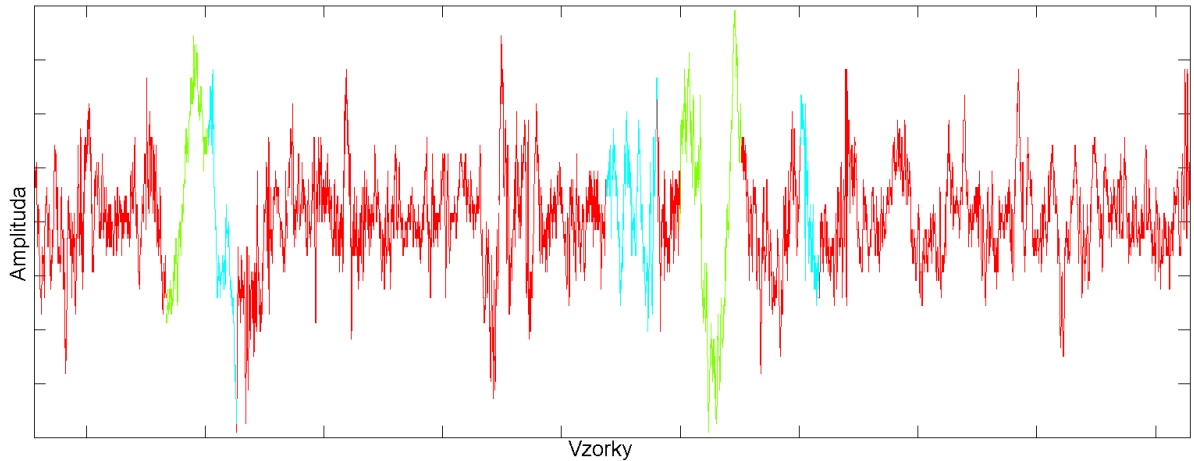
Obrázek 5.7: Ilustrace použití zoomu

**Browse Signal Plot:** Poslední ovládací prvek v této části, po jeho stisknutí se otevře klasické MATLAB Figure okno s výřezem vybrané části signálu, opět obarvené, jak ukazuje Obr. 5.8. Všechny ovládací prvky v tomto okně jsou aktivní a uživatel tak může libovolně procházet a zobrazovat tento signál.

Dále se v tomto okně nacházejí dvě toolboxová menu, *File* s možností vytisknutí výsledků a uzavření okna a *Data*, které umožňuje exportovat výsledky shlukové analýzy.

### 5.3 Export výsledků

Export výsledků probíhá z nabídky *Data* -> *Export* v okně vizualizace. Po vybrání této funkce je uživatel dotázán na jméno a umístění souboru, do kterého se mají výsledky uložit.



Obrázek 5.8: Výřez signálu

### 5.3.1 Formát výstupních dat

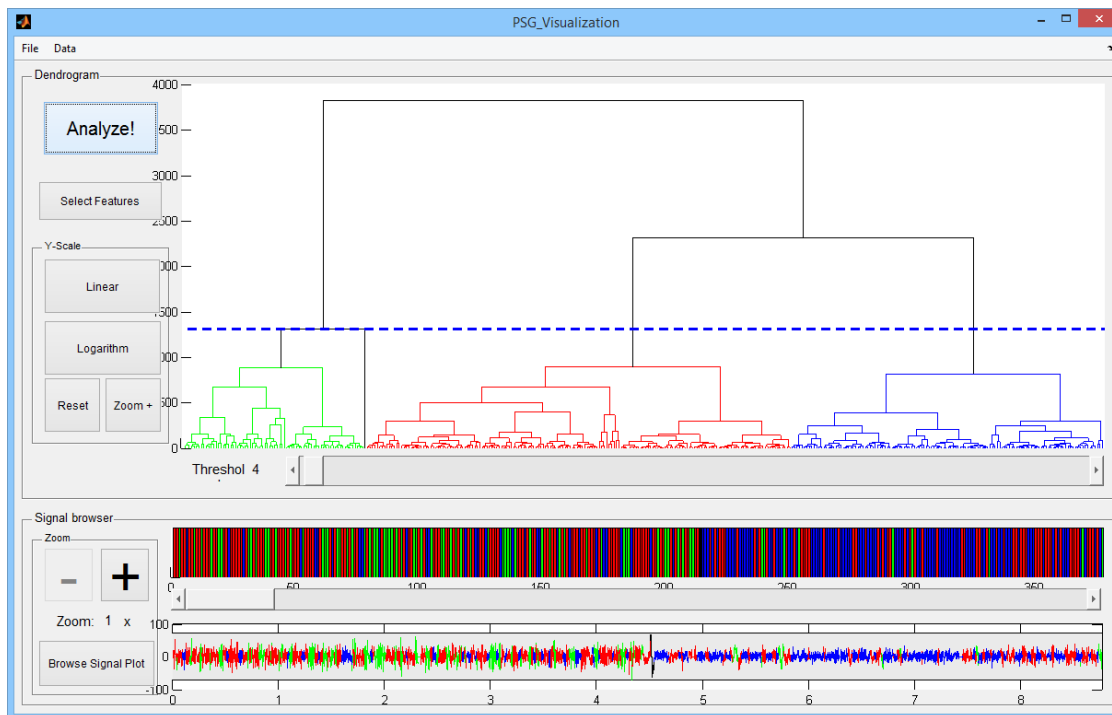
Aby program poskytl uživateli co největší užitek z výsledků, exportuje maximální množství dat, které se snaží seřadit co neoptimálněji. Nejdříve jsou do pole s příznaky přiřazeny dva sloupce s hranicemi segmentů a vzniká tak nové pole příznaků o velikosti  $[m \times n+2]$ . Toto pole je uloženo v řádcích 2 až  $m + 1$ . Na první řádek se zapíše hodnota  $m$ , která po otevření exportovaného souboru slouží jako vodítko určující počet segmentů. Dále se do řádků  $m + 2$  až  $2m$  uloží proměnná **TREE** a do posledního řádku číselné hodnoty signálu ze souboru Channel-xxx.mat.

## 5.4 Popis funkcí využívaných programem

### 5.4.1 Vytvoření dendrogramu

Ke spočítání binárního stromu slouží funkce *linkage*[22] s názvem shlukovací metody zadávaným jako nepovinný parametr. Pokud není zadán, je vybrána metoda nejbližšího souseda. Implementace této funkce počítá s Eukleidovskou metrikou. Strom pak slouží jako vstup funkce *dendrogram*[23]. Dalšími vstupy jsou číselné parametry **P** a **threshold**. Výstupem funkce *dendrogram* jsou tři matice **T**, **H** a **outperm**.

**P:** Číslo určující konečný počet listů dendrogramu. Pokud není zadán, implicitně je nastaven na číslo 30. Pokud je zadán jako 0, počet listů se rovná počtu zkoumaných objektů, což je případ, se kterým tento program pracuje.



Obrázek 5.9: Celkový pohled na analýzu s prahem 4

**Threshold:** Udává hranici, práh, pod kterým má každý shluk unikátní barvu. Jeho hodnota je reálné číslo a znamená vzdálenost od nultého rozkladu. Pro aplikaci v tomto programu není tato definice vhodná, protože by uživatel musel zadávat maximální vzdálenost, po kterou chce jednotlivé listy obarvovat. Tuto vzdálenost většinou nezná. Proto je definice mírně upravena. Práh je spočítán jako  $k-2$ . řádek ve třetím sloupci matice **TREE**, kde  $k$  je počet shluků, které chce uživatel zobrazit.

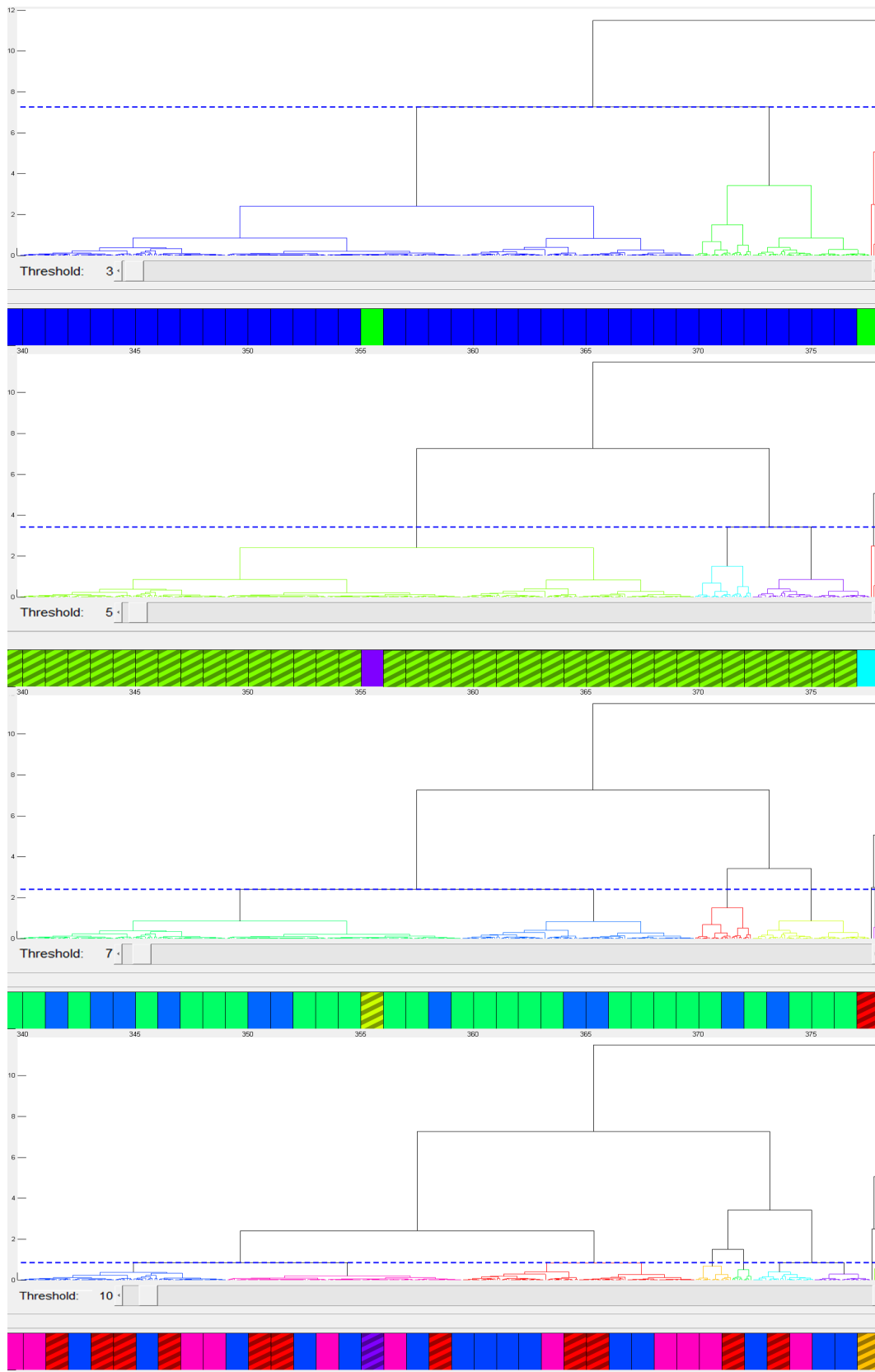
**T:** Matice obsahující  $m$  čísel listů dendrogramu, které odpovídají pořadí objektů v původní sadě. V tomto případě se neuplatní, jelikož každý list obsahuje právě jeden objekt.

**H:** Matice obsahující handles<sup>4</sup> objektů v dendrogramu.

**Outperm:** Matice obsahující pořadí vykreslených listů zleva doprava.

<sup>4</sup>handle je identifikátor objektů se kterými pracuje grafická nadstavba MATLABu. Žádný český ekvivalent se běžně neuzívá.





Obrázek 5.10: Znázornění zvyšování počtu shluků

### 5.4.2 Nastavení prahu

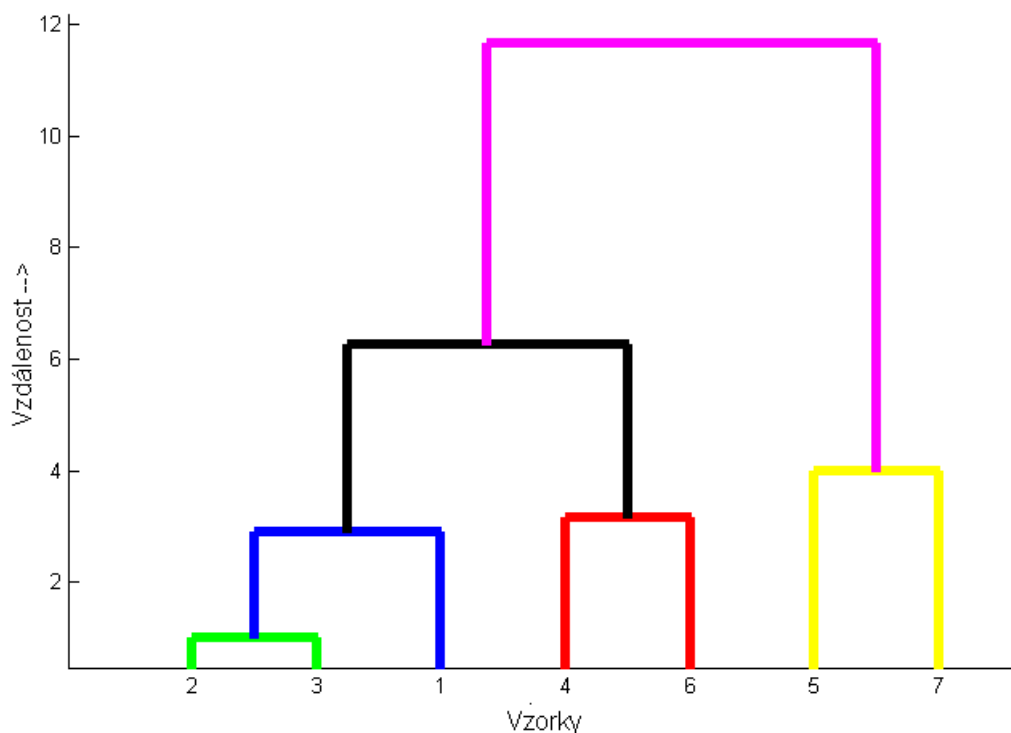
Hodnota prahu je důležitým parametrem definujícím vzhled výsledků analýzy. Při každé její změně dochází k překreslení a přebarvení všech grafů. Obr. 5.10 ukazuje, jak se mění dendrogram a přiblížená koncová část segmentového grafu při posunu prahu z hodnoty 3 na hodnotu 10.

Obrázek je oproti skutečnosti mírně upraven. Určité segmenty jsou šrafovány. Jsou to ty, u kterých při zvýšení hodnoty prahu nedošlo proti původní hodnotě k dělení. Program samotný možnost šrafování neobsahuje, v ostatních ohledech vypadají výsledky totožně.

Zmínku si zaslouží také obarvení objektů v segmentovém grafu. Nabízejí se dvě varianty, jak jej řešit. V programu použitá varianta je přebírání barev, které vytváří funkce MATLABu *dendrogram*. Její nevýhoda je zřejmá z Obr. 5.10. Shluky si při změně hodnoty prahu neuchovávají svou barvu, ale každému objektu jim přiřazena barva odpovídající barvě stejného shluku v dendrogramu. Druhá varianta by byla implementovat řešení, které u neměnicích se shluků zachová barvu, což by vedlo k nesouvislosti s barvou v dendrogramu. Změna tvorby barev ve funkci *dendrogram* by si vyžádala zásadní zásahy do jejího kódu.

### 5.4.3 Získání barev listů dendrogramu

Aby bylo možné obarvit segmenty v segmentovém grafu i v grafu signálu stejnou barvou, jako jsou obarveny v dendrogramu, je třeba zjistit barvy listů. Zjištění barvy objektu v grafu se provede příkazem `get(handle, 'color')`. Problém nastává ve skutečnosti, že matice **H** neobsahuje handles pouze k listům, ale ke všem objektům, kterých je **P**-1. Každý je ve tvaru obráceného „U“, jak je znázorněno na Obr. 5.11. Na něm je každý objekt obarven unikátní barvou. Lze si všimnout, že zeleně, červeně a žlutě jsou obarveny objekty, jejichž oba konce jsou listy. Modře je obarven objekt, který obsahuje pouze jeden list a černý a fialový objekt nemá list žádný. Skutečnost, že z matice **H** nelze odlišit objekty obsahující listy, se dá obejít jednoduchým způsobem. Příkaz `get(handle, 'ydata')` zjistí y-ové souřadnice konců objektů. Pokud je alespoň jedna z nich rovna nule, jedná se o list. Dále lze obdobným voláním příkazu `get(handle, 'xdata')` zjistit x-ové souřadnice pro zjištěné listy a příkazem `get(handle, 'Color')` zjistit jejich barvu. Barvy a x-ové souřadnice listů jsou uloženy do řádků matice **xdc** v pořadí odpovídajícímu pořadí segmentů.

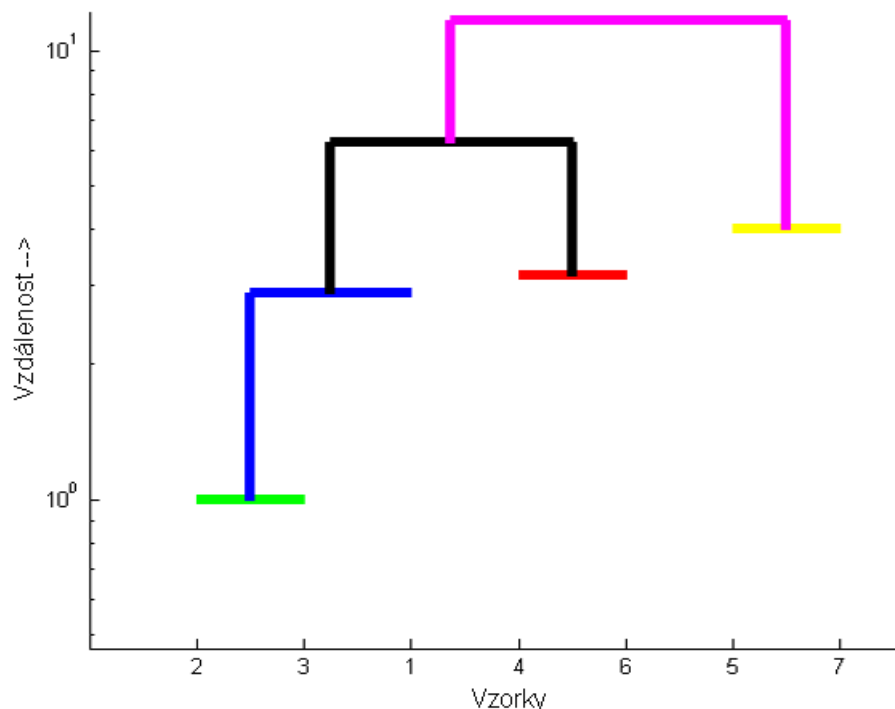


Obrázek 5.11: Demonstrace obarvení objektů v grafu dendrogramu

#### 5.4.4 Logaritmické měřítko

Při programování funkce pro změnu měřítka y-ové osy došlo ke zjištění nekonzistentnosti teorie a praxe grafické nadstavby MATLABu. Lze ji demonstrovat na stejném případě jako znázorňuje Obr. 5.11. Po manuálním přepnutí na logaritmické měřítko se stejný dendrogram zobrazí jako ukazuje Obr. 5.12. Z něj je zřejmé, že ani jedna y-ová hodnota červeného, žlutého, modrého, ani zeleného objektu se nerovná nule. Tomu neodpovídá volání `get(handle, 'ydata')`, které vrací pro konce objektů nulové hodnoty. V důsledku dochází k zneprůhlednění dendrogramu, které se ještě prohlubuje s jeho narůstajícími horizontálními rozměry. Důsledek této chyby lze odstranit vynuceným přičtením nenulové konstanty vhodné velikosti<sup>5</sup> k y-ovým souřadnicím objektů. Po této úpravě výsledný dendrogram vypadá obdobně jako na Obr. 5.5b.

<sup>5</sup>Empiricky se jeví dostatečná hodnota konstanty v řádu setin jednotek.



Obrázek 5.12: Příklad dendrogramu s logaritmickým měřítkem

### 5.4.5 Vykreslení segmentového grafu

Vykreslení segmentového grafu je zajištěno jedním `for` cyklem, který prochází řádky matice `xdc` a používá funkci `patch(X,Y,Color)`, kde `X` a `Y` jsou čtyřprvkové vektory určující vodorovné, respektive svislé hranice polygonu, v tomto případě obdélníka, `Color` určuje barvu segmentu, která se zjišťuje z příslušného řádku matice barev. Svislá velikost polygonu není důležitá, vždy zabírá celou výšku grafu. Pozice polygonů na x-ové ose odpovídá pořadí segmentů v signálu.

### 5.4.6 Vykreslení grafu signálu

Vykreslení grafu signálu probíhá obdobně jako vykreslení segmentového grafu s tím rozdílem, že bere v úvahu tvar signálu. `For` cyklus prochází vektor `outperm`. Pro každou jeho hodnotu z matice segmentů zjistí hranice odpovídajícího segmentu a odpovídající barvu z matice `xdc`. Touto barvou pak vykreslí hodnoty signálu v daných mezích segmentů.

## 5.5 Otestování na klinických datech

Testování programu probíhalo na dvou sadách dat:

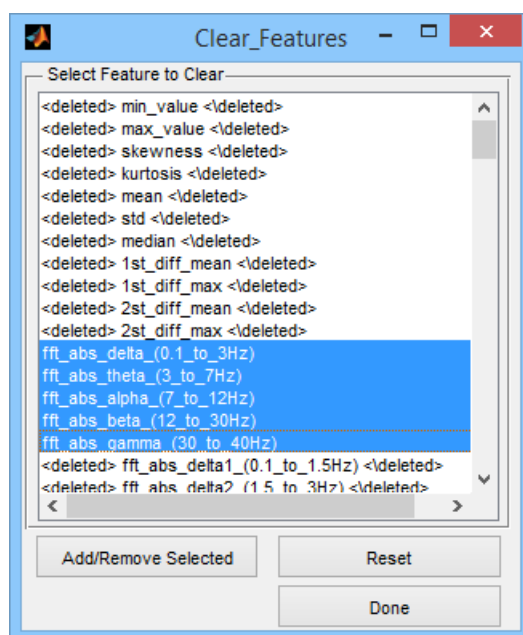
**Novorozenecké EEG.** První sada se skládá z osmi naměřených EEG signálů. Jedná se o reálná data snímaná z hlavy novorozence. Každý signál obsahuje 87554 vzorků, což při vzorkovací frekvenci  $f = 256$  Hz odpovídá necelým 6 minutám záznamu. Každý z kanálů byl adaptivní segmentací rozdělen na 348 až 451 segmentů a pro každý z nich bylo spočítáno 157 příznaků, jejichž jména jsou uvedena v tabulce v **příloze A** a vysvětlena v [20].

**Data s artefakty.** Druhá sada dat obsahuje 19 EEG signálů. Záznamy jsou kratší než v první sadě, obsahuje pouze 11745 vzorků a byly adaptivní segmentací rozděleny do 70 segmentů, opět se 157 spočítanými příznaky. Tyto záznamy byly vybrány cíleně kvůli obsahu artefaktů dvou druhů. Očních a pohybových. Dále je k dispozici expertní ohodnocení záznamu rozdělující segmenty na tři druhy podle příslušnosti ke skupině záznamu relevantní aktivity, očních artefaktů nebo pohybových artefaktů.

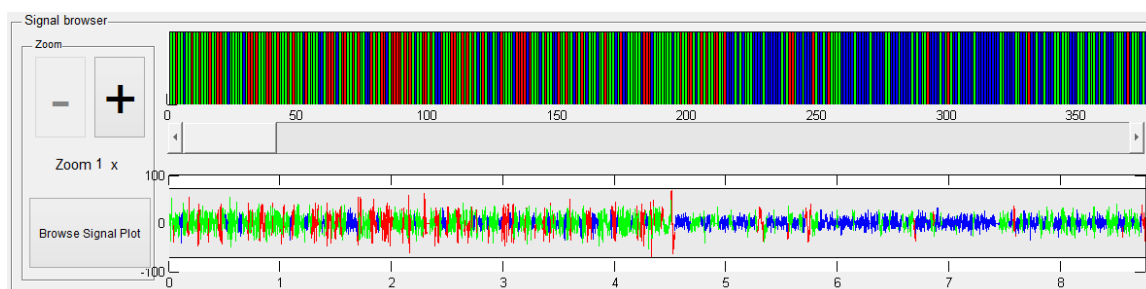
### 5.5.1 Demonstrace práce programu při detekci paradoxního spánku

Charakteristickou vlastností první výše zmíněné datové sady je fakt, že jedna část EEG záznamu reprezentuje aktivní (paradoxní) spánek, zatímco druhá klidný spánek. V této podkapitole je popsán test, zda si program dokáže poradit s rozlišením těchto dvou fází spánku.

Jak je popsáno v kapitole 2.3.1, jedna z určujících vlastností aktivního spánku je posun EEG do vyšších frekvenčních pásem, která odpovídají zejména alfa a beta rytmu. Program tedy dostane pokyn, aby pracoval pouze s vlastnostmi závislými na změřené frekvenci, jak demonstruje obrázek 5.13. Očekávaný výsledek je zřejmý. Záznam by se měl rozdělit na dvě fáze, kde každá bude obsahovat charakteristickou barvu nebo barvy. Jak lze vidět na Obr. 5.14, už při dělení na tři shluky je zřejmé, že v první přibližně polovině záznamu dominuje zelená barva s občasným střídáním červenou a řídkým výskytem modré barvy. Naopak v druhé polovině záznamu je výrazně dominantní modrá barva a červená se nevyskytuje téměř vůbec.



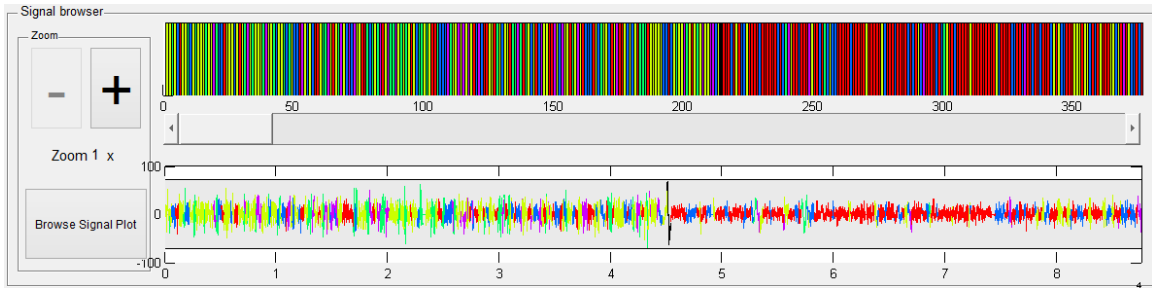
Obrázek 5.13: Výběr příznaků pro detekci klidného/aktivního spánku



Obrázek 5.14: Výsledek detekce klidného/aktivního spánku pro 3 shluky

Obr. 5.15 pak ukazuje rozdělení stejného záznamu se stejnými parametry do 6 shluků. První polovina záznamu je stejně jako v prvním případě viditelně pestřejší než druhá. Ve výsledku lze prohlásit, že hierarchické shlukování založené na analýze frekvenčních vlastností EEG záznamu může sloužit k rozlišení dvou fází spánku. Jaká část odpovídá klidnému a jaká aktivnímu spánku však pouze z výše zmíněného demonstrativního příkladu s jistotou učit nelze. K tomu je třeba klasifikace odborníkem nebo lékařem, případně rozbor jiných biosignálů, například EOG.

Obdobně lze hierarchické shlukování použít i pro detekci jiných fází, stačí na základě znalostí charakteristických vlastností zvolit vhodné příznaky.



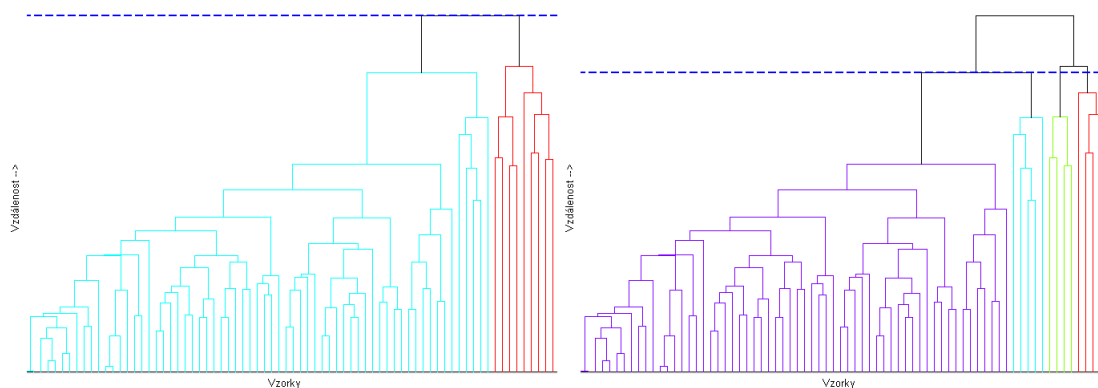
Obrázek 5.15: Výsledek detekce klidného/aktivního spánku pro 6 shluků

### 5.5.2 Demonstrace práce programu při detekci artefaktů

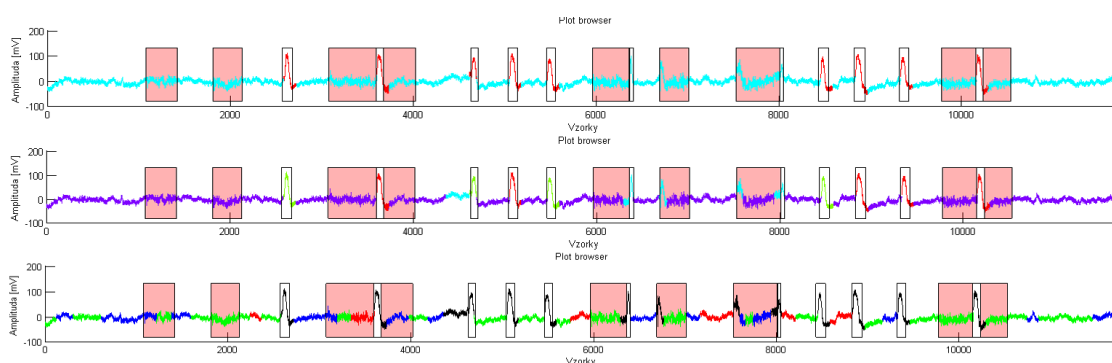
Druhý test hodnotí použitelnost programu pro oddělení biologických artefaktů od zbytku signálu. Jako testovací sada dat slouží druhá sada popsaná v podkapitole 5.5. K expertnímu ohodnocení segmentů program nepřihlíží, slouží pouze ke zhodnocení výsledků. Segmenty identifikované jako artefakty jsou v následujících grafech vyznačeny obdélníky. Červené pozadí reprezentuje artefakty pohybové a bezbarvé artefakty oční.

Dendrogram na Obr. 5.16a ukazuje, že pokud jsou do výpočtu shlukování zahrnuty všechny příznaky, stačí už dva shluky k odhalení určité odlišné skupiny obsahující segmenty, u nichž je pravděpodobnost, že jsou artefakty. Obr. 5.17a ukazuje, že tomu tak opravdu je. Program úspěšně odhalil devět z jedenácti očních artefaktů, aniž by se dopustil jediné falešné klasifikace. Pokud je zvýšen počet shluků na 4 (Obr. 5.16b), odhaleny jsou všechny oční artefakty, avšak za cenu jedné falešné klasifikace (Obr. 5.17b). Při tomto relativně malém počtu shluků nevede analýza k odhalení na pohled méně výraznějších pohybových artefaktů. Při zvyšování počtu shluků jsou sice pohybové artefakty odhalovány, ale za cenu nepřiměřeného počtu falešných klasifikací. Z Obr. 5.17c nelze určit, zda modrá barva reprezentuje artefakty a zelená relevantní záznam nebo naopak.

Z výsledků výše uvedených postupů je zřejmé, že pro detekci pohybových artefaktů je nutné použít sofistikovanější řešení. Podobně jako v kapitole 5.5.1 je jím shlukování beroucí v úvahu pouze určité příznaky charakteristické pro tyto artefakty. Výsledky shlukování všech příznaků jsou zásadně ovlivněny množstvím příznaků popisujících frekvence mozkových vln, které nejsou v tomto testu relevantní.



Obrázek 5.16: Dendrogramy pro (a) 2 a (b) 4 shluky



Obrázek 5.17: Odhalování artefaktů pro (a) 2, (b) 4 a (c) 17 shluků

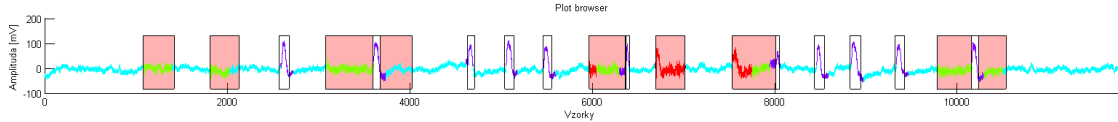
Po zahrnutí pouze vybraných příznaků<sup>6</sup> dochází program k velmi uspokojivým výsledkům opět již při dělení do pouze 4 shluků (Obr 5.18).

Tato metoda uspěla v detekci devatenácti z dvaceti artefaktů, přičemž se nedopustila ani jedné falešné klasifikace. Dokonce se jí podařilo zařadit deset z jedenácti očních artefaktů do jednoho shluku. Lze tedy prohlásit, že program PSG-HC se může stát podpůrnou pomůckou odhalování artefaktů.

Obr. 5.17 a 5.18 dále ukazují, že si ne vždy přesně odpovídají hranice obarvení signálu s hranicemi obdélníky vyznačených artefaktů. Nejedná se o chybu při vykreslování grafů, ale o odchylky způsobené adaptivní segmentací. Segmentace byla aplikována na každý z devatenácti vzorových kanálů zvlášť, přičemž expertní klasifikace vycházela ze zprůměrovaných záznamů.

<sup>6</sup>rms, line\_length, nonlin\_energy, entropy\_shannon, entropy\_log, hjorth2, hjorth3, ampl\_range, mean\_vv\_ampl, var\_vv\_ampl, mean\_vv\_time, var\_vv\_time, mean\_vv\_slope, var\_vv\_slope. Jejich popis v [20]





Obrázek 5.18: Odhalování artefaktů při vybraných příznacích

Obr. 5.17 a 5.18 jsou pro lepší přehlednost znovu zařazeny na konci **Přílohy A** ve větších rozměrech.

### 5.5.3 Použití aplikace nad jinými typy dat

Vzhledem k tomu, že PSG-HC je testován a optimalizován pro výše uvedený typ signálů, v některých jiných aplikacích by byl v dosavadní formě nepoužitelný ze dvou základních důvodů.

Prvním je strojový čas a prostor nutný pro analýzu hierarchického shlukování. Jelikož se komplexita hierarchických metod z rodiny *linkage* pohybuje mezi  $O(n^2)$  až  $O(n^2 \log(n))$  [21], jsou tyto metody nevhodné pro velké datové sady. Druhým problémem je horizontální velikost dendrogramu. Už při zobrazování přibližně 380 segmentů jako na Obr. 5.5 si lze jen těžko představit další nárůst počtu segmentů, ke kterému by jistě při aplikacích na jiné delší signály mohlo dojít. Například v polysomnografii je třeba měřit signály po dobu několika hodin, což je čas nepoměrně delší k testovaným datům a je tedy vhodné použít jinou metodu zobrazení.

Alternativou k používání metod z rodiny *linkage* mohou být metody založené na jiných principech, navržené přímo pro analýzu velkých datových sad, například: *random sampling*, *grid-based approaches*, *divide and conquer* nebo jiné uvedené v kapitole 8 v [21]. Některé z nich dosahují časové komplexity lineární s velikostí vstupních dat, ale jejich podrobnější definice přesahuje rozsah této práce. Tabulka 5.5.3 přináší přehled některých dalších shlukovacích algoritmů, které nejsou součástí této práce, včetně vhodných pro velké sady dat.

Druhý nedostatek by bylo možné odstranit buď prodloužením průměrné periody segmentace nebo zobrazením jen části dendrogramu. U obou těchto řešení by se však mohly objevit jiné nevýhody, buď v přesnosti analýzy nebo v přehlednosti zobrazení.

| Cluster Algorithm        | Complexity   | Suitable for High Dimensional Data |
|--------------------------|--|------------------------------------|
| <i>K</i> -means          | $O(NKd)$ (time) $O(N + K)$ (space)   | No                                 |
| Fuzzy c-means            | Near $O(N)$  | No                                 |
| Hierarchical clustering* | $O(N^2)$ (time and space)  | No                                 |
| PAM                      | $O(K(N - K)^2)$  | No                                 |
| CLARA <sup>†</sup>       | $O(K(40 + K)^2 + K(N - K))$ (time)   | No                                 |
| CLARANS                  | Quadratic in total performance   | No                                 |
| BIRCH                    | $O(N)$ (time)  | No                                 |
| DBSCAN                   | $O(N \log N)$ (time)   | No                                 |
| CURE                     | $O(N_{sample}^2 \log N_{sample})$ (time) $O(N_{sample})$ (space)           | Yes                                |
| WaveCluster              | $O(N)$ (time)  | No                                 |
| DENCLUE                  | $O(N \log N)$ (time)   | Yes                                |
| FC                       | $O(N)$ (time)  | Yes                                |
| STING                    | $O(\text{Number of cells at the bottom layer})$                            | No                                 |
| CLIQUE                   | Linear with the number of objects, Quadratic with the number of dimensions | Yes                                |
| OptiGrid                 | Between $O(Nd)$ and $O(Nd \log N)$   | Yes                                |
| ORCLUS                   | $O(K_0^3 + K_0Nd + K_0^2d^3)$ (space) $O(K_0d^2)$ (space) #                | Yes                                |

\*Include single-linkage, complete-linkage, average-linkage, etc.

<sup>†</sup>Based on the heuristic for drawing a sample from the entire data set (Kaufman and Rousseeuw, 1990)

# $K_0$  is the number of initial seeds.

Tabulka 5.1: Vybrané shlukovací algoritmy, převzato z [21]

# Kapitola 6

## Závěr

Účelem této práce bylo vytvořit program, který si obstojně poradí s aplikací hierarchické shlukové analýzy na předzpracovaný signál EEG včetně přehledného grafického uživatelského rozhraní.

Toto bylo splněno implementací sedmi metod hierarchického shlukování. Grafická nadstavba byla vytvořena s úmyslem poskytnout co nejvíce prostoru na obrazovce počítače pro zobrazení výsledků analýzy. Výsledky je rovněž možno exportovat v \*.mat souboru a dále zpracovávat jiným způsobem.

Teoretická část této práce seznamuje čtenáře se základními pojmy a metodami měření a zpracování biomedicínských signálů a uvádí do problematiky shlukové analýzy s důrazem na hierarchické shlukování.

Výsledkem praktické části této práce je program nazvaný PSG-HC, což lze chápat jako modul rozšiřující PSGLab o možnost provedení shlukové analýzy (aj. *Hierarchical Clustering*). Pro zachování konzistentnosti PSGLabu je GUI programu kompletně psáno v angličtině. Práce s PSG-HC začíná vybráním tří souborů s daty odpovídajícími jednomu signálu. Poté uživatel zvolí požadovanou metodu hierarchické analýzy. Její výsledky jsou zobrazeny v novém okně ve třech grafech různých typů: Dendrogramem, segmentovým grafem a grafem signálu. Uživateli je umožněno posuvníkem určovat, do kolika shluků se má signál rozdělit. Tyto shluky jsou poté zobrazeny v různých barvách. PSG-HC se neomezuje pouze na práci se signálem EEG, z principu může pracovat s jakýmkoli biomedicínským nebo i jiným podobným signálem.

V budoucnu je možné PSG-HC upravit a rozšířit o několik funkcí. Zejména se jedná o možnost současné práce s více kanály. Dalším vylepšením by mohla být implementace shlukového algoritmu, který bude vhodnější pro delší záznamy dat.



# Literatura

- [1] KRAJČA, Vladimír a MOHYLOVÁ, Jitka: *Číslíkové zpracování neurofyziologických signálů*. České vysoké učení technické, 2011, ISBN 978-80-01-04721-7.
- [2] EEG Reviewer: *Polysomnography*. 2014, dostupné online <http://eegreviewer.blogspot.cz/>
- [3] MEDICALLOOK.COM: *Brain Waves*. 2014, dostupné online <http://medicallook.com>
- [4] LEE-CHIONG, Teofilo: *Sleep: A Comprehensive Handbook*. Hoboken: John Wiley, 2005. ISBN 978-047-1751-717
- [5] DEWEY, Russ: *Stages of Sleep*. 2007, dostupné online <http://www.intropsych.com>
- [6] POKORNÝ, Jan: *Evokované potenciály*. Přednáška, Fakulta biomedicínského inženýrství, ČVUT, dostupné online [http://fbmi.cvut.cz/files/nodes/657/public/EVOKOVANÉ PO-TENCIÁLY.pdf](http://fbmi.cvut.cz/files/nodes/657/public/EVOKOVANÉ_PO-TENCIÁLY.pdf)
- [7] WIKISKRIPTA, projekt sítě lékařských fakult MEFANET: *Poruchy spánku*, dostupné online <http://wikiskripta.eu/>
- [8] SAEGELING MEDIZINTECHNIK: *Polysomnografický systém Alice 6LDxS*, dostupné online <http://www.saegeling-mt.cz>
- [9] SVATOŠ, Josef, *BIOLOGICKÉ SIGNÁLY I - GENEZE, ZPRACOVÁNÍ A ANALÝZA*. České vysoké učení technické, 1992. ISBN 80-01-00884-3.

- [10] DIYTDCS.COM: *EEG Standard*. 2014, dostupné online <http://diytdcs.com>
- [11] FIELDING DSP: *Aliasing*. 2014, dostupné online <http://www.fieldingdsp.com>
- [12] ROMESBURG, Charles: *Cluster Analysis for Researchers*. Lulu Press, 2004. ISBN 14-1160617-5
- [13] MELOUN, Milan: *Přednosti analýzy shluků ve vícerozměrné statistické analýze*. Univerzita Pardubice, 2014, dostupné online <http://meloun.upce.cz>
- [14] HYNAR, Martin: *Shlukování 1*. VŠB-TU, 2003, dostupné online <http://www.fit.vutbr.cz/study/courses/ZZD/>
- [15] DESGRAUPES, Bernard: *Clustering Indices*. University Paris Ouest, 2013, dostupné online <http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf>
- [16] KUČERA, Jiří: *Metody kategorizace dat*. Bakalářská práce. Masarykova univerzita, 2008. Dostupné online <http://is.muni.cz/th/172767>
- [17] MELOUN, Milan: *Analýza shluků CLU*. 2014, dostupné online <http://meloun.upce.cz>
- [18] KAUFMAN, Leonard; ROUSSEEUW, Peter J: *Finding Groups in Data: An Introduction to Cluster Analysis*. 9th ed. Hoboken, NJ: Wiley-Interscience, 2005. ISBN 978-047-0317-488
- [19] GUHA, Sudipto, RASTOGI, Rajeev, SHIM, Kyuseok: *CURE: An Efficient Clustering Algorithm for Large Databases*. 2014, dostupné online <http://www.cs.bu.edu/fac/gkollios/ada05/LectNotes/guha98cure.pdf>
- [20] GERLA, Václav: *Automated Analysis of Long-Term EEG Signals*. Dizertační práce. České vysoké učení technické v Praze. 2012, dostupné online <http://bio.felk.cvut.cz/psglab/disertace/disertace-2012-02-29.pdf>
- [21] XU, Rui a WUNSCH, Don: *Clustering*. Hoboken, N.J.: Wiley. 2009, ISBN 978-0-470-27680-8.

- [22] The MathWorks, Inc. *Linkage*. Dostupné online <http://www.mathworks.com/help/stats/linkage.html>
- [23] The MathWorks, Inc.: *Dendrogram*. Dostupné online <http://www.mathworks.com/help/stats/dendrogram.html>
- [24] STANFORD NATURAL LANGUAGE PROCESSING GROUP, The: *Single-Link, Complete-Link & Average-Link Clustering*. Stanford University, dostupné online: <http://nlp.stanford.edu/IR-book/completelink.html>
- [25] STACE, Clive A.: *Plant Taxonomy and Biosystematics*, 2nd ed. New York: Cambridge Univ. Press, 1994. ISBN 978-052-1427-852
- [26] TOWERS, Sherry: *K-means clustering*. 2014, dostupné online <http://sherrytowers.com>
- [27] HUMAN ORIENTED SYSTEMS: *Hierarchical diagram*. 2014, dostupné online <http://humanoriented.com>
- [28] SAS Institute, Inc.: *Proximity Measures*. 2014, dostupné online <http://support.sas.com>
- [29] McLEISH, Erin: *Manhattan distance*. 2014, dostupné online <http://cgm.cs.mcgill.ca/mcleish/644>





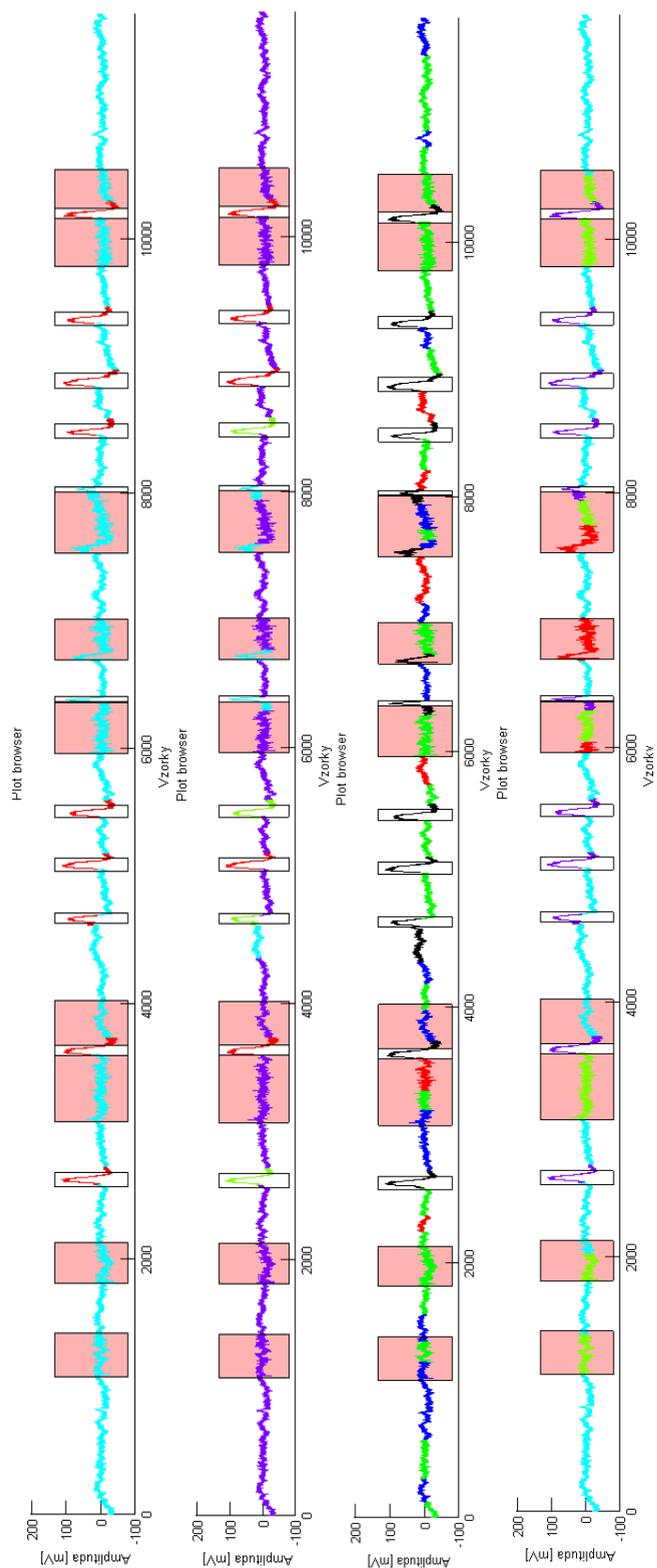
# Příloha A

## Seznam příznaků, graf hledání artefaktů

|                               |                               |                                 |
|-------------------------------|-------------------------------|---------------------------------|
| min_value                     | fft_abs_alpha1_(7_to_9.5Hz)   | fft_rel_gamma1_(30_to_35Hz)     |
| max_value                     | fft_abs_alpha2_(9.5_to_12Hz)  | fft_rel_gamma2_(35_to_40Hz)     |
| skewness                      | fft_abs_beta1_(12_to_21Hz)    | fft_delta_theta_ratio           |
| kurtosis                      | fft_abs_beta2_(21_to_30Hz)    | fft_delta_alpha_ratio           |
| mean                          | fft_abs_gamma1_(30_to_35Hz)   | fft_theta_alpha_ratio           |
| std                           | fft_abs_gamma2_(35_to_40Hz)   | fft_dt_alpha_ratio              |
| median                        | fft_abs_whole                 | fft_spectral_roll_off_(1_to_12) |
| 1st_diff_mean                 | fft_rel_delta_(0.1_to_3Hz)    | fft_sef95_(1_to_12)             |
| 1st_diff_max                  | fft_rel_theta_(3_to_7Hz)      | min_wav_value_opseg1_db4        |
| 2st_diff_mean                 | fft_rel_alpha_(7_to_12Hz)     | min_wav_value_opseg2_db4        |
| 2st_diff_max                  | fft_rel_beta_(12_to_30Hz)     | min_wav_value_opseg3_db4        |
| fft_abs_delta_(0.1_to_3Hz)    | fft_rel_gamma_(30_to_40Hz)    | min_wav_value_opseg4_db4        |
| fft_abs_theta_(3_to_7Hz)      | fft_rel_delta1_(0.1_to_1.5Hz) | min_wav_value_opseg5_db4        |
| fft_abs_alpha_(7_to_12Hz)     | fft_rel_delta2_(1.5_to_3Hz)   | max_wav_value_opseg1_db4        |
| fft_abs_beta_(12_to_30Hz)     | fft_rel_theta1_(3_to_5Hz)     | max_wav_value_opseg2_db4        |
| fft_abs_gamma_(30_to_40Hz)    | fft_rel_theta2_(5_to_7Hz)     | max_wav_value_opseg3_db4        |
| fft_abs_delta1_(0.1_to_1.5Hz) | fft_rel_alpha1_(7_to_9.5Hz)   | max_wav_value_opseg4_db4        |
| fft_abs_delta2_(1.5_to_3Hz)   | fft_rel_alpha2_(9.5_to_12Hz)  | max_wav_value_opseg5_db4        |
| fft_abs_theta1_(3_to_5Hz)     | fft_rel_beta1_(12_to_21Hz)    | mean_wav_value_opseg1_db4       |
| fft_abs_theta2_(5_to_7Hz)     | fft_rel_beta2_(21_to_30Hz)    | mean_wav_value_opseg2_db4       |

---

|                               |                                  |                                   |
|-------------------------------|----------------------------------|-----------------------------------|
| mean_wav_value_opseg3_db4     | wav_band_ratio_opseg1_opseg3_db4 | line_length                       |
| mean_wav_value_opseg4_db4     | wav_total_energy_db4             | nonlin_energy                     |
| mean_wav_value_opseg5_db4     | entropy_log_wav_opseg1_db4       | entropy_shannon                   |
| median_wav_value_opseg1_db4   | entropy_log_wav_opseg2_db4       | entropy_log                       |
| median_wav_value_opseg2_db4   | entropy_log_wav_opseg3_db4       | hjorth2                           |
| median_wav_value_opseg3_db4   | entropy_log_wav_opseg4_db4       | hjorth3                           |
| median_wav_value_opseg4_db4   | entropy_log_wav_opseg5_db4       | energy_percent_wav_opseg1_db4     |
| median_wav_value_opseg5_db4   | entropy_wav_opseg1_db4           | energy_percent_wav_opseg2_db4     |
| std_wav_value_opseg1_db4      | entropy_wav_opseg2_db4           | energy_percent_wav_opseg3_db4     |
| std_wav_value_opseg2_db4      | entropy_wav_opseg3_db4           | energy_percent_wav_opseg4_db4     |
| std_wav_value_opseg3_db4      | entropy_wav_opseg4_db4           | energy_percent_wav_opseg5_db4     |
| std_wav_value_opseg4_db4      | entropy_wav_opseg5_db4           | wav_zero_crossing_opseg1_db4      |
| std_wav_value_opseg5_db4      | 1st_diff_wav_mean_opseg1_db4     | wav_zero_crossing_opseg2_db4      |
| skewness_wav_value_opseg1_db4 | 1st_diff_wav_mean_opseg2_db4     | wav_zero_crossing_opseg3_db4      |
| skewness_wav_value_opseg2_db4 | 1st_diff_wav_mean_opseg3_db4     | wav_zero_crossing_opseg4_db4      |
| skewness_wav_value_opseg3_db4 | 1st_diff_wav_mean_opseg4_db4     | wav_zero_crossing_opseg5_db4      |
| skewness_wav_value_opseg4_db4 | 1st_diff_wav_mean_opseg5_db4     | zero_crossing                     |
| skewness_wav_value_opseg5_db4 | 1st_diff_wav_max_opseg1_db4      | min_max_number                    |
| kurtosis_wav_value_opseg1_db4 | 1st_diff_wav_max_opseg2_db4      | coeff_of_variation                |
| kurtosis_wav_value_opseg2_db4 | 1st_diff_wav_max_opseg3_db4      | wav_coeff_of_variation_opseg1_db4 |
| kurtosis_wav_value_opseg3_db4 | 1st_diff_wav_max_opseg4_db4      | wav_coeff_of_variation_opseg2_db4 |
| kurtosis_wav_value_opseg4_db4 | 1st_diff_wav_max_opseg5_db4      | wav_coeff_of_variation_opseg3_db4 |
| kurtosis_wav_value_opseg5_db4 | 2nd_diff_wav_mean_opseg1_db4     | wav_coeff_of_variation_opseg4_db4 |
| wav_band_opseg1_db4           | 2nd_diff_wav_mean_opseg2_db4     | wav_coeff_of_variation_opseg5_db4 |
| wav_band_opseg2_db4           | 2nd_diff_wav_mean_opseg3_db4     | ampl_range                        |
| wav_band_opseg3_db4           | 2nd_diff_wav_mean_opseg4_db4     | mean_vv_ampl                      |
| wav_band_opseg4_db4           | 2nd_diff_wav_mean_opseg5_db4     | var_vv_ampl                       |
| wav_band_opseg5_db4           | 2nd_diff_wav_max_opseg1_db4      | mean_vv_time                      |
| wav_band_relative_opseg1_db4  | 2nd_diff_wav_max_opseg2_db4      | var_vv_time                       |
| wav_band_relative_opseg2_db4  | 2nd_diff_wav_max_opseg3_db4      | mean_vv_slope                     |
| wav_band_relative_opseg3_db4  | 2nd_diff_wav_max_opseg4_db4      | var_vv_slope                      |
| wav_band_relative_opseg4_db4  | 2nd_diff_wav_max_opseg5_db4      |                                   |
| wav_band_relative_opseg5_db4  | rms                              |                                   |



Obrázek A.1: Grafy detekce artefaktů



# Příloha B

## Obsah CD

- Program PSG-HC
- Ukázka dat, na kterých byl program testován
- Elektronická verze této práce ve formátu PDF