České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra matematiky
Obor: Matematické inženýrství
Zaměření: Aplikované matematicko-stochastické metody

Využití FFT při diagnostice Alzheimerovy choroby z EEG

Use of FFT in the diagnosis of Alzheimer's disease from EEG

BAKALÁŘSKÁ PRÁCE

Vypracoval: Nikol Kopecká
Vedoucí práce: doc. Ing. Jaromír Kukal, Ph.D.
Rok: 2013

Před svázáním místo téhle stránky vložíte zadání práce s podpisem děkana (bude to jediný oboustranný list ve Vaší práci) !!!!

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracovala samostatně a použila jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v přiloženém seznamu.

V Praze dne .....

.....
Nikol Kopecká

Poděkování

Jako první bych chtěla poděkovat svému vedoucímu práce, doc. Ing. Jaromíru Kukalovi, Ph.D. za mnoho cenných nápadů a odborné vedení, jež se podepsalo na úrovni této práce. Dále bych chtěla poděkovat doc. MVDr. Šimonu Vaculínovi, Ph.D., Ing. Janu Dudákovi a Bc. Pavlu Bártovi za praktickou ukázku použití EEG a za názorné vysvětlení jeho principu. Za pomoc s korekturou práce jsem velmi vděčná Mgr. Miroslavě Kubešové a Mgr. Zdeně Burgerové. A nakonec bych ráda poděkovala všem svým přátelům a své rodině za jejich neutuchající podporu.

Nikol Kopecká

Název práce:
Využití FFT při diagnostice Alzheimerovy choroby z EEG

Autor: Nikol Kopecká

Obor: Matematické inženýrství
Druh práce: Bakalářská práce

Vedoucí práce: doc. Ing. Jaromír Kukal, Ph.D.
Katedra softwarového inženýrství v ekonomii, Fakulta jaderná a fyzikálně inženýrská, České vysoké učení technické v Praze
Konzultant: —

Abstrakt: Cílem této práce byla diagnostika Alzheimerovy choroby na základě EEG signálu pacientů. Neinvazivní charakter a jednoduchost EEG by umožnilo snadné vyšetření rizikových skupin obyvatelstva. Včasná identifikace a započetí léčby jsou podstatné pro zpomalení nástupu nemoci.
K testování byla k dispozici sada 28 nemocných a 146 zdravých pacientů. Pro vyhodnocení EEG signálu bylo použito několik příznakových modelů – relativní spektrální výkon pásem mozkových vln, vyhlazené Fourierovo spektrum, cepstrum a autoregresní model. Pro vybrání optimálního příznaku, či nejlepší a nejjednodušší lineární kombinace, byla použita LDA metoda regularizovaná L^1 normou.
Nejspolehlivějšího oddělení bylo dosaženo pro cepstrum a Fourierovo spektrum pro elektrody na temeni hlavy. Pro svoji jednoduchost a robustnost je nejlepším příznakem kombinace cepstra na quefrenci 0.2 a 0.3 s v poměru 5:1. Bylo dosaženo 80% oddělení pacientů pro AUC 0.88 a p-hodnotu 10^{-60}.

Klíčová slova: Alzheimerova choroba, analýza EEG, FFT, Cepstrum

Title:
Use of FFT in the diagnosis of Alzheimer's disease from EEG

Author: Nikol Kopecká

Abstract: The main aid of this thesis is to diagnose the Alzheimer's disease, based on the EEG signal of the patients. Noninvasive character and simplicity of EEG method allows easy examination of the predisposed groups of people. Early diagnosis is critical in order to make the cure effective.
The evaluation of the EEG signals was based on the several classification models – relative spectral power of the certain brain waves, smoothed Fourier power spectra, cepstrum and autoregressive model. These models were trained on the set of 28 diseased patients and testing set of 146 healthy patients. The optimal binary classificatory was chosen by the LDA method regularized by the L^1 norm.
The most successful classification was achieved by the cepstrum and the smoothed Fourier spectrum with the electrodes situated on the vertex of the patients head. For its simplicity and robustness was as the post promising classificatory found combination of the cepstrum at quefrency 0.2 and 0.3 in ratio 5:1. This classification provided the 80% classification rate for AUC 0.88 and p-value 10^{-60}.

Key words: Alzheimer disease, EEG analysis, FFT, cepstrum

Obsah

Úvod

Alzheimerova choroba (AD) se stává čím dále častějším neduhem vyskytujícím se především u starší generace. Příčinou je jednak celosvětové stárnutí populace v rozvinutých zemích, ale také mohou mít vliv další těžko identifikovatelné civilizační faktory. A poslední, umělá příčina nárůstu počtu pacientů jsou stále se zlepšující diagnostické metody umožňující důvěryhodnou identifikaci této choroby.

V současnosti není medicína schopná AD vyléčit, ale při včasné detekci lze její postup výrazně zpomalit.

Dnes se AD v ranné fázi diagnostikuje na základě poruch krátkodobé paměti a pomocí podrobného neurologiského a neuropsychologického vyšetření [21]. Také musí být též vyloučeny ostatní možné příčiny, jako jsou některé typy demence a cerebrální patologie. K tomuto účelu se používají pokročilé zobrazovací techniky mozku, jako je PET (Positron Emission Tomography), CT (Computed Tomography) a MRI (Magnetic Resonance Imaging). Často také dochází k odebrání mozkomíšního moku a testům na přítomnost specifickýh bílkovin - prionů.

Konečná diagnóza je založená na vyhodnocení více různých metod a vyloučení všech alternativ. Jedná se tedy o vcelku komplikovanou diagnostiku, často též invazivní, zcela nevhodnou například pro plošný screening statisticky nejohroženějších skupin obyvatelstva.

Proto je cílem této práce vyvinout metody umožňující co nejdůvěryhodnější diagnostiku pomocí EEG měření.

Electroencephalography (EEG)

První měření elektrické akktivity mozku bylo provedeno německým psychiatrem Hansem Bergerem roku 1924 [1]. Tehdy byl také poprvé pozorován periodický elektrický signál o napětí kolem 100 μV s frekvencí mezi 1 a 60 Hz, vydávaný mozkem.

Dnes se EEG používá jako základní, v principu jednoduchá a neinvazivní diagnostika lidského mozku. Funguje na základě měření velmi slabých elektrických potenciálů na povrchu kůže lidské hlavy. Slabá elektrická pole jsou vytvářená korelovanou neurální aktivitou velkých skupin neuronů a mohou projevovat jako periodické změny potenciálu na povrchu kůže. Výhody této metody jsou současně i její limitací. Velká vzdálenost sond od zdrojů signálu – neuronů způsobuje ztrátu a prostorové rozmazání signálu. Svalová aktivita mimických svalů způsobuje falešné signály, které

se dají jen mimořádně obtížně eliminovat [9]. Je možné měření přímo na povrchu mozku, ale obvykle se z praktických důvodů neprovádí.

Běžně se EEG používá například k výzkumu epilepsie či k diagnostice pacienta v kómatu [25]. Další využití je třeba ke sledování mozkové aktivity během spánku. V tabulce 1 se nachází obvykle pozorované mozkové vlny a v jakých případech jsou u pacientů obvykle pozorovány.

typ vln	pásmo	popis
delta	0.5–4 Hz	U bdělého dospělého člověka jsou vždy patologickým jevem. Projevují se například v hlubokém kómatu, transu, hypnóze nebo nádoru.
theta	4–8 Hz	Mohou indikovat patologický jev, jsou-li alespoň 2× větší než alfa vlny. Vyskytují se třeba v lehkém bezvědomí.
alfa	8–13 Hz	Nejintenzivnější vlny v bdělém stavu. Největší intenzita je, je-li pacient v klidu (bez duševní činnosti) se zavřenýma očima.
beta	13–30 Hz	Typické vlny pro soustředění a logicko-analytické myšlení nebo intenzivní pocity.

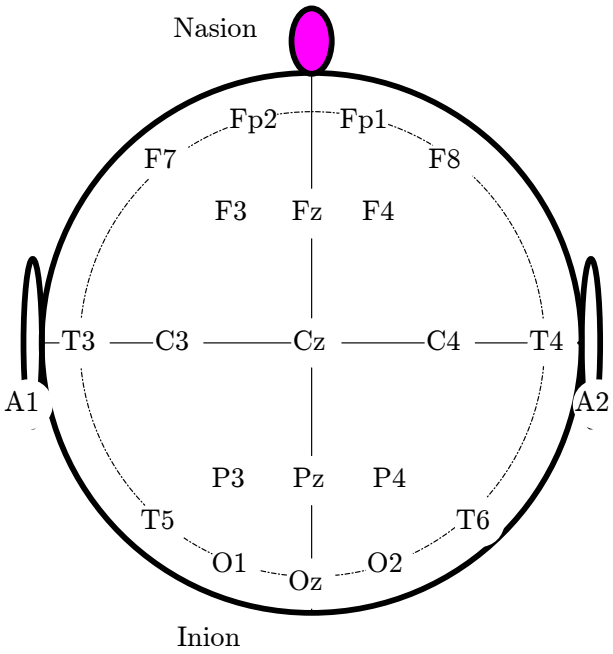
Tabulka 1: Základní klasifikace mozkových vln podle jejich frekvence a případy, kdy se jejich výskyt obvykle pozoruje. [13]

Popis EEG měření

Při EEG neření je na povrch hlavy umístěna sada elektrod. Pomocí vodivého gelu je zajištěno spojení s pokožkou i bez toho aby musel být pacient přetím zbaven vlasů. Elektrody jsou uspořádány ve standartizovaném mezinárodním "10-20 systému". V pohledu svrchu jsou elektrody zobrazeny v Obr. 1. Obvykle se používá 21 elektrod, z čehož dvě, A1 a A2 jsou připevněné na ušlí boltce a slouží k uzemění. Signál z elektrod je poté měřen se vzorkovací frekvencí minimálně dvojnásobnou než je frekvence frekvence pozorovaných vln. Navíc před záznamem jsou pomocí filtrů potlačeny frekvence obsahujích šum, například oblast kolem 50 Hz. Způsob, jakým byly kanály v námi zpracované datové sadě přiřazeny elektrodám je uveden v tanulce 2.

Tabulka 2: Seznam čísel kanálů a příslušné zkratky určující polohu na schematickém nákresu v Obr. 1

číslo signálu	1	2	3	4	5	6	7	8	9	10	11
jméno signálu	Fp1	Fp2	F7	F3	Fz	F4	F8	T3	C3	Cz	C4
číslo signálu	12	13	14	15	16	17	18	19	20	21	
jméno signálu	T4	T5	P3	Pz	P4	T6	O1	O2	A1	A2	



Obrázek 1: Schéma standardizovaného mezinárodního rozložení elektrod 10-20 na skalpu hlavy v pohledu shora

Použití EEG k identifikaci Alzheimerovy choroby

A v neposlední řadě lze použít právě k identifikaci Alzheimerovy choroby. U pacienta s Alzheimerovou chorobou dochází jednak k poklesu komplexnosti signálu,k poklesu intenzity ve vysokofrekvenční složce, a také k nárůstu nízkofrekvenční složky EEG signálu [9]. A právě změna ve frekvenčním rozdělení signálu se dá určit z jeho Fourierovy transformace. Ovšem rozlišení pacientů je mimořádně obtížný problém, o jehož spolehlivé řešení se pokoušejí vědci už mnoho let. Problém je už jen v tom, že každý pacient je unikátní s různou neurální aktivitou a zatímco u jednoho může být změna ve spektru vln již signifikantním znamením počátků Alzheimerovy choroby, u druhého tomu tak být nemusí. Proto nebude cílem této práce jednoznačně rozlišit tyto dvě skupiny, ale nalézt takové veličiny, které umožní nejlepší spolehlivé oddělení a identifikování oblastí, kde prostě oddělení jen na základě dostupných dat z EEG není možné.

Přehled práce

V první kapitole této práce je definovaná spojitá a diskrétní Fourierova transformace a jsou shrnuty jejich základní vlastnosti. V následující kapitole je popsána příprava reálného signálu spolu s popisem použitých příznakových modelů. Ve třetí kapitole jsou popsány otestované metody použité k nalezení příznaků a nakonec v poslední kapitole jsou tyto metody, využité k finálnímu statickému vyhodnocení, srovnání s výsledky prací na podobné téma provedených na stejné datové sadě a nebo její podmnožině [14, 24].

Kapitola 1

Fourierova transformace

Samotná Fourierova transformace byla zavedena Josephem Fourierem (1768 – 1830) k popisu rovnice vedení tepla. Dnes má nespočetné množství praktických využití ve zpracování signálů, obrazu, řešení diferenciálních rovnic a v mnohém dalších [5]. Navíc byla postupně rozšířena i pro zobecněné funkce a definována v elegantním tvaru Fourierova operátoru v komplexním vektorovém Schwarzově prostoru [3]. Tato definice zde bude uvedena a následně pomocí zobecněných funkcí odvozen i diskrétní tvar Fourierovy transformace, který je mnohem vhodnější pro praktické aplikace.

Na závěr kapitoly je uvedena rychlá Fourierova transformace (*Fast Fourier Transformation* FFT), která má nezastupitelné místo v jakékoli moderní metodě na zpracování signálu.

1.1 Spojitá Fourierova transformace

Obecná Fourierova transformace je ve své nejobecnější podobě definována jako ortonormální zobrazení na Schwarzově prostoru funkcí. Tento prostor, též nazývaný prostor rychle klesajících funkcí, je definován následujícím vztahem

$$S\left(\mathbb{R}^n\right) = \left\{f \in C^\infty(\mathbb{R}^n) \mid \|f\|_{\alpha,\beta} < \infty \quad \forall \alpha, \beta\right\},$$

kde α, β jsou libovolné multiindexy, $\mathbb{C}^\infty(\mathbb{R}^n)$ je množina hladkých funkcí z $\mathbb{R}^n$ do $\mathbb{C}$ a norma $\|\cdot\|_{\alpha,\beta}$ je definovaná následovně

$$\|f\|_{\alpha,\beta} = \sup_{x\in\mathbb{R}^n} \left|x^\alpha \partial_x^\beta f(x)\right|.$$

Nejtypičtějším příkladem funkce splňující tyto podmínky je například

$$p_n(x)e^{-\|x\|^2} \in S(\mathbb{R}^n),$$

kde $p_n(x)$ je libovolný polynom n-tého řádu. Navíc i jakákoli hladká funkce s konečným supportem (prvek prostoru testovacích funkcí [16]) tam také patří.

Fourierova transformace je poté definována následovně

$$\mathcal{F}[f](x) = (2\pi)^{-\frac{n}{2}} \int_S f(x)\ e^{-2\pi i(x,\xi)}\ \mathrm{d}x, \text{ for } \forall f \in S \text{ and } \forall \xi \in \mathbb{R}^n \tag{1.1}$$

a inverzní transformace

$$\mathcal{F}^{-1}[\hat{f}](\xi) = (2\pi)^{-\frac{n}{2}} \int_S \hat{f}(\xi)\ e^{2\pi i(x,\xi)}\ \mathrm{d}x.$$

1.1.1 Příklady užitečných funkcí

Lze přímým výpočtem ukázat, že vlastní čísla tohoto operátoru jsou $e^{i\pi n/4} \quad \forall \in \hat{4}$ a vlastní vektory jsou *Hermitovy polynomy* [3]

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}.$$

Příkladem je třeba Gaussova funkce, která je vlastně Hermitovým polynomem 0-tého řádu $H_0(x)$, a proto

$$\mathcal{F}[e^{x/2}](\xi) = e^{\xi/2}.$$

Dalším užitečným příkladem, který dále použijeme, je tzv. vzorkovací funkce (*Dirac comb*), definována jako periodická řada Dirackových delta funkcí

$$\Delta_T(t) \ \stackrel{\text{def}}{=} \sum_{k=-\infty}^{\infty} \delta(t - kT) \tag{1.2}$$

$$\sum_{n=-\infty}^{\infty} \delta(t - nT) \quad \stackrel{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{-i2\pi fn} = \frac{1}{T} \sum_{k=-\infty}^{\infty} \delta\left(f - k\right).$$

kde T je vzorkovací frekvence.

1.1.2 Základní vlastnosti

Začneme se základními, přesto velmi důležitými vlastnostmi $\mathcal{F}$:

- **Aditivita a homogenita**

  $$\mathcal{F}[\alpha f + g] = \alpha \mathcal{F}[f] + \mathcal{F}[g] \quad \forall f, g \in S \quad \forall \alpha \in \mathbb{C}$$

  která plyne přímo z definice integrálu.

- **Symetrie**

  Pokud $h(x)$ a $\hat{h}(\xi)$ představují funkci z S a její Fourierovu transformaci, potom

  $$\hat{h}(\xi) \quad \stackrel{\mathcal{F}}{\longleftrightarrow} \quad h(-x).$$

- **Škálování**

$$h(kx) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{|k|}\hat{h}\left(\frac{\xi}{k}\right),$$

pro $k \in \mathbb{C}$, a to samé ovšem platí obráceně:

$$\frac{1}{|k|}h\left(\frac{x}{k}\right) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \hat{h}(\xi).$$

- **Posunutí**

V případě posuvu o konstantu $a \in \mathbb{C}$ dojde ke změně fáze

$$f(x-a) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad e^{-ia\xi}\hat{f}(\xi)$$

$$e^{iax}f(x) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \hat{f}(\xi-a)$$

- **Derivace**

$$D^{\alpha}(\mathcal{F}(f)) = (-i)^{|\alpha|}\mathcal{F}(x^{\alpha}f),$$
$$\mathcal{F}(D^{\alpha}f)(\xi) = i^{|\alpha|}\xi^{\alpha}\mathcal{F}(f)(\xi),$$

kde α je libovolný multiindex a $|\alpha|$ je celkový počet derivací.

Tyto vlastnosti plynou přímo u definice $\mathcal{F}$ (1.1). Několik dalších důležitých vlastností také není těžké dokázat [5]:

- **Parsevalova rovnost**

$$\|\mathcal{F}[f]\|_2 = \|f\|_2, \tag{1.3}$$

což ovšem spolu s tím, že $\mathcal{F}$ zobrazuje z S na S plyne z toho, že se jedná o ortonormální operátor.

- **Konvoluční teorém**

$$(g * h)(x) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \hat{g}(\xi)\hat{f}(\xi). \tag{1.4}$$

kde g a h jsou prvky Schwartzova prostoru.

Parsevalova rovnost nám umožní spojit celkový vyzářený výkon signálu s normou jeho Fourierovy transformace. A vzhledem k tomu, že pomocí Fourierovy transformace se každý signál dá zapsat jako superpozice projekcí do ortonormálních vektorů $v_y(x) = e^{i\pi(x,y)}$, bude i možné přiřadit kvadrátu amplitudy Fourierovy transformace význam spektrální hustoty výkonu signálu.

Konvoluční teorém je naopak velmi významný pro jednoduchý výpočet konvoluce a ve spojení s rychlou FFT představuje základní nástroj pro analýzu signálu.

1.2 Diskrétní Fourierova transformace

Diskrétní Fourierova transformace zobrazuje periodickou funkci vzorkovanou v konečně mnoha bodech s konstantní vzorkovací frekvencí na diskrétní, periodický vzorkovaný obraz. Obecně se jedná o zobrazení z komplexních čísel opět do komplexních čísel. Máme-li konečnou množinu $\{x_i \in \mathbb{C} \quad | i \in \hat{N}\}$, diskrétní FT je definována následujícím vztahem

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1. \tag{1.5}$$

a inverzní transformace je definována téměř symetricky až na faktor $1/N$

$$x_k = \frac{1}{N} \sum_{n=0}^{N-1} X_n e^{i2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1. \tag{1.6}$$

Diskrétní Fourierovu transformaci lze odvodit také jako speciální případ spojité Fourierovy transformace. Vynásobíme-li spojitou a periodickou funkci $u(x)$ vzorkovací funkce definovanou vzorcem (1.2), vyjde nám opět diskrétní periodická funkce, ale na frekvenční doméně

$$u(x) \cdot \sum_{n=-\infty}^{\infty} \delta(t - nT) \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{T} U(f) \cdot \sum_{k=-\infty}^{\infty} \delta\left(f - k/T\right).$$

Stejnou cestou by šlo získat i Fourierovu řadu. Provedeme-li totiž spojitou Fourierovu transformaci spojité, ale periodické funkce získáme diskrétní nekonečnou řadu. A i naopak Fourierovou transformaci diskrétní nekonečné řady násobené vzorovací funkcí získáme spojitou, ale periodickou funkci. Tato korespondence mezi spojitou a diskrétní Fourierovou transformací nám umožňuje používat všechny vlastnosti spojité transformace dokázané v předchozích sekcích, také u diskrétní transformace.

Provedeme-li DFT reálných čísel, získáme Hermitovsky symetrické spektrum, to znamená, že $X_i = X^*_{-i}$. Polovina spektra tedy nepřináší žádnou dodatečnou informaci, a proto se zpravidla nepoužívá. Frekvence vyšší než je polovina vzorkovací frekvence signálu nemohou být DFT správně určeny, jak uvádí Nyquistův – Shannonův teorém [5, 22]. Kvadrát absolutní hodnoty X_i odpovídá podle Parsevalova teorému výkonu detekovanému na dané frekvenci. Fáze odpovídá tangensu podílu kosinové a sinové složky této vlny.

1.3 Rychlá Fourierova transformace (FFT)

Algoritmus umožňující efektivní a rychlý výpočet diskrétní Fourierovy transformace byl poprvé objeven C. F. Gaussem [12], ale potom na více než 100 let upadl v zapomnění. Skutečný rozmach přišel až s počítači a *rychlou Fourierovou transformací* (FFT) znovuobjevenou pomocí J. Cooleym a J. Turkeym v roce 1965 [7]. Tento algoritmus je založený na principu "rozděl a panuj". Při klasickém výpočtu vzorce

(1.6) je nutné spočíst N^2 součtů a také N^2 součinů. Ale vypočítávají-li se koeficienty rekurzivně ve správním pořadí, je možné snížit výpočetní náročnost na $O(n \log n)$. Konkrétní postup výpočtu je například důkladně vysvětlen v knize [5]. Ale samotné implementace není jednoduchá a často se provádí náročné výpočetní optimalizace pro dosažení nejlepších výsledků na konkrétním počítači. To ale řeší programy jako MATLAB sami, bez zásahu uživatele.

1.3.1 Výpočetní náročnost FT

Výpočetní náročnost $O(n \log n)$ platí, jen má-li signál délku mocniny dvou. Zobecnění pro vektory o jiné délce také existuje, ale složitost výpočtu je vyšší, je přibližně úměrná součtu prvočíselného rozkladu násobeného počtem prvků [5]. Takže například FFT pole délky 127 prvků je $10\times$ pomalejší než délky 128. Proto je nezbytné signál buď dostatečně zkrátit, anebo prodloužit nulami tak, aby bylo dosaženo požadované délky a maximální rychlosti výpočtu.

1.3.2 Okrajové efekty FT

Diskrétní FT předpokládá, že vstupní signál je periodický. K výpočtu ovšem stačí jen jediná jeho perioda. Stejně tak je obraz ve frekvenční doméně diskrétní a periodický. Z tohoto důvodu mohou vznikat okrajové jevy, které mohou ztížit vyhodnocení signálu ve frekvenční doméně. To, že použijeme pouze signál konečné délky, je to ekvivalentní, kdybychom původní nekonečně dlouhý signál vynásobili obdélníkovou funkcí příslušné délky. Ovšem Fourierova transformace součinu dvou funkcí v časové doméně je konvolucí ve frekvenční doméně. A obdélníková funkce způsobí výrazné rozkmitání FT původní funkce. Proto se často na signál aplikuje tzv. *windowing* signálu [5].

1.4 Výkon signálu

Vraťme se nyní k reálné aplikaci FT na skutečném signálu. Často nás zajímá *výkon signálu*. Předpokládejme pro zjednodušení, že měříme napětí $u(t)$ na odporu R. Potom je okamžitý výkon roven

$$P(t) = u(t)i(t) = \frac{1}{R}u^2(t).$$

Zcela analogicky můžeme zadefinovat výkon jakéhokoli abstraktního signálu jako

$$P(t) = x^2(t). \tag{1.7}$$

Často nás zajímá frekvenční výkon signálu. Ovšem ten má smysl definovat, jen když zkoumáme *stacionární* signál, jinak získáme jen střední hodnotu přes zkoumaný časový interval.

Předpokládejme, že měříme pouze harmonický signál o frekvenci ω a amplitudě A_0. Potom je střední výkon přes periodu roven

$$\hat{P} = \frac{1}{T}\int_0^T |Ae^{i\omega_0 t}|^2 \mathrm{d}t = |A|^2.$$

To vede k domněnce, že frekvenční výkon takového signálu by měl být

$$P(w) = A^2\delta(\omega - \omega_0).$$

Je-li signál superpozicí více vln, $x(t) = \int A(\omega)e^{i\omega t}\mathrm{d}\omega$, pak z definice FT, její linearity a poznatku, že $\int e^{i(\omega_1-\omega_2)t}\mathrm{d}t = \delta(\omega_1 - \omega_2)$ získáme vztah

$$P(\omega) = |A(\omega)|^2 = |\mathcal{F}[x(t)]|^2.$$

Celkový frekvenční výkon je potom podle Parsevalovy rovnosti (1.3) roven celkovému výkonu v čase

$$\int |x(t)|^2 \mathrm{d}t = \int |\mathcal{F}[x(t)]|^2(\omega)\mathrm{d}\omega.$$

V případě výpočtu pomocí FFT je důležité nezapomenou na přenormování pomocí faktoru $1/\sqrt{N}$.

Kapitola 2

Metody vyhodnocení EEG signálu

2.1 Popis databáze EEG signálů

Ke statistické analýze jsme použili soubor dat pořízený v oblastní nemocnici v Rychnově nad Kněžnou. Celý soubor obsahuje 28 pacientů s Alzheimerovou nemocí (dále značení AD) v různých fázích a 146 zdravých pacientů z kontrolní skupiny (dále značení CN) obdobného stáří jako AD skupina. Po odstranění příliš poškozených souborů se počty zredukovaly na 141 CN a 26 AD pacientů.

U každého pacienta bylo provedeno měření délky 5–10 min se vzorkovací frekvencí 200 Hz (tedy 60 000–120 000 vzorků). Konfigurace elektrod na hlavě byla provedena podle 10-20 systému s vynecháním Oz elektrody. Bylo tedy k dispozici celkem 21 kanálů.

Pacienti se během celého měření nacházeli v klidu na lůžku se zavřenýma očima a bez přítomnosti jakýchkoli vnějších podmětů.

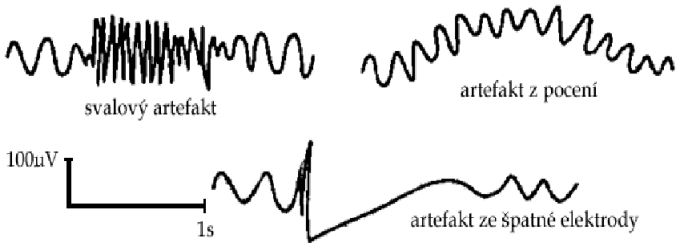
2.1.1 Příprava dat

V EEG signálu se vyskytují artefakty obvykle dělené do dvou kategorií

- Biologické
- Technické

V první kategorii jsou například nechtěné signály způsobené lidskou aktivitou – mrkání, svalová aktivita (mimika), pocení. Ale také různá aktivita mozku, zatímco jeden pacient může být v klidu, druhý může být stresovaný. Do druhé kategorie řadíme další vlivy technického charakteru, jako je vnější elektromagnetické rušení, špatně připojená elektroda, spuštění přístroje atd. Technické artefakty se dají identifikovat a odstranit celkem snadno – 50 Hz signál ze zásuvky se dá odfiltrovat pásmovou propustí nebo dodatečně odstranit při počítačovém zpracování a vadnou elektrodu jde identifikovat a vyřadit ji z dalšího zpracování. Naopak biologické

artefakty jsou odstranitelné jen velmi obtížně. Několik postupů založených na simultárním využití více dalších diagnostik je shrnuto v článku [9], ale žádný z nich není použitelný v našem případě, kdy nemáme dostupné další diagnostiky. Proto nezbývá než doufat, že svalová aktivita a mrkání byla víceméně u všech pacientů stejná a spektrum je ovlivněno stejným způsobem. V případě, kdy se tento artefakt objevil jako výrazný záchvěv na sledovaných signálu, odstranění bylo již možné, a také bylo pečlivě provedeno.



Obrázek 2.1: Příklady artefaktů vyskytujících se v EEG signálu [13]

Poslední problém, který je nutné vzít v úvahu, je že intenzita měřeného signálu značně závisí na vlhkosti pacientovy pokožky nebo způsobu připevnění elektrod a může se velmi výrazně lišit pacient od pacienta, či dokonce elektrodu od elektrody. Proto je nezbytné signál přenormovat a odstranit konstantní odchylku signálu od nulové hladiny. Otázkou může být, jakou normalizaci zvolit. Můžeme použít rozptyl (STD standart deviation) a nebo například MAD (Median Absolute Deviation), případně Mean Absolute Deviation, lišící se od předchozího tím, že použijeme místo mediánu průměr. Samotná příprava dat se skládala ze dvou fází, napřed se provedla korekce v časové a poté až ve frekvenční doméně.

Shrňme tedy krok po kroku postup, který byl použit pro prvotní přípravu dat:

1. Byly odstraněny signály A_1 a A_2 odpovídající kanálům 20 a 21, neboť se jedná jen o zemnící elektrody na ušních boltcích.
2. Byl odstraněn začátek a konec dat, kdy docházelo k zapínání a vypínání přístroje. Bylo odstraněno přesně 20 s na začátku a 10 s na konci.
3. Od signálu byl odečten jeho lineární trend (fce `detrend` v MATLABu).
4. Nyní mohla být provedena renormalizace signálu pomocí MAD.
5. Následný krok byla detekce skoků v signálu způsobená pravděpodobně svalovou aktivitou pacienta.
6. Malá okolí skoků byla nahrazena nulou, neboť takto bude nejméně ovlivněno frekvenční spektrum signálu. Jako alternativa byla otestována i neúplná fourierova trasnformace [19], ale nebyla pozorována žádná významná změna.
7. Opět byl odečten lineární vývoj signálu, ale tentokrát mezi každým skokem zvlášť.

8. Poslední krok byla opětovná normalizace opraveného signálu pomocí MAD.

Ukázka signálu před a po této korekci je v Obr. 2.2. Bez odstranění skoků byla u mnoha pacientů pozorována nevysvětlitelná mozková aktivita mezi 0.1–1 Hz, která se po této korekci ztratila.

Ve frekvenční doméně bylo nezbytné udělat korekce vzhledem k tomu, že signál mezi 45–55 Hz a od 61 Hz výš byl potlačen o cca 20 dB pomocí frekvenčních filtrů. V rozsahu 45–55 Hz se nachází nízkofrekvenční signál od zásuvky, zatímco na frekvencích nad 61 Hz se nejspíše nachází šum dalších elektrických přístrojů. Z tohoto důvodu nebyly tyto oblasti použity k další analýze. Příklad jednoho Fourierova spektra, získaného jako průměr všech kanálů jediného pacienta, je vykreslený v Obr. 2.3.

2.2 Windowing

Často řešeným problémem Fourierovy transformace jsou okrajové jevy. Předpokládejme, že máme nekonečně dlouhý diskrétní signál x_n. Tento signál vynásobíme obdélníkovou funkcí R_n^M, jejíž nenulová oblast má délku M, tzv. *oknem*. Získáme tím vlastně výřez signálu x_n^M délky M, neboť s nekonečně dlouhým signál se špatně pracuje. Zajímá nás, jaký to bude mít efekt na spektrum původního signálu. Matematicky to lze zapsat jako

$$x_n^M = R_n^M x_n.$$

Nás ale zajímá $\mathcal{F}[x^M]$:

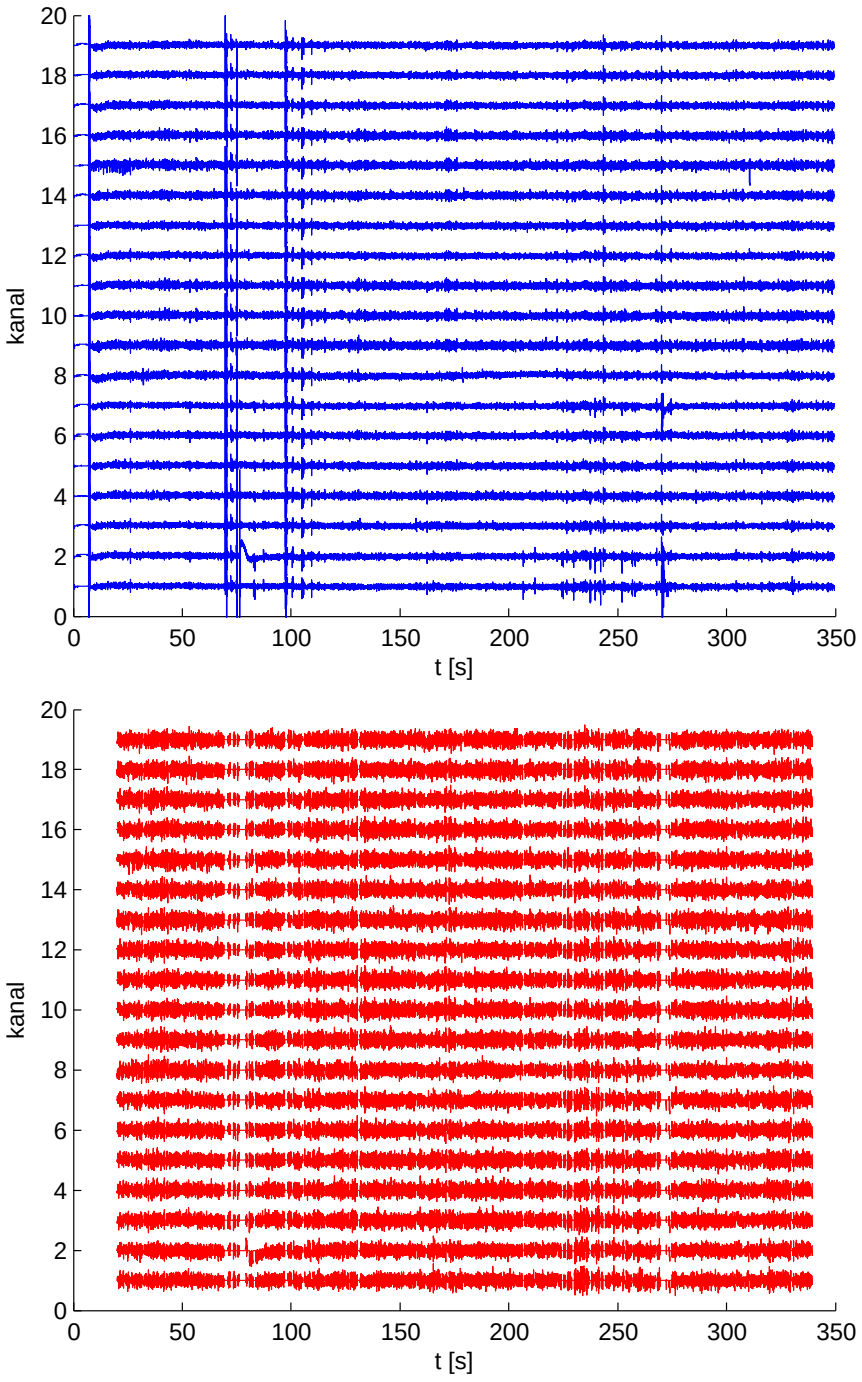
$$\mathcal{F}[x^M] = \mathcal{F}[R^M x] = \mathcal{F}[R^M] * \mathcal{F}[x].$$

Vynásobení obdélníkovým (nebo libovolným jiným) oknem v časové doméně způsobí rozmazání spektra originálního signálu konvolucí s Fourierovou transformací tohoto okna. Otázkou je, k jak významnému rozmazání dojde, a zda-li bude mít volba okna vliv na dosažené výsledky v analýze EEG signálu. V obrázku 2.4 je vykreslena absolutní hodnota FT dalších běžně používaných oken, očividně mají všechny řádově podobnou šířku.

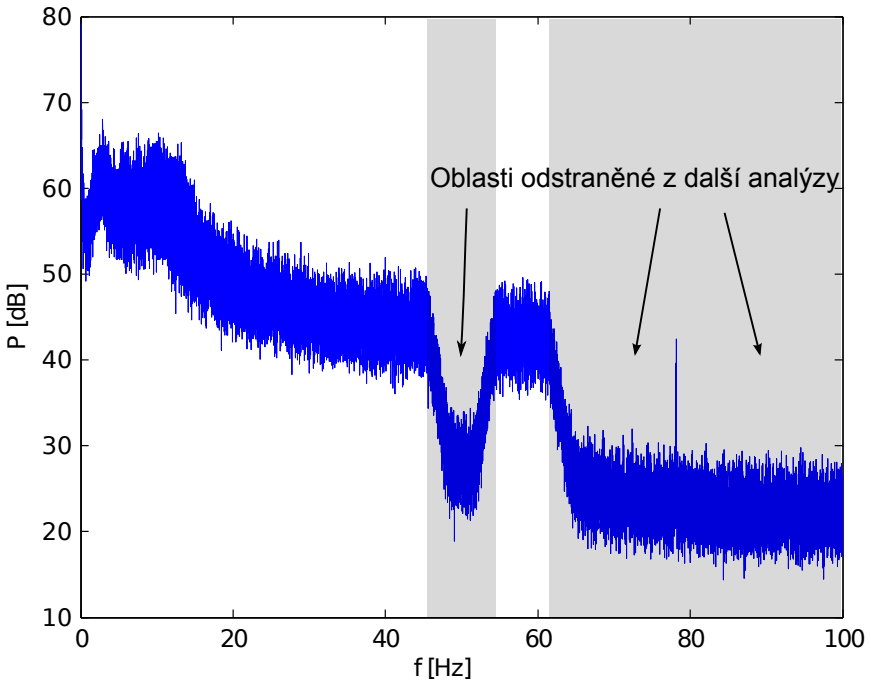
Pro svoji jednoduchost analyzujme Gaussovské okno, pro další okna bude výsledek řádově podobný. Platí, že

$$e^{-\frac{x^2}{2\sigma^2}} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{\sigma} e^{\frac{\sigma^2\omega^2}{2}}.$$

Například, použijeme-li okno šířky 10 000 vzorků, 50 s, což je méně než nejužší okno použité ve všech následujících analýzách, zjistíme, že šířka okna ve spektrální doméně je $1/50 = 0.02$ Hz. To také představuje odhad limitu rozlišení daný tímto oknem. Významné rysy pozorované ve spektru měly šířku alespoň 1 Hz, tedy daleko větší, než-li je limit rozlišení. Ale i přesto bylo na každý signál před Fourierovou transformací aplikováno Hammingovo okno.



Obrázek 2.2: Srovnání EEG signálu před provedením korekcí (1. graf) a po provedení korekcí (2.graf)



Obrázek 2.3: Příklad spektra mozkové aktivity jednoho ze zdravých pacientů.

2.3 Příznakové modely

Cílem příznakových (klasifikačních) modelů je popsat reálný signál způsobem, který umožní další zpracování statistickými metodami. Cílem je především snížit dimenzi zkoumaného systému. Každý pacient je totiž popsán řádově 10^5 body, přičemž skutečná informace je velmi řídce skrytá "někde uvnitř". Prvním krokem u většiny následujících metod je přechod do vhodnější báze pomocí Fourierovy transformace. Signál je totiž složen nejen ze šumu, ale i z velkého množství mozkových vln různých frekvencí. Analýza výkonnostního spektra signálu neumožňuje sice analyzovat jednotlivé tyto vlny, ale po správném znormování získáme hustotu pravděpodobnosti výskytu dané frekvence ve spektru. Píky, které poté ve spektru vidíme, vznikly vystřeďováním velkého množství jednotlivých vln.

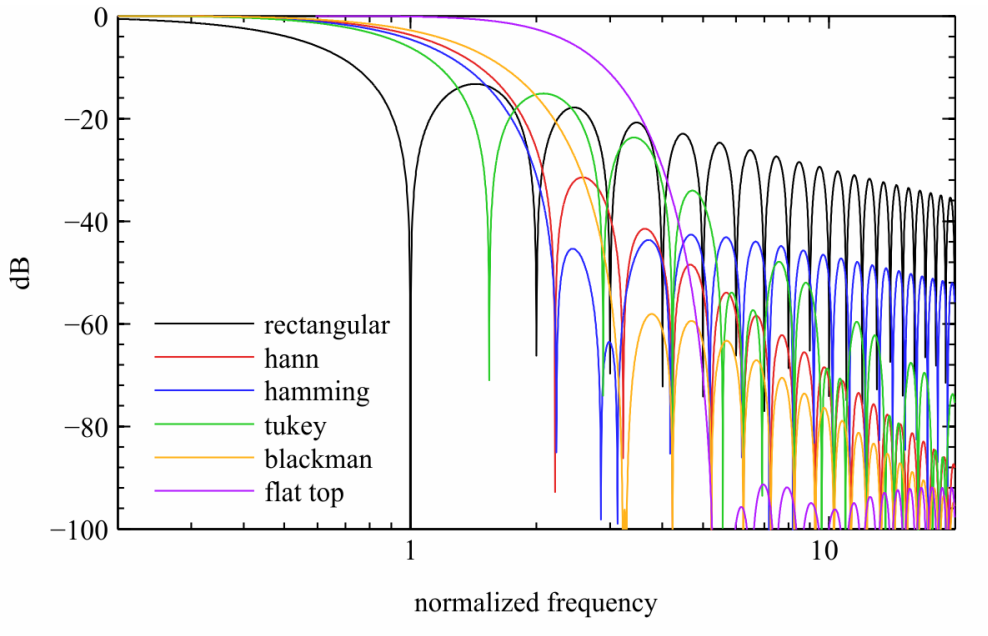
2.3.1 Relativní spektrální výkon

Nejprve byla provedena nejjednodušší analýza založená na spočtení relativního výkonu ve vybrané spektrální oblasti. Relativní výkon byl spočten podle vzorce

$$P_{\langle a,b\rangle} = \frac{\int_{f_{\min}}^{f_{\max}} \chi_{\langle a,b\rangle} |X(f)|^2 \mathrm{d}f}{\int_{f_{\min}}^{f_{\max}} |X(f)|^2 \mathrm{d}f}, \tag{2.1}$$

kde $X(f)$ je Fourierova transformace EEG signálu a $\chi_{\langle a,b\rangle}$ je charakteristická funkce intervalu hledaných mozkových vln. Intervaly $\langle a, b\rangle$ byly použity z tabulky 1.

Tato metoda byla použita především, aby bylo možno provést srovnání s ostatními



Obrázek 2.4: Tvar absolutní hodnoty Fourierovy transformace nejběžněji používaných oken. Obrázek byl převzat z `en.wikipedia.org/wiki/Window_function`

pracemi na toto téma.

2.3.2 Cepstrum

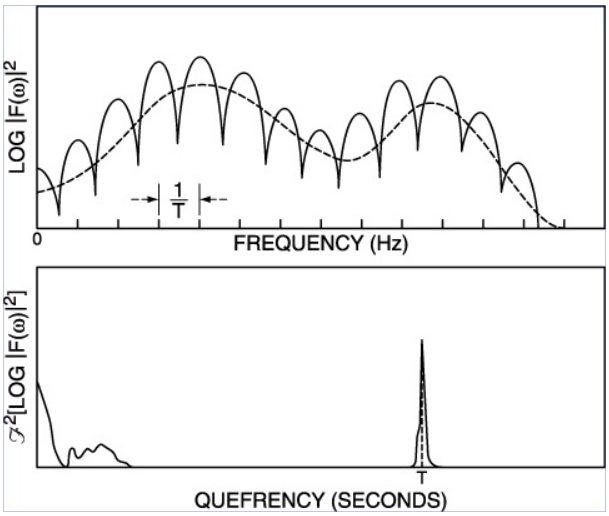
Cepstrum je definováno následujícím způsobem [4]

$$C[f](T) = \left|\mathcal{F}^{-1}\left[\log\left(|\mathcal{F}[f(t)]|^2\right)\right]\right|^2,$$

jedná se tedy o kvadrát absolutní hodnoty inverzní Fourierovy transformace logaritmu výkonnostního spektra. Veličinou, na níž cepstrum závisí, již není frekvence, ale tzv. quefrence, jejíž jednotkou je sekunda. Cepstrum analýza je primárně založená na hledání periodicity ve výkonnostním spektru.

Liftering

Stejně jako na klasické Fourierovo spektrum lze na cepstrum lze aplikovat "quefrenční" filtr. I tato operace má speciální název - *liftering*. Aplikujeme-li na cepstrum nejjednodušší nízkofrekvenční filtr, Fourierovou transformací takového cepstra získáme ve frekvenční doméně hladké spektrum, u kterého lze snížit rozlišení a tím snížit počet analyzovaných dimenzí.



Obrázek 2.5: Ilustrační příklad transformace výkonnostního spektra na cepstrum převzatý z [15]

2.3.3 Vyhlazené FFT spektrum s adaptivní šířkou okna

Myšlenka, jak snížit dimenzi problému, založená na lifteringu, je v principu správná, ale moc dobře nefunguje, neboť zkoumáme signály přes velmi široký rozsah frekvencí od 1 do 60 Hz. Provedením lifteringu se buď ztratí důležité rysy na nízkých frekvencích, anebo se neodfiltruje veškerý šum na vysokých frekvencích. Pro naše účely by bylo vhodné provést vyhlazení spektra, které má nejen zlogaritmovanou amplitudu, ale i frekvenci.

Toho cíle bylo dosaženo definicí filtru vyhlazujícího data pomocí gaussovského okna konstantní relativní šířkou okna

$$\hat{P}_i = \frac{\sum_j w_{ij} P_j}{\sum_j w_{ij}},$$

kde váhová funkce w_{ij} byla definována jako

$$w_{ij} = \exp\left(-\left(\frac{f_j - F_i}{F_i \gamma}\right)^2\right).$$

f_j je lineární frekvenční vektor příslušný ke spektru P_j, F_i je exponenciálně rostoucí frekvenční vektor od $f_{\min}$ po $f_{\max}$ a nakonec γ je shlazovaní faktor, udávající šířku okna.

2.3.4 Autoregresivní model

Autoregresivní model, nazývaný též lineární prediktivní model, je založený na předpokladu, že z k posledních měření $Y_{n-1} \ldots, Y_{n-k}$ lze predikovat budoucí hodnotu Y_n na základě lineárního vztahu

$$Y_n = \sum_{i=1}^{k} \beta_i Y_{n-i} + \varepsilon_n, \tag{2.2}$$

kde β_i jsou neznámé parametry modelu a ε_n je náhodná složka Y_n s normálním rozdělením. Pro nalezení optimálních hodnot β_i použijeme našeho EEG signálu délky N. Protože je to konečný signál, je nutné dodat okrajové podmínky. Nejjednodušší volbou je cyklická okrajová podmínka $Y_{-k} = Y_{N-k} \quad \forall k \in \hat{N}$. Poté je možné rovnici (2.2) zapsat v maticovém zápisu

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} Y_0 & Y_{-1} & Y_{-2} & \dots Y_{-p} \\ Y_1 & Y_0 & Y_{-1} & \dots Y_{-p+1} \\ Y_2 & Y_1 & Y_0 & \dots Y_{-p+2} \\ \vdots & \vdots & \vdots & \ddots \\ Y_{N-1} & Y_{N-2} & Y_{N-3} & \dots Y_{N-p} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

kde vztah mezi vektorem $\mathbf{Y} \in R^N$ a vektorem $\beta \in R^p$ určuje matice $\mathbb{A} \in R^{N,p}$. Ovšem problém je, že my matici $\mathbb{A}$ neznáme, hodnoty Y_k, které jsou změřené, obsahují také náhodnou chybu. Je vícero možností řešení, buď je možné tento fakt, ignorovat anebo lze provést více průchodovou metodu, kdy předchozí průchod bude použit k lepšímu odhadu hodnot Y_n v matici $\mathbb{A}$. A jako počáteční odhad se použijí surová data. Tato metoda se nazývá Burgova metoda [6].

Za předpokladu, že náhodné chyby ε_i jsou nezávislé, mají nulovou střední hodnotu a mají stejný rozptyl, a navíc, pokud je hodnota matice $\mathbb{A}$ větší než p, můžeme k nalezení optimálních parametrů β_i použít obyčejnou metodu nejmenších čtverců. Cílem této metody je nalezení takových koeficientů β_i, které budou minimalizovat kvadratickou odchylku dat od predikce, tzv. reziduum

$$\min \|\mathbb{A}\beta - \mathbf{Y}\|_2^2,$$

derivací podle vektoru β získáme vztah

$$\beta = (\mathbb{A}^T\mathbb{A})^{-1}\mathbb{A}^T\mathbf{Y},$$

který lze vyřešit pro malé p i přímou inverzí matice $\mathbb{A}^T\mathbb{A}$. Případně je možné použít standardní metody na řešení metody nejmenších čtverců založené na nalezení singulárních hodnot (SVD), anebo QR dekompozici. Nalezení optimální hodnoty p lze provést tak, aby pro větší hodnoty p se hodnota rezidua již téměř neměnila.

Spektrum autoregresního modelu

Na vzorec (2.2) lze pohlížet i jako na diskrétní konvoluci vektoru Y_k s vektorem β_{p-k}. Vektor β je tady zároveň i filtrem, a proto zkusme zjistit, jak se bude tento model chovat v přítomnosti signálu $Z_n = Ae^{2\pi nf}$ o frekvenci f

$$|Z_n - \sum_{i=1}^{k} \beta_i Z_{n-i}|^2 = \langle \varepsilon_n^2 \rangle,$$

kde $\langle \varepsilon_n^2 \rangle$ je rovno rozptylu veličiny ε_n. Dosazením za Z_n a vyjádřením A, vyjde

$$A^2(f) = \frac{\sigma_n^2}{|1 - \sum_{i=1}^{k} e^{-2\pi i f}|^2}.$$

Hodnota veličiny $A^2(f)$ představuje výkonnostní spektrum autoregresního modelu. V principu vypadá obdobně jako výkonnostní spektrum Fourierovy transformace.

Kapitola 3

Statistické vyhodnocení klasifikačních modelů

Provedení Fourierovy transformace EEG signálu je pouze prvním z mnoha kroků nezbytných k dosažení našeho cíle – rozlišení mezi pacienty s Alzheimerovou nemocí a zdravými pacienty.

Následujícím krokem je nalezení optimálního *klasifikačního modelu*. Klasifikační (někdy nazývaný *příznakový*) model je zobrazení, které každého pacienta popíše jediným bodem v parametrickém prostoru s minimálním počtem dimenzí (pokud možno jedinou). Dalším krokem je *binární klasifikace* – nalezením hranice nejlépe oddělující tyto dvě skupiny. Ovšem z dosavadních výsledků dosažených v této oblasti to vypadá, že jednoznačné oddělení není možné [9]. Prvním z důvodů je, že nelze přesně říci, v jaké fázi nemoci se pacient nachází. Tudíž tento problém není zcela nejvhodnější pro binární klasifikaci. Lepší by bylo rozdělení, kdy by vznikly tři oblasti – zdraví, nemocní a potenciálně nemocní. A druhý problém je, že i další choroby mozku spojené se stárnutím mohou mít podobné projevy na EEG signálu.

Pro začátek se ale omezíme jen na základní přístup založený na oddělení v jediné dimenzi a statistickém vyhodnocení.

3.1 Studentův test

Nejběžněji používané statistické vyhodnocení modelu je založené na předpokladu, že optimální model by měl být schopen s největší pravděpodobností prokázat rozdíl mezi skupinou AD a CN pacientů na kontrolním vzorku dat.

Matematicky to lze formulovat pomocí testování nulové hypotézy H_0 oproti alternativní hypotéze H_1. Tato metoda neumožňuje potvrzení hypotézy H_0, ale za určitých předpokladů může dojít k jejímu zamítnutí ve prospěch H_1 na hladině významnosti $\alpha \cdot 100\%$.

V našem případě definujeme H_0 a H_1 následujícím způsobem:

$\mathbf{H}_0$: střední hodnota deskriptivní veličiny modelu je totožná pro CN a AD skupinu

$\mathbf{H}_1$: střední hodnoty se liší

K testování lze použít buď jednoduchý dvouvýběrový Studentův t-test daný vzorcem

$$t^* = \sqrt{\frac{mn(m+n-2)}{n+m}} \frac{\mu_{\mathrm{AD}} - \mu_{\mathrm{CN}}}{\sqrt{(n-1)s_{\mathrm{AD}}^2 + (m-1)s_{\mathrm{CN}}^2}}, \tag{3.1}$$

kde t^* je veličina se Studentovým rozdělením a $m+n-2$ stupni volnosti, μ_X je průměr a s_X výběrový rozptyl veličiny. V případě, že by nám šlo pouze o to rozhodnout, zda-li se tyto dvě skupiny staticky významně liší, mohli bychom nalézt kritickou hodnotu t-testu pro určitou hladinu spolehlivosti a provést vyhodnocení. My ale naopak nalezneme tzv. p-hodnotu testu, tedy hodnotu pravděpodobnosti, na které by došlo k zamítnutí.

Ovšem mezi zcela základní předpoklady Studentova t-testu patří normální rozdělení měření kolem střední hodnoty. Tento předpoklad není možné na dostupných datech zajistit. Z tohoto důvodu byl použit Mann-Whitney-Wilcoxon (MWW test) neparametrický test, který normální rozdělení nepředpokládá. Citlivost tohoto testu je pouze o 5% nižší než má t-test [18], ale pro rozdělení, která nejsou normální, dosahuje vyšší účinnosti.

Nicméně princip je totožný, výsledkem (též jako u klasického Studentova testu) je p-hodnota, kterou lze použít pro srovnávání modelů.

3.2 ROC křivka

Na druhou stranu p-hodnota není zcela vhodná pro porovnávání datových sad rozdílné velikosti, neboť její velikost klesá s $m+n$ nezávisle na "kvalitě oddělení" těchto dvou skupin. Hodnota t^* daná rovnicí (3.1) totiž roste pro $m \approx n$ jako $\sqrt{m}$, ačkoli μ_{X} a s_{X} konvergují ke konstantní hodnotě. Proto i p-hodnota bude nevyhnutelně klesat. MWW má asymptoticky totožné chování, a proto se jeho p-hodnota bude chovat obdobně. I pro nepříliš dobré testy vycházely hodnoty 10^{-30} až 10^{-60}. A nakonec, z praktického hlediska nás nezajímá, zda-li měly dvě skupiny dané stejný průměr, naopak nás zajímá, jak dobře jsou skupiny oddělené, kolik procent nemocných pacientů bylo neidentifikováno, a naopak kolik zdravých bylo diagnostikováno s pozitivním výsledkem.

Platí, že tyto dva údaje jsou pevně svázány a zlepšením jednoho dochází ke zhoršení druhého z této dvojice parametrů. Vztah mezi nimi je popsán tzv. ROC (Receiver operating characteristic) křivkou.

Předpokládejme, že máme nějaký klasifikační model, jehož výstupem je jedno jediné spojité číslo popisující datový soubor, na nějž byl tento model aplikován. Příkladem může být třeba relativní výkon alfa vln. V takovém případě je nutné najít optimální

práh separující nejlépe tyto dvě zkoumané skupiny. Výsledkem klasifikace mohou potom být 4 výsledky:

- Nemocný pacient byl správně označen jako nemocný - tzv. TP (*True Positive*) skupina.
- Zdravý pacient byl mylně správně označen jako nemocný - tzv. FP (*False Positive*) skupina, označována též jako chyba 1. typu.
- Zdravý pacient byl správně označen jako zdravý - tzv. TN (*True Negative*) skupina.
- Nemocný pacient byl nesprávně označen jako zdravý - tzv. FN (*False Negative*) skupina, chyba 2. typu.

Ještě je nezbytné si nadefinovat několik pojmů, které budeme dále používat:

Sensitivita , také označována jako TPR (True Positive Rate), je definovaná následovně

$$TPR = TP/(TP + FN)$$

Specificita , také označována jako TNR (True Negative Rate), je definovaná následovně

$$TNR = TN/(FP + TN)$$

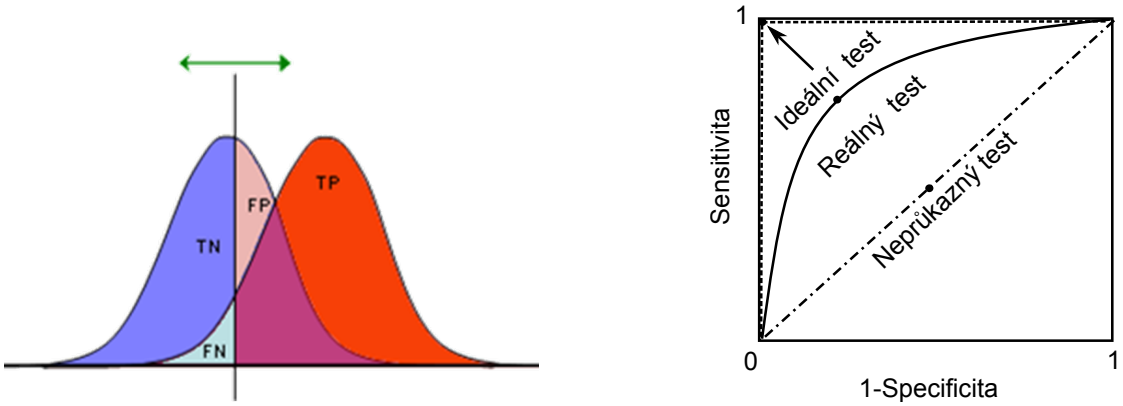
FPR False Positive Rate, je špatná identifikace zdravých vůči všem zdravým

$$FPR = 1 - TNR = FP/(FP + TN)$$

ROC křivka je potom graf závislosti sensitivity (TPR) na 1-specificitě (FPR). Ilustrační příklad je v Obr. 3.1. Změnou prahu pro detekci se budeme pohybovat po ROC křivce, která je právě tímto prahem parametrizována. V případě neprůrazného testu, tedy například, když se distribuční funkce veličin vrácených klasifikačním modelem zcela překrývají, anebo si experimentátor místo pracného měření jen házel mincí, jsou si specificita a 1-sensitivita rovny a ROC křivka je diagonálou. Naopak v ideálním testu jsou sensitivita i specificita rovny jedné a bod (0,1) v grafu ROC křivky se nazývá *dokonalé oddělení* .

3.2.1 Plocha pod ROC křivkou

Důležitý poznatek, že tvar ROC křivky je nezávislý na volbě detekčního prahu a záleží jen na klasifikačním modulu, lze využít k volbě optimálního modelu. K tomuto účelu se používá právě plocha pod ROC křivkou, označovaná též AUC (Area Under Curve), která udává pravděpodobnost, že klasifikační model ohodnotí náhodně vybraného nemocného pacienta vyšší hodnotou než náhodně vybraného zdravého pacienta [10].



Obrázek 3.1: Ilustrační příklad dvou oddělovaných skupin a příslušná ROC křivka udávající specificitu a sensitivitu v závislosti na hodnotě prahu.

AUC se dá vypočítat pomocí vzorce

$$AUC = \sum_{k=1}^{n} \frac{1}{2} \left(TPR_k + TPR_{k-1}\right) \left(FPR_k - FPR_{k-1}\right) \approx \int_{-\infty}^{\infty} y(\rho) \frac{dx}{d\rho}(\rho) d\rho,$$

jedná se o diskrétní aproximaci spojité integraci přes parametr prahu ρ.

Srovnejme teď vlastnosti p-hodnoty a AUC:

Vlastnost	p-hodnota	AUC
Význam	pravděpodobnost, že střední hodnoty obou skupin jsou totožné	pravděpodobnost, že náhodný pozitivní vzorek bude větší než náhodný negativní vzorek.
Závislost na velikosti vzorku dat	exponenciálně klesá jako $\propto \operatorname{erfc} \alpha\sqrt{n}$	konverguje ke konstantě
Rozsah	$(0, 1\rangle$	$(0, 1)$, ale prakticky $\langle 0.5, 1)$
optimální hodnota	konverguje “exponenciálně” k 0	“lineární” konvergence k 1

Jako hlavní výhodu p-hodnoty lze považovat to, že je obecně používaná v odborných medicínských publikacích na podobná témata. Ale z důvodů uvedených výše ji nelze použít pro srovnávání různých databází.

3.3 Lineární diskriminantní analýza (LDA)

Často dochází k tomu, že příznakový (klasifikační) model nevrátí jediné číslo, ale bod v vektorovém prostoru. Na takový případ již nejde použít Studentův test, ani jednorozměrnou verzi ROC křivky. Měřená data pak představují realizaci vícerozměrné náhodné veličiny. Otázkou je, jak je nejlépe oddělit.

Pro zjednodušení předpokládejme, že hustoty podmíněných pravděpodobností toho, že bod $\vec{x}$ je AD, $p(\vec{x}|\text{AD})$ resp. $\vec{x}$ je CN, $p(\vec{x}|\text{CN})$, mají vícerozměrným normální

rozdělení Σ (pro zjednodušení předpokládejme regularitu Σ)

$$\mathcal{N}(\vec{\mu_i}, \Sigma_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu_i})\Sigma_i^{-1}(\vec{x} - \vec{\mu_i})^T\right) \qquad i \in \{\mathrm{AD}, \mathrm{CN}\}, \quad (3.2)$$

kde $\vec{\mu_i}$ je střední hodnota a Σ_i je kovariantní matice. Za těchto předpokladů je Bayesovsky optimálním řešením, pokud je hodnota logaritmu věrohodnostní funkce $L = p(\vec{x}|\mathrm{AD})p(\vec{x}|\mathrm{CN})$ větší než nějaký práh T [11]. Zapíšeme-li to ve formě nerovnice a dosadíme-li za $p(\vec{x}|\mathrm{AD})$, $p(\vec{x}|\mathrm{CN})$ normální rozdělení (3.2), získáme

$$(\vec{x} - \vec{\mu}_{\mathrm{AD}})^T\Sigma_{\mathrm{AD}}^{-1}(\vec{x} - \vec{\mu}_{\mathrm{AD}}) + \ln|\Sigma_{\mathrm{AD}}| - (\vec{x} - \vec{\mu}_{\mathrm{CN}})^T\Sigma_{\mathrm{CN}}^{-1}(\vec{x} - \vec{\mu}_1) - \ln|\Sigma_{\mathrm{CN}}| \ < \ T,$$

jedná se očividně o předpis kvadriky, a bez dalších předpokladů by řešení vedlo na *kvadratickou diskriminantní analýzu* (QDA). Ale budeme-li předpokládat homoskedasticitu, tedy že kovariance obou souborů dat jsou stejné, předchozí vzorec se výrazně zjednodušší

$$\vec{w} \cdot \vec{x}^T < -\frac{1}{2}(T - \vec{\mu}_{\mathrm{AD}}^T\Sigma^{-1}\vec{\mu}_{\mathrm{AD}} + \vec{\mu}_{\mathrm{CN}}^T\Sigma^{-1}\vec{\mu}_{\mathrm{CN}}) = c, \quad (3.3)$$

kde váhový vektor $\vec{w}$ je definován jako $\vec{w} = \Sigma^{-1}(\vec{\mu}_{\mathrm{CN}} - \vec{\mu}_{\mathrm{AD}})$. Získáváme tedy podmínku, že dané dvě skupiny budou odděleny plochou, jejíž normálový vektor je určený váhovým vektorem $\vec{w}$. Tato metoda se nazývá lineární diskriminantní analýza. Teoreticky je možné i zobecnění na klasifikaci více než dvou skupin [20].

Je ale nezbytné nezapomínat, že každá metoda automatické klasifikace je jenom tak dobrá, jak kvalitní jsou použité příznaky. Bez kvalitního klasifikačního modelu je aplikace LDA zbytečná.

3.3.1 Regularizovaná LDA

Při reálné aplikaci předchozího postupu se projevil problém zcela typický pro klasifikační problémy, došlo k tzv. *přefitování* (overfitting), neboť použitím prosté inverze (resp. *pseudoinverze*) matice Σ bylo dosaženo 99% separování obou skupin, což je zcela nereálný výsledek bez jakékoli prediktivní schopnosti. Z toho důvodu se muselo přistoupit k *regularizaci* kovariantní matice Σ. Regularizace znamená, že nalezneme řešení splňující ještě další omezení, než jen podmínku $\min_w \|\vec{w} - \Sigma\Delta\vec{\mu}\|$.

Byly otestovány dvě metody.

L1 regularizace

$$\min_w \|\vec{w} - \Sigma\Delta\vec{\mu}\|_2 + \lambda \|\vec{w}\|_1$$

implementována s pomocí algoritmu na lineární programování [17]. λ je regularizační parametr udávající, jak moc výrazná má být regularizace.

Druhá možnost regularizovaná inverze byla provedna s pomocí TSVD (Truncated Singular Value Decomposition) metody, kdy byly z Σ odstraněny před inverzí vlastní

čísla (resp. *singulární hodnoty*) menší než zvolený práh. Tato metoda byla realizována funkcí `pinv` v MATLABu.

L1 regularizace preferuje řídký tvar váhového vektoru s maximem nulových složek. TSVD také tlačí složky váhového vektoru k nule, ale výsledkem je jednoduchá hladká křivka.

3.3.2 Další obtíže

Aplikaci LDA provázelo ještě několik obtíží. Předně nebyl dostatečný počet pacientů, aby šlo vůbec udělat smysluplný odhad kovariantní matice. Z toho důvodu byly všechny signály rozsekány na kusy o délce 2^{14} a ty byly vyšetřovány samostatně. Vzhledem k tomu, že v EEG signály nejsou zcela stacionární [13], zahrneme tak do modelu i rozptyl příznaků v rámci jediného pacienta, což lze považovat za přínosný krok.

Druhý problém je nalezení optimální hodnoty regularizačního parametru λ. Za správný postup by šlo považovat například rozdělení datové sady do učící, trénovací a testovací sady [23]. Učící sada by se použila k nalezení $\vec{w}$, cross-variací[1] prvků, mezi učící se a trénovací sadou by se nalezlo λ s nejlepší schopností predikce. A v dalším kroku by se využila testovací sada na to, aby se zkontrolovalo, jestli model skutečně dobře funguje.

Ovšem na provedení takto složité procedury nebyl dostatek dat, a tak byla λ odhadnuta většinou tak, aby vznikl nejjednodušší možný model, který je ještě schopen predikce, tedy kdy váhový vektor w ještě nebyl zcela nulový. Navíc bylo otestováno, že stejný model vzniká i na všech ostatních kanálech, které představují alespoň částečně nezávislé měření.

3.4 Pravděpodobnostní sigmoid

Žádná z předchozích metod nebyla schopna dosáhnout úplného oddělení obou skupin. Proto by bylo užitečné určit nejen do které ze skupin testovaný pacient patří, ale i pravděpodobnost, že tam leží. To ale přímo LDA metoda, ani žádná jiná z testovaných metod, nedovede. Nejjednodušší cestou, jak toho docílit, je *naivní Bayesova metoda* [2]. Chtějme třeba určit, že pacient leží v AD. Dále předpokládejme, že AD i CN skupina mají normální rozdělení klasifikačního parametru L kolem střední hodnoty μ_{AD} a μ_{CN}, a ještě předpokládejme homoskedasticitu. Klasická Bayesova věta potom tvrdí, že

$$p(\mathrm{AD}|L) = \frac{p(L|\mathrm{AD})}{p(L|\mathrm{AD}) + p(L|\mathrm{CN})},$$

[1] cross-validace je postup, kdy se k učení rozhodovacího algoritmu použije náhodná část dat a zbytek se tímto spočteným modelem predikuje. Cílem je maximalizovat úspěšnost predikce

kde pravděpodobnosti $p(L, \mathrm{AD})$ a $p(L, \mathrm{CN})$ známe. Dosazením předpokladu o normálnosti a homoskedasticitě těchto dvou rozdělení vyjde

$$p(\mathrm{AD}|L) = \frac{\exp(-(L-\mu_{\mathrm{AD}})^2/(2\sigma^2))}{\exp(-(L-\mu_{\mathrm{AD}})^2/(2\sigma^2)) + \exp(-(L-\mu_{\mathrm{CN}})^2/(2\sigma^2))} =$$

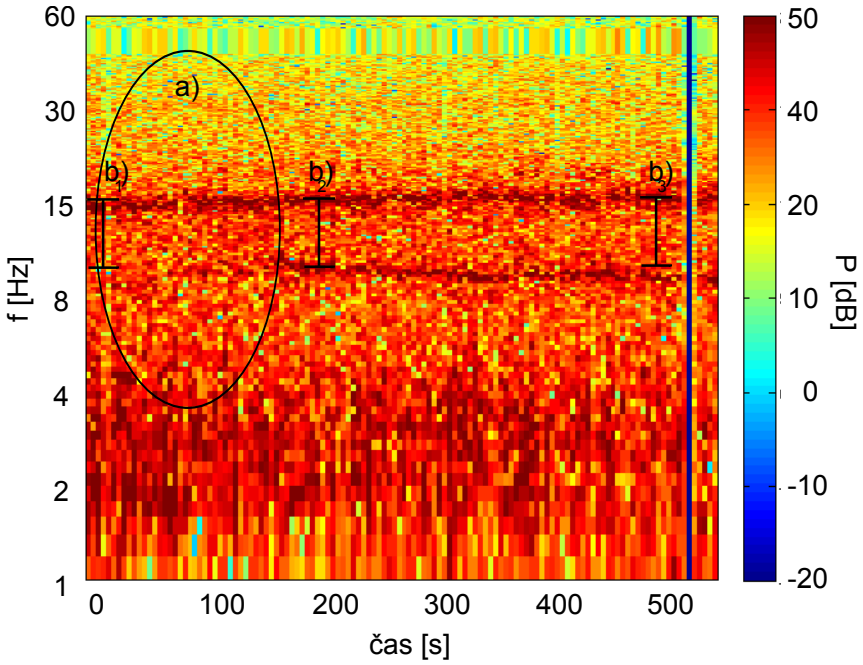
$$= \frac{1}{1+\exp\left(\frac{-(L-\mu_{\mathrm{CN}})^2+(L-\mu_{\mathrm{AD}})^2}{2\sigma^2}\right)} = \frac{1}{1+\exp\left(\frac{\mu_{\mathrm{CN}}-\mu_{\mathrm{AD}}}{\sigma^2}\left(L-\frac{\mu_{\mathrm{AD}}+\mu_{\mathrm{CN}}}{2}\right)\right)}$$

získali jsme tzv. pravděpodobnostní sigmoid $f(x) = \frac{1}{1+\exp(x)}$ se středem v průměru μ_{AD} a μ_{CN}.

Kapitola 4

Analýza EEG signálů

V této kapitole budou shrnuty výsledky získané pečivou analýzou EEG signálů. Postupně budou otestovány všechny metody z kapitoly 2.3. Aby bylo možné srovnání mezi různými metodami, byly všechny vypočteny stejnou cestou. Signály pacientů, které měly různou délku od 60 000–120 000 vzorků, byly rozděleny na úseky délky $2^{14} = 16384$ a všechny kanály byly analyzovány zvlášť. Toto řešení bylo zvoleno ze dvou důvodů, předně tím byla do databáze zahrnuta i variabilita signálu v rámci jednoho pacienta. EEG signál totiž není stacionární, a docházelo k malým, ale potencionálně důležitým změnám ve spektru. Pro ilustraci je v Obr. 4.1 znázorněn spektrogram jednoho z pacientů, kdy tyto změny byly výrazně vidět. Druhý důvod byl, že by se jinak musel EEG signál zkrátit na signál nejkratšího z pacientů, anebo nastavovat nulami.



Obrázek 4.1: Spektrogram CN pacienta č. 116 – ukázky nestacionarity signálu a) občas se objevily vlny, které tam předtím nebyly, b) celkem často docházelo k pomalé změně jejich frekvence i o 20-30%

4.1 Relativní spektrální výkon

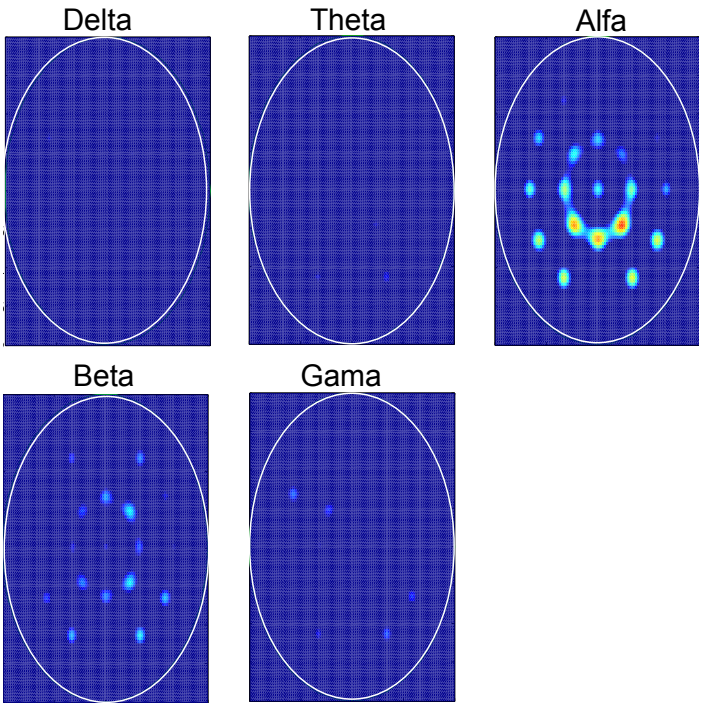
Jako první a základní analýza byl spočten pro každého pacienta relativní spektrální výkon na základních pásmech mozkových vln. Hodnoty AUC parametru spočtené pro všechny kanály a všechny pásma jsou uvedeny v tabulce 4.1.

Tabulka 4.1: Výsledky z L1 regularizované LDA analýzy aplikované na kanál cepstra zvlášť. AUC hodnota 0.5 je u kanálů, kde vyšel váhový vektor nulový

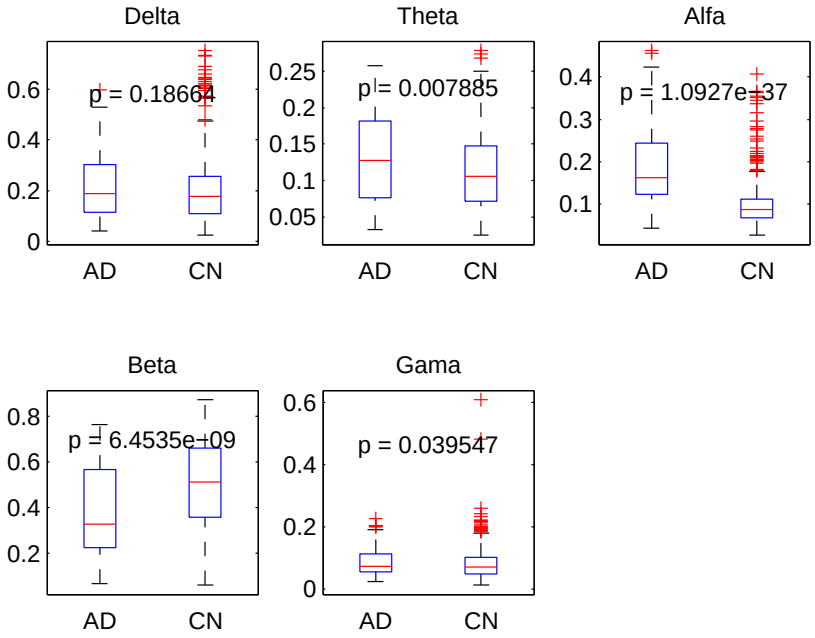
kanál	jméno	AUC delta	AUC theta	AUC alfa	AUC beta	AUC gama
1	Fp1	0.557	0.5759	0.637	0.656	0.561
2	Fp2	0.576	0.545	0.582	0.686	0.525
3	F7	0.616	0.534	0.711	0.567	0.672
4	F3	0.535	0.496	0.679	0.612	0.612
5	Fz	0.533	0.522	0.660	0.644	0.500
6	F4	0.567	0.533	0.623	0.679	0.510
7	F8	0.508	0.525	0.622	0.624	0.565
8	T3	0.573	0.517	0.753	0.553	0.525
9	C3	0.555	0.511	0.731	0.584	0.567
10	Cz	0.505	0.489	0.715	0.610	0.557
11	C4	0.521	0.505	0.712	0.613	0.567
12	T4	0.486	0.508	0.693	0.612	0.495
13	T5	0.538	0.556	0.788	0.652	0.573
14	P3	0.528	0.552	0.793	0.632	0.523
15	Pz	0.502	0.560	0.793	0.632	0.535
16	P4	0.527	0.574	**0.825**	0.676	0.554
17	T6	0.496	0.580	**0.801**	0.700	0.652
18	O1	0.529	0.626	**0.804**	0.704	0.637
19	O2	0.561	0.640	**0.814**	0.728	0.670

Nejvýraznější vliv je vidět v pásmu alfa vln, což odpovídá i pozorování v článku [8], ale narozdíl od tohoto článku nebyl pozorován žádný rozdíl mezi pacienty v theta vlnách. Nejlepšího oddělení bylo dosaženo na kanálech 16–19 nacházejících se na temeni hlavy. Pro názornost byly AUC hodnoty také vykresleny do schematického zobrazení hlavy v Obr. 4.2. Stojí za zmínku, že v případě alfa vln lze pozorovat téměř spojitý pokles se vzdáleností směrem od maximální hodnoty na detektoru P4 a Pz. Krabicové diagramy pro kanál P4 jdou v Obr. 4.3, p-hodnota je rovna 10^{-37}, takže můžeme předpokládat, že průměr těchto dvou skupin se nejspíše skutečně liší.

Vzhledem k tomu, že se jedná o nejjednodušší možnou metodu, lze očekávat, že pokročilejší metody by měly dosáhnout ještě lepších výsledků.



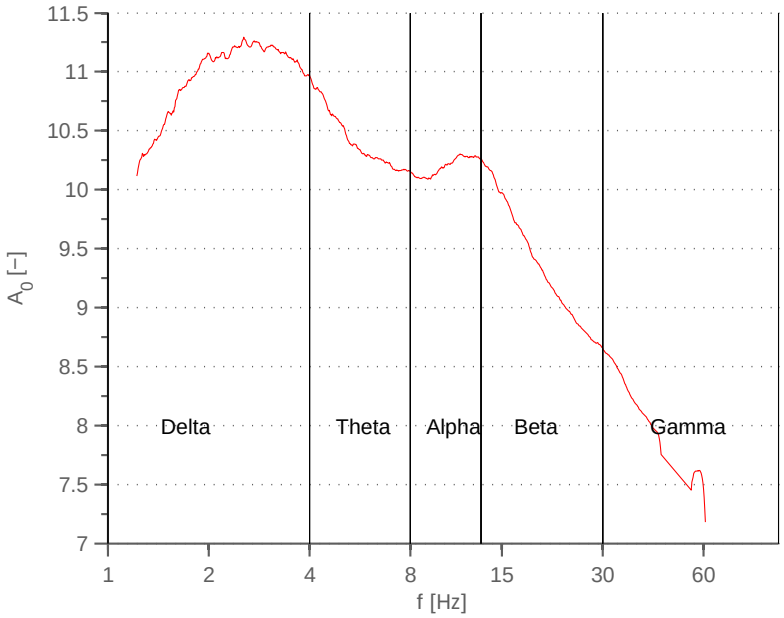
Obrázek 4.2: Hodnoty AUC jednotlivých pásem mozkových vln vykreslené do schematického nákresu hlavy definovaného v Obr. 1. Kromě alfa pásma a malého rozdílu v beta pásmu se AD a CN pacienti téměř neliší.



Obrázek 4.3: Krabicové diagramy relativního výkonu jednotlivých pásem vykreslené pro kanál P4 (16).

4.2 Fourierovo spektrum

Tím, že zkoumáme průměr spektrálního výkonu přes celé pásmo, ztrácíme velké množství informací. Proto je nezbytné prozkoumat Fourierovo spektrum detailněji, abychom zjistili, co se tam skutečně odehrává. K tomuto účelu byla využita metoda vyhlazení Fourierova spektra definovaná v kapitole 2.3.3. Protože nás zajímá pouze rozdíl mezi skupinami, byl ode všech spekter odečten jejich celkový průměr. Toto průměrné pozadí je vykresleno v Obr. 4.4. Odečtení neovlivní výsledky lineární analýzy (třeba LDA), ale umožní přehlednější zobrazení spekter.

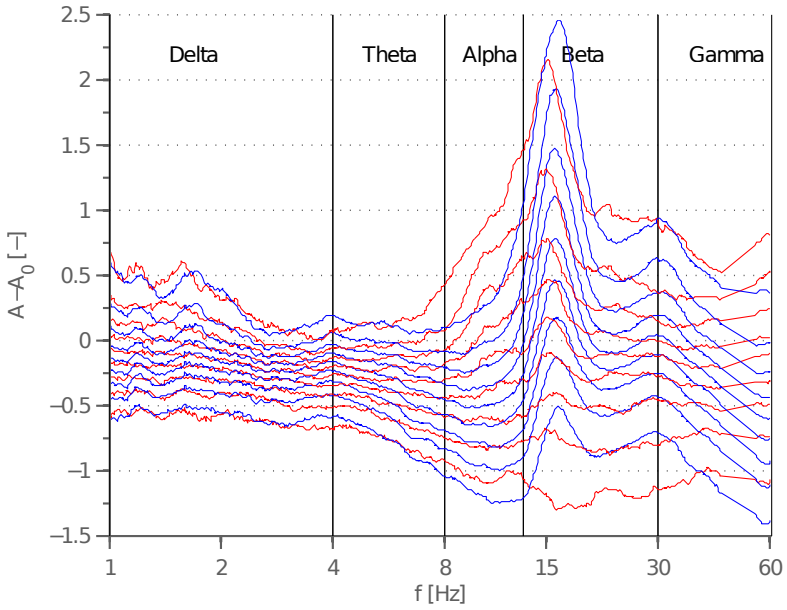


Obrázek 4.4: Konstantní pozadí odečtené ode všech spekter. Tato operace je vlastně ekvivalentní přenormování amplitudy celkovým harmonickým průměrem.

Distribuční funkce podmíněná frekvencí f je vykresená v grafu 4.5. Pro jednu pevnou frekvenci popisuje 9 křivek 10%,... 90% kvantil. Největší rozdíl je vidět právě v alfa vlnách, což odpovídá předchozímu pozorování v Tab. 4.1. Významný rozdíl je také v beta vlnách, ale ten se při předchozí analýze nejspíše ztratil zprůměrováním přes celou oblast.

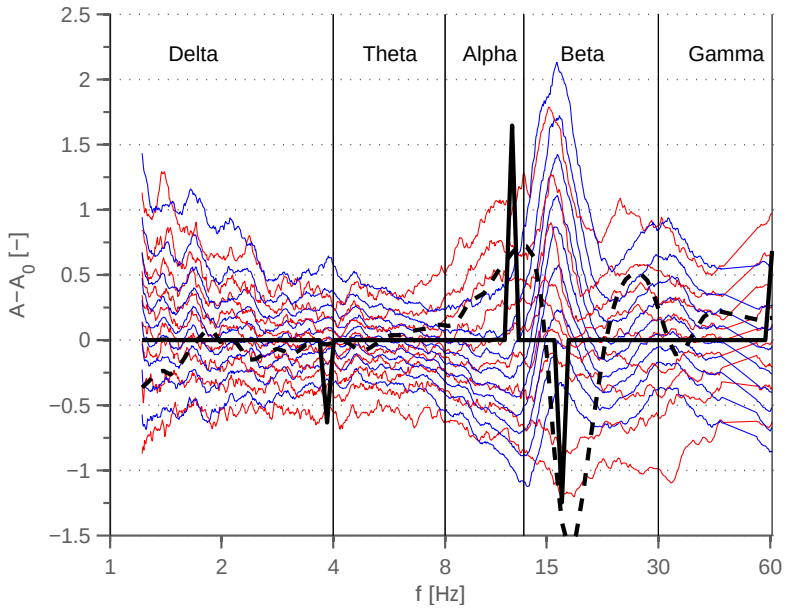
Pro každou jednotlivou diskrétní frekvenci f_n představuje funkce $p(A, f_n)$ jedno jednorozměrné náhodné rozdělení a všechny dohromady tvoří náhodné, konečně rozměrné, rozdělení $p(x_1, \ldots, x_n)$ pro diskrétní frekvence $f_1, \ldots, f_n$. Složky $p(x_1, \ldots, x_n)$ nejsou nezávislé, ale to pro další analýzu není překážkou. Navíc lze toto rozdělení dobře aproximovat mnohorozměrným normálním rozdělením (3.2). Rozdíl v rozptylu mezi AD a CN skupinou je menší než 20%, takže předpoklad homoskedasticity je také přibližně splněn. To stačí pro to, abychom mohli provést analýzu pomocí LDA.

V grafu 4.6 jsou vykresleny váhové vektory získané L1 regularizací a TSVD regularizací pro 11. kanál EEG signálu. LDA algoritmus dokázal přesně určit, které oblasti jsou zajímavé. Váhový vektor na ostatních kanálech vyšel téměř identický. V případě L1 regularizace občas chyběl váhový faktor na 4 Hz a 60 Hz, a naopak se



Obrázek 4.5: Podmíněná distribuční funkce $p(A - A_0|f)$ zdravých (modří) a nemocných (červení) pacientů.

někdy objevil na 25 Hz. To záleží na volbě regularizační konstanty. Váhové faktory v alfa a beta vlnách byly přítomny vždy na stejném místě na všech kanálech.

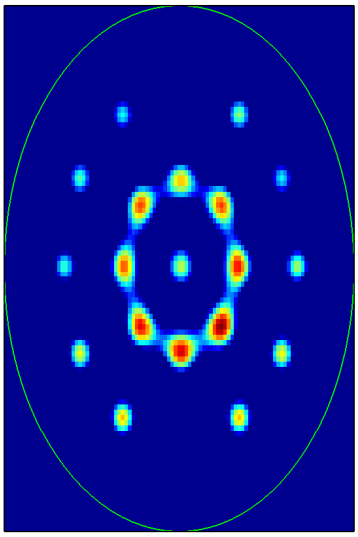


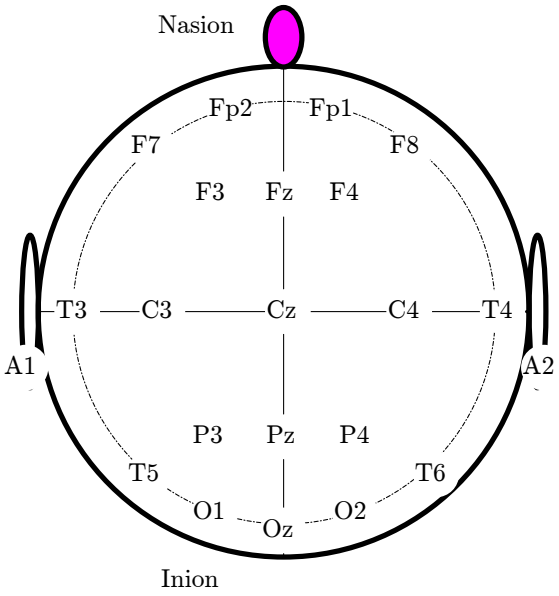
Obrázek 4.6: Váhové vektory získané regularizovanou LDA metodou 3.3.1. Celá čára odpovídá L1 regularizaci, čárkovaná TSVD regularizaci.

Výsledky získané analýzou všech kanálů EEG pro $\lambda = 0.1$ jsou v tabulce 4.2. Nejlepšího výsledku bylo opět dosaženo na kanálu P4, a došlo k 20% zlepšení oproti původní metodě. Hodnoty AUC vykreslené v závislosti na poloze elektrod na povrchu hlavy jsou v Obr. 4.7. Opět jsou nejlepší výsledky v zadní části temene hlavy stejně jako v Obr. 4.2.

Tabulka 4.2: Výsledky z lineárně regularizované LDA analýzy aplikované na kanál EEG zvlášť.

kanál	jméno	AUC	p-hodnota	kanál	jméno	AUC	p-hodnota
1	Fp1	0.792	$1.37{\cdot}10^{-32}$	11	C4	0.847	$8.44{\cdot}10^{-44}$
2	Fp2	0.825	$1.00{\cdot}10^{-34}$	12	T4	0.825	$2.06{\cdot}10^{-46}$
3	F7	0.808	$6.07{\cdot}10^{-43}$	13	T5	0.843	$4.12{\cdot}10^{-47}$
4	F3	0.836	$2.90{\cdot}10^{-36}$	14	P3	0.855	$4.40{\cdot}10^{-51}$
5	Fz	0.805	$1.83{\cdot}10^{-33}$	15	Pz	0.860	$7.04{\cdot}10^{-49}$
6	F4	0.837	$4.16{\cdot}10^{-38}$	16	P4	0.880	$1.54{\cdot}10^{-57}$
7	F8	0.787	$1.89{\cdot}10^{-27}$	17	T6	0.849	$3.21{\cdot}10^{-52}$
8	T3	0.800	$7.22{\cdot}10^{-43}$	18	O1	0.866	$2.95{\cdot}10^{-48}$
9	C3	0.827	$1.13{\cdot}10^{-37}$	19	O2	0.868	$1.30{\cdot}10^{-52}$
10	Cz	0.816	$4.06{\cdot}10^{-35}$				

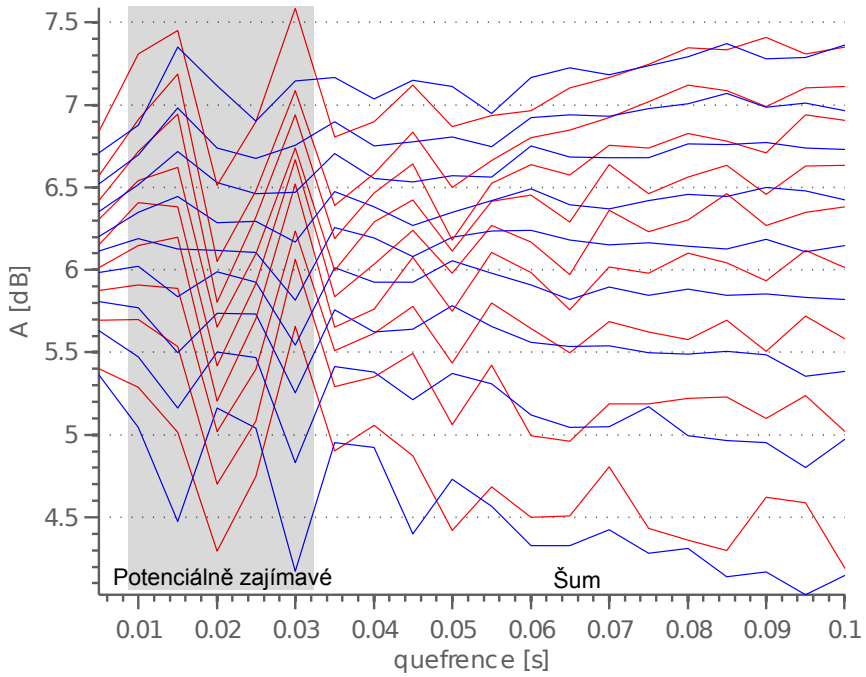




Obrázek 4.7: AUC hodnoty z tabulky 4.2 vykreslené na povrchu hlavy a schematický nákres vpravo pro jednodušší interpretaci.

4.3 Cepstrum

Další testovaný příznakový model bylo cepstrum. Kromě několika prvních prvků odpovídajích nejnižším quefrencím Fourierova spektra se v cepstru nacházel pouze náhodný šum. Což může být výhoda, informace z příznaků může být "zahuštěnější". V grafu 4.8 je vykresleno prvních 20 složek cepstra z 4096, po odečtení průměru obou skupin. Jediná potenciálně zajímavá oblast se nachází kolem quefrence 0.02 s.



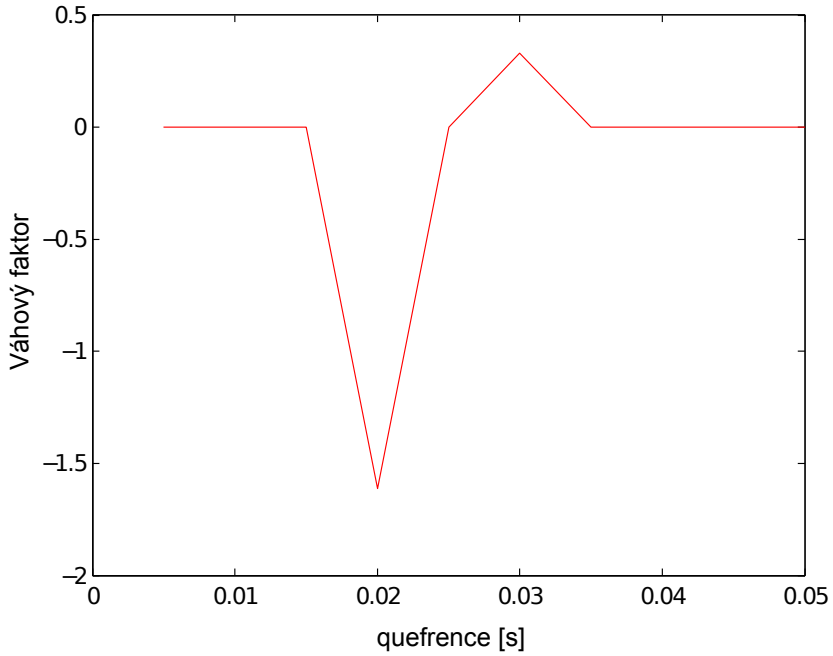
Obrázek 4.8: Hustota pravděpodobnosti několika prvních složek cepstra EEG signálu.

Následná analýza byla provedena stejnou cestou, jako u Fourierova spektra. Pomocí LDA byl vypočten optimální váhový vektor. Tento vektor byl vždy nenulový jen pro quefrenci 0.02 s a 0.03 s, což odpovídá pozorování v grafu 4.8. Pro většinou kanálů s nejvyšším AUC měl téměř identické váhové koeficienty s variací v řádech procent.

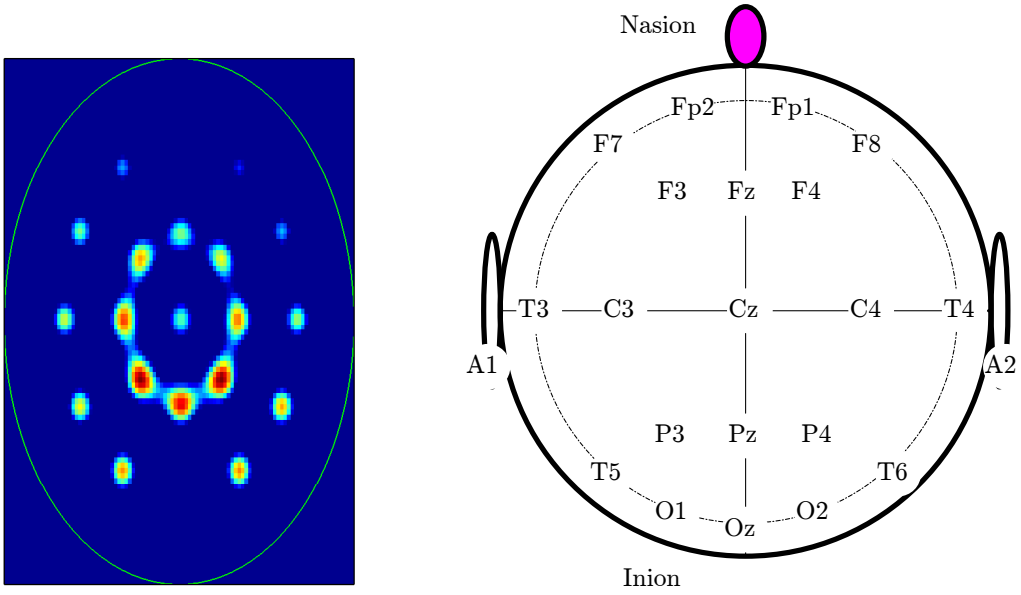
V tabulce 4.3 jsou vypsány konkrétní výsledky získané pro každý kanál zvlášť. A nakonec jsou výsledky stejně jako pro Fourierovo spektrum vykresleny v závislosti na poloze kanálu 4.10.

Tabulka 4.3: Výsledky z L1 regularizované LDA analýzy aplikované na kanál cepstra zvlášť.

kanál	jméno	AUC	p-hodnota	kanál	jméno	AUC	p-hodnota
1	Fp1	0.766	$8.69{\cdot}10^{-19}$	11	C4	0.821	$1.02{\cdot}10^{-31}$
2	Fp2	0.734	$4.76{\cdot}10^{-15}$	12	T4	0.827	$7.27{\cdot}10^{-33}$
3	F7	0.817	$2.26{\cdot}10^{-34}$	13	T5	0.850	0
4	F3	0.813	$4.64{\cdot}10^{-28}$	14	P3	0.877	0
5	Fz	0.762	$2.12{\cdot}10^{-19}$	15	Pz	0.856	0
6	F4	0.788	$6.43{\cdot}10^{-22}$	16	P4	0.882	0
7	F8	0.769	$3.20{\cdot}10^{-21}$	17	T6	0.872	0
8	T3	0.838	$7.34{\cdot}10^{-38}$	18	O1	0.879	0
9	C3	0.835	$4.25{\cdot}10^{-34}$	19	O2	0.882	0
10	Cz	0.791	$3.22{\cdot}10^{-22}$				



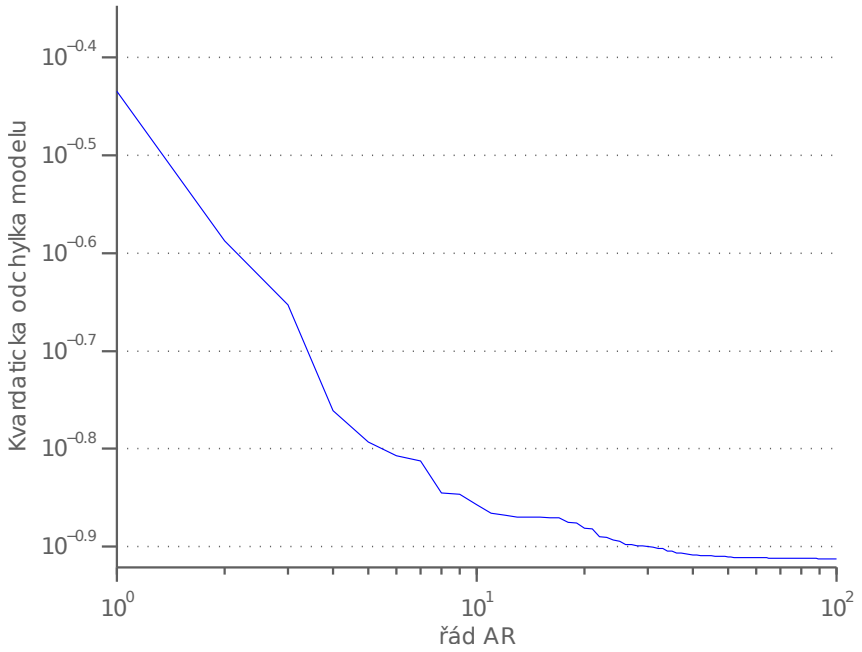
Obrázek 4.9: Váhový vektor získaný z LDA cepstra. Pro kanály 8–19 (ty s nejvyšším AUC) byl až na varianci v řádu procent identický.



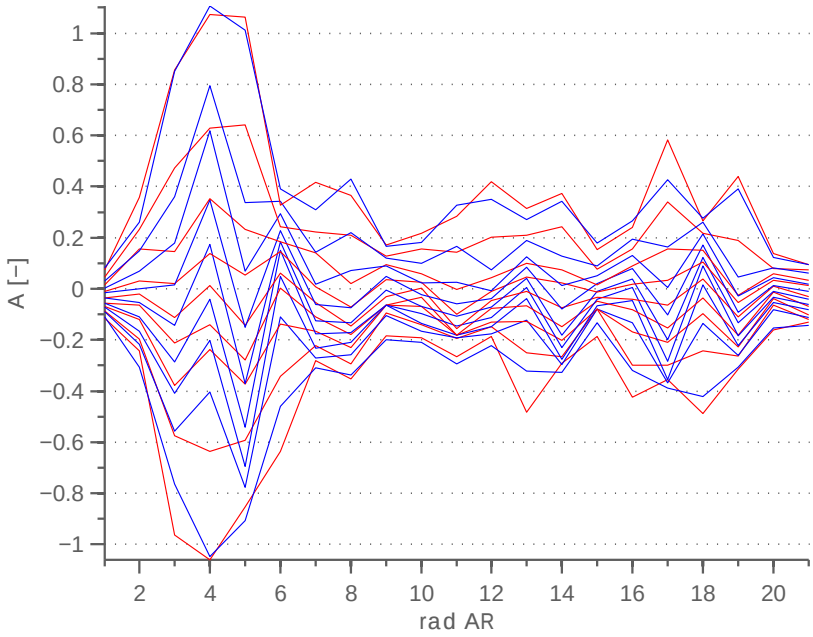
Obrázek 4.10: Prostorové rozdělení hodnot AUC získané z cepstrum analýzy pro jednotlivé kanály.

4.4 Autoregresní model

Posledním testovaným příznakem byl autoregresní model. K nalezení optimálního počtu parametrů bylo vypočteno rezidum mezi predikcí a signálem. Toto reziduum je vykresleno v grafu 4.11. Rychlost poklesu rezidua s rostoucím počtem parametrů velmi rychle klesá, a proto byl počet parametrů zvolen na 20. To odpovídá autokorelačnímu času $10/f_0 = 0.2\,\mathrm{s}$. Abychom mohli identikovat, který z parametrů AR modelu má největší přínos k identifikaci AD pacientů, byla, stejně jako v předchozích případech, vykreslená hustota pravděodobnosti v závislosti pro obě dvě skupiny do grafu 4.12. Napřed byly ovšem opět odečteny průměrné hodnoty AR koeficientů. Jak je z grafu 4.12 vidět, rozdíl mezi skupinami je malý. Pokud tam bude nějaký rozdíl, tak nejspíše ve 4. a 5. koeficientu. Abychom to ověřili, byla vypočtena regularizovaná LDA pro každý kanál zvlášť. Jak je patrné z tabulky 4.4, oddělení má ještě nižší kvalitu, než měla nejjednodušší metoda založená na sledování alfa vln. Navíc váhový vektor vypočtený pomocí LDA má jiný tvar pro každý kanál a i z tohoto důvodu je prediktivní hodnota tohoto testo téměř nulová.



Obrázek 4.11: Odchylka predikce autoregresního modelu od surových dat 1. kanálu EEG signálu v závislosti na počtu parametrů tohoto modelu.



Obrázek 4.12: Pravděpodobnostní rozdělení koeficientů autokorelačního modelu pro zdravé (modří) a nemocné (červení) pacienty.

Tabulka 4.4: Výsledky predikce regularizovanou LDA založené na autoregresních koeficientech EEG signálu.

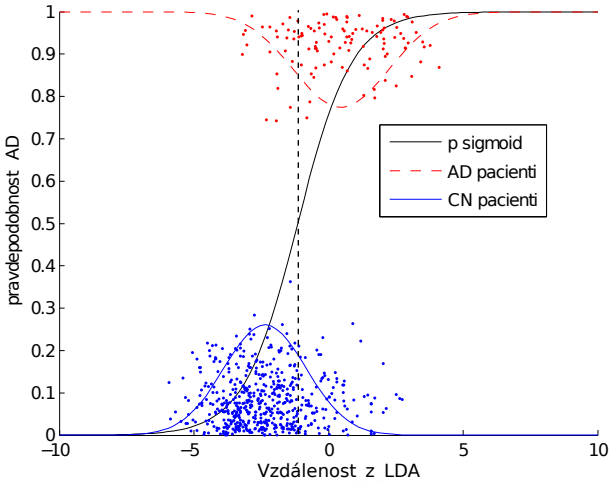
kanál	jméno	AUC	p-hodnota	kanál	jméno	AUC	p-hodnota
1	Fp1	0.656	$1.99{\cdot}10^{-09}$	10	Cz	0.709	$8.58{\cdot}10^{-11}$
2	Fp2	0.673	$1.80{\cdot}10^{-10}$	10	Cz	0.709	$8.58{\cdot}10^{-11}$
3	F7	0.698	$2.86{\cdot}10^{-11}$	12	T4	0.698	$2.00{\cdot}10^{-11}$
4	F3	0.707	$1.60{\cdot}10^{-13}$	13	T5	0.691	$1.20{\cdot}10^{-11}$
5	Fz	0.704	$1.79{\cdot}10^{-11}$	14	P3	0.686	$1.47{\cdot}10^{-10}$
6	F4	0.710	$6.69{\cdot}10^{-13}$	15	Pz	0.709	$4.16{\cdot}10^{-13}$
7	F8	0.658	$4.69{\cdot}10^{-8}$	16	P4	0.696	$2.16{\cdot}10^{-12}$
8	T3	0.615	$1.51{\cdot}10^{-3}$	17	T6	0.745	$6.27{\cdot}10^{-19}$
9	C3	0.686	$2.70{\cdot}10^{-10}$	18	O1	0.720	$1.30{\cdot}10^{-14}$
				19	O2	0.730	$1.37{\cdot}10^{-17}$

4.5 Pravděpodobnostní sigmoid

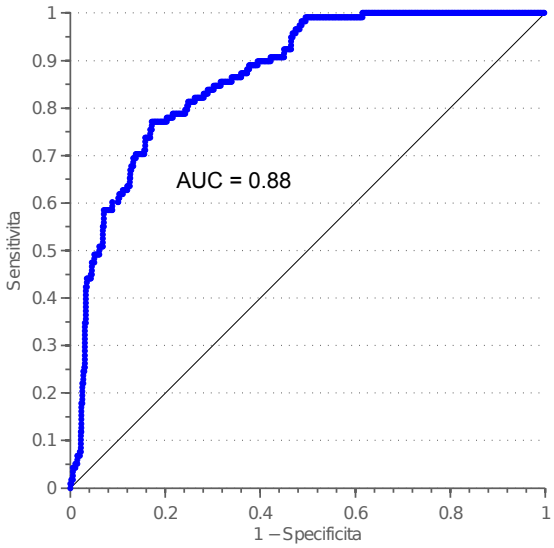
Abychom ke každému pacientovi přiřadili i pravděpodonost s jakou patří do přiřazené skupiny, byl vypočten pravděpodobnostní sigmoid, definovaný v kapitole 3.4. Sigmoid vypočtený pro 16. kanál EEG signálu analyzovaného za pomocí cepstra je vykreslený v Obr. 4.13. Konkrétní předpis tohoto sigmoidu je

$$S(x) = (1 + \exp(-1.38(x + 0.84)))^{-1}.$$

Je zřejmé, že jen malý zlomek bodů je určený s pravděpodobností vyšší než 90% že jsou v AD nebo CN skupině. Proto ani nejlepší nalezený test neumožňuje skutečně spolehlivé oddělení obou skupin.



Obrázek 4.13: Pravděpodobnostní sigmoid vypočtený pro 16. kanál LDA z Cepstra za pomocí váhového vektoru z Obr. 4.9.



Obrázek 4.14: ROC křivka nejlepšího identifikovaného příznaku - 16. kanálu cepstra

Závěr

V rámci této bakalářské práce bylo navrženo, implementováno a otestováno několik příznakových modelů pro analýzu EEG signálu. Navíc všechny tyto modely byly naprogramovány jako jednoduché skripty v MATLABu. A nakonec bylo provedeno pečlivé a sjednocené statistické vyhodnocení všech těchto modelů.

Jako základní model byl zvolen běžně používaný model založený na relativní intenzitě alfa vln. V klidovém stavu se zavřenýma očima, ve kterém se všichni pacienti nacházeli, lze očekávat nárůst amplitudy právě alfa vln. Ale u zdravých pacientů byla pozorována statisticky významně větší intenzita než u nemocných. Statistická významnost daná p-hodnotou MWW testu byla 10^{-37}. Další testované příznakové modely byly: vyhlazené FFT spektrum, cepstrum a autoregresní model. Koeficienty získané těmito modely byly analyzovány pomocí LDA (lineární diskriminantní analýzy), aby byly určeny nejpodstatnější dimenze a byla nalezena optimální lineární kombinace příznaků k dosažení robustního a kvalitního oddělení.

Srovnáme-li tyto čtyři testované příznakové modely, nejlepších výsledků bylo dosaženo a za pomocí FFT spektra a cepstra, zatímco klasifikace pomocí autoregresního modelu zcela ,v porovnání se základní metodou, selhala. Jako nejvhodnější poloha měření pro klasifikaci se u všech metod ukázalo temeno hlavy, kdy byla AUC hodnota obou metod mezi 0.87 a 0.88 a také oběma metodama byl identifikován jako zcela nejlepší kanál č. 16 (P4) ležící na pravé zadní straně temene hlavy. Tato oblast hlavy by měla být zodpovědná za rozpoznávací schopnosti, matematické problémy, neverbální vyjadřování. ROC hodnota tohoto kanálu byla téměř identická, 0.880 pro vyhlazené Fourierovo spektrum a 0.882 pro cepstrum. Přesto lze považovat za lepší volbu cepstrum, neboť, model určený pomocí LDA byl jednodušší a téměř identický pro všechny kanály na temeni hlavy. Proto lze očekávat lepší prediktivní schopnost tohoto modelu. Obě dvě skupiny pacientů byly odděleny touto metodou s úspěšností kolem 80%.

Za největší přínos této práce lze považovat efektivní využití regularizované LDA analýzy pro identifikaci podstatných dimenzí. To spolu s vykreslením pomocí hustot pravděpodobnosti umožňuje získat velmi názornou představu o tom, co se skutečně v signálu odehrává. Navíc vykreslení hodnot ROC přímo na skalpu hlavy umožňuje velmi přehledně určit oblasti mozku vhodné pro analýzu. To je zcela opačný přístup, než byl zvolen v práci A. Hraběte [14], kde byla metodou Monte Carlo nalezena optimální volba příznakového modelu, kanálu a frekvence. Ale i touto cestou byl identifikovaný stejný kanál i podobná frekvence jako pomocí LDA.

4.6 Možnosti pokračování

Z pohledu analýzy signálu a hledání optimálního příznakového modelu by bylo zajímavé stejnou cestou otestovat i další modely. Jako velmi nadějného kandidáta lze považovat komplexní cepstrum, kde je možné využít nejen amplitudu, ale i fázi jednotlivých složek.

Pro samotnou klasifikaci by šlo použít i pokročilejších nástrojů než jen základní LDA analýzy, ale bylo by nezbytné se zabývat rozsáhlým problémem učících se algoritmů, který přesahuje rozsah této práce.

A nakonec pro dosažení skutečně v praxi použitelných výsledků je nezbytné získát lepší datovou sada. Zvláště důležitá je pečlivější klasifikace nemocných pacientů do kategorií podle závažnosti Alzheimerovy choroby. Naopak do testovací sady je nezbytné přidat skupiny pacientů s nemocemi mozku, které postihují běžně starší lidi a mohou mít také vliv na EEG spektrum. Je totiž dost pravděpodobné, že člověk s podezřením na AD může jednou z nich trpět. Po získání dostatečně rozsáhlé a velmi důvěryhodné datové sady je teprve možné pro vyhodnocení zkusit používat pokročilé klasifikační algoritmy.

Seznam použitých zdrojů

[1] Berger, H.: Über das Elektroenkephalogram des Menschen. *Arch. f. Psychiat*, vol. 87, 1927: p. 527–70.

[2] Bishop, C. M.: *Pattern recognition and machine learning (information science and statistics)*. Springer, 2007.

[3] Blank, J.; Řezáčová, K.; Exner, P.; et al.: *Lineární operátory v kvantové fyzice*. Karolinum, 1993.

[4] Bogart, B.; Healy, M.; Tukey, J.: The quefrency analysis of time-series for echoes. In *Proc Symp. Time Series Analysis, Wiley, NY*, 1963, p. 209–243.

[5] Brigham, E.: *The Fast Fourier Transform and its applications*. Prentice Hall, 1988.

[6] Brockwell, P. J.; Dahlhaus, R.; Trindade, A. A.: Modified Burg algorithms for multivariate subset autoregression. *Statistica Sinica*, vol. 15, no. 1, 2005: p. 197–213.

[7] Cooley, J. W.; Tukey, J. W.: An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, vol. 19, no. 90, 1965: p. 297–301.

[8] Dauwels, J.; Srinivasan, K.; Ramasubba Reddy, M.; et al.: Slowing and loss of complexity in Alzheimer's EEG: two sides of the same coin? *International journal of Alzheimer's disease*, vol. 2011, 2011.

[9] Dauwels, J.; Vialatte, F.; Cichocki, A.: Diagnosis of alzheimers disease from eeg signals: Where are we standing? *Current Alzheimer Research*, vol. 7, no. 6, 2010: p. 487–505.

[10] Fawcett, T.: An introduction to ROC analysis. *Pattern recognition letters*, vol. 27, no. 8, 2006: p. 861–874.

[11] Fisher, R. A.: The use of multiple measurements in taxonomic problems. *Annals of eugenics*, vol. 7, no. 2, 1936: p. 179–188.

[12] Gauss, C. F.: Nachlass: Theoria interpolationis methodo nova tractata. *Carl Friedrich Gauss, Werke*, vol. 3, 1866: p. 265–303.

[13] Gerla, V.; Krajča, V.; Lhotská, L.; et al.: Metody zpracování dat z dlouhodobých EEG záznamů. *LÉKAŘ A TECHNIKA*, 2008: str. 10.

[14] Hrabě, A.: *Statistické charakteristiky signálu EEG a jejich využití*. Diplomová práce, CVUT FJFI, 2011.

[15] Jamieson, L. H.: Speech Processing by Computer, Notes. online.

[16] Krbálek, M.: *Rovnice matematické fyziky*. CVUT, 2012.

[17] Kwangmoo, K.; Seung-Jean, K.; Boyd, S.: L1-Regularized Least Squares Problem Solver. online, 2007.

[18] Lehmann, E.: *Elements of large-sample theory*. Springer, 1998.

[19] Liepins, V.: Extended Fourier analysis of signals. *arXiv preprint arXiv:1303.2033*, 2013.

[20] Rao, C. R.: The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10, no. 2, 1948: p. 159–203.

[21] Regnault, M.: *Alzheimerova choroba: Pruvodce pro blizke nemocnych*. Praha: Portál, sro, 2011.

[22] Shannon, C. E.: Communication in the presence of noise. *Proceedings of the IRE*, vol. 37, no. 1, 1949: p. 10–21.

[23] Vapnik, V.: *Statistical learning theory*. Wiley New York, 1998.

[24] Zachová, B. D.: *Biomedicínské využití robustní predikce signálu*. Diplomová práce, CVUT FJFI, 2010.

[25] Zschocke, S.; Hansen, H.-C.: *Klinische Elektroenzephalographie*. Springer DE, 2012.

Přílohy

Appendix A

Obsah CD

Tabulka A.1: Obsah přiloženého CD

Adresář/soubor	Popis
`./BP_Kopecka.pdf`	elektronická verze bakalářské práce
`./MATLAB/`	skripty použité k přípravě a analýze EEG signálu
`./MATLAB/Animace_mozku.m`	vykreslení vybrané frekvence jako animace přímo na skalpu hlavy
`./MATLAB/Autoregrese.m`	Autoregresní model
`./MATLAB/Cepstrum.m`	Cepstrum
`./MATLAB/KorekceEEGSignalu.m`	příprava EEG signálu
`./MATLAB/LDA.m`	Lineární diskriminantní analýza
`./MATLAB/Music.m`	MUSIC algoritmus
`./MATLAB/RelativniVykon.m`	Relativní výkon na intervalech mozkových vln
`./MATLAB/Spectrogram.m`	vypočet spektrogramů pro všechny pacienty a všechny kanály
`./MATLAB/Windowing.m`	vypočtení všech běžných typů windowingu signálu
`./MATLAB/` ostatní	další pomocné rutiny nezbytné pro běh