VYSOKÁ ŠKOLA
CHEMICKO-TECHNOLOGICKÁ
V PRAZE

Fakulta chemicko-inženýrská

Ústav počítačové a řídicí techniky

POUŽITÍ ADAPTIVNÍCH SYSTÉMŮ PŘI ANALÝZE DAT

TEZE DISERTAČNÍ PRÁCE

AUTOR	**Ing. Jan Vrba**
ŠKOLITEL	**doc. Ing. Jan Mareš, Ph.D.**
ŠKOLITEL SPECIALISTA	**doc. Ing. Pavel Hrnčiřík, Ph.D.**
STUDIJNÍ PROGRAM	Chemické a procesní inženýrství (čtyřleté)
STUDIJNÍ OBOR	Technická kybernetika
ROK	**2020**

Souhrn

Dizertační práce se zabývá použitím adaptivních systémů v oblasti detekce novosti. Tento přístup v oblasti detekce novosti v datech se stal v posledních letech slibným směrem výzkumu. V rámci této práce je navržen nový algoritmus pojmenovaný jako Extreme Seeking Entropy. Tento algoritmus je založen na vyhodnocování přírůstku adaptivních parametrů systémů pomocí zobecněného Paretova rozdělení. Navržený algoritmus byl otestován na celé řadě typů syntetických dat, která reprezentují různé druhy novosti. Pro detekci změny trendu a skokové změny generátoru signálu pak bylo provedeno i vyhodnocení úspěšnosti detekce. Dále byla provedena experimentální studie zabývající se časovou náročností výpočtu algoritmu, vyhodnocena ROC (Receiver Operating Characteristic) křivka pro detekci změny trendu, a provedena studie detekce epilepsie v záznamu EEG myši.

Klíčová slova

adaptivní systémy, detekce novosti, časové řady, extreme seeking entropy

Summary

The dissertation deals with the use of adaptive systems in the field of novelty detection. The use of adaptive systems for detecting novelty in data has become a promising direction of research in recent years. In this work, a new algorithm was designed using adaptive systems, called Extreme Seeking Entropy. This algorithm is based on evaluating the increment of adaptive parameters of systems using a generalized Pareto distribution. The proposed algorithm has been tested on a number of types of synthetic data that represent different types of novelty. To detect a change in the trend and a step-change in the signal generator, an evaluation of the detection success was performed. Furthermore, an experimental study was performed dealing with the time consumption of the algorithm, the ROC curve was evaluated to detect a change in the trend, and a study was performed to detect epilepsy in the EEG mouse record.

Keywords

adaptive systems, novelty detection, time series, extreme seeking entropy

Obsah

1 Úvod

Tato disertační práce je věnována problematice využití adaptivních systémů při analýze dat. Vzhledem k exponenciálnímu celosvětovému nárůstu objemu dat [1] a ke zvyšování jejich variability roste i potřeba tato data analyzovat, kategorizovat a vytěžovat. Analýzou dat rozumíme proces, kdy z nezpracovaných naměřených dat získáme nějakou interpretovatelnou informaci, s kterou pak lze dále pracovat. Jedna z možných důležitých interpretací nově získaných dat je, zda-li se nově získaná data nějakým zásadním způsobem odlišují od předchozích dat. Této problematice se věnuje obor detekce novosti, neboli anomálií, který spadá do oblasti vytěžování dat a strojového učení. Úspěšná detekce novosti pak může být využita k vícero účelům. Například k diagnostice sledovaného procesu, ke změně struktury nebo parametrů adaptivního modelu za účelem zlepšení predikce, z konkrétních aplikací pak k odhalení neoprávněného vniknutí do sítě nebo zneužití dat, v lékařství se detekce novosti používá k diagnostickým účelům, z průmyslových aplikací pak k detekci poruchy a monitoringu stavu strojů, senzorů, ve zpracování textových dat k detekci nových témat, originálnosti textů atd. Spektrum využití je velice široké.

V oblasti detekce novosti byla v posledních desetiletích intenzivního vývoje navrhnuta celá řada algoritmů. Vzhledem k rostoucímu výpočetnímu výkonu a rozmanitosti analyzovaných dat rostla i potřeba nových algoritmů. Nové algoritmy typicky předčily ostatní v rámci jedné aplikace, respektive v rámci jednoho typu dat. Doposud se však nepodařilo vytvořit algoritmus, který by ve všech (nebo alespoň ve významné části) oblastech použití předčil již publikované algoritmy. I proto vznikají v oblasti detekce novosti neustále nové přístupy, které navíc umožňují analyzovat nové typy dat.

1.1 Cíle dizertační práce

Cíle předkládané dizertační práce jsou:

1) **Návrh algoritmu pro detekci novosti v datech s využitím adaptivních systémů**

   Navrhovaný algoritmus pro detekci novosti bude využívat adaptivní systémy a bude mít interpretovatelný výstup. Jeho použití by mělo být možné v kombinaci libovolným

adaptivním systémem. Algoritmus bude implementován v jazyce Python, který je v současné době jedním z nejrozšířenějších programovacích jazyků.

2) **Otestování navrženého algoritmu na syntetických datech**

Navržený algoritmus bude otestován na syntetických datech, která budou simulovat různé typy novosti a budou obsahovat šum. Mezi testovanými daty budou i data časových řad obsahujících trend. Výsledky algoritmu budou porovnány se obdobnými soudobými metodami využívajícími adaptivní systémy, konkrétně s algoritmem Learning Entropy a Error and Learning Based Novelty Detection.

3) **Provedení případových studií na reálných datech z oblasti biomedicíny a chemie**

Pro otestování navrženého algoritmu je zásadní provedení případových studiích na reálných datech.

4) **Vyhodnocení kvality navrženého algoritmu z pohledu úspěšnosti detekce novosti**

Pro vyhodnocení kvality navrženého algoritmu budou určeny dosažené přesnosti detekce v různých scénářích. Dále budou pro vybrané scénáře určeny ROC křivky a vyhodnocena plocha pod nimi. Vyhodnocení bude provedeno pro scénáře s různými poměry signál-šum. Výsledky budou porovnány s výsledky algoritmů Learning Entropy a Error and Learning Based Novelty Detection.

2 Algoritmus Extreme Seeking Entropy

V této kapitole je představen algoritmus pro detekci novosti nazvaný Extreme Seeking Entropy (ESE), který tvoří hlavní teoretický výsledek předkládané dizertační práce (publikováno v [V1]). Navržený algoritmus, vychází z předpokladu, že novost v datech se projeví neobvykle velkými přírůstky vah adaptivního filtru, který danou řadu dat modeluje.

Nejprve uvažujme myšlenkový experiment, ve kterém máme dokonale nastavený adaptivní filtr, jehož chyba predikce e je nulová pro všechny vstupní hodnoty. Potom přírůstky adaptivních vah tohoto filtru budou nulové. V případě, že dojde k nějaké změně v generátoru dat pro tento filtr, začne se filtr opět adaptovat což vyústí v nenulové změny adaptivních vah, které budou reflektovat novost způsobenou změnou vlastností daného generátoru.

V publikacích [2, 3], kde autoři představují algoritmus Learning Entropy, je zmíněna obecná míra snahy adaptivního filtru o adaptaci, L, která slouží k vyhodnocení neobvykle velkých přírůstků adaptivních vah a je definovaná jako

$$L(k) = A(f(\Delta \mathbf{w}))) \tag{2.1}$$

kde A je obecně nějaká agregační funkce a funkce f je funkce která nějakým způsobem kvantifikuje odchylku v adaptaci adaptivních parametrů filtru.

Nejprve uvažujme, že hodnota funkce f, která slouží k vyhodnocení neobvykle velkých přírůstků by měla mít neobvykle velkou (nebo neobvykle malou) hodnotu v okamžiku, kdy přírůstky adaptivních vah filtru $\Delta \mathbf{w}$ jsou neobvykle velké. Dalším požadavkem je, aby tato funkce zohledňovala i nějakou historii těchto přírůstků. Přirozeně se tedy nabízí nějaká vhodná distribuční funkce f_{cdf}. S ohledem na požadavek, že neobvykle velké přírůstky vah by měli být reflektovány neobvykle velkou hodnotou míry snahy o adaptaci, nabízí se jako vhodná funkce A, kterou uvažujeme ve tvaru

$$A(f(\Delta|\mathbf{w}|(k)) = -\log \prod_{i=1}^{n} (1 - f_{cdf_i}(|\Delta w_i(k)|)). \tag{2.2}$$

Uvedená agregační funkce A má pro neobvykle velké přírůstky vah neobvykle velkou kladnou hodnotu, pokud jsou hodnoty f_{cdf} blízké 1. Člen $1-f_{cdf}$ je vlastně komplementární distribuční funkcí (funkce přežití, spolehlivostní funkce). Výhodou uvedeného přístupu je absence potřeby nastavovat několik prahů pro detekci.

Jak již bylo uvedeno, cílem je tedy vyhodnotit neobvykle velké přírůstky vah adaptivního systém. Nejprve je nutné určit nějaký práh z, podle kterého můžeme přírůstky adaptivních vah filtru rozdělit do dvou množin. Množinu, která bude obsahovat přírůstky menší než je zvolený práh z označíme L. Přírůstek, který je větší nebo roven hodnotě prahu z označíme H. Uvažujme, že obě množiny existují pro každou adaptivní váhu, potom pro i-tou adaptivní váhu zvolíme práh z_i tak, že velikosti přírůstku této váhy náleží do jedné ze dvou množin, tak, že:

$$\forall|\Delta w_i| < z_i \in L_i \tag{2.3}$$

$$\forall|\Delta w_i| \geq z_i \in H_i \tag{2.4}$$

Vzhledem k výše zmíněnému předpokladu o velikosti změn vah adaptivního filtru a novosti v datech, uvažujme, že přírůstky náležející množině L_i pravděpodobně neobsahují informaci o novosti během adaptace, a proto nebudou vyhodnocovány. Množina H_i by měla obsahovat přírůstky adaptivních vah filtru, u kterých lze očekávat, že mohou nést nějakou informaci o novosti v datech.

Nyní můžeme zavést novou míru, kterou nazvěme Extreme Seeking Entropy (ESE), definovanou jako

$$ESE(|\Delta \mathbf{w}(k)|) = -\log \prod_{i=1}^{n_w} (1 - f_{cdf_i}(|\Delta w_i(k)|)) \tag{2.5}$$

kde

$$f_{cdf_i}(|\Delta w_i(k)|) = \begin{cases} 0, \, |\Delta w_i(k)| \in L_i \\ F_{(\xi_i,\mu_i,\sigma_i)}(|\Delta w_i(k)|), \, |\Delta w_i(k)| \in H_i. \end{cases} \tag{2.6}$$

a funkce $F_{(\xi_i,\mu_i,\sigma_i)}$ je distribuční funkce zobecněného Paretova rozdělení (GPD). Přírůstky adaptivních vah, které jsou menší než hodnota prahu z získaného metodou Peak-Over-Threshold, nezmění hodnotu ESE. Přírůstky vah, které mají velice malou pravděpodobnost a spíše nesou nějakou informaci o novosti v datech, způsobí velký nárůst hodnoty ESE. Zásadním aspektem navrhovaného algoritmu je, že mohou být vyhodnocovány buď všechny získané přírůstky vah, anebo lze vyhodnocovat pouze nejnovějších n_s vzorků a na ně aplikovat metodu POT. Novost potom můžeme vnímat v kontextu těchto n_s vzorků.

Navrhovaný algoritmus pro detekci novosti v datech je popsán následujícím pseudokódem.

Algoritmus 1 Extreme Seeking Entropy

1: nastavení n_s a výběr metody POT
2: počáteční nastavení parametrů ξ_i, μ_i, σ_i GPD pro každý adaptivní parametr
3: **for** vzorek $y(k)$ **do**
4: výpočet změny adaptivních parametrů filtru $\Delta\mathbf{w}(k)$
5: aplikace metody POT
6: **if** $|\Delta w_i|(k) \in H_i$ **then**
7: výpočet parametrů ξ_i, μ_i, σ_i pro příslušné GPD
8: **end if**
9: výpočet hodnoty ESE podle rovnice 2.5
10: **end for**

Výhodou navrhovaného algoritmu je, že jediným volitelným parametrem je parametr n_s, který ale v případě potřeby můžeme vynechat a zpracovávat všechny historicky dostupné přírůstky adaptivních parametrů. Dále je nutné zvolit vhodnou metodu POT.

Určitou limitací uvedeného algoritmu je, že pro získání první hodnoty ESE potřebujeme nějakou apriorní informaci o parametrech GPD rozdělení. Konkrétněji, pro každou i-tou adaptivní váhu musíme mít k dispozici odhad parametrů ξ_i, μ_i, σ_i. Vzhledem k tomu, že adaptivní filtr má n_w různých adaptivních vah, potřebujeme získat odhad $3 \cdot n_w$ parametrů aby začal algoritmus ESE poskytovat první hodnoty. Pokud není k dispozici žádná apriorní informace o hodnotách těchto parametrů, potřebujeme získat alespoň n_s vzorků, na jejichž základě můžeme získat první hodnotu ESE. Dalším úskalím je proměnlivost parametrů nebo typu rozdělení pravděpodobnosti přírůstků adaptivních vah v čase.

3 Výsledky detekce novosti algoritmem Extreme Seeking Entropy

V dalším textu jsou shrnuty výsledky algoritmu Extreme Seeking Entropy, kterých bylo v rámci práce na dizertační práci dosaženo. Výsledky použití ESE pro detekci pertubace v chaotické časové řadě Mackey-Glass je uvedena v podkapitole 3.1. Dále jsou uvedeny výsledky pro: detekci změny rozptylu šumu v náhodném datovém toku (podkapitola 3.2), detekci skokové změny parametrů generátoru signálu (podkapitola 3.3), detekci náhle absence šumu (podkapitola 3.4), detekci změny trendu (podkapitola 3.5) a detekci epilepsie v záznamu myšího EEG (viz podkapitola 3.6). Dále jsou uvedeny výsledky vyhodnocení úspěšnosti detekce skokové změny parametrů generátoru signálu (podkapitola 3.7) a úspěšnosti detekce skokové změny trendu (podkapitola 3.8). V posledních dvou kapitolách je potom vyhodnocení ROC křivky (receiver operating characteristic) pro detekci změny trendu (podkapitola 3.9) a experimentální vyhodnocení výpočetní náročnosti metod odhadu parametrů zobecněného Paretova rozdělení (podkapitola 3.10).

Výsledky v podkapitolách 3.1-3.8 byly publikovány v [V1]. Výsledky v podkapitolách 3.9 a 3.10 pak v publikacích [V2],[V3].

3.1 Chaotická časová řada Mackey-Glass a detekce pertubace

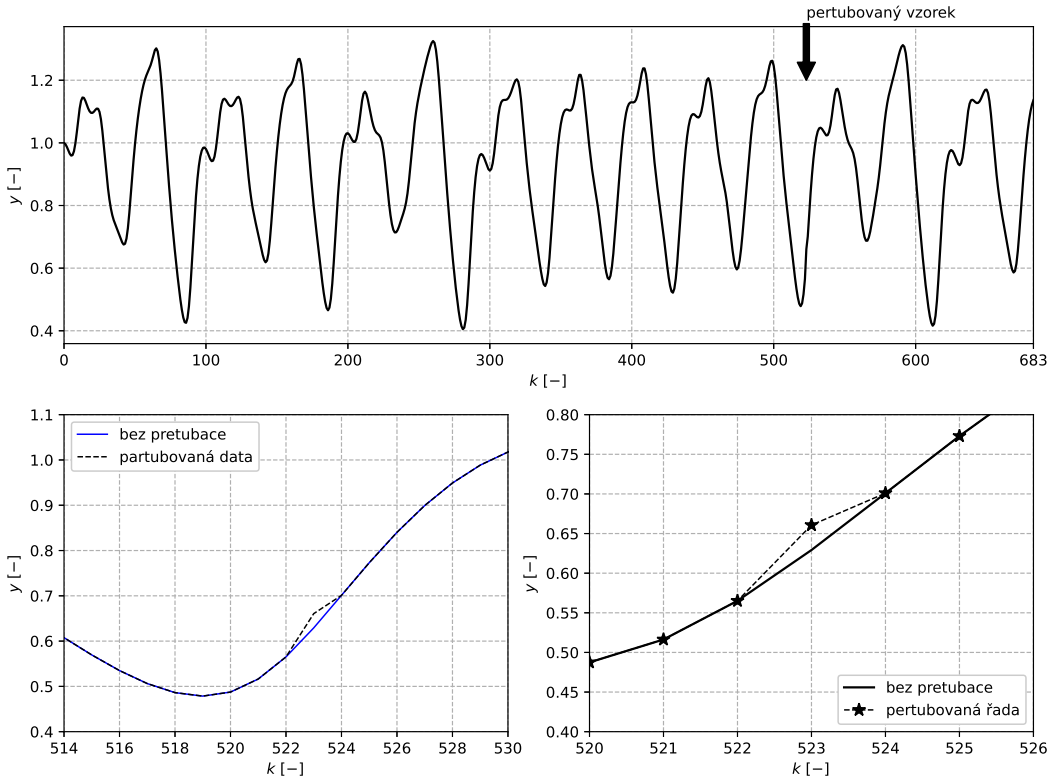
Tento experiment byl proveden pro porovnání s výsledky uvedenými v publikaci [2], která je první publikací o algoritmu Learning Entropy. Experiment spočívá v detekci pertubovaného vzorku v chaotické časové řadě, která je výsledkem řešení Mackey-Glassovy rovnice [7].

$$\frac{dy(t)}{dt} = \beta \cdot \frac{y(t-\tau)}{1+y^{\alpha}(t-\tau)} - \gamma y(t) \tag{3.1}$$

přičemž parametry $\alpha = 10$, $\beta = 0.2$, $\gamma = 0.1$ and $\tau = 17$ byly vybrány tak, aby řešením této rovnice byla chaotická časová řada. Celkem bylo vygenerováno 701 vzorků. Data v diskrétní časový okamžiku $k = 523$ pak byly pertubovány podle následujícího předpisu.

$$y(523) = y(523) + 0.05 \cdot y(523) \tag{3.2}$$

Výsledná časová řada a detail pertubace je znázorněn na obrázku 3.1. Výstup adaptivního filtru a chyba predikce jsou znázorněny na obrázku 3.2. Výsledky metod detekce novosti jsou zobrazeny na obrázku 3.3. Z obrázku je patrné, že globální maximum průběhu ESE odpovídá pertubovaným datům. Globální maximum metod $ELBND$ a LE odpovídá vzorku u něhož byla největší chyba predikce.



Obrázek 3.1: Horní graf zobrazuje celou datovou řadu. Spodní grafy zobrazují detail pertubovaného vzorku v diskrétní časový okamžik $k = 523$.

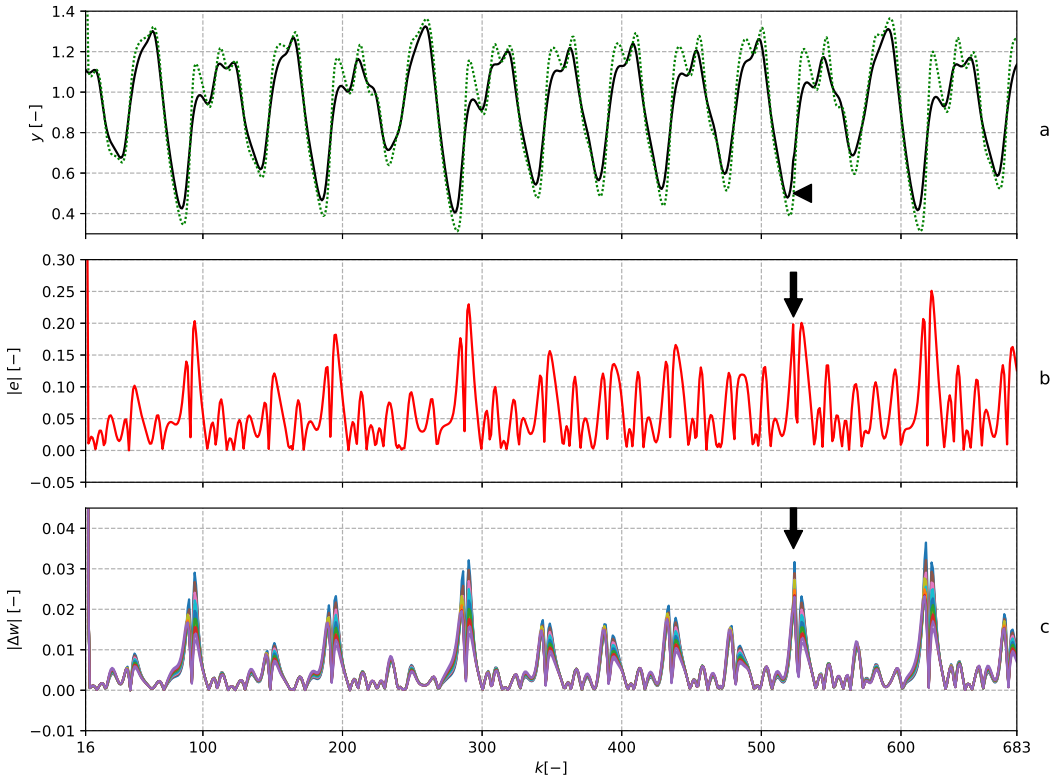
3.2 Detekce změny rozptylu šumu v náhodném datovém toku

Tato případová studie je navržená na základě problému, který se vyskytuje v použití hybridních navigačních systémů využívajících GPS (Global Positioning System) senzory pro navigaci výpočtem [8]. Smyslem experimentu je demonstrovat možnost využití algoritmu ESE pro detekci změn rozptylu šumu v náhodných datech.

Uvažujme dva vstupy $x_1(k)$ a $x_2(k)$ a výstup generátoru signálu $y(k)$ takový, že

$$y(k) = x_1(k) + x_2(k) + x_1(k) \cdot x_2(k) + v(k) \tag{3.3}$$

kde člen $v(k)$ reprezentuje aditivní Gaussovský šum který je přidán k výstupu generátoru $y(k)$.

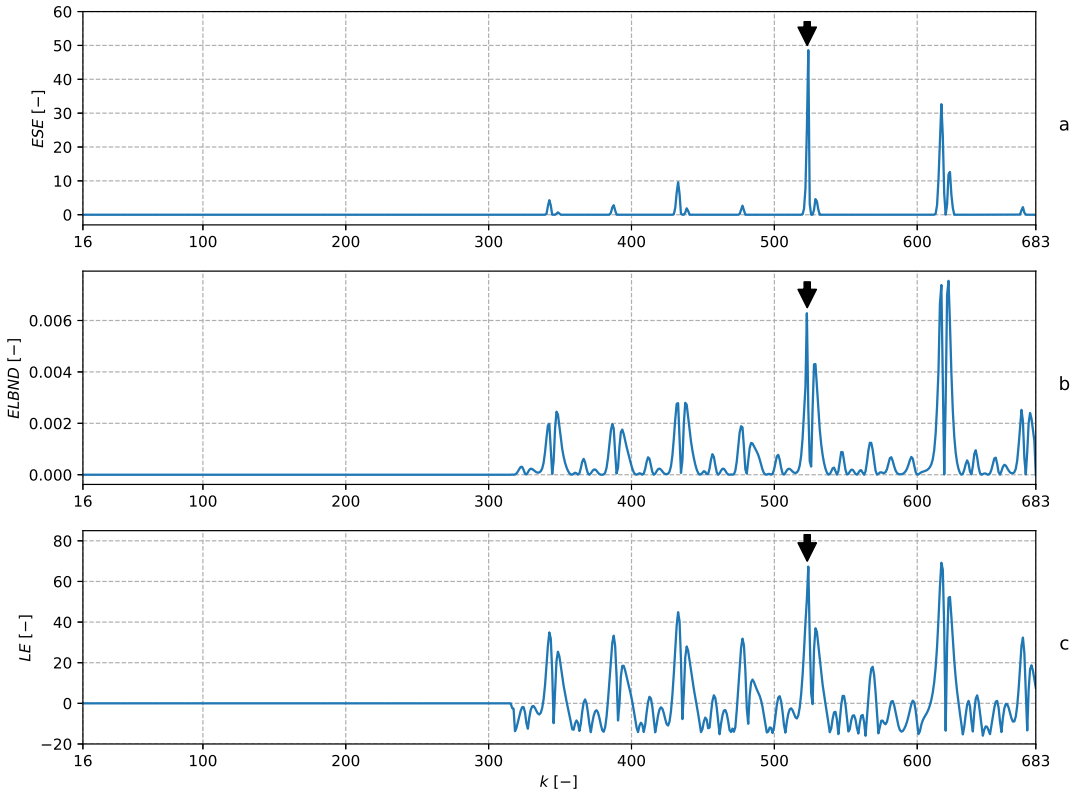


Obrázek 3.2: Graf (a) zobrazuje datovou řadu s pertubací (černá plná čára) a výstup adaptivního filtru (tečkovaná zelená čára). Pertubovaný vzorek je označen černou šipkou. Graf (b) zobrazuje velikost chyby predikce e (resp. její absolutní hodnotu). Na grafu (c) jsou znázorněny přírůstky adaptivních vah filtru (resp. absolutní hodnotu těchto přírůstků).

Přidaný šum má nulovou střední hodnotu a směrodatnou odchylku $\sigma_n = 0.1$, takže $v(k) \sim N(0, 1)$. Hodnoty vstupů jsou v každém diskrétním časovém okamžiku vybrány náhodně z rovnoměrného rozdělení na intervalu $\langle 0, \rangle$. V diskrétním časovém okamžiku $k = 500$ dojde ke změně směrodatné odchylky šumu na hodnota $\sigma_n = 0.2$, $v(k) \sim N(0, 0.2)$. Adaptivní filtr v tomto experimentu byl QNU ve tvaru

$$\hat{y}(k) = w_1 \cdot x_1(k) + w_2 \cdot x_2(k) + w_3 \cdot x_1(k) \cdot x_2(k) \tag{3.4}$$

tak, že jeho struktura odpovídá struktuře generátoru signálu. Adaptivní parametry filtru byly adaptovány algoritmem GNGD. Výsledky experimentu jsou zobrazeny na obrázku 3.4. Apriorní hodnoty parametrů GPD byly stanoveny na základě 500 vzorků, které nejsou v následujícím obrázku 3.4 zobrazeny. Globální maximum ESE odpovídá změně směrodatné odchylky šumu σ_n. Detekce pomocí algoritmů LE a ELBND je o několik vzorků opožděná.



Obrázek 3.3: Graf (a) zobrazuje hodnotu ESE. Prvních 300 vzorků je hodnota ESE nulová, protože délka okna pro vyhodnocování novosti $n_s = 300$. Graf (b) zobrazuje výsledky algoritmu ELBND (Error and Learning Based Novelty Detection [4–6]). Prvních 300 výsledků ELBND je pro názornost vynecháno. Graf (c) zobrazuje výsledky algoritmu LE.

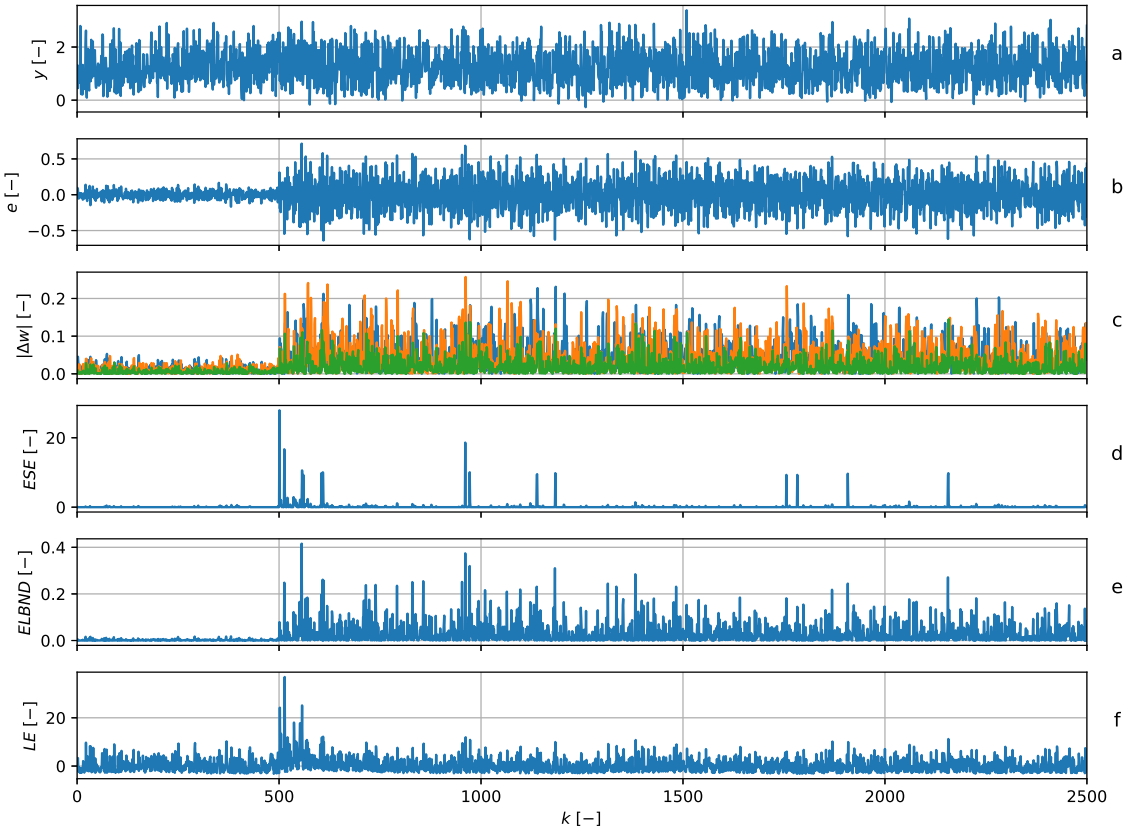
3.3 Detekce skokové změny parametrů generátoru signálu

Tato případová studie je motivována problémem, který vzniká při sledování vícero náhodných datových toků [9] u kterých se kontroluje, zda nedošlo ke změně vlastností jejich generátoru. Uvažujme opět dva vstupy $x_1(k)$, $x_2(k)$ a výstup generátoru signálu ve tvaru

$$y(k) = x_1(k) + x_2(k) + x_1(k) \cdot x_2(k) + v(k) \tag{3.5}$$

kde člen $v(k)$ reprezentuje gaussovský aditivní šum s nulovou střední hodnotou a směrodatnou odchylkou $\sigma_n = 0.1$, $v \sim N(0, 0.1)$. Hodnoty vstupů $x_1(k)$ a $x_2(k)$ jsou v každém diskrétním časovém okamžiku náhodně vybrány z rovnoměrného rozdělení, $x_1(k) \sim U(0, 1)$ resp. $x_2(k) \sim U(0, 1)$. V časový diskrétní okamžik $k = 500$ dojde ke změně parametrů generátoru, a výstup generátoru přejde do tvaru

$$y(k) = 0.4 \cdot x_1(k) + 1.6x_2(k) + 0.99x_1(k) \cdot x_2(k) + v(k). \tag{3.6}$$

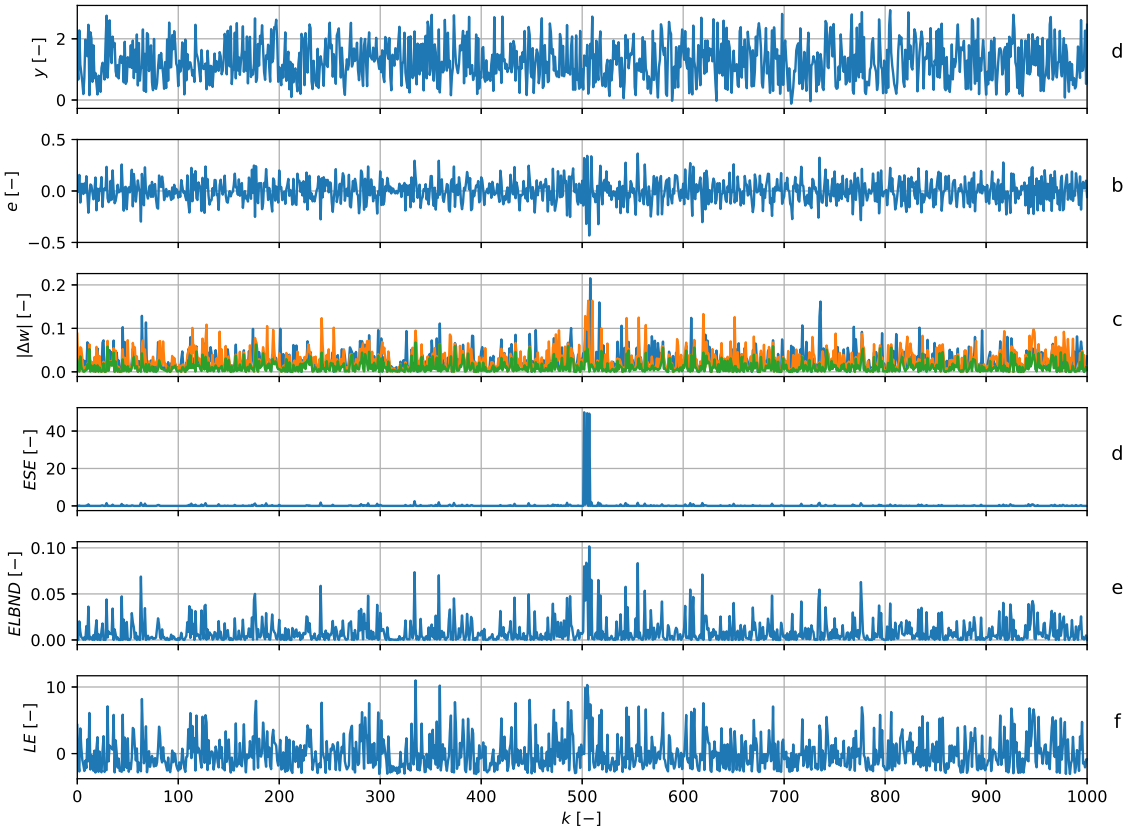


Obrázek 3.4: Detekce změny rozpylu šumu. Na grafu (a) jsou zobrazeny data z generátoru (modrá) a výstup z adaptivního filtru (zelená). Na grafu (b) je vynesena chyba filtru $e(k)$. Graf (c) zobrazuje velikosti přírůstků adaptivních parametrů filtru. Na grafu (d) jsou zobrazeny hodnoty ESE. V diskrétní časový okamžik $k = 500$ je patrný značný nárůst v ESE, který reflektuje změnu směrodatné odchylky aditivního šumu. Na dalších grafech (e) a (f) jsou zobrazeny výsledné hodnoty algoritmů ELBND a LE.

Jako adaptivní filtr byl v tomto případě zvolen QNU, jehož struktura odpovídá generátoru dat. Výstup tohoto filtru je tedy

$$\hat{y}(k) = w_1 \cdot x_1(k) + w_2 \cdot x_2(k) + w_3 \cdot x_1(k) \cdot x_2(k) \tag{3.7}$$

přičemž adaptivní parametry uvedeného filtru jsou adaptovány algoritmem GNGD. Apriorní hodnota parametrů GPD a dat pro LE byla získána použitím 500 vzorků dat (vygenerovaných podle rovnice 3.6). Výsledky experimentu jsou znázorněny na obrázku 3.5. Z obrázku je patrné, že pomocí algoritmu ESE se podařilo detekovat skokovou změnu parametrů, čemuž odpovídá výrazné globální maximum v ESE. Globální maximum v LE neodpovídá skokové změně parametrů generátoru a detekce pomocí algoritmu ELBND je opožděná.



Obrázek 3.5: Detekce skokové změny generátoru signálu.Na grafu (a) je zobrazena původní časová řada (modrá). Graf (b) zobrazuje chybu filtru e. Na grafu (c) jsou zobrazeny velikosti přírůstků adaptivních vah filtru. Na grafu (d) jsou pak výsledky algoritmu ESE, přičemž k skokové změně parametrů generátoru signálu došlo v diskrétní časový okamžik $k = 500$. Je tedy vidět globální maximum v ESE odpovídající úspěšné detekci. Na grafech (e) a (f) jsou pak výsledky metod ELBND a LE. Detekci algoritmem ELBND lze považovat za úspěšnou. Detekce algoritmem LE byla neúspěšná. Globální maximum LE je v diskrétním časovém okamžiku $k = 338$, který neodpovídá skokové změně parametrů generátoru signálu.

3.4 Detekce náhlé absence šumu

V této kapitole je ukázáno, že lehce modifikovaný algoritmus ESE může být využit také k detekci neobvykle malých změn parametrů adaptivního filtru. Oproti standardní variantě ESE budeme vyhodnocovat neobvykle malé přírůstky vah adaptivního filtru. Takže jediná změna v algoritmu je, že metodou POT budeme vybírat pouze nejmenší změny adaptivních vah a budeme odhadovat parametry GPD z takto vybraných hodnot.

Uvažujme dva vstupy $x_1(k)$ a $x_2(k)$ jejichž hodnoty jsou v každém diskrétním časovém okamžiku k vybrány z rovnoměrného rozdělení, takže $x_1(k) \sim U(0,1)$ a $x_2(k) \sim U(0,1)$.

Výstup generátoru dat $y(k)$ je definován jako

$$y(k) = x_1(k) + x_2(k) + x_1(k) \cdot x_2(k) + v(k) \tag{3.8}$$

kde člen $v(k)$ reprezentuje aditivní gaussovský šum s nulovou střední hodnotou a směrodatnou odchylkou $\sigma_n = 0.1$. V diskrétním časovém okamžiku dojde k odstranění aditivního šumu a výstup generátoru signálu přejde do tvaru

$$y(k) = x_1(k) + x_2(k) + x_1(k) \cdot x_2(k). \tag{3.9}$$

který platí pro všechna $k \geq 500$.

Jako adaptivní filtr byl zvolen QNU, jehož výstup je definován

$$\hat{y}(k) = w_1 \cdot x_1(k) + w_2 \cdot x_2(k) + w_3 \cdot x_1(k) \cdot x_2(k) \tag{3.10}$$

takže jeho struktura odpovídá generátoru signálu. Parametry toho filtru jsou adaptovány algoritmem GNGD. Na obrázku 3.6 jsou zobrazeny výsledky experimentu. Maximum v ESE odpovídá detekci vymizení šumu z generátoru signálu. Výsledky metod ELBND a LE jsou uvedeny pouze pro ilustraci.

3.5 Detekce změny trendu

Cílem této kapitoly je demonstrovat použití algoritmu ESE při detekci změny trendu, což je úloha, která se často vyskytuje v oblasti detekce poruch a diagnostice [10]. Uvažujme opět dva vstupy $x_1(k) \sim U(0,1)$ a $x_2(k) \sim U(0,1)$ a výstup generátoru dat $y(k)$ takový, že

$$y(k) = x_1(k) + x_2(k) + 0.01 \cdot k + v(k) \tag{3.11}$$

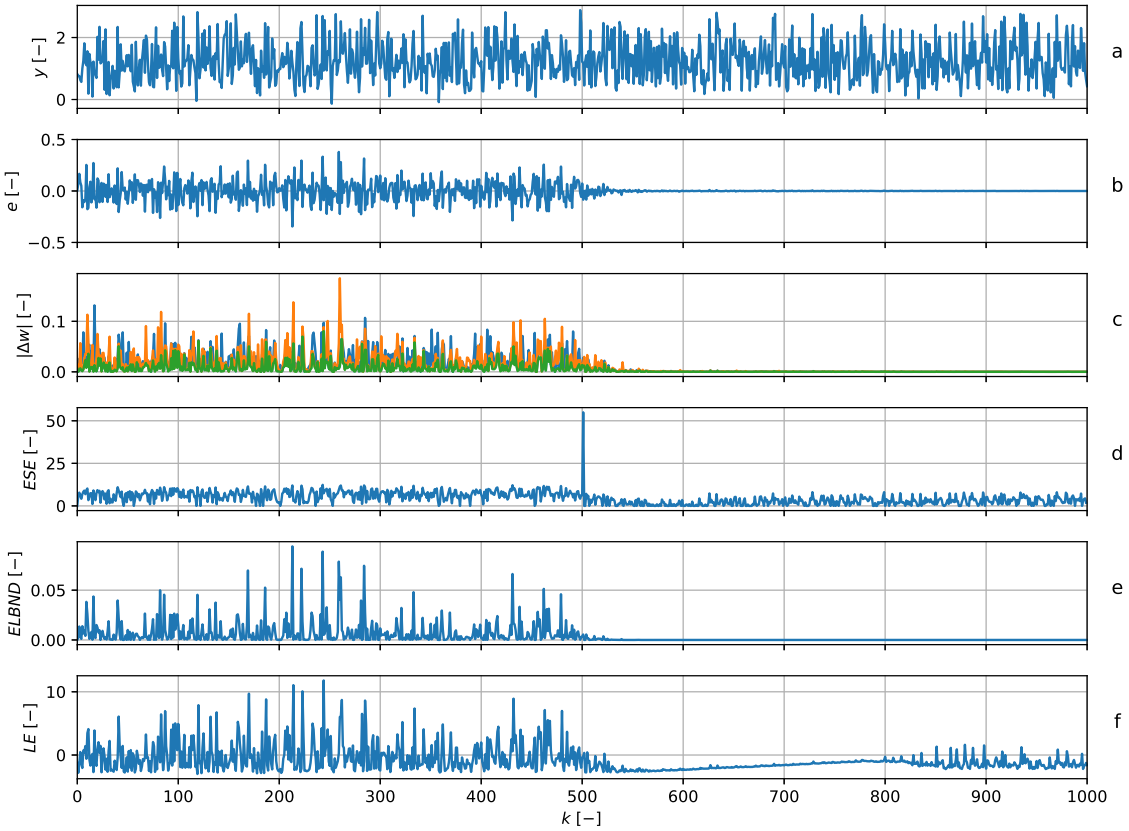
kde člen $v(k)$ reprezentuje aditivní gaussovský šum s nulovou střední hodnotou a směrodatnou odchylkou $\sigma_n = 0.1$. V diskrétním časovém okamžiku $k = 500$ nastane změna trendu. Výstup generátoru signálu se změní, tak, že

$$y(k) = x_1(k) + x_2(k) + 0.0105 \cdot k + v(k), \tag{3.12}$$

pro $k \geq 500$.

Pro zpravování signálu byl použit filtr typu LNU s třemi vstupy, takže výstup uvedeného filtru je ve tvaru

$$\hat{y}(k) = w_1 \cdot x_1(k) + w_2 \cdot x_2(k) + w_3 \tag{3.13}$$



Obrázek 3.6: Detekce vymizení šumu ze signálu. Na grafu (a) je zobrazena původní časová řada (modrá). Graf (b) zobrazuje chybu filtru e. Na grafu (c) jsou zobrazeny velikosti přírůstků adaptivních vah filtru. Na grafu (d) jsou pak výsledky modifikovaného algoritmu ESE, přičemž k odstranění šumu ze signálu došlo v diskrétní časový okamžik $k = 500$. Je tedy vidět globální maximum v ESE odpovídající úspěšné detekci vymizení šumu ze signálu. Na grafech (e) a (f) jsou pak výsledky získané pomocí metod ELBND a LE.

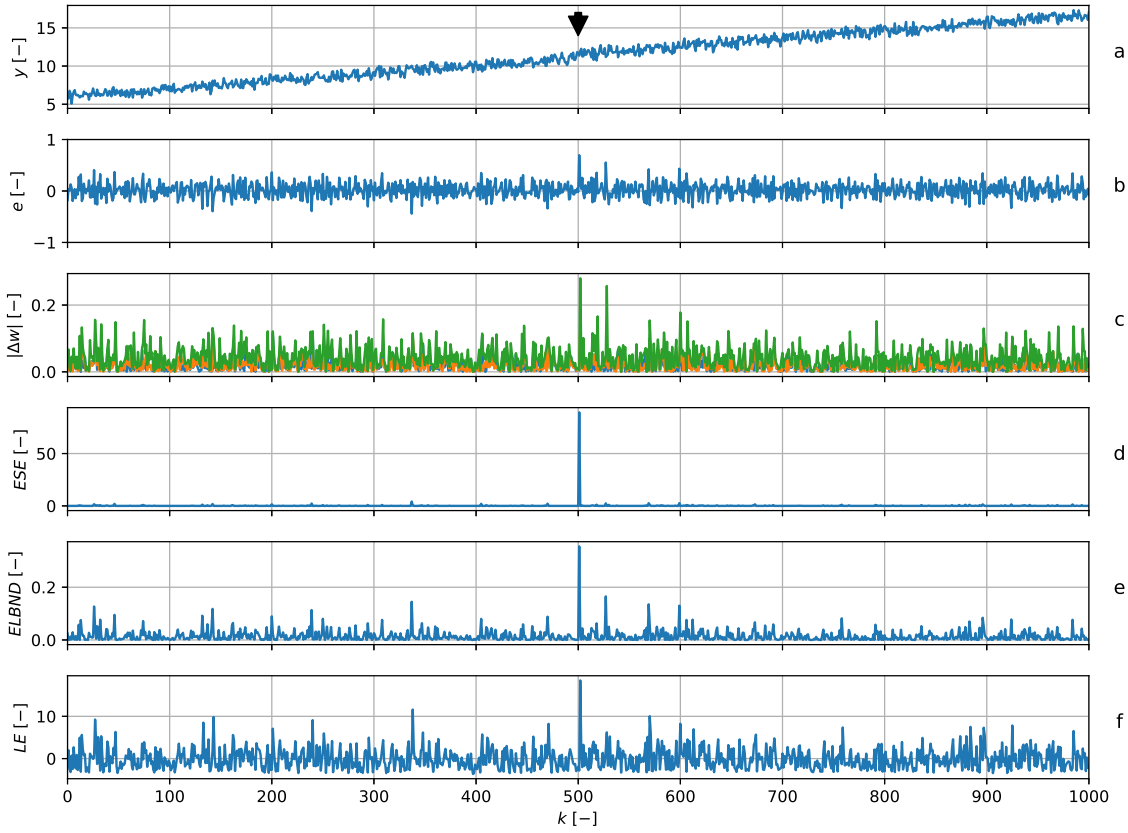
takže odpovídající vektor vstupů je

$$\mathbf{x}(k) = [x_1(k), x_2(k), 1]. \tag{3.14}$$

Struktura LNU byla vybrána tak, aby co nejlépe odpovídala struktuře generátoru signálu. Parametry adaptivního filtru byly v tomto experimentu adaptovány algoritmem GNGD. Na obrázku 3.7 jsou zobrazeny výsledky experimentu. Globální maximum ESE odpovídá okamžiku změny trendu. Algoritmy LE a ELBND úspěšně změnu trendu také detekovali.

3.6 Detekce epilepsie v EEG záznamu myši

Poslední případová studie je věnována detekci epileptického záchvatu v signálu EEG myši pomocí algoritmu EEG. Standartizovaná data ze tří vybraných kanálů EEG, ve kterých byl



Obrázek 3.7: Detekce změny trendu při použití algoritmu GNGD. Na grafu (a) jsou zobrazena data z generátoru signálu (modré). Černá šipka znázorňuje okamžik ve kterém došlo ke změně trendu. Na grafu (b) je zobrazena chyba adaptivního filtru e. Na grafu (c) jsou znázorněny velikosti přírůstků adaptivních vah filtru. Grafy (d), (e) a (f) znázorňují výsledky detekce novosti pomocí algoritmů ESE, ELBND a LE. Všechny tři algoritmy vykazují úspěšnou detekci změny trendu, která koresponduje s jejich maximální hodnoty během experimentu.

expertem stanoven začátek epileptického záchvatu přibližně v čase $k \approx 1700$, jsou zobrazeny na obrázku 3.8. Standartizace byla provedena podle předpisu

$$y = \frac{x - \mu_x}{\sigma_x} \tag{3.15}$$

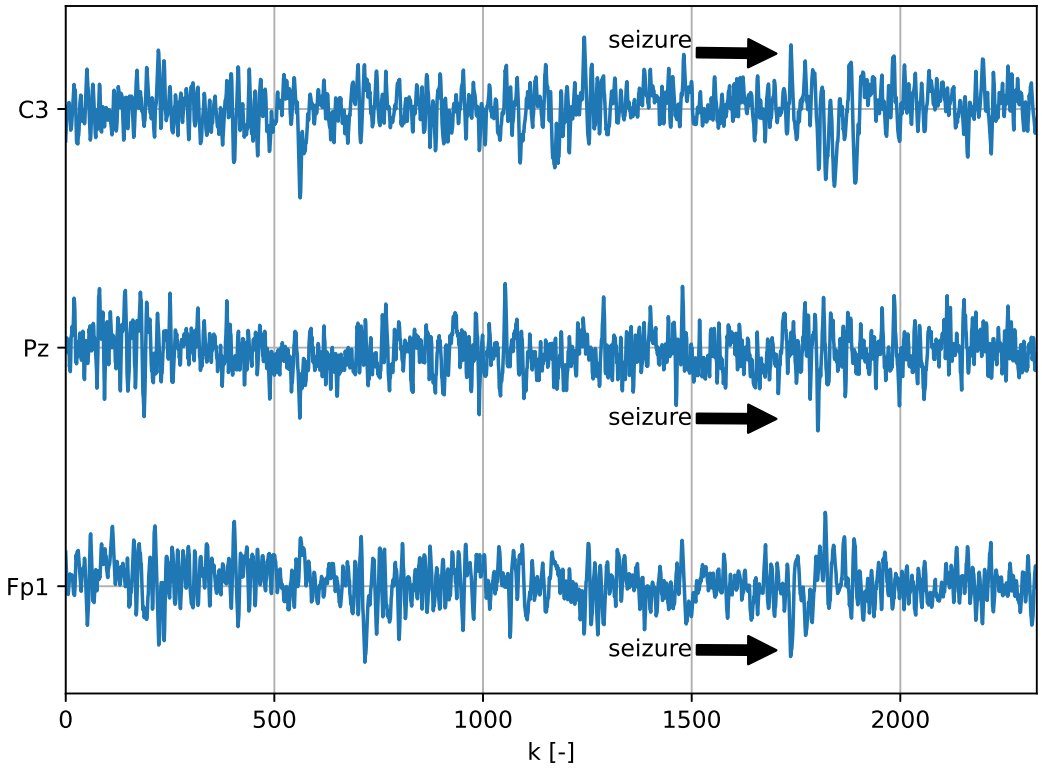
kde y je výsledná standartizovaná hodnota, x je původní hodnota, μ_x je průměrná hodnota původních dat daného kanálu a σ_x je jejich původní směrodatná odchylka.

Jako adaptivní filtr byl, na základě experimentů, zvolen FIR filtr délky 10. Vstupem je vektor dat

$$\mathbf{x} = [x(k-1, x(k-2), \ldots, x(k-10)] \tag{3.16}$$

takže filtr má 10 adaptivních parametrů. Filtr byl adaptován algoritmem NLMS. Výsledky detekce algoritmem ESE jsou zobrazeny na obrázku 3.9. Pozice globálního maxima ESE v

kanálu C3 (obzvláště signifikantní) je v diskrétním časovém okamžiku $k = 1735$, v kanálu Pz je to $k = 1698$ a v kanálu Fp1 je to v $k = 1727$.



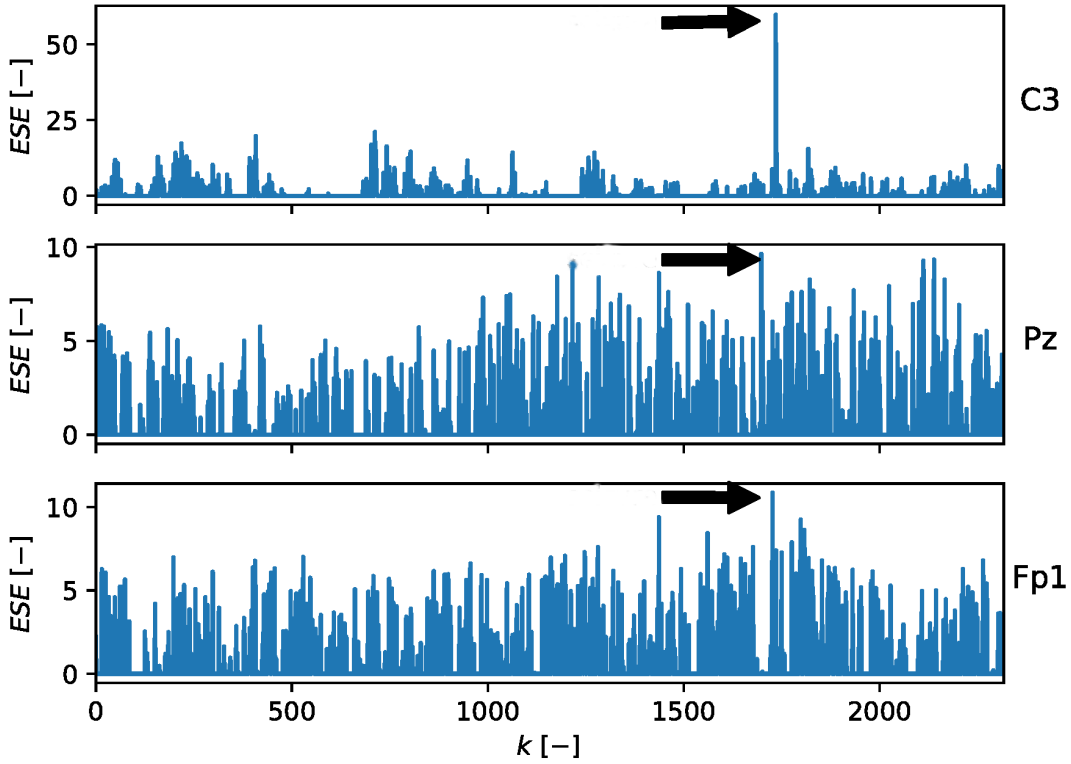
Obrázek 3.8: Vybrané kanály myšího EEG na kterých je patrný epileptický záchvat. Data byly standartizovány. Začátek záchvatu je přibližně v $k \approx 1700$, což znázorňuje černá šipka.

3.7 Vyhodnocení úspěšnosti detekce skokové změny parametrů generátoru signálu

Pro vyhodnocení úspěšnosti skokové změny parametrů generátoru signálu uvažujme generátor signálu s dvěma vstupy $x_1(k)$ a $x_2(k)$ a výstupem $y(k)$ ve tvaru

$$y(k) = a_1 \cdot x_1(k) + a_2 \cdot x_2(k) + a_3 \cdot x_1(k) \cdot x_2(k) + v(k) \tag{3.17}$$

kde člen $v(k)$ reprezentuje gaussovský aditivní šum s nulovou střední hodnotou a směrodatnou odchylkou σ. Počáteční hodnoty parametrů a_1, a_2 a a_3 jsou vygenerovány z rovnoměrného rozdělení $U(-1, 1)$. V diskrétním časovém okamžiku $k = 200$, dojde ke skokové změně těchto parametrů a jejich nová hodnota je opět náhodně vygenerována z rovnoměrného rozdělení $U(-1, 1)$. Celkový počet vzorků experimentu je 400. Použitý adaptivní filtr je stejný jako v předchozí případové studii detekce skokové změny parametrů generátoru signálu, viz kapi-



Obrázek 3.9: Hodnota ESE pro vybrané kanály se záznamem myšího EEG ve kterých je patrný epileptický záchvat. V kanále C3 je v ESE výrazný nárůst po začátku záchvatu (přibližně v $k \approx 1700$), v porovnání s ostatními kanály. Černá šipka znázorňuje přibližný začátek epileptického záchvatu.

tola 3.3. Parametry tohoto adaptivního filtru byly adaptovány algoritmem GNGD. Apriorní informace o parametrech GPD byla pro každý experiment získána pomoci 1200 vzorků, s počátečními hodnotami parametrů a_1, a_2 a a_3. Pro každý experiment byla vyhodnocena hodnota SNR jako

$$SNR = 10 \log_{10} \frac{\sigma_s^2}{\sigma_n^2} \tag{3.18}$$

kde σ_s je hodnota směrodatné odchylky výstupu generátoru signálu během experimentu a σ_n je směrodatná odchylka aditivního gaussovského šumu. Vyhodnocení přesnosti detekce bylo provedeno následujícím způsobem:

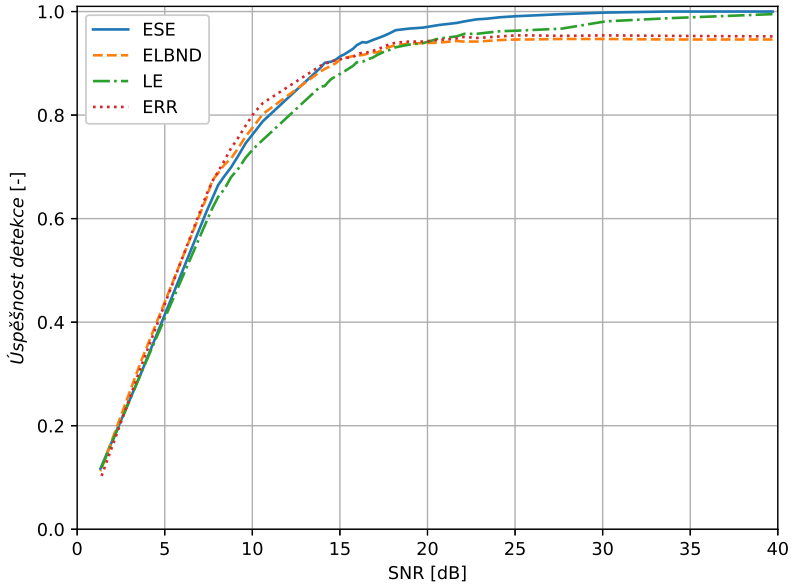
1. nastavení hodnoty směrodatné odchylky šumu σ_n
2. pro zvolenou hodnotu směrodatné odchylky σ_n se provede 1000 experimentů, přičemž pro každý experiment jsou nově vygenerovány počáteční hodnoty parametrů generátoru signálu a_1, a_2 a a_3.

3. pro každý experiment je vyhodnocena úspěšnost detekce. Za úspěšnou detekci je považováno, pokud globální maximum ESE, ELBND, EL respektive chyby filtru je v mezích $k \geq 200$ a $k \leq 210$.

4. vypočte se celková úspěšnost detekce pro danou hodnotu směrodatné odchylky (poměr počtu úspěšných detekcí k celkovému počtu experimentů)

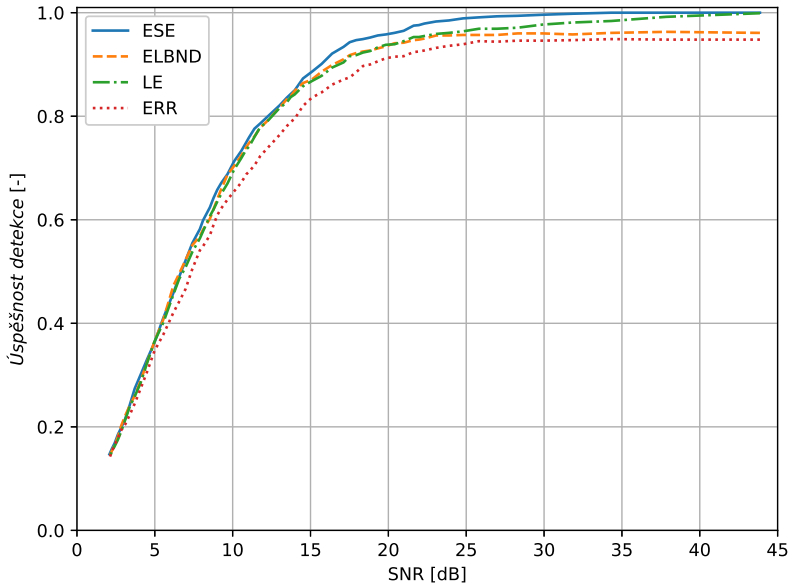
5. pro každý experiment se vyhodnotí SNR podle 3.18 a pak se pro zvolenou hodnotu σ vypočítá průměrná hodnota SNR pro všechny experimenty

Vyhodnocení úspěšnosti detekce bylo vyhodnoceno pro dva případy. V prvním případě, byly hodnoty vstupů $x_1(k)$ a $x_2(k)$ generovány z rovnoměrného rozdělení $U(-1, 1)$. Výsledky úspěšnosti detekce pro různé hodnoty směrodatných odchylek šumu σ_n jsou zobrazeny na obrázku 3.10. Pro porovnání jsou zvoleny metody ELBND , LE s oknem $n_s = 1200$ a velikost chyby adaptivního filtru e (v grafu označeno jako ERR).



Obrázek 3.10: Úspěšnost detekce skokové změny parametrů generátoru signálu. Hodnoty vstupů generátoru signálu jsou generovány z rovnoměrného rozdělení $U(-1, 1)$. Pro hodnoty $SNR > 15\ dB$ dosáhl algoritmus ESE vyšší úspěšnost než algoritmy LE, ELBND a vyhodnocení pomocí chyby filtru (ERR). Pro $SNR > 33\ dB$ dosáhl algoritmus ESE 100% úspěšnost detekce.

Ve druhém případě byly hodnoty vstupů $x_1(k)$ a $x_2(k)$ generovány z normálního rozdělení. Vyhodnocení úspěšnosti detekce bylo provedeno stejně jako v případě popsaném výše. Výsledky úspěšnosti detekce pro různé hodnoty směrodatných odchylek šumu σ_n jsou zobrazeny na obrázku 3.11.



Obrázek 3.11: Úspěšnost detekce skokové změny parametrů signálu. Hodnoty vstupů generátoru signálu jsou generovány z normálního rozdělení $N(0,1)$. Pro hodnoty $SNR > 8\ dB$ dosáhl algoritmus ESE lepší úspěšnosti detekce než algoritmy LE, ELBND a vyhodnocení pomocí velikosti chyby predikce (ERR). Pro $SNR > 34\ dB$ dosáhl algoritmus ESE 100% úspěšnosti detekce.

3.8 Vyhodnocení úspěšnosti detekce skokové změny trendu

Pro vyhodnocení úspěšnosti změny trendu uvažujme výstup generátoru signálu $y(k)$ se dvěma vstupy $x_1(k)$ a $x_2(k)$ jehož výstup je definován jako

$$y(k) = x_1(k) + x_2(k) + 0.01 \cdot k + v(k) \tag{3.19}$$

kde člen $v(k)$ reprezentuje aditivní gaussovký šum s nulovou střední hodnotou a směrodatnou odchylkou σ_n. V diskrétním časovém okamžiku $k = 200$ se změní výstup generátoru signálu

$$y(k) = x_1(k) + x_2(k) + (0.01 + a) \cdot k + v(k) \tag{3.20}$$

přičemž parametr a je vygenerován v každém experimentu z rovnoměrného rozdělení $U(-0.02, 0.02)$. Počet vzorků experimentu je 400.

Struktura adaptivního filtru byla zvolena stejně jako v předcházející případové studii detekce změny trendu (viz kapitola 3.3). Adaptivní parametry filtru byly adaptovány algoritmem GNGD. Apriorní informace o parametrech GPD byla získána na základě 1200 vzorků, ve kterých nedošlo ke změně trendu.

Vyhodnocení přesnosti detekce změny trendu bylo provedeno následujícím způsobem:

1. nastavení hodnoty směrodatné odchylky šumu σ_n

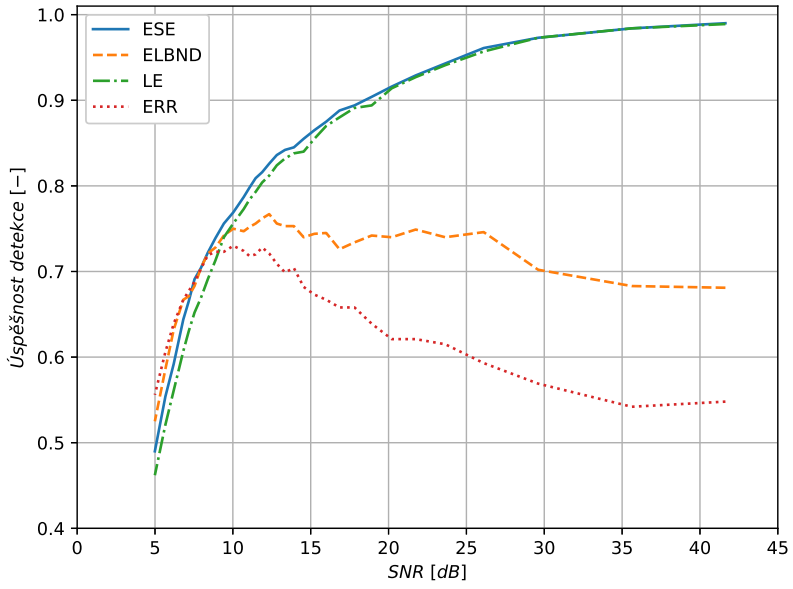
2. pro zvolenou hodnotu směrodatné odchylky σ_n se provede 1000 experimentů. V každém experimentu dojde v diskrétním časovém okamžiku $k = 200$ k novému vygenerování hodnoty parametru a z rovnoměrného rozdělení $U(-0.02, 0.02)$.

3. pro každý experiment je vyhodnocena úspěšnost detekce. Za úspěšnou detekci je považováno, pokud globální maximum ESE, ELBND, EL respektive chyby filtru je v mezích $k \geq 200$ a $k \leq 210$.

4. vypočte se celková úspěšnost detekce pro danou hodnotu směrodatné odchylky šumu σ_n (poměr počtu úspěšných detekcí k celkovému počtu experimentů)

5. pro každý experiment se vyhodnotí SNR podle 3.18 a pak se pro zvolenou hodnotu σ_n vypočítá průměrná hodnota SNR pro všechny experimenty

Vyhodnocení úspěšnosti detekce změny trendu bylo vyhodnoceno pro hodnoty vstupů $x_1(k)$ a $x_2(k)$ vygenerovány z rovnoměrného rozdělení $U(-1, 1)$. Výsledky úspěšnosti detekce pro různé hodnoty směrodatných odchylek šumu σ_n jsou zobrazeny na obrázku 3.12. Pro porovnání jsou zvoleny metody ELBND, LE s oknem $n_s = 1200$ (výpočet podle rovnice a velikost chyby adaptivního filtru e (v grafu označeno jako ERR).



Obrázek 3.12: Úspěšnost detekce změny trendu. Hodnoty vstupů generátoru signálu byly generovány z rovnoměrného rozdělení $U(-1, 1)$. Pro hodnoty $SNR > 8\ dB$ dosáhl algoritmus ESE větší úspěšnosti detekce než LE, ELBND a vyhodnocení pomocí velikosti chyby filtru.

3.9 Evaluace ROC křivky pro detekci změny trendu

Protože úspěšná detekce novosti pomocí algoritmu ESE je závislá na volbě hodnoty, od které budeme považovat hodnotu ESE za „novost“, byl proveden experiment detekce změny trendu a vyhodnocena ROC (Receiver Operating Characteristics) křivka [11]. ROC křivka poskytuje vhodný způsob jak vizualizovat schopnost binárního klasifikátoru klasifikovat správně data na základě proměnlivé velikosti prahu, který klasifikaci určuje (v případě algoritmu ESE je to hodnota ESE) a zároveň umožňuje objektivně jednotlivé klasifikátory porovnávat [12] (více viz následující podkapitola 3.9.2). Pro porovnání algoritmu ESE byly opět zvoleny algoritmy LE, ELBND a klasifikátor, který klasifikuje vzorky náhodně.

Výsledky uvedené v této kapitole byli publikovány v [V3].

3.9.1 Popis experimentu

Stejně jako v případě vyhodnocení přesnosti detekce změny trendu (viz podkapitola 3.8) i v tomto experimentu uvažujeme dva vstupy $x_1(k)$ a $x_2(k)$ výstup generátoru signálu $y(k)$ ve tvaru

$$d(k) = x_1(k) + x_2(x) + 0.01 \cdot k + v(k) \tag{3.21}$$

$$0 \leq k < 200$$

kde člen $v(k)$ reprezentuje gaussovský aditivní šum s nulovou střední hodnotou a směrodatnou odchylkou σ_n. V diskrétním časovém okamžiku $k = 200$ přejde výstup generátoru signálu do tvaru

$$y(k) = x_1(k) + x_2(x) + (0.01 + a) \cdot k + v(k) \tag{3.22}$$

$$200 \leq k \leq 399$$

přičemž hodnota parametru a je vygenerována z rovnoměrného rozdělení $U(-0.02, 0.02)$ a pro všechna $200 \leq k \leq 399$ je během daného experimentu konstantní. Hodnoty vstupů $x_1(k)$ a $x_2(k)$ jsou generovány z rovnoměrného rozdělení $U(-1, 1)$.

Jako adaptivní filtr byl zvolen QNU, jehož struktura odpovídá struktuře generátoru signálu. Výstup adaptivního filtru je ve tvaru

$$\hat{y}(k) = w_1 \cdot x_1(k) + w_2 \cdot x_2(k) + w_3 \cdot x_1(k) \cdot x_2(k) \tag{3.23}$$

a parametry toho adaptivního filtru byly adaptovány algoritmem GNGD.

Apriorní hodnota parametrů GPD pro algoritmus ESE je získána pomocí 1200 vzorků, získaných z výstupu generátoru signálu, který je dán rovnicí 3.21. Během experimentů byla délka okna $n_s = 1200$. Výsledky algoritmu LE byly získány pro okno délky $M = 1200$. Pro každou hodnotu σ bylo provedeno 10000 experimentů na jejichž základě byla zkonstruována

ROC křivka. Hodnoty směrodatných odchylek σ_n byly vybrány takto:

$$\sigma_n = \{0.1, 0.2, 0.5, 1.0, 2.0, 2.5\} \tag{3.24}$$

a pro každou hodnotu σ_n byla pro všech 10000 experimentů určená průměrná hodnota SNR, která byla pro každý experiment vypočtena podle rovnice 3.18.

3.9.2 Konstrukce ROC křivky

Pro konstrukci ROC křivky je důležité, aby množina výsledků byla vyvážená. Tedy aby obsahovala stejný počet pozitivních i negativních vzorků. Pro získání vyvážené množiny výsledků byl nejdřív každý experiment převzorkován podle následujícího předpisu

$$ND_r(i) = max\{ND(i \cdot 10), ND(i \cdot 10 + 1), ND(i \cdot 10 + 2), \ldots, ND(i \cdot 10 + 9)\} \tag{3.25}$$

$$i = 0, 1, \ldots, 39$$

kde ND reprezentuje hodnotu detektoru novosti (resp. hodnoty algoritmu ESE, ELBND, LE). Z každé převzorkované datové řady jsou vygenerovány dvě množiny. Množina P obsahuje pozitivní vzorek, takový, že $P = \{ND_r(20)\}$ (protože v diskrétním časovém okamžiku $k = 200$ došlo ke změně trendu). Množina N obsahuje obsahuje zbylých 39 negativních vzorků, takže $N = \{ND(0), \ldots ND(19), ND(21), \ldots ND(39)\}$. Při konstrukci ROC křivky jsou pro každý experiment vybrány dva vzorky. Jeden vzorek z množiny P a jeden náhodně vybraný vzorek z množiny N. Pro vyhodnocení ROC je klíčové zjistit, jestli jsou pozitivní vzorky (prvky množiny P) pro daný práh správně klasifikovány jako pozitivní (True Positive) a zda-li jsou negativní vzorky klasifikovány jako falešně pozitivní (False Positive). Pro danou velikost prahu se určí úspěšnost detekce skutečně pozitivních (True Positive Rate) jako

$$TPR = \frac{TP}{P} = \frac{TP}{10000} \tag{3.26}$$

kde TP je počet správně pozitivně klasifikovaných vzorků a P je celkový počet skutečně pozitivních vzorků. Dále je potřeba určit poměr falešně pozitivních (False Positive Rate) vzorků pro daný práh jako

$$FPR = \frac{FP}{N} = \frac{FP}{10000} \tag{3.27}$$

kde FP je počet vzorků klasifikovaných jako falešně pozitivní a N je celkový počet skutečně negativních vzorků. ROC pak zobrazuje závislost úspěšnost detekce skutečně pozitivních vzorků (TPR) v závislosti na poměru falešně pozitivních vzorků (FPR).

Hodnota TPR bývá nazývána také jako sensitivita. Komplementární hodnotou k FPR je potom specifita (True Negative Rate TNR), která určuje, kolik opravdu negativních vzorku

(TN) je klasifikováno jako negativní. Komplementární ve smyslu

$$TNR = \frac{TN}{N} = 1 - FPR. \tag{3.28}$$

Komplementární k sensitivitě je hodnota míry falešně negativních (FNR), která určuje poměr falešně negativních k (FN) k celkovému počtu pozitivních, tedy

$$FNR = \frac{FN}{P} = 1 - TPR. \tag{3.29}$$

Pro každou ROC křivku je možné určit plochu pod touto křivkou AUROC (Area Under ROC), která vypovídá o schopnosti klasifikátoru rozlišovat mezi jednotlivými třídami. Čím větší plocha pod křivkou, tím víc klasifikátor správně klasifikuje pozitivní případy jako pozitivní a negativní případy jako negativní. Plocha AUROC ideálního klasifikátoru bude 1, zatímco plocha nejhoršího možného klasifikátoru bude rovna 0 (tento klasifikátor, ale bude dokonalým klasifikátorem, pokud zaměníme označení negativní třídy za pozitivní). Plocha AUROC náhodného klasifikátoru bude 0.5, neboť tento klasifikátor nedokáže vůbec rozlišovat mezi pozitivními a negativními případy.

3.9.3 Výsledky experimentu

Výsledné ROC křivky pro různé hodnoty SNR jsou zobrazeny v obrázcích 3.13-3.18. Modrá čára zobrazuje výsledky algoritmu ESE, zelená tečkovaná čára zobrazuje výsledky algoritmu LE, červená přerušovaná čára výsledky algoritmu ELBND a černá čerchovaná čára zobrazuje výsledky náhodného klasifikátoru.

Pro každou ROC křivku byla vypočtena hodnota plochy pod touto křivkou, AUROC, pomocí lichoběžníkové metody, jako

$$AUROC \approx \sum_{j=1}^{n_t} \frac{TPR(FPR(j)) + TPR(FPR(j+1))}{2} \cdot (FPR(j+1) - FPR(j)) \tag{3.30}$$

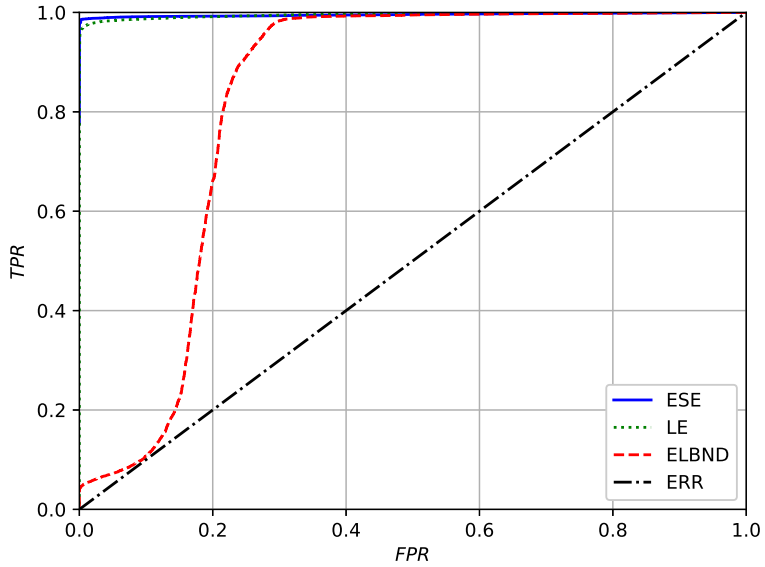
kde n_t reprezentuje počet vyhodnocovaných prahů zmenšený o 1. Výsledné plochy pod křivkami ROC jsou pro jednotlivé metody a směrodatné odchylky šumu uvedené v následující tabulce 3.1. Tučně je zvýrazněna nejvetší hodnota AUROC. Podle uvedených hodnot při průměrném $SNR = 35.8$ nejlépe rozlišuje mezi pozitivními a negativními případy algoritmus LE. Pro nižší hodnoty SNR je nejlépe separujícím algoritmem ESE. V další tabulce 3.2 je uvedená úspěšnost klasifikace, kde za úspěšnou klasifikaci je považován případ, kdy maximální hodnota ESE, ELBND nebo LE během experimentu je v intervalu $200 \leq k \leq 210$, tedy do deseti vzorků po změně trendu. Tučně jsou zvýrazněny hodnoty nejvyšší úspěšnosti detekce. Z výsledků je patrné, že pro průměrné $SNR = 35.8$ má nejvyšší úspěšnost algoritmus LE. Pro nižší hodnoty SNR je algoritmem s nejvyšší úspěšností detekce algoritmus ESE.

Tabulka 3.1: *AUROC* pro detekci změny trendu

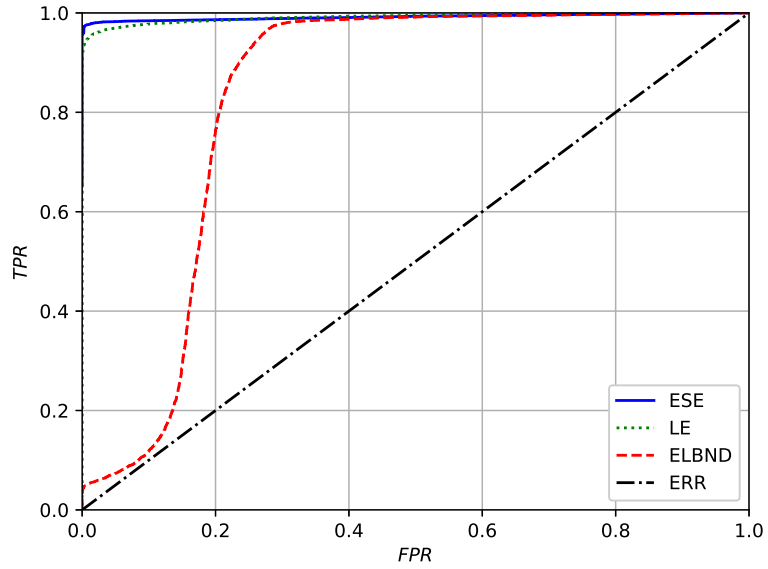
		AUROC		
σ_n	*SNR [dB]*	*ESE*	*LE*	*ELBND*
0.1	35.8	**0.9954**	0.9952	0.8234
0.2	30.0	**0.9920**	0.9912	0.8299
0.5	21.7	**0.9816**	0.9777	0.8288
1.0	16.2	**0.9576**	0.9496	0.8263
2.0	10.8	**0.9286**	0.9214	0.8397
2.5	9.2	**0.9134**	0.9056	0.8446

Tabulka 3.2: Úspěšnost detekce změny trendu

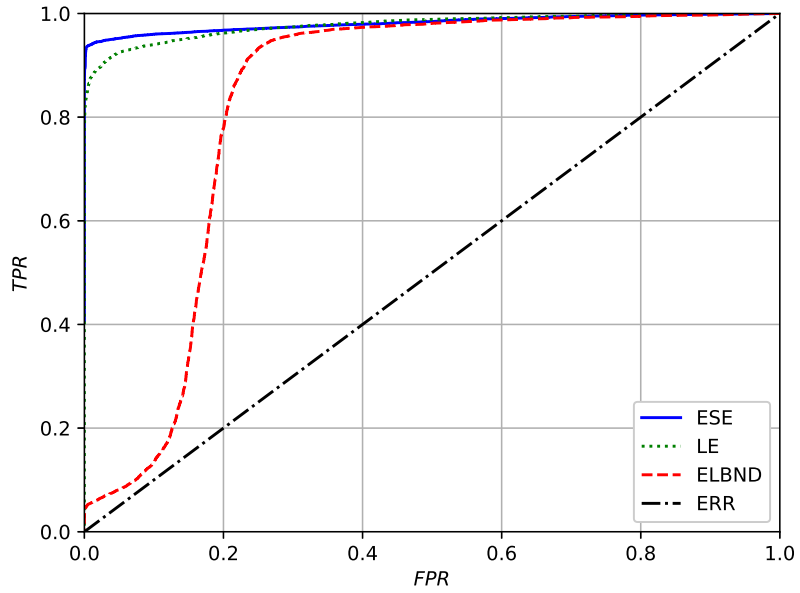
		Úspěšnost detekce		
σ_n	*SNR [dB]*	*ESE*	*LE*	*ELBND*
0.1	35.8	98.88	**98.92**	60.00
0.2	30.0	**98.14**	98.03	59.61
0.5	21.7	**95.18**	95.08	59.65
1.0	16.2	**90.42**	89.96	57.67
2.0	10.8	**81.27**	78.51	57.69
2.5	9.2	**75.86**	71.56	57.16



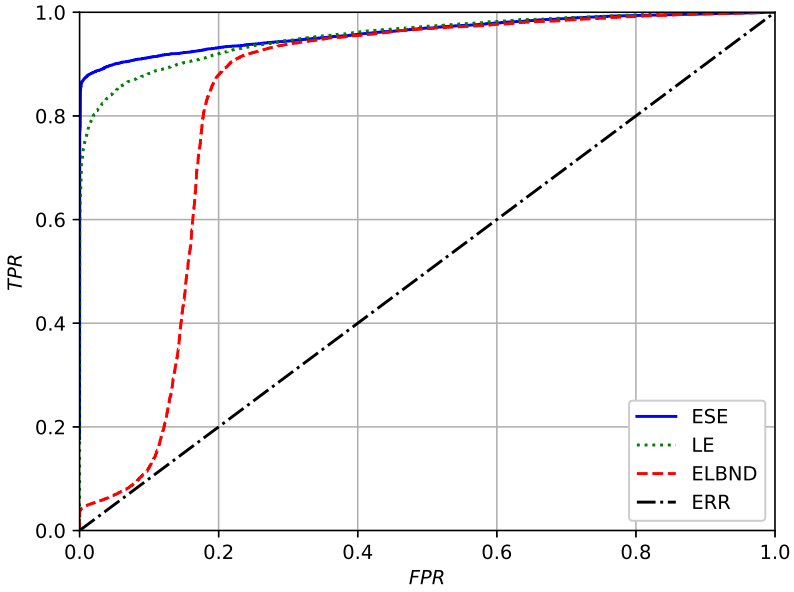
Obrázek 3.13: ROC křivky v případě detekce změny trendu signálu obsahujícího aditivní gaussovský šum se směrodatnou odchylkou $\sigma_n = 0.1$. Průměrná hodnota SNR experimentů byla $SNR = 35.80\,dB$. Černá čerchovaná čára (RANDOM) reprezentuje náhodný klasifikátor.



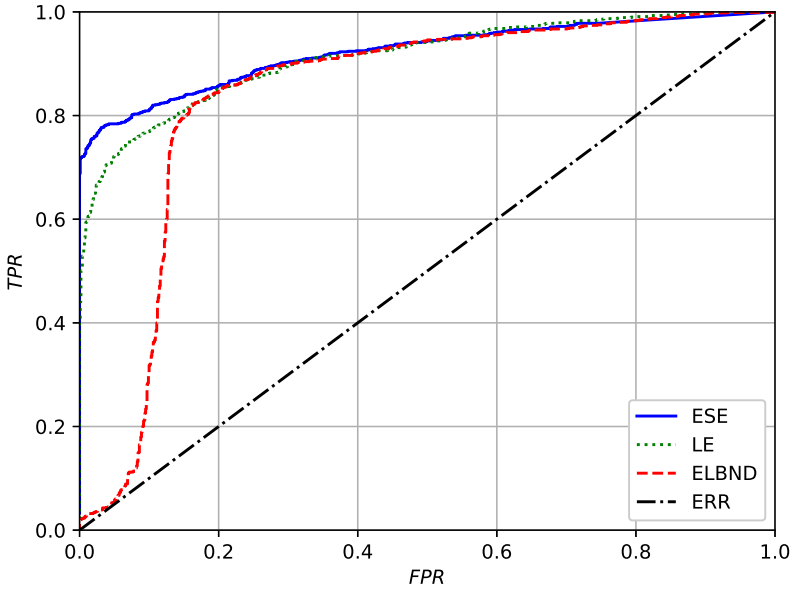
Obrázek 3.14: ROC křivky v případě detekce změny trendu signálu obsahujícího aditivní gaussovský šum se směrodatnou odchylkou $\sigma_n = .2$. Průměrná hodnota SNR experimentů byla $SNR = 30.00\,dB$. Černá čerchovaná čára (RANDOM) reprezentuje náhodný klasifikátor.



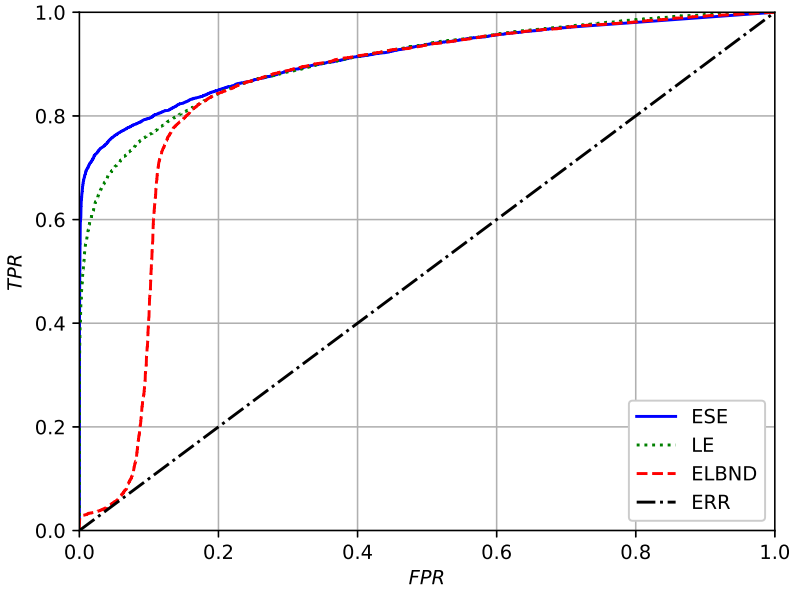
Obrázek 3.15: ROC křivky v případě detekce změny trendu signálu obsahujícího aditivní gaussovský šum se směrodatnou odchylkou $\sigma_n = 0.5$. Průměrná hodnota SNR experimentů byla $SNR = 21.70\,dB$. Černá čerchovaná čára (RANDOM) reprezentuje náhodný klasifikátor.



Obrázek 3.16: ROC křivky v případě detekce změny trendu signálu obsahujícího aditivní gaussovský šum se směrodatnou odchylkou $\sigma_n = 1.0$. Průměrná hodnota SNR experimentů byla $SNR = 16.20\,dB$. Černá čerchovaná čára (RANDOM) reprezentuje náhodný klasifikátor.



Obrázek 3.17: ROC křivky v případě detekce změny trendu signálu obsahujícího aditivní gaussovský šum se směrodatnou odchylkou $\sigma_n = 2.0$. Průměrná hodnota SNR experimentů byla $SNR = 10.88\,dB$. Černá čerchovaná čára (RANDOM) reprezentuje náhodný klasifikátor.



Obrázek 3.18: ROC křivky v případě detekce změny trendu signálu obsahujícího aditivní gaussovský šum se směrodatnou odchylkou $\sigma_n = 2.5$. Průměrná hodnota SNR experimentů byla $SNR = 9.20\ dB$. Černá čerchovaná čára (RANDOM) reprezentuje náhodný klasifikátor.

3.10 Vyhodnocení výpočetní náročnosti metod odhadu parametrů zobecněného Paretova rozdělení

Výsledky v této podkapitole byli publikovány v [V2]. Cílem bylo určit výpočetní čas výpočtu parametrů GPD v typické aplikaci pro použití algoritmu ESE, který byl, v tomto případě, testován ne experimentu detekce skokové změny parametrů generátoru signálu.

3.10.1 Motivace

Detekce novosti v reálném čase je úloha, která nalézá své uplatnění nejen v oblasti detekci a diagnostiky v průmyslových aplikacích [13], ale také např. v detekci narušení počítačových sítí [14] nebo v zabezpečovacích systémech [15]. Další oblastí uplatnění je např. mobilní robotika, která je specifická tím, že robot má k dispozici pouze limitovaný výpočetní výkon [16, 17]. Pro metody detekce novosti v reálném čase je tedy důležité, aby vynikali dostatečně nízkou výpočetní náročností. Z tohoto důvodu byli otestovány tři různé metody výpočtu parametrů GPD, protože tento výpočet je z hlediska použití algoritmu ESE potenciálně limitující z hlediska využitelnosti v aplikacích detekce v reálném čase. Jmenovitě byli otestovány tyto metody: metoda maximální věrohodnosti (ML), metoda momentů (MOM) a metoda kvazi-maximální věrohodnosti (QML). Výpočetní čas potřebný k určení parametrů GPD pomocí těchto metod byl vyhodnocen při experimentu, ve kterém dojde ke skokové změně parametrů

generátoru signálu.

3.10.2 Specifikace experimentu

Vzhledem k povaze experimentu, který slouží k vyhodnocení výpočetní náročnosti různých metod určení parametrů GPD, a nikoliv k detekci novosti v nějakém komplexním procesu, byl zvolen jednoduchý lineární kombinační filtr (LNU), jehož výstup v diskrétním časovém okamžiku k je definován jako

$$\hat{y}(k) = w_1 \cdot x_1(k) + w_2 \cdot x_2(k) + w_3 \cdot x_3(k) \tag{3.31}$$

a tento filtr je adaptován algoritmem NLMS.

Pro výstup generátoru signálu platí vztah

$$y(k) = x_1(k) + x_2(k) + x_3(k) + v(k) \tag{3.32}$$

pro všechny $1 \leq k \leq 200$. Člen $v(k)$ reprezentuje aditivní gaussovský šum s nulovou střední hodnotou a směrodatnou odchylkou $\sigma_n = 0.1$. V diskrétním časovém okamžiku $k = 201$ dojde ke změně generátoru signálu a jeho výstup přejde do tvaru

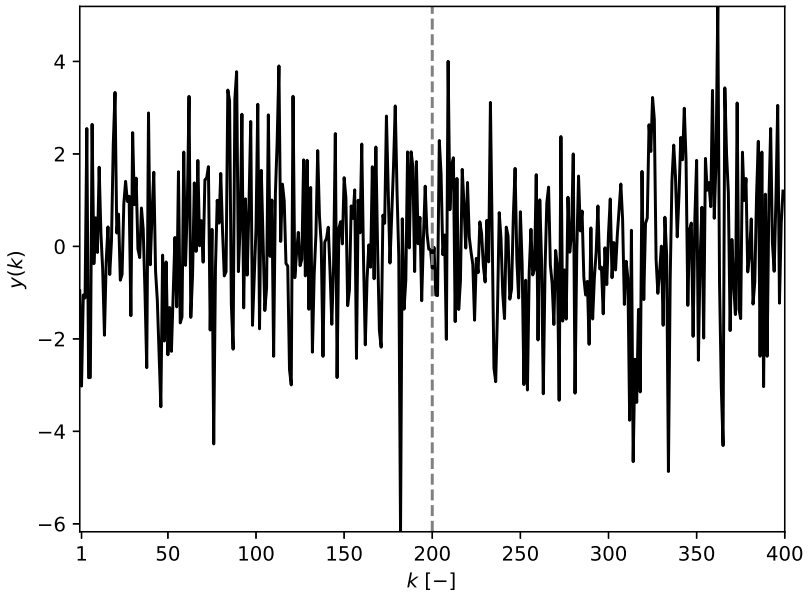
$$y(k) = 0.7 \cdot x_1(k) + 1.2 \cdot x_1(k) + 1.1 \cdot x_1(k) + v(k) \tag{3.33}$$

pro $201 \leq k \leq 400$. Hodnota všech vstupů generátoru signálu je v každém časovém okamžiku k vybrána ze standartního rozdělení normálního rozdělení, takže i-tý vstup $x_i \sim \mathcal{N}(0, 1)$. Změna parametrů signálu byla vybrána tak, aby nedošlo ke změně střední hodnoty signálu $y(k)$.

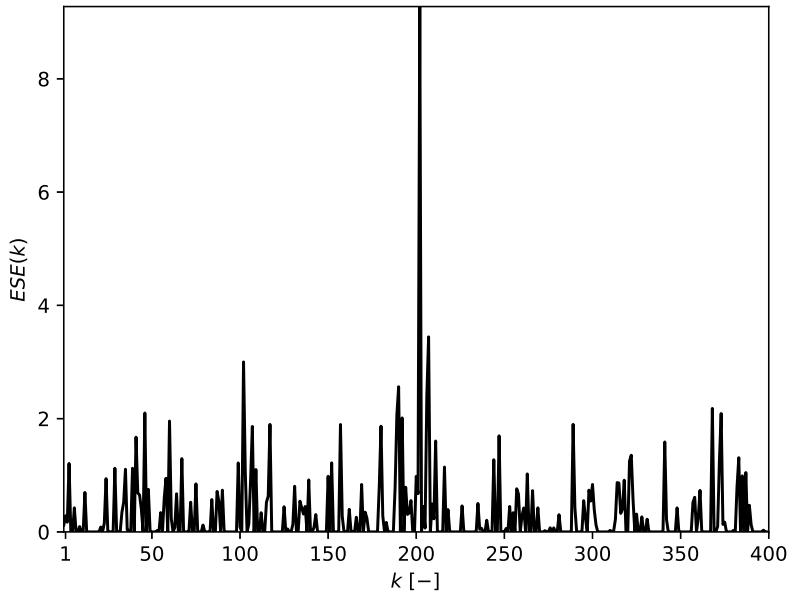
Délka okna pro odhad parametrů GPD byla během experimentu nastavena na $n_s = 1200$. Před experimentem bylo pořízení 1200 vzorků vygenerovaných generátorem signálu definovaným vztahem 3.31, na něž byla použita metoda POT, tak aby při experimentu v diskrétní časový okamžik $k = 1$ byla hodnota ESE relevantní.

Průběh výstupní hodnoty filtru je zobrazen na obrázku 3.19, hodnota ESE potom na obrázku 3.20. Hodnoty parametrů ξ, σ, μ GPD, pro všechny tři adaptivní váhy, během experimentu jsou zobrazeny na obrázcích 3.22, 3.23 a 3.21.

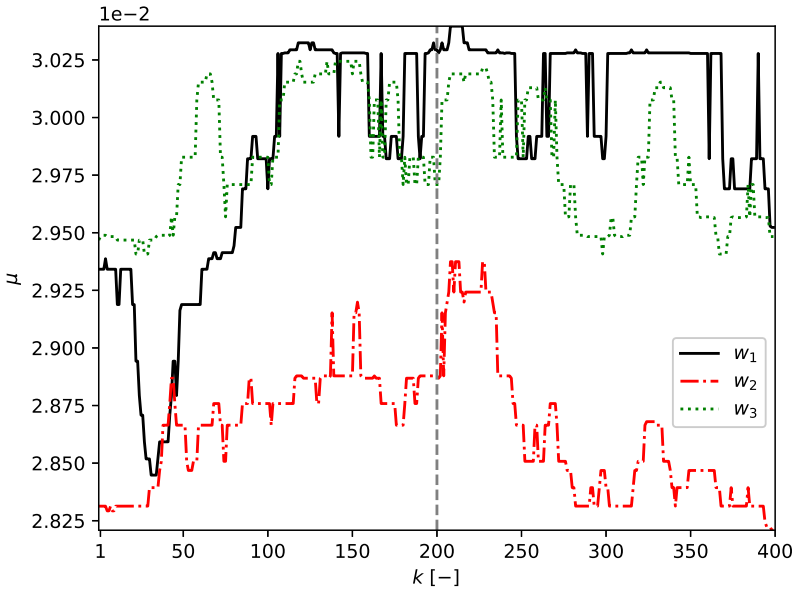
Experiment byl proveden na PC s procesorem Intel(R) Core(TM) i5-7400 se 4mi jádry s taktovací frekvencí 3001 MHz a operační pamětí o velikosti 32 GB. Operační systém byl Windows 10 Pro, 64-bitová verze 10.0.18362. Kód byl napsán v Python 3.6.1 [18] a byly použity knihovny Numpy 1.17.0 [19] a Scipy 1.4.1 [20].



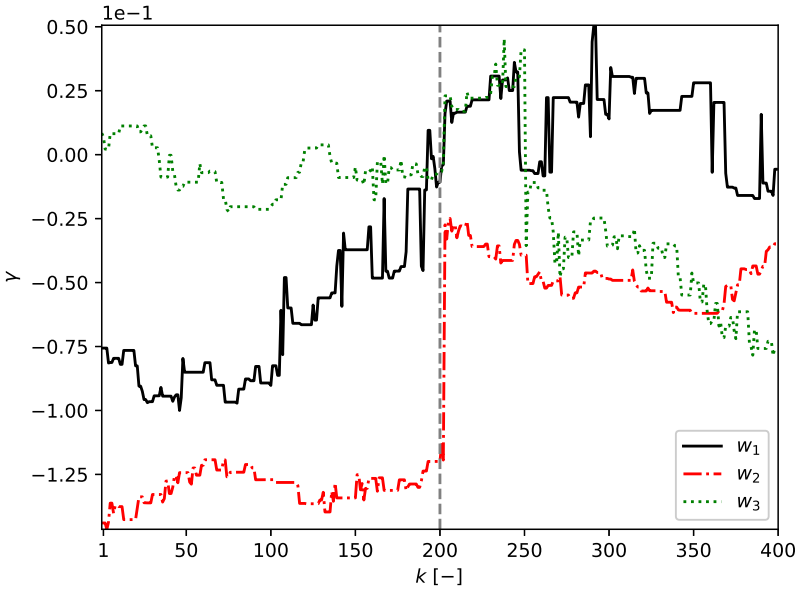
Obrázek 3.19: Výstup adaptivního filtru během experimentu. Skoková změna parametrů generátoru signálu je zvýrazněná svislou vodorovnou čarou v diskrétním časovém okamžiku $k = 200$.



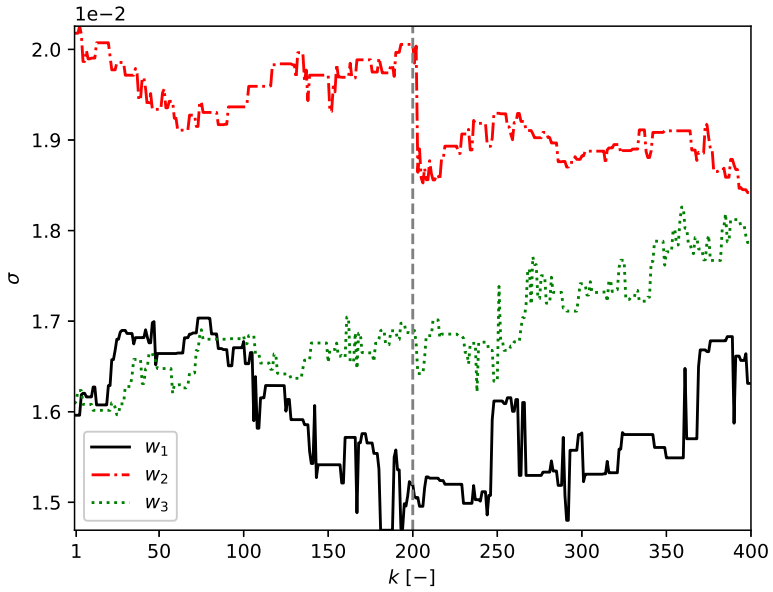
Obrázek 3.20: Hodnota ESE během experimentu. Globální maximum odpovídá změně parametrů generátoru signálu, resp. úspěšné detekci novosti.



Obrázek 3.21: Hodnota parametru μ GPD pro všechny tři adaptivní váhy w_1, w_2, w_3 během experimentu detekce změn parametrů generátoru signálu. Svislá čára v diskrétním časovém okamžiku $k = 200$ znázorňuje skokovou změnu parametrů generátoru signálu.



Obrázek 3.22: Hodnota parametru γ GPD pro všechny tři adaptivní váhy w_1, w_2, w_3 během experimentu detekce změn parametrů generátoru signálu. Svislá čára v diskrétním časovém okamžiku $k = 200$ znázorňuje skokovou změnu parametrů generátoru signálu.



Obrázek 3.23: Hodnota parametru σ GPD pro všechny tři adaptivní váhy w_1, w_2, w_3 během experimentu detekce změn parametrů generátoru signálu. Svislá čára v diskrétním časovém okamžiku $k = 200$ znázorňuje skokovou změnu parametrů generátoru signálu.

3.10.3 Výsledky a diskuze

Průměrný čas výpočtu $\bar{t}$ parametrů všech tří GPD (počet GPD odpovídá počtu adaptivních parametrů filtru) a odpovídající směrodatné odchylky σ_t jsou uvedeny v následující tabulce 3.3. Čas výpočtu je určený pro jeden experiment (400 vzorků).

Tabulka 3.3: Tabulka průměrných časů výpočtu pro jednotlivé adaptivní váhy a odpovídajících směrodatných odchylek vybraných metod výpočtu parametrů GPD

	Metoda	$\bar{t}$ [ms]	σ_t [ms]
w_1	ML	26.198	3.396
	QML	0.354	0.478
	MOM	**0.076**	0.264
w_2	ML	26.718	2.302
	QML	0.337	0.471
	MOM	**0.064**	0.244
w_3	ML	24.982	1.964
	QML	0.395	0.489
	MOM	**0.060**	0.238

Z uvedených výsledků je patrné, že nejrychlejší metoda je MOM. Podstatnou nevýhodou této metody pro využití v aplikacích, které vyhodnocují data v reálném čase je její omezení na hodnoty parametrů GPD. Pokud parametry uvedené omezení nesplňují, vypočtené hodnoty

nepřesné (resp. nesmyslné) a tedy nepoužitelné pro algoritmus ESE, který začne produkovat nepřesné výsledky. Z pohledu úlohy detekce novosti je diskutabilní, zda-li můžeme garantovat, že sledovaný proces po celou dobu bude splňovat uvedené omezení.

Metoda, jejíž výpočetní čas byl nejvyšší je ML, což je vzhledem k iterativnímu určení parametrů GPD očekávatelné. Nevýhoda použití této metody tkví v nemožnosti určit minimální resp. maximální počet iterací. Jednou z možností jak zrychlit nalezení parametrů je využití apriorní informace o hodnotách těchto parametrů. V rámci experimentu však byla využita pouze apriorní informace o parametru μ, který odpovídá nejmenší hodnotě přírůstku vah, které byli získány metodou POT aplikovanou na plovoucí okno délky n_s.

Dobrým kompromisem mezi výše uvedenými metodami je použití metody QML. Výpočetní čas této metody byl v uvedeném experimentu o dva řády kratší než ML a asi pětkrát delší než MOM. Pro každou z vyhodnocovaných vah byl kratší než 500 μs.

Z provedeného experimentu je patrné, že použití algoritmu ESE pro aplikace v reálném čase je limitováno počtem parametrů filtru a rychlostí vzorkování monitorovaného procesu.

4 Závěr

Předložená dizertační práce je věnována použití adaptivních systémů při analýze dat. Za zásadní výsledek lze považovat nový originální algoritmus pro detekci novosti, který vyhodnocuje přírůstky adaptivních vah filtru, Extreme Seeking Entropy [V1] (viz kapitola 2). Tento algoritmus byl otestován v následujících případových studiích (viz kapitola 3): detekce pertubace v chaotické časové řadě získané řešením Mackey-Glassovy rovnice, detekce změny rozptylu šumu v náhodném datovém toku, detekce skokové změny parametrů generátoru signálu, detekce náhlé absence šumu, detekce změny trendu a při detekci epilepsie v myším EEG. Pro detekci skokové změny trendu a skokové změny parametrů signálu byla vyhodnocena úspěšnost této detekce a výsledky porovnány s výsledky algoritmů Learning Entropy a Error and Learning Based Novelty Detection, přičemž v obou případech byla úspěšnost detekce algoritmu ESE vyšší pro téměř všechny vyhodnocované hodnoty SNR. Pro hodnotu $SNR > 34\ dB$ dosáhl algoritmus při detekci skokové změny parametrů generátoru signálu 100% úspěšnost. Při detekci změny trendu měl algoritmus ESE pro hodnoty $SNR > 8\ dB$ větší úspěšnost detekce než srovnávané algoritmy LE a ELBND. Výše uvedené výsledky byly publikovány v [V1].

Pro možné použití v aplikacích detekce v reálném čase byla experimentálně zjišťována výpočetní časová náročnost různých metod odhadů parametrů zobecněného Paretova rozdělení v případě použití algoritmu ESE při detekci skokové změny parametrů [V2]. Výsledkem je porovnání 3 různých metod odhadu parametrů. Limitujícím faktorem použití ESE v reálném čase je v zásadě počet adaptivních parametrů filtru, které je potřeba vyhodnocovat a samozřejmě rychlost vzorkování monitorovaného signálu.

Pro odhad úspěšnosti detekce novosti pomocí algoritmu ESE byla také vyhodnocena ROC křivka v případě detekce změny trendu signálu s různými poměry SNR a byly určeny příslušné plochy pod těmito ROC křivkami [V3]. Dosažené výsledky byly opět porovnány s algoritmy LE a ELBND a bylo ověřeno, že pro hodnoty $SNR \leq 30\ dB$ dosahuje algoritmus ESE lepších výsledků. Pro vyšší hodnoty SNR pak byly výsledky ESE srovnatelné s výsledky LE. Cílem této studie bylo zjistit jak dobře dokáže algoritmus ESE separovat nová data v závislosti na volbě prahu, který rozhoduje o tom zda data obsahují novost či nikoliv.

Stanovené cíle dizertační práce (viz kapitola 1.1) tak lze, na základě výše uvedených výsledků, považovat za splněné.

4.1 Možné směry budoucího výzkumu

V budoucnu se nabízí rozvíjet téma využití adaptivních systémů ve zpracování dat několika směry. Potenciální využití vyhodnocení změn vah adaptivních systémů lze využít v optimalizaci velikosti datasetů v oblasti hlubokého učení, což by mohlo výrazně snížit časovou náročnost učení hlubokých sítí. Za účelem snížení výpočetního času algoritmu ESE je potřeba vyzkoušet další metody odhadu parametrů zobecněného Paretova rozdělení a vyzkoušet adaptivní metody volby velikosti prahu pro metodu Peak-over-threshold. Zajímavým tématem je také vliv šumu a jeho typu na velikost přírůstků adaptivních vah filtrů a jejich pravděpodobnostní rozdělení. V neposlední řadě se nabízí otázka, jak ovlivní typ kriteriální funkce pro optimalizaci adaptivního filtru výsledky algoritmu ESE a je-li možné různé kriteriální funkce využívat k detekci novosti, případně pomocí nich typ novosti klasifikovat.

Publikace autora

[V1] VRBA, Jan; MAREŠ, Jan. Introduction to Extreme Seeking Entropy. *Entropy*, 2020, 22.1: 93.

[V2] VRBA, Jan; MAREŠ, Jan. Computational Performance of the Parameters Estimation in Extreme Seeking Entropy Algorithm. In: *2020 International Conference on Applied Electronics (AE)*. IEEE, 2020. p. 1-4.

[V3] VRBA, Jan; MAREŠ, Jan. ROC Analysis of Extreme Seeking Entropy for Trend Change Detection. In: *2020 International Conference on Applied Electronics (AE)*. IEEE, 2020. p. 1-4.

[V4] VRBA, Jan. Využití fuzzy systémů a algoritmu learning entropy pro detekci změn stavů bioprocesu. In: *Automatizacia a riadenie v teorii a praxi ARTEP 2017*. Technická univerzita Košice, 2017.

[V5] VRBA, Jan. *XLIII. Seminář ASŘ - Adaptivní metoda detekce* [přednáška]. Ostrava: VŠB TU Ostrava, 27.4.2018.

[V6] VRBA, Jan. Adaptive Novelty Detection with Generalized Extreme Value Distribution. In: *2018 International Conference on Applied Electronics (AE)*. IEEE, 2018. p. 1-4.

[V7] BUKOVSKÝ, Ivo, et al. Study of learning entropy for onset detection of epileptic seizures in EEG time series. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016. p. 3302-3305.

[V8] OSWALD, Cyril, et al. Novelty Detection in System Monitoring and Control with HONU. In: *Applied Artificial Higher Order Neural Networks for Control and Recognition*. IGI Global, 2016. p. 61-78.

[V9] VRBA, Jan, et al. An Automated Platform for Microrobot Manipulation. In: *International Workshop on Soft Computing Models in Industrial and Environmental Applications*. Springer, Cham, 2020. p. 255-265.

[V10] BÍLA, Jiří; VRBA, Jan. The Detection and Interpretation of Emergent Situations in ECG Signals. In: *International Conference on Soft Computing-MENDEL*. Springer, Cham, 2016. p. 264-275.

[V11] MOJZES, Matej, et al. Feature selection via competitive levy flights. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016. p. 3731-3736.

[V12] BÍLA, Jiří; NOVÁK, Martin; VRBA, Jan. Detection of emergent situations in complex systems represented by algebras of transformations. In: *MATEC Web of Conferences*. EDP Sciences, 2016. p. 02035.

[V13] BUKOVSKÝ, Ivo, OSWALD, Cyril, VRBA, Jan. Případová studie použití entropie učení pro adaptivní detekci při řízení spalování tuhých paliv. In: *Automatizácia a riadenie v teórii a praxi 2015*. Technická univerzita Košice, 2015.

Literatura relevantní k tezím

[1] SHARMA, Anish and ANDREWS, Rebecca. Managing-Exponential-Data-Growth-and-Application-Modernization. *IBM* [online]. 11 November 1999. [cit. 11.8.2020]. Dostupné z: https://www.ibm.com/cloud/blog/managing-exponential-data-growth-and-application-modernization

[2] BUKOVSKY, Ivo. Learning entropy: Multiscale measure for incremental learning. *Entropy*, 2013, 15.10: 4159-4187.

[3] BUKOVSKY, Ivo; KINSNER, Witold; HOMMA, Noriyasu. Learning Entropy as a Learning-Based Information Concept. *Entropy*, 2019, 21.2: 166.

[4] CEJNEK, Matous; BUKOVSKY, Ivo. Concept drift robust adaptive novelty detection for data streams. *Neurocomputing*, 2018, 309: 46-53.

[5] CEJNEK, Matous; BUKOVSKY, Ivo. Influence of type and level of noise on the performance of an adaptive novelty detector. In: *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2017. p. 373-377.

[6] CEJNEK, Matous; BUKOVSKY, Ivo; VYSATA, Oldrich. Adaptive classification of EEG for dementia diagnosis. In: *2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*. IEEE, 2015. p. 1-5.

[7] MACKEY, Michael C.; GLASS, Leon. Oscillation and chaos in physiological control systems. *Science*, 1977, 197.4300: 287-289.

[8] SPANGENBERG, Mariana, et al. Detection of variance changes and mean value jumps in measurement noise for multipath mitigation in urban navigation. *Navigation*, 2010, 57.1: 35-52.

[9] L'ECUYER, Pierre. History of uniform random number generation. In: *2017 Winter Simulation Conference (WSC)*. IEEE, 2017. p. 202-230.

[10] MAURYA, Mano Ram; RENGASWAMY, Raghunathan; VENKATASUBRAMANIAN, Venkat. Fault diagnosis using dynamic trend analysis: A review and recent developments. *Engineering Applications of artificial intelligence*, 2007, 20.2: 133-146.

[11] EGAN, James P. *Signal detection theory and ROC-analysis*. Academic press, 1975.

[12] FAWCETT, Tom. An introduction to ROC analysis. *Pattern recognition letters*, 2006, 27.8: 861-874.

[13] GERTLER, Janos. *Fault detection and diagnosis in engineering systems*. CRC press, 1998.

[14] YU, Kangqing, et al. Real-time Outlier Detection over Streaming Data. In: *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2019. p. 125-132.

[15] RAMEZANI, Ramin; ANGELOV, Plamen; ZHOU, Xiaowei. A fast approach to novelty detection in video streams using recursive density estimation. In: *2008 4th International IEEE Conference Intelligent Systems. IEEE, 2008*. p. 14-2-14-7.

[16] MARSLAND, Stephen; NEHMZOW, Ulrich; SHAPIRO, Jonathan. On-line novelty detection for autonomous mobile robots. *Robotics and Autonomous Systems*, 2005, 51.2-3: 191-206.

[17] NEHMZOW, Ulrich, et al. Novelty detection as an intrinsic motivation for cumulative learning robots. In: *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer, Berlin, Heidelberg, 2013. p. 185-207.

[18] VAN ROSUM, G.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[19] OLIPHANT, Travis E. *A guide to NumPy (Vol. 1)*. Trelgol Publishing USA, 2006.

[20] VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020, 17(3), 261-272.