Posudek oponenta bakalářské práce

Název: Wild Binary Segmentation
Autor: Jakub Lasota

Shrnutí obsahu práce

Bakalárska práca študenta Jakuba Lasotu sa venuje problému hľadania a odhadovania počtu a lokácie neznámych bodov zmien (tzv. change-pointov) v polohe zašumeného signálu (t.j., neznámej strednej hodnote) a využíva k tomu metódy založené na postupnej segmentácii pozorovaných dat (konkrétne tzv. *binary segmentation* postup a jeho modifikáciu známu ako *wild binary segmentation*).

Práca pozostáva z troch hlavných častí: V prvej kapitole autor stručne vysvetľuje základný princíp kumulativných súčtových (CUSUM) štatistík, ktoré tvoria teoretický základ pre binárnú segmentáciu aj pre jej divodkú (wild) modifikáciu. V druhej kapitole je predstavený samotný (algoritmický) princíp fungovania binárnej segmentácie a obe metódy sú ilustrované na praktických príkladoch. V poslednej kapitole sú prezentované metódy aplikované na reálne trhové data logaritmických denných výnosov spoločnosti *Zoom Video Communications (ZOOM)*.

Tématický považujem prácu za veľmi zaujímavú a určite vhodnú pre bakalársku prácu. Z odborného hľadiska sa jedná o netriviálny text, ktorý vyžadoval naštudovanie a pochopenie matematickej, pravdepodobnostnej a štatistickej teórie, ktorá ide výrazne nad rámec bakalárskeho študia. Samotné vypracovanie v podaní autora ale žiaľ pôsobí rozpačito a neúplne. Druhá a tretia kapitola dokonca pôsobia dojmom, akoby boli písané na poslednú chvíľu a výrazne by potrebovali ešte uhladiť. V práci nie je ani jedná časť, ktorá by sa dala považovať za solídnu. Z formálneho hľadiska pôsobí divne napr. zaradenie appendixu pred záverečnú kapitolu a zoznam literatúry, používanie tabuliek bez popiskov, alebo chybné odkazovanie na obrázky (`\eqref{}` namiesto `\ref{}`). Z matematického hľadiska autor často nerozlišuje medzi vektorom, funkciou, postupnosťou, či jednorozmernou kvantitou, zamieňa intervaly a konečné množiny, alebo používa značenie, ktoré nie je zavedené, alebo vysvetlené. Formulácia matematického textu je často nesprávna, resp. minimálne hodne neintuitívna (zvláštne je hlavne nerozlíšovanie medzi pojmami *estimate*, *estimator* a *estimation*) a myslím si, že aj formulácia nematematického (t.j., anglického) textu a celková úroveň práce by šla výrazne zlepšiť (napr., nekonzistentné používanie výrazov *change-point* a *change point*, niektoré neobratné anglické formulácie, ktoré pôsobia ako doslovný preklad, či výrazná absencia používania členov).

Vzhľadom k vyššie uvedenému hodnotím prácu skôr ako podpriemernú, ale stále ju doporučujem štátnicovej komisii uznať ako bakalársku prácu na MFF UK. Celkové hodnotenie bude závisieť na výslednej prezentácii pri obhajobe práce a na schopnosti autora odpovedať na príslušné otázky.

Otázky & pripomienky k obhajobe

- Teoretický model uvažovaný v práci (str.2) má tvar $Y_t = f_t + \varepsilon_t$, kde $t \in \{1, \dots, T\}$ predstavuje pozorovania ($T \in \mathbb{N}$) a $f_t \in \mathbb{R}$ sú nejaké neznáme reálne hodnoty. Autor ale v práci explicitne píše, že "*f_t je deterministická, jednorozmerná a po častiach konštantná funkcia.*" Funkcia je ale definovaná ako predpis, ktorý každej hodnote z definičného oboru funkcie priradí práve jednú hodnotu z príslušného oboru hodnôt funkcie. Ako je potrebné zápis modelu v (2.1) opraviť tak, aby sa dalo korektne hovoriť o deterministickej, jednorozmernej a po častiach konštantnej funkcii v zmysle definície?

- V úvode Sekcie 1.2 (str.3) je napísané, že "*CUSUM štatistika je definovaná ako skalárný súčin medzi vektorom pozorovaní* $(X_s, \ldots, X_e)^\top$ *a konkrétnym vektorom* $\tilde{X}^b_{s,e}$", ktorý je následne definovaný výrazom (2.2). Hodnota $\tilde{X}^b_{s,e}$ je ale zjavne jednorozmerná (náhodná) veličina, nie vektor. Ako teda správne chápať tvrdeniu, že sa jedná o skalárný súčin?

- Autor následne uvádza, že náhodná veličina $\tilde{X}^b_{s,e}$ je *vektor kontrastných váh (vector of 'contrast' weights)*. Nejedná sa ale o vektor. V akom zmysle sa teda jedna o váhy (resp. váhu – singulár)? A pripúšťajú sa aj záporné hodnoty pre váhy?

- Ako presne vyzerá množina $\mathcal{F}^b_{s,e}$? Autor v práci tvrdí (str.3), že je to "*množina všetkých vektorov na intervale* $[s, e]$"?

- Čo presne má autor na mysli formuláciu, že "*estimation of this change-point is very likely to be correct*" (str. 4)? Bolo by možné toto tvrdenie matematicky formalizovať?

- Na str.5 je uvedené, že najmenšie medzery medzi dvoma následnými bodmi zmeny (*minimum spacing between change-points*) sú alespoň (*greater or equal than*) δ_T, kde $\delta_T \leq CT^\Theta$, pre nejaké $C > 0$ a $\Theta \leq 1$. Inými slovami, obmedzenie na minimálne medzery medzi dvoma bodmi zmeny môže byť ľubovoľne malé?

- Binárna segmentácia a aj jej divoká (wild) modifikácia sú určené na hľadanie bodov zmien v signále, ktorý je po častiach konštantný. V úvode aplikácie (str.22) ale autor uvádza, že metódy využije na hľadanie a analýzu trendov v denných logaritmických výnosoch spoločnosti Zoom Video Communications (čím sa väčšínou mysli po častiach lineárny signál). U logaritmických výnosov sa síce všeobecne predpokladá viac-menej konštantný (a dokonca) nulový priebeh, takže z tohto pohľadu je aplikácia a data ok, ale príde mi použitá formulácia trochu zvlaštna. Minimálne sa ale ponúka otázka, či ilustrácia metód určených na analýzu po častiach konštantného (informativného) signálu na data logaritmických výnosov s predpokladanou nulovou strednou hodnotou (t.j., neinformatívnym signálom) je vhodnou voľbou.

- Všetky body zmeny detekované algoritmom WBS (Wild Binary Segmentation), prezentované od str.22, mi prídu patologické a z praktického hľadiska nezmyselné. V prvom rade, dva po sebe následujúce body zmeny (napr. detekované change-pointy v bodoch 233 a 234, 246 a 347, alebo 843 a 844, atď) jednak nekorespondujú s teoretickým predpokladom modelu (ohľadom minimálneho rozostupu medzi change-pointmi) a za druhé, čo mi príde zásadné, tak z pohľadu na data a obrázky je hneď zrejmé, že detekované change-pointy nekorespondujú so zmenou magnitúdy po častiach konštantného signálu, ale skôr len identifikujú jednorázove extrémne logaritmické výnosy—čim vlastne ukazujú na značnú nerobustnosť celého WBS algoritmu.

Drobné poznámky

- Na str.2/3 je uvedené, že ε je premenná, ktorá popisuje chybu (error variable/noise). Chýba ale informácia o tom, že sa jedná o *náhodnú* veličinu. Navyše veličina ε sa v rovnici modelu na str.2 ani v rovnici v (2.1) vôbec nevyskytuje—asi by malo byť správne ε_t, pre $t = 1, \ldots, T$;

- Na str.3 autor uvádza, že X_t *predstavuje data* ("*we often call* X_t *'data'.*"). Hodnota X_t ale predstavuje pouze jednu realizáciu z celkového datového súboru. Data by správne mali byť reprezentované zápisom $\{X_t\}_{t=1}^T$, prípadne ako množina $\{X_1, \ldots, X_T\}$;

- ❏ V poslednom výraze na str.3 sa objavujú symboly, ktoré nie sú zavedené/definované—konkrétne hodnoty X_s^e a $\overline{f}_{s,e}^b$. Vo výraze je navyše uvedené, že $\overline{f}_{s,e}^b \in \mathcal{F}_{s,e}^b$, kde $\mathcal{F}_{s,e}^b$ je definované ako *množina vektorov* (viď vyššie). Aky rozmer má teda vektor $\overline{f}_{s,e}^b$ a ako je definovaná binárna operácia vo výraze $X_s^e - \overline{f}_{s,e}^b$ (resp. čo predstavuje hodnota X_s^e)?

- ❏ Obecně by asi bolo vhodnejšie používať 'arg min' (`\arg\min`) namiesto 'arg min' (`arg\min`);

- ❏ Je zrejmé, že $f_t \in \mathbb{R}$ (hoci autor vo svojej práci výraz f_t označuje ako *funkcia*—v pôvodnom článku autori používajú pojem 'signal'). Ako správne chápať, resp. vysvetliť/popísať výraz $\{f_t\}_{t=1}^T$ na str.5, ktorý autor opäť označuje ako funkcia?

- ❏ Zápis $\{\varepsilon_t\}_{t=1}^T \sim N(0,1)$ mi nepríde úplne korektný—asi by bolo vhodnejšie použiť zápis $\varepsilon_t \sim N(0,1)$, pre $t = 1, \ldots, T$; Navyše, ak je uvažované rozdelenie náhodných chýb štandardné normálne (i.e., jednotkový rozptyl), prečo následne autor píše, že "*budeme predpokladať, že rozptyl* $Var\varepsilon_t$ *je známy*"?

- ❏ Čo presne znamená výraz $|f_t| < \overline{f} < \infty$ ak $t \in [1, T]$ (namiesto $t \in \{1, \ldots, T\}$)? Čo predstavuje hodnota $\overline{f}$ (nie je v práci definovaná)?

- ❏ Vo Vete 1, 2 a 3 (Theorem 1, 2, 3) je uvedený predpoklad, že "X_T *follows the model (2.1)*". Naozaj pre tvrdenie viet postačuje, aby posledné poorovanie X_T bolo z modelu (2.1)? Není náhodou potrebné, aby všetky pozorovania, t.j. $\{X_t\}_{t=1}^T$, korespondovali s modelom (2.1)?

- ❏ Na konci str.6 je uvedená veta, že "*we can surely say, that the number of true change-points is* $N = 1$", pričom sa autor odkazuje na Obr.3.1. Na základe čoho autor poskytuje takéto silné tvrdenie? Z pohľadu na Obr.3.1. to podľa mňa nie je zďaleka také evidentné;

- ❏ Na str.7 je uvedené, že hodnoty CUSUM štatistiky $\tilde{X}_{s,e}^b$ boli spočítané pre VŠETKY hodnoty (each) $b \in [s, e]$. Nemá autor skôr na mysli všetky hodnoty $b \in \{s, s+1, \ldots, e\}$, kde $s, e \in \{1, \ldots, N\}$?

- ❏ Na str.11 je uvedené, že σ značí rozptyl štandardného normálneho šumu. Naozaj sa jedná o rozptyl, alebo o smerodatnú chybu?

- ❏ Na str.13 je v súvislosti s wild segmentáciou uvedený algoritmus, kde sa píše, že počiatočné aj koncové body 'intervalu $[s_m, e_m]$' boli vybrané náhodne, nezávislé a rovnomerne z množiny $\{1, \ldots, T\}$. Ak bol výber počiatočného bodu s_m a koncového bodu e_m nezávislý a z tej istej množiny, ako autor garantoval, že $s_m < e_m$?

Praha, 11.06.2024

Doc. RNDr. Matúš Maciak, Ph.D.
maciak@karlin.mff.cuni.cz