## Posudek oponenta diplomové práce

**Název:** Hlavní komponenty
**Autor:** Anna Zavadilová

### Shrnutí obsahu práce

Autorka v práci poskytuje prehľad literatúry zameranej na voľbu optimálneho počtu hlavných komponent v analýze viacrozmerných dát. V prvej časti popisuje jednoduché heuristické metódy, neskôr odvodzuje (zrejme) originálnu metódu založenú na testoch hypotéz o vlastných číslach náhodných matíc, a nakoniec poskytuje prehľad ďalších zložitejších metód z literatúry.

### Celkové hodnocení práce

**Téma práce.** Téma práce je zaujímavá a vhodná, ale tiež veľmi široká. Autorka spracovala informácie z množstva odbornej literatúry. Zadanie práce sa, s výhradami, dá považovať za splnené.

**Vlastní příspěvek.** Text je sprevádzaný radou jednoduchých numerických ilustrácií a krátkych simulačných štúdií, ktoré vhodne doplňujú výklad. Za vlastný príspevok sa dá zrejme považovať najmä kapitola 4, v ktorej sú navrhnuté testy hypotéz o vlastných číslach Wishartovej matice. Tieto testy sú použité pre voľbu počtu hlavných komponent v prípade, keď sú pôvodné náhodné veličiny takmer nekorelované. Táto časť práce sa však javí ako problematická (otázky 1–3).

**Matematická úroveň.** Text obsahuje iba minimum rigorózne odvodených tvrdení. Typicky sa jedná o krátke dôkazy najjednoduchších vzťahov.

**Práce se zdroji.** Vzhľadom k tomu, že väčšina textu je kompiláciou výsledkov z literatúry, práca obsahuje veľké množstvo referencií. Zdroje sa zdajú byť citované prevažne správne. Niektoré časti práce však pôsobia ako doslovný preklad príslušných pasáží z literatúry (pripomienka 4).

**Formální úprava.** Úprava je uspokojivá. Text ale obsahuje radu drobných chýb, preklepov, a nekonzistencií.

### Připomínky a otázky

1. Verzia nového testu popísaná v sekcii 4.2 je založená na "podozrení"(str. 38), že aj pre iné vlastné čísla ako $\hat{\lambda}_1$ platí vzťah obdobný Vete 10. Toto tvrdenie je podložené výlučne obrázkom 3.7 a jedinou simuláciou. Neodporovalo by ale takéto tvrdenie Vete 11 o rozdelení najmenšieho vlastného čísla?

2. Vlastné čísla náhodnej matice sú určite závislé. Akým spôsobom je tento fakt zohľadnený v analýze navrhnutej v sekcii 4.2? Dokážeme niečo povedať o hladine navrhnutého testu?

3. Čo presne je mienené hypotézou $H_0 : V_1 = V_2$ (pre $p = 2$), pre $V_1$ a $V_2$ náhodné veličiny v sekcii 4.4? Ako presne súvisí model broken stick a test rovnosti vlastných čísel?

4. Celá sekcia 3.2 sa zdá byť z veľkej časti preložená z Mardia a kol. (1979, sekcie 8.3.1 a 8.3.2).

5. Kapitola 5 pôsobí nedokončene. V časti 5.1 sa autorka obmedzuje na algoritmický popis metódy, bez uvedenia hlbšej motivácie alebo zaujímavejších komentárov. V metóde popísanej v časti 5.2 je v skratke uvedený iba prvý krok postupu.

6. Interpretácia biplotu v sekcii 2.2 sa nezdá byť dostatočná. Aký má zmysel odvodenie s parametrom $\alpha$, ak v ďalšom texte vždy pracujeme iba s $\alpha = 1$? Ako sa dajú slovne interpretovať skóry a záťaže hlavných komponent? Čo reprezentujú čísla na osiach biplotu? Kedy a prečo je vhodnejšia analýza hlavných komponent na základe variančnej matice, a kedy na základe korelačnej matice?

7. str. 30: ako z Vety 8 plynie asymptotická nezávislosť výberových vlastných čísiel? Ako plynie nestrannosť odhadov vlastných vektorov matice $\mathbb{U}$?

8. str. 36, Tvrdenie 9: ak je matica $\mathbb{X}$ tvorená centrovaným náhodným výberom, jej riadky sú závislé. Je teda Tvrdenie 9 a jeho dôkaz v poriadku?

9. str. 13, posledná rovnica: tvrdenie o Mahalanobisovej vzdialenosti nie je dokázané. Skutočne platí aj pre aproximácie $\tilde{\boldsymbol{g}}_j$ vektorov $\boldsymbol{g}_j$?

10. str. 41: odkaz na dôkaz Vety 11 je uvedený iba pre $p$ párne (sudé). Platí teda Veta 11 tak ako je uvedená?

Nasledujú niektoré ďalšie, menej závažné pripomienky a pozorovania.

1. str. iii: hlavička by mala obsahovať popisky *Název práce* atď.

2. str. 5, Definice 1: musíme predpokladať $n \geq p$.

3. Na str. 6, r. -3 má byť $\boldsymbol{S\Gamma}$ namiesto $\boldsymbol{SG\Gamma}$; v (2.6) na str. 12 sa jedná o riadky, nie stĺpce matíc $\mathbb{G}$ a $\mathbb{H}$; obr. 2.1 neobsahuje červenú šípku ako sa píše na str. 14; v poslednom prvku tabuľky 2.1 chýba des. čiarka; prvému odstavcu sekcie 2.4.4 nie je rozumieť; vo formuli (3.2) je nesprávne $\exp^{-x_i/2}$; na obr. 3.4 nie sú čiary dobre viditeľné; na str. 33 hore určite nie je pre rozdelenie $\boldsymbol{S\Gamma}$ volené $\mu = 1$; rozmery matíc vo vzťahu (5.1) nesúhlasia; nie je jasné či sú matice $\mathbb{W}$ a $\mathbb{H}$ vo vzťahu (5.1) deterministické alebo náhodné; na str. 50, r. -4 nesúhlasí index riadku; str. 50, r. -1 by index $k$ mal ísť asi až do $m$; str. 51, r. 1 ide o prvky diagonály matíc; atď.

4. Reálne čísla sa na str. 6 označujú ako $\mathbb{R}$ a na str. 30 ako $\mathbb{R}$; dimenzia náhodných vektorov je v kapitole 3 označená ako $p$, v kapitole 5 ako $d$ aj $p$ (str. 51); na str. 31 symbol $\approx$ označuje približnú rovnosť, na str. 30 (asi) asymptotickú ekvivalenciu; v tabuľkách v prílohe je výberový priemer značený najprv ako $\bar{x}$, neskôr ako $\mu$; na vzorce (3.11), (5.9), (5.12), (5.13), a ďalšie nie je v texte odkazované; atď.

5. Diskusia o generovaní náhodných čísel v sekcii 2.4.4, ani výsledky o aproximácii rozdelenia $\mathcal{TW}_1$ (sekcia 3.3) sa nikde v ďalšom texte nepoužívajú.

6. Ako vzťah (5.13) závisí na $k$? Má byť hustota na pravej strane skutočne podmienená maticou $\mathbb{H}$?

7. Kniha Mardia a kol. je z roku 1979, nie z roku 2003.

Závěr

Práca pôsobí pomerne nesúrodým dojmom. Kapitoly 1, 2 a 5 sú kompilačné, v kapitolách 3 a 4 vidíme pokus o odvodenie novej metódy. Celkove, obe časti by bolo možné výrazne zlepšiť a doplniť. Napriek tomu, prácu nepovažujem za zlú. Po presvedčivej obhajobe ju bude možné odporučiť uznať ako diplomovú prácu.

Stanislav Nagy
KPMS MFF UK
14. augusta 2018