Oponentský posudek na diplomovou práci

Bc. Radek Solnický: Metody statistické inference založené na matici vzdáleností

Předložená práce se zabývá statistickými metodami, z nichž mnohé jsou v současnosti značně populární ve společenských vědách (psychologie, sociologie), přičemž popisované postupy jsou v práci ilustrovány na datech z oblasti ekologie, což je další významná aplikační oblast pro metody založené na matici vzdáleností. Téma práce je tedy poměrně aktuální a jeho zpracování uchazečem je zcela jistě přínosné pro jeho další profesionální život.

Práce začíná úvodem, v kterém je nastíněna motivace pro statistiku založenou na matici vzdáleností. V druhé kapitole je popsána celá řada koeficientů nepodobnosti, resp. možností, jak měřit vzdálenost v p-rozměrném eukleidovském prostoru a tudíž vytvářet matici vzdáleností, jež tvoří primární vstup pro dále popisované metody. Následně jsou vysvětleny principy vícerozměrného škálování (MDS) (autorem nazývaného podle mne ne úplně hezky jako metoda hlavních koordinát), na kterou lze v jistém smyslu pohlížet jako na inverzní operaci k převodu původně pozorovaných dat na matici vzdáleností. Ve třetí kapitole jsou na testu o nulovosti korelačního koeficientu vysvětleny principy permutačních testů, jichž je dále v práci hojně využíváno k vlastnímu testování a výpočtu P-hodnot. Následně se autor věnuje Mantelovu testu nezávislosti, který je v jistém smyslu vystavěn právě na testu o nulovosti korelačního koeficientu. Čtvrtá kapitola se poté zabývá zobecněním klasických vícevýběrových testů na situaci, kdy vstupem jsou matice vzdáleností. Poměrně značná část diplomové práce (přibližně jedna třetina) je věnována prezentaci výsledků několika simulačních studií v kapitole 5, přičemž nemalé úsilí je věnováno odlišení situací, kdy vícevýběrový test zamítá shodu rozdělení kvůli změně v posunutí od situací, kdy test zamítá shodu rozdělení kvůli změně měřítka.

Svou povahou je práce spíše kompilačního charakteru, což však s ohledem na fakt, že popisované metody nejsou obsahem žádného standardně vyučovaného kurzu v rámci bakalářského ani magisterského studia na MFF UK, není ani v nejmenším na závadu. Přímý vlastní přínos autora, v přiměřené kvalitě, nalezneme zcela jistě v aplikacích popisovaných metod na reálná data získaná v rámci výzkumného záměru MŠMT a v simulačních studiích uvedených v rámci páté kapitoly.

Práce je logicky dobře sestavena a může posloužit zájemci o seznámení se s metodami založenými na matici vzdáleností jako poměrně slušný odrazový text pro další studium. Závažné chyby se v práci nevyskytují. Za poměrně významný nedostatek však považuji styl, jakým je práce napsána a kvůli kterému nelze práci považovat za lepší než průměrnou. Většina textu je napsána bez významnějšího členění a nebýt vyskytujících se vzorců, čtenář spíše nabývá dojmu, že se jedná o román a nikoliv o diplomovou práci na studijním programu matematika. Je velice obtížné odlišit, kde končí popis testu, kde začíná ilustrace na datech atp. Obdobně je dosti obtížné zjistit, kde končí tvrzení o nějaké vlastnosti a kde začínají úvahy, které vysvětlují, proč dané tvrzení platí. Na nemálo místech jsou navíc jednotlivé postupy prezentovány až příliš algoritmickým způsobem (alá kuchařka) bez uvedení důkazů, resp. alespoň motivačních úvah pro dané kroky (např. popis vícerozměrného škálování na str. 8). Jisté nepřesnosti, resp. nedůslednosti lze nalézt též v některých čistě matematických formulacích (např. v definici 4.2 si čtenář musí domyslet, že F_U a F_V označují distribuční funkce příslušných náhodných veličin).

Typograficky je práce na slušné úrovni. V práci se sice vyskytuje jisté množství překlepů („ddruhy“ na str. 16, „původnách“ na str. 20, „Obrázek 5.6 je ilustruje“ na str. 43, „obsahuyjících“ na str. 59), nicméně s ohledem na povahu a rozsah práce je jejich počet na přijatelné úrovni. Z jazykového hlediska bych nicméně práci vytknul používání anglických výrazů i na místech, na kterých to podle mého názoru není vůbec nutné (Chord distance na str. 5, χ^2 distance, Hellinger distance na str. 6, Canberra metric, Manhattan distance na str. 7)

Práci **lze** nepochybně uznat jako práci diplomovou pro studijní obor Pravděpodobnost, matematická statistika a ekonometrie na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze a **doporučuji** ji k obhajobě.

V Praze dne 15. srpna 2011

RNDr. Arnošt Komárek, Ph.D.
oponent diplomové práce

Výběr konkrétních připomínek a dotazů

1. Prosím o důkaz vztahu $\Delta_{EU}(\mathbf{u}_i, \mathbf{u}_j) = \Delta(\mathbf{x}_i, \mathbf{x}_j)$ ze str. 9.

2. Prosím o vysvětlení, proč lze koeficienty R_1, resp. R_2 interpretovat jako podíl variability zachycený reprezentací (viz str. 10 nahoře)?

3. Na str. 38 nahoře se tvrdí, že se simuluje z normálního rozdělení proto, že to umožňuje od sebe oddělit parametr polohy a parametr měřítka. S jinými rozděleními není možné od sebe oddělit parametr polohy a měřítka?