Posudek dizertační práce

Název dizertační práce: Normality Test of the Gene Expression Data

Autor: Mgr. Bobosharif Shokirov

Úvodní kapitola 1 postupně přechází od vysvětlení pojmu genová exprese a souvisejících biologických procesů a poznatků k vysvětlení některých statistických postupů, které byly v poslední době navrženy pro analýzu dat genové exprese (kap. 1.4 a 1.5). Zejména jde o postupy rozvíjené prof. Klebanovem, které budou využity v následujících kapitolách.

V kapitole 2 autor shrnuje charakterizační věty pro normální rozdělení, které jsou převzaty z literatury. V diskusi k tvrzení, které dokázal Sakata (1977b) pro $k \geq 3$, autor podrobně popisuje příklad pro $k = 2$. Tím tedy vyvrací, že by Sakatovo tvrzení platilo pro $k = 2$. Tyto úvahy lze chápat jako přípravu na originální postupy v kapitole 3.

Kapitola 3 pojednává o testování normality pomocí testu uniformity (rovnoměrného rozdělení) transformovaných dat na sféře. Autor definuje svůj nový test v kapitole 3.2.2, byť zde není formálně označen jako definice. Základní myšlenkou je nahradit test normality pozorovaných dat ekvivalentním postupem, který stojí na teoretických výsledcích pro $\Re$-vzdálenosti a který lze provést na vhodně transformovaná data. B. Shokirov také navrhl algoritmus pro výpočet nově navrženého testu. Na str. 34 pak uvádí, že sice publikoval další výsledky k testu sférické uniformity, ale nezahrnul je do předložené dizertace (proč?). U ilustrací v Tabulce 3.1 a 3.2 není uvedeno, který ze dvou uvažovaných datových souborů (HYPERDIP, TEL) byl použit.

V kapitole 4 dizertace směřuje k jinému testu normality. Nejprve autor formuluje a dokazuje originální teoretické výsledky, zejména charakterizační věty pro normální rozdělení. Navržený test normality pak vychází z těchto tvrzení. Důležitým autorovým výsledkem je Věta 4.2.2. pro rozdělení testové statistiky U pro náhodný výběr $X_1, \ldots, X_n$ (jednorozměrných) náhodných veličin, které pocházejí z normálního rozdělení.

Samotný test normality je pak navržen v kapitole 4.3. Pozorované hodnoty veličin

$$X_{1,2j}, \ldots, X_{n,2j} \quad \text{a} \quad X_{1,2j-1}, \ldots, X_{n,2j-1} \tag{1}$$

jsou nejprve nahrazeny jedinou veličinou

$$X_{1,2j} - X_{1,2j-1}, \ldots, X_{n,2j} - X_{n,2j-1}. \tag{2}$$

Tyto tzv. δ-posloupnosti, přestože jsou obecným nástrojem, byly použity právě v kontextu genových expresí prof. Klebanovem a dalšími. Mgr. Shokirov je sice popsal v kap. 1.4 a 1.5, ale očekával bych, že zmíní i jejich nevýhody, a to vzhledem k tomu, že výsledek výpočtu pak záleží na počátečním uspořádání proměnných. Další krok testu slučuje sousední pozorování do skupin po devíti. Tento krok není dostatečně zdůvodněn, přestože jednotlivá pozorování $i = 1, \ldots, n$ mohou být v datové matici uspořádaná zcela náhodně. Následně autorův nově navržený test vyžaduje náhodně vybrat určitý počet s jednotlivých devític a pro ně spočítat hodnoty U. Ty pak vstupují jako pozorování do Kolmogorovovy-Smirnovovy testové statistiky.

Celkově by test v kap. 4 zasloužil podrobnější diskusi. Provedení testu vzbuzuje některé otázky, například jak v praxi zvolit hodnotu s nebo jaký je vliv samotných náhodných výběrů na výsledek. Sám B. Shokirov přitom v příkladech uznává, že parametr s je klíčový pro výsledek testu. Dizertace se nevyjadřuje ani k síle navrženého testu. Zřejmě by však bylo velmi obtížné ji studovat teoreticky.

Autor ilustroval chování nového testu na datových souborech s reálnými daty (HYPERDIP, TEL). Jejich analýzu uzavírá s tím závěrem, že pomocí nově navrženého testu nedokáže posoudit, zda data mají nebo nemají normální rozdělení. Taková analýza ale působí nedokončeným dojmem. Autor mohl pro studovaná data prezentovat aspoň základní popisné statistiky a pro porovnání i jiné existující testy jednorozměrné normality. Kromě toho je škoda, že neilustruje nový test pomocí simulací.

Pokud jde o formulaci nulové a alternativní hypotézy, pak kapitola 1.2 uvádí, že cílem práce je test předpokladu normálního rozdělení každé z uvažovaných náhodných veličin (tj. jednorozměrný test), nikoli test mnohorozměrné normality. Přestože to pak další kapitoly opomíjejí zopakovat, domnívám se, že celou předloženou práci je třeba vnímat jako příspěvek k dlouho otevřenému problému, jak testovat společnou normalitu pro p náhodných veličin v situaci, kdy každá z nich má **odlišné** parametry, tj. svou vlastní střední hodnotu a rozptyl. Toto mělo být zdůrazněno v úvodní kapitole 1, například i spolu s diskusí, jak je předpoklad normality potřebný či vhodný pro různé klasifikační metody, které se pro data o genové expresi běžně používají.

V duchu této interpretace pak vnímám samotný test z kapitoly 4 hlavně jako ilustraci složitých teoretických výsledků, které se týkají zejména charakterizace normálního rozdělení. Tyto teoretické výsledky z kapitoly 3 a 4 považuji za hlavní přínos dizertace, kterým Mgr. Shokirov prokázal své předpoklady k samostatné tvořivé práci.

Pokud se autorovi podaří doplnit podstatné nejasnosti a srozumitelně vysvětlit motivaci pro nové postupy, pak práce splní očekávané požadavky a bude možné předloženou práci uznat jako dizertační práci na MFF UK.

RNDr. Jan Kalina, Ph.D.
Ústav informatiky AV ČR, v.v.i.
Odd. medicínské informatiky a biostatistiky
Pod Vodárenskou věží 2
182 07 Praha 8
kalina@cs.cas.cz

25.8. 2015