Posudek dizertační práce

Název dizertační práce:
Multidimensional statistics and applications to study genes

Autor: Mgr. Peter Bubelíny

Předložená dizertační práce pojednává o statistickém testování hypotéz v různých situacích. Cíl práce je uveden jako studium různých statistických problémů, které se buď přímo týkají anebo souvisí s analýzou molekulárně genetických experimentů za pomoci technologie mikročipů. Práce je strukturovaná jako souhrn autonomních kapitol. Konkrétními tématy je testování hypotéz pro jednorozměrná data (kapitola 2), dvouvýběrový problém pro mnohorozměrná data (kapitola 6) a test nezávislosti dvou proměnných (kapitola 7). Mnohonásobnému testování hypotéz pro jednotlivé proměnné se pak věnují kapitoly 3, 4 a 5.

Kapitola 1 představuje úvod do celé dizertace a spolu s Předmluvou stručně pojednává o vybraných problémech při analýze mikročipů. Vzhledem k tomu, že některé kapitoly souvisí s tématem jen volně, bych zde očekával, že autor lépe popíše, jak jednotlivé kapitoly přispívají k hlavnímu tématu práce. V kapitole 1 jsou představeny i reálné datové soubory, které jsou pak v celé práci používány pro ilustrování jednotlivých pojmů a metod. Kap. 1.6 přitom upozorňuje, že cílem práce není řešení aplikovaných problémů zpracování reálných dat genové exprese.

V kapitole 2 je studován Kolmogorovův-Smirnovův (K-S) test pro jednorozměrná data. Je zde ukázáno, že jednovýběrový i dvouvýběrový K-S test nemusí být obecně nestranný. Dále je zde navržena modifikace jednovýběrového K-S testu, která je nestranná.

Kapitola 3 využívá myšlenku prof. Klebanova (2007) transformovat data tak, aby se získala posloupnost nových náhodných veličin, které by byly porovnatelné mezi jednotlivými mikročipy i aby měly blíže k předpokladu nezávislosti. Kapitola se zabývá problémem závislosti p-hodnot v případě, kdy se provádí mnohonásobné testování. Autor vyzkoušel různé možnosti, které dokážou tento problém řešit. Mezi ně patří i myšlenka nahradit množinu uvažovaných genů tak, že se nejprve uspořádají (např. podle svojí variability) a následně se uvažuje jen posloupnost podílů vždy mezi dvěma následujícími geny. Tak se vytvoří méně korelované veličiny, i když je problém je interpretovat z genetického hlediska. Přestože se v kap. 3 hovoří o normalizaci, až kap. 5 podrobně vysvětluje, co se normalizací rozumí.

Kapitola 4 shrnuje obecné známé postupy pro mnohonásobné testování, podrobně dokazuje jejich známé vlastnosti a porovnává je na reálných i simulovaných datech. Jde spíš o ilustraci známých postupů, které již byly v literatuře porovnány (např. Dziuda (2010), Hastie & Tibshirani (2009)).

Kapitola 5 se věnuje normalizaci dat, což je téma, kterému se při analýze genetických dat věnuje značná pozornost. Kapitola vychází z pojmu δ-sekvence (Klebanov, 2007). Autor na simulovaných i reálných datech ověřil, že vhodné normalizace odstraňují problém se závislostí p-hodnot. Zde navrhl i vhodnou interpretaci takových postupů. Ukázal však, a to i pro klasickou kvantilovou normalizaci, že p-hodnoty jsou normalizací ovlivněny způsobem, který neodpovídá intuici. U výroku *The most known normalizations are quantile normalization*

and global normalization schází citace. Tento výrok je i v rozporu s knihou Rueda (2014), kap. 1.5.2.

Kapitola 6 pojednává o testování složené hypotézy, kdy je žádoucí porovnat několik genů najednou mezi dvěma skupinami. Autor poukázal na překvapivé chování Hotellingovy T^2 statistiky pro dva nezávislé normálně rozdělené výběry, kdy jednotlivé proměnné v první (i druhé) skupině jsou silně korelovány (pro speciální předpoklady). Ukázal, že i v této situaci je T^2 test vhodný.

Kapitola 7 se věnuje testu nezávislosti dvou náhodných výběrů. Na simulacích studuje chování permutačního testu, který je založen na porovnání charakteristických funkcí. Vysvětluje také některé nedávno navržené přístupy k modelování závislosti (*A-dependence, hidden regulator dependence*). Zde chybí motivace, proč je v genetice důležité testovat nezávislost genů. Nakonec v kapitole 8 autor shrnuje hlavní myšlenky své práce.

Společným jmenovatelem kapitol 3, 4 a 5 je mnohonásobné testování hypotéz. To má za cíl nalezení diferenciálně exprimovaných genů. Mnohonásobnému testování se věnuje v matematické statistice velká pozornost a autor vychází z monografie Dudoit & van der Laan (2008). Autor se zdá být dobře obeznámen zejména s přístupem a pracemi prof. Klebanova, ale neuvádí přitom širší kontext. V genetice se často testují hypotézy o jednotlivých genech ne s cílem přímo zjistit p-hodnotu testu pro jednotlivý gen, ale spíš s cílem uspořádat geny podle jejich schopnosti diskriminovat mezi skupinami. Tvrzení, že t test patří mezi nejčastější metody v daném kontextu v genetice (str. 4), je uvedeno bez citace. Dále autor neuvádí obecné výhrady k mnohonásobnému testování jednorozměrných veličin. Např. Dziuda (2010) uvádí, že tato analýza může sloužit pouze jako pomůcka pro seznámení se s daty spíše než k zodpovězení seriózních vědeckých otázek. Dizertace neuvádí alternativní postupy pro selekci nejvíce diskriminativních genů. Ty mohou být založeny na regularizovaných testových statistikách anebo na iterativních postupech (např. Auffarth et al., 2010).

Práci lze vytknout méně obratný způsob vyjadřování, některé partie neodpovídají strukturovanému matematickému textu, práce by zasloužila lepší srozumitelnost. Jednotlivým kapitolám by prospěl podrobnější úvod i uceleně formulované závěry, a to vzhledem k tomu, že každá kapitola pojednává o poměrně samostatném tématu.

Některé další připomínky:

- Str. 35: problematické značení, např. $H_i' : G_i^H = G_i^H$.
- Str. 70: Pojmem *irregular matrix* se zřejmě myslí *singular matrix*.
- Výsledky některých výpočtů jsou poněkud stručně komentovány. U analýzy dat nejsou někde srozumitelně shrnuty závěry, nebo někde autor formuluje obecné závěry, aniž si uvědomuje, že dané vlastnosti pozoroval pouze na svých datech a nemusejí mít obecnou platnost.
- Není jasné, proč si autor vybral pro podrobnou analýzu zrovna datové soubory HYPERDIP a TEL.
- Z jazykového hlediska obsahuje anglický text některé gramatické chyby, ale jen v menším počtu.

Peter Bubelíny vypracoval práci na důležité téma. Dosáhl nových teoretických výsledků v kapitolách 2 a 6. V aplikovaných kapitolách vynaložil velké úsilí při

analýze reálných genetických dat či v simulacích. Zde uceleně aplikoval nově navržené metody jiných autorů nebo i některé známé pojmy, aby je mezi sebou porovnal či ilustroval jejich chování. Kladně hodnotím, že autor publikoval své články samostatně. Tím prokázal, že je schopen samostatné vědecké práce. Doporučuji přijmout předloženou práci jako dizertační práci.

Literatura

[1] Auffarth B., López M., Cerquides J. (2010): Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images. In: Advances in Data Mining, Applications and Theoretical Aspects. *Lecture Notes in Computer Science* 6171, Springer: Berlin, 248 – 262.

[2] Dziuda D.M. (2010): *Data mining for genomics and proteomics: Analysis of gene and protein expression data.* Wiley, New York.

[3] Hastie T., Tibshirani R., Friedman J. (2009): *The elements of statistical learning.* 2nd ed. Springer, New York.

[4] Rueda L. (2014): *Microarray image and data analysis. Theory and practice.* CRC Press, Boca Raton.

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

14.4.2014