Posudek oponenta na disertační práci Petra Bubelínyho

Multidimensional statistics and applications to study genes

Celkový přehled:

Disertační práce je věnována statistickým metodám použitelným ke studiu micro-array dat, která se vyznačují mimořádně velkou dimenzí při neúměrně malém počtu pozorování, navíc s vysokou závislostí. Této problematice je v literatuře věnována značná pozornost, jak se strany lékařů a biologů, tak statistiků a informatiků. Interakce těchto skupin je velmi důležitá, protože biologové někdy model zjednodušují a zacházejí s ním jako s modelem plné hodnosti, zatímco matematikové si někdy neuvědomují závažnost hledaných závěrů a cílů. Exaktní metody lze v této problematice těžko použít. Nejprve je třeba analyzovat kvalitu dat a provést potřebnou filtraci; nato lze aplikovat metody klasifikace a klastrování, s využitím různé normalizace dat. Data se dále často zpracovávají pomocí statistických metod učení a rozhodovacích stromů. Problematika řešená v práci je tedy velmi obtížná pro první rozsáhlou samostatnou práci doktoranda, navíc bez jeho předchozího pobytu na biostatistickém pracovišti.

Doktorand se zaměřil hlavně na studium statistických testů mnohonásobného srovnávání a nezávislosti, použitelných v těchto modelech. Jednu kapitolu věnuje srovnání různých typů normalizace dat, a další studiu vlastností některých neparametrických jedno- a dvouvýběrových testů v jednorozměrném modelu. Práce obsahuje dílčí teoretické výsledky, ale většinou vychází z prací školitele a jeho spoluautorů. Doktorand některé teoretické výsledky upřesnil a dále specifikoval, ale hlavně se zaměřil na jejich ilustraci a ověřování na simulovaných datech a na skutečných datových souborech HYPERDIP a TEL o dětské leukemii. Proti původní předběžné verzi, se kterou jsem se seznámila, je nová verze častěji doplněna o heuristickou motivaci některých kroků a postupů. Autorovi by prospělo srovnání jeho postupů s metodami rozvíjenými na jiných pracovištích, ale neměl možnost taková pracoviště navštívit.

Celé práci předchází rozsáhlý úvod o povaze micro-array a genetických dat převzatý z odborné literatury. Následuje kapitola o jednorozměrných neparametrických testech (zde N-test a Kolmogorov-Smirnovův test). Autor přidává své samostatné výsledky o nestrannosti Kolmogorova-Smirnovova testu proti některým oboustranným alternativám. Jako hlavní charakteristiky testů se v celé práci využívají p-hodnoty; jedna kapitola je věnována testům o závislosti p-hodnot dílčích testů o jednotlivých genech. Protože se většinou uvažují testy o jednotlivých složkách, je další kapitola

věnována mnohonásobným testům a jejich kombinování pomocí Bonferroniho nerovnosti a dalších metod, a jednotlivé postupy jsou numericky porovnány. Další kapitola se zabývá testováním genových skupin; zde se numericky srovnává Hotellingův test, případně t-test, s N-testem a testem založeným na Mahalanobisově vzdálenosti. V kapitole 7 se uvažuje test nezávislosti založený na empirických charakteristických funkcích, který je porovnán s Csörgöho testem. Výsledky a závěry jsou přehledně shrnuty v kapitole 8.

Práce je dvouoborová a tedy náročná jak na praktické znalosti statistiky, tak znalosti problematiky genetických dat. Proti původní předběžné verzi, kterou jsem dříve s autorem konzultovala, dosáhla práce značného zlepšení. Autor se již nevyhýbal samostatným motivacím v úvodech k jednotlivým kapitolám, jakož i interpretacím výsledků. Dobře porozuměl idejím jednotlivých postupů, které pak ilustroval na simulovaných i skutečných datech. Spíše než na matematických závěrech se tvrzení opírají o numerické simulace a o numerické srovnání jednotlivých testů. Právě řada numerických ilustrací a srovnání je hlavním přínosem celé práce, a doplněna vhodnými motivacemi, citacemi a výklady po konzultaci s odborníky v genetice by se práce mohla publikovat v některém mezioooborovém časopise.

Dílčí přípomínky:

K práci mám některé drobné připomínky, ale proti původní verzi je jich méně:

- Často se píše o odhadech, ale převážně jde jen o odhady na základě simulací.
- str. 3, ř. 4: Jaký význam má alternativa "$F \leq G$ nebo $F \geq G$" pro mnohorozměrná data ?
- str. 4, ř. (-18): Kdy je zaručeno, že "about half of marginal distributions are equally shifted" ?
- str. 16, ř. (-8): Kromě této alternativy je mnoho dalších, proti kterým oboustranný KS test není nestranný.
- str. 19, ř. (-15): Je tato modifikace zamýšlena jen pro specifické rozdělení pravděpodobností G ?
- str. 36, ř. 10: Hypotéza v této formě nedává smysl.
- str. 41, ř. 7: Není jasné, jakou hypotézu vlastně testujeme.
- str. 43: True hypothesis a false hypothesis: Pro čtenáře by snad bylo jasnější valid and non-valid hypothesis. Ale to je jen formalita.
- str. 51: Stále se kontroluje chyba 1. druhu, a co síla ?
- str. 66, ř. 3: Spíš variability než variations.

- str. 85, ř. 3: "It was shown that normalization of gene expressions data makes these data almost uncorrelated," kde to bylo ukázáno?
- str. 89: Jako nejvhodnější se doporučuje N-test, ale ten také není nutně nestranný.

Závěr: Přes tyto připomínky je patrné, že autor věnoval práci značné úsilí, a hlavně svými numerickými studiemi přispěl k rozvoji problematiky. Proto navrhuji, aby práce byla přijata k obhajobě jako doktorská disertační práce na MFF UK.

V Praze dne 20.4.2014

Prof. RNDr Jana Jurečková, DrSc
katedra pravděpodobnosti
a matematické statistiky MFF UK