

Posudek vedoucího diplomové práce

Jaroslav Ševčík: „Směšové modely pravděpodobnostních distribucí“

KPMS MFF UK v Praze

Práce je věnována nesnadným problémům spojeným s teorií i použitím pravděpodobnostních směsí, ať už k modelování složitých distribucí nebo k analýze heterogenity zkoumaného souboru (vlastně k shlukové analýze). Nesnadným problémům proto, že vlastně teprve v průběhu práce na diplomce jsme (oba) objevovali nám neznámé výsledky, a to zejména pokud jde o určení počtu komponent směsi. A přitom nejde o výsledky nové, práce Keribina je 6 let stará. Je ale příznačné, že v množství literatury na toto téma napsané se ty skutečně zásadní výsledky objevují poměrně okrajově, a pokud jsou citovány, tak čistě „oznamovacím“ způsobem, jako by si autoři nedali práci reference vyhledat a přečíst. J. Ševčík si tu práci dal, a navazuje na nedávnou diplomovou práci J. Němečka postoupil dále v porozumění proč a jaké volit penalizační členy, aby se došlo i ke konzistentnímu odhadu počtu komponent. Po přehledu mnohých kritérií (penalizací), která různí autoři vymysleli pro jimi řešení specifické situace, se vždy nakonec ukáže, že standardní BIC je asi nejpraktičtější, jenom o jeho konzistenci v této situaci byly pochyby. Ale přes pokusy s různými kritérii penalizujícími např. vzdálenost distribucí typu Kolmogorova Smirnova nakonec Keribin (1999) a další z francouzské „analytické“ školy prokázali konzistenci i pro BIC a soubor podobných kritérií penalizujících log-věrohodnost. Toto vše je obsahem druhé kapitoly, hlavně závěrečných částí. Z tohoto hlediska je diplomová práce spíše kompilační, ale při prostudování velkého množství literatury autor výsledky i logicky uspořádal a vlastně nás provádí vývojem metod analýzy směšových modelů.

Následující kapitola 3 je zcela věnována výpočetním aspektům analýzy směsí, podává přehled algoritmů, které jsou k dispozici v knihovnách R. Autor samozřejmě vytvořil i řadu vlastních procedur právě proto, aby mohl srovnat procedury standardní s metodami podle nových teoretických výsledků. Toto srovnání pak provádí na oblíbených cvičných datech jednorozměrných směsí. Samozřejmě pro konečné rozsahy dat, takže výsledné porovnání vypovídá opět o citlivosti kritérií na určité specifické situace. Nechybí ani ukázka práce s reálnými daty.

Celkově práce obsahuje bohatý materiál dobře zpracovaný, utříděný, vyzkoušený a okomentovaný, a její součástí jsou i vlastní algoritmy. Soudím, že splňuje nároky kladené na diplomovou práci a navrhuji ji hodnotit známkou výborně.

12. září 2006

doc. Petr Volf, CSc.

