Oponentský posudek diplomové práce

Modely pro data s nadbytečnými nulami

Bc. Dominika Matuly

Autor se v práci zabývá regresními modely s diskrétní (celočíselnou) odezvou s nadměrným výskytem nul. To znamená v situaci, kdy například sledujeme výskyt poruch a připouštíme existenci „perfektního“ stavu, ve kterém k poruchám nedochází. Pak obvykle nelze použít běžná rozdělení (Poissonovo, binomické, negativně binomické), neboť dávají mnohem menší pravděpodobnost nuly, než jaká máme pozorování.

Autor v první kapitole popisuje různé přístupy k modelování dat s nadbytečným počtem nul, dále stručně zavádí zobecněné lineární modely (v práci uvažuje nejenom sledování četností, ale též regresorů) a udává základní vzorce pro odhad parametrů metodou maximální věrohodnosti, případně kvazi-věrohodnosti tam, kde maximálně věrohodný odhad použít nelze. Pro výpočet odhadu navrhuje použít EM algoritmus. Druhá kapitola je věnována obecným modelům bez nulové odezvy, což znamená oddělit odhad pravděpodobnosti nulového stavu a parametrický model pro pravděpodobnosti nenulových stavů. V dalších kapitolách se autor věnuje modelům s nadbytečnými nulami, obecně či pro specifické modely (Poissonovo rozdělení, dvojí parametrizace negativně binomického rozdělení) a nakonec hradbovým modelům. Tyto dva přístupy se liší tím, že v modelu s nadbytečnými nulami uvažujeme směs dvou rozdělení, jednoho degenerovaného v nule a rozdělení na nezáporných celých číslech, v případě hradbového modelu uvažujeme směs rozdělení degenerovaného v nule a rozdělení na kladných celých číslech. V sedmé kapitole je představena simulační studie.

Práce je psaná pečlivě a téměř bez chyb. Je celkem rozsáhlá jak v teoretické, tak i v simulační části. Přesto se nečte úplně snadno, čtenář musí proniknout do množství zkratek a značení. Nicméně po překonání tohoto kroku se ukazuje logická struktura práce, kde je postupně vysvětleno mnoho přístupů k modelování dat s nadbytečnými nulami a mnoho odkazů na literaturu, takže práce může sloužit jako rozsáhlý přehled a odkazová příručka. **Velmi oceňuji snahu o českou terminologii**, která je v celé práci přítomna a podle mého soudu se jedná o promyšlené a vhodně navržené názvosloví.

Na druhou stranu právě rozsáhlost a styl práce mohou být považovány za slabou stránku práce. Můj celkový dojem je ten, že je práce pojatá čistě technicky, modelů je možná až moc, ke každému sice najdeme v práci podrobné vzorečky pro odhady, ale už méně diskuse o výhodách a nevýhodách a srovnání s ostatními. To konečně platí i pro simulace. Všechna simulovaná data vycházejí z jednoho modelu a ostatní modely jsou použité pro odhady. Simulace jsou celkem rozsáhlé, svým provedením však také hodně obdobné. Trochu tak tápu, co si z jejich výsledků mohu a mám odnést za poučení. Toto bych rád probral při obhajobě.

K práci mám několik drobných připomínek

- Ve větě 3 je rozpor mezi body (i) a (iii). Nejspíš jde o překlep, jak zní věta správně?
- Na straně 26 jsou zavedeny tři různé funkce ℓ_c. To mi přijde matoucí, zejména s ohledem na M-krok optimalizace na straně 28, kde rozlišení těchto funkcí se děje jen na základě argumentu (β, nebo γ).
- Jak rychlé jsou do sebe vnořené iterace v EM algoritmu na straně 28?
- Na straně 44 můžeme číst „mírně zmíněný přístup modifikujeme“.
- Na straně 48 je napsáno „rozdělení bez nulové odezvy již obecně nejsou rozděleními exponenciálního typu“. To vyvolává dojem, že rozdělení s nulovou odezvou jsou obecně exponenciálního typu.
- V simulacích vychází překvapivé shody ve výsledcích pro parametry γ u PHM a NBHM modelů. Tyto modely jsou přitom označeny za nedobré, proto je tato úplná shoda dost překvapivá. Dá se nějak vysvětlit?

Přes uvedené drobné výhrady si myslím, že předložená práce splnila očekávání, je pečlivě napsaná a přínosná. Proto práci doporučuji **uznat jako diplomovou práci**.

Daniel Hlubinka
v Praze 29.8.2016