

POSUDOK OPONENTA DIPLOMOVEJ PRÁCE

Název: Řídké regresní modely

Autor: Bc. Samir Bessiso

Predložená diplomová práca sa zaoberá veľmi aktuálnou a relatívne novou problematikou štatistickej inference/učenia, metódami hľadania tzv. málopočetných (“riedkych”) riešení v lineárnych regresných modeloch. Vo svojej podstate je názov práce tak trochu zavádzajúci: prísne vzato, nejde tu o “iné” modely — modely zostávajú nezmenené, také, ako sa používali a používajú v klasickej regresnej analýze — ale ide tu o novú filozofiu pri ich odhadovaní: hľadá sa riešenie za predpokladu, že väčšina regresorov nemá prediktívnu hodnotu, čo sa predikcie odozvy týka — inými slovami, väčšina “skutočných” hodnôt odhadovaných parametrov je nulová. A toto veľmi často v situácii, keď je systém nedourčený: pozorovaní je menej ako regresorov, situácia v literatúre vyznačovaná ako “ $n < p$ ”.

Práca je písaná matematicky sofistikovaným jazykom a formalizmom — možno trochu až príliš; snaha po maximálnej reči formulí miestami ide na úkor čitateľnosti a zrozumiteľnosti. Konkrétne, dosť ma rozladil popis simulačného experimentu a jeho sumarizácia v Obrázku 2.5 — tým viac, že ide o časť práce, ktorá nebola prevzatá z citovanej literatúry, ale je očividne samostatným prínosom autora. Ako vidno zo všetkých mojich otázok, ktoré chcem aby boli položené autorovi pri obhajobe (viď nižšie), popis nielenže ponecháva veľa nevy povedaného, ale celková organizácia vo forme farebných kódov je málo intuitívna a kladie na čitateľa, ak tento chce na ich základe dospieť aj k nejakému samostatnému úsudku (nielen tomu, ktorý mu podsúva autor) dosť veľké nároky. Inak povedané, nejde tu len o potenciálne nejasnosti, no celej prezentácii simulačného experimentu by prospelo, ak by bola poňatá úplne inak.

Na druhej strane, autor pomerne na malej ploche podáva formálne správny prehľad pomerne veľkého počtu výsledkov z danej problematiky, a tiež zopár zaujímavých ilustratívnych obrázkov. Možno by opäť nebolo na škodu, keby sa štatistickému významu Obrázkov 1.2 a 3.1 venovala väčšia pozornosť — a tiež prípadne sa uviedli príslušné dôkazy (ktoré nie sú komplikované, a predstavujú cenný vhlád do problematiky).

Práca je vypracovaná starostlivo, formálne chyby v matematike ani citovanej literatúre sa mi nepodarilo nájsť, grafická kultúra je na úrovni. Jediné drobné prehrešky sú jazykového charakteru:

- veľké/malé písmená — str. 29, riadok 13 zhora “náhodnou”, riadky 8 a 12 “Gibbsov”, “Gibbsova”;
- čeština: Bayesov odhad (veľké B) je správne, ale bayesovská štatistika má byť malé b.

Záver: práca je dosť dobrá, a tým pádom ju samozrejme odporúčam k obhajobe — ale zdráham sa ju označiť ako vynikajúcu, vzhľadom na nedostatky, ktoré spomínam vyššie.

OTÁZKY KU OBHAJOBE

Na str. 28 sa píše, že pre každú trojicu (p, n, σ) boli 200-krát vygenerované dáta: ako konkrétne, čo sa β^* týka? To isté β^* 200-krát? Ak áno, tak ako bolo zvolené? Ak nie, a β^* boli zvolené po každý krát rôzne, ako?

Ďalej, ak rozumiem správne, p je dimenzia β^* a n je počet pozorovaní. Ako bola regulovaná málopočetnosť (“riedkosť”), t.j. počet nenulových členov v β^* ? (Je známym faktom, že metódy uvažované v práci sú relevantné len vtedy, ak je počet nenulových členov výrazne menší než p .)

Praha, 2.6.2024

IVAN MIZERA
KPMS MFF UK
imizera@me.com