

Oponentský posudek na disertační práci:

TEREZA ŠIMKOVÁ: STATISTICKÁ INFERENCE POMOCÍ L-MOMENTŮ

Hlavním přínosem předložené disertační práce je představení a matematicky korektní odvození statistických metod založených na L-momentech s důrazem na analýzu skutečných hydrometeorologických dat.

Úvodní kapitola představuje L-momenty, jejich základní vlastnosti a dvě jejich robustní zobecnění (TL- a LQ-momenty). Ve druhé kapitole jsou pomocí výběrových L-, TL- i LQ-momentů odvozené odhady parametrů a kvantilů pro GP a GEV rozdělení; vlastnosti těchto odhadů jsou pak v simulační studii porovnané s maximálně věrohodným odhadem. Zajímavá a prakticky přínosná je navazující třetí kapitola, kde autorka pomocí delta metody odvozuje asymptotické konfidenční intervaly pro odhady kvantilů založené na momentech, L-momentech a na metodě maximální věrohodnosti (pouze pro GEV rozdělení); vlastnosti těchto konfidenčních intervalů jsou pak opět porovnané pomocí simulační studie.

Ve druhé i třetí kapitole jsou některé hodnoty parametrů předem vyloučené (typicky se předpokládá, že $k \neq 0$) a není mi jasné, jestli lze tento předpoklad nějak zdůvodnit. Může se například stát, že konfidenční interval obsahuje bod 0, přestože se při jeho konstrukci předpokládá nenulovost odhadovaného parametru?

Další tři kapitoly se věnují vícerozměrným datům. Po zavedení vícerozměrných L-momentů a dalších potřebných pojmů (např. kopule) se autorka ve čtvrté kapitole věnuje testům homogenity pro dvourozměrná data a podrobně vysvětluje také konstrukci dvourozměrných kvantilových křivek založených na kopulích pro roční maximální 1- a 5-ti denní srážkové úhrny pozorované v průběhu 47 let asi na 200 stanicích v 6 regionech ČR. Pátá kapitola popisuje použití testů homogenity i pro třírozměrná data; zde jsou pak popsány i vícerozměrné kopule.

Pro praktické použití navržených testů homogenity může být problematický předpoklad nezávislosti, který pro skutečná data nemusí být splněn. Tento problém je prodiskutován a částečně vyřešen v kapitole 6 pomocí modifikace metody Monte Carlo pro nenulové prostorové korelace. Není zde však vysvětleno, jestli mohou být za nezávislá považována i pozorování v po sobě následujících letech (tj. jestli pozorování nemohou tvořit časovou řadu).

Nejdůležitější vlastní výsledky jsou:

1. odvození konfidenčních intervalů založených na L-momentech a jejich porovnání s jinými metodami ve třetí kapitole,
2. odhady kvantilových křivek a testy homogenity pro vícerozměrná nezávislá i prostorově korelovaná data v kapitolách 4 až 6.

K těmto výsledkům mám několik otázek:

- Jak by se mělo postupovat, pokud není možné předem vyloučit možnost, že parametr k (například ve Větě 3) se rovná nule? Lze tento předpoklad zdůvodnit například při analýze meteorologických dat? Jak se podobné předpoklady projeví při konstrukci konfidenčních intervalů v kapitole 3?
- Jak se na vlastnostech odhadů a konfidenčních intervalů ve druhé a třetí kapitole projeví porušení předpokladu nezávislosti? Lze splnění předpokladu nezávislosti otestovat?
- Postupy ve čtvrté a páté kapitole také zřejmě závisí na předpokladu nezávislosti jednotlivých pozorování. Vliv prostorových korelací se sice rozebírá v kapitole 6, ale není mi jasné, jestli skutečná pozorování nemohou být korelovaná i v čase. Jak by se postupovalo, pokud by data tvořila spíše časovou řadu než náhodný výběr?

Celá práce je napsaná srozumitelnou angličtinou: vytkl bych snad jenom opakované použití termínu ‘parameters estimators’ a některé neobratné formulace, které se objevovaly spíše v úvodních kapitolách (např. str. 33, řádek 1 nebo sekce 2.4, řádek 6). Jinak je předložená disertační práce napsaná velice pečlivě, srozumitelně a s minimem překlepů.

Další připomínky a otázky:

kapitola 3 Proč neobsahuje třetí kapitola konfidenční intervaly i pro odhady založené na LQ- a TL-momentech?

str. 56, poslední řádek Chybná dimenze normálního rozdělení?

str. 58, věta 10 Byly uvedené vzorce spočítané na počítači? Jaká software byl k těmto výpočtům použitý?

str. 73, obr. 3.1 Čím mohou být způsobené špatné vlastnosti konfidenčního intervalu např. pro $k = -0.4$?

str. 104, obr. 4.6 & str. 117, obr. 4.11 Prosím o vysvětlení praktického přínosu nakreslené kvantilové křivky v porovnání s ‘naivní’ křivkou (tj. kvantilovou křivkou, která by vůbec neměla ‘proper part’). Jak dobře je tato kvantilová křivka použitelná v praxi?

str. 105, řádek 7 Co znamená ‘ungaged’?

str. 114, řádek 1 Podle tabulky 4.5 se homogenita zamítá v regionu 1.

str. 169, řádek 3 Pro jaká další data mohou být tyto metody vhodné?

Předložená práce se zabývá statistickými metodami založenými na L-momentech, které se ukazují jako velice užitečné hlavně v meteorologii. Dosažené teoretické výsledky jsou motivované skutečnými problémy: konfidenční intervaly pro kvantily GP a GEV

rozdělení v kapitole 3 a testy homogenity a odhady kvantilových křivek (metoda RFA) v kapitolách 4 až 6. Odvozené metody by v budoucnu mohly být rozšířené i pro časové řady (pokud se časové závislosti v těchto datech skutečně mohou vyskytnout). Vhodné by mohlo být i použití bootstrapu zmíněné na str. 169.

Shrnutí: Dosažené výsledky již byly publikovány ve čtyřech odborných člancích (z toho tři v mezinárodních časopisech s impakt faktorem) a další publikace se připravuje. Předložená disertační práce i kvalitní publikace podle mého názoru jednoznačně prokazují předpoklady autorky k samostatné tvořivé vědecké práci a předloženou práci tedy doporučuji uznat jako práci disertační.



Doc. RNDr. Zdeněk Hlávka, Ph.D.

Univerzita Karlova, MFF

Katedra pravděpodobnosti a mat. statistiky

Sokolovská 83, 186 75 Praha 8

Posudek oponentky na disertační práci

Autorka: Mgr. Tereza Šimková

Název: Statistická inference pomocí L-momentů

Školitel: prof. RNDr. Jan Pícek, CSc.

Předložená disertační práce se zabývá použitím robustních momentů (převážně L-momentů) v modelování pravděpodobnostních rozdělení a závislostí mezi náhodnými veličinami. V souladu s nejčastější aplikací doplňuje teoretické výsledky aplikací na regionální frekvenční analýzu meteorologických dat. Téma na pomezí statistického a matematického modelování je velmi užitečné a aktuální, je vhodné pro obor Matematické modely a jejich aplikace. Umožňuje věnovat se teoretickým základům přístupu, stejně jako řešit velmi složité problémy s aplikací postupů s potenciálním přínosem v praxi.

Statistické úsudky založené na robustních momentech, jako jsou v práci používané L, TL a QL momenty, jsou moderním přístupem, který nabízí možnost vypořádat se s problémy, které má metoda maximální věrohodnosti a na ní založené postupy v případě zešikmených rozdělení s těžkými konci. Náhodné výběry z takových rozdělení obsahují často extrémní nebo odlehlá pozorování. Autorka se nejprve věnuje jednorozměrné analýze, ve druhé části již používá vícerozměrné L-momenty. Autorka aplikuje zkoumané postupy na jednorozměrnou, dvourozměrnou a trojrozměrnou regionální frekvenční analýzu. Obdobné aplikace jsou s úspěchem využívány v meteorologických a klimatologických studiích, proto autorka publikuje, kromě statistických časopisů, také v odborných časopisech věnovaných těmto tématům, jako jsou Water Resource Research a Meteorological Applications.

Práce je poměrně obsáhlá, ale její rozsah odpovídá odbornému obsahu. Je přehledně členěna do šesti částí, obsahuje výklad postupů, simulační ilustrace a aplikaci na reálná data; výpočty jsou provedeny v programu R. Autorka bohužel své cíle formulovala velmi povšechně (strany 15, 16). Takové cíle by splnily práce málo kvalitní i vysoce kvalitní, autorka si mohla předsevzít velmi ambiciózní a konkrétní cíle své disertace. Cíle, které si autorka předsevzala, byly splněny.

Práce obsahuje přehled existujících výsledků (od prvních prací do těch nejnovějších) a obsahuje jednoznačně originální a nové výsledky dosažené autorkou. Autorka dobře a zodpovědně cituje, ale ani v textu není úplně jasné, co je ale přece jen bych ocenila vymezení přínosu autorky, například v závěru práce (též otázka k diskusi 4 níže). Práce obsahuje velké množství referencí, které jsou vhodně vybrány tak, aby obsáhly všechny oblasti, které jsou v práci obsaženy.

V prvních třech kapitolách se autorka zabývá jednorozměrnou analýzou. Snaží se posoudit, kdy je momentová metoda založená na robustních momentech úspěšnější než maximálně věrohodný odhad, jehož optimální vlastnosti v klasickém parametrickém přístupu k odhadům jsou známy a odhady jsou široce implementovány ve statistických programech. Zkoumaná rozdělení (rozdělení extrémních hodnot (GEV) a zobecněné Paretovo rozdělení (GP)) se používají pro modelování pravděpodobnostního rozdělení extrémních hodnot a umožňují odhad vysokých kvantilů rozdělení. Kromě představení základních pojmů (L-momentů, QL-momentů a TL-momentů) jsou odvozeny explicitní vztahy pro tyto momenty a jejich bodové a intervalové odhady. Je také třeba ocenit snahu o optimalizaci volby parametrů v definici QL momentů. Ve třetí kapitole je navíc představena konstrukce intervalů spolehlivosti pro kvantily zkoumaných rozdělení – otázka odhadu hodnot vysokých kvantilů rozdělení je důležitou praktickou úlohou, kterou autorka ve své práci řeší.

Další tři kapitoly jsou věnovány vícerozměrným L-momentům. Autorka představuje problémy spojené se statistickou inferencí o jevech s velmi řídkým výskytem na základě relativně krátkých časových řad (vzhledem k periodicitě jejich výskytu) a s jejich řešením v regionální frekvenční analýze. Použití kopulí je velmi vhodným a obtížným nástrojem pro popis vícerozměrných závislostí. Kapitoly 4 a zvláště pak 5 a 6 považuji za stěžejní výsledky disertační práce. Autorka se kromě teoretických východisek věnuje praktickému popisu dvourozměrných rozdělení (aplikace na maximální roční jednodenní a pětidenní objemy srážek) a trojrozměrných dat (jednodenní, třídní a sedmidenní srážky) pro trojrozměrné závislosti.

Autorka píše jasně a přehledně, výklad může čtenář dobře sledovat. Je třeba ocenit, že autorka našla vhodný kompromis mezi přehledným výkladem a přesností nutnou v odborném matematickém a statistickém textu. Tento cit pro výklad se například jasně projevuje v částech 4-6 při popisu zkoumaných statistických testů. Práce obsahuje velké množství vzorců, které jsou napsány pečlivě a bez chyb, proto se čtenář může na vzorce spolehnout. Grafická úroveň práce je na vysoké úrovni, obrázky výklad dobře doplňují a jasně prezentují výsledky. Jedině bych snad ocenila, aby i v případě matic grafů byly použity větší fonty písma tak, aby byly čitelné popisky os (zvláště v případě, kdy jsou v různých grafech různé).

K práci nemám výhrady, autorka se drží pečlivě zvoleného přístupu k problému. Občas postrádám nějakou interpretaci z hlediska statistického náhledu na problém. Například diskusi o rozděleních v jednorozměrné úloze autorka nijak neinterpretuje význam parametrů, asi i z tohoto důvodu nepohlídala značení (parametry α a σ). Klasické momenty lze interpretovat i vzhledem k existenci L-momentů; některé vzorce nemají smysl pro parametry, které patří do

autorkou uváděného parametrického prostoru. Čtenář by patrně uvítal již při definici rozdělení i další vlastnosti, které jsou pak dále využívány (hustoty pro věrohodnostní funkci, kvantilovou funkci a hlavně obecné i centrované momenty) a uvedeny až při prvním výskytu.

U kopulí je diskutován parametr θ , který například vzorec pro normální případ (nečíslováno str. 91) neobsahuje, obrázek 4.1 již je jasný. K přehlednějším zápisům by pomohlo využívání různých sčítacích indexů, než je i a j , čtenář musí přemýšlet o mezích v konkrétních vzorcích.

V předložené práci je uveden přehled publikační činnosti autorky. Jsou uvedeny pouze příspěvky v časopisech, předpokládám, že autorka také absolvovala konference, kde referovala o výsledcích své práce. Publikované články jsou ve shodě s tématem disertační práce, publikační činnost je velmi kvalitní, autorka publikovala teoretické práce i aplikace zkoumaných metod.

Náměty k diskusi při obhajobě:

1. Můžete uvést další reálné situace, kde by se, podle Vašeho názoru, mohla metodika z Vaší disertační práce používat? Proč se nestala běžně používanou alternativou k metodě maximální věrohodnosti při modelování zešikmených rozdělení s těžkými konci?
2. Jaký je význam parametrů u vámi zkoumaných rozdělení ve 2. kapitole a s čím souvisejí jejich hodnoty z hlediska vzniku? Jaké praktické i teoretické problémy přináší práce s podobnými pravděpodobnostními rozděleními při použití metody maximální věrohodnosti?
3. Metody jsou využívány pro časové řady objemu srážek. Je nějaký rozdíl mezi aplikací zkoumaných postupů na náhodný výběr z rozdělení a na časovou řadu?
4. Můžete specifikovat svůj přínos k problematice? Které konkrétní výsledky své práce považujete za stěžejní?

Disertační práci Mgr. Šimkové považuji za velmi zdařilou, autorka prokázala, že je schopna samostatně vědecky pracovat, řešit konkrétní problém na vysoké teoretické, aplikační, výpočetní i interpretační úrovni. Autorka s přehledem a velmi kompetentně formuluje problémy, odvozuje teoretické vztahy, pracuje s reálnými daty a provádí simulace a ukazuje tak komplexní schopnosti a kompetence v práci s daty a modely. Předložený text splňuje všechny předpoklady kladené na disertační práce, a proto disertační práci Mgr. Šimkové doporučuji k obhajobě před komisí pro obhajoby disertačních prací.

V Praze, 12.4.2019

doc. RNDr. Ivana Malá, CSc.

Ivana Malá