Posudek oponenta diplomové práce

Adéla Jalovcová: Prostorová epidemiologie

Předložená práce je věnovaná analýze epidemiologických dat s využitím metod prostorové statistiky. Data jsou uvažovaná ve formě počtů případů agregovaných za určité územní celky v daném časovém období. Konkrétně jsou zpracována data o počtech pozitivně testovaných na COVID-19 podle okresů ČR v jednotlivých dne po dobu 89 týdnů. Autorka nejprve provádí testy významnosti prostorové autokorelace použitím tradičních postupů založených na Moranově a Gearyho indexu. Dále se pak zabývá odhadováním vhodného prostorového modelu. Využívá přitom bayesovský přístup a metodu INLA (integrated nested Laplace approximation). Rovněž zmiňuje rozšíření modelu o zahrnutí dalších kovariát nebo časové složky.

Jde o aplikačně zaměřenou práci, která je svým rozsahem poměrně útlá. Některé části by si zasloužily víc rozvinout. Mělo by se jasněji stanovit, jaké praktické otázky se řeší. Důkladněji by se měly vysvětlit použité výpočetní postupy. Také interpretace a okomentování získaných výstupů mohlo být podrobnější. Samotná práce je přehledně uspořádaná a celkem srozumitelně napsaná. Ovšem někdy nepůsobí úplně uceleným dojmem, trochu mi chybí lepší propojení jednotlivých sekcí. Například metoda INLA je popsaná obecně, pak je představen konkrétní model, ale už není rozebráno, jak INLA funguje pro tento model.

Práce je rozdělena do čtyř kapitol. V první kapitole jsou představena analyzovaná data. Druhá kapitola se zabývá testováním přítomnosti prostorové autokorelace. Hlavní část práce spočívá ve třetí kapitole. Ve čtvrté kapitole je krátce diskutováno zohlednění časového vývoje do analýzy.

Téma práce je aktuální a zajímavé. Bylo zpracováno v souladu se zadáním práce, i když vzhledem k povaze dat nedošlo na otázky spojené se statistikou bodových procesů. Studentka se musela seznámit s modelem latentního gaussovského markovského náhodného pole a nastudovat metodu INLA, kterou použila pro přibližnou bayesovskou inferenci. Hlavním vlastním přínosem autorky je implementace algoritmu a aplikace na reálná data. Práce obsahuje minimum matematických odvození, které neobsahují závažné chyby, ale na některých místech mohly být předpoklady korektněji zformulovány. Použité zdroje jsou citovány správně. Některé nedostatky v citacích, gramatice a ve formální a typografické úpravě zmiňuji níže.

Otázky

1. V první kapitole se uvádí, že se bude pracovat s daty o počtu nakažených. Jsou taková data skutečně dostupná? Nemělo by spíše jít o počty případů s prokázanou nákazou?
2. K jakému datu jsou údaje o počtech obyvatel v jednotlivých okresech?
3. Jestliže M_{obs} je napozorovaná hodnota testové statistiky, tak $\frac{M_{obs}-E_g M}{\sqrt{var_g M}}$ je číslo. Jak pak chápat tvrzení, že má asymptoticky normální rozdělení?
4. Jaké jsou rozměry parametrů vyskytujících se ve vyjádření (3.1)? Co znamená, že gaussovské pole je vektor?
5. Je funkce f z kapitoly 3.2.1 stejná jako v předpisu (3.1)?
6. Jak se v kroku 1 algoritmu z podkapitoly 3.2.2 použije $\tilde{\pi}(\psi|y)$ ke stanovení K-tice bodů? Co znamená vysoký odhad? A jaké K je voleno v aplikaci na data?

7. V podkapitole 3.2.1 je popsána aproximace určitého integrálu. Jak přesně se tato aproximace využije k odhadu hustoty $\pi(\theta_i|\psi, y)$ v kroku 3 algoritmu z podkapitoly 3.2.2? A jak tento odhad závisí na k?

8. Má v bodu 4 v podkapitole 3.2.2 opravdu být $\pi(\theta, \psi|\theta)$? A jakým způsobem se integruje podle vektorů ψ a θ_{-i}?

9. V podkapitole 3.2.2 je dvakrát zmíněná zjednodušená Laplaceova aproximace, a to pokaždé s jiným zdrojem. Je tato zjednodušená aproximace v práci nakonec použita? A pokud ano, tak jakým způsobem?

10. Dalo by se odvodit, proč u v podkapitole 3.3 má zrovna rozdělení $N(0, (I - \alpha W)S^2)$? Odkud se vzala konstanta α?

11. V definici 2 se zavádí aposteriorní prediktivní rozdělení nebo jeho hustota?

12. Je v pořádku, že nulová hypotéza na str. 20 závisí na parametrech odhadnutých z dat?

13. Proč se ve výstupech v podkapitole 3.5 objevuje položka `data.suicides`? Co představuje?

14. Po vzorci (4.1) se zmiňuje člen $Temp_i$, který se v (4.1) nevyskytuje. Podle vyjádření (4.2) by $Temp_t$ mělo záviset také na i. Nemělo by se tedy spíš používat značení $Temp_{it}$?

15. Obrázek 4.2 by si zasloužil detailnější vysvětlení. Není například jasné, co přesně je myšleno trendem nákazy v čase. Dalo by se matematicky zformulovat, co je znázorněno na obou obrazcích? A jak to souvisí s číselnými výstupy uvedenými na str. 32 a 33?

Připomínky

Opomenutí

- V definici 1 chybí složené závorky.
- V tabulce 2.1 chybí týden 25.
- U odkazů na některé číslované rovnice chybí závorky (str. 15, 16, 17, 19).
- V definici p^{erl} na str. 21 mají být tučná R.
- Obrázky 3.7 a 4.2 nejsou v textu práce citovány.

Nevysvětlené symboly

- U definice binárních vah se mluví o sousedech x_i, x_j, které se nikde nevyskytují.
- V definici 1 se objevují symboly L a ω, které nejsou definovány.
- V integrálu na str. 18 nahoře není vysvětleno θ_{-i}.
- Dimenze d vektoru $\mathbf{R}_k$ na str. 21 není zavedena, zřejmě má jít o n.
- Rozdělení *logGamma* a *Poisson* jsou uvedena jen ve vzorci a nejsou vysvětlena v textu.

Jazyková úprava

- velká písmena místo malých (Index – str. 8, Gaussovský – str. 15, Nested – str. 16, Chain – str. 16)
- chybějící tečka na konci věty (str. 15, 16, 18, 23, 25, 31)
- chybějící čárka ve větě (str. 22, 23)
- překlepy (absoluntí – str. 18, kapitol – str. 19, rozvání – str. 30, přírůtků – str. 32, Beság – str. 35)
- anglické výrazy by se dalo nahradit českými ekvivalenty (quasi-Newtonova, goodness of fit)

Typografické nedostatky

- způsob psaní trojtečky
- zápis kalendářních dat (str. 7)
- chybějící mezera za tečkou (str. 8)
- nepoužití matematického fontu pro matematické symboly (str. 11, 19, 28, 30)
- odsazení textu po vzorci (3.6)
- mezera před čárkou (str. 18)
- log není jako matematická funkce (str. 19)
- chybějící mezera před závorkou (str. 24)
- počítačové výstupy nejsou číslovány ani nemají legendu a někdy jsou nevhodně zalomeny (str. 24)

Nekonzistence

- velké nebo malé písmeno při odkazování na kapitolu knihy (str. 2, 20)
- Covid 19 nebo Covid-19
- Monte Carlo nebo Monte-Carlo
- v druhé kapitole konkrétně 77 veličin nebo obecně rozsah n
- $\mathbb{E}$ nebo E pro střední hodnotu
- některé odstavce jsou odsazené, jiné ne
- diferenciály stojatě nebo skloněně
- prvky matice sousednosti jsou ω_{ij} nebo w_{ij}, na str. 18 dokonce i W_{ii} a a_{ii}

- počet územních celků je I na str. 19 oproti n ve zbytku práce
- extreme length rank nebo extreme length measure, ani jednomu neodpovídá zkratka erl

Seznam literatury

- Místo příjmení autorů jen iniciály (Beaglehole a kol., Illian a kol., Komenda a kol.).
- Některá vlastní jména nebo zkratky malými písmeny (bayesian, new york, r-inla, laplace, malawi, gaussian).
- Neúplný nebo chybějící rozsah stránek (Cliff a Ord, Wood).
- Neúplný název knihy (Illian a kol.).
- Chybějící ročník časopisu (Mrkvička a kol., Wood).
- Chybný rok vydání (Mrkvička a kol., Schrödle a Held, Sen, Wood).
- Přebytečné TEXové symboly (Schrödle a Held).

Závěr

Diplomovou práci Adély Jalovcové považuji za podprůměrnou, ale **doporučuji ji uznat jako diplomovou práci na MFF UK**.

V Praze, 4. září 2022

doc. RNDr. Zbyněk Pawlas, Ph.D.
KPMS MFF UK