Oponentský posudek na diplomovou práci

Název: Testy normality časových řad
Autor: David Stibůrek

Shrnutí:

Diplomová práce Davida Stibůrka se zabývá testováním normality závislých dat, zejména dat pocházejících ze stacionárních procesů. Práce začíná popisem čtyř metod vyvinutých různými autory na různé speciální případy: v kapitole 1 je popsána Lomnického metoda (1961), v kapitole 2 Eppsova metoda (1987), v kapitole 3 Hinichova metoda (1982) a ve čtvrté kapitole Pierceova metoda (1985). Lomnického metoda, spočívající v porovnání výběrové šikmosti a špičatosti dat z lineárního stochastického procesu s hodnotami odpovídajícími normálnímu rozdělení, je doprovázena podrobným odvozením vlastností výběrových momentů spočítaných ze závislých dat. Eppsova metoda je založena na harmonických funkcích pozorování stacionárního procesu a vyžaduje odhad spektrální hustoty. I této metodě je věnován poměrně obsáhlý výklad. Hinichova metoda a Pierceova metoda je popsána relativně stručně.

Druhá část práce je věnována simulačním studiím hladiny a síly testů normality určených pro nezávislá data a testů studovaných v prvních třech kapitolách. Data jsou generována z řady různých AR, MA a ARMA modelů s normálním rozdělením chyb. Alternativní hypotéza (data nemají normální rozdělení) je zkoumána výhradně za exponenciálního rozdělení chyb.

Práce Davida Stibůrka se zabývá zajímavým tématem a lze ji pochválit za podrobné zpracování Lomnického a Eppsovy metody a rozsáhlé simulační studie. Vytkl bych jí velké množství chyb a nepřesností, z nichž některé jsou závažné, chybějící vysvětlení testů na nezávislá data, které se používají v simulacích, špatnou srozumitelnost některých pasáží, nevhodnou volbu alternativní hypotézy při simulacích a nepřehlednou prezentaci simulačních výsledků.

Předloženou práci Davida Stibůrka celkově hodnotím jako uspokojivou a doporučuji ji uznat za práci diplomovou.

Hlavní připomínky:

1. Chybí přehledné srovnání studovaných metod, jejich předpokladů a přístupů k řešení problému. Kapitoly 1–4 jsou od sebe navzájem zcela izolované.
2. Důkaz věty 2.9 obsahuje chyby (viz níže).
3. Chybí přehled testů normality na nezávislá data. Tyto testy jsou označovány názvy funkcí z R (`cvm.test`, `lillie.test`, atd.).
4. Prezentace simulačních výsledků na 123 za sebou jdoucích stránkách tabulek a grafů bez jakékoli struktury je zcela nepřehledná.
5. V simulacích použité exponenciální rozdělení chyb porušuje více předpokladů než pouze normalitu. Exponenciální rozdělení má nenulovou střední hodnotu a tudíž negeneruje stacionární proces, veškeré studované metody přitom předpokládají stacionaritu. Existuje řada daleko vhodnějších rozdělení (t_k, Gumbelovo, šikmé normální), které bylo možné zvolit.
6. V simulacích se každý test vyhodnocoval na nově vygenerovaných datech. To je jednak plýtvání a jednak to do porovnání testů přidává zbytečnou variabilitu. Pro daný test se přitom používala v sekvenci simulací ze stejného modelu vždy stejná posloupnost chyb, ale pro různé testy byly tyto sekvence jiné.
7. V simulacích se jako zamítnutí nulové hypotézy počítaly i případy, kdy p-hodnota vyšla vyšší než 1 (??) a případy, kdy test nebyl schopen p-hodnotu spočítat.

Drobné připomínky:

- 10 asi ne t_2, ale t_n
- 15_1 místo $R(i-j)$ má být $R(q)$
- 14 co to je *caesarovská suma*?
- 25_6 „jaku"
- 31_9 na levé straně chybí $N^{1/2}$; také by bylo třeba uvažovat konvergenci sdruženého rozdělení (M_2, M_4)

- 31_6 spíše dle (1.43) a (1.44)
- 33 Nebyla spočítána kovariance cov (G_1, G_2). Pro test nelze tyto výsledky požít, pokud není k disposici odhad var G_1 a var G_2.
- 33–34 Příklad: není zřejmé, zdali jde o teoretický příklad, simulaci, nebo reálná data. Odkud jsou známy všechny údaje v tabulce?
- 35^7 Proč je N z předchozí kapitoly najednou přeznačeno jako T, jež mělo předtím úplně jiný význam, a N nyní také značí něco jiného než předtím?
- 35_8 Význam parametru θ není v tomto okamžiku vůbec patrný.
- 35_1 Tady nemá být $\mathrm{E}\, g(X_1, \lambda)$, ale $\int g(x, \lambda)\phi((x-\mu)/\sigma)\, dx$.
- 37–38 Lemma 2.2: K je definováno na $\langle -1, 1\rangle$, ale používá se pouze na $\langle 0, 1\rangle$. Jaký význam funkce K v lemmatu vlastně má a co znamenají podmínky (2.8)–(2.11)? Co znamená $\hat{K}$ ve (2.14)? Jak se volí funkce K, jak se volí M_T, a jaký vliv tyto volby mají na výsledný test?
- 38 Jak se postupuje při minimalizaci (2.15), aby výsledná θ_T byla „co nejblíže" výběrovému průměru a rozptylu?
- 38 Na této stránce se objevují čtyři varianty značení odhadu spektrální hustoty: $\hat{f}$, $\tilde{f}_T$, $\tilde{f}$, $\hat{f}_T$. Znamenají skutečně všechny totéž?
- 39 Podmínka (D): Kterou omezenou podmnožinu Θ volíte pro střední hodnotu a rozptyl normálního rozdělení? Když počítáte tento test, snažíte se určit množinu $\Theta_0(\lambda)$?
- 44^2 a níže: T je již zahrnuto v Z_T, výrazy nemají být vynásobeny T
- 44_{11} Konvergence G_T^- vyžaduje ověření spojitosti pseudoinverse
- 44–45, druhá polovina důkazu V. 2.9: Místo $O_P(1)$ má být na některých místech $o_P(1)$. Ty členy musí konvergovat k nule, omezenost nestačí.
- 44_5 Nevidím, z čeho plyne (2.28) — kde se vzaly absolutní hodnoty na pravé straně? Levá strana může být záporná, pravá je kladná. Kam se ztratilo $Z_T(\theta_0, \cdot)$?
- 45^1 Z čeho plyne, že D je spojitá v obou argumentech?
- 45^4 „Zbývá nám ještě ukázat, že první člen je řádu $O_P(1)$" [má být $o_P(1)$]: Jestli to chápu dobře, argument je, že jestliže dvě posloupnosti veličin konvergují v distribuci k témuž rozdělení, pak jejich rozdíl konverguje v pravděpodobnosti k 0. Nebo je to míněno jinak?
- 45, 2.10: Chybí kapitoly 2.7–2.9
- 45, 2.10: Nezdá se mi, že zde uvedený postup řeší problém zmiňovaný na konci kap. 2.5 (dimenze N příliš velká, rozdíly mezi sousedními λ_j příliš velké nebo příliš malé). I po podělení výběrovou směrodatnou odchylkou je N vybráno subjektivně uživatelem a rozdíly mezi sousedními $l_T(\lambda_j)$ jsou determinovány původní, taktéž subjektivní, volbou λ_j.
- 48 $h(t)$ má být h_t
- 48 Zde byly definovány $\omega_0, \ldots, \omega_{N-1}$, ale vzápětí se používá ω_{j+N} pro nespecifikované j. Tomu nerozumím.
- 54^2 $\sum_{i=1}^{n}[(\hat{\gamma}_n - \gamma) + (\hat{\rho}_n - \rho)]$: nerozumím, co je $\hat{\gamma}_n$ a γ a přes jaké i se to sčítá.
- 54^4 Chybí $1/\sigma^*$, místo e_i má být e_i^*, místo σ má být σ^*.
- 70^3 „kvadratitcké"
- 70, 6.3: Eppsův test vyžaduje spoustu uživatelských vstupů, které lze zadávat téměř libovolně (K, M_T, λ_j, dokonce g a l_T). Jak jste tyto parametry volil a uvažoval jste vliv těchto voleb na chování testu?

Otázky k obhajobě:

1. K lemmatu 2: Prosím vysvětlete význam funkce K pro odhad spektrální hustoty a diskutujte volbu K a M_T. Uveďte, jak jste toto (a volbu λ_j) řešil při simulacích.
2. Jak se postupuje při minimalizaci kvadratické formy (2.15), aby výsledná θ_T byla „co nejblíže" výběrovému průměru a rozptylu?
3. Je možné spravit důkaz Věty 2.9?

doc. Mgr. Michal Kulich, Ph.D.
KPMS MFF UK
6. května 2010