

Posudek oponenta diplomové práce

Jan Hanousek: Dense zeros

Téma práce:

Práce se zabývá modelováním časových řad s vysokým výskytem nulových hodnot. Pro tento účel vhodným způsobem modifikuje klasické modely řad s podmíněným rozptylem či podmíněnou střední hodnotou. V úvodu by mělo být pro jistotu zdůrazněno, že práce se nezabývá řadami s řídkými pozorováními (sparse observations nebo irregular observations), neboť nuly se chápou jako legitimní pozorování.

Shrnutí obsahu:

Práce nejprve uvádí přehled klasických GARCH a MEM modelů, které pak modifikuje pro případ časových řad s vysokým výskytem nulových hodnot. Součástí práce je numerická studie pro vhodně vybraná reálná data včetně výpočetního kódu uvedeného v příloze.

Vlastní příspěvek:

Autor navrhuje zajímavé netriviální modifikace klasických modelů, vyšetřuje jejich teoretické vlastnosti a ukazuje jejich využití v praxi.

Matematická úroveň:

Práce je na vysoké matematické úrovni, autor dokonce navrhuje v rámci dané problematiky vlastní modely. Na druhé straně v matematických zápisech se objevují určité nekorektnosti, kterých si měl autor všimnout.

Zdroje a formální úprava:

Bibliografie je adekvátní a korektně citovaná. Grafická úprava práce je dobrá. Práce někdy používá beletristickou angličtinu, která není příliš vhodná pro odborný text (namátkou 2₁₉: ... called for duty; 13³: ... chapter dives into heart of our proposed solution ..., 13¹²: ...informed predictions ...; 27¹⁷: název sekce 3.4.2 Success Echoes Through History a další).

Připomínky a otázky:

- 4₁₁₋₁₀: normalized → standardized
- 5₁₇₋₁₆: Není vhodný příklad: při živelní katastrofě se obvykle nenavýšuje volatilita, ale skokovitě roste objem škod.
- 5₉: Není vysvětleno, co jsou *past errors* e_{t-k} .
- 7¹³: Případ platykurtických rozdělání se zápornou (excess) špičatostí se neuvažuje?
- 8₁: Rovnice volatility v GJR-modelu (1.24) není dle mého názoru správně zapsaná.
- 9₄: Má být σ_{t-1}^2 .
- 10-11: Předpoklad (1.33) je jediný předpoklad o časové řadě chybových členů $\{\varepsilon_t\}$ (nepředpokládá se *iid*, jako je tomu na str. 5₈₋₇)?

- 11: V (1.34) a (1.35) má být y_t místo x_t (včetně poznámky 4 pod čarou), viz také (1.38).
- 12: Výraz (1.42) není podmínka.
- 13₃: Proč se uvádí podmínky na pozitivitu parametrů α_0 , α_1 a β (vzhledem k formulaci rovnice volatility pomocí jejího logaritmu jsou tyto podmínky redundantní a jen zbytečně komplikují příslušný odhadový software)? To samé platí rovněž pro DZ-EGARCH1 a DZ-EGARCH2 (nebo tam se pozitivita parametrů nepředpokládá?), viz také připomínku níže.
- 14: Proč se vztahy (2.3) a (2.4) pro ε_t jednoduše nemotivují tím, že cílem je modelovat proces s polospojitém (semi-continuous) rozdělením (viz také odkaz na Hudecová a Pešta (2024))? Navíc není nutné předpokládat nezávislost procesů $\{v_t\}$ a $\{b_t\}$ minimálně pro to, aby platil vztah (2.6)? To samé platí pro (2.19) v modelu DZ-GMEM.
- 16¹³⁻¹⁴: Daná parametrická množina není kompaktní, takže předpoklad (A1) pro ni není splněn (v Eukleidovských prostorech je kompaktnost shodná s omezeností množiny).
- 24: V tab. 3.1 by měly být také odpovídající hodnoty vypočtené z původních dat v testovacím vzorku.
- 26: V tab. 3.3 chybí odhad parametru γ v modelu EGARCH1. Navíc by bylo užitečné uvést odhad parametru p , který vypovídá o četnosti výskytu nulových hodnot datech, viz také připomínku níže.
- Řada úvah a vztahů v práci vychází z implicitního předpokladu, že uvažované časové řady jsou až na „dense zeros“ kladné (viz 2¹²). Nebylo by tedy přirozenější přijmout tento předpoklad explicitně? Tím by se navíc vyřešil problém s podmínkami positivity parametrů. Není totiž jasné, zda pro řady modelující podmíněnou střední hodnotu a nikoli volatilitu (viz např. (1.32) v MEM, (2.18) v DZ-GMEM a (2.22) v DZ-NMEM21) má smysl taková omezení na parametry požadovat.
- Provedená numerická studie se více zaměřuje na srovnání numerických výstupů jednotlivých modelů mezi sebou než na jejich porovnání s originálními daty (včetně dat z testovacího vzorku). Co se týče provedeného „optického“ porovnání výstupů v obr. 3.2, je dost obtížné (zvláště když na vertikálních osách jsou různá měřítka). Případné porovnání kvantilů v tab. 3.1 a 3.2 není také příliš korektní, neboť dané tabulky nepracují se shodnými vzorky dat (kdybychom tento fakt pominuli vzhledem k deklarované stacionaritě dat, pak z hlediska odhadnutých kvantilů si uvažované modely s výjimkou NMEM21 nestojí příliš dobře).
- Osobně mě překvapuje, že v rámci provedené studie se nevěnuje větší pozornost frekvenci predikovaných nulových hodnot a vůbec odhadu parametru p , který s tím úzce souvisí. Nevyvíjí se v čase (s nárůstem počtu provozovaných vozidel v čase se počet bezškodných dnů může zmenšovat), není tam určitá sezónnost (ta je pro pojištění motorových vozidel typická)? V práci jsem nezaregistroval dokonce ani odhadnutou numerickou hodnotu tohoto parametru.

Závěr:

Autor prokázal schopnost pracovat jak s netriviální matematikou v rámci analýzy časových řad, tak aplikovat teorii na reálná data. Proto doporučuji, aby předložená práce byla uznána jako diplomová práce.