Posudek oponenta diplomové práce

Název: Modelování durací mezi finančními transakcemi
Autor: Andrea Voráčková

Shrnutí obsahu práce

Hlavním předmětem práce je zkoumání ACD (autoregressive conditional duration) modelu, který se často používá pro popis posloupnosti délek nepravidelných časových intervalů mezi obchodními transakcemi. Autorka se zabývá problémem odhadu parametrů a konstrukcí předpovědí. Velká pozornost je věnována simulační studii. Na závěr je uvedena aplikace na reálná data.

Celkové hodnocení práce

Práce je celkem rozsáhlá, obsahuje 94 stránek, ovšem 32 z nich jsou opakující se podobné výstupy simulační studie pro různé speciální případy a 21 stránek tvoří skript pro počítačové výpočty. Bohužel se v práci vyskytuje dost nepřesností, a to v teoretické i praktické části. Rovněž není jasně specifikováno, v čem spočívá vlastní přínos autorky.

Struktura práce. Práce je rozdělena do čtyř kapitol. První dvě obsahují potřebné teoretické základy. Nejprve jsou uvedeny základní vlastnosti ARMA a GARCH náhodných procesů. Následně se v druhé kapitole studentka věnuje samotnému ACD modelu a jeho speciálním případům. Třetí kapitola nabízí podrobné výsledky simulační studie. Ve čtvrté kapitole jsou vyložené postupy použity na vysokofrekvenční data získaná z časů transakcí s akciemi společnosti IBM v jeden určitý den. Práce navíc obsahuje přílohu s výpisem použitého kódu programu R.

Téma práce. Téma považuji za velmi vhodné pro studijní obor *Finanční a pojistná matematika*. Téma bylo zpracováno v souladu se zadáním práce. Jenom je trochu nejasné, do jaké míry byla naplněna věta ze zadání: „provede potřebná odvození, která nejsou do detailu rozpracovaná v literatuře“.

Vlastní příspěvek. Teoretická část je zpracována podle citované literatury. Z práce není zřejmé, jestli některá odvození prováděla autorka podrobněji sama. Vlastní je simulační část (kapitola 3), která mohla být zvládnuta přehledněji, úsporněji a hlavně pečlivěji. Aplikace na reálná data (kapitola 4) se dá také považovat za vlastní příspěvek, i když se dost výrazně vychází z postupů uvedených v Tsay (2002).

Matematická úroveň. Matematický text je formulovaný především v prvních dvou kapitolách, kde se nachází několik definic a vět. Důkaz je proveden pouze u věty 4. Na některých místech by mělo být matematické vyjadřování přesnější. Například hned v první definici a první větě chybí předpoklady na náhodný proces a uvažuje se pouze lineární predikce, což neodpovídá tvaru z věty 1. V definici 3 není moc šťastné psát, že proces má pro každé t konstantní střední hodnotu. Také není vhodné definovat r_t pomocí r_t (definice 5). Na začátku podkapitoly 1.3.2 by $\hat{\sigma}_h^2(l)$ spíš mělo predikovat hodnotu σ_{h+l}^2. S filtrací $\mathcal{F}_t$ se zachází dost neurčitě, na některých místech by ji bylo třeba lépe specifikovat. Překlepy ve vzorcích jsou např. na posledním řádku str. 7 nebo ve vzorci (2.11), ve vzorci (4.1) chybí β_0.

Práce se zdroji. Seznam použité literatury obsahuje 7 položek, přičemž 3 z nich jsou internetové zdroje, u kterých by bylo vhodné přidat datum citování.

Formální úprava. Některé formální nedostatky jsou typografického rázu, např. záměna spojovníku, pomlčky a minusu, psaní výpustky nebo uvozovek. Místo desetinných teček by měly být desetinné čárky. Na některých místech dochází k přetečení řádků. Zbytečná je prázdná strana nadepsaná *Seznam použitých zkratek*. Po jazykové stránce je práce sepsána kvalitně, obsahuje minimum překlepů a gramatických chyb, snad až na používání spojení *zda-li*.

Připomínky a otázky

V diskusi při obhajobě navrhuji několik otázek. Podle časových možností by se uchazečka měla k některým z nich vyjádřit.

1. Proč je potřeba v definici 1 uvažovat $h > l$? Připouští se nulová hodnota h, jak se má tento případ rozumět? Připouští se nulová hodnota l, co znamená predikce o 0 kroků?

2. Jaký je smysl toho, že autokorelační funkce se pokládá rovna nule, když neplatí $s, t \neq 0$?

3. Jak se mají chápat v definicích 4, 5 a 6 veličiny se zápornými indexy?

4. Můžete podrobněji zdůvodnit druhou rovnost v (1.5), případně poslední rovnost ve (2.9)?

5. Výpočet rozptylu η_i na straně 11 nedává správný výsledek.

6. Předpovědi v podkapitole 2.3 obsahují členy $\tau_{h-m+1}, \ldots, \tau_h$, které v praxi nejsou k dispozici. Nikde není vysvětleno, jakým způsobem se toto řeší.

7. Jakého typu jsou konvergence s H_0 nad šipkou na str. 15? V prvním případě při $T \to \infty$ by limita $t^*(T)$ neměla záviset na T.

8. V simulační části se nejprve vykresluje histogram a hustota náhodných chyb ϵ_i. Tyto náhodné chyby byly vždy generované z příslušného rozdělení (exponenciální, Weibullovo, zobecněné gama), a tedy není vůbec překvapivé, že pro delší řady se histogram víc blíží teoretické hustotě. Přijde mi zbytečné uvádět 7 podobných obrázků, které nedávají skoro žádnou zajímavou informaci.

9. Na straně 25 je uveden vzorec pro průměrnou absolutní chybu predikce. V něm se normuje každá absolutní odchylka pomocí l místo 10, což by se dalo očekávat, když se mluví o průměrné absolutní odchylce. Navíc podle přiloženého kódu je pro MAE použit jiný vzorec (bez jakéhokoli normování).

10. V modelu ACD(1, 1) má predikce tvar $\hat{\tau}_h(1) = \alpha_0 + \alpha_1\tau_h + \beta_1 z_h$ a $\hat{\tau}_h(l) = \alpha_0 + (\alpha_1 + \beta_1)\hat{\tau}_h(l-1)$ pro $l \geq 2$. Odtud je vidět, proč má monotónní průběh a blíží se ke střední hodnotě procesu, zatímco simulované hodnoty můžou dost fluktuovat, což je vidět i z přiložených obrázků. Proto není moc vhodné posuzovat kvalitu predikce z jediné realizace. Ta je sice v jistém smyslu „průměrná“, ale pořád hodně závisí na tom, jak dopadlo v dané realizaci generování náhodné složky. Nebyla snaha zkusit spočítat předpovědi pro každou simulovanou řadu a poté zjistit průměr MAE hodnot? Navíc způsob, jakým je predikce v kapitole 3 počítána, není ideologicky správný. Při simulacích se využívají známé hodnoty α_0, α_1, β_1 a τ_h, které by v praxi bylo třeba odhadnout.

11. Jaký cíl mělo simulovat k (100, 500 a 1000) řad? Vzhledem k tomu, že řady jsou generovány nezávisle na sobě, tak je jasné, že čím víc jich použijeme, tím přesnější budou výsledné odhady (rozptyl průměru klesá úměrně $1/k$).

12. K tabulce 4.1 mi chybí vysvětlení, jaký test se vlastně provádí. Jak jsou spočteny testové statistiky a příslušné p-hodnoty? Podobná poznámka se týká více míst, kde by se slušela větší diskuse toho, co a jak se testuje. Zdá se mi, že často jsou testy bezhlavě používány, aniž by se autorka snažila zamyslet se nebo ověřit, zda jsou splněny jejich předpoklady.

13. Nezkoušela jste na data použít model s vyššími řády, např. ACD$(1, 2)$?

14. Reálné durace, které se mají předpovědět, jsou všechny nejmenší možné (1 sekunda), takže se dá celkem očekávat, že předpovědi nebudou příliš přesné. Navíc v programu je několik chyb, které způsobují, že výsledky prezentované v obrázku 4.9 nejsou v pořádku. Předně osmý řádek na str. 92 způsobí, že se `simDur` přepíše hodnotami z `condDur`, takže skutečné pozorované (očištěné) hodnoty vůbec do funkce pro predikci nevstupují. Dále pak při zpětném výpočtu $\hat{\Delta}_h(l)$ má být všude u `beta7` na str. 93 znaménko plus.

15. Řada for cyklů v přiloženém kódu je zbytečná. Stejného výsledku by se dalo dosáhnout efektivněji.

Závěr

Vzhledem k několika uvedeným nedostatkům považuji diplomovou práci Andrey Voráčkové za spíše podprůměrnou, ale i tak ji **doporučuji uznat jako diplomovou práci na MFF UK**.

V Praze, 17. ledna 2018

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