Moderní přístupy k testování periodicity v časových řadách

Veronika Ročková
Matematicko-fyzikální fakulta, Univerzita Karlova v Praze

UNIVERZITA KARLOVA V PRAZE
matematicko-fyzikální fakulta

Abstrakt

V příspěvku je diskutován problém testování periodicity v časových řadách. Stručně jsou připomenuty dva klasické testy, Fisherův a Siegelův test. Hlavní pozornost je věnována přístupu, který využívá informační kritéria na odhad řádu regresního modelu. Jsou prezentovány výsledky některých simulací, které naznačují, že testy pomocí kritérií dosahují lepších výsledků než klasický Fisherův test. Rozdíl je patrný především u modelů, ve kterých je periodicita překryta náhodnou složkou. Tento přístup umožňuje nejen efektivně testovat přítomnost periodicity, ale také odhadnout, kolik periodických tendencí řada má, a tím získat úplnější obraz o periodickém chování řady.

Úvod do problému

Máme náhodné veličiny $X_1, \dots, X_N$ a uvažujeme následující regresní model:

$$X_t = \mu + \sum_{k=1}^{p} (a_k \cos t\lambda_k + b_k \sin t\lambda_k) + \varepsilon_t, \quad t = 1, \dots, N, \quad (1)$$

kde μ, a_j, b_j, λ_j jsou neznámé parametry, přičemž λ_j jsou navzájem různé frekvence z intervalu $(0, \pi)$ a ε_t je striktní bílý šum s normálním rozdělením $\mathrm{N}(0, \sigma^2)$. Parametr σ^2 také neznáme. Klíčový problém spočívá v odhadu počtu periodických složek p a odhadu frekvencí $\lambda_j,\ j = 1, \dots, p$. Pokud bychom totiž počet složek a všechny frekvence znali, situace by se redukovala pouze na odhad zbývajících parametrů pomocí lineární regrese. Účinným nástrojem při odhadu frekvencí je periodogram.

Definice Mějme konečnou posloupnost náhodných veličin $X_1, \dots, X_n$. Periodogramem této posloupnosti rozumíme funkci

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} X_t e^{-it\lambda} \right|^2, \quad -\pi \le \lambda \le \pi.$$

Pokud řada $X_t,\ t = 1, \dots, N,$ obsahuje výraznou periodicitu o frekvenci λ_0, pak je hodnota $I(\lambda_0)$ velká v porovnání s ostatními. Pokud má ale periodogram velké množství menších lokálních maxim, v řadě periodicita zpravidla není. Skutečnost, zda lze spojovat výrazné hodnoty periodogramu s přítomností periodicity, je třeba ověřit.

Fisherův test

Pomocí Fisherova testu můžeme ověřit statistickou významnost maxima periodogramu počítaného pouze přes síť tzv. Fourierových frekvencí $\lambda_r = \frac{2\pi r}{N},\ r = 1, \dots, \lfloor N/2 \rfloor$. Symbolem $\lfloor x \rfloor$ rozumíme celou část čísla x. Fisherův test tedy používá jen některé hodnoty periodogramu. Fourierovy frekvence jsou ekvidistantní a hustota jejich rozmístění závisí na délce řady. Za platnosti nulové hypotézy Fisherova testu předpokládáme, že řada délky $N = 2m + 1$ je tvořena posloupností nezávislých stejně rozdělených náhodných veličin s normálním rozdělením. Hodnoty periodogramu ve Fourierových frekvencích uspořádáme sestupně podle velikosti a označíme V_1 největší atd. až V_m nejmenší z nich. Testová statistika Fisherova testu je

$$W = \frac{V_1}{V_1 + \dots + V_m}.$$

Pokud maximální hodnota přes síť Fourierových frekvencí bude v porovnání s ostatními v rámci této sítě výrazně větší, testová statistika bude blízko jedné. Nulovou hypotézu tedy zamítneme, pokud W překročí kritickou hodnotu g_F. Test je podrobně popsán v (Anděl, 1976).

Siegelův test

Pokud je v řadě více periodických složek, tj. $p > 1$, může se mezi hodnotami periodogramu ve Fourierových frekvencích vyskytnout více velkých hodnot. Tyto hodnoty přispívají do jmenovatele Fisherovy testové statistiky a snižují tak její velikost. Tím je snížena také celková sílu testu. Siegel navrhl test, který měl tento nedostatek odstranit. Testová statistika je rovna

$$T_\lambda = \sum_{j=1}^{m} (Y_j - \lambda g_F)_+, \quad \text{kde} \quad Y_j = \frac{V_j}{V_1 + \dots + V_m},$$

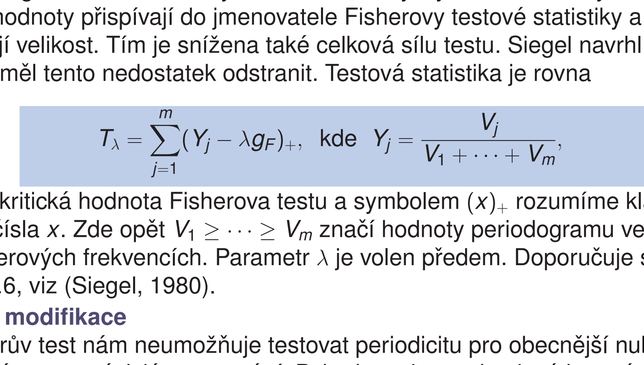
g_F je kritická hodnota Fisherova testu a symbolem $(x)_+$ rozumíme kladnou část čísla x. Zde opět $V_1 \ge \dots \ge V_m$ značí hodnoty periodogramu ve Fourierových frekvencích. Parametr λ je volen předem. Doporučuje se volit $\lambda = 0.6$, viz (Siegel, 1980).

Další modifikace

Fisherův test nám neumožňuje testovat periodicitu pro obecnější nulovou hypotézu, pro závislá pozorování. Pokud za platnosti nulové hypotézy předpokládáme, že řada je tvořena lineárním procesem generovaným striktním bílým šumem s normálním rozdělením, lze použít test, který navrhl Hannan, viz (Hannan, 1978) .

Periodogram spojité proměnné

Odhady frekvencí $\lambda_1, \dots, \lambda_p$ lze získat jako ty Fourierovy frekvence, které odpovídají p lokálním maximům periodogramu přes síť Fourierových frekvencí. Pokud je řada krátká, Fourierovy frekvence jsou daleko od sebe a mezi dvěma sousedními hodnotami se může vyskytnout lokální maximum, viz obrázek 1. Na obrázku 1 vlevo je červeně vykreslen periodogram spojité proměnné a modře jsou spojeny hodnoty periodogramu ve Fourierových frekvencích. Řada je délky 25 a zahrnuje dvě periodické složky o frekvencích $\frac{19\pi}{25}$, $\frac{11\pi}{25}$. Bílý šum má rozdělení $\mathrm{N}(0, 1)$. Při odhadech frekvencí je proto výhodné přihlédnout k periodogramu jako k funkci spojité proměnné. Za odhady frekvencí pak můžeme vzít argumenty p lokálních maxim periodogramu přes celý interval $(0, \pi)$.



Obrázek 1: Periodogram řady délky 25 obsahující periodické složky o frekvencích $19\pi/25$ a $11\pi/25$, vlevo červeně vykreslený jako spojitá funkce a modře jsou spojeny hodnoty ve Fourierových frekvencích, vpravo hodnoty periodogramu pro síť frekvencí která je desetkrát hustší než Fourierovy frekvence

Informační kritéria

Pomocí informačních kritérií můžeme odhadnout počet regresních parametrů modelu (1). Těmito parametry jsou koeficienty $a_j,\ b_j$ a frekvence λ_j. Nepřímo tak dokážeme odhadnout i počet složek p. Každé periodické složce $a_k \cos t\lambda_k + b_k \sin t\lambda_k,\ t = 1, \dots, N,$ totiž odpovídají právě tři regresní parametry. Odhad počtu parametrů pomocí informačních kritérií získáme jako takové $k \le K$, které dané kritérium minimalizuje. Číslo K je přitom maximální počet parametrů, který jsme ochotni připustit.

Informační kritéria

K nejužívanějším patří následující informační kritéria:

$$AIC(k) = \ln \frac{RSS}{N-k} + 2\frac{k}{N} \quad \text{(Akaikovo kritérium)},$$
$$BIC(k) = \ln \frac{RSS}{N-k} + \frac{k \ln N}{N} \quad \text{(Bayesovské kritérium)},$$
$$HQ(k) = \ln \frac{RSS}{N-k} + 2c\frac{k \ln(\ln N)}{N},\ c > 1, \quad \text{(Hannan-Quinnovo kritérium)}.$$

V simulacích byla použita tato klasická informační kritéria, Hannanovo-Quinnovo pro $c = 2$, a jedno nově navržené kritérium

$$A(k) = \ln \frac{RSS}{N-k} + \frac{k\sqrt{k}}{N}.$$

Test periodicity pomocí informačních kritérií

Pokud nám kritérium označí za odhad počtu parametrů modelu (1) číslo jedna, znamená to, že model vyjádříme pouze pomocí parametru μ. Jinými slovy v řadě není prokázána periodicita. Pokud ale získáme odhad počtu parametrů roven například číslu 4 (resp. 7), model můžeme popsat pomocí jedné (resp. dvou) periodických složek, a lze tedy zamítnout nulovou hypotézu. Jeden parametr je vždy střední hodnota μ a každé periodické složce přísluší právě tři parametry, koeficienty $a_k,\ b_k$ a frekvence λ_k. Pro implementaci této metody použijeme algoritmus nelineární regrese. Pro výpočet je třeba zadat počáteční odhady frekvencí. Odhady nebudeme hledat v rámci Fourierových frekvencí, protože takové odhady mohou být nepřesné, viz obrázek 1.

Počáteční odhady frekvencí

Nejprve odhadneme frekvenci periodické složky, která je v řadě zastoupena nejvýrazněji.

- Za odhad frekvence vezmeme argument maxima periodogramu přes síť frekvencí, která je desetkrát hustší než Fourierovy frekvence. (Důvodem je poměrně snadná implementace. Stanovit globální maximum spojitého periodogramu totiž znamenalo upřesnit interval, ve kterém se toto maximum nachází. Navíc je argument maxima přes naší hustší síť zpravidla velmi blízko argumentu globálního maxima periodogramu přes celý interval $(0, \pi)$, viz obrázek 1 vpravo.)
- V dalším kroku odhadneme zbývající regresní parametry periodické složky o této frekvenci.
- Tuto složku z řady odečteme.

Opakujeme celý postup s takto upravenou řadou. Tak získáme odhad frekvence druhé nejvýraznější periodické složky atd.

Fisherův test versus informační kritéria

Chování testů založených na informačních kritériích bylo zkoumáno na modelech, u kterých se měly projevit slabiny Fisherova testu. Víme, že taková situace nastane, pokud

- všechny frekvence jsou voleny mezi dvěma sousedními Fourierovými frekvencemi,
- rozptyl náhodné složky je relativně velký v porovnání s amplitudami periodických složek,
- v modelech se uplatňuje periodicita složená z více periodických složek o různých frekvencích,
- v modelech se složenou periodicitou jsou ampliduty všech složek stejně velké.

V simulacích byly uvažovány řady konstantní délky 75 s jednou, dvěma a třemi periodickými složkami. Velikost všech amplitud byla rovna 3. Hodnoty směrodatné odchylky bílého šumu ε_t byly uvažovány v rozmezí od 1 do 5. Pro každý model bylo provedeno 50 opakování. Maximální počet parametrů, který připouštíme, je 13. Tomu odpovídají 4 periodické složky. Uvedeme výsledky jen pro řady s jednou a dvěma periodickými složkami. Detailnější rozpis výsledků simulací lze nalézt v (Ročková, 2007).

Jedna periodická složka

Uvažujme model s jednou periodickou složkou ve tvaru

$$X_t = 3\cos(19\pi/75\, t) + \varepsilon_t,\ t = 1, \dots, 75.$$

V tabulce 1 je uveden počet, kolikrát z celkových padesáti opakování kritéria odhadla správně počet periodických složek, kolikrát řád modelu nadhodnotila a podhodnotila. V tabulce 1 je uveden počet zamítnutí hypotézy, že řada neobsahuje periodicitu, Fisherovým testem a testem pomocí kritérií. Z hlediska zamítání hypotézy si nejlépe vedlo kritérium *AIC*. Na druhé straně mělo tendenci řád modelu nadhodnocovat. Kritérium *A* například u modelu s velikostí směrodatné odchylky 5 odhadlo správně řád modelu v 39 případech a zamítlo nulovou hypotézu ve 42 případech. Fisherův test u samého modelu zamítl jen v 6 případech.

Tabulka 1: Model s jednou periodickou složkou, počet správně a chybně odhadnutého počtu periodických složek z možných 50 opakování

σ	Správně				Nad				Pod			
	A	AIC	BIC	HQ	A	AIC	BIC	HQ	A	AIC	BIC	HQ
2	44	12	50	49	6	38	0	1	0	0	0	0
3	46	12	43	49	4	37	0	1	0	1	7	0
4	38	7	33	20	11	43	3	2	1	0	14	28
5	39	12	16	5	3	34	1	0	8	4	33	45

Tabulka 2: Model s jednou periodickou složkou, počet zamítnutí hypotézy testem pomocí kritérií a Fisherovým testem z možných 50 opakování

σ	Test kritériem				Fisher
	A	AIC	BIC	HQ	
2	50	50	50	50	50
3	50	49	43	50	34
4	49	50	36	22	11
5	42	46	17	5	6

Dvě periodické složky

Uvažujme model se dvěma periodickými složkami

$$X_t = 3\cos(19\pi/75\, t) + 3\cos(11\pi/75\, t) + \varepsilon_t,\ t = 1, \dots, 75.$$

Obdobně jako u modelu s jednou periodickou složkou se rozdíl mezi Fisherovým testem a testy pomocí kritérií se prohluboval s rostoucím rozptylem náhodné složky. U modelu s nejvyšší uvažovanou směrodatnou odchylkou 5 byla u Fisherova testu úspěšnost v odhalení periodicity 20%, zatímco u kritéria *A* téměř 100%. V tabulkách 3 a 4 jsou výsledky shrnuty.

Tabulka 3: Model se dvěma periodickými složkami, počet správně a chybně odhadnutého počtu periodických složek z možných 50 opakování

σ	Správně				Nad				Pod			
	A	AIC	BIC	HQ	A	AIC	BIC	HQ	A	AIC	BIC	HQ
2	46	9	46	49	4	41	4	1	0	0	0	0
3	44	8	43	41	6	42	5	2	0	0	2	7
4	35	10	27	11	1	40	1	0	14	0	22	39
5	17	4	9	2	3	44	2	0	30	2	39	48

Tabulka 4: Model se dvěma periodickými složkami, počet zamítnutí hypotézy testem pomocí kritérií a Fisherovým testem z možných 50 opakování

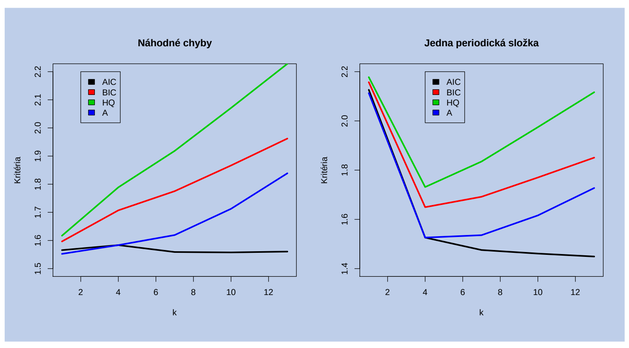
σ	Test kritériem				Fisher
	A	AIC	BIC	HQ	
2	50	50	50	50	46
3	50	50	50	48	26
4	50	50	40	25	15
5	49	50	31	14	10

Závěr

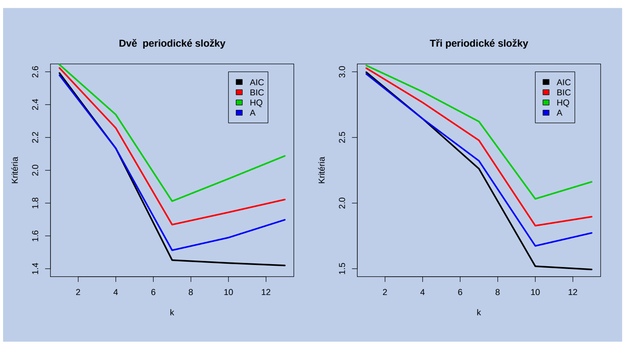
Výsledky simulačních studií ukázaly, že přístup k testování periodicity pomocí informačních kritérií dává lepší výsledky než klasický Fisherův test. Rozdíly se projevily především u modelů, ve kterých byly amplitudy periodických složek relativně malé oproti velikosti rozptylu bílého šumu. Testy pomocí kritérií jsou tedy mnohem citlivější a dokáží odhalit i periodicitu, která je zastoupena méně výrazně. Hlavní přínos tohoto přístupu spočívá v tom, že dokážeme efektivně testovat přítomnost periodicity a také odhadnout celkový počet periodických složek.

Obrázková příloha

Na následujících obrázcích jsou znázorněna kritéria v závislosti na počtu regresních parametrů modelu (1) pro různý počet periodických složek. Z obrázků je vidět, že kritérium *AIC* má tendenci nadhodnocovat řád modelu a nehodí se na testování přítomnosti periodicity.



Obrázek 2: Kritéria v závislosti na počtu parametrů k vlevo pro model tvořený pouze striktním bílým šumem s normálním rozdělením, vpravo je model s jednou periodickou složkou, v obou případech platí $\sigma = 2$



Obrázek 3: Kritéria v závislosti na počtu parametrů k vlevo pro model se dvěma periodickými složkami, vpravo se třemi periodickými složkami, v obou případech $\sigma = 2$

Poděkování

Ráda bych na tomto místě poděkovala firmě Median za grant, který mi umožnil účast na letní škole Robust 2008.

Odkazy

- Anděl J.: *Statistická analýza časových řad*, Nakladatelství technické literatury, Praha, 1976.
- Bhansali R. J.: *A Derivation of the Information Criteria for Selecting Autoregressive Models*, Adv. Appl. Prob. **18** (1986) 360–387.
- Hannan E. J.: *Testing for a Jump in the Spectral Function*, J. Appl. Prob. **15** (1978) 774–789.
- Ročková V.: *Testy periodicity v časových řadách*, (2007) Bakalářská práce, MFF UK.
- Siegel A. F.: *Testing for periodicity in a time series*, J. Am. Stat. Assoc. **75** (1980) 345–348.

Mail: rockova.ver@centrum.cz