Modellezés magasabb rendű Markov-láncokkal

Egyed Tünde
Témavezető: Csiszár Villő

1. Bevezetés

A sztochasztikus folyamatok modellezése érdekes kérdéseket vet fel. Jó módszer lehet a modellezésükre, hogy a folyamatra Markov-láncként tekintünk, vagyis azt feltételezzük, hogy a jövőbeli érték, csak a jelenbeli értéktől függ, a múltbeli értékektől nem. Ez formálisan azt jelenti, hogy

$$P(X_{n+1} = x_{n+1}|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

A dolgozatomban azt vizsgálom, hogy bizonyos esetekben nem lenne-e hasznosabb, ha a folyamatot magasabb rendű Markov-lánccal modelleznénk, azaz hosszabb memóriát feltételeznénk. Ennek azonban egyik hátránya, hogy a modell rendjének növelésével nő a modell komplexitása, ezáltal több adatra van szükség a megbízható modellezéshez, ami nem mindig áll rendelkezésre. A projektem célja bemutatni, hogy hogyan érdemes meghatározni egy Markov-lánc rendjét.

A munkám során több cikket tanulmányoztam a témában, melyek különböző módszerekkel, mutatókkal vizsgálták hálózatok rendjét.

A [1] cikkben az internetes oldalakon való böngészést figyelték meg. Itt a kutatás célja a hálózat emlékezetének meghatározása volt. Mivel a felhasználók linkeken keresztül közlekednek, ezért azt várjuk, hogy a legalább elsőrendű modellek jobban teljesítenek, mint a nullarendű modell, azonban az eredményekből az látszik, hogy weboldalak szintjén a nullarendű modell mutatói bizonyultak a legjobbnak. Ennek oka valószínűleg az, hogy a komplexebb modellek megtalálásához nem áll rendelkezésre elegendő adat. De ha a weboldalakat témakörök szerint csoportosítjuk, akkor már a mutatókban is megfigyelhető a magasabb rendű modellek eredményessége.

A [2] cikkben különböző hálózatokon vizsgálták az első- és a másodrendű modellek közti különbséget.

A múlt félévi munkámat folytatva a beszámolómban ismertetek a cikkekben leírt módszerek közül néhányat, majd ezeket egy valós mintán mutatom be.

Ahogy már említettem, a magasabb rendű modelleknél gondot okoz a paraméterek számának gyors növekedése, ezért ebben a félévben a [3] és a [4] cikkek alapján bemutatok egy modellt, ami kiküszöböli ezt a problémát. Ezt a módszert alkalmazva szintén összehasonlítom a valós mintán a különböző rendű modelleket.

2. LAMP modell

A magasabb rendű Markov-láncok modellezésének egyik nehézsége, hogy a rend növelésével a paraméterek száma exponenciálisan nő, ami magasabb rendű modelleknél azt eredményezheti, hogy nem áll rendelkezésre elegendő adat a paraméterek becsléséhez. Ezt a problémát kezeli a LAMP (Linear Additive Markov Process) modell. A hagyományos k-ad rendű Markov modellhez hasonlóan ez a folyamat is az előző k állapotra emlékszik vissza, azonban megőrizve az elsőrendű modell egyszerűségét az átmenetmátix mérete nem nő a rend növelésével, helyette k darab átmenetmátrixot használunk fel.

Legyen az állapottér n elemű, a minta t-edik elemét jelölje i_t. Legyen adott egy ψ eloszlás az $\{1, 2 \ldots, k\}$ halmazon, továbbá k darab $n \times n$-es átmenetmátrix, ezek közül a μ-ediket jelölje a^μ. Az $a^\mu(i_t|i_{t-\mu})$ jelölje az $i_{t-\mu}$ i_t átmenet valószínűségét az a^μ átmenetmátrix szerint. Ekkor a LAMP modell az átmenetmátrixot az alábbi képlettel határozza meg:

$$P(i_t|i_{t-1}, i_{t-2}, \ldots, i_{t-k}) = \sum_{\mu=1}^{k} \psi(\mu) a^\mu(i_t|i_{t-\mu})$$

A modellben a ψ eloszlás határozza meg, hogy a múltat milyen gyorsan "felejtjük el", míg az a^μ mátrix reprezentálja a μ-lépéses átmenetvalószínűségeket.

2.1. Paraméterek becslése

A módszer alkalmazásához első lépésben meg kell becsülnünk a ψ eloszlást és az a átmenetmátrixokat. Ehhez EM algoritmust használtam az alábbi iterációs lépésekkel, ahol I az $\{i_1, i_2, \ldots, i_L\}$ L hosszú mintát jelöli.

$$\psi(\mu) = \frac{\sum_t P(x_t = \mu|I)}{\sum_{t,\nu} P(x_t = \nu|I)}$$

$$a^\mu(i'|i) = \frac{\sum_t P(x_t = \mu, i_{t-\mu} = i, i_t = i'|I)}{\sum_t P(x_t = \mu, i_{t-\mu} = i|I)}$$

A fenti képleteket a következő összefüggés segítségével számolhatjuk ki:

$$P(x_t = \mu|I) = \frac{\psi(\mu) a^\mu(i_t|i_{t-\mu})}{\sum_\nu \psi(\nu) a^\nu(i_t|i_{t-\nu})}$$

A paraméterek meghatározásánál további kérdés, hogy milyen kezdőértékkel indítjuk el az iterációt. A valós mintán a paraméterek becslése során több kezdőértékkel is próbálkoztam, és összehasonlítottam a becslés végén kapott paraméterekhez tartozó likelihood értékeket.

3. Optimális rend kiválasztásának módszerei

Ebben a fejezetben az optimális rend kiválasztásának néhány lehetséges módját mutatom be.

3.1. Likelihood-hányados módszer

A likelihood-hányados módszer során felírjuk a k-rendű θ_k átmenetmátrixhoz tartozó likelihood függvényt. Ezt első rendű modell esetén az alábbi képlettel kapjuk:

$$L(\theta_k, x) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2}) \cdots p(x_2|x_1) = p(x_1) \prod \prod p_{ij}^{n_{ij}}$$

ahol p_{ij} az átmenet valószínűsége x_i állapotból x_j állapotba θ_k átmenetmátrix mellett, n_{ij} az átmenetek száma a mintában x_i-ből x_j-be. Belátható, hogy az átmenetmátrix maximum likelihood becslése a relatív gyakoriság.

A modell rendjének növelésével ugyan nő a likelihood érték, de ezzel együtt nő a modell komplexitása is. Éppen ezért vizsgálni kell, hogy a magasabb rendű modellhez tartozó likelihood érték szignifikánsan nagyobb-e. Ennek eldöntésére a valószínűségi hányados próbát alkalmazzuk. Nullhipotézisként azt tesszük fel, hogy a modell k rendű, alternatív hipotézisként m rendet feltételezünk, ahol $m > k$. A próbastatisztika értéke a következő:

$$_k\eta_m = -2(\log L(\theta_k, x) - \log L(\theta_m, x))$$

Az így kapott $_k\eta_m$ statisztika a nullhipotézis mellett χ^2 eloszlást követ, melynek szabadságfoka a Markov modell esetén $(|S|^m - |S|^k)(|S|-1)$, a LAMP modellt alkalmazva pedig $(m-k)|S|(|S|-1) + m - k$. Ahol S a lehetséges állapotok halmazát jelöli.

Ennek a módszernek az egyik hátránya, hogy egyszerre csak két modellt tudunk tesztelni.

3.2. Akaike-féle információs kritérium

A különböző rendű modellek összehasonlítására segítségünkre lehetnek különféle információs kritériumok, ezek közül az egyik az Akaike-féle információs kritérium, melyet a következő képlettel kapunk:

$$AIC = -2 \log L + 2k,$$

ahol k a becsült paraméterek száma, L a modell likelihood függvényének maximum értéke.

Tehát az információs kritérium i-ed rendű Markov modell esetén:

$$AIC(i) = -2 \log L + 2(|S|^i)(|S| - 1)$$

Az i-ed rendű LAMP modellre pedig:

$$AIC(i) = -2 \log L + 2(i|S|(|S| - 1) + i - 1)$$

A modell annál jobb, minél alacsonyabb AIC értéket kapunk.

Az Akaike-féle információs kritérium egy változata, amikor kiválasztunk egy referenciamodellt, majd a vizsgált modellre a fenti módon számolt információs kritériumból kivonjuk a referenciamodellre számolt információs kritériumot. Ettől a módosítástól természetesen nem változik meg, hogy melyik modellhez tartozik a legkisebb AIC érték, ugyanakkor ez a meghatározás jobban szemlélteti a modellek közti különbséget.

3.3. Bayes-féle információs kritérium

Egy másik gyakran használt információs kritérium a Bayes-féle információs kritérium n elemű mintára:

$$BIC = -2 \log L + k \log n,$$

ahol k szintén a paraméterek számát jelöli.

Ahogy az előző módszerben, itt is a kritérium csökkenése jelenti a modell javulását. A Bayes-féle információs kritériumnak szintén létezik a referenciamodellel összehasonlított változata.

3.4. Cross Validation

Első lépésben felosztjuk a mintát két részre, egy tanító halmazra, amin megbecsüljük az átmenetvalószínűségeket, és egy validáló halmazra, amin ellenőrizhetjük az eredményt. Miután a tanító halmazon megbecsültük az átmenetvalószínűségeket, minden i, j állapotpárhoz definiálunk egy r_{ij} rangértékeket, vagyis hogy az x_i állapotból hányadik legvalószínűbb, hogy az x_j állapotba kerülünk. Ha két állapotba ugyanakkora valószínűséggel kerülhetünk, akkor mindkét állapot a nagyobb rangot kapja. Például ha az x_i állapotból 4 különböző állapotba léphetünk $3/8, 1/4, 1/4, 1/8$ valószínűségekkel, akkor az r_{ij} rangértékek rendre $1, 3, 3, 4$.

A rend növelésével nő a lehetséges állapotok száma, így az egyes állapotokból kevesebb fordul elő a mintában. Ez azt eredményezheti, hogy egy állapot a validáló halmazban előfordul, de a tanító halmazban nem. Ebben az esetben mindegyik állapot a legnagyobb rangot kapta.

Végül kiszámoljuk a modellhez tartozó átlagos rangot az alábbi képlet alapján:

$$\frac{\sum_i \sum_j n_{ij} r_{ij}}{\sum_i \sum_j n_{ij}}$$

ahol n_{ij} az átmenetek számát jelölöi x_i-ből x_j-be a validáló halmazon.

A különböző rendű modellek közül a legkisebb átlagos rangút választjuk.

3.5. Kétlépéses visszatérés

Ebben a módszerben megbecsüljük a k-rendű modellel annak a valószínűségét, hogy két lépés múlva ugyanabba az állapotba térünk vissza, majd az így kapott eredményt összevetjük a mintában szereplő kétlépéses visszatérések relatív gyakoriságával.

Első lépésben megbecsüljük az átmenetmátrixot a k-rendű modellel, majd ebből kiszámoljuk a staconárius eloszlást. Ezután felírjuk a $P(x_{n+2} = i, x_{n+1} = j | x_n = i)$ valószínűséget minden i-re, és ezeket a valószínűséget összegezzük súlyozva a stacionárius eloszlással:

$$\sum_i P(X_n = i, X_{n+2} = i) = \sum_i \pi(i) \sum_j P(X_{n+2} = i, X_{n+1} = j | X_n = i)$$

Az így kapott érték egy becslés a kétlépéses visszatérés valószínűségére. Ezt összevetjük a mintában szereplő kétlépéses valószínűségek relatív gyakoriságával. A modell annál jobb, minél közelebb áll egymáshoz a két érték.

3.6. Entrópia

A megfelelő rend megtalálásához bevezetjük az entrópia fogalmát. Az első rendű Markov folyamat betűnkénti entrópiája az alábbi képlettel számolandó:

$$H(X_{t+1}|X_t) = -\sum_{j,k} \pi(j) p_{j,k} \log(p_{j,k})$$

ahol π a stacionárius eloszlás.

Magasabb rendű modell esetén a képlet hasonló annyi módosítással, hogy a feltételbe miden megjegyzett állapotot figyelembe kell venni. Tehát például másodrendű modellre:

$$H(X_{t+1}|X_t X_{t-1}) = -\sum_{i,j,k} \pi(ij) p_{ij,jk} \log(p_{ij,jk})$$

Az optimális rend megtalálását jelzi, ha a rend további növelésével már nem csökken tovább az entrópia.

4. Modellezés valós adatokon

A fent bemutatott módszereket valós adatokon is kipróbáltam. A felhasznált minta napkitörések erősségét tartalmazza 2002. január és 2019. május között. Az adatok https://www.kaggle.com/datasets/laudiomachadopaulo/solar-flare-list-over-12-years oldalról származnak. A minta 13 870 napkitörést tartalmazott. Az eredeti adatsorban a napkitörések erősségét 211 különböző kategóriba sorolák, azonban ennyi állapot nehezen kezelhető a modellezés során, illetve nagyon sok állapot csak kevésszer szerepelt a mintában, ezért az állapotokat 11 kategóriába vontam össze. Az összevont kategóriák gyakoriságát az 1 ábra mutataja, a 2 ábrán pedig az elsőrendű átmenetgyakoriságok mátrixa látható.

A modellezéshez R programnyelvet használtam. Az elemzéshez használt forráskódok a https://github.com/TundeEgyed/Markov oldalon találhatók.

A projekt során az első-, másod- és harmadrendű Markov modelleket hasonlítottam össze, magasabb rendű modellek vizsgálatára nem volt lehetőség, mivel ekkor már nagyon sok állapot csak egyszer fordult volna elő a mintában.

1. ábra.

	To										
From	C1	C2	C3	C4	C5	C6	C7	C8	C9	M	X
C1	3398	1088	459	250	157	94	83	61	43	327	18
C2	1114	631	321	167	109	70	46	37	36	167	17
C3	459	298	166	99	58	37	49	28	25	141	9
C4	238	172	96	67	47	35	25	21	17	95	8
C5	131	114	66	51	41	28	19	9	15	75	5
C6	101	69	48	28	16	12	13	7	9	51	4
C7	67	49	37	31	17	15	11	11	8	63	7
C8	59	45	25	14	15	11	12	4	7	32	1
C9	53	33	22	17	14	5	3	4	3	32	5
M	340	202	117	92	76	50	52	38	28	237	15
X	18	15	12	4	4	1	3	5	0	27	6

2. ábra.

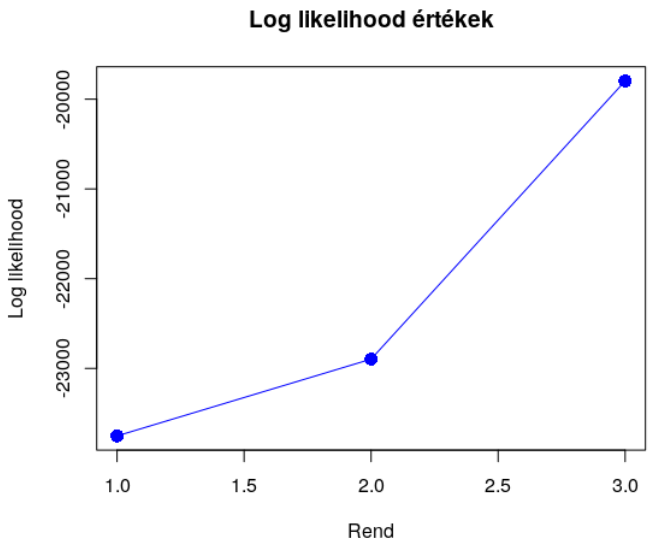
4.1. Likelihood

A loglikelihood értékeket a különböző rendekre a 3 ábra mutatja. Mivel egymásba ágyazott modellekről van szó, ezért vizsgáljuk, hogy a magasabb rendű modell szignifikánsan jobb-e. Az 1 táblázatban láthatóak a rendek összehasonlítása során kapott

	$k=1,\ m=2$	$k=1,\ m=3$	$k=2,\ m=3$
${}_k\eta_m$	1 712	7 912	6 200
Szabadságfok	1 100	13 200	12 100
p-érték	0	1	1

1. táblázat.

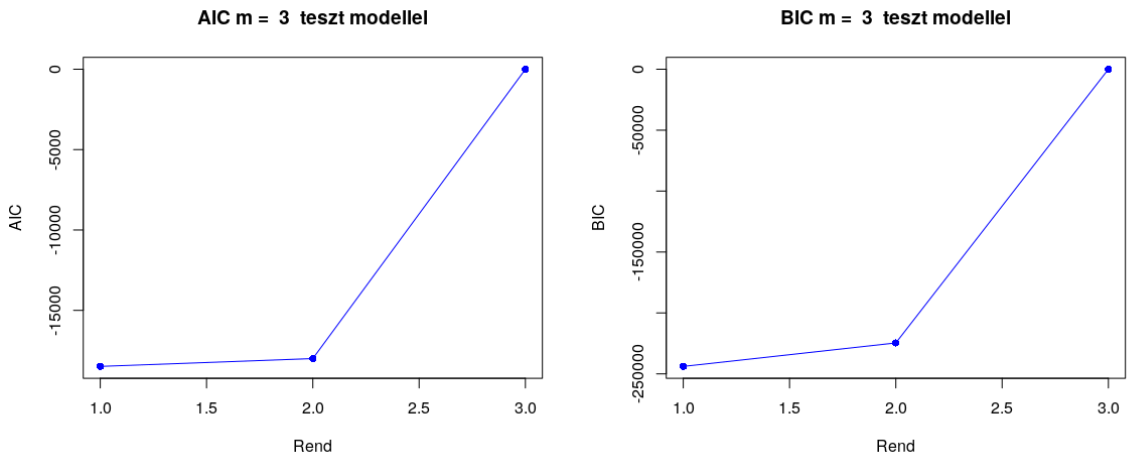
próbastatisztika értékek, szabadságfokok és p értékek. A próba alapján azt mondhatjuk, hogy csak a másodrendű modell javít jelentősen az eredményen, a harmadrendű már nem.



3. ábra.

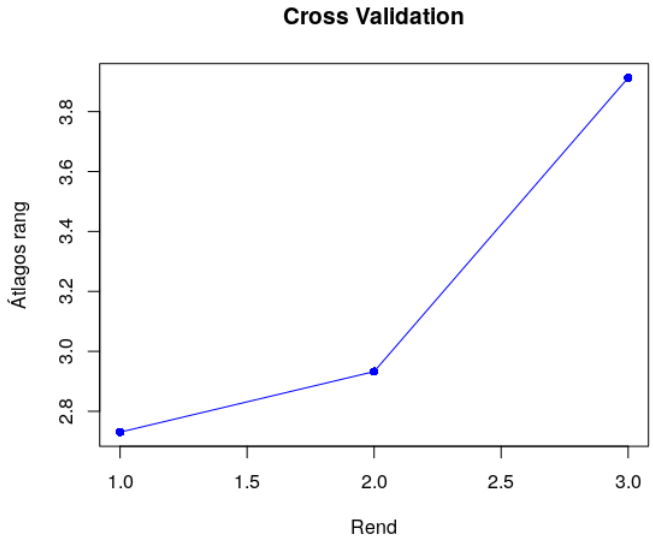
4.2. Információs kritériumok

A modelleket összehasonlítottam az Akaike és Bayes-féle információs kritériumok segítségével is. Ezek eredményei a 4.2 ábrákon láthatóak. Mindkét információs kritérium azt mutaja, hogy az első rendű modell a legjobb.



4.3. Cross Validation

A cross validation módszer alkalmazása során tanuló halmazt a minta első 10 000 eleme alkotta, a maradék 3 870 elem került a validáló halmazba. Az eredményeket a 4 ábra mutatja. Ez a módszer megerősíti az információs kritériumokkal kapott eredményt, és az elsőrendű modellt javasolja.



4. ábra.

4.4. Kétlépéses visszatérés

A kétlépéses visszatérési valószínűségeket vizsgálva azonban azt tapasztaljuk, hogy a másod- vagy harmadrendű modell közelíti legjobban a valóságot. Ugyanis a mintában

annak relatív gyakorisága, hogy két lépés múlva ugyanabba az állapotba kerülünk $0,31$, a különböző rendű modellekkel a 5 táblázatban szereplő értékeket kapjuk. Itt azt látjuk, hogy a másod és harmadrendű modellel kapott érték van legközelebb a kétlépéses visszatérés relatív gyakoriságához.

Order	Two-Step Return
1	0.27
2	0.31
3	0.31

5. ábra.

4.5. Entrópia

A különböző rendű modellekkel kapott entrópia értékek a 6 táblázatban láthatóak. Ennél a módszernél azt tapasztaljuk, hogy még a harmadrendű modell esetén is jelentősen csökken az entrópia, tehát ez alapján akár a harmadrendű modelllel is érdemes lehet dolgozni.

Order	Entropy Rate
1	0.34
2	0.14
3	0.07

6. ábra.

4.6. LAMP modell

Az adatokon kipróbáltam a LAMP modellt is, majd ennek eredményeit összevetettem az előző modellekkel.

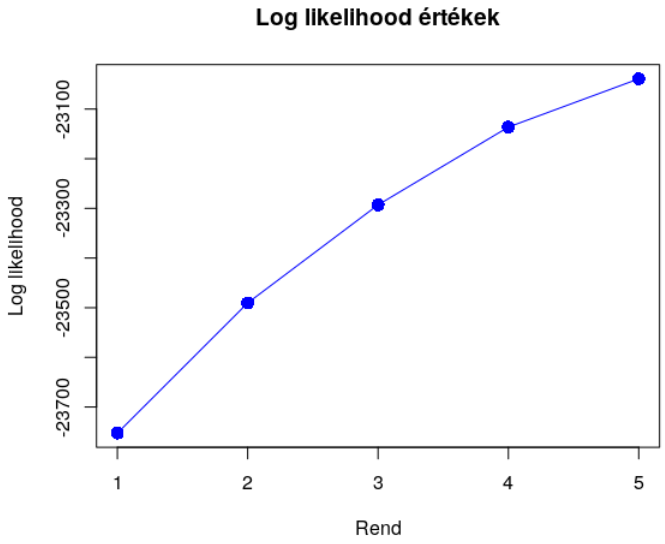
Első lépésben megbecsültem a ψ vektort és az a mátrixokat. Az EM algoritmust többféle kezdőértékkel is elindítottam, és azt tapasztaltam, hogy a kapott ψ vektorok és az a mátrixok eltérhehetnek ugyan, de ezek szorzatösszegeként kapott mátrixok az indítástól függetlenül nagyságrendileg egyenlőek lesznek. Az algoritmust 100 ismétlésszámmal futtattam le, de az figyelhető meg, hogy a likelihood értékek 10-15 ismétlésszám után már nem nőnek jelentősen.

A paramméterek becslése után kiszámoltam a loglikelihood értékeket ötödrendű modellig, ami a 7 ábrán látaható. Ezután megvizsgáltam a modellekhez tartozó Akaike-féle és a Bayes-féle információs kritériumokat. Ezeket az eredményeket összehasonlítottam egymással és a Markov modellel, az Akaike-féle információs kritéium

Modell	Paraméterek száma	Loglikelihood	AIC	BIC
Markov 1	110	-23 753	47 725	48 554
Markov 2	1 210	-22 897	50 320	57 334
Markov 3	13 310	-19 797	74 605	166 538
LAMP 1	110	-23 753	47 725	48 554
LAMP 2	221	-23 490	47 423	49 089
LAMP 3	332	-23 293	47 250	49 752
LAMP 4	443	-23 136	47 158	50 498
LAMP 5	554	-23 039	47 186	51 365

2. táblázat.

alapján azt kaptam, hogy a negyedrendű LAMP modell a legjobb, míg a Bayes-féle az elsőrendű modellt javasolja. A loglikelihood értékek és az információs kritériumok a 2 táblázatban láthatóak.



7. ábra.

Hivatkozások

[1] Singer, Philipp, et al, "Detecting memory and structure in human navigation patterns using markov chain models of varying order," PloS one 9,7 (2014): e102070,

[2] Rosvall, Martin, et al, "Memory in network flows and its effects on community detection, ranking, and spreading," Ecology 19 (2014): 30,

[3] Ravi Kumar, Maithra Raghu, Tamás Sarlós, Andrew Tomkins, "Linear Additive Markov Processes", International World Wide Web Conference Committee 2017, April 3–7, 2017, Perth, Australia,

[4] Lawrence K. Saul, Michael I. Jordan "Mixed Memory Markov Models: Decomposing Complex Stochastic Processes as Mixtures of Simpler Ones", 1999 Kluwer Academic Publishers