Modellezés magasabb rendű Markov láncokkal

Egyed Tünde
Témavezető: Csiszár Villő

1. Bevezetés

Az előző félévben elkezdett munkámat folytatva, sztochasztikus folyamatok magasabb rendű Markov láncokkal való modellezésével foglalkoztam. A félév során megismerkedtem különböző módszerekkel, melyek segítségével vizsgálható a Markov láncok rendje. A megismert módszereket szimulált mintán próbáltam ki, majd néhány módszert valós adatokon is alkalmaztam.

2. Elmélet

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

2.1. Likelihood-hányados módszer

A likelihood-hányados módszer során felírjuk a k-rendű θ_k átmenetmátrixhoz tartozó likelihood függvényt:

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

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. 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 $(|S|^m - |S|^k)(|S| - 1)$ szabadságfokkal, ahol S a lehetséges állapotok halmaza.

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

2.2. 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:

$$AIC(k) = {}_k\eta_m - 2(|S|^m - |S|^k)(|S| - 1)$$

Itt m a legmagasabb ésszerű rend, k a vizsgált modell rendje. A modell annál jobb, minél alacsonyabb AIC értéket kapunk.

2.3. Bayes-féle információs kritérium

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

$$BIC(k) = {}_k\eta_m - (|S|^m - |S|^k)(|S| - 1)\log n$$

Ahogy az előző módszerben, itt is a kritérium csökkenése jelenti a modell javulását.

2.4. 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 training 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. 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.

2.5. 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.

2.6. Entrópia

A megfelelő rendet megtalálásához definiáljuk az entrópia fogalmát:

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

ahol π a stacionárius eloszlás.

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

3. Modellezés valós adatokon

A fent bemutatott módszerek közül a likelihood-hányados módszert és az információs kritériumokat valós adatsorokon is kipróbáltam. Ennek eredményeit ismertetem ebben a fejezetben.

Az első minta üzleti ciklusokat tartalmazott, azonban a minta a rövidsége és az állapotok ritka változása miatt nem volt alkalmas magasabb rendű Markov modellekkel való elemzésre.

A második adathalmaz 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 211 különböző kategória szerepelt. Mivel ennyi állapot nehezen kezelhető, illetve sok állapot nagyon kevésszer szerepelt a mintában, ezért a kategóriákat 11 állapotba vontam össze. Az így kapott adatsorhoz tartozó egylépéses gyakorisági mátrix a 1. ábrán látható.

A modellezés során R programnyelvet használtam.

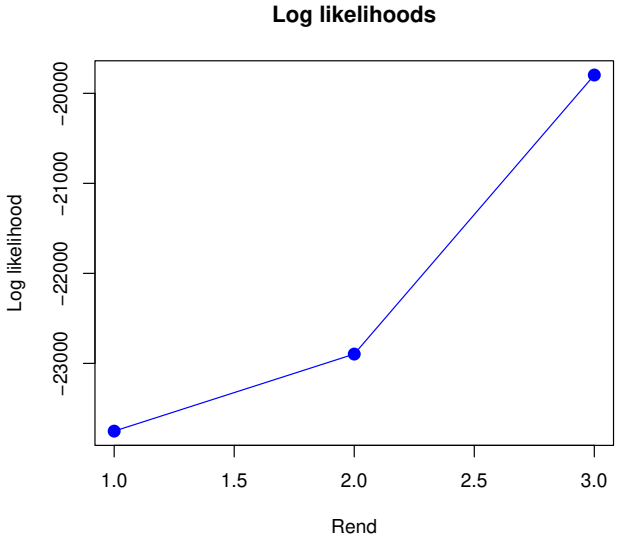
3.1. Likelihood

A likelihood módszerrel az 2. ábrán látható log likelihood értékeket kaptam.

Az elsőrendű és a másodrendű modell összehasonlításaként a valószínűségi hányados próbával $p = 0$ értéket kaptam, vagyis a másodrendű modell valóban jobb, mint az elsőrendű. A harmadrendű modell vizsgálata során viszont már nem nőtt jelentősen a loglikelihood érték. Az eredmények a 1. táblázatban láthatók.

From \ To	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

1. ábra.



2. ábra.

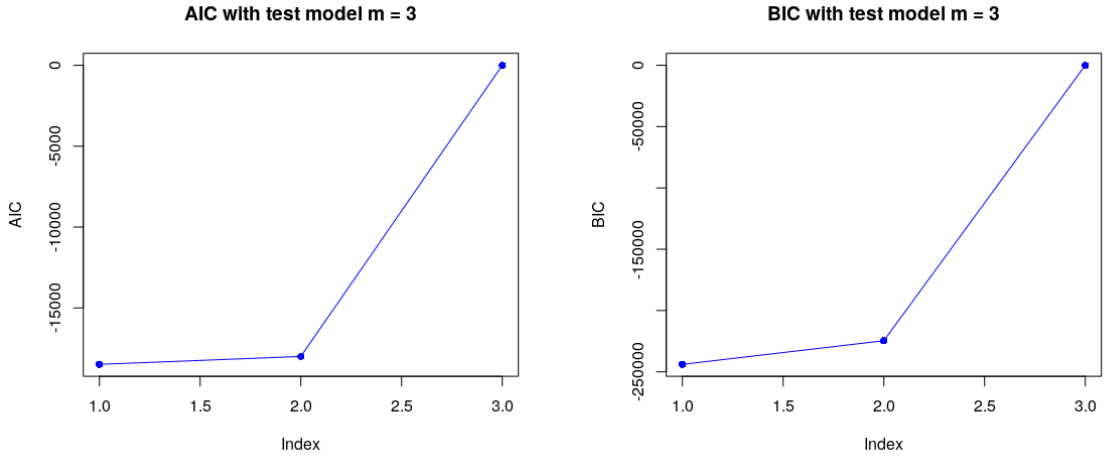
3.2. Információs kritériumok

A likelihood-hányados módszerrel szemben az Akaike és a Bayes-féle információs kritériumok nem támasztják alá a másodrendűséget. A következő ábrákon látszik, hogy a másodrendű modell esetén magasabb információs kritériumok adódnak, ugyanakkor

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

1. táblázat.

a növekedés nem jelentős, az így kapott eredmény tehát nem mond teljesen ellent a likelihood módszernek.



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.