Modellezés magasabb rendű Markov láncokkal

Egyed Tünde
(Támavezető: Csiszár Villő)

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Önálló projekt 2
2022. május, Budapest

Bevezetés

- Cél: Markov lánc optimális rendjének megtalálása
- Néhány módszer bemutatása az optimális rend megtalálásához
- A rend vizsgálata egy valós mintán

Likelihood módszer

- Megbecsüljük az átmenetvalószínűségeket a relatív gyakoriságokkal
- Felírjuk a likelihood függvényt:

$$L(\theta_k, x) = p(x_1) \prod_i \prod_j p_{ij}^{n_{ij}}$$

- Valószínűség-hányados próba a szignifikancia tesztelésére H_0 : a modell rendje k, H_1 : a modell rendje m hipotézisek mellett a próbastatisztika:

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

- Ez H_0 esetén χ^2 eloszlású $(|S|^m - |S|^k)(|S| - 1)$ szabadságfokkal, ahol S a lehetséges állapotok halmaza

Információs kritériumok

- Akaike-féle információs kritérium:

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

- Bayes-féle információs kritérium:

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

  ahol n a minta elemszáma

- A kisebb információs kritériumok jobb modellt jeleznek

Cross Validation

- A mintát két részre osztjuk, egy tanító halmazra és egy validáló halmazra
- A tanítóhalmazon megbecsüljük az átmenetvalószínűségeket
- r_{ij}: az x_i állapotból hányadik legvalószínűbb, hogy x_j-be kerülünk
- n_{ij}: az átmenetek száma x_i-ből x_j-be a validáló halmazon
- Kiszámoljuk az átlagos rangot:

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

- Minél kisebb az átlagos rang, annál jobb a modell

Kétlépéses visszatérés

- Kétlépéses visszatérési valószínűségek becslése
- Stacionárius eloszlás kiszámítása
- A kétlépéses visszatérési valószínűségek összege a stacionárius eloszlással súlyozva

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

- Eredmények összevetése a mintában szereplő kétlépéses visszatérések relatív gyakoriságával

Entrópia

Az entrópiát az alábbi képlettel számolhatjuk ki:

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

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

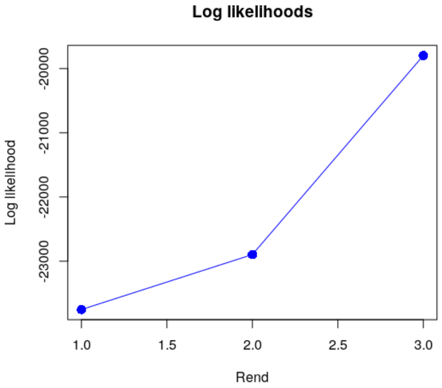
Minta

- A vizsgált minta napkitörések erősségét tartalmazza 2002. január és 2019. május között
- 13 870 napkitörés
- Eredetileg 211 állapotot tartalmazott, amit 11 állapotba vontam össze

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

Eredmények likelihood módszerrel

- A vizsgálat során az első-, a másod- és a harmadrendű modelleket hasonlítottam össze
- A log likelihood értékek:

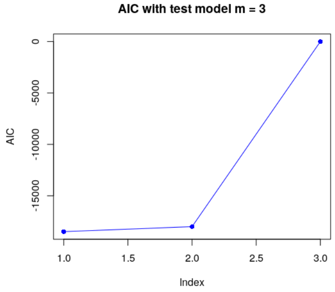


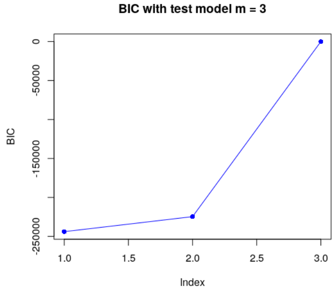
Eredmények likelihood módszerrel

A másodrendű modell jobb, mint az elsőrendű, de a harmadrendű már nem javít jelentősen.

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

Eredmények információs kritériumokkal





Az információs kritériumok nagyobbak a másodrendű modellre, mint az elsőrendűre, de nem nagy a növekedés.

Köszönöm a figyelmet!