Modellválasztás információs kritériumokkal famodellekre és Markov mezőkre

Talata Zsolt

A Ph.D. értekezés tézisei

Témavezető: Professzor Csiszár Imre
Magyar Tudományos Akadémia
Rényi Alfréd Matematikai Kutatóintézet

Budapesti Műszaki és Gazdaságtudományi Egyetem
Matematika Intézet
Budapest

2004

1. Bevezetés

Ez a mű a *Modellválasztás információs kritériumokkal famodellekre és Markov mezőkre* című Ph.D. értekezés vázlata. Az értekezés három fejezetből áll. Az első áttekintést nyújt a modellválasztási problémákról. A két további fejezet az értekezés új eredményeit tartalmazza: a második fejezet a kontextusfák becslésével foglalkozik, a harmadik fejezet a Markov mezők esetében tűzi ki célul a modellválasztást.

Az értekezés minden része publikálásra került. A három fejezet három publikációnak felel meg. Lényegében az 1., 2. és 3. fejezet sorra a (Talata, 2004), a (Csiszár and Talata, 2004b) és a (Csiszár and Talata, 2004a) publikáció.

A tézisfüzet a 2. fejezetben bemutatja a modellválasztási problémát, és a 3. fejezetben összegzi az irodalomban található azon eredményeket, amelyek az értekezés új eredményeit ösztönözték. A két új eredmény a 4. és 5. fejezetben kerül megfogalmazásra, jelezve a bizonyítások módszereit. Az irodalomjegyzék tartalmazza az értekezés összes hivatkozását.

2. A modellválasztási probléma

Legyen adott egy $\{ X_t,\, t \in T \}$ sztochasztikus folyamat, ahol minden egyes X_t egy valószínűségi változó $a \in A$ értékekkel, és T egy indexhalmaz. Az X_t, $t \in T$ valószínűségi változók együttes eloszlására a folyamat eloszlásaként fogunk hivatkozni és Q-val fogjuk jelölni. A folyamat egy *modell*je meghatározza a folyamat egy hipotetikus eloszlását vagy hipotetikus eloszlások egy csoportját. Egy modellt tipikusan egy, valamilyen $\mathcal{K}$ halmazbeli k struktúra-paraméter és egy $\theta_k \in \Theta_k \subset \mathbb{R}^{d_k}$ paramétervektor határoz meg; ezt a modellt M_{θ_k}-val jelöljük. Adottak a folyamat lehetséges modelljei, ezeket modellosztályokba lehet rendezni a struktúra-paraméter szerint: $\mathcal{M}_k = \{ M_{\theta_k},\, \theta_k \in \Theta_k \subset \mathbb{R}^{d_k} \}$. A folyamatról statisztikai következtetést annak egy $\{ x_t,\, t \in T \}$ realizációjának az $R_n \subset T$ tartományban adott megfigyelése alapján vonunk le, ahol R_n növekszik n-nel. Így az n-edik *minta* $x(n) = \{ x_t,\, t \in R_n \}$. Az alábbiakban felsorolunk néhány jellemző példát folyamatokra és azok modelljeire.

Sűrűségfüggvény becslése esetén $T = \mathbb{N}$, és az X_t, $t \in \mathbb{N}$ valószínűségi változók függetlenek és azonos eloszlásúak (i.i.d.) az f_{θ_k} sűrűségfüggvény szerint. Az n-edik minta $\{ x_i,\, i = 1, \ldots, n \}$.

Polinomillesztés során $T \subseteq \mathbb{R}$, ahol T egy megszámlálható halmaz, $A = \mathbb{R}$, és a model

$$X_t = \theta_k[0] + \theta_k[1]\, t + \theta_k[2]\, t^2 + \ldots + \theta_k[k-1]\, t^{k-1} + Z_t,$$

ahol Z_t, $t \in T$ független valószínűségi változók normális eloszlással, nulla várható értékkel, ismeretlen közös szórással, és $\theta_k[i]$ az i-edik komponense a k-dimenziós θ_k paramétervektornak. Most a $k \in \mathbb{N}$ struktúra-paraméter a $\theta_k[0]+\theta_k[1]\, t+\theta_k[2]\, t^2+\ldots+\theta_k[k-1]\, t^{k-1}$ polinom fokszáma plusz 1, és az n-edik minta $\{ x_t,\, t \in \{t_1, \ldots, t_n\} \subset T \}$.

A folyamat, amikor $T = \mathbb{N}$, $A = \mathbb{R}$, k rendű *autoregresszív (AR) folyamat*, ha

$$X_t = \sum_{i=1}^{k} a_i X_{t-i} + Z_t,$$

ahol Z_t, $t \in \mathbb{N}$ független valószínűségi változók normális eloszlással, nulla várható értékkel, ismeretlen közös szórással, és $a_i \in \mathbb{R}$, $i = 1, \ldots, k$ alkotja a θ_k paramétervektort. Most a $k \in \mathbb{N}$ struktúra-paraméter az a_i együtthatók száma, és az n-edik minta $\{ x_i,\, i = 1, \ldots, n \}$.

Az *autoregresszív mozgó átlag (ARMA) folyamat* hasonló az AR folyamathoz. Ebben az esetben

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + Z_t + \sum_{j=1}^{q} b_j Z_{t-j}.$$

A paramétervektor $\theta_k = \{ a_1, \ldots, a_p, b_1, \ldots, b_q \} \in \mathbb{R}^{p+q}$, és a k struktúra-paraméter kétkomponensű: $k = (p, q) \in \mathbb{N}^2$.

A folyamat, amikor $T = \mathbb{N}$, $|A| < \infty$, k rendű *Markov lánc*, ha

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{i=k+1}^{n} Q(x_i |\, x_{i-k}^{i-1}), \quad n \geq k,\ x_1^n \in A^n,$$

alkalmas $Q(\,.\,|\,.\,)$ átmenet-valószínűségekkel. Az x_i^j az $x_i, x_{i+1}, \ldots, x_j$ sorozatot jelöli. Mivel minden egyes $a_1^k \in A^k$-ra a $\{ Q(a|\, a_1^k),\, a \in A \}$ vektor valószínűségi eloszlás az A-n, a $\theta_k \in \mathbb{R}^{d_k}$ paramétervektor $d_k = (|A| - 1)\,|A|^k$ számú $Q(a|\, a_1^k)$, $a \in A^*$, $a_1^k \in A^k$ átmenet-valószínűségből áll, ahol $|A^*| = |A| - 1$. Most a $k \in \mathbb{N}$ struktúra-paraméter azon szekvenciák hossza, amelyektől az átmenet-valószínűségek a második változójukban függenek. Az n-edik minta $\{ x_i,\, i = 1, \ldots, n \}$.

Az AR és ARMA folyamatok és a Markov láncok példák arra az esetre, amikor a modell nem határozza meg egyértelműen a folyamat hipotetikus eloszlását. Konkrétan, a k rendű AR folyamatok vagy Markov láncok estén a modell csak egy hipotetikus feltételes eloszlást határoz meg $X_{k+1}, X_{k+2}, \ldots$-re adott $X_1, \ldots, X_k$ mellett.

A lehetséges k struktúra-paraméterek $\mathcal{K}$ halmaza egy rendezett vagy félig rendezett halmaz az $\mathcal{M}_k$ modellosztályok tartalmazására nézve. Amikor a k struktúra-paraméterű M_{θ_k} modell megfelel a folyamat valódi Q eloszlásának, egy összetettebb modell (a fenti értelemben) nagyobb k' struktúra-paraméterrel szintén megfelelhet a Q eloszlásnak egy alkalmas $\theta_{k'}$ paramétervektorral. Például bármely k rendű AR folyamat vagy Markov lánc egyben k' rendű is, minden $k' > k$-ra. A *valódi modell*en, amit M_{θ_0} fog jelölni, a legkisebb modellt értjük a valódi Q eloszlásnak megfelelő modellek között, azaz amelyre nem létezik a valódi Q eloszlásnak megfelelő másik modell a fenti értelemben kisebb struktúra-paraméterrel. E valódi modell struktúra-paraméterét k_0 fogja jelölni.

A *modellválasztási probléma* a k_0 valódi struktúra-paraméternek a folyamat $x(n)$ statisztikai megfigyelése alapján történő becslésében áll.

Az *alulbecslés* kifejezés arra az esetre utal, amikor a valódi k_0-nál egy kisebb k struktúra-paramétert választunk. Ilyen esetben $\theta_0 \notin \Theta_k$, ennélfogva a valódi modell nem lehet pontosan becsülve, a paramétervektor becslése torzítást hordoz.

A *felülbecslés* kifejezés arra az esetre utal, amikor a valódi k_0-nál egy nagyobb k struktúra-paramétert választunk. Ilyen esetben $M_{\theta_0} \in \mathcal{M}_{k_0} \subset \mathcal{M}_k$, így $M_{\theta_0} = M_{\theta_k}$ valamely $\theta_k \in \Theta_k$-ra, azonban θ_k-nak több komponense van, mint θ_0-nak, ennélfogva nehezebb becsülni a valódi értéket, a paramétervektor becslésének nagyobb lesz a szórása.

Az értekezés a modellválasztási problémát tárgyalja az információs kritérium fogalmán keresztül. Egy *információs kritérium (IC)* az $x(n)$ minta alapján hozzárendel egy valós számot minden egyes modellosztályhoz: $IC : \mathcal{K} \times \{x(n)\} \to \mathbb{R}$, a k_0 becslője egyenlő azzal a struktúra-paraméterrel, amelyre a kritériumnak a legkisebb az értéke:

$$\hat{k}(x(n)) = \arg\min_{k \in \mathcal{K}} IC_k(x(n)).$$

3. Előzmények

Különféle folyamatokra bizonyítást nyert, hogy a *Bayes-féle Információs Kritérium (BIC)* (Schwarz, 1978) és a *Legkisebb Leíró Hossz (MDL)* kritérium (Rissanen, 1978, 1983a, 1989) a struktúra-paraméter erősen konzisztens becslésére vezet. Ez azt jelenti, hogy a BIC-t illetve az MDL-et minimalizáló struktúra-paraméter jelölt egyenlő a valódi struktúra-paraméterrel, 1 vsz-gel elég nagy n esetén. Itt és a továbbiakban az „1 vsz-gel elég nagy n esetén" azt jelenti, hogy 1 valószínűséggel létezik egy olyan (az $\{ x_t,\, t \in T \}$ realizációtól függő) n_0 küszöbszám, hogy az állítás érvényes minden $n \geq n_0$-ra.

Konzisztencia bizonyítások rendelkezésre állnak például i.i.d. folyamatokra az exponenciális családba tartozó eloszlással (Haughton, 1988), AR folyamatokra (Hannan and Quinn, 1979), ARMA folyamatokra (Hannan, 1980), Markov láncokra (Finesso, 1992) és famodellekre (Willems, Shtarkov and Tjalkens, 1993, 2000).

E bizonyítások mindegyike magában foglalja azt a feltételezést, hogy a jelölt struktúra-paraméterek száma véges, azaz van egy ismert *felső korlát* a struktúra-paraméterre. Ez a feltételezés technikai természetű és a bizonyítást egyszerűsíti. Mindamellett nemkívánatos, mert a gyakorlatban nem szokott előzetes információ rendelkezésre állni a struktúra-paraméterről, továbbá amikor növekvő mennyiségű adatunk van, egyre összetettebb hipotetikus modellosztályokat szeretnénk jelöltként számításba venni. Ezért indokolt célkitűzés elvetni a struktúra-paraméter előzetes korlátjának feltételét.

A k rendű Markov láncra a Bayes-féle információs kritérium a következő alakú:

$$\mathrm{BIC}_k(x_1^n) = -\log \mathrm{ML}_k(x_1^n) + \frac{(|A|-1)|A|^k}{2} \log n,$$

ahol ML_k jelöli a maximum likelihoodot. Itt $(|A|-1)|A|^k$ egyenlő a paraméterek számával. A maximum likelihoodra kapjuk, hogy

$$\mathrm{ML}_k(x_1^n) = \prod_{a_1^{k+1}:\, N_n(a_1^{k+1}) \geq 1} \left[\frac{N_n(a_1^{k+1})}{N_n(a_1^k)} \right]^{N_n(a_1^{k+1})},$$

ahol $N_n(a_1^k)$ jelöli az $a_1^k \in A^k$ előfordulásainak számát az x_1^n mintában.

Az MDL kritérium felírható

$$\mathrm{MDL}_k(x_1^n) = -\log P_k^{(n)}(x_1^n) + L(k)$$

alakban, ahol $P_k^{(n)}$ egy, a k rendű Markov láncok osztályára szabott kódoló eloszlást jelöl, és $L(k)$ jelöli a k rend kódhosszát. A szokásos kódoló eloszlások a *Kricsevszkij – Trofimov (KT)* (Krichevsky and Trofimov, 1981) és a *Normalizált Maximum Likelihood (NML)* eloszlások.

A k rendű KT eloszlás explicit alakja:

$$\mathrm{KT}_k(x_1^n) = \frac{1}{|A|^k} \prod_{a_1^k : N_n(a_1^k)\geq 1} \frac{\prod_{a_{k+1}:N_n(a_1^{k+1})\geq 1} \left[\left(N_n(a_1^{k+1}) - \frac{1}{2}\right) \left(N_n(a_1^{k+1}) - \frac{3}{2}\right) \ldots \left(\frac{1}{2}\right) \right]}{\left(N_n(a_1^k) - 1 + \frac{|A|}{2}\right) \left(N_n(a_1^k) - 2 + \frac{|A|}{2}\right) \ldots \left(\frac{|A|}{2}\right)}.$$

Az NML eloszlás definíciója a következő:

$$\mathrm{NML}_k^{(n)}(x(n)) = \mathrm{ML}_k(x_1^n) \Big/ \sum_{x_1^n \in A^n} \mathrm{ML}_k(x_1^n) \; .$$

Shtarkov (1977) megmutatta, hogy az $\mathrm{ML}_k(x_1^n)$ maximum likelihoodoknak az összes lehetséges x_1^n mintára vett összege aszimptotikusan (amint n tart végtelenbe, rögzített k mellett) egyenlő $(|A|-1)|A|^k\, 2^{-1} \log n$-nel. Ennélfogva az MDL kritérium NML módozata aszimptotikusan ekvivalens a BIC-vel, ha a jelölt k rendek száma véges. Amint a következő eredmények mutatják, ez az ekvivalencia nem érvényes amikor nincs előzetes korlát a k rendre.

Csiszár és Shields (2000) bebizonyították, hogy a Markov láncok rendjének BIC becslője erősen konzisztens még akkor is, ha a rend előzetes konstans korlátjának feltételét elvetjük, és az n-edik x_1^n mintára az összes lehetséges $0 \leq k < n$ rendet jelölt rendnek tekintjük.

Ugyanakkor Csiszár és Shields (2000) megmutatták, hogy ugyanez az eredmény nem lehet érvényes az MDL becslőre. Tekintsük az egyenletes eloszlású i.i.d. folyamatot. Ez a folyamat 0 rendű Markov lánc. Az MDL kritériumra, amikor a $P_k^{(n)}$ kódoló eloszlás akár KT_k, akár $\mathrm{NML}_k^{(n)}$, és a k rend $L(k)$ kódhosszára teljesül $L(k) = o(k)$, azt kapjuk, hogy

$$\hat{k}(x_1^n) = \arg \min_{0\leq k\leq \alpha \log n} \mathrm{MDL}_k(x(n)) \;\; \to +\infty \quad \text{ha } n \to \infty,$$

ahol $\alpha = 4/\log|A|$. Ez az ellenpélda azt mutatja, hogy az MDL becslő nem konzisztens, ha a rendre vonatkozó korlátot teljesen elvetjük.

Csiszár (2002) bizonyította a Markov láncok BIC rendbecslőjének erős konzisztenciáját akkor, amikor a jelölt rendek halmaza növekedhet az n mintamérettel, mégpedig a számításba vett rendek korlátja $o(\log n)$ a KT esetben, és $\alpha \log n$, ahol $\alpha < 1/\log|A|$, az NML esetben. Figyeljük meg, hogy ezen MDL becslők nem igényelnek előzetes korlátot a valódi rendre. A konzisztencia az $L(k)$ tag nélküli MDL kritériumra került bizonyításra, ami egy erősebb eredmény.

Az értekezés a modellválasztási problémát az alábbiakban ismertetett két modell esetében tűzi ki célul. A struktúra-paraméterekre erősen konzisztens becslőket adunk. A fenti eredményektől ösztönözve, a jelölt modellosztályok száma növekedhet a minta méretével, így nincs szükség előzetes korlátra a struktúra-paraméterre vonatkozóan.

4. Kontextusfa becslése nem szükségképpen véges memóriájú folyamatokra, BIC és MDL útján

Egy véges A halmaz esetén a számosságát jelölje $|A|$. Egy $s = a_m a_{m+1} \ldots a_n$ *sztring*et (ahol $a_i \in A$, $m \leq i \leq n$) a_m^n-nel is jelölünk, a hossza $l(s) = n - m + 1$. Az

üres sztringet $\varnothing$ jelöli, ennek hossza $l(\varnothing) = 0$. Egy u és v sztring összeláncolását uv jelöli. Azt mondjuk, hogy a v sztring *utótag*ja az s sztringnek, ezt $s \succeq v$ jelöli, ha létezik egy u sztring amelyre $s = uv$. Egy valódi utótagra, azaz amikor $s \neq v$, azt írjuk, hogy $s \succ v$. Egy egyirányban végtelen $a_{-\infty}^{-1} = \dots a_{-k} \dots a_{-1}$ sorozat utótagja hasonlóképpen definiált. Megjegyezzük, hogy az irodalomban $\succ$ gyakrabban előtag relációt jelöl.

*Fá*nak nevezzük sztringek – és akár egyirányban végtelen sorozatok – egy $\mathcal{T}$ halmazát, ha egyik $s_1 \in \mathcal{T}$ sem utótagja semelyik másik $s_2 \in \mathcal{T}$-nek.

Minden $s = a_1^k \in \mathcal{T}$ sztring egy úttal szemléltethető egy levéltől a gyökérig (a gyökeret legfelülre rajzolva), amely út k számú, az $a_1 \dots a_k$ szimbólumokkal címkézett élből áll. Egy egyirányban végtelen $a_{-\infty}^{-1} \in \mathcal{T}$ sorozat egy gyökérbe tartó végtelen úttal szemléltethető. Az $s \in \mathcal{T}$ sztringeket egyúttal a $\mathcal{T}$ fa leveleivel is azonosítjuk: az s *levél* az a levél, amelyik a gyökérhez a fentiek szerint megjelenített s út által kapcsolódik. Hasonlóképpen, a $\mathcal{T}$ fa *csomópontja*it az összes (véges vagy végtelen) $s \in \mathcal{T}$ véges utótagjaival azonosítjuk, a gyökeret pedig a $\varnothing$ üres sztringgel. Az s csomópont *gyermekei* azok az as, $a \in A$ sztringek, amelyek önmagukban is csomópontok, azaz utótagjai valamely $s' \in \mathcal{T}$-nek.

A $\mathcal{T}$ fa *teljes*, ha minden csomópontnak a levelek kivételével pontosan $|A|$ számú gyermeke van. Egy gyengébb tulajdonság, amire majd szükségünk lesz, az *irreducibilitás*, ami azt jelenti, hogy egyik $s \in \mathcal{T}$ sem helyettesíthető egy valódi utótagjával a fa tulajdonság megsértése nélkül. Az irreducibilis fák családját $\mathcal{I}$ fogja jelölni.

A $\mathcal{T}_1$ és $\mathcal{T}_2$ fákra azt írjuk, hogy $\mathcal{T}_2 \succeq \mathcal{T}_1$, ha minden egyes $s_2 \in \mathcal{T}_2$-nek van egy $s_1 \in \mathcal{T}_1$ utótagja, és minden egyes $s_1 \in \mathcal{T}_1$ utótagja valamely $s_2 \in \mathcal{T}_2$-nek. Amikor hangsúlyozzuk, hogy $\mathcal{T}_2 \neq \mathcal{T}_1$, akkor azt írjuk, hogy $\mathcal{T}_2 \succ \mathcal{T}_1$.

Jelölje $d(\mathcal{T})$ a $\mathcal{T}$ fa mélységét: $d(\mathcal{T}) = \max\{\, l(s), s \in \mathcal{T} \,\}$. Jelölje $\mathcal{T}\big|_K$ a $\mathcal{T}$ fa elmetszését a K szinten:

$$\mathcal{T}\big|_K = \{\, s' : s' \in \mathcal{T} \text{ és } l(s') \leq K, \text{ vagy } s' \text{ egy } \lfloor K \rfloor \text{ hosszúságú utótagja valamely } s \in \mathcal{T}\text{-nek} \,\}. \tag{1}$$

Tekintsünk egy stacionárius ergodikus sztochasztikus $\{\, X_i, -\infty < i < +\infty \,\}$ folyamatot véges A ábécével. Legyen

$$Q(a_m^n) = \mathrm{Prob}\{\, X_m^n = a_m^n \,\},$$

és, ha $s \in A^k$-ra $Q(s) > 0$, írjuk azt, hogy

$$Q(a|\, s) = \mathrm{Prob}\{\, X_0 = a \mid X_{-k}^{-1} = s \,\}.$$

A fenti folyamatra Q folyamatként fogunk hivatkozni.

4.1. Definíció. *Egy $s \in A^k$ sztring a Q folyamat egy* kontextusa, *ha $Q(s) > 0$ és*

$$\mathrm{Prob}\{\, X_0 = a \mid X_{-\infty}^{-1} = x_{-\infty}^{-1} \,\} = Q(a|\, s) \quad \textit{minden } a \in A\textit{-ra},$$

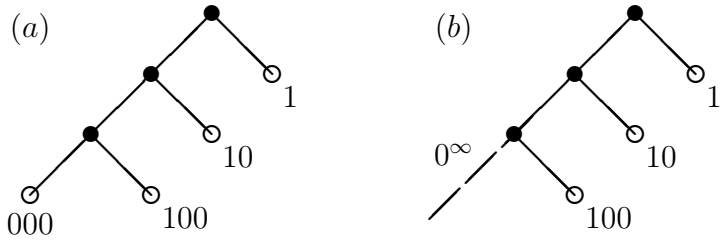
amennyiben s utótagja az egyirányban végtelen $x_{-\infty}^{-1}$ sorozatnak, és s-nek nincs valódi utótagja ugyanezzel a tulajdonsággal. Egy végtelen kontextus *egy olyan egyirányban végtelen $x_{-\infty}^{-1}$ sorozat, amely x_{-k}^{-1}, $k = 1, 2, \dots$ utótagjai pozitív valószínűségűek, de egyikük sem kontextus.*

Nyilvánvaló, hogy az összes kontextus halmaza egy fát alkot. Ezt a Q folyamat *kontextusfá*jának fogjuk nevezni, és $\mathcal{T}_0$ fogja jelölni.

4.2. Megjegyzés. A $\mathcal{T}_0$ kontextusfa szükségképpen teljes, ha $Q(s) > 0$ minden s sztringre. Általános esetben a $\mathcal{T}_0$ kontextusfa egyes s csomópontjainak pontosan azok az as, $a \in A$-k a gyermekei, amelyekre $Q(as) > 0$. Továbbá, a 4.1. definícióból következik, hogy $\mathcal{T}_0 \in \mathcal{I}$ mindig. □

Amikor a kontextusfa mélysége $d(\mathcal{T}_0) = k_0 < \infty$, a Q folyamat k rendű Markov lánc. Ebben az esetben a kontextusfa egy takarékos leírását adja a folyamatnak, mert $(|A|-1)|\mathcal{T}_0|$ számú átmenet-valószínűség elegendő a folyamat leírásához, szemben az $(|A|-1)|A|^{k_0}$ számúval. Megjegyezzük, hogy egy i.i.d. folyamat kontextusfája egyedül a $\varnothing$ gyökérből áll, így $|\mathcal{T}_0| = 1$.

4.3. Példa. *(Felújítási folyamat).* Legyen $A = \{0,1\}$ és tételezzük fel, hogy az 1-ek előfordulása közötti távolság i.i.d.. Jelölje p_j annak a valószínűségét, hogy ez a távolság j, azaz $p_j = Q(10^{j-1}1)$. Ekkor $k \geq 1$-re $Q(10^{k-1}) = \sum_{i=k}^{\infty} p_i \triangleq q_k$, $Q_k = Q(1|\,10^{k-1}) = p_k/q_k$. Legyen $Q_0 = Q(1) \triangleq q_0$. Jelölje k_0 azt a legkisebb egész számot, amelyre Q_k állandó azon $k \geq k_0$-k esetén, amikor $q_k > 0$, illetve legyen $k_0 = \infty$ ha nem létezik ilyen egész szám. Így a kontextusok az 10^{i-1}, $i \leq k_0$ sztringek, valamint a 0^{k_0} sztring (amennyiben $k_0 < \infty$) vagy az egyirányban végtelen 0^{∞} sorozat (amennyiben $k_0 = \infty$), ld. 1. ábra. □



1. ábra. A felújítási folyamat kontextusfája. (a) $k_0 = 3$. (b) $k_0 = \infty$.

Az értekezésben a $\mathcal{T}_0$ kontextusfának az x_1^n mintából, az X_1^n egy realizációjából, történő statisztikai becslésével foglalkozunk. Megköveteljük a becslés erős konzisztenciáját. Ezen a $d(\mathcal{T}_0) < \infty$ esetben azt értjük, hogy a becsült kontextusfa egyenlő $\mathcal{T}_0$-val, 1 vsz-gel elég nagy n-re, míg egyébként azt, hogy a becsült kontextusfa bármely rögzített K szinten elmetszve egyenlő $\mathcal{T}_0\big|_K$-val, 1 vsz-gel elég nagy n-re, ld. (1). Itt és a továbbiakban az „1 vsz-gel elég nagy n-re" azt jelenti, hogy 1 valószínűséggel létezik egy olyan (a kétirányban végtelen $x_{-\infty}^{\infty}$ realizációtól függő) n_0 küszöbszám, hogy az állítás érvényes minden $n \geq n_0$-ra.

Jelölje $N_n(s,a)$ az $s \in A^{l(s)}$ sztring azon előfordulásainak számát az x_1^n mintában, amikor az $a \in A$ betű követi, ahol feltételezzük, hogy az s hossza legfeljebb $\log n$, és – technikai okokból – csak az $i > \log n$ helyen lévő betűket tekintjük:

$$N_n(s,a) = \left| \left\{ i: \ \log n < i \leq n,\, x_{i-l(s)}^{i-1} = s,\, x_i = a \right\} \right|.$$

A logaritmusok e alapúak. Az s ilyen előfordulásait jelölje $N_n(s)$:

$$N_n(s) = \left| \left\{ i: \ \log n < i \leq n,\, x_{i-l(s)}^{i-1} = s \right\} \right|.$$

Adott egy x_1^n minta, egy *alkalmas fa* olyan $\mathcal{T}$ fa $d(\mathcal{T}) \leq \lceil \log n \rceil$ mélységgel, amelyre $N_n(s) \geq 1$ minden $s \in \mathcal{T}$ esetén, és minden egyes s' sztring amelyre $N_n(s') \geq 1$, vagy egy utótagja valamely $s \in \mathcal{T}$-nek, vagy van egy $s \in \mathcal{T}$ utótagja. Egy alkalmas $\mathcal{T}$ fát *r-gyakori*nak nevezünk, ha $N_n(s) \geq r$ minden $s \in \mathcal{T}$-re. Az összes alkalmas illetve r-gyakori fák családját $\mathcal{F}_1(x_1^n)$ illetve $\mathcal{F}_r(x_1^n)$ jelölje.

Világos, hogy

$$\sum_{a \in A} N_n(s,a) = N_n(s) \quad \text{és} \quad \sum_{s \in \mathcal{T}} N_n(s) = n - \lceil \log n \rceil$$

bármely alkalmas $\mathcal{T}$ fára. Egy ilyen $\mathcal{T}$ fát hipotetikus kontextusfának tekintve az x_1^n minta valószínűsége a következőképpen írható:

$$Q(x_1^n) = Q(x_1^{\lceil \log n \rceil}) \prod_{s \in \mathcal{T},\, a \in A} Q(a|\, s)^{N_n(s,a)}.$$

A kifejezést egy kissé szabadon használva egy hipotetikus $\mathcal{T} \in \mathcal{F}_1(x_1^n)$ kontextusfára az $\mathrm{ML}_{\mathcal{T}}(x_1^n)$ maximum likelihoodot úgy definiáljuk, mint a második fenti tényező maximumát $Q(a|\, s)$-ben. Ezáltal

$$\log \mathrm{ML}_{\mathcal{T}}(x_1^n) = \sum_{s \in \mathcal{T},\, a \in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)}.$$

A $\mathcal{T}_0$ becsléséhez két információs kritériumot fogunk használni, mindkettő az MDL elvből származtatható. Egy információs kritérium egy értéket rendel hozzá minden egyes hipotetikus modellhez (jelen esetben kontextusfához) a minta alapján, és a becslő az a modell lesz, amelyre ez az érték a legkisebb.

4.4. Definíció. *Adott az x_1^n minta, a* BIC *egy alkalmas $\mathcal{T}$ fára*

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) = -\log \mathrm{ML}_{\mathcal{T}}(x_1^n) + \frac{(|A|-1)|\mathcal{T}|}{2} \log n.$$

4.5. Megjegyzés. A BIC jellegzetessége a szabad paraméterek száma osztva kettővel szorozva $\log n$-nel ún. büntetéstag. Most egy $\mathcal{T}$ kontextusfájú Q folyamatot a $Q(a|\, s)$, $a \in A$, $s \in \mathcal{T}$ feltételes valószínűségekkel írunk le, és ezek közül $(|A|-1)|\mathcal{T}|$ számú a szabad paraméter, ha a $\mathcal{T}$ fa teljes. Egy nem teljes kontextusfájú folyamat esetén bizonyos sztringek valószínűsége szükségképpen 0, ennélfogva a szabad paraméterek száma tipikusan kisebb, mint $(|A|-1)|\mathcal{T}|$, ha $\mathcal{T}$ nem teljes. Így a 4.4 definíció némileg eltér a BIC szokásos kifejezésétől. Megjegyezzük, hogy a 4.4 definícióban az $(|A|-1)/2$ helyettesítése bármely $c > 0$-val nem befolyásolná az eredményeket és azok bizonyításait. □

Ismert (Csiszár and Shields, 2000), hogy a Markov láncok rendjének becslésére a BIC becslő konzisztens a hipotetikus rend bármilyen megszorítása nélkül. Az alábbi tétel igényel korlátot a hipotetikus kontextusfa mélységére. Mégis, minthogy ez a korlát növekszik a minta n méretével, nincs szükség egy előzetes korlátra az ismeretlen $\mathcal{T}_0$ méretéről, sőt $d(\mathcal{T}_0) = \infty$ is megengedett. Megjegyezzük, hogy e korlát jelenléte csökkenti a számítási bonyolultságot.

4.6. Tétel. *Abban az esetben, amikor* $d(\mathcal{T}_0) < \infty$, *a*

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \arg \min_{\mathcal{T}\in\mathcal{F}_1(x_1^n)\cap\mathcal{I},\, d(\mathcal{T})\le D(n)} \mathrm{BIC}_{\mathcal{T}}(x_1^n)$$

BIC becslőre, ha $D(n) = o(\log n)$, *teljesül*

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \mathcal{T}_0$$

1 *vsz-gel elég nagy* n*-re.*

Általános esetben a

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \arg \min_{\mathcal{T}\in\mathcal{F}_{n^\alpha}(x_1^n)\cap\mathcal{I},\, d(\mathcal{T})\le D(n)} \mathrm{BIC}_{\mathcal{T}}(x_1^n)$$

BIC becslőre, ha $D(n) = o(\log n)$ *és* $0 < \alpha < 1$ *tetszőleges, teljesül bármely* K *konstans esetén*

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n)\big|_K = \mathcal{T}_0\big|_K$$

1 *vsz-gel elég nagy* n*-re.*

A 4.6 és 4.9 tételek bizonyítása. Ennek és a következő tételnek a bizonyítása a Markov láncokra vonatkozó nagymintás tipikussági eredményekre (Csiszár, 2000) épül. □

4.7. Megjegyzés. Itt és a következő 4.9 tételben a jelzett minimum biztosan felvétetik, mivel az alkalmas fák száma véges, de a minimalizáló fa nem szükségképpen egyértelmű; ebben az esetben bármelyik minimalizáló fa vehető arg min-ként. □

Az általunk vizsgált másik információs kritérium a Kricsevszkij – Trofimov kódhossz (ld. (Krichevsky and Trofimov, 1981), (Willems, Shtarkov and Tjalkens, 1995)). Jegyezzük meg, hogy egy, az alábbi $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ hosszfüggvényű kód minimalizálja a legrosszabb esetbeli átlagos redundanciát egy additív konstans erejéig, a $\mathcal{T}$ kontextusfájú folyamatok osztályára.

4.8. Definíció. *Adott az* x_1^n *minta, a* KT *kritérium egy alkalmas* $\mathcal{T}$ *fára*

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = -\log P_{\mathrm{KT},\mathcal{T}}(x_1^n),$$

ahol

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^{\lceil \log n\rceil}} \prod_{s\in\mathcal{T}} \frac{\prod_{a:N_n(s,a)\ge 1}\left[\left(N_n(s,a)-\frac{1}{2}\right)\left(N_n(s,a)-\frac{3}{2}\right)\dots\left(\frac{1}{2}\right)\right]}{\left(N_n(s)-1+\frac{|A|}{2}\right)\left(N_n(s)-2+\frac{|A|}{2}\right)\dots\left(\frac{|A|}{2}\right)}$$

a $\mathcal{T}$ *szerinti KT-valószínűsége az* x_1^n*-nek.*

A Markov láncok rendjének becslésére bizonyítást nyert a KT becslő konzisztenciája akkor, amikor a hipotetikus rendek nagyságrendje $o(\log n)$ (Csiszár, 2002), míg a rendre vonatkozó bármilyen korlát nélküli esetben, vagy egy olyan korlát esetén, ami egyenlő egy elég nagy konstans szorozva $\log n$-nel, ismert egy ellenpélda a KT becslő konzisztenciájára (Csiszár and Shields, 2000).

4.9. Tétel. *Abban az esetben, amikor* $d(\mathcal{T}_0) < \infty$, *a*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{KT}_{\mathcal{T}}(x_1^n)$$

BIC becslőre, ha $D(n) = o(\log n)$, *teljesül*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \mathcal{T}_0$$

1 *vsz-gel elég nagy* n*-re.*

Általános esetben a

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{KT}_{\mathcal{T}}(x_1^n)$$

KT becslőre, ha $D(n) = o(\log n)$ *és* $0 < \alpha < 1$ *tetszőleges, teljesül bármely* K *konstans esetén*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n)\big|_K = \mathcal{T}_0\big|_K$$

1 *vsz-gel elég nagy* n*-re.*

4.10. Megjegyzés. Szigorúan véve az MDL elv értelmében minimalizálni a $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ „kódhossznak” és egy tagnak, a „$\mathcal{T}$ kódhosszának” (ezt a $\mathcal{T}$ költségének nevezte Willems, Shtarkov és Tjalkens (1995)) az összegét kellene. Az utóbbi tagot figyelmen kívül hagyjuk, mivel ez nem befolyásolja a konzisztencia eredményt. □

A gyakorlatban kivihetetlen a becslőket úgy kiszámítani, hogy az információs kritérium értékét minden egyes modellre kiszámoljuk, mert a hipotetikus kontextusfák száma nagyon nagy. Azonban az értekezésbeli algoritmusok lehetővé teszik a tárgyalt becslők kiszámítását megvalósítható számítási bonyolultsággal.

A megszokott módon feltételezzük, ld. (Baron and Bresler, 2004), (Martín, Seroussi and Weinberger, 2004), hogy a számításokat $O(\log n)$ regisztermérettel végezzük.

Egyaránt tekintünk off-line és on-line módszereket. Megjegyezzük, hogy a becslő on-line számítása hasznos amikor a minta mérete nem rögzített, hanem folyamatosan mintavételezünk amíg a becslő „stabillá” nem válik, mondjuk állandó marad ha a minta mérete kétszeresére nő.

4.11. Tétel. *A 4.6 és 4.9 tételekbeli BIC becslőnek és KT becslőnek egy adott* x_1^n *mintára történő meghatározásához szükséges számítások száma* $O(n)$, *ami elérhető* $O(n^\varepsilon)$ *számú adat tárolásával, ahol* $\varepsilon > 0$ *tetszőleges.*

4.12. Tétel. *Adott az* x_1^n *minta, a 4.9 tételbeli KT becslőnek az összes* x_1^i, $i \le n$ *részmintára történő meghatározásához szükséges számítások száma* $o(n \log n)$, *ami elérhető minden időpontban* $O(n^\varepsilon)$ *számú adat tárolásával, ahol* $\varepsilon > 0$ *tetszőleges.*

Ugyanez érvényes a 4.6 tételbeli BIC becslőre a BIC definíciójának kis módosításával. Konkrétan, jelölje k_m, $m \in \mathbb{N}$ *azt a legkisebb* k *egész számot, amelyre* $D(k) = m$, *és cseréljük ki* n*-et a büntetéstagban a 4.4 definícióban a* $\{k_m\}$ *sorozat legkisebb olyan elemére, ami nagyobb* n*-nél.*

A 4.11 és 4.12 tételek bizonyítása. Ezeket a tételeket a Kontextusfa Maximalizáló (CTM) eljárás (Willems, Shtarkov és Tjalkens, 1993, 2000) kiterjesztésének útján bizonyítjuk. □

5. Alapkörnyezet konzisztens becslése Markov mezőkre

Tekintsük a d-dimenziós $\mathbb{Z}^d$ *rács*ot. A $i \in \mathbb{Z}^d$ pontokat rácspontoknak nevezzük, és $\|i\|$ jelöli az i maximum normáját, azaz a koordinátái abszolútértékeinek maximumát. Egy véges Δ halmaz számosságát $|\Delta|$ jelöli. A tartalmazás és a szigorú tartalmazás $\subseteq$ és $\subset$ jelöléseit megkülönböztetjük az értekezésben.

Egy *véletlen mező* a rács rácspontjai által indexelt valószínűségi változók egy családja: $\{ X(i) : i \in \mathbb{Z}^d \}$, ahol minden egyes $X(i)$ egy valószínűségi változó egy véges A halmazbeli értékekkel. A rács egy $\Delta \subseteq \mathbb{Z}^d$ tartományára azt írjuk, hogy $X(\Delta) = \{ X(i) : i \in \Delta \}$. Az $X(\Delta)$ realizációira az $a(\Delta) = \{ a(i) \in A : i \in \Delta \}$ jelölést használjuk. Ha Δ véges, a $|\Delta|$ komponensű $a(\Delta) \in A^\Delta$-kra *blokkok*ként fogunk hivatkozni.

Az $X(i)$ valószínűségi változók együttes eloszlását jelölje Q. Feltételezzük, hogy ennek a véges dimenziós marginálisai szigorúan pozitívak, azaz

$$Q(a(\Delta)) = \mathrm{Prob}\{ X(\Delta) = a(\Delta) \} > 0 \quad \text{minden véges } \Delta \subset \mathbb{Z}^d\text{-re, } a(\Delta) \in A^\Delta.$$

E szokásos feltételezés lehetővé teszi a feltételes valószínűségek egyértelmű definícióját:

$$Q(a(\Delta)|\, a(\Phi)) = \mathrm{Prob}\{ X(\Delta) = a(\Delta) \mid X(\Phi) = a(\Phi) \}$$

minden véges diszjunkt Δ és Φ tartományra.

Egy Γ *környezet*en (amely a 0 origó környezete) a rácspontok egy véges, középpontosan szimmetrikus halmazát értjük, ahol $0 \notin \Gamma$. Ennek a sugara $r(\Gamma) = \max_{i \in \Gamma} \|i\|$. Bármely $\Delta \subseteq \mathbb{Z}^d$-re a Δ eltoltját, amint a 0-t az i-be toljuk, Δ^i-vel jelöljük. Egy Γ környezet Γ^i eltoltját az i rácspont egy Γ környezetének fogjuk nevezni, ld. 2. ábra.

Egy *Markov mező* egy olyan fenti véletlen mező, amelyre létezik egy Γ környezet, amit *Markov környezet*nek fogunk nevezni, hogy minden $i \in \mathbb{Z}^d$ esetén

$$Q(a(i)|\, a(\Delta^i)) = Q(a(i)|\, a(\Gamma^i)) \quad \text{ha } \Delta \supset \Gamma,\ 0 \notin \Delta, \tag{2}$$

ahol az utóbbi feltételes valószínűség eltolás-invariáns.

Ez a fogalom azonos a véges kölcsönhatás-távolságú Gibbs mezőével, ld. Georgii (1988). E tény által ösztönözve a

$$Q_\Gamma = \left\{ Q_\Gamma(a|\, a(\Gamma)) : a \in A,\ a(\Gamma) \in A^\Gamma \right\}$$

mátrixot, amely meghatározza a (pozitív, eltolás-invariáns) feltételes valószínűségeket (2)-ben, *egypontú specifikáció*nak fogjuk nevezni. Az olyan eloszlásokat $A^{\mathbb{Z}^d}$-en, amelyek kielégítik (2)-t egy adott Q_Γ feltételes-valószínűség mártixszal, *Gibbs eloszlások*nak nevezzük Q_Γ egypontú specifikációval. Az adott Markov mező Q eloszlása egy ezek közül; Q nem szükségképpen eltolás-invariáns.

A következő lemma összefoglal néhány jól ismert tényt.

5.1. Lemma. *A rácson értelmezett fenti Markov mezőre létezik egy olyan Γ_0 környezet, amellyel a Markov környezetek pontosan azok lesznek, amelyek tartalmazzák Γ_0-at. Továbbá, a*

$$Q(a(\Delta)|\, a(\mathbb{Z}^d \setminus \Delta)) = Q(a(\Delta)|\, a(\cup_{i \in \Delta} \Gamma_0^i \setminus \Delta))$$

globális Markov tulajdonság teljesül minden egyes véges $\Delta \subset \mathbb{Z}^d$ tartományra. Ezek a feltételes valószínűségek eltolás-invariánsak és egyértelműen meghatározottak a Q_{Γ_0} egypontú specifikáció által.

Az 5.1 lemmabeli legkisebb Γ_0 Markov környezetet *alapkörnyezet*nek fogjuk nevezni.

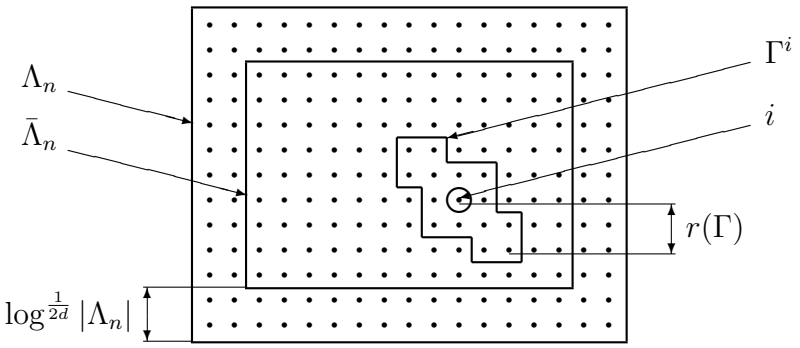
Az értekezésben a Γ_0 alapkörnyezet statisztikai becslésével foglalkozunk, a Markov mező egy realizációjának növekvő véges $\Lambda_n \subset \mathbb{Z}^d$, $n \in \mathbb{N}$ tartományokon adott megfigyeléséből; így az n-edik minta $x(\Lambda_n)$.

A statisztikai következtetést egy lehetséges Γ alapkörnyezetről az $x(\Lambda_n)$ mintában előforduló $a(\Gamma) \in A^\Gamma$ blokkok alapján fogjuk levonni. Technikai okokból csak azokat a blokkokat fogjuk tekinteni, amelyek középpontja a Λ_n egy $\bar{\Lambda}_n$ résztartományába esik, ahol $\bar{\Lambda}_n$ azon $i \in \Lambda_n$ rácspontokból áll, amelyekre az i középpontú és $\log^{\frac{1}{2d}} |\Lambda_n|$ sugarú gömb Λ_n-be esik:

$$\bar{\Lambda}_n = \left\{ i \in \Lambda_n : \left\{ j \in \mathbb{Z}^d : \|i - j\| \le \log^{\frac{1}{2d}} |\Lambda_n| \right\} \subseteq \Lambda_n \right\},$$

ld. 2. ábra. A logaritmusok e alapúak. Csupán azt fogjuk feltételezni a Λ_n mintatartományokról, hogy

$$\Lambda_1 \subset \Lambda_2 \subset \ldots; \quad |\Lambda_n| \,/\, \left|\bar{\Lambda}_n\right| \to 1.$$



2. ábra. Az i rácspont Γ környezete és a Λ_n mintatartomány.

Minden egyes $a(\Gamma) \in A^\Gamma$ blokkra jelölje $N_n(a(\Gamma))$ az $a(\Gamma)$ blokk azon előfordulásainak számát az $x(\Lambda_n)$ mintában, amikor a középpontja $\bar{\Lambda}_n$-ben van:

$$N_n(a(\Gamma)) = \left| \left\{ i \in \bar{\Lambda}_n : \Gamma^i \subseteq \Lambda_n,\, x(\Gamma^i) = a(\Gamma) \right\} \right|.$$

Azokat a blokkokat, amelyeket a középpontjaikkal kiegészített Γ környezetek alkotják, röviden $a(\Gamma, 0)$-val fogjuk jelölni. Hasonlóan az előzőekhez, minden egyes $a(\Gamma, 0) \in A^{\Gamma \cup \{0\}}$-ra azt írjuk, hogy

$$N_n(a(\Gamma, 0)) = \left| \left\{ i \in \bar{\Lambda}_n : \Gamma^i \subseteq \Lambda_n,\, x(\Gamma^i \cup \{i\}) = a(\Gamma, 0) \right\} \right|.$$

Az $a(\Gamma, 0) \in x(\Lambda_n)$ jelölés azt fogja jelenteni, hogy $N_n(a(\Gamma, 0)) \ge 1$.

A $\Gamma^i \subseteq \Lambda_n$ megkötés a fenti definíciókban automatikusan teljesül, amikor $r(\Gamma) \leq \log^{\frac{1}{2d}} |\Lambda_n|$. Ennélfogva minden környezet esetén, kivéve a nagyon nagyokat, ugyanannyi blokkot veszünk számításba:

$$\sum_{a(\Gamma)\in A^\Gamma} N_n(a(\Gamma)) = \left|\bar{\Lambda}_n\right| \quad \text{ha } r(\Gamma) \leq \log^{\frac{1}{2d}} |\Lambda_n|.$$

A Markov mezőkre a likelihood függvényt nem lehet explicit alakban meghatározni. Helyette az alábbiakban definiált pszeudo-likelihoodot fogjuk használni.

Adott az $x(\Lambda_n)$ minta, egy Γ környezetre a *pszeudo-likelihood* függvény az alábbi függvénye a Q'_Γ mátrixnak, ahol Q'_Γ-t egy olyan hipotetikus Markov mező egypontú specifikációjának tekintjük, amelyre Γ egy Markov környezet:

$$\mathrm{PL}_\Gamma(x(\Lambda_n), Q'_\Gamma) = \prod_{i\in\bar{\Lambda}_n} Q'_\Gamma(x(i)|\, x(\Gamma^i)) = \prod_{a(\Gamma,0)\in x(\Lambda_n)} Q'_\Gamma(a(0)|\, a(\Gamma))^{N_n(a(\Gamma,0))}.$$

Megjegyezzük, hogy nem minden olyan Q'_Γ mátrix lehetséges egypontú specifikáció, amelyre teljesül

$$\sum_{a\in A} Q'_\Gamma(a(0)|\, a(\Gamma)) = 1, \quad a(\Gamma) \in A^\Gamma,$$

egy egypontú-specifikáció mátrix elemeinek eleget kell tenniük különféle algebrai összefüggéseknek, amelyeket itt nem részletezünk. Mindazonáltal a pszeudo-likelihoodot olyan Q'_Γ-kre is definiáljuk, amelyek nem elégítik ki ezeket az összefüggéseket, megengedve akár azt, hogy Q'_Γ néhány eleme 0 legyen.

A pszeudo-likelihood maximuma $Q'_\Gamma(a(0)|\, a(\Gamma)) = \frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))}$ esetén éretik el. Így, adott $x(\Lambda_n)$ mintára, a *maximum pszeudo-likelihood* logaritmusa egy Γ környezetre

$$\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) = \sum_{a(\Gamma,0)\in x(\Lambda_n)} N_n(a(\Gamma,0)) \log \frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))}.$$

Meg tudunk fogalmazni egy olyan kritériumot a Bayes-féle információs kritérium analógiájára, amelyet ki lehet számítani a mintából.

5.2. Definíció. *Adott az $x(\Lambda_n)$ minta, a Pszeudo Bayes-féle Információs Kritérium, röviden* PIC, *a Γ környezetre*

$$\mathrm{PIC}_\Gamma(x(\Lambda_n)) = -\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) + |A|^{|\Gamma|} \log |\Lambda_n|\,.$$

5.3. Megjegyzés. *A fenti büntetéstagban a lehetséges $a(\Gamma) \in A^\Gamma$ blokkok $|A|^{|\Gamma|}$ száma helyettesíti a BIC-ben megjelenő „szabad paraméterek számának a felét", amelyre nem áll rendelkezésre egyszerű kifejezés. Megjegyezzük, hogy az eredményeink érvényesek maradnak ugyanazokkal a bizonyításokkal, ha a fenti büntetéstagot beszorozzuk tetszőleges $c > 0$-val.* □

A Γ_0 alapkörnyezet PIC becslőjét úgy definiáljuk, mint azt a hipotetikus Γ-t, amelyre a kritérium értéke a legkisebb. Fontos jellemzője a becslőnknek, hogy a hipotetikus Γ-k családja növekedhet amint $n \to \infty$, így nincs szükség előzetes felső

korlátra az ismeretlen Γ_0-ról. A főeredményünk szerint a PIC becslő erősen konzisztens amikor a hipotetikus Γ-k azok, amelyekre $r(\Gamma) \leq r_n$, ahol r_n megfelelően lassan növekszik.

Erős konzisztencián azt érjük, hogy a becsült alapkörnyezet egyenlő Γ_0-val, 1 vsz-gel elég nagy n-re. Itt és a továbbiakban az „1 vsz-gel elég nagy n-re" azt jelenti, hogy 1 valószínűséggel létezik egy olyan (a végtelen $x(\mathbb{Z}^d)$ realizációtól függő) n_0 küszöbszám, hogy az állítás érvényes minden $n \geq n_0$-ra.

5.4. Tétel. *A*

$$\widehat{\Gamma}_{\mathrm{PIC}}(x(\Lambda_n)) = \arg \min_{\Gamma:\, r(\Gamma)\leq r_n} \mathrm{PIC}_\Gamma(x(\Lambda_n))$$

PIC becslőre, ha

$$r_n = o\left(\log^{\frac{1}{2d}} |\Lambda_n|\right),$$

teljesül

$$\widehat{\Gamma}_{\mathrm{PIC}}(x(\Lambda_n)) = \Gamma_0$$

1 *vsz-gel elég nagy n-re.*

Bizonyítás. A túlbecslés kizárhatóságát Besag „kódolási technikája" (Besag, 1974) és nagy eltérések felhasználásával bizonyítjuk. Az alulbecslés kizárhatóságát entrópia számítással bizonyítjuk. □

5.5. Megjegyzés. *Valójában az állítást olyan r_n-re fogjuk bizonyítani, ami egyenlő egy konstans szorozva $\log^{\frac{1}{2d}} \left|\bar{\Lambda}_n\right|$-kel. Azonban, minthogy ez a konstans függ az ismeretlen Q eloszlástól, a konzisztenciát csak akkor lehet biztosítani, ha*

$$r_n = o\left(\log^{\frac{1}{2d}} \left|\bar{\Lambda}_n\right|\right) = o\left(\log^{\frac{1}{2d}} |\Lambda_n|\right).$$

Nyitott az a kérdés, hogy vajon a konzisztencia igaz marad-e akkor, amikor a hipotetikus környezetek gyorsabban nőhetnek, vagy akár a hipotetikus környezetekre vonatkozó mindenfajta feltétel nélkül. □

6. Irodalomjegyzék

Az értekezés részeit tartalmazó publikációk

Csiszár, I. and Talata, Zs. (2004a). Consistent Estimation of the Basic Neighborhood of Markov Random Fields. *Ann. Statist.* Elfogadva.

Csiszár, I. and Talata, Zs. (2004b). Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL. *IEEE Trans. Inform. Theory.* Benyújtva.

Talata, Zs. (2004). Model Selection via Information Criteria. *Period. Math. Hungar.* Felkért publikáció.

Az értekezésben előforduló hivatkozások

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203–217.

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory,* Supplement to Problems of Control and Information Theory (B. N. Petrov and F. Csáki, eds.) 267–281. Akadémia Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19** 716–723.

Akaike, H. (1977). On entropy maximization principle. In *Application of Statistics* (P.R. Krishnaiah, ed.) 27–41. North-Holland, Amsterdam.

Anderson, T.W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **33** 255–265.

Anderson, T.W. (1963). Determination of the order of dependence in normally distributed time series. In *Time series analysis* (M. Rosenblatt, ed.) 425–446. Wiley, New York.

Azencott, R. (1987). Image analysis and Markov fields. In *Proceedings of the First International Conference on Applied Mathematics, Paris* (J. McKenna and R. Temen, eds.) 53–61. SIAM, Philadelphia.

Baron, D. and Bresler, Y. (2004). An $O(N)$ semipredictive universal encoder via the BWT. *IEEE Trans. Inform. Theory* **50** 928–937.

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44** 2743–2760.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

Bühlmann, P. and Wyner, A.J. (1999). Variable length Markov chains. *Ann. Statist.* **27** 480–513.

Comets, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.* **20** 455–468.

Csiszár, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory* **48** 1616–1629.

Csiszár, I. and Shields, P.C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28** 1601–1619.

Davisson, L.D. (1965). Prediction error of stationary Gaussian time series of unknown variance. *IEEE Trans. Inform. Theory* **19** 783–795.

Dobrushin, R.L. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab. Appl.* **13** 197–224.

Finesso, L. (1992). Estimation of the order of a finite Markov chain. In *Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing, I* (H. Kimura and S. Kodama, eds.) 643–645. Mita Press, Tokyo.

Geman, S. and Graffigne, C. (1987). Markov random fields image models and their applications to computer vision. In *Proceedings of the International Congress Mathematicians* (A. M. Gleason, ed.) **2** 1496–1517. Amer. Math. Soc., Providence, R.I.

Georgii, H.O. (1988). *Gibbs Measures and Phase Transitions.* de Gruyter, Berlin.

Gerencsér, L. (1987). Order estimation of stationary Gaussian ARMA processes using Rissanen's complexity. Working paper, Computer and Automation Institute of the Hungarian Academy of Sciences.

Gidas, B. (1988). Consistency of maximum likelihood and pseudolikelihood estimators for Gibbs distributions. *Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl.* **10** 129–145.

HAMERLY, E.M. and DAVIS, M.H.A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.* **17** 941–946.

HANNAN, E.J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081.

HANNAN, E.J. and QUINN, B.G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

HAUGHTON, D. (1988). On the choice of model to fit data from an exponential family. *Ann. Statist.* **16** 342–355.

KRICHEVSKY, R.E. and TROFIMOV, V.K. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27** 199–207.

MALLOWS, C. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the IMS, Manhattan, Kansas.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MARTÍN, A., SEROUSSI, G. and WEINBERGER, M.J. (2004). Linear time universal coding and time reversal of tree sources via FSM closure. *IEEE Trans. Inform. Theory* **50** 1442–1468.

NEYMAN, J. and PEARSON, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* **20A** 263–294.

PICKARD, D.K. (1987). Inference for discrete Markov field: The simplest non-trivial case. *J. Amer. Statist. Assoc.* **82** 90–96.

RÉNYI, A. (1970). *Probability Theory.* American Elsevier Publishing Co., Inc., New York.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

RISSANEN, J. (1983a). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

RISSANEN, J. (1983b). A universal data compression system. *IEEE Trans. Inform. Theory* **29** 656–664.

RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore.

RISSANEN, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory* **42** 40–47.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126.

SHTARKOV, J. (1977). Coding of discrete sources with unknown statistics. In *Topics in Information Theory* (I. Csiszár and P. Elias, eds.) 559–574. North-Holland, Amsterdam.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47.

WEINBERGER, M.J., LEMPEL, A. and ZIV, J. (1992). A sequentional algorithm for the universal coding of finite memory sources. *IEEE Trans. Inform. Theory* **38** 1002–1014.

WEINBERGER, M.J., RISSANEN, J. and FEDER, M. (1995). A universal finite memory source. *IEEE Trans. Inform. Theory* **41** 643–652.

WILLEMS, F.M.J. (1998). The context-tree weighting method: Extensions. *IEEE Trans. Inform. Theory* **44** 792–798.

WILLEMS, F.M.J., SHTARKOV, Y.M. and TJALKENS, T.J. (1993). The context-tree weighting method: Basic properties. *Tech. Rep., EE Dept., Eindhoven University.* An earlier unabridged version of (Willems, Shtarkov and Tjalkens, 1995).

WILLEMS, F.M.J., SHTARKOV, Y.M. and TJALKENS, T.J. (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inform. Theory* **41** 653–664.

WILLEMS, F.M.J., SHTARKOV, Y.M. and TJALKENS, T.J. (2000). Context-tree maximizing. In *Proc. 2000 Conf. Information Sciences and Systems* TP6-7–TP6-12. Princeton, NJ.