Véletlen permutációk statisztikai vizsgálata

Doktori értekezés

Készítette: Csiszár Villő
Témavezető: Tusnády Gábor
tudományos tanácsadó,
az MTA rendes tagja

Matematika Doktori Iskola
Iskolavezető: Laczkovich Miklós
Alkalmazott Matematika Doktori Program
Programvezető: Michaletzky György

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2008

Tartalomjegyzék

Ábrák jegyzéke

Táblázatok jegyzéke

Jelölések

$[n]$:	$[n] = \{1,2,\ldots,n\}$
S_n:	az n-edfokú szimmetrikus csoport
$\overline{U}$:	$[n] \setminus U$, az U halmaz komplementere
$\pi(U)$, $\pi\{U\}$:	lásd (1)
$\pi(i..j)$, $\pi\{i..j\}$:	lásd (15)
$U \perp \overline{U} \mid \emptyset$:	$\pi(U)$ és $\pi(\overline{U})$ feltételesen függetlenek $\pi\{U\}$-ra nézve
$\mathbf{F}(M)$:	az M mátrix szerint faktorizálódó eloszlások, lásd (2)
$\mathbf{E}(M)$:	az M-hez tartozó exponenciális család, lásd (3)
$\mathrm{Supp}(v)$:	a v vektor tartója
$\mathrm{cl}(\cdot)$:	lezárás az euklideszi térben
X_M:	az M-hez tartozó nemnegatív torikus varietás, lásd (4)
I_M:	az M-hez tartozó torikus ideál
$\chi\{A\}$:	indikátorfüggvény: 1, ha A teljesül, különben 0
$\mathbf{L}$ ill. $\mathbf{L}'$:	az L- ill. invertálva L-felbontható eloszlások családja
$\mathbf{L}_K$:	a K-beli k-knál felbontható eloszlások családja
M_L ill. $M_{L'}$:	az $\mathbf{L}$ ill. $\mathbf{L}'$ torikus modellhez tartozó mátrix
$\mathbf{B}$:	a duplán L-felbontható eloszlások családja
F ill. F':	az $M_L^\top$ ill. $M_{L'}^\top$ operátor képtere, $F \cap F' = G$
$U \perp_\cap V$:	az U és V altér merőlegesen metszi egymást
Pr_U:	az U altérre való merőleges vetítés
$\lvert\pi(\mathcal{D} \times \mathcal{R})\rvert$:	a π durvítása a szorzatpartícióra, lásd (24)
$U_\mathcal{P}$:	a $\mathcal{P}$ partícióhoz rendelt altér, lásd (25)
$\mathcal{L}(\mathcal{P}_1,\ldots,\mathcal{P}_s)$:	hierarchikus modell, lásd 9.2. Definíció
$\mathcal{D}' \succeq \mathcal{D}$:	a $\mathcal{D}'$ partíció finomabb $\mathcal{D}$-nél
M_B:	a $\rho_{aq}^{k\ell}$ sorokból alkotott mátrix, lásd (32)
σ_r:	$\sigma_r = (n(n-1)\ldots 21)$ az id identitás megfordítása
$\sigma_{(12)}$:	$\sigma_{(12)} = (213\ldots n)$
$\phi_{\circ\sigma}$ ill. $\phi_{\sigma\circ}$:	a σ-val való jobbról- ill. balról szorzás

I. rész

Előkészületek

1. Az értekezés témája és felépítése

Értekezésem a véletlen permutációk statisztikai vizsgálatához kapcsolódó eredményeimet összegzi. Tegyük fel, hogy valamilyen véletlen kísérlet eredményeként rögzített n hosszúságú permutációkat kapunk adatként. Legalább három alapvető esetet különböztethetünk meg aszerint, hogy a π adat-permutációk mit fejeznek ki.

- A permutáció tömör leírása lehet két azonos elemszámú halmaz összepárosításának. Ha az A és B halmaz elemeit 1-től n-ig számozzuk, akkor $\pi(i) = j$ fejezheti ki azt, hogy az A halmaz i. eleme a B halmaz j. elemével áll párban (vagy fordítva). Ha egy táncos rendezvényen ugyanannyi férfi és nő vesz részt, és mindenki mindig táncol, feljegyezhetjük az egyes táncok során a párosításokat. Ha nem mindig táncol mindenki, akkor csak részleges párosításaink lesznek.

- A permutáció kifejezheti egy halmaz elemeinek sorbarendezését is. Ismét jelöljük a halmaz elemeit az $1, \ldots, n$ számokkal. Ekkor beszélhetünk a π sorrend-permutációról, amikor $\pi(i) = j$ azt jelenti, hogy a sorrend i. pozíciójában a halmaz j-vel jelölt eleme áll, vagy ennek inverzéről, a π^{-1} helyezés-permutációról. Azaz $\pi^{-1}(i) = j$ jelentése: az i-vel jelölt elem a j. helyen szerepel a sorrendben. Ilyen adatokat kapunk például akkor, amikor bírálók pályázatokat rangsorolnak. Ebben az esetben is előfordulhat, hogy nem teljes permutációt kapunk, ha például a holtverseny is megengedett, vagy a bírálók csak az általuk legjobbnak tartott néhány pályázatot rangsorolják. Az irodalomban a legtöbbet ezzel az esettel foglalkoztak.

- A harmadik esetben a permutáció egy rendezett halmaz átrendezését fejezi ki. $\pi(i) = j$ jelentheti azt, hogy az eredeti sorrend szerinti i. elem

az új sorrendben a j. helyen áll (vagy fordítva). Gondoljunk például arra, hogy egy irodában halomban áll n dosszié. A titkárnő mindig kikeresi az éppen szükségeset, majd a halom legtetejére teszi vissza. Kérdezhetjük, hogy a nap végére hogyan változik meg a dossziék eredeti sorrendje.

Az adatelemzés első lépése minden esetben az adatokkal való ismerkedés: az adatok grafikus megjelenítése, alapstatisztikák kiszámítása. Az adatmegjelenítés permutációk esetében a magas dimenzionalitás miatt nem rutinfeladat. Ha az n hosszú permutációkat, mint $\mathbb{R}^n$-beli vektorokat tekintjük, akkor konvex burkuk az úgynevezett permutáció-politóp (Yemelichev et al. [62]). A politóp csúcsai egy $n-1$ dimenziós gömb felszínén helyezkednek el. Egy permutációkból álló adatsort úgy ábrázolhatunk, hogy a politóp csúcsaiba gömböket helyezünk el, melyek sugara a megfigyelt gyakoriság monoton növő függvénye (Thompson [58] például azt javasolja, hogy a sugár a gyakoriság 5/7-edik hatványával legyen arányos). A dimenzionalitás miatt ez a módszer csak $n \leq 4$ esetén igazán hasznos. Magasabb dimenzióban a fenti ábrázolás helyett annak érdekes alacsony dimenziós vetületeit jeleníthetjük meg, az érdekes vetületek meghatározása hasonlóan történhet, mint általában a sokdimenziós adatok esetében. Az adatmegjelenítés kérdésével ebben az értekezésben a továbbiakban nem foglalkozom.

Az adatokkal való első ismerkedés után a megfigyelésekre elfogadható modellt keresünk. Paraméteres modell illesztésekor először megkeressük az adott modellen belül a mintához leginkább megfelelő paramétereket (paraméterbecslés), majd vizsgáljuk a modell illeszkedésének jóságát (hipotézisvizsgálat). Értekezésemben egyrészt az irodalomban jelen lévő modellekkel kapcsolatban vezetek le új eredményeket, másrészt új modelleket vezetek be, és azok tulajdonságait vizsgálom.

A 2. fejezetben foglalom össze azokat az eszközöket, melyekre a későbbiekben szükség lesz. Az algebrai statisztika alapészrevétele az, hogy számos paraméteres eloszláscsalád algebrai varietást alkot, azaz a család elemei polinomiális egyenleteket elégítenek ki. Ezek a polinomiális egyenletek algoritmikusan megtalálhatók, és felhasználhatók például Monte Carlo algoritmusok futtatásához. A második eszköz a véges eseménytéren definiált eloszlások

exponenciális családjainak, illetve hierarchikus modelljeinek elmélete. Exponenciális családokban általános tételek szólnak a maximum likelihood becslés létezéséről és aszimptotikus tulajdonságairól. A ML becslés kiszámítására a hierarchikus modellek esetében egyszerűen programozható iteratív eljárás az iteratív arányos illesztés. Végül bemutatom az EM- és az MM-algoritmust, melyek sok esetben használhatók a likelihood numerikus maximalizálására.

A II. részben a véletlen permutációkra illeszthető ismert modelleket tárgyalom, illetve ezekkel kapcsolatos néhány új eredményt mutatok be. Először a 3. fejezetben áttekintem a modelleket, illetve ezzel párhuzamosan szót ejtek néhány fontos alapstatisztikáról, hiszen ezek szerepet játszanak egyes modellek paramétereinek becslésénél. Természetesen nem célom a létező összes modell felsorolása (ez nem is lenne lehetséges), a legfontosabbakat, illetve a disszertáció további részéhez leginkább kapcsolódókat igyekeztem összegyűjteni.

A 4. fejezetben McCullagh egy sejtését bizonyítom. A fejezet eredményeit a [21] dolgozatban írtam le. McCullagh [49] egy új modellcsaládot vezetett be véletlen permutációkra, de azt csak sejtésként mondta ki, hogy a modellek általa megadott paraméterei identifikálhatók (azaz különböző paraméterekhez kölönböző eloszlások tartoznak). A sejtés bizonyítása a következő állításon múlik. Definiáljunk S_n-en egy irányított gráfot. Akkor vezet él π-ből σ-ba, ha π egy elemét „helyre rakva” (a többi elemet odébb csúsztatva) σ-t kapjuk. Például a $\pi = (24531)$ permutációból a 4-et „helyre rakva” a $\sigma = (25341)$ permutációt kapjuk. Az állítás az, hogy ebben a gráfban nincs irányított kör. Ezt a gráfot, illetve az általa definiált részbenrendezést (amennyiben ez a részbenrendezés valóban új) úgy érzem, érdemes lenne tovább vizsgálni, bár ezek a vizsgálatok már messze vezetnének a statisztika témakörétől. Az 5. fejezet kis kitérőt tesz a szimmetrikus csoporton megadható, statisztikai szempontból is érdekes részbenrendezések világába.

A 6. fejezetben bemutatok néhány EM algoritmust általánosított Bradley-Terry modellekre, illetve a rendezett minta modellre. Ezekre az eredményekre a [22] dolgozatban történik utalás. A Bradley-Terry modellben két vagy több versenyző sorrendjét a versenyzőkhöz tartozó, független, különböző paraméterű exponenciális eloszlású változók sorrendje határozza meg. Ezt

általánosítani lehet például úgy, hogy csapatok versenyeznek egymással. Ezekben a modellekben a paraméterek maximum likelihood becslésének kiszámítására Hunter [40] MM algoritmusokat javasolt. Megmutatom, hogy ezekre a feladatokra EM algoritmusok is megadhatók, bár ezek – szimulációs vizsgálatok szerint – az MM algoritmusoknál lassabban konvergálnak. Az EM algoritmus viszont arra az esetre is alkalmazható, amikor a Bradley-Terry modell exponenciális változóit tetszőleges eloszlásokkal helyettesítjük, erről szól a 7. fejezet.

A III. rész az értekezés leghosszabb és legfontosabb része. Központi gondolata a feltételes függetlenség és a hierarchikus modell véletlen permutációkra való alkalmazása. Egy véletlen permutáció n dimenziós diszkrét valószínűségi változó, minden koordinátája az $[n] = \{1, \ldots, n\}$ halmaz eleme. A koordinátáknak azonban különbözőknek kell lenniük. Eszünkbe juthat itt a kontingenciatáblák elemzése, amikor olyan valószínűségi változókat vizsgálunk, melyek egy $[m_1] \times \ldots \times [m_n]$ szorzathalmazban veszik fel értékeiket. Ezekre szokás úgynevezett hierarchikus modelleket illeszteni, melyekben egy cella valószínűsége bizonyos marginálisaihoz tartozó paraméterek szorzata. Ha ezeket a modelleket közvetlenül szeretnénk a permutációkra alkalmazni, akkor az összes olyan $(i_1, \ldots, i_n)$ cellát strukturális nullának kellene vennünk, melynek nem minden koordinátája különböző.

1.1. Példa. Legyen X olyan diszkrét valószínűségi változó, melyre $P(X \in \in [n]^n) = 1$. X-re a teljes függetlenség (hierarchikus) modellje:

$$P(X = (i_1, \ldots, i_n)) = \prod_{k=1}^{n} c_k(i_k), \quad 1 \le i_k \le n,$$

valamilyen $c_k(i)$ paraméterekre, melyekre $\sum_{i=1}^{n} c_k(i) = 1$ minden k-ra. A fenti strukturális nullák bevezetésével kapjuk a Π véletlen permutációra a kvázi-függetlenség modelljét:

$$P(\Pi = \pi) = K \prod_{k=1}^{n} c_k(\pi(k)), \quad \pi \in S_n,$$

ahol S_n az n-edfokú szimmetrikus csoport, K pedig normáló tényező. ■

Alkalmazhatjuk azonban a hierarchikus modelleket nem közvetlenül a permutáció koordinátáira. Lehetőségünk van arra, hogy S_n elemeit kölcsönösen egyértelműen megfeleltessük olyan vektoroknak, melyek már egy szorzathalmazban veszik fel értékeiket. A π permutáció leírható például az $r_\pi \in [1] \times \times [2] \times \ldots \times [n]$ vektorral, ahol $r_\pi(i)$ azt mutatja meg, hogy $\pi(i)$ hányadik legnagyobb eleme a $\{\pi(1), \ldots, \pi(i)\}$ halmaznak. Hasonló megfeleltetés érhető el ortogonális kontrasztok segítségével (Marden [46], illetve a 3. fejezet). Ennek a módszernek az lehet a hátránya, hogy a kapott modellek kevésbé jól értelmezhetők.

Az értekezésben egy másik lehetséges módszert vizsgálok. Először is, minden $U \subseteq [n]$-re és $\pi \in S_n$-re legyen

$$\pi(U) = (\pi(u) : u \in U), \quad \pi\{U\} = \{\pi(u) : u \in U\}. \tag{1}$$

Legyenek $U, V, W \subset [n]$ diszjunkt részhalmazok, és jelölje $U \perp V \,|\, W$ azt az állítást, hogy $\pi(U)$ és $\pi(V)$ feltételesen függetlenek, ha ismerjük $\pi(W)$-t, *valamint* a $\pi\{U\}$ és $\pi\{V\}$ halmazokat. Az általunk vizsgált modellek mindegyikében elegendő az $U \perp \overline{U} \,|\, \emptyset$ alakú relációkat használni, ahol $\overline{U} = [n] \setminus U$ az U halmaz komplementere.

1.2. Példa. Legyen $n = 4$. Az $\{1{,}2\} \perp \{3{,}4\} \,|\, \emptyset$ reláció azt fejezi ki, hogy $(\pi(1), \pi(2))$ és $(\pi(3), \pi(4))$ feltételesen független, ha ismerjük a $\{\pi(1), \pi(2)\}$ halmazt. ■

A 8. fejezetben azokat az eloszlásokat vizsgálom, amelyek eleget tesznek a $[k] \perp \overline{[k]} \,|\, \emptyset$ relációnak minden k-ra. Ezek az L-felbontható eloszlások. Az L-felbonthatóságot, mint tulajdonságot Critchlow, Fligner és Verducci [15] vezette be. Ebben a fejezetben az összes L-felbontható eloszlást, mint modellt vizsgálom. Az L-felbontható családban könnyű megadni a paraméterek explicit maximum likelihood becslését, melyeknek nem csak az aszimptotikus, hanem a pontos eloszlása is kiszámítható. Megmutatom, hogyan lehet a rögzített elégséges statisztikával rendelkező mintákból egyenletes eloszlás szerint generálni közvetlenül, illetve Monte Carlo módszerrel egy egyszerű Markov bázis segítségével. Ezen másodlagosan generált minták használhatók

az illeszkedés jóságának mérésére. Az eredmények egy kicsit általánosabb eloszláscsaládra, a korlátozottan L-felbontható eloszlásokra is átvihetők. A fejezet eredményei a [18] és a [19] dolgozatokban találhatók meg.

A 9. fejezetben azt vizsgálom, hogy mikor lesz π és π^{-1} eloszlása is L-felbontható. Az ilyen eloszlásokat duplán L-felbonthatónak nevezve, meghatározom a szigorúan pozitív duplán L-felbontható eloszláscsalád szabad paramétereinek számát, megadom két paraméterezését, és egy algoritmust a maximum likelihood becslés meghatározására. Megvizsgálom, mit mondhatunk abban az esetben, ha nem csak szigorúan pozitív eloszlásokat engedünk meg: ebben segít a család Markov bázisának meghatározása az $n = 4$ esetben. Ez a fejezet szintén a [18] és a [19] dolgozatokra, valamint a [22] publikációra épül.

A 10. fejezet néhány további, a felbonthatósággal kapcsolatos kérdést vizsgál. Ilyen például a 9. fejezet fő eredményének bizonyításában fontos szerepet játszó S-felbonthatóság, mely az L-felbonthatóságnál erősebb tulajdonság. Foglalkozom továbbá azokkal az eloszlásokkal, amelyek eleget tesznek a $U \perp \overline{U} \mid \emptyset$ relációnak minden U-ra, ezeket teljesen L-felbontható, vagy TL-felbontható eloszlásoknak hívom. Megmutatom, hogy a szigorúan pozitív TL-felbontható eloszlások éppen a kvázi független alakúak (lásd az 1.1. Példát). Ez az eredmény a [21] dolgozatból való.

Végül a 11. fejezetben egy, az irodalomban „állatorvosi lónak" tekintett adatsorra, az Amerikai Pszichológiai Társaság 1980-as elnökválasztásának adataira (röviden APA adatokra) illesztem az összes felbontható modellt.

Itt szeretném megköszönni témavezetőmnek, Tusnády Gábornak a közel nyolc évnyi közös munkát, a tőle kapott ismereteket, de még inkább azt a matematikai szemléletmódot és emberi hozzáállást, amit folyamatosan sugároz. Köszönöm családom és kollégáim segítségét, támogatását, és nem utolsó sorban azt a sok nógatást, ami nélkül ez a disszertáció nem jött volna létre.

2. Felhasznált eszközök

2.1. Algebrai statisztika

Algebrai módszereket a statisztikában elsőként Diaconis és Sturmfels [32] használt, és éppen kontingenciatáblák elemzésére. Geiger, Meek és Sturmfels [37] diszkrét tereken definiált torikus modelleket vizsgált, különös tekintettel a grafikus modellekre, Diaconis és Eriksson [30] pedig Markov lánc Monte Carlo technikával generálta a minta feltételes eloszlását az elégséges statisztikára nézve. Érdekes evolúciós alkalmazások találhatók például Seth Sullivant PhD disszertációjában [56]. Az alábbiakban összefoglaljuk azokat az eredményeket, melyeket később felhasználunk. További olvasmányként ajánljuk a [13, 55] könyveket.

Legyen $\mathcal{X} = \{x_1, \ldots, x_s\}$ véges halmaz, $M = (m_{ij})$ pedig egy $t \times s$ méretű, nemnegatív egész elemű mátrix. Azt mondjuk, hogy a $p = (p(x_1), \ldots, p(x_s))$ valószínűség-eloszlás az M modellhez tartozik, vagy p az M szerint faktorizálódik, ha léteznek olyan nemnegatív $\lambda_1, \ldots, \lambda_t$ paraméterek, melyekre

$$p(x_i) = c(\lambda) \prod_{j=1}^{t} \lambda_j^{m_{ji}},\ 1 \leq i \leq s. \tag{2}$$

Az összes M szerint faktorizálódó eloszlás halmazára az $\mathbf{F}(M)$ jelölést használjuk. Azokat az eloszláscsaládokat, melyek $\mathbf{F}(M)$ alakúak valamilyen alkalmas M-mel, torikus modelleknek hívjuk. Az $\mathbf{E}(M)$ diszkrét exponenciális család pedig álljon azokból a $p = (p(x_1), \ldots, p(x_s))$ eloszlásokból, melyekre

$$p(x_i) = c(\theta) \exp \sum_{j=1}^{t} m_{ji}\theta_j, 1 \leq i \leq s, \text{ ahol } \theta = (\theta_1, \ldots, \theta_t) \in (-\infty, \infty)^t. \tag{3}$$

A továbbiakban mindig feltesszük, hogy M sorvektorainak tere tartalmazza az $\mathbf{1} = (1, \ldots, 1)^\top$ vektort, így a normáló konstansok el is hagyhatók.

$\mathbf{F}(M)$ szigorúan pozitív elemei éppen $\mathbf{E}(M)$-et alkotják, és általában

$$\mathbf{E}(M) \subseteq \mathbf{F}(M) \subseteq \mathrm{cl}(\mathbf{E}(M)),$$

ahol cl() a lezárást jelöli (az euklideszi topológiában). A második tartalmazás akkor és csak akkor szigorú, ha $\mathbf{F}(M)$ nem zárt. Vegyük észre, hogy $\mathbf{E}(M)$ csak az M sorai által kifeszített altértől függ, de mint látni fogjuk, ugyanez nem igaz $\mathbf{F}(M)$-re. A t hosszú $v = (v_1, \ldots, v_t)$ vektor tartójára vezessük be a

$$\mathrm{Supp}(v) = \{1 \leq k \leq t : v_k \neq 0\}$$

jelölést, és az M mátrix i. oszlopvektorát jelölje m_i.

2.1. Definíció. *Legyen M $t \times s$ méretű, nemnegatív egész elemű mátrix. A $T \subseteq \{1, \ldots, s\}$ halmaz M-megvalósítható (M-feasible), ha minden $i \notin T$-re*

$$\mathit{Supp}(m_i) \not\subseteq \cup_{j \in T} \mathit{Supp}(m_j).$$

Vezessük be a szintén $x_1, \ldots, x_s$-sel jelölt változókat, a továbbiakban az $R[x] = R[x_1, \ldots, x_s]$ valós együtthatós polinomgyűrűben dolgozunk. Az $u = (u_1, \ldots, u_s)$ nemnegatív egész elemű vektorra definiáljuk az $x^u = x_1^{u_1} x_2^{u_2} \cdots x_s^{u_s}$ s változós monomot (egy tagú polinomot). Az M mátrixhoz rendeljük hozzá az X_M nemnegatív torikus varietást:

$$X_M = \{x \in \mathbb{R}^s_{\geq 0} : x^u - x^v = 0 \quad \forall u, v \in \mathbb{N}^s \text{ melyre } Mu = Mv\}. \tag{4}$$

X_M-et azért nevezik torikus varietásnak, mert az őt definiáló polinomok mindegyike kéttagú. Az $\mathcal{X}$-en adott p eloszlást írjuk vektor alakba, azaz legyen $p_i = p(x_i)$. Érvényes a következő tétel.

2.2. Tétel. (Geiger et al. [37]) *$cl(\mathbf{F}(M)) = X_M$. Továbbá $p \in X_M$-re $p \in \mathbf{F}(M)$ akkor és csak akkor, ha p tartója M-megvalósítható.*

Nyilván X_M és ezzel együtt $\mathrm{cl}(\mathbf{F}(M))$ is csak az M sorai által kifeszített altértől függ, hiszen az $Mu = Mv$ megoldásai csak ettől függnek. Az M-megvalósítható hamazok, és így $\mathbf{F}(M)$ azonban már nem csak ettől az altértől függ. Rapallo [52] bizonyítja a következő tételt.

2.3. Tétel. (Rapallo [52]) *Minden M-hez van olyan $M_{\max}$ maximális reprezentáció, melyre $cl(\mathbf{F}(M)) = \mathbf{F}(M_{\max})$.*

Adott p esetén könnyű eldönteni, hogy tartója M-megvalósítható-e. Azt, hogy eleme-e az X_M varietásnak, már nehezebb. A (4)-ben szereplő kéttagú polinomok egy I_M torikus ideált generálnak. Hilbert bázis-tétele azt mondja ki, hogy minden ideál végesen generált, sőt, algoritmikusan kiszámítható az ideált generáló bázis. Ezeknek az algoritmusoknak is kiterjedt az elmélete, a mi szempontunkból elég annyi, hogy a világhálón elérhetők ezeket az algoritmusokat implementáló ingyen használható programcsomagok. Ezután már csak azt kell ellenőrizni, hogy p gyöke-e ennek a véges sok bázispolinomnak. Megemlítjük, hogy az $M_{\max}$ maximális reprezentáció is algoritmikusan számolható.

Adott modellből származó minta esetén a minta feltételes eloszlása az elégséges statisztikára nézve már nem függ az ismeretlen paraméterektől. Ez a feltételes eloszlás használható például illeszkedésvizsgálatra, ezért lényeges kérdés, hogy tudunk-e belőle mintát generálni. Legyen adva a (2) által definiált $\mathbf{F}(M)$ torikus modell, $X = (X_1, \ldots, X_m)$ pedig legyen a modellből származó iid minta. Jelölje f_X a gyakoriságvektort, azaz $x \in \mathcal{X}$-re $f_X(x) = |\{1 \leq i \leq m : X_i = x\}|$. Ezzel a jelöléssel az Mf_X statisztika elégséges λ-ra. Továbbá X eloszlása az $Mf_X = u$ feltétel mellett egyenletes az

$$\mathcal{Y}_u = \{y \in \mathcal{X}^m : Mf_y = u\}$$

halmazon. A gyakoriságvektorra átfogalmazva, f_X eloszlása az $Mf_X = u$ feltétel mellett hipergeometrikus az

$$\mathcal{F}_u = \{f : \mathcal{X} \to \mathbb{N} : Mf = u\}$$

halmazon, azaz

$$P(f_X = f | Mf_X = u) = \frac{m!}{|\mathcal{Y}_u|} \prod_x (f(x)!)^{-1}, \quad f \in \mathcal{F}_u.$$

Ezekből a feltételes eloszlásokból általában nem tudunk közvetlenül generálni, de Markov lánc Monte Carlo technikák alkalmazhatók, ha a feladatra találunk egy Markov bázist.

2.4. Definíció. *Az $f_1, \ldots, f_L : \mathcal{X} \to \mathbb{Z}$ függvények az $\mathbf{F}(M)$ modell Markov bázisát alkotják, ha $Mf_i = 0$ minden i-re, továbbá minden u-ra és $f, f' \in \in \mathcal{F}_u$-ra található olyan $(\epsilon_1, f_{i_1}), \ldots, (\epsilon_A, f_{i_A})$ sorozat, ahol $\epsilon_i = \pm 1$, valamint*

$$f' = f + \sum_{j=1}^{A} \epsilon_j f_{i_j}, \text{ és } f + \sum_{j=1}^{a} \epsilon_j f_{i_j} \geq 0 \text{ minden } 1 \leq a \leq A\text{-ra.}$$

A Markov bázis segítségével irreducibilis Markov láncot definiálhatunk az $\mathcal{F}_u$ állapottéren. Minden lépésben válasszunk egy I-t egyenletesen $\{1, \ldots, L\}$-ből, és legyen $\epsilon = \pm 1$, $1/2 - 1/2$ valószínűséggel. A jelenlegi f állapotból próbáljunk az $f' = f + \epsilon f_I$-be lépni. Ha f' nemnegatív, tegyük meg a lépést, ellenkező esetben helyben maradunk. A Metropolis algoritmus segítségével módosíthatjuk a láncot, hogy a kívánt stacionárius eloszláshoz konvergáljon.

Diaconis és Sturmfels [32] megmutatja, hogyan lehet az ideálbázisokat megkereső algebrai algoritmusokat Markov bázisok keresésére használni. Jelölje az f függvény pozitív (negatív) részét f^+ (f^-), azaz $f = f^+ - f^-$, foka pedig legyen $\deg f = \max(\sum_x f^+(x), \sum_x f^-(x))$.

2.5. Tétel. (Diaconis és Sturmfels [32]) *Az $f_1, \ldots, f_L$ függvények akkor és csak akkor alkotják az $\mathbf{F}(M)$ modell Markov bázisát, ha az $x^{f_i^+} - x^{f_i^-}$ polinomok generálják az I_M ideált.*

Megjegyezzük, hogy [32] arra az esetre is ad algoritmust, ha a Markov bázis kiszámítása méretproblémák miatt nem kivitelezhető. Tegyük fel, hogy tudjuk, hogy a feladatra van olyan $f_1, \ldots, f_L$ Markov bázis, melyre $\max_i(\deg f_i) \leq d$, bár magát a Markov bázist nem tudjuk kiszámolni. Az $\mathcal{Y}_u$ állapottéren dolgozva, minden lépésben válasszunk ki egyenletes eloszlás szerint d különböző koordinátát. Számoljuk ki a kiválasztott koordináták gyakoriságvektorának elégséges statisztikáját, majd helyettesítsük a kiválasztott koordinátákat az ugyanilyen elégséges statisztikájú d-esek közül egyenletesen választott társasággal. Ha d elég kicsi, akkor a lehetséges új d-esekből elég kevés van, így a módszer alkalmazható.

2.2. Exponenciális családok, hierarchikus modellek

A jelen disszertáció szempontjából az előző szakasz (3) alakú diszkrét exponenciális családjai lesznek érdekesek. Szorzatalakú állapottér esetén ezek speciális esetei a hierarchikus, illetve grafikus modellek.

Legyen $X = (X(1), \dots, X(n))$ diszkrét valószínűségi változó, lehetséges értékeinek halmaza az $I = I_1 \times \cdots \times I_n$ véges szorzathalmaz. Minden $x = (x(1), \dots, x(n))$ vektorra és $A \subseteq [n]$-re legyen $x(A) = (x(i) : i \in A)$ az x vektor A-marginálisa. Legyen $\mathcal{A} \subset 2^{[n]}$ az $[n]$ részhalmazainak egy családja. Az $\mathcal{A}$ generátorokkal definiált hierarchikus modell a

$$p(x) = \prod_{A \in \mathcal{A}} \theta_A(x(A)) \quad \forall x \in I \tag{5}$$

alakú p eloszlásokból áll, ahol θ_A alkalmas paraméterek. Nyilván feltehetjük, hogy $\mathcal{A}$-nak nincs két egymást tartalmazó eleme. A hierarchikus modell *grafikus*, ha megadható az $1, \dots, n$ csúcspontokon olyan G gráf, hogy $\mathcal{A}$ éppen G klikkjeinek (maximális teljes részgráfjainak) halmaza.

A hierarchikus modellek torikus modellek, mivel $\mathbf{F}(M_{\mathcal{A}})$ alakúak egy alkalmas $M_{\mathcal{A}}$ $0-1$ mátrixra. A grafikus modellek elméletének egyik szép eredménye a Hammersley–Clifford tétel. Eszerint (5) (ahol $\mathcal{A}$ most a G klikkjeinek halmaza) akkor és csak akkor teljesül a p szigorúan pozitív eloszlásra, ha p rendelkezik a globális Markov tulajdonsággal, azaz bármely $U, V, S \subseteq [n]$ diszjunkt részhalmazokra, $X(U)$ és $X(V)$ feltételesen függetlenek $X(S)$-re nézve, valahányszor S elválasztja U-t és V-t a G gráfban. A Markov tulajdonság átírható polinomiális egyenletekké, ezért a 2.2. Tétel a Hammersley–Clifford tétel általánosításának tekinthető a torikus modellek esetére. A 2.2. Tétel szerint ugyanis a szigorúan pozitív p eloszlás akkor és csak akkor eleme az $\mathbf{E}(M)$ exponenciális családnak, ha p kielégíti az I_M ideált generáló polinomokat. Ehhez kapcsolódó további eredmények találhatók a [37, 52, 56] munkákban.

A grafikus modelleken belül különösen szépen viselkednek a felbontható (decomposable) modellek, melyekhez tartozó gráfokban nincs húr nélküli legalább négy hosszúságú kör. Érvényes a következő tétel.

2.6. Tétel. (Geiger et al. [37]) *Legyen adott a G gráfhoz tartozó (diszkrét) grafikus modell. A következő állítások ekvivalensek:*
(i) A grafikus modell felbontható.
(ii) A G szerint faktorizálódó eloszlások családja zárt.
(iii) A cellagyakoriságok maximum likelihood becslései tetszőleges minta esetén racionális számok.
(iv) A modellhez tartozó torikus ideálnak van másodfokú polinomokból álló Gröbner bázisa (és ezek a polinomok a globális Markov tulajdonság átírásából adódnak).

A hierarchikus és grafikus modellekhez kapcsolódó további irodalom [10, 14, 38, 42, 60].

Legyen $X_1, \ldots, X_m$ iid minta az $\mathcal{A}$ generátorhalmazzal definiált hierarchikus modellből, és jelölje $r(x)$ az $x \in I$ érték mintabeli relatív gyakoriságát. A valódi eloszlás, illetve paramétereinek maximum likelihood becslése általában nem explicit. Fontos kivételt képez a felbontható grafikus modellek osztálya (lásd a 2.6. Tételt). Az általános esetben egyik leggyakrabban használt numerikus módszer az iteratív arányos illesztés (iterative proportional scaling (IPS) vagy Deming–Stephan algoritmus) [24, 26]. Ez az algoritmus garantáltan konvergál a modell lezártjában egyértelműen létező maximum likelihood becsléshez, $\hat{p}$-hoz. Ez az eloszlás azzal a tulajdonsággal karakterizálható, hogy a $\hat{p}(x(A)) = \sum_{y \in I:\, y(A)=x(A)} \hat{p}(y)$ valószínűségek megegyeznek a megfelelő $r(x(A))$ tapasztalati valószínűségekkel, minden $A \in \mathcal{A}$-ra és $x(A)$ vektorra. Ha a $\hat{p}$ eloszlás tartója nem teljes, akkor azt mondjuk, hogy a minta tartalmaz strukturális nullát (a strukturális nullának ez a fogalma különbözik a modellbeli strukturális nulla fogalmától).

Az IPS algoritmus során ciklikusan illesztjük az $\mathcal{A}$-beli marginálisok eloszlását. Legyen $p^{(0)}$ a hierarchikus modell tetszőleges szigorúan pozitív eleme (pl. az egyenletes eloszlás). A $(t+1)$. iterációs lépésben legyen

$$p^{(t+1)}(x) = \frac{r(x(A))}{p^{(t)}(x(A))} p^{(t)}(x), \quad x \in I$$

ahol A ciklikusan befutja $\mathcal{A}$ elemeit.

A későbbiekben olyan exponenciális családokkal fogunk foglalkozni, ahol,

a hierarchikus modellekhez hasonlóan, a $t \times s$ méretű M mátrix minden eleme 0 vagy 1. Tegyük még fel, hogy M sorai lineárisan függetlenek, azaz az $\mathbf{E}(M)$ modell reguláris, minimális exponenciális család θ kanonikus paraméterekkel, és a $T_j(x_i) = m_{ji} = \chi\{x_i \in A_j\}$ statisztikákkal, ahol $A_j \subseteq \mathcal{X}$ alkalmas halmazok. A következő tétel az exponenciális családokra vonatkozó sokkal általánosabb tétel (pl. [7], 2.28.6. Tétel) speciális esete.

2.7. Tétel. *Legyen adott a* (3) *diszkrét exponenciális család, ahol az* M *mátrix teljes rangú, és minden eleme* 0 *vagy* 1. *Legyen* $X_1, \ldots, X_m$ *a családbeli* p_θ *eloszlásból származó iid minta. Ekkor, amint* $m \to \infty$, *a* $\hat{\theta}^{(m)}$ *maximum likelihood becslés* 1*-hez tartó valószínűséggel egyértelműen létezik, és*

$$\sqrt{m}(\hat{\theta}^{(m)} - \theta) \to N(0, I(\theta)^{-1}) \text{ eloszlásban},$$

ahol $I(\theta)$ *a Fisher információs mátrix,* $N(\mu, \Sigma)$ *pedig a* μ *várható értékű,* Σ *kovariancia-mátrixú normális eloszlás. Továbbá*

$$(I(\theta))_{ij} = P_\theta(A_i \cap A_j) - P_\theta(A_i)P_\theta(A_j).$$

2.3. EM- és MM-algoritmusok

Az EM algoritmusok szerteágazó elméletéből itt csak annyit ismertetünk, amennyit a későbbiekben fel fogunk használni. Az algoritmust hiányos megfigyelések esetén fogjuk alkalmazni: tegyük fel, hogy a teljes megfigyelés Z, melynek eloszlását a β paramétervektor írja le. Azonban Z helyett annak csak valamilyen X függvényét figyeljük meg, ez a hiányos megfigyelés. A feladat a β paramétervektor maximum likelihood becslése. Az algoritmus valamilyen $\beta^{(0)}$ kezdőértékből indul, a $(t+1)$. iterációban pedig az alábbi két lépést végzi el:

1. E-lépés (E = expectation): Az X hiányos megfigyelés és a $\beta^{(t)}$ aktuális paraméterek mellett számítsuk ki a teljes megfigyelés log-likelihoodjának várható értékét:

$$Q(\beta, \beta^{(t)}) = \mathbb{E}(\log L(Z, \beta)|X, \beta^{(t)}).$$

2. M-lépés (M = maximization): A $Q(\beta, \beta^{(t)})$ függvényt β-ban maximalizálva kapjuk az új $\beta^{(t+1)}$ paramétervektort.
Az algoritmust ebben az általánosságban Dempster et al. [27] vezette be. Megmutatta, hogy az algoritmus során az $L(X, \beta^{(t)})$ likelihood monoton növekszik. Ezért, ha a likelihood függvény felülről korlátos, akkor $L(X, \beta^{(t)})$ egy L^* értékhez konvergál. Általánosságban azonban nem garantált, hogy L^* globális maximum, sőt a $\beta^{(t)}$ értékek akár a likelihood függvény egy nyeregpontjához is konvergálhatnak. Az EM algoritmus konvergenciáját vizsgálata például Csiszár és Tusnády [17] és Wu [61]. Az EM algoritmusról és annak általánosításairól szól McLachlan és Krishnan [50] könyve. Az EM algoritmus, amennyiben konvergál, akkor is lassú: konvergenciája lineáris, sebessége pedig a hiányzó információ hányadától függ. Ennek ellenére népszerű, mivel általában könnyen programozható és kevés memóriát igényel.

Az MM (minorization-maximization) algoritmusok egy bővebb halmazt alkotnak, azaz az EM algoritmusok speciális MM algoritmusok. Bár ilyen típusú algoritmusokat már jóval korábban is vizsgáltak, az elnevezést Hunter és Lange [41] vezette be. A feladat a β paramétervektor maximum likelihood becslése az X mintából (most nincs teljes és hiányos minta). Az algoritmus valamilyen $\beta^{(0)}$ kezdőértékből indul, a $(t+1)$. iterációban pedig az alábbi két lépést végzi el:
1. M(inorization)-lépés: Előállítunk egy olyan $Q_t(\beta)$ függvényt, melyre

$$Q_t(\beta) \leq \log L(X, \beta) \, \forall \beta, \text{ és } Q_t(\beta^{(t)}) = \log L(X, \beta^{(t)}).$$

2. M(aximization) lépés: A $Q_t(\beta)$ függvényt β-ban maximalizálva kapjuk az új $\beta^{(t+1)}$ paramétervektort.
Ez az algoritmus is rendelkezik azzal a tulajdonsággal, hogy az $L(X, \beta^{(t)})$ likelihood monoton növekszik. Természetesen a Q_t függvények jó megválasztásán áll vagy bukik minden: Q_t-t sokszor úgy választják, hogy „szétválassza" a β paramétervektor koordinátáit, azaz a β-ban vett maximalizálás koordinátánként legyen elvégezhető. A konvergencia itt is csak bizonyos feltételek mellett igazolható.

II. rész

Különféle modellek

3. Statisztikák és modellek áttekintése

Elöljáróban három olyan munkát említünk, melyek együttesen átfogó képet adnak a véletlen permutációk elemzéséről. Marden [48] könyve a grafikai megjelenítéstől kezdve, a modellalkotáson át, a becslésig és hipotézisvizsgálatig végigvezeti az olvasót a véletlen permutációk statisztikai eszköztárán. A Fligner és Verducci által szerkesztett [36] kötet az 1990-ben Amherstben rendezett „Probability Models and Statistical Analyses for Ranking Data" című konferencia fontosabb előadásainak anyagát tartalmazza. A kötet érdekességét egyrészt sokszínűsége adja, másrészt az, hogy a terület 1990-beli aktuális állapotát tükrözi. Végül Critchlow, Fligner és Verducci [15] cikke a különböző modellek olyan tulajdonságait vizsgálja, mint a megfordíthatóság, címke-invariancia, L-felbonthatóság, unimodalitás, és teljes konszenzus.

Valós (vagy vektor) értékű adatoknál a két leggyakrabban használt statisztika az átlag és a szórás. Permutációk esetében is számolható átlag, de az általában nem lesz maga is permutáció. Jelölje S_n az n hosszú permutációk halmazát, és legyen adva S_n-en egy d távolságfüggvény. Ekkor a $\pi_1, \ldots, \pi_m$ minta d-középértéke a $\sum_{i=1}^{m} d(\rho, \pi_i)$ összeget minimalizáló $\rho \in S_n$ lesz, a szóródást pedig maga az összeg (megfelelően normalizált) értéke méri. S_n-en számos olyan d távolság definiálható, melyek a statisztikai alkalmazások szempontjából érdekesek és fontosak, ezeket hamarosan ismertetjük (lásd még pl. [28], 6. fejezet). Egy középérték helyett kereshetünk többet is, azaz klaszterosíthatjuk az adatokat néhány klaszterközéppont körül.

Röviden ismertetjük a permutáció-adatok spektrális analízisét (Diaconis [29] és az ottani hivatkozások). Adott S_n beli $\pi_1, \ldots, \pi_m$ mintából számítsuk ki az f gyakoriságvektort, azaz legyen $f(\pi) = |\{i : \pi_i = \pi\}|$. Az f vektort, mint az $\mathbb{R}^{n!}$ euklideszi tér elemét tekintjük. A csoportok reprezentációelméletének egyik eredménye szerint ez a tér egymásra merőleges V_λ *izotipikus alterekre* bomlik, ahol λ az $[n]$ halmaz partícióit futja be. Az „izotipikusság"

itt azt jelenti, hogy a V_λ alterek invariánsak az S_n elemeivel való balról illetve jobbról szorzásra nézve. Pontosabban ez azt jelenti, hogy $f \in V_\lambda$ és $\sigma \in S_n$ esetén az

$$f_{\sigma\circ}(\pi) = f(\sigma\pi) \text{ és } f_{\circ\sigma}(\pi) = f(\pi\sigma) \quad (\pi \in S_n)$$

balról illetve jobbról szorzott vektorok is V_λ-beliek, ahol a permutációk egymás után írása a csoportszorzást jelöli. Továbbá minden V_λ izotipikus altér felbomlik $k(\lambda)$ darab egymásra merőleges *invariáns altérre*, melyek mind izomorfak, és $k(\lambda)$ dimenziósak. Az „invariáns" kifejezés jelentése, hogy a balról szorzásokra invariánsak. A $k(\lambda)$ értékek például a hook-length képletből számolhatók, további részletek a [28, 29, 30] hivatkozásokban találhatók. Az f gyakoriságvektornak az izotipikus alterekre vett vetületeinek hossza (figyelembe véve az egyes alterek dimenzióját is), érdekes információt tartalmaz az adat struktúrájáról.

Térjünk most rá a leggyakrabban használt modellek ismertetésére! A modellek első csoportja a **rendezett minta modellek** (order statistics models). Képzeljük el, hogy valakinek hangokat kell erősség szerint sorrendbe állítania (leghalkabbtól leghangosabbig). Tegyük fel, hogy az i. hang észlelt erősségét egy folytonos F_i eloszlású X_i valószínűségi változó írja le. Ekkor a π sorrend valószínűsége

$$p(\pi) = P(X_{\pi(1)} < \cdots < X_{\pi(n)}).$$

Ezt a példát először Thurstone [59], majd később Daniels [23] tanulmányozta. Fel szokás tenni, hogy X_i-k függetlenek (de nem mindig, például azt az esetet is sokat vizsgálták, amikor az X_i-k együttes eloszlása normális). Ha az eloszlások csak eltolásparaméterben különböznek, azaz $F_i(x) = F(x - \mu_i)$ (ahol F ismert), akkor Thurstone modellről beszélünk. A Gumbel eloszlás esetében, azaz mikor $F(x) = 1 - \exp(-\exp x)$, a modell neve Luce vagy **Plackett–Luce modell**. Luce [44] ezt a modellt egy másik alakban vezette le. Elsőnek felállította a sorbarendezési posztulátumot (ranking postulate). Tegyük fel, hogy n elemet akarunk sorbarendezni, a legjobbtól a legrosszabbig. A posztulátum azt mondja ki, hogy a sorrend úgy keletkezik, hogy minden lépésben kiválasztjuk a legjobbat a még fennmaradó elemek közül. Azaz az

elemek minden C részhalmazára, és minden $x \in C$ elemre adott annak $p_C(x)$ valószínűsége, hogy C-ből x-et választjuk legjobbnak. Ezen kiválasztási valószínűségek alapján a π sorrend esélye

$$p(\pi) = \prod_{k=1}^{n} p_{C_k}(\pi(k)), \tag{6}$$

ahol $C_k = \{\pi(k), \ldots, \pi(n)\}$ a k. lépésben még választható elemek halmaza. Ezután Luce a kiválasztási axiómát (choice axiom) alkalmazta a $p_C(x)$ értékekre: az axióma azt mondja ki, hogy két elem kiválasztásának egymáshoz viszonyított valószínűsége nem függ attól, hogy a többi elem kiválasztható-e vagy sem, azaz a $p_C(x)/p_C(y)$ hányados ugyanannyi minden x, y-t tartalmazó C-re (IIA: independence from irrelevant alternatives). Ez pedig csak úgy lehetséges, ha $p_C(x) = \lambda_x / \sum_{y \in C} \lambda_y$ alakú. Kaptuk tehát, hogy

$$p(\pi) = \prod_{k=1}^{n} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{n} \lambda_{\pi(j)}}. \tag{7}$$

Könnyű ellenőrizni, hogy a Gumbel eloszlású Thurstone model ugyanezeket a valószínűségeket adja.

A **Babington Smith modellt** Babington Smith [3] javasolta, ez páros összehasonlításokból építi fel a sorrendet. A páros összehasonlításoknak külön elmélete van, ezt lehetett összeházasítani a teljes sorrendek elméletével: vegyük sorra az n elemből alkotható összes párt, és mindegyik párból válasszuk ki a nekünk szimpatikusabb elemet. Az n pontú teljes gráf éleit irányítsuk a választásainknak megfelelően. Ha ebben a gráfban nincs irányított kör, akkor választásaink konzisztensek egy egyértelműen meghatározott sorbarendezéssel. A körmentességre feltételt véve kapjuk, hogy

$$p(\pi) = c(\theta) \prod_{i<j} \theta_{\pi(i)\pi(j)},$$

ahol θ_{xy} annak a valószínűsége, hogy x és y páros összehasonlításában x győz. A modellből vett m elemű $\pi_1, \ldots, \pi_m$ minta esetén a paraméterekre elégséges

statisztika a $\widehat{K}$ pármátrix, melynek (i,j). eleme

$$\widehat{K}_{ij} = \frac{1}{m}\left|\left\{k : (\pi_k)^{-1}(i) < (\pi_k)^{-1}(j)\right\}\right|$$

azt fejezi ki, hogy a minta hányad részében kapott az i sorszámú elem jobb helyezést, mint a j sorszámú.

A Babington Smith modell speciális esete a **Mallows–Bradley–Terry modell**, melyben $\theta_{xy} = \frac{\alpha_x}{\alpha_x+\alpha_y}$. A valószínűségek ilyen alakját Bradley és Terry [9] vetette fel, a páros összehasonlítás modellben való használatát pedig Mallows [45] javasolta. A π sorrend esélye ebben a modellben tehát

$$p(\pi) = c(\alpha)\prod_{i=1}^{n}\alpha_{\pi(i)}^{n-i} = c(\alpha)\prod_{j=1}^{n}\alpha_j^{n-\pi^{-1}(j)}.$$

A modellből vett m elemű $\pi_1,\ldots,\pi_m$ minta esetén a paraméterekre elégséges statisztika a $\widehat{V} = (\widehat{V}(1),\ldots,\widehat{V}(n))$ átlaghelyezés vektor, ahol

$$\widehat{V}(j) = \frac{1}{m}\sum_{k=1}^{m}(\pi_k)^{-1}(j)$$

a j sorszámú elem helyezésének mintabeli átlaga.

A következő csoportot a **távolság-alapú modellek** (distance-based models) alkotják. Legyen d ismét egy S_n-en adott távolságfüggvény, ennek segítségével az alábbi egyszerű modellt definiálhatjuk a π *helyezésekre*:

$$p(\pi) = K(\theta)e^{-\theta d(\pi,\pi_0)},$$

ahol $\pi_0 \in S_n$ az eloszlás (ismert vagy ismeretlen) módusza, $\theta \geq 0$ pedig paraméter. Fontos megjegyezni, hogy a „távolság” definíciójába nem feltétlenül értjük bele a háromszög-egyenlőtlenség teljesülését. Ha azt szeretnénk, hogy a modell ne függjön az elemek számozásától, akkor fel kell tenni, hogy a d távolság jobbról invariáns, és ezt mindig fel is szokták tenni. Az 1. táblázatban a legfontosabb távolságokat gyűjtöttük össze ($\chi\{\cdot\}$ az indikátor függvény). Megjegyezzük, hogy a Kendall τ, Cayley, és Ulam távolságoknak mind van olyan típusú definíciója, hogy hány megengedett lépéssel lehet egyik per-

1. táblázat. Néhány fontos távolság definíciója

Hamming	$d_H(\pi,\rho)$	$=$	$\sum_{k=1}^{n}\chi\{\pi(k)\neq\rho(k)\}$
p-távolság	$d_p(\pi,\rho)$	$=$	$\sum_{k=1}^{n}\lvert\pi(k)-\rho(k)\rvert^p$
Maximum	$d_M(\pi,\rho)$	$=$	$\max_{k=1}^{n}\lvert\pi(k)-\rho(k)\rvert$
Kendall τ	$d_K(\pi,\rho)$	$=$	$\sum_{i<j}\chi\{(\pi(i)-\pi(j))(\rho(i)-\rho(j))<0\}$
Cayley	$d_C(\pi,\rho)$	$=$	$n-$ Ciklusok száma $(\pi\rho^{-1})$-ben
Ulam	$d_U(\pi,\rho)$	$=$	$n-$ Max. mon. növő rész hossza$(\pi\rho^{-1})$-ben

mutációt a másikba transzformálni, ahol a megengedett lépések halmaza a három távolságnál természetesen más és más. A statisztikai alkalmazások szempontjából érdekes, hogy mely távolságok invariánsak balról, melyek invariánsak a megfordításra vagy az invertálásra. Ezeket a tulajdonságokat részletesen tárgyalja pl. [28, 48].

A **kvázi független modellről** már beszéltünk, itt csak azt tesszük hozzá, hogy a modellből vett m elemű $\pi_1,\ldots,\pi_m$ minta esetén a paraméterekre elégséges statisztika az $\widehat{M}$ marginális mátrix, melynek (i,j). eleme

$$\widehat{M}_{ij}=\frac{1}{m}\,\lvert\{k:\pi_k(i)=j\}\rvert$$

annak tapasztalati valószínűségét adja meg, hogy a j sorszámú elem az i. helyezést kapja. Itt jegyezzük meg, hogy a $\widehat{K}$, $\widehat{V}$, $\widehat{M}$ statisztikák a modellektől függetlenül is sokat segítenek az adatokkal való ismerkedésben.

A következő modell a **többlépcsős helyezési modell** (multistage ranking model), melyet Fligner és Verducci [35] vezetett be. Itt a halmaz elemei (ezeket gyakran jelölteknek hívjuk) 1-től n-ig vannak számozva. A helyezéseket egyesével osztjuk ki. A k. lépésben a legjobb $k-1$ helyezést már kiosztottuk, és most választjuk ki a k. helyezett jelöltet, de a kiválasztásnál csak a maradék jelöltek relatív számozását vesszük figyelembe. Ha tehát a $j_1<\cdots<j_{n-k+1}$ jelöltek maradtak, akkor j_i-t $\theta(i,k)$ valószínű-

séggel választjuk, ahol $\theta(i,k)$ a $\sum_{i=1}^{n-k+1}\theta(i,k) = 1$ összefüggést kielégítő paraméterek.

Doignon, Pekeč és Regenwetter [33] egy hasonló modelljében a jelölteket számozásuk szerint vesszük sorba, és az új jelöltet az eddigiek sorrendjébe szúrjuk be. Minden k-ra adottak tehát a $\theta(i,k)$, $i = 1,\ldots,k$ beszúrási valószínűségek, melyek összege 1. A k. jelöltet k helyre illeszthetjük be: szúrjuk az eddigi sorrend $(i-1)$. és i. helye közé $\theta(i,k)$ valószínűséggel. Az így adódó **ismételt beszúrások modellje** (repeated insertion model) a többlépcsős helyezési modell „duálisa" abban az értelemben, hogy a helyezések és a jelöltek szerepét felcseréltük.

Végül ismertetjük a Chung és Marden [11] által bevezetett **ortogonális kontrasztok modelljét**. Egy C kontraszt nem más, mint a jelöltek néhány csoportjának összehasonlítása, pl. $C = (I_1,\ldots,I_K)$, ahol I_j nemüres diszjunkt részhalmazai az $[n]$ jelölthalmaznak. A C kontraszt értéke a π helyezésvektoron megadja az egyes I_j csoportokba eső jelöltek relatív helyezéseit a $C^U = \cup_j I_j$ halmazhoz képest. Formálisan,

$$C(\pi) = (\{\overline{\pi}(i) : i \in I_1\},\ldots,\{\overline{\pi}(i) : i \in I_K\}),$$

ahol $\overline{\pi}(i)$ azt mutatja meg, hogy $\pi(i)$ hányadik legkisebb a $\{\pi(j) : j \in C^U\}$ halmazban. Legyen például $n = 5$, a helyezésvektor pedig $\pi = (24513)$. A $C = (\{1{,}2\},\{3{,}4\})$ kontraszt az első kettő és a második kettő jelöltet hasonlítja össze. C értéke π-n $C(\pi) = (\{2{,}3\},\{1{,}4\})$, azaz mindkét első jelölt a két következő közé van rangsorolva.

Két kontrasztot akkor nevezünk ortogonálisnak, ha az általuk definiált összehasonlítások nem keverednek egymással. Formálisan, $C = (I_i)$ és $D = (J_j)$ akkor ortogonálisak, ha vagy (i) $C^U \cap D^U = \emptyset$, vagy (ii) $C^U \subseteq J_k$ valamilyen k-ra, vagy (iii) $D^U \subseteq I_k$ valamilyen k-ra. Ortogonális kontrasztok esetén a lehetséges kontraszt-értékek akárhogy kombinálhatók egymással, ezért javasolta Marden [47] a kontraszt-értékek modellezésére a kontingenciatáblák esetére használt modelleket. Ennek speciális esete a **ϕ-modell**, melyben a $C_1,\ldots,C_s$ ortogonális kontrasztok értékei függetlenek, sőt, az egyes

permutációk valószínűségei

$$p(\pi) = K(\theta) \exp\{\sum_{i=1}^{s} \theta_i d(C_i(\pi))\}$$

alakúak, ahol $\theta = (\theta_i)$ paramétervektor, $K(\theta)$ normáló tényező, és ha $C = = (I_1, \ldots, I_K)$, akkor

$$d(C(\pi)) = \sum_{1 \le i < j \le K} \sum_{r \in I_i} \sum_{s \in I_j} \chi\{\pi(r) > \pi(s)\}$$

a Jonckheere-Terpstra statisztika.

4. McCullagh inverziós modellje

Peter McCullagh [49] vezette be a páros összehasonlítások (Babington Smith) modelljének következő általánosítását. Minden $C \subseteq [n]$ halmazra álljon I_C a C halmaz elemeinek olyan permutációiból, melyek C egy elemét sem teszik a nagyság szerint növekvő sorrendben elfoglalt helyére. Ha például $n = 5$, és $C = \{2, 4, 5\}$, akkor I_C az (524) és a (452) permutációkból áll. McCullagh ezeket $(k-1)$-*ed rendű inverzióknak* nevezi, ha $|C| = k$. A továbbiakban C-t az I_C-beli inverziók alaphalmazának nevezzük. Az I_C halmaz elemszámát könnyű felírni a szita formula segítségével, és az is egyszerűen adódik, hogy az összes (tetszőleges rendű) inverzió száma $n!-1$. Azt mondjuk, hogy egy $\pi \in S_n$ permutáció tartalmazza az I_C-beli σ_C inverziót (jelölésben $\sigma_C \subseteq \pi$), hogyha π elemeiből csak a C-belieket megtartva éppen σ_C-t kapjuk. Ismét $n = 5$-tel, az (15324) permutáció az (524), (534), (32), (52), (53), (54) inverziókat tartalmazza.

Az inverziók segítségével definiálhatunk torikus modelleket. Válasszunk ki az inverziók közül s darabot, legyenek ezek $\sigma^1_{C_1}, \ldots, \sigma^s_{C_s}$. Konstruáljuk meg hozzájuk azt az $s \times n!$ méretű M incidencia-mátrixot, melynek (j, π)-dik eleme 1, ha π tartalmazza $\sigma^j_{C_j}$-t, egyébként pedig 0. Tekintsük az ehhez tartozó $\mathbf{F}(M)$ modellt.

McCullagh olyan eseteket vizsgált, amikor az inverziók közül az összes

legfeljebb k rendűt vesszük be a modellbe, valamilyen k-ra. A modell szabad paramétereinek száma (szigorúan pozitív esetben) attól függ, hogy az M mátrix sorai között vannak-e lineáris összefüggések. McCullagh azt a sejtést fogalmazta meg, hogy ilyen összefüggések nincsenek. Pontosabban, vegyük be a modellbe az összes inverziót, és még az **1** sort is. Így egy $n!$ méretű M_n négyzetes mátrixot kapunk, melyről McCullagh azt sejtette, hogy determinánsa ± 1 (attól függően, hogy milyen sorrendben soroljuk fel a sorokat/oszlopokat). Sejtését $n \leq 4$-re ellenőrizte, és azt is észrevette, hogy a sorok és oszlopok megfelelő átrendezésével háromszögmátrix kapható, melynek főátlójában csupa 1-es áll. Marden [48] minden $n \leq 7$-re ellenőrizte a sejtést.

Ha a fenti torikus modellbe csak az elsőrendű inverziókat vesszük bele, akkor visszakapjuk a Babington Smith modellt. Az ennél bonyolultabb modellek felépítése analógiát mutat a kontingenciatáblákra alkalmazott hierarchikus modellekkel, ahol egyre magasabb rendű kölcsönhatásokat építünk be a jobb illeszkedés kedvéért. A k-adrendű inverziók a k-adrendű interakciók analógjai. Tegyük fel, hogy egy elsőrendű inverziós modellben (azaz amely csak elsőrendű inverziókat tartalmaz), szerepel a (kj) $(j < k)$ inverzió. Ekkor, feltéve hogy egy permutációban a j és a k szomszédos, annak esélye, hogy k jön előbb, a permutáció többi elemétől független:

$$\frac{p(*_1 kj *_2)}{p(*_1 jk *_2)} = c_{kj},$$

ahol $*_1, *_2$ a számlálóban és nevezőben azonos elem-sorozatok.

Hasonlóan, ha egy másodrendű modellben szerepel az (ljk) $(j < k < l)$ inverzió, akkor

$$\log\left(\frac{p(*_1 kjl *_2)}{p(*_1 jkl *_2)}\right) - \log\left(\frac{p(*_1 lkj *_2)}{p(*_1 ljk *_2)}\right) = c_{ljk}.$$

A paraméterek interpretációját nehezíti, hogy a fenti egyenletek csak akkor igazak, ha a vizsgált elemek szomszédosak a permutációban.

$n \leq 4$-re McCullagh felsorolta az összes olyan modellt, melyek bizonyos peremfeltételeknek eleget tesznek. Ezek a peremfeltételek biztosítják, hogy a fenti paraméterek jobban interpretálhatóak legyenek. $n = 3$-ra kilenc ilyen

modell van, azonban n növekedésével egyre nehezebbnek tűnik az ilyen típusú modellek áttekintése.

A továbbiakban bebizonyítjuk McCullagh sejtését, elemi meggondolásokkal. Mondjuk ki először az állítást!

4.1. Tétel. *Legyen M_n az az $n!$ méretű négyzetes mátrix, melynek első $n!-1$ sora a σ_C inverziókhoz tartozó $\chi\{\sigma_C \subseteq \pi\}$ indikátorvektorokat tartalmazza, utolsó sora pedig* **1**. *Ekkor M_n sorai és oszlopai minden n-re átrendezhetők úgy, hogy alsó háromszögmátrixot kapjunk, melynek főátlójában csupa* 1*-es áll.*

A tételt bizonyítás előtt fogalmazzuk át! Ehhez azonosítsuk az inverziókat és a permutációkat. A C alaphalmazú σ_C inverziót azonosítsuk azzal a σ permutációval, mely a C elemeit σ_C szerint rakja sorba, míg $[n]\setminus C$ elemeit fixen hagyja. Például $n = 5$-re az (524) inverziót az (15324) permutációval identifikáljuk. Ahhoz, hogy S_n minden elemét megkapjuk, vezessük be az üres inverziót (melynek alaphalmaza $\emptyset$) úgy, hogy ezt az inverziót minden permutáció tartalmazza. Az üres inverziót az *id* identitással azonosítjuk. Ha a π permutáció tartalmazza a σ_C inverziót, és a σ_C-vel azonosított permutáció σ, akkor ezt a $\sigma \preceq \pi$ relációval jelöljük. Így a 4.1. Tételben szereplő M_n mátrix megegyezik (a sorok és oszlopok sorrendjétől eltekintve) a $\preceq$ reláció $M_{\preceq}$ indikátormátrixával: $M_{\preceq}(\sigma, \pi) = \chi\{\sigma \preceq \pi\}$. Megmutatjuk, hogy van olyan $S_n = \{\pi_1, \ldots, \pi_{n!}\}$ felsorolás, hogy minden $1 \leq i < j \leq n!$ esetén $\pi_j \not\preceq \pi_i$. Ha M_n sorait és oszlopait is ebben a sorrendben soroljuk fel, akkor M_n alsó háromszög lesz, a főátlóban pedig 1-esek állnak, mivel $\pi \preceq \pi$ minden $\pi \in S_n$-re. Könnyű látni, hogy ilyen felsorolás akkor és csak akkor létezik, ha a relációt leíró $G_{\preceq}$ irányított gráfban nincs irányított kör (a hurokéleken kívül). Ennek a gráfnak S_n elemei a csúcsai, és akkor vezet π-ből σ-ba irányított él, ha $\sigma \preceq \pi$. A szokásos módon bevezethetjük a $\prec$ relációt: $\sigma \prec \pi$ akkor és csak akkor, ha $\sigma \preceq \pi$ és $\sigma \neq \pi$. A hozzá tartozó $G_{\prec}$ gráf csak annyiban különbözik $G_{\preceq}$-től, hogy nincsenek benne hurokélek. Mivel a $\prec$ reláció nem tranzitív, a következő, 4.1. Tétellel ekvivalens állítás nem triviális.

4.2. Tétel. *A $G_{\prec}$ gráfban nincs irányított kör.*

A 4.2. Tétel bizonyítása előtt egy másik tételt fogunk belátni. Definiálunk az S_n halmazon egy másik, G_H gráfot. Egy π permutációból vezessen irányított él minden olyan permutációba, melyet π-ből úgy kapunk, hogy egy elemet a helyére rakunk, a többit pedig odébb toljuk, ha szükséges. Nézzünk megint egy példát $n = 5$-re: a $\pi = (24351)$ permutációból az (12435), (42351), (23541), (24315) permutációkba vezet él, rendre az 1,2,4,5 elemeket tettük helyre (π-ben 3 a helyén van, ezért őt nem akarjuk már helyre tenni: így ebben a gráfban nem lesz hurokél).

4.3. Tétel. *A G_H gráfban nincs irányított kör.*

Bizonyítás. n szerinti teljes indukcióval bizonyítunk. $n = 2$-re az állítás triviálisan teljesül. Tegyük fel, hogy minden $h \leq (n-1)$-re már tudjuk az állítást, bizonyítsuk be n-re! Tegyük fel indirekt, hogy van a gráfban egy $\pi_1 \to \ldots \to \pi_k \to \pi_1$ irányított kör.

Vegyük először észre, hogy egy helyrerakó lépés (H-lépés) nem növelheti a $v(\pi) = \sum_{i=1}^{n} |\pi^{-1}(i) - i|$ értéket. Itt $|\pi^{-1}(i) - i|$ azt mutatja meg, hogy az i elem milyen messze van a saját helyétől a π permutációban. $v(\pi)$ azért nem nőhet, mert ha a j elemet rakjuk helyre, akkor az összeg j. tagja $|\pi^{-1}(j) - j|$-vel csökken. A helyrerakás közben összesen $|\pi^{-1}(j) - j|$ elemet kellett eggyel odébb tolni, és az összeg eltolt elemekhez tartozó tagjai legfeljebb eggyel nőnek. Az összeg többi tagja pedig nem változik.

Ha van egy irányított körünk, akkor a fentiek szerint a $v(\pi)$ értékek konstansok a körben. Ez csak úgy lehet, ha minden H-lépésben az összes eltolt elem *távolodik* a saját helyétől. Nézzük meg először, hogy mi történik n-nel a kör során. Ha egyszer n-et a helyére rakjuk, akkor ott is marad, mert nincs senki, aki ki tudná tolni a helyéről. Emiatt két eset lehetséges: (i) n a saját helyén van az egész kör alatt, (ii) n sosincs a saját helyén a kör alatt. Az (ii) esetben tegyük fel, hogy a kör elején és végén n a $k < n$ helyen van. Mivel már beláttuk, hogy n a kör során csak balra tolódhat, a helyére ugrani pedig nem fog, így csak az lehetséges, hogy n a kör alatt végig a k. helyen marad. Tehát mindkét esetben van a permutációnak legalább egy olyan eleme (maga n), mely végig egy helyben marad a kör során.

Jelölje $x_1 < \ldots < x_f$, $n > f \geq 1$, azokat a *helyeket*, ahol az elemek az egész kör alatt fixen maradnak, azaz $\pi_i(x_j) = c_j$ minden i, j-re. Minden elem, amely nem ezeken a helyeken áll, legalább egyszer a helyére kerül, és legalább egyszer eltolódik. Ha ugyanis csak tolódna, akkor folyamatosan távolodna a saját helyétől, és a kör nem zárulna be. Ha pedig csak annyi történne vele, hogy egyszer a helyére raknánk, akkor szintén nem záródna a kör. Vegyünk most egy $[x_g + 1, \ldots, x_{g+1} - 1]$ intervallumot, és nézzük meg, hogy ezeken a helyeken mi történik. Az ezen helyeken álló elemeket legalább egyszer a helyükre tesszük, azonban ezzel nem hagyhatják el a vizsgált intervallumot, hiszen akkor vagy az x_g, vagy az x_{g+1} helyen álló elem odébb tolódna. Beláttuk tehát, hogy az $[x_g + 1, \ldots, x_{g+1} - 1]$ intervallumot a kör mindegyik π_i permutációja önmagára képezi. Az ilyen intervallumok között van legalább egy nem-üres, melynek hossza $2 \leq h \leq n - 1$. Ha az egész kört megszorítjuk erre az intervallumra (és átszámozzuk az elemeket), irányított kört kapunk az S_h-hoz tartozó G_H gráfban, ami az indukciós feltevés szerint ellentmondás. ■

Most már bebizonyíthatjuk a 4.2. Tételt.

Bizonyítás. (4.2. Tételé) Azt látjuk be, hogy $G_\prec$ tartalmazza a G_H összes élét, és $G_\prec$ összes többi $\sigma \to \pi$ éléhez van G_H-ban $\sigma \to \rho_1 \to \cdots \to \rho_k \to \pi$ irányított út. Ebből a tétel már következik, hiszen minden $G_\prec$-beli kör finomítható lenne G_H-beli körré, amiről már beláttuk, hogy nincs.

Egyrészt világos, hogy ha π-ből σ egy H-lépéssel kapható, akkor σ nem-fixpontjai ugyanolyan sorrendben vannak σ-ban és π-ben, azaz $\sigma \prec \pi$. Másrészt meg kell mutatnunk, hogy ha σ nem-fixpontjai ugyanolyan sorrendben vannak σ-ban és π-ben, akkor π-ből σ megkapható véges sok H-lépéssel. Ehhez csak annyit kell tenni, hogy π azon elemeit, melyek σ-ban fixpontok, a helyükre rakosgatjuk, egész addig, míg mindegyik a helyére nem kerül. Előfordulhat, hogy egy elemet többször is mozgatni kell, mert az első helyrerakás után eltolódik, de mivel G_H-ban nincs kör, véges sok lépésben garantáltan megkapjuk σ-t. ■

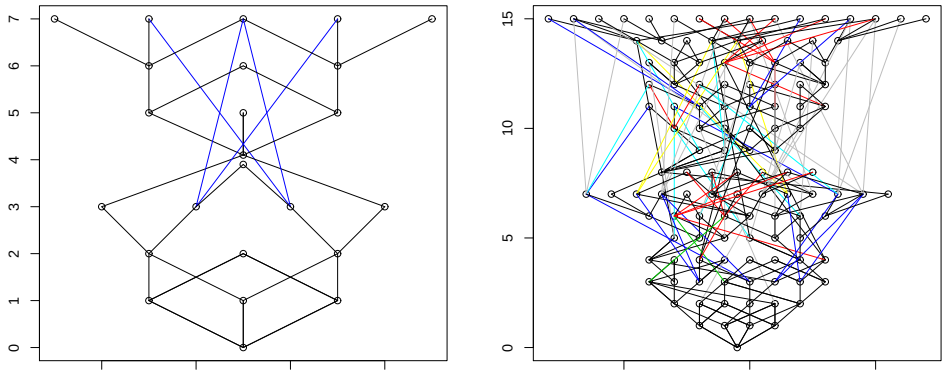
Beláttuk McCullagh sejtését, és közben találtunk két olyan gráfot S_n-en, melyben nincs irányított kör. Minden ilyen G gráf definiál egy $\preceq_G$ részbenrendezést S_n-en: $\sigma \preceq_G \pi$ akkor és csak akkor, ha $\sigma = \pi$, vagy π-ből el lehet

jutni σ-ba G-beli irányított úton. A 4.2. Tétel bizonyítása szerint a $G_{\prec}$ és a G_H gráf ugyanazt a részbenrendezést definiálja, jelölje ezt $\preceq_H$. Ezt a részbenrendezést tehát a H-lépés definiálja. Az S_n halmazon természetesen sok más (körmentes) lépés definiálható, és vizsgálhatók a hozzájuk tartozó részbenrendezések. A következő szakaszban röviden bemutatunk majd néhány olyat, melyeknek statisztikai alkalmazása is van.

Egy körmentes irányított gráf csúcsait szintekre oszthatjuk. A nulladik szint álljon azokból a csúcsokból, melyekből nem vezet ki él. Ha az első $k-1$ szintet már definiáltuk, akkor a k. szint álljon a maradék csúcsok közül azokból, melyekből csak az első $k-1$ szintre vezetnek ki élek. Ezzel a felbontással a k. szintre azok a csúcsok kerülnek, melyekből a nulladik szintre vezető leghosszabb út éppen k hosszúságú. A legmagasabb szint sorszáma pedig a gráf leghosszabb irányított útjának hossza.

A G_H gráfban a nulladik szint csak az identitásból áll. Az első szinten azok a permutációk foglalnak helyet, melyek az identitástól csak két szomszédos elem felcserélésével térnek el. A magasabb szintek leírása már sokkal bonyolultabbnak tűnik: nem találtunk olyan módszert, mellyel egy π permutációról könnyen eldönthető lenne, hogy hányadik szinten van. Azt is tudjuk, hogy a G_H gráfban van $2^{n-1}-1$ hosszúságú út, és azt sejtjük, hogy ennél hosszabb nincs. $2^{n-1}-1$ hosszú utat például úgy kapunk, ha a $(234\ldots n1)$ permutációból indulva, mindig a legelső helyen álló elemet tesszük a helyére. Ennek az útnak a $(k+1)$. tagját, π_{k+1}-et úgy kapjuk $(0 \leq k \leq 2^{n-1}-1)$, hogy a k számot kettes számrendszerbe írjuk (az elejét nullákkal feltöltve, hogy $n-1$ hosszú sorozat keletkezzék), jelölje ezt v^k. π_{k+1}-ben pontosan akkor lesz a $j>1$ fixpont, ha v^k-nak az $(n+1-j)$. koordinátája 1-es. Az 1 csak akkor lesz fixpont, ha $k=2^{n-1}-1$, azaz az út legvégén, hiszen az út az identitásban végződik. A $k<2^{n-1}-1$ esetben pedig a fixpontok rögzítése után írjuk az utolsó szabad helyre az 1-et, a többi szabad helyet pedig töltsük fel a maradék elemekkel növekvő sorrendben. Elemi meggondolással belátható, hogy az így megkonstruált π_k legelső elemének helyrerakásával valóban a konstrukció szerinti π_{k+1}-et kapjuk.

4.4. Példa. Legyen $n=7$, ekkor a fenti út hossza 63. Keressük meg ennek az útnak 23. tagját! A 22-t kettes számrendszerbe írjuk: 010110. Ezért a



1. ábra. A G_H gráf váza $n = 4$ és $n = 5$ esetén.

keresett permutáció fixpontjai 3,4,6 lesznek. A maradék helyeket a leírás szerint kitöltve, $\pi_{23} = (2534761)$. ■

Ha a G_H gráfnak csak a szintjeire, illetve a csúcsok közötti leghosszabb utakra vagyunk kíváncsiak, akkor elég, ha a gráf „vázát” vizsgáljuk. Ezt úgy kapjuk, hogy az éleken egyesével sorbamenve, minden olyan $\pi \to \sigma$ élt elhagyunk, melyre a gráfban van π-ből σ-ba vezető egynél hosszabb út. Az, hogy mely éleket fogjuk így elhagyni, nyilván nem függ attól, hogy milyen sorrendben vesszük őket végig. A G_H gráf vázát szemlélteti az 1. ábra $n = 4$-re és $n = = 5$-re. Helyhiány miatt a csúcsok mellé nem írtuk oda a hozzájuk tartozó permutációkat. A függőleges tengelyen a szint sorszáma látható, egy szinten belül a csúcsok elrendezése viszont nem követ különösebb logikát. $n = 4$-re úgy kaphattunk szimmetrikus ábrát, hogy a negyedik szinten található két csúcsot egymás fölé rajzoltuk, kis eltolással. Az éleket aszerint szineztük, hogy milyen távoli szintek között futnak.

A leghosszabb út mellett a legrövidebbre is kíváncsiak lehetünk. Jelölje $\ell(\pi)$ a legrövidebb, π-ből az identitásba vezető út hosszát. Erre a mennyiségre csak alsó és felső becslést tudunk adni, pontos értékére nem tudunk könnyű kiszámítási módot. Jelölje $s_1(\pi)$ a π leghosszabb monoton növő részsorozatának hosszát:

$$s_1(\pi) = \max\{s |\ \exists i_1 < \ldots < i_s : \pi(i_1) < \ldots < \pi(i_s)\}.$$

Hasonlóan, jelölje $s_2(\pi)$ a π leghosszabb, egyesével növekedő részsorozatának hosszát:

$$s_2(\pi) = \max\{s|\ \exists i_1 < \ldots < i_s : \pi(i_k) = \pi(i_{k-1}) + 1\ \forall k\}.$$

A következő egyenlőtlenségek teljesülnek.

4.5. Tétel.

$$n - s_1(\pi) \leq \ell(\pi) \leq n - s_2(\pi) \quad \forall \pi \in S_n.$$

Bizonyítás. Az első egyenlőtlenség azért igaz, mert minden H-lépés legfeljebb eggyel növelheti $s_1(\pi)$-t A másodikhoz legyen $k, k+1, \ldots, k+s_2(\pi)-1$ egy leghosszabb eggyel növekvő részsorozat π-ben. Rakjuk először helyre az 1,2,$\ldots, k-1$ elemeket, azután pedig az $n, n-1, \ldots, k+s_2(\pi)$ elemeket. Így az identitást kapjuk legfeljebb $n - s_2(\pi)$ H-lépésben. ■

A szakasz befejezéseként megemlítünk egy általánosítást. Legyen T egy n csúcsú fa, a csúcsok legyenek 1-től n-ig számozva. Legyen még n objektumunk is, szintén 1-től n ig számozva. Ha a fa minden csúcsába pontosan egy objektumot rakunk, akkor az objektumok elhelyezkedését egy $\pi \in S_n$ permutáció írja le. Definiálhatjuk a fa szerinti H-lépést: válasszunk egy olyan objektumot, mely nem a saját helyén van, és tegyük át a saját helyére úgy, hogy az eddigi helyét és a saját helyét összekötő fa-beli út mentén a többi objektumot eggyel elcsúsztatjuk. Ez a H-lépés is körmentes, a 4.3. Tétel bizonyítása (ami az a speciális eset, amikor a fa egyetlen út) most is működik. Annyi a módosítás, hogy n szerepét a fa egy tetszőlegesen választott levele veszi át, az $[x_g + 1, \ldots, x_{g+1} - 1]$ intervallumok helyett pedig azok a részfák szerepelnek, melyek a kör során változatlan objektumú csúcsok elhagyásával keletkeznek. A tetszőleges fához tartozó H-lépésre is feltehetők az eddigi kérdések a szintek számáról, a legrövidebb utakról, stb. Ezek azonban már végképp nem vágnak a jelen értekezés témájába.

Ehhez kapcsolódó valószínűségszámítási feladat a következő. Induljunk ki valamilyen konfigurációból, és minden lépésben véletlenül válasszuk ki a helyrerakandó objektumot, mondjuk az i. objektumot $p(i)$ valószínűséggel.

Jelölje X azt, hogy hány lépés múlva lesz minden objektum a helyén. Keresendő X eloszlása.

5. Részbenrendezések

Ebben a szakaszban áttekintünk néhány ismert eredményt az S_n-en megadható – statisztikai szempontból is érdekes – részbenrendezések, és a velük szorosan összefüggő távolságok témaköréből.

A távolságokhoz hasonlóan, az S_n bizonyos részbenrendezéseinek is van fontos statisztikai alkalmazása. Az alapkérdés itt az, hogy a H_1 és H_2 kétváltozós eloszlások közül melyikben erősebb a két változó közötti pozitív összefüggőség. Itt most nem azt a megközelítést használjuk, hogy minden kétváltozós eloszláshoz hozzárendelünk egy, a pozitív összefüggőség erősségét mérő számot, hanem a kétváltozós eloszlások halmazán egy részbenrendezést definiálunk. Pontosabban, az $M(F,G)$ halmozon definiáljuk a részbenrendezést, mely mindazon H-kból áll, melyek X-marginálisa F, Y-marginálisa pedig G. Ha a H_1 és H_2 eloszlású (X_1,Y_1) és (X_2,Y_2) változók marginálisai különböznek, összehasonlításuk még mindig lehetséges, amennyiben

$$F_1(X_1) \sim F_2(X_2) \text{ és } G_1(Y_1) \sim G_2(Y_2), \tag{8}$$

ahol $\sim$ azt jelenti, hogy a két változó eloszlása megegyezik. Ekkor az eredeti eloszlások helyett az $(F_1(X_1), G_1(Y_1)$ és az $(F_2(X_2), G_2(Y_2))$ párok H_1^* és H_2^* eloszlásait hasonlíthatjuk össze. Az irodalomban többféle ilyen részbenrendezés is használatos (lásd pl. Schriever [53] vagy Lehmann [43]): H_1 lehet H_2-nél jobban konkordáns (concordant, $\preceq_c$), asszociált (associated, $\preceq_a$), sor regresszió összefüggő (row regression dependent, $\preceq_{rr}$) vagy oszlop regresszió összefüggő (column regression dependent, $\preceq_{cr}$). Ezek pontos definíciójára itt nem térünk ki, mivel az értekezés központi témájához csak lazán kapcsolódnak.

Tegyük most fel, hogy a H_1 és a H_2 kétváltozós eloszlásokból megfigyelünk egy-egy n elemű mintát, és a mintaelemek mindkét mintában különbözőek. Ekkor a $\hat{H}_1$ és $\hat{H}_2$ tapasztalati eloszlásfüggvényekre teljesül a (8) feltétel, és a

$\hat{H}_1^*$, $\hat{H}_2^*$ transzformált tapasztalati eloszlásfüggvények marginálisai egyenletesek lesznek az $\{1/n, 2/n, \ldots, 1\}$ halmazon. Így az eloszlásfüggvények közötti, pozitív összefüggőséget mérő $\preceq$ részbenrendezések természetes módon definiálnak S_n-en részbenrendezéseket: ha a $\hat{H}_i^*$ eloszlás a $(k/n, \pi_i(k)/n)$ pontokhoz rendel $1/n$ súlyokat, ahol $k = 1, \ldots, n$, és $i = 1{,}2$, π_i pedig S_n-beli permutáció, akkor legyen például $\pi_1 \preceq_c \pi_2$ akkor és csak akkor, ha $\hat{H}_1^* \preceq_c \hat{H}_2^*$.

A [4, 5] cikkekben Block, Chhetry, Fang és Sampson az S_n így keletkező négy részbenrendezését vizsgálják. Ekvivalens megfogalmazásokat adnak ezekre a részbenrendezésekre, illetve módszereket adnak arra, hogyan dönthető el, hogy két permutáció relációban áll-e. Ehhez kapcsolódik Block et al. [6] valamint Diaconis és Graham [31] munkája, melyekben a szerzők a részbenrendezések és a távolságok kapcsolatát elemzik. Kicsit részletesebben, jelöljön $\preceq_i$ négy részbenrendezést S_n-en, ahol $i = 1, \ldots, 4$. A $\preceq_i$ relációk inverzió-megszűntető transzpozíciókkal vannak definiálva. Akkor mondjuk, hogy $\pi \preceq_i \rho$, ha ρ-ból π elérhető olyan (i-től függő típusú) transzpozíciósorozattal, mely az inverziók számát minden lépésben csökkenti. Az $i = 1$ esetben minden transzpozíció megengedett. A többi i-re csak bizonyos transzpozíciók megengedettek. $i = 2$-re csak szomszédos sorszámú (azaz k, $k + 1$ alakú) elemeket cserélhetünk fel, míg $i = 3$-ra csak szomszédos helyen álló elemeket. Az $i = 4$ esetben az $i = 2, 3$ esetekhez tartozó transzpozíciók vannak megengedve. A következő eredményt Block et al. [5] tartalmazza:

$$\hat{H}_1^* \preceq_c \hat{H}_2^* \iff \pi_1 \preceq_1 \pi_2, \quad \hat{H}_1^* \preceq_{rr} \hat{H}_2^* \iff \pi_1 \preceq_2 \pi_2,$$
$$\hat{H}_1^* \preceq_{cr} \hat{H}_2^* \iff \pi_1 \preceq_3 \pi_2, \quad \hat{H}_1^* \preceq_a \hat{H}_2^* \iff \pi_1 \preceq_4 \pi_2.$$

Az $i \leq 3$ esetekben a szerzők egyszerűen ellenőrizhető feltételeket adnak arra, hogy két permutáció mikor van relációban.

A [6] cikkben a szerzők négy távolságot vizsgálnak S_n-en, melyeket rendre $d_i(\pi, \rho)$ jelöl ($i = 1, \ldots, 4$). Jelölje továbbá $I(\pi)$ a π inverzióinak számát. Megmutatják, hogy

$$\pi \preceq_i \rho \iff d_i(\pi, \rho) = I(\rho) - I(\pi).$$

Az $i > 1$ esetekben $d_i(\pi, \rho)$ a π-t és ρ-t összekötő legrövidebb, i-típusú transz-

pozíciókat használó út hossza (most nem kötjük ki, hogy az inverziók száma csökkenjen). Ezek közül d_2 és d_3 könnyen számolható:

$$d_2(\pi, \rho) = I(\pi\rho^{-1}), \quad d_3(\pi, \rho) = I(\pi^{-1}\rho),$$

ezek éppen a Kendall-féle τ-távolság, illetve annak inverze. A d_4 távolságra azonban nem adnak egyszerű kiszámítási módot. Végül az $i = 1$ esetben $d_1(\pi, \rho)$ a π-t és ρ-t összekötő legrövidebb olyan transzpozíció-sorozat hossza, melyben az inverziók száma minden lépésben pontosan eggyel változik.

Fogalmazzuk meg az eddig látottakat kissé általánosabban! Először is, minden $G = (S_n, E)$ irányítatlan, összefüggő gráf definiál S_n-en egy d_G távolságot: $d_G(\pi, \rho)$ a π-t és ρ-t összekötő legrövidebb út hossza. Ezen kívül legyen $M : S_n \to \mathbb{N}$ egy függvény, ennek segítségével részbenrendezés is definiálható: $\pi \preceq_{G,M} \rho$, ha ρ-ból vezet olyan G-beli út π-be, amely mentén M szigorúan monoton csökken. Tegyük még fel, hogy

$$|M(\pi) - M(\rho)| \leq 1 \quad \forall (\pi, \rho) \in E, \tag{9}$$

valamint hogy

$$d_G(\pi, \rho) = M(\rho^{-1}\pi) \quad \forall \pi, \rho \in S_n. \tag{10}$$

A (9) teljesülése esetén (10) teljesüléséhez például elegendő, hogy d_G balról (vagy jobbról) invariáns (azaz $d_G(\pi, \rho) = d_G(\sigma\pi, \sigma\rho)$), $M(\pi) = 0$ csak $\pi = id$-re teljesüljön, valamint hogy minden π-nek legyen nála eggyel kisebb M-értékkel rendelkező szomszédja.

5.1. Tétel. *Teljesüljön* (9) *és* (10). *Ekkor* $\pi \preceq_{G,M} \rho$ *akkor és csak akkor, ha* $M(\rho^{-1}\pi) = M(\rho) - M(\pi)$.

Bizonyítás. Az egyik irányban, ha $\pi \preceq_{G,M} \rho$, akkor ρ-ból vezet olyan út π-be, mely mentén M értéke szigorúan monoton csökken. (9) miatt ennek az útnak a hossza $M(\rho) - M(\pi)$, másrészt, szintén (9) miatt ennél rövidebb út nincs. (10) szerint tehát

$$M(\rho) - M(\pi) = d_G(\pi, \rho) = M(\rho^{-1}\pi).$$

A másik irányban viszont tegyük fel, hogy

$$d_G(\pi, \rho) = M(\rho^{-1}\pi) = M(\rho) - M(\pi).$$

Ez azt jelenti, hogy van ρ-ból π-be vezető $M(\rho) - M(\pi)$ hosszú út. (9) miatt azonban egy ilyen út mentén M szükségképpen szigorúan monoton csökken. ■

Ebben az esetben tehát az M függvény megadja két permutáció távolságát, és az is eldönthető általa, hogy két permutáció relációban áll-e. Ez az állítás alkalmazható a fenti $\preceq_2$ és $\preceq_3$ részbenrendezésekre. Az $i = 1{,}4$ eset ennél bonyolultabb: itt csak a (9) feltétel teljesül, ennek megfelelően az állítás is más, és a bizonyításhoz is plusz meggondolások kellenek.

Nézzük meg, hogy a H-lépéshez milyen távolság tartozik. Sajnos a H-lépés nem megfordítható, azaz a hozzá tartozó gráf irányított lesz. Az egyik lehetőség az, hogy elhagyjuk az irányítást, azaz olyan lépéseket is megengedünk, amikor a permutáció egy fixpontját tetszőleges másik helyre elmozdítjuk. Ekkor összefüggő, irányítatlan gráfot kapunk, és vizsgálhatjuk az így keletkező távolságot S_n-en.

A másik lehetőség, hogy minden olyan lépést megengedünk, amikor egy tetszőleges elemet más helyre rakunk át, a közbenső elemeket elcsúsztatva. Ez az áthelyező-lépés (Á-lépés) megfordítható, és éppen a d_U Ulam távolságot definiálja:

$$d_U(\pi, \rho) = M_U(\rho^{-1}\pi) = n - s_1(\rho^{-1}\pi),$$

ahol s_1 most is egy permutáció leghosszabb monoton növő részsorozatának hosszát jelöli. Ez éppen a (10) egyenlet megfelelője, igazolásához például könnyű ellenőrizni, hogy a (10) után megfogalmazott feltételek teljesülnek. Az Ulam távolsághoz tartozó részbenrendezés szerint $\pi \preceq_U \rho$, ha ρ-ból π elérhető olyan Á-lépések sorozatával, melyek mindegyike növeli a leghosszabb növekvő részsorozat hosszát. Az 5.1. Tételt alkalmazva kapjuk, hogy

$$\pi \preceq_U \rho \iff s_1(\rho^{-1}\pi) = n + s_1(\rho) - s_1(\pi).$$

Végül megjegyezzük, hogy ahogy $\preceq_3$ a $\preceq_2$ inverze, úgy a $\preceq_H$ részbenrendezésnek is definiálható az inverze. Ehhez a H-lépés inverzét kell definiálni,

azaz megnézni, hogy a π-n ható H-lépés hogyan hat π^{-1}-en. Legyen $\pi \in$ $\in S_n$ permutáció. Akkor hajtunk végre rajta H'-lépést, ha egy $k \neq \pi(k)$ nem-fixpontot kiválasztva, ezt fixponttá tesszük, azaz a k. helyre k-t írunk. Ezután, ha $\pi(k) > k$, akkor a permutáció $k, \ldots, \pi(k) - 1$ elemeihez egyet hozzáadunk, ha pedig $\pi(k) < k$, akkor a $\pi(k) + 1, \ldots, k$ elemekből egyet kivonunk. Ha például $\pi = (251364)$, akkor $k = 2$-t választva a H'-lépés eredménye (321465), ha pedig $k = 3$, akkor az eredmény (153264).

6. Plackett-Luce-féle modellek

Ebben a szakaszban a Plackett-Luce modell, illetve vele rokon modellek paramétereinek maximum likelihood becslésével foglalkozunk. A Plackett-Luce modell megfelelője páros összehasonlítások esetére a Bradley-Terry modell [9]: tegyük fel, hogy n objektumunk van, és egy megfigyelés két objektum összehasonlításából áll. Feltesszük, hogy minden i objektumhoz tartozik egy λ_i paraméter, mely az objektum értékét méri. Modellünk szerint annak valószínűsége, hogy az i és j objektumok összehasonlításakor i „nyer",

$$P(i \text{ legyőzi } j\text{-t}) = \frac{\lambda_i}{\lambda_i + \lambda_j}.$$

Általában Plackett-Luce modellnek nevezhetjük azt az esetet, amikor minden megfigyelés az n objektum valamely részhalmazának sorbarendezéséből áll, és a sorrendek valószínűségei (7) alakúak.

A Bradley-Terry modellnek számos egyéb általánosítása született. Agresti [2] modellje annyiban tér el az eredeti modelltől, hogy a páros összehasonlítások kimenetele függ attól, hogy a két objektum közül melyik az első, és melyik a második (pl. sportmérkőzéseknél az eredmény függ attól, hogy melyik csapat játszik otthon). Rao és Kupper [51] modelljében az összehasonlítások kimenetele döntetlen is lehet. Ezekben a modellekben a paraméterek maximum likelihood becslésére különböző algoritmusokat adtak a szerzők. Hunter [40] egységesen tárgyalja ezen modellek maximum likelihood becslését az MM algoritmus segítségével, és megadja, hogy milyen feltételek mellett konvergálnak az algoritmusok.

Huang et al. [39] további általánosítása az, hogy az objektumok két (vagy több) részhalmazát (pl. csapatok) hasonlítjuk össze. Ebbe a modellbe is beleférhet a döntetlen kimenetel, illetve a hazai pálya hatása. A szerzők ebben a cikkben is MM algoritmusokat javasolnak a maximum likelihood becslés megtalálására, és bizonyos feltételek mellett konvergenciát bizonyítanak.

A következőkben azt mutatjuk meg, hogy ugyanezekre a modellekre az EM algoritmus is használható, mivel belefférnek a „hiányos megfigyelések" keretébe. Sőt, egyes esetekben kétféle természetes módon is definiálhatjuk a teljes megfigyelést, így két különböző EM algoritmust kapunk. Mi itt csak az algoritmusok kiszámítására szorítkozunk, konvergenciájukkal, konvergencia sebességükkel nem foglalkozunk. Ennek részben az az oka, hogy a kapott algoritmusok formálisan nagyon hasonlítanak a már említett MM algoritmusokhoz, azaz implementációjuk ugyanolyan bonyolultságú, viszont szimulációs tapasztalataink szerint az EM algoritmusok az MM algoritmusoknál lassabban konvergálnak. Ezért az EM algoritmusoknak mindössze elméleti érdekessége van.

Összesen tehát négy modellt vizsgálunk. Minden esetben n objektumunk van, a továbbiakban ezeket játékosoknak nevezzük. Mindegyik modellben ugyanazok a λ_i, $i = 1, \ldots, n$ paraméterek szerepelnek, ahol λ_i az i játékos képességét méri (jobb képességhez nagyobb paraméterérték tartozik), jelölje $\lambda = (\lambda_1, \ldots, \lambda_n)$ ezt a paramétervektort. Feltesszük még, hogy $\sum \lambda_i = 1$. Két modellben ezeken kívül egy pozitív θ paraméter is lesz. A megfigyelések minden esetben kettő vagy több játékos vagy csapat sorbarendezéseiből állnak.

Plackett-Luce: Egy $I \subset [n]$ halmazba eső játékosokat rakjuk sorba. Az I halmaz π sorrendjének valószínűsége

$$p(\pi) = \prod_{k=1}^{|I|} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{|I|} \lambda_{\pi(j)}}, \tag{11}$$

ahol $\pi(1)$ az első helyezett, $\pi(|I|)$ az utolsó helyezett.

Hazai pálya: Az i és j játékosokat hasonlítjuk össze. A modell szerint

$$P(i \text{ legyőzi } j\text{-t}) = \begin{cases} \theta\lambda_i/(\theta\lambda_i + \lambda_j) & \text{ha } i \text{ van otthon,} \\ \lambda_i/(\lambda_i + \theta\lambda_j) & \text{ha } j \text{ van otthon,} \end{cases}$$

ahol a $\theta > 0$ paraméter a hazai pálya nyújtotta előny vagy hátrány erősségét fejezi ki.

Rao-Kupper: Itt az i és j játékosok mérkőznek egymással, de a végeredmény döntetlen is lehet. A valószínűségek:

$$P(i \text{ legyőzi } j\text{-t}) = \lambda_i/(\lambda_i + \theta\lambda_j),$$
$$P(j \text{ legyőzi } i\text{-t}) = \lambda_j/(\lambda_j + \theta\lambda_i),$$
$$P(i \text{ és } j \text{ döntetlent játszik}) = (\theta^2 - 1)\lambda_i\lambda_j/[(\lambda_i + \theta\lambda_j)(\lambda_j + \theta\lambda_i)],$$

ahol $\theta > 1$ az úgynevezett küszöb paraméter.

Csapatmérkőzés: Az I és J csapatok mérkőznek egymással, ahol $I, J \subset [n]$, és $I \cap J = \emptyset$. Ekkor

$$P(I \text{ legyőzi } J\text{-t}) = \frac{\sum_{i\in I} \lambda_i}{\sum_{i\in I} \lambda_i + \sum_{j\in J} \lambda_j}, \tag{12}$$

azaz a valószínűségek olyan alakúak, mint a Bradley-Terry modellnél, ha egy csapat képességét a játékosok képességeinek összege méri.

A modelleket a következőképpen fogalmazhatjuk meg exponenciális eloszlású valószínűségi változókkal.

6.1. Lemma. *Legyenek a Z_i $(i = 1, \ldots, n)$ valószínűségi változók függetlenek, λ_i paraméterű exponenciális eloszlásúak. Ekkor*

$P(Z_{\pi(1)} < \ldots < Z_{\pi(\lvert I\rvert)}) = \prod_{k=1}^{\lvert I\rvert} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{\lvert I\rvert} \lambda_{\pi(j)}}$	*Plackett-Luce*
$P(Z_i < Z_j/\theta) = \frac{\lambda_i}{\lambda_i + \theta\lambda_j}$	*Hazai pálya, Rao-Kupper*
$P(Z_j/\theta < Z_i < Z_j\theta) = \frac{(\theta^2-1)\lambda_i\lambda_j}{(\lambda_i+\theta\lambda_j)(\lambda_j+\theta\lambda_i)}$	*Rao-Kupper*
$P(\min_{i\in I} Z_i < \min_{j\in J} Z_j) = \frac{\sum_{i\in I}\lambda_i}{\sum_{i\in I}\lambda_i + \sum_{j\in J}\lambda_j}$	*Csapatmérkőzés*

A lemma egyszerűen igazolható az exponenciális eloszlás örökifjúságából, illetve abból, hogy független exponenciális változók minimuma is exponen-

ciális eloszlású. Látjuk tehát, hogy ha minden megfigyelésre ismernénk a megfigyelésben résztvevő játékosokhoz tartozó Z_i változókat, és ezek értékei alapján állítanánk fel a sorrendet, akkor a kívánt alakú valószínűségeket kapnánk. A hazai pálya és a Rao-Kupper esetben ez persze csak akkor igaz, ha θ ismert. Ha θ ismeretlen, akkor a hazai pálya esetben könnyen megoldható a probléma, míg a Kupper-Rao modellben nem látszik ilyen egyszerűen a megoldás.

A Plackett-Luce és a csapatmérkőzés modellben más változókat is használhatunk teljes megfigyelésként.

6.2. Lemma. *Tegyük fel, hogy egy urnában az $i = 1, \ldots, n$ egészekkel megszámozott golyók vannak. Visszatevéssel húzunk, egy húzásnál az i-golyót λ_i valószínűséggel húzzuk ki.*
Plackett-Luce*: Legyen I a sorbarendezendő játékosok halmaza. Addig húzunk, amíg mindegyik $i \in I$ golyó legalább egyszer kijött, és jelölje π, hogy az I elemei milyen sorrendben jöttek ki. Ekkor π valószínűségét éppen* (11) *adja meg.*
Csapatmérkőzés*: Legyen I és J a két csapat. Addig húzunk, amíg először $I \cup J$-beli golyót húzunk. Ekkor annak valószínűségét, hogy I-beli golyó jön előbb, éppen* (12) *jobb oldala adja meg.*

A bizonyítást, egyszerűsége miatt, most sem részletezzük. Ha tehát minden megfigyelésre ismernénk az urnából való húzások sorozatát, és a golyók megjelenési sorrendje alapján állítanánk fel a sorrendet, akkor a kívánt alakú valószínűségeket kapnánk. Miután definiáltuk a teljes és a hiányos megfigyeléseket, kiszámíthatjuk az EM algoritmus egy iterációját. A számolást itt nem részletezzük, csak az eredményeket mondjuk ki.

6.1. Plackett-Luce

Tegyük fel, hogy m megfigyelésünk van, az r-edikben az I_r halmazba eső játékosok π_r sorrendjét figyeljük meg. Jelölje minden $i \in I_r$-re $\alpha_r(i)$, hogy az i hányadik helyen áll a π_r sorrendben. Jelölje továbbá m_i azon megfigyelések számát, melyben az i játékos szerepel.

Ekkor az „exponenciális” teljes megfigyeléshez tartozó EM algoritmus paraméterfrissítésének képlete:

$$\lambda_i^{(t+1)} = m_i \left[\sum_{r:i\in I_r} \sum_{k=1}^{\alpha_r(i)} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} \right]^{-1} \qquad 1 \le i \le n.$$

Hunter [40] MM algoritmusa hasonló, csak a számlálóból le kell vonni u_i-t, ahol u_i azt mutatja, hogy hány sorrendben lett i az utolsó, a nevezőből pedig le kell vonni $u_i/\lambda_i^{(t)}$-t. A „húzássorozat” teljes megfigyeléshez tartozó EM algoritmus paraméterfrissítésének képlete pedig:

$$\lambda_i^{(t+1)} = m_i + \lambda_i^{(t)} \left[\sum_{r=1}^{m} \sum_{k=1}^{|I_r|} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} - \sum_{r:i\in I_r} \sum_{k=1}^{\alpha_r(i)} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} \right], \; 1 \le i \le n.$$

Ez a második EM algoritmus már kevésbé hasonlít Hunter algoritmusára, és tapasztalatunk szerint lassabb az előző EM algoritmusnál.

6.2. Hazai pálya

Legyen most is m megfigyelésünk, jelölje most is m_i az i játékos mérkőzéseinek számát. Jelölje m_{ik}, hogy hány mérkőzést játszott i és k úgy, hogy i volt otthon. Legyen még HV_i az i által hazai pályán elvesztett meccsek száma, IV_i pedig az i által idegenben elszenvedett vereségek száma. Az EM algoritmus M lépése nem oldható meg explicit, de helyette használható a következő kétlépcsős iteráció (mellyel úgynevezett GEM algoritmust kapunk):

$$\theta^{(t+1)} = \frac{m}{\sum_{i\neq k} \frac{m_{ik}\lambda_i^{(t)}}{\theta^{(t)}\lambda_i^{(t)}+\lambda_k^{(t)}} + \frac{\sum_{i=1}^{n} HV_i}{\theta^{(t)}}}, \tag{13}$$

$$\lambda_i^{(t+1)} = \frac{m_i}{\sum_{k\neq i} \left(\frac{m_{ik}\theta^{(t+1)}}{\theta^{(t+1)}\lambda_i^{(t)}+\lambda_k^{(t)}} + \frac{m_{ki}}{\theta^{(t+1)}\lambda_k^{(t)}+\lambda_i^{(t)}} \right) + \frac{HV_i+IV_i}{\lambda_i^{(t)}}}, \; 1 \le i \le n. \tag{14}$$

A Hunter által megadott MM algoritmus most is hasonló. Az ő algoritmusát úgy kapjuk, ha (13)-ban kivonjuk a számlálóból a $\sum_{i=1}^{n} HV_i$ mennyiséget,

a nevezőből pedig a $\frac{\sum_{i=1}^{n} HV_i}{\theta^{(t)}}$ tagot, és (14)-ban kivonjuk a számlálóból a $HV_i + IV_i$ mennyiséget, a nevezőből pedig a $\frac{HV_i+IV_i}{\lambda_i^{(t)}}$ tagot.

6.3. Rao-Kupper

Ennél a modellnél feltesszük, hogy θ ismert. Még mindig m megfigyelésünk van, és m_i az i által játszott mérkőzések száma. Jelölje még N_{ik}, V_{ik}, D_{ik} azt, hogy i hányszor nyert, veszített, illetve játszott döntetlent k ellen. Ekkor az EM algoritmus egy iterációja:

$$\lambda_i^{(t+1)} = \frac{m_i}{\sum_{k\neq i}\left(\frac{N_{ik}+D_{ik}}{\lambda_i^{(t)}+\theta\lambda_k^{(t)}} + \frac{\theta(V_{ik}+D_{ik})}{\lambda_k^{(t)}+\theta\lambda_i^{(t)}} + \frac{V_{ik}}{\lambda_i^{(t)}}\right)}, \; 1 \leq i \leq n.$$

Legyen $V_i = \sum_{k\neq i} V_{ik}$ az i vereségeinek száma. Ha a fenti képlet számlálójából V_i-t, nevezőjéből $V_i/\lambda_i^{(t)}$-t kivonunk, akkor megkapjuk Hunter MM algoritmusát.

6.4. Csapatmérkőzés

Most kicsit más jelöléseket használunk: az I_r játékoshalmaz az I_r^1, I_r^2 diszjunkt csapatokra bomlik ($r = 1, \ldots, m$). Tegyük fel, hogy I_r^1 és I_r^2 d_r darab mérkőzést játszik egymással, tehát összesen $\sum_{r=1}^{m} d_r$ megfigyelésünk van. Minden $i \in I_r$-re jelölje $N_r(i)$ ($V_r(i)$) azt, hogy az I_r-beli csapatfelállásban hányszor nyer (veszít) az i játékos csapata. Legyen még $q_r(i) = \sum_{j\in I_r^*} \lambda_j$, ahol I_r^* az I_r^1, I_r^2 csapatok közül azt jelöli, amelyikben i benne van. $q_r(i)$ tehát az I_r csapatfelállásban i csapatának összképessége. Továbbá legyen $q_r = \sum_{j\in I_r} \lambda_j$. Ismét m_i jelöli az i által játszott mérkőzések számát. Ekkor az „exponenciális" teljes megfigyeléseken alapuló EM algoritmus egy iterációja:

$$\lambda_i^{(t+1)} = \frac{m_i}{\sum_{r:\, i\in I_r} \frac{d_r}{q_r^{(t)}} + \left(V_r(i) + N_r(i)\frac{q_r^{(t)}(i)-\lambda_i^{(t)}}{q_r^{(t)}(i)}\right)\frac{1}{\lambda_i^{(t)}}}, \; 1 \leq i \leq n.$$

A „húzássorozatok" teljes megfigyelésen alapuló EM algoritmus paraméter-

frissítése pedig

$$\lambda_i^{(t+1)} = \left\{ \sum_{r:\, i \in I_r} \frac{N_r(i)}{q_r^{(t)}(i)} + \sum_{r:\, i \notin I_r} \frac{d_r}{q_r^{(t)}} \right\} \lambda_i^{(t)},\ 1 \le i \le n.$$

Érdekes, hogy ebben az esetben ez az EM algoritmus hasonlít inkább a Huang, Weng és Lin által adott MM algoritmushoz, melyben az iteráció

$$\lambda_i^{(t+1)} = \left(\sum_{r:\, i \in I_r} \frac{N_r(i)}{q_r^{(t)}(i)} \right) \left(\sum_{r:\, i \in I_r} \frac{d_r}{q_r^{(t)}} \right)^{-1} \lambda_i^{(t)},\ 1 \le i \le n$$

alakú.

7. Rendezett minta modell

Ebben a szakaszban azt a kérdést vizsgáljuk, hogy ha a rendezett minta modellben csak azt tesszük fel, hogy az X_i $(1 \le i \le n)$ valószínűségi változók F_i eloszlásai függetlenek egymástól, akkor hogyan lehet az F_i eloszlásfüggvényeket a $\pi_1, \ldots, \pi_m$ mintából becsülni. Erre a kérdésre egyáltalán nincs kielégítő válaszunk, mindössze egy próbálkozást szeretnénk itt bemutatni. Az ötlet az, hogy megint az EM algoritmust használjuk. Az egyszerűség kedvéért tegyük fel, hogy az X_i valószínűségi változók mindegyike csak véges sok értéket vehet fel. Ha azt akarjuk, hogy mindig permutáció keletkezzék, fel kell tennünk, hogy a tartók diszjunktak. Ha X_i tartója T_i, akkor mindig feltehető, hogy

$$T_i = \{s_{ij} = j + i/n : j = 1, \ldots, J\}, \quad i = 1, \ldots, n,$$

azzal a plusz feltétellel, hogy a tartók egyes pontjai nulla valószínűségűek is lehetnek. Legyen tehát $p(i,j) = P(X_i = s_{ij})$, feladatunk ezen paraméterek maximum likelihood becslése. Legyen a teljes minta $\{X_{i,r} : 1 \le i \le n, 1 \le \le r \le m\}$. Ekkor

$$Q(p,p^*) = \sum_{i=1}^{n}\sum_{j=1}^{J} \log p(i,j) \sum_{r=1}^{m} P(X_{i,r} = s_{ij}|\pi_r, p^*).$$

Ezt a kifejezést p-ben maximalizálva kapjuk az iterációs lépést:

$$p^{(t+1)}(i,j) = \frac{1}{m}\sum_{r=1}^{m} P(X_{i,r} = s_{ij}|\pi_r, p^{(t)}).$$

Nézzük meg, hogyan számítható ki a $P(X_i = s_{ij}|\pi)$ feltételes valószínűség, adott p paraméterek mellett. Ehhez nyilván elég a $P(X_i = s_{ij}, \pi)$ valószínűségeket kiszámolni.

$$P(X_i = s_{ij}, \pi) = A_\pi(\pi^{-1}(i), j)p(i,j)B_\pi(\pi^{-1}(i), j),$$

ahol

$$A_\pi(k,j) = \sum_{(j_1,\ldots,j_{k-1})\in\mathcal{J}(k,\pi,j)} \prod_{\ell=1}^{k-1} p(\pi(\ell), j_\ell)$$

annak valószínűsége, hogy $X_{\pi(1)} < \ldots < X_{\pi(k-1)} < s_{\pi(k)j}$. Ennek megfelelően

$$\mathcal{J}(k,\pi,j) = \{(j_1,\ldots,j_{k-1}) : s_{\pi(1)j_1} < \ldots < s_{\pi(k-1)j_{k-1}} < s_{\pi(k)j}\}.$$

Hasonlóan írható fel $B_\pi(k,j)$, ami annak valószínűsége, hogy $s_{\pi(k)j} < X_{\pi(k+1)} < \ldots < X_{\pi(n)}$. Adott π mellett az A_π, B_π mátrixok rekurzívan kitölthetők (lásd a következő oldalt). Így az EM algoritmus könnyen programozható és gyorsan fut. Azonban nem várható, hogy globális maximumhoz konvergáljon, mivel számos lokális szélsőértékhely lehet. Érdemes lenne az elkövetkezőkben ezzel a feladattal alaposabban foglalkozni.

Algoritmus A_π kitöltésére:
Inicializáció:
legyen $A_\pi(1,j) = 1$ $(1 \le j \le J)$ és $A_\pi(k,0) = 0$ $(1 \le k \le n)$.
Rekurzió:
for $\ell = 3$ to $J + n$
for $k = \max(2, \ell - J)$ to $\min(n, \ell - 1)$
if $\pi(k) < \pi(k-1)$:
$A_\pi(k,\ell-k) = A_\pi(k,\ell-k-1)+A_\pi(k-1,\ell-k-1)p(\pi(k-1),\ell-k-1)$
else:
$A_\pi(k,\ell-k) = A_\pi(k,\ell-k-1) + A_\pi(k-1,\ell-k)p(\pi(k-1),\ell-k)$.

Algoritmus B_π kitöltésére:
Inicializáció:
legyen $B_\pi(n,j) = 1$ $(1 \le j \le J)$ és $B_\pi(k,J+1) = 0$ $(1 \le k \le n)$.
Rekurzió:
for $\ell = J + n - 1$ downto 2
for $k = \max(1, \ell - J)$ to $\min(n-1, \ell - 1)$
if $\pi(k) > \pi(k+1)$:
$B_\pi(k,\ell-k) = B_\pi(k,\ell-k+1)+B_\pi(k+1,\ell-k+1)p(\pi(k+1),\ell-k+1)$
else:
$B_\pi(k,\ell-k) = B_\pi(k,\ell-k+1) + B_\pi(k+1,\ell-k)p(\pi(k+1),\ell-k)$.

III. rész

Feltételes függetlenség és hierarchikus modellek

8. L-felbonthatóság

Legyen $v = (v(1), \ldots, v(s))$ egy vektor, melynek koordinátáit az olvashatóság kedvéért most nem alsó indexekkel jelöljük. Az i.-től a j.-ig terjedő koordináták halmazát illetve vektorát jelölje

$$v\{i..j\} = \{v(i), \ldots, v(j)\},\ v(i..j) = (v(i), \ldots, v(j)), \quad 1 \leq i \leq j \leq s. \quad (15)$$

Jelölje S_n $(n \geq 1)$ az n-edfokú szimmetrikus csoportot, ennek elemei a $\pi = (\pi(1), \ldots, \pi(n))$ permutációk. Egy S_n-en adott valószínűségeloszlást jelöljön $p = \{p(\pi) : \pi \in S_n\}$, és legyen $\Pi = (\Pi(1), \ldots, \Pi(n))$ p eloszlású valószínűségi változó, azaz $P(\Pi = \pi) = p(\pi)$.

Mint a bevezetésben már említettük, az L-felbonthatóságot, mint tulajdonságot, Critchlow et al. [15] vezette be. Az „L" Luce nevére utal, hiszen látni fogjuk, hogy pontosan azok az eloszlások L-felbonthatóak, melyek eleget tesznek Luce sorbarendezési posztulátumának, azaz (6) alakúak. Definiáljuk tehát az L-felbonthatóságot! Rögtön négy különböző ekvivalens definíciót is adunk.

8.1. Definíció. (Critchlow et al. [15]) *Legyen Π véletlen permutáció, eloszlása pedig p. Π vagy p L-felbontható, ha a következő négy ekvivalens feltétel teljesül.*
1. Minden $1 \leq k \leq n-1$ és $\{x_1, \ldots, x_k, y\} \subseteq [n]$ mellett az

$$f(y|x_1, \ldots, x_k) = P\left(\Pi(k+1) = y \mid \Pi(1) = x_1, \ldots, \Pi(k) = x_k\right)$$

(esetleg nem definiált) függvény szimmetrikus az $x_1, \ldots, x_k$ argumentumokban.

2. Minden $1 \leq k \leq n-1$ *és* $\{x_1, \ldots, x_k, y\} \subseteq [n]$ *mellett a fenti* f*-re*

$$f(y|x_1, \ldots, x_k) = P\left(\Pi(k+1) = y \mid \Pi\{1..k\} = \{x_1, \ldots, x_k\}\right).$$

3. A $Z_k = \Pi\{1..k\}$, $k = 1, \ldots, n$ *véletlen halmazok (mint diszkrét valószínűségi változók) Markov láncot alkotnak.*
4. Megadható olyan Λ *nemnegatív függvény az*

$$(x, C) : C \subseteq \{1, \ldots, n\},\ x \notin C \tag{16}$$

párokon, mellyel

$$p(\pi) = \prod_{k=0}^{n-1} \Lambda(\pi(k+1), \pi\{1..k\}) \quad \forall \pi \in S_n. \tag{17}$$

Bizonyítás. Azt bizonyítjuk, hogy a fenti négy definíció valóban ekvivalens. 2. és 3. nyilvánvalóan csak egymás átfogalmazása. Legyen ugyanis $z_i = \{x_1, \ldots, x_i\}$, ha $1 \leq i \leq k$, és $z_{k+1} = \{x_1, \ldots, x_k, y\}$. Ekkor $f(y|x_1, \ldots, x_k) = P(Z_{k+1} = z_{k+1}|Z_1 = z_1, \ldots, Z_k = z_k)$, míg $P(\Pi(k+1) = = y \mid \Pi\{1..k\} = \{x_1, \ldots, x_k\}) = P(Z_{k+1} = z_{k+1}|Z_k = z_k)$. Azaz 2. éppen a Markov tulajdonságot fejezi ki.

Az is nyilvánvaló, hogy 2.-ből következik 1. Fordítva pedig azt használjuk, hogy ha $B = \cup_i B_i$ diszjunkt felbontás, és egy A eseményre $P(A|B_i)$ i-től független konstans, akkor $P(A|B) = P(A|B_i)$.

1.-ből triviálisan következik 4. Nevezzük ugyanis a 4.-ben szereplő Λ függvényt p L-felbontásának. Ha 1. teljesül, akkor egy kitüntetett, úgynevezett kanonikus L-felbontást kapunk a

$$\Lambda(x, C) = P(\Pi(|C|+1) = x \mid \Pi\{1..|C|\} = C), \tag{18}$$

ha a feltétel valószínűsége pozitív, egyébként pedig $\Lambda(x, C) = 0$ függvény formájában.

Végül belátjuk, hogy 4.-ből következik 1. Megjegyezzük, hogy ez szerepel Critchlow et al. [15] cikkében. Legyen megint $z_i = \{x_1, \ldots, x_i\}$, $z_{k+1} = z_k \cup$

$\cup y$, valamint $[n] \setminus z_{k+1}$ elemeinek egy tetszőleges felsorolása $l_1, \ldots, l_{n-k-1}$, és $l_{n-k} = y$. Ekkor 4. felhasználásával

$$f(y|x_1, \ldots, x_k) = \frac{\sum_{\pi:\pi(1..k)=x_{(1..k)},\pi(k+1)=y} \prod_{t=0}^{n-1} \Lambda(\pi(t+1), \pi\{1..t\})}{\sum_{\pi:\pi(1..k)=x_{(1..k)}} \prod_{t=0}^{n-1} \Lambda(\pi(t+1), \pi\{1..t\})} =$$

$$\frac{\prod_{t=0}^{k-1} \Lambda(x_{t+1}, z_t)\Lambda(y, z_k)(k+1)! \sum_{\sigma\in S_{n-k-1}} \prod_{t=0}^{n-k-2} \Lambda(l_{\sigma(t+1)}, z_{k+1} \cup l_{\sigma\{1..t\}})}{\prod_{t=0}^{k-1} \Lambda(x_{t+1}, z_t)k! \sum_{\sigma\in S_{n-k}} \prod_{t=0}^{n-k-1} \Lambda(l_{\sigma(t+1)}, z_k \cup l_{\sigma\{1..t\}})} =$$

$$(k+1)\Lambda(y, z_k)\frac{\sum_{\sigma\in S_{n-k-1}} \prod_{t=0}^{n-k-2} \Lambda(l_{\sigma(t+1)}, z_{k+1} \cup l_{\sigma\{1..t\}})}{\sum_{\sigma\in S_{n-k}} \prod_{t=0}^{n-k-1} \Lambda(l_{\sigma(t+1)}, z_k \cup l_{\sigma\{1..t\}})},$$

ez pedig valóban csak a z_k halmaztól függ, $x_1, \ldots, x_k$ sorrendjétől nem. ■

Ahogy már említettük, az L-felbonthatóság ekvivalens Luce sorbarendezési posztulátumával, hiszen a (6) és a (17) egyenletek ugyanazt fejezik ki. Másrészt, a $Z_k = \Pi\{1..k\}$ halmazok markovitása szerint, ha ismerjük a $\Pi\{1..k\}$ jelenlegi állapotot, akkor a múltbeli állapotokat definiáló $\Pi(1..k)$ és a jövőbeli állapotokat definiáló $\Pi(k+1..n)$ értékek már függetlenek egymástól, azaz a korábban használt jelöléssel $[k] \perp \overline{[k]} \,|\, \emptyset$.

Világos, hogy az L-felbonthatóságot gyengíthetjük úgy, hogy a $[k] \perp \overline{[k]} \,|\, \emptyset$ relációt csak bizonyos k-kra követeljük meg.

8.2. Definíció. *Legyen Π véletlen permutáció, eloszlása p. Π vagy p felbontható k-nál, ha $[k] \perp \overline{[k]} \,|\, \emptyset$.*

Ezzel a definícióval p akkor és csak akkor L-felbontható, ha minden k-nál felbontható. Minden $K \subseteq [n]$ esetén tekinthetjük a K-beli k-knál felbontható eloszlások $\mathbf{L}_K$ családját. Az ilyen családokat korlátozottan L-felbontható modelleknek hívjuk. Ha $K = \{k_1 < \ldots < k_j\}$, akkor a modellt a

$$P(\Pi(k_i + 1..k_{i+1}) = x|\Pi\{1..k_i\} = C)$$

valószínűségek paraméterezik, ahol x most megfelelő hosszúságú vektor, és x, C, i minden lehetséges értéket befut (i = 0-t is megengedve). Az egyes permutációk valószínűségei ilyen feltételes valószínűségek szorzataként állnak elő.

Még ennél is általánosabb felbonthatóságokat is definiálhatnánk, például valamilyen $\mathcal{U}$ halmazrendszerre tekinthetnénk azt az $\mathbf{L}_{\mathcal{U}}$ családot, melynek elemeire $U \perp \overline{U} \,|\, \emptyset$ minden $U \in \mathcal{U}$-ra. Mivel ilyen bonyolult modellek alkalmazását a gyakorlati megfontolások úgysem támogatnák, ezzel a kérdéssel nem foglalkozunk. (Illetve valamilyen formában később felbukkannak majd ilyen modellek.)

8.1. Markov bázis

Jelölje a továbbiakban $\mathbf{L}$ az összes S_n-en adott L-felbontható eloszlás családját. Itt n mindig egy tetszőleges rögzített pozitív egész szám. A 8.1. Definícióból látszik, hogy $n \leq 3$ esetén minden eloszlás L-felbontható, így igazából csak az $n \geq 4$ esetek érdekesek.

A (17) felírásból könnyen látszik, hogy $\mathbf{L}$ egy torikus modell, azaz található olyan M_L mátrix, mellyel $\mathbf{L} = \mathbf{F}(M_L)$. Jelölje a_n a (16)-ban szereplő (x, C) párok számát, azaz

$$a_n = \sum_{k=0}^{n-1} \binom{n}{k} (n-k) = n2^{n-1}.$$

Az M_L mátrix mérete $a_n \times n!$ lesz, oszlopait indexeljük a $\pi \in S_n$ permutációkkal, sorait indexeljük az (x, C) párokkal, és legyen az (x, C). sor és π. oszlop találkozásában álló elem

$$M_L((x, C), \pi) = \chi\{\pi\{1..|C|\} = C,\, \pi(|C|+1) = x\},$$

ahol χ az indikátor függvény.

Mivel az L-felbonthatóságot bizonyos feltételes valószínűségek egyenlőségével definiáltuk, az $\mathbf{L}$ család zárt. Azaz a 2.2. Tétel szerint egy p eloszlás akkor és csak akkor L-felbontható, ha zérussá teszi az I_{M_L} ideált generáló polinomokat. Ezeket fogjuk most megkeresni. Legyen $2 \leq k \leq n-2$, és a π_{11} és π_{22} permutációkra teljesüljön $\pi_{11}\{1..k\} = \pi_{22}\{1..k\}$, $\pi_{11}(1..k) \neq \pi_{22}(1..k)$,

és $\pi_{11}(k+1..n) \neq \pi_{22}(k+1..n)$. Definiáljuk a „keresztezett" permutációkat:

$$\pi_{12}(i) = \begin{cases} \pi_{11}(i) & \text{ha} \quad i \leq k \\ \pi_{22}(i) & \text{ha} \quad i > k \end{cases}$$
$$\pi_{21}(i) = \begin{cases} \pi_{22}(i) & \text{ha} \quad i \leq k \\ \pi_{11}(i) & \text{ha} \quad i > k \end{cases}$$

Minden ilyen választásra készítsünk el egy polinomot:

$$x_{\pi_{11}} x_{\pi_{22}} - x_{\pi_{12}} x_{\pi_{21}}. \tag{19}$$

Ezek a polinomok nyilvánvalóan I_{M_L}-hez tartoznak. Érvényes továbbá a következő.

8.3. Lemma. *A p eloszlás akkor és csak akkor L-felbontható, ha p zérushelye minden* (19)*-beli polinomnak.*

Bizonyítás. Csak az egyik irányt kell bizonyítani: tegyük fel, hogy p zérushelye minden (19)-beli polinomnak. Megmutatjuk, hogy p kielégíti a 8.1. Definíció első feltételét. Legyen x, x' két permutációja az $x_1, \ldots, x_k$ elemeknek. A feltételben szereplő feltételes valószínűségeket kiírva, azt kell megmutatni, hogy

$$\sum_{\eta} p(x, y, \eta) \sum_{\zeta} p(x', \zeta) = \sum_{\eta} p(x', y, \eta) \sum_{\zeta} p(x, \zeta), \tag{20}$$

ahol η az $\{x_1, \ldots, x_k, y\}$ halmaz komplementerének permutációit futja be, ζ pedig az $\{x_1, \ldots, x_k\}$ halmaz komplementerének permutációit futja be. Feltettük azonban, hogy minden η, ζ-ra

$$p(x, y, \eta) p(x', \zeta) - p(x', y, \eta) p(x, \zeta) = 0,$$

így (20) teljesül. ■

Azt is megmutatjuk, hogy a (19)-beli polinomok generálják az I_{M_L} ideált. Ez egyébként nem következik minden további nélkül a 8.3. Lemmából, mivel az

X_M *nemnegatív* torikus varietás esetleg az I_M bázisánál szűkebb egyenletrendszerrel is jellemezhető.

8.4. Tétel. *A* (19) *polinomok generálják az* I_{M_L} *ideált.*

Bizonyítás. A 2.5. Tételt fogjuk használni, azaz megmutatjuk, hogy a (19)-beli polinomokhoz tartozó f_i függvények Markov bázist alkotnak. Legyen indexhalmazunk

$$\mathcal{I} = \{i = i(C, \pi_1, \pi_2, \rho_1, \rho_2) : C \subseteq [n],\ 2 \leq |C| \leq n-2,\\ \pi_1, \pi_2 \in S_C,\ \pi_1 \neq \pi_2,\ \rho_1, \rho_2 \in S_{[n]\setminus C},\ \rho_1 \neq \rho_2\},$$

ahol S_C a C-beli elemek permutációinak halmazát jelöli. Ha $i \in \mathcal{I}$, akkor legyen $f_i : S_n \to \mathbb{Z}$ a következő függvény:

$$\begin{aligned} f_i(\pi_1, \rho_1) = f_i(\pi_2, \rho_2) &= -1, \\ f_i(\pi_1, \rho_2) = f_i(\pi_2, \rho_1) &= 1, \\ f_i(\sigma) &= 0 \text{ egyébként.} \end{aligned} \tag{21}$$

Világos, hogy $M_L f_i = 0$ minden i-re, tehát csak a lánc irreducibilitását kell igazolni. Legyen $u, v \in \mathbb{N}^{n!}$ két gyakoriságvektor, melyekre $M_L u = M_L v$. Megkonstruálunk egy u-ból v-be vezető, f_i lépéseket használó utat. Ha a $\mathbf{j} = (j_1, \ldots, j_k)$ vektor elemei $[n]$-beliek és mind különböznek, akkor jelölje $B_{\mathbf{j}} \subseteq S_n$ a $\mathbf{j}$-vel kezdődő S_n-beli permutációk halmazát, azaz azokat a π-ket, melyekre $\pi(s) = j_s$ ha $1 \leq s \leq k$. Vezessük be az $u(B_{\mathbf{j}}) = \sum_{\pi \in B_{\mathbf{j}}} u(\pi)$ egyszerűsítő jelölést. Vegyük észre, hogy $u(B_j) = v(B_j)$ minden j-re, hiszen $u(B_j) = (M_L u)(j, \emptyset)$, és erről feltettük, hogy u-ra és v-re megegyezik. Tegyük fel indukcióval, hogy az f_i lépések segítségével u-ból már eljutottunk egy olyan u^k gyakoriságvektorhoz, melyre $u^k(B_{\mathbf{j}}) = v(B_{\mathbf{j}})$ minden legfeljebb k hosszúságú $\mathbf{j}$ vektorra.

Legyen most C egy k elemű részhalmaza $[n]$-nek, és készítsük el u^k-ra és v-re külön-külön azt a $k! \times (n-k)$ méretű kontingenciatáblát, melynek sorait S_C elemeivel indexeljük, oszlopait pedig $[n] \setminus C$ elemeivel. A $(\mathbf{j}, x)$. cellába pedig írjuk be az $u^k(B_{\mathbf{j},x})$ illetve a $v(B_{\mathbf{j},x})$ értéket. Az u^k- és a v-táblának azonosak a marginálisai: a sorösszegek az indukciós feltevés miatt egyeznek

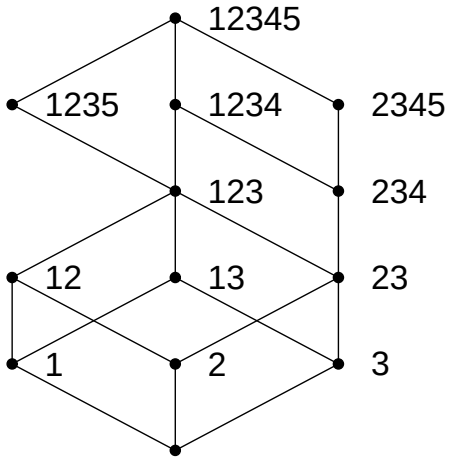
meg, az oszlopösszegek pedig az $(M_L u^k)(x,C) = (M_L v)(x,C)$ összefüggés miatt.

Ismert (pl. Diaconis és Sturmfels [32]), hogy a rögzített marginálisú kétdimenziós kontingenciatáblák esetén a $\begin{matrix} - & + \\ + & - \end{matrix}$ lépések Markov bázist alkotnak. Minden $r_1 \neq r_2$ és $c_1 \neq c_2$ választáshoz tartozik egy ilyen lépés: kivonunk egyet a tábla (r_1, c_1) és (r_2, c_2) celláiból, és hozzáadunk egyet az (r_1, c_2) és (r_2, c_1) cellákhoz. Ilyen lépésekkel tehát az u^k-tábla átvihető a v-táblába. Ezeket a lépéseket végrehajthatjuk az f_i függvényekkel: ha a kontingenciatábla-lépés a $\mathbf{j}$ és $\mathbf{j}'$ sorokon, és az x, x' oszlopokon hat, ahol $u^k(B_{\mathbf{j},x}) > 0$ és $u^k(B_{\mathbf{j}',x'}) > 0$ (azaz a lépés megléphető), akkor léteznek az $[n]\backslash(C \cup x)$ illetve $[n]\backslash(C \cup x')$ halmazoknak $\mathbf{g}$ illetve $\mathbf{g}'$ permutációi, melyekre $u^k(\mathbf{j}, x, \mathbf{g}) > 0$ és $u^k(\mathbf{j}', x', \mathbf{g}') > 0$. Az $i = i(C, \mathbf{j}, \mathbf{j}', (x, \mathbf{g}), (x', \mathbf{g}'))$ választással az f_i lépés éppen a kívánt kontingenciatábla-lépést eredményezi. Ha egymás után minden k elemű C halmazra egymásba alakítjuk a kontingenciatáblákat, eljutunk a következő u^{k+1} gyakoriságvektorhoz. Végül $u^n = v$. ■

Itt jegyezzük meg, hogy S_n elemei megfeleltethetők az n-dimenziós egységkocka, mint irányított gráf, maximális útjainak. A kocka minden csúcsa egy n hosszú 0-1 vektor, és akkor vezet él v-ből v'-be, ha v' úgy kapható v-ből, hogy egy 0 koordinátát 1-re cserélünk. Így minden permutáció a gráf egy **0**-ból **1**-be vezető irányított útjának felel meg: **0**-ból indulva sorra 1-re cseréljük a $\pi(1).$, $\pi(2).$, stb. koordinátákat.

Legyen p (17) alakú L-felbontható eloszlás, és (v, v') a kockagráf egy éle. Ha C jelöli v 1-koordinátáinak halmazát, és v'-t az x. koordináta 1-re cserélésével kaptuk, akkor írjuk a (v, v') élre a $\Lambda(x, C)$ paramétert. Így a π permutáció $p(\pi)$ valószínűsége az általa bejárt élekre írt paraméterek szorzata.

Vegyük észre, hogy a 8.4. Tétel fenti bizonyítása általánosabb esetben is működik. Tegyük fel, hogy egy G irányított gráfban nincs irányított kör, és egyetlen F forrás és egyetlen N nyelő van benne. Álljon az eseménytér az F-ből N-be vezető utakból. Ezen a téren tekintsük azt a torikus modellt, ahol egy út valószínűsége az általa bejárt élekhez rendelt nemnegatív paraméterek szorzata. A modellhez tartozó I_G ideált az „átkeresztezett utak” generálják,



2. ábra. Az S_5 egy 16 elemű részhalmazának gráfja (8.5. Példa)

vagyis az olyan $x_{s_{11}}x_{s_{22}} - x_{s_{12}}x_{s_{21}}$ polinomok, ahol az s_{11}, s_{22} utak valahol találkoznak, és ebben a találkozási pontban átkeresztezve őket az s_{12}, s_{21} utakat kapjuk. A bizonyításban arra kell csak figyelni, hogy a csúcsokat az F-ből hozzájuk vezető utak maximális hossza szerint vegyük sorra.

Ebből az észrevételből következik, hogy az $\mathbf{L}_K$ (szintén zárt torikus) modell Markov bázisa azokból a (19)-beli polinomokból áll (illetve az ezekhez tartozó függvényekből), melyeknél a szereplő permutációk K-beli helyen kereszteződnek át. Hiszen az $\mathbf{L}_K$-beli eloszlásokra is igaz, hogy egy π permutáció valószínűsége egy megfelelő gráf éleire írt paraméterek szorzata. Annyi csak a különbség, hogy ebben a gráfban többszörös élek is vannak, de ez nem okoz semmilyen gondot. Megjegyezzük még, hogy a család szabad paramétereinek száma ebben az esetben is könnyen kiszámolható.

8.5. Példa. Tekintsük például az S_5 kockagráfjának a 2. ábrán látható részgráfját! A gráfot úgy rajzoltuk le, hogy minden él alulról felfelé legyen irányítva, ezért az irányítást nem is jelöltük külön. A csúcsok mellé az 1-koordináták halmazát írtuk: a legalsó csúcs a **0**, a legfelső az **1**. A **0**-ból **1**-be vezető „legjobboldalibb" út a $\pi = (32451)$ permutációnak felel meg.

A gráfban 16 alulról felfelé irányított út van, ezek 16 permutációnak felelnek meg. A 2. táblázatban találjuk ezeket a permutációkat, valamint azokat

2. táblázat. Az u és v gyakoriságvektorok összekötése a Markov bázis elemeivel (8.5. Példa)

	u	{2,3}	{2,3,4}	{1,2,3}	{1,2,3}	{1,2,3}	v
12345	4			-1			3
12354	2			$+1$			3
13245	5			$+1$		-3	3
13254	1			-1		$+3$	3
21345	4				-1		3
21354	2				$+1$		3
23145	1	$+1$			$+1$		3
23154	4				-1		3
31245	0					$+3$	3
31254	6					-3	3
32145	4	-1					3
32154	3						3
23415	2		$+1$				3
23451	5	-1	-1				3
32415	4		-1				3
32451	1	$+1$	$+1$				3

az u és v gyakoriságvektorokat, melyeket a 8.4. Tételben szereplő Markov bázis elemeivel össze szeretnénk kötni. A táblázatban már maga az összekötés is szerepel, nézzük meg, hogyan is adódnak a lépések!

A 8.4. Tétel bizonyításában leírtak szerint sorra vesszük a gráf csúcsait, elemszám szerint növekvő sorrendben. Az első olyan csúcs, melyhez nemtriviális kontingencia-táblázat tartozik (azaz a sorok és oszlopok száma is legalább kettő), a {2,3} csúcs. A megfelelő 2×2-es u-kontingenciatábla:

	1	4
23	5	7
32	7	5

mivel u-ban pl. 5 darab 231-gyel kezdődő permutáció van. A megfelelő v-kontingenciatábla minden cellájában 6 van. Ezért a 231-gyel és a 324-gyel kezdődő permutációk gyakoriságaihoz 1-et kell hozzáadni, a 234-gyel és

321-gyel kezdődőek gyakoriságából pedig 1-et ki kell vonni. Báziselemeink segítségével ezt most négyféleképpen tehetnénk meg, hiszen ennyiféleképpen választhatunk ki egy-egy {1,2,3}-ból illetve {2,3,4}-ből {1,2,3,4,5}-be vezető utat. Válasszuk mondjuk a 45 és 51 utakat! A bázislépést a 2. táblázat u utáni első oszlopa mutatja. Ezután a {2,3,4} csúcsot vehetjük, melyhez szintén egy 2×2-es kontingenciatáblázat tartozik, és ezt is egy lépéssel a kívánt táblába transzformálhatjuk. Végül az {1,2,3} csúcs következik, 6×2-es kontingenciatáblákkal. Itt már több transzformációs lépésre van szükség. ■

A (21)-beli f_i Markov bázis lépéseit a klasszikus Metropolis algoritmus szerint módosíthatjuk úgy, hogy az $\mathcal{F}_u$ állapottéren hipergeometrikus stacionárius eloszlású, irreducibilis, aperiodikus, megfordítható Markov láncot kapjunk. Ehhez egy adott u állapothoz válasszuk $i \in \mathcal{I}$-t egyenletesen, és legyen $\epsilon = \pm 1$, $1/2$ valószínűséggel, i-től függetlenül. Ha $u' = u + \epsilon f_i$ nemnegatív, akkor lépjünk u'-be

$$\min\left(\prod_{\pi \in C_i} \frac{u(\pi)!}{u'(\pi)!}, 1\right)$$

valószínűséggel, ahol $C_i \subseteq S_n$ azokat a permutációkat jelöli, melyekre $f_i(\pi) \neq 0$. Minden más esetben maradjunk u-ban. Ez az algoritmus nagy n esetén is könnyen futtatható, anélkül, hogy a báziselemeket memóriában kellene tárolni. A $\mathcal{I}$ halmazból úgy választhatunk ki egy elemet, ha először választunk egy $2 \leq k \leq n-2$ számot, majd egy k elemszámú C részhalmazt, majd C két π_1, π_2 sorbarendezését, és $[n] \setminus C$ két ρ_1, ρ_2 sorbarendezését. Ahhoz, hogy a választott elem egyenletes eloszlású legyen $\mathcal{I}$-n, a k elemszámot válasszuk a $[k!-1][(n-k)!-1]$ kifejezéssel arányos valószínűséggel, az összes többi választás pedig legyen egyenletes.

Egy másik algoritmus (Diaconis és Sturmfels [32], Lemma 2.2), amely hosszabb lépéseket tesz, a következő. Legyen ismét f_i a Markov bázis véletlenszerűen választott eleme, u pedig a lánc jelenlegi állapota. Határozzuk meg azon egész j értékek halmazát, melyekre $u + j f_i \geq 0$, és lépjünk át egy ilyen

$u + jf_i$ állapotba a

$$\prod_{\pi \in C_i} \frac{1}{(u(\pi) + jf_i(\pi))!}$$

mennyiséggel arányos valószínűséggel, ahol C_i ugyanaz, mint az előbb. Ez a Markov lánc is irreducibilis, aperiodikus, megfordítható, és hipergeometrikus stacionárius eloszlású. Szerencsére ebben a konkrét esetben ez a lánc is könnyedén üzemeltethető: ha választásunk az $i = i(C, \pi_1, \pi_2, \rho_1, \rho_2)$ indexű báziselemre esett, akkor a

$a = u(\pi_1, \rho_1)$	$b = u(\pi_1, \rho_2)$
$c = u(\pi_2, \rho_1)$	$d = u(\pi_2, \rho_2)$

kontingenciatáblát egy ugyanilyen sor- és oszlopösszegekkel rendelkező nemnegatív táblával kell helyettesítenünk, melyet a hipergeometrikus eloszlás szerint kell kiválasztanunk. Ezt megtehetjük például úgy, hogy egyenletes eloszlás szerint generálunk egy $\pi' \in S_{a+b+c+d}$ permutációt, és a-t az $a' = |\{1 \le k \le a+b : 1 \le \pi'(k) \le a+c\}|$ értékkel helyettesítjük.

A szakasz végén megemlítjük, hogy elméleti szempontból lehet jelentősége annak, hogy egy Markov bázis minimális-e, illetve hogy hány minimális Markov bázis van. Egy f_i függvényekből álló Markov bázist akkor nevezünk minimálisnak, ha bármelyik f_i elemet elhagyva, a maradék függvények már nem alkotnak Markov bázist. Minimális Markov bázisokat vizsgált például Takemura és Aoki [57].

8.6. Tétel. *Az $n = 4$ és $n = 5$ esetben az* **L** *modell minimális Markov bázisa egyértelmű, és megegyezik a* (21) *képletben szereplő bázissal.*
Ha $n \ge 6$, akkor a (21) *képlet bázisa nem minimális, és a minimális Markov bázis nem egyértelmű.*

Bizonyítás. Az $n = 4$ vagy 5 eset: Azt kell megmutatni, hogy a megadott bázis minimális. Legyen f_i az $i = i(C, \pi_1, \pi_2, \rho_1, \rho_2)$ indexhez tartozó báziselem, és legyen u és v két gyakoriságvektor:

$$u(\pi_1, \rho_1) = u(\pi_2, \rho_2) = 1, u(\sigma) = 0 \text{ egyébként},$$
$$v(\pi_1, \rho_2) = v(\pi_2, \rho_1) = 1, v(\sigma) = 0 \text{ egyébként}.$$

Világos, hogy ezt a két gyakoriságvektort csak az f_i lépéssel lehet összeköt-ni, hiszen nincs is rajtuk kívül más gyakoriságvektor, mely ugyanezekkel az elégséges statisztikákkal rendelkezne. Azaz f_i nem hagyható el a bázisból.

Az $n \geq 6$ eset: Legyen i, u és v ugyanaz, mint az előbb, és $|C| = k$. Az előző esethez képest annyi a különbség, hogy a (π_1, ρ_1) és a (π_2, ρ_2) permutációkkal előfordulhat, hogy nem csak a 0. és a k. elem után „válnak szét". Például $n = 6$-ra az (123456) és a (214365) permutációk a $0., 2., 4.$ elemek után válnak szét. Általában, ha a szétválások száma h, akkor 2^{h-1} gyakoriságvektor rendelkezik ugyanazokkal az elégséges statisztikákkal, mint u (és persze v). Egy minimális Markov bázis lépései ezeket a gyakoriságvektorokat egy fává kapcsolják össze. Ezt a fát pedig az eredeti báziselemekből többféleképpen kiválaszthatjuk. Az előző példát tekintve, az $i = i(\{1,2\}, 12, 21, 3456, 4365)$ indexhez tartozó lépést elhagyhatjuk, mert ez a lépés helyettesíthető az

$$\begin{aligned} i_1 &= i(\{1,2,3,4\}, 1234, 2143, 56, 65), \\ i_2 &= i(\{1,2\}, 12, 21, 3465, 4356), \\ i_3 &= i(\{1,2,3,4\}, 1243, 2134, 56, 65) \end{aligned}$$

indexekhez tartozó három lépéssel. ■

Tehát az egyértelmű minimális Markov bázis $n = 4$-re 6 elemű, $n = 5$-re pedig 270 elemű.

8.2. Paraméterbecslés

Ebben a szakaszban az L-felbontható eloszláscsalád paramétereinek maximum likelihood (ML) becslésével foglalkozunk. Legyen $\pi_1, \ldots, \pi_m$ iid minta a (18) kanonikus felbontású p L-felbontható eloszlásból. Jelölje a minta gyakoriságvektorát $(f(\pi) : \pi \in S_n)$. Legyen (x, C) a kocka élgráfjának egy éle, és legyen $g_f(x, C) = \sum\{f(\pi) : (x, C) \in \pi\}$, ahol $(x, C) \in \pi$ azt jelenti, hogy π átmegy az (x, C) élen. Vezessük még be a

$$h_f(C) = \sum_{x \notin C} g_f(x, C) = \sum_{y \in C} g_f(y, C \setminus y)$$

jelölést a C csúcson átmenő mintabeli utak (permutációk) számára. Az $\mathbf{L}$ családot a (18)-beli kanonikus L-felbontás paraméterezi, azaz a szabad paraméterek száma

$$b_n = \sum_{k=2}^{n} \binom{n}{k} (k-1) = 2^n(n/2-1)+1.$$

Az $\mathbf{L}$ modellben a paraméterek jól interpretálhatóak: feltételes valószínűségeket jelentenek. A paraméterek maximum likelihood becslése pedig éppen a mintából számolt megfelelő tapasztalati feltételes valószínűség, azaz a $\Lambda(x,C)$ kanonikus paraméter maximum likelihood becslése $g_f(x,C)/h_f(C)$. Ha a mintánkban található összes permutáció útját pirossal kihúzzuk a gráfon, akkor a maximum likelihood becslésként kapott eloszlás pontosan azokhoz a permutációkhoz rendel pozitív valószínűséget, melyek csak piros éleken mennek át (ez a legszűkebb olyan M_L-megvalósítható halmaz, mely a mintabeli permutációkat tartalmazza). Az eloszlás $\hat{p}_f$ ML becslése tehát:

$$\hat{p}_f(\pi) = \prod_{k=0}^{n-1} \frac{g_f(\pi(k+1), \pi\{1..k\})}{h_f(\pi\{1..k\})}. \tag{22}$$

Ebben az esetben a ML becslés pontos eloszlása kiszámítható. Annak valószínűsége ugyanis, hogy a minta gyakoriságvektora éppen f legyen:

$$P(f) = \frac{N!}{\prod_\pi f(\pi)!} \prod_{(x,C)} \Lambda(x,C)^{g_f(x,C)}.$$

Ebből pedig annak valószínűsége, hogy a ML becslés pont $\hat{p}$ legyen:

$$P(\hat{p}_f = \hat{p}) = P(g_f = g) = H(g) \prod_{(x,C)} \Lambda(x,C)^{g(x,C)},$$

ahol

$$H(g) = \sum_{f: g_f = g} \frac{N!}{\prod_\pi f(\pi)!} = \sum_{\pi_1,\ldots,\pi_m : g_f = g} 1.$$

$H(g)$ azt mondja meg, hogy hány olyan m elemű minta van, amely a g elégséges statisztikát produkálja. Felhasználtuk továbbá, hogy $\hat{p}$ és g között

kölcsönösen egyértelmű megfeleltetés van.

8.7. Lemma.

$$H(g) = \prod_{C \subseteq [n]} \frac{h(C)!}{\prod_{x \notin C} g(x, C)!}.$$

Bizonyítás. Meg kell számolnunk, hogy hány olyan $\pi_1, \ldots, \pi_m$ minta van, melynek elégséges statisztikája g, azaz minden (x, C) párra $g(x, C)$ darab $1 \leq \leq i \leq m$ indexre teljesül, hogy $\pi_i\{1..|C|\} = C$ és $\pi_i(|C|+1) = x$. A következő eljárás az összes ilyen mintát előállítja (egyszeres multiplicitással). Először is legyen a $\pi_1(1), \ldots, \pi_m(1)$ sorozat $g(1, \emptyset)$ darab 1-es, ..., $g(n, \emptyset)$ darab n-es egy tetszőleges ismétléses permutációja. Ezután, ha a mintaelemek első k koordinátája már megvan, akkor az $\{i : \pi_i\{1..k\} = C\}$ indexű mintaelemek $k+$ $+1$. koordinátáinak sorozata legyen $g(1, C)$ darab 1-es, ..., $g(n, C)$ darab n-es egy tetszőleges ismétléses permutációja, ahol definíció szerint $g(x, C) = 0$, ha $x \in C$. A lemma állítása ezután az ismétléses permutációk darabszámainak összeszorzásával adódik. ■

Megkaptuk tehát az elégséges statisztika, illetve a ML becslés pontos eloszlását.

8.8. Tétel. *Legyen p a* (18)*-beli kanonikus felbontású L-felbontható eloszlás, $\hat{p}_f$ pedig az eloszlásból származó, f gyakoriságvektorú iid mintához tartozó* (22) *ML becslés. Ekkor*

$$P(\hat{p}_f = \hat{p}) = P(g_f = g) = \prod_{C \subseteq [n]} h(C)! \prod_{x \notin C} \frac{\Lambda(x, C)^{g(x,C)}}{g(x, C)!}. \tag{23}$$

Láttuk, hogy ha Π eloszlása L-felbontható, akkor a $\Pi(1..k)$ és a $\Pi(k+1..n)$ véletlen vektorok feltételesen függetlenek a $\Pi\{1..k\}$ véletlen halmazra nézve. Hasonló állítás a ML becslésre is érvényes, ezt nevezi Dawid és Lauritzen [25] hiper-Markov tulajdonságnak, mely felbontható grafikus modellek esetén is teljesül.

8.9. Tétel. *Legyen $v \in S_n$, $1 \leq k \leq n-1$ pedig rögzített. Jelölje f egy m elemű iid minta gyakoriságvektorát, $\hat{P}$ pedig a* (22)*-beli ML becslés szerinti*

valószínűségeket. Ekkor

$$\begin{aligned}
\hat{P}(\Pi(1..k) = v(1..k)) &= \prod_{j=0}^{k-1} \frac{g_f(v(j+1), v\{1..j\})}{h_f(v\{1..j\})}, \\
\hat{P}(\Pi(k+1..n) = v(k+1..n)) &= \prod_{j=k}^{n-1} \frac{g_f(v(j+1), v\{1..j\})}{h_f(v\{1..j\})} \frac{h_f(v\{1..k\})}{m}, \\
\hat{P}(\Pi\{1..k\} = v\{1..k\}) &= \frac{h_f(v\{1..k\})}{m}.
\end{aligned}$$

Valamint $\{\hat{P}(\Pi(1..k) = u)\}_u$ *és* $\{\hat{P}(\Pi(k{+}1..n) = v)\}_v$ *feltételesen függetlenek* $\{\hat{P}(\Pi\{1..k\} = C)\}_C$*-re, ahol* u, v, C *az összes lehetőséget befutja.*

Bizonyítás. A három egyenlet könnyen igazolható (pl. indukcióval). A feltételes függetlenséghez: $\{\hat{P}(\Pi(1..k) = u)\}_u$ a $\{g(x,C)\}_{|C|\le k-1}$ változók függvénye, $\{\hat{P}(\Pi(k+1..n) = v)\}_v$ pedig a $\{g(x,C)\}_{|C|\ge k}$ változók függvénye. Végül $\hat{P}(\Pi\{1..k\} = C)\}_C$ kölcsönösen egyértelmű kapcsolatban van a $\{h(C)\}_{|C|=k}$ változókkal. A (23) egyenletből pedig látszik, hogy ezek között fennáll a feltételes függetlenség. ■

Vegyük észre, hogy a 8.7. Lemma bizonyításában leírt eljárás lehetőséget ad arra, hogy a $\pi_1, \ldots, \pi_m$ mintának a g_f elégséges statisztikára vett feltételes eloszlásából (ami éppen az egyenletes eloszlás az összes ilyen elégséges statisztikájú mintán) közvetlenül generáljunk. A leírt módszerrel az $n \times m$-es adatmátrixot, melynek oszlopai a $\pi_1, \ldots, \pi_m$ permutációk, soronként generálhatjuk. Ugyanilyen jó módszer az oszloponkénti generálás: menjünk először végig a kockagráfon úgy, hogy minden csúcsból az élgyakoriságokkal arányos valószínűséggel megyünk tovább. Így kapjuk a π_1 mintaelemet. A π_1 által bejárt élek gyakoriságából vonjunk le egyet, és folytassuk az eljárást.

A direkt generálás kivitelezhetősége nem feltétlenül teszi feleslegessé az előző szakasz Markov bázis számításait. Egyrészt, nagyobb n, m esetén a Monte Carlo módszert könnyebb lehet üzemeltetni, másrészt, látni fogjuk, hogy az előző szakaszban megtalált Markov bázis a bonyolultabb modelleknél is megjelenik majd. Megjegyezzük még, hogy ennek a szakasznak minden eredménye könnyen átfogalmazható az $\mathbf{L}_K$ családra, illetve tetszőleges irányított forrás-nyelő gráf által definiált torikus modellre.

8.3. Illeszkedésvizsgálat

Az **L** modell illeszkedésének vizsgálatára többféle módszert alkalmazhatunk. Problémát jelent, hogy a $H_0 : p \in \mathbf{L}$ hipotézis különböző tartójú exponenciális családok uniója, és ezek az exponenciális családok különböző paraméterszámmal rendelkeznek. Gyakran természetes a $H_0 : p \in \mathbf{E}(M_L)$ hipotézist feltenni, azaz a szigorúan pozitív L-felbontható eloszlásokra koncentrálni. Azonban a gyakorlatban az $n!$-hoz képest kis mintaelemszámok esetén gyakran előfordul, hogy az $\mathbf{E}(M_L)$ családban nem létezik ML becslés, azaz az **L**-beli ML becslés nem teljes tartójú. Ha a ML becslés teljes tartójú, akkor használhatjuk a becsléses illeszkedésvizsgálatra vonatkozó klasszikus χ^2-próbát, melynek szabadsági foka $n! - 2^n(n/2 - 1) - 2$.

Kis mintaelemszám esetén az illeszkedés jóságát Monte Carlo módszerrel vizsgálhatjuk, a korábban leírt módon. Azaz másodlagos mintákat generálunk a megfigyelt elégséges statisztikával rendelkező minták teréből, egyenletes eloszlás szerint, és a másodlagos minták χ^2-statisztikáit vetjük össze az eredeti mintáéval. Ebben az esetben a χ^2 statisztika helyett más, az illeszkedés jóságát mérő statisztikát is használhatunk. Ahogy korábban leírtuk, másodlagos mintákat közvetlenül, vagy Markov lánc módszerrel is generálhatunk.

Az $n = 4$ eset különösen egyszerű. Ekkor csak egy felbonthatóságot kell ellenőrizni, a $k = 2$ vágáshoz tartozót. Ez azt jelenti, hogy a kételemű részhalmazoknak megfelelő 6 darab 2×2-es kontingenciatáblázatban kell a függetlenségnek teljesülni: a χ^2 statisztika e 6 táblázat egy szabadsági fokú, független χ^2 statisztikáinak összege. Ebben az esetben az adott elégséges statisztikával rendelkező gyakoriságvektorok is könnyen felsorolhatók: a hat darab, rögzített marginálisú 2×2-es kontingenciatáblát kell nemnegatív egész számokkal kitölteni.

Erre az esetre ad példát a Croon [16] cikk adatsora, melyet többek között [8] és [32] is elemez. Egy kutatás során 2262 német állampolgár a következő négy politikai célt állította preferencia-sorrendbe: (1) A rend fenntartása, (2) Kapjanak az emberek több beleszólást az ország ügyeibe, (3) Az infláció megfékezése, (4) A szólásszabadság védelme. A kutatás abból a hipotézisből indult ki, hogy az emberek két csoportra oszthatók: a „liberálisok" inkább a

(2)-es és (4)-es célokat részesítik előnyben, míg a „konzervatívok" az (1)-est és a (3)-ast. A következő 6 kontingenciatáblában a sorok a sorrend első két elemét, az oszlopok az utolsó két elemét jelentik, azaz pl. 137-en állították fel az (1234) sorrendet.

A	34	43
12	137	29
21	48	23

B	24	42
13	309	255
31	330	294

C	23	32
14	52	93
41	21	30

D	14	41
23	61	55
32	117	69

E	13	31
24	33	59
42	29	52

F	12	21
34	70	34
43	35	27

Az A, B, C, D, E, F táblák χ^2 statisztikája rendre 6.47, 0.43, 0.46, 3.14, 0.00, 1.97. Feltűnő, hogy a B, E táblák statisztikája a legkisebb, ezek azokat a válaszadókat tartalmazzák, akik a „liberális" célokat az első két vagy az utolsó két helyre rangsorolják. A függetlenség egyedül az A tábla esetén nem elfogadható: láthatjuk, hogy a függetlenség esetén vártnál többen választották az (1234) sorrendet. Elképzelhető, hogy egyes válaszadók gondolkodás nélkül ezt a sorrendet adták meg. Az is kiszámolható, hogy összesen 692723631600 gyakoriságvektor adja ezeket a táblamarginálisokat.

Nézzünk ezután néhány szimulációs eredményt! A permutációk hosszúsága $n = 5$ illetve 6 lesz, a mintanagyságot pedig jelölje most is m. Háromféle eloszlásból generálunk mintát: az első az egyenletes, a második a Plackett-Luce eloszlás, $\lambda_i = i$ választással. Ez a két eloszlás L-felbontható. A harmadik eloszlást véletlen kereséssel választottuk úgy, hogy viszonylag messze legyen $\mathbf{L}$-től, külön $n = 5$-re és $n = 6$-ra. A divergencia $D(P||\mathbf{L}) = 0.1748$ az $n = 5$ esetben, és $D(P||\mathbf{L}) = 0.1899$ az $n = 6$ esetben. Itt jegyezzük meg, hogy érdekes lenne tudni, hogy melyik eloszlás(ok) van(nak) a legmesszebb az $\mathbf{L}$ modelltől (divergencia értelemben). Minden esetben 50 mintát generáltunk, és minden mintához 500 Monte Carlo mintánk volt (direkt generálással). Az aszimptotikus kritikus érték az $n = 5$ esetben 90.53, az $n = 6$ esetben 647.62 ($\alpha = 0.05$). Az eredményeket a 3., 4. és 5. táblázatokban foglaltuk össze. A

3. táblázat. Monte Carlo illeszkedésvizsgálat: egyenletes eloszlás

	$n=5$			$n=6$			
$m \to$	60	120	240	180	360	720	1440
$\#\{\chi^2 > c_{\text{asz}}\}$	0	4	3	0	2	2	2
$\#\{\chi^2 > c_{\text{MC}}\}$	5	4	2	1	1	2	2
átl. c_{MC}	73.22	90.29	91.32	608.61	652.01	649.10	649.54
átl. tartó	96	117	120	654	717	720	720

4. táblázat. Monte Carlo illeszkedésvizsgálat: Plackett-Luce eloszlás

	$n=5$			$n=6$			
$m \to$	60	120	240	180	360	720	1440
$\#\{\chi^2 > c_{\text{asz}}\}$	0	0	0	0	0	0	2
$\#\{\chi^2 > c_{\text{MC}}\}$	3	2	0	1	1	3	1
átl. c_{MC}	52.06	70.52	83.61	397.59	522.62	606.93	641.76
átl. tartó	69	93	110	423	561	661	707

$\#\{\chi^2 > c_{\text{asz}}\}$ sor azt mutatja, hogy az 50 mintából hányszor utasítottuk el az L-felbonthatóság hipotézisét az aszimptotikus χ^2 próbával. A következő, $\#\{\chi^2 > c_{\text{MC}}\}$ sor pedig a Monte Carlo módszerrel elutasított minták számát adja meg. Az ezután következő „átl. c_{MC}" mennyiség a Monte Carlo kritikus értékek átlaga az 50 mintából, „átl. tartó" pedig a ML becslés tartójának átlagos elemszáma (egészre kerekítve).

Látható, hogy az egyenletes eloszlás esetén az $n = 5$ esetben már $m = 120$, az $n = 6$ esetben már $m = 360$ mintanagyságra a kétféle próba majdnem ugyanazt az eredményt adja, és a ML becslések tartója majdnem teljes. Hasonlót mondhatunk a nem L-felbontható eloszlásra is. A Plackett-Luce eloszlás esetén azonban nagyobb elemszámra van szükség ahhoz, hogy a kétféle próba hasonló eredményt adjon.

Adott minta esetén feladatunk lehet annak eldöntése, hogy az elméleti eloszlás melyik k-knál felbontható. Egy adott k-nál a felbonthatóságot köny-

5. táblázat. Monte Carlo illeszkedésvizsgálat: nem L-felbontható eloszlás

	$n=5$			$n=6$			
$m \to$	60	120	240	180	360	720	1440
$\#\{\chi^2 > c_{\text{asz}}\}$	2	37	50	2	41	50	50
$\#\{\chi^2 > c_{\text{MC}}\}$	16	41	50	13	42	50	50
átl. c_{MC}	72.63	88.83	91.24	596.82	650.45	650.56	648.86
átl. tartó	94	115	120	638	715	720	720

nyű vizsgálni, ahogy az $n = 4$ esetnél már láttuk: a függetlenségnek $\binom{n}{k}$ darab $k! \times (n-k)!$ méretű kontingenciatáblán kell teljesülnie. Mivel ezen táblák χ^2 statisztikái függetlenek egymástól, a statisztikákat összeadva aszimptotikusan

$$\binom{n}{k}[(k!-1)((n-k)!-1)]$$

szabadsági fokú χ^2 eloszlású próbastatisztikát kapunk a felbonthatóság hipotézise mellett. Amennyiben kevés adatunk van, akkor Monte Carlo módszert használhatunk, azaz a rögzített marginálisú táblákon Markov láncot futtatva értékeljük ki az illeszkedés jóságát. Ha egyszerre több k-nál szeretnénk a felbonthatóságot vizsgálni, akkor a különböző elemszámú csúcsokhoz tartozó táblák χ^2 statisztikái már nem függetlenek. A korlátozottan L-felbontható modellek közül a legjobb megtalálásához ajánlható például a következő módszer: először egyesével teszteljük a felbonthatóságokat, majd illesszük azt a modellt, mely az így elfogadott felbonthatóságokat tartalmazza. Ezután még megnézhetjük, hogy jobb modellt kapunk-e egy-egy felbonthatóság elhagyásával, vagy bevételével.

9. Duplán L-felbonthatóság

Ahogy a bevezetésben már említettük, egy párosítási, sorbarendezési, vagy átrendezési kísérlet eredményét ugyanúgy megadhatjuk egy permutációval vagy annak inverzével. Vizsgálhatjuk, hogy egyik vagy másik megadás L-felbontható eloszlást eredményez-e S_n-en, de azzal a hipotézissel is élhetünk, hogy mindkét megadás L-felbontható.

9.1. Definíció. *A Π véletlen permutáció (vagy eloszlása) duplán L-felbontható, ha Π és Π^{-1} is L-felbontható.*

Jelölje a duplán L-felbontható eloszlások családját a továbbiakban $\mathbf{B}$. $\mathbf{L}$-hez hasonlóan definiálhatjuk az invertálva L-felbontható eloszlások $\mathbf{L}'$ családját, azaz

$$\mathbf{L}' = \{p : \mathrm{inv}(p) \in \mathbf{L}\},$$

ahol $(\mathrm{inv}(p))(\pi) = p(\pi^{-1})$. Természetesen $\mathbf{L}'$ is torikus modell, a hozzá tartozó $M_{L'}$ mátrixot úgy kapjuk M_L-ből, hogy a (π, π^{-1}) inverzpároknak megfelelő oszlopokat páronként felcseréljük. Látni fogjuk, hogy $\mathbf{B}$, amely az $\mathbf{L}$ és $\mathbf{L}'$ torikus modellek metszete, maga nem torikus modell, ezért ez a család nehezebben kezelhető, mint $\mathbf{L}$.

Mostantól egy darabig csak a szigorúan pozitív duplán L-felbontható eloszlásokkal fogunk foglalkozni, vagyis az $\mathbf{E}(M_L) \cap \mathbf{E}(M_{L'})$ metszettel. Vezessünk be két új jelölést! Legyen $F = \mathrm{Im}M_L^\top$ az $M_L^\top : \mathbb{R}^{a_n} \to \mathbb{R}^{n!}$ lineáris operátor képtere, és $F' = \mathrm{Im}M_{L'}^\top$. Ezt a két alteret egyszer és mindenkorra rögzítsük! Ezzel a jelöléssel $p \in \mathbf{E}(M_L)$ akkor és csak akkor, ha $\log p \in F$, ahol $\log(\cdot)$ koordinátánként értendő. Ebből adódik, hogy p akkor és csak akkor lesz szigorúan pozitív duplán L-felbontható eloszlás, ha $\log p \in F \cap F'$, és ezek az eloszlások is exponenciális családot alkotnak. Egy darabig azon az ártatlannak tűnő feladaton fogunk dolgozni, hogy meghatározzuk az $F \cap F'$ altér dimenzióját, illetve egyszerű bázisát.

A feladat megoldását a merőlegességek teszik majd lehetővé. Az U és V alterekről azt mondjuk, hogy merőlegesen metszik egymást, ha U-nak V-re vett merőleges vetülete éppen $U \cap V$, vagy ami ezzel ekvivalens, ha V-nek U-ra vett merőleges vetülete éppen $U \cap V$. Jelölje az U-ra való merőleges

vetítés operátorát Pr_U. A merőleges metszéssel ekvivalens az a tulajdonság is, hogy a Pr_U és Pr_V operátorok kommutálnak. A merőleges metszésre a $\perp_\cap$ jelölést használva kapjuk, hogy

$$U \perp_\cap V \iff Pr_U V = U \cap V \iff Pr_V U = U \cap V \iff Pr_U Pr_V = Pr_V Pr_U.$$

Azt fogjuk megmutatni, hogy F és F' merőlegesen metszik egymást, sőt, mind F-et, mind F'-t fel tudjuk majd bontani páronként merőleges F_k és F'_ℓ alterekre úgy, hogy, minden (F_k, F'_ℓ) altér-pár merőlegesen messe egymást. Ezután elég lesz az $F_k \cap F'_\ell$ alacsony dimenziós alterek dimenzióját és bázisát meghatározni.

9.1. Hierarchikus modellek permutációkra

Először egy kicsit általánosabb vizekre evezünk: definiáljuk, hogy mit értünk hierarchikus modellen véletlen permutációk esetén. A klasszikus hierarchikus modellek terminológiáját használva, a modellt generátorok fogják definiálni. Egy-egy generátor pedig az $[n] \times [n]$ halmaz egy-egy szorzatpartíciója lesz.

Az $[n]$ halmaz egy partíciója

$$\mathcal{D} = \{D_1, \ldots, D_d\} : \cup_{i=1}^{d} D_i = [n],\ D_i \cap D_j = \emptyset\ \forall i \neq j.$$

A D_i halmazok a partíció osztályai, ezekről mindig feltesszük, hogy nem üresek. Ha $\mathcal{D}$ és $\mathcal{R}$ két partíció $[n]$-en, akkor a $\mathcal{D} \times \mathcal{R}$ szorzatpartíció az $[n] \times [n]$ halmazt partícionálja.

Legyen $\mathcal{D}$ (illetve $\mathcal{R}$) az $[n]$ halmaz egy d (illetve r) osztályú partíciója. A π permutáció *durvítása* a $\mathcal{D} \times \mathcal{R}$ szorzatpartícióra a következő $d \times r$-es mátrix:

$$|\pi(\mathcal{D} \times \mathcal{R})| = (t_{ij}), \quad t_{ij} = |\{1 \leq s \leq n : s \in D_i,\ \pi(s) \in R_j\}|. \tag{24}$$

Minden $\mathcal{P} = \mathcal{D} \times \mathcal{R}$ szorzatpartícióhoz hozzárendelhető egy $U_\mathcal{P} \subseteq \mathbb{R}^{n!}$ lineáris altér a következőképpen. Legyenek az $\mathbb{R}^{n!}$ euklideszi tér elemei a $v =$

$= (v(\pi) : \pi \in S_n)$ vektorok, a kifeszített altérre pedig használjuk a Span$(\cdot)$ jelölést. Ekkor

$$U_{\mathcal{P}} = \{v \in \mathbb{R}^{n!} : |\pi(\mathcal{P})| = |\sigma(\mathcal{P})| \Rightarrow v(\pi) = v(\sigma)\}, \tag{25}$$

azaz $v \in U_{\mathcal{P}}$ akkor és csak akkor, ha létezik a $d \times r$-es mátrixokon olyan θ függvény, mellyel $v(\pi) = \theta(|\pi(\mathcal{P})|)$. Most már definiálhatjuk a hierarchikus modellt.

9.2. Definíció. *Legyenek $\mathcal{P}_1, \ldots, \mathcal{P}_s$ az $[n] \times [n]$ szorzatpartíciói. Az S_n-en adott szigorúan pozitív p eloszlás akkor tartozik a $\mathcal{P}_1, \ldots, \mathcal{P}_s$ generátorok definiálta hierarchikus modellhez, jelölésben $p \in \mathcal{L}(\mathcal{P}_1, \ldots, \mathcal{P}_s)$, ha*

$$\log p(\pi) = \sum_{i=1}^{s} \theta_i(|\pi(\mathcal{P}_i)|) \quad \forall \pi \in S_n$$

valamilyen θ_i függvényekre. Ezzel ekvivalens, hogy

$$\log p \in \mathit{Span}(U_1, \ldots, U_s),$$

ahol az $U_{\mathcal{P}_i} = U_i$ egyszerűsítő jelölést használtuk.

Ismert, hogy a klasszikus hierarchikus modellek esetén az $\mathcal{A}$ és az $\mathcal{A}'$ generátorrendszerű modellek metszete szintén hierarchikus modell, melynek generátorai az $\{A \cap A' : A \in \mathcal{A}, A' \in \mathcal{A}'\}$ halmazok (a hierarchikus modell alatt most csak a szigorúan pozitív eloszlásokat értjük). Ez a következőképpen látható be. Az $\mathcal{A}$-hierarchikus modell A generátoraihoz hasonló módon rendelhetők U_A lineáris alterek, mint a permutációk hierarchikus modelljei esetében. A 9.3. Lemma alapján könnyű belátni, hogy minden $A, A' \subseteq [n]$ generátorpárra U_A és $U_{A'}$ merőlegesen metszi egymást, valamint $U_A \cap U_{A'} = = U_{A \cap A'}$. Az állítás ezután a 9.5. Lemmából következik.

Permutációk esetén a helyzet nem ilyen egyszerű, mivel nem minden $\mathcal{P}, \mathcal{P}'$ partíció-párra metszik a megfelelő alterek merőlegesen egymást. Elégséges feltételt adunk azonban arra, hogy két hierarchikus modell metszete hierarchikus modell legyen, melynek generátorai is azonosíthatók.

9.3. Lemma. *Legyen $(\Omega, \mathcal{A}, P)$ valószínűségi mező, és jelölje $L_2(\mathcal{A})$ a négyzetesen integrálható valószínűségi változók Hilbert terét. A $\mathcal{B} \subseteq \mathcal{A}$ σ-algebrára jelölje $L_2(\mathcal{B})$ a $\mathcal{B}$-mérhető valószínűségi változók zárt lineáris alterét $L_2(\mathcal{A})$-ban. Legyen $\mathcal{B}_1$, $\mathcal{B}_2 \subseteq \mathcal{A}$. Ekkor $L_2(\mathcal{B}_1) \perp_\cap L_2(\mathcal{B}_2)$ akkor és csak akkor, ha $\mathcal{B}_1$ és $\mathcal{B}_2$ feltételesen független $\mathcal{B}_1 \cap \mathcal{B}_2$-re nézve.*

Bizonyítás. Mivel $L_2(\mathcal{B}_1 \cap \mathcal{B}_2) = L_2(\mathcal{B}_1) \cap L_2(\mathcal{B}_2)$, az $L_2(\mathcal{B}_1)$ és $L_2(\mathcal{B}_2)$ terek akkor és csak akkor metszik egymást merőlegesen, ha minden $f \in L_2(\mathcal{B}_1)$, $g \in L_2(\mathcal{B}_2)$ esetén

$$\mathbb{E}\big(\big[f - \mathbb{E}(f \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big]\big[g - \mathbb{E}(g \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big]\big) = 0.$$

A feltételes függetlenség teljesülése esetén a következő erősebb egyenlőség is igaz:

$$\mathbb{E}\big(\big[f - \mathbb{E}(f \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big]\big[g - \mathbb{E}(g \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big] \mid \mathcal{B}_1 \cap \mathcal{B}_2\big) = 0.$$

Fordítva, ha a két altér merőlegesen metszi egymást, akkor legyen $E_1 \in$ $\in \mathcal{B}_1$, $E_2 \in \mathcal{B}_2$ két esemény, és jelölje C azt az eseményt, hogy

$$P(E_1 \cap E_2 \mid \mathcal{B}_1 \cap \mathcal{B}_2) - P(E_1 \mid \mathcal{B}_1 \cap \mathcal{B}_2)P(E_2 \mid \mathcal{B}_1 \cap \mathcal{B}_2) > 0.$$

Bevezetve az $f = \chi(E_1)\chi(C)$ és $g = \chi(E_2)\chi(C)$ jelöléseket,

$$\begin{aligned}
&\mathbb{E}\big(\big[f - \mathbb{E}(f \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big]\big[g - \mathbb{E}(g \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big]\big) = \\
&\qquad \mathbb{E}\left(\mathbb{E}(fg \mid \mathcal{B}_1 \cap \mathcal{B}_2) - \mathbb{E}(f \mid \mathcal{B}_1 \cap \mathcal{B}_2)\mathbb{E}(g \mid \mathcal{B}_1 \cap \mathcal{B}_2)\right) = \\
&= \mathbb{E}\big(\chi(C)\big[P(E_1 \cap E_2 \mid \mathcal{B}_1 \cap \mathcal{B}_2) - P(E_1 \mid \mathcal{B}_1 \cap \mathcal{B}_2)P(E_2 \mid \mathcal{B}_1 \cap \mathcal{B}_2)\big]\big) = 0.
\end{aligned}$$

Ez csak úgy lehetséges, ha $P(E_1 \cap E_2 \mid \mathcal{B}_1 \cap \mathcal{B}_2) - P(E_1 \mid \mathcal{B}_1 \cap \mathcal{B}_2)P(E_2 \mid$ $\mid \mathcal{B}_1 \cap \mathcal{B}_2) \leq 0$ teljesül 1 valószínűséggel. A fordított egyenlőtlenség hasonlóan igazolható, azaz E_1 és E_2 feltételesen függetlenek a $\mathcal{B}_1 \cap \mathcal{B}_2$ σ-algebrára. ■

A partíciók között tekintsük a következő részbenrendezést. A $\mathcal{D}' =$ $= (D'_1, \ldots, D'_{d'})$ partíció finomabb a $\mathcal{D} = (D_1, \ldots, D_d)$ partíciónál (vagy

$\mathcal{D}$ durvább $\mathcal{D}'$-nél), ha minden i-hez van olyan j, hogy $D'_i \subseteq D_j$. Jelölje ezt a relációt $\mathcal{D}' \succeq \mathcal{D}$. Világos, hogy ebből $U_{\mathcal{D}'} \supseteq U_{\mathcal{D}}$ következik.

9.4. Lemma. *Legyenek $\mathcal{D}' \succeq \mathcal{D}$ és $\mathcal{R}' \succeq \mathcal{R}$ az $[n]$ partíciói. Ekkor*

$$U_{\mathcal{D}\times\mathcal{R}'} \perp_\cap U_{\mathcal{D}'\times\mathcal{R}} \quad \textit{és} \quad U_{\mathcal{D}\times\mathcal{R}'} \cap U_{\mathcal{D}'\times\mathcal{R}} = U_{\mathcal{D}\times\mathcal{R}}. \tag{26}$$

Bizonyítás. Alkalmazzuk a 9.3. Lemmát az S_n-en egyenletes eloszlásra. Az egyenletesség miatt az L_2-beli merőlegesség az $\mathbb{R}^{n!}$-beli merőlegességgel ekvivalens. Legyen $\mathcal{P}$ az $[n] \times [n]$ halmaz partíciója. Jelölje a továbbiakban $\sigma(\mathcal{P})$ az

$$X_{\mathcal{P}} : \pi \mapsto |\pi(\mathcal{P})|$$

valószínűségi változó által generált σ-algebrát S_n-en. Ezt talán pontosabb lenne $\sigma(X_{\mathcal{P}})$-vel jelölni, azonban reméljük, hogy ez a jelölés sem vezet félreértéshez. Ekkor az $U_{\mathcal{P}}$ altér éppen azokból a $v : S_n \to \mathbb{R}$ függvényekből áll, melyek $\sigma(\mathcal{P})$-mérhetők, azaz $U_{\mathcal{P}} = L_2(\sigma(\mathcal{P}))$.

A (26)-beli második állítás a $\sigma(\mathcal{D}' \times \mathcal{R}) \cap \sigma(\mathcal{D} \times \mathcal{R}') = \sigma(\mathcal{D} \times \mathcal{R})$ könnyen ellenőrizhető összefüggés miatt teljesül. Az első állításhoz pedig azt kell belátni, hogy $|\Pi(\mathcal{D}' \times \mathcal{R})|$ és $|\Pi(\mathcal{D} \times \mathcal{R}')|$ feltételesen függetlenek a $|\Pi(\mathcal{D} \times \times \mathcal{R})|$ rögzítése mellett, ha Π egyenletes eloszlású véletlen permutáció. Ez ismét könnyen ellenőrizhető. ■

Még két lineáris algebrai lemmára lesz szükségünk. A merőleges felbontást jelölje $U = U_1 \oplus U_2$, ennek pontos jelentése tehát az, hogy $U = \mathrm{Span}(U_1, U_2)$, ahol U_1 és U_2 merőleges alterek.

9.5. Lemma. *Legyen $U = Span(U_i : i \in I)$ és $V = Span(V_j : j \in J)$ két altér. Tegyük még fel, hogy $U_i \perp_\cap V_j$ minden i, j párra. Ekkor $U \perp_\cap V$, és $U \cap V = Span(U_i \cap V_j : i \in I, j \in J)$.*

Bizonyítás. Feltevésünk szerint $Pr_{V_j} U_i \subseteq U_i$ minden i, j-re, ezért $Pr_{V_j} U \subseteq$ $\subseteq U$ minden j-re, tehát U merőlegesen metsz minden V_j alteret. Emiatt $Pr_U V_j \subseteq V_j$ minden j-re, amiből már következik a bizonyítandó $Pr_U V \subseteq$ $\subseteq V$ tartalmazás. Másrészt, legyen $W = \mathrm{Span}(U_i \cap V_j : i \in I, j \in J)$.

Erre $Pr_{V_j}U_i \subseteq W$, továbbá $Pr_U V_j = Pr_{V_j}U \subseteq W$, amiből $Pr_U V \subseteq W$ következik. ■

9.6. Lemma. *Legyen $U = U_1 \oplus U_2$ és $V = V_1 \oplus V_2$ két altér. Ha $U \perp_\cap V$, $U_1 \perp_\cap V_1$, $U \perp_\cap V_1$ és $U_1 \perp_\cap V$ teljesülnek, akkor $U_2 \perp_\cap V_2$ is igaz, és*

$$U \cap V = (U_1 \cap V_1) \oplus (U_1 \cap V_2) \oplus (U_2 \cap V_1) \oplus (U_2 \cap V_2).$$

Bizonyítás. Az első állításhoz azt használjuk fel, hogy $U \perp_\cap V$ akkor és csak akkor, ha a vetítés operátorok kommutálnak, azaz $Pr_U Pr_V = Pr_V Pr_U$. Mármost

$$Pr_{U_2}Pr_{V_2} = (Pr_U - Pr_{U_1})(Pr_V - Pr_{V_1}) = \\ Pr_U Pr_V - Pr_{U_1}Pr_V - Pr_U Pr_{V_1} + Pr_{U_1}Pr_{V_1},$$

ahol a feltétel szerint mind a négy tag operátorai felcserélhetők. A felcseréléseket elvégezve éppen a $Pr_{V_2}Pr_{U_2}$ operátort kapjuk. A második állítás pedig a 9.5. Lemmából következik. ■

A 9.4. Lemma és a 9.5. Lemma azonnali következménye az alábbi.

9.7. Következmény. *Legyen $\mathcal{L}(\mathcal{D}_i \times \mathcal{R} : i = 1, \ldots, s)$ és $\mathcal{L}(\mathcal{D} \times \mathcal{R}_j : j = = 1, \ldots, t)$ két hierarchikus modell, ahol $\mathcal{D} \succeq \mathcal{D}_i$ és $\mathcal{R} \succeq \mathcal{R}_j$ minden $1 \leq \leq i \leq s$, $1 \leq j \leq t$ esetén. Ekkor a két modell metszete az $\mathcal{L}(\mathcal{D}_i \times \mathcal{R}_j : i = = 1, \ldots, s, j = 1, \ldots, t)$ hierarchikus modell.*

9.2. A család paraméterezése

Ebben a szakaszban megmutatjuk, hogy a szigorúan pozitív L-felbontható és invertálva L-felbontható eloszlások egy-egy hierarchikus modellt alkotnak. Ez a két család ráadásul olyan szerencsés, hogy alkalmazható rájuk a 9.7. Következmény is, így könnyen megkapjuk a szigorúan pozitív duplán L-felbontható eloszlások hierarchikus reprezentációját is. Azonban a modell szabad paraméterei számának meghatározásához ennél kicsit több munkára lesz szükség.

Definiáljuk először a *vastag vágásokat*, ezek az $[n]$ halmazt három intervallumra partícionálják, melyek közül a középső csak egy pontból áll. A k. vastag vágás tehát

$$\Phi_k = (\{1..k-1\}, \{k\}, \{k+1..n\}), \quad 2 \leq k \leq n-1. \tag{27}$$

Kényelmes lesz a Φ_k jelölést $k = 1$-re és $k = n$-re is használni, bár ezek a vastag vágások csak két részre bontják $[n]$-et. Azt a partíciót, mely $[n]$-et elemeire hasítja, *teljes partíciónak* nevezzük, jelölésben

$$\Psi = (\{1\}, \{2\}, \ldots, \{n\}).$$

A (17) egyenletből világos, hogy az $\mathbf{E}(M_L)$ L-felbontható exponenciális család egy hierarchikus modell:

$$\mathbf{E}(M_L) = \mathcal{L}(\Phi_k \times \Psi : 1 \leq k \leq n). \tag{28}$$

Ez amiatt van, hogy a $|\pi(\Phi_k \times \Psi)|$ $0-1$-mátrix ekvivalens a $(\pi\{1..k-1\}, \pi(k))$ párral.

A permutációk invertálása azt jelenti, hogy a szorzatpartíciókban a szorzást fordított sorrendben kell elvégezni. Ebből kapjuk, hogy az invertálva L-felbontható exponenciális családra

$$\mathbf{E}(M_{L'}) = \mathcal{L}(\Psi \times \Phi_k : 1 \leq k \leq n).$$

Legyen a (k, ℓ). *vastag kereszt*

$$\Sigma_{k\ell} = \Phi_k \times \Phi_\ell,$$

ez tehát a (27)-ban definiált vastag vágások szorzata. Vezessük be a szigorúan pozitív duplán L-felbontható eloszlások családjára az $\mathbf{E}(M_B)$ jelölést, hiszen tudjuk, hogy ezek exponenciális családot alkotnak, bár az M_B mátrixot még nem ismerjük. A 9.7. Következmény szerint

$$\mathbf{E}(M_B) = \mathcal{L}(\Sigma_{k\ell} : 1 \leq k, \ell \leq n). \tag{29}$$

A (29) reprezentáció által máris ölünkbe hull egy M_B mátrix, ez azonban nem lesz teljes rangú. A továbbiakban azon dolgozunk, hogy ennek a mátrixnak a rangját, azaz az $\mathbf{E}(M_B)$ család szabad paramétereinek számát meghatározzuk.

A (28)-beli hierarchikus modell egy részmodelljére is szükség lesz a számoláshoz. Ezt durvább partíciók generálják. *Vékony vágásnak* nevezünk egy olyan partíciót, mely két intervallumra bontja $[n]$-et. A k. vékony vágás tehát

$$\widetilde{\Phi}_k = (\{1..k\}, \{k+1..n\}), \quad 1 \leq k \leq n-1. \tag{30}$$

Vegyük észre, hogy $\Phi_1 = \widetilde{\Phi}_1$ és $\Phi_n = \widetilde{\Phi}_{n-1}$. Az egyszerűbb jelölés kedvéért a $\widetilde{\Phi}_n$ triviális partíciót is vezessük be. Az $\mathbf{E}(M_L)$ egy részmodellje az

$$\mathbf{E}(M_S) = \mathcal{L}(\widetilde{\Phi}_k \times \Psi : 1 \leq k \leq n-1)$$

hierarchikus modell. Az ilyen eloszlásokat *S-felbonthatónak* nevezzük, ahol S az angol „set” (halmaz) szóból jön, mivel ebben a modellben minden $C \subseteq [n]$ részhalmazhoz tartozik egy paraméter. Ez a modell önmagában is érdekes lehet, bár nekünk csak mint segédmodell bukkan fel. Később majd részletesebben megvizsgáljuk ezt a modellt. Hasonlóan definiáljuk az invertálva S-felbontható eloszlások $\mathbf{E}(M_{S'})$ családját. Felidézve a szorzatpartíciókhoz rendelt alterek (25)-beli definícióját, vezessünk be néhány újabb jelölést a leggyakrabban előforduló altereinkre. Legyen

$$U_{\Phi_k \times \Psi} = U_k, \; U_{\widetilde{\Phi}_k \times \Psi} = \widetilde{U}_k.$$

A partíciók között fennálló reláció miatt $\widetilde{U}_k \subseteq U_k$, jelölje $\widetilde{U}_k$ ortogonális kiegészítő alterét U_k-ban F_k. Hasonlóan, $\widetilde{U}_k \subseteq U_{k+1}$ is teljesül. Ezért

$$\mathrm{Span}(U_k : 1 \leq k \leq n) = \mathrm{Span}(F_k : 1 \leq k \leq n, \widetilde{U}_n).$$

Mivel $\Phi_1 = \widetilde{\Phi}_1$, így $F_1 = \{\mathbf{0}\}$. Továbbá, mivel $\widetilde{\Phi}_n$ a triviális partíció, $\widetilde{U}_n = \mathrm{Span}(\mathbf{1})$. Ezért az $\mathbf{E}(M_L)$ L-felbontható hierarchikus modellhez tartozó

altér

$$F = \mathrm{Span}(U_k : 1 \le k \le n) = \mathrm{Span}(F_k : 2 \le k < n, \mathbf{1}). \tag{31}$$

Megmutatjuk, hogy a (31) jobb oldalán álló alterek az F altér ortogonális felbontását adják.

9.8. Lemma. *Az F_k $(2 \le k \le n)$ alterek merőlegesek egymásra és az $\mathbf{1}$ vektorra.*

Bizonyítás. Ismét használjuk, hogy az egyenletes eloszlás melletti L_2-beli merőlegesség ekvivalens az $\mathbb{R}^{n!}$-beli euklideszi merőlegességgel. Ebben a bizonyításban a

$$\sigma_k = \sigma(\Phi_k \times \Psi), \quad \widetilde{\sigma}_k = \sigma(\widetilde{\Phi}_k \times \Psi)$$

jelölést használjuk. Az F_k elemei $f_1 = f - \mathbb{E}(f \mid \widetilde{\sigma}_k)$ alakban állnak elő, ahol f σ_k-mérhető. Az $\mathbf{1}$-re való merőlegesség az $\mathbb{E}(f_1) = 0$ összefüggésből adódik. A másik állításhoz pedig legyen $g_1 \in F_j$, ahol $j > k$. Könnyű ellenőrizni, hogy egyenletes eloszlás esetén σ_k és σ_j feltételesen függetlenek a $\widetilde{\sigma}_k$ σ-algebrára nézve. Ezért

$$\mathbb{E}(f_1 g_1) = \mathbb{E}[\mathbb{E}(f_1 g_1 \mid \widetilde{\sigma}_k)] = \mathbb{E}[\mathbb{E}(f_1 \mid \widetilde{\sigma}_k)\mathbb{E}(g_1 \mid \widetilde{\sigma}_k)] = 0,$$

hiszen 1 valószínűséggel $\mathbb{E}(f_1 \mid \widetilde{\sigma}_k) = 0$. ■

Az $\mathbf{E}(M_L)$ család szabad paramétereinek b_n száma könnyen számolható, többek között a (31) ortogonális felbontásból kapjuk, hogy

$$b_n = \dim(F) - 1 = \sum_{k=2}^{n} \dim(F_k) = \sum_{k=2}^{n} \binom{n}{k}(k-1) = 2^n(n/2 - 1) + 1,$$

ahol F_k dimenziójának kiszámítása könnyű feladat (b_n-et egyébként már korábban is meghatároztuk). Azonban a (31) felbontás igazi haszna abban áll, hogy segít kiszámolni a $G = F \cap F'$ altér dimenzióját.

Tartozzanak az $U'_\ell, \widetilde{U}'_\ell, F'_\ell$ alterek az $\mathbf{E}(M_{L'})$ modellhez ugyanúgy, ahogy $U_k, \widetilde{U}_k, F_k$ az $\mathbf{E}(M_L)$ modellhez tartozik. Tehát a (25)-beli jelöléssel,

$$U_{\Psi\times\Phi_\ell} = U'_\ell,\ U_{\Psi\times\widetilde{\Phi}_\ell} = \widetilde{U}'_\ell.$$

Ekkor a 9.8. Lemma szerint az invertálva L-felbontható hierarchikus modellhez tartozó altér:

$$F' = \oplus_{\ell=2}^{n} F'_\ell \oplus \mathrm{Span}(\mathbf{1}).$$

A 9.4. Lemma szerint minden

$$U \in \{U_k, \widetilde{U}_k : \ 2 \le k \le n\}, \ U' \in \{U'_\ell, \widetilde{U}'_\ell : \ 2 \le \ell \le n\}$$

altér-párra $U \perp_\cap U'$, hiszen az U-alterekhez tartozó szorzatpartíciók második tényezője, míg az U'-alterekhez tartozó szorzatpartíciók első tényezője az $[n]$ teljes partíciója. A 9.6. Lemmából kapjuk, hogy $F_k \perp_\cap F'_\ell$ minden k, ℓ esetén. Továbbá a 9.5. Lemma alapján a $G = F \cap F'$ altér egy ortogonális felbontása

$$G = \oplus_{2 \le k, \ell \le n} (F_k \cap F'_\ell) \oplus \mathbf{1}.$$

Most már csak az $F_k \cap F'_\ell$ alterek dimenzióját és bázisát kell meghatároznunk. Ehhez rögzítsük k-t és ℓ-et. A 9.4. Lemma alapján a $\Sigma_{k\ell}$ vastag kereszthez tartozó altér $U_k \cap U'_\ell$. Vegyük észre, hogy $F_k \cap F'_\ell$ pontosan azokból az $U_k \cap \cap U'_\ell$-beli vektorokból áll, melyek merőlegesek az $\widetilde{U}_k$ és az $\widetilde{U}'_\ell$ terekre.

Emlékezzünk vissza, hogy az $U_k \cap U'_\ell$ tér vektorainak π. koordinátája csak a (24)-ben definiált $|\pi(\Sigma_{k\ell})|$ statisztikától függ. A $|\pi(\Sigma_{k\ell})| = (t_{ij})$ 3×3-as mátrix elemeinek sorösszegei rendre $k-1$, 1, $n-k$, oszlopösszegei rendre $\ell - 1$, 1, $n - \ell$ kell legyenek. Ezért elég a mátrixnak a bal felső 2×2-es részét megadni. Mivel a t_{12}, t_{21}, t_{22} elemek csak 0 vagy 1 értékűek lehetnek, a $|\pi(\Sigma_{k\ell})|$ statisztikát még tömörebben leírhatjuk egy $(a^{k\ell}(\pi), q^{k\ell}(\pi))$ párral. Legyen tehát

$$a^{k\ell}(\pi) = t_{11} + t_{12} + t_{21} + t_{22},$$

és $q^{k\ell}(\pi)$ a 6. táblázat szerinti.

Mivel az $U_k \cap U'_\ell$ altér vektorainak π. koordinátája csak a $|\pi(\Sigma_{k\ell})|$ statisztikától függ, az $U_k \cap U'_\ell$ altér egy ortogonális bázisa a

$$\rho^{k\ell}_{aq}(\pi) = \chi\{a^{k\ell}(\pi) = a, \ q^{k\ell}(\pi) = q\} \tag{32}$$

indikátorvektorokból áll, ahol a, q minden olyan lehetséges értéket befut,

6. táblázat. A $|\pi(\Sigma_{k\ell})|$ statisztika lehetséges értékeinek kódolása

$(t_{ij})_{1\le i,j\le 2}$	$q^{k\ell}(\pi)$	$(t_{ij})_{1\le i,j\le 2}$	$q^{k\ell}(\pi)$
$\begin{pmatrix} a-1 & 1 \\ 0 & 0 \end{pmatrix}$	1	$\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$	4
$\begin{pmatrix} a-2 & 1 \\ 1 & 0 \end{pmatrix}$	2	$\begin{pmatrix} a-1 & 0 \\ 0 & 1 \end{pmatrix}$	5
$\begin{pmatrix} a-1 & 0 \\ 1 & 0 \end{pmatrix}$	3		

melyre $\rho_{aq}^{k\ell}$ nem azonosan nulla. Ehhez például szükséges, hogy

$$\max(0, k+\ell-n) \le a \le \min(k,\ell),$$

míg q általában 1-től 5-ig bármi lehet, kivéve ha $a \in \{0,1,k,\ell\}$.

Legyen most $u = \sum_{b,q} c_{bq}\rho_{bq}^{k\ell}$ az $U_k \cap U'_\ell$ altér általános eleme, meg kell határoznunk, hogy mikor lesz merőleges az $\widetilde{U}_k$ és $\widetilde{U}'_\ell$ alterekre. Legyen először $v(\pi) = \chi(\pi\{1..k\} = C)$ az $\widetilde{U}_k$ tér egy báziseleme, és legyen $|\, C \cap \{1..\ell\} \,| = a$. Ha még $h = (k-1)!\,(n-k)!$, akkor a skaláris szorzatra

$$(u,v) = \begin{cases} c_{a1}(k-a)h + c_{a2}(a-1)h + c_{a5}h & \text{ha } \ell \in C, \\ c_{a3}ah + c_{a4}(k-a)h & \text{ha } \ell \notin C. \end{cases}$$

Hasonlóan, ha $v(\pi) = \chi(\pi^{-1}\{1..\ell\} = D)$ az $\widetilde{U}'_\ell$ altér báziseleme, valamint $|\, D \cap \{1..k\} \,| = a$, és $g = (\ell-1)!\,(n-\ell)!$, akkor

$$(u,v) = \begin{cases} c_{a3}(\ell-a)g + c_{a2}(a-1)g + c_{a5}g & \text{ha } k \in D, \\ c_{a1}ag + c_{a4}(\ell-a)g & \text{ha } k \notin D. \end{cases}$$

Az $F_k \cap F'_\ell$ altér tehát olyan $\sum_{q=1}^{5} c_{aq}\rho_{aq}^{k\ell}$ vektorok lineáris kombinációiból

áll, melyekre

$$c_{a1}(k-a)+c_{a2}(a-1)+c_{a5}=c_{a3}a+c_{a4}(k-a)=$$
$$c_{a3}(\ell-a)+c_{a2}(a-1)+c_{a5}=c_{a1}a+c_{a4}(\ell-a)=0.$$

A négy feltételből három lineárisan független választható ki, így általában két lineárisan független megoldás van. Az $a=0{,}1,\min(k,\ell)$ eseteket külön kell végignézni. Az $a=0$ esetben csak az azonosan nulla megoldás, míg az $a=$ $=1,\min(k,\ell)$ esetekben egy nemnulla megoldás van. Jelölje $N_a^{k\ell}$ a lineárisan független megoldások számát, azaz $N_a^{k\ell}$ nulla, egy vagy kettő. Megmutatjuk, hogy minden $1\le i\le n$-re

$$|\{(k,\ell):\dim(F_k\cap F_\ell^{-1})\ge i\}|=(n-i)^2.$$

Keressük meg először azon k,ℓ párokat, melyekre $\dim(F_k\cap F'_\ell)\ge 2j+2$. Ez úgy lehet, ha az $N_a^{k\ell}$ mennyiségek között vagy két 1-es és legalább j 2-es van, vagy egy 1-es és legalább $(j+1)$ 2-es.

Az első lehetőség akkor következik be, ha $\ell+k\le n+1$ és $\min\{k,\ell\}\ge j+$ $+2$, a második pedig akkor, ha $\ell+k\ge n+2$ és $\max\{k,\ell\}\le n-j-1$. De ha $\ell+k\le n+1$ és $k,\ell\ge j+2$, akkor $k,\ell\le n-j-1$ is igaz. Hasonlóképp, ha $\ell+k\ge n+2$ és $k,\ell\le n-j-1$, akkor egyszersmind $k,\ell\ge j+3>j+2$. Tehát $\dim(F_k\cap F_\ell^{-1})\ge 2j+2$ akkor és csak akkor, ha $j+2\le k,\ell\le n-j-1$, ilyen párból pedig $[n-(2j+2)]^2$ van.

Most keressük meg azon k,ℓ párokat, melyekre $\dim(F_k\cap F'_\ell)\ge 2j+1$. Ez úgy történhet, ha az $N_a^{k\ell}$ értékek között vagy két 1-es és legalább j 2-es van, vagy egy 1-es és legalább j 2-es.

Az első lehetőség akkor következik be, ha $\ell+k\le n+1$ és $\min\{k,\ell\}\ge$ $\ge j+2$, a második pedig akkor, ha $\ell+k\ge n+2$ és $\max\{k,\ell\}\le n-j$. De ha $\ell+k\le n+1$ és $k,\ell\ge j+2$, akkor $k,\ell\le n-j-1<n-j$ is teljesül. Ha pedig $\ell+k\ge n+2$ és $k,\ell\le n-j$, akkor $k,\ell\ge j+2$ is fennáll. Tehát $\dim(F_k\cap F'_\ell)\ge 2j+1$ akkor és csak akkor, ha $j+2\le k,\ell\le n-j$, ilyen párból pedig $[n-(2j+1)]^2$ van.

Mivel

$$\sum_{k,\ell} \dim(F_k \cap F_\ell^{-1}) = \sum_{i \geq 1} |\{(k,\ell) : \dim(F_k \cap F_\ell^{-1}) \geq i\}|,$$

és az $\mathbf{E}(M_B)$ család szabad paramétereinek száma $G = F \cap F'$ dimenziója mínusz egy, kaptuk, hogy

9.9. Tétel. *Az $\mathbf{E}(M_B)$ család szabad paramétereinek száma*

$$d_n = \sum_{i=1}^{n-1} i^2.$$

Továbbá beláttuk a következő tételt.

9.10. Tétel. *Legyen*

$$\begin{aligned} \mu_{a1}^{k\ell} = \;& -\rho_{a2}^{k\ell} + (a-1)\rho_{a5}^{k\ell} \\ \mu_{a2}^{k\ell} = \;& -(\ell-a)a\rho_{a1}^{k\ell} + (k-a)(\ell-a)\rho_{a2}^{k\ell} - (k-a)a\rho_{a3}^{k\ell} + \\ & + a^2\rho_{a4}^{k\ell} + (k-a)(\ell-a)\rho_{a5}^{k\ell}. \end{aligned}$$

Ekkor az

$$\{\mathbf{1}\} \cup \{\mu_{ai}^{k\ell} : 2 \leq k,\ell \leq n,\ \max(0, k+\ell-n) \leq a \leq \min(k,\ell),\ i = 1,2\}$$

vektorok közül az azonosan nullákat elhagyva, a G altér ortogonális bázisát kapjuk.

A merőlegesség különösen kényelmes tulajdonság, ha egy szigorúan pozitív duplán L-felbontható eloszlásnak a 9.10. Tételben megadott paraméterezését keressük. Ugyanígy könnyen meghatározhatjuk egy szigorúan pozitív eloszlás logaritmusának a G altértől vett euklideszi távolságát, illetve a hozzá euklideszi értelemben legközelebb eső duplán L-felbontható eloszlást. Ez természetesen különbözni fog a maximum likelihood becsléstől.

Most a G térnek egy indikátor-vektorokból álló bázisát is megadjuk. Tet-

szőleges k, ℓ, a-ra vezessük be a

$$\nu_a^{k\ell} = \sum_{q=1}^{5} \rho_{aq}^{k\ell}$$

jelölést, ahol emlékeztetünk $\rho_{aq}^{k\ell}$ (32)-beli definíciójára.

9.11. Tétel. *Az* $\mathbf{1}$ *és a*

$$\begin{aligned} \nu_a^{k\ell}: &\quad 1 \le k, \ell \le n-1, \max(0, k+\ell-n) < a \le \min(k, \ell), \\ \rho_{a5}^{k\ell}: &\quad 1 \le k, \ell \le n-1, \max(1, k+\ell-n) < a \le \min(k, \ell) \end{aligned} \tag{33}$$

vektorok a G *altér bázisát alkotják.*

Bizonyítás. Két dolgot mutatunk meg. Egyrészt triviális kiszámolni, hogy a tételbeli vektorok elemszáma $d_n + 1$. Másrészt megmutatjuk, hogy a $\rho_{aq}^{k\ell}$ $(2 \le k, \ell \le n)$ vektorok benne vannak a tételbeli vektorok által generált G_ν altérben. Rögzített k, ℓ-re a $\nu_a^{k\ell}$ és $\rho_{a5}^{k\ell}$ vektorrendszerek mindegyikéből egy vektort kizártunk, méghozzá a lehető legkisebb a értékhez tartozót (ha pl. $\min(k, \ell) = 1$, akkor a $\rho_{a5}^{k\ell}$ vektorok között csak az $a = 1$ fordul elő, amit ki is zárunk). Ezt azért tehetjük meg, mert ezek már benne vannak G_ν-ben, amint azt most megmutatjuk. Először is, mivel

$$\sum_{a=\max(0,k+\ell-n)}^{\min(k,\ell)} \nu_a^{k\ell} = \mathbf{1},$$

a $\nu_a^{k\ell}$ vektorok $a = \max(0, k+\ell-n)$-re is G_ν-ben vannak. A $\rho_{a5}^{k\ell}$ vektorok esetében pedig elég belátni, hogy a $\{\pi(k) = \ell\}$ esemény indikátorvektora, azaz a

$$v^{k\ell} = \sum_{a=\max(1,k+\ell-n)}^{\min(k,\ell)} \rho_{a5}^{k\ell}$$

vektor G_ν-beli. Ezt indukcióval mutatjuk meg. Az indukciós lépésben feltesszük, hogy $v^{ij} \in G_\nu$ minden $i \le k$, $j \le \ell$, $(i, j) \ne (k, \ell)$ esetén, majd ebből megmutatjuk, hogy $v^{k\ell} \in G_\nu$ is igaz. Így az (1,1) párból elindulva minden (k, ℓ) párhoz eljutunk. Az indukció kezdetéhez vegyük észre, hogy $v^{11} = \nu_1^{11} \in G_\nu$.

Az indukciós lépés pedig a

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell} v^{ij} = \sum_{a} a\nu_a^{k\ell} \in G_\nu$$

összefüggés miatt végezhető el.

Rögzítsük most a $2 \le k, \ell \le n-1$ értékeket, és mutassuk meg, hogy $\rho_{aq}^{k\ell} \in G_\nu$ minden $a \le \min(k,\ell)$, $1 \le q \le 4$ választásra. A bizonyítás lényegét a következő egyenletek adják:

$$\begin{aligned}
\rho_{a1}^{k\ell} &= \nu_a^{k-1,\ell} - \nu_a^{k-1,\ell-1} + \epsilon_1 \\
\rho_{a2}^{k\ell} &= \nu_a^{k\ell} + \nu_a^{k-1,\ell-1} - \nu_a^{k-1,\ell} - \nu_a^{k,\ell-1} - \rho_{a5}^{k\ell} + \epsilon_2 \\
\rho_{a3}^{k\ell} &= \nu_a^{k,\ell-1} - \nu_a^{k-1,\ell-1} + \epsilon_3 \\
\rho_{a4}^{k\ell} &= \nu_a^{k-1,\ell-1} + \epsilon_4
\end{aligned} \tag{34}$$

ahol

$$\epsilon_i = d_i \rho_{a+2,2}^{k\ell} + \sum_{q=1}^{5} c_{iq} \rho_{a+1,q}^{k\ell} \tag{35}$$

valamilyen d_i, c_{iq} együtthatókkal, melyeknek pontos értéke nem fontos. Ezért a szerinti visszafelé haladó indukcióval bizonyíthatunk: ha már tudjuk, hogy $\rho_{bq}^{k\ell} \in G_\nu$ minden $b \ge a+1$-re, akkor a (35) egyenletben $\epsilon_i \in G_\nu$, és így $\rho_{aq}^{k\ell} \in G_\nu$ is, mivel a (34) egyenletek jobb oldalainak többi vektora a tételbeli (33) bázis eleme. (Az indukció elindításával sincs gond, mivel $a = \min(k,\ell)$ esetén $\epsilon_i = 0$.)

Végül nézzük azt az esetet, amikor $\max(k,\ell) = n$. Tegyük fel, hogy $\ell < k = n$, ekkor $a = \ell$ lehet csak, és

$$\rho_{\ell 1}^{n\ell} = \nu_\ell^{n-1,\ell}, \quad \rho_{\ell 2}^{n\ell} = \nu_{\ell-2}^{n-1,\ell-1}, \quad \rho_{\ell 5}^{n\ell} = \mathbf{1} - \nu_\ell^{n-1,\ell} - \nu_{\ell-2}^{n-1,\ell-1}.$$

Ha pedig $k = \ell = n$, akkor $a = n$, és

$$\rho_{n2}^{nn} = \mathbf{1} - \nu_{n-1}^{n-1,n-1}, \quad \rho_{n5}^{nn} = \nu_{n-1}^{n-1,n-1}.$$

■

Nevezzük azokat az eloszlásokat, melyek S-felbonthatók és invertálva S-felbonthatók is, duplán S-felbontható eloszlásoknak. A 9.7. Következmény szerint a pozitív duplán S-felbontható eloszlások hierarchikus modellt alkotnak, melynek generátorai a

$$\widetilde{\Sigma}_{k\ell} = \widetilde{\Phi}_k \times \widetilde{\Phi}_\ell$$

szorzatpartíciók, melyeket *vékony kereszteknek* hívunk. Emlékeztetünk a szorzótényezők (30)-beli definíciójára. Ebben az esetben fix k, ℓ-re a $\widetilde{\Sigma}_{k\ell}$ partícióhoz tartozó $U_{\widetilde{\Sigma}_{k\ell}}$ altérnek a $\nu_a^{k\ell}$ vektorok (ahol a minden lehetséges értéket befut) ortogonális bázisát adják. A 9.11. Tétel szerint a különböző (k, ℓ) párokhoz tartozó $U_{\widetilde{\Sigma}_{k\ell}}$ alterek „majdnem lineárisan függetlenek", a közöttük lévő lineáris összefüggés csak abból származik, hogy az $\mathbf{1}$ vektor mindegyikükben benne van. Ebből a következő tételt kapjuk.

9.12. Tétel. *A pozitív duplán S-felbontható eloszlások hierarchikus modellt alkotnak, mely szabad paramétereinek száma*

$$e_n = \sum_{j=0}^{\lfloor (n-1)/2 \rfloor} (n-2j-1)^2.$$

A modellhez tartozó altér egy bázisa az $\mathbf{1}$ *vektorból és a* $\nu_a^{k\ell}$ *vektorokból áll, ahol* k, ℓ, a *ugyanott fut, mint* (33)*-ban.*

A (33)-beli $\rho_{a5}^{k\ell}$ vektorok tehát a duplán L-felbontható és a duplán S-felbontható eloszlások közötti „különbséget" adják meg. A teljesség kedvéért megjegyezzük, hogy a pozitív S-felbontható eloszláscsalád szabad paramétereinek száma

$$c_n = \sum_{i=1}^{n} \left[\binom{n}{i} - 1 \right] = 2^n - n - 1.$$

Végül megemlítjük, hogy elmetszhetjük egymással az $\mathbf{L}_K$ és az $\mathbf{L}'_{K'}$ családokat is, azaz kereshetjük azokat az eloszlásokat, melyek felbonthatók minden $k \in K$-ra és invertálva felbonthatók minden $k' \in K'$-re, ahol $K, K' \subseteq$

$\subseteq [n]$ tetszőleges halmazok. Az ilyen pozitív eloszlások is hierarchikus modellt alkotnak (9.7. Következmény), melynek paraméterszáma minden konkrét n, K, K'-re lineáris algebrai módszerekkel meghatározható.

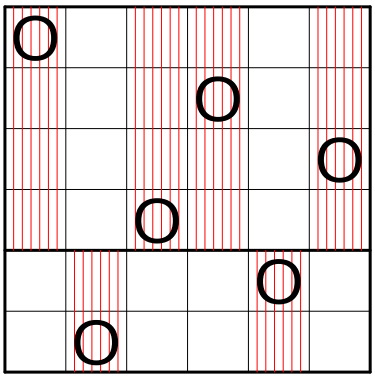
9.3. Két szemléltetési mód

Szólunk pár szót arról, hogy a véletlen permutációkat, a $|\pi(\mathcal{P})|$ statisztikákat, illetve a feltételes függetlenséget hogyan lehet szemléletesen ábrázolni. Két ábrázolásmódot mutatunk be: a páros gráfot és a sakktáblát.

Minden $\pi \in S_n$-hez hozzárendelhetjük a $G(\pi) = ([n], [n], E(\pi))$ páros gráfot. Itt $[n]$ jelöli mind a két csúcsosztályt, azaz mindkét osztályban n darab, 1-től n-ig számozott csúcs van. Az élek halmaza pedig $E = \{(k, \pi(k)) : k = = 1, \dots, n\}$, azaz az első osztály k. csúcsát a második osztály $\pi(k)$. csúcsával kötjük össze. Ebben a gráfban tehát minden pont foka pontosan egy. Ha most $\mathcal{D}$ és $\mathcal{R}$ az $[n]$ partíciói, akkor a $G(\pi)$ gráf első csúcsosztályának minden D_i részhalmazát vonjuk össze egy i^* csúccsá, és második csúcsosztályának minden R_j részhalmazát vonjuk össze egy j^* csúccsá úgy, hogy az összevont csúcsok megöröklik a régi csúcsok éleit. A $|\pi(\mathcal{D} \times \mathcal{R})| = (t_{ij})$ mátrix t_{ij} eleme ekkor azt adja meg, hogy a kapott multigráfban hány él vezet az első osztály i^* és a második osztály j^* csúcsa között.

Tegyük fel, hogy a két csúcsosztályt két oszlopban rendezzük el, balra az első osztályt, jobbra a második osztályt helyezzük, és a csúcsok számozása lefelé növekszik. Az élek pedig legyenek egyenes szakaszok. Ekkor például a $|\pi(\Sigma_{k\ell})| = (t_{ij})$ statisztikát „lokálisan" leolvashatjuk a gráfból. Azt kell meghatároznunk, hogy (i) a $(k, \pi(k))$ él a (k, ℓ) szakasz alatt, fölött, vagy rajta fut, (ii) a $(\pi^{-1}(\ell), \ell)$ él a (k, ℓ) szakasz alatt, fölött, vagy rajta fut, és (iii) hány E-beli él metszi a (k, ℓ) szakaszt. (iii)-hoz azt vegyük észre, hogy a (k, ℓ) szakaszt metsző E-beli élek száma $t_{13}+t_{31}$, ebből pedig t_{11} kiszámítható, ha ismerjük még a $t_{12}, t_{21}, t_{22}, t_{23}, t_{32}$ mátrixelemeket.

A sakktábla reprezentációhoz vegyünk egy $n \times n$-es sakktáblát. Minden $\pi \in S_n$-hez helyezzünk el n darab bástyát a táblán, a k. bástya kerüljön a k. sor $\pi(k)$. oszlopába. A sakk csak úgy jön elő, hogy ezek a bástyák nem ütik egymást. Ha most $\mathcal{D}$ és $\mathcal{R}$ az $[n]$ partíciói, akkor a tábla sorainak minden D_i



3. ábra. L-felbonthatóság a sakktáblán, $n = 6$. Feltéve, hogy $\Pi\{1..4\} = \{1,3,4,6\}$, $\Pi(1..4)$ és $\Pi(5..6)$ független.

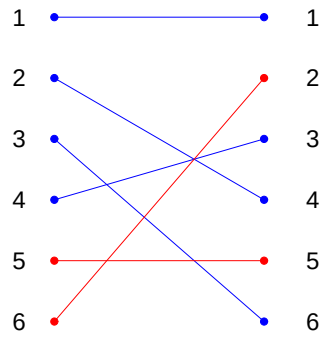
részhalmazát vonjuk össze egy i^* sorrá, és oszlopainak minden R_j részhalmazát vonjuk össze egy j^* oszloppá úgy, hogy az összevonás utáni (i^*, j^*) tábla-mező megörökli a régi mezők bástyáit. A $|\pi(\mathcal{D} \times \mathcal{R})| = (t_{ij})$ mátrix t_{ij} eleme ekkor azt adja meg, hogy az új (i^*, j^*) mezőben hány bástya áll.

A sakktábla reprezentáció segítségével szemléltethetjük az L-felbonthatóságot. Ha a sakktáblát vízszintesen elvágjuk két sora között, akkor a felső és alsó részben a bástyák elhelyezkedése független lesz, feltéve, hogy tudjuk, mely oszlopokat foglalják el a felső és alsó rész bástyái. (lásd a 3. ábrát). Ez akkor is igaz, ha a táblát több vízszintes részre vágjuk. Az invertálva L-felbonthatóságot úgy kapjuk, ha a sorok és oszlopok szerepét felcseréljük.

Ugyanígy a páros gráfon is szemléltethetjük az L-felbonthatóságot. Ha az első csúcsosztályt valahol elvágjuk, akkor a vágás feletti és alatti csúcsokból kiinduló élek függetlenek lesznek, feltéve, hogy ismerjük a vágás feletti csúcsok szomszédainak halmazát (lásd a 4. ábrát). Az invertálva L-felbonthatóságot úgy kapjuk, ha a két csúcsosztály szerepét felcseréljük.

9.4. Markov bázis $n = 4$-re

Az L-felbontható torikus modellre könnyű volt egy Markov bázist megadni, általános n-re. Adódik a kérdés, hogy a duplán L-felbontható esetben is megoldható-e ez a feladat. Szeretnénk tehát megkeresni, hogy melyek azok



4. ábra. L-felbonthatóság a páros gráfon, $n = 6$. Feltéve, hogy $\Pi\{1..4\} = \{1,3,4,6\}$, $\Pi(1..4)$ és $\Pi(5..6)$ független.

a polinomiális összefüggések, melyek a $\mathrm{cl}(\mathbf{F}(M_B))$ zárt torikus modell minden elemére teljesülnek. Ezt általános n-re nem sikerült tisztázni. Sőt, az egyetlen nemtriviális n, melyre a nyílt hozzáférésű programcsomagok eredményesek voltak, az $n = 4$. Az $n = 5$ már túl nagy feladatnak bizonyult ezen programcsomagok számára. Mi a 4ti2 [1] programcsomagot használtuk, mely egy adott modellmátrixhoz tartozó torikus ideál minimális generátorrendszerét számolja ki algoritmikusan. Korábban láttuk, hogy az $n = 4$ esetben az I_{M_L} minimális generátorrendszere hat másodfokú polinomból áll, ugyanígy az $I_{M_{L'}}$-é is. Két polinom azonban közös, így összesen tíz másodfokú polinomot kapunk, melyek nyilván az I_{M_B} ideálnak is elemei. Ezeken túl az I_{M_B} minimális generátorrendszerében még nyolc negyedfokú polinom szerepel:

$$
\begin{array}{ll}
1 & x_{1324}x_{2431}x_{3241}x_{4123} - x_{1423}x_{2341}x_{3124}x_{4231} \\
2 & x_{1324}x_{2431}x_{3214}x_{4132} - x_{1432}x_{2314}x_{3124}x_{4231} \\
3 & x_{1324}x_{1432}x_{3241}x_{4213} - x_{1342}x_{1423}x_{3214}x_{4231} \\
4 & x_{1324}x_{2341}x_{4213}x_{4132} - x_{1342}x_{2314}x_{4231}x_{4123} \\
5 & x_{1342}x_{2413}x_{3241}x_{4123} - x_{1423}x_{2341}x_{3142}x_{4213} \\
6 & x_{1342}x_{2413}x_{3214}x_{4132} - x_{1432}x_{2314}x_{3142}x_{4213} \\
7 & x_{1432}x_{2413}x_{3241}x_{3124} - x_{1423}x_{2431}x_{3214}x_{3142} \\
8 & x_{2314}x_{2431}x_{3142}x_{4123} - x_{2341}x_{2413}x_{3124}x_{4132}
\end{array}
\tag{36}
$$

Ez a nyolc polinom egyetlen orbitot alkot a következő értelemben. A következő szakaszban belátjuk, hogy az $\mathbf{E}(M_B)$ család, és így a lezártja is, invariáns bizonyos transzformációkra nézve. Ezek a transzformációk valamilyen S_n-en adott transzformáció szerint átrendezik a $p(\pi)$ valószínűségeket. Konkrétan, legyen $\phi : S_n \to S_n$ kölcsönösen egyértelmű leképezés, p pedig S_n-en adott eloszlás. p-nek ϕ szerinti átrendezése a $p_\phi(\pi) = p(\phi(\pi))$-vel definiált p_ϕ eloszlás. Akkor mondjuk, hogy egy eloszláscsalád invariáns ϕ-re, ha minden p elemével együtt p_ϕ is eleme a családnak. Ha ϕ az invertálás, vagy bizonyos permutációkkal való jobbról vagy balról szorzás, akkor $\mathbf{E}(M_B)$ invariáns ϕ-re (lásd a következő szakaszt). A ϕ transzformáció természetes módon hat az x_π változókból felépített polinomokra is. Ha egy polinomiális összefüggés igaz egy családban, és a család invariáns ϕ-re, akkor a ϕ szerint transzformált polinomiális összefüggés is érvényes a családban. Egy polinomiális összefüggés orbitját az összes, a családot invariánsan hagyó ϕ-k szerint transzformált polinomok képezik. A (36) nyolc polinomja ebben az értelemben teljes orbitot alkot.

Mivel $\mathbf{E}(M_L) \cap \mathbf{E}(M_{L'}) = \mathbf{E}(M_B)$, így szigorúan pozitív x_π változók esetén a (36)-beli polinomok levezethetők az I_{M_L} és $I_{M_{L'}}$ bázisának tíz másodfokú összefüggéséből, a következőképpen. Az $x_{1324}x_{2413} = x_{2314}x_{1423}$ és az $x_{3214}x_{4123} = x_{3124}x_{4213}$ polinomok $I_{M_{L'}}$-höz tartoznak. Ezeket összeszorozva kapjuk, hogy $x_{1324}x_{2413}x_{3214}x_{4123} = x_{3124}x_{4213}x_{2314}x_{1423}$, azaz

$$\frac{x_{1324}x_{2413}x_{3214}x_{4123}}{x_{3124}x_{4213}x_{2314}x_{1423}} = 1.$$

Azonban az L-felbonthatóság miatt

$$\frac{x_{2413}}{x_{4213}} = \frac{x_{2431}}{x_{4231}} \text{ és } \frac{x_{3214}}{x_{2314}} = \frac{x_{3241}}{x_{2341}}.$$

Ez utóbbit az előbbibe helyettesítve, és a nevezővel visszaszorozva éppen a (36) polinomok közül az elsőt kapjuk.

Összességében azt mondhatjuk, hogy a pozitív duplán L-felbontható

eloszlásokra a következő alakú összefüggések érvényesek:

$$\begin{aligned}
\left(\frac{x_{13ab}}{x_{31ab}}\right) / \left(\frac{x_{14cd}}{x_{41cd}}\right) &= \left(\frac{x_{23ef}}{x_{32ef}}\right) / \left(\frac{x_{24gh}}{x_{42gh}}\right) \\
\left(\frac{x_{ab13}}{x_{ab31}}\right) / \left(\frac{x_{cd14}}{x_{cd41}}\right) &= \left(\frac{x_{ef23}}{x_{ef32}}\right) / \left(\frac{x_{gh24}}{x_{gh42}}\right) \\
\left(\frac{x_{1a2b}}{x_{2a1b}}\right) / \left(\frac{x_{1cd2}}{x_{2cd1}}\right) &= \left(\frac{x_{e12f}}{x_{e21f}}\right) / \left(\frac{x_{g1h2}}{x_{g2h1}}\right) \\
\left(\frac{x_{3a4b}}{x_{4a3b}}\right) / \left(\frac{x_{3cd4}}{x_{4cd3}}\right) &= \left(\frac{x_{e34f}}{x_{e43f}}\right) / \left(\frac{x_{g3h4}}{x_{g4h3}}\right),
\end{aligned}$$

ahol az a, b, c, d, e, f, g, h tetszőleges értékek. Ezek az összefüggések ugyanúgy vezethetők le, mint a (36) első polinomja: valamilyen $a, \ldots, h$ értékekre az összefüggés két I_{M_L}-beli vagy $I_{M_{L'}}$-beli polinom összeszorzásával és átrendezésével kapható meg. Ebből pedig, ismét felhasználva az L-, illetve invertálva L-felbonthatóságot, az összefüggés tetszőleges $a, \ldots, h$ választásra is teljesül. Ha a fenti összefüggésekben eltűntetjük a nevezőket, megkapjuk azokat a polinomokat, melyeket minden $\mathrm{cl}(\mathbf{F}(M_B))$-beli eloszlás kielégít.

9.5. Maximum likelihood becslés

A permutációk hierarchikus modelljei esetében a maximum likelihood becslés ugyanúgy megkapható az IPS algoritmussal, mint a kontingenciatáblák esetében. A ML becslés pedig a 2.7. Tételben kimondott tulajdonságokkal rendelkezik.

Az IPS algoritmus során a generátor-partíciókat ciklikusan vesszük sorra, és az éppen soron levő $|\pi(\mathcal{P})|$ statisztika elméleti eloszlását átskálázással egyenlővé tesszük a megfelelő tapasztalati eloszlással. A duplán L-felbontható hierarchikus modell esetén egy alternatív algoritmus kínálkozik, mivel az $\mathbf{L}$ és $\mathbf{L}'$ családokban a ML becslés explicit. A következő algoritmusban, melyet nevezzünk iteratív vetítésnek, felváltva számoljuk az $\mathbf{L}$-beli és $\mathbf{L}'$-beli ML becsléseket. Ismét r jelöli a minta tapasztalati eloszlását.

Iteratív vetítés:

$p_0 := r$, $k := 0$.

Ismételjük:

$q :=$ a p_k eloszlás ML becslése $\mathbf{L}$-ben,

$k := k + 1$, $p_k := q$,

$q :=$ a p_k eloszlás ML becslése $\mathbf{L}'$-ben,

$k := k + 1$, $p_k := q$,

amíg az eljárás konvergál.

Az iteratív vetítési algoritmusról egyelőre nem tudtuk bizonyítani, hogy a $\mathrm{cl}(\mathbf{F}(M_B))$-beli ML becsléshez tart, ami biztosan nem is igaz teljes általánosságban. Amint azt később látni fogjuk, $\mathrm{cl}(\mathbf{E}(M_B)) \subsetneq \mathbf{L} \cap \mathbf{L}'$. Így ha az r tapasztalati eloszlás $(\mathbf{L} \cap \mathbf{L}') \setminus \mathrm{cl}(\mathbf{E}(M_B))$-beli, akkor az iteratív vetítés nem mozdul el r-ből, míg az IPS algoritmus a $\mathrm{cl}(\mathbf{E}(M_B))$-beli ML becsléshez konvergál. Azt sejtjük azonban, hogy szigorúan pozitív r-ekre az iteratív vetítés ugyanoda konvergál, mint az IPS algoritmus. Nézzünk erre egy numerikus példát! Legyen r a (38)-beli permutációkon egyenletes eloszlás ($n = 4$), erre $r \in (\mathbf{L} \cap \mathbf{L}') \setminus \mathrm{cl}(\mathbf{E}(M_B))$. A hozzá tartozó $\mathrm{cl}(\mathbf{E}(M_B))$-beli ML becslés (az IPS algoritmussal) a következő:

$$\begin{aligned} p(1324) = p(2431) = p(3241) = p(4321) &= 0.1123, \\ p(2341) = p(3124) = p(4231) &= 0.1734, \\ p(1423) &= 0.0306. \end{aligned}$$

Módosítsuk most kicsit r-et! A (38)-beli permutációkhoz rendeljünk egyforma valószínűséget, az összes többi permutációhoz pedig egy kis pozitív valószínűséget (például $\epsilon = 0.0001$). Ekkor az IPS algoritmus és az iteratív vetítés ugyanahhoz a határponthoz konvergál. Az iteratív vetítés azonban sokkal lassabb.

Gyakorlati szempontból is érdekes lehet a következő kérdés. Tegyük fel, hogy két hierarchikus modell metszetében szeretnénk a ML becslést megtalálni, ahol a ML becslés a két modellben külön-külön explicit kiszámolható, azonban a két modell nem teljesíti a 9.7. Következmény feltételeit. Ekkor a metszetükről nem tudjuk, hogy hierarchikus modell (esetleg nem is az), így

nem alkalmazhatjuk minden további nélkül az IPS algoritmust. Vajon az iteratív vetítés ekkor elvezet-e a ML becsléshez? Később még visszatérünk arra, hogy mit mondhatunk két hierarchikus modell metszetéről általában.

9.6. A modell lezártja

Ebben a szakaszban áttérünk a nem szigorúan pozitív eloszlások vizsgálatára. Emlékeztetünk, hogy az L-felbontható eloszlások $\mathbf{L}$ családja megegyezik az $\mathbf{F}(M_L)$ zárt torikus modellel. Már megvizsgáltuk a pozitív duplán L-felbontható eloszlásokat, és azt kaptuk, hogy ezek egy $\mathbf{E}(M_B)$ exponenciális családot alkotnak, ahol az M_B mátrix explicit megadható. Ebben a szakaszban három modellt szemlélünk meg közelebbről: (i) az exponenciális család $\mathrm{cl}(\mathbf{E}(M_B))$ lezártját, (ii) az $\mathbf{F}(M_B)$ torikus modellt, (iii) az összes duplán L-felbontható eloszlás $\mathbf{L} \cap \mathbf{L}'$ családját. Megmutatjuk, hogy ez a három modell szigorúan bővülő sorozatot alkot. Mivel az $\mathbf{F}(M_B)$ család nem csak az M_B sorai által kifeszített altértől függ, meg kell mondanunk, hogy pontosan mi legyen ez a mátrix. Legyenek M_B sorai a $\rho_{aq}^{k\ell}$ vektorok, minden lehetséges k, ℓ, a, q választásra. Tudjuk, hogy ezek nem lineárisan függetlenek, de ez most nem is kell nekünk.

A következő lemma hasznos lesz, ha a megfigyeléseinket $n = 4$-ről ki akarjuk terjeszteni $n > 4$-re. A lemmában * helyére L, invertálva L, vagy duplán L írható.

9.13. Lemma. *(i) Legyen $n < m$, p pedig S_n-en adott eloszlás. Legyen még σ az $n+1, \ldots, m$ egészek tetszőleges permutációja. A p eloszlás S_m-re való σ-felemeltje a következő q eloszlás:*

$$q(\pi) = \begin{cases} p(\pi(1..n)) & \text{ha } \pi\{1..n\} = \{1..n\} \text{ és } \pi(n+1..m) = \sigma, \\ 0 & \text{egyébként.} \end{cases}$$

*Ekkor, ha p *-felbontható, akkor q is az.*
(ii) Legyen $n < m$, és q olyan eloszlás S_m-en, melyre $\sum_{\rho\{1..n\}=\{1..n\}} q(\rho) > 0$.

Legyen még σ az $n+1,\dots,m$ egészek olyan permutációja, melyre

$$\sum_{\rho(n+1..m)=\sigma} q(\rho) > 0.$$

Ekkor a q eloszlás S_n-re való σ-megszorítása a következő p eloszlás:

$$p(\pi) = c \cdot q(\pi, \sigma),$$

*ahol c normáló tényező. Ezekkel a definíciókkal, ha q *-felbontható, akkor p is az.*

A lemma bizonyítása triviális, ezért eltekintünk tőle.

9.14. Tétel. *A szakasz elején definiált modellek között a következő szigorú tartalmazások állnak fenn:*

$$\mathbf{E}(M_B) \subsetneqq \mathbf{F}(M_B) \subsetneqq cl(\mathbf{E}(M_B)) \subsetneqq \mathbf{L} \cap \mathbf{L}'.$$

Bizonyítás. Az első reláció triviális. A másik kettőnél is csak azt kell megmutatnunk, hogy egyenlőség nem állhat fenn. Mindkét esetben mutatunk egy eloszlást a két modell különbségében $n = 4$-re, majd ezt felemeljük az általános $n > 4$ esetre.

Az első reláció esetén $n = 4$-re legyen $T = T_4 \subset S_4$ a következő 16 permutációból álló halmaz:

$$\begin{array}{cccccccc} 1234, & 1243, & 1324, & 1342, & 1423, & 2134, & 2143, & 2341, \\ 3241, & 3412, & 3421, & 4231, & 4213, & 4321, & 4312, & 4123. \end{array} \tag{37}$$

Könnyen adódik, hogy T nem M_B-megvalósítható, mivel az $\cup_{\pi\in T}\mathrm{Supp}(b(\cdot,\pi))$ halmaz M_B minden sorát lefedi ($b(\cdot,\pi)$ most az M_B mátrix π. oszlopát jelöli). A T-n egyenletes p eloszlás tehát nem eleme $\mathbf{F}(M_B)$-nek. Viszont eleme $\mathrm{cl}(\mathbf{E}(M_B))$-nek. Közvetlenül látszik ugyanis, hogy a $p_m \in \mathbf{E}(M_B)$ eloszlások konvergálnak p-hez $m \to \infty$ esetén, ahol $\log p_m = mv + c(m)\mathbf{1}$, ahol v a következő 24 hosszúságú vektor:

$$v = \nu_1^{11} - \nu_1^{12} - \nu_1^{21} + 2\nu_2^{22} - \rho_{25}^{22} + \rho_{25}^{23} + \rho_{25}^{32} - \nu_3^{33} + \rho_{35}^{33},$$

ahol, mint korábban is, $\nu_a^{k\ell} = \sum_{q=1}^{5} \rho_{aq}^{k\ell}$.

Térjünk át az utolsó tartalmazási relációra, $n = 4$-et még mindig rögzítve. A 2.2. Tétel fényében azt kell belátnunk, hogy

$$X_{M_L} \cap X_{M_{L'}} \supsetneqq X_{M_B}.$$

Ezt pedig már tudjuk, hiszen mind a három nemnegatív varietásra meghatároztuk már a rajtuk zérus polinomok ideáljának bázisát. Könnyű felírni például olyan eloszlást, mely az I_{M_L}-t és $I_{M_{L'}}$-t generáló összesen tíz másodfokú polinomot kielégíti, de a (36)-beli első negyedfokú polinomot nem (mely az I_{M_B} ideál eleme). Egy ilyen eloszlás az alábbi permutációkból álló $Z_4 \subset S_4$ halmazon egyenletes eloszlás:

$$1324, \quad 2341, \quad 2431, \quad 3241, \quad 3124, \quad 4231, \quad 4123. \tag{38}$$

Ha most $n > 4$, akkor rögzítsünk egy tetszőleges σ permutációt az $5, \ldots, n$ egészeken. Nézzük az első bizonyítandó relációt! A $T_n = T_4 \times \sigma = \{(\pi, \sigma) : : \pi \in T_4\}$ halmazon egyenletes eloszlás nem eleme $\mathbf{F}(M_B)$-nek, mivel a T_n-et tartalmazó legszűkebb M_B-megvalósítható halmaz az $S_4 \times \sigma = \{(\pi, \sigma) : \pi \in \in S_4\}$ permutációkból áll. Vegyük ugyanis észre, hogy a $k, \ell \leq 4$ esetben

$$\{|\pi(\Sigma_{k\ell})| : \pi \in T_n\} = \{|\pi(\Sigma_{k\ell})| : \pi \in S_4 \times \sigma\},$$

míg ha k vagy ℓ nagyobb négynél, akkor a $|\rho(\Sigma_{k\ell})|$ statisztika konstans az $S_4 \times \sigma$ halmazon (ahol a konstans függ σ-tól)

$$|\rho(\Sigma_{k\ell})| = c(k, \ell, \sigma)\, \forall \rho \in S_4 \times \sigma, \text{ ha } k > 4 \text{ vagy } \ell > 4.$$

Másrészt megmutatjuk, hogy a T_n-en egyenletes eloszlás eleme $\mathrm{cl}(\mathbf{E}(M_B))$-nek. Legyen

$$v_n = \nu_1^{11} - \nu_1^{12} - \nu_1^{21} + 2\nu_2^{22} - \rho_{25}^{22} + \rho_{25}^{23} + \rho_{25}^{32} - \nu_3^{33} + \rho_{35}^{33},$$

ahol a vektorok koordinátái most az S_n-beli permutációkkal vannak indexel-

ve. Minden $\pi \in S_4$-re $v_n(\pi, \sigma) = v_4(\pi)$. Ha most $k, \ell > 4$, akkor definiáljuk a

$$w_n^{k\ell} = \mathbf{1} - \rho^{k\ell}_{a^{k\ell}(\pi,\sigma), q^{k\ell}(\pi,\sigma)}$$

vektorokat, ahol $\pi \in S_4$ tetszőlegesen választható, ettől nem függ a $w_n^{k\ell}$ definíciója. Ezek a vektorok tehát nullák az $S_4 \times \sigma$-beli koordinátákon, de minden más $\rho \in S_n$-re van legalább egy olyan k, ℓ pár, hogy $w_n^{k\ell}(\rho) = 1$. Ha tehát p_m az az $\mathbf{E}(M_B)$-beli eloszlás, melyre

$$\log p_m = m(v_n - c \sum_{k,\ell>4} w_n^{k\ell}) + c(m)\mathbf{1},$$

és c elég nagy, akkor p_m a T_n-en egyenletes eloszláshoz tart.

Nézzük a másik bizonyítandó állítást. Legyen most $Z_n = Z_4 \times \sigma$. A 9.13. Lemma (i) része szerint a Z_n-en egyenletes eloszlás duplán L-felbontható. Megmutatjuk azonban, hogy nem eleme a $\mathrm{cl}(\mathbf{E}(M_B))$ családnak. Tegyük ugyanis fel, hogy eleme. Ekkor lenne hozzá tartó $q_m \in \mathbf{E}(M_B)$ részsorozat, melynek S_4-re való σ-megszorítása, p_m, a 9.13. Lemma (ii) része miatt benne lenne $\mathbf{E}(M_B)$-ben $n = 4$-re. A p_m sorozat viszont a Z_4-en egyenletes eloszláshoz tart, ami ellentmondás. ■

9.15. Következmény. $\mathbf{L} \cap \mathbf{L}'$ *nem torikus modell, így nem is exponenciális család lezártja.*

Az $\mathbf{L} \cap \mathbf{L}'$ metszet pontos leírásával nem rendelkezünk.

Megmutattuk, hogy az $\mathbf{F}(M_B)$ torikus modell nem zárt, van azonban olyan maximális $M_B^{\max}$ reprezentáció (M_B kiegészítve néhány sorral), melyre $\mathbf{F}(M_B^{\max})$ már zárt. A 4ti2 programcsomaggal kiszámoltunk egy ilyen maximális reprezentációt. Egy 32 sorból álló $0-1$ mátrixot kaptunk, melyből 24 sor $\rho^{k\ell}_{aq}$ alakú volt valamilyen k, ℓ, a, q négyesre. A 8 új sor mindegyike nyolc 1-est tartalmazott, az egyik sor például pontosan a (37)-beli permutációkon volt nulla. Ez a maximális reprezentáció is bizonyítja, hogy a (38)-beli permutációkon egyenletes eloszlás nincs a $\mathrm{cl}(\mathbf{E}(M_B))$ családban, mivel a (38) halmaz nem $M_B^{\max}$-megvalósítható, csak akkor válik azzá, ha hozzávesszük az (1423) permutációt.

A $\mathbf{B} = \mathbf{L} \cap \mathbf{L}'$ család nyilván jellemezhető az $\mathbf{L}$ és az $\mathbf{L}'$ családokat jellemző feltételes függetlenségi relációk összességével. E szakasz végén egy másik, „szimmetrikusabb" jellemzést is adunk. Minden $\pi \in S_n$ permutációra jelölje $\delta_\pi = \{(i, \pi(i)) : 1 \leq i \leq n\}$ a π-hez tartozó bástyaelrendezést az $[n] \times [n]$ négyzetben (azaz a sakktábla-reprezentációt). Továbbá jelölje Pr_1 (illetve Pr_2) $[n] \times [n]$-ben az első (illetve második) koordinátára való vetítést, azaz tetszőleges $\delta \subset [n] \times [n]$ részhalmazra

$$\mathrm{Pr}_1(\delta) = \{i \in [n] : \exists j \in [n] \text{ melyre } (i, j) \in \delta\}.$$

9.16. Definíció. *Legyen* Π *véletlen permutáció,* $\mathcal{D}$ *és* $\mathcal{R}$ *pedig* d *és* r *osztályú partíciói* $[n]$*-nek. Azt mondjuk, hogy* Π *majdnem független növekményű* $\mathcal{D} \times \times \mathcal{R}$*-en, ha a*

$$\delta_\Pi \cap (D_i \times R_j), \quad 1 \leq i \leq d,\, 1 \leq j \leq r$$

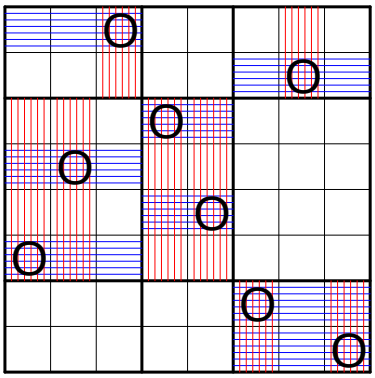
véletlen ponthalmazok feltételesen függetlenek, ha a

$$Pr_1(\delta_\Pi \cap (D_i \times R_j)),\, Pr_2(\delta_\Pi \cap (D_i \times R_j)), \quad 1 \leq i \leq d,\, 1 \leq j \leq r$$

vetületek mind ismertek.

9.17. Tétel. *A* Π *véletlen permutáció akkor és csak akkor duplán L-felbontható, ha majdnem független növekményű minden olyan* $\mathcal{D} \times \mathcal{R}$*-en, ahol* $\mathcal{D}, \mathcal{R}$ *intervallum-partíciók (azaz minden elemük intervallum).*

Bizonyítás. Az egyik irány triviális: a duplán L-felbonthatósághoz elég olyan partíció-párokra tudni a feltételes függetlenségeket, ahol az egyik partíció triviális. A másik irányhoz tegyük fel, hogy Π duplán L-felbontható. Az L-felbonthatóság miatt a $\Gamma_i^1 = \delta_\Pi \cap (D_i \times [n])$ $(1 \leq i \leq d)$ ponthalmazok feltételesen függetlenek, ha az $\Omega_i^1 = \mathrm{Pr}_2(\delta_\Pi \cap (D_i \times [n]))$ $(1 \leq i \leq d)$ vetületek ismertek. A feltételes függetlenség megmarad, ha a feltételbe az $\Omega_j^2 = \mathrm{Pr}_1(\delta_\Pi \cap ([n] \times R_j))$ $(1 \leq j \leq r)$ vetületeket is bevesszük, hiszen ezek a Γ_i^1 változók értékét egyesével korlátozzák. Beláttuk tehát, hogy a Γ_i^1 változók feltételesen függetlenek, ha az Ω_i^1, Ω_j^2 vetületek mind ismertek. Invertálással kapjuk, hogy ugyanez igaz a $\Gamma_j^2 = \delta_\Pi \cap ([n] \times R_j)$ $(1 \leq j \leq r)$ halmazokra is.



5. ábra. Duplán L-felbonthatóság a sakktáblán: feltéve, hogy a téglalapokban ismerjük a rácsokat, a bástyák elhelyezkedése az egyes rácsokon független (9.17. Tétel)

Legyen E az Ω_i^1, Ω_j^2 vetületek által generált σ-algebra egy atomja. Ekkor

$$P(\Pi = \pi | E) = \prod_{i=1}^{d} P(\Gamma_i^1 = \gamma_i^1 | E).$$

Továbbá

$$\Gamma_i^1 = (\delta_\Pi \cap (D_i \times R_j) : 1 \leq j \leq r),$$

ahol $\Gamma_{ij} = \delta_\Pi \cap (D_i \times R_j)$ a Γ_j^2 függvénye, így kapjuk, hogy

$$P(\Gamma_i^1 = \gamma_i^1 | E) = \prod_{j=1}^{r} P(\Gamma_{ij} = \gamma_{ij} | E),$$

és ezt kellett bizonyítani. ■

A tétel tartalmát az 5. ábra szemlélteti.

10. Egyéb felbonthatóságok

10.1. S-felbonthatóság

Térjünk kicsit vissza az S-felbontható és duplán S-felbontható modellekre! Ha nem szorítkozunk a szigorúan pozitív eloszlásokra, akkor a következő definíció tűnik természetesnek.

10.1. Definíció. *Az S_n-en adott p eloszlás, illetve a p eloszlású Π véletlen permutáció* S-felbontható, *ha léteznek $\Lambda(C) \geq 0$ $(C \subseteq [n])$ paraméterek, melyekkel*

$$p(\pi) = \prod_{k=1}^{n} \Lambda(\pi\{1..k\}), \quad \pi \in S_n.$$

Továbbá p, illetve Π duplán S-felbontható, ha Π és Π^{-1} is S-felbontható.

Az S-felbonthatóság másik megfogalmazása az, hogy a $Z_k = \Pi\{1..k\}$, $k = 1, \ldots, n$ véletlen halmazok kvázi-függetlenek. Az interpretációt a következő tétel is segíti.

10.2. Tétel. *A p szigorúan pozitív eloszlás (illetve a Π permutáció) akkor és csak akkor S-felbontható, ha L-felbontható és léteznek olyan $\Lambda'(C) > 0$ $(\emptyset \neq C \subseteq [n])$ paraméterek, melyekre*

$$P(\Pi(k+1) = x \,|\, \Pi\{1..k\} = C) = \frac{\Lambda'(C \cup x)}{\sum_{y \notin C} \Lambda'(C \cup y)}. \tag{39}$$

Bizonyítás. Az egyik irányban, ha p L-felbontható és (39) teljesül, akkor p S-felbontható, mivel

$$p(\pi) = \frac{1}{\sum_{j=1}^{n} \Lambda'(j)} \prod_{k=1}^{n-1} \frac{\Lambda'(\pi\{1..k\})}{\sum_{y \notin \pi\{1..k\}} \Lambda'(\pi\{1..k\} \cup y)} \cdot \Lambda'(\pi\{1..n\}).$$

A másik irányban, tegyük fel, hogy p S-felbontható. Ekkor p nyilván L-fel-

bontható, és

$$P(\Pi(k+1)=x|\Pi\{1..k\}=C)=\\ \frac{\sum_{C_j}\prod_{j=1}^{k-1}\Lambda(C_j)\cdot\Lambda(C)\Lambda(C\cup x)\cdot\sum_{D_j}\prod_{j=k+2}^{n}\Lambda(C\cup x\cup D_j)}{\sum_{C_j}\prod_{j=1}^{k-1}\Lambda(C_j)\cdot\Lambda(C)\cdot\sum_{y\notin C}\Lambda(C\cup y)\sum_{E_j}\prod_{j=k+2}^{n}\Lambda(C\cup y\cup E_j)},$$

ahol a C_j halmazok az összes $C_1 \subsetneq C_2 \subsetneq \cdots \subsetneq C_{k-1} \subsetneq C$ láncot futják be, a D_j halmazok az $[n]\setminus(C\cup x)$ halmaz hasonló láncait futják be, az E_j halmazok pedig az $[n]\setminus(C\cup y)$ láncait futják be. Egyszerűsítve látjuk, hogy (39) teljesül a $\Lambda'(C\cup x)=\Lambda(C\cup x)\cdot\sum_{D_j}\prod_{j=k+2}^{n}\Lambda(C\cup x\cup D_j)$ választással. ■

A fenti tétel alakja emlékeztet a Plackett-Luce modellre: emlékeztetünk, hogy ott

$$P(\Pi(k+1)=x|\,\Pi\{1..k\}=C)=\frac{\lambda_x}{\sum_{y\notin C}\lambda_y}.$$

Azaz minden jelölthöz tartozik egy paraméter, és a következő lépésben a szóbajövő jelöltek közül ezen paraméterekkel arányos valószínűséggel választunk. Az S-felbontható modellben viszont minden halmazhoz tartozik egy paraméter, és a következő lépésben a szóbajövő halmazok közül ezen paraméterekkel arányos valószínűséggel választunk. Képzeljük el a következő feladatot. Tegyük fel, hogy egy bizottságnak egy csapatversenyre kiküldendő csapatot kell kiválasztani, de még nem lehet tudni, hány fős csapat utazhat. A bizottság ezért felállít egy sorrendet, melyből az első valahány jelölt alkotja majd a csapatot. A bizottság minden C lehetséges csapathoz hozzárendel egy $\Lambda'(C)$ kívánatossági értéket. Tegyük fel, hogy a sorrend első $k-1$ eleme már megvan, jelölje az ő halmazukat C. Ekkor annak a valószínűsége, hogy a k. elem x lesz, arányos a $\Lambda'(C\cup x)$ értékkel.

Ha már a Plackett-Luce modellt említettük, akkor bevezethetünk egy analóg S-felbontható modellt. Tegyük fel, hogy minden jelöltnek van egy pozitív λ_i paramétere, mely a jelölt „jóságát” méri, és egy C csapat kívánatossága egyszerűen a tagok paramétereinek összege, azaz $\Lambda(C)=\sum_{i\in C}\lambda_i$.

Ekkor kapjuk, hogy

$$P(\Pi(k+1) = x | \, \Pi\{1..k\} = C) = \frac{\sum_{i \in C} \lambda_i + \lambda_x}{\sum_{y \notin C} (\sum_{i \in C} \lambda_i + \lambda_y)}.$$

Azaz a Plackett-Luce modellel ellentétben, a véletlen sorrend vége felé a jelöltek közötti különbségek csökkennek, a feltételes eloszlás a még választható jelöltek halmazán egyre egyenletesebb lesz.

Ugyanúgy, mint az L-felbontható esetben, a 10.1. Definícióból kiolvasható az az M_S mátrix, mellyel az S-felbontható modell éppen az $\mathbf{F}(M_S)$ torikus modellel egyezik meg. Megkérdezhetjük, hogy ez a torikus modell zárt-e, illetve kiszámíthatjuk Markov bázisát. $n = 3$-ra mind az S-felbonthatóság, mind a duplán S-felbonthatóság lényegében a kvázi-függetlenséggel ekvivalens, tehát nem annyira érdekes.

Az $n = 4$ esetben a 4ti2 programmal a maximális reprezentációt kiszámolva adódik az eredmény, hogy az $\mathbf{F}(M_S)$ modell nem zárt. A programmal a minimális Markov bázist kiszámolva pedig azt kapjuk, hogy az egyrészt az L-felbontható család hat darab másodfokú polinomjából áll, ezen kívül pedig még 64 darab harmadfokú, és 93 darab negyedfokú polinom is van benne. A harmadfokú polinomokat viszonylag könnyű jellemezni. Minden $i \in [4]$-re legyen $C_1^i, C_2^i, C_3^i \subset [4]$ az i-t tartalmazó három darab kételemű részhalmaz, $D_1^i, D_2^i, D_3^i \subset [4]$ pedig az i-t tartalmazó három darab három elemű részhalmaz. Az S_4 gráfja erre a hat csúcsra megszorítva hat élt tartalmaz, mely két darab teljes párosítás uniója. Ha $j = 1,2,3$-ra választunk egy-egy tetszőleges C_j^i-hez vezető utat a gráfban, akkor a polinom egyik tagjában az egyik teljes párosításon megyünk tovább, a másik tagban a másikon. Például $i = 1$-re a C_j^i halmazok: $\{1,2\}, \{1,3\}, \{1,4\}$, válasszuk hozzájuk az $12, 31, 41$ utakat. Ebből a bázisbeli polinom:

$$x_{1234} x_{3142} x_{4123} - x_{1243} x_{3124} x_{4132}.$$

Ezzel $4 \cdot 8 = 32$ polinomot kaptunk. A másik 32 pedig úgy kapható, hogy minden x_π változó indexébe π helyett annak megfordítottját írjuk, azaz pl.

az előző polinomból

$$x_{4321}x_{2413}x_{3214} - x_{3421}x_{4213}x_{2314}$$

lesz. A negyedfokú bázispolinomok hasonlóak: az összes kételemű részhalmaz közül hagyjunk el egyet (legyen ez A), és annak komplementerét, jelölje a maradék négy kételemű részhalmazt $C_j^{\{A,\overline{A}\}}$ $(1 \leq j \leq 4)$. A D_j halmazok pedig legyenek a háromelemű részhalmazok $(1 \leq j \leq 4)$. Az S_4 gráfja erre a nyolc csúcsra megszorítva egy nyolc hosszúságú kör. Ha választunk egy-egy tetszőleges $C_j^{\{A,\overline{A}\}}$-hez vezető utat a gráfban, akkor a polinom mindkét tagjában a kör négy nem-szomszédos élén megyünk tovább. Az előző példánál maradva, ha az $13, 41, 32, 24$ utakat választjuk, akkor a polinom

$$x_{1342}x_{4123}x_{3214}x_{2431} - x_{1324}x_{4132}x_{3241}x_{2413}.$$

Ez összesen $3 \cdot 16 = 48$ polinom, megfordítva még 48 polinomot kapunk, ami összesen 96 negyedfokú polinom. Az így kapott bázis azonban nem minimális. Válasszunk megint egy kételemű A halmazt, és S_4 gráfjából hagyjuk el az $A, \overline{A}$ csúcsokat. Tekintsük most azokat a mintákat, melyekre a maradék gráf minden C $(0 < |C| < 4)$ csúcsán pontosan egy mintaelem halad át. Ilyen mintából négy darab van, melyeket a felsorolt báziselemek egy körré kapcsolnak össze. A Markov bázis akkor lesz minimális, ha e négy báziselemből egyet (méghozzá bármelyiket) elhagyjuk. Mivel három $A, \overline{A}$ pár van, a 96-ból 3-at kivonva kapjuk, hogy a minimális bázisban 93 negyedfokú polinom szerepel.

Hasonló eredményeket kapunk a duplán S-felbontható eloszlások esetére is. Itt is – mint a duplán L-felbontható esetben – három modell különül el: legszűkebb a torikus modell, ennél bővebb annak lezártja, legbővebb pedig a 10.1. Definíció modellje. Szinte biztos, hogy kevés erőfeszítéssel precízen lehetne bizonyítani a 9.14. Tételnek megfelelő eredményt erre a három modellre is. Mindenesetre $n = 4$-re a 4ti2 programmal kiszámoltuk, hogy a torikus modell nem zárt. A minimális Markov bázis egyrészt a duplán L-felbontható család tíz darab másodfokú polinomjából áll. Ezen kívül tartalmaz még 104 darab harmadfokú, és 33 darab negyedfokú polinomot is. A ne-

gyedfokú polinomok mindegyike az S-felbontható család Markov bázisából, valamint azok invertáltjaiból származik, és a harmadfokú polinomok többsége is ilyen. Összesen 8 darab új harmadfokú polinom van, egyikük:

$$x_{2341}x_{3412}x_{4123} - x_{2413}x_{3142}x_{4321}\,.$$

Ez a nyolc új polinom is egyetlen orbitot alkot ugyanabban az értelemben, mint a duplán L-felbontható család esetében.

Az S-felbontható modell egy (nem minimális) Markov bázisa általános n-re is megadható. Tartalmazza az L-felbontható modell Markov lépéseit, plusz olyan polinomokat, melyekben a permutációkat egy „alternáló kör mentén átkötjük". A kérdés feltehető általános forrás-nyelő gráfok esetében is, amennyiben a forrás-nyelő utak valószínűsége az út csúcsaiban ülő paraméterek szorzata. Most azonban a bizonyítás nem megy át erre az esetre teljes általánosságban.

10.3. Tétel. *Jelölje G_n az S_n gráfját. A gráf k. szintje álljon azokból a $C \subseteq [n]$ csúcsokból, melyekre $|C| = k$. Legyen K a G_n gráfban egy olyan (páros hosszú) kör, melynek minden csúcsa a k. vagy a $(k+1)$. szinthez tartozik valamilyen $1 \leq k \leq n-2$-re. Jelölje a K kör egymás utáni csúcsait $C_1, D_1, C_2, D_2, \ldots, C_j, D_j$, ahol a C_i csúcsok a k., a D_i csúcsok a $(k+1)$. szinten vannak. Legyenek még $\pi_i \in S_{C_i}$ a C_i halmazok elemeinek permutációi, és $\rho_i \in S_{[n]\setminus D_i}$ az $[n] \setminus D_i$ halmazok elemeinek permutációi. Minden ilyen választásra készítsük el a*

$$\prod_{i=1}^{j} x_{(\pi_i, D_i \setminus C_i, \rho_i)} - \prod_{i=1}^{j} x_{(\pi_i, D_{i-1} \setminus C_i, \rho_{i-1})} \tag{40}$$

polinomot, ahol $D_0 = D_j$ és $\rho_0 = \rho_j$. Ekkor a (19) és a (40) polinomok együttesen generálják az I_{M_S} ideált.

Bizonyítás. Megmutatjuk, hogy a mondott polinomokhoz tartozó Markov lépésekkel tetszőleges két olyan u és v gyakoriságvektor összeköthető, melyekre a G_n gráf minden csúcsán ugyanannyi permutáció megy át. Az u gyakoriságvektorból elindulva, elég a (40) lépések segítségével egy olyan u' gyako-

riságvektorba eljutni, hogy a G_n gráf minden élén ugyanannyi permutáció megy át u' és v szerint. Hiszen ekkor már tudjuk, hogy u'-ből elérhető v a (19) lépések felhasználásával.

Azt tudjuk, hogy a 0. és 1. szintek közötti éleken ugyanannyi permutáció megy át u és v szerint. Tegyük fel, hogy eljutottunk egy olyan u_k vektorba, hogy a $(j-1)$. és j. szintek közötti éleken ugyanannyi permutáció megy át u_k és v szerint minden $j \leq k$-ra. Nézzük most a k. és $(k+1)$. szintek közötti éleket! Ehhez készítsünk el egy-egy $\binom{n}{k} \times \binom{n}{k+1}$ méretű kontingenciatáblát u_k-ra és v-re. A táblák sorait indexeljük az $[n]$ alaphalmaz k elemű részhalmazaival, oszlopait pedig a $k+1$ elemű részhalmazokkal. A (C, D) cellába pedig írjuk be, hogy hány permutáció megy át a $C \to D$ élen az adott gyakoriságvektor szerint. A feltevés szerint e két tábla marginálisai megegyeznek. A tábláknak minden olyan (C, D) cellája strukturális nulla, melyre $C \not\subset D$. Szerencsére a strukturális nullákkal rendelkező, rögzített marginálisú kétdimenziós kontingenciatáblák Markov bázisa is ismert (pl. Diaconis és Sturmfels [32], Remark 3.4). Ezeket a kontingenciatábla-báziselemeket lehet most is a permutációkra átvinni úgy, hogy a korábbi élek elégséges statisztikáját ne rontsuk el, pont ugyanúgy, mint a 8.4. Tétel bizonyításában. Ebből következik az állítás. ■

10.2. Bal-jobb szorzások hatása a modellekre

Egy A és B halmaz közötti véletlen párosítást leíró Π véletlen permutáció L-felbonthatósága függhet attól, hogy a halmazok elemeit hogyan számozzuk. Tegyük fel, hogy valamely számozás (címkézés) mellett a permutáció Π_0, azaz $\Pi_0(k)$ az A-beli k-címkéjű elem B-beli párjának címkéje. Cseréljük most minden $i \in [n]$-re az A-beli i címkét $\sigma(i)$-re, ahol $\sigma \in S_n$, és a B-beli i címkét $\rho(i)$-re, ahol $\rho \in S_n$. Az új címkézéssel az eredeti párosítást a Π_1 permutáció írja le. A két permutáció között a $\Pi_1 = \rho\Pi_0\sigma^{-1}$ összefüggés áll fenn. Ezért az eloszlásukra

$$p_0(\pi) = P(\Pi_0 = \pi) = P(\Pi_1 = \rho\pi\sigma^{-1}) = p_1(\rho\pi\sigma^{-1}).$$

Ebben a szakaszban azt vizsgáljuk, hogy az L-felbonthatóság megőrződik-e ilyen átszámozások során.

Korábban már elmondtuk, hogy mikor nevezünk egy eloszláscsaládot invariánsnak egy $\phi : S_n \to S_n$ kölcsönösen egyértelmű transzformációra nézve. Vezessünk be néhány ilyen transzformációt! Minden $\sigma \in S_n$-re legyen $\phi_{\circ\sigma} : : \pi \mapsto \pi\sigma$, illetve $\phi_{\sigma\circ} : \pi \mapsto \sigma\pi$ a σ-val való jobbról, illetve balról szorzás. Jelölje $\sigma_{(12)}$ az 1-et és 2-t felcserélő permutációt (inverziót), és σ_r a megfordító permutációt, azaz melyre $\sigma_r(k) = n + 1 - k$.

Az e szakasz elején vizsgált véletlen párosítások esetében az A, B halmazoknak lehet, hogy van természetes számozása, de lehet, hogy a számok csak címkék, melyeknek értéke semmiféle jelentést nem hordoz. A sorbarendezések esetében azonban a sorban elfoglalt egymás utáni helyek számozása adott, és csak a sorbarendezett objektumok számozása variálható. Az L-felbonthatóság definíciójából azonnal látszik, hogy egy ilyen véletlen sorbarendezés L-felbonthatósága nem függ az objektumok számozásától. Ezt úgy is megfogalmazhatjuk, hogy az L-felbontható eloszláscsalád *címke-invariáns*, azaz invariáns a $\phi_{\sigma\circ}$ balról szorzásokra. Az invertálva L-felbontható család ennek megfelelően a jobbról szorzásokra invariáns. Bár azt mondtuk, hogy a sorbarendezési szituációban nem variáljuk a sorrend helyeit, matematikailag értelmes kérdés, hogy az L-felbontható család invariáns-e a jobbról szorzásokra. Különösen érdekelhet minket a sorrend helyeinek megfordítása, azaz amikor az objektumokat nem a legjobbtól a legrosszabbig rakjuk sorba, hanem fordítva, a legrosszabbtól a legjobbig. Ha egy sorbarendezési modell invariáns erre a transzformációra, azaz a $\phi_{\circ\sigma_r}$ jobbról szorzásra, akkor a modell *megfordítható*. A címke-invariancia és a megfordíthatóság teljesülését Critchlow et al. [15] számos modellre ellenőrizte. A következő tétel azt mondja ki, hogy az L-felbontható család megfordítható, és ezen kívül lényegében csak egy másik jobbról szorzásra invariáns.

10.4. Tétel. *Legyen $n \geq 4$. Az* **L** *család az összes $\phi_{\sigma\circ}$ balról szorzásra invariáns. A jobbról szorzások közül pontosan azokra a $\phi_{\circ\sigma}$ transzformációkra invariáns, ahol σ eleme a σ_r és $\sigma_{(12)}$ által generált nyolc elemű csoportnak.*

Bizonyítás. A balról szorzások esete triviális. A $\phi_{\circ\sigma_r}$ esetben

$$p_{\circ\sigma_r}(\pi) = \prod_{k=0}^{n-1} \Lambda(\pi(n-k), \pi\{n-k+1..n\}) = \prod_{j=0}^{n-1} \Lambda^r(\pi(j+1), \pi\{1..j\}),$$

ahol $\Lambda^r(x, C) = \Lambda(x, \overline{C} \setminus x)$, ez mutatja az invarianciát. A $\phi_{\circ\sigma_{(12)}}$ esetben pedig

$$p_{\circ\sigma_{(12)}}(\pi) = \Lambda(\pi(2), \emptyset) \cdot \Lambda(\pi(1), \pi(2)) \cdot \prod_{k=2}^{n-1} \Lambda(\pi(k+1), \pi\{1..k\}) = \prod_{k=0}^{n-1} \Lambda^{(12)}(\pi(k+1), \pi\{1..k\}),$$

ahol

$$\Lambda^{(12)}(x, C) = \begin{cases} 1 & \text{ha} \quad C = \emptyset \\ \Lambda(x, \emptyset) \cdot \Lambda(x, C) & \text{ha} \quad |C| = 1 \\ \Lambda(x, C) & \text{ha} \quad |C| \geq 2 \end{cases},$$

azaz megint készen vagyunk.

Meg kell még mutatni, hogy **L** semmilyen más jobbról szorzásra nem invariáns. Ehhez belátjuk, hogy minden más $\sigma \in S_n$-hez van olyan (szigorúan pozitív) duplán L-felbontható p eloszlás, hogy $p_{\circ\sigma}$ nem L-felbontható. Azt már tudjuk, hogy a szigorúan pozitív L-felbontható eloszlások a

$$\frac{p(\pi_{11})}{p(\pi_{12})} = \frac{p(\pi_{21})}{p(\pi_{22})} \tag{41}$$

összefüggésekkel jellemezhetők, ahol π_{ij}, $1 \leq i, j \leq 2$ „átkeresztezett" permutációk.

A tétel állításában szereplő nyolc elemű részcsoport tagjai:

$$id,\ \sigma_r,\ \sigma_{(12)},\ \sigma_r\sigma_{(12)},\ \sigma_r\sigma_{(12)}\sigma_r,\ \sigma_{(12)}\sigma_r,\ \sigma_r\sigma_{(12)}\sigma_r\sigma_{(12)},\ \sigma_{(12)}\sigma_r\sigma_{(12)}.$$

Ha σ nem eleme ennek a részcsoportnak, akkor az inverze sem, és ezért van

olyan $2 \leq a \leq n-2$, melyre

$$\sigma^{-1}\{1..a\} \neq \{1..a\}, \{n-a+1..n\}.$$

Rögzítsünk egy ilyen a-t. Van tehát $c, e \in \{1..a\}$ és $d, f \notin \{1..a\}$, melyekre

$$c^* = \sigma^{-1}(c) > \sigma^{-1}(d) = d^*, \quad e^* = \sigma^{-1}(e) < \sigma^{-1}(f) = f^*.$$

Az α, β, γ számokra azt mondjuk, hogy α elválasztja β-t és γ-t, ha $\beta < \alpha < \gamma$ vagy $\beta > \alpha > \gamma$. Ha most $d^* \geq f^*$, akkor d^* (és f^* is) elválasztja c^*-ot és e^*-ot. Ha pedig $d^* < f^*$, akkor vagy valamelyikük elválasztja c^*-ot és e^*-ot, vagy c^* (és e^* is) elválasztja d^*-ot és f^*-ot. Ezért a következő két eset valamelyike teljesül:

1. $\exists \quad c, e \in \{1..a\}, \qquad d \notin \{1..a\}: \quad d^*$ elválasztja c^*-ot és e^*-ot
2. $\exists \quad\ \ c \in \{1..a\}, \quad d, f \notin \{1..a\}: \quad c^*$ elválasztja d^*-ot és f^*-ot

A két eset ugyanúgy kezelhető, a bizonyítást az elsőre részletezzük. Legyen $f \notin \{1..a\}$, $f \neq d$ tetszőleges, és $f^* = \sigma^{-1}(f)$. Emlékezzünk vissza a (32) definícióra, és legyen $p = c(d^*) \exp\{\rho_{d^*5}^{d^*d^*}\}$, ez egy pozitív duplán L-felbontható eloszlás. Legyen $\pi_{11} = \sigma^{-1}$, melyből π_{22}-t két pár elem felcserélésével kapjuk:

$$\pi_{22}(c) = e^*, \pi_{22}(e) = c^*, \pi_{22}(d) = f^*, \pi_{22}(f) = d^*.$$

Készítsük el a π_{12} és π_{21} a-nál átkeresztezett permutációkat. Megmutatjuk, hogy erre a négy permutációra $p_{\circ\sigma}$ nem elégíti ki a (41) egyenletet. Egyrészt ugyanis $\pi_{11}\sigma = id$, melyre $\rho_{d^*5}^{d^*d^*}(id) = 1$. Másrészt viszont a $\pi_{12}\sigma$ és $\pi_{21}\sigma$ koordinátákon $\rho_{d^*5}^{d^*d^*}$ nullát vesz fel, mivel $\pi_{12}\sigma$-nak d^* nem fixpontja, és $\pi_{21}\sigma$ első d^* koordinátája között van d^*-nál nagyobb. Ezzel bizonyításunk teljes. ■

Végül megjegyezzük, hogy az S-felbontható eloszláscsalád is rendelkezik a 10.4. Tételben megfogalmazott invarianciákkal, és valószínűleg be lehet bizonyítani, hogy csak azokkal.

10.3. Teljesen L-felbontható eloszlások

Ebben a szakaszban karakterizáljuk a teljesen L-felbontható (TL-felbontható) pozitív eloszlásokat: ezek azok a p eloszlások S_n-en, melyekre a $p_{\circ\sigma}$ jobbról szorzott eloszlás L-felbontható minden $\sigma \in S_n$-re. Másképpen, a TL-felbontható eloszlások pontosan azok, melyek majdnem független növekményűek minden $\mathcal{D} \times \mathcal{R}$ szorzatpartíción.

Megint csak az $n \geq 4$ eset az érdekes. Belátjuk, hogy ekkor minden szigorúan pozitív TL-felbontható eloszlás kvázi-független, azaz léteznek $c_i(x)$, $1 \leq i, x \leq n$ paraméterek, melyekkel

$$p(\pi) = \prod_{i=1}^{n} c_i(\pi(i)) \quad \forall \pi \in S_n. \tag{42}$$

10.5. Tétel. *Legyen $n \geq 4$. Az S_n-en adott szigorúan pozitív p eloszlás akkor és csak akkor TL-felbontható, ha kvázi-független, azaz* (42) *alakú.*

A 10.5. Tétel $n = 3$-ra nem igaz, hiszen ekkor minden eloszlás TL-felbontható, míg ismert, hogy S_3-on a kvázi függetlenséget a

$$p(123)p(231)p(312) = p(132)p(321)p(213)$$

egyenlőség karakterizálja. Ebből kapjuk, hogy a 10.5. Tétel nem marad igaz, ha a pozitivitás feltételét elhagyjuk.

10.6. Példa. Legyen $n \geq 4$, és legyen q tetszőleges eloszlás S_3-on. Rögzítsük a $\{4, \ldots, n\}$ számok egy tetszőleges σ permutációját. Definiáljuk S_n-en a következő p eloszlást:

$$p(\pi) = \begin{cases} q(\rho) & \text{ha } \pi = (\rho, \sigma) \\ 0 & \text{egyébként,} \end{cases}$$

ahol (ρ, σ), mint eddig, a két permutáció egymás után írását jelöli. Ekkor p TL-felbontható, de nem kvázi független, hacsak q nem az. ■

A 10.5. Tétel egyik iránya triviális: könnyű látni, hogy minden kvázi-független eloszlás TL-felbontható. A másik irányt lemmák sorozatán keresztül

bizonyítjuk.

10.7. Lemma. *Legyen p szigorúan pozitív TL-felbontható eloszlás S_n-en. Ekkor vannak olyan $d_{ij}(x,y)$ paraméterek (ahol $1 \le i < j \le n$ és $1 \le x < y \le n$), melyekkel*

$$p(\pi)/p(\sigma) = d_{ij}(x,y), \tag{43}$$

ha $\pi(i) = x$, $\pi(j) = y$, $\sigma(i) = y$, $\sigma(j) = x$, és $\pi(k) = \sigma(k)$ ha $k \neq i,j$.

Bizonyítás. Legyenek π, σ a lemmában szereplő permutációk. Meg kell mutatnunk, hogy a $p(\pi)/p(\sigma)$ hányados csak az i,j,x,y négyestől függ. A TL-felbonthatóságot felhasználva kapjuk, hogy

$$p(\pi) = p_i(x) \cdot p_{j|i}(y|x) \cdot p_{k_1,\dots,k_{n-2}|i,j}(\pi(k_1),\dots,\pi(k_{n-2})|\{x,y\}),$$

ahol $k_1,\dots,k_{n-2}$ az $[n] \setminus \{i,j\}$ halmaz tetszőleges felsorolása, és az utolsó feltételes valószínűség feltételében x,y helyett $\{x,y\}$-t írhattunk a TL-felbonthatóság miatt. $p(\sigma)$-t is ilyen alakban felírva, a hányadosra

$$\frac{p(\pi)}{p(\sigma)} = \frac{p_i(x)p_{j|i}(y|x)}{p_i(y)p_{j|i}(x|y)}$$

adódik, azaz a lemmát bebizonyítottuk. ■

Vegyük észre, hogy ha p kvázi-független, akkor

$$\frac{p(\pi)}{p(\sigma)} = \frac{c_i(x)c_j(y)}{c_i(y)c_j(x)}, \tag{44}$$

minden olyan π, σ esetén, mely a 10.7. Lemma állításában szerepel. Ennek megfordítása is igaz.

10.8. Lemma. *Legyen p szigorúan pozitív eloszlás S_n-en. Tegyük fel, hogy léteznek olyan $c_i(x)$ paraméterek (ahol $1 \le i,x \le n$), melyekkel (44) teljesül, valahányszor $\pi(i) = x$, $\pi(j) = y$, $\sigma(i) = y$, $\sigma(j) = x$, és $\pi(k) = \sigma(k)$ ha $k \neq i,j$. Ekkor*

$$p(\pi) = K \prod_{i=1}^{n} c_i(\pi(i)) \quad \forall \pi \in S_n$$

valamilyen K konstanssal.

Bizonyítás. Legyen $id = (12\dots n)$ az identitás permutáció, és tegyük fel, hogy (44) teljesül. Az id permutációból kiindulva, jussunk el π-hez transzpozíciók valamilyen sorozatával, azaz kapjuk az $id = \rho_0,\ \rho_1, \dots,\ \rho_k = \pi$ sorozatot. A sorozat szomszédos tagjai csak egy transzpozícióban különböznek, azaz alkalmazható rájuk a (44) összefüggés. Tehát

$$p(\pi) = p(id) \cdot \frac{p(\rho_1)}{p(id)} \cdot \frac{p(\rho_2)}{p(\rho_1)} \cdots \frac{p(\pi)}{p(\rho_{k-1})}.$$

A hányadosok mindegyike (44) alakú valamilyen i, j, x, y-ra. Rögzített i, x párra a $c_i(x)$ tényező akkor szerepel egy hányados nevezőjében, ha az adott transzpozíció x-et elmozdítja az i. pozícióból. Hasonlóan, a $c_i(x)$ tényező akkor szerepel valamelyik hányados számlálójában, ha az adott transzpozíció x-et az i. pozícióba mozgatja. Ezek a tényezők csak azokra az i, x párokra nem ejtik ki egymást, melyekre $\pi(i) \neq i$, és $x = i$ vagy $x = \pi(i)$. Kapjuk tehát, hogy

$$p(\pi) = \frac{\prod_{i:\pi(i)\neq i} c_i(\pi(i))}{\prod_{i:\pi(i)\neq i} c_i(i)} p(id) = \frac{p(id)}{\prod_{i=1}^{n} c_i(i)} \prod_{i=1}^{n} c_i(\pi(i)).$$

■

10.9. Lemma. *Legyenek $d_{ij}(x, y)$ pozitív számok (ahol $1 \leq i < j \leq n$ és $1 \leq x < y \leq n$). Akkor és csak akkor léteznek a*

$$d_{ij}(x, y) = \frac{c_i(x)c_j(y)}{c_i(y)c_j(x)} \tag{45}$$

egyenleteket kielégítő $c_i(x)$ $(1 \leq i, x \leq n)$ paraméterek, ha a $d_{ij}(x, y)$ számok kielégítik a következő két egyenletrendszert:

$$\begin{aligned} d_{ij}(x, y)d_{jk}(x, y) &= d_{ik}(x, y) \quad \forall i < j < k,\ x < y \\ d_{ij}(x, y)d_{ij}(y, z) &= d_{ij}(x, z) \quad \forall x < y < z,\ i < j. \end{aligned} \tag{46}$$

Bizonyítás. A „csak akkor” irány triviális. Az „akkor” irányhoz tegyük fel, hogy a (46) egyenletek teljesülnek. Legyen $c_i(x) = 1$, ha $\min(i, x) = 1$, és

$c_i(x) = d_{1i}(1,x)$, ha $i,x > 1$. Könnyű ellenőrizni, hogy ezzel a definícióval (45) teljesül. ■

A következő lemma teljessé teszi a 10.5. Tétel bizonyítását.

10.10. Lemma. *Legyen p szigorúan pozitív TL-felbontható eloszlás S_n-en, ahol $n \geq 4$. Ekkor a (43) egyenlettel definiált $d_{ij}(x,y)$ mennyiségek kielégítik a (46) egyenletrendszereket.*

Bizonyítás. Vegyük elsőnek észre, hogy a TL-felbontható eloszlások családja definíció szerint invariáns az összes $\phi_{\circ\sigma}$ jobb- és az összes $\phi_{\sigma\circ}$ bal-szorzásra. Ezen túl az eloszláscsalád invariáns a $\phi_{-1} : \pi \mapsto \pi^{-1}$ invertálásra is. Ezért elég a

$$d_{12}(1,2)d_{23}(1,2) = d_{13}(1,2) \tag{47}$$

egyenlőséget belátni. A (46) első sorának többi egyenlete bal- és jobbszorzásokkal következik ebből, míg a (46) második sorának egyenletei invertálással adódnak. A (47) egyenlőséget kiírva, bizonyítandó a

$$p(123\sigma')p(231\sigma')p(312\sigma') = p(132\sigma')p(321\sigma')p(213\sigma') \tag{48}$$

egyenlőség, ahol σ' a $4,\ldots,n$ egészek tetszőlegesen rögzített permutációja. Vezessük be a $q(\pi) = \log p(\pi)$ jelölést. A pozitív TL-felbontható eloszlások logaritmusai abban a lineáris altérben ülnek, melyet a különböző permutált L-felbonthatóságokat kifejező lineáris feltételek jelölnek ki. Azt kell belátni, hogy a (48) egyenlet logaritmusaként kapott lineáris feltétel a korábbi lineáris feltételek következménye. Most pontosan megadjuk, hogy mely lineáris feltételekből vezethető le (48) logaritmusa. Megjegyezzük, hogy az alábbi levezetést a számítógép adta, $n = 4$-re. Válasszuk σ'-t olyannak, hogy első eleme 4, azaz $\sigma' = (4,\sigma)$, ahol σ az $5,\ldots,n$ egészek permutációja. A TL-felbonthatóság miatt q-ra teljesülnek a 7. táblázatban felsorolt egyenlőségek. Az első három összefüggés például abból következik, hogy a $(\Pi(2),\Pi(3),\Pi(5),\ldots)$ koordináták függetlenek a $(\Pi(1),\Pi(4))$ koordinátáktól (ezeket vastagon jelöltük), feltéve, hogy ismert a $\{\Pi(1),\Pi(4)\}$ halmaz. A többi összefüggés hasonlóan következik, vastaggal jelöltük, hogy mely

7. táblázat. Egy TL-felbontható eloszlás q logaritmusára teljesülő összefüggések

$q(\mathbf{3}41\mathbf{2}\sigma)$	+	$q(\mathbf{2}14\mathbf{3}\sigma)$	−	$q(\mathbf{3}14\mathbf{2}\sigma)$	−	$q(\mathbf{2}41\mathbf{3}\sigma)$	=	0
$q(3241\sigma)$	+	$q(1423\sigma)$	−	$q(3421\sigma)$	−	$q(1243\sigma)$	=	0
$q(2431\sigma)$	+	$q(1342\sigma)$	−	$q(2341\sigma)$	−	$q(1432\sigma)$	=	0
$q(\mathbf{43}1\mathbf{2}\sigma)$	+	$q(\mathbf{12}4\mathbf{3}\sigma)$	−	$q(\mathbf{13}4\mathbf{2}\sigma)$	−	$q(\mathbf{42}1\mathbf{3}\sigma)$	=	0
$q(2341\sigma)$	+	$q(4123\sigma)$	−	$q(4321\sigma)$	−	$q(2143\sigma)$	=	0
$q(4231\sigma)$	+	$q(3142\sigma)$	−	$q(3241\sigma)$	−	$q(4132\sigma)$	=	0
$q(\mathbf{14}32\sigma)$	+	$q(\mathbf{41}23\sigma)$	−	$q(\mathbf{41}32\sigma)$	−	$q(\mathbf{14}23\sigma)$	=	0
$q(4231\sigma)$	+	$q(2413\sigma)$	−	$q(2431\sigma)$	−	$q(4213\sigma)$	=	0
$q(3421\sigma)$	+	$q(4312\sigma)$	−	$q(4321\sigma)$	−	$q(3412\sigma)$	=	0
$2q(\mathbf{3}12\mathbf{4}\sigma)$	+	$2q(\mathbf{4}21\mathbf{3}\sigma)$	−	$2q(\mathbf{3}21\mathbf{4}\sigma)$	−	$2q(\mathbf{4}12\mathbf{3}\sigma)$	=	0
$2q(2314\sigma)$	+	$2q(4132\sigma)$	−	$2q(2134\sigma)$	−	$2q(4312\sigma)$	=	0
$2q(1234\sigma)$	+	$2q(4321\sigma)$	−	$2q(1324\sigma)$	−	$2q(4231\sigma)$	=	0

koordináta-részhalmazok feltételes függetlenségét használjuk. A 7. táblázat sorait összeadva, majd kettővel osztva kapjuk a kívánt

$$q(3124\sigma) + q(2314\sigma) + q(1234\sigma) - q(3214\sigma) - q(2134\sigma) - q(1324\sigma) = 0$$

egyenlőséget.

■

Megkérdezhetjük, hogy melyek azok a szigorúan pozitív eloszlások, melyek teljesen S-felbonthatók (TS-felbonthatók). Egy ilyen eloszlás TL-felbontható is, azaz a 10.5. Tétel szerint kvázi-független. Azonban könnyen látszik, hogy a kvázi-független eloszlások S-felbonthatók, mert kielégítik a 10.3. Tételben szereplő polinomokat. Emiatt persze TS-felbonthatók is, hiszen tetszőleges jobbról szorzottjuk is kvázi-független. Azaz azt kaptuk, hogy a szigorúan pozitív eloszlások körében a TL-felbonthatóság és a TS-felbonthatóság ekvivalens fogalmak. Természetesen a kvázi-független eloszlások duplán S-felbonthatók is. Az, hogy a kvázi-független eloszlások duplán S-felbonthatók, közvetlenül is látszik: ezen eloszlások logaritmusai a 9.11. Tétel bizonyításában bevezetett $v^{k\ell}$ vektorok lineáris kombinációi. Szintén e tétel

bizonyítása során megmutattuk, hogy a $v^{k\ell}$ vektorok felírhatók a $\nu_a^{k\ell}$ vektorok lineáris kombinációiként. Végül a $\nu_a^{k\ell}$ vektorok a duplán S-felbontható modellhez tartozó altér elemei.

10.4. Hierarchikus modellek metszete

Tegyük fel, hogy van két (szigorúan pozitív) hierarchikus modellünk:

$$\mathcal{L}_1 = \mathcal{L}(\mathcal{P}_1, \ldots, \mathcal{P}_s), \quad \mathcal{L}_2 = \mathcal{L}(\mathcal{R}_1, \ldots, \mathcal{R}_t).$$

Legyen $\mathcal{D}_{ij}$ a $\mathcal{P}_i$ és az $\mathcal{R}_j$ partíciók közös durvítása. Ekkor triviálisan

$$\mathcal{L}(\mathcal{D}_{ij} : 1 \leq i \leq s, 1 \leq j \leq t) \subseteq \mathcal{L}_1 \cap \mathcal{L}_2. \tag{49}$$

A 9.7. Következmény egy olyan esetről szól, amikor a fenti tartalmazás egyenlőséggel teljesül. Az egyik kérdés, melyre nem tudjuk a választ, hogy mindig igaz-e, hogy a (49) egyenlet baloldalán álló $\mathcal{L}_3$ hierarchikus modell a legbővebb olyan hierarchikus modell, melyet a metszet tartalmaz. Egy másik kérdés, hogy az $\mathcal{L}_1 \cap \mathcal{L}_2$ metszet, mint exponenciális család, megadható-e nulla-egy mátrixszal, azaz van-e olyan csupa $0-1$ elemű M mátrix, hogy $\mathcal{L}_1 \cap \mathcal{L}_2 = \mathbf{E}(M)$. Sajnos erre a kérdésre sem tudjuk a választ.

Egy viszonylag kis méretű esetben a következőképpen járhatunk el. Jelölje az $\mathcal{L}_i$ elemeinek logaritmusai által kifeszített alteret $V_i \subseteq \mathbb{R}^{n!}$, $i = 1,2,3$. Két kérdésünk van tehát: 1) Teljesül-e, hogy $V_1 \cap V_2 = V_3$? 2) Ha nem, akkor van-e a $V_1 \cap V_2$ altérnek indikátor vektorokból álló bázisa? Az első kérdés megválaszolásához elég az alterek dimenzióját kiszámolni. A második kérdés megválaszolásához pedig meg kell keresni a $V_1 \cap V_2$ altér összes indikátorvektorát. Ez utóbbi feladatot könnyítheti meg a következő tétel.

10.11. Tétel. *Legyen M olyan $0-1$ mátrix, melyre az $\mathbf{F}(M)$ torikus modell zárt. Ekkor $M^\top$ képterének minden indikátorvektora előáll, mint M néhány sorának összege.*

Bizonyítás. Jelölje $M^\top$ képterét V, emlékeztetünk, hogy mindig feltesszük, hogy $\mathbf{1} \in V$. Legyen $v \in V$ indikátorvektor, jelölje $S = \mathrm{Supp}(v)$ a v vektor

tartóját, és legyen $\overline{S}$ az S halmaz komplementere. Először megmutatjuk, hogy az $\overline{S}$ halmaz M-megvalósítható. Jelölje ugyanis M^* azt a mátrixot, melyet M-ből úgy kapunk, hogy hozzávesszük a v vektort utolsó, $(t+1)$. sorként. Ekkor nyilván $\mathrm{cl}(\mathbf{F}(M)) = \mathrm{cl}(\mathbf{F}(M^*))$. A $\lambda_{t+1} = 0$, $\lambda_k \neq 0$ $(1 \leq k \leq t)$ paraméterválasztással a (2) egyenlet szerint előállított $p_\lambda \in \mathbf{F}(M^*)$ eloszlásra $\mathrm{Supp}(p_\lambda) = \overline{\mathrm{Supp}(v)}$. Ha $\mathbf{F}(M)$ zárt, akkor $p_\lambda \in \mathbf{F}(M)$, amiből a 2.2. Tétel szerint kapjuk, hogy $\mathrm{Supp}(p_\lambda)$ M-megvalósítható.

Jelölje az M mátrix sorait $v_1, \ldots, v_t$. Azt kell belátnunk, hogy $v = \sum_{k \in K} v_k$ alakú. Vezessük be a $T(S) = \cup_{b \in S} \mathrm{Supp}(m_b)$ jelölést. Mivel az $\overline{S}$ halmaz M-megvalósítható, minden $a \notin \overline{S}$-re $\mathrm{Supp}(m_a) \subsetneq T(\overline{S})$. Átfogalmazva, minden $a \in S$-re van olyan $k \in \mathrm{Supp}(m_a)$, melyre $k \in T(S) \setminus T(\overline{S})$. Másrészt, minden $k \in T(S) \setminus T(\overline{S})$-re teljesül, hogy $\mathrm{Supp}(v_k) \subset S = \mathrm{Supp}(v)$.

A $v = \mathbf{0}$ eset nem érdekes, hiszen akkor $v = \sum_{k \in \emptyset} v_k$. Minden más esetben $S \neq \emptyset$, azaz a fentiek szerint van olyan k, hogy $v - v_k$ is indikátorvektor, melynek kevesebb 1-es koordinátája van, mint v-nek. Innen indukcióval kapjuk, hogy v felbontható v^k-k összegére. (Sőt, vegyük észre, hogy az is elérhető, hogy mindig az első nem-nulla koordinátát nullázzuk ki.) ■

Ha tehát az $\mathcal{L}_1$ modellhez tartozó M_1 mátrix olyan, hogy $\mathbf{F}(M_1)$ zárt, akkor az összes V_1-beli indikátorvektor az M_1 mátrix valahány, diszjunkt tartójú sorának összege. Ha $\mathbf{F}(M_1)$ nem zárt, akkor pedig meg kell keresni M_1 maximális reprezentációját. Az is egy nyitott kérdés, hogy az általunk vizsgált mátrixok maximális reprezentációja mindig $0-1$ mátrix-e.

10.12. Példa. Korábban szóltunk már a duplán L-felbontható modell $M_B^{\max}$ maximális reprezentációjáról az $n = 4$ esetben. Ennek 24 sora bizonyos vastag kereszteknek felel meg. Nyilvánvaló, hogy minden vastag kereszt indikátorvektora felírható, mint az M_L mátrix néhány sorának összege. A 10.11. Tétel szerint ezzel a tulajdonsággal a maradék nyolc sor is rendelkezik. Említettük, hogy az egyik extra sor éppen a (37) halmaz komplementerének v indikátorvektora. A v vektor az M_L mátrix következő (x, C) párokkal indexelt sorainak összege:

$$(1, \{3\}), (4, \{2\}), (3, \{1{,}4\}), (1, \{2{,}3\}).$$

Természetesen a v indikátorvektorunkat az M_L' mátrix néhány sorának összegeként is fel kell tudni írni: ebben az esetben ugyanezeket a sorokat kell összeadni, mivel a (37) halmaz komplementere öninverz.

10.13. Példa. Legyen $n = 4$, vizsgáljuk meg az $\mathcal{M} = \mathbf{E}(M_L) \cap \mathbf{E}(M_L)\sigma$ modellt, ahol

$$\mathbf{E}(M_L)\sigma = \{p_{\circ\sigma} : p \in \mathbf{E}(M_L)\},$$

és $\sigma = (1324)$. Tehát két egyszerű hierarchikus modell metszetére vagyunk kíváncsiak. A (49) képlet szerinti, a közös durvításoknak megfelelő hierarchikus modell könnyen láthatóan a kvázi-független exponenciális családdal egyezik meg, melynek szabad paraméterszáma 9 (a hozzá tartozó altér dimenziója 10).

Ezzel szemben az $\mathcal{M}$ modell szabad paraméterszáma 12, a hozzá tartozó altér dimenziója pedig 13. Tehát a (49) egyenletben szigorú tartalmazás van.

A metszet altérben van indikátorvektorokból álló bázis, például a következő 65 vektor kifeszíti a teret. A vektorok között szerepel természetesen az **1**, a maradék 64 pedig négy darab 16 elemű osztályba sorolható. Ezek a vektorok a 10.11. Tételnek megfelelően mind az M_L mátrix néhány sorának összegeként állnak elő. Mind a 64 nulla-egy vektor M_L három sorának összege. A négy csoport leírásában $\{i, j, k, \ell\} = \{1, 2, 3, 4\}$ valamilyen sorrendben. Könnyen ellenőrizhető, hogy az alábbi négy osztály mindegyike 16 vektort definiál. Azt adjuk meg, hogy a metszetbeli indikátorvektor az M_L mátrix melyik (x, C) párokkal indexelt sorainak összege:

$$\begin{array}{lccc} 1: & (i,\{j\}) & (i,\{k\}) & (\ell,\{j,k\}) \\ 2: & (j,\{i\}) & (k,\{i\}) & (i,\{j,k\}) \\ 3: & (j,\{i\}) & (\ell,\{i,k\}) & (k,\{i,\ell\}) \\ 4: & (j,\{i\}) & (i,\{j,k\}) & (i,\{j,\ell\}) \end{array}$$

■

10.14. Példa. Legyen még mindig $n = 4$, most vizsgáljuk meg az $\mathcal{M} = \mathbf{E}(M_B) \cap \mathbf{E}(M_B)\sigma$ modellt, ahol

$$\mathbf{E}(M_B)\sigma = \{p_{\circ\sigma} : p \in \mathbf{E}(M_B)\},$$

és $\sigma = (1324)$. A $\mathcal{M}$ modell szabad paraméterszáma 11, a hozzá tartozó altér dimenziója pedig 12. A (49) képlet szerinti, a közös durvításoknak megfelelő hierarchikus modell viszont ismét a kvázi-függeten exponenciális család. Megnéztük, hogy a kvázi-független modell mátrixához hogyan lehet még két sort hozzáadni úgy, hogy a keletkező mátrix rangja 12 legyen. A következő két sor pl. megfelel:

$$v_1 = \rho_{24}^{33} + \rho_{14}^{22} + \rho_{11}^{32} + \rho_{25}^{22}, \quad v_2 = \rho_{24}^{33} + \rho_{14}^{22} + \rho_{11}^{32} + \rho_{11}^{23}.$$

Itt az M_B mátrix sorainak korábban bevezetett jelölését használtuk. Tehát ismét azt kaptuk, hogy a két hierarchikus modell metszetének van indikátorvektorokból álló bázisa. ■

10.5. Ismert eloszláscsaládok felbonthatósága

Ebben a szakaszban megvizsgáljuk, hogy azok az eloszláscsaládok, melyeket a gyakorlatban (és az irodalomban) gyakran illesztenek permutációadatokra, rendelkeznek-e valamilyen felbontható tulajdonsággal. Az itt szereplő modellek többségének L-felbonthatóságát már Critchlow et al. [15] is vizsgálta.

A rendezett minta modellek, illetve a Thurstone modellek általában nem L-felbonthatók. Fontos kivételt képez a Plackett-Luce modell, amely a sorbarendezési axióma szerint L-felbontható, de általános paraméterek mellett nem duplán L-felbontható, és nem is S-felbontható.

Ugyanez érvényes a Babington Smith modellre, az L-felbonthatóság a következő felírásból látszik:

$$p(\pi) = c(\theta) \prod_{i=1}^{n-1} \prod_{y \notin \pi\{1..i\}} \theta_{\pi(i)y}.$$

Viszont ennek speciális esete, a Mallows-Bradley-Terry modell, kvázi-független, így a jelen értekezésben tárgyalt összes felbonthatósággal rendelkezik.

A többlépcsős helyezési modell (THM) és az ismételt beszúrások modellje (IBM) az eddigiekkel szemben duplán L-felbontható. A két modellben egy-

egy L-felbontás:

$$\begin{aligned} \text{THM:} \quad & \Lambda(x, C) = \theta(|\overline{C} \cap \{1..x\}|, |C| + 1), \\ \text{IBM:} \quad & \Lambda(x, C) = \theta(|(C \cup x) \cap \{1..x\}|, x). \end{aligned}$$

A duplán L-felbonthatóság pedig a

$$\begin{aligned} \text{THM:} \quad & \log(p) = \textstyle\sum_{k,\ell,a} (\log \theta(n + 1 - a, k)) \rho_{a5}^{k\ell}, \\ \text{IBM:} \quad & \log(p) = \textstyle\sum_{k,\ell,a} (\log \theta(a, \ell)) \rho_{a5}^{k\ell} \end{aligned}$$

felírásokból következik.

Térjünk rá a távolságalapú modellekre! Vegyük először észre, hogy ha a távolság $d(\pi, \rho) = \sum_{k=1}^{n} c_k(\pi(k), \rho(k))$ alakba írható, akkor a megfelelő távolságalapú modell kvázi-független. Ez az 1. táblázatból a Hamming, p, és maximum távolságokra teljesül.

A maradék három esetben először azt szeretnénk vizsgálni, hogy a sorrendek eloszlása L-felbontható-e, ezért az eredeti definícióval összhangban most legyen a π *sorrend* valószínűsége

$$p(\pi) = K(\theta) e^{-\theta d(\pi^{-1}, \pi_0^{-1})}, \quad \pi \in S_n. \tag{50}$$

A $\theta = 0$ paraméter mellett persze minden távolságalapú modell az egyenletes eloszlást adja, ami minden felbonthatósággal rendelkezik. Az viszont, hogy egy $\theta \neq 0$ paraméter mellett az (50) eloszlás L-felbontható-e, csak a távolságtól függ, θ és π_0 értékétől nem. Érvényes a következő tétel.

10.15. Tétel. (Critchlow et al. [15]) *A sorrendeken megadott* (50) *eloszlás akkor és csak akkor L-felbontható, ha a d távolság additíven felbontható (additively decomposable), azaz minden $2 \leq k \leq n$-re léteznek olyan f_k és g_k függvények, melyekkel*

$$d(\pi, id) = f_k(\pi(1..k-1)) + g_k(\pi(k..n)).$$

Ebből következik, hogy a Kendall távolsághoz tartozó modell is L-felbontható, míg a Cayley és Ulam távolságokhoz tartozó modellek nem. Megjegyez-

zük, hogy a tétel egyszerűen levezethető az L-felbontható modell Markov bázisának ismeretéből.

A Cayley-, illetve Ulam távolságon alapuló modell akkor invertálva L-felbontható adott π_0-ra, ha

$$T(\pi_{11}\pi_0) + T(\pi_{22}\pi_0) = T(\pi_{12}\pi_0) + T(\pi_{21}\pi_0)$$

teljesül minden keresztező π_{11}, π_{22} permutáció-párra, ahol T az első esetben a ciklusok száma, a második esetben pedig a leghosszabb monoton növő részsorozat hossza. A Cayley távolság esetében, mivel balról is invariáns, feltehető, hogy $\pi_0 = id$. Ekkor, ha $n = 4$, akkor a $\pi_{11} = (3412), \pi_{22} = (4321)$ választással a bal oldal 4, a jobb oldal 2 lesz, és ezt az ellenpéldát tetszőleges $n > 4$-re ki lehet terjeszteni. Az Ulam távolság nem balról invariáns, ezért az invertálva L-felbonthatóság függhet a π_0 választásától. Egy olyan π_0-t sem találtunk, amely mellett a modell rendelkezne a mondott tulajdonsággal. A Kendall távolság sem balról invariáns, tehát ugyanaz vonatkozik rá, mint az Ulam távolságra. Ebben az esetben azt találtuk, hogy a modell olyan π_0 permutációkra lesz invertálva L-felbontható, melyekre a $\pi_0^{-1}\{1..k\}$ halmaz minden $1 \leq k \leq n$-re intervallum, vagy π_0 egy ilyen permutációból a $\sigma_r, \sigma_{(12)}$ elemekkel való balról szorzásokkal adódik.

A szakasz lezárásaként bemutatunk egy olyan távolságot, melyhez tartozó modell S-felbontható. Legyen $\pi, \sigma \in S_n$ két sorrend-vektor. Legyen

$$V_i(\pi, \sigma) = |\pi\{1..i\} \setminus \sigma\{1..i\}| = |\sigma\{1..i\} \setminus \pi\{1..i\}|$$

azon jelöltek száma, akik π-ben az első i hely egyikén vannak, de σ-ban nem. Könnyű látni, hogy $d_1(\pi^{-1}, \sigma^{-1}) = 1/2 \cdot \sum_{i=1}^{n} V_i(\pi, \sigma)$. Általánosabban, válasszunk θ_i pozitív paramétereket, és legyen

$$d_\theta^V(\pi, \sigma) = \sum_{i=1}^{n} \theta_i V_i(\pi, \sigma).$$

Ez balról invariáns távolság S_n-en (teljesíti a háromszög-egyenlőtlenséget),

és definiálható vele a

$$p(\pi) = c(\theta)e^{-d_\theta^V(\pi,\pi_0)}$$

távolság-alapú modell a π sorrendekre. Ez a modell nyilván S-felbontható, sőt, pl. $\pi_0 = id$ esetén duplán S-felbontható.

11. APA adatsor elemzése

Ebben a szakaszban egy "híres" adatsort vizsgálunk felbonthatóság szempontjából. Az adatok az Amerikai Pszichológiai Társaság (American Psychological Association, APA) 1980-as elnökválasztásához kapcsolódnak. Az APA tagjainak minden évben öt elnökjelölt sorbaállításával kell szavazniuk. 1980-ban körülbelül 15 ezren szavaztak, de ebből csak 5738 szavazó rangsorolta mind az öt jelöltet (a többiek csak az általuk legjobban preferált egy, két vagy három jelöltet nevezték meg). Az érdekesség kedvéért megjegyezzük, hogy az APA a Hare-rendszer szerint jelöli ki a győztest. Ha valamelyik jelölt a szavazók több, mint felétől első helyezést kap, akkor nyer. Ha nincs ilyen jelölt, akkor a legkevesebb első helyezést kapott jelöltet törlik. A törölt jelöltre leadott első helyezéseket szétosztják a maradék jelöltek között, mégpedig aszerint, hogy a törölt jelöltet első helyre rangsoroló szavazók kit tettek a második helyre. Ha valahány jelöltet már töröltek, akkor minden szavazó rangsorából az első nem törölt jelöltet veszik figyelembe. Az eljárást addig folytatják, amíg meg nem találják a győztest. Ez egyike a számos szavazatszámlálási rendszernek, melyeknek komoly elmélete van. A Hare rendszer előnyeit és hátrányait pl. Fishburn [34] elemzi.

Az APA adatsor egyike a legtöbbet vizsgáltaknak, lásd például [12, 29, 48, 49, 54]. Érdekessége, hogy viszonylag nagy az elemszáma az 5!-hoz képest, így nem könnyű rá jól illeszkedő modellt találni. 1980-ban az 5 jelölt ABC sorrendben: W. Bevan, I. Iscoe, C. Kiesler, M. Siegle, és L. Wright volt. Máris egy olyan számozást rögzítünk, amely mellett a felbontható eloszláscsaládok legjobban illeszkednek:

$$1 - \text{Bevan},\ 2 - \text{Kiesler},\ 3 - \text{Siegle},\ 4 - \text{Iscoe},\ 5 - \text{Wright}.$$

8. táblázat. APA elnökválasztás: az egyes sorrendek gyakorisága

12345	45	24513	52	43125	42	12354	50	24531	35	43152	34
12435	102	25341	43	43512	45	12453	95	25314	38	43521	46
12543	70	25431	34	41325	26	12534	70	25413	38	41352	42
13245	24	25143	87	41235	40	13254	17	25134	45	41253	30
13425	35	32145	28	41523	24	13452	28	32154	27	41532	36
13542	48	32415	35	45312	50	13524	35	32451	22	45321	50
14325	27	32541	24	45132	30	14352	35	32514	24	45123	40
14235	30	31245	28	45213	31	14253	28	31254	21	45231	25
14523	29	31425	52	52341	37	14532	34	31452	52	52314	34
15342	52	31542	53	52431	22	15324	40	31524	34	52413	30
15432	35	34125	75	52143	62	15423	37	34152	64	52134	41
15243	51	34215	51	53241	29	15234	36	34251	24	53214	31
21345	96	34521	44	53421	57	21354	79	34512	66	53412	91
21435	172	35142	133	53142	107	21453	186	35124	61	53124	71
21543	162	35412	84	54321	54	21534	106	35421	49	54312	58
23145	35	35241	67	54231	26	23154	36	35214	54	54213	24
23415	35	42315	16	54123	34	23451	26	42351	29	54132	41
23541	28	42135	34	51342	63	23514	30	42153	30	51324	35
24315	40	42513	22	51432	45	24351	50	42531	11	51423	53
24135	74	43215	23	51243	44	24153	82	43251	19	51234	40

Az adatok elemzése során több elemző is észrevette, hogy a szavazatok az APA megosztottságát tükrözik: éles választóvonal van a kutatók (1-es és 2-es jelölt) és a klinikusok (3-as és 5-ös jelölt) között, illetve egy harmadik, kisebb csoportot alkotnak a közösségi pszichológusok (4-es jelölt). Nyilván a különböző csoportok nagyon más szempontok alapján szavaznak, mindenki a saját jelöltjeit részesíti előnyben. A 8. táblázat tartalmazza a nyers adatokat: a jelöltek minden sorrendjére megadja, hogy hány szavazó választotta az adott sorrendet (a teljes szavazatot leadók közül).

Mielőtt a felbontható modelleket illesztenénk, tekintsük át, hogy néhány más modell hogyan illeszkedik az adatokra! Az eredményeket a jelöltek általunk rögzített számozásával közöljük. A nagy mintaelemszám és a szavazók heterogenitása miatt a legegyszerűbb, legfeljebb öt paramétert tartalmazó modellek nagyon rossz illeszkedést adnak. Ennek egyik megoldása az, hogy több ilyen egyszerű modell keverékével próbálkozunk, és az irodalomban találunk is ilyen megközelítést. Mi most mégis inkább azokra az elemzésekre koncentrálunk, ahol egy, de bonyolultabb modellt illesztettek a kutatók.

Marden [48] az adatokra log-lineáris modelleket illeszt. A kvázi-független

modell nem ad jó illeszkedést: az illeszkedést mérő, aszimptotikusan χ^2 eloszlású statisztika értéke $GOF = 973.8$, szabadsági foka $df = 103$. Itt

$$GOF = 2m \sum_{\pi \in S_n} r(\pi) \log \frac{r(\pi)}{\hat{p}(\pi)},$$

ahol m a minta elemszáma, r a tapasztalati eloszlás, $\hat{p}$ pedig a ML becslés. Marden ezután a kvázi-független modellhez egyesével vesz hozzá interakciós tagokat. Az általa legjobbnak ítélt modellben a π sorrend valószínűségére

$$\log p(\pi) = \sum_{k,i} \alpha_k^i \chi\{\pi(i) = k\} + \sum_{(k,\ell) \in A} \sum_{i,j} \alpha_{k\ell}^{ij} \chi\{\pi(i) = k, \pi(j) = \ell\},$$

ahol $A = \{(1,2), (1,4), (3,5), (4,5)\}$. Erre a modellre $GOF = 66.82$, a szabadsági fok pedig 59.

Chung és Marden [12] az ortogonális kontrasztokra épülő hierarchikus modellt illeszti az adatokra. Először a kontrasztokat választja ki, egyrészt előzetes elemzések, másrészt próbálkozás alapján. Emlékeztetünk, hogy minden kontraszt a jelöltek bizonyos részhalmazainak halmaza. A választott kontrasztok tehát: *I*= (4, 1235), *II*= (12,35), *III*= (1,2) és *IV*= (3,5), ezek éppen a kutató/klinikus/közösségi felbontásnak felelnek meg. Mivel a kontrasztok ortogonálisak, az egyes kontrasztok értékei tetszőlegesen kombinálhatók egymással, és minden kombináció egyértelműen meghatároz egy sorrendet. A kapott $5 \times 6 \times 2 \times 2$ méretű kontingencia-táblára a legjobban illeszkedő hierarchikus modell generátorai: {*I, II, III*} és {*I, II, IV*}, azaz ha tudjuk az *I, II* kontrasztok értékét, akkor a *III* és *IV* kontrasztok értékei feltételesen függetlenek. Az illeszkedést mérő GOF statisztika értéke 32.78, szabadsági foka 29.

McCullagh [49] az inverziós modelleket illesztette az APA adatokra. Az elsőrendű (Babington Smith) modellre $GOF = 1527.9$, $df = 109$ adódott, ami persze nagyon rossz illeszkedést mutat. A teljes másodrendű modell (melyben az összes első- és másodrendű inverzió szerepel) esetében $GOF = 246.5$ adódott, 89 szabadsági fokkal. A paraméterek becsült értékei alapján végül

McCullagh a következő modellt javasolta az 1,2,3,4,5 jelöltek π *helyezéseire*:

$$\log p(\pi) = c_{14}\chi\{\pi(1) < \pi(4)\} + \sum_{\{j,k\}} c_{\{j,k\}}|\pi(j) - \pi(k)|.$$

Az illeszkedés próbastatisztikája $GOF = 290.5$, szabadsági foka 109.

A fenti három modell közül az első kettő valóban jó illeszkedést ad az adatokra. Közös bennük, hogy a modell szerkezetét mindkét esetben próbálkozás útján választották ki a sok lehetséges szerkezet közül, ami általában jellemző a hierarchikus modellek illesztésénél. Mindkét modell meglehetősen sok paramétert használ, míg a harmadik esetben kevesebb a paraméter, de az illeszkedés is sokkal rosszabb. Illesszük most a sorrend-adatokra először a hat "teljes" felbontható modellt: L- és S-felbonthatóból is a simát, az invertáltat, és a duplát.

A sima felbonthatóságok invariánsak a balról szorzásokra (jelen esetben a jelöltek átszámozására), a jobbról szorzásokra (jelen esetben a helyezések átszámozására) viszont nem. A helyezések számozása nem csak címkézés, hanem egy valódi rendezést tükröz, az érdekesség kedvéért mégis a sorrendek valamennyi jobbról szorzott változatára illesztettük a modelleket (ez 15 eset az ekvivalenciaosztályok miatt). "Szerencsére" a helyezések eredeti számozása adta a legjobb illeszkedést.

Az invertálva felbonthatóságok esetében a helyzet éppen fordított: itt az illeszkedés jósága a jelöltek számozásától függ. Mivel a jelöltek számozása csak címkézés, ebben az esetben természetes megközelítés a számozások mind a 15 ekvivalenciaosztályát kipróbálni, és a legjobb illeszkedést adó számozást elfogadni. A legjobb eredményt a már említett számozás (illetve azzal ekvivalens másik hét) adta, mind az L-, mind az S-esetben. Vegyük észre, hogy az ekvivalens számozások is három csoportra osztják az öt jelöltet: $\{\{1,2\},\{3\},\{4,5\}\}$, ez a felosztás azonban csak részben egyezik meg a kutató/ klinikus/ közösségi felosztással.

A duplán felbontható esetekben mind a balról-, mind a jobbról szorzás érdekes, tehát $15 \cdot 15$ különböző illesztés végezhető el. Itt is, mind az L-, mind az S-esetben az előző két bekezdésben leírt számozások adták a legjobb eredményt. (A kvázi-független modell esetén, melynek illeszkedéséről már

9. táblázat. Felbontható eloszlások illeszkedése az APA adatokra

Modell	L	χ^2 (df)	u
L-felbontható	49	98.9 (70)	2.44
S-felbontható	72	144.8 (93)	3.80
invertálva L-felbontható	62	126.5 (70)	4.78
invertálva S-felbontható	85	171.7 (93)	5.77
duplán L-felbontható	75	151.8 (89)	4.71
duplán S-felbontható	89	180.1 (99)	5.76

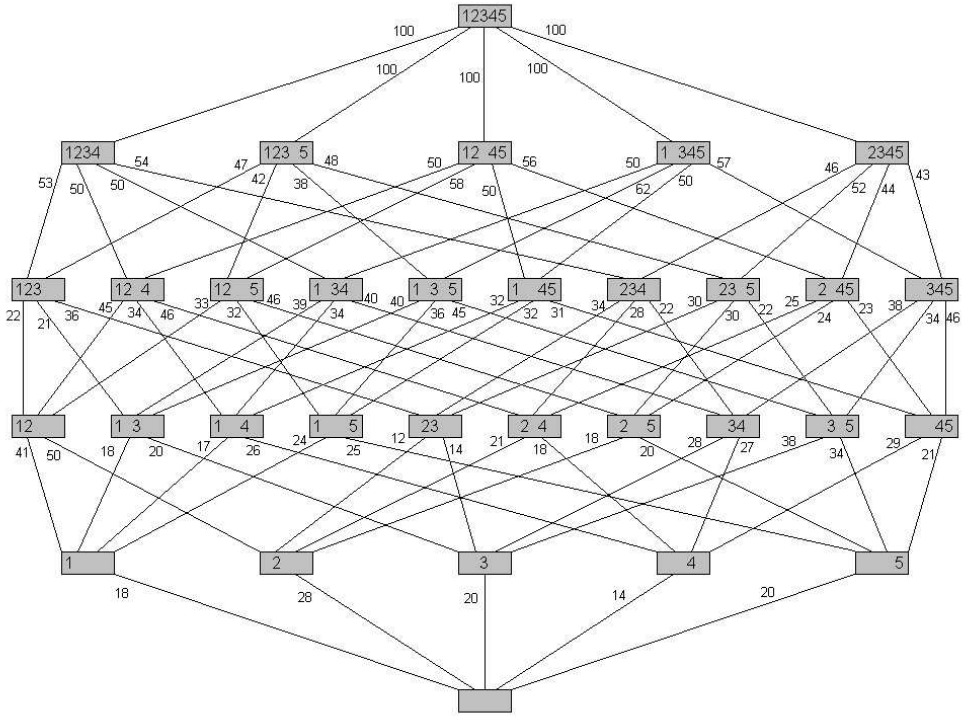
szóltunk, természetesen minden számozás ugyanazt az eredményt adja.)

Az illesztések eredményét a 9. táblázatban foglaltuk össze. Az eloszlás nemparaméteres maximum likelihood becslése a tapasztalati eloszlás (telített modell), e mellett a minta log-likelihoodja -26612. A jobb áttekinthetőség kedvéért a modellek log-likelihoodját ehhez az értékhez viszonyítva adjuk meg, azaz L jelöli, hogy az egyes modellek maximális log-likelihoodja menynyivel kisebb a telített modell értékénél. Megadjuk azután az illeszkedésvizsgálat χ^2-próbájának próbastatisztikáját (χ^2), ahol nem vontunk össze osztályokat, azaz 120 osztállyal dolgoztunk. Zárójelben szerepel a szabadsági fok, azaz 119 mínusz a becsült paraméterek száma. Az utolsó oszlopban a χ^2 statisztika standardizáltját, az $u = (\chi^2 - df)/\sqrt{2\,df}$ értéket tűntettük fel. Látható, hogy a vizsgált modellek közül az L-felbontható modell adta a legjobb illeszkedést, erre $p = 0.013$.

A modellek közül az L-felbontható illeszkedik a legjobban. Ebben az esetben a paraméterek ML becslését is megadjuk: a 6. ábra az S_5 gráfját mutatja, az élekre pedig a ML becslés szerinti $\Lambda(x, C) = P(\Pi(k+1) = x | \Pi\{1..k\} = C)$ feltételes valószínűségeket írtuk.

Ha szeretnénk jobban illeszkedő modellt találni, és nem bánjuk, ha ehhez több paramétert kell használnunk, illeszthetünk korlátozottan L-felbontható modellt. Ha például a $k = 2$,3-nál való felbonthatóságok közül csak a $k = 2$-t tesszük fel, akkor

$$p(\pi) = P(\Pi(1..2) = \pi(1..2))P(\Pi(3..5) = \pi(3..5) | \Pi\{1..2\} = \pi\{1..2\}).$$



6. ábra. Az L-felbontható ML becslés kanonikus paraméterei az APA adatra

Ennek a modellnek 69 szabad paramétere van. A becsléses illeszkedésvizsgálatot elvégezve kapjuk, hogy a χ^2 statisztika 57.98, szabadsági foka 50. Megjegyezzük, hogy ez a modell ekvivalens az első két és az utolsó három koordináta kvázi-függetlenségével. Ha pedig csak a $k = 3$ felbonthatóságot tesszük fel, akkor a χ^2 statisztika értéke 64.32.

Hivatkozások

[1] 4ti2 team: *4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces.* Available at www.4ti2.de.

[2] Agresti, A.: *Categorical Data Analysis.* Wiley, New York (1990).

[3] Babington Smith, B.: Discussion of Professor Ross's paper. *J. Roy. Statist. Soc. Ser. B* **12** (1950), 153-162.

[4] Block, H. W., Chhetry, D., Fang, Z. and Sampson, A. R.: Partial orderings on permutations. In: Block, H. W., Sampson, A. R. and Savits, T. H. (eds.): *Topics in Statistical Dependence.* IMS Lecture Notes–Monograph Series **16** (1990), 45-56.

[5] Block, H. W., Chhetry, D., Fang, Z. and Sampson, A. R.: Partial orders on permutations and dependence orderings on bivariate empirical distributions. *Ann. Statist.* **18** (1990), 1840-1850.

[6] Block, H. W., Chhetry, D., Fang, Z. and Sampson, A. R.: Metrics on permutations useful for positive dependence. *J. Statist. Plann. Inference* **62** (1997), 219-234.

[7] Borovkov, A. A.: *Matematikai Statisztika.* Typotex, Budapest (1999).

[8] Bökenholt, U.: Applications of Thurstonian models to ranking data. In: [36] 157-172.

[9] Bradley, R. A. and Terry, M. A.: Rank analysis of incomplete block designs, I. *Biometrika* **39** (1952), 324-345.

[10] Christensen, R.: *Log-linear models.* Springer-Verlag, New York (1990).

[11] Chung, L. and Marden, J. I.: Use of nonnull models for rank statistics in bivariate, two-sample, and analysis-of-variance problems. *J. Amer. Statist. Assoc.* **86** (1991), 188-200.

[12] Chung, L. and Marden, J. I.: Extensions of Mallows' ϕ model. In: [36] 108-139.

[13] Cox, D., Little, J., O'Shea, D.: *Ideals, Varieties, and Algorithms.* Springer, New York (1992).

[14] Cox, D. R. and Wermuth, N.: *Multivariate dependencies.* Chapman and Hall, London (1996).

[15] Critchlow, D. E., Fligner, M. A. and Verducci, J. S.: Probability models on rankings. *J. Math. Psych.* **35** (1991), 294-318.

[16] Croon, M.: Latent class models for the analysis of rankings. In: De Solte, G., Feger, H. and Klauer, K. C. (eds.): *New developments in Psychological choice modeling.* North-Holland, Amsterdam (1989), 99-121.

[17] Csiszár, I. and Tusnády, G.: Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplementary Issue No. 1 (1984), 205-237.

[18] Csiszár, V.: Conditional independence relations and log-linear models for random matchings. *Acta Math. Hungar.*, Online First (2008).

[19] Csiszár, V.: Markov bases of conditional independence models for permutations. *Kybernetika*, közlésre elfogadva.

[20] Csiszár, V.: On L-decomposability of random permutations. *J. Math. Psych.*, átdolgozás alatt.

[21] Csiszár, V.: An acyclic operation on the symmetric group, benyújtva.

[22] Csiszár, V., Rejtő, L. and Tusnády, G.: Statistical Inference on Random Structures. In: *Horizons of Combinatorics*, Bolyai Society Mathematical Studies **17**, Springer (2008), 37-66.

[23] Daniels, H. E.: Rank correlation and population models. *J. Roy. Statist. Soc. Ser. B* **12** (1950), 171-181.

[24] Darroch, J. N. and Ratcliff, D.: Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43** (1972), 1470-1480.

[25] Dawid, A. P. and Lauritzen, S. L.: Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** (1993), 1272-1317.

[26] Deming, W. E. and Stephan, F. F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11** (1940), 427-444.

[27] Dempster, A. P., Laird, N. and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** (1977), 1-38.

[28] Diaconis, P.: *Group Representations in Probability and Statistics.* Institute of Mathematical Statistics, Hayward, California (1988).

[29] Diaconis, P.: A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17** (1989), 949-979.

[30] Diaconis, P. and Eriksson, N.: Markov bases for noncommutative Fourier analysis of ranked data. *J. Symbolic Comput.* **41** (2006), 173-181.

[31] Diaconis, P. and Graham, R. L.: Spearman's foot rule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B* **39** (1977), 262-268.

[32] Diaconis, P. and Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** (1998), 363-397.

[33] Doignon, J.-P., Pekeč, A. and Regenwetter, M.: The repeated insertion model for rankings: missing link between two subset choice models. *Psychometrika* **69** (2004), 33-54.

[34] Fishburn, P.: *The Theory of Social Choice.* Princeton University Press, Princeton, N.J. (1973).

[35] Fligner, M. A. and Verducci, J. S.: Multi-stage ranking models. *J. Amer. Statist. Assoc.* **83** (1988), 892-901.

[36] Fligner, M. A. and Verducci, J. S. (eds.): *Probability Models and Statistical Analyses for Ranking Data.* Springer-Verlag, New York (1993).

[37] Geiger, D., Meek, C. and Sturmfels, B.: On the toric algebra of graphical models. *Ann. Statist.* **34** (2006), 1463-1492.

[38] Haberman, S. J.: *The analysis of frequency data.* University of Chicago Press, Chicago (1974).

[39] Huang, T.-K., Weng, R. C. and Lin, C.-J.: Generalized Bradley-Terry models and Multi-class probability estimates. *J. Mach. Learn. Res.* **4** (2006), 85-115.

[40] Hunter, D. R.: MM algorithms for generalized Bradley-Terry models. *Ann. Statist.* **32** (2004), 384-406.

[41] Hunter, D. R. and Lange, K.: Rejoinder to discussion of „Optimization transfer algorithms using surrogate objective functions." *J. Comput. Graph. Statist.* **9** (2000), 52-59.

[42] Lauritzen, S.: *Graphical Models.* Clarendon Press, Oxford (1996).

[43] Lehmann, E. L.: Some concepts of dependence. *Ann. Math. Statist* **37** (1966), 1137-1153.

[44] Luce, R. D.: *Individual choice behavior.* Wiley, New York (1959).

[45] Mallows, C. L.: Non-null ranking models, I. *Biometrika* **44** (1957), 114-130.

[46] Marden, J. I.: Use of orthogonal contrasts in analyzing rank data. *Technical Report* Dept. Statist., University of Illinois at Urbana-Champaign (1990).

[47] Marden, J. I.: Use of nested orthogonal contrasts in analyzing rank data. *J. Amer. Statist. Assoc.* **87** (1992), 307-318.

[48] Marden, J. I.: *Analyzing and Modelling Rank Data.* Chapman&Hall, London (1995).

[49] McCullagh, P.: Permutations and regression models. In: [36] 196-215.

[50] McLachlan G. J. and Krishnan, T.: *The EM Algorithm and Extensions.* John Wiley&Sons, New York (1997).

[51] Rao, P. V. and Kupper, L. L.: Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *J. Amer. Statist. Assoc.* **62** (1967), 194-204.

[52] Rapallo, F.: Toric statistical models: parametric and binomial representations. *Ann. Inst. Statist. Math.*, Online First (2006).

[53] Schriever, B. F.: An ordering for positive dependence. *Ann. Statist.* **15** (1987), 1208-1214.

[54] Stern, H.: Probability models on rankings and the electoral process. In: [36] 173-195.

[55] Sturmfels, B.: *Gröbner bases and convex polytopes.* Amer. Math. Soc., Providence, RI (1996).

[56] Sullivant, S.: *Toric Ideals in Algebraic Statistics.* PhD thesis, University of California, Berkeley (2005).

[57] Takemura, A. and Aoki, S.: Some characterizations of minimal Markov basis for sampling from discrete conditional distributions. *Ann. Inst. Statist. Math.* **56** (2004), 1-17.

[58] Thompson, G. L.: Generalized permutation polytopes and exploratory graphical methods for ranked data. *Ann. Statist.* **21** (1993), 1401-1430.

[59] Thurstone, L. L.: A law of comparative judgement. *Psychological Reviews* **34** (1927), 273-286.

[60] Whittaker, J.: *Graphical models in applied multivariate statistics.* John Wiley and Sons, Chichester (1990).

[61] Wu, C. F. J.: On the convergence properties of the EM algorithm. *Ann. Statist.* **11** (1983), 95-103.

[62] Yemelichev, V. A., Kovalev, M. M. and Kravtsov, M. K.: *Polytopes, Graphs and Optimisation.* Cambridge University Press (1984).

Összefoglalás

Az értekezés a véletlen permutációk statisztikai vizsgálatának néhány kérdésével foglalkozik. Az előkészületek után a második részben bemutattuk a sorbarendezési adatokra az irodalomban leggyakrabban használt modelleket. McCullagh inverziókon alapuló torikus modelljében bizonyítottuk a paraméterek identifikálhatóságát, ami azon az észrevételen múlt, hogy a helyre rakó, úgynevezett H-lépés körmentes gráfot indukál a szimmetrikus csoporton. Ezután a Plackett-Luce-féle modellek paramétereinek maximum likelihood becslésére vezettünk le EM algoritmusokat.

A harmadik rész a feltételes függetlenség és a hierarchikus modellek kapcsolatát vizsgálja a permutációs felállásban. Megmutattuk, hogy az L-felbontható eloszlások családja több szempontból szépen viselkedik: zárt hierarchikus modell, melyben a maximum likelihood becslés explicit kiszámolható, és a minta feltételes eloszlása az elégséges statisztikára nézve egyaránt generálható direkt módszerrel, illetve egy egyszerű, másodfokú Markov bázis segítségével.

Ezután azokat az úgynevezett duplán L-felbontható véletlen permutációkat vizsgáltuk, melyek nem csak L-felbonthatók, hanem az inverzük is az. Megmutattuk, hogy szigorúan pozitív esetben a szabad paraméterek száma n^3 nagyságrendű, ahol n a permutáció hossza. Ezek a vizsgálatok vezettek minket a véletlen permutációk hierarchikus modelljeinek bevezetésére, melyek olyan torikus modellek, melyeket az $\{1, \ldots, n\}$ halmaz partíció-párjai generálnak. Kiderült, hogy a duplán L-felbontható modell nem rendelkezik az L-felbontható modell „szép” tulajdonságaival. A Markov bázis meghatározása például rendkívül nehéz a gyors méretnövekedés miatt: az elérhető algoritmusok csak az $n = 4$ esettel tudtak megbirkózni. Még egy érdekes fogalmat, az S-felbonthatóságot is, bár rövidebben, de tanulmányoztuk.

Megkérdeztük, hogy melyek azok a véletlen permutációk, melyeknek bármely részvektora és annak komplementere feltételesen független, ha a részvektorokba eső elemek halmazát ismerjük. Megmutattuk, hogy ha minden permutáció valószínűsége pozitív, akkor ezek éppen a kvázi-független eloszlású permutációk. Végül egy ismert adatsorra illesztettük a felbontható modelleket.

Summary

This dissertation addresses a number of questions connected to the statistical analysis of random permutations. After the preliminaries, in Part II., we introduced the models for rank-ordered data most frequently used in the literature. We demonstrated that the parameters in McCullagh's inversion-based model are identifyable. This result followed from the fact that the graph on the symmetric group, induced by the so-called „move-home-step" is cycle free. Then we obtained EM algorithms for the maximum likelihood estimates of the parameters in the Plackett-Luce model and its generalizations.

Part III. deals with conditional independence and hierarchical models in the permutation setting. We showed that the L-decomposable family has some „nice" properties: namely, it is a closed hierarchical model in which the ML estimate is explicite, and the conditional distribution of the sample, given the sufficient statistics, can be generated either directly, or by means of a simple, degree-2 Markov basis.

We turned our attention to the so-called bi-L-decomposable random permutations, whose distribution, as well as that of their inverse, is L-decomposable. We calculated the number of parameters to be (in the strictly positive case) of order n^3, where n denotes the length of the permutation. These investigations led naturally to the definition of hierarchical models, which are generated by partition-pairs of the set $\{1, \ldots, n\}$. It turned out that the bi-L-decomposable model possesses none of the „nice" properties of the L-decomposable model. For example, the calculation of a Markov basis becomes quickly infeasible due to the size of the problem: the available algorithms could solve only the case $n = 4$. We briefly looked at another interesting property called S-decomposability.

We asked which random permutations had the property that every subvector and its complement are conditionally independent, given the set of elements in each subvector. We proved that if every permutation has positive probability, then these random permutations are exactly the quasi-independent ones. Finally, we fit our decomposable models to a real dataset.