SZTOCHASZTIKUS SZÁMÍTÁSTECHNIKA

Tusnády Gábor

Debrecen, KLTE
1996 - 1997

Készült a Művelődési és Közoktatási Minisztérium támogatásával
a 179/1995 számú kutatási program keretében

Tartalom

Előszó 5
1. Adatstruktúrák 7
2. Randomizálas 13
3. Eloszlások 21
4. Hipotézisek vizsgálata 28
5. Statisztikai becslések 32
6. Nem paraméteres módszerek 37
7. Markov láncok 43
8. Idősorok 50
9. Matematikai genetika 57
10. Sztochasztikus kapcsolatok 68
11. Sztochasztikus kontroll 78
12. Sztochasztikus automaták 86
13. Sztochasztikus mezők 94
14. Sztochasztikus optimalizálás 97
15. Fraktálok 98
16. Káosz 99
Kulcs 99
Áttekintő tábla 105
AS Algoritmusok 105
Tárgymutató 106

.

Előszó

A sztochasztikus számítástechnika a matematikának aránylag fiatal ága, és napjainkban nagyon gyorsan fejlődik. Két jól elkülöníthető területe van.

(A) A sztochasztika komolyabb számítástechnikai problémát jelentő feladatai: ezeket a sztochasztika elmélete folyamatosan termeli, miközben a számítástechnika fejlődő lehetőségei révén lépésről lépésre megoldások keletkeznek sokszor olyan feladatokra is, amelyeket korábban megközelíthetetlennek hittek. Külön említhetőek azok az elméleti feladatok, amelyeket jelenleg kizárólag számítástechnikai eszközökkel lehet megoldani.

(B) A számítástechnika azon feladatai, amelyek függetlenek a sztochasztikától, megoldásukra azonban használhatunk sztochasztikus eszközöket. Itt elsősorban a véletlen keresésekre, optimalizálásokra gondolok, de említhető például a mátrixjátékok nyeregpont tétele is.

A közölt feladatok önálló munkához adnak anyagot. Céljuk számítógépes kísérletezésre buzdítás akkor is, ha ez explicite nincs a szövegükben. És megfordítva: a "papiron ceruzával" történő munka akkor is hasznos lehet hozzájuk, ha a feladat első pillanatra mechanikusnak tűnik. Többségüket megoldottam abban a laza értelemben, hogy foglalkoztam velük. De van, amelyikről csak képzelem, hogy meg tudnám oldani, és olyan is van, amelyiknek az elméleti megoldásával hosszú ideje másokkal együtt sikertelenül próbálkozom. A feladatok egy részére a nemzetközi forgalomban beszerezhető programok léteznek. Néhány feladatot átöltöztettem, nem azért, hogy ne lehessen felismerni, hanem hogy a munkát könyebb legyen elkezdeni. Másoknak csak a címét adom meg, ezekről egyszerűbbnek tartom, hogy személyesen beszélgessünk, vagy aki meg akarja oldani, az irodalomban keresse ki őket. Minden feladatra érvényes, hogy az a kör, amit a megoldás jelent, laza, nem formális, és csak bizonyos kötetlen "baráti" egyetértés esetén remélhető, hogy világos, egyáltalán mit is várok a megoldásban. Általában csak annyit mondhatok, hogy mindig "értelmes" feladatot kivántam megfogalmazni, sose listák mechanikus előállítását. Azt várom, hogy aki megold egy feladatot, maga értelmezze az eredményt, ne tőlem várja ezt. Ha kifejezetten nem kérdezem is, mindig érdemes meggondolni, hogy a tervezett algoritmus elég effektív-e. A statisztika nagy adatmezőkkel dolgozik, sok feladat megoldása valóban reménytelenül hosszú futásokat igényel. Néha lehet találni egzakt, vagy heurisztikus "rövidítést", ezeket mindig

érdemes keresni. És ha elméletileg tisztáható, hogy ilyenek nincsenek, azt is jó tudni.

Az irodalom nagyobb része az általam írt összefoglalásokban található meg, ezeket a listákat itt takarékosságból nem ismétlem meg. A külön említett anyag elsősorban azért szerepel itt, mert az összefoglalások megírása után került a kezembe. Kapcsolatokat többnyire nem szervezek: a feladatok egymással is, a közölt tematikával is, és az irodalommal is sokszor nagyon szoros kapcsolatban állnak (például az Nkv feladat megoldása explicite megtalálható a Tusnády-Czeizel-Telegdi(1981) cikkben). Van egy folyóirat, Journal of the Royal Statistical Society, Section C, Applied Statistics, ebben önálló, lekódolt statisztikai algoritmusokat közölnek hosszú ideje. Néhányat közülük külön kiemelek ennek a tematikának a legvégén, hogy a tájékozódást megkönnyítsem. A feladatok hárombetűs kódjai az azonosításukat kívánják szolgálni, ezeket a tematika végén kétféle összesítésben is megadom. Itt derül ki az is, hogy ezek a kódok mit rövidítenek.

1. Adatstruktúrák

Témák: a statisztikai munka felépítése, kérdőívek és kísérletek tervezése, leíró statisztikák, tájékozódás többdimenziós adatmezőkben, többdimenziós skálázás, relációs adatbázisok, expert systems, hashing.

A statisztikai munka felépítése A statisztikus sztochasztikus adatmezőket értékel ki: meghatározza a véletlen szerepét az adatok kialakulásában. Szerencsés esetben a statisztikus is részt vesz az adatmező létrehozásában, ez estben az is gyakran előfordul, hogy a véletlent részben a statisztikus generálja. A klasszikus statisztikának nagyon sok aránylag automatikus eljárása van, ezek kész programcsomagban is megtalálhatóak, mégis célszerű, ha a statisztikus megtanulja azt a területet, ahonnan az adatok fakadnak. Egy-egy munkában nagyon fontos világosan megfogalmazni a célt, aminek az érdekében azt a konkrét munkát megtervezték. A statisztikus összefüggéseket keres az adatok között: helyes, ha tudja, melyek az ismert összefüggések, és melyek azok, amelyeket az adatokból kell kiolvasni. Az egyértelműen a statisztikus feladata, hogy eldöntse, ki lehet-e azokat az eredményeket olvasni az adatokból. Jó, ha azt is ellenőrizni tudja, hogy az ismertnek vélt adatok valóban ismertek lehetnek-e. Különösen az együttműködések kezdeti szakaszában tipikus, hogy a felhasználó olyan ismeretekre is hivatkozik, amiket korábbi statisztikai munka hozott létre, de nem értvén a statisztikát, pontatlanul idézi őket.

A munka első fázisában ellenőrizni kell az adatokat, meg kell határozni a megbízhatóságukat, és fel kell térképezni a struktúrájukat. Ha az ember saját programmal dolgozik, ebben a fázisban kell egyáltalán először elovasni az adatokat, és szükség esetén átkódolni azokat. Meg kell határozni a kilógó adatokat, a hiányzó adatokat, és el kell dönteni, hogy az egyes adatok milyen típusúak. A statisztikus adatokat elemez: ahogy egy tájról, szinházi előadásról, politikai mozgalomról egymásra épülő fogalmak segítségével számolunk be azoknak, akik azt közvetlenül nem ismerhették meg, a statisztikus szavakká transzformálja az adatokat, leírja azokat, feltárja a belső összefüggéseiket, és meghatározza azoknak a következményeknek a körét, amiket az adatokból ki lehet olvasni. Ennek előfeltétele, hogy ő maga képes legyen adatokat olvasni. Minden összesítés, származtatott szám "statisztika", és ha jól választjuk meg, többet mond a nyers adatokról mint ami közvetlenül látható bennük.

A leggyakoribb kérdés egy adatmezővel kapcsolatban az, hogy hogyan keletkezett. Az adatok kialakulásának vannak determinisztikus és vannak sztochasztikus mozzanatai. Az utóbbiak véletlen számokkal helyettesíthetőek, a minta ilyen másolatai

a számítástechnika eszközeivel tetszés szerinti példányszámban előállíthatóak. Ezzel a technikával ellenőrizhetjük, hogy helyes volt-e az adatok kialakulására vonatkozó elképzelésünk. Csak azt kell megnéznünk, hogy az új adatok hasonlítanak-e az eredetiekre. De a módszer ennél többet ad: az esetek többségében valamit ki tudunk ugyan olvasni az adatokból, de soha sem lehetünk egészen biztosak abban, jó-e az olvasatunk. Az általunk előállított adatok segítségével ezt is ellenőrizhetjük, hiszen tudjuk, egyáltalán mit írtunk be az adatokba. Egyúttal azt is meghatározhatjuk, mekkora az eljárásunk hibája.

A valódi statisztikai munka annak a sztochasztikus modellnek a létrehozása, amely az adatmezőt kialakította. Ennek során először magát a modellt kell felépíteni, majd ellenőrizni kell az alapfeltevéseit, végül meg kell határozni az ismeretlen paramétereit.

Egy modell akkor jó, ha segítségével a minta újrageneralható: azokon a pontokon, ahol véletlen szerepel benne a mintát kialakító véletlent új, mechanikusan generált véletlennel kell helyettesíteni. Az új mintán meg kell ismételni az eredeti eljárást, és ellenőrizni kell az eredmények stabilitását.

Kérdőívek és kísérletek tervezése Minden konkrét együttműködésben célszerű arra törekedni, hogy a közös munka ne akkor kezdődjön, amikor elkészült az adatmező, hanem a statisztikus vegyen részt már annak előkészítésében is. Itt általában a következő szempontok fontosak:

- az adatmező mérete: ezt az a két ellentétes hatás alakítja ki, hogy a kapacitás korlátai csökkenteni szeretnék, a megkívánt pontosság viszont növelni igyekszik;

- az adatmező struktúrája: nagyon gyakori, hogy az adatokban "dominó hatás" érvényesül, egyik adat meghatározza mások számát és szintaxisát;

- magának a vizsgálatnak a célja: a felhasználó ugyanis általában nem érzékeli, hogy az őt érdeklő kérdések megválaszolásához milyen jellegű adatokra van szükség, avagy egyáltalán lehet-e olyan adatokat előállítani, amelyek alapján a kérdéseire válaszolni lehet;

- az adatok tervezett feldolgozása: ha eleve adott az a statisztikai modell, amelyet el fogunk fogadni a munka során, azon belül optimalizálhatjuk az adatfelvételt a statisztikai effektívitás szerint - ez a szórásanalízisen belül külön ággá nőtt, ott kísérletek tervezésének nevezik.

Egy vizsgálat során a legfontosabb először azt a célt világosan megfogalmazni, amiért a vizsgálatot végezzük. Az ember mér, különböző mennyiségeket határoz

meg. De az esetek többségében a mérés nem közvetlen. A vizsgálat tárgya nem elérhető, mint a csillagászatban, nem megfogható, mint a részecskefizikában, még csak nem is definiálható egyértelműen, mint az emberi viselkedéssel foglalkozó tudományokban. A munka csak akkor lehet erdeményes, ha ismerjük a hatásmechanizmust, a mérés bemenet-kimenet dinamikáját, ha meg tudjuk mondani, hogy a nem mérhető mennyiségek hogyan hatnak a mérési eredményekre.

A méhető mennyiségekből aztán bizonyos nem mérhető mennyiségekre lehet következtetni, másokra nem. Ezeket a vizsgálat előtt élesen szét kell választani, és a rendelkezésre álló eszközöket úgy kell felhasználni, hogy az eredmény a lehető legpontosabb legyen. Ezt csak akkor tudjuk megtenni, ha ismerjük az adatok feldolgozásának a módját. Itt egy bizonyos mértékű körforgás kialakulhat: ez utóbbihoz ugyanis már jó lenne ismerni az adatokat. Gyakran érdemes több lépcsőre bontani a munkát, az adatok felvétele és értékelésük váltakozhat. Közben azt is a fokozatosan kezünkbe kerülő adatokból olvashatjuk ki, egyáltalán meddig tartson a vizsgálat.

Leíró statisztikák Az adatokkal való első ismerkedést az egyszerű leíró statisztikák szolgálják, amelyek az eredmények publikálása esetén a kölönböző irodalmi közlések összehasonlítását is megkönnyítik. Nagyon hasznosak a különböző grafikus megjelenítések (például a többdimenziós adatok emberi arcokkal való szemléltetése). Ezek segítik a hibák kiszűrését is, a kiugró adatok lokalizálását, és a hiányzó adatok elhelyezkedésének a feltérképezését. Néhány módszert említeni fogok én is (projection pursuit, ACE, beágyazások).

Tájékozódás többdimenziós adatmezőkben Egy komplex jelenséget sokfé adat együttese képes csak leírni, ezek sokszor jellegükben is eltérnek. A szokásos megkülönböztetés szerint vannak tulajdonságokra vonatkozó adatok, ezek valamilyen rendezett vagy rendezetlen halmaz elemei. Vannak továbbá intervallum jellegű adatok, amelyek a vizsgált mennyiségnek csak a határait jelölik ki, vannak diszkrét, többnyire egész értékű adatok, és vannak valós értékű adatok. Ez utóbbiak többnyire valamilyen skála függvényei amelynek a kezdőpontja és léptéke megváltoztatható. Ha egyformák, homogének egy jelenség adatai, tekinthetjük a jelenséget az euklideszi tér pontjának, az adatmezőt a tér véges ponthalmazának.

Ebben az esetben a legfontosabb kérdés az adatok együttesének térbeli elhelyezkedése. A pontok páronkénti távolsága, az egyes pontokból a legközelebbi pontba mutató élekből álló irányított gráf, a ponthalmaz konvex burkának a térfogata segíthet az első tájékozódásban. Itt is igaz, hogy effektívebb teljes eloszlásokat előállítani, mint egyes statisztikákat.

A klasszikus módszerek a normalitást tételezik fel, ami durva első közelítésben azt jelenti, hogy elliptikusnak képzeljük az adatmezőt. Ettől a leggyakrabba a következő két eltérést tapasztaljuk. Egyes adatok elszigetelődnek, "kiugróak", és a teljes adatmezőnek bonyolultabb, összetettebb az alakja. Sokszor nem töltik ki az adatok a rendelkezésükre álló teret, aminek az az oka, hogy túl sok, egymással szoros összefüggésben álló adatot határoztunk meg. A valódi, fügetlen komponensek megtalálása sokszor a feldolgozás végső célja is egyben.

Többdimenziós skálázás Az úgynevezett cluster analízis során merül fel, hogy különböző objektumok között távolságokat (vagy közelségeket, hasonlóságot) definiálunk. Ekkor is érdemes az objektumokat több dimenzióban, elsősorban síkon megjeleníteni. Azt szeretnénk, ha az új objektumok közti euklideszi távolságok valamilyen alkalmasan választott monoton függvénye jól közelítené az eredeti (általában nem euklideszi geometriából származó) távolságokat.

Relációs adatbázisok A számítástechnikai gyakorlatban ezen a kifejezésen egy bizonyos programrendszert értenek, amely elsősorban mátrixok kezelését szolgálja. Én abban az értelemben használom a kifejezést, hogy a statisztikai munka mindig azzal jár, hogy feltárjuk az adatmező belső összefüggéseit. Ezeket csak akkor tudjuk megtalálni, ha tudjuk, mit keresünk, és vannak eszközeink a kapcsolatok megkeresésére. Ezért mindig hasznos, ha az általános statisztikai programcsomagok mellet magunk által írt programokkal is dolgozunk, mert tapasztalatom szerint csak ezekkel lehet igazán megismerni egy adatmezőt. Sokszor tapasztaltam, hogy egy kezdő nem tud mit kezdeni az adatokkal, amikor felszólítom, hogy "nézegesse" az adatokat. Való igaz, hogy a mai méretek mellett az adatmezők közvetlen "olvasgatása" lehetetlen. Létre kell hozni a relációkat az adatmező különféle részei között, és ezekkel addig kell tömöríteni statisztikailag az adatokat, amíg már áttekinthető összesítésekre jutunk.

Expert systems Lásd: irodalom.

Hashing Lásd: irodalom.

Feladatok:

1.1. **(Hpr)** Legyen N pozitív egész. Keresendő az első N egész szám $P_1, \ldots, P_N$ permutációja, amelyre a következő mennyiség minimális. Vegyük a permutációk minden lehetséges N elemű ismétléses variációját. Ezek mindegyikében $k = 1, \ldots, N$

mellett határozzuk meg a k-adik permutációnak azt a legelső elemét, amelyiket korábban nem láttunk. A szóban forgó mennyiség a kezünkbe kerülő elemek száma összegezve először a permutációkra, majd az összes N^N lehetséges ismétléses variációra.

1.2. **(Rhs)** Ciklikusan elhelyezkedő urnákba egyenletes eloszlás szerint egymástól független golyókat helyezünk el a következő eljárás szerint. Az urnák adott körforgása szerint minden egyes golyót a beérkezéséhez legközelebbi üres urnába teszünk. Számoljuk össze a közben meglátogatott urnákat. Mit mondhatunk n urna esetében az m golyó elhelyezése során meglátogatott urnák számáról? (Legyen például $n = 1000$, $m = 950$.)

1.3. **(Kct)** Legyenek n, m pozitív egészek, és legyenek $\varepsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, m$, "szabályosan" véletlen és független ± 1-ek. Legyen $S_{kj} = \sum_{i=1}^{k} \varepsilon_{ij}$, és T_k legyen olyan valós számokból álló sorozat, melyre $\sum_{j=1}^{m} (\max_k \mid S_{kj} - T_k \mid)^2$ minimális. Határozzuk meg e minimum eloszlását.

1.4. **(Tsk)** Egy tetszőleges sokdimenzós pontrendszerhez keressünk alacsonyabb dimenziókban pontokat amelyek távolságainak monoton függvénye négyzetes középben jól közelíti az eredeti távolságokat.

1.5. **(Opt)** Egy sokváltozós függvény értékeit egy közelebbről nem részletezett eljárás adja. Feltéve, hogy a függvény "sima" (másodrendben jól közelíthető), adjunk általános eljárást a függvény lokális minimumának a meghatározására.

1.6. **(Rho)** Rádió-hullámhosszak optimális megválasztása. Rendezzük el az első 56 pozitív egészet egy 7 sorból és 8 oszlopból álló mátrixba úgy, hogy az egy sorban levő számok páronkénti különbségeinek az abszolut értékei különbözőek legyenek, és mindegyik különbség abszolut értéke legalább kettő legyen. Mennyi a különbségek abszolut értékeinek a megfelelő matrixokban fellépő maximumának a megfelelő mátrixokra vett minimuma?

1.7. **(Tac)** Adott téglalapokat rendezzünk el úgy, hogy az őket tartalmazó négyzet minimális legyen.

1.8. **(Mtn)** Adott négyzetek közül válasszunk ki diszjunktakat úgy, hogy az együttes területük maximális legyen.

1.9. **(Dmb)** Bontsuk dominókra egy kockás papir dominókra bontható részét.

1.10. **(Kör)** Döntsük el, hogy található-e egy adott irányított gráfban önmagába visszatérő út.

1.11. **(Ügm)** Legyen d tetszőleges pozitív egész. Adott a d-dimenziós térben véges sok pont. Mondjuk azt, hogy egy gömb a térben üres, ha nincs benne

adott pont, és mondjuk azt, hogy egy üres gömböt a pontok közrefognak, ha a gömb középpontja benne van a pontok konvex burkában. Keresendő a legnagyobb közrefogott üres gömb.

1.12. **(Mvg)** Legyen d tetszőleges pozitív egész. Adott a d-dimenziós térben véges sok pont. Tetszőleges sík mellett határozzuk meg a sík két oldalán a pontok konvex burkának a térfogatát, és e térfogatok összegével osszuk el a teljes konvex burok térfogatát. Határozzuk meg e hányadosok maximumát.

1.13. **(Tlt)** Legyen d tetszőleges pozitív egész. Adott a d-dimenziós térben véges sok pont. Határozzuk meg a következő mennyiséget. Írjunk minden egyes pont köré akkora sugarú gömböt amekkora a pontnak a hozzá legközelebbi ponttól mért távolságának a fele, és vegyük ennek a gömbnek a pontok konvex burkába eső részének a térfogatát. Adjuk össze ezeket a térfogatokat és az összeget osszuk el a konvex burok térfogatával. Milyen értékek között ingadozhat ez a szám?

1.14. **(Gfs)** Egy tetszőleges összefüggő gráfban határozzuk meg a következő mennyiséget. Minden egyes csúcshoz keressük meg azt a másik csúcsot, amelyik tőle a legmesszebb van, ha a távolságot a két csúcs között futó legkevesebb élből álló út éleinek a számával mérjük. Legyen ez a szám a csúcs "sugara". A keresett mennyiség e sugarak minimuma osztva a csúcsok számával. Milyen értékek között ingadozhat ez a szám?

1.15. **(Ptf)** Adottak a d-dimenziós térben a $P_1, \ldots, P_n, Q_1, \ldots, Q_n$ pontok. Keresendő az az egybevágóság a térben, amelyik a P_i pontokat olyan P_i' pontokba viszi, amelyekre $\sum_{i=1}^{n} \parallel P_i' - Q_i \parallel^2$ minimális.

1.16. **(Nsd)** Legfeljebb hány pontot lehet megadni a d-dimenziós tér egységkockájában, ha a pontok távolsága nem lehet s-nél kisebb?

1.17. **(Sza)** Készítsünk algoritmust a következő feladatra. Szavakat kapunk a felhasználótól, és a felhasználó megadja a szavak asszociációs rendszerét abban a formában, ahogyan mi azt kérjük. A program tetszés szerinti szóhoz meghatározza az adott szavaknak azt a sorrendjét, amelyben azok az asszociációs rendben távolodnak az adott szótól.

1.18. **(Jjb)** Jancsi és Juliska a számegyenesen bolyong. El lehet-e érni, hogy mind a ketten szabálkyos bolyongást végezzenek, és a találkozási idejük várható értéke véges legyen?

1.19. **(Fpl)** A fizika törvényei szerint pattogó labda.

Irodalom:

J.W. Tukey: Exploratory data analysis, Addison-Wesley, 1977

Lovász László-Gács Péter: Algoritmusok, Műszaki Könyvkiadó, 1978

Futó Péter-Frank Lajos(1979): Egy bűvös számtáblázat nyomában, Középiskolai Matematikai Lapok, 58, 56-60

M. Eigen-R. Winkler: A játék, Gondolat, 1981

Tusnády Gábor(1986): Hashing, Matematikai Lapok, 33/1-3, 143-148

P. Frankl: The shifting technique in extremal graph theory, in: Surveys in combinatorics, 1987, Ed. C. Whitehead, 81-110

J.A. Rice: Mathematical statistics and data analysis, Duxbury Press, 1988

G.R. Loftus-E.F. Loftus: Essence of statistics, A.A.Knopf, 1988

W.H. Press-B.P. Flannery-S.A. Teukolsky-W.T. Vetterling: Numerical recepies, The art of scientific computing, Cambridge University Press, 1988

R.J. Schalkoff: Artificial intelligence: An engineering approach, McGraw-Hill, 1990

L. Babai-P. Frankl: Linear algebra methods in combinatorics, Kézirat, 1992

Bognár Jánosné-Göndöcs Ferenc-Kászonyi László-Kováts Antal-Michaletzky György-Móri Tamas-Somogyi Árpád-Szeidl László-Székely J.Gábor: Matematikai statisztika, ELTE TTK, Nemzeti Tankönyvkiadó, 1995

2. Randomizálas

Témák: egyenletes eloszlású véletlen számok generálása, diszkrét eloszlások generálása, folytonos eloszlások generálása, véletlen permutációk, konvoluciók, titkosítás, a minták véletlen felújítása (resampling).

Egyenletes eloszlású véletlen számok generálása Egy diszkrét értékkészletű véletlen szám egyenletes eloszlású, ha az értékeit egyforma valószínűséggel veszi fel. A generált véletlen számoktól azt is meg akarjuk követelni, hogy az egymás után generált számok sorozatában az egyes elemek egymástól teljesen függetlenek legyenek, ami azt jelenti, hogy ha a generált számokból adott hosszúságú diszjunkt blokkokat formálunk, ez az új sorozat is egyenletes eloszlású legyen a blokkok hosszának tetszés szerinti megválasztása mellett. A számítógépeken kétféle véletlen generátor van: fizikai és algebrai. A fizikai generálás általában a gép órajelét használja, az algebraiak valamilyen rekurzív sorozat elemeit számolják. Ez utóbbiak közül a legegyszerűbb és egyben a legeffektívebb is az

$$X_k = a * X_{k-1} \quad mod(b),$$

$$Y_k = X_k \quad mod(c),$$

képletekkel generált Y sorozat, ahol b valamilyen nagy prim, és a, c b-nél kisebb számok. Ez a $(0, 1, \ldots, c-1)$ egészek felett generál az ideális véletlen számokat jól közelítő sorozatot, ha b nagy. (Szokás a különböző algoritmusokkal generált véletlen számokat pszeudo véletlennek nevezni.)

Noha a számítógépeken csak diszkrét számábrázolás van, beszélhetünk folytonos eloszlású véletlen számokról, a kettő közti átmenet biztosítását a konkrét gépi megvalósításra bízva. Azt mondjuk, hogy egy folytonos eloszlású valószínűségi változó az (a, b) intervallumban egyenletes eloszlású, ha

$$P(X < t) \quad = \quad \frac{t-a}{b-a}, \qquad \text{ha} a \quad a < t < b.$$

Általános szokás, hogy tisztán véletlenszerűnek azt mondjuk, ami egyenletes eloszlású: a lottószámok tisztán véletlenszerűek, mert az 1 és 90 közötti számokból képezett ötösök között egyenletes eloszlásúak, egy gráf tisztán véletlen, ha az éleket egyenletes eloszlás szerint és függetlenül választjuk, egy permutáció tisztán véletlenszerű, ha a generálásban az összes lehetséges sorrend egyformán valószínű. Elvileg ezek generálása ekvivalens feladat, a gyakorlatban azonban érdemes az adott helyzetnek megfelelő módszert választani, például a permutációk esetében effektívebb a $(0, 1)$-ben egyenletes eloszlású (és persze független) számokat nagyság szerint rendező permutációt használni. Más esetekben eleve az sem egyértelmű, hogy milyen halmazon tekintjük az egyenletességet, például mit értünk tisztán véletlen polinomon, tisztán véletlen mátrixon, vagy tisztán véletlen fán.

Vannak eljárások, amelyek segítségével "javítható" egy véletlen sorozat. Egy ezek közül a következő. Használjunk egy fix méretű puffert, ezt induláskor töltsük fel valahogy. Adjunk meg egy leképezést a generált számok értékkészletéről a puffer pozicióira úgy, hogy az eredmény lehetőleg egyenletes eloszlású legyen. Ezek után a generált sorozat elemeivel rendre "üssük" ki a puffer elemeit: ültessük be őket annak az elemnek a helyére, amelynek pozicióját hozzájuk rendeltük, és a transzformált sorozat soron következő eleme az a szám legyen, amelyik azon a pozición ült. Ezzel az eljárással javítható a függetlenség. Ha a generált számokból blokkokat képezünk, és a blokk elemeit összeadjuk, majd vesszük az összeg maradékát az értékkészlet modulusára, akkor az egyenletesség javul. Általánosan csak annyi mondható, hogy minden konkrét véletlen sorozat bizonyos célokra jó, másokra nem, és a konkrét felhasználás előtt érdemes egyrészt ellenőrizni, hogy arra alkalmas-e egyáltalán a sorozat, és ha lehet, növelni kell az alkalmasságát.

Az effektívitás érdekében a generálások általában egyszerűek. Szokás definiálni egy sorozat bonyolultságát a legrövidebb program hosszával, amely képes azt a

sorozatot előállítani.

DISZKRÉT ELOSZLÁSOK GENERÁLÁSA Legyenek a $(p_0, p_1, \ldots, p_n, \ldots)$ sorozat elemei nem negatívak, és legyen a sorozat elemeinek az összege 1. Ez a sorozat egy diszkrét eloszlást definiál, amelynek az elemeit a "dominó" szabállyal állíthatjuk elő a $(0, 1)$-ben egyenletes eloszlású véletlen számokból. Képzeletben minden k természetes számhoz p_k hosszúságú dominót rendelünk, és dominóinkat a 0-tól elindulva egymás után berakjuk a $(0, 1)$ intervallumba. (Mivel az összeg 1, éppen elérjük az 1-et.) A generálandó szám az a k lesz, amelyhez rendelt dominóra a soron levő véletlen szám esik. A megfelelő dominó kiválasztását érdemes megfelelően előkészített pointerek használatával gyorsítani. Az úgynevezett szűréssel visszavezethetjük egyik sorozat generálását egy másikra: ha $q_k = k\mathit{onstans} * p_k * \rho_k$, ahol $0 \leq \rho_k \leq 1$, és a k$\mathit{onstans}$ úgy van megválasztva, hogy a q_k-k összege is 1 legyen, akkor a q_k eloszlású sorozat elemei úgy kaphatóak a p_k eloszlású véletlen számokból, hogy azokat csak ρ_k valószínűséggel tartjuk meg. Ez persze rontja az effektívitást. Viszont kényelmes lehet, ha az első sorozat generálása már megfelelően elő van készítve. Külön haszon, hogy nem kell meghatározni a normáló konstans értékét.

A véletlen számokkal általában a nagy számok szoktak társulni valamilyen formában, ezért mindig gondolni kell a túlcsordulásra. Így van ez már a legegyszerűbb eloszlások, a binomiális, hipergeometrikus, negatív binomiális és Poisson eloszlás esetében is. Itt a faktoriális csordul általában túl, amit elkerülhetünk logaritmálással, de célszerűbb az eloszlásokat a maximális elemükből kiindulva a szomszédos elemek hányadosa alapján rekurzíven számolni.

FOLYTONOS ELOSZLÁSOK GENERÁLÁSA Univerzális, de nem mindig effektív módszer a kvantilis transzformáció: egyszerűen behelyettesítjük a $(0, 1)$-ben egyenletes eloszlású számot a generálandó eloszlás inverzébe. (A dominó módszer is ebből az eljárásból fakad.) Így például logaritmálással standard exponenciális eloszlású változót kapunk, amit aztán tetszés szerinti pozititív számmal átszorozva kívánt paraméterűvé transzformálhatunk. (Közben érdemes nem elfeledni, hogy a konstans szorzó a várható értékkel egyenlő, ami viszont a szokásos paraméterezés mellett a paraméter reciproka.) A normális eloszlásra már nem alkalmazható a módszer, de itt felhasználható az a tulajdonság, hogy a kétdimenziós standard normális eloszlás polárkoordinátás alakjában a szög és a rádiusz függetlenek, a szög egyenletes eloszlású, és a rádiusz négyzete 2 várható értékű exponenciális eloszlású véletlen szám. (Csak zárójelen belül jegyzem meg, hogy nekem ez az elvileg kristálytiszta módszer sosem működött - valószínűleg rossz volt az egyenletes

generátorom - ezért én is, mint mindenki, az IBM "rossz" normális generátorát használom: összeadok 12 darab $(0,1)$-ben egyenletes eloszlású véletlen számot, majd az összegből levonok 6-ot. Ha már az ember ilyen eljárást alkalmaz, lehet ezt is javítani: elvégezhetjük az összeadás előtt az egyenletes eloszlássú számokon az $(ln(x) - ln(1-x))$ transzformációt.)

Gamma eloszlást szűréssel állíthatunk elő exponenciális eloszlásból, ha az α alakparaméter legalább 1: ha az exponenciális eloszlású szám értéke x, akkor

$$(\frac{x}{\alpha})^{\alpha} e^{\alpha - x}$$

valószínűséggel szűrhetünk. A béta eloszlás előállítható a gammából: két független gamma közül az egyiket osztjuk a kettő összegével. Hasonlóan állítható elő F eloszlású véletlen szám két független, χ^2 eloszlású változó hányadosaként (a χ^2 eloszlás speciális gamma).

Véletlen permutációk Nem csak a valós számokon vagy vektorokon adhatunk meg eloszlásokat, hanem tetszés szerinti struktúrához értelmezhetünk valamilyen paraméterekkel módosítható véletlen mechanizmust, amely a struktúra elemeit generálja. Természetesen az a cél, hogy az általunk generált eloszlások közel tudjanak kerülni a gyakorlatban előforduló véletlen elemekhez. A permutációk a rangsoroláskor keletkeznek: gyakori, hogy anélkül, hogy a szóban forgó dolgokat valamilyen skálán el tudnánk helyezni, valahogy a sorrendjüket meg tudjuk határozni. Ennek egyik módja a páros összehasonlítás, a tapasztalat szerint ezt úgyanis a szakértők megbízhatóbban tudják megadni.

Rendeljünk az objektumokhoz valós számokat, legyen mondjuk az x objektumhoz rendelt szám $Q(x)$. Adjunk meg ezek alapján egy véletlen irányított gráfot az objektumokon úgy, hogy az x-ből y-ba futó élt

$$p(x,y) = \frac{e^{Q(x)-Q(y)}}{e^{Q(x)-Q(y)} + e^{Q(y)-Q(x)}}$$

valószínűséggel y felé, és $1-p(x,y) = p(y,x)$ valószínűséggel x felé irányítjuk. Ezek után keressük meg azt a permutációt, amelyik a legközelebb van a kapott irányított gráfhoz a távolságot a különbözően irányított párok számával mérve. (Ha több extremális permutáció van, azok között válasszunk tisztán véletlenszerűen.)

Egy másik lehetőség az, hogy eloszlásokat rendelünk az objektumokhoz, azokból generálunk független véletlen számokat, és az objektumokat a számuk szerint rakjuk sorba.

TITKOSÍTÁS A randomizálás egyik legfontosabb terűlete a titkosítás, annak biztosítása, hogy egy anyag csak azok kezébe kerülhessen, akiket illet. Szokás a zárak nyitását véletlen jelszóhoz kötni. Legyen S az érdekelt személyek halmaza, és $\mathcal{G}$ az S részhalmazainak a halmaza, vagyis legyen $\mathcal{G}$ S felett hipergráf. El szeretnénk érni, hogy ha S tagjai közül olyan sokan vannak együtt, hogy köztük $\mathcal{G}$ valamely elemének minden tagja megtalálható, akkor a csoport elő tudja állítani a jelszót, különben nem. Ennek érdekében keressünk olyan $(X_s, s \in S, Y)$ többdimenziós eloszlást, amelyben

$$H(X_s,\ s \in A, Y) = H(X_s,\ s \in A), \quad \text{vagy}$$

$$H(X_s,\ s \in A, Y) = H(X_s,\ s \in A) + H(Y)$$

aszerint, hogy S-nek a szóban forgó A részhalmazának $\mathcal{G}$ valamely eleme része-e, vagy sem (tartalmazásként az identitást is megengedve, H az entrópiát jelöli). Ez a követelmény csak akkor teljesíthető, ha ellentmondás-mentes: S minden A részhalmazáról eldönthető, hogy része-e $\mathcal{G}$ valamely elemének, vagy sem, vagyis $\mathcal{G}$ elemei között nem fordulhat elő olyan, amelyik valódi része $\mathcal{G}$ valamelyik másik elemének. (Azt feltehetjük, hogy $\mathcal{G}$ elemei különbözőek.) Ezt a tulajdonságot a továbbiakban mindig feltesszük $\mathcal{G}$-ről.

Feltehetjük, hogy Y véletlen bit, mondjuk ± 1 $\frac{1}{2}$ valószínűséggel. Ha $\mathcal{G}$-nek egyetlen eleme van, jó konstrukció a következő. Legyenek az X_s változók is szabályos (egyenletes eloszlású) és egymástól független véletlen ± 1-ek, és legyen Y értéke a $\mathcal{G}$ egyetlen elemében levő elemekhez rendelt előjelek szorzata. (Vegyük észre, hogy a $\mathcal{G}$-beli és Y-beli előjelek együttes eloszlása szimmetrikus.) Mit tegyünk, ha $\mathcal{G}$-ben egynél több elem van? Az első elemre alkalmazhatjuk a mondott eljárást, aztán a többire megfordíthatjuk a dolgot úgy, hogy az egyik elemet elhagyjuk, a többihez független és véletlen előjeleket rendelünk, majd az elhagyott elemhez ezek és Y szorzatát rendeljük. (Az egész eljárás során minden új véletlen előjel független a korábbitól.) Belátható, hogy ez az eljárás jó. Egy baj van vele: a generált eloszlás entrópiája nagy, ha $\mathcal{G}$-nek sok eleme van (pédául $\mathcal{G}$ tartalmazza S összes $\mid S \mid /2$ elemű részhalmazát, ahol $\mid S \mid$ S elemeinek a számát jelöli - legyen az mondjuk páros).

Nem tudjuk, hogyan lehet a szóba jöhető eloszlások között a minimális entrópiájút megtalálni. Ez a kérdés kapcsolódik egy másikhoz. Felejtsük el az Y-t, tekintsük az összes $X_s, s \in S$ többdimenziós eloszlást, és ezek mindegyikéhez rendeljük hozzá a $2^{|S|}$ dimenziós térnek azt a pontját, amelynek a koordinátái a $H(X_s, s \in$

A) entrópiák, ahol A befutja S részhalmazait (valahogy rendeljük ezeket az első $\mid S \mid$ természetes számhoz). Nem tudjuk, hogyan lehet a keletkező halmazt karakterizálni. Jelöljük $\mid S \mid = n$ mellett ezt a halmazt E_n-nel. Bármely $k < n$ mellett E_n-ből $\binom{n}{k}$-féképpen állítható elő vetítéssel E_k. Ez szükséges feltételek rendszerét adja. Nem tudjuk, melyek azok az n-ek, amelyekre ezek a feltételek egyben elégségesek is.

A MINTÁK VÉLETLEN FELÚJÍTÁSA (RESAMPLING) Egy statisztikai eljárás megbízhatóságáról képet kapunk abból, ha az egész eljárást a mintavételezéssel együtt többször, egymástól függetlenül azonos körülmények között megismételjük. Ha megtehetjük, ez a legjobb. Ez a szokás a mérésekkel is, így kapunk képet a mérés hibájáról. Régóta foglalkoztatja a statisztikusokat az a kérdés, hogyan lehetne "egy rókáról több bőrt lehúzni": hogyan lehetne ezt az ellenőrzést egyetlen minta alapján elvégezni? Triviális, de nem effektív megoldás a minta véletlenszerű kis részekre bontása. Ennél hatékonyabb, ha a módszer elemenként ellenőrizhető, az úgynevezett jackknife (bugylibicska) eljárás: minden egyes elemet rendre elhagyva az eljárást a maradékon végezzük el, és aztán az elhagyott elemen ellenőrizzük. Jól alkalmazható ez a regressziószámításban és a diszkriminamcia-analízisben. Tíz - tizenöt éve jött divatba az úgynevezett bootstrap (csizmahúzó - az a kis bőrdarab, amelynél fogva a csizma felhúzáskor megfogható) eljárás: az új mintát úgy kapjuk, hogy az eredeti minta elemei közül egyenletes eloszlás szerint kiveszünk egy-egy elemet, az egyes választások egymástől függetlenek, és a kiválasztott elemeket képzeletben mindig visszatesszük az eredeti mintába, tehát azok újraválaszthatóak. (A névnek kicsit az a - szándékolt - íze, hogy a statisztikus a saját hajánál fogva húzza ki magát a mocsárból: olyan eredményeket "facsar" ki a mintából, amilyenek azok abban benne sincsenek.)

FELADATOK:

2.1. **(Flf)** Egy este fiúk és lányok táncoltak. Az est jegyzőkönyve alapján összeállítandó olyan véletlen táncrend, amelyben mindenki pontosan annyiszor táncol mint eredetileg, de a párválasztás tisztán véletlenszerű.

2.2. **(Prp)** Hagyjuk el egy tetszés szerinti szövegben a szóközöket és az irásjeleket. A visszamaradó betűsorozatban határozzuk meg minden egyes rendezett betűpár előfordulásának a gyakoriságát. Hogyan lehetne a betűsorozatot véletlenszerűen átrendezni úgy, hogy ezek a páros gyakoriságok változatlanok maradjanak?

2.3. **(Eck)** Legyen n pozitív egész, $X_1, \ldots, X_n$ legyenek független, és az $(1, \ldots, n)$

egészeken egyenletes eloszlású valószínűségi változók. Határozzuk meg az

$$(i, X_i),\ i = 1, \ldots, n$$

élekből álló irányított gráfban a legnagyobb kör méretének az eloszlását.

2.4. **(Pck)** Legyen n pozitív egész, és $X_1, \ldots, X_n$ legyen egy (tiszta) véletlen permutáció. Határozzuk meg az $(i, X_i),\ i = 1, \ldots, n$ élekből álló irányított gráfban a legnagyobb kör méretének az eloszlását. Mi annak a valószínűsége, hogy ez a méret n?

2.5. **(Hdg)** Adjunk a Huffman kód alapján eljárást, amellyel tetszés szerinti diszkrét eloszlásból kevés lépésben lehet véletlen számot generálni.

2.6. **(Zwt)** Egy tetszőleges intervallumból kiindulva iteratíven alkalmazzuk a következő eljárást. Vegyünk az intervallumban egy egyenletes eloszlású véletlen töréspontot, és az általa létrehozott két darab közül vegyük a nagyobbat. Határozzuk meg a századik lépés után keletkező darab hosszának az eloszlását.

2.7. **(Lhú)** Legyenek n, m pozitív egészek, és legyenek $X_1, \ldots, X_m,\ Y_1, \ldots, Y_m$ független, az $(1, \ldots, n)$ egészeken egyenletes eloszlású valószínűségi változók. Határozzuk meg az $(X_i, Y_i),\ i = 1, \ldots, m$ élekből álló gráfban a leghosszabb út méretének az eloszlását.

2.8. **(Prt)** Legyenek n, m pozitív egészek, és legyenek $X_1, \ldots, X_n$ nem negatív egészek, amelyek összege m. Adjunk eljárást az $X_1, \ldots, X_n$ elemek (tisztán) véletlen előállítására.

2.9. **(Run)** Legyen n pozitív egész, $X_1, \ldots, X_n$ legyenek független, a $(0, 1)$ intervallumban egyenletes eloszlású valószínűségi változók. Mondjuk azt, hogy (a, b) monoton blokk, ha $X_i \leq X_{i+1}$ midőn $i = a, \ldots, b-1$. Legyen $(b - a + 1)$ az (a, b) blokk "hossza". Határozzuk meg a leghosszabb monoton blokk hosszának az eloszlását.

2.10. **(Bft)** Válasszunk egy adott körben véletlen húrt a következő eljárások szerint:

(A) a kör kerületén véletlenszerűen egymástól függetlenül választunk két pontot, ezek a húr végpontjai,

(B) a kör belsejében véletlenszerűen egymástól függetlenül választunk két pontot, ezek a húr pontjai,

(C) a kör belsejében véletlenszerűen választunk egy pontot, ez a húr felezőpontja.

Határozzuk meg mindhárom esetben a húr hosszának az eloszlását.

2.11. **(Vss)** Válasszuk egy háromszög csúcsait egy háromszög belsejében egyenletes eloszlás szerint, egymástól függetlenül. Határozzuk meg a háromszög köré

írható kör középpontja és a háromszög magasságpontja által meghatározott szakasz hosszának az eloszlását. Hogyan lehetne ezt a feladatot magasabb dimenzióra általánosítani?

2.12. **(Knv)** Írjunk eljárást a természetes számokon értelmezett eloszlások konvolúciójának effektív kiszámolására.

2.13. **(Luk)** Legyen n pozitív egész, és legyenek $(\varepsilon_i,\ i = 1,\dots,n)$ független szabályosan véletlen ± 1-ek. Legyen $S_t = \sum_{i=1}^{t} \varepsilon_i, M = \max_{1\le t\le n} S_t$, és $k = 1,\dots,M$ mellett legyen T_k értéke a legkisebb t index, melyre $S_t = k$. Legyen $T_0 = 0$, és $1 \le k \le M$ mellett legyen $U_k = T_k - T_{k-1}$, továbbá legyen $U_{M+1} = n - T_M$. Határozzuk meg az $(U_1,\dots,U_{M+1})$ mennyiségek maximumának az eloszlását. Hogyan lehetne ezt a kérdést magasabb dimenzióra általánosítani?

2.14. **(Etr)** Hogyan lehetne olyan véletlen téglalapokat generálni, amelyek területe egyenletes eloszlású?

2.15. **(Gms)** Legyenek m, n pozitív egészek. Dobjunk be m golyót véletlenszerűen n urnába. Határozzuk meg az egy urnába kerülő golyók számának a minimumát, majd e minimum eloszlását.

2.16. **(Mtv)** Adott a d-dimenziós térben n pont. Határozzuk meg lehetőleg kevés művelettel a köztük fellépő távolságok mininimumát.

2.17. **(Ákk)** Áttörhetetlen kulcs készítése.

Irodalom:

B.A Wichmann-I.D. Hill(1982): Algorithm AS 183, An efficient and portable pseudo-random number generator, Journal of the Royal Statistical Society, Section C, Applied Statistics, 31, 188-190

A.I. McLeod(1985): Reamark AS R58, A remark on Algorithm AS 183, An efficient and portable pseudo-random number generator, Journal of the Royal Statistical Society, Section C, Applied Statistics, 34, 198-200

L. Devroy: Non-uniform random variate generation, Springer, 1986

D.E. Knuth: A számítógépprogramozás művészete 2, Műszaki Könyvkiadó, 1987

I. Deák: Random number generators and simulation, Akadémiai Kiadó, 1990

H. Niederreiter: Random number generation and quasi-Monte Carlo methods, SIAM 1992

3. Eloszlások

Témák: többdimenziós normális eloszlás, lineáris regresszió, diszkriminancia analízis, hierarchikus modellek, szórásanalízis, monoton regresszió, görbevonalú

szórásanalízis.

TÖBBDIMENZIÓS NORMÁLIS ELOSZLÁS Azt mondjuk, hogy az X véletlen szám standard normális eloszlású, ha sűrűségfüggvénye

$$\varphi(x) \quad = \quad \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Azt mondjuk, hogy az X véletlen vektor standard normális eloszlású, ha koordinátái függetlenek és standard normális eloszlásúak. A d-dimenziós standard normális eloszlású vektor sűrűségfüggvénye tehát

$$f(x) \quad = \quad (2\pi)^{-d/2} e^{-\|x\|^2/2},$$

ahol most x d-dimenziós vektor. Mivel a forgatás mérték- és normatartó, a standard normális eloszlású vektor tetszés szerinti forgatottja is standard normális eloszlású. Ezért a standard normális eloszlású vektor egységvektora egyenletes eloszlású az egységgömbön, és független a vektor normájától. Ez utóbbi négyzete χ^2 eloszlású d szabadságfokkal. Ha $X_1, \ldots, X_d$ független d-dimenziós standard normális vektorok, és $Y_1, \ldots, Y_d$ az ortonormáltjuk (Schmidt módszer szerint, vagy optimális ortonormálás szerint, vagyis $Y_1, \ldots, Y_d$ minimalizálja a

$$\sum_{i=1}^{d} \| X_i - Y_i \|^2$$

mennyiséget), akkor $Y_1, \ldots, Y_d$ egyenletes eloszlású az ortonormált vektorok körében. A Schmidt ortonormálás során keletkező normanégyzetek mind χ^2 eloszlásúak egyesével csökkenő szabadságfokkal, és függetlenek egymástól. Szorzatuk az $X_1, \ldots, X_d$ vektorokól mint oszlopvektorokból összeállított mátrix determinánsának az abszolut értéke.

Azt mondjuk, hogy az n-dimenziós vektor normális eloszlású, ha

$$X \quad = \quad AY \; + \; b,$$

ahol Y d-dimenziós standard normális eloszlású véletlen vektor, és A $n \times d$ méretű determinisztikus mátrix, b n-dimenziós determinisztikus vektor. Ez az előállítás nem egyértelmű. Ha Q tetszőleges d-dimenziós forgatás, akkor A és Y helyett nyilván írhatunk AQ^T-t illetve QY-t, ahol T a transzponálás jele. Ettől eltekintve az előállítás egyértelemű, feltéve, hogy d minimális. Ekkor d egyenlő a

$$\Sigma \quad = \quad E(X-b)\,(X-b)^T \quad = \quad A\,A^T$$

mátrix rangjával. Itt $b = EX$ a várható érték, és Σ a kovariancia mátrix. A normális eloszlást ez a két mennyiség egyértelműen jellemzi, és tetszés szerinti véges második momentumú véletlen vektorhoz található vele megegyező váható értékű és kovarianca mátrixú normális eloszlású vektor. A centrális határeloszlás tétele szerint ez az illető vektorral egyező eloszlású és független vektorok normált átlagainak a határeloszlása.

Azt mondjuk, hogy az X, Y véletlen vektorok együttesen normális eloszlásúak, ha az összefűzésükkel keletkező

$$\begin{pmatrix} X \\ Y \end{pmatrix}$$

vektor normális eloszlású. A többdimenziós standard normális eloszlás forgatással szembeni invarianciája alapján belátható, hogy X és Y függetlenek, ha kovarianciájuk nulla. Tegyük fel, hogy X és Y várható értéke nulla, és válasszuk meg a B mátrixot úgy, hogy X terét Y terébe vigye, továbbá X és $(Y - BX)$ kovarianciája nulla legyen:

$$E\,(Y - BX)\,X^T \quad = \quad \Sigma_{YX} \quad - \quad B\,\Sigma_{XX} \quad = \quad 0,$$

ahol $\Sigma_{YX} = E\,Y\,X^T$, és $\Sigma_{XX} = E\,X\,X^T$. Ha ez utóbbi invertálható, akkor $B = \Sigma_{YX}\,\Sigma_{XX}^{-1}$. Ez Y-nak X-re vonatkozó regressziós együtthatója. A regresszió hibája

$$\Sigma_{Y|X} \quad = \quad \Sigma_{YY} \quad - \quad \Sigma_{YX}\,\Sigma_{XX}^{-1}\,\Sigma_{XY}$$

egyenlő X és Y együttes kovariancia mátrixa inverzének jobb alsó eleme inverzével, amint az a vektorok együttes sűrűségfüggvényéből is kiolvasható B fenti alakjával együtt.

Lineáris regresszió Legyen

$$y \quad = \quad X\,\beta \,+\, \sigma\,\varepsilon,$$

ahol y egy mérés n-dimenziós eredménye, X ismert $n \times k$ méretű mátrix, β ismeretlen k-dimenziós, σ ismeretlen skalár paraméter, és ε n-dimenziós standard normális eloszlású vektor. A β paraméter-vektort a legkisebb négyzetek elve alapján becsülhetjük: keressük azt a $\hat{\beta}$ k-dimenziós vektort, melyre

$$\|\, y \quad - \quad X\,\hat{\beta}\,\|^2$$

minimális. Ez az

$$X^T\,X\,\hat{\beta} \quad = \quad X^T\,y$$

úgynevezett normál egyenlet megoldása:

$$\hat{\beta} \quad = \quad (X^T X)^{-1} X^T y.$$

E becslés kovarianciamátrixa

$$\Sigma(\hat{\beta}) \quad = \quad (X^T X)^{-1} \sigma^2,$$

és itt σ^2 becslése

$$\hat{\sigma}^2 \quad = \quad \frac{\| y - X\hat{\beta} \|^2}{n-k}.$$

Ha az $(X \ y)$ mátrix önmagával képezett diadikus szorzatát D-vel jelöljük, a normál egyenlet megoldását megkapjuk a következő eljárással.

for t:=1 to k do begin

for j:=t+1 to m do d[t,j]:=d[t,j]/d[t,t];

for i:=1 to m do if(i<>t) then

for j:=t+1 to m do d[i,j]:=d[i,j]-d[i,t]*d[t,j]; end;

Itt m=k+1 és $\hat{\beta}$ D utolsó oszlopában keletkezik, ahol a legalsó elem egyenlő a $\hat{\sigma}^2$ fenti alakjában a számlálóban álló "maradék hibával".

Diszkriminancia analízis Lásd: irodalom.

Hierarchikus modellek Lásd: irodalom.

Szórásanalízis Legyen x egy d-dimenziós vektor, amelynek a koordinátái valamilyen véges ABC elemei. Jelöljük az első d egész halmazát S-sel, és S valamely G részhalmaza mellett legyen x_G az a D-dimenziós vektor, amelynek G-beli koordinátái megegyeznek x megfelelő koordinátáival, de a többi valamilyen új, az ABC-ben nem szereplő jellel egyenlő, amit "üresnek" fogunk mondani, magát x_G-t pedig az x G-re redukált értékének nevezzük.

Ha adott S részhalmazainak valamilyen $\mathcal{G}$ hipergráfja, és $G \in \mathcal{G}$ elemein adottak az $F_G(x_G)$ függvények, velük egy új függvényt definiálhatunk:

$$f(x) \quad = \quad \sum_{G \in \mathcal{G}} f_G(x_G).$$

Tegyük fel, hogy minden lehetséges x vektorhoz megmérjük az $Y(x)$ skalár menynyiséget, ezek függetlenek, normális eloszlásúak $f(x)$ várható értékkel és σ szórással. A feladatunk az $f_G(x_G)$ függvények és σ becslése. Itt alkalmazható lenne a legkisebb négyzetek módszere, de a feladatra közvetlen és aránylag egyszerű megoldás is

található: első közelítésben azt mondhatjuk, hogy megátlagoljuk mindazokat az $Y(x)$ értékeket, amelyekhez tartozó x G-re redukált értéke x_G. Ez az eljárás azonban csak bizonyos egyszerű $\mathcal{G}$ hipergráfokra jó. Általában az eredeti $f_G(x_G)$ függvényeket is, a minta átlagait is kanonikus alakra kell hoznunk: $\mathcal{G}$-be fel kell vennünk minden $G \in \mathcal{G}$-beli H halmazt, ezeken is értelmezni kell a $h_H(x_H)$ függvényeket úgy, hogy belőlük az eredeti függvények

$$f_G(x_G) \quad = \quad \sum_{H \subseteq G} h_H(x_H)$$

alakban előállíthatóak legyenek, és minden $A \subset H$ és minden a mellett teljesüljön rájuk a

$$\sum_{x_H : x_A = a} h_H(x_H) \quad = \quad 0$$

feltétel.

Monoton regresszió A lineáris regresszió a legegyszerűbb esete egy általános függvénykereső feladatkörnek. Ebben a nagy családban egy másik aránylag egyszerű esetet képvisel az a feladat, amikor a keresett függvény monoton. Ez vezet a következő feladatra: adottak az $(a_1, \ldots, a_n)$ számok, és keressük az $(x_1, \ldots, x_n)$ számokat úgy hogy rájuk $x_1 \leq x_2 \leq \cdots \leq x_n$ teljesüljön, és e mellett a feltétel mellett $\sum_{i=1}^n (a_i - x_i)^2$ minimális legyen.

Görbevonalú szórásanalízis Az előző két pont ötvözeteként kapjuk azt a feladatot, hogy az $Y(x)$ mérési eredményeket $\psi(b + \sum_{i=1}^n h_i(x_i))$ alakban közelítjük, ahol ψ montoton növő függvény.

Feladatok:

3.1. **(Nkv)** Legyenek X, Y ρ korrelációjú standard normális változók, és T tetszőleges valós szám. Határozzuk meg a

$$P(Y > T \mid X > T)$$

feltételes valószínűséget.

3.2. **(Nwt)** Legyenek az $X_{ij}, 1 \leq i, j \leq N$ mátrix elemei $i < j$ mellett független standard normálisok, $i = j$ mellett egymástól és az előzőektől független nulla várható értékű, $\sqrt{2}$ szórású normálisok, és legyen $i > j$ mellett $X_{ij} = X_{ji}$. Meghatározandó a mátrix saját értékeinek gyakoriság-hisztogramja.

3.3. **(Ewt)** Legyenek az $X_{ij}, 1 \leq i, j \leq N$ mátrix elemei független ± 1-esek. Határozzuk meg a mátrix determinánsának az eloszlását.

3.4. **(Gsa)** Legyen Y folytonos, és legyenek $X_1, \ldots, X_N$ diszkrét adatok. Keresendő az a monoton növő f függvény, és azok az $f_i, 1 \leq i \leq N$ függvények, amelyekre

$$\sum \left(Y \; - \; f \left(\sum_{i=1}^{N} f_i(X_i) \right) \right)^2$$

minimális, ahol az első összegezés a mintára vonatkozik.

3.5. **(Nkc)** Legyen $X_1, \ldots, X_n$ d-dimenziós standard normális minta, és legyenek $C_1, \ldots, C_k$ tetszés szerinti pontok. Vegyük minden egyes mintapontnak a hozzá legközelebbi C_i-től mért távolságának a négyzetét, és adjuk ezeket össze. Határozzuk meg azokat a $C_1, \ldots, C_k$ pontokat, amelyekre ez az összeg minimális. Mi ezeknek az empirikus értékenek az elméleti megfelelőjük?

3.6. **(Ndc)** Legyen $X_1, \ldots, X_n$ d-dimenziós standard normális minta, és minden egyes mintaponthoz rendeljük hozzá a tér azon pontjainak a halmazát, amelyekhez a minta pontjai közül az adott pont van legközelebb. Mondjuk azt, hogy egy mintapont véges, ha a hozzárendelt térrész korlátos. Határozzuk meg a véges mintapontok számának az eloszlását.

3.7. **(Ndg)** Legyen $X_1, \ldots, X_n$ d-dimenziós standard normális minta, és minden egyes mintaponthoz rendeljük hozzá a tér azon pontjainak a halmazát, amelyekhez a minta pontjai közül az adott pont van legközelebb. Mondjuk azt, hogy két mintapont szomszédos, ha halmazaik szomszédosak, és tekintsük azt a gráfot, amelynek a csúcsai a mintapontok, és amelyikben a szomszédosakat köti össze él. Határozzuk meg e gráfban a csúcsok fokának az eloszlását.

3.8. **(Ndf)** Legyen $X_1, \ldots, X_n$ d-dimenziós standard normális minta, és minden egyes mintaponthoz rendeljük hozzá a tér azon pontjainak a halmazát, amelyekhez a minta pontjai közül az adott pont van legközelebb. Mondjuk azt, hogy két mintapont szomszédos, ha halmazaik szomszédosak, és tekintsük azt a gráfot, amelynek a csúcsai a mintapontok, és amelyikben a szomszédosakat köti össze él. Határozzuk meg e gráfban a felfúvási szám eloszlását (cf. 1.14. Gfs).

3.9. **(Nkb)** Legyen $X_1, \ldots, X_n$ d-dimenziós standard normális minta. Határozzuk meg a konvex burok csúcs-számának az eloszlását.

3.10. **(Npt)** Legyen $X_1, \ldots, X_n$ d-dimenziós standard normális minta. Állítsuk elő azt a mátrixot, amelyiknek az i-edik sorának a j-edik eleme az i-edik mintaelemnek a j-ediktől mért távolságnégyzete. Határozzuk meg a sajátértékek hisztogramjának aszimptotikus viselkedését.

3.11. **(Wtb)** Legyen n tetszés szerinti egynél nagyobb egész. Keresendő az az eloszlás, amelynek a várható értéke 0, szórása 1, és az n elemű mintában a

páronkénti különbségek abszolut értékének a logaritmusának a várható értékének az összege maximális.

3.12. **(Önl)** Legyen c_{ij} egy tetszés szerinti szimmetrikus $n \times n$ méretű mátrix. Megadható-e az n-dimenziós ± 1 koordinátájú $X^T = (x_1, \ldots, x_n)$ vektor eloszlása úgy, hogy tetszőleges $1 \leq i \leq n$ mellett

$$\log\left(\frac{P\left(x_i = 0 \mid \{x_j, j \neq i\}\right)}{P\left(x_i = 1 \mid \{x_j, j \neq i\}\right)}\right) = \sum_{j \neq i} c_{ij} x_j$$

legyen?

3.13. **(Trm)** Mondjuk azt, hogy egy ± 1 elemű négyzetes mátrix valamelyik részmátrixa "tiszta", ha minden eleme $+1$. Határozzuk meg a véletlen mátrix legnagyobb tiszta részmátrixának a méretének az eloszlását.

3.14. **(Ind)** Egy páratlan sok csúcsú négyzetes rács csúcsaiban indánok ülnek, akik egymástól függetlenül $\frac{1}{2}$ valószínűséggel vagy ébren vannak, vagy alszanak, de a centrális indián ébren van, és valami fontosat mond azoknak a szomszédainak, akik ébren vannak. Azok ugyanezt teszik. Mérjük az indiánok távolságát sor és oszlop koordinátáik abszolut különbségének a maximumával. Határozzuk meg a hírt megtudó legtávolabbi indiánnak a centrumtól mért távolságának az eloszlását.

3.15. **(Bvt)** Mondjuk azt a szabályos egydimenziós bolyongásban, hogy az origóba történő valamely visszatérés "boldog", ha utunk során sehol máshol nem jártunk többször addig, mint az origóban. Határozzuk meg az n-lépéses bolyongás boldog visszatéréseinek a számának az eloszlását.

3.16. **(Mtg)** Adott egy irányított gráf. Keresendő az a sokdimenziós pontrendszer (ha egyáltalán ilyen van), amelyben az egyes pontokból a hozzájuk legközelebbi pontba futó élek együttese az adott irányított gráf.

Irodalom:

N.L. Johnson-S. Kotz: Discrete distributions, J. Wiley, 1969

N.L. Johnson-S. Kotz: Contonuous univariate distributions, J. Wiley, 1970

Tusnády Gábor(1979): Mátrixok szinguláris felbontása, Alkalmazott Matematikai Lapok 5, 375-384

G. Tusnády-A. Czeizel-L. Telegdi(1981): ML-fitting of multifactorial threshold models, Periodica Mathematica Hungarica, 12/3, 205-216

Y.Lee-J.A.Nelder(1996): Hierarchical generalized linear models, J.R.Statist. Soc.B 58/4, 619-678

4. Hipotézisek vizsgálata

TÉMÁK: a χ^2-próba, a függetlenség tesztelése, faktor analízis, log-lineáris modellek, Boole faktoranalízis, logisztikus regresszió, szekvenciális módszerek, döntésfüggvények, statisztikai programcsomagok: BMDP, SPSS, SAS, GENSTAT.

A χ^2-PRÓBA A statisztika legfontosabb feladata sztochasztikus hipotézisek felállítása, és ezek ellenőrzése. A legelső feladat, amivel egy statisztikus találkozik egy $(A_1, \ldots, A_k)$ teljes eseményrendszer valószínűségeinek az ellenőrzése. Jelöljük ezeket rendre p_i-vel, és tegyük fel, hogy n megfigyelés során A_i ν_i-szer következett be. Ezt a ν_i számot "kapott" értéknek nevezzük, és a "várt" értékével hasonlítjuk össze, ami $n\,p_i$. Magát az összehasonlítást két mennyiség alapján végezhetjük el, egyik

$$\frac{(\text{KAPOTT - VÁRT})^2}{\text{VÁRT}},$$

a másik

$$2\ \text{KAPOTT}\ \ln(\text{KAPOTT/VÁRT}).$$

Ha n a k-hoz képest nagy, a két mennyiség közel van egymáshoz, és eloszlásuk $(k-1)$ szabadságfokú χ^2 eloszlással közelíthető. Határértékben tehát az eloszlás nem függ a $(p_1, \ldots, p_k)$ számoktól. A véges eloszlás különben mindentől függ, és megbízható eredményt megfelelő szimulációval kaphatunk. Magát a második mennyiséget divergenciának hívják.

A FÜGGETLENSÉG TESZTELÉSE Ha az $(A_1, \ldots, A_k)$, és $(B_1, \ldots, B_m)$ teljes eseményrendszerekre n elemű közös megfigyelésünk van, és ezekben A_i B_j-vel ν_{ij}-szer következett be együtt, akkor legyen

$$\nu_{i.} \quad = \quad \sum_{j=1}^{m} \nu_{ij}, \qquad \nu_{.j} \quad = \quad \sum_{i=1}^{k} \nu_{ij}.$$

Ezek alapján akkor is becslést adhatunk ν_{ij} "várt" értékére, ha az $A_i,\ B_j$ események valószínűségét nem ismerjük:

$$\frac{\nu_{i.}\nu_{.j}}{n}.$$

A statisztikák ugyanazok, mint az előbb, most a határeloszlás szabadságfoka

$$(k-1)(m-1).$$

Tapasztalatom szerint a második mennyiség jobb eredményt ad, de ez is csak akkor használható, ha a ν_{ij} gyakoriságok között nincs nagyon sok nulla. Egy bizonyos

határig még maga a statisztika használható, ha a határeloszlása már nem is ad elfogadható közelítést, de egy ponton túl maga a statisztika semmitmondóvá válik. Ez még nem jelenti azt, hogy maga a függetlenség ellenőrizhetetlen lenne, a használható eljárásokra később visszatérünk.

FAKTOR ANALÍZIS Lásd: irodalom.

LOG-LINEÁRIS MODELLEK A szórásanalízis egyszerűsége vezetett arra, hogy a diszkrét adatok elemzéséhez valami hasonlót keressünk. Itt a marginális eloszlások az elemi építőkövek: ha nagyon sok változó esetén a változók bizonyos csoportjait kijelöljük, azokban azt a marginális eloszlást szeretnénk elfogadni, ami a mintában található, akkor ezekhez meg szeretnénk határozni azt a "legegyszerűbb" együttes eloszlást az összes változón, amelyiknek ezek a marginálisai.

Maga a numerikus munka egy aránylag egyszerű iteráció: minden egyes lépésben úgy szorozzuk át a kezünkben levő közelítést, hogy valamelyik marginális szerint tökéletes legyen a minta és a modell egyezése. Meglepő módon ez az egyszerű, bizonyos értelemben mohó algoritmus konvergál.

BOOLE FAKTORANALÍZIS Bináris adatok elemzésére sokféle speciális eljárás alakult ki. Ezek közül az egyik a szokásos faktoranalízis modelljét írja át Boole műveletekre. Szerintem a modell azzal egészítendő ki, hogy minden egyes koordinátát egy zajos csatornán keresztül figyelünk meg.

LOGISZTIKUS REGRESSZIÓ Ha bináris adatokból bináris adatokra akarunk következtetni, célszerű a magyarázandó változóban a két lehetséges érték valószínűségeinek a hányadosának a logaritmusát előállítani a magyarázó változók lineáris függvényeként.

SZEKVENCIÁLIS MÓDSZEREK Lásd: irodalom.

DÖNTÉSFÜGGVÉNYEK Lásd: irodalom.

STATISZTIKAI PROGRAMCSOMAGOK: BMDP, SPSS, SAS, GENSTAT Lásd: irodalom.

FELADATOK:

4.1. **(Mfp)** Adott $X_1, \ldots, X_N$ valószínűségi változóknak keresendő az a maximális particiója, amelyben a csoportok egymástól függetlenek.

4.2. **(Bfa)** Boole faktoranalízis.

4.3. **(Lgr)** Logisztikus regresszió.

4.4. **(Dhg)** Mondjuk azt, hogy egy hipergráf dekomponálható, ha elemei sorba állíthatóak úgy, hogy ha ebben a sorrendben az elemek $H_1, \ldots, H_N$, akkor bármely $1 \leq k < N$ mellett teljesül, hogy

$$\left(\cup_{i=1}^{k} H_i\right) \cap \left(\cup_{i=k+1}^{N} H_i\right) = H_k \cap H_{k+1}.$$

Hogyan lehet eldönteni, hogy egy hipergráf dekomponálható-e?

4.5. **(Flt)** Teszteljük a 2.1. feladatban előállított függetlenséget.

4.6. **(Shg)** Mondjunk szabadnak egy hipergráfot, ha rendelkezik a következő tulajdonsággal. Akárhogy adunk meg az élein egymással kompatibilis eloszlásokat, azokhoz található olyan eloszlás, amelynek az adott eloszlások mind marginális eloszlásai. Karakterizáljuk a szabad hipergráfokat.

4.7. **(Knk)** Legyen n tetszőleges pozitív egész, és $\rho = 0.5$. Legyenek $(X_i, Y_i, i = 1, \ldots, n)$ olyan standard normális eloszlású valószínűségi változó párok, amelyek egymástól függetlenek, de a párokon belüli elemek korrelációja ρ. Meg tudjuk-e különböztetni ezt a $2n$ elemű mintát a teljesen független standard normális mintától a rendezett minta alapján (tehát akkor, ha a $2n$ számot nagyság szerint növekvő sorrendben kapjuk meg)?

4.8. **(Neg)** Hogyan lehetne azt a hipotézist tesztelni, hogy az eloszlás 3 várható értékű $1/\sqrt{n}$ szórású normális eloszlás azzal az ellenhipotézissel szemben, hogy az eloszlás hat darab $(0, 1)$-beli egyenletes eloszlású véletlen szám összege? Mekkora minta mellett lehet biztosítani, hogy a valódi mintát csak 0.05 valószínűséggel utasítsuk el, de a tévedést 0.95 valószínűséggel észrevegyük?

4.9. **(Nex)** Jancsi és Juliska véletlen számokkal játszanak. Jancsi többdimenziós standard normálisa vektorokat generál. Amikor nem veszi észre, Juliska mindegyiket az erdetivel azonos irányú, de exponenciális eloszlású nagyságú vektorra cseréli ki, ahol a hosszak négyzetének várható értékét úgy választja meg, hogy az egyenlő legyen a standard normális norma-négyzetének a várható értékével. Hogyan tudja Jancsi ellenőrizni, hogy nem nyúlt-e Juliska a mintájához? Mit tehet, ha "megrágják az egerek" a mintáját: az csak a tér egy adott darabjában használható?

4.10. **(Ntc)** Jancsi és Juliska véletlen számokkal játszanak. Jancsi többdimenziós standard normálisa vektorokat generál. Amikor nem veszi észre, Juliska mindegyiket kicseréli a hozzá legközelebbire vett tükörképével. Hogyan tudja Jancsi ellenőrizni, hogy nem nyúlt-e Juliska a mintájához?

4.11. **(Nst)** Független, egyforma eloszlású, egységnyi szórású normálisak sorozatában akarjuk a várható érték előjelét szekvencális módszerrel tesztelni. A várható érték abszolut értéke ismert, és a hibák adottak, legyen mondjuk mindkettő 0.05. Viszont a szükséges mintaelemek számának a várható értékét a mellett a hipotézis mellett szeretnénk minimalizálni, hogy a várható érték nulla.

4.12. **(Ria)** Független, egységnyi szórású normálisak sorozatában akarjuk a várható értéket ellenőrizni. Tudjuk, hogy annak értéke kezdetben nulla, de egyszer csak +1-re változik, és attól kezdve annyi is marad. Minden egyes értéket megfigyelünk, és egyszer csak azt mondhatjuk, hogy állítsák le a folyamatot. Ilyenkor valahogy minden pontosan kiderül, és ha még nulla a várható érték, akkor A büntetést fizetünk, ha viszont már s lépésben +1 volt a várható érték, akkor sB a büntetés. Mit tegyünk?

4.13. **(Ntk)** Független, egyforma eloszlású, egységnyi szórású normálisak sorozatában akarjuk a várható érték előjelét tesztelni. A várható érték abszolut értéke ismert, és adottak a hibák költségei, valamint a mintavétel költsége. Mit tegyünk?

4.14. **(Bfp)** Két homogén, független, egyforma szórású normális eloszlású mintában akarjuk a várható értékek egyformaságát tesztelni. Mit tegyünk?

4.15. **(Vst)** Véletlen számok tesztelése.

IRODALOM:

Móri F. Tamás-Székely J. Gábor: Többváltozós statisztikai analízis, Műszaki Könyvkiadó, 1986

5. STATISZTIKAI BECSLÉSEK

TÉMÁK: a maximum likelihood módszer, elégséges statisztikák, exponencális család, szekvenciális módszerek, EM algoritmus, keverékek felbontása.

A statisztikus feladata a véletlennel terhelt megfigyelések elemzése. Az elemzés eszköze annak a sztochasztikus modellnek a kialakítása, amely a megfigyelés eredményéhez hasonló eredményekre vezet. A modell nem csak egyetlen eredményt képes előállítani, hiszen valahol szerepel benne a véletlen. Valahogy mérnünk kell a minta és az elmélet távolságát, minősítenünk kell, mennyire felelnek meg a modell tulajdonságai a mintáéinak. Általában a modell nincs egyértelműen meghatározva, ismeretlen mennyiségek, paraméterek szerepelnek benne, amelyeket a minta alapján kell meghatároznunk. Ezt az eljárást hívjuk becslésnek.

Egy aránylag egyszerű becslési eljárás a momentumok módszere. Annyi statisztikát választunk, ahány paraméterünk van, és a paramétereket úgy választjuk meg, hogy a statisztikák várható értéke pontosan annyi legyen, amennyi a mintában kapott értékük. Ez a módszer különösen szemléletes, ha a paraméter eleve valamilyen momentumot jelöl, ha például a normális eloszlás momentumait kell becsülnünk. Ebben az esetben a kezdő különösen hajlamos arra, hogy zavarba jöjjön: mit is jelenthet ez a tevékenység? Nehezíti a helyzetét, hogy elvileg bárminek bármi lehet a becslése, ha bizonyos triviális feltételeket kielégít.

Ugyancsak meglepő lehet, hogy általában a feladat megoldása nem egyértelmű. Minősítenünk kell a különböző megoldásokat. E minősítésben az egyetlen szempontunk az lehet, hogy a véletlen hogyan befolyásolja az egyes becsléseket. A torzítatlanság azt jelenti, hogy a becslés várható értéke maga a paraméter. A konzisztencia azt, hogy a becslés sztochasztikusan tart a becsült paraméterhez, ha a minta mérete tart a végtelenbe. A minimális szórásúság pedig azt jelent, hogy a becslés szórása a lehető legkisebb. Ezek ugyan természetes feltételek, de nem mindig teljesíthetőek. Például, ha X, Y független egységnyi szórású és normális eloszlású mennyiségek, a nagyobbik várható értékére már nem lehet egyértelműen meghatározott jó becslést adni.

A MAXIMUM LIKELIHOOD MÓDSZER A legáltalánosabban használt becslési eljárás. Népszerűségét jó aszimptotikus viselkedésének köszönheti. Legyenek $X_1, \ldots, X_n$

független azonos eloszlású valószínűségi változók, és az eloszlásuk ismeretetlen k-dimenziós paraméterét jelöljük ϑ-val. A minta likelihood függvénye:

$$L(\vartheta) \;=\; \sum_{i=1}^{n} \log f(X_i, \vartheta).$$

Fejtsük ezt Taylor sorba a valódi paraméter körül, amit jelöljünk ϑ_0-lal:

$$L(\vartheta) \;=\; L(\vartheta_0) + \;<\frac{1}{\sqrt{n}} L'(\vartheta_0), \Delta_n> \;+\frac{1}{2}<\frac{1}{n} L''(\vartheta_0)\Delta_n, \Delta_n> \;+R,$$

ahol R a maradéktag, $\Delta_n = \sqrt{n}\,(\vartheta - \vartheta_0)$, és$<.,.>$ a skaláris szorzást jelöli. Itt az $Y_n = \frac{1}{\sqrt{n}} L'(\vartheta_0)$ mennyiség a centrális határeloszlástétel miatt aszimptotikusan normális eloszlású, a $\Phi_n = \frac{1}{n} L''(\vartheta_0)$ mennyiség a nagy számok törvénye miatt egy Φ konstans mátrixhoz tart, ami csak a feladattól, és azon belül ϑ_0-tól függ. Mivel a maradéktag nullához tart, Δ_n aszimptotikusan $(-\Phi_n^{-1} Y_n)$-nel egyenlő, tehát a becslés hibája $1/\sqrt{n}$ rendű, és aszimptotikusan normális eloszlású. Belátható, hogy Y_n határeloszlásának a kovarianciamátrixa $-\Phi$, tehát Δ_n-ben ezt "túlosztjuk": Δ_n határeloszlásának a kovarianciamátrixa $-\Phi^{-1}$. Az úgynevezett Cramer-Rao egyenlőtlenség szerint viszont ennél kisebb kovariancia nem érhető el.

Abból, hogy $L(\vartheta)$ aszimptotikusan egy olyan kvadratikus alakkal egyenlő, amelyben a mátrix egy kovariancia (-1)-szerese, az is következik, hogy ha ez a kovariancia nem elfajuló (azaz jól van a feladat paraméterezve), akkor a

$$\hat{\vartheta} = \vartheta_0 - \frac{1}{\sqrt{n}} \Phi_n^{-1} Y_n$$

helyen aszimptotikusan maximuma van $L(\vartheta)$-nak.

A maximum értéke $\frac{1}{2} < \Phi_n^{-1} Y_n, Y_n >$-nel nagyobb $L(\vartheta_0)$-nál, ami aszimptotikusan $\frac{1}{2}\chi_k^2$ eloszlású, tehát a növekmény határeloszlása csak az ismeretlen paraméterek számától függ. Ennek alapján konfidencia intervallumot szerkeszthetünk az ismeretlen paraméterre: annak a valószínűsége ugyanis, hogy az ismeretlen paraméter benne van a

$$\{\vartheta : L(\vartheta) \;>\; \max_u L(u) - \chi_k^2(\alpha)\}$$

halmazban aszimptotikusan $(1-\alpha)$, ahol $\chi_k^2(\alpha)$ a χ_k^2 eloszlás megfelelő kvantilisét jelöli.

Elégséges statisztikák Lásd: irodalom.

Exponenciális család Lásd: irodalom.

Szekvenciális módszerek A hipotézisek vizsgálatán és becsléseken túl egy harmadik statisztikai feladat a konfidencia intervallumok szerkesztése: a minta alapján olyan intervallumot, tartományt jelölünk ki az ismeretlen paraméterekre, amelybe a valódi paraméter valamilyen előírt nagy valószínűséggel benne van. A megoldás gyakran csak annyi, hogy megadjuk a becslés szórásának a becslését. Maga a feladat keveréke a hipotézisek vizsgálatának és a becsléseknek, hiszen egy általános megoldása az, hogy összegyűjtjük mindazokat a paramétereket, amelyeket a minta az adott szignifikancia határ mellett elfogad valódi paraméternek.

Gyakori, hogy a konfidencia intervallum méretére felső korlát van: nem szeretnénk nagyon pontatlan eredményt kiadni a kezünkből. Ilyenkor azt mondjuk, hogy a minta még nem elég nagy, növelni kell. Amikor ezt a lehetőséget előre eltervezzük, a statisztikai eljárás részének tekintjük, szekvenciális módszerekről beszélünk. Igen hatékonyak a szekvenciális minőségvizsgáló módszerek.

EM algoritmus A maximum likelihood becslés kiszámítása általában bonyolult optimalizálást jelent. Csak néhány esetben adható explicit megoldás. Például az egyszerű eloszlások paramétereit általában a megfelelő momentumok becsülik, vagy a legkisebb négyzetek elve a maximum likelihood speciális eseteként ugyancsak explicit megoldást ad. Gyakori eset, hogy azért nincs explicit megoldás, mert nem rendelkezünk elegendő információval. Van egy nagy és sok részletet tartalmazó minta, amelyben van explicit becslés, de mi ennek csak töredékeit ismerjük. Például egy teljes kísérleti terv kivitelezése során bizonyos mérések sikertelenek voltak.

Hogyan lehetne ezt a körülményt kihasználni? A redukált, a mi rendelkezésünkre álló minta likelihood feladata bonyolult, sokszor maga a likelihood függvény is csak bonyolultan írható fel. Természetes gondolat, hogy valahogy rekonstruálni kellene a hiányzó adatokat. Itt egy körforgás alakul ki: ha ismernénk a paramétereket, a rekonstrukciót a feltételes eloszlás alapján el tudnánk végezni. Akkor viszont jobb becslést adhatunk a paraméterekre és a dolog kezdhető előlről.

A rekonstrukció során a likelihood függvény feltételes várható értékét kell kiszámolni, ez az eljárás E (expected value) része. Majd meg kell határozni ennek a maximumát, ez az M lépés.

Keverékek felbontása Lásd: irodalom.

Feladatok:

5.1. **(Bkf)** Bayesi keverékfelbontás.

5.2. **(Cca)** Adott egy $C_{ij}, 1 \leq i \leq N, 1 \leq j \leq M$ mátrix, amelyben az elemek nem negatívak, és az oszlopok összege 1. Lépésről lépésre megtehetjük a következőt. Tetszés szerint kiválasztunk egy sort, ennek az elemeit töröljük. A többi sor elemeit végigszorozhatjuk tetszés szerinti pozitív számokkal. Az egyetlen feltétel, hogy szorzás után az oszlopok öszege 1 alatt maradjon. Végül a kiválasztott sor elemeit feltöltjük úgy, hogy az oszlopok összege újra 1 legyen. Maximalizálandó az összes lépésben alkalmazott szorzó szorzata.

5.3. **(See)** Adott egy d-dimenziós szimplex, S. Megfigyeljük az S-en egyenletes eloszlású független $(X_1, \ldots, X_N)$ mintát. Adjunk $(X_1, \ldots, X_N)$ alapján becslést S-re.

5.4. **(Ekb)** Adott egy d-dimenziós szimplex, S. Legyenek $(X_1, \ldots, X_N)$ S-en egyenletes eloszlásúak és függetlenek. Határozzuk meg $(X_1, \ldots, X_N)$ konvex burkának a csúcsainak a számát. Mit mondhatunk ennek eloszlásáról?

5.5. **(Srt)** Hogyan lehetne az u.n. stepwise regression eljárásban a szignifikanciát tesztelni?

5.6. **(Gpp)** Legyenek μ, ρ ismeretlen paraméterek, n pozitív egész. Legyen $i = 1, \ldots, n$ mellett Y_i $\mu\rho^i$ paraméterű Poisson eloszlású valószínűségi változó. Adjunk az Y_i-k alapján becslést a μ, ρ paraméterekre.

5.7. **(Dfm)** Legyenek $(X_1, \ldots, X_n,\ Y_1, \ldots, Y_m)$ független, nem egyforma paraméterű normális eloszlású valószinűségi változók. A változókat nem tudjuk megfigyelni, csak az $\{\varepsilon_{ij} = \mathcal{I}(X_i < Y_j), i = 1, \ldots, n,\ j = 1, \ldots, m\}$ indikátor változókat, ahol $\mathcal{I}(.)$ értéke 1 vagy 0 aszerint, hogy a zárójelben álló esemény bekövetkezett-e vagy sem. Hogyan lehetne az eredeti X, Y változók paramétereit becsülni?

5.8. **(Mb1)** Legyen n pozitív egész, és legyenek $(T_i, V_i, i = 1, \ldots, n)$ ismeretlen paraméterek. Legyenek $(A_{ij}, K_{ij}, i, j = 1, \ldots, n, i \neq j)$ független, $\exp(T_i - V_j)$, illetve $\exp(V_i - T_j)$ paraméterű Poisson eloszlású valószínűségi változók. Adjunk becslést az ismeretlen paraméterekre, ha bizonyos (i, j) párokra ismertek az (A, K) mennyiségek.

5.9. **(Vpr)** Definiáljunk olyan véletlen mechanizmust, amely permutációkat generál valamilyen jól azonosítható paraméterek alapján, és adjunk becslést e paraméterekre.

5.10. **(Neb)** Legyen d pozitív egész, és μ ismeretlen d- dimenziós vektor, Z pedig d-dimenziós standard normális. Meg lehet-e μ normáját becsülni, ha ismerjük a $\mu + Z$ vektor koordinátáinak az előjelét?

5.11. **(Nse)** Tekintsük egy tetszőleges várható értékű és kovariancia mátrixú

többdimenziós normális eloszlás sűrűségfüggvényét az egységkocka csúcsain. Hogyan lehetne ezeket a paramétereket abból a diszkrét eloszlásból becsülni, amelynek a valószínűségei arányosak ezekkel a számokkal?

5.12. **(Nkf)** Adjunk becslést a többdimenziós normálisak keverékének a paramétereire.

5.13. **(Nvk)** Független, egyforma eloszlású, ismeretlen szórású normálisak sorozatában akarunk a várható értékre adott megbízhatóságú és adott szélességű konfidencia intervallumot szerkeszteni. Mit tegyünk?

5.14. **(Ghb)** Független, egyforma eloszlású minta elemeiről annyit tudunk, hogy ismeretlen paraméterű gamma eloszlású valószínűségi változók valamilyen ismeretlen, de egyforma kitevőjű hatványai. Adjunk becslést a paraméterekre.

5.15. **(Kln)** Legyen k pozitív egész, és legyenek $n_1, \dots, n_k$ szintén pozitív egészek. Legyenek $Y_1, \dots, Y_k$ ismeretlen, de egyforma paraméterű független normális eloszlású valószínűségi változók, ezeket nem tudjuk megfigyelni. Helyettük rendelkezésünkre állnak az

$$X_{ij} = Y_i + Z_{ij}, \quad j = 1, \dots, n_i, i = 1, \dots, k$$

megfigyelések, ahol a Z_{ij} mérési hibák függetlenek, normális eloszlásúak, nulla várható értékűek és a szórásuk egyforma, de ennek az értéke ismeretlen. Adjunk becslést az ismeretlen paraméterekre. Adható-e a hibák szórásnégyzetére torzítatlan becslés akkor, ha a becslés értéke nem lehet negatív?

Irodalom:

Tusnády Gábor-Roknich György(1969): A matematikai staisztika egy speciális alkalmazása a mérnöki gyakorlatban: a gammahatvány eloszlás, Mélyépítéstudományi Szemle 19/10, 472-478

A.P. Dempster-N.M. Laird-D.B. Rubin(1977): Maximum likelihood from incomplete data via the EM algorithm, J. R. Statist. Soc. B 39, 1-22

O. Barndorff-Nielsen: Information and exponential families in statistical theory, Wiley 1978

R. Pick-G. Tusnády(1980): Decomposition of mixtures, Studia Scientiarum Mathematicarum Hungarica, 15, 31-37

Tusnády Gábor(1982): Keverékek felbontása, Matematikai Lapok 30/1-3, 59-67

I. Csiszár-G. Tusnády(1984): Information geometry and alternating minimization procedures, Statistics and Decissions, Supplement Issue 1, 205-237

A. Perczel-M. Hollósi-G. Tusnády-G.D. Fasman(1991): Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins, Protein Engineering 4/6, 669-679

6. Nem paraméteres módszerek

Témák: túlélésbecslések, Kaplan Meier becslés, Cox regresszió, nem paraméteres módszerek, projection pursuit, matching, bin packing, a biztosításmatematika alapjai.

Túlélésbecslések A statisztika önálló ága az orvosi statisztika vagy biometria. Ez utóbbi kicsit általánosabb, minden egyéb biológiai statisztika is ide értendő, például a mezőgazdasági statisztika. A határok persze nem élesek, az egyes területeket részben a megvizsgált anyag, részben a speciális statisztikai feladat különíti el. Az orvosi statisztika egyik központi kérdése a betegek követése, és e követések kiértékelése. Amikor például a szívátültetéseket elkezdték, természetesen minden érdeklődő elsősorban arra volt kíváncsi, meddig maradnak életben a frissen operált betegek. Mint a statisztika minden egyes alkalmazásakor, itt is bonyolult, sok tényezős kérdésről van szó, amelynek a statisztika csupán az egyik része.

Statisztikai szempontból azonos feladatot jelent az ipari minőségvizsgálatok élettartamokra vonatkozó része. Inkább ez utóbbi terület alapfogalma a meghibásodási ráta, a $Q(t) = P(X > t)$ túlélésfüggvény logaritmikus deriváltjának a (-1)-szerese:

$$\rho(t) = \frac{f(t)}{Q(t)},$$

ahol $f(t) = -Q'(t)$. Ennek szemléletes jelentése a

$$\lim_{\delta \to 0} P(X < t + \delta \mid X \geq t)$$

határérértékből fakad: ha egy alkatrész t-ig használható, közelítőleg $\delta\rho(t)$ a valószínűsége annak, hogy t és $t + \delta$ között meghibásodik. A tapasztalat szerint a $\rho(t)$ függvény szemléletesebb képet ad az alkatrész megbízhatóságáról, nevezetesen jól értelmezhetőek a változásai: ha állandó, az exponenciális eloszlásról van szó, ha nő, az alkatrész öregszik, ha fogy, az alkatrész fiatalodik. Gyakran U-alakú ez a függvény: kezdetben nagy, de csökkenő intenzitású a meghibásodás, majd ideális esetben huzamosan állandó, végül lassan növekedni kezd, ami esetleg felgyorsul. Ilyen az ember élettartama is, az utolsó szakaszban a múlt század elején élt orvos, Gompertz megfigyelése szerint $\rho(t)$ exponenciálisan nő. Első közelítésben az igaz,

hogy annak a valószínűsége hogy valaki a következő évben meghal e^{-H}, ahol H a száz éves koráig még hátralevő évtizedeinek a száma.

KAPLAN MEIER BECSLÉS Természetes velejárójuk a túlélésbecsléseknek a cenzorálás. (Érdemes a csonkítást és cenzorálást megkülönböztetni: csonkításkor úgy vész el egy adat, hogy nyoma sem marad, cenzoráláskor annyit megtudunk róla, hogy milyen határok között van.) Operált betegek követésekor ez két okból is előfordulhat: a vizsgálat lezártakor a beteg még él és tünetmentes, vagy már korábban kivált a beteg a vizsgálatból, külföldre távozott, más orvoshoz ment, vagy meghalt, de nem az operálás következtében. Nem tehetjük meg, hogy az ilyen betegek adatait figyelmen kivül hagyjuk, noha nem ismerjük az élettartamuk végét.

Legyen a vizsgált X valószínűségi változó túlélésfüggvénye $P(X \geq t) = Q(t)$, és egy n-elemű megfigyelés során legyenek a megfigyeléseink a $((Z_i, \delta_i), i = 1, \ldots, n)$ párok, ahol $\delta_i = 1$, ha Z_i egy pontosan megfigyelt X_i-vel egyenlő, és legyen $\delta_i = 0$, ha csak annyi derült ki X_i-ről, hogy $X_i \geq Z_i$.

Rendezzük nagyság szerint a megfigyeléseinket. Jelöljük a szükséges permutációt $(\nu(1), \ldots, \nu(n))$-nel, és a kapott rendezett minta elemeit U_i-vel: $U_i = Z_{\nu(i)}$. Vigyük a társult változókat is az új helyükre, itt jelöljük őket v_i-vel: $v_i = \delta_{\nu(i)}$.

Keressük az ismeretlen $Q(t)$ eloszlásfüggvényt atomos alakban: tegyük fel, hogy X-nek csak az U_i-k a lehetséges értékeik, egyetlen U_n-nél nagyobb tetszőleges U_{n+1} érték kivételével. Jelöljük a $Q(U_i)$ ismeretleneket q_i-vel, és legyen $p_i = q_{i+1}/q_i, (i = 1, \ldots, n)$.

Ha $v_i = 1$, akkor a $\nu(i)$-edik megfigyelés valószínűsége

$$P(X = U_i) = Q(U_i) - Q(U_{i+1}) = q_i - q_{i+1} = (1 - p_i) \prod_{j=1}^{i-1} p_j,$$

ha pedig $v_i = 0$, akkor

$$P(X \geq U_i) = Q(U_i) = q_i = \prod_{j=1}^{i-1} p_j.$$

A teljes minta valószínűsége tehát

$$\prod_{i=1}^{n} (1 - p_i)^{v_i} \prod_{j=1}^{i-1} p_j \quad = \quad \prod_{i=1}^{n} (1 - p_i)^{v_i} \; {p_i}^{n-i}.$$

Ezt kell maximálnunk a $(p_1, \ldots, p_n)$ ismeretlenekben. "Szerencsénk" van: a változókban egymástól függetlenül lehet maximalizálni. Mivel $(1 - p)^A \; p^B$ maximuma

a $p = B/(A+B)$ helyen van (ez a relatív gyakoriság), $p_i = (n-i)/(n-i+v_i)$, tehát $q_i = \prod_{j=1}^{i-1}(n-j)/(n-j+v_j)$. Rejtélyes okból ezt product limit becslésnek hívják.

Cox regresszió A betegek követésekor nem csak a túlélési időt szokás megfigyelni: a betegek kórtörténete, kezelése nem egyforma. Az orvost elsősorban az érdekli, az eltérések hogyan befolyásolják a túlélési időt. Itt tehát első ránézésre regressziós feladatról van szó. A szokásos regressziós feladatban azonban számokból számokra következtetünk, itt pedig számokból eloszlásokra szeretnénk következtetni. Meg kell tehát találni ennek az alakját. A sok lehetőség közül a legegyszerűbb, és emiatt a legelterjedtebb a következő, Coxtól származó alak. Feltesszük, hogy minden szóban forgó túlélési függvény egy közös $Q(t)$ függvény valamilyen hatványa. Legyen a kitevő w: ha $w > 0$, akkor Q-val együtt Q^w is monton fogy, formálisan tehát újra megfelelő függvényt kapunk. Mivel pedig w szám, ezt már kereshetjük a szokásos regressziós alakok szerint. Mivel w-nek pozitívnek kell lenni, a logaritmusát közelítjük a kísérő változók lineáris függvényeként.

Az így kapott feladatban ismeretlen paraméterek és egy ismeretlen függvény szerepel: az ilyen feladatokat mostanában szemiparametrikus feladatnak mondják. Ha a lineáris együtthatókat ismertnek vesszük, akkor ismerjük a w kitevőket. Ekkor - még cenzorálás esetén is - a keresett függvény explicite felírható a Kaplan-Meier becslés levezetésekor alkalmazott gondolatmenettel. Ennek segítségével pedig a teljes feladat közönséges többváltozós optimalizálássá redukálható.

Elképzelhető, hogy hasonló módon oldható meg a következő feladat is. Legyenek az $(X_i, i = 1, \ldots, n)$ k dimenziós véletlen vektorok függetlenek, egyforma eloszlásúak, hasonlóan legyenek az $(Y_i, i = 1, \ldots, n)$ k dimenziós véletlen vektorok függetlenek, egyforma eloszlásúak, és legyen ez a két rendszer független egymástól. Tegyük fel, hogy csak az $(Y_i, i = 1, \ldots, n)$ véletlen vektorokat és az $(< X_i, Y_i >, i := 1, \ldots, n\ \ k)$ skaláris szorzatokat tudjuk megfigyelni. Hogyan becsüljük az X_i-k eloszlását?

Tegyük fel bizonyos értelemben általánosabban, hogy egy ismeretlen $(p_1, \ldots, p_N)$ eloszlással kapcsolatban független megfigyeléseket végezhetünk úgy, hogy minden megfigyelés előtt tetszés szerint kijelölhetjük az első N pozitív egész valamelyik részhalmazát, és annyit tudunk meg, hogy a soron következő véletlen szám abban benne van-e, vagy sem. Hogyan válasszuk a halamazokat, és hogyan becsüljük az ismeretlen eloszlást, ha azt akarjuk, hogy például a becslésünk divergenciája a valódi eloszlástól a lehető leggyorsabban tartson nullához?

NEM PARAMÉTERES MÓDSZEREK Miután a statisztika hőskorában minden feladatot a normalitás feltételezése mellett oldtak meg, lassan terjedni kezdtek a normalitást nem használó módszerek. Itt két egymást átfedő, egymással szoros kapcsolatban álló terület említhető:

- eloszlásmentes eljárások keresése, nevezetesen a rendezett mintán alapuló, úgynevezett rendstatisztikák vizsgálata, és ezek alapján a marginális eloszlások normálissá való transzformálása, az úgynevezett skálázás, vagy scoring;

- magának az eloszlásnak a vizsgálata, mint például a fentiekben tárgyalt túlélésvizsgálatok, vagy az eloszlásokra vonatkozó hipotézisek vizsgálata, például a normalitás, vagy általában valamilyen családhoz való tartozás vizsgálata, vagy annak eldöntése, elfogadható-e Cox fenti hipotézise.

Ha egy adott eloszlásra vonatkozó hipotézist akarunk ellenőrizni, tiszta illeszkedésvizsgálatról beszélünk, ha csak az eloszlás típusát ismerjük, és néhány paramétert a mintából határozunk meg, akkor becsléses illeszkedésvizsgálatról van szó.

PROJECTION PURSUIT Pár éve hirtelen divatba jött a többdimenziós adatmezők effektív automatikus vizsgálata, amelynek az alapgondolata az, hogy vetítéssel a magas dimenziós adatmezőt alacsony dimenzióssá alakítjuk, alkalmazunk a vetületen valamilyen kifejezetten alacsony dimenziós eljárást (például ha az alacsony dimenzió kettő, megnézzük az adatokat), majd az egészet "minden lehetséges vetítésre" megismételjük, és meghatározzuk a legmarkánsabb vetítést.

MATCHING Magas dimenzióban egy aránylag egyszerű illeszkedésvizsgálati módszer a következő. Akkora mintát generálunk a valódi eloszlásból, mint az eredeti minta, és "összeházasítjuk" a két mintát: úgy állítjuk párba az elemeket, hogy a páronkénti távolságok négyzetösszege minimális legyen. Az illeszkedés jóságát a kapott minimummal mérjük.

BIN PACKING Ha az egységnégyzeten egyenletes eloszlású mintát ellenőrzünk a fenti módszerrel, a kapott statisztika kapcsolatban áll a következő feladat heurisztikus algoritmusainak a műveleti sebességének a becslésével. Egy véletlen input $(0,1)$-beli, mondjuk ott egyenletes eloszlású, és egymástól független számokból áll. A számokat egységnyi térfogatú csuprokba rakjuk, és lehetőleg kevés csuprot szeretnénk felhasználni. A feladat nehézsége abból fakad, hogy "on line" kell megoldanumk: minden egyes szám helyét érkezésekor ki kell jelölnünk, később azon már nem változtathatunk.

A BIZTOSÍTÁSMATEMATIKA ALAPJAI Ha az élettartamom az X valószínűségi változó ismert $Q(t)$ túlélés függvénnyel, kérdés Δ folytonos kamatozás mellett mekkora összeget kell most a biztosítónak fizetnem, ha halálomkor egységnyi kifizetést szeretnék a családomnak biztosítani? A válasz $E\ e^{-\Delta X}$, és ha ezt a mennyiséget κ-val jelöljük, továbbá β annak a folytonos befizetési intenzitásnak a mértéke, amely ugyanezt a célt szolgálja, és μ az egységnyi befizetésre járó folytonos kifizetés intenzitása, akkor belátható, hogy

$$\beta \quad = \quad \frac{\kappa\Delta}{1-\kappa},$$

és $\beta = \kappa\mu$.

FELADATOK:

6.1. **(Qmt)** Legyenek $X_1, \ldots, X_N,\ Y_1, \ldots, Y_N$ az egységnégyzetben független, egyenletes eloszlású pontok. Meghatározandó az a $P_1, \ldots, P_N$ permutáció, amelyre

$$\sum_{i=1}^{N} \parallel X_i \ - \ Y_{P_i} \parallel^2$$

minimális, majd e minimum eloszlása.

6.2. **(Hmt)** Legyenek $X_1, \ldots, X_N,\ Y_1, \ldots, Y_N$ az egységnégyzetben független, egyenletes eloszlású pontok. Állítsuk párba a pontokat úgy, hogy minden egyes párban az X pont koordinátái kisebbek legyenek az Y pont koordinátáinál, és a párok száma maximális legyen. Mi a párok számának az eloszlása?

6.3. **(Env)** Egydimenziós normalitás-vizsgálat tetszés szerinti módszerrel.

6.4. **(Cxr)** Legyen W H eloszlású pozitív valószínűségi változó, és legyen X-nek W-re vett feltételes eloszlása

$$P(X > t \mid W) \ = \ (1 - F(t))^W,$$

ahol F ugyancsak egy pozitív valószínűségi változó eloszlása. Becsülendő az X, W párra vonatkozó független, egyforma eloszlású minta alapján F.

6.5. **(Npp)** Többdimenziós normalitás-vizsgálat projection pursuit módszerrel.

6.6. **(Mrg)** Monoton regresszió.

6.7. **(Mrk)** Legyen $(a_i,\ i = 1, \ldots, n)$ ismeretlen monoton növő sorozat, és legyen σ ismeretlen pozitív szám. Adjunk az $(Y_i = A_i + \sigma Z_i,\ i = 1, \ldots, n)$ megfigyelések alapján konfidencia intervallumot az a_i számokra, feltéve, hogy a Z_i-k független standard normálisak.

6.8. **(Mri)** Adott egy körmentes irányított gráf, és a csúcsoknak egy tetszés szerinti értékelése. Adjunk erre az értékelésre négyzetes eltérésben legjobb közelítést, mely ekvimonoton az adott gráffal.

6.9. **(Ext)** Az exponencialitás tesztelése.

6.10. **(Gat)** A gamma eloszlás tesztelése.

6.11. **(Eta)** Legyen n ismert pozitív egész, és $\pi = (p_1, \dots, p_n)$ egy ismeretlen eloszlás, $X_i, i = 1, \dots$ független, π eloszlású valószínűségi változók. Magukat az X_i-ket nem tudjuk megfigyelni, de lépésről lépésre minden egyes $i = 1, 2, \dots$ pozitív egész mellett kijelölhetjük az első n pozitív egész tetszés szerinti A_i részhalmazát, és annyit megtudunk, hogy X_i benne van-e az A_i részhalmazban, vagy sem. Miután ezt megtudtuk, kijelölhetjuk az A_{i+1} halmazt, és így tovább. Hogyan válasszuk az A_i halmazokat, ha meg szeretnénk becsülni a p_i valószínűségeket?

6.12. **(Ebn)** Legyen $X = Y + Z$, ahol Y, Z fügetlenek, Z standard normális. Hogyan becsülhető Y eloszlása az X-re vonatkozó, független, egyforma eloszlású minta alapján?

6.13. **(Gzr)** Legyen $X_1, \dots, X_n \dots$ független, egyforma eloszlású valószínűségi változók végtelen sorozata, és $S_n = X_1 + \dots + X_n$. Hogyan lehetne becslést adni az X_n-ek eloszlására, ha ismerjük az S_n összegek egész részét?

6.14. **(Emi)** Hogyan lehetne a sokdimenziós eloszlásokat konzisztensen és eloszlásmentesen tesztelni?

6.15. **(Skp)** Legyen $X_1, \dots, X_n$ ismert eloszlású sokdimenziós független minta, és legyen $(d_{kj}, j = 1, \dots, n-1)$ az X_k-tól különböző mintapontok X_k-tól mért távolságának rendezett mintája, p_{kj} az X_k középpontú, d_{kj} sugarú gömb elméleti valószínűsége az adott eloszlás szerint. Mit mondhatunk a $\max_{kj} \mid p_{kj} - j/n \mid$ statisztika aszimptotikus eloszlásáról?

IRODALOM:

E.L. Kaplan-P. Meier(1958): Nonparametric estimation from incomplete data, J. of the Amer. Stat. Assoc. 53, 457-481

P. Erdős-A. Rényi(1970): On a new law of large numbers, Jour. Analyse Mathématique, 22, 103-111

Major Péter-Tusnády Gábor(1973): Normalitás-vizsgálat, MTA III. Osztály Közleményei,22, 257-281

D.R. Cox-D. Oakes: Analysis of survival data, Chapman and Hall, 1984

M. Ajtai-J. Komlós-G. Tusnády(1984): On optimal matchings, Combinatorica 4/4, 259-264

C.H. Zhang(1990): Fourier methods for testing mixing densities and distributions, Ann. Stat. 18, 806-831

P.W. Shor(1992): How to pack better then best fit: tight bounds for average-case on-line bin packing, Proceedings of 32nd Annual Symposium on Foundations of Computer Sciences, 752-766

P.K. Andersen-Ø. Borgan-R.D. Gill-N. Keiding: Statistical models based on counting processes, Springer 1992

7. MARKOV LÁNCOK

TÉMÁK: elágazó folyamatok, sorban állás, entrópia, kódolás, csatorna-kapacitás, Ziv-algoritmus, fehérjék, aligning, többdimenziós integrál.

Egy Markov láncnak a szokásos szóhasználat mellett "állapotai" vannak. Tegyük fel, hogy ezek száma véges, jelöljük a számukat N-nel, magukat az állapotokat $S_1, \ldots, S_N$-nel, az állapotok halmazát $\mathcal{S}$-sel. A Markov lánc olyan fix időközönként változó sztochasztikus folyamat, amelynek a lehetséges értékei $\mathcal{S}$-beliek, és amelyre a

$$P(X_t \mid X_{t-1}, \ldots, X_0)$$

feltételes valószínűségek csak t-től és X_{t-1}-től függnek. Ha a folyamat homogén, akkor ezek a feltételes valószínűségek t-től sem függnek, ami még nem jelenti azt, hogy a

$$p_t(i_0, i_1, \ldots, i_k) \;=\; P(X_{t+j} = S_{i_j}, \quad j = 0, 1, \ldots, k)$$

valószínűségek sem függenek t-től, csak ha a

$$\pi_i \;=\; P(X_0 \;=\; S_i)$$

kezdeti eloszlást úgy választjuk meg, hogy X_1 eloszlása megegyezzen X_0 eloszlásával. Könyen látható, hogy ez mindig elérhető, és ha az

$$a_{ij} \;=\; P(X_1 = S_j \mid X_0 = S_i)$$

átmenet valószínűségek mind pozitívak, akkor tetszőleges kezdeti eloszlásból kiindulva X_t eloszlása tart az egyértelműen meghatározott stacionárius kezdeti eloszláshoz. Általában ha ez így van, a Markov láncot ergodikusnak hívjuk. Magukat az

$$a_{ij}(t) \;=\; P(X_t = S_j \mid X_0 = S_i)$$

t-lépéses átmenet valószínűségeket rekurzíven számolhatjuk ki. Egy Markov láncban a "jelen" ismerete mellett a "múlt" és "jövő" feltételesen függetlenek: ez egy szimmetrikus tulajdonság, tehát az idő irányításának nincs kitüntetett szerepe. Ha kell, megfordíthatjuk a Markov láncot, és a jövőből a múltba vivő feltételes valószínűségekkel dolgozhatunk. Ezek kiszámolásához azonban nem csak az átmenet valószínűségeket kell ismernünk, hanem a kezdeti eloszlást is.

Statisztikai szempontból a véges állapotú, homogén és ergodikus Markov láncok vizsgálata nem jelent nehezebb feladatot mint a független azonos eloszlású változók vizsgálata: az átmenet valószínűséget ugyanúgy relatív gyakoriságokkal kell becsülni, mint magát az eloszlást. Nehezebb a helyzet, ha a folyamat csak egy zajos csatornán keresztül figyelhető meg. Ha maga a csatorna emlékezet nélküli, akkor a

$$b_{ij} \quad = \quad P(Y_t \;=\; O_j \mid X_t \;=\; S_i)$$

átviteli valószínűséget határozzák meg, ahol $(O_1, \dots O_M)$ a csatorna kimeneti ABC-jének az elemei, és $Y_0, Y_1, \dots, Y_T$ a megfigyelhető folyamat értékei. A megfigyelhető folyamat általában nem Markov, de az $Y_0, Y_1, \dots, Y_T$ változók együttes eloszlása hasonló rekurzióval számolható ki, mint a többlépéses átmenet valószínűségek, az a_{ij}, b_{ij} paraméterek pedig az EM algoritmussal becsülhetőek. Ezeket a folyamatokat rejtett Markov folyamatoknak hívjuk, a paraméterek becslése a Baum-Welch algoritmus.

Elágazó folyamatok Tegyük fel, hogy egy populáció egyedei egyformák, egységnyi ideig élnek, és haláluk előtt egymástól függetlenül p_k valószínűséggel k utóduk lesz, ahol $(p_k, k = 0, 1, \dots)$ tetszőleges eloszlás a természetes számokon. Ha a populáció megfigyelésekor azt is meg tudjuk figyelni, hogy melyik egyednek hány utóda lett, akkor a megfigyelések Markov láncot alkotnak. Ha csak az egyedek

számát tudjuk meghatározni, a megfigyelések rejtett Markov folyamatot alkotnak és az utódok számának az eloszlását ismét az EM algoritmussal becsülhetjük.

SORBAN ÁLLÁS Független, azonos eloszlású X_t változók részletösszegei felújítási folyamatot alkotnak:

$$S_t \;=\; X_1 + \cdots + X_t.$$

Ha egy borbélyhoz vendégek érkeznek, az érkezési időket közelíthetjük felújítási folyamattal. Ha a borbély minden egyes vendéggel ugyanolyan eloszlású véletlen időt tölt, és ezek az idők egymástól is, az érkezési folyamattól is függetlenek, úgynevezett sorbanállási folyamatot kapunk. Jelöljük ebben a k-adik vendég kiszolgálási idejét Y_k-val, a kiszolgálásának a végét T_k-val. Ekkor

$$T_k \quad = \quad \max(T_{k-1}, S_k) \;+\; Y_k,$$

hiszen csak akkor kezdhet a borbély a k-adik vendéggel foglalkozni, ha a $(k-1)$-edikkel már végzett, és a k-adik vendég megérkezett. Kérdés, meg tudjuk-e becsülni az X_i, Y_i változók eloszlását, ha csak a várakozók számának az alakulását ismerjük? Ez utóbbi jelentheti azt is, hogy minden egyes változást regisztrálni tudunk, de jelentheti azt is, hogy időegységenként megnézzük, hányan várakoznak a borbélyra.

ENTRÓPIA Azt mondjuk, hogy egy diszkrét értékű folyamat stacionárius, ha a

$$P(X_{t+j} \;=\; S_j, j \;=\; 0, 1, \ldots, k)$$

együttes eloszlások nem függenek t-től. Ha az állapotok száma véges, ezek az együttes eloszlások is végesek, és az entrópiájuk ugyanúgy meghatározható, mint a marginális eloszlásoké. Belátható, hogy az együttes eloszlások entrópiáját a figyelembe vett változók számával osztva monoton fogyó sorozatot kapunk (hiszen a feltételes entrópia fogy, ha a feltétel nő). E monoton fogyó sorozat határértéke a folyamat egy jelre jutó entrópiája. Maga az átlagos entrópiákból álló sorozat sokat mond egy statisztikusnak: ha egy tagtól kezdve állandó, akkor a folyamat véges rendű Markov folyamat, ami azt jelenti, hogy ha az állapotaiból képezett blokkokat tekintjük új állapotoknak, akkor Markov a folyamat. A véges rendű Markov folyamatok rendjét általában az úgynevezett büntető függvényes maximum likelihood becsléssel határozhatjuk meg: az eredeti likelihood függvényből levonjuk a rend valamilyen alkalmasan választott monoton növő függvényét, és ezt a különbséget maximalizáljuk egyszerre a rendben és a hozzá tartozó paraméterekben.

A büntető tag általában a tagszám logaritmusának konstansszorosa. Ezt az eljárást helyettesítheti az entrópia valamilyen konziszens becslése: ennek alapján ugyanis elegendő a rendet addig növelni, amíg az illesztett Markov-modell entrópiája egyenlővé nem válik a folyamat entrópiájával.

KÓDOLÁS Shannon tétele alapján minden stacionárius folyamat átkódolható bináris folyamattá úgy, hogy a kódszavak átlaga az entrópiát tetszőleges pontossággal megközelítse. Közben az eredeti folyamat blokkjait kódoljuk, és a kódszavak hossza változó. Ha a kód jó, a keletkezett bináris folyamat közelítőleg tisztán véletlenszerű: ha ez nem így van, megpróbálhatjuk a kódolt folyamatból magukat a kódokat kiolvasni. Ebben a modellben az a legfontosabb, hogy gyökeresen átalakítja a folyamat időskáláját, és ha a kódolás "rossz", látványos összefüggéseket is generálhat, miközben a folyamat lényegében független és azonos eloszlású elemekből építkezik.

CSATORNA-KAPACITÁS Az információelméletben tárgyalt csatorna ugyanolyan Markov-maggal adható meg, mint egy Markov folyamat, csak a bemeneti és kimeneti állapotok különbözhetnek. A csatorna kapacitása e kettő kölcsönös információjának a maximuma.

ZIV-ALGORITMUS Egy tetszőleges karakter-sorozatot aránylag rövid blokkokra tördelhetünk rekurzíven. Dobáljuk képzeletben egy zsákba a keletkező darabokat. Egy törés után addig megyünk el a sorozatban, amíg olyan blokk nem alakul ki, ami már nincs a zsákban. Ezt a blokkot letörjük, és beledobjuk a zsákba. Induláskor definició szerint törés van és a zsák üres. Például a

VSSPGFSPTSPTYSPTSPAYSPTS

karaktersorozat törési pontjai:

V•S•SP• G•F•SPT•SPTY•SPTS• P•A•Y•SPTS.

A sorozat végén nincs törés, dehát ott vége van a sorozatnak. A darabokból egy fa állítható össze, amit érdemes az algoritmus üzemeltetésekor adminisztrálni. Maga a fa ismeretlen folyamatokkal való első ismerkedéskor használható a statisztikában: ha az ágak közel egyforma hosszúak, nem túl erős a folyamat struktúrája, ha nagy a hosszak szórása, az mindig erős belső összefüggésekre utal. Eredetileg adattömörítésre, és az entrópia meghatározására született az algoritmus, ez utóbbira saját tapasztalatom szerint nem nagyon használható a konvergencia lassúsága miatt.

FEHÉRJÉK, ALIGNING A fehérjék 20 aminosavból épülnek fel, ezeket az ABC betűivel szokás jelölni. Az alábbi két 24 elemű blokk egy-egy nagyobb fehérjének a része:

VSSPGFSPTSPTYSPTSPAYSPTS,
SPSYSPTSPCYSPTSPSYSPTSPN.

A blokkokat bizonyos szisztematikus keresés emelte ki egy körülbelül egymillió elemű adatbázisból. Kérdés: hogyan lehet ilyen párokat találni? Ha az elsőben az első két, a másodikban az utolsó két betűt elhagyjuk, a két blokk még jobb fedésbe hozható:

SPGFSPTSPTYSPTSPAYSPTS,
SPSYSPTSPCYSPTSPSYSPTS.

Ez persze így nagyon speciális: általában kisebb méretű az egyezés. Például ugyanebben az adatmezőben található a

KSSKGGPGSAVSPYPTFNPSSDVA

blokk is, ez már aránylag messzebb van az első blokktól:

V S S • • • P G F S P T S P T Y S P T • S P A Y S P T S • • • ,
K S S K G G P G • S A V S P • Y • P T F N P • • S • • S D V A .

Itt a következő a feladat: helyezzünk el tetszőleges számú • jelet az adott szövegekbe úgy, hogy az egy oszlopban álló párok között az azonosak száma maximális legyen. Egy másik példa ebből a családból:

T S • P G F G V S S P G F S P T S P T Y S P T S P • ,
A S S P G • G A S • P N Y S P S S P N Y S P T S P L .

Itt világos, hogy az YSPTSPS blokk ciklikus ismétlődése adja a család elemei között a kapcsolatot. Vagy egy kicsit kócosabb példa:

A F N K E A N E L • • A K V A • • A • • T G • D A A • A V K A Q F G K V G Q • ,
• M N K • • N E L V S A • V A E K A G L T K S D A A S A V D A V F D • V V Q A .

Általában nehéz megmondani, mit is érdemes keresni egy ilyen adatmezőben: nyilván nagyobb családok struktúráját nehezebb áttekinteni, és csak fokozatos, nem csupán számítástechnikai lépések során kristályosodik ki a valódi cél. Ennek megfelelően ennek a feladatkörnek (aligning) kiterjedt és szerteágazó irodalma van, amit külön színez az a körülmény, hogy ezeknek a nagy molekuláknak "kristályos" szerkezetük van, és épp ez a szerkezet volt az, ami az élet kialakulása és fejlődése során az egyes családokat a mutáció és szelekció egyensúlya révén létrehozta.

Az úgynevezett multiple aligning feladat egy lehetséges megfogalmazása a következő. Adott szövegeket szóközök beszúrásával rendezzünk mátrixba úgy, hogy az

egyes szövegek a mátrix egy-egy sorába kerüljenek. Adjuk össze a mátrix oszlopainak az entrópiáit és vonjuk ki ebből a felhasznált szóközök számának a konstansszorosát. Ezt a mennyiséget kell maximalizálni. A dinamikus programozás alkalmazható erre a feladatra, de csak két-három szövegre kivitelezhető a mai számolási sebesség mellett. Ezért heurisztikus algoritmusok terjedtek el: az egyes szövegeket egy közös szöveghez igazítják, majd az új mátrix alapján átírják a közös szöveget.

TÖBBDIMENZIÓS INTEGRÁL A közönséges kvadratura formulák műveletigénye a tér dimenziójának a függvényében exponenciálisan nő. Elvileg egy Monte Carlo szimuláció lépésszáma nem függ a dimenziótól, a módszer hatékonysága azonban a tipikus feladatokon nem kielégítő. Lovász és Simonovits a vizsgált tartományba helyezett rácson való Markov típusú bolyongással elérték, hogy a műveletigény a dimenzió fix hatványa alatt maradjon (jelenleg a legjobb kitevő 5).

FELADATOK:

7.1. **(Mab)** Legyenek $X_1, \ldots, X_N$, $Y_1, \ldots, Y_M$ diszkrét értékű sorozatok. A két sorozat tetszőleges két A, B elemű blokkjára legyen D azoknak a beszúrásoknak és törléseknek a minimális száma, amelyekkel az egyik blokkból a másik előállítható. Keresendő az a blokkpár, amelyre $A + B - 2D$ maximális.

7.2. **(Ali)** Legyen S egy tetszőleges szöveg, és lépésről lépésre állítsuk elő egymástól függetlenül belőle véletlen törlésekkel, beszúrásokkal és betűmódosítással a

$$T_1, \ldots, T_N$$

szövegeket. Állítsuk vissza ezekből S-et, amennyire ez lehetséges. Hogyan függ a visszaállítás hibája a körülményektől?

7.3. **(Cst)** Egy sztochasztikus csatorna kapacitásának a kiszámolása.

7.4. **(Kiv)** Határozzuk meg az elágazó folyamat kipusztulásának a valószínűségét.

7.5. **(Kii)** Határozzuk meg az elágazó folyamatban a család élettartamának az eloszlását.

7.6. **(Qma)** Legyenek az $n \times n$ méretű M mátrix elemei nem negatívak, és tegyük fel, hogy a sorok, oszlopok összege pozitív. Tekintsük az $S = \{(x(1), \ldots, x(n)) : x(i) \geq 0, i = 1, \ldots, n; \sum_{i=1}^{n} x(i) = 1\}$ halmazon az $\mathcal{M} : S \mapsto S$ leképezést, melyre

$$(\mathcal{M}x)(i) = y(i)/\kappa, \quad \text{ahol} \quad y(i) = \sum_{j=1}^{n} M_{ij} x(j)^2 \quad \text{és} \quad \kappa = \sum_{i=1}^{n} y(i).$$

Megadható-e az M mátrix úgy, hogy a $\cap_{n=1}^{\infty}\mathcal{M}^n S$ halmaz egy görbe legyen?

7.7. **(Rmb)** Hogyan lehetne egy Markov folyamat paramétereit becsülni, ha a folyamat állapotait csak egy zajos, emlékezet nélküli csatornán át tudjuk megfigyelni?

7.8. **(Lsi)** Lovász és Simonovits többdimenziós integrálja.

7.9. **(Tkv)** Tegyük fel, hogy egy populációban több különböző fajta él, és ezek mindegyike időegység után egy, a fajtára jellemző többdimenziós eloszlás szerint különböző fajtájú utódokat hoz létre. Határozzuk meg a populáció kipusztulásának a valószínűségét.

7.10. **(Tki)** Tegyük fel, hogy egy populációban több különböző fajta él, és ezek mindegyike időegység után egy, a fajtára jellemző többdimenziós eloszlás szerint különböző fajtájú utódokat hoz létre. Határozzuk meg a populáció kipusztulásának az idejét.

7.11. **(Mrb)** Markov folyamat rendjének becslése.

Irodalom:

I. Csiszár-G. Katona-G. Tusnády(1969): Information sources with different cost scales and the principle of conservation of entropy, Zeitschrifts für Warscheinlichkeitstheorie und verwandte Gebiete, 12, 185-222

J. D. Watson: A gén molekuláris biológiája, Medicina Könyvkiadó, 1980

I. Csiszár-J. Körner: Information theory: coding theorems for discrete memoriless systems, Akadémiai Kiadó, 1986

L. R. Rabimer(1989): A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 77, 257-286

J.D. Watson-M. Gilman-J. Witkowski-M. Zoller: Recombinant DNA, Freeman, 1992

L. Lovász-M. Simonovits(1993): Random walks in a convex body and an improved volume algorithm, Random Structures and Algorithms, 4, 359-412

8. Idősorok

Témák: mozgó átlag folyamatok, elsőrendű autoregresszív folyamatok, ARMA folyamatok, állapotteres leírás, Kálmán szűrés.

Elvileg nincs nagy különbség a többdimenziós analízis és az idősorok analízise között. Formálisan persze mondható, hogy az idősorokban az adatok az időtől is függenek, de ez sincs mindig így, például a fehérjékben az aminosavak sorrendje statisztikai szempontból ugyanúgy kezelhető, mint valamely információforrás által kibocsátott jelek sorozata. A tradicionális idősorokról annyi minden esetre elmondható, hogy általában homogének, a változó időben az adatok struktúrája csak kis mértékben változik, maga a dimenzió flexibilisebb, és bizonyos értelemben nagyobb, mint ami a sokváltozós analízisben szokásos. Én itt most kizárólag Gauss folyamatokról fogok beszélni, ez többé kevésbé megfelel az általános szóhasználatnak.

Mozgó átlag folyamatok Legyenek az $(u_t, t = 0, \pm 1, \dots, \pm n, \dots)$ kétirányú végtelen sorozat elemei független skalár standard normális változók. Ezt a sorozatot fehér zajnak nevezik. Legyen q tetszőleges természetes szám, és legyenek

$$(b_0, b_1, \dots, b_q)$$

tetszőleges valós számok. Az

$$y_t \;=\; \sum_{n=0}^{q} b_n \, u_{t-n}$$

folyamatot q-adrendű mozgó átlag folyamatnak hívják és MA(q)-val jelölik. Ez a folyamat mindig stacionárius, ami Gauss folyamatok esetén azt jelenti, hogy $E\, y_t$ és $E\, y_t\, y_{t+s}$ nem függ t-től. Esetünkben $E\, y_t \;=\; 0$, ekkor az utóbbi a folyamat autokovariancia függvénye, ezt c_s-sel fogjuk jelölni. Definició szerint $c_{-s} \;=\; c_s$, és ha $0 \le s \le q$, akkor

$$c_s \;=\; \sum_{n=0}^{q-s} b_n \, b_{n+s},$$

ha pedig $s > q$, akkor $c_s = 0$. Így értelmezhető a

$$C(z) \;=\; \sum_{s=-q}^{q} c_s\, z^s$$

függvény, és $C(z) = B(z)B(1/z)$, ahol $B(z) \;=\; \sum_{n=0}^{q} b_k z^{q-n}$ a folyamat generátor függvénye.

Mivel a normális eloszlást a momentumai egyértelműen meghatározzák, azokra az együtthatókra, amelyekhez ugyanaz a $C(z)$ függvény tartozik, a mozgó átlag folyamat is ugyanaz lesz. Egy idősornál az egyik tipikus statisztikai feladat az előrejelzés, vagy predikció: megfigyeljük a folyamatot egy bizonyos időszakban, és meg szeretnénk mondani, mi következik a jövőre nézve a megfigyeléseinkből. Az elnevezés és a dolog értelme is ugyanaz, mint az időjárás előrejelzése.

Egy előrejelzés csak a megfigyelt adatoktól függhet, esetünkben legyen mondjuk az $(y_0, \ldots, y_t)$ minta a megfigyelés. Ezek az értékek a fehér zaj $(u_{-q}, \ldots, u_t)$ szakaszától függenek. Ha ezeket is felhasználhatnánk, nyilván nem csökkenne a predikciós hibánk. Ha viszont ezeket ismerjük, maga az eredeti minta már nem javíthatja a predikciót. Ez utóbbiak alapján az y_{t+1}-et előállító összeg minden tagja ismert az utolsó kivételével: ez a csonkított összeg lesz a predikciónk, és a hiba az elhagyott tag, b_0 abszolut értéke lesz. Ez tehát egy alsó becslés a predikció hibájára. Ha egy folyamatot többféle együttható rendszerből is előállíthatunk, akkor azok mindegyikéből kapunk alsó becslést a predikció hibájára. Határozzuk meg ezek maximumát.

Jelöljük $B(z)$ gyökeit z_j-vel, $j = 1, \ldots, q$, (ezek nem feltétlenül valósak). Ha gyökök között nincs 0-val egyenlő, akkor

$$C(z) = b_0^2 \prod_{j=1}^{q} (z - z_j)(1/z - z_j) \;=\; b_0^2 z^{-q} \prod_{j=1}^{q} (-z_j) \prod_{j=1}^{q} (z - z_j)(z - 1/z_j),$$

tehát az $M = b_0^2 \prod_{j=1}^{q} (-z_j)$ mennyiség nem függ a $B(z)$ megválasztásától, hiszen épp a fenti alakból láthatóan egy gyököt mindig kicserélhetünk a reciprokára (komplex esetben a konjugáltjával együtt) anélkül, hogy magát a folyamatot megváltoztatnánk, ha közben b_0 értékét úgy igazítjuk, hogy M változatlan maradjon. Ez viszont azt jelenti, hogy érdemes az egynél nagyobb abszolut értékű gyököket a reciprokukra cserélni: akkor lesz b_0 maximális, ha $B(z)$-nek nincs a komplex egységkörön kivül gyöke.

Ha $q = 1$, ez azt jelenti, hogy például $u_t + 2u_{t-1}$ és $2u_t + u_{t-1}$ ugyanaz a folyamat, az elsőből 1, a másodikból 2 adódik a predikciós hibára, ez az utóbbi a

jobb becslés. De ki tudjuk-e számolni a zaj elemeit a folyamatból? Megpróbálhatjuk u_t-t kiszámolni: $u_t = 0.5y_t - 0.5u_{t-1}$. Hasonlóan $u_{t-1} = 0.5y_{t-1} - 0.5u_{t-2}$. E kettő együtt azt adja, hogy $u_t = 0.5y_t - 0.25y_{t-1} + 0.25u_{t-2}$, vagy általában

$$u_t \;=\; \sum_{n=0}^{s-1} \rho^{n+1} y_{t-n} \;+\; \rho^s u_{t-s},$$

ahol $\rho = -1/2$. Véges minta alapján nem tudjuk az utolsó tagot meghatározni, de ha a megfigyelések száma nagy, a hiba elhanyagolható. Ez érvényes minden 1-nél kisebb gyökre, és ha $q > 1$, ezek a visszaszámlálások egymás után hajthatóak végre. Ha pedig van az egységkörön is gyök, a folyamat közelíthető olyan folyamattal, amelyikben ezt a gyököt kicsit elmozdítjuk az egységkör belseje felé. Tehát a predikciós hiba aszimptotikusan mindig b_0 abszolut értéke, feltéve, hogy $B(z)$-nek nincs az egységkörön kivüli gyöke.

Elsőrendű autoregresszív folyamatok Ahogy az előbb a zajt meghatároztuk a folyamatból, úgy származhat maga a folyamat is a zajból: legyen

$$y_t \;=\; \rho y_{t-1} + b_0 u_t.$$

Ha $\mid b_0 \mid < 1$, akkor ez a rekurzió stacionárius folyamatot határoz meg, amely a fehér zajból az

$$y_t \;=\; b_0 \sum_{k=0}^{\infty} \rho^k u_{t-k}$$

végtelen sorral határozható meg (belátható, hogy ez 1 valószínűséggel konvergens). Tehát $E\, y_t \;=\; 0$ és

$$E\, y_t^2 \;=\; b_0^2 \sum_{k=0}^{\infty} \rho^2 \;=\; \frac{b_0^2}{1-\rho^2}.$$

Láttuk, hogy a rekurzió tetszőleges $s > 0$ mellett írható

$$y_t \;=\; \sum_{k=0}^{s-1} \rho^k u_{t-k} \;+\; \rho^s y_{t-s}$$

alakban is, és itt a jobb oldalon álló tagok függetlenek egymástól. Emiatt

$$E y_t\, y_{t-s} \;=\; \rho^s\, E\, y_{t-s}^2 \;=\; \frac{\rho^s\, b_0^2}{1-\rho^2},$$

tehát annak ellenére, hogy a folyamatot definiáló rekurzióban csak két tag szerepel, most az autokovariancia függvény értéke sehol sem nulla (feltéve, hogy ρ nem nulla).

Mindez formálisan változtatás nélkül komplex értékű valószínűségi változókra is elmondható. Csak egy dologra kell vigyázni: ha azt akarjuk, hogy a normális eloszlást továbbra is egyértelműen meghatározzák a momentumai, most fel kell tennünk, hogy a valós és komplex részek kovariancia mátrixa megegyezik, és a két rész kereszt-kovarianca mátrixa antiszimmetrikus. Ez így is lesz, ha a komplex standard normális változót úgy definiáljuk, hogy a valós és képzetes része független standard normálisok $1/\sqrt{2}$-ed része. Ekkor a valós részek kovarianciája egyenlő lesz a kovarianciák valós részével. Ha $\rho = r\,e^{i\omega}$, akkor $\rho^s = r^s\,e^{is\omega}$, tehát az autokovariancia függvény exponenciálisan lecsengő trigonometrikus függvény.

ARMA folyamatok Legyen általában

$$\sum_{n=0}^{p} a_n y_{t-n} \;=\; \sum_{n=0}^{q} b_n u_{t-n},$$

ahol az $A(z) = \sum_{n=0}^{p} a_n z^{p-n}$ polinom gyökei a komplex egységkör belsejében vannak (és $a_0 = 1$), a $B(z) = \sum_{n=0}^{q} b_n z^{q-n}$ polinomnak nincs az egységkörön kivül gyöke (mint láttuk, ez nem korlátozó feltevés). Akkor először is a jobb oldalon álló folyamat MA(q), és az $A(z)$ polinomra tett feltevés mellett ebből lépésről lépésre egy-egy geometriai sor által adott együtthatókkal végtelen mozgó átlag formájában újabb és újabb stacionárius folyamatok állíthatóak elő, végül is kapjuk magát az y_t folyamatot: ezt ARMA folyamatnak hívjuk, és ARMA(p, q)-val jelöljük. Ha $q = 1$, speciálisan a p-ed rendű autoregresszív folyamatot kapjuk, ezt AR(p)-vel jelöljük.

Állapotteres leírás Szokás többdimenziós ARMA folyamatokról beszélni: a fenti alakban a folyamatok vektorok, az együtthatók mátrixok. Kálmán Rudolftól származik ezeknek egy egyszerűbb és szemléletesebb leírása. Induljunk ki a p-dimenziós autoregresszív folyamatból: legyen

$$x_t \;=\; Ax_{t-1} \,+\, Ku_t,$$

ahol u_t független, q-dimenziós standard normálisok sorozata, A $p \times p$, K $p \times q$ méretű mátrixok. Ezt az alakot rekurzíven alkalmazva kapjuk, hogy

$$x_t \;=\; A^s x_{t-s} \,+\, \sum_{n=0}^{s-1} A^n K u_{t-n}.$$

Meg kell tehát vizsgálnunk, hogy mi történik, ha egy négyzetes mátrixot hatványozunk. Ha az A mátrix saját értékei különbözőek, felírható $S\Lambda S^{-1}$ alakban, ahol

Λ diagonális. Ennek k-adik hatványa $S\Lambda^k S^{-1}$, tehát a konvergencia feltétele az, hogy a saját értékek a komplex egységkör belsejében legyenek. Ez általában is a konvergencia feltétele, és ezt a továbbiakban mindig feltesszük. Ekkor

$$x_t \ = \sum_{n=0}^{\infty} A^n K u_{t-n},$$

és a jobb oldalon álló összeg egy valószínűséggel konvergens. Ez a folyamat stacionárius Markov folyamat, és autokovariancia függvénye a fenti előállítás alapján

$$Ex_t x_{t-s}^T \ = \ A^s P,$$

ahol $P = Ex_t x_t^T$, és T a transzponálás jele, hiszen a fenti előállítás szerint folyamat jelene független a zaj jövőjétől. Maga a P kovariancia mátrix az eredeti rekurzióból és a stacionaritásból fakadó

$$P \ = \ APA^T + KK^T$$

úgynevezett Ljapunov egyenletből számolható ki. Ez lineáris egyenlet, ami iterációval is megoldható. Legyen C $q \times p$ méretű mátrix, és legyen $q \leq p$. Határozzuk meg az x_t folyamat

$$y_t \ = \ Cx_t$$

alakú lineáris függvényét. Ez

$$y_t \ = \sum_{n=0}^{\infty} CA^n K u_{t-n},$$

alakban állítható elő, és autokovariancia mátrixa

$$Ey_t y_{t-s}^T \ = \ CA^s PC^T \ = \ CA^s B,$$

ahol $B = PC^T$.

KÁLMÁN SZŰRÉS Tegyük fel, hogy egy q-dimenziós stacionárius y_t folyamatról tudjuk, hogy autokovariancia függvénye

$$Ey_t y_{t-s}^T \ = \ CA^s B$$

alakú alkalmas $q \times p, p \times p, p \times q$ méretű C, A, B mátrixok mellett. Az $x_0 = 0$ kezdeti értékből kiindulva lépésről lépésre meghatározzuk az $y_1, \ldots, y_t$ vektoroknak azt az x_t lineáris függvényét, amelyre

$$Ex_t y_{t-s}^T \ = \ A^s B, \quad s = 0, \ldots, n-1$$

teljesül, és amelyből y_t maga Cx_t alakban olvasható ki. Miért lehetséges ez? Egyáltalán, milyen kereszt-kovariancia írható elő egy adott vektor-rendszer lineáris függvényére? Ha minden vektort egyetlen hosszú vektorrá fűzünk össze, akkor a régi és új rendszer eloszlását a nagy vektor kovariancia mátrixa határozza meg. Látható, hogy az előállíthatóság szükséges és elégséges feltétele az, hogy az új kovariancia mátrix rangja ne legyen nagyobb a régiénél: ekkor ugyanis az új mátrix kereszt-kovariancia részében az oszlopvektorok előállíthatóak a régi oszlopvektorok lineáris függvényeként, és épp ez a lineáris kombináció adja az új vektor koordinátáit a régiek lineáris függvényében.

Ha a magyarázó változók kovariancia mátrixa teljes rangú, akkor a kereszt-kovariancia tetszés szerint választható. Tegyük fel, hogy az $y_1, \ldots, y_t$ vektorok együttes kovariancia mátrixa minden $t > 0$ mellett teljes rangú. Ekkor a kívánt x_t létezik, és mivel $C\,x_t$ kereszt-kovarianciája az $(y_s,\, 1 \leq s \leq t)$ változókkal ugyanaz, mint az y_t vektoré, e két változó megegyezik. Hasonlóan látható, hogy (Ax_{t-1}, CAx_{t-1}) kereszt-kovarianciái az $(y_s,\, 1 \leq s < t)$ változókkal ugyanazok, mint az (x_t, y_t) változóké, tehát az (x_t, y_t) változóknak az $(y_s,\, 1 \leq s < t)$ változókra vonatkozó $(\hat{x}_t, \hat{y}_t)$ feltételes várható értéke (Ax_{t-1}, CAx_{t-1}).

Jelöljük x_t kovariancia mátrixát P_t-vel, $\hat{x}_t$ kovariancia mátrixát $\hat{P}_t$-vel. Akkor

$$\hat{P}_t \;=\; AP_{t-1}A^T.$$

Jelöljük a $v_t \;=\; y_t - CAx_{t-1}$ innovációs hiba kovariancia mátrixát D_t-vel, az x_t változóval való kereszt-kovarianciájat B_t-vel. Akkor

$$B_t \;=\; B - \hat{P}_t C^T,$$

és

$$D_t \;=\; CB_t.$$

Legyen továbbá $u_t \;=\; D_t^{-1/2} v_t$ a standardizált innováció, akkor a rendszer egyenlete

$$x_t \;=\; Ax_{t-1} + K_t u_t,$$

ahol

$$K_t \;=\; B_t D_t^{-1/2}.$$

Végül az állapotteres leírás kovariancia mátrixának a rekurziója

$$P_t \;=\; \hat{P}_t + K_t K_t^T.$$

Ezek Kálmán egyenletei. Vegyük észre, hogy a levezetésükhöz nem használtuk fel az eredeti (stacionárius) állapotteres leírást. A jelölés talán kicsit félrevezető is: ez az x_t nem ugyanaz, mint ami a stacionárius leírásban szerepel, csak ha t tart a végtelenbe, akkor tűnik el a kettő között a különbség. A K_t Kálmán-mátrix esetében a jelölés is tükrözi a különbséget: ez tart K-hoz, P_t tart P-hez, D_t tart a végtelen múltra vonatkozó predikciós hiba kovariancia mátrixához.

A fenti egyenletek statisztikai jelentősége a folyamat likelihood függvényének a meghatározásában van: a standardizált innováció alapján ez egyszerűen felírható.

FELADATOK:

8.1. **(Álb)** Az állapotteres leírás paramétereinek a becslése.

8.2. **(Hka)** Adottak a pozitív definit $A_1, \ldots, A_k \quad N \times N$ méretű mátrixok, ahol $k \leq N$. Meghatározandó az N-dimenziós tér $X_1, \ldots, X_k$ ortonormált rendszere úgy, hogy

$$\sum_{i=1}^{k} X_i^T A_i X_i$$

maximális legyen, ahol T a transzponálás jele.

8.3. **(Gme)** Adjunk becslést egy Gauss-Markov folyamat paramétereire, ha csak a folyamat értékeinek az egész részét tudjuk megfigyelni.

8.4. **(Fdm)** Adjunk becslést egy folytonos idejű ARMA folyamat paramétereire, ha a folyamatot valamilyen (nem egyenletes) $t_1, t_2, \ldots, t_n$) időpontokban tudjuk megfigyelni.

8.5. **(Wlg)** Wiener lepedő generálása.

IRODALOM:

Tusnády Gábor-Ziermann Margit: Idősorok analízise, Műszaki Könyvkiadó, 1986

Michaletzky György-Tusnády Gábor(1987): Többdimenziós idősorok állapotteres leírása, Alkalmazott Matematikai Lapok, 13/3-4, 231-284

P.E. Caines: Linear stochastic systems, Wiley 1988

9. MATEMATIKAI GENETIKA

TÉMÁK: a fehérjék kódolása, Mendel törvényei, mérhető mennyiségek öröklődése, a küszöb modell, genetikai tanácsadás, az általános modell, mutáció és szelekció.

T C | | | | | T C
| T | C | A | G |

A G								A G
	F F	S S	Y Y	C C				
T								T
	L L	S S	- -	- W				
	L L	P P	H H	R R				
C								C
	L L	P P	Q Q	R R				
	I I	T T	N N	S S				
A								A
	I M	T T	K K	R R				
	V V	A A	D D	G G				
G								G
	V V	A A	E E	G G				
T C								T C
	T	C	A	G				
A G								A G

A FEHÉRJÉK KÓDOLÁSA A DNS egy négybetűs ABC-ben van írva, a betűi: T, C, A és G. Ezek közül kettő-kettő szorosan összetartozik: a T és A illetve a C és a G, ezek a kettős spirálban egymás párjai. A négy betű közül kettő valamivel kisebb vegyületet jelöl, ezek a T és C, és ezeket pirimidineknek nevezik, a másik kettő, a C és G nagyobb, ezek a purinok. Ez a négy betű egy körülbelül húsz jelből álló indító jel után hármasával fehérjéket kódol az úgy nevezett genetikai kód szerint.

MENDEL TÖRVÉNYEI Ez a kódtábla húsz-harminc éve ismert, amikor Mendel a múlt század második felében a törvényeit kimondta még gyakorlatilag semmit sem lehetett tudni a DNS-ről. A törvények azonban ma is érvényesek, ha a géneket a fehérjékkel azonosítjuk bennük. Ennek tükrében kezdetben az ember nem érti, hogyan lehetséges hogy a géneknek csak kevés, általában két lehetséges értékük van, hiszen egy-egy fehérje több száz aminosavból áll. A magyarázat az, hogy az élő szervezetek konzervatívok: egy-egy funkcióra csak nagyon kevés verziót, mutánst fogadnak el.

A továbbiakban néhány egyszerűsítő feltevéssel élünk: feltesszük, hogy egy-egy funkcióra csak két lehetséges gén van, a géngyakoriságok nem változnak a populációban, és az ivarsejteket kialakító úgynevezett számcsökkentő osztódásban a különböző helyeken (locusokon) elhelyezkedő gének egymástól függetlenül, $\frac{1}{2}$ valószínűséggel választódnak ki a DNS két ágán ülő gének közül. Feltesszük még, hogy nincs irányított párválasztás.

Tegyük fel továbbá, hogy az a tulajdonság, amit vizsgálunk szintén kétértékű, és azonosítható a két lehetséges génnel. Az egyszerű szóhasználat kedvéért ezeket fehérnek és feketének mondjuk. Ez azt jelenti, hogy ha a DNS mindkét ágán fehér gén ül, akkor az egyed fehér lesz, ha mindkettő fekete, az egyed is fekete lesz. Kérdés, milyen lesz az egyed, ha az egyik gén fehér, a másik fekete? Eddig még szimmetrikusak voltak a színek, most tegyük fel, hogy az ilyen tarka génállománynak fehér egyedek felelnek meg. Ez azt jelenti, hogy a fehérség a domináns: mihelyt az egyik gén fehér, az egyed máris fehér lesz, és a feketeség a recesszív: csak akkor lesz az egyed fekete, ha mindkét gén fekete.

Vizsgáljuk meg most e két tulajdonság öröklődését. Jelöljük a fehér gén gyakoriságát p-vel, a feketéét q-val, a fehérség populációbeli valószínűségét P-vel, a feketeségét Q-val. Akkor $P = p^2 + 2pq$, és $Q = q^2$. Ha egy egyed fehér, akkor $\alpha = p^2/P$ valószínűséggel mindkét génje fehér, vagyis az egyed homozygota, és $\beta = 1 - \alpha = 2pq/P$ valószínűséggel az egyed heterozygota, vagyis egyik génje fehér, a másik fekete. Ha egy egyed fekete, akkor biztosan homozygota. Ha mindkét szülő fehér, akkor $\alpha^2 + 2\alpha\beta$ valószínűséggel legalább az egyik homozygota, és akkor a gyerekeik biztosan fehérek lesznek. A maradék β^2 valószínűséggel mindketten heterozygoták, és ekkor a gyerekük $\frac{1}{4}$ valószínűséggel lehet fekete. Ha az egyik szülő fehér, a másik fekete, a gyerekük $\alpha + \beta/2 = p/P$ valószínűséggel lesz fehér. Két fekete szülőnek minden gyereke fekete lesz.

Mérhető mennyiségek öröklődése Legyen X egy genetikaliag meghatározott mérhető mennyiség, és jelöljük az X-et meghatározó locusokon ülő gének mátrixát G-vel. Ez utóbbi egy véletlen mátrix, amelynek az elemei két lehetséges értéket vesznek fel. Jelöljük ezeket 0-val és 1-gyel, a locusok számát L-lel. Akkor G-nek L sora és 2 oszlopa van. Jelöljük X-nek a G i-edik sorában és j-edik oszlopában álló elemére vett feltételes várható értékét a_{ij}-vel. Tegyük fel, hogy X egyenlő ezek összegével, és az egyes locusokon ülő gének függetlenek egymástól. Definició szerint az a_{ij} mennyiségek kétértékű függvények; $a_{ij} = f_{ij}(g_{ij})$, ahol g_{ij} G i-edik sorának és j-edik oszlopának az elemét jelöli. Legyen G' a G génállományú egyed valamely

rokonának a génállománya, és X' a megfelelő mérhető mennyiség. Akkor

$$COV(X, X') = \sum_{i=1}^{L} \sum_{j=1}^{2} \mathrm{C}OV(f_{ij}(g_{ij}), f_{ij}(g'_{ij}),$$

ahol COV a kovarianciát, és g'_{ij} G' i-edik sorának és j-edik oszlopának az elemét jelöli.

Mit mondhatunk általában G és G' együttes eloszlásáról? Mivel az elemek egymástól függetlenek, csak az azonos pozicióban álló elemek között lehet sztochasztikus összefüggés. Definició szerint azokat az egyedeket tekintjük rokonoknak, akiknek van közös ősük. A szülők egy-egy konkrét génüket $\frac{1}{2}$ valószínűségel vagy átadják az utóduknak, vagy nem. Tovább menve egy k-adik ős egy-egy konkrét génjét $(\frac{1}{2})^k$ valószínűségel vagy átadja az utódjának, vagy nem. Két rokon egy-egy kiszemelt génje vagy ugyanattól a közös ősüktől származik, és ekkor az a két gén azonos, vagy nem, és akkor az a két gén független egymástól.

Az azonosság valószínűségét a következő módon határozhatjuk meg. Kössük össze képzeletben éllel az elsőfokú rokonokat: a testvéreket és a szülő-gyerek párokat. Menjünk el az így kapott gráfban az egyik egyedtől a másikig a legrövidebb úton, és jelöljük a lépések számát R-rel. Azt mondjuk, hogy a vizsgált egyedek R-edfokú rokonok. Akkor $(\frac{1}{2})^R$ a valószínűsége annak, hogy a két rokon valamely kiszemelt génje a közös őstől származik. Ha ugyanis egyeneságú rokonságról van szó, ezt már láttuk. Ha különben a közös ősig U illetve U' lépés vezet az egyedektől, akkor a közös ős egy-egy génje $(\frac{1}{2})^U$ illetve $(\frac{1}{2})^{U'}$ valószínűséggel található meg az utódokban, de most a közös ősnek mind a két génjét számításba kell vennünk $(R = U + U' - 1)$.

(Feltesszük, hogy nincs vérrokonság, vagyis két egyednek legfeljebb egy közös őse van. Ez a közös ős persze valójában egy házaspár. Feltesszük, hogy az utódokban identifikálható és elkülöníthető az apai és anyai gén.)

Ezzel beláttuk, hogy a mondott feltételek mellett az R-edfokú rokonok mérhető mennyiségei közti korrelációs együttható $(\frac{1}{2})^R$. Ha a vizsgált mennyiség a környezettől is függ, felbontható $X = Y + Z$ alakban, ahol Y a genetikailag meghatározott rész, és Z a környezet Y-tól független hatása. Feltesszük továbbá, hogy a rokonokat érő környezeti hatások függetlenek. Ez persze vitatható, hiszen az életmód közös elemei sok szálon teremthetnek sztochasztikus kapcsolatot. Ennek a nehézségnek egy lehetséges megoldása az ikrek kutatása: ha sikerül különválasztani az egypetéjűeket a kétpetéjűektől, akkor reálisan külön lehet választani a környezti hatásokban jelentkező korrelációt a genetikailag meghatározott korrelációtól.

Az X mérhető mennyiség és a benne levő Y genetikailag meghatározott rész korrelációs együtthatóját az öröklődés mértékének hívják, és h^2-tel jelölik. Általában az R-edfokú rokonok mérhető mennyiségének a korrelációs együtthatója $h^2/2^R$.

A KÜSZÖB MODELL A bináris mennyiségek öröklődését visszavezethetjük az mérhető mennyiségek öröklődésére, ha feltételezzük, hogy van egy genetikailag meghatározott nem mérhető mennyiség, és egy küszöb, amelynek az egyik oldalán a bináris mennyiség az egyik értékét veszi fel, a másikon a másikat. Ha feltesszük, hogy az érintett locusok száma nagy, akkor feltehetjük azt is, hogy ez a virtuális mérhető mennyiség normális eloszlású. Ekkor egy kiterjedt család tagjaihoz tartozó virtuális mennyiségek együttes eloszlását meghatározzák a korrelációs együtthatók (azt is feltehetjük, hogy a szórások egységnyiek), tehát csak a küszöböt és az öröklődés mértékét kell meghatározni.

Ha a vizsgált bináris mennyiség valamilyen a gyerekkel veleszületett rendellenesség, akkor erre a célra családvizsgálatokat szokás végezni: amilyen széles körben csak lehet össze kell gyűjteni a rendellenes gyerekek rokonai körében a rendellenességekre vonatkozó adatokat. Az ilyen felméréseket hasznos lenne családonként kiértékelni, ami azonban igen munkaigényes feladat. Helyette elfogadható eredményt kapunk, ha az adatokat rokonsági típusok szerint csoportosítjuk.

GENETIKAI TANÁCSADÁS Egy genetikai tanácsadáson a szülők leendő gyermekük várható rendellenességét szeretnék megtudni. Elmondják a családban meglevő rendellenességeket, és ezek alapján kell kiszámolni annak a feltételes valószínűségét minden egyes rendellenességre, hogy a gyermeknek az a rendellenessége megleszе. Maga a küszöb modell kiterjeszthető több rendellenességre: a háttérváltozóknak kell meghatározni az egymás közti korrelációs együtthatóját. Ha ez r, akkor a megfelelő együttható R-edfokú rokonokra $rh^2/2^R$. A technikai nehézséget itt a sokdimenziós normális eloszlásra vonatkozó feltételes eloszlása okozza, ez jelenleg nincs megnyugtatóan megoldva.

AZ ÁLTALÁNOS MODELL Mendell törvényei és a küszöb modell között a véges sok locustól függő rendellenességek széles skálája található. Legyen egy adott rendellenességre $p(G)$ annak a valószínűsége, hogy egy G genetikai állományú egyednek kifejlődik az adott rendellenessége. A genetikai tanácsadáskor egy adott családfán elhelyezkedő egyedekre kell meghatároznuk annak a valószínűségét, hogy pont azoknak legyen meg a szóban forgó rendellenességük, akiknek megvan. A családfán ülő egyedek genetikai állományának az együttes eloszlását annak alapján határozhatjuk

meg, hogy az ivarsejteket kialakító számcsökkentő osztódáskor G-ből olyan L-dimenziós bináris H oszlopvektor keletkezik, amelynek i-edik sorában $\frac{1}{2}$ valószínűséggel g_{i1}, $\frac{1}{2}$ valószínűséggel g_{i2} áll, és az egyes koordináták kiválaztása egymástól függetlenül történik. (A valóságban ez nincs így az úgynevezett "crossing over" miatt.) Tehát az ősöktől kiindulva és az utódok felé haladva felépíthetjük a teljes családfa genetikai állományának együttes eloszlását. Testvérekre a H-kat előállító randomizálások függetlenek egymástól.

Ez az eljárás azonban kombinatorikusan felrobban. Kellő óvatossággal redukálni lehet a lépésszámot a dolgot a széleken kezdve és mindig csak annyi feltételes valószínűséget tartva a kezünkben, amennyire a továbbhaladáshoz feltétlenül szükségünk van. Ez az eljárás ugyanazon a gondolaton alapszik, mint a rejtett Markov láncokra alkalmazott EM algoritmus számolása és a Kálmán szűrés.

Ha több rendellenességek vizsgálunk egyidejűleg, minden változatlan, csak $p(G)$ vektor lesz.

MUTÁCIÓ ÉS SZELEKCIÓ Amit idáig elmondtam, az akkor igaz, ha a vizsgált tulajdonságoknak nincs szelektív hatásuk: az hogy egy egyednek lesz-e utóda, és ha igen, mennyi, az nem függ a genetikai állományától. Ez azonban általában nincs így. Tegyük fel, hogy $p(G)$ annak a valószínűsége, hogy egy G genetikai állományú egyednek lesz utóda, és az egyszerűség kedvéért tegyük fel, hogy mindenkinek maximum egy utóda lehet, és a populáció végtelen, benne a generációk fix időközökben szinkronizáltan váltják egymást.

Része a kérdésnek, hogy kötelezően társulnia kell-e a szelekciónak mutációval: bizonyos esetekben ugyanis ha a "rossz" gének mint egy fürdőkádból kifolynak a populációból, és mi mégis olyan modellt szeretnénk találni, amelyben van stacionárius állapota a populációnak, és abban pozitív valószínűséggel lesznek rendellenes egyedek, akkor szükség lehet mutációra. Legyen α annak a valószínűsége, hogy egy 0 értékű génből 1 lesz, és legyen β annak a valószínűsége, hogy egy 1 értékű génből 0 lesz. (Az egyszerűség kedvéért mondhatjuk, hogy ezek nem függenek a locusoktól, de ennek nincs különösebb jelentősége.) Tegyük fel, hogy az egyes géneket érő mutációk függetlenek egymástól.

Ezen feltételek mellett a populációban található ivarsejteknek (haploidoknak) minden egyes generációban lesz egy, az arra generációra jellemző $Q(H)$ eloszlásuk: ha L locus van, ez 2^L nem negatív szám, amelyek összege 1. Az egyedek génállományának az eloszlása akkor $P(G) = Q(H_1)Q(H_2)$, ha $G = (H_1, H_2)$. Legyen $M(H_1, H_2, H)$ annak a feltételes valószínűsége, hogy a (H_1, H_2) génállományú egyed-

nek a számcsökkentő osztódás során a mutációt is figyeklembe véve H génállományú ivarsejtje keletkezik. Akkor az új generáció ivarsejtjeinek az eloszlása

$$Q'(H) = \sum_{H1,H2} Q(H_1)Q(H_2)p((H_1,H_2))M(H_1,H_2,H).$$

Ennek alapján eloszlások sorozatát állíthatjuk elő úgy, hogy a $Q = Q_n$ eloszláshoz előbb a fenti formulával meghatározzuk a Q' mennyiségeket, majd ezeket leosztjuk az összegükkel, feltéve, hogy az nem nulla. Ez az új eloszlás lesz Q_{n+1}.

Kérdés, mi történik az iteráció során? Konvergens-e az eloszlások sorozata? Én erre nem ismerek elégséges feltételt. Ellenpéldát sem tudok, amikor ne volna konvergens (például aszimptotikusan ciklizálna az iteráció). Azt sem tudom, hogyan lehet a stacionárius eloszlásokat megkeresni. Az a módszer nyilván működik, hogy különböző kezdeti eloszlásokból kiindulva meghatározzuk a határértékeket, de ez valószínűleg nem elég hatékony.

Egy speciális eset Egy aránylag egyszerű eset a következő. Tegyük fel, hogy ha eltérő homozygoták vannak a génállományban, akkor nincs utód, különben van. Tegyük fel, hogy $\alpha = \beta$, és jelöljük ezek közös értéket ε-nal. Mivel a locusok ekvivalensek, a haploidot meghatározza a 0 gének száma, és a generációk iterációja a 0 és L közötti egészeken adott eloszlásokat transzformálja. Ha $L = 1$, nincs szelekció, és ε minden pozitív értéke mellett egyetlen stacionárius megoldás van: $Q(0) = Q(1) = 0.5$. Ha $L > 1$, akkor van egy L-től függő ε_L konstans, és ha $\varepsilon < \varepsilon_L$, akkor három stacionárius megoldás van: az egyik szimmetrikus, és ennek két oldalán megjelenik két egymásra szimmetrikus a 0-át illetve L-et preferáló eloszlás, amelyekben az antiszimmetria mértéke annál nagyobb, minél kisebb ε ($\varepsilon_2 = 0.07, \varepsilon_3 = 0.14, \varepsilon_4 = 0.20, \varepsilon_5 = 0.26, \varepsilon_6 = 0.30$). Ha $\varepsilon > \varepsilon_L$, akkor csak a szimmetrikus stacionárius eloszlás létezik. Ezeket a tapasztalatokat a következő programmal szereztem. Érdekes futási tapasztalat, hogy ε_L környékén nagyon lassú a stacionárius eloszláshoz való konvergencia, de csak az aszimmetrikus oldalon.

```
program gen;
uses crt;
const maxl=20;
      kicsi=-70;
      mineps=1.0E-30;
      minszum=1.0E-24;
      minszim=1.0E-18;
```

```
      maxlep=32000;

type bint=array[0..maxl,0..maxl] of real;
      fakt=array[0..maxl] of real;
      magt=array[0..maxl,0..maxl,0..maxl] of real;

var mag:magt;
      bin:bint;
      fak,p,q:fakt;
      nlep,i,j,k,l,m,n,mut:integer;
      x,y,z,u,v,lnee,lne,eps,lnk:real;
      bfut,bszim:boolean;
      sum,szum,szumm,szim:real;

procedure faktor;
begin
fak[0]:=0; for i:=1 to maxl do fak[i]:=fak[i-1]+ln(i);
end;

procedure binom;
begin
for i:=0 to maxl do for j:=0 to i do bin[i,j]:=fak[i]-fak[j]-fak[i-j];
end;

procedure atmenet;
var r,s,ss,t,z,u,v,w,ii,jj,kk:integer;
      fordit:boolean;
begin
for i:=0 to maxl do for j:=0 to maxl do for k:=0 to maxl do mag[i,j,k]:=0;
for i:=0 to l do for j:=0 to l do begin fordit:=false;
ii:=i;jj:=j;s:=i+j;ss:=s;
if(s>l) then begin fordit:=true;ii:=l-i;jj:=l-j;ss:=ii+jj;end;
t:=l-ss; for z:=0 to ss do for w:=0 to z do begin
r:=ss-z; u:=t+r; for v:=0 to u do begin
kk:=w+v;k:=kk; if fordit then k:=l-kk;
x:=bin[l,ss]+bin[ss,z]+bin[z,w]+bin[u,v]
+(w+u-v)*lnee+(v+z-w)*lne-ss*lnk-bin[l,i] -bin[l,j];
```

```
if(x<kicsi) then x:=kicsi; x:=exp(x); mag[i,j,k]:=mag[i,j,k]+x;
end; end; end;end;

procedure lep;
begin nlep:=nlep+1; for i:=0 to l do q[i]:=0;
for i:=0 to l do for j:=0 to l do for k:=0 to l do
q[k]:=q[k]+p[i]*p[j]*mag[i,j,k];
sum:=0; for k:=0 to l do sum:=sum+q[k]; szum:=0; for k:=0 to l do begin
x:=q[k]/sum;szum:=szum+sqr(p[k]-x);p[k]:=x;end;
if(nlep=100*(nlep div 100)) then begin gotoxy(1,mut);
writeln(nlep:6,' SZUM',szum:30,szumm-szum:30); szumm:=szum;
end;if(szum<l*minszum) then bfut:=false;
if(nlep>maxlep) then bfut:=false;
end;

{FOPROGRAM}
begin
lnk:=ln(2);clrscr; faktor;binom;
repeat
write(' LOCUS : ');readln(l);
if(l>0) then begin write(' EPSILON : ');readln(eps);
if(eps<mineps) then eps:=mineps; lnee:=ln(1-eps);lne:=ln(eps);
atmenet; clrscr;
bfut:=true; for i:=0 to l do p[i]:=0;p[0]:=1;nlep:=0; mut:=1;szumm:=2.0;
repeat lep; if keypressed then bfut:=false;
until not bfut; writeln;
writeln(nlep:7,' LOCUS',l:5,' EPSZILON',eps:12:8,' SZUM',szum:30);
for i:=0 to l do writeln(i:5,p[i]:15:9);
szim:=0.0;for i:=0 to l do szim:=szim+sqr(p[i]-p[l-i]);
writeln(' SZIMMETRIA',szim:30);
if(szim<l*minszim) then bszim:=true else bszim:=false;
if not bszim then begin
bfut:=true; for i:=0 to l do p[i]:=0;p[l]:=1;nlep:=0; mut:=8;szumm:=2.0;
repeat lep; if keypressed then bfut:=false;
until not bfut; writeln;
writeln(nlep:7,' LOCUS',l:5,' EPSZILON',eps:12:8,' SZUM',szum:30);
```

```
for i:=0 to l do writeln(i:5,p[i]:15:9);
bfut:=true; for i:=0 to l do p[i]:=1/(1+l);nlep:=0; mut:=15;szumm:=2.0;
repeat lep; if keypressed then bfut:=false;
until not bfut; writeln;
writeln(nlep:7,' LOCUS',l:5,' EPSZILON',eps:12:8,' SZUM',szum:30);
for i:=0 to l do writeln(i:5,p[i]:15:9);end; end;
until(l=0); end.
```

Általában feltehetjük, hogy csak a tiszta homozygota génállomány képes utódképzésre: ekkor minden haploid önmagában egy stacionárius eloszlást reprezentál, körülötte a mutáció mértékétől függően lecsengő eloszlásokkal, és ezek egymásba átkódolhatóak. Így egyetlen közös mutációs küszöb van, amely ezeket az aszimmetrikus eloszlásokat elválasztja a szimmetrikus stacionárius eloszlástól.

Monoton eset Természetes feltevés, hogy a két lehetséges gén közül az egyik jó, a másik rossz. Ilyenkor értelmezhető a génállomány rendezése is: egyik génállomány rosszabb a másiknál, ha benne mindenütt rossz gén ül, ahol a másikban rossz gén van. Feltehetjük, hogy rosszabb génállomány mellett kisebb az utódképzés valószínűsége. Az egyszerűség kedvéért azt is feltesszük, hogy most csak jó génből lesz rossz a mutáció során, de a rossz gének rosszak maradnak. Ekkor van egy triviális stacionárius megoldás: minden gén rossz. Kérdés, van-e más? Ha az utódképzés valószínűsége csak a rossz gének számától függ, ismét csak a 0 és L közötti egészeken adott eloszlásokat kell iterálni. A numerikus tapasztalatok szerint mindig van egy egyértelműen meghatározott nem triviális stacionárius eloszlás is.

Feladatok:

9.1. **(Gtp)** A genetikai tanácsadás programja.

9.2. **(Dme)** Egy N-szintű diadikus fa csúcsán levő 2^N fészekbe elhelyezünk tetszés szerinti eloszlásból származó független egyforma eloszlású nem negatív egészeket, az összes többi csúcs értékét nullával indítjuk. A fán lefelé haladva lépésről lépésre minden csúcsban végrehajtjuk a következő eljárást. Ha a csúcs értéke legalább 2, abból kivonunk egyet, és a maradékot hozzádjuk a csúcs alatti csúcs számához, különben nem csinálunk semmit. Egy szintre csak akkor lépünk, ha a felette levő szinten már minden csúccsal végeztünk. Meghatározandó a gyökérben keletkező szám eloszlása.

9.3. **(Dbp)** Egy N-szintű diadikus fa csúcsán levő 2^N fészekbe elhelyezünk tetszés szerinti természetes számokat. Lépésről lépésre minden csúcsban végrehajt-

juk a következő eljárást. Feldobunk annyi érmét amennyi a csúcs feletti két szám összege, és a fejek számához hozzáadunk egy tőlük és minden mástól független λ paraméterű Poisson eloszlású véletlen számot. Meghatározandó a gyökérben keletkező szám feltételes eloszlása arra a feltételre, hogy minden szóba jövő szám értéke legfeljebb K, ahol K egy adott pozitív egész.

9.4. **(Cms)** Legyen T pozitív egész, μ pozitív valós szám. Legyen $\mathcal{P}$ azon $\pi = (p_0, p_1, \dots)$ végtelen sorozatok halmaza, melyekre

$$p_k \geq 0, k = 0, 1, \dots, \sum_{k=0}^{\infty} p_k = 1, \sum_{k=0}^{T} p_k > 0$$

teljesül. Tekintsük a következő három operátort $\mathcal{P}$-n:

Legyen S a szelekció operátora: ez a π eloszlású X valószínűségi változóhoz a $q(k) = P(X = k \mid X \leq T), k = 0, 1, \dots$ feltételes eloszlást rendeli.

Legyen M a meiozis és mutáció együttes hatásának az operátora: ez a π eloszlású X valószínűségi változóhoz az $U = Y + Z$ valószínűségi változó eloszlását rendeli, ahol Y, Z függetlenek, $Y \quad X, \frac{1}{2}$ paraméterű binomiális, Z μ paraméterű Poisson eloszlású valószínűségi változó.

Legyen C a megtermékenyítés operátora: ez a π eloszlású X valószínűségi változóhoz az $Z = X+Y$ valószínűségi változó eloszlását rendeli, ahol Y az X-től független, vele azonos eloszlású valószínűségi változó.

E három operátor egymás utáni alkalmazása legyen G: $G = CMS$, és legyen $\pi_t = G^t\pi$.

Adjunk numerikus tájékozódás alapján választ a következő kérdésekre:

– konvergens-e a π_t sorozat?

– függ-e a π_t sorozat határértéke π-től?

– hogyan függ a konvergencia sebessége a T, μ paraméterektől?

– hogyan függ a határeolszlás a T, μ paraméterektől?

– egyértelmű-e a $G\pi = \pi$ egyenlet megoldása?

9.5. **(Szk)** Legenek a s_{ij} valós számok minden nem negatív egész (i, j) mellett nem negatívak, és legyen a $\sum_{j=0}^{\infty} s_{ij}$ összegek értéke minden i-re legfeljebb 1. Mondjuk azt, hogy a $\left(q_i = \sum_{j=0}^{\infty} s_{ij}p_j, i = 0, \dots\right)$ sorozat a $(p_i,\ i = 0, \dots)$ sorozat szűréséből származik. Ha $(p_i,\ i = 0, \dots)$ eloszlás, és a szűréséből származó sorozat elemeinek összege pozitív, azzal osztva ismét eloszlást kapunk. Mi történik, ha ezt az operátort és a konvolúciót váltogatva végtelen sokszor alkalmazzuk?

Irodalom:

Tusnády Gábor(1969): A multifaktoriális öröklődés, Matematikai Lapok, 20/3-4, 389-396

A. Czeizel-G. Tusnády: Aetiological studies of isolated common congenital abnormalities in Hungary, Akadémiai Kiadó, 1984

A. Czeizel-L. Telegdi-G. Tusnády: Multiple congenital abnormalities, Akadémiai Kiadó, 1986

J. Komlós-A. Odlyzko-L. Ozarow-L.A. Shepp(1991): On the properties of a tree-structured server process, The Annals of Applied Probability, 1/1, 118-125

S.A.Kauffman: The origin of life, Oxford University Press, 1993

Tusnády Gábor: Mutacio és szelekció, Magyar Tudomány, megjelenés alatt

10. Sztochasztikus kapcsolatok

Témák: gráfok, clustering, ACE, sztochasztikus folyamatok beágyazása, többszempontú optimalizálás,

Gráfok Legyen U és V két véges halmaz, és legyen $\mathcal{P}$ az (U,V) pár feletti páros gráf, vagyis legyen $\mathcal{P}$ az (U,V) pár $((u,v), u \in U, v \in V)$ rendezett párjaiból álló halmaz. Az (u,v) rendezett párt $\mathcal{P}$ élének mondjuk. $\mathcal{P}$-t reprezentálhatjuk egy B mátrixszal, amelynek a sorai U elemeihez, oszlopai V elemeihez vannak rendelve, és az u-hoz rendelt sor és v-hez rendelt oszlop b_{uv} eleme 1 vagy 0 aszerint, hogy az (u,v) pár benne van-e $\mathcal{P}$-ben, vagy sem. Kicsit általánosabban feltehetjük, hogy B elemei tetszőleges nem negatív számok: ekkor ezek a megfelő él "súlyát" jelentik. Vagy egész általánosan feltehetjük, hogy két mérhető tér szorzata felett van egy valószínűségi mértékünk.

Definíció. Azt mondjuk, hogy az $(U,\mathcal{A}),(V,\mathcal{B})$ mérhető terek szorzata feletti P valószínűségi mértéknek az $X : U \mapsto R^k,\ Y : V \mapsto R^k$ mérhető leképezés–pár a beágyazása, ha $EXX^T = I$. Egy (X,Y) beágyazás optimalis, ha tetszőleges (X_1,Y_1) beágyazás mellett

$$E \parallel X - Y \parallel^2 \leq E \parallel X_1 - Y_1 \parallel^2$$

teljesül (R^k a k–dimenziós valós Euklideszi teret, I a $k \times k$ méretű egységmátrixot jelöli).

Maradjunk a mátrixok mellett egy pillanatra. Jelöljük a B mátrix u-adik sorában álló elemek összegét s_u-val, a v-edik oszlop összegét r_v-vel, és tegyük fel, hogy ezek mind pozitívak. Most az $(x_{ui}, y_{vi}, u \in U, v \in V, i = 1,\dots,k)$ számokat keressük úgy, hogy a

$$\sum_{u \in U} s_u x_{ui} x_{uj} \ = \ \delta_{ij}$$

kényszerfeltétel mellett (amelyben δ_{ij} értéke 1 vagy 0 aszerint, hogy $i = j$ teljesül-e vagy sem) az

$$\sum_{u \in U, v \in V} b_{uv} \sum_{i=1}^{k} (x_{ui} - y_{vi})^2$$

összeg minimális legyen.

Adott x_{ui} számok melett az y_{vi} számokban egymástól függetlenül optimalizálhatunk: az $\sum_{u\in U} b_{uv}(x_{ui}-y_{vi})^2$ összeget kell minimalizálnunk, tehát y_{vi} ezeknek az x_{ui} számoknak a súlyozott átlaga: $y_{vi} = (1/r_v)\sum_{u\in U} b_{uv}x_{ui}$. Így az x_{ui} számokban a

$$\sum_{u\in U} s_u \sum_{i=1}^{k} x_{ui}^2 - \sum_{v\in V} r_v \sum_{i=1}^{k} y_{vi}^2 = trX(S-BR^{-1}B^T)X^T$$

célfüggvényt kell minimalizálnunk a $trXSX^T = I$ feltétel mellett, ahol tr a mátrix diagonálisában álló számok összegét jelöli, S, R olyan négyzetes mátrixok, amelyek diagonálisában az s_u, r_v számok állnak és X u-adik sorának i-edik eleme x_{ui}. Látni fogjuk, hogy ekkor a minimumot az $(S-BR^{-1}B^T)$ mátrix k legkisebb saját értékéhez tartozó saját vektor adja.

Egy hipergráfot reprezentálhatunk páros gráffal úgy, hogy B-ben a sorokat a hipergráf csúcsainak, az oszlopokat az éleknek feleltetjük meg. Láttuk, hogy ekkor az éleket a csúcsaik beágyazott értékének a súlypontjába kell beágyaznunk. Közönséges gráfokra az élek költsége az általuk összekötött csúcsok beágyazott képének távolságnégyzetének a fele. Ezt ismét általánosíthatjuk szorzatterek feletti mértékekre.

Definíció. Azt mondjuk, hogy az $(U,\mathcal{A}),(U,\mathcal{A})$ terek szorzata feletti Q mérték szimmetrikus, ha tetszőleges $A, A' \in \mathcal{A}$ mellett $Q((A,A')) = Q((A',A))$. Azt mondjuk, hogy az $(U,\mathcal{A}),(U,\mathcal{A})$ terek szorzata feletti szimmetrikus Q mértéknek az $X : U \mapsto R^k$ mérhető leképezés a beágyazása (röviden: X a Q beágyazása), ha $EXX^T = I$. A beágyazás költsége

$$C(X) = E \parallel X - X' \parallel^2,$$

ahol az X' véletlen vektort úgy kapjuk, hogy az X leképezést a tér második koordinátájára alkalmazzuk (maga X az első koordinátához rendelt vektor). Egy beágyazás optimális, ha

$$C(X) \le C(X_1)$$

teljesül tetszőleges X_1 beágyazásra.

Ha egy közönséges gráfot az A mátrix úgy ír le, hogy $a_{uv} = 1$ vagy 0 aszerint hogy az u, v csúcsok összse vannak-e kötve vagy sem, és s_u az u csúcs foka, akkor a $\sum_{u\in U} s_u x_{ui}x_{uj} = \delta_{ij}$ kényszerfeltétel mellett kell a

$$\sum_{u\in U}\sum_{v\in U} a_{uv}\sum_{i=1}^{k}(x_{ui}-x_{vi})^2$$

célfüggvényt minimalizálni.

Mit jelent ez a feladat? A gráf élei "össze szeretnék húzni" a beágyazásban az egyes csúcsokhoz rendelt pontokat, a kényszerfeltétel pedig ellenáll ennek, "szét szeretné feszíteni" a pontok rendszerét. A két ellentétes hatás kompromisszuma szerint a csúcsok beágyazott értékei úgy helyezkednek el, hogy azok a csúcsok, amelyek között sok él fut közel legyenek egymáshoz. Ha ez lehetséges, ha el tudjuk így helyezni a csúcsokat az euklideszi térben, önmagában az a tény is jellemzi a gráfokat, hogy ez a beágyazás milyen jól képes megmutatani a valódi struktúrájukat. A legegyszerűbb eset az, amikor csoportosíthatóak a csúcsok úgy, hogy az élek többnyire csak egy-egy csoporton belül húzódnak.

CLUSTERING Mondjuk azt hogy, egy k-dimenziós véletlen vektor szimpliciális, ha lehetséges érékeinek a száma k. Ha egy k-dimenziós X véletlen vektorra teljesül, hogy $E\ XX^T = I$, ahol I az egységmátrix, és T a transzponálás jele, akkor X szimplicialítása legyen az $E \parallel X - Z \parallel^2$ mennyiségek infimuma, ahol Z szimpliciális és $E\ ZZ^T = I$. X szimplicialítását $S(X)$-szel jelöljük.

Jelöljük Q optimális k-dimenziós beágyazásának a költségét $B_k(Q)$-val, a szimpliciális beágyazások költségeinek a minimumát $C_k(Q)$-val. Ekkor

$$B_k(Q) \leq C_k(Q)$$

teljesül tetszőleges szimmetrikus Q mértékre, hiszen C_k-ban csak a szimpliciális beágyazásokat engedjük meg.

Természetes kérdés, hogy mennyire éles ez a becslés, azaz becsülhető-e $C_k(Q)$ felülről $B_k(Q)$ alkalmas függvényével. Valószínűleg nem, mert ha Q a "szalag"-gráfból származik (amelyben a $0, 1, ..., n$ csúcsok közül $i = 1, ..., n$ mellett $(i-1, i)$ között fut él), akkor a két mennyiség egyre távolabb kerül egymástól növekvő n mellett. Vizsgáljuk meg rézletesebben ezeket a szimpliciális beágyazásokat!

Jelöljük az $(U, \mathcal{A}), (U, \mathcal{A})$ szorzattér kétváltozós és Q szerint négyzetesen integrálható függvények terét H-val, közülük azok alterét amelyek csak az első koordinátájuktól függnek G-vel, azokét, amelyek a másodiktól, G'-vel. Ezeket az altereket rendre Q teljes illetve koordináta alterének hívjuk.

DEFINÍCIÓ. Ha G és G' a H Hilbert tér tetszőleges alterei, bennük az $\{a_\gamma, \gamma \in \Gamma\}$, $\{a'_\gamma, \gamma \in \Gamma\}$ ortonormált bázisokat csatoltnak nevezzük ha minden $\gamma \in \Gamma$ mellett a_γ-nak G'-n levő merőleges vetülete $\alpha_\gamma a'_\gamma$ és a'_γ-nak G-n levő merőleges vetülete $\alpha_\gamma a_\gamma$ (ilyenkor a két együttható nyilván megegyezik).

LEMMA. Ha $\{a_\gamma, \gamma \in \Gamma\}$, $\{a'_\gamma, \gamma \in \Gamma\}$ a Q koordináta tereinek csatolt bázisai, és Γ valamely k elemü Γ_k részhalmazára

$$\alpha_\gamma \geq \alpha_{\gamma'}$$

teljesül menden $\gamma \in \Gamma_k$,$\gamma' \in \Gamma$ mellett, akkor az a k-dimenziós X leképezés optimális beágyazás, amelynek a koordinátái az $\{a_\gamma, \gamma \in \Gamma_k\}$ függvények.

BIZONYÍTÁS. Mivel a csatolt bázisok elemei ortonormáltak, X beágyazás, és a költsége

$$B_k(X) = \sum_{\gamma \in \Gamma_k} (1 - \alpha_\gamma).$$

Irjuk fel egy tetszőleges beágyazás koordinátáit az $\{a_\gamma, \gamma \in \Gamma\}$, bázisban: kapjuk a $\{C_{i\gamma}, 1 \leq i \leq k, \gamma \in \Gamma\}$ együttható mátrixot, amelyben

$$\sum_{\gamma \in \Gamma} C_{i\gamma}^2 = 1, 1 \leq i \leq k,$$

és a költség

$$2 \sum_{i=1}^{k} \sum_{\gamma \in \Gamma} C_{i\gamma}^2 (1 - \alpha_\gamma) = 2 \sum_{\gamma \in \Gamma} (1 - \alpha_\gamma) W_\gamma,$$

ahol a $W_\gamma = \sum_{i=1}^{k} C_{i\gamma}^2$ súlyok összege k, és mindegyikük legfeljebb 1. Így valóban X költsége minimális.

A lemma állítása szerint

$$B_k(Q) = \sum_{\gamma \in \Gamma_k} (1 - \alpha_\gamma).$$

Rendezzük az $(1-\alpha_\gamma)$ számokat növekvő sorrendbe, és jelöljük őket ebben a sorban $\lambda_1, \lambda_2, \ldots, \lambda_k$-val, az $a_1, a_2, \ldots, a_n$ koordinátájú vektort X_n-nel. Ez utóbbi minden $1 \leq n \leq k$ mellett optimális beágyazás, ilyenkor tehát magasabb dimenzióba lépve mindössze egy újabb koordinátával bővül a korábbi optimális beágyazás. A költségek rendre

$$B_n = \sum_{i=1}^{n} \lambda_i,$$

ahol $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$, tehát a B_n sorozat konvex abban az értelemben, hogy a növekményei nőnek. Mivel a konstans optimális beágyazás $n = 1$ mellett, itt $\lambda_1 = 0$, de elvileg elképzelhető, hogy több λ értéke is 0, vagyis $B_n = 0$ valamilyen $n > 1$ mellett. Jelöljük a legnagyobb ilyen n-et N-nel. Mivel $B_N = 0$, ezért $X_N = X'_N$

teljesül 1 valószínűséggel: az N-dimenziós tér tetszőleges n- elemű $V_1, V_2, \ldots, V_n$ particiója az U tér olyan $U_1, U_2, \ldots, U_n$ particióját határozza meg az

$$U_i = \{u : X_N \in V_i\}, 1 \leq i \leq n$$

összefüggés alapján, amelyre $Q(U_i, U_j) = 0$,ha $i \neq j$. Megfordítva, egy ilyen partició alapján definiálhatjuk X_n-et az

$$X_n(u) = e_i / Q^{\frac{1}{2}}(U_i), \text{ ha } u \in U_i, 1 \leq i \leq n$$

összefüggéssel, ahol $e_1, e_2, \ldots, e_n$ az n-dimenziós tér tetszőleges ortonormált bázisa. (Itt feltesszük, hogy az U_i halmazok valószínűsége pozitív.) Ennek a beágyazásnak is 0 a költsége tehát $n \leq N$, hiszen N a legnagyobb index volt, amelyre $B_n = 0$. Viszont

$$EX_N X_N^T = I$$

miatt fel lehet bontani az N-dimenziós teret N pozitív mértékü részre. Ellenkező esetben ugyanis egy finomodó rácsfelbontásból lépésről lépésre csak a pozitív valószínűségű szemeket meghagyva végül is N-nél kevesebb pozitív mértékű határpontot kapunk, ami nem feszíti ki az N-dimenziós teret. Van tehát rájuk merőleges irány: ezen X_N eloszlásának a vetülete az origóra volna koncentrálva így a négyzetintegrálja nem lehetne 1.

A továbbiakban azt vizsgáljuk, mi vihető át ebből az egyszerű képből az általános esetre.

Definíció. A U tér $\pi = (U_1, U_2, \ldots, U_k)$ particióját a Q szimmetrikus mérték k színnel való színezésének mondjuk. Egy Z k-dimenziós beágyazást a π színezés beágyazásának mondunk, ha mérhető a Q által generált σ-algebrára, vagyis ha π szemein konstans. A színezés beágyazásának a költségét röviden a színezés költségének mondjuk, és $C(\pi)$-vel jelöljük. A k színnel való színezések költségének infimumát $C_k(Q)$-val, vagy röviden C_k-val jelöljük ($B_k(Q)$-t pedig B_k-val).

Lemma. Egy színezés beágyazásának a költsége nem függ a beágyazás megválasztásától, és $B_k \leq C_k$.

Bizonyítás. Az $EZZ^T = I$ feltétel miatt ha Z lehetséges értékeinek a száma k, azok csak ortogonálisak lehetnek és hosszúkat a megfelelő szem valószínűsége egyértelmüen meghatározza. Továbbá

$$C(\pi) = E \parallel Z - Z' \parallel^2 = \sum_{i=2}^{k} \sum_{j=1}^{k-1} Q(U_i, U_j)(\frac{1}{Q(U_i)} + \frac{1}{Q(U_j)}).$$

A színezések beágyazása egyben az eredeti mérték beágyazása is, tehát költségeik nem lehetnek B_k-nál kisebbek.

A π partició k^2 szemet határoz meg az $U \times U$ szorzat térben: ezek közül k "tiszta", az (U_i, U_i) szemek ezek $1 \leq i \leq k$ mellett, ezek mértéke nem szerepel a költségben. A "tarka" (U_i, U_j) szemek $Q(U_i, U_j)$ költsége $i \neq j$ mellett

$$\frac{1}{Q(U_i)} + \frac{1}{Q(U_j)}$$

súllyal növeli a költséget. Ennek a súlynak a szemléletes tartalma az, hogy nagyobb valószínűségű halmazok között "olcsóbb", természetesebb a tarkaság. Ha Q szerint a koordináta terek függetlenek, akkor a költség független a particiótól:

$$C(\pi) = \sum_{i \neq j} (Q(U_i) + Q(U_j)) = (k-1).$$

Ez elég nagy, hiszen egy tetszőleges beágyazásra igaz, hogy

$$C(X) = E \parallel X - X' \parallel^2 \leq 4k.$$

Ennek alapján tesztelhetjük a függetlenséget.

Ha a Q mérték egy gráf éleinek az adjacencia mátrixából származik, a színezés a csúcsokra vonatkozik, és a költség a tarka éleket bünteti. Ha a gráf k összefüggő komponensből áll, ezek csúcsai lesznek megegyező színüek, és egyáltalán nem lesz tarka él. Ilyenkor $B_k = C_k$, a kétféle költség megegyezik. Az élek súlyozásával szabályozhatjuk a tarka élek büntetését, így azokban az esetekben amikor "lényegében" k komponens van, vagyis az összefüggő komponensek között csak néhány él fut keresztbe, egyrészt remélhetjük, hogy ezeket a "szabad" beágyazás is detektálja, másrészt kérdezhetjük, mennyire képes a kétféle mérték eltávolodni egymától. Elsősorban akkor szeretné az ember ezt tudni, ha B_k kicsi: kérdés, várható-e ekkor, hogy C_k is kicsi?

Természetesen nem: hiszen C_k csak a két koordináta tér korrelációját méri: állhat például Q az egymástól független (u_i, v_i) valószínűségi változó párokból ahol a változók maguk mondjuk standard normálisak, és a korrelácoójuk $r_i, i = 1, 2, \ldots, k$. Ha az r_i-k közel vannak 1-hez, B_k nyilván kicsi, pedig C_k egyáltalán nem az. Vagy gráfok esetében eleve megfordíthatjuk a sorrendet: előbb elhelyezzük a pontokat a térben, aztán a már elhelyezett pontok távolságától tesszük függővé, hogy két pont között fusson-el vagy sem. Közeliekre legyen ez a valószínűség nagy, távoliakra kicsi. Vagy mégegyszerűbben: egy négyzetrács legyen a gráf. Az ember mégis úgy érzi, van a két mennyiség között kapcsolat. Ezt többféleképpen

kereshetjük. Egyrészt kérdezhetjük adott B_k mellet C_k maximumát, esetleg a térre vonatkozó mellékfeltételek mellet. Ezt itt nem részletezzük csak megemlítjük a következő egyszerű állítást, amit itt használni lehet.

LEMMA. Tegyük fel, hogy k-dimenzióban adottak pontok úgy, hogy a tér tetszőleges egyenesén a vetületükben a szomszédos pontok között mindig található legalább egységnyi távolság. Akkor a pontok kiszínezhetőek $(k+1)$ színel úgy, hogy mindegyik színt felhasználjuk és a különböző színüek távolsága legalább egységnyi.

DEFINÍCIÓ. Egy k-dimenziós X véletlen vektor Lipschitz-szabadsága az EY^2 mennyiségek szuprémuma, ahol $Y = f(X)$, $\mid f(u)-f(v) \mid \leq \parallel u-v \parallel^2$, és $EXY = 0$. Ezt a mennyiséget $L(X)$-szel jelöljük.

Kérdés: mi a kapcsolat $S(X)$ és $L(X)$ között?

TÉTEL. Ha X optimális k-dimenziós beágyazás, akkor

$$L(X) \leq \frac{B_k}{\lambda_{k+1}}.$$

BIZONYÍTÁS. Cauchy nyeregpont tétele szerint az X-et $(k+1)$-dimenzióssá kiegészítő Z valószínűségi változóra

$$\lambda_{k+1} \leq E(Z - Z')^2$$

teljesül. Y-ból a $Z = Y/v$ normálással alkalmas kiegészítést kapunk, ahol $v^2 = EY^2$. Mivel Y eleget tesz a mondott Lipschitz feltételnek,

$$E(Y - Y')^2 \leq B_k,$$

vagyis $\lambda_{k+1} \leq B_k/v$.

SEJTÉS. Van olyan f függvény, hogy

$$L(X) \geq f(S(X)).$$

Ez a sejtés az előző tétellel együtt azt adná, hogy $S(X)$ jól becsülhető a $\frac{B_k}{\lambda_{k+1}}$ mennyiséggel. Vagyis ha kis B_k után hirtelen egy nagy λ_{k+1} saját érték következik, akkor $S(X)$ kicsi, X jól clusterosíható. Ezt eddig még nem sikerült bizonyítani, de a fordított állítást be tudjuk látni.

TÉTEL.

$$L(X) \leq \frac{S(X)}{1 - kS(X)}.$$

BIZONYÍTÁS. Mivel

$$EY^2 \leq E(Y - E(Y \mid Z))^2 + EE(Y \mid Z)^2,$$

ha Z szimpliciális, itt az első tag S-sel, a második SL-lel becsülhető. E két állítás közül az első Y Lipschitz tulajdonságából, a második Z szimplicialításából következik. Ez utóbbi szerint ugyanis $E(Y \mid Z) = Z^T \times EYZ$, és

$$(EYZ)^2 = (EY(X-Z))^2 \leq EY^2 \times E \parallel X - Z \parallel^2 .$$

Azt várjuk tehát, hogy ha egy gráf beágyazási költségeiben egy ponton nagy ugrás van, akkor a gráf jól clusterosítható. De honnan ismerhetjük fel azt, hogy a gráf jól ágyazható be euklideszi térbe? Egyáltalán, ha egy értékelt élű gráfra adott az élek öszsúlya, mennyi lehet egy adott dimenzióban a maximális költség? És melyik gráfra éretik ez el? Belátható, hogy mindig a teljes gráf adja a szélső értéket, hiszen a teljes gráfra csak a triviális nulla saját érték marad meg kis saját értéknek, az összes többi egyenlő egymással. Azt viszont nem tudom, hogy egy adott beágyazásról hogyan dönthető el, hogy van-e gráf, amelynek ő a beágyazása, és ha igen, hogyan lehet ezt a gráfot megtalálni?

ACE Az $(U, \mathcal{A}), (V, \mathcal{B})$ mérhető terek szorzata feletti P valószínűségi mérték k-dimenziós beágyazása megmutatja a két tér között a mérték által létesített sztochasztikus kapcsolat jellegét, szorosságát. Az optimális beágyazást egy önadjungált operátor saját vektorai határozzák meg. Ezeket kereshetjük direkt iterációval: nem csak Y helyére írhatjuk X-nek $\mathcal{B}$-re vett feltételes várható értékét, hanem X helyére is írhatjuk Y-nek $\mathcal{A}$-re vett feltételes várható értékét. Belátható, hogy ez is csökkenti a célfüggvényt akkor is, ha "visszanormáljuk": beszorozzuk kovariancia mátrixának $(-1/2)$-edik hatványával. Ezt az eljárást "ACE"-nak, alternating conditional expectation képzésnek hívják. Akkor is használható, ha $(V, \mathcal{B})$ helyére a $\prod_{s=1}^{m}(V_s, \mathcal{B}_s)$ szorzatot írjuk, és ezzel egyidejüleg Y helyére a $(V_s, \mathcal{B}_s)$ tereket R^k-ba vívő Y_s leképezéseket írjuk, és a $\parallel X - \sum_{s=1}^{m} Y_s \parallel^2$ hiba várható értékét kívánjuk minimalizálni úgy, hogy az X-re vonatkozó kényszerfeltételt változatlanul hagyjuk.

Ez így nem paraméteres regresszió: egy tér elemeit mások összegével közelítjük. De nem sokat változtat az eljáráson, ha X-et beolvasztjuk az Y_s-ek közzé: egy tetszőleges többdimenziós eloszlásban a koordináta változókat k-dimenziós valós véletlen vektorokkal helyettesítjük úgy, hogy e vektorok összege lehetőleg kicsi legyen. Ez a kicsi összeg tükrözi a változókban rejlő összefüggést: minél szorosabb az összefüggés, annál kisebb az összeg. Az eljárás emlékeztet a vektoralgebra lineáris függetlenség fogalmára: azok a vektorok függnek össze, amelyeknek a lineáris kombinációjaként a nulla előállítható. Mindkét esetben csak annyit tudunk meg, hogy van összefüggés, az, hogy ez az összefüggés voltaképpen a változók mely csoportja

között áll fenn, az az együtthatókból olvasható ki. Maga az algoritmus ugyanaz mint $s = 1$ mellett: ciklikusan valamelyik ismeretlen leképezést úgy határozzuk meg, mint a többikek feltételes várható értéke.

Az eljárás nem csak összefüggések keresésére használható: lévén többdimenziós beágyazás egyszerűen visszavezeti az absztrakt feladatot a többdimenziós statisztika keretei közé: miután beágyaztuk a koordinátákat, átkódoltuk azokat, az új változókon tetszés szerinti statisztikai eljárást végrehajthatunk. Persze ha tudjuk, hogy mire akarjuk használni a beágyazást, a cél függvényében magát a beágyazást is megváltoztathatjuk. Például egy orvosi vizsgálatban betegek kezelés előtti és utáni adatait hasonlítjuk össze, de úgy, hogy a két adatsor között semmilyen szemantikus összefüggés nincs, tehát belépéskor egész más adatokat vettek fel a betegekről, mint kilépéskor. El szeretnénk dönteni, van-e egyáltalán kapcsolat a kétféle adat között. Most a kovarianciára vonatkozó kényszerfeltételt az egyik adatsor beágyazott értékeinek az összegére célszerű tenni.

Sztochasztikus folyamatok beágyazása Hasonlóan más típusú beágyazást kapunk, ha az eredeti adatok idősort alkotnak: most az a cél, hogy a beágyazás után kapott idősorban a predikciós hiba kicsi legyen.

Többszempontú optimalizálás Ismét más természetű beágyazásra jutunk, ha az adataink alapján bizonyos objektumokat sorba akarunk rendezni (például a betegeket a betegségük súlyossága alapján). Ebben az esetben a többdimenziós tér félig rendezett volta segíthet az objektumok között létesíthető rendezési relációk áttekintésében. Itt is gondot okozhat a statisztikai elemzések általános nehézsége, az tudni illik, hogy ha sok önálló stastisztikai eljárást hajtunk végre, az egyes lépések között semmifajta harmónia nem várató el: önmaguktól az egyes statisztikai döntések nem fognak illeszkedni egymáshoz. Minden féle rendezettséget csak akkor kapunk, ha ezt eleve megköveteljük az eljárástól.

A több szempontú optimalizálás során különböző formájú megfigyeléseket végzünk egy véletlen permutációra: a sorbarendezendő objektumok valódi sorrendjére. Egyáltalán hogyan lehet véletlen permutációkra eloszlásokat modellezni? Egy lehetőség a következő. Rendeljünk az objektumokhoz eloszlásokat. Minden objektumhoz generáljunk egy-egy véletlen számot egymástól függetlenül, és állítsuk ezeket nagyság szerint sorba. Ez legyen az objektumok sorrendje. Nem tudom, hogyan lehet itt az eloszlásokat becsülni. Feltehetjük, hogy az eloszlások exponenciálisak. Ekkor csak a paramétereiket kell becsülni. Belátható hogy ez a modell ekvivalens azzal, hogy

magukok az objektumokon adunk meg egy eloszlást, ebből generálunk véletlen és független számokat mindaddig, ameddig mindegyik objektum sorszáma sorra nem kerül, és most legyen az objektumok sorrendje az, ahogyan a sorszámuk a generálás során felbukkan. Mindkét modellben alkalmazható az EM algoritmus, de érdekes módon különböző iterációra vezet.

FELADATOK:

10.1. **(Grb)** Gráfok beágyazása. Legyen $a_{ij}, 1 \leq i, j \leq N$ egy tetszőleges mátrix. Keresendők a d dimenziós Euklideszi térben az $X_1, \ldots, X_N$ pontok úgy, hogy a

$$\sum_{i=1}^{N} X_i X_i^T$$

diádösszeg a d-dimenziós egységmátrixszal legyen egyenlő, és

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} \parallel X_i - X_j \parallel^2$$

minimális legyen (T a transzponálás jele).

10.2. **(Igs)** Állítsuk sorba valamilyen értelmes szempont szerint egy irányított gráf csúcsait.

10.3. **(Ace)** Rendeljük hozzá egy diszkrét adatmátrix elemeihez a d-dimenziós tér elemeit úgy, hogy a kapott pontok kovariancia-mátrixa egységnyi legyen, és az egy rekordban álló elemekhez rendelt mennyiségek összegének a normanégyzetének az összege maximális legyen.

IRODALOM:

D.M. Cvetković-M. Doob-H. Sachs: Spectra of graphs, Theory and application, Academic Press, 1979

L. Breiman-J.H. Friedman(1985): Estimating optimal transformations for multiple regression and correlation, J. Amer. Stat. Assoc. 46, 580-619

Bakonyi Árpád-Meszéna György-Tusnády Gábor(1988): A "Termékek összehasonlító vizsgálata" című pályázat tapasztalatai, Minőség és Megbízhatóság, 22/6, 41-50

D. Cvetković-M. Doob-I. Gutman-A. Torğasev: Recent results in the theory of graph spectra, North-Holland, 1988

L. Lovász: Combinatorical problems and exercises, Akadémiai Kiadó, 1993

M. Bolla-G. Tusnády(1994): Spectra and optimal partitions of weighted graphs, Discrete Matematics, 128, 1-20

11. Sztochasztikus kontroll

Témák: stopping rules, Bellman egyenlete, scheduling, keresések, Gittins indexe, optimális portfolio, dinamikus üzletpolitikák.

A sztochasztikus kontroll azzal foglalkozik, hogyan lehet rákényszeríteni az akaratunkat a véletlenre, vagy szelídebben fogalmazva, hogyan lehet irányítani, befolyásolni a véletlen jelenségeket. Akaratunk megközelítését célfüggvénnyel mérjük, nyilván nincs jelentősége annak, hogy maximalizálni, vagy minimalizálni akarunk-e.

Képzeljünk el egy nagy fát, amelynek az ágaira valószínűségeket írtak, az ágak végén a fészkekben pedig ajándékokat helyeztek el. A gyökértől kiindulva felváltva léphetünk egyszer mi, egyszer a véletlen. Minden lépésben csak felfelé kúszhatunk, de amikor mi lépünk, azt szabadon dönthetjük el, melyik ágon kapaszkodunk egy emeletnyit. Aztán a következő lépés a véletlené: ő az ágakra írt valószínűségek alapján dönti el, mi merre kússzunk a következő lépésben. Onnan megint mi választhatjuk meg az utunkat. Végül felérve a csúcsra megkapjuk azt az ajándékot, amit ott a fészekben találunk. Mindent pontosan tudunk már induláskor, az ajándékok elhelyezkedését, a valószínűségeket, a játék szabályait. Mit csináljunk?

Stopping rules Szinbád, a nagy tengeri utazó egyik kalandja során megmentette a szultán életét. Ezért a szultán elhatározta, hogy Szinbádot a következőképpen jutalmazza meg. Véletlenszerűen elvonultatja előtte mind a 300 feleségét (egyesével), akik között ugyan egyértelmű szépségi sorrend állapítható meg, de Szinbádnak, lévén távoli földek lakója, fogalma sincs a feleségek szépségéről. Szinbád választhat egyet a feleségek közül, mondván: "Hatalmas szultán! Szerintem ő a legszebb feleséged." Mindig csak azt választhatja, akit éppen lát, és ha valóban a legszebbet választotta, a szultán neki ajándékozza a legszebb feleségét. Ha nem, mivel ez egy keleti mese, lefejezteti. Mekkora Szinbád túlélési esélye? Hogyan változna meg ez, ha a szultánnak 3000 felesége lenne?

Mi történne, ha a feleségek felsorakoztatása helyett azt csinálnák, hogy közösen választanának egy folytonos eloszlást, a szultán abból generálna független, egyforma eloszlású mintát, és Szinbádnak most a legnagyobb számot kellene menet közben ("on line") felismernie? Mit kell ebben az esetben Szinbádnak tennie, ha sikerülne rábeszélnie a szultánt, hogy egyszerűen adjon neki annyi aranyat, amennyi az általa választott véletlen szám értéke?

Mi történik, ha a szultánnak van egy egészen gonosz varázslója, aki miután Szinbád megválasztotta a stratégiáját, ezt kiolvassa a gondolataiból, és ennek megfele-

lően megváltoztatja az eloszlást?

Ezekben a feladatokban általában egy $(x_1, \ldots, x_n)$ sztochasztikus folyamatról van szó, amelyiknek ismerjük (többé kevésbé) az eloszlását. Minden egyes $t = 0, \ldots, n-1$ időpontban a folyamat addig látott (vagy $t = 0$ esetén nem látott) $(x_1, \ldots, x_t)$ darabja alapján arról az egyetlen dologról dönthetünk, hogy "megálljunk-e", vagy sem. Csak egyetlen egyszer állhatunk meg. Aztán ettől függetlenül realizálódik a teljes folyamat, és a nyereményünk (vagy veszteségünk) értéke az előre meghatározott $C(T, x_1, \ldots, x_n)$ függvény, ahol T az a pillanat, amikor megálltunk. Ennek értéke persze lehet n is, de akkor ezt azáltal döntjük el, hogy sose állunk meg amikor ezt még megtehetnénk.

A feladatban nagyon fontos, hogy a $(T \leq t)$ eseménynek mérhetőnek kell lennie az $(x_1, \ldots, x_t)$ vátozók által generált σ-algebrára, ami azt jelenti, hogy amikor megállunk, csak annyit látunk a folyamatból, amennyi addig realizálódott belőle. Ez jellemző lesz a bonyolultabb feladatokra is, ez a dinamikus jelleg az, ami elsősorban karakterizálja a sztochasztikus kontroll feladatainak a többségét.

Az optimális megállási szabályt általában a "backwards induction" nevű eljárással határozhatjuk meg. Ha $t = n-1$, már csak arról dönthetünk, ha még nem álltunk meg, hogy utolsó lehetőségünket megragadjuk-e. Ismervén a folyamat eloszlását, kiszámolhatjuk az

$$M(x_1, \ldots, x_{n-1}) \;=\; E(C(n-1, x_1, \ldots, x_n) \mid x_1, \ldots, x_{n-1}),$$

$$F(x_1, \ldots, x_{n-1}) \;=\; E(C(n, x_1, \ldots, x_n) \mid x_1, \ldots, x_{n-1}),$$

feltételes várható értékeket, és azt fogjuk választani, amelyik a kettő közül nagyobb. Jelöljük ennek az értékét $K_1(x_1, \ldots, x_{n-1})$-val:

$$K_1(x_1, \ldots, x_{n-1}) \;=\; \max(M(x_1, \ldots, x_{n-1}), F(n, x_1, \ldots, x_{n-1})).$$

Ha már ezt a függvényt ismerjük, léphetünk eggyel előre. Most is kiszámolhatjuk az

$$M(x_1, \ldots, x_{n-2}) \;=\; E(C(n-2, x_1, \ldots, x_n) \mid x_1, \ldots, x_{n-2})$$

$$F(x_1, \ldots, x_{n-2}) \;=\; E(K_1(x_1, \ldots, x_{n-1}) \mid x_1, \ldots, x_{n-2})$$

feltételes várható értékeket és azt fogjuk választani, amelyik a kettő közül nagyobb. Jelöljük ennek az értékét $K_2(x_1, \ldots, x_{n-2})$-val:

$$K_2(x_1, \ldots, x_{n-1}) \;=\; \max(M(x_1, \ldots, x_{n-1}), F(n, x_1, \ldots, x_{n-1})).$$

Most már talán mondhatjuk, hogy "és így tovább": mindig tudjuk, mi vár ránk, ha folytatjuk a folyamat megfigyelését, és azt is, ha megállunk. Végül a legelső lépesben el tudjuk dönteni, hogy érdemes-e egyáltalán az egészet elkezdeni.

Szinbád esetében ha a választott eloszlás $(0,1)$-ben egyenletes, aszimptotikusan $(1-\frac{c}{k})$-nál nagyobb értéknél kell megállni, ahol k a még hátra levő lépések száma, és a legnagyobb szám eltalálásának a valószínűsége az

$$\int_0^1 \int_{1-x}^1 \frac{1}{y} e^{-\frac{cy}{x}} dydx \;=\; 0.580$$

értékhez tart, ahol c a

$$\sum_{j=1}^{\infty} \frac{c^j}{jj!} \;=\; 1$$

egyenlet gyöke.

Bellman egyenlete Egy sztochasztikus automata az $M(z \mid s,u)$ úgynevezett átmenet mátrixszal adható meg, ahol s,z az S állapottér elemei, u az U input-ABC eleme. Az átmenet mátrix elemei nem negatívak, és z-re vett összegük tetszőleges s,u mellett 1. Feltesszük, hogy S és U véges halmazok. Tegyük fel továbbá, hogy ha az u input hatására az automata az s állapotból a z állapotba megy át, akkor ezért $C(z \mid s,u)$ jutalmat kapunk. Adott s_0 kezdeti állapot mellett úgy kell az $(u_1,\ldots,u_n)$ bemeneti jeleket meghatároznunk, hogy az összes nyereményünk várható értéke maximális legyen. Ez a következőket jelenti. Nekünk először is meg kell adnunk az $\mathcal{U}_t(s_0,\ldots,s_{t-1}), t \;=\; 1,\ldots,n$ függvényeket, amelyek lépésröl lépésre megadják a stratégiánkat. Egy ilyen függvény-rendszer már egyértelműen meghatározza az $(s_1,\ldots,s_n \mid s_0)$ feltételes eloszlást: erre a feltételes eloszra vonatkozóan kell a nyereményeink összegének a feltételes várható értékét meghatározni.

A backwards induction gondolata itt is alkalmazható. Jelöljük $K_t(s)$-sel a várható nyeremények maximumát ha az automata az s állapotban van és még t lépés van hátra. $K_0(s)$ értéke nulla: ha az automata nem lép, nincs jutalom. Ha $t>0$, akkor

$$K_t(s) \;=\; \max_{u\in U}\Big(\sum_{z\in S} M(z \mid s,u)(C(z \mid s,u) + K_{t-1}(z)),$$

és az az optimális input, amelyik ezt maximalizálja. Tehát nem kell a múltat figyelembe venni, de annak ellenére, hogy a feladat homogén, az optimális input az állapotnak nem homogén függvénye, hiszen véges időhorizont mellett vizsgáljuk a feladatot.

SCHEDULING Tegyük fel, hogy egy borbélyüzletben két borbély dolgozik, és az idő diszkrét módon változik. Mondjuk tíz percenként mindkét borbély feldob egy-egy érmét, és az eredménytől függően vagy elkészül azzal a vendéggel, akinek a hajával éppen foglalatoskodik, vagy nem. Legyen az elkészülés valószínűsége p_1 az egyik borbélyra, p_2 a másikra. A borbélyok pénzdobálása között félidőben a külvilág dob fel egy érmét, és q valószínűséggel beküld az üzletbe egy új vendéget, $(1-q)$ valószínűséggel nem küld. Minden randomizálás független. A vendégeket a borbélyok érkezési sorrendben szolgálják ki, de minden egyes vendégnek az érkezésekor azonnal meg kell választania a borbélyát. Most az egész rendszer büntetést fizet percenként és vendégenként amíg a vendégek ki nem lépnek az üzletből.

Jelöljük a külvilág t-edik randomizálásakor az első borbélyra várakozók számát azt a vendéget is hozzájuk számítva, akin esetleg a borbély dolgozik x_1-gyel, a második borbélyra legyen ez a szám x_2. Most ugyan az állapotok száma végtelen, de Bellman egyenlete nyilván érvényes erre az esetre is. Ha véges időhorizonttal dolgozunk, valahogy még le kell zárni a folyamatot. Mondjuk azt, hogy mint a Csipkerózsikában, a külvilág n-edik randomizálása után hirtelen minden megmerevedik, és csak az addigi költségeket kell megfizetnünk. Ha a borbélyok elég gyorsan dolgoznak, és n nagy, lassan kialakul egy stacionárius állapot, az optimális stratégiának is van határértéke. Ennek az az érdekessége, hogy mellette nem mindig úgy kell választaniuk az egyes vendégeknek, hogy az ő személyes várakozási idejük várható értéke minimális legyen. Értelmezhetjük ezt a jelenséget úgy, hogy a büntetést a borbélyok fizetik, mivel várakoztatják a vendégeiket, és ennek megfelelően ők is irányítják a vendégeket azt figyelembe véve, hogy a globális érdekek érvényesüljenek.

Ebben a feladatban eredetileg folytonos volt az idő, és a kiszolgálások ideje, és a vendégel érkezés közti idők exponenciális eloszlásúak voltak. Ezt az esetet határátmenettel kaphatjuk a fenti diszkretizált változatból. Térjünk most át eleve az exponenciális eloszlásokra. Tegyük fel, hogy k borbély van, ők ugyan egyformák, de most a vendégek legyenek különbözöek: mindegyik exponenciális eloszlású kiszolgálást igényel, de ezek paramétere minden vendégre más. Reggel minden vendég egyszerre belép az üzletbe, és paramétereik ismeretében fel kell sorakoztatnunk őket az egyes borbélyokhoz. Aztán a munka előrehaladtával már nem lehet a vendégeket átszervezni. Változatlanul a vendégeknek a borbélyüzletben eltöltött összidejéért kell büntetést fizetnünk. Most a feladat "off line": egyszer kell döntenünk. Az optimális stratégia egyszerűen megadható: a várható kiszolgálási idő szerint növekvő

rendben ciklikusan kell a vendégeket az egyes borbélyokhoz rendelni. Érdekes módon épp a fordított sorrend optimális, ha a teljes munka idejének a várható értékét akarjuk minimalizálni. (Vagyis azt az időpontot kívánjuk várhatóan a lehető legkorábbivá tenni, amikor az utolsó vendég is eltávozik, és be lehet a boltot csukni.)

GITTINS INDEXE Van három aranbányánk. Napról napra el kell döntenünk, melyikbe menjünk dolgozni. Az elsőről tudjuk, hogy mostantól kezdve a t-edik munkanap során x_t aranyat tudunk felhozni belőle, a másodikból y_t-t, a harmadikból z_t-t. Ezt úgy értjük, hogy ismerjük az

$$(x_1, \dots, x_t, \dots), (y_1, \dots, y_t, \dots), (z_1, \dots, z_t, \dots)$$

sztochasztikus folyamatok eloszlását. (A folyamatok függetlenek.) A kibányászott aranyat esténként bevisszük a városba, betesszük a bankba, és ott naponta az értéke a ρ-szorosára nő ($\rho > 1$). Hogyan alakítsuk ki a bányáink művelési rendjét, ha az n-edik nap végére várható vagyonunkat szeretnénk maximalizálni? Ha a t-edik munkanap u_t aranyának az értéke helyett most tennénk be a bankba pénzt, nyilván $\rho^{-t}u_t$ értéket kellene ott elhelyeznünk, ha azt szeretnénk, hogy a t-edik napon pontosan annyi pénzünk legyen, mint amennyit akkor bányászunk. Mondhatjuk azt is, hogy ma nekünk az az arany ennyit ér. Ezt az átszámítást diszkontálásnak hívják. Ennek a segítségével végtelenre tágíthatjuk a feladat horizontját, és kérdezhetjük, melyik az a stratégia, amelyik az $\sum_{t=1}^{\infty} \rho^{-t}u_t$ összeg várható értékét minimalizálja.

J.C. Gittins vette észre, igaz csak Markov láncokra, hogy a folyamatok elhelyezhetőek egy abszolut skálán: minden folyamatnak van egy értéke, ez egy jól meghatározott valós szám, és ha bármely két folyamatra a fenti ütemezési feladatot kell megoldanunk, mindig abból a folyamatból kell az első értéket választanunk, amelyiknek az indexe nagyobb. Ez természetesen nem jelenti azt, hogy örökké azt a folyamatot kell választanunk. Egy napi munka után megváltozik a folyamat. Ha például az első nap az első bányába megyek dolgozni, a következő napon az

$$(x_2, \dots, x_t, \dots), (y_1, \dots, y_t, \dots), (z_1, \dots, z_t, \dots)$$

folyamatok között kell választanom. A másik kettő persze indexestül változatlan marad, de az első indexe megváltozott azáltal, hogy egy napig dolgoztam benne. És általában: akárhány folyamatra is kell egyidejüleg munkarendet készítenünk, a folyamatokat mindig indexeik szerint kell sorra vennünk. Eközben a nagy indexű folyamatokat művelvén, azok "kimerülhetnek", mint a bányák: lecsökkenhet az

indexük, és akkor végre a többiek is sorra kerülnek. Ami azonban nem jelenti azt, hogy az egész folyamat egyre értéktelenebb lesz, hiszen az index növekedhet is.

Hogyan lehet ezt az indexet meghatározni? Ha igaz az állítás, és ez valóban egy abszolut skála, tesztelhetőek a folyamatok speciális folyamatokkal, erre a célra a legalkalmasabb a konstans. Tekintsük az

$$(x_1, \dots, x_t, \dots), (w_1, \dots, w_t, \dots)$$

páros összehasonlítási feladatot, ahol w_t nem függ sem az időtől, sem a véletlentől: $w_t = c$, ahol c egy tetszőleges valós szám. Ennek a konstansnak az értékétől függően két eset lehetséges:

az optimális stratégia szerint a konstans folyamatból kell az első elemet venni, vagy

az optimális stratégia szerint az eredeti $(x_1, \dots, x_t, \dots)$ folyamatból kell az első elemet venni.

Az első esetben a helyzet nyilván változatlan marad a további lépések során. A másodikban előfordulhat, hogy egszer csak megváltozik a helyzet. Számunkra viszont most lényegesebb az, hogy a két esetet nyilvánvalóan egy c_0 küszöb választja el: ha $c > c_0$, akkor az első eset fordul elő, ha $c < c_0$, akkor a második. Ez a c_0 a $(x_1, \dots, x_t, \dots)$ folyamat Gittins indexe. Az elmondottakból látható, hogy értéke a

$$\sup_T \frac{E \sum_{t=1}^{T} \rho^{-t} x_t \mid x_1)}{E \sum_{t=1}^{T} \rho^{-t} \mid x_1)}$$

szupremum, ahol T befutja az összes lehetséges megállási szabályt: ez ugyanis az egy napra jutó maximális várható átlagos nyereség: ezek azok a kezdeti szakaszai az $(x_1, \dots, x_t, \dots)$ folyamatnak, amikor őt üzemeltetjük, és nem a konstanst.

Optimális portfolio Legyenek most az

$$(x_1, \dots, x_t, \dots), (y_1, \dots, y_t, \dots), (z_1, \dots, z_t, \dots)$$

folyamatok nem negatív értékűek, és változatlanul függetlenek. (Az aranybányák esetében előfordulhatott, hogy egész napi munkámmal csak veszteséget termeltem: eltörött a fúró, teljesen feleslegesen elhasználtam egy csomó dinamitot.) Tegyük fel, hogy van egységnyi vagyonom. Ezt első lépésként $\alpha_1, \beta_1, \gamma_1$ arányban osztom meg a három folyamat között, ahol $\alpha_1, \beta_1, \gamma_1$ nem negatívok, és az összegük 1. Az első nap végére a vagyonom

$$v_1 = \alpha_1 x_1 + \beta_1 y_1 + \gamma_1 z_1$$

lesz, másnap ezt $\alpha_2, \beta_2, \gamma_2$ arányban osztom meg a lehetséges befektetések között, lesz

$$v_2 \;=\; (\alpha_1 x_1 \;+\; \beta_1 y_1 \;+\; \gamma_1 z_1) v_1$$

vagyonom, és így tovább. Hogyan kell a vagyonomat megosztani? Kezdetben csak ösztönösen, aztán egyre tudatosabban azt javasolták a kérdés felvetői, hogy a vagyon logaritmusának a várható értékét kell lépésről lépésre maximalizálni. Kiderült ugyanis, hogy elég általános feltételek mellett így lehet elérni a legnagyobb exponensű növekedést.

Dinamikus üzletpolitikák Legyen most az $(x_1, \dots, x_t, \dots)$ folyamat a napi pénzforgalmam: legyen x_t a t-edik nap előjeles bevétele a számomra, az most egyáltalán nem lényeges, honnan származik ez a bevétel, csak az a fontos, hogy ez lehet pozitív is, negatív is. Világomban egyetlen "zsugori" bank van: bármikor tehetek bele pénzt, az ott naponta meg-ρ-szorozódik ($\rho > 1$), ha viszont pénzt akarok kivenni belőle, akkor minden kivett forintért Q forint tartozást számol fel ($q > \rho$) akkor is, ha a saját pénzemet veszem ki. Különben szabadon és korlátlanul ad kölcsönt. Nekem viszont a készpénzem sosem lehet negatív: ha a napi forgalmam negatív, vagyis valamiért fizetnem kell, akkor azt valóban ki kell fizetnem akár a készpénzemből, akár a bankból kivett pénzből. Kérdés, mikor mennyi pénzt tartsak magamnál. Évek óta nem tudom ezt a feladatot megoldani. Bellman egyenletei persze felírhatóak, formálisan a backwards induction is működik, de az egész kombinatórikusan felrobban, és ezért számolhatatlan. Ez persze bármikor előfordulhat. A baj az, hogy nem tudom eldönteni, a feladat valóban megoldhatatlan a mai viszonyok között, vagy van az optimális stratégiának olyan karakterizációja, amelyik alapján már számolható.

Feladatok:

11.1. **(Mnp)** Mátrix-játék nyeregpontjának a meghatározása.

11.2. **(Git)** Gittins indexének a kiszámolása.

11.3. **(Lqc)** Adottak az $A, B, Q, R \quad N \times N, N \times M, M \times M, N \times N$ méretű mátrixok, az N-dimenziós tér X_0, Y elemei, és egy n pozitív egész. Meghatározandóak az M-dimenziós tér $U_1, \dots, U_n$ elemei úgy, hogy

$$\sum_{i=1}^{n} U_i^T Q U_i \;+\; (X_n - Y)^T R (X_n - Y)$$

minimális legyen, ahol T a transzponálás jele és X_n-et az

$$X_k = A X_{k-1} + B U_k, \quad k = 1, \dots, n$$

egyenletek határozzák meg.

11.4. **(Bpm)** Tekintsük a következő játékot. Egy adott n-szer n-es mátrix elemein bolyongunk úgy, hogy az irányt minden lépésben szabadon választhatjuk, de csak $\frac{1}{2}$ valószínűséggel léphetünk oda, különben $\frac{1}{4}$ valószínűséggel jobbra vagy balra lépünk a választott irányhoz viszonyítva. A cél az utunk során látott értékek összegének a várható értékének a maximalizálása.

11.5. **(Elt)** Szerkesztendő egy véges memóriájú automata, amely a lehető legjobban képes követni a független ± 1-ek összegének az előjelét.

11.6. **(Düp)** Tegyük fel, hogy egyetlen eloszlásra kötnek az emberek életbiztosítást, eszerint maximum n napig élhetnek, és az egyes napokon rendre

$$P(1), P(2), \ldots, P(n)$$

valószínűséggel halnak meg. A biztosítás megkötésekor ξ forintot fizetnek, és halálukkor 1 forintot adunk a családjuknak. Naponta maximum egy ember akar biztosítást kötni, ennek $R(1)$ a valószínűsége, tehát $R(0) = 1 - R(1)$ valószínűséggel nem akar senki biztosítást kötni. Mi csak arról dönthetünk, mekkora legyen ξ értéke, és ha pozitív a készpénzünk, abból mennyit tegyünk a bankba. Ha nincs pénzünk, kölcsönt egy "zsugori" banktól kérhetünk: ha A forintot kapunk, attól kezdve QA forinttal tartozunk, és az adósságunk naponta meg-ρ-szorozódik. (ρ 1-nél nagyobb ismert konstans.) Ugyanebbe a bankba tehetjük be a megtakarított pénzünket is, az ugyanígy kamatozik, és ha abból veszünk ki A forintot, a bankbeli pénzünk akkor is QA forinttal csökken. Dönteni minden egyes nap arról dönthetünk, hogy mennyi pénzt tegyünk a bankba. A cél a T-edik napra várható vagyon maximalizálása. Ezen a T-edik napon már egybeszámolhatjuk a készpénzünket és a bankbeli pénzünket az utóbbi Q-val való osztása nélkül, de ha a készpénzünk negatív volna, még az utolsó napon is fel kell venni a megfelelő kölcsönt.

11.7. **(Szi)** Szinbád.

Irodalom:

Tusnády Gábor(1973): Egy kockajátékról, A Matematika Tanítása, 20, 25-27, 75-80

P. Whittle: Optimisation over time, Wiley, 1980

D.P. Heyman-M.J. Sobel: Stochastic modells, North Holland, 1991

P. Whittle: Systems in stochastic equilibrium, Wiley, 1986

J. Komlós-L. Rejtő-G. Tusnády(1993): Learning with finite memory, Studia Scientiarum Mathematicarum Hungarica, 28, 167-172

12. Sztochasztikus automaták

Témák: ergodicitás, monotonitás, coupling, Gács tétele.

Ergodicitás Vizsgálódásaink tárgya egyforma lineárisan rendezett sztochasztikus automaták sorozata lesz. Legtöbbet azt az esetet vizsgáljuk majd, amikor az automatáknak mindössze két állapotuk van: a 0 és az 1, ezeket üresnek és telinek mondjuk majd. Automatáink az állapotaikat a két szomszédjuk állapotának a függvényében időben folytonosan változtatják. Egy teli automata rendre

$$U(0,0), U(0,1), U(1,0), U(1,1)$$

intenzitással ürül ki, ha szomszédjainak állapota rendre $(0,0), (0,1), (1,0)$, vagy $(1,1)$, amin azt értjük, hogy ha például a bal szomszéd üres, és a jobb teli, akkor Δt idő alatt $U(0,1)\Delta t + o(\Delta t)$ valószínűséggel ürül ki a szóban forgó automata. Amíg az automata ki nem ürül, és szomszédai állapota megváltozik, ezt az automata késedelem nélkül érzékeli, és rögtön megváltoztatja a kiürülési intenzitását. Magához az állapotváltoztatáshoz egyébiránt nem kell idő.

Ha üres az automata, akkor szomszédainak az állapotától függően

$$T(0,0), T(0,1), T(1,0), T(1,1)$$

intenzitással telik meg. A teljes automatasor állapotát a t időben az

$$\{A_n(t), \quad n = 0, \pm 1, \pm 2, \dots\}$$

sorozat írja le, itt $A_n(t)$ értéke 1 vagy 0 aszerint, hogy az n–edik automata teli van vagy üres. A teljes sor állapotát $A(t)$–vel jelöljük:

$$A(t) = \{A_n(t), \quad n = 0, \pm 1, \pm 2, \dots\},$$

és az állapotok ilyen sorozatát szuperállapotnak mondjuk, magát az n–edik automatát A_n–nel jelöljük.

Az automatasor viselkedését nem negatív t mellett vizsgáljuk, az $A(0)$ induló szuperállapotot A–val jelöljük. Az átmeneti intenzitásokon kívül ez határozza meg az automatasor jövőjét. Belátható, hogy ha t tart végtelenbe, mindíg van a folyamatnak határeloszlása. Kérdezhetjük, hogyan jellemezhető ez? Minket most elsősorban az fog érdekelni, hogy a határeloszlás függ–e A–tól. Erre az egyszerű kérdésre nem ismeretes a válasz: általában nem tudjuk megmondani adott 8 átmeneti intenzitás mellett, hogy hányféle határeloszlás alakulhat ki. Van viszont egy aránylag

természetes feltétel, amely mellett általános válasz adható. Mi a következőkben végig fel fogjuk tenni, hogy az intenzitások pozitívak, ami nem jelenti azt, hogy érdektelen volna az az eset, amikor közöttük 0 is lehet.

Azt mondjuk, hogy az automatasor attraktív, ha az $U(0,0), U(0,1), U(1,0), U(1,1)$ számok között $U(0,0)$ a legnagyobb, $U(1,1)$ a legkisebb, és a $T(0,0), T(0,1), T(1,0), T(1,1)$ számok között $T(0,0)$ a legkisebb, $T(1,1)$ a legnagyobb. Ennek a feltételnek az értelme az, hogy mellette a hasonló állapotok "vonzzák" egymást: egy teli automata akkor ürül ki a leggyorsabban, ha mindkét szomszédja üres, és ez a folyamat akkor a leglassúbb, ha a két szomszéd telített. Hasonló a helyzet a megteléssel.

Kétállapotú folyamatok állapotait sokféleképpen jelölhetjük, szemléltethetjük. Szokás az állapotokat a mágnesség mintájára "észak"–nak, "dél"–nek, vagy "fel"–nek, "le"–nek, "+"–nak, "—"–nak mondani. Itt most a két állapot lényegében szimmetrikus szerepű, ezért talán zavaró lesz, hogy aszimmetrikusan jelöljük őket. Indoklásul csak annyi mondható, hogy sokszor a szimmetria is zavaró, itt most több előnye lesz az állapotok irányított voltának. Képzelhetjük őket akár "jó"–nak, "rossz"–nak, "élő"–nek, "holt"–nak is, a különböző szemléletek szavai akaratlanul is átcsúsznak egymásba. Vigyázni kell azonban, hogy az állapotok irányítottsága, a számegyenes irányitása, és az idő irányított volta össze ne keveredjenek.

Ha egy automatasor attraktív, feltehetjük, hogy $U(0,0) + T(1,1) = 1$, hiszen ez csak az óra skálázását érinti, amellyel a folyamat idejét mérjük. A következő konstrukció megmutatja a vizsgált folyamat létezését, de ezen túlmenően az egész bizonyítás alapja lesz. Legyenek az $X(n,k)$ valószínűségi változók függetlenek és a $(0,1)$ intervallumban egyenletes eloszlásúak, az $Y(n,k)$ valószínűségi változók pedig tőlük is és egymástól is független 1 paraméterű exponenciális eloszlású valószínűségi változók, ahol n tetszőleges, k nem negatív egész. Konstrukciónk (és a bizonyítás) alapvető tulajdonsága az, hogy az automatasor viselkedését ez a két független és egyformal eloszlású elemekből álló, egymástól is független mátrix határozza meg.

Állítsuk elő az

$$S(n,k) = \sum_{i=1}^{k} Y(n,i)$$

összegeket: ezek minden egyes n–re megadják azokat az időpontokat, amelyekben A_n–nel egyáltalán valami történhet. Könnyen látható, hogy az

$$S(n,k) = S(m,j)$$

események valószínűsége tetszőleges n, k, m, j mellett 0, tehát ezek együttes valószí-

nűsége is 0. Mi a továbbiakban kizárjuk ezt az eseményt, feltesszük, hogy minden állapotváltozás időpontja különböző. Azt is feltesszük, hogy az $\{S(n,k), k = 1,2,\dots\}$ sorozatoknak nincs torlódáspontjuk. Maguk az $S(n,k)$ időpontok csak potenciálisan jelentenek állapotváltozást. Minden egyes $t = S(n,k)$ időpontban az $X(n,k)$ változó értéke határozza meg $A_n(t)$ értékét. Ha t előtt A_n üres, és $X(n,k) < U(0,0)$ akkor nem történik semmi. Különben akkor telik meg az A_n automata, ha

$$X(n,k) > 1 - T(A_{n-1}(t), A_{n+1}(t)).$$

Ha t előtt A_n teli van, és $X(n,k) > 1-T(1,1)$, akkor nem történik semmi. Különben akkor ürül ki A_n, ha

$$X(n,k) < U(A_{n-1}(t), A_{n+1}(t)).$$

Mivel nincsenek egyidejű állapotváltozások, a folyamat így valóban meg van konstruálva. Pontosabban mondva szükség van még arra a feltételre is, hogy az

$$\{S(n,k), k = 0,1,\dots\}$$

sorozatoknak nincs torlódáspontjuk: ez biztosítja azt, hogy t–ben a szomszédok állapota egyértelműen meg van határozva.

Monotonitás Első ránézésre a konstrukció felesleges fontoskodásnak tűnik: váltakozó intenzitású Poisson folyamatokat természetesen előállíthatunk homogén ritkitásával, de nem világos, miért kell ennek adminisztrálására az $X(n,k)$ változókat is felhasználni. Különösen nem látja az ember a születés–halál (telítődés–kiürülés) mesterséges szétválasztását, a két randomizálásnak ugyanabból az egyenletes eloszlású változóból történő előállításának a hasznát. A teljes megértést csak a bizonyítás egésze adja. A kezdet a konstrukcióból fakadó monotonitás.

Indítsuk az automatasort az A, B szuper állapotokból, és jelöljük a kialakuló folyamatokat $A(t), B(t)$–vel. Maguk az átmenetek megadják ezeket a folyamatokat, de szabadon hagyják e folyamatok együttes eloszlását. Ezt azzal az egyszerű, ártatlannak tűnő technikai fogással alakítjuk ki, hogy UGYANAZT az X, Y rendszert használjuk a két folyamathoz. Sőt, akárhány indítás is jön majd szóba itt, azokhoz mindíg egy és ugyanazt az X, Y rendszert fogjuk felhasználni. Így a folyamatot egycsapásra minden lehetséges indulás mellett meghatároztuk.

Mondjuk azt, hogy az A szuperállapot üresebb B–nél, ha B–ben minden egyes automata teli van, amelyik A–ban telített, és A–ban mindegyik üres, amelyik B–ben üres, vagyis ha $A_n \le B_n$ teljesül minden n–re. Helyezzük képzeletben ilyenkor

az A sort B fölé, vagyis tekintsük az $\binom{A_n(t)}{B_n(t)}$ párokból álló $C_n(t)$ automatasort. Általában ennek négy lehetséges állapota lehetne: $\binom{0}{0}, \binom{0}{1}, \binom{1}{0}, \binom{1}{1}$. Akkor mondjuk $A_n(t)$–t üresebbnek $B_n(t)$–nél, ha $C_n(t)$–ben sehol sem fordul elő az $\binom{1}{0}$ állapot.

Mondhatjuk szemléletesen azt, hogy a $\binom{1}{0}$ állapot azért természetellenes, mert felülről a teli automata "lecsorog" az üresbe. Belátható, hogy konstrukciónk mellett ha $A(0)$ üresebb $B(0)$–nál, akkor $A(t)$ is üresebb $B(t)$–nél. Az attraktivitás nélkül is igaz az a nagyon fontos észrevétel, hogy ha a két szinten 3 szomszéd egyenlő, akkor a középső pár egyenlő marad mindaddig, amíg a szomszédjaiban a két szint állapotai megegyeznek.

Az a konstrukció folyománya, hogy a "kevert" $\binom{1}{0}, \binom{0}{1}$ állapotok csak a "tiszta" $\binom{0}{0}, \binom{1}{1}$ állapotokba mehetnek át az egyesített, két szintű automatában, hiszen egy állapotváltozás során egy helyen csak egyféle átmenet mehet végbe: vagy telítődés vagy kiürülés. Az attraktivitás esetén ha egy helyen a két szint állapota megegyezik, és a két szomszédban felül üresebb az állapot mint alul, akkor a vizsgált helyen is ez lesz az eredmény a következő állapotváltozás után. De ennél több is igaz. Az egymás fölé helyezett A, B automatasorokból kialakuló C automatában még a $\binom{0}{1}$ állapot is megszűnik fokozatosan. Az $\binom{0}{0}, \binom{1}{1}$ "tiszta" állapotokból kialakuló blokkokba ugyanis csak a széleken hatolhatnak be ezek a $\binom{0}{1}$ "kevert" állapotok. Mondhatjuk most a tisztaságot egészségnek, a kevert állapotot fertőzésnek. A kép azért pontos, mert fertőzött csak úgy lehet C–ben egy kétszintű automata, ha korábban valamelyik szomszédja az volt. Indítsuk el ezt a kétszintű rendszert úgy, hogy kezdetben minden szem fertőzött legyen. Ebben a különleges indításban jelöljük a felső szint automatáit U–val, az alsóét T–vel. Legyen tehát

$$U_n(0) = 0, \quad n = 0, \pm 1, \ldots$$

$$T_n(0) = 1, \quad n = 0, \pm 1, \ldots .$$

Ezeket a szélsőséges sorokat üresnek és telinek nevezzük, ennek megfelelően jelöljük őket $U_n(t)$–vel, $T_n(t)$–vel, az $\binom{U_n(t)}{T_n(t)}$ kétszintű folyamatot pedig jelöljük $W_n(t)$–vel. (Az idők során természetesen U–ban is lesznek telítettek és T–ben üresek, épp azt szeretnénk belátni, hogy a két szint azonosul, de a nevüket végig megtartjuk.)

A monotonitás miatt W sorai afféle csendőr-pár szerepét játszák: mivel $U(0)$ minden lehetséges $A(0)$–nál üresebb és minden lehetséges $A(0)$ üresebb $T(0)$–nál, a rendezettség az $\begin{pmatrix} U \\ A \\ T \end{pmatrix}$ három szintű folyamatban is megmarad, és ha itt az

$\binom{U}{T}$ pár tiszta, akkor $\binom{U}{A}, \binom{A}{T}$ is azok. Elegendő tehát azt belátni, hogy W–ben fokozatosan megszünnek a kevert állapotok. Ennek bizonyításához azonban további változatokra is szükségünk lesz. Ahogy a fertőzést, betegséget lázmérővel mutatjuk ki, a kitisztulás folyamatát segédfolyamatokkal fogjuk regisztrálni. Lesznek az üres és teli folyamatoknak is kevert, felemás változatai.

Jelöljük $U_n^k(t)$–vel azt a folyamatot, amelyben

$$U_n^k(0) = \begin{cases} 0, & \text{ha} \quad n < k, \\ 1, & \text{ha} \quad n \geq k, \end{cases}$$

és $T_n^k(t)$–vel azt, amelyben

$$T_n^k(0) = \begin{cases} 1, & \text{ha} \quad n < k, \\ 0, & \text{ha} \quad n \geq k, \end{cases}$$

Ezeket a folyamatokat rendre kevert üresnek, illetve kevert tisztának mondjuk. Itt az elnevezést az indokolja, hogy balról jobbra olvasva a kevert üres folyamatok "kezdetben" (kis n–ekre) üresek, a kevert teliek "kezdetben" teliek.

VÁGÁSOK A kevert üres folyamatban $t = 0$ mellett a szuperállapotban egyetlen 01 szomszéd pár található: $U_{k-1}^k(0) = 0$ és $U_k^k(0) = 1$, ezért azt mondjuk, hogy itt $(k - \frac{1}{2})$–ben "vágás" van, ez lesz egy új $U^k(t)$ folyamat kezdő értéke: $U^k(0) = k - \frac{1}{2}$. Minden egyes k egész mellett úgy alakítjuk ki az $U_n^k(t)$ folyamat alapján az $U^k(t)$ folyamatot, hogy értékei szomszédos egészek számtani közepei legyenek, és ha $U^k(t) = m - \frac{1}{2}$, akkor $U_{m-1}^k(t) = 0$, $U_m^k = 1$ legyen. Ezt azzal érjük el, hogy $U^k(t)$ állapotait is az $S(j, i)$ időpontokban változtatjuk, de csak akkor, ha $U_n^k(t)$–ben a t előtti vágásban szereplő automatapár valamelyikének változik az állapota. Ha a pár első tagja megtelik, $U^k(t)$ addig ugrik balra, amíg az első üres automatát el nem éri, ha pedig a második kiürül, $U^k(t)$ addig ugrik jobbra, amíg az első teli automatát el nem éri. A $T_n^k(t)$ kevert teli folyamat alapján hasonlóan követjük a $T^k(t)$ vágásokkal a kezdő 10 pár mozgását. Folyamatainkban persze sok más 01, illetve 10 pár is kialakul majd, lényeges a "jogfolytonosság" a definícióban, az "egyenes ágú öröklődés".

Mondjuk a három szintű $\begin{pmatrix} U_n(t) \\ U_n^k(t) \\ T_n(t) \end{pmatrix}$ folyamatban az $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ állapotokat alsó keverteknek, az $\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ állapotokat felső keverteknek. Ha $t = 0$, az $U^k(0)$ vágás előtt minden kevert állapot alsó, utána mindegyik felső, és általában vágás előtt a felső két szint automatái azonos állapotban vannak, a vágás után pedig az alsó két szint

automatái vannak azonos állapotban. Belátható a konstrukció alapján, hogy ez a tulajdonság minden t mellett megmarad. Sőt, ha $m < k$ akkor tekinthetjük a négy szintű $\begin{pmatrix} U_n(t) \\ U_n^k(t) \\ U_n^m(t) \\ T_n(t) \end{pmatrix}$ folyamatot, ebben most már háromféle kevertség fordulhat elő, de ezek is végig rendezettek maradnak: $U^m(t)$ előtt a felső három szint azonos, $U^m(t)$ sose ugorhatja át $U^k(t)$–t, köztük a felső kettő és alsó kettő szint azonos, végül $U^k(t)$ után az alsó 3 szint azonos. A legsérülékenyebbek az $U^m(t)$ és $U^k(t)$ vágások közötti keverékek: ha egyszer kipusztulnak nincs ki ujjászülné őket, ezért ha $U^m(t)$ és $U^k(t)$ valamilyen t–re megegyezik, attól kezdve egyenlő marad.

Kellően bő fantáziával végtelen hosszú tornyot rakhatunk az automatákból a k–adik sorba $U_n^k(t)$–t helyezve (a sorokat felül nagy pozitív számokkal számozzuk, és a sorszámok lefelé fogynak). "Legfelülre" aztán feléjük helyezhetjük $U_n(t)$–t, "legalulra" $T_n(t)$–t. Ebben a $+\infty$ és $-\infty$ jelű koordinátákkal kiegészitett vektorban minden egyes koordináta vagy 0 vagy 1, és a koordináták sora felülről lefelé haladva növekvő. Tehát egy oszlop lehet azonosan 0, ez a W-ben is tiszta üres állapot, lehet az oszlop azonosan 1, ez a W–ben is tiszta teli állapot, és lehet az oszlop felülről lefelé haladva egy darabig 0, aztán 1. Automatatornyunk tehát viszonylag egyszerű struktúrájú, végtelen sok sor helyett jellemezhető ezzel a váltáshellyel. Ez a kép ott van a bizonyítás hátterében, de nem segíti igazán annak megértését. A $T_n^k(t)$ automatákból persze hasonló torony rakható, csak benne a sorokat fordítva kell számozni.

Belátható, hogy a

$$P(U^m(t) = U^k(t))$$

valószínűségek tetszőleges $m < k$ párokra monoton nőnek és 1-hez tartanak ha t tart a végtelenbe.

Coupling A bizonyításban alkalmazott technika a "coupling", párosítás: bizonyos értelemben inverze annak, amiről a sztochasztikus kapcsolatok vizsgálatánál volt szó: ott két mérhető tér szorzatán volt adva egy mérték, és azt vizsgáltuk, milyen sztochasztikus kapcsolatot képes teremteni az általa összekötött terek között. Most a mértéknek csak a két térre eső marginális mértékei adottak, és azt keressük, azok között a mértékek között, amelyek marginálisai az adott mértékek melyik képes a két mérték között a legszorosabb kapcsolatot létesíteni. Itt a kapcsolat szorosságát természetesen még konkrétan meg kell határozni. Diszkrét folyamatok esetében ez gyakran egyszerűen azt jelenti, hogy a két folyamat egy idő után legyen

identikus, mintegy lehessen a két folyamatot összeragasztani.

Ezzel a technikával nagyon egyszerű például a Markov folyamatok ergodicitását bizonyítani: két tetszőleges kezdeti értékből indítva üzemeltessünk egy darabig ugyanazzal az átmenet mátrixszal két független realizációt. Ha az átmenet mátrix minden eleme pozitív, és az állapotok száma véges, akkor minden lépésben egy fix számnál nagyobb valószínűséggel fognak a folyamataink találkozni, és amíg a folyamataink függetlenek egymástól, ezek az alsó becslések össze is szorozhatóak egymással. Előbb vagy utóbb tehát egy valószínűséggel találkozik a két folyamat. Onnan kezdve meg járhatnak együtt is, hiszen az átmenet mátrixaik azonosak.

GÁCS TÉTELE Visszatérve Gray tételére felvethető a kérdés, van-e egyáltalán olyan automatasor, amelyik nem ergodikus. Ez a kérdés nem csak a számegyenes egészeiben ülő automatákra vethető fel, hanem bármilyen más struktúrára. Például síkrácsra. Furcsa módon a számegyenes esete sokkal nehezebb, mint más struktúráké. Az ember kezdetben azt hiszi, borzasztó egyszerű ellenpéldát konstruálni: ha az automaták lehetséges állapotainak a száma nagy, valahol rögzíthetjük az "igazságot", és ha valahol hiba lépne fel, a környező automatáknak egyszerűen csak "meg kell győzniük" a helyes útról letért automatát a tévedéséről, és minden rendbejön. Az egyetlen baj a stacionaritás: a renegát is hiheti önmagáról, hogy ő ismeri a helyes utat. Ezen az ember úgy próbálhat segíteni, hogy valahogy szavazástól teszi függővé az igazságot: az a jobb, amit többen mondanak. És itt ütközik az egyszerű konstrukció az állapotok számának a végességébe.

Az eredetileg Kurdjumovtól származó konstrukció alapgondolata az, hogy az automatákból blokkonként velük ekvivalens automatákat kell felépíteni. Ezekbe a logikai értelemben létező automatákba is beszivárog a legalsó szint hibája, de csökkenő mértékben, és mivel a számegyenes mindkét irányban végtelen, a logikai hierarchia "felfelé", egyre nagyobb blokkok felé korlátlanul folytatható. Ha a külső szemlélő ki szeretné olvasni a bizonyos ideje működő automatasorból az igazságot, csupán ezt a logikai láncot kell befutnia.

FELADATOK:

12.1. **(Eas)** Ergodikus automaták a síkrácson.

12.2. **(Gcv)** Gray couplingjának vizsgálata.

IRODALOM:

L. Gray(1982): The positive rates problem for attractive nearest-neighbor spin systems on Z, Zeitschrifts für Wahrscheinlichkeitstheorie und verwandte Gebiete

61, 389-404

P. Gács(1986): Reliable computation with cellular automata, Journal of Computer System Science 32, 15–78

Komlós János-Móri Tamás-Tusnády Gábor: Valószínűségi mértékek beágyazása, Egyetemi jegyzet, ELTE, TTK, 1993

13. Sztochasztikus mezők

Témák: Curie-Weiss modell, Ising folyamatok, Gibbs mezők, pontfolyamatok, kriging, sztochasztikus geometria, képanalízis,

Bevezetés Az idősorokat kétféleképpen általánosíthatjuk: valamilyen többdimenziós dologgal helyettesíthetjük magát az időt (ami természetes módon egydimenziós mennyiség), vagy az idő mellé felvehetünk más argumentumot is, amitől a folyamat függhet. Az így kapott véletlen függvényeket hívjuk sztochasztikus mezőknek.

Curie-Weiss modell Legyen x k-dimenziós, θ p-dimenziós vektor, f és g a k dimenziós tér valamely X részén értelmezett függvények, f értékei legyenek valós számok, g értékei p-dimenziós vektorok. Tegyük fel, hogy az

$$R_n(\theta) = \sum_{\{\frac{x}{n} \in X\}} \exp(n(f(\frac{x}{n}) - (\theta, g(\frac{x}{n}))))$$

összeg (ahol (.,.) a skaláris szorzást jelöli, és x a k-dimenziós tér egész koordinátáju pontjain fut) véges minden n-re, és állítsuk elő azt az eloszlást, amely az egészkoordinátájú x-hez a

$$\exp(n(f(\frac{x}{n}) - (\theta, g(\frac{x}{n}))))/R_n(\theta)$$

tömeget rendeli. Kérdés θ becslésének aszimptotikus viselkedése, ha n tart végtelenben.

A legegyszerübb speciális eset, ahol ezt az általános kérdést megvizsgálták a Curie-Weiss modell. Ez egy urna-modell: m urnába n golyót dobunk, $\nu_1, \ldots, \nu_m$ az egyes urnákbe kerülő golyók száma, és a valószínüség logaritmusa a ν_i gyakoriságok kvadratikus alakja.

Ising folyamatok Ültessünk a sík rácspontjaiba véletlen ± 1-eket úgy, hogy a szomszédos előjelek között legyen "kölcsönhatás": az egész konfiguráció valószínűsége függjön az egymás melletti eltérő előjelek számától is.

Ha mondjuk $\mid i \mid \leqq n, \mid j \mid \leqq n$, legyen $h(i,j)$ lehetséges értéke $+1$ vagy -1, és egy ilyen $h(i,j)$ függvényhez rendeljük hozzá az értékeinek S összegét, és a szomszédos eltérő értékű helyek számát jelöljük T-vel. Tegyük fel, hogy $h(i,j)$ valószínüsége az

$$exp(-(\alpha S + \beta T))$$

mennyiséggel arányos. Kérdés, van-e ennek a mezőnek határmezeje, ha n tart a végtelenbe? Általában rögzíthetjük a folyamat értékeit az értelmezési tartomány peremén, és kérdezhetjük az ilyen rögzítések különböző sorozataira ugyanezt, és azt, hogy a limesz függ-e a választott sorozattól.

GIBBS MEZŐK A fenti lehetőségek általánosításai. Preston(1974,1976) könyvei elég egyszerű bevezetést adnak, Liggett(1985) könyve elég nehezen olvasható, ami persze nem a szerző hanem a téma hibája.

PONTFOLYAMATOK A legegyszerübb az u.n. Strauss folyamat. Ez egy olyan időben változó térbeli Poisson folyamat, amelyikben a pontoknak exponenciális élettartamuk van, ez a környezetük izotróp függvénye, és a pontok születési intenzitása is a környezettől függ. Tehát ismét van két függvény, mondjuk f és g, és a születési intenzitás $\exp(const + \sum f(d(x,x_i))$, ahol az összegezés a folyamat pontjain fut, és d a távolság, x pedig az a pont, ahol az intenzitást ki akarjuk számolni, az élettartam paramétere hasonlóan $\exp(const + \sum g(d(x,x_i))$, ahol most x persze egy eleven pont.

KRIGING Itt egy izotróp Gauss mező értékeit ismerjük néhány pontban, és kíváncsiak vagyunk másutt a folyamat értékeire. A dolgot az teszi nehézzé, hogy nem ismerjük a kovarianciastruktúrát. (Ismerős ez a helyzet a Kálmán szűrésből.) A "falusi kislány" módszere persze itt is "működik" abban az értelemben, hogy eddig még senki nem tiltotta meg a használatát: először megbecsüljük a kovarianciát, majd úgy használjuk, mintha a valódi lenne.

SZTOCHASZTIKUS GEOMETRIA Egy egyszerű, az egész témát jellemző tétel a következő. Vegyünk független, egyenletes eloszlású pontokat egy centrálszimmetrikus tartományban. Akkor annak a valószínűsége, hogy a kivett pontok konvex burka tartakmazza az origót nem füg a tartomány alakjától, csak a dimenziótól, és a pontok számától, és valamilyen binomiális eloszlásra emlékeztető alakja van.

KÉPANALÍZIS A sztochasztikus mezők vizsgálatával érintkező tőle független terület: vagy maga a vizsgált kép közelíthető ismert struktúrájú mezővel, vagy a képek kapcsolata.

FELADATOK:

13.1. **(Dip)** Előjelezzük a síkrács jobb felét véletlenül. Egy tetszőleges bolyongás pályájához rendeljük hozzá az útbaejtett előjelek összegét. Mekkora ezek maximuma?

13.2. **(Isi)** Előjelezzük meg véletlenül a síkrács egy négyzet alakú részének a pontjait. Legyen P a pozitív előjelú rácspontok száma, és legyen E azoknak a szomszédos rácspontoknak a száma, amelyek előjele megegyezik. Határozzuk meg P és E együttes eloszlását.

13.3. **(Sgl)** Legyenek az $X_{ij}, 1 \le i, j \le N$ mátrix elemei $i < j$ mellett független standard normálisok, $i = j$ mellett legyen $X_{ij} = 0$, és $i > j$ mellett legyen $X_{ij} = X_{ji}$. Legyen továbbá T tetszőleges pozitív valós szám, és legyen

$$\kappa(T) \quad = \quad \sum \exp\left(\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}\varepsilon_i\varepsilon_j\right),$$

ahol a jobb oldalon az összegezés az összes lehetséges ± 1 elemű N-dimenziós vektorra vonatkozik. Mit mondhatunk a

$$P(\varepsilon = y) \quad = \quad \frac{1}{\kappa(T)}\exp\left(\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{N} X_{ij}y_iy_j\right),$$

eloszlású véletlen bináris vektorról, ahol y tetszőleges ± 1 elemű N-dimenziós vektor?

13.4. **(Pot)** Potts modell.

13.5. **(Krt)** Egy $n \times n$ méretű kertben négyféle növény nő: zöld, sárga, piros és kék. Ha valamelyik növény éppen nincs a kertben, akkor az minden egyes üres cellában valamilyen intenzitással kinőhet. A zöld intenzitása a legnagyobb, a többieké a mondott sorrendben fogy. Ha egy cellában van valamilyen növény, az a távolság négyzetével lecsengő intenzitással bizonyos más színű cellát atváltoztathat a saját színére. A zöldből sárga lesz, abból piros, és mindhármukból lehet kék. A kék önmagát pusztítja ugyancsak a távolság négyzetével lecsengő intenzitással.

13.6. **(Pkr)** Mi történik az előző feladatban, ha a szomszédság-gráfot véletlenszerűen választjuk, de távolságnak az eredeti Euklideszi távolságot használjuk?

13.7. **(Gsz)** Keressünk a csillagos égen a Göncöl Szekérhez hasonló érdekes alakzatokat.

13.8. **(Skt)** Sztochasztikus képtisztítás.

IRODALOM:

Fritz József: Az alakfelismerés statisztikai módszerei, Révész Pál és Fritz József előadásai, MTA Mat. Kut. Int., 1974

C.J. Preston: Gibbs states on countable sets, Cambridge Univ. Press 1974

C.J. Preston: Random fields, Springer, 1976

T.M. Liggett: Interactive particle systems, Springer, 1985

I. Bárány-C. Buchta(1993): Random polytopes in a convex polytope, independence of shape, and concentracion of vertices, Math. Ann. 297, 467-497

C.W. Gardiner: Handbook of stochastic methods, Springer 1990

14. Sztochasztikus optimalizálás

Témák: sztochasztikus approximáció tanulás modellek, simulated annealing, neural networks, genetikai optimalizálás.

Sztochasztikus approximáció Lásd: irodalom.

Tanulás modellek Lásd: irodalom.

Simulated annealing Lásd: irodalom.

Neural networks Lásd: irodalom.

Genetikai optimalizálás Lásd: irodalom.

Feladatok:

14.1. **(Usa)** Az utazó ügynök feladatának megoldása simulated annealinggel.

14.2. **(Unh)** Az utazó ügynök feladatának megoldása neuron hálózattal.

14.3. **(Ugo)** Az utazó ügynök feladatának megoldása genetikai optimalizálással.

14.4. **(Rsa)** Rádió-hullámhosszak optimális megválasztása simulated annealinggel (cf.1.6 Rho).

14.5. **(Rnh)** Rádió-hullámhosszak optimális megválasztása neuron hálózattal (cf.1.6 Rho).

14.6. **(Rgo)** Rádió-hullámhosszak optimális megválasztása genetikai optimalizálással (cf.1.6 Rho).

14.7. **(Mta)** Miért tojás alakú a tojás?

Irodalom:

G. Tusnády(1974): On the updated maximum likelihood estimators, Studia Scientiarum Mathematicarum Hungarica, 9, 377-389

T.M. Cover-M.A. Freedman-M.E. Hellman(1976): Optimal finite memory learning algorithms for the finite sample problem, Information and Control 30, 49-85

J.J. Hopfield(1982): Neural networks and physical systems with emergent collective computational abilities, Proc. Nat. Acad. Sci. USA 79, 2554-2558

P.J.M. Van Laarhoven-E.H.L. Aarts: Simulated annealing: theory and applications, D. Reidel, Dordrecht-Boston-Lancaster-Tokyo, 1987

D.E. Goldberg: Genetic algoritms in search, optimization and machine learning, Addison-Wesley, Reading, 1989

S. Forrest(1993): Genetic algoritms: principles of natural selection applied to computation, Science 261, 872-878

15. Fraktálok

Témák: Julia halmazok, Mandelbrot halmaz, kontrakciók lánca.

Julia halmazok Lásd: irodalom.

Mandelbrot halmaz Lásd: irodalom.

kontrakciók lánca Lásd: irodalom.

Irodalom:

B.B. Mandelbrot: The fractal geometry of nature, W.H. Feeman, 1983

16. Káosz

Témák: Komplexitás, érzékenység, ciklikusság, Feigenbaum tétele.

Komplexitás Lásd: irodalom.

Érzékenység Lásd: irodalom.

Ciklikusság Lásd: irodalom.

Feigenbaum tétele Lásd: irodalom.

Feladat:

16.1. **(Rek)** Tekintsük az $x_{n+1} = x_n - 1/x_n$ rekurziót, és legyen x_1 egy paraméterű exponenciális eloszlású valószínűségi változó. Határozzuk meg x_{100} eloszlását.

IRODALOM: P.Whittle: Systems in stochastic equilibrium, Wiley 1986

R.D. Devaney: An introduction to chaotic dynamical systems, Addison-Wesley 1989

S.N.Rasband: Chaotic dynamics of nonlinear systems, Wiley 1990

C.Robinson: Dynamical systems. Stability, symbolic dynamics, and chaos, CRC Press An Arbor 1995

KULCS

Ace	10.3	Alternating Conditional Expectation
Ads	1.0	ADatStruktúrák
Ákk	2.17	Áttörhetetlen Kulcs Készítése
Álb	8.1	Állapotteres Leírás Becslése
Ali	7.2	ALIgning
Aut	12.0	sztochasztikus AUTomaták
Bec	5.0	statisztikai BECslések
Bfa	4.2	Boole Faktor Analízis
Bfp	4.14	Behrens-Fisher Probléma
Bft	2.10	BuFfon Tű
Bkf	5.1	Bayesi KeverékFelbontás
Bpm	11.4	Bolyongás Plusz-Minusz mátrixon
Bvt	3.15	Boldog VisszaTérés
Cca	5.2	Convex Constraint Analysis
Cms	9.4	Conception-Mutation-Selection
Cst	7.3	CSaTorna kapacitás
Cxr	6.4	CoX Regresszió
Dbp	9.3	Diadikus fa Binomiális és Poisson eloszlással
Dfm	5.7	Diákok FeladatMegoldása
Dhg	4.4	Dekomponálható HiperGráfok
Dip	13.1	DIrected Polimers
Dmb	1.9	DoMinókra Bontás
Dme	9.2	Diadikus fa Minusz Egy operátorral
Düp	11.2	Dinamikus ÜzletPolitikák
Eas	12.1	Ergodikus Automaták a Síkrácson
Ebn	6.12	EloszlásBecslés Normális hibával
Eck	2.3	Egy CiKlus

Ekb	5.4	Egyenletes eloszlású pontok Konvex Burka
Elo	3.0	ELOszlások
Elt	11.5	ELőjelTanulás
Emi	6.14	EloszlásMentes Illeszkedésvizsgálat
Env	6.3	Egydimenziós Normalitás-Vizsgálat
Eta	6.11	Eloszlás TAnulása
Etr	2.14	Egyenles TeRületű téglalapok
Ewt	3.3	Egyenletes Wigner Tétel
Ext	6.9	EXponencialitás Tesztelése
Fdm	8.4	Folytonos idejű folyamat Diszkrét idejű Megfigyelése
Flf	2.1	Fiúk-Lányok Független tánca
Flt	4.5	Fiúk-Lányok Tesztelése
Fpl	1.19	A fizika törvényei szerint Pattogó Labda
Frk	15.0	FRaKtálok
Gat	6.10	GAmma eloszlás Tesztelése
Gcv	12.2	Gray Couplingjának Vizsgálata
Gen	9.0	matematikai GENetika
Gfs	1.14	Gráfok Felfúvási Száma
Ghb	5.14	Gamma Hatvány Becslése
Git	11.2	GITtins indexének a kiszámolása
Gme	8.3	Gauss-Markov folyamat Egész része
Gms	2.15	Golyók Minimális Száma
Gpp	5.6	Geometriai Paraméterű Poisson eloszlások
Grb	10.1	GRáfok Beágyazása
Gsa	3.4	Görbevonalú Szórás Analízis
Gsz	13.7	Göncöl SZekér keresése
Gtp	9.1	Genetikai Tanácsadás Programja
Gzr	6.13	GejZíR
Hdg	2.5	Huffman kód alapján Diszkrét Generálás
Hip	4.0	HIPotézisek vizsgálata
Hka	8.2	Heterogén Kvadratikus Alakok
Hmt	6.2	Hierarchic MaTching
Hpr	1.1	Hashing PeRmutációi
Ids	8.0	IDőSorok
Igs	10.2	Irányított Gráf csúcsainak Sorrendje

Ind	3.14	INDiánok
Isi	13.2	ISIng
Jjb	1.18	Jancsi és Juliska Bolyongása
Kap	10.0	sztochasztikus KAPcsolatok
Kct	1.3	K-CenTrum
Kii	7.5	KIpusztulás Ideje
Kiv	7.4	KIpusztulás Valószínűsége
Kln	5.15	KétLépcsős Normális
Knk	4.7	Korrelált Normálisak Keverése
Knv	2.12	KoNVolució
Kör	1.10	KÖRmentesség tesztelése irányított gráfban
Krt	13.5	KeRT
Ksz	16.0	KáoSZ
Ktr	11.0	sztochasztikus KonTRoll
Lgr	4.3	LoGisztikus Regresszió
Lhú	2.7	LegHosszab Út
Lqc	11.3	Linear-Quadratic Control
Lsi	7.8	Lovász-Simonovits Integrál
Luk	2.13	leghosszab LUK a bolyongásban
Mab	7.1	MAximális Blokkpár
Mb1	5.1	MBI Poisson modellje
Mez	13.0	sztochasztikus MEZők
Mfp	4.1	Maximális Független Partició
Mkv	7.0	MarKoV láncok
Mnp	11.1	Mátrix-játék NyeregPontjának a meghatározása
Mrb	7.11	Markov folyamat Rendjének Becslése
Mrg	6.6	Monoton ReGresszió
Mri	6.8	Monoton Regresszió Irányított gráfon
Mrk	6.7	Monoton Regresszió Konfidencia intervalluma
Mta	14.7	Miért Tojás Alakú a tojás?
Mtg	3.16	Minimális Távolság Gráfja
Mtn	1.8	Maximális Területű diszjunkt Négyzetek
Mtv	2.16	Minimális TáVolság
Mvg	1.12	Maximális VáGás
Ndc	3.6	Normális eloszlás Dirichlet Cellái

Ndg	3.8	Normális eloszlás Dirichlet celláinak a gráfjának a Felfúvási száma
Ndg	3.7	Normális eloszlás Dirichlet celláinak a Gráfja
Neb	5.10	Normális eloszlás Előjelei alapján Becslés
Neg	4.8	Normálisak és EGyenletesek tesztelése
Nex	4.9	Normálisak és EXponenciálisak tesztelése
Nkb	3.9	Normálisak Konvex Burka
Nkc	3.5	Normálisak K-Centruma
Nkf	5.12	Normálisak KeverékFelbontása
Nkv	3.1	Normális Küszöb Valószínűség
Npp	6.5	többdimenziós Normalitás-vizsgálat Projection Pursuit módszerrel
Npr	6.0	Nem PaRaméteres módszerek
Npt	3.10	Normálisak Páronkénti Távolsága
Nsd	1.16	N pont S távolságra D dimenzióban
Nse	5.11	Normális Sűrűség az Egységkockán
Nst	4.11	Normálisok Szekvenciális Tesztelése
Ntc	4.10	Normálisakat a Tükörképükre Cseréljük
Ntk	4.13	Normálisak Tesztelése Költséggel
Nvk	5.13	Normálisak Várható értékének a Konfidencia intervalluma
Nwt	3.2	Normális Wigner Tétel
Opt	1.5	OPTimalizálás
Önl	1.5	ÖNLogisztikus eloszlás
Pck	2.4	Permutáció CiKlusa
Pkr	13.6	Permutált KeRt
Pot	13.4	POTts modell
Prp	2.2	PáRos Permutáció
Prt	2.8	PaRTició
Ptf	1.15	Pontok Térbeli Forgatása
Opt	14.0	sztochasztikus OPTimalizálás
Qma	7.6	Quadratic MArkov
Qmt	6.1	Quadratic MaTching
Rek	16.1	REKurzió
Rgo	14.6	Rádió-hullámhosszak optimális megválasztása Genetikai Optimalizálással
Rho	1.6	Rádió-Hullámhosszak Optimális megválasztása
Rhs	1.2	Rossz HaShing
Ria	4.12	RIAsztás

Rmb	7.7	Rejtett Markov Becslése
Rnd	2.0	RaNDomizálás
Rnh	14.5	Rádió-hullámhosszak optimális megválasztása Neuron Hálózattal
Rsa	14.4	Rádió-hullámhosszak optimális megválasztása Simulated Annealing algoritmussal
Run	2.9	Leghosszabb monoton szakasz
See	5.3	Szimplexen Egyenletes Eloszlás
Sgl	13.3	Spin GLass
Shg	4.6	Szabad HiperGráfok
Skp	6.15	Sokdimenziós Kolmogorov Próba
Skt	13.8	Sztochasztikus KépTisztítás
Srt	5.5	Stepwise Regresszió Tesztelése
Sza	1.17	SZAvak
Szi	11.7	SZInbád
Szk	9.5	SZűrés-Konvolució
Tac	1.7	TACtics
Tki	7.10	Többdimenziós Kipusztulási Idő
Tkv	7.9	Többdimenziós Kipusztulási Valószínűség
Tlt	1.13	TeLíTettség
Trm	3.13	Tiszta RészMátrix
Tsk	1.4	Többdimenziós SKálázás
Ugo	14.3	Utazó ügynök feladatának megoldása Genetikai Optimalizálással
Unh	14.2	Utazó ügynök feladatának megoldása Neuron Hálózattal
Usa	14.1	Utazó ügynök feladatának megoldása Simulated Annealing algoritmussal
Ügm	1.11	Üres GöMb
Vpr	5.9	Véletlen PeRmutációk
Vss	2.11	Véletlen Simpson Szakasz
Vst	4.15	Véletlen Számok Tesztelése
Wlg	3.11	Wiener Lepedő Generálása
Wtb	3.11	Wigner tételének a bizonyítása
Zwt	2.6	van ZWeT eljárása

Áttekintő tábla

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ads	Rnd	Elo	Hip	Bec	Npr	Mkv	Ids	Gen	Kap	Ktr	Aut	Mez	Opt	Frk	Ksz

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	Hpr	Flf	Nkv	Mfp	Bkf	Qmt	Mab	Álb	Gtp	Grb	Mnp	Eas	Dip	Usa	1	Rek	1
2	Rhs	Prp	Nwt	Bfa	Cca	Hmt	Ali	Hka	Dme	Igs	Git	Gcv	Isi	Unh	2	2	2
3	Kct	Eck	Ewt	Lgr	See	Env	Cst	Gme	Dbp	Ace	Lqc	3	Sgl	Ugo	3	3	3
4	Tsk	Pck	Gsa	Dhg	Ekb	Cxr	Kiv	Fdm	Cms	4	Bpm	4	Pot	Rsa	4	4	4
5	Opt	Hdg	Nkc	Flt	Srt	Npp	Kii	Wlg	Szk	5	Elt	5	Krt	Rnh	5	5	5
6	Rho	Zwt	Ndc	Shg	Gpp	Mrg	Qma	6	6	6	Düp	6	Pkr	Rgo	6	6	6
7	Tac	Lhú	Ndg	Knk	Dfm	Mrk	Rmb	7	7	7	Szi	7	Gsz	Mta	7	7	7
8	Mtn	Prt	Ndf	Neg	Mb1	Mri	Lsi	8	8	8	8	8	Skt	8	8	8	8
9	Dmb	Run	Nkb	Nex	Vpr	Ext	Tkv	9	9	9	9	9	9	9	9	9	9
10	Kör	Bft	Npt	Ntc	Neb	Gat	Tki	10	10	10	10	10	10	10	10	10	10
11	Ügm	Vss	Wtb	Nst	Nse	Eta	Mrb	11	11	11	11	11	11	11	11	11	11
12	Mvg	Knv	Önl	Ria	Nkf	Ebn	12	12	12	12	12	12	12	12	12	12	12
13	Tlt	Luk	Trm	Ntk	Nvk	Gzr	13	13	13	13	13	13	13	13	13	13	13
14	Gfs	Etr	Ind	Bfp	Ghb	Emi	14	14	14	14	14	14	14	14	14	14	14
15	Ptf	Gms	Bvt	Vst	Kln	Skp	15	15	15	15	15	15	15	15	15	15	15
16	Nsd	Mtv	Mtg	16	16	16	16	16	16	16	16	16	16	16	16	16	16
17	Sza	Ákk	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17
18	Jjb	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
19	Fpl	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19
	Ads	Rnd	Elo	Hip	Bec	Npr	Mkv	Ids	Gen	Kap	Ktr	Aut	Mez	Opt	Frk	Ksz	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	

AS Algoritmusok

Journal of the Royal Statistical Society, Section C

Applied Statistics

144 random R C tables

159 random R C tables

183 random numbers

187 incomplete gamma

189 ML for beta

191 ARMA
195 multivariate normal probabilities
197 ARMA
203 ML for mixtures
206 isotonic regression
207 general loglinear model
217 DIP test
244 decomposability of log-linear models
248 goodness-of-fit tests for empirical distribution function
285 multivariate normal probabilities
288 exact Smirnov test
191 DNS frequencies
292 Fisher information

Tárgymutató

(A számok a fejezetek sorszámai)

ACE 10
alakfelismerés 13
aligning 7
állapotteres leírás 8
autoregresszív folyamatok 8
Bellman egyenlete 11
bin packing 6
biztosításmatematika 6
Boole faktoranalízis 4
bootstrap 2
χ^2-próba 4
coupling 12
Cox regresszió 6
csatorna-kapacitás 7
Curie-Weiss modell 13
elágazó folyamatok 7
EM algoritmus 5
entrópia 7

ergodicitás 12
fehérjék kódolása 9
fehér zaj 8
függetlenség tesztelése 4
gamma eloszlás 2
genetikai optimalizálás 14
genetikai tanácsadás 9
Gibbs mezők 13
Gittins indexe 11
görbevonalú szórásanalízis 3
gráfok beágyazása 10
hashing 1
Ising folyamatok 13
jackknife 2
Kálmán szűrés 8
Kaplan Meier becslés 6
képanalízis 13
keverékek felbontása 5
kísérletek tervezése 1
kódolás 7
Kriging 13
küszöb modell 9
kvantilis transzformáció 2
lineáris regresszió 3
Ljapunov egyenlet 8
logisztikus regresszió 4
Log-lineáris modellek 4
Markov láncok 7
matching 6
mátrix-játék 11
maximum likelihood 5
Mendel törvényei 9
mérhető mennyiségek öröklődése 9
momentumok módszere 5
monoton regresszióó 3

mozgó átlag folyamatok 8
mutáció és szelekció 9
nem mérhető mennyiségek öröklődése 14
neural networks 14
nem paraméteres módszerek 6
normál egyenlet 3
normalitásvizsgálat 6
optimális portfolio 11
Poisson folyamatok 13
predikció 8
projection pursuit 6
rejtett Markov modell 7
relációs adatbázisok 1
scheduling 11
simulated annealing 14
sorban állás 7
statisztikai programcsomagok 4
stopping rules 11
szekvenciális módszerek 5
sztochasztikus approximáció 14
sztochasztikus automaták 12
szűrés 2
többdimenziós integrál 7
tanulás modellek 14
titkosítás 2
többdimenziós skálázás 1
többszempontú optimalizálás 10
túlélésbecslések 6
véletlen permutációk 2
Ziv-algoritmus 7