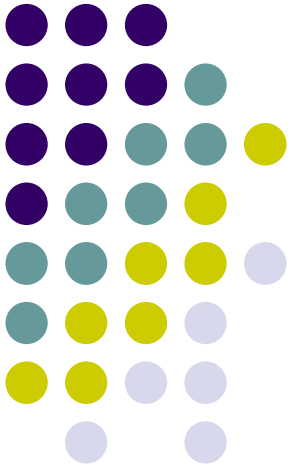
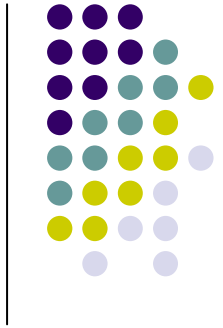
A sokváltozós statisztikai módszerek egymást kiegészítő alkalmazásai

Meszéna György 80. születésnapjára



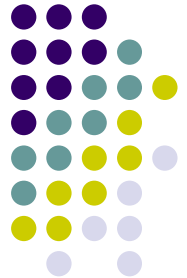
Kapcsolatunk



- 50 éves volt 2010-ben a Terv-mat szak
- Meszéna György vezette a Gazdaságmatematika szakágazatot és
- Matematikai statisztikát oktatott
  - Szeminárium vezető lettem mellette

- TDK témám: faktorelemzés
- szakdolgozatom: lineáris szeparáció
- dr univ, MDS
- CSc, több módszer együttes alkalmazása
- 1997: Térstatisztika könyv **lektorálás**
- 2004: Alakfelismerés – szerzőtárs

Murphy szerint Booker törvénye:



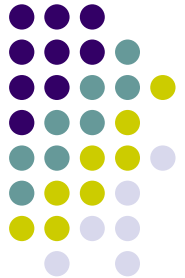
Egy gramm alkalmazás felér egy tonna elvonatkoztatással.

- Érdekes alkalmazás: az emberi arc 22 pontját leírja adatokkal és felismeri a felhasználót
  - telefonáló
  - bankkártya tulajdonos
  - ajtó kinyílik

Google-ben - Meszéna és szerzőtársai - könyvek megtalálhatóak:

- **Dr. *Meszéna György* könyvei | Nálunk megvan, vagy megszerezzük!**

Elemzői szemléletmód szerint



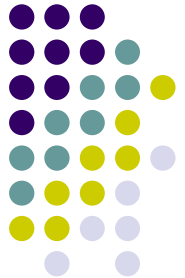
Feltáró (exploratív)

1. Leíró statisztikák
2. Hierarchikus klaszterezés
3. Lépésenkénti regresszió
4. Főkomponens elemzés
5. Lépésenkénti diszkriminancia

Megerősítő (konfirmatív)

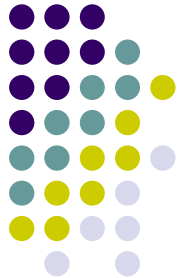
1. Kereszttábla
2. Nem-hierarchikus klaszterezés
3. Regresszió (Enter)
4. Faktorelemzés
5. Diszkriminancia (Enter)

Kombinált alkalmazások – 3 út



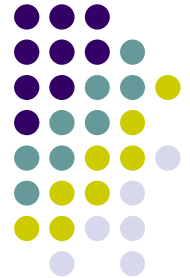
1. Adatvezérelt módszerek:
   - feltáró (vagy robusztus) módszerek: PCA, MDS, M-esztimátorok
   - információ kinyerő eljárások (pl: bootstrap (n elemből ismétléses n elemű minta) és jackknife (n*(n-1) elemű minta) is
   - diagnosztikai elemzések: extrémek sorkihagyása regresszióban, diszkriminancia elemzésben
2. Modellvezérelt elemzések
3. Szintézis

2. Modellvezérelt elemzések



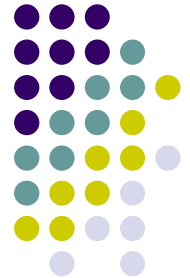
- A modellekből indulunk ki
- alkalmazáshoz meghatározott előfeltételek ellenőrzése (pl. eloszlás, linearitás)
- GLM: általánosított lineáris modell - részei
  - a (kétváltozós) lineáris regresszió, szóráselemzés
  - logisztikus regressziós (logit) modellt

3. Szintézis: elemzési láncok



- adat- és modellvezérelt eljárások kombinált alkalmazása
- **egy módszercsalád** több eljárásának alkalmazása (hierarchikus klaszter eljárások)
- több adatvezérelt módszer egymást követő alkalmazása (főkomponensből diszkrimináló függvény)
- Feltáró és megerősítő elemzések szinergiája: kevés az „illeszkedést” mérő mutató

Az alkalmazás problémái: előfeltételek, ajánlások vannak



A változók

- száma,
- eloszlása,
- függetlensége

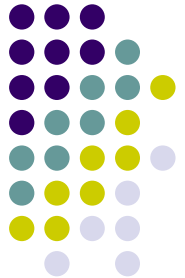
Példa: azonos kovariancia struktúra feltételezése

A megfigyelések

- száma
- szerkezete

Példa: homogén adathalmaz vagy előre ismert/ feltételezett csoportok vannak

Nagy adathalmaz öröme?



1. Hi-cluster és MDS (max 100*100)

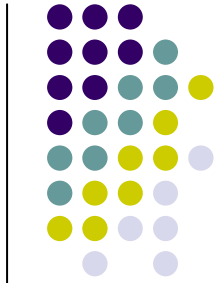
csak a változók kapcsolatrendszerének leírására alkalmazhatóak

2. K-közép klaszter: extrémek egyedül, és hatalmas méretű csoportok képződnek

3. Faktor, regresszió

Korrelációs kapcsolatok: „kis érték” is szignifikáns lesz

Sok – és nem független - változó?



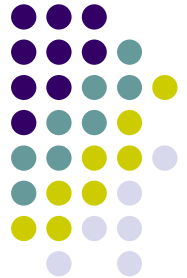
Dimenziócsökkentés hatékonysága

- Főkomponens vagy faktorelemzés – megőrzött összvariancia hányad
- MDS skálatérkép Stress függvény
- Kanonikus diszkrimináló tér kanonikus korreláció

Változó-szelekció

- Lépésenkénti változó bevonás
- Hátránya: a többi változó hatása *nagyrészt* figyelmen kívül marad
- Előnye: könnyebb értelmezni

A megfigyelések tere



Homogén adathalmazunk van

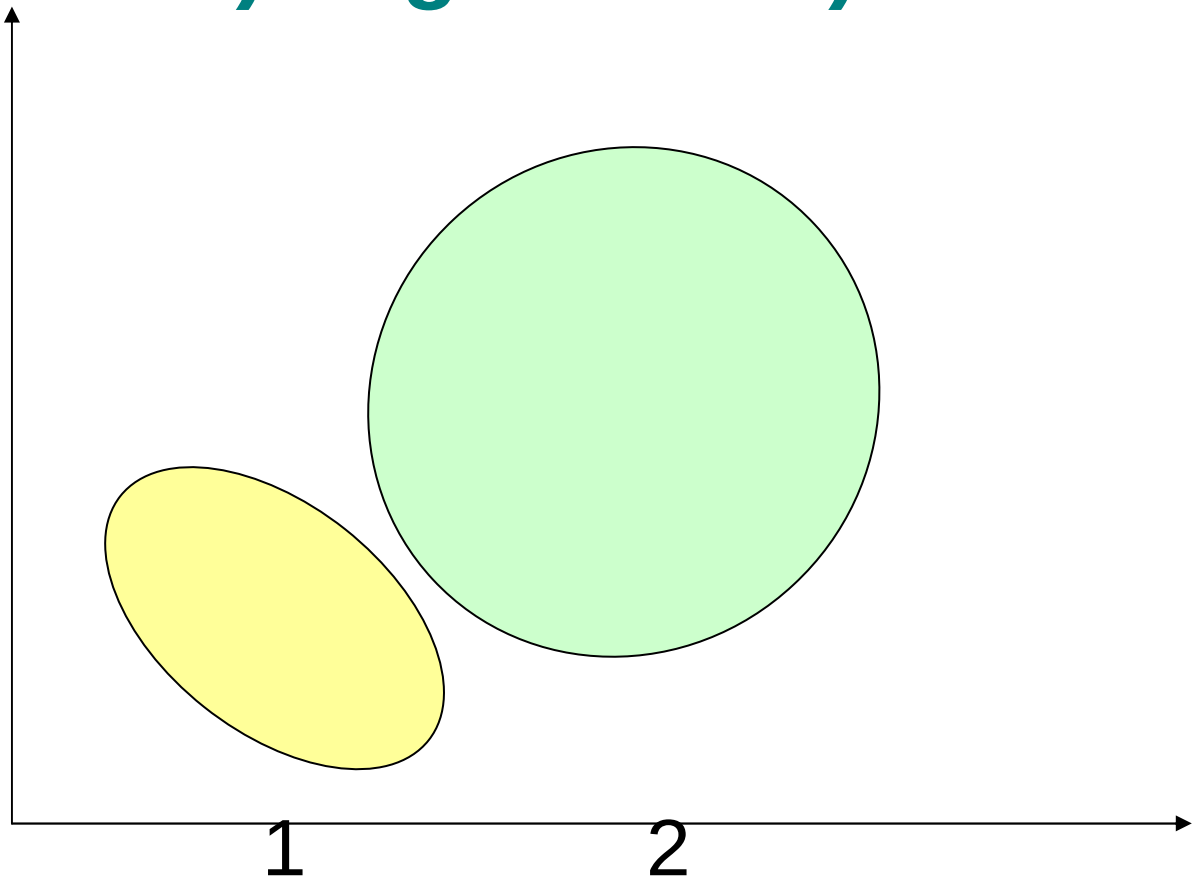
- Regresszió
- Faktorelemzés
- Sokdimenziós skálázás

Almintákból áll/ vagy feltesszük, hogy tagolt adathalmazunk van

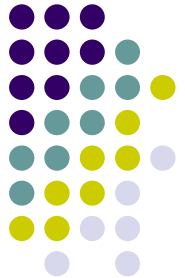
- Klaszter
- Diszkriminancia
- Logisztikus regresszió

Alminták együtt pozitív korreláció, erős első komponens

külön: 1) negatív és 2) korrelálatlan



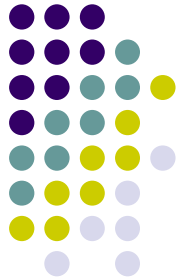
Hol hibázhatunk?



- Faktortérben klaszterezés
  - homogén adathalmazt feltételező faktorokat állítunk elő - és
  - ezek terében belső tagozódást keresünk

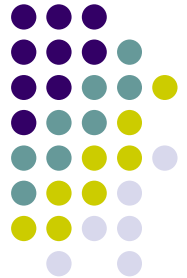
- Nominális/ordinális szinten mért változókra lineáris korreláció, regresszió, faktorelemzés
  - ezen adatok átlaga, szórása nem értelmezhető

Faktor-score előny/hátrány?



- Általában nem követ normális eloszlást a faktor score
- A változók nem azonos előjelű korrelációja miatt két póluson mér a komponens, nehezebb az origó körül az értelmezés
- Sztenderdizált faktor koordináta megtartása vagy a különböző varianciájú faktor a fontos?

Faktortér és skálatérkép egymást erősítheti - 1



PCA koordináta

- Input: X (n*p)-ből

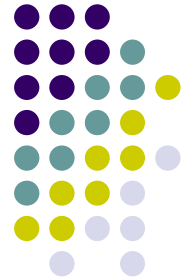
R (p*p) korrelációs mátrix

- R sajátérték-sajátvektor felbontása
- Y=XA score-k
- var(y)= sajátérték

MDS koordináták

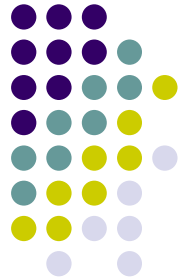
- Input: X(n*p) vagy D(n*n) euklideszi távolságok mátrixa
- B (n*n) sajátérték-sajátvektor felbontása
- $Y=AL^{1/2}$ koordináták
- Metrikus skálázás

Faktortér és skálatérkép egymást erősítheti - 2



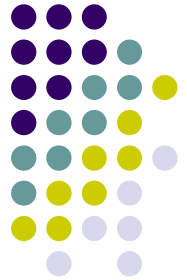
- Azonos matematikai eszköztár, de más mátrix dekompozíciója
- Sajátvektor előjele tetszőleges, emiatt is eltérhetnek
- Csak normalizált sajátvektorokra egyeznek meg az eredmények
- Nem lineáris kapcsolat (vagy gyenge korrelációk) esetén MDS robusztusabb

Faktortér és skálatérkép egymást erősítheti - 3



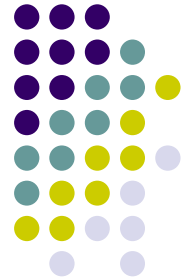
- Változók kapcsolódása, dimenziócsökkentés összevetése az MDS térképpel
- Input: X(n*p) vagy D(p*p) euklideszi távolságok mátrixa
- B (p*p) sajátérték-sajátvektor felbontása
- Változók egymással bezárt szöge látható: Korreláció= bezárt szög cosinusa
- k<p dimenziós illeszkedés jósága mérhető

Klaszterezés és skálázás egymást erősítheti



- Valóban csoportokra tagolódik az adathalmaz?
  - Skála térképen látható
- Hány csoportra bontható?
  - Hierarchikus klaszter, dendrogram ábra mutatja
  - Skála térképen is látható
  - McQuen-féle k-közép klaszterrel összevethető
- A változók kapcsolódását látjuk
  - Dendrogramon
  - MDS térképen

Két csoport szétválasztható?
Melyik csoportba, milyen valószínűséggel?



Diszkriminancia elemzés

Előfeltételei:

1. Többdim. normális eloszlás
2. Azonos kovar. strukt.
3. Intervallum skálán mért független változók

Előny: több csoportra is

Logisztikus regresszió

- Előfeltétel?

Előnyei:

1. „kvalitatív” változó is magyarázó változó lehet
2. Interakció is megengedett

Házas- nem házas ügyfelek szétválasztása

- Diszkriminancia elemzés

  (AUC=0,918) és logisztikus regresszió (AUC=0,921) is használható!

SPSS: Insurance-claims.sav