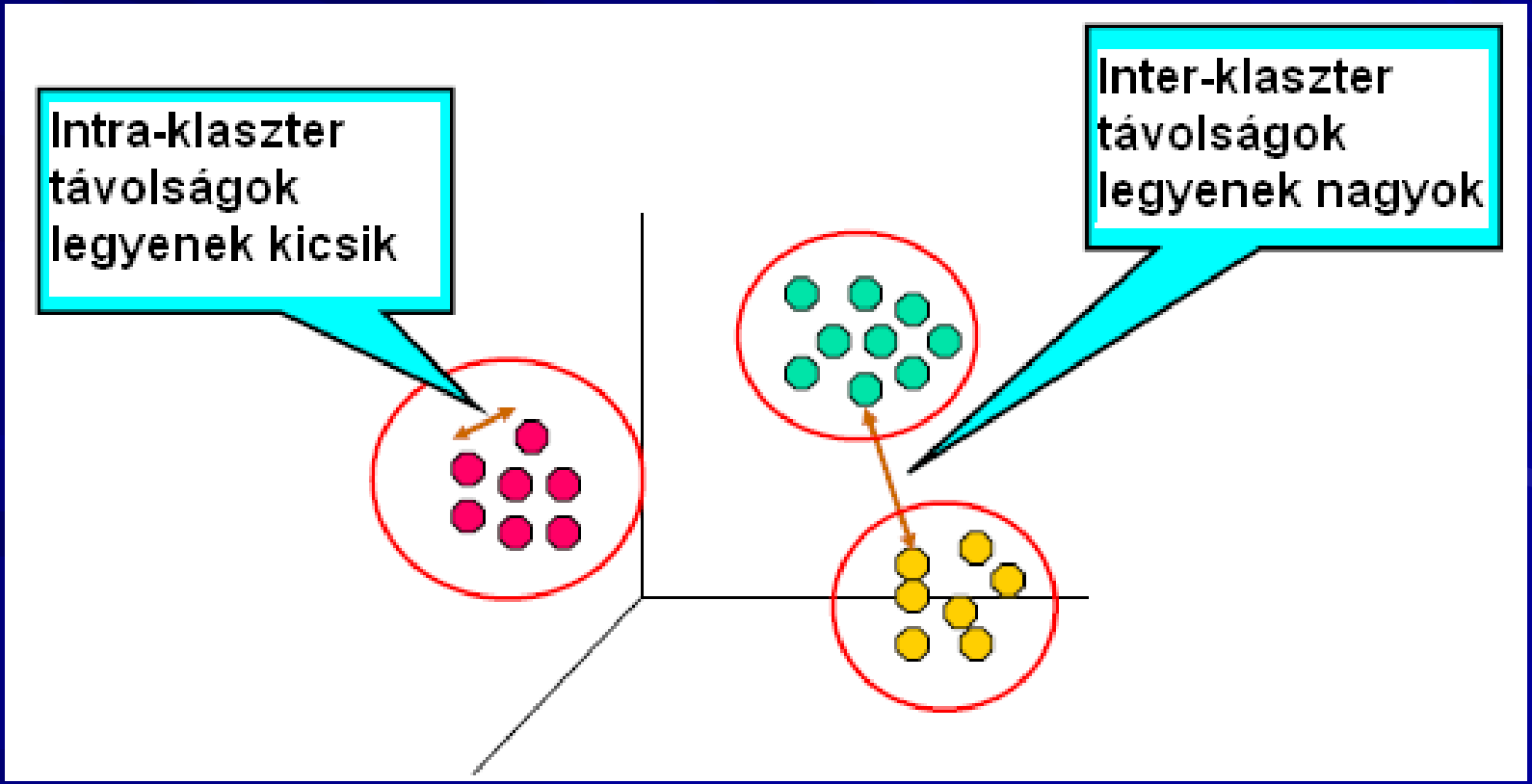
# Izgalmas újdonságok a klaszteranalízisben

Vargha András
KRE és ELTE, Pszichológiai Intézet

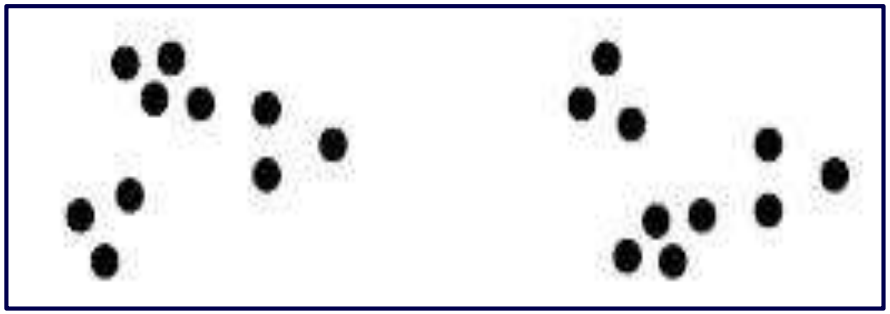
# Mi a klaszteranalízis (KLA)?

Keressük a személyek (vagy bármilyen objektumok) olyan csoportjait, ahol az egy csoportba tartozók egymásra „hasonlítanak”, de más csoportokba tartozó személyektől „különböznek”.

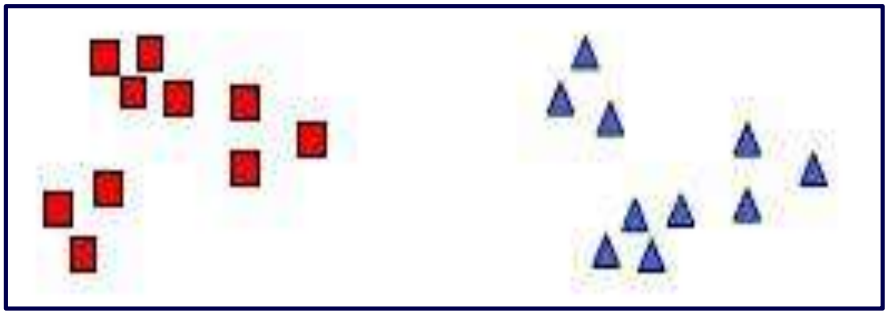


# A klaszterek néha nem egyértelműek

**Hány klaszter van itt?**



**2 klaszter?**



**6 klaszter?**

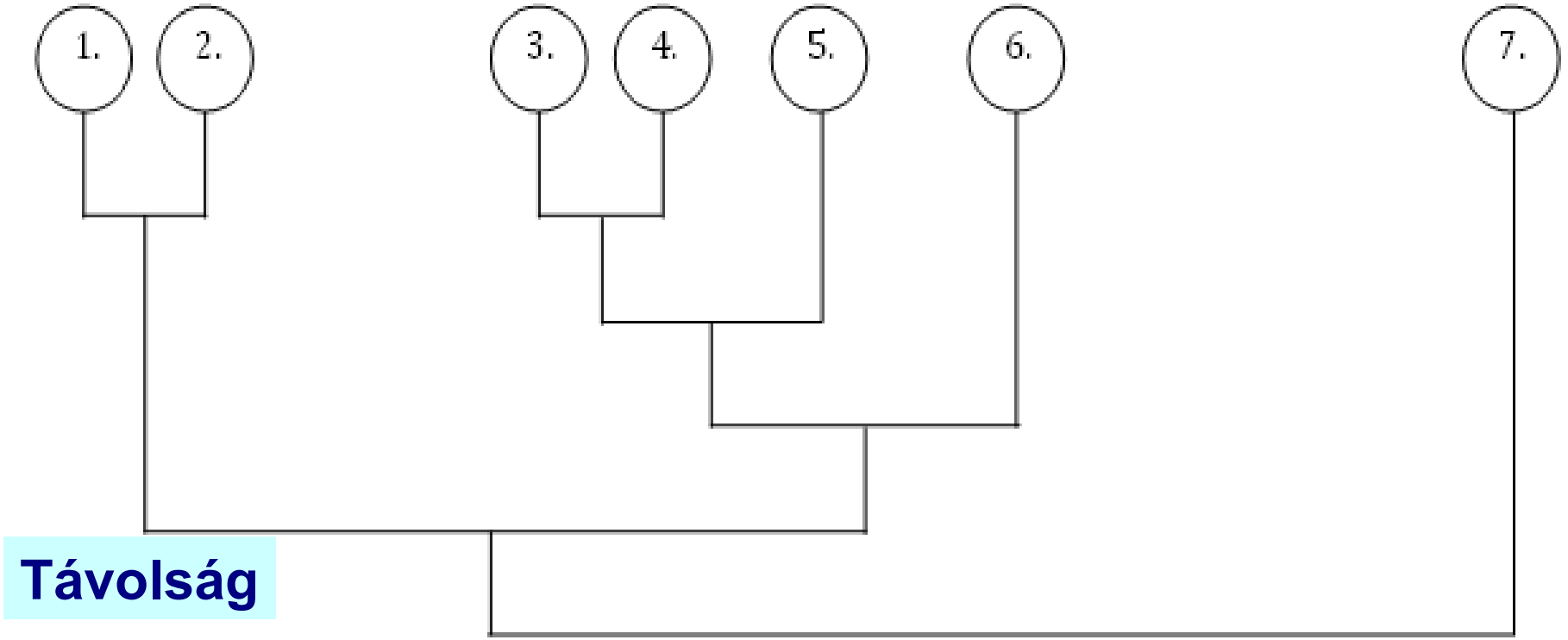
# Alapvető kritérium a személyek hasonlósága, távolsága

1) A mintázatok hasonlósága mérhető két személy adatsorának a korrelációjával. Minél nagyobb a korreláció (előjelesen!), annál hasonlóbb a két személy.
2) A hasonlóságot (távolságot) leggyakrabban az adatsorbeli értékek átlagos távolságával szokták mérni (Euklideszi távolság). Egy ilyen variáns az ASED (átlagos négyzetes eltérés, Average Squared Euclidian Distance).

# Klaszteranalízisek fő típusai

- Hierarchikus (HKA): egymásra épülő klasszifikációk rendszere, melyhez úgy jutunk, hogy lépésenként egyesítünk vagy felbontunk klasztereket
  - Agglomeratív: sok kis klaszterrel kezdünk, egy naggyal végzünk (ez a szokásos)
  - Osztódó: egy nagy klaszterrel kezdünk, sok kicsivel végzünk
- Nemhierarchikus (k-központú, particionáló): Esetek optimális besorolása k számú csoportba (k fix)
- Modell-alapú: Feltételezett keverék-eloszlás felbontása komponensekre

# A HKA szemléltetése



# Klasztertávolság az összevonásnál

- Minimális távolság módszere (Single linkage method)
- Maximális távolság módszere (Complete linkage method)
- Átlagos távolság módszere (Average linkage between groups)
- Medián módszer (Median method)
- Centroidok távolsága (Centroid method)
- Indirekt hierarchikus módszerek
  - Ward-féle módszer
  - Átlagos távolság módszere

# Egy HKA 10 és 2 klaszter között

| k | EESS% | PB | SilCoef | HCátlag | HC min-max | Egyesítendő klaszterek |
|---|---|---|---|---|---|---|
| 10 | 78,30 | 0,360 | 0,529 | 0,351 | 0,15-0,82 | 17 (22), 18 (44) |
| 9 | 76,90 | 0,360 | 0,527 | 0,373 | 0,15-0,82 | 19 (33), 23 (56) |
| 8 | 74,81 | 0,359 | 0,521 | 0,406 | 0,15-0,84 | 9 (190), 47 (131) |
| 7 | 72,68 | 0,449 | 0,564 | 0,440 | 0,16-0,84 | 16 (44), 50 (76) |
| 6 | 69,97 | 0,444 | 0,567 | 0,482 | 0,16-1,11 | 20 (162), 36 (88) |
| 5 | 66,97 | 0,477 | 0,602 | 0,530 | 0,28-1,11 | 17 (66), 19 (89) |
| 4 | 62,93 | 0,472 | 0,616 | 0,594 | 0,28-1,11 | 16 (120), 17 (155) |
| 3 | 52,87 | 0,413 | 0,612 | 0,753 | 0,28-1,60 | 9 (321), 20 (250) |
| 2 | 40,07 | 0,583 | 0,700 | 0,957 | 0,65-1,60 | 9 (571), 16 (275) |

Ennek alapján próbálunk dönteni arról, hogy mely k klaszterszám megoldása tetszik

# Klaszter kvalifikációs mutatók (QC-k)

- EESS%: Klaszterek által magyarázott varianciaarány (Explained Error SS)
- PB: Esetpárok mintájában az egy klaszterbe tartozás és az esettávolság közti korreláció
- SilCoef (Silhouette eh.): A klaszterstruktúra szeparáltságának egyik mértéke
- HC (homogenitási együttható): Egy-egy klaszterben az esetek átlagos távolsága
- HCátlag: A k klaszter HC-értékének átlaga
- HCmin-max: Legkisebb és legnagyobb HC

# EESS%

$$EESS\% = 100 \cdot \frac{SS_{total} - SS_{cluster}}{SS_{total}}$$

- Többdimenziós eta-négyzet (magyarázott varianciaarány: MV%)

# Klaszter pont-biszeriális korreláció (PB)

$$PB = r(X, Y)$$

- r(X,Y)-t az N személyből kiválasztható összes (n = N(N-1)/2) számú) pár mintáján számítjuk ki
- X: Egy klaszterbe esnek? (0: nem, 1: igen)
- Y: A pár két tagjának Euklideszi távolsága (ASED)

# Alternatív formula PB-re

$$PB = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

- PB jelzi, hogy átlagosan mennyivel vannak távolabb egymástól a különböző klaszterbe tartozó személyek, mint az egyazon klaszterbe tartozók

# Klaszter delta együttható (CLdelta)

$$\text{CLdelta} = \frac{M_1 - M_0}{s_{n-1}}$$

- A PB mutató Cohen-delta-féle változata, értelmezése hasonló PB-éhez

# Silhouette együttható (SC)

Az i-edik személyre legyen

$$SC_i = (B_i - A_i)/\max(A_i, B_i),$$

ahol

$A_i$ = átlagos távolság a saját társaktól,

$B_i$ = átlagos távolság az idegen társaktól.

SC: az összes egyedi $SC_i$ érték átlaga

- Értelmezés: a személyek mennyivel vannak közelebb saját klasztercentrumukhoz, mint a legközelebbi idegen klaszter centrumához

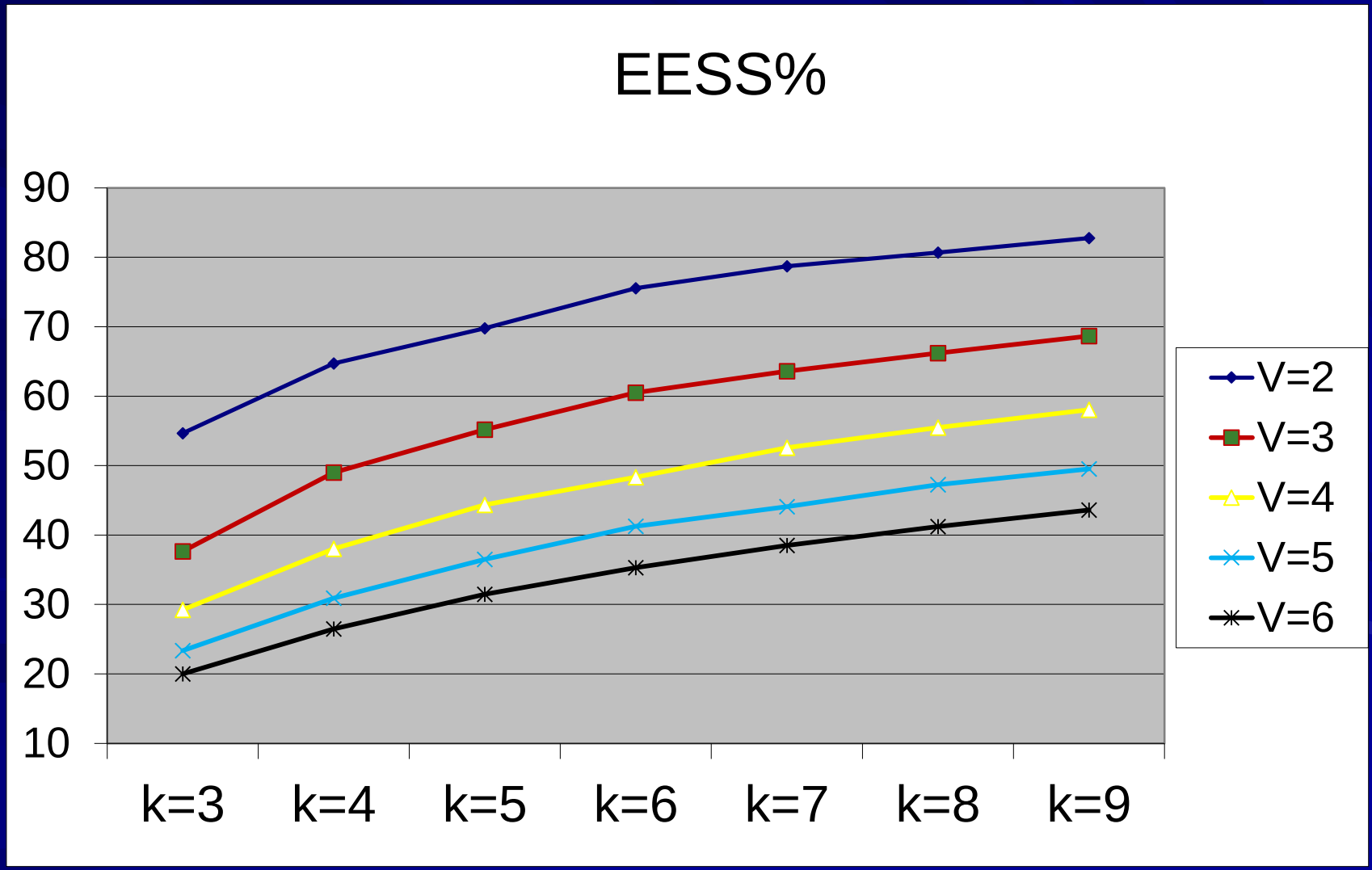
# A QC mutatók típusai

- Homogenitást, kohéziót mérik: EESS%, HCátlag
- Szeparációt mérik: SC, XBmod, GDI24
- Mindkettőt mérik: PB, CLdelta
- QC-kről bővebben:
  - Desgraupes, B. (2013).
  - Vargha–Torma–Bergman, 2015
  - Vargha–Bergman–Takács, 2016

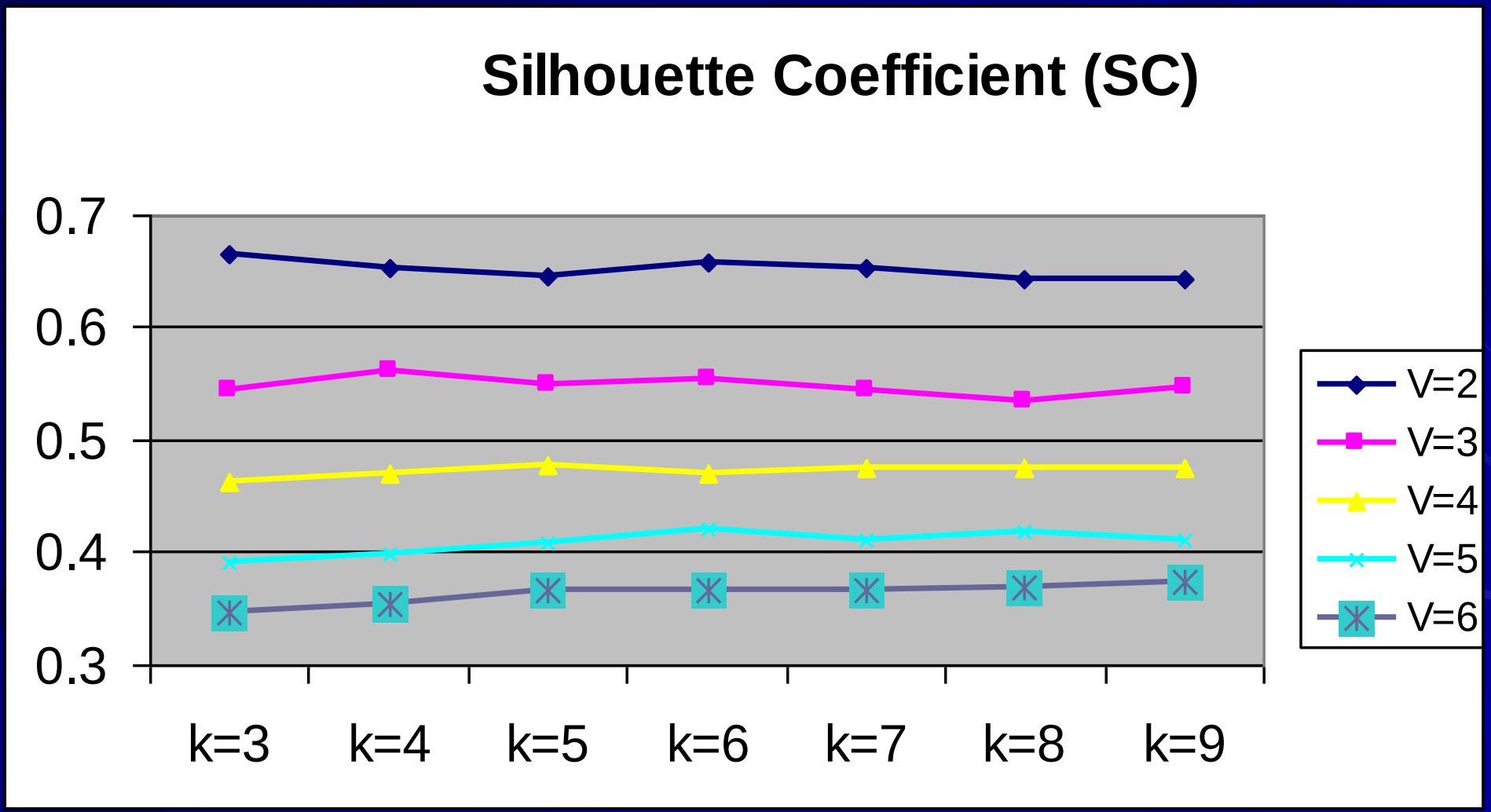
# Nemhierarchikus k-központú klaszteranalízis (KKA)

- A klaszterszámot (k) előre rögzítjük
- Optimalizálásos módszer több iterációval
- A kezdő besorolás véletlenszerű is lehet
- Eseteket ide-oda tesszük (relokáció), amíg EESS% (~MV%) maximális nem lesz
- A végén homogénebb struktúrát kapunk, mint a HKA-ban
- Kiértékelés QC-k segítségével

# Probléma a QC-kkel: erősen függnek a klaszter- és a változószámtól



# A Silhouette-együttható erősen függ a változószámtól



# Megoldás: QC viszonyítása random mintán kapott struktúra QC-jéhez

$$\mathrm{MORI} = \frac{\mathrm{QC} - \mathrm{QC}_{\mathrm{rand}}}{\mathrm{QC}_{\mathrm{best}} - \mathrm{QC}_{\mathrm{rand}}}$$

MORI = Measure of Relative Improvement
Ez a kapott struktúra belső validitásának (internal validity) legfontosabb mutatója

(Vargha–Bergman–Takács, 2016)

# Részletek

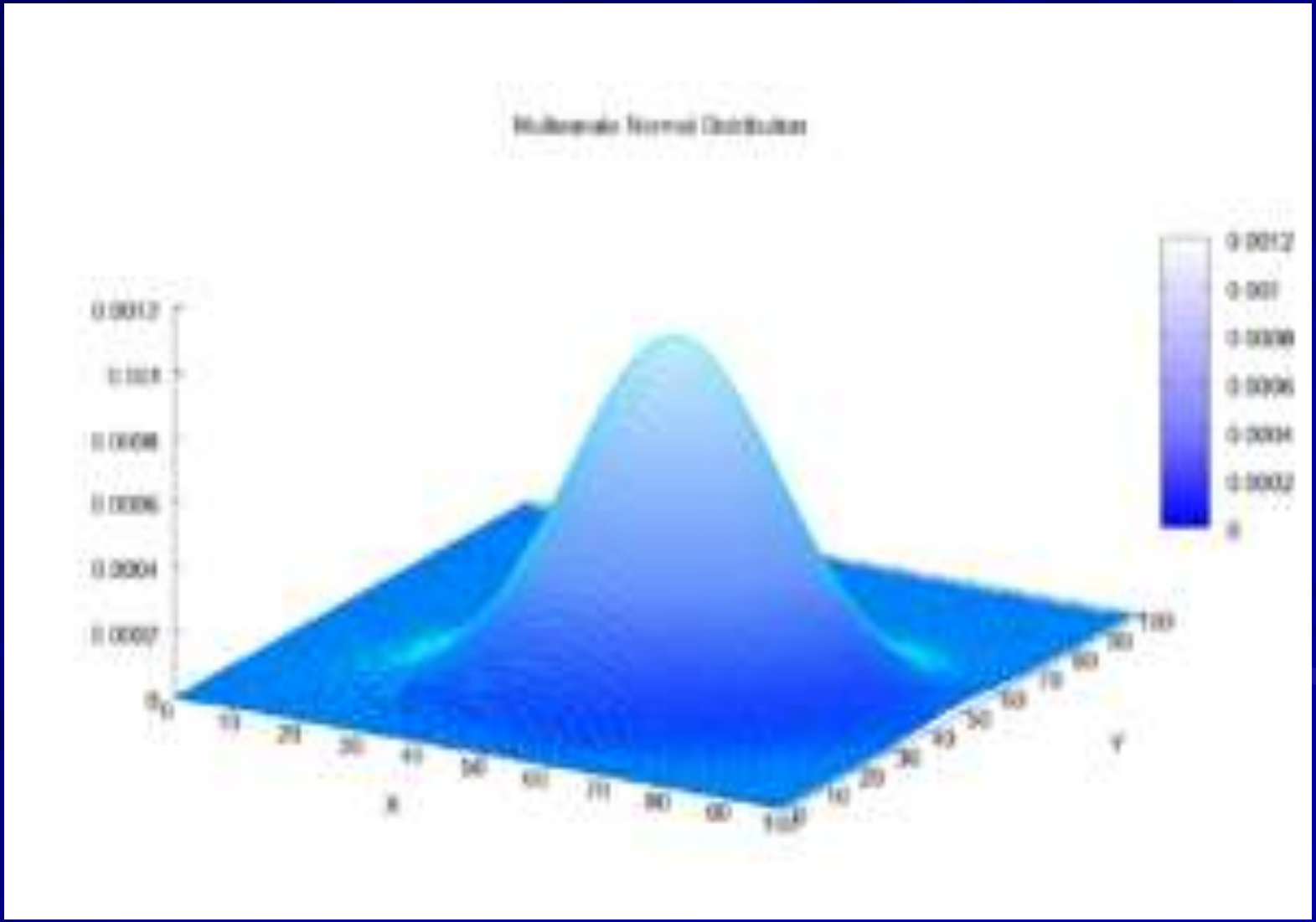
- Generálunk random adatokat ugyanannyi változóval, egymás után többször (100-szor)
- Minden random adatállományon HKA-t vagy KKA-t futtatunk (ugyanannyi klaszterrel) és kiszámítunk a struktúra jóságának a mérésére egy sor QC-t.
- A kapott QC-ket átlagoljuk ($QC_{rand}$) és összevetve a tesztelendő struktúra QC-értékeivel, képezzük a MORI indexeket.

# Kontrollként használt random változótípusok

1. Az aktuálissal megegyező eloszlású, de független változók (adatok random permutálásával változónként)
2. Független random normális eloszlású változók
3. **Korreláló random normális változók az input változók páronkénti korrelációi alapján (változók FA-ja, súlymátrix!)**

(Vargha–Borbély, 2017; Vargha–Bergman–Kövi, in preparation)

# Kétdimenziós normális eloszlás



# Normalitás és klaszteranalízis

- Ha az input változók többdimenziós normális eloszlásúak, csak egyetlen csúcsa van az eloszlásnak. Ilyenkor nincs értelme több típust keresni.
- Ha a normalitás sérül, megjelenhetnek a nemlineáris kapcsolatok.
- Emiatt az input változók interkorrelációit reprodukáló, random normális korreláló kontroll az egyik leginformatívabb validáló eszköz.

# Klaszterstruktúrák összehasonlítása k = 5 és k = 7 között hat QC-vel, korreláló random normális kontrollal

MORI-indexek (rep = 100)

| | EESS% | PB | XBmod | SilCoef | HCátlag | CLdelta |
|---|---|---|---|---|---|---|
| k = 5 | 0,21 | 0,18 | -0,03 | 0,21 | 0,21 | 0,28 |
| k = 6 | 0,25 | 0,22 | 0,21 | 0,29 | 0,25 | 0,30 |
| k = 7 | 0,26 | 0,20 | 0,07 | 0,24 | 0,25 | 0,30 |

A 6-klaszteres struktúra tűnik a legjobbnak

# A belső validitás QC és MORI mutatóinak használata

- Megítélhetjük velük egy struktúra jóságát
- Segítséget nyújthatnak a helyes klaszterszám megállapításához
- Összehasonlíthatunk velük különböző algoritmusokat
- Összehasonlíthatunk velük különböző klasztermegoldásokat (struktúrákat)

# Két klasztermegoldás összehasonlítása a ROPstatban

- Centroid
  - Klaszterközéppontok (centroidok) páronkénti távolságainak a kiszámítása
- Exacon:
  - Kódváltozók keresztgyakorisági táblázata
  - Hasonlósági mutatók (Cramér-féle V, Jaccard index, Rand, ARand)

# Külső validitás (external validity) vizsgálata

A kapott klasztereket megpróbáljuk olyan változókkal összefüggésbe hozni, amelyek nem szerepeltek a klaszteranalízis modelljében (pl. nem, kor, iskolázottság), de segítenek értelmezni és ezáltal érvényessé tenni a klasztereket

# KÖSZÖNÖM A FIGYELMET