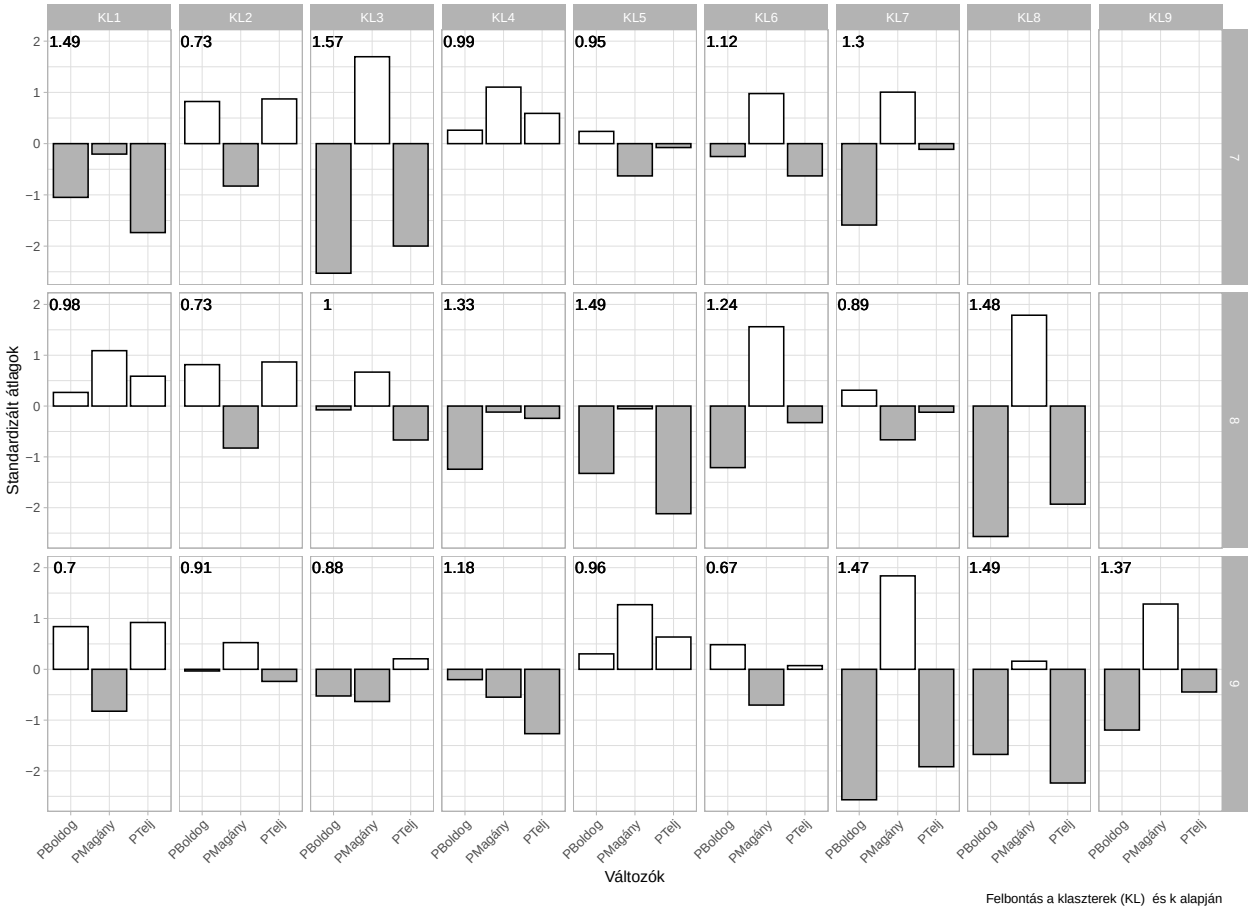
Házi feladatok megoldása 9.

k-középpontú klaszteranalízis R-ben

Smahajcsik-Szabó Tamás, M9IJYM

1. Végezz k-közép elemzést R-ben a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással k = 7 és 9 között!

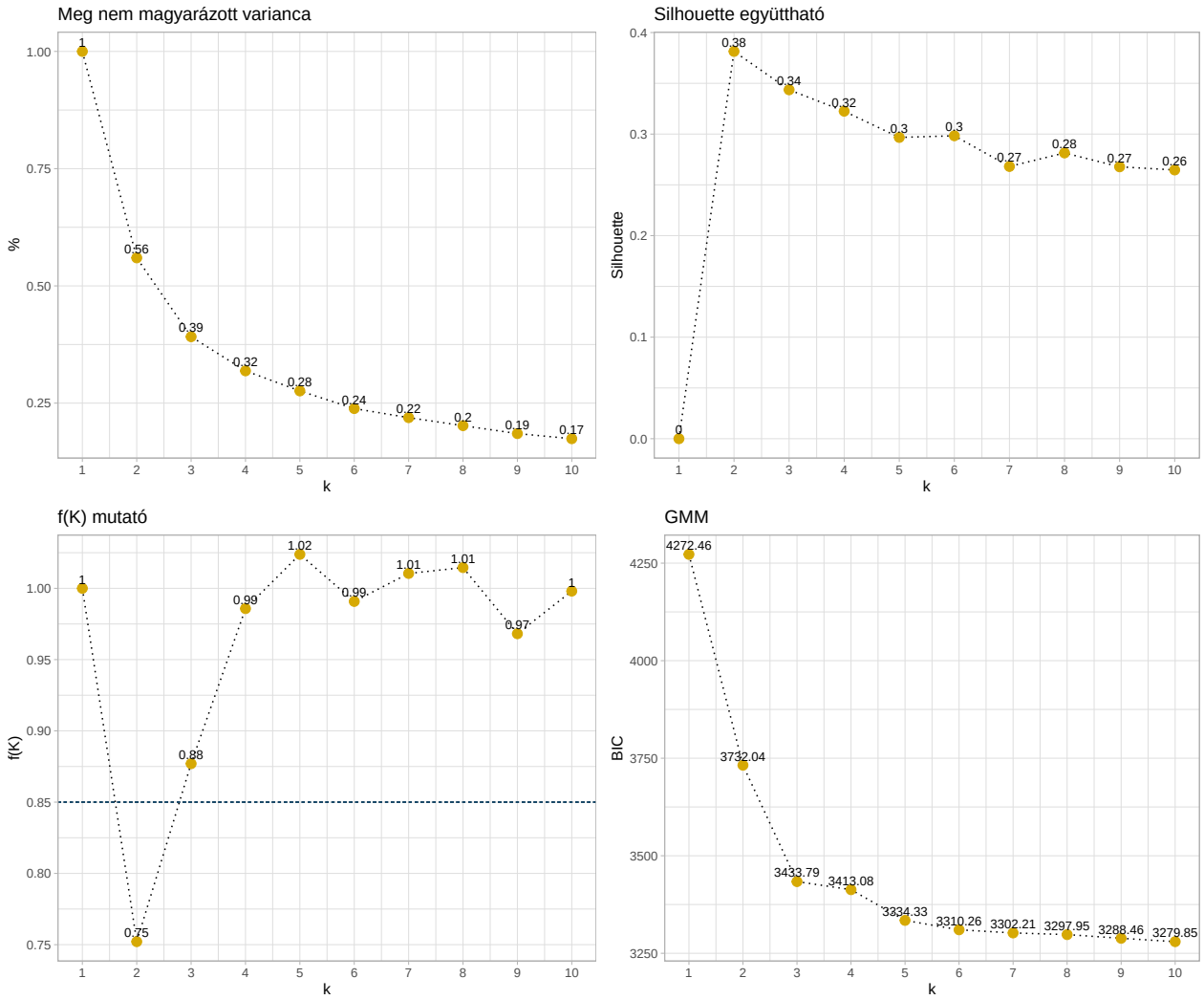
A k-közép elemzéseket 5 kezdeti centroid struktúrával, maximális 20 iterációval végeztem (MacQueen-féle algoritmussal). Az eredményekről az alábbi áttekintő ábra tájékoztat. A képződött klaszterek standard átlagait, és a homogenitási együtthatókat is feltüntettem.



1.ábra A klaszterstruktúra áttekintése

2. Hány klaszteres megoldás tűnik a legjobbnak az 1. feladat változói esetében a 6.4.1. alpontban leírt R-beli módszerek alapján (vö. 6.6-6.10. ábrák)?

A fenti ábra kedvezőtlenebb homogenitási indexeivel összhangban, a k=7 és k=9 közötti megoldás nem optimális a segítő ábrák alapján sem.



2. ábra: Segítő ábrák

A meg nem magyarázott varianca könyökábrája (bal felül), illetve a Silhouette együttható (jobb felül) egy k=2 megoldás fölényét erősíti a k = 7, 8 illetve 9 megoldásokkal szemben. Az *f(K)* mutató (bal alul) a k=2 megoldást emeli ki, k=2-nél ereszkedik f(k) értéke a 0.85 küszöb alá. A k=7 és k=9 megoldások közül a k=9 esetén kedvezőbb kissé a mutató, de mindegyikre nézve suboptimális a jelzés. GMM függvénnyel tesztelve az adatokat a BIC értéke egyaránt alacsony k=7 és k=9 között, de egyrészt nem optimális a BIC ezen struktúrák mellett, másrészt minden más segítő ábra a k=7, 8 vagy 9 megoldások nem kielégítő voltát erősíti.

3. Mentsd el az 1. feladat klaszterváltozóit k = 7 és 9 között, tedd át ROPstatba és számítsd ki a Validálás modullal a főbb QC mutatókat! Melyik klaszterszám megoldása tűnik a legjobbnak?

A *Validálás* modullal végzett számítások eredményét az alábbi táblázat összegzi.

EESS%	Pontbisz	XBmod	Sil.eh.	HCatlag	CLdelta	GDI24	HCmin-HCmax	k
78.39	0.355	0.494	0.661	0.440	0.893	0.384	0.21-1.11	7
79.75	0.347	0.518	0.659	0.413	0.894	0.412	0.21-1.02	8
81.34	0.313	0.353	0.639	0.382	0.875	0.283	0.18-1.01	9

1.táblázat Klaszterstruktúra validitás mérése ROPStattal

A magyarázott variancia (EESS%) növekszik **k** értékének emelkedésével, de nem láthatunk kiugró javulást. A Pontbiszeriális együttható értéke csökken, a klaszterszám emelkedésével. Optimális szintjént k=7-nél éri el. k=9 esetén nemcsak a PB-index, de a módosított Xien Beni mutató is kedvezőtlen, 0.353 értéket vesz fel, legjobb k=8 esetén. A Silhouette index mindegyik megoldás esetén kedvező, k=7-nél tetőzik, majd folyamatosan csökken. A CLdelta k=8 esetén a legjobb a három struktúra közül, elfogadható szinten. A HCÁtlag természetszerűleg csökken **k** emelkedésével A GDI24 index mindegyik megoldás esetén kedvezőtlen struktúrát jelez.

Mindezek alapján a k=7 vagy k=8 struktúra tűnik megfelelőnek.

4. Végezz k-medoid elemzést R-ben a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel, standardizálással k = 7 és 9 között!

5. Mentsd el a 4. feladat klaszterváltozóit k = 7 és 9 között, tedd át ROPstatba és számítsd ki a Validálás modullal a főbb QC mutatókat! Melyik klaszterszám megoldása tűnik a legjobbnak?