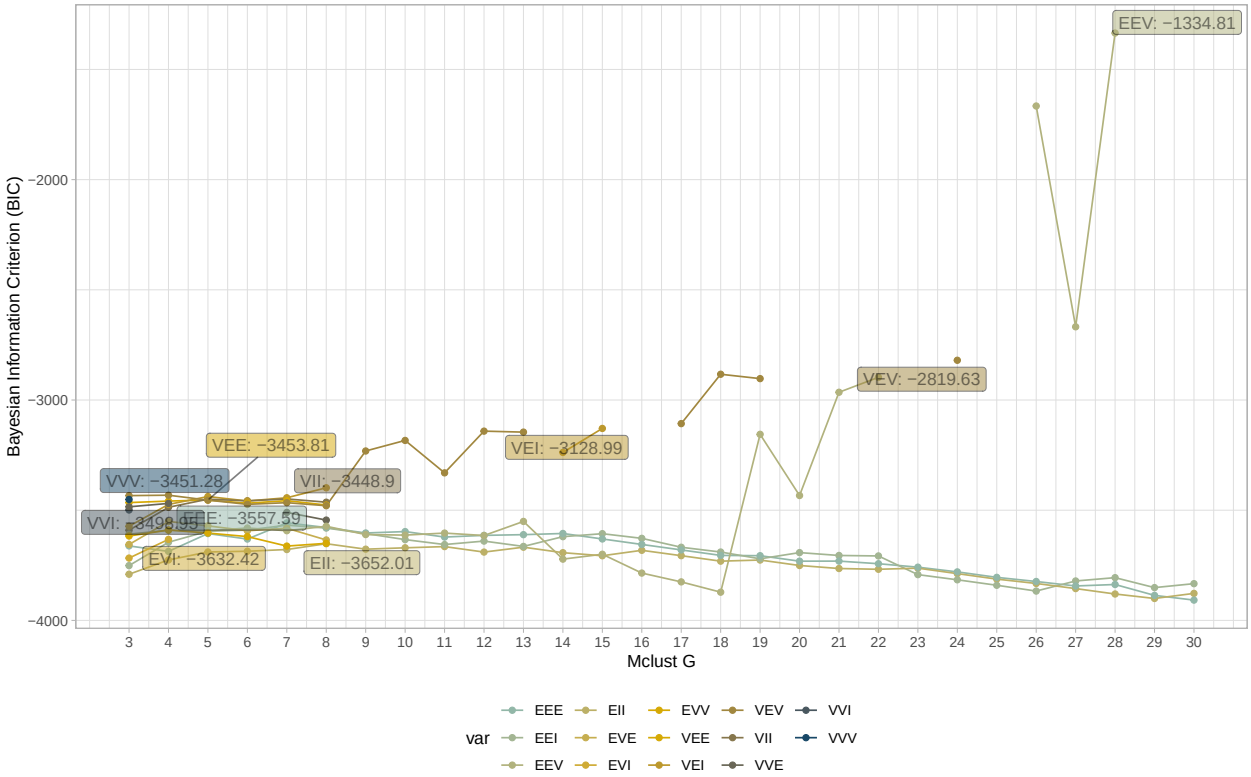
Házi feladatok megoldása 10.

Modell-alapú klaszteranalízis (MKA) R-ben

Smahajcsik-Szabó Tamás, M9IJYM

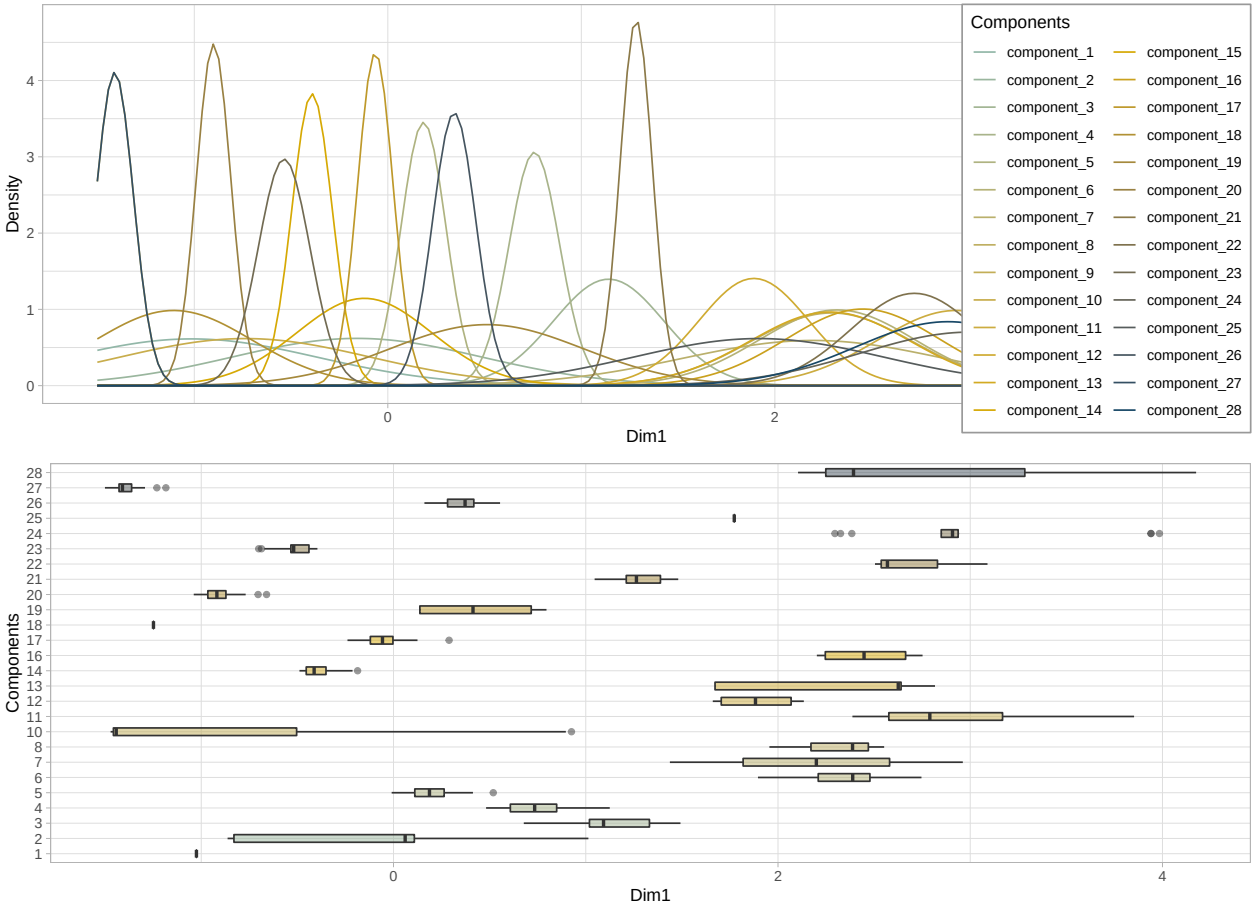
1. Végezz MKA-t a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel! Melyik megoldás tűnik a legjobbnak a BIC-grafikon alapján?

Az *mclust* programcsomag *Mclust()* függvényét használva, különböző **G** érték-konfigurációkat teszteltem a BIC legnagyobb értékét, a képződő struktúra értelmezhetőségét is szem előtt tartva. Első körben, 3 és 30 között vizsgálódva elmondható, hogy alapvetően három olyan keverékeloszlás típus mutatkozott meg, melynél a Bayes-féle Információs Kritérium a legjobb értéket érte el. Ebben a nagy tartományban a G=28 esetben az EEV típus mutatta a lejobb BIC értéket, utána visszafelé haladva, a VEV áll G=24-nél, majd pedig a VEI típus illeszkedésénél legjobb BIC G=15-nél.



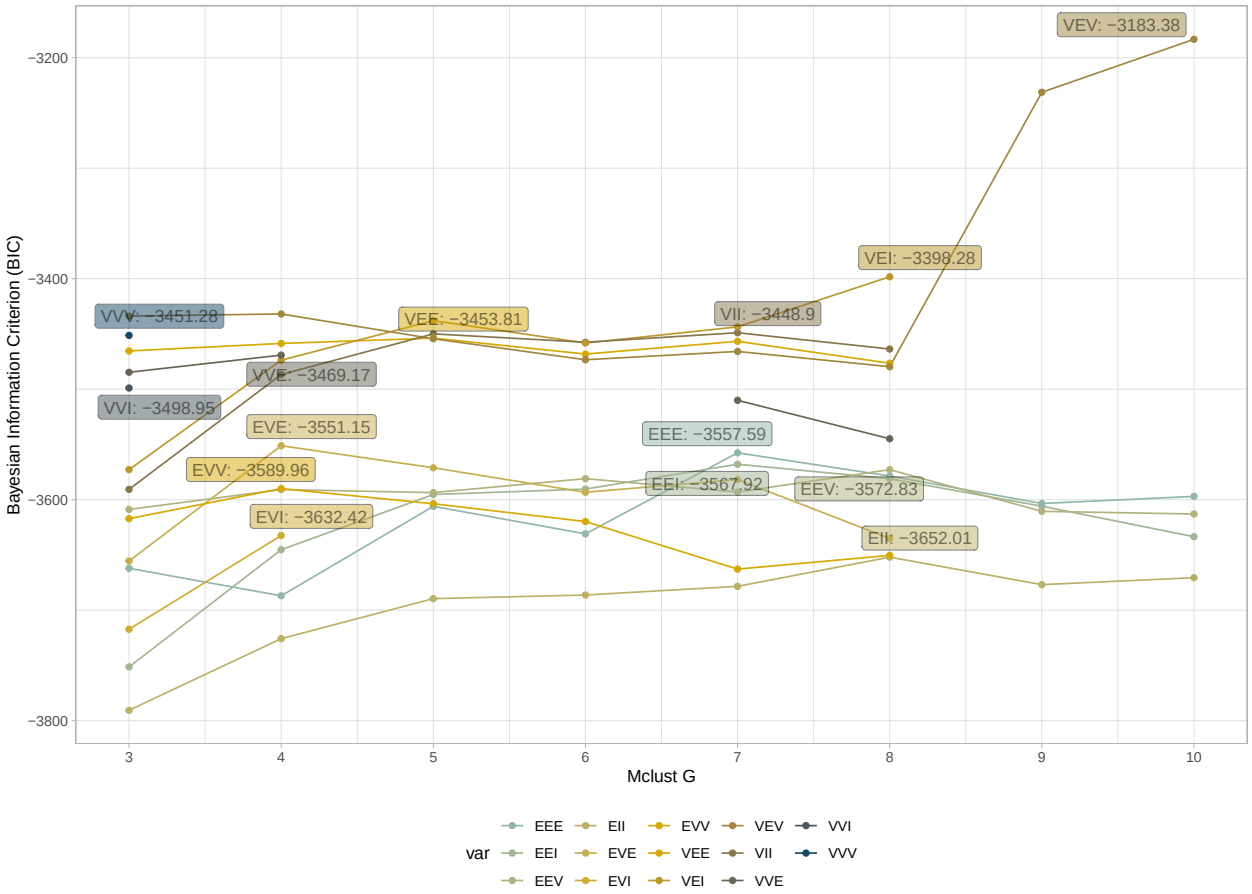
1. ábra MKA BIC eredmények G=3 és G=30 között

Az alábbi ábrán az egyes keverék-komponensek sűrűségét, illetve box-whiskers eloszlásait látjuk.



2. ábra MKA sűrűség és eloszlási eredmények G=28 között

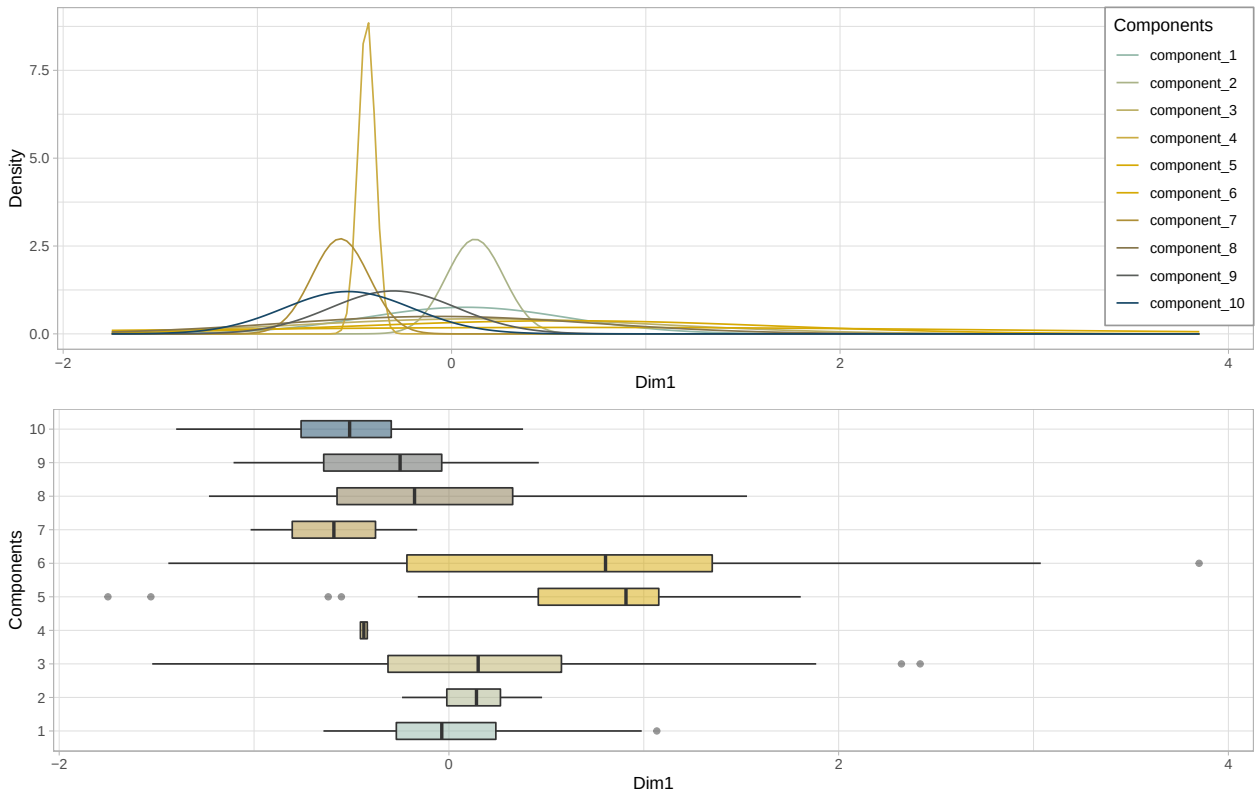
Mindezeken túl azonban, az értelmezhetőség végett az elemzést a G=3 és G=10 tartományra is megismételtem.



3. ábra MKA BIC eredmények G=3 és G=10 között

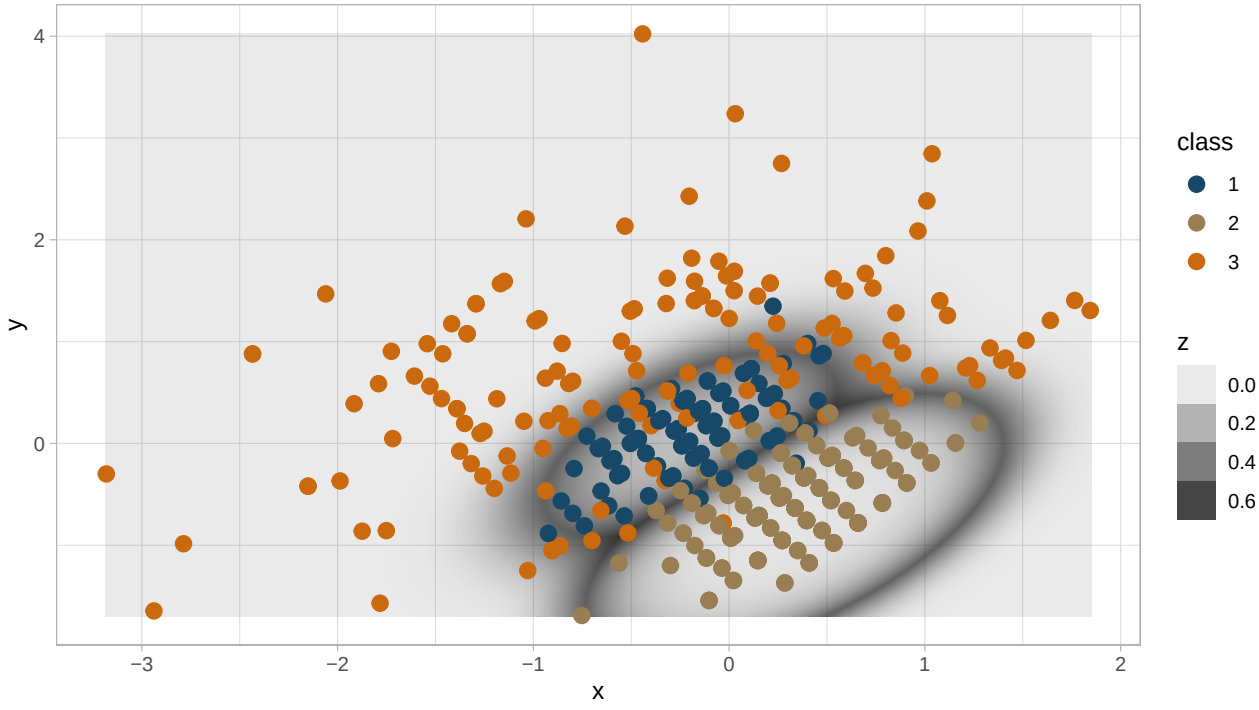
Ezen elemzés nyomán azt látjuk, a VEV eloszlástípus bizonyul a legjobbnak, mely a fenti, kiterjesztett elemzésnél is a legjobbak között volt, így az képződött struktúra értelmezhetőségének reményében a G=10, VEV struktúrát fogadom el.

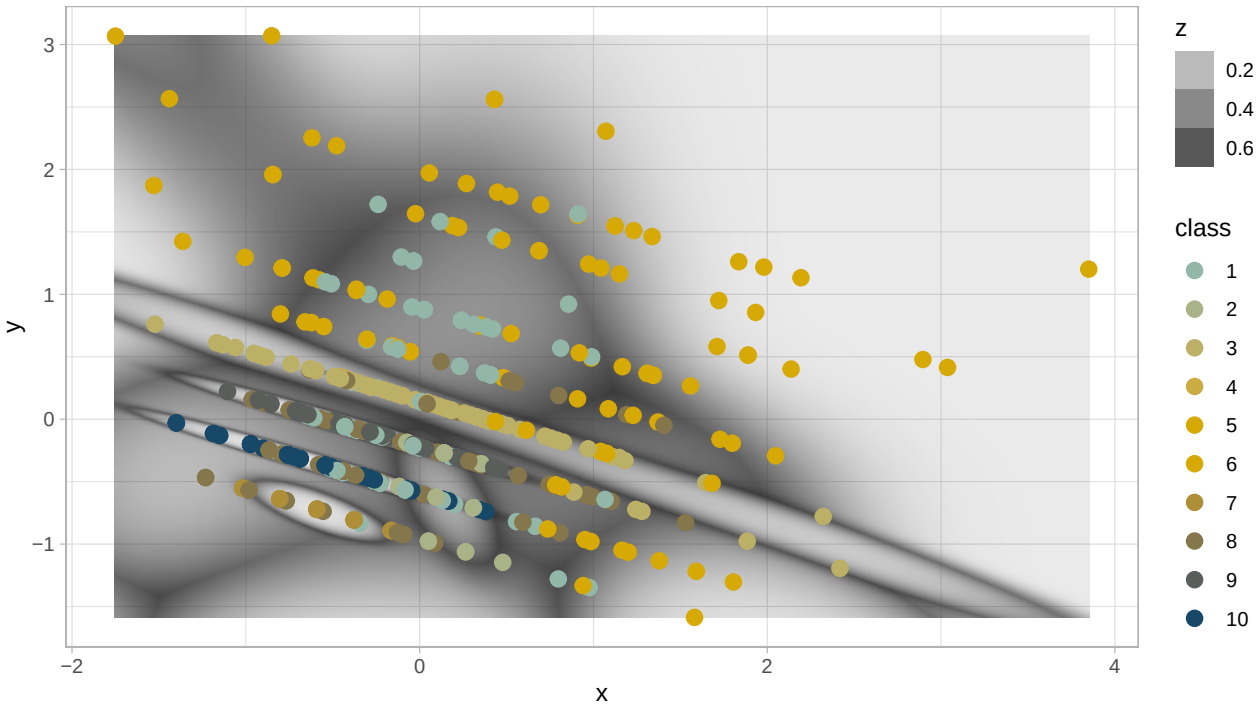
Ha az keverékkomponensek sűrűségét tekintjük (alábbi, 4. ábra)



4. ábra MKA sűrűség és eloszlási eredmények G=10 között

A keverékkomponensek sűrűsödés vizsgálata három jobban elkülönölő struktúrát jelez, így mindez felveti hipotézisként, vajon mennyire értelmezhetőbb egy G=3 struktúra az információveszteség ellenére a G=10-zel szemben.



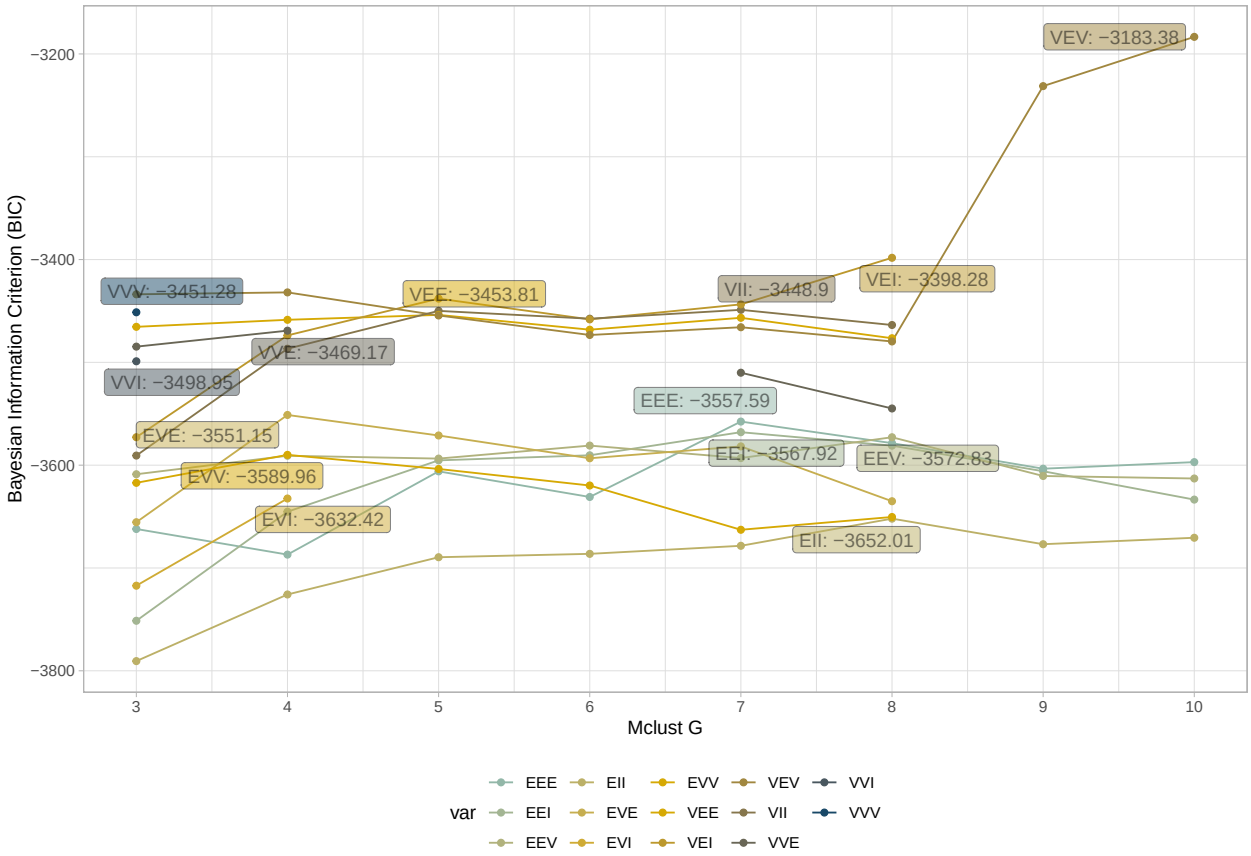


5.ábra A G=3 (felső ábra) és a G=10 (alsó ábra) struktúrák összevetése az adatok első két főkomponense mentén képzett két dimenziós síkban; a klaszterhatárok bizonytalanságát a háttér árnyalata (*z-paraméter*) jelzi.

Noha értelmezhetőbb struktúrát kapunk G=3 értékkel, ez jelentősen rosszabb BIC struktúra mint a G=10.

G = 10 esetén jelentős átfedés is megfigyelhető a eloszlás-komponensek között.

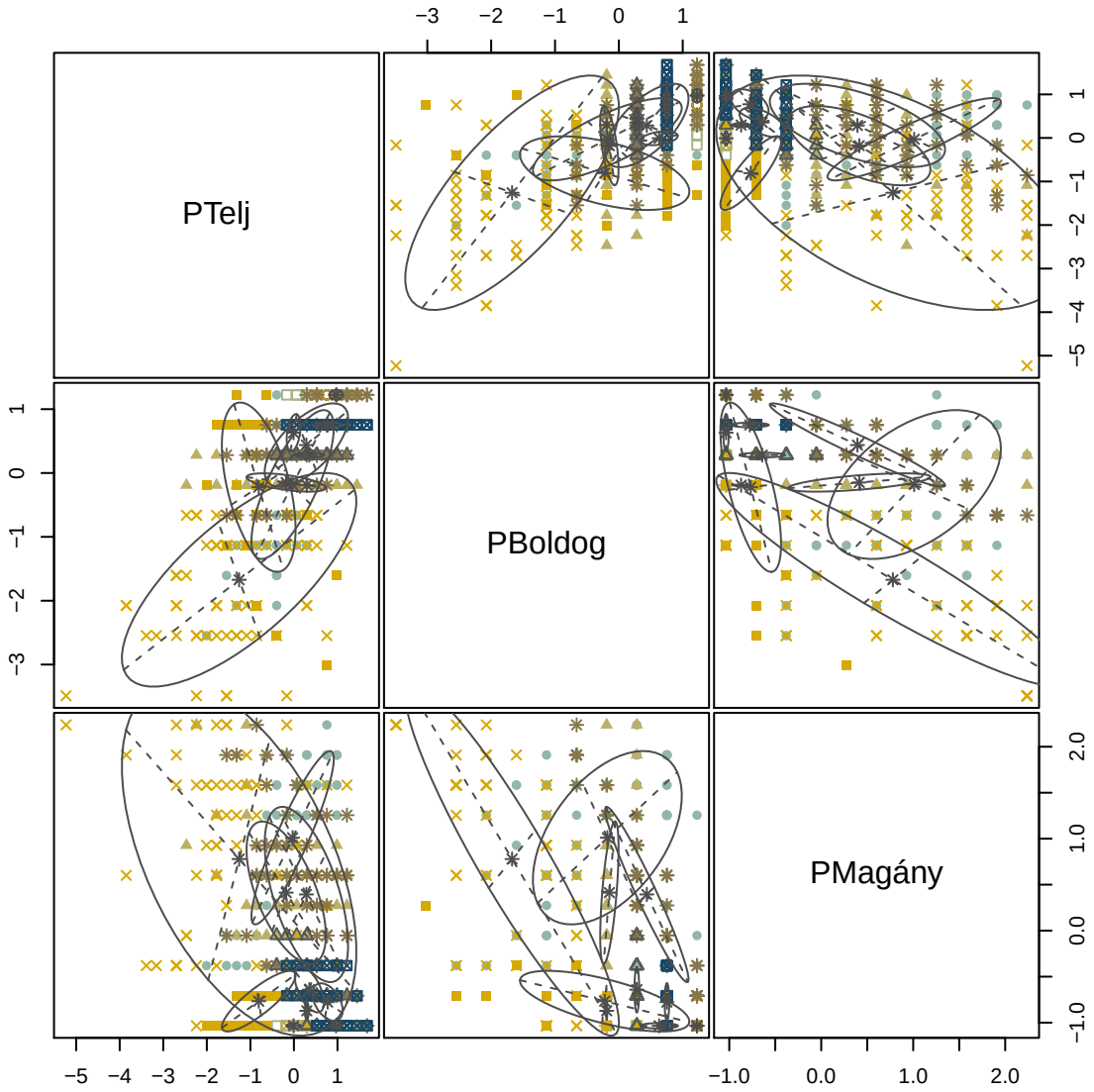
2. Készítsd el az 1. feladat BIC-grafikonját k = 3 és 10 között!



6. ábra MKA sűrűség és eloszlási eredmények G=10 között

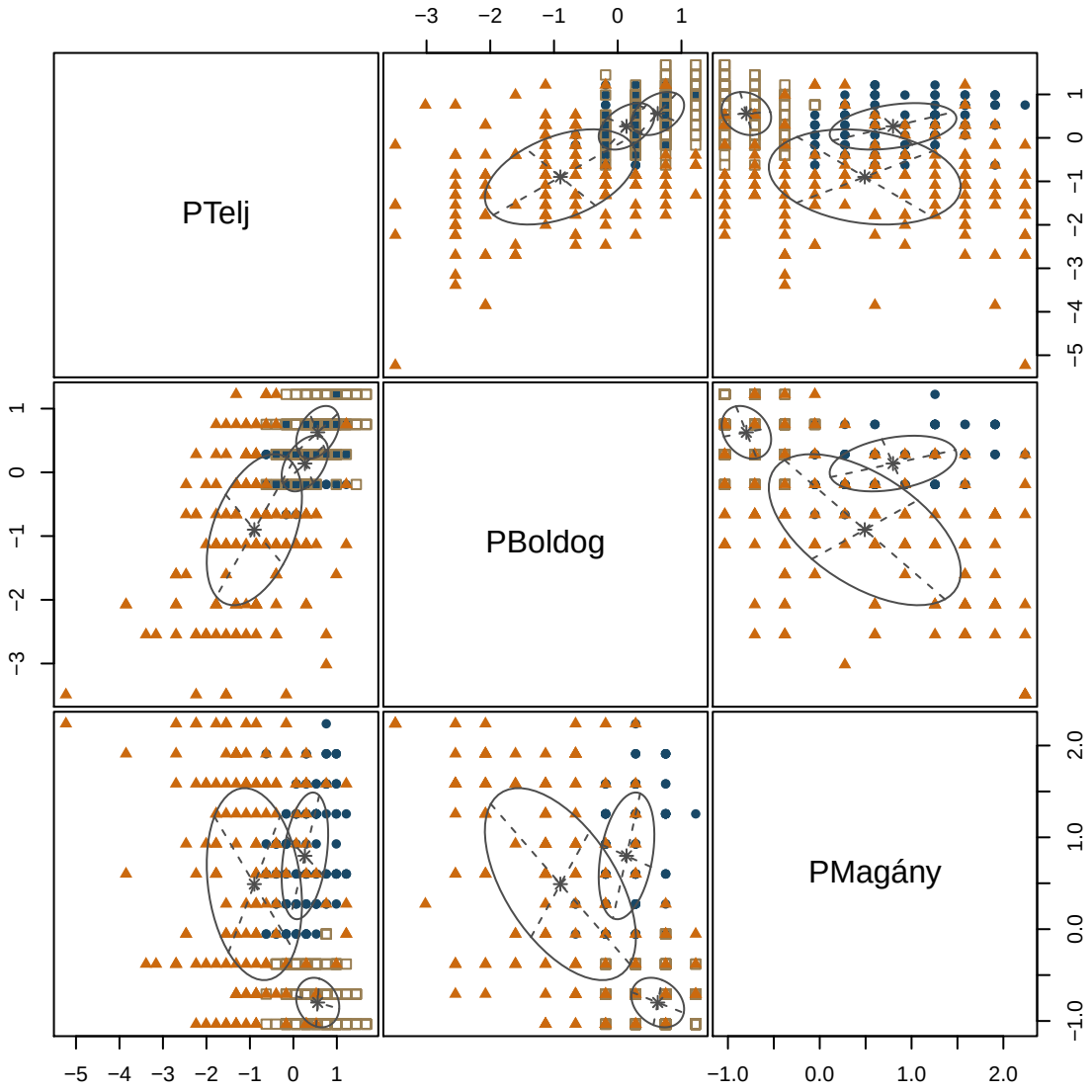
BIC ábráimon a szövegdobozok az adott keveréktípus maximuális értékénél állnak, azaz G azon értékénél, melynél az adott típus BIC-értéke a legnagyobb szintet éri el az adott modellezési folyamat során. Ennek értelmében G=10-nél a legjobbnak tűnő eloszlástípus a VEV, melynek BIC értéke -3183.38.

3. Készítsd el az 1. feladat legjobb BIC megoldásának classification ábráját!



7. ábra A G=10 megoldás klasszifikációs ábrája

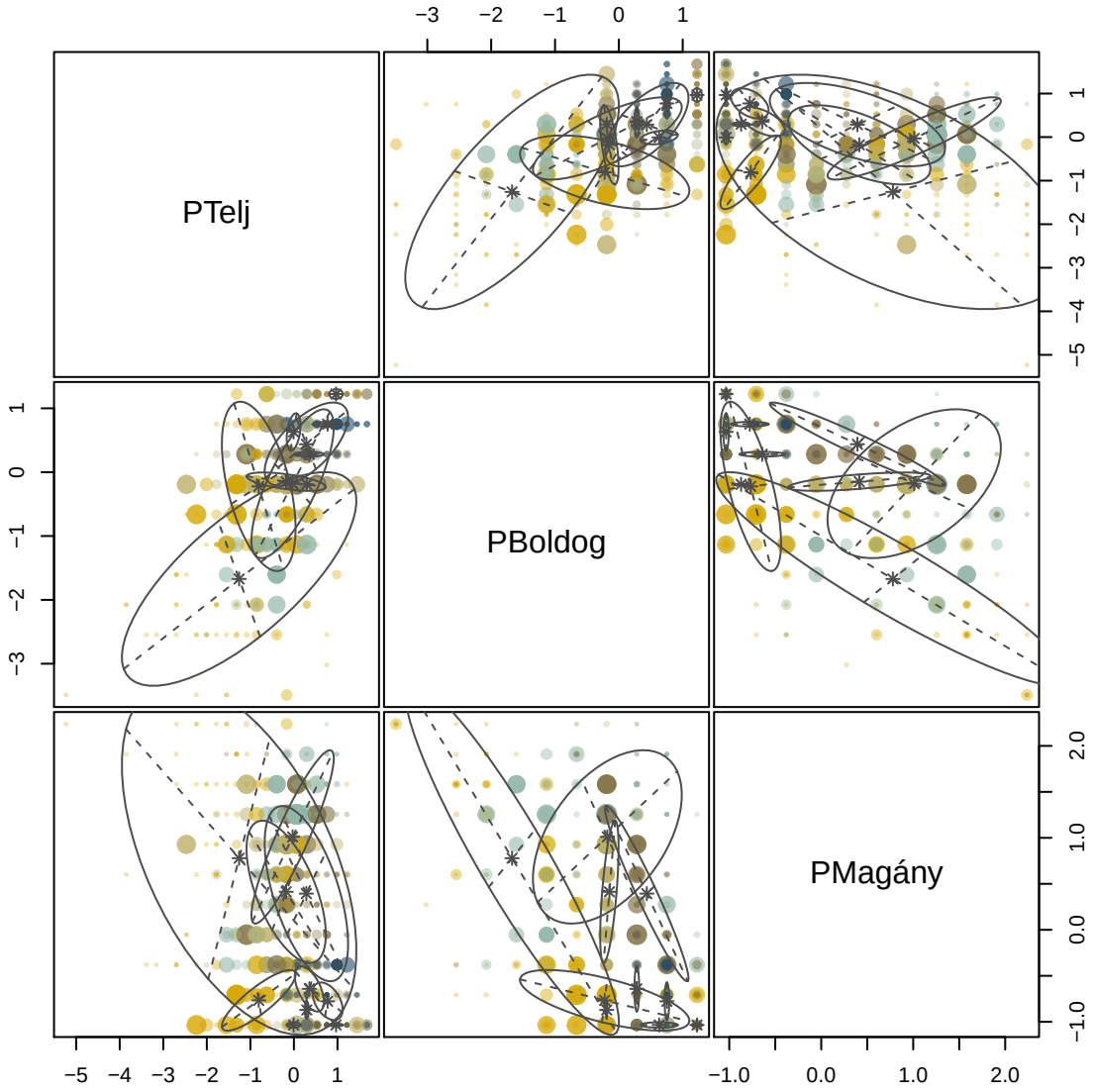
Jól látható három nagyobb klaszter elkülönülése (például a Boldogság és a Magányosság szeletében), és több kisebb, részben átfedő struktúra is. Ez ismét felveti a kérdést, mennyiben értelmezhetőbb egy G=3 struktúra. Különösen a Boldogság és Magányosság szeleteiben láthatóak pusztán csak néhány esetet magukban foglaló apró klaszterek.

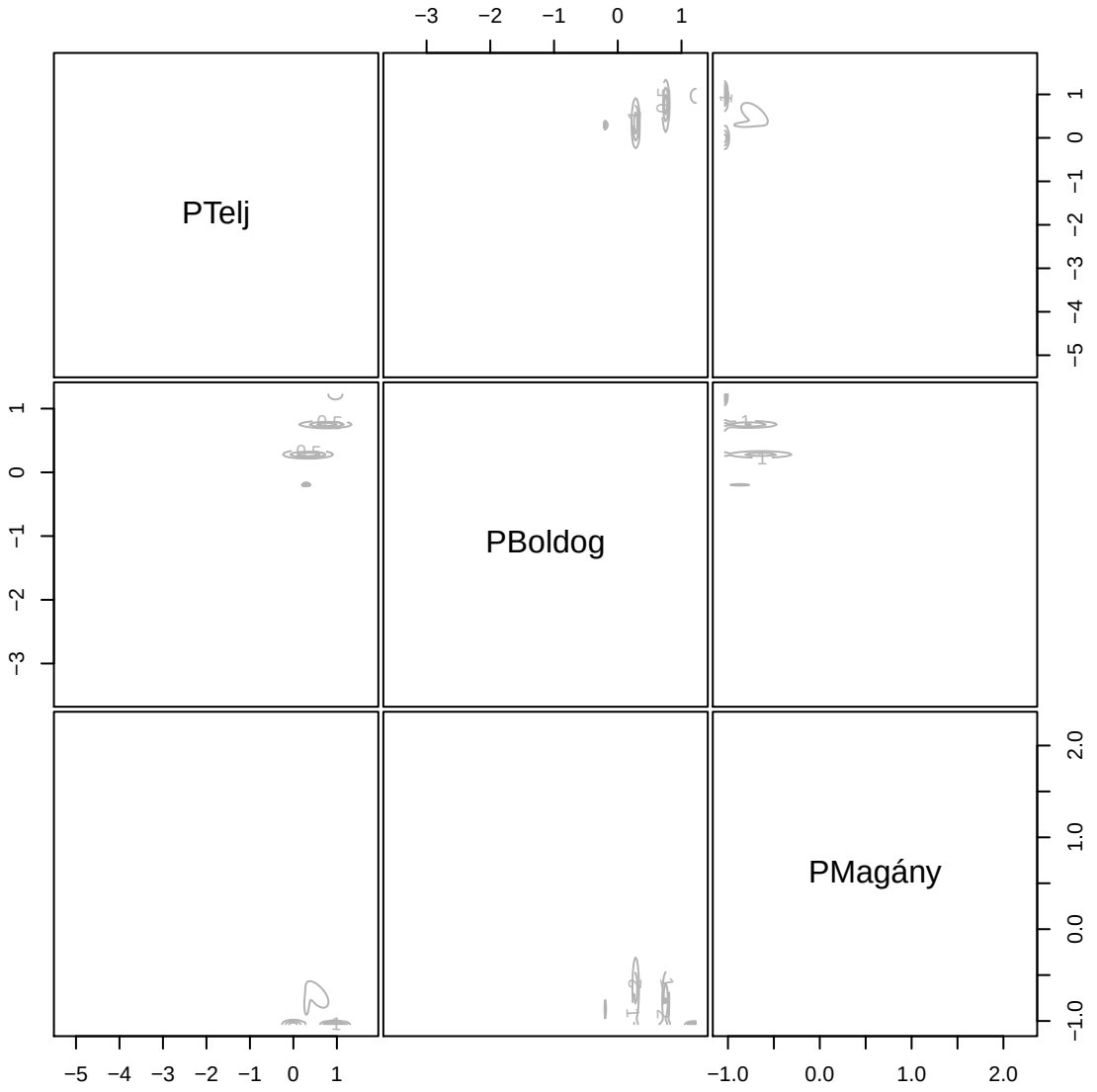


8. ábra A G=3 megoldás klasszifikációs ábrája

A BIC-értékben való csökkenés ellenére egy jobban értelmezhető, kevésbé átfedő, kevésbé redundáns megoldást kapunk. Ugyanakkor számos eset adott, melynél nem világos a besorolás.

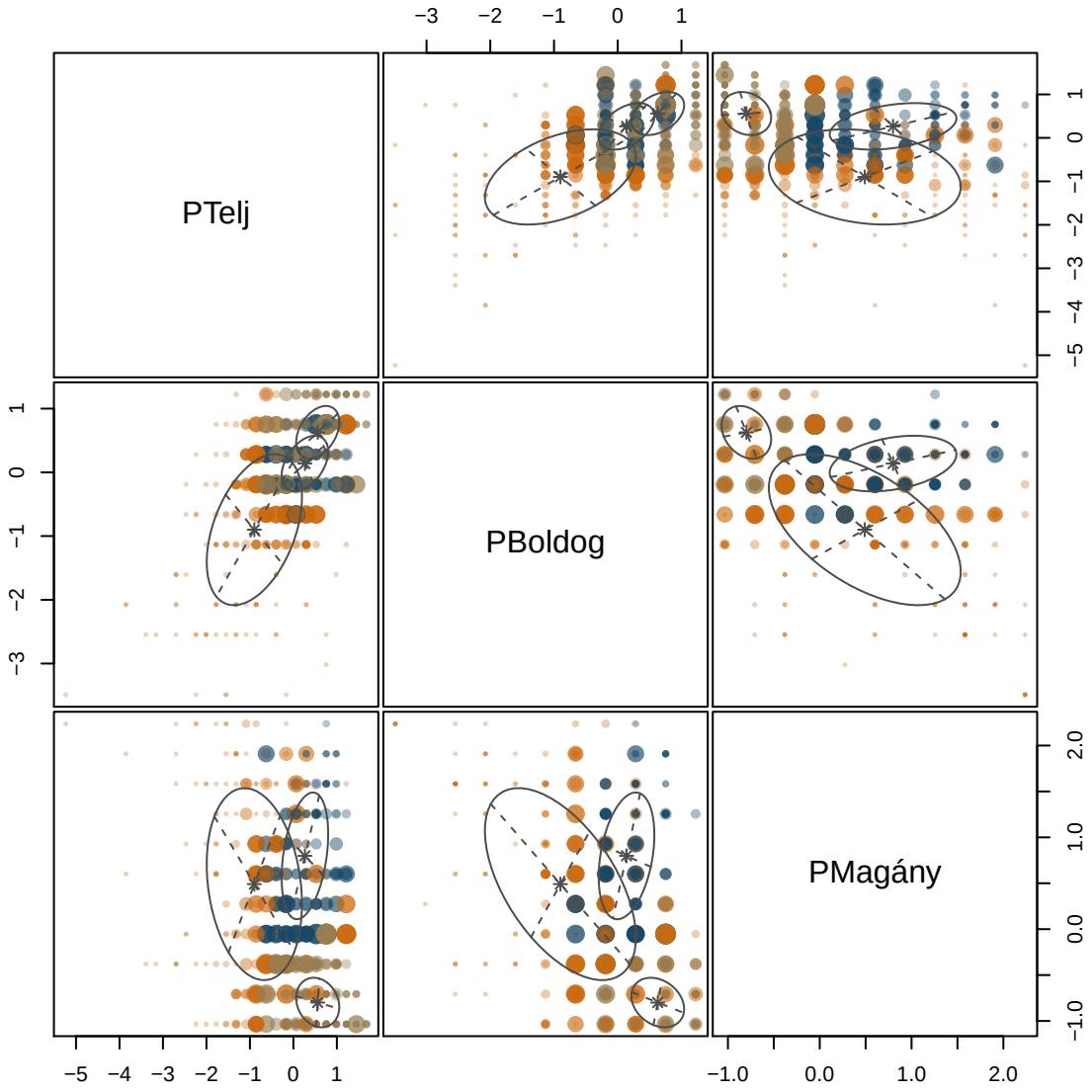
4. Készítsd el az 1. feladat legjobb BIC megoldásának uncertainty és density ábráját!

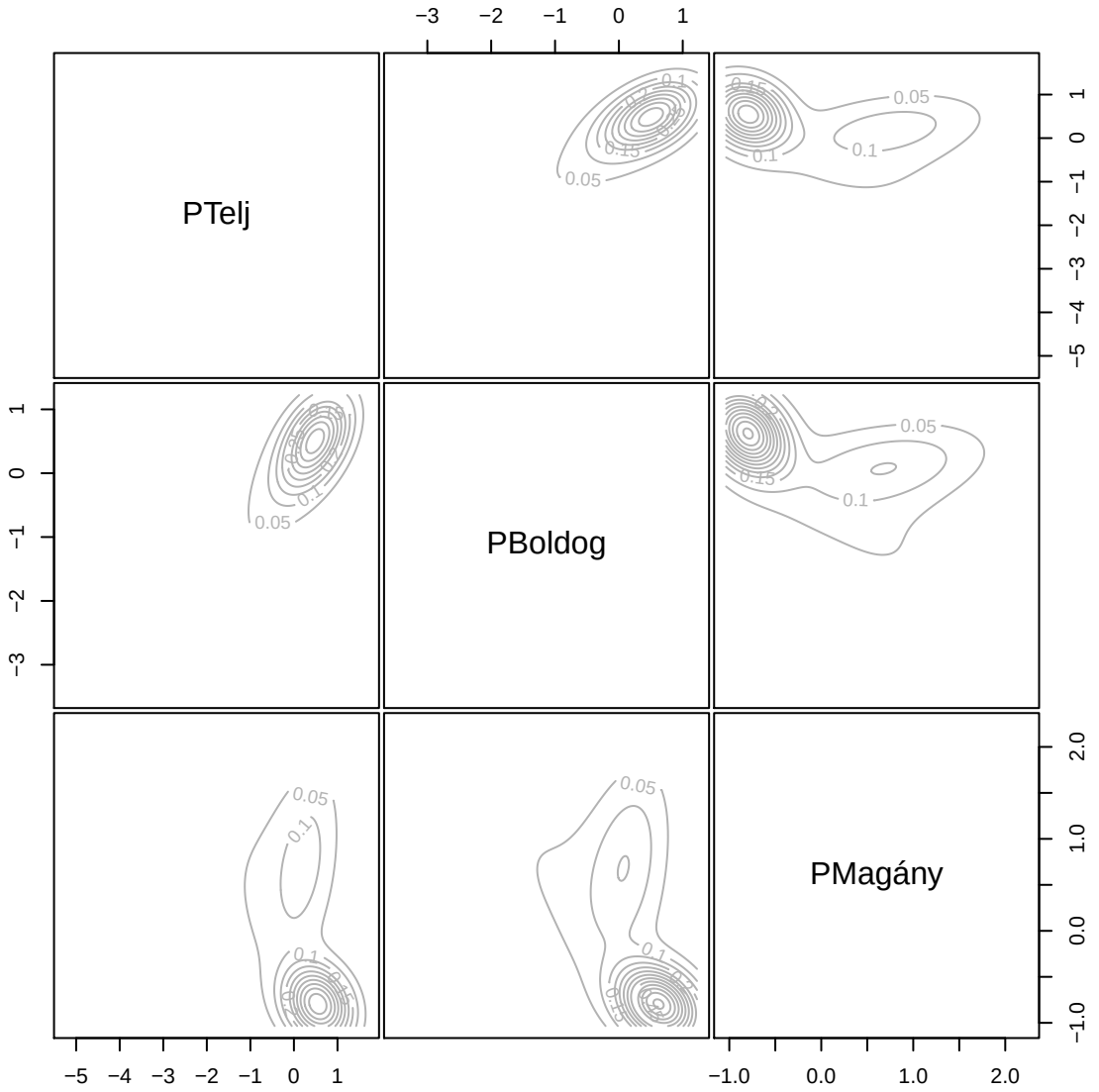




9. ábra A G=10 megoldás “uncertainty” és sűrűsödés ábrája

G=10 esetében a klaszterek átfednek, több esetben is a klaszterbe tartozás bizonytalansága emelkedett. Különösen a Magányosság és a Teljesítmény változók szeletiben látunk nehezen értelmezhető, átfedő struktúrákat.

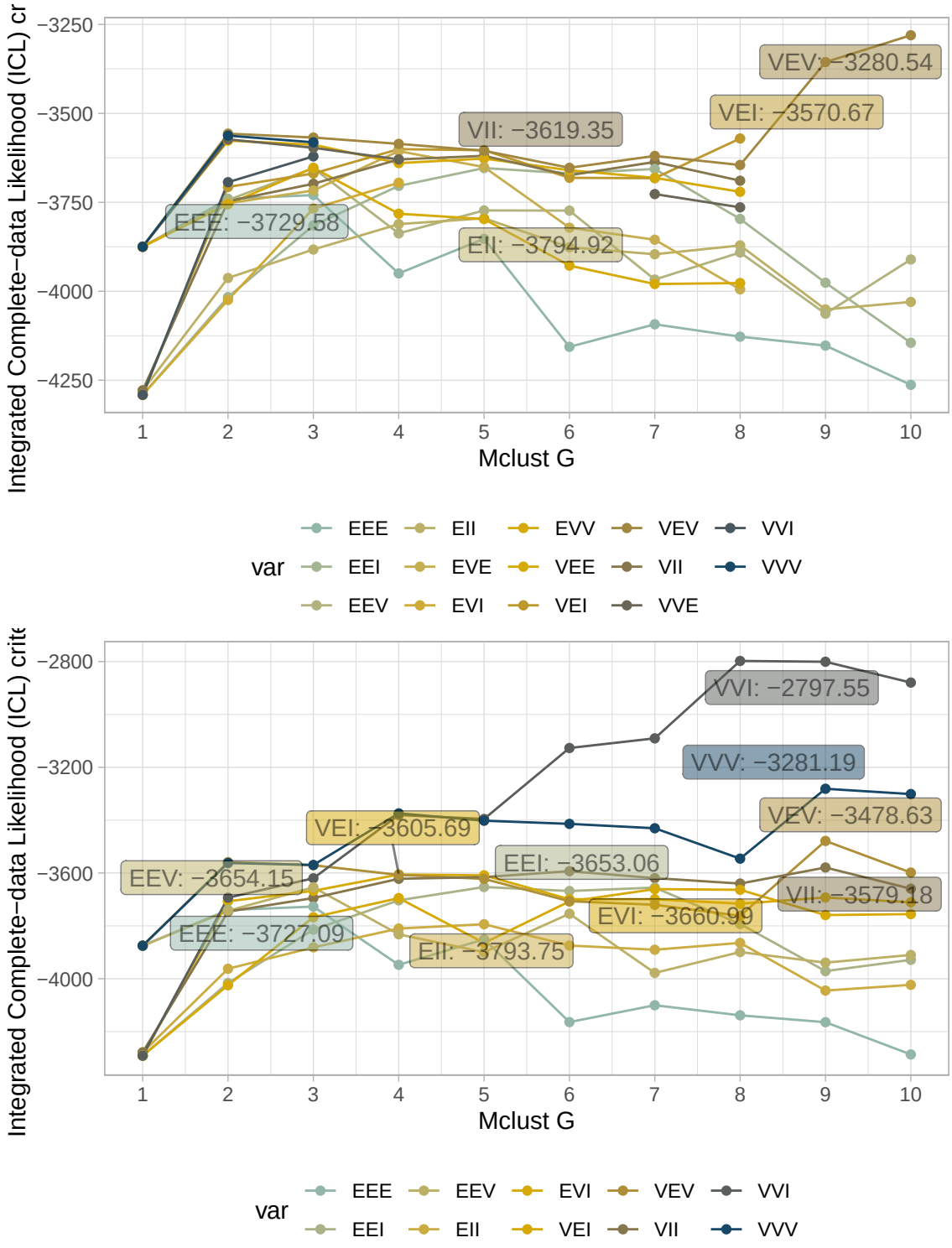

PTelj
PBoldog
PMagány
−3 −2 −1 0 1
1 0 −1 −2 −3 −4 −5
1 0 −1 −2 −3
2.0 1.0 0.0 −1.0
−5 −4 −3 −2 −1 0 1
−1.0 0.0 1.0 2.0



10. ábra A G=3 megoldás “uncertainty” és sűrűsödés ábrája

Több esetben emelkedett besorolási bizonytalanság mutatkozik meg, de a struktúra értelmezhetőbb, világosabban elkülönülpő eloszláskomponenseket látok.

5. Készítsd el az 1. feladat ICL-grafikonját k = 1 és 9 között! Ugyanaz a modell tűnik a legjobbnak, mint a BIC-grafikon alapján?



11. ábra A G=1 és G=10 közti megoldások ICL ábrái (a felső a prior opció nélkül, az alsó pedig ezzel kiegészített lefutás)

A *priorControl()* hangolási opció nélkül, a felső ábra szerint hasonlóképpen a VEV keveréktípus a legjobb az ICL információs kritérium szeriunt is, akár G=9, akár G=10 opciót tekintjük is.

A finomhangolással együtt azonban (alsó ábra) a VVI eloszlástípus mutat kedvezőbb illeszkedést (G=8 beállítás mellett). Ezt úgy értelmezem, hogy méretükben, tengelyeik hosszában eltérő, de a főtengelyekkel és egymás tengelyirányultságában egyező eloszlásokat modellez a legjobb ICL-lel leírható modell.

A VVI típus G=8 struktúra esetén tetőzik, így ezt fogadom el a legjobb megoldásnak, mely az alacsonyabb komponensszám miatt az értelmezhetőségben is kedvezőbb.

6. Mentsd el a legjobb BIC-megoldást, tedd át ROPstatba és számítsd ki a Validálás modullal a főbb QC mutatókat! Hasonlítsd össze a kapott értékeket a 8. óra 1. feladatában kapott QC-értékekkel!

Az alábbi táblázatban mutatom be a validálás eredményét.

A G=3, 9, 10 és 28 MKA konfigurációkat vetettem össze a korábbi k-középpontú elemzésekével, ahol k értéke 6, 7, 8 és 9 között mozgott.

Jól láthatóan a G=3 megoldás, noha magyarázott varianciája 51.52%-os, mind a PB, mind ott Xie-Beni, mind a Solhouette mutatók, tekintetében egy a KKA-s elemzésekhez hasonlatos eredménnyel szolgált, noha átlagos homogenitása kedvezőtlenebb, GDI24 és CL mutatói is jobbnak mondhatóak mint a többi MKA megoldásé. Emögött azt gondolom, hogy a jól elhatárolható három eloszláskomponens áll, alacsony átfedéssel.

A G = 9, 10 MKA megoldások redundáns eloszláskomponens-szerkezete, a sok klaszterátfedés folytán egy a magyarázott varianciában alig jobb (54.99%-58.49%), és nagyon rossz szeparációs mutatókkal jellemezhető struktúrák. Hasonlóképpen, noha BIC-szempontból az exploratív elemzésnél kiemelkedő volt a G=28 megoldás, EESS%-a alig éri el a közepes KKA elemzésést, miközben minden egyéb QC-jában alul teljesít, talán CLdeltája kivétel egyedül.

EESS%	Pontbisz	XBmod	Sil.eh.	HCátlag	CLdelta	GDI24	HCmin-HCmax	G/K	Type
74.70	0.369	0.491	0.614	0.513	0.888	0.345	0.24-1.44	6	KKA
76.83	0.370	0.534	0.610	0.471	0.895	0.484	0.24-1.03	7	KKA
78.72	0.355	0.456	0.598	0.434	0.893	0.381	0.24-1.03	8	KKA
80.26	0.351	0.496	0.590	0.404	0.898	0.381	0.24-1.03	9	KKA
81.48	0.342	0.473	0.592	0.379	0.896	0.342	0.18-1.03	10	KKA
51.52	0.362	0.513	0.631	0.974	0.754	0.447	0.33-2.22	3	MKA
54.99	0.189	-0.795	0.181	0.915	0.555	0.039	0.09-2.37	9	MKA
58.49	0.176	-0.776	0.216	0.842	0.544	0.041	0.02-2.25	10	MKA
79.50	0.233	-0.877	0.136	0.438	0.789	0.011	0.00-2.21	28	MKA

1. táblázat Adekvációs mutatók összevetése