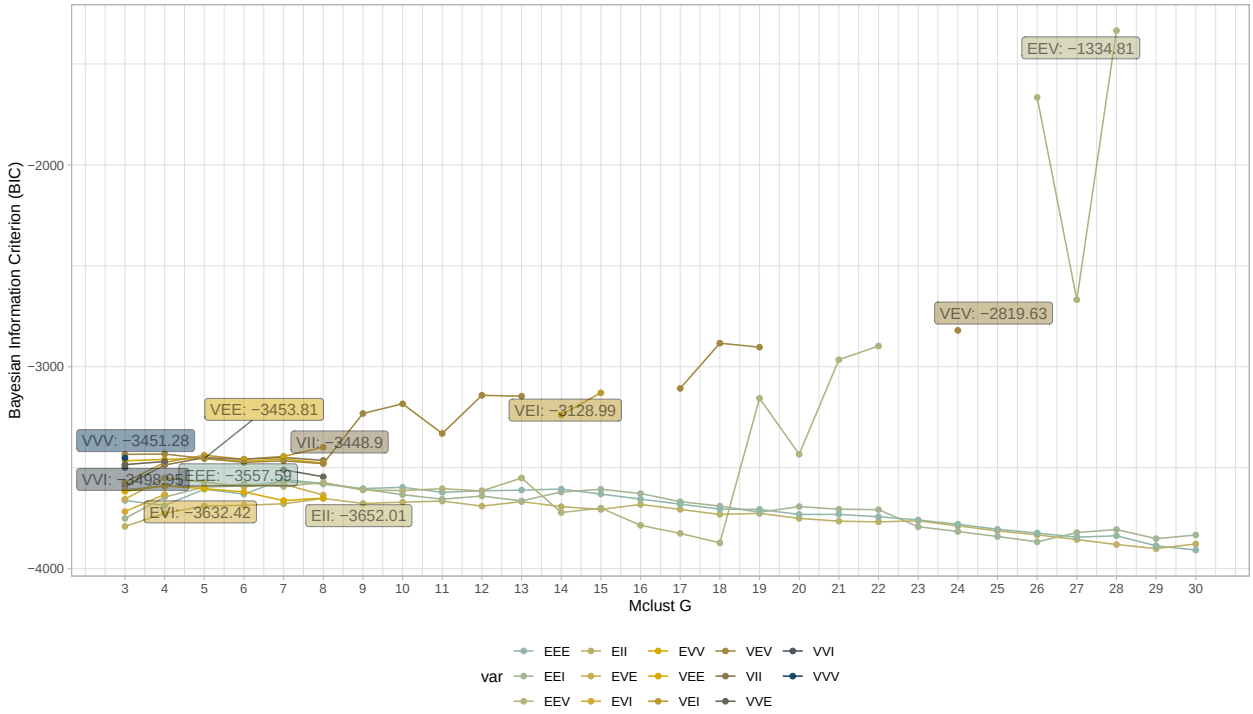
# Házi feladatok megoldása 10.

Modell-alapú klaszteranalízis (MKA) R-ben

Smahajcsik-Szabó Tamás, M9IJYM

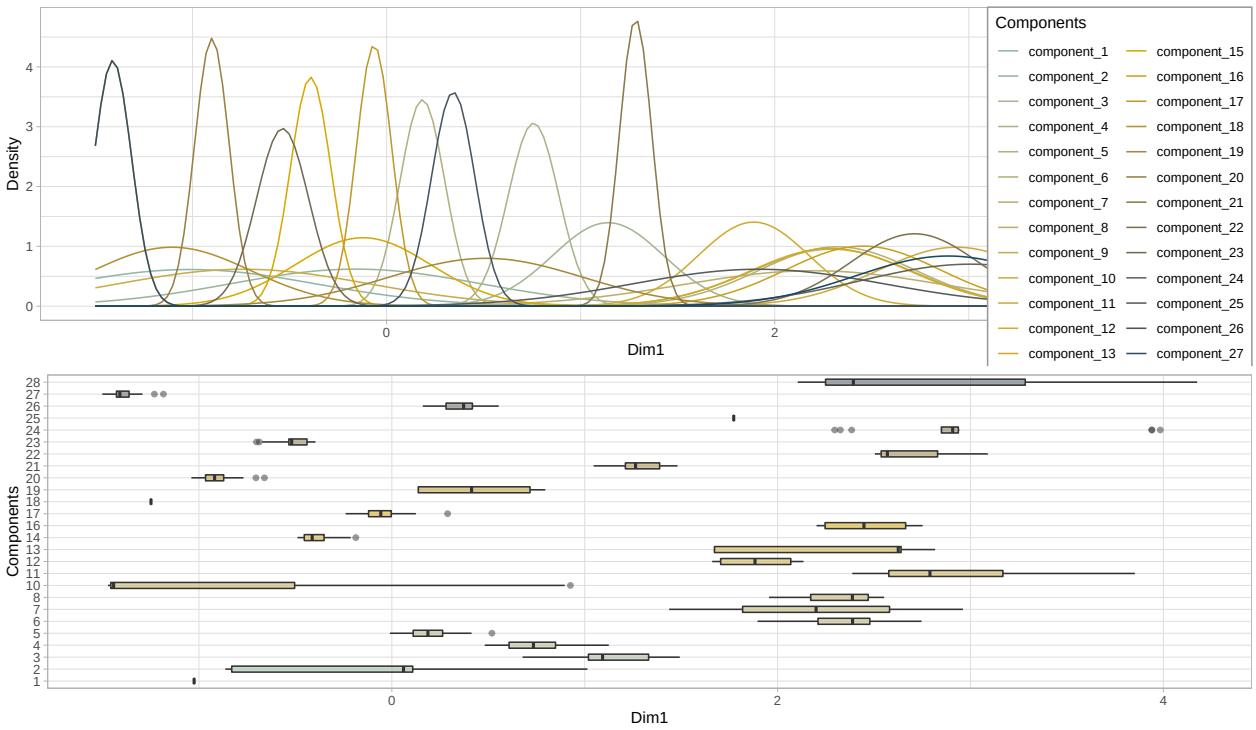
## 1. Végezz MKA-t a PTELJ, Pboldog, Pmagány input változókkal, outlier kiszűréssel! Melyik megoldás tűnik a legjobbnak a BIC-grafikon alapján?

Az *mclust* programcsomag *Mclust()* függvényét használva, különböző **G** érték-konfigurációkat teszteltem a BIC legnagyobb értékét, de ugyanakkor a képződő struktúra értelmezhetőségét is szem előtt tartva. 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.



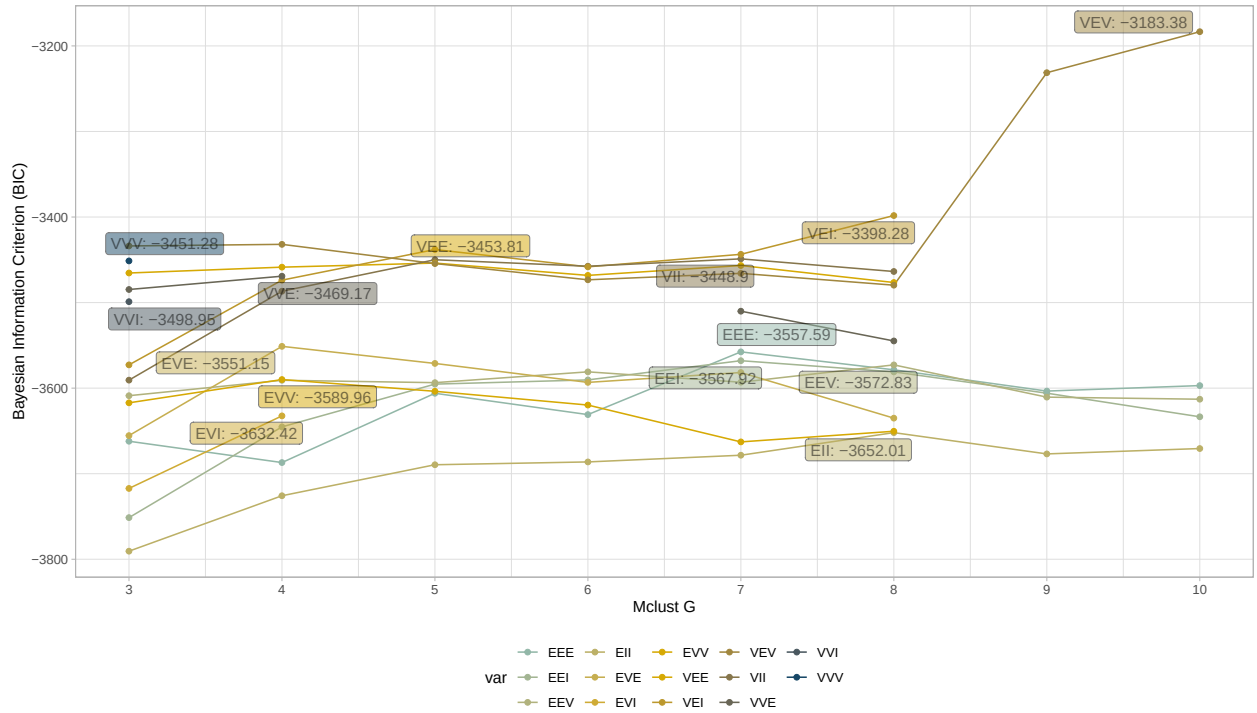

**1. ábra** MKA BIC eredmények G=3 és G=30 között

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




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

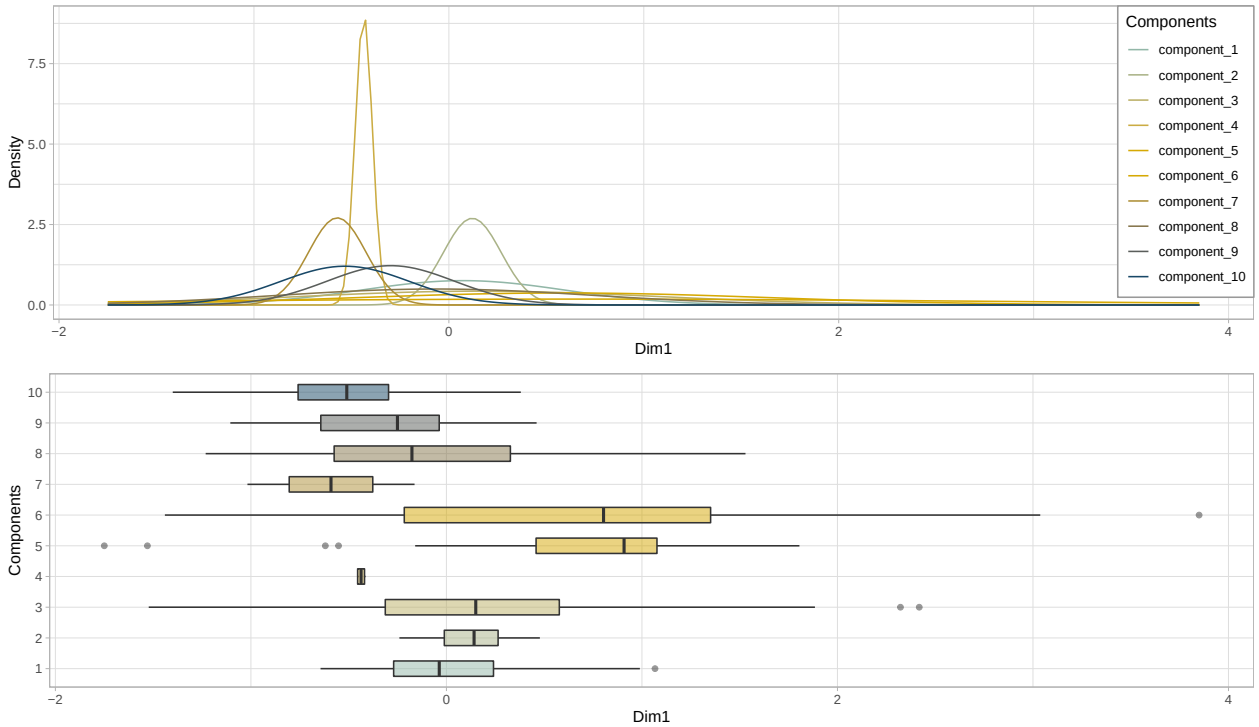
Mindezeken túl azonban, az értelmezhetőség végett az elemzést a G=3 és G=10 tartományra is megismételtem.




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

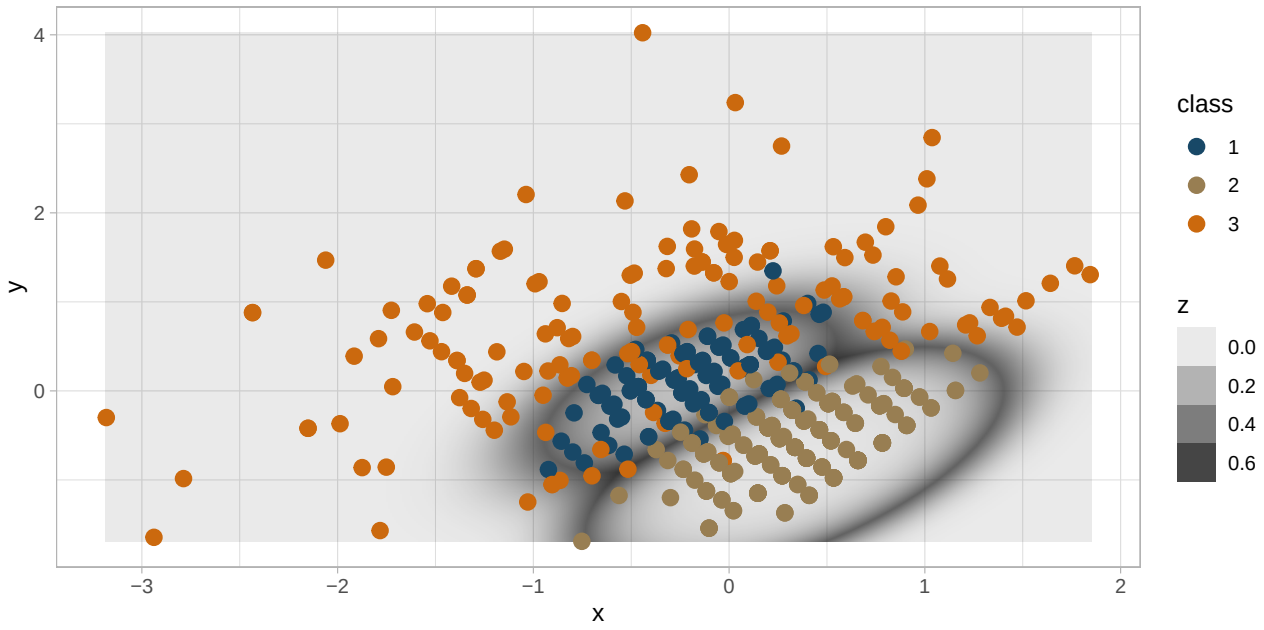
Ezen elemzés nyomán azt látjuk, a VEV eloszlástípus bizonyul a legjobbnak, mely a fenti, kiterjesztett elemzésnél is a legjobbak között volt, így az képződött struktúra értelmezhetőségének reményében a G=10, VEV struktúrát fogadom el.

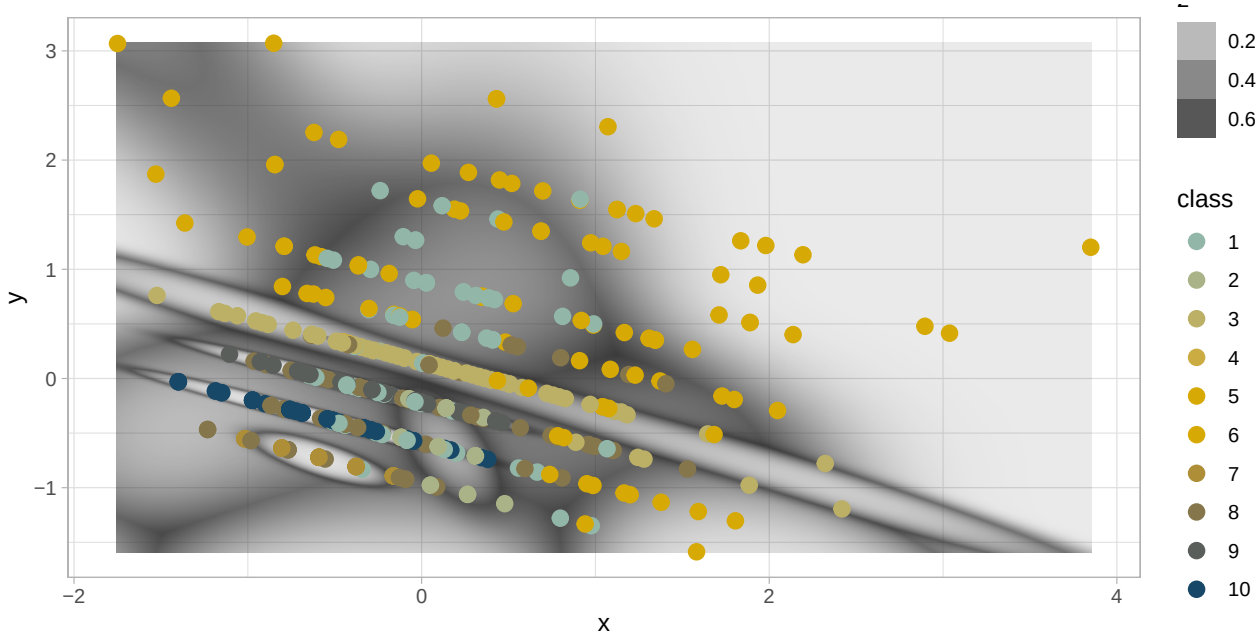
Ha az keverékkomponensek sűrűségét tekintjük (alábbi, 4. ábra)




**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.

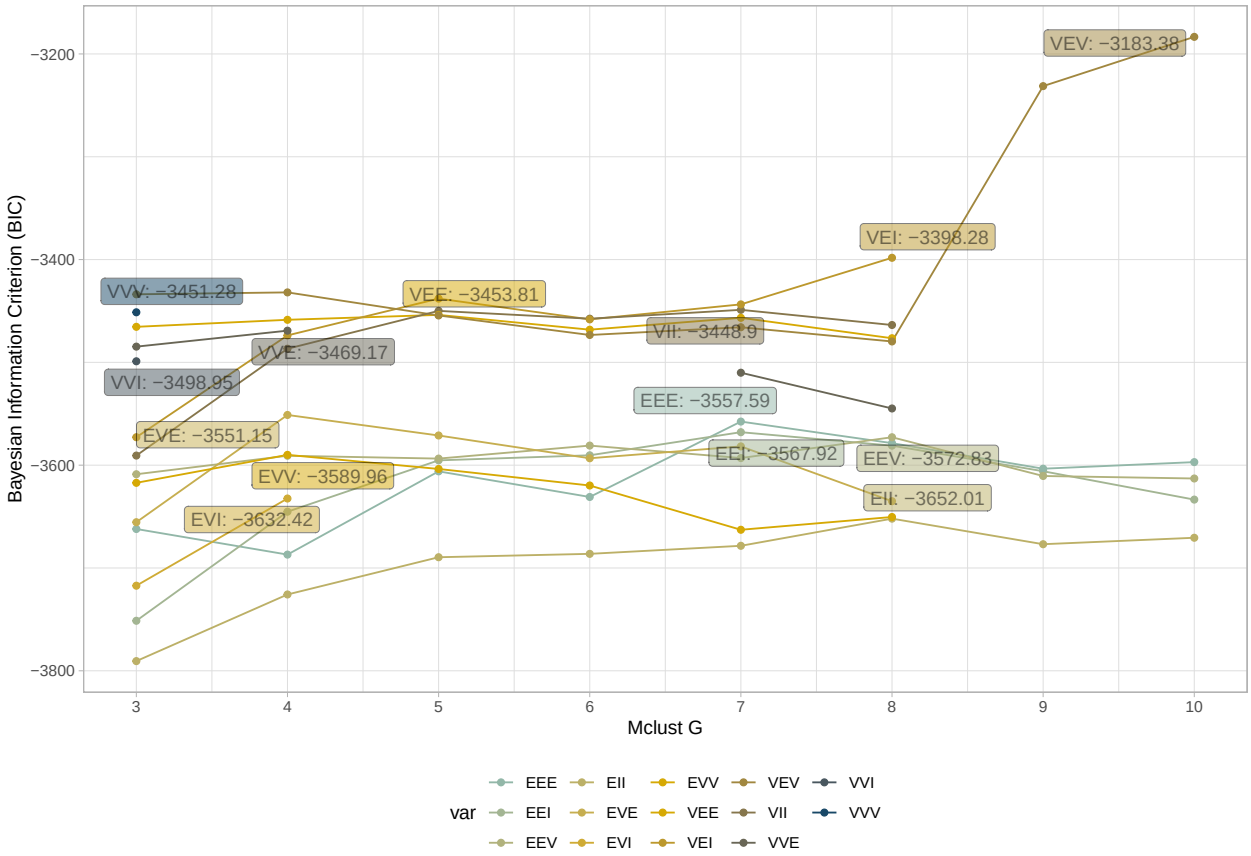





**5.ábra** A G=3 és a G=10 struktúrák összevetése az adatok első két főkomponense mentén képzett két dimenziós síkban; a klaszterhatárok bizonytalanságát a háttér árnyalata (*z-paraméter*) jelzi.

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

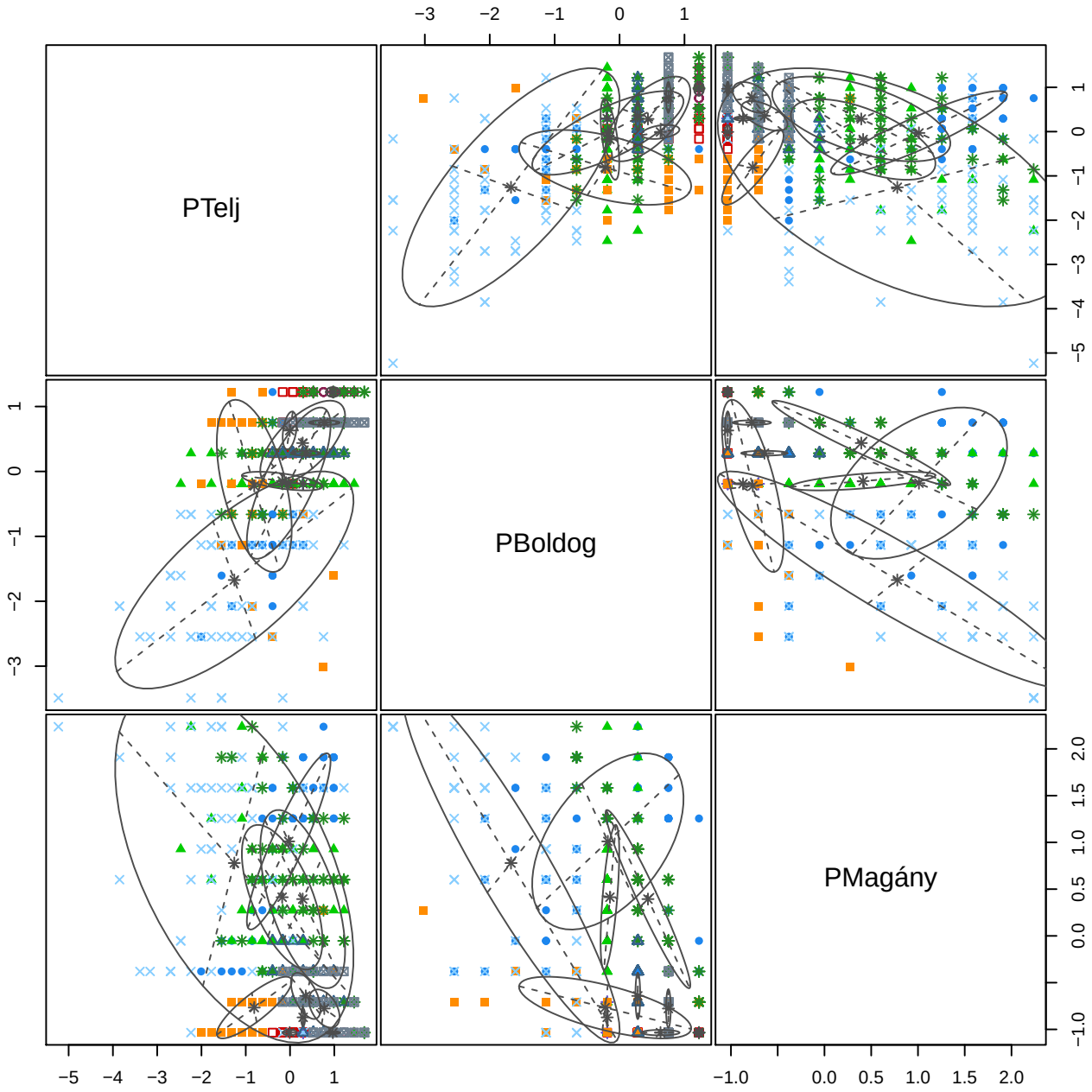
## 2. Készítsd el az 1. feladat BIC-grafikonját k = 3 és 10 között!




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

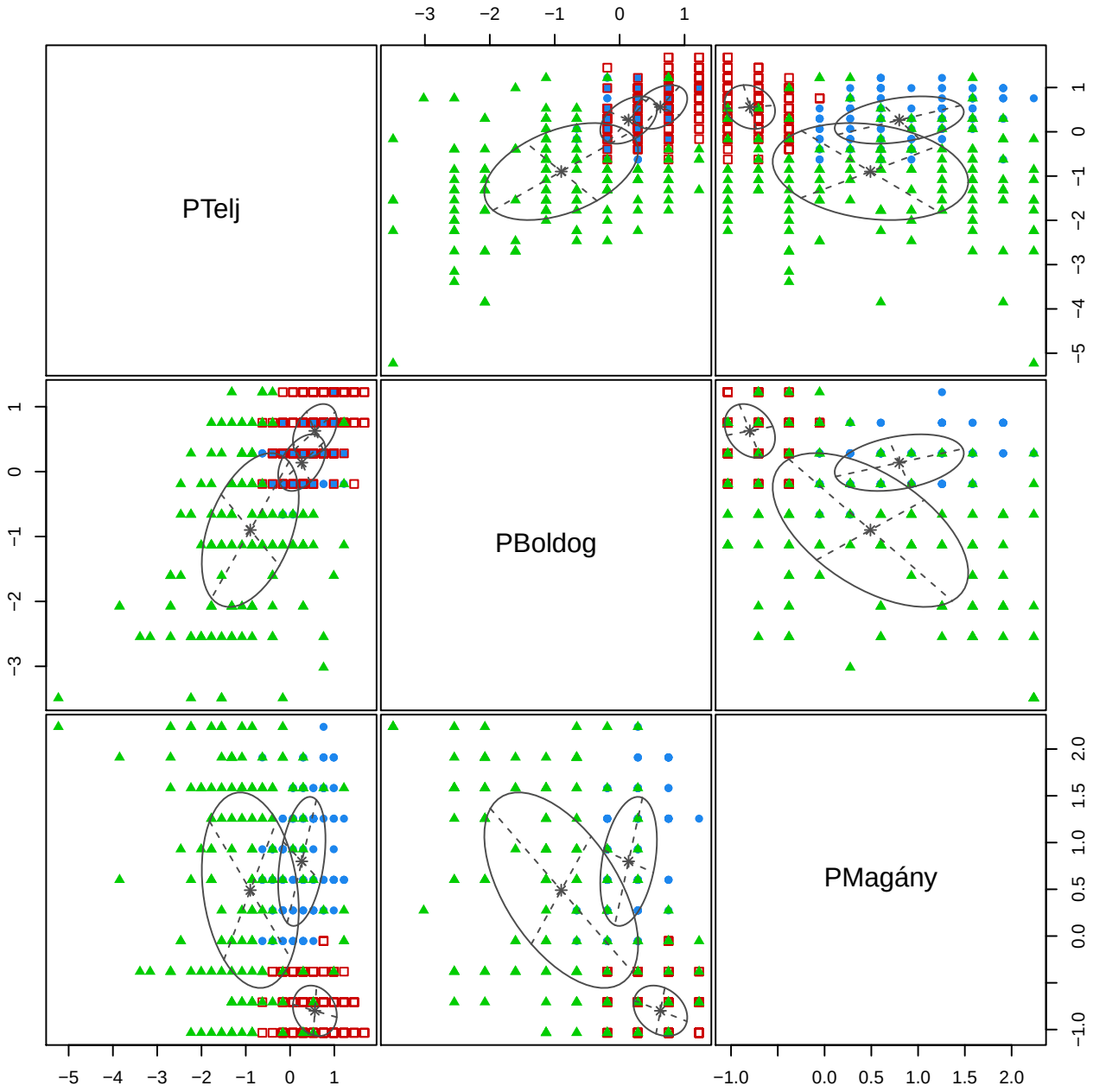
BIC ábráimon a szövegdobozok az adott keveréktípus maximuális értékénél állnak. Ennek értelmében G=10-nél a legjobbnak tűnő eloszlástípus a VEV, melynek BIC értéke -3183.38.

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




**7. ábra** A G=10 megoldás klasszifikációs ábrája

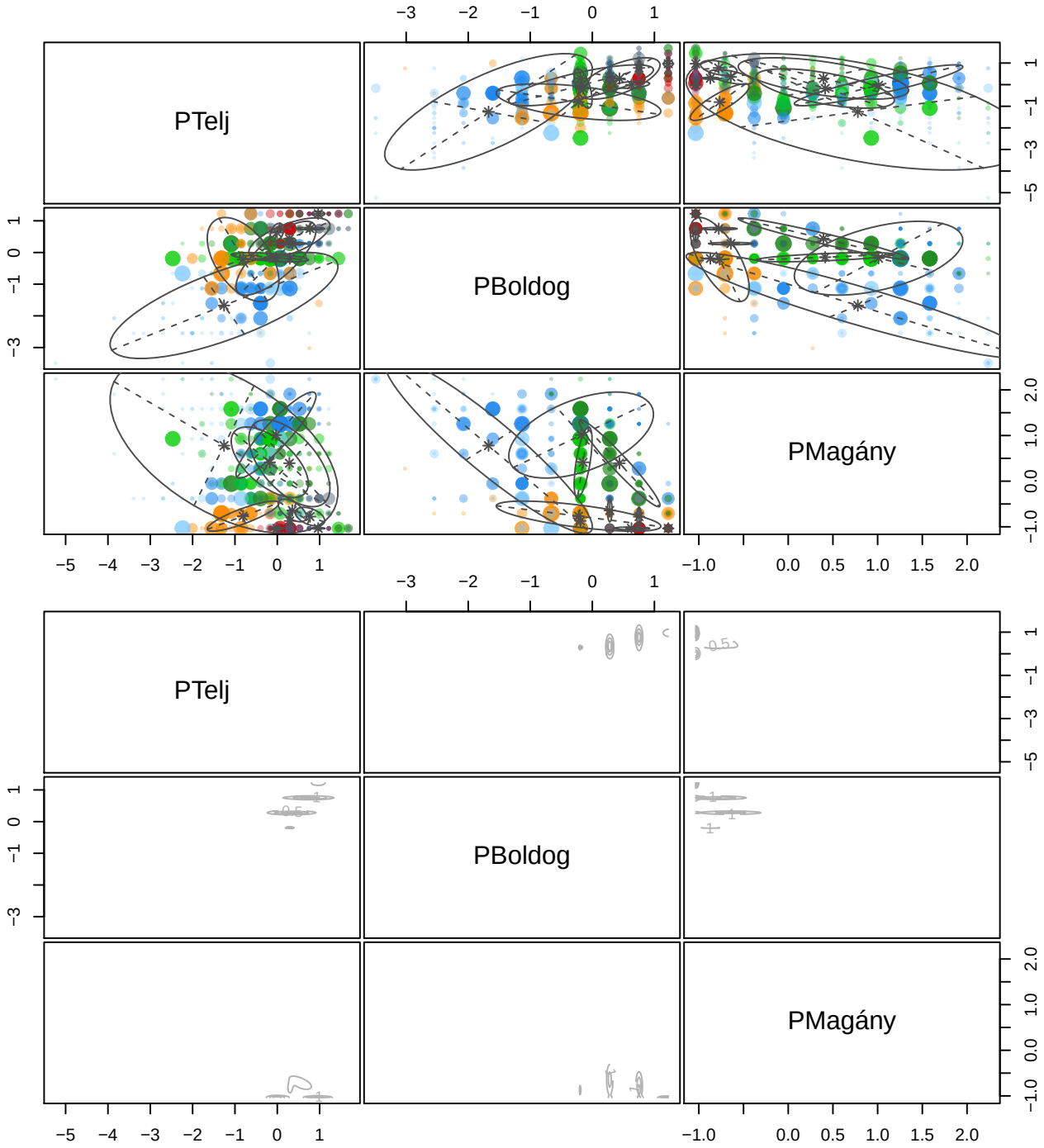
Jól látható három nagyobb klaszter elkülönülése, és több kisebb, részben átfedő struktúra is. Ez ismét felveti a kérdést, mennyiben értelmezhetőbb egy G=3 struktúra. Különösen a Boldogság és Magányosság szeleteiben láthatóak néhány személyt magukban foglaló apró klaszterek.




**8. ábra** A G=3 megoldás klasszifikációs ábrája

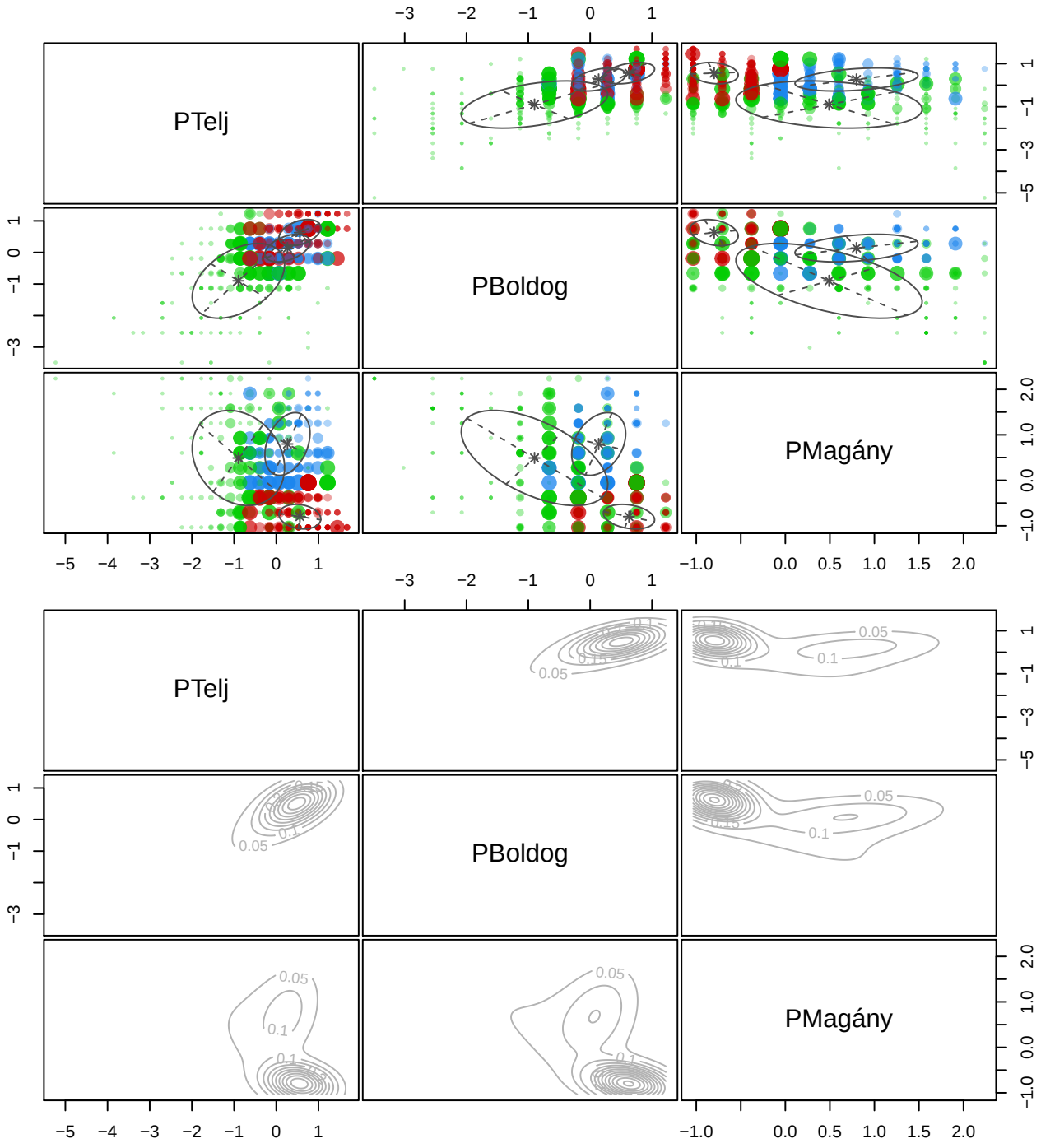
A BIC-értékben való csökkenés ellenére egy jobban értelmezhető, kevésbé átfedő, kevésbé redundáns megoldást kapunk.

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



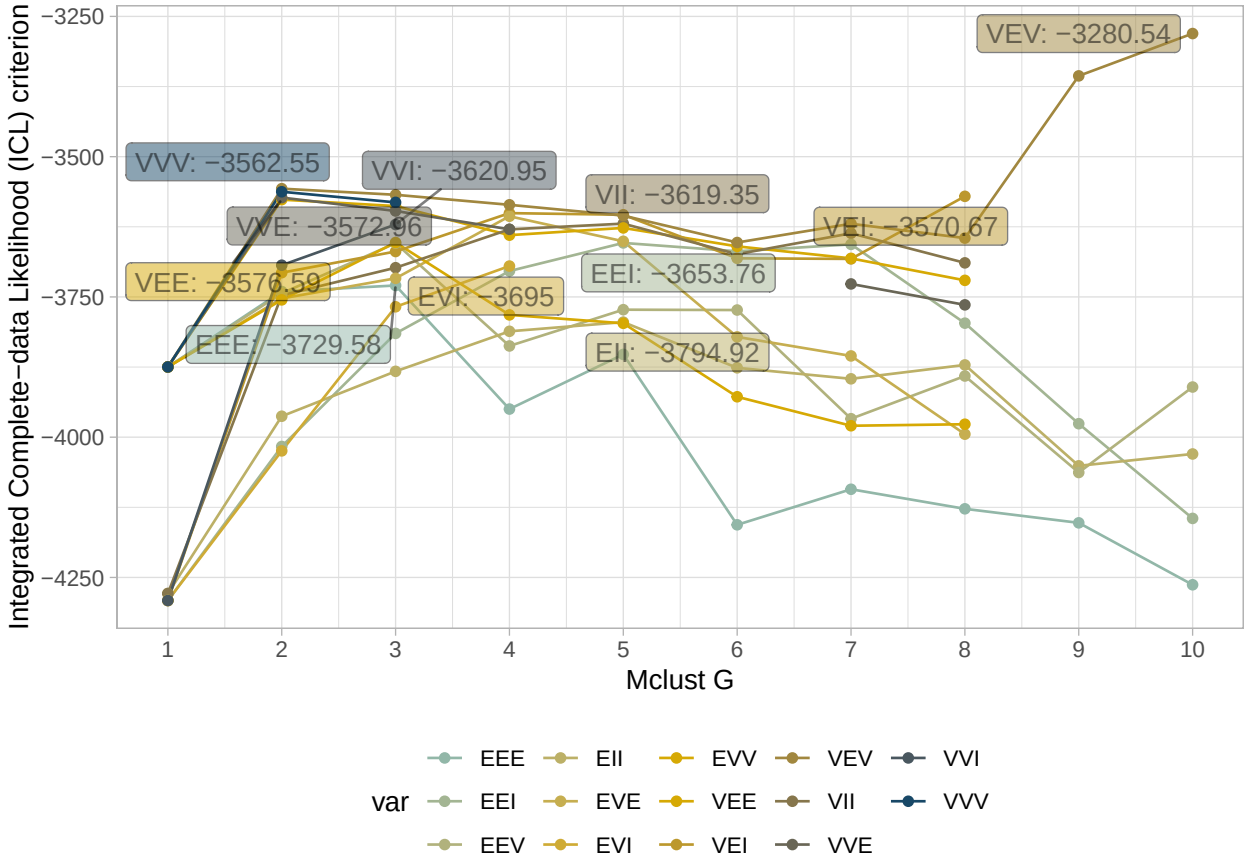

**9. ábra** A G=10 megoldás "uncertainty" és sűrűsödés ábrája

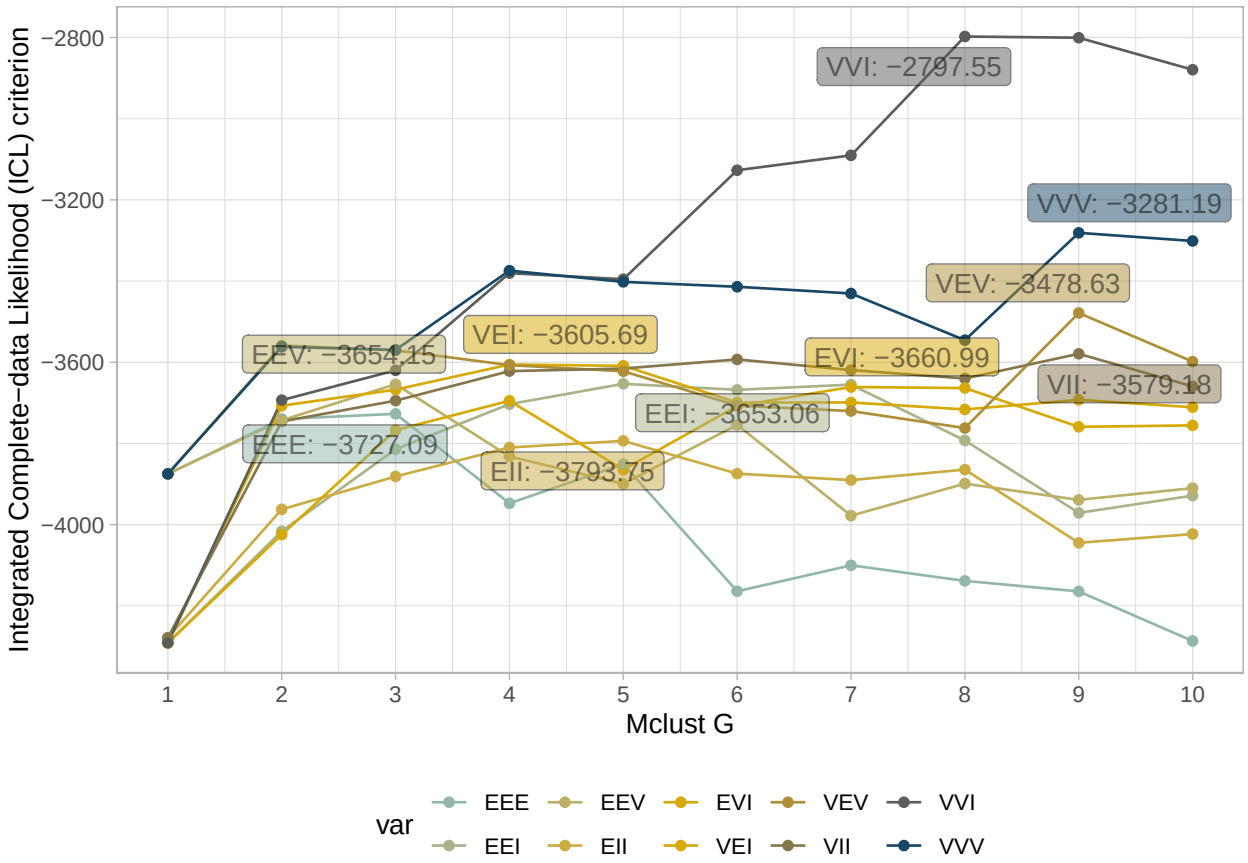
G=10 esetében a klaszterek átfednek, több esetben is a klaszterbe tartozás bizonytalansága emelkedett.




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

**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. Ezt úgy értelmezem, hogy méretükben, tengelyeik hosszában eltérő, de a főtengelyekkel és egymás tengelyirányultságában egyező eloszlásokat modellez a legjobb ICL-lel leírható modell.

## 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!