Domača naloga 2

Neža Kržan, Tom Rupnik Medjedovič

1 Cilj naloge

Želiva preučiti 3 različne metode razvrščanja v skupine. Primerjaval bova metodo voditeljev(*k-means*), razvrščanje na podlagi modelov s pomočjo BIC kriterija in Wardowo metodo hierarhičnega razvrščanja v skupine. Zanima nas katera bo najboljša na podatkih, generiranih iz bivariatne multivariatne normalne porazdelitve.

Zanima naju tudi, kako na metode vpliva dodajanje nepomembnih spremenljivk, torej tistih, ki imajo enako porazdelitev v vseh skupinah.

Vse te tri metode sva se odločila na podlagi jihovih predpostavk, ker so nekatere dokaj podobne, npr. vse predpostavljajo, da so skupine dovolj ločene oz. ni prekomernega prekirvanja med njimi, homogenosti variance znotraj skupine oz. podatki so v skupinah azmeroma homogeno razporejeni, razvrščanje na podlagi modelov s pomočjo BIC kriterija pa celo predpostavlja normalno porazdelitev podatkov znotrah skupine.

2 Generiranje podatkov

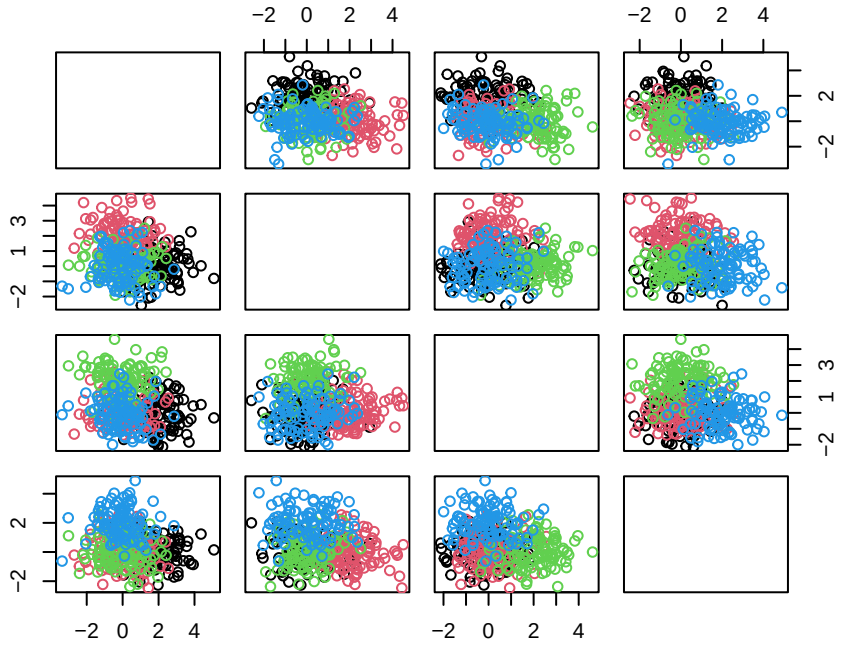
Podatke sva generirala tako, da je njihova porazdelitev bivariatna multivariatna normalna. Zanima naju kako se bodo metode obnesle glede na to kako so si skupine med seboj različne. V ta namen sva si izbrala parameter, ki prilagaja povprečja v skupini, tj. $diff = (0.5, 1, 2, 10)$. Želiva si, da imava primere, ko so si skupine zelo različne med seboj in ne tako zelo različne. Torej bo pri porazdelitvi povprečja generirana s pomočjo faktorja *diff* in število skupin, kovariančna matrika pa bo po diagonali vsebovala število spremenljivk.

Faktorji, ki jih bova še spreminjala so:

- število skupin, $k = (4, 8, 10)$,
- velikosti skupin, $n = (20, 100, 200)$, ri čemer bodo imele vse skupine vedno enako velikost in
- število spremenljivk, $v = (8, 16, 2)$.

Faktorji so bili izbrani na podlagi tega, da si želiva rezultate, ki bodo dobri in slabi oziroma da bodo za nekatere metode dobri za druge pa slabi.

Pri generiranju podatkov bo število spremenljivk enako številu skupin, vse ostale spremenljivke bodo neinformativne, ker nas zanima tudi kako vpliva dodajanje nepomembnih oz. neinformativnih spremenljivk.



Slika 1: Primer generiranih podatkov za 4 skupine, velikosti n = 100, 8 spremenljivk in diff = 2.