Domača naloga 2

Neža Kržan, Tom Rupnik Medjedovič

1 Cilj naloge

Želiva preučiti 2 različni metodi razvrščanja v skupine. Primerjala bova metodo voditeljev(*k-means*) in razvrščanje na podlagi modelov. Za metodo razvrščanje na podlagi modelov sva pustila, da izbere najboljši model na podlagi BIC vrednosti. Število skupin pri kateri se računa BIC vrednost sva določila glede na trenutne nastavitve (`settings`) v iteraciji. Zanima nas katera bo najboljša na podatkih, generiranih iz bivariatne multivariatne normalne porazdelitve.

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

Za metodi sva se odločila na podlagi njunih predpostavk, ker so nekatere dokaj podobne, npr. predpostavljajata, da so skupine dovolj ločene oz. ni prekomernega prekrivanja med njimi, homogenosti variance znotraj skupine oz. podatki so v skupinah razmeroma homogeno razporejeni in zahtevata vnaprejšnjo določitev števila skupin, poleg tega pri razvrščanju na podlagi modelov zahtevamo v predpostavkah, da so podatki generirani iz multivariatnih normalnih porazdelitev.

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. $diff = (1, 2, 4, 10)$. Želiva si, da imava primere, ko so si skupine zelo različne med seboj in ne tako zelo različne. Ker si želiva rezultate, kjer bo ena metoda delovala bolje od druge, bova spreminjala tudi korelacije med skupinami, $cor = c(0, 0.2, 0.9)$.

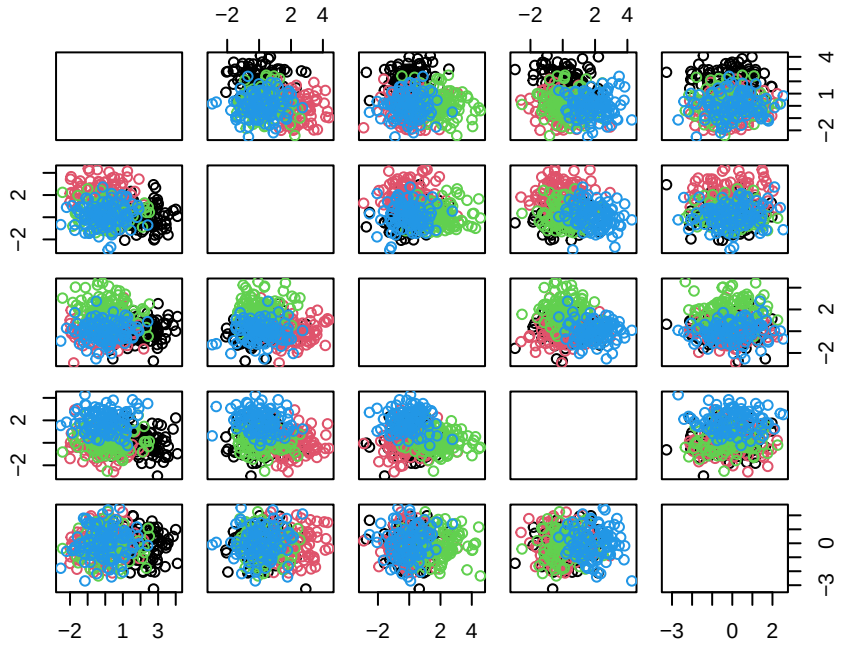
Faktorji, ki jih bova še spreminjala so:

- število skupin, $k = (4, 8, 14)$,
- velikosti skupin, $n = (20, 100, 200)$, pri čemer bodo imele vse skupine vedno enako velikost in
- število neinformativnih spremenljivk, $v = (1, 4, 10)$.

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.

Torej če opišemo postopek generiranja podatkov. Število informativnih spremenljivk sva nastavila na število skupin, število neinformativnih spremenljivk pa bova tekom simulacije spreminjala, saj naju tudi zanima vpliv dodajanja le-teh. Vse spremenljivke generirava torej iz bivariatne multivariatne normalne porazdelitve, kjer imajo informativne spremenljivke povprečje enako parametru `diff`, neinformativne pa enako 0. Standardni odklon pa določiva s pomočjo parametra `cor`, za vse spremenljivke enako(informativne in neinformativne). Število elementov oziroma velikost vzorca, ki ga generirava za spremenljivko pa je določen s pomočjo velikosti skupin.

V naslednjih dveh primerih je število **vseh** spremenljivk enako 5, kjer je 1 spremenljivka **neinformativna**, torej je število skupin enako 4. Na grafu so vidne le 4 spremenljivke, ker ena od vseh ni informativna.



Slika 1: Primer generiranih podatkov za 4 skupine, velikosti n = 100, 5 spremenljivk ter diff = 2 in cor = 0.

3 Simulacija

Izvedla sva simulacijo s 1000 ponovitvami (za večje število ponovitev se nisva odločila zaradi časovne zahtevnosti) in uporabila t.i. paralelno računanje(angl. *parallel computing*). V simulaciji sva generirala podatke in potem izvedla obe metodi razvrščanja v skupine(*metodo voditeljev* in *razvrščanje na podlagi modelov*).

Za obe metodi sva izračunala mero prilagojeni Randov indeks(ARI), ki sva jo uporabila za analizo in primerjavo metod med seboj.

Prilagojeni Randovi indeksi zavzemajo vrednosti na intervalu $[-1, 1]$ in želimo si, da so čim bližje 1, torej da gre za dobro ujemanje med razvrstitvami, kar je boljše od naključnega (vrednosti blizu 0).

Pri primerjavi metod sva se osredotočala predvsem na prilagojeni Randov indeks(ARI), saj so nekatere druge mere kot naprimer WSS pristranske in jo nekatere metode optimizirajo(ravno metoda *kmeans*).