Лабораторна робота №4
з дисципліни
"Статистичний аналіз багатовимірних даних"
Студента 2 курсу магістратугри
групи "Статистика"
Варіант №4

Горбунов Даніел

18 грудня 2022 р.

Зміст

1 Вступ. **2**

2 Хід роботи. **2**

2.1 Метод головних компонент. 2

2.2 Ієрархічна кластеризація на початкових даних. 5

2.2.1 Ієрархічна кластеризація на основі евклідової метрики. 5

2.2.2 Ієрархічна кластеризація на основі метрики Махаланобіса. 8

2.3 Спектральна кластеризація. 11

3 Висновки. **22**

1 Вступ.

Дана робота присвячена використанню класичних методів ієрархічної кластеризації та використанню техніки спектральної кластеризації, результати якої було порівняно із попередніми.

2 Хід роботи.

2.1 Метод головних компонент.

Будемо працювати з таблицею з текстового файлу "F4p.txt". Таблиця складається з 160 спостережень та 10 колонок. Поточна задача – спробувати виявити цікаві закономірності між змінними, використовуючи проектування даних на головні компоненти, щозберігають найбільшу частку дисперсії.
Спочатку подивимося, чи можна спостерігати цікаві закономірності на попарних діаграмах розсіювання.

```
data <- read.table("F4p.txt", header=F)
plot(data, cex=0.25)
```

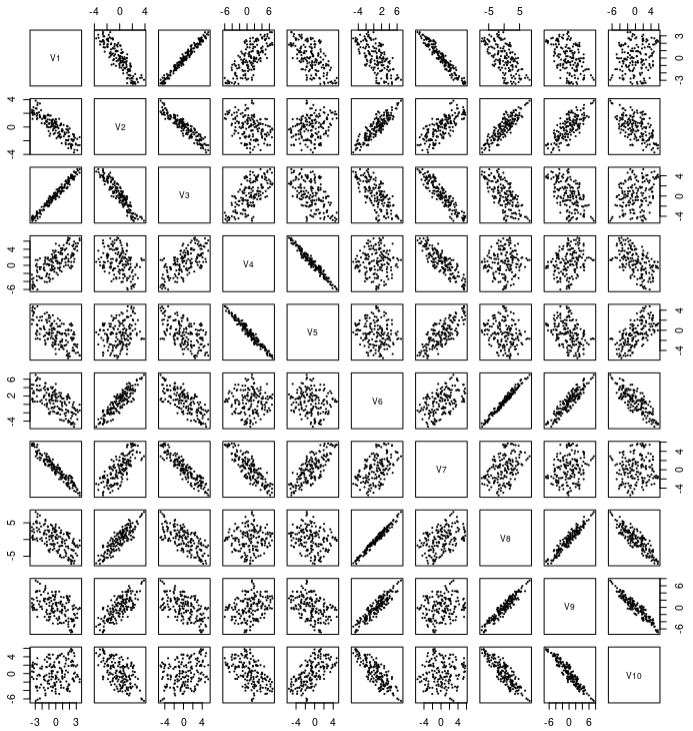

Рис. 1: Попарні діаграми розсіювання початкових даних.

Те, що вперше кидається на око, то це навність лінійних залежностей між деякими змінними, різниця помітна у характері розкиду даних (наприклад, розкиданість точок за змінними V1 та V3 менша у порівнянні з розкидом точок при виборі V1 та V9). Інших геометричних сюрпризів на діаграмах не помітно.

Використаємо техніку головних компонент, але треба визначитися з вибором матриці: коваріаційна чи кореляційна? Легко побачити неоднорідність розкиду:

```
> print(diag(cov(data)))
       V1        V2        V3        V4        V5
 3.500510  2.646443  7.260174  9.071737  5.728481
       V6        V7        V8        V9       V10
 7.294878  7.541027 10.943434  9.026291  8.156558
```

Зокрема попередньо у нас немає інформації про шкали вимірювання кожної з десяти змінних. Тому здається доречним використання кореляційної матриці для подальших обчислень:

```
data.princomp <- princomp(data, cor=T)
plot(data.princomp)
```

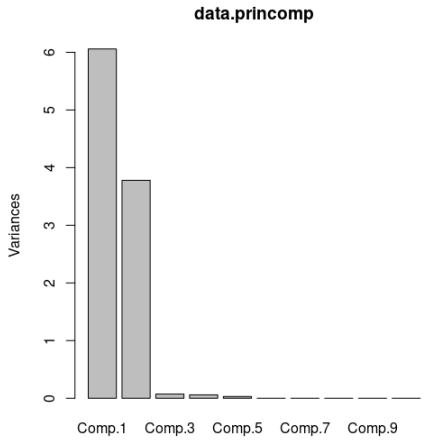

Рис. 2: Діаграма власних чисел за кореляційною матрицею.

За діаграмою власних чисел можна побачити "впливовість" на інформацію про розкид перших двох компонент. Набагато менше впливають наступні дві компоненти, хоча частка збереженого розкиду така, що можна брати ці компоненти до уваги. Також видно, що вплив третьої та четвертої компонент більш-менш однаковий. Корисно буде вивести чисельну інформацію про те, яка частка дисперсії припадає на кожну з компонент:

```
> print(summary(data.princomp))
Importance of components:
                          Comp.1    Comp.2      Comp.3      Comp.4
Standard deviation     2.4613328 1.9442128 0.270127303 0.242424405
Proportion of Variance 0.6058159 0.3779963 0.007296876 0.005876959
Cumulative Proportion  0.6058159 0.9838122 0.991109116 0.996986075
                           Comp.5       Comp.6       Comp.7
Standard deviation     0.173606597 2.739156e-08 2.384756e-08
Proportion of Variance 0.003013925 7.502974e-17 5.687061e-17
Cumulative Proportion  1.000000000 1.000000e+00 1.000000e+00
                             Comp.8       Comp.9 Comp.10
Standard deviation     2.291657e-08 1.467275e-08       0
Proportion of Variance 5.251693e-17 2.152896e-17       0
Cumulative Proportion  1.000000e+00 1.000000e+00       1
```

Як видно, при використанні перших трьох (чотирьох) компонент зберігається приблизно 99% інформації про розкид початкових даних. Далі переходимо до візуалізації.
Першочервого варто сконцентруватися на перших двох компонентах. Діаграма розсіювання проекції на відповідні напрями виглядає таким чином: Особливих закономірностей не видно,

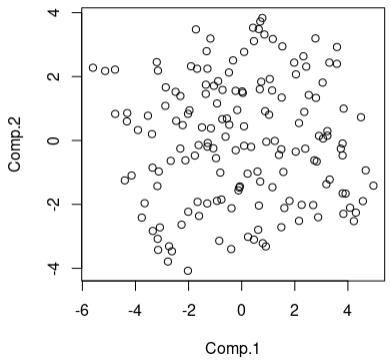

Рис. 3: Діаграма розсіювання даних на перші дві головні компоненти.

лише форма хмари нагадує якусь подушку. Але попереднє спостереження не дає відповіді про можливе розмежування даних. Треба виходити на розмірність вище. Побудуємо просторові діаграми розсіювання для кожних можливої трійки серед осей, що відповідають першим чотирьом головним компонентам. Цікаво, що при виборі третьої компоненти виокремлюється

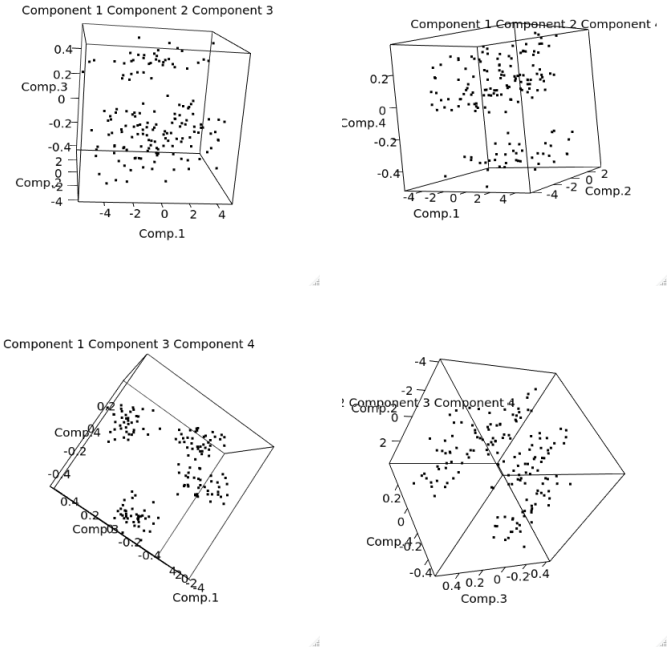

Рис. 4: Діаграма розсіювання даних на перші дві головні компоненти.

дві великі хмарини, а при четвертій – цілих чотири. Можна припустити, що логічним вибором кількості кластерів буде, відповідно 2 та 4 (або навіть 3 в одному з випадків). Питання: чи здатна вхопити спектральні особливості даних техніка ієрархічної кластеризації?

2.2 Ієрархічна кластеризація на початкових даних.

Застосуємо ієрархічну кластеризацію на основі двох метрик:

1. Евклідова метрика: $\rho_e(x,y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2}$;

2. Метрика Махаланобіса: $\rho_m(x,y) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2/\hat{S}_j^2}$, де $\hat{S}_j^2$ – оцінка дисперсії j-ої координати, x, y – реалізації векторів з відповідного розподілу.

2.2.1 Ієрархічна кластеризація на основі евклідової метрики.

Продемонструємо дендрограми кластеризації для трьох методів зв'язку: "Одного", "Повного" та "Середнього".

```
# Euclidean distance

d.euclidean <- dist(data, method="euclidean")

clust.hierarchy.single.e <- hclust(d.euclidean, method="single")
plot(clust.hierarchy.single.e, labels=F)

clust.hierarchy.complete.e <- hclust(d.euclidean, method="complete")
plot(clust.hierarchy.complete.e, labels=F)

clust.hierarchy.average.e <- hclust(d.euclidean, method="average")
plot(clust.hierarchy.average.e, labels=F)
```

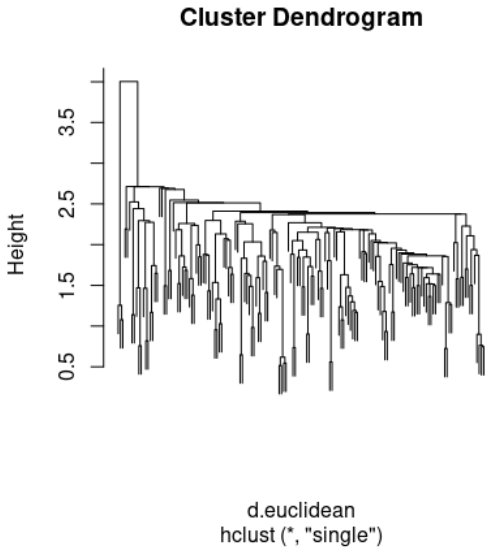

Рис. 5: Дендрограма кластеризації на основі ρ_e з використанням одного зв'язку.

Метод одного (найближчого) зв'язку на евклідовій метриці виділяє купу дрібних кластерів, зливаючи їх майже одразу в два, де кількість у першому набагато вища за другу. Якщо згадати просторову діаграму розсіювання, то кількість точок у кожній з двох хмарин була помірною, що не узгоджується з отриманим результатом. Кофенетична кореляція становить ≈ 0.41937.

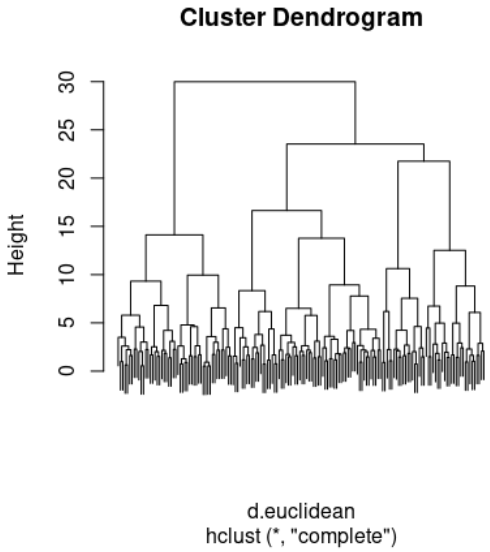

Рис. 6: Дендрограма кластеризації на основі ρ_e з використанням повного зв'язку.

Метод повного зв'язку, на відміну від одного зв'язку, вбачає певні особливості у положенні даних, а тому й зміг згрупувати у декілька кластерів. Добре видно, що утворюється два великих кластери, зокрема кожен з них має по два виразних підкластери. В цілому, по мірі зменшення коефенетичної відстані, спостерігається така кількість кластерів, яка добре виділяється на фоні інших, дрібніших розбиттів: 2, 3, 5. Однак отримані розбиття не узгоджуються з геометричними формами, які спостерагілися раніше (див. нижче рисунки). Отримана кофенетична кореляція становить ≈ 0.58788.

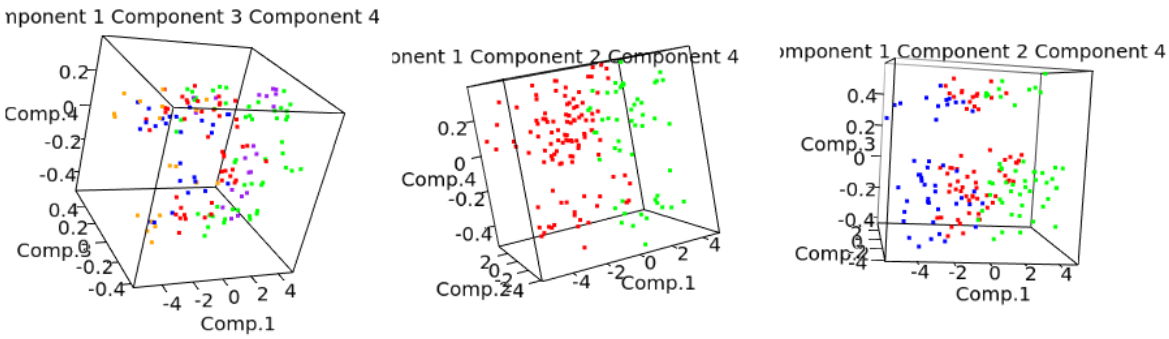

Рис. 7: Просторові діаграми розсіювання даних на головні компоненти з розміткою.

Добре видно, що ніякої узгодженості з геомемтричними формами, що вдалося побачити, немає. Алгоритми ієраріхчної кластеризації на основі повного зв'язку для початкових даних не вбачає їхньої спектральної "структури".

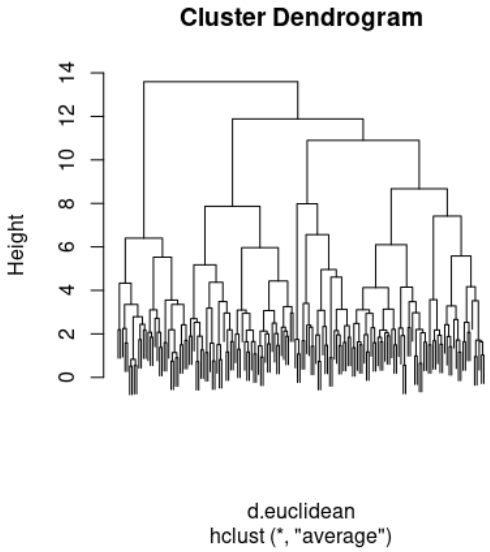

Рис. 8: Дендрограма кластеризації на основі ρ_e з використанням середнього зв'язку.

Метод середнього зв'язку теж виділяє виразно два, три, чотири кластери, як видно на дендрограмі крастеризації. Але чи узгодиться відповідне розбиття зі спосетережуваною геометрію, то треба перевірити. Кофенетична кореляція становить десь 0.60771.

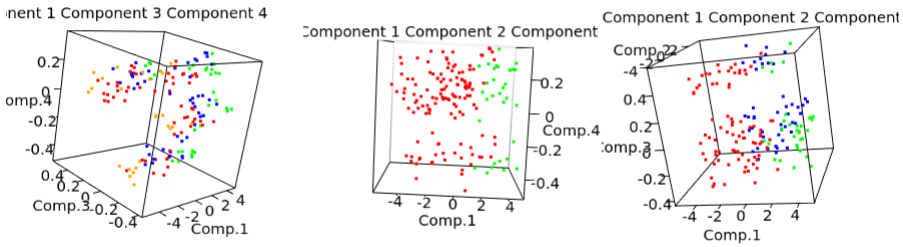

Рис. 9: Просторові діаграми розсіювання даних на головні компоненти з розміткою.

Картинка вийшла дуже схожою з тим, що виходило при використанні повного зв'язку при двох та трьох кластерах. У цьому можна додатково переконатись, якщо побудувати таблиці невідповідностей та обчисливши індекс Ренда:

```
> rand.index(average.e.2, complete.e.2)
[1] 0.7893868
> rand.index(average.e.3, complete.e.3)
[1] 0.7221698
> table(average.e.2, complete.e.2)
           complete.e.2
average.e.2   1   2
          1 110  19
          2   0  31
```

```
> table(average.e.3, complete.e.3)
           complete.e.3
average.e.3  1  2  3
          1 19 19 45
          2  0 31  0
          3 46  0  0
```

В цілому, висновок збігається із попереднім: ієрархічна кластеризація на початкових даних не вхопила спектральні геометричні особливості. Можливо, якщо замінити евклідову метрику на іншу, щось зміниться?

2.2.2 Ієрархічна кластеризація на основі метрики Махаланобіса.

Продемонструємо дендрограми кластеризації для трьох методів зв'язку: "Одного", "Повного" та "Середнього".

```
# x, y -- vectors from sample
# s.vect -- sample variances by factor
mahalanobis.dist <- function(x, y, s.vect)
{
  delta.sq <- data.frame((x - y)^2)
  sqrt(apply(delta.sq / s.vect, 1, sum))
}

mahalanobis.dist.matr <- function(x.matr)
{
  s.x <- apply(x.matr, 2, var)
  m.d <- function(i, j) { mahalanobis.dist(x.matr[i,], x.matr[j,], s.x) }
  idx <- 1:nrow(x.matr)
  d.result <- as.dist(outer(idx, idx, m.d))
  d.result
}

# Mahalanobis distance

d.mahalanobis <- mahalanobis.dist.matr(data)

clust.hierarchy.single.m <- hclust(d.mahalanobis, method="single")
plot(clust.hierarchy.single.m, labels=F)

clust.hierarchy.complete.m <- hclust(d.mahalanobis, method="complete")
plot(clust.hierarchy.complete.m, labels=F)

clust.hierarchy.average.m <- hclust(d.mahalanobis, method="average")
plot(clust.hierarchy.average.m, labels=F)
```

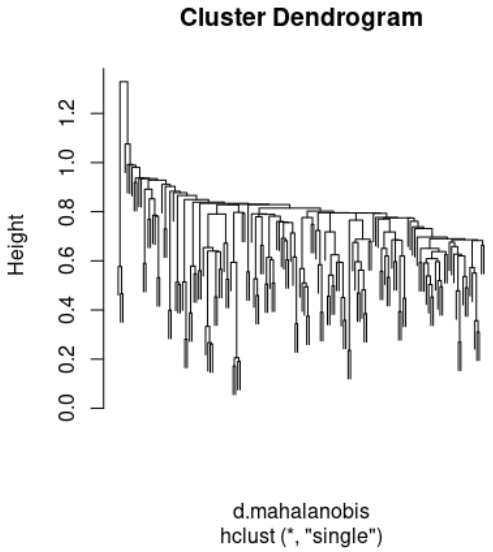

Рис. 10: Дендрограма кластеризації на основі ρ_m з використанням одного зв'язку.

Метод одного (найближчого) зв'язку і при використанні метрики Махаланобіса не може вдало розділити дані хоча б на якісь кластери помірного розміру. Кофенетична кореляція становить приблизно 0.38776.

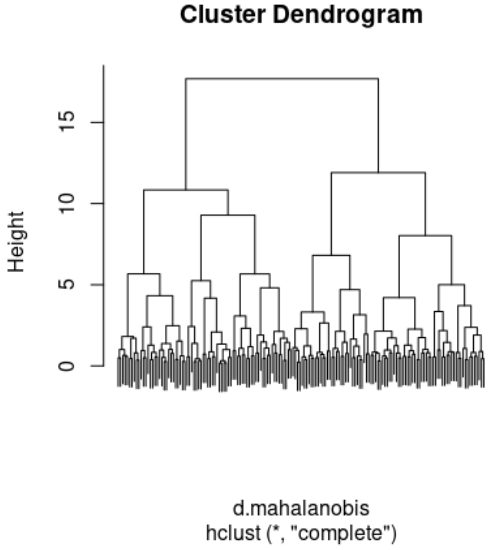

Рис. 11: Дендрограма кластеризації на основі ρ_m з використанням повного зв'язку.

Метод повного зв'язку виокремлює два великих кластери, а в кожному з них ще є розбиття на два підкластери. Розбиття виходить приблизно таким, що виходило при використанні евклідової відстані. Коефенетична кореляція становить 0.49461.

```
> rand.index(complete.e.2, complete.m.2)
[1] 0.6103774
> rand.index(average.e.2, complete.m.2)
[1] 0.5931604
> rand.index(average.e.4, complete.m.4)
[1] 0.7546384
```

```
> table(complete.e.2, complete.m.2)
            complete.m.2
complete.e.2  1  2
           1 76 34
           2  8 42
> table(average.e.2, complete.m.2)
           complete.m.2
average.e.2  1  2
          1 84 45
          2  0 31
> table(average.e.4, complete.m.4)
           complete.m.4
average.e.4  1  2  3  4
          1 15 15 28  0
          2  0 31  0  0
          3 16  0  0 30
          4 19  0  6  0
```

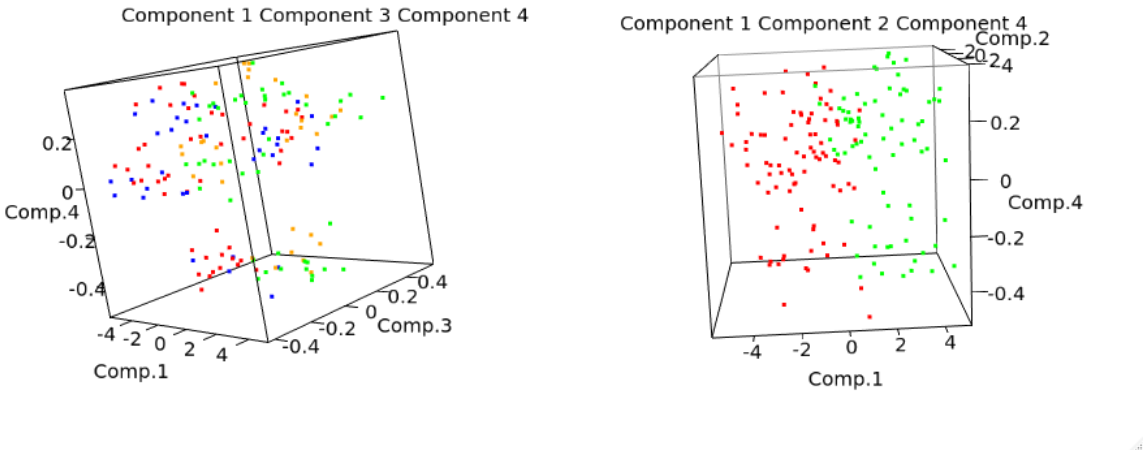

Рис. 12: Просторові діаграми розсіювання даних на головні компоненти з розміткою.

Метод середнього зв'язку дає не добре структуровану дендрограму кластеризації. Коефенетична кореляція становить 0.57175. Враховуючи попередні результати, то і в цьому випадку узгодженості з геометричними формами не буде. Іншими словами, метрика Махаланобіса не допомагає алгоритму побачити потрібні візуальні форми.

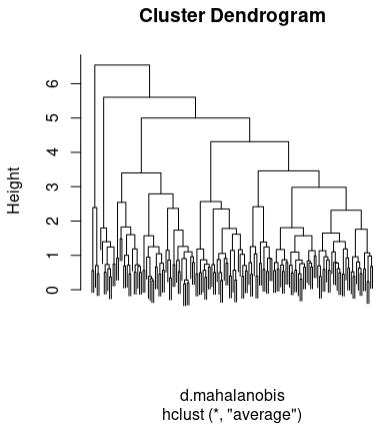

Рис. 13: Дендрограма кластеризації на основі ρ_m з використанням середнього зв'язку.

2.3 Спектральна кластеризація.

Перевіримо, чи зможе вхопити геометричні властивості початкових даних метод спектральної кластеризації. Використаємо його реалізацію з пакета "kernlab".

```
clust.spectral.2 <- specc(data, centers=2)
plot3d(axes.selected[, c(1,2,3)], main="On raw",
       col=c("red", "blue")[clust.spectral.2])
clust.spectral.2.s <- specc(axes.selected[,1:4], centers=2)
plot3d(axes.selected[, c(1,2,3)], main="On spectral",
       col=c("red", "blue")[clust.spectral.2.s])

clust.spectral.4 <- specc(data, centers=4)
plot3d(axes.selected[, c(1,3,4)], main="On raw",
       col=c("red", "blue", "green", "orange")[clust.spectral.4])
clust.spectral.4.s <- specc(axes.selected[,1:4], centers=4)
plot3d(axes.selected[, c(1,3,4)], main="On spectral",
       col=c("red", "blue", "green", "orange")[clust.spectral.4.s])
```

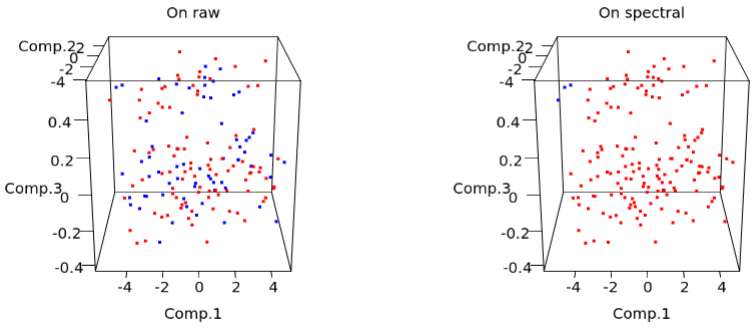

Рис. 14: Просторова діаграма розсіювання на 1, 3 та 4 компоненти з розміткою.

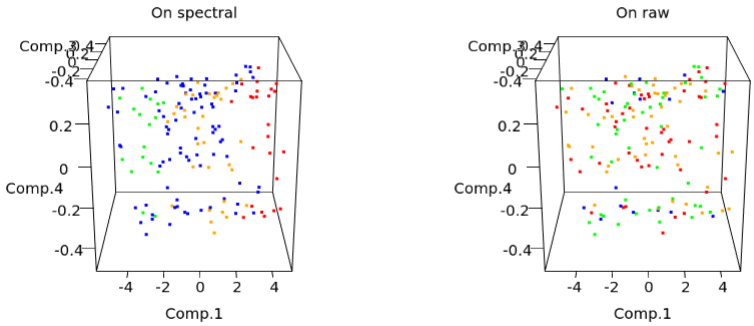

Рис. 15: Просторова діаграма розсіювання на 1, 3 та 4 компоненти з розміткою.

Нормальне розбиття навіть не виходить при використанні безпосередньо спроектованих на головні напрямки дані, що досить дивно. Можливо, варто використовувати іншу інформацію, тобто комбінувати інші напрямки, які містять менше даних про розкид?
Тут ще справа в тому, що початкові дані не містять ніяку геометрію, яка хоча б віддалено нагадувала спектральну. Тому доречно робити кластеризацію на даних після проектування на власні вектори. Подивимося на попарну діаграму розсіювання за всіма головними напрямками.

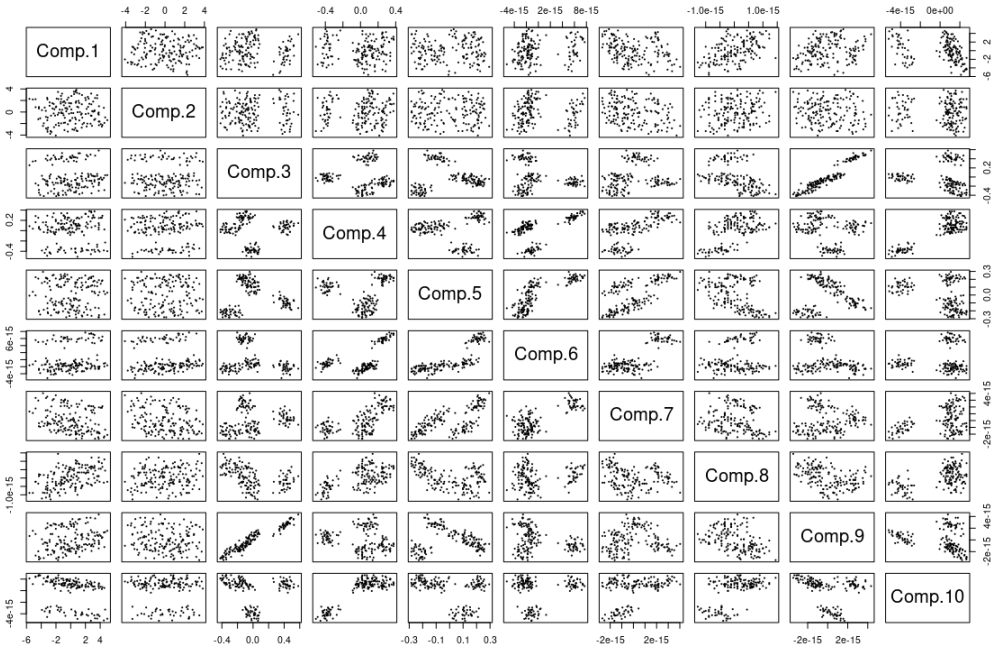

Рис. 16: Діаграма розсіювання за всіма можливими парами головних напрямків.

Напевно, треба було подивитися на початку роботи саме на цю діаграму, аби визначитися з цікавими геометричними формами. Здебільшого легко бачити, що на багатьох змінних виділяється по три-чотири кластери. Оберемо ті головні напрямки, які добре відображають відмінність між положеннями кластерів (бо, здається, розкиданість відносно перших двох головних компонент псувала процес кластеризації).

Наприклад, можна взяти до уваги такі напрямки, як за компонентами 3, 6 та 10. Зобразимо просторову діаграму даних, спроектованих на відповідні напрямки:

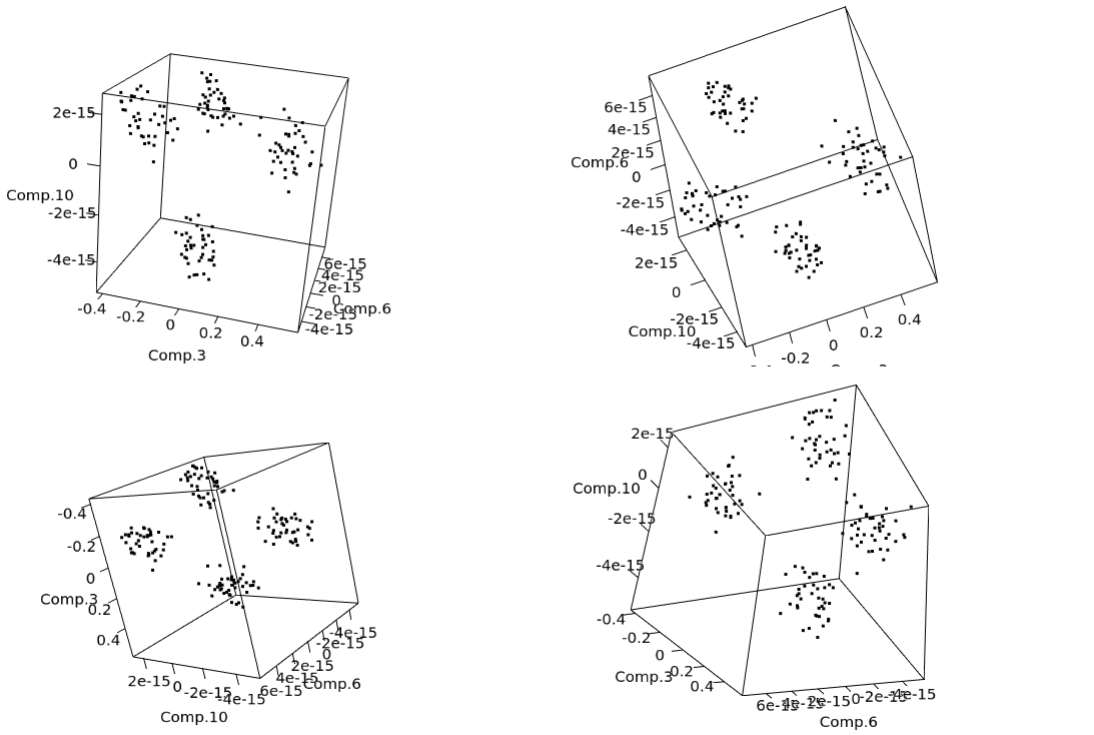

Рис. 17: Просторова діаграма розсіювання на напрямки 3, 6 та 10 головних компонент.

За цими напрямками добре відокремлюються чотири групки даних, які спостерігалися на матричній діаграмі розсіювання. Сформовані кластери є компактними та добре віддалені один від одного. Тому, здається, окрім спектральної кластеризації, використання одного з класичних алгоритмих на кшталт методу центроїдів може повернути хороші результати.
Спочатку використаємо метод центроїдів на обрані компоненти. Звіт з підгонки наступний:

```
> kmeans.spec <- kmeans(axes.selected[,c(3,6,10)], centers = 4)
> print(kmeans.spec)
K-means clustering with 4 clusters of sizes 47, 46, 27, 40
Cluster means:
        Comp.3        Comp.6       Comp.10
1 -0.152230153  2.816832e-15  7.019944e-16
2 -0.008048798  5.829425e-16 -2.725288e-15
3 -0.341377544 -2.542004e-15  1.662026e-15
4  0.418556390 -2.258610e-15  1.138326e-15
Clustering vector:
[...]
Within cluster sum of squares by cluster:
[1] 0.09994440 0.09902231 0.05469977 0.17505512
 (between_SS / total_SS =  96.3 %)
```

Візуалізація буде далі. Варто відмітити, що кластеризація "пояснює" приблизно 96% від усього розкиду даних, що може бути знаком того, що отримана кластеризація узгодиться з фактичною.

Було б цікаво відмітити, яка оптимальну кількість розбиттів бачить метод центроїдів, максимізуючи середній силует.

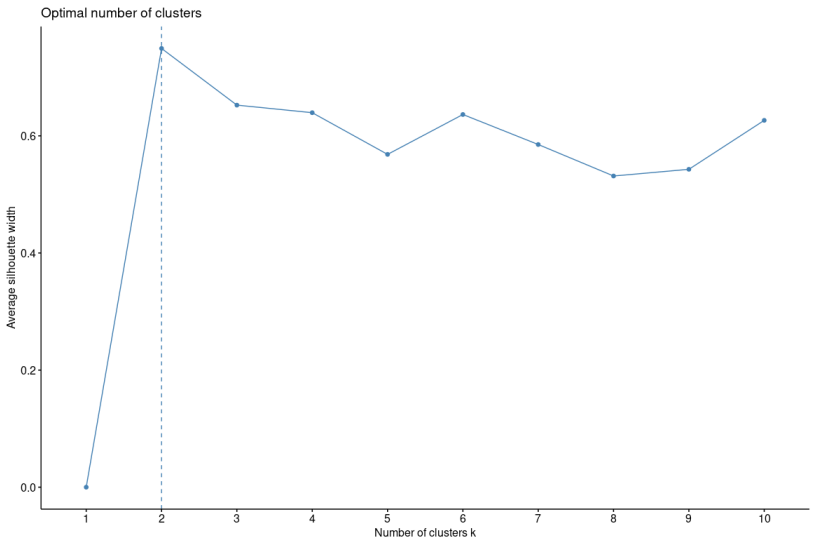

Рис. 18: Діаграма середніх силуетів на основі кластеризації методом центроїдів.

Локальні максимуми спостерігаються при розбиттях на 2, 3, 4, 5 та 10 штук. Візуально, на діаграмах розсіювання, всюди виднілися до чотирьох хмар даних. Можна надалі перевірити кожну з кластеризацій від 2 до 4 підмножин та порівняти узгодженість з геометричними формами на інших головних напрямках.
Спочатку беремо до уваги кластеризацію центроїдами на два розбиття. На матричній діаграмі розсіювання, якщо розмалювати точки отриманими номерами кластерів, спостерігається наступна картина:

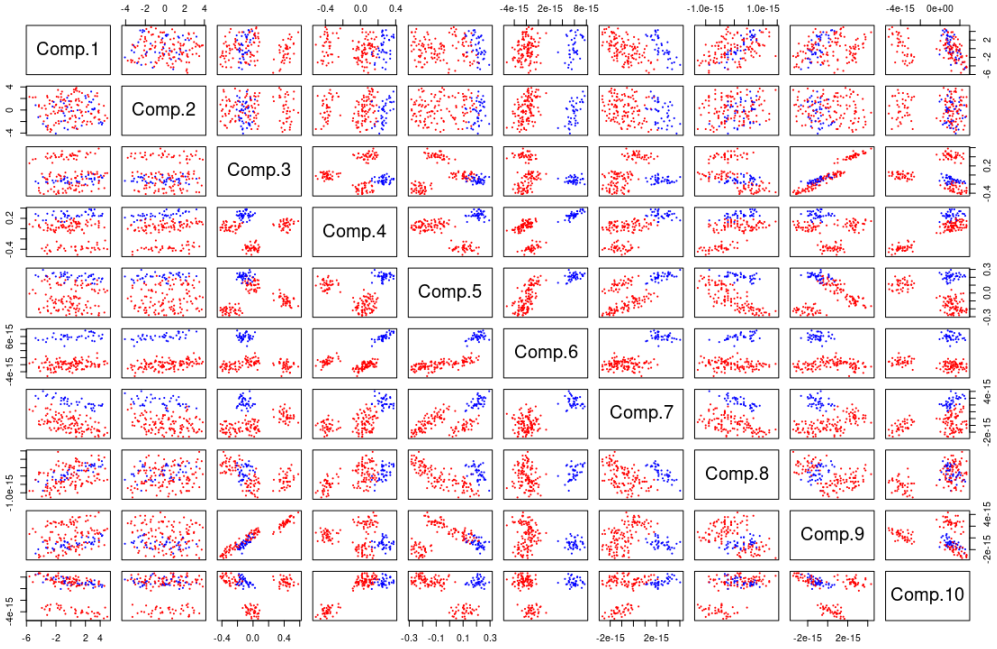

Рис. 19: Діаграма розсіювання за всіма можливими парами головних напрямків.

Хаотичної розмітки, як було при деяких техніках, немає, розмітка виглядає більш-менш організованою. Зрозуміліше буде, якщо подивитися на проекції на ті напрямки, де чітко видно розділення даних на дві частини. Наприклад, можна розглянути проекцію на перші три компоненти. Або ж, зауважимо, на проекції на 1 та 6 напрямки (або 2 та 6) чітко видне розмежовування на два кластери. Також можна долучити напрямок десятої компоненти, там попередньо є схожість на розділення на дві групи. Що вийде у просторі – зараз побачимо.

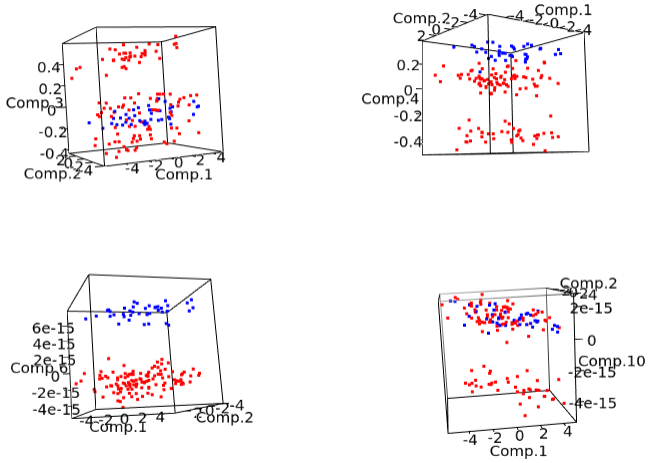

Рис. 20: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

На проекції на 1, 2 та 6 компоненти розмітка відповідає геометричній дійсності. На іншій проекції, де беруться до уваги 1, 2 та 4 компоненти, розмітка теж виходить адекватною, але на око видно, що там можна виділити замість червоного ще два окремих кластери. На інших проекціях (з використанням 3 або 10 компоненти) цікаво спостерігати як один кластер розміщається всередині іншого. Тут нашої помилки немає, бо ми враховували вимірність конкретно обраних напрямків, а отримані підгрупи в інших вимірах можуть вже по-різному розміщуватися.
Тепер час подивитися на розбиття по три кластери.

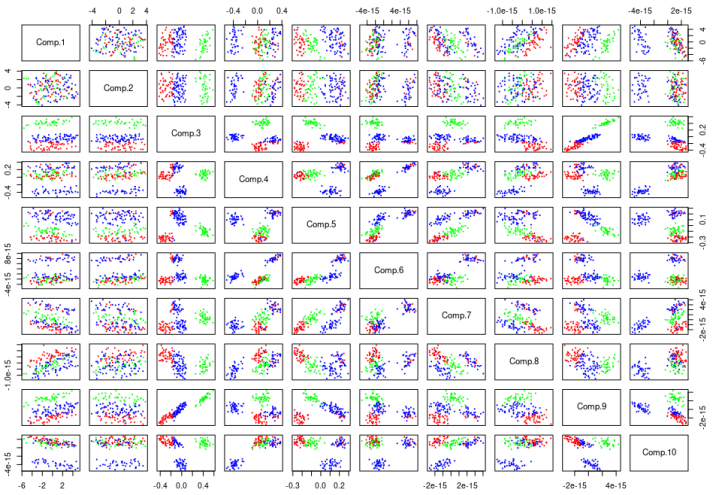

Рис. 21: Діаграма розсіювання за всіма можливими парами головних напрямків.

Якщо робити висновки неозброєним оком, то здається, що отримана розмітка застосовна добре. В тому сенсі, там де можна спостерігати по три кластери, то розмітка це вхоплює. Продемонструємо це на просторових діаграмах даних з проекцією на такі трійки головних компонент: 1,2,3; 1,7,8; 1,2,9; 1,2,10.

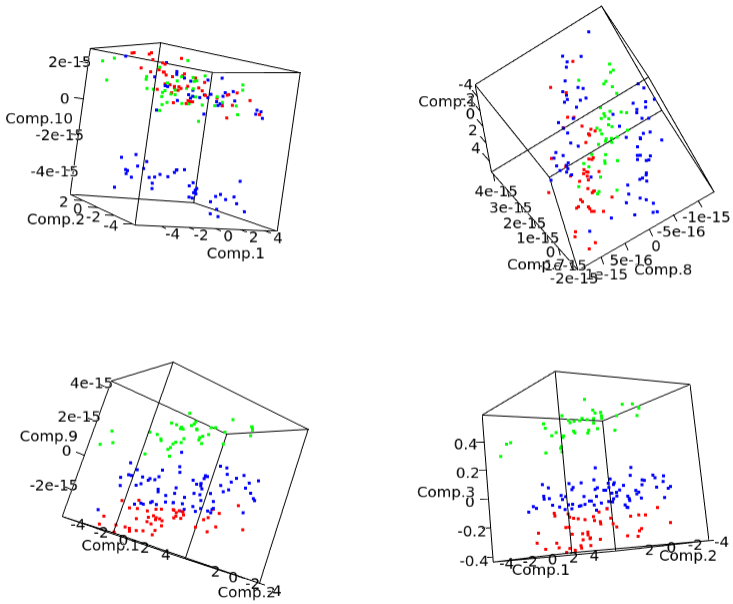

Рис. 22: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

На проекціях на 1,2,3 або 1,2,9 компоненти розмітка узгоджується із розбиттям на три великих хмари даних. На проекції з використанням головних напрямів 1,7,8 компонент насправді можна побачити чотири розтягнутих хмарини. Розмітка "на три" викоремила у різні кластери дві з них, інша об'єднує в один кластер. Ситуація досить хаотична у просторі 1,2 та 10 головних компонент.
Залишається дослідити розбиття на 4 підгрупи.

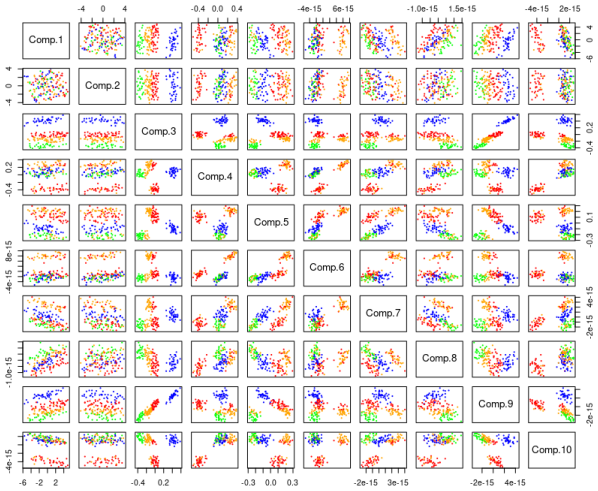

Рис. 23: Діаграма розсіювання за всіма можливими парами головних напрямків.

Візьмемо до уваги проекції на такі трійки компонент: 1,2,3; 1,7,8; 3,4,5; 3,7,9; 1,2,4; 2,3,4; 3,6,9; 3,6,10.

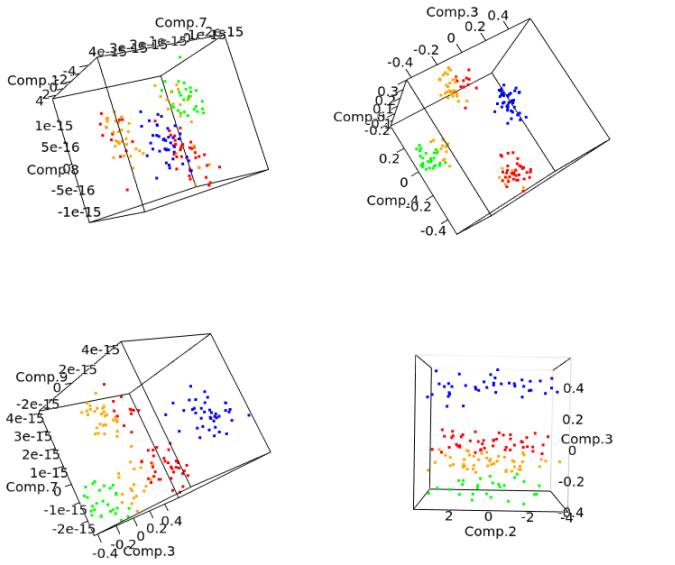

Рис. 24: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

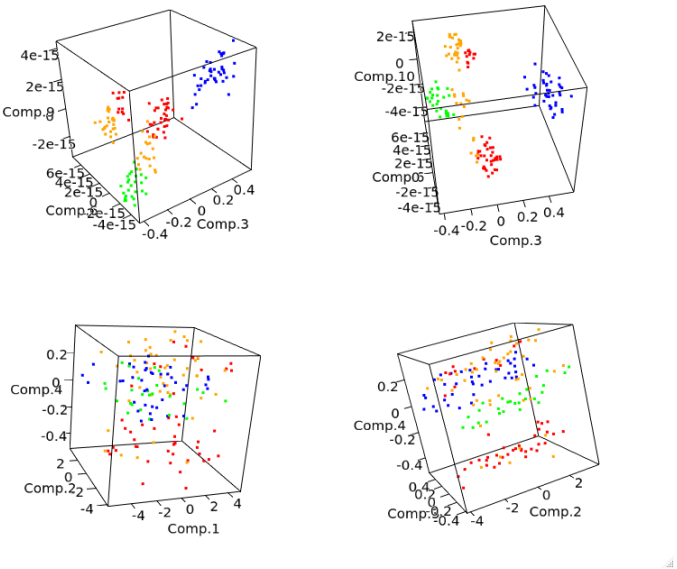

Рис. 25: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

Місцями розмітка вдало лягає на справжню геометричну структуру даних у просторі головних компонент, а в деяких або спостерігається перемішування, або ж розмітка одного кластера чіпляється до сусідніх хмаринок. В принципі, такий результат можна було очікувати, враховуючи те, що ми обмежувалися лише трьома напрямками, що характеризують відмінність у положенням та формах відповідних груп.

Тепер застосуємо ті самі дані, але замість методу центроїдів використаємо техніку спектральної кластеризації. Будемо порівнювати із тим, що виходило за допомогою центроїдів, використовуючи індекс Ренда і таблицю невідповідностей.

```
> # Спектральна кластеризація на основі проекції на 3,6,10 головні напрямки
> specc.2 <- specc(x=axes.selected[,c(3,6,10)], centers=2)
> specc.3 <- specc(x=axes.selected[,c(3,6,10)], centers=3)
> specc.4 <- specc(x=axes.selected[,c(3,6,10)], centers=4)
> rand.index(specc.2, km.2$cluster)
[1] 0.4968553

> table(specc.2, km.2$cluster)
specc.2  1  2
      1 40  0
      2 80 40

> rand.index(specc.3, km.3$cluster)
[1] 0.7925314

> table(specc.3, km.3$cluster)
specc.3  1  2  3
      1  0  0 40
      2 29 74  0
      3 17  0  0

> rand.index(specc.4, km.4$cluster)
[1] 0.8971698

> table(specc.4, km.4$cluster)
specc.4  1  2  3  4
      1  0  0 27  6
      2  1  0  0 30
      3 45  0  0 11
      4  0 40  0  0
```

Кластеризація методом центроїдів стає більш схожою на спектральну, судячи з отриманих невідповідностей. Продемонструємо матричні діаграми розсіювання із розмітками, отриманими спектральною кластеризацією.

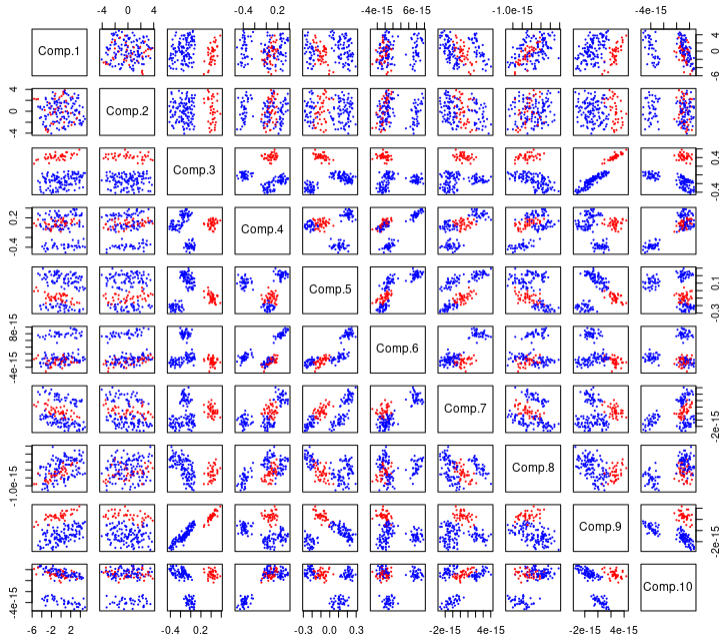

Рис. 26: Діаграма розсіювання за всіма можливими парами головних напрямків.

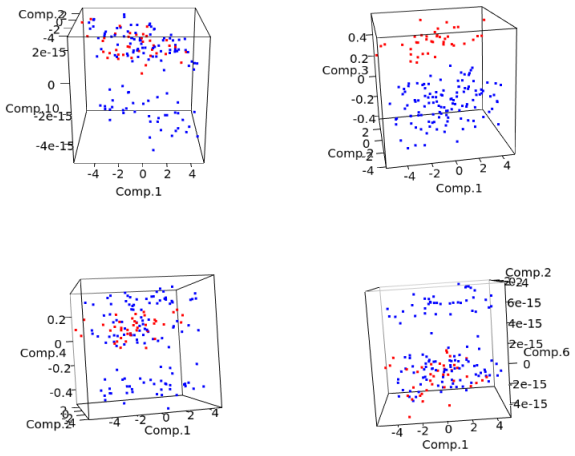

Рис. 27: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

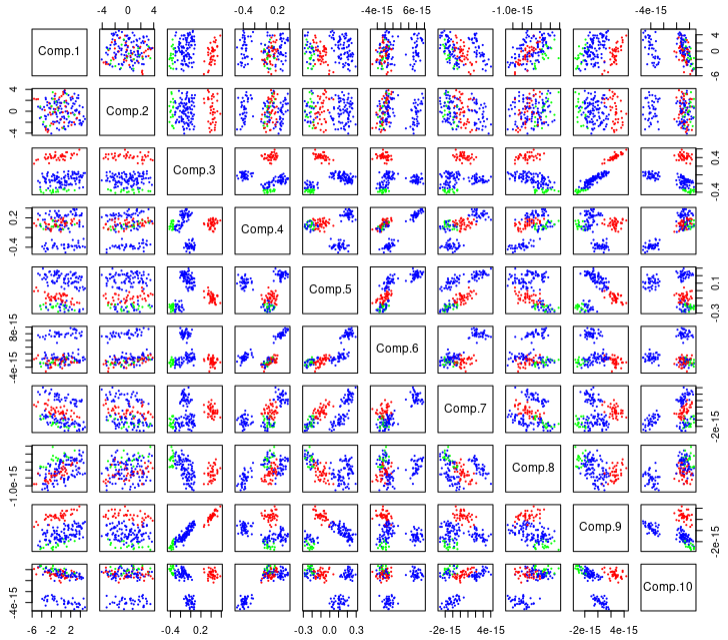

Рис. 28: Діаграма розсіювання за всіма можливими парами головних напрямків.

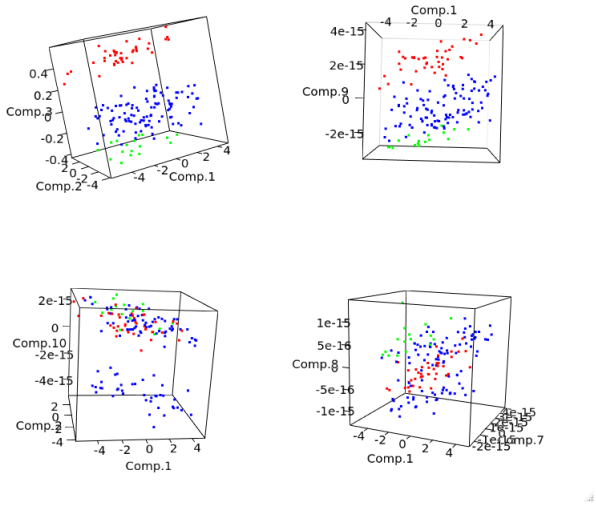

Рис. 29: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

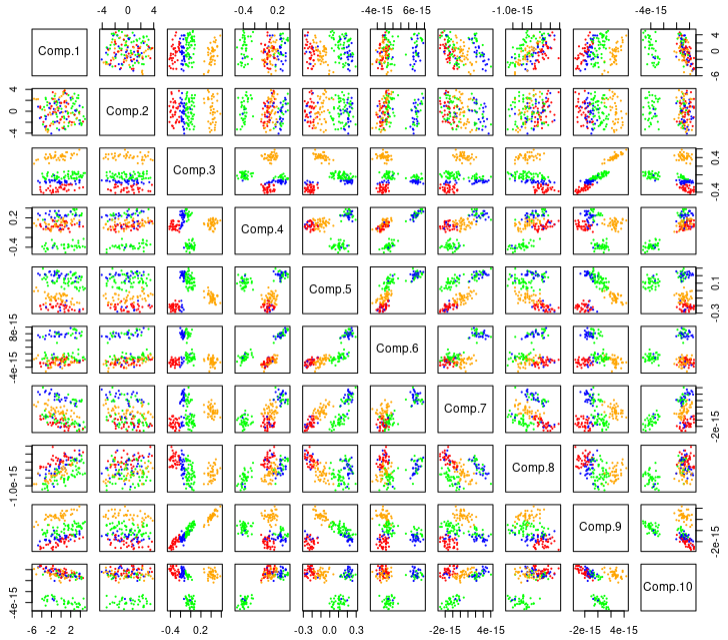

Рис. 30: Діаграма розсіювання за всіма можливими парами головних напрямків.

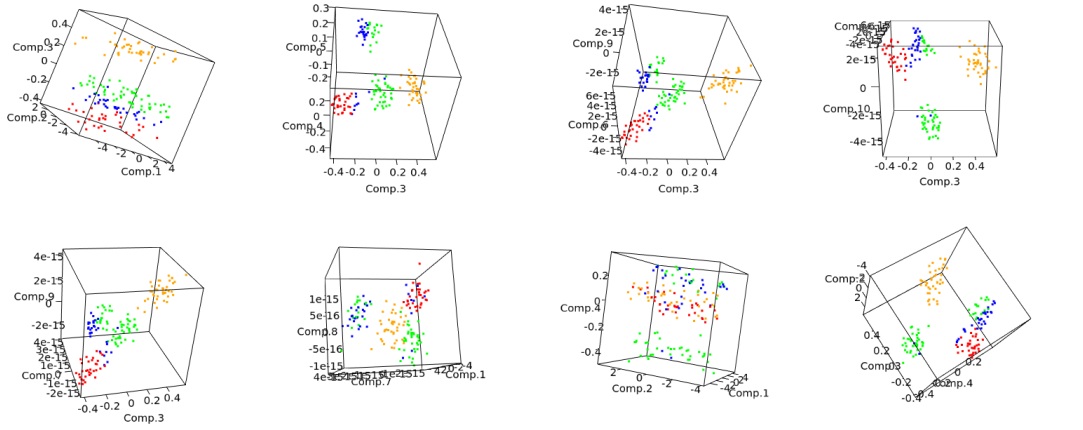

Рис. 31: Просторова діаграма розсіювання на деякі трійки напрямів головних компонент.

Техніка спектральної кластеризації розбить розмітки, які на деяких проекціях краще узгоджуються, ніж це виходило при використанні техники кластеризації центроїдами. Втім, внаслідок того, що обрані три напрямки не охоплюють достатньо інформації про відмінності спостережуваних груп, на деяких проекціях все ж таки можна побачити перемішування, перекриття, дотикання до чужих хмар тощо.

3 Висновки.

У роботі було виявлено різноманіття спектральної геометрії початкових даних, які представляли з себе, в основному, або два-три-чотири розтягнутих вздовж деякої вісі хмарини, або ж по декілька хмар, які нагадують диски. Таким чином, за спектральною формою дані можна розгрупувати декількома способами: або на дві групи, або на три, або ж на чотири. В залежності від того, які геометричні відмінності цікавлять, відповідним чином можна робити групування. Було показано, що при виборі тривимірної проекції на ті головні напрямки, які містять інформацію про відмінність груп, що були знайдені, можна отримати досить непогані результати методом центроїдів. Техніка спектральної кластеризації добре застосовна на тих даних, де вже попередньо можна побачити потрібну геометрію (як-то було показано, що на початкових даних кластеризація не вхоплювала спектральну геометрію, але при переході у простір головних компонент можна було отримати задовільний результат). Зокрема, варто зазначити, спектральна кластеризація візуально спрацьовує краще у порівнянні з технікою центроїдів. Для отримання більш чітких розміток, треба додати до проекції хоча б ще декілька напрямків головних компонент, які мають інформацію про відмінностями між групами. Також було показано, що використання лише тих компонент, які зберігають найбільшу частку інформації про загальний розкид даних, не було ефективним у задачі вхоплення спектральних форм за допомогою кластеризації – ці особливості можна було зробити лише за допомогою розширення "діапазону" спостерігаємих компонент.