Лабораторна робота №5
з дисципліни
"Статистичний аналіз багатовимірних даних"
Студента 2 курсу магістратугри
групи "Статистика"
Варіант №4

Горбунов Даніел

18 грудня 2022 р.

Зміст

1 Вступ.

Дана робота присвячена прикладному застосуванню моделі гауссової суміші на модельованих даних та оцінюванню параметрів у ній за допомогу ЕМ-алгоритму. Підігнана модель використовується надалі в якості кластеризації початкових даних. Демонструються результати розбиття візуально.

2 Хід роботи.

2.1 Геометрична форма даних.

Дані подано у табличному вигляді з файлу "v4.txt", всього 5 змінних. Продемонструємо попарні діаграми розсіювання змінних на площині.

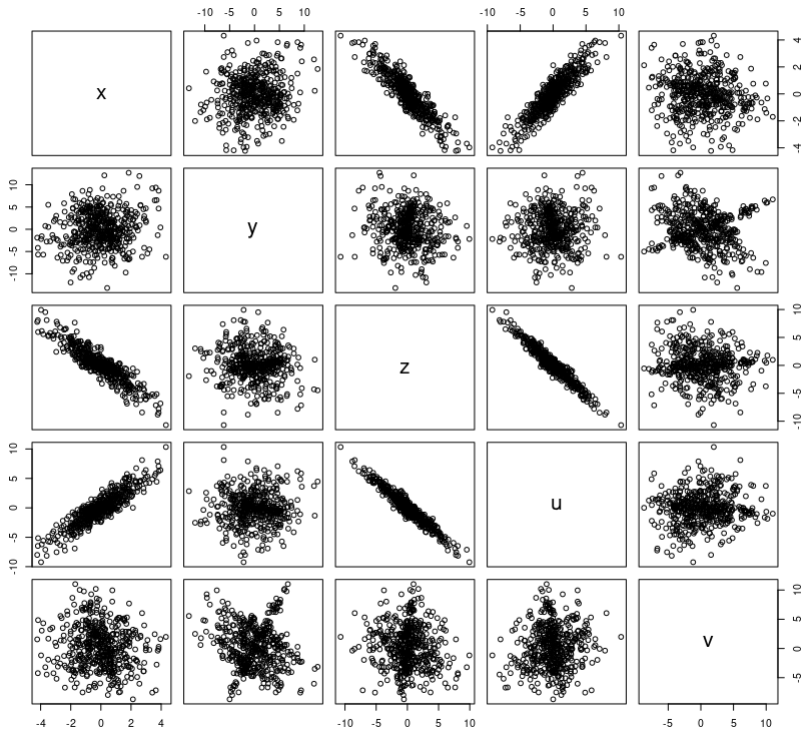

Рис. 1: Діаграми розсіювання пар змінних.

На деяких діаграмах можна помітити деякі особливості. Наприклад, якщо взяти проекцію на вісі y та v, то можна виокремити чотири сильно витягнутих еліпсоїди. На інших двовимірних проекціях з фігуруванням змінної v, неозброєним оком можна побачити дві хмарини точок, що перекриваються. На інших проекціях особливі геометричні форми не спостерігаються. Можна додатково розглянути проекції на три напрями, побудувавши просторову діаграму розсіювання. Враховуючи попереднє спостереження, побудуємо зобразимо діаграму розсіювання даних на вісі y, u, v (замість u можна було б взяти, наприклад, x або z, головне аби виділялися чотири еліпсоїди).

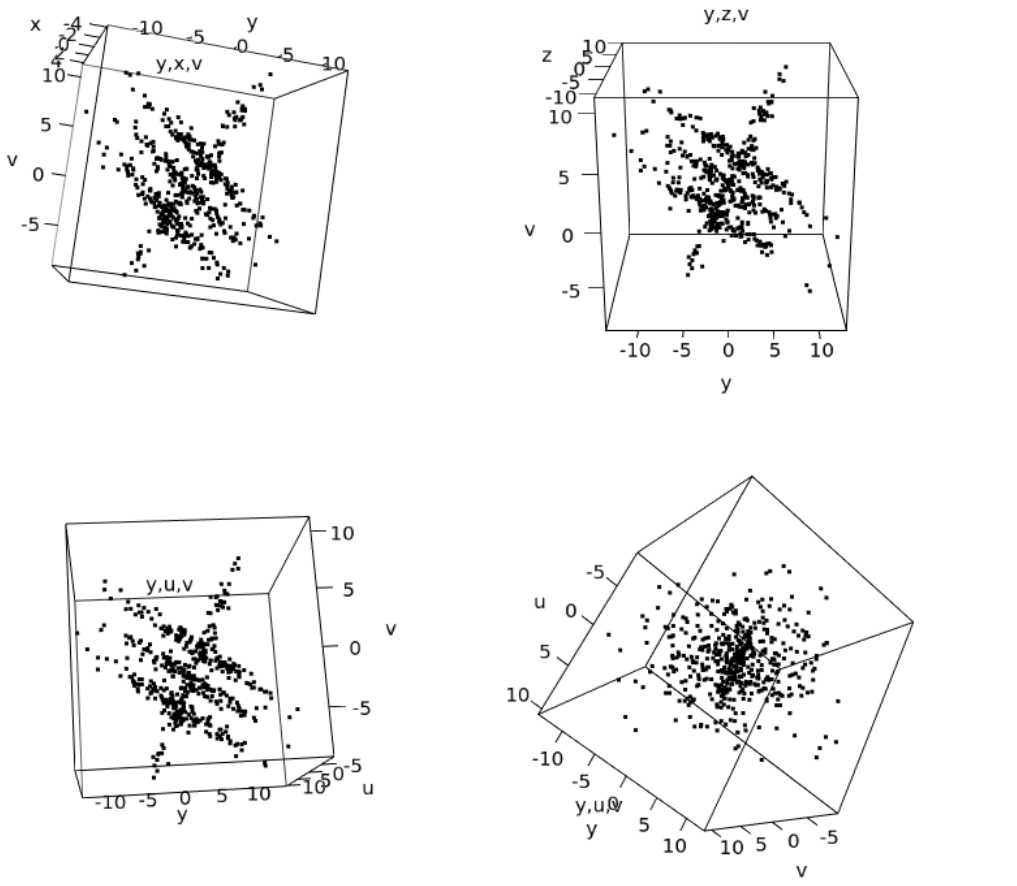

Рис. 2: Просторові діаграми розсювання з виділенням особливої геометричної форми.

Дуже добре виділяються еліпсоїди у просторі: три з них відрізняються лише центрами, але мають спільну орієнтацію, форму та об'єм (пласкі млинці); останній розташовано так, наче "перетинає" всі три хмарини водночас. Форма та положення останнього еліпсоїда відрізняється від трьох попередніх (більш сплющений та розтягнутий вздовж відповідних осей, нагадує палицю). Якщо припустити, що кожна з хмарин точок відповідає реалізаціям гауссових векторів, тоді можна спробувати підігнати модель гауссової суміші, використовуючи при цьому EM-алгоритм. Для цього скористаємося функціоналом пакета "mclust" в R.

2.2 Підгонка моделі гауссової суміші.

Підгонку параметрів у моделі суміші реалізує функція Mclust(data, ...). Можна очікувати, що оптимальним вибором буде чотири гауссових компоненти, де умови на коваріаційні матриці відсутні (алгоритм в Mclust перебирає можливі обмеження на коваріаційні матриці умовних гауссових розподілів, які накладаються на форму, об'єм та орієнтацію).

```
library('mclust')
data <- read.table("v4.txt", header=T)
data.mclust <- Mclust(data=data)
summary(data.mclust)
```

Виведемо звіт з результатів підгонки параметрів моделі.

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------
Mclust VVV (ellipsoidal, varying volume, shape, and orientation)
model with 4 components:
 log-likelihood   n df       BIC       ICL
      -4437.148 500 83 -9390.109 -9421.577
Clustering table:
  1   2   3   4
115 128 127 130
```

Перше, на що можна звернути увагу, так це кількість компонент, які алгоритм вважає "оптимальною", базуючись на максимізації байєсівського інформаційного критерія. У звіті також наведена інформація про інтегровану вірогідність за повними даними (integrated complete-data likelihood, ICL), але в рамках роботи розглядати його не будемо. Режим "VVV" означає, що найкращим вибором в сенсі максимізації інформаційного критерія є відсутність обмеженість на форми коваріаційних матриць умовних розподілів. Цікаво дослідити те, як змінювалося значення інформаційного критерія для інших режимів та кількості компонент.

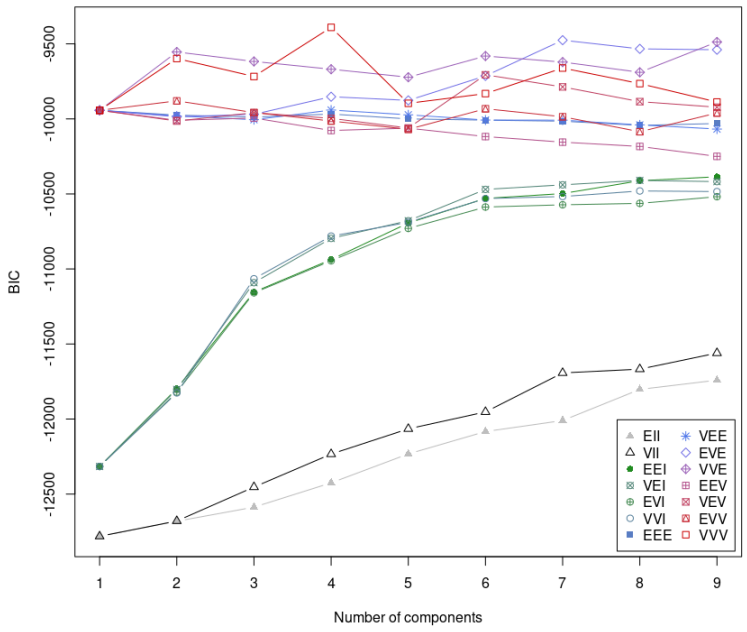

Рис. 3: Діаграма значень BIC в залежності від обмежень на коваріації та кількості компонент.

Легко бачити, як кількісно переважає вибір вищезазначених "параметрів" підгонки у порівнянні з іншими. На другому місці можна побачити кобмінування невеликої (або необгрунтовано великої) кількості компонент з режимом VVE – допускається довільність форм та об'ємів, однак робиться обмеження на орієнтацію. Обмеженна на орієнтацію здається неприроднім, враховуючи попереднє дослідження просторових діаграм, де було чітко видно, що тонкий еліпс мав зовсім іншу оірентацію у порівнянні з еліпсами-млинцями. Також локальну перемогу здобували варіанти, де робилося обмеження на об'єм та орієнтацію (останнє, знову, не добре "лягає" на справжні дані, принаймні так здається з візуальних причин).

2.3 Зменшення розмірності.

Для початку продемонструємо як алгоритм розмітив точки на одній з просторових діаграм, де було чітко видно чотири еліпсоїди.

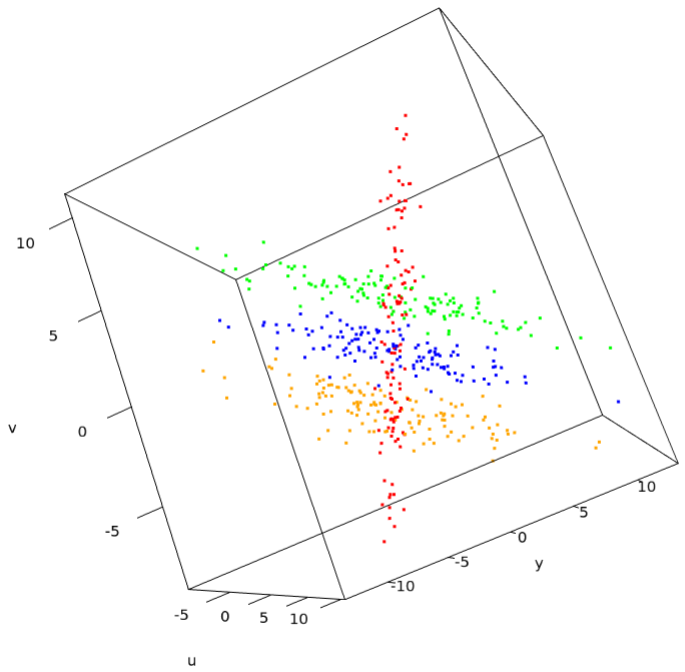

Рис. 4: Просторова діаграми розсювання з розмальовкою відповідно до компонент.

Отриманий результат співпадає з очікуваним з попередніх досліджень. Тепер постає така задача: чи можна обрати такі напрями, аби проекція даних на них максимально відображала особливості кожної з компонент суміші? Відповідь однозначна: можна, за допомогою техніки зменшення розмірності, яка базується на збільшенні розкиданості середніх та дисперсій компонент відносно загального середнього та дисперсії відповідно. Для реалізації можна скористатися функцією MclustDR(object, lambda, ...), де object – об'єкт, отриманий викликом функції Mclust, lambda – параметр, який дозволяє керувати вибором осей в залежності від того, що необхідно виокремити (наприклад, надавати більше увги розкиданості середніх, ніж дисперсій, або навпаки). Нас буде цікавити рівномірний вплив на розкиданість параметрів, тому покладемо lambda = 0,5.

```
data.mclust.dr <- MclustDR(object=data.mclust, lambda=0.5)
```

Далі наведемо графік власних чисел для критерія, що реалізовано у заданій функції, для вибору оптимальних напрямів проектування.

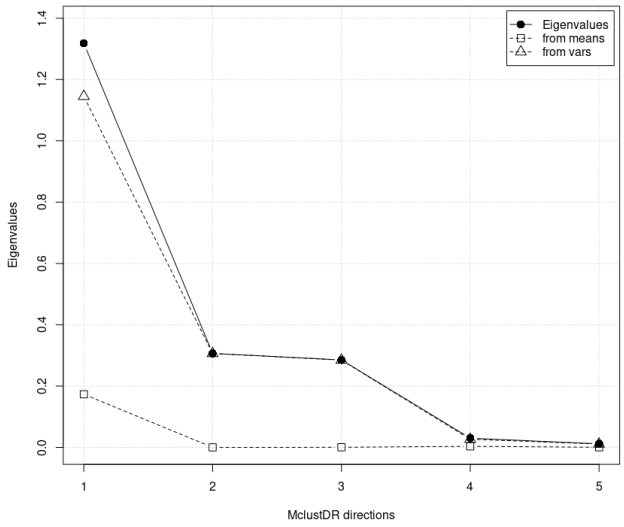

Рис. 5: Діаграма власних чисел для критерія з MclustDR.

На графіку власних чисел можна побачити, що всі відмінності між середніми містяться у першій координаті проекції. Більша частка інформації про відмінності між коваріаціями зберігаються у перших трьох координатах (зокрема у четвертій, але мало).

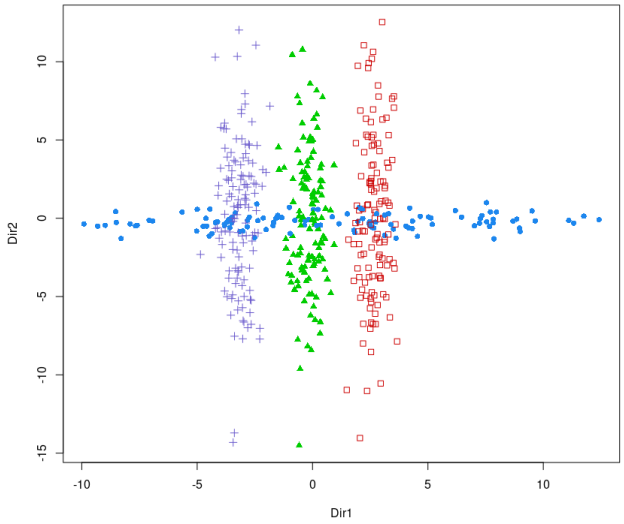

Рис. 6: Діаграма розміювання перших двох напрямів проекції.

Якщо спроектувати дані на перші два напрями, то уся відмінність між положеннями та варіацією добре відображена на відповідній діаграмі. Спостерігається дуже схожі форми, які вдалося помітити на деяких проекціях просторових діаграм (зокрема на одній з двовимірних проекцій). Далі покажемо візуалізацію на інші можливі пари напрямків:

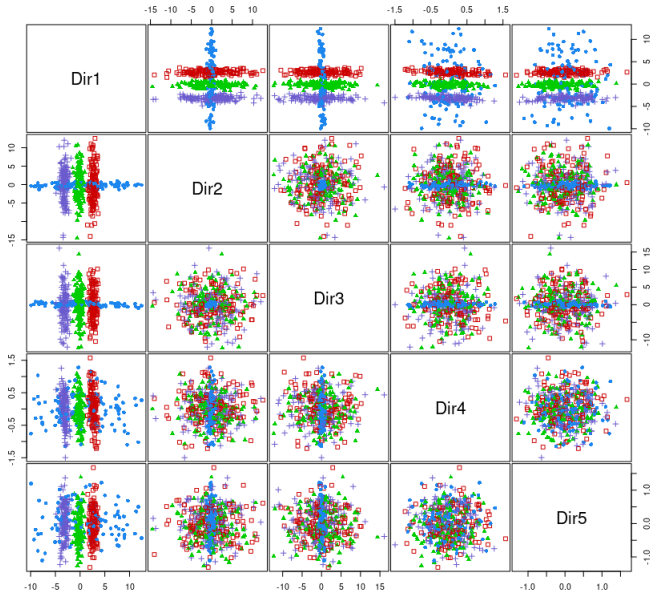

Рис. 7: Попарна діаграма розміювання напрямів проекції.

Як було зрозуміло з міркувань над графіком власних чисел, то видно, що проекції з участю першого напрямку містять інформацію про відмінності середніх. В усіх інших випадках акцентується увага безпосередньо на форми розкиду (більш-менш спільні для еліпсівмлинців, та звужений розкид для еліпса-палиці). Оскільки більшість інформації про відмінності характеристик положення та розкиду містяться у перших трьох координатах, то доцільно буде зобразити просторову проекцію даних на відповідні напрямки:

```
dr.b <- data.mclust.dr$basis
plot3d((data.matrix(data)%*%dr.b)[,1:3], col=pal[data.mclust$classification])
```

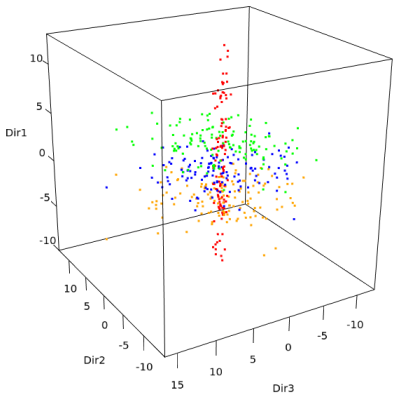

Рис. 8: Попарна діаграма розміювання напрямів проекції.

Просторова проекція нагадує проекцію на відповідні вісі початкових даних, за виключенням нахилу еліпсоїдів, що на поточній діграмі не спостерігається.

2.4 Вимірність кластерів.

З початкових досліджень є гіпотеза, що еліпси-млинці містяться у площинах, паралельних між собою, а еліпс-палиця розташований вздовж деякої прямої, що перетинає ці площини. Інакше кажучи, можна очікувати, що розмірність палиці буде рівна 1, а розмірність млинців буде дорівнювати 2. Переконаємося у цьому, дослідивши власні числа коваріаційної матриці кожної з чотирьох компонент.

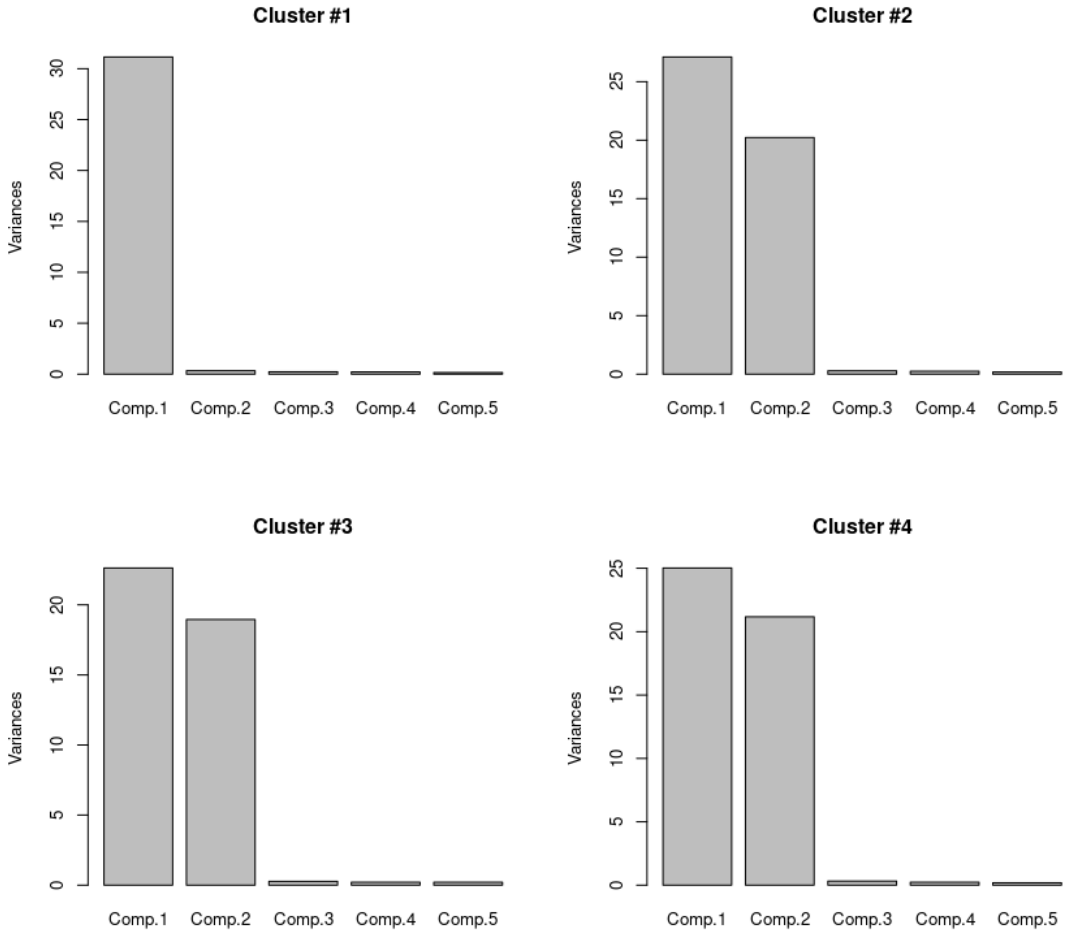

Рис. 9: Діаграми власних чисел за коваріаційною матрицею кожного кластера.

На діаграмі власних чисел за коваріаційною матрицею першого кластера можна зробити висновок, що розкид здебільшого пояснюється вздовж одного напряму. Таким чином, розмірність кластера-палиці дорівнює 1, як і очікувалося. Аналогічні міркування можна провести над іншими діаграмами власних чисел із поправкою на те, що інформація про розкид максимально поюяснюється не однією, а двома компонентами.

3 Висновки.

Підгонка моделі гауссової суміші за спостережуваними даними мало місце, оптимальна кількість компонент (кластерів) становила чотири штуки. Було визначено, що три кластери мають спільну коваріаційну матрицю, але мають різні центри (математичні сподівання). Ці кластери виявилися плоскими, розмірність дорівнювала 2. Останній кластер, що перетинає три зазначені, має більш видовжену форму (та, відповідно, іншу коваріаційну матрицю) та має розмірність 1.