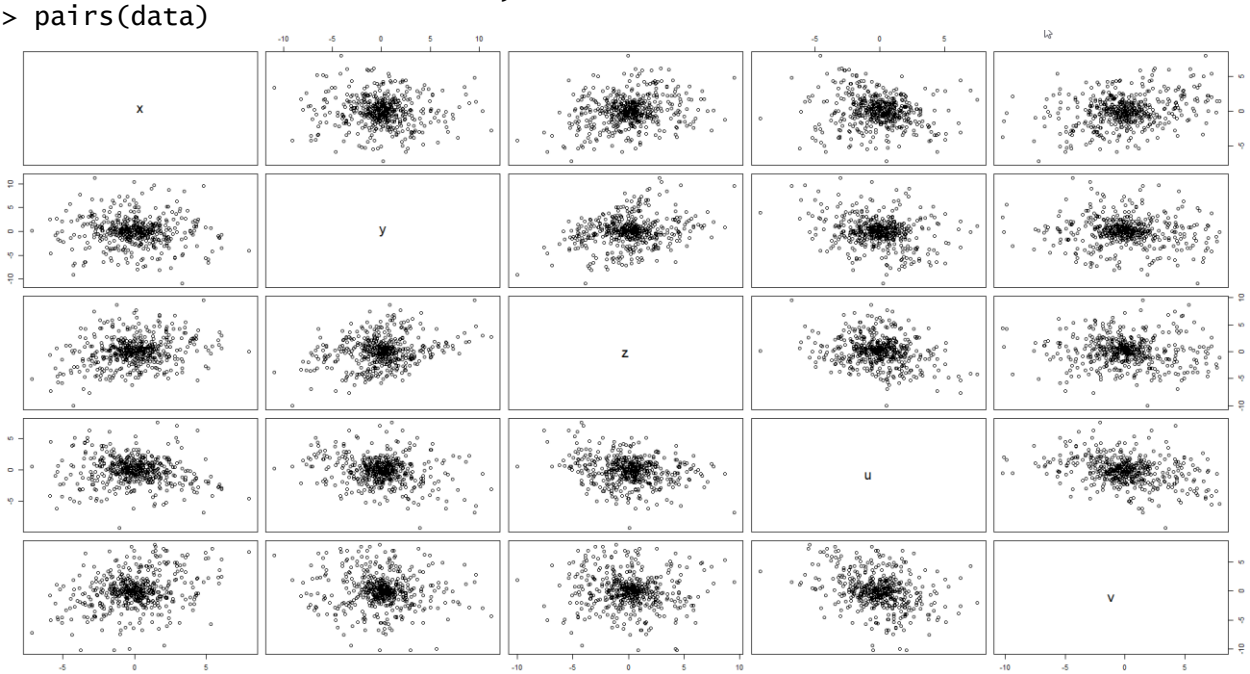
ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ №5
«КЛАСТЕРИЗАЦІЯ НА ОСНОВІ МОДЕЛІ ГАУСОВОЇ СУМІШІ»

Ломако О., 2 к. маг, «статистика», варіант 9

В даній роботі матимемо справу з даними з файлу v9.txt, в якому вміщено 500 спостережень по 5 змінним.

Спершу виведемо матричну діаграму розсіювання.

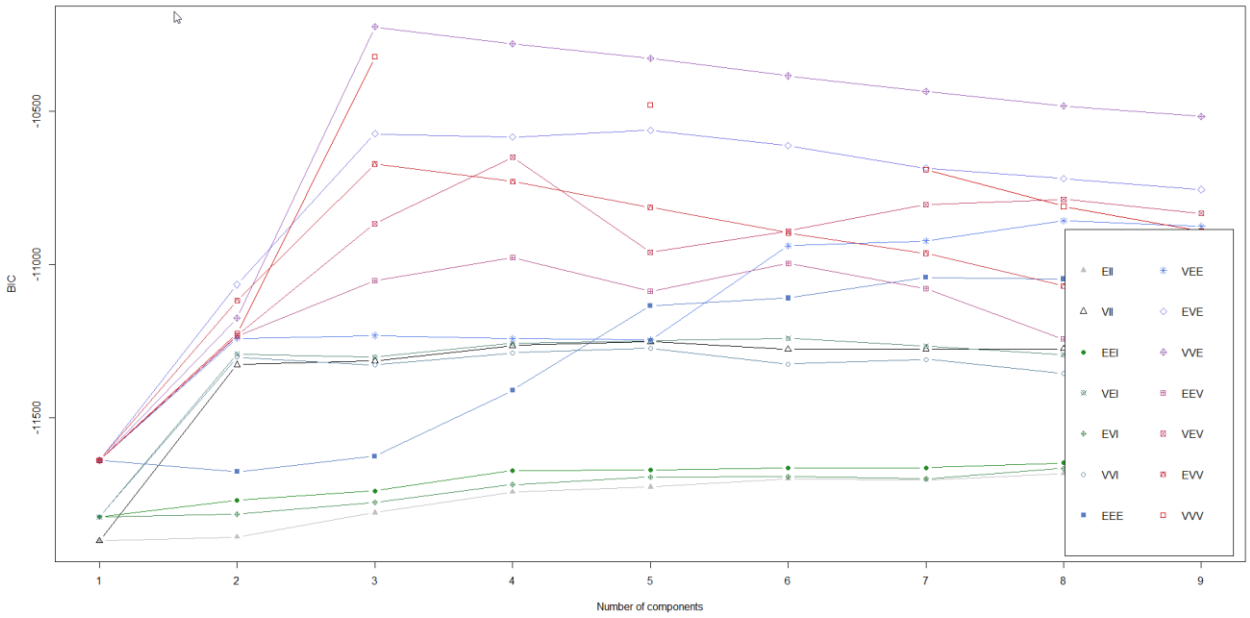
```
> library(rgl)
> library(mclust)
> data <- read.table('C:\\Users\\Razor\\Desktop\\дистанційне навчання\\статис
тичний аналіз багатовимірних даних\\lab5\\v9.txt',
+                    header = T)
> pairs(data)
```



На матричній діаграмі, чесно кажучи, я не можу виявити якої-небудь закономірності.

Проведемо кластеризацію, застосовуючи функцію *Mclust*

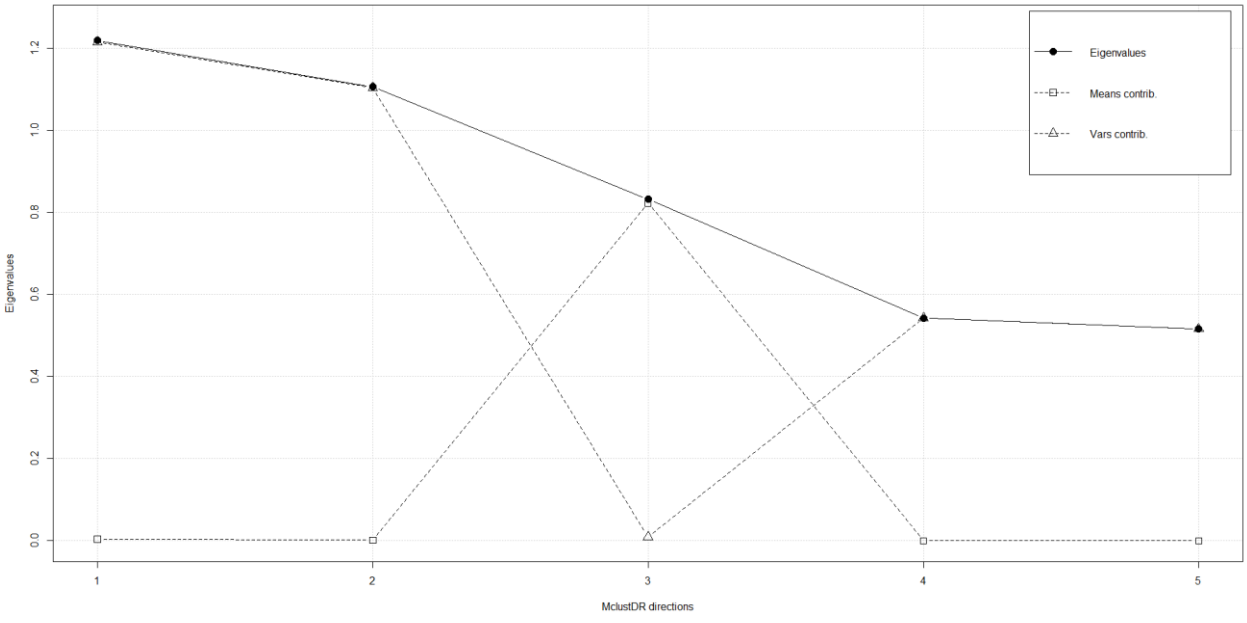
```
> mod <- Mclust(data)
fitting ...
  |=======================================================================
===============| 100%
> summary(mod)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVE (ellipsoidal, equal orientation) model with 3 components:

 log-likelihood   n df      BIC       ICL
      -4981.594 500 42 -10224.2 -10236.98

Clustering table:
  1   2   3
195 154 151
> plot(mod, what = "BIC")
```



З точки зору мінімізації BIC, найкращою виявилась модель з трьома кластерами VVE, тобто з неізотропними (еліптичними) коваріаційними матрицями однакового напряму *(distribution – ellipsoidal, volume – variable, shape – variable, orientation – equal).*

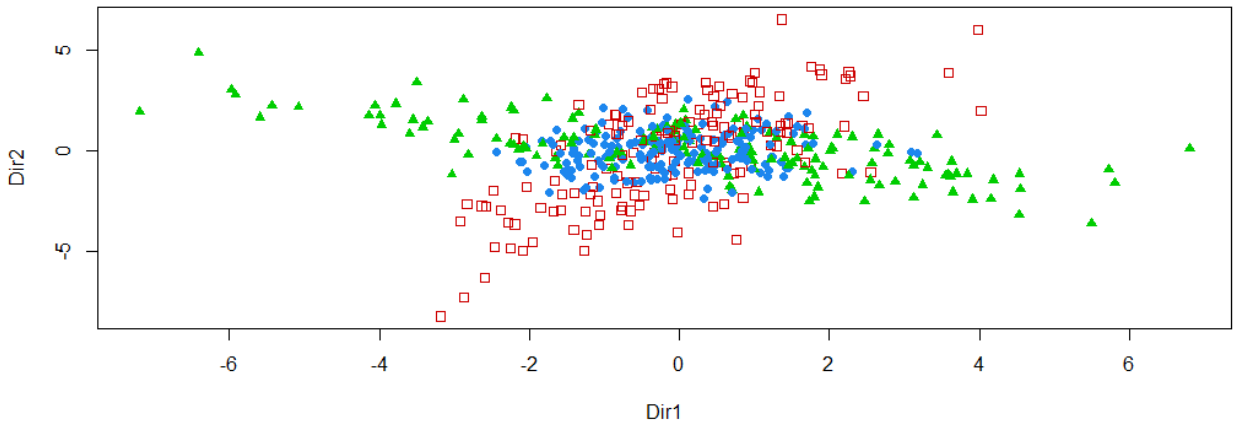
Далі розглянемо оптимальну модель, використовуючи техніку зниження розмірності за допомогою функції *MclustDR*

```
> datamod <- MclustDR(mod)
> plot(datamod, what = 'evalues')
```



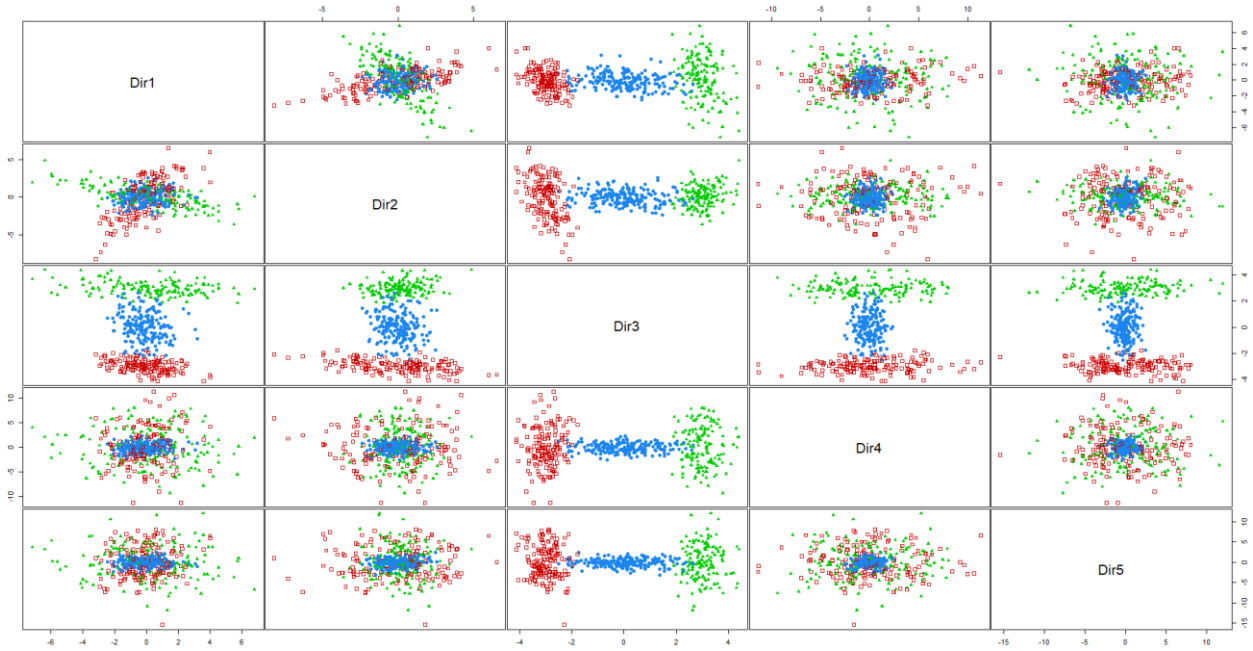
Виведемо діаграму розсіювання перших двох напрямків проекції.

```
> plot(datamod, what = 'scatterplot')
```



По такому рисунку важко зробити чіткі висновки, здається, червоні квадратики витягуються у плоску фігуру по діагоналі, сині кружечки згруповані десь всередині, і зелені трикутники розташувались певною мірою «хвилясто».

Подивимось тепер на матричну діаграму розсіювання по всім напрямкам проекції.

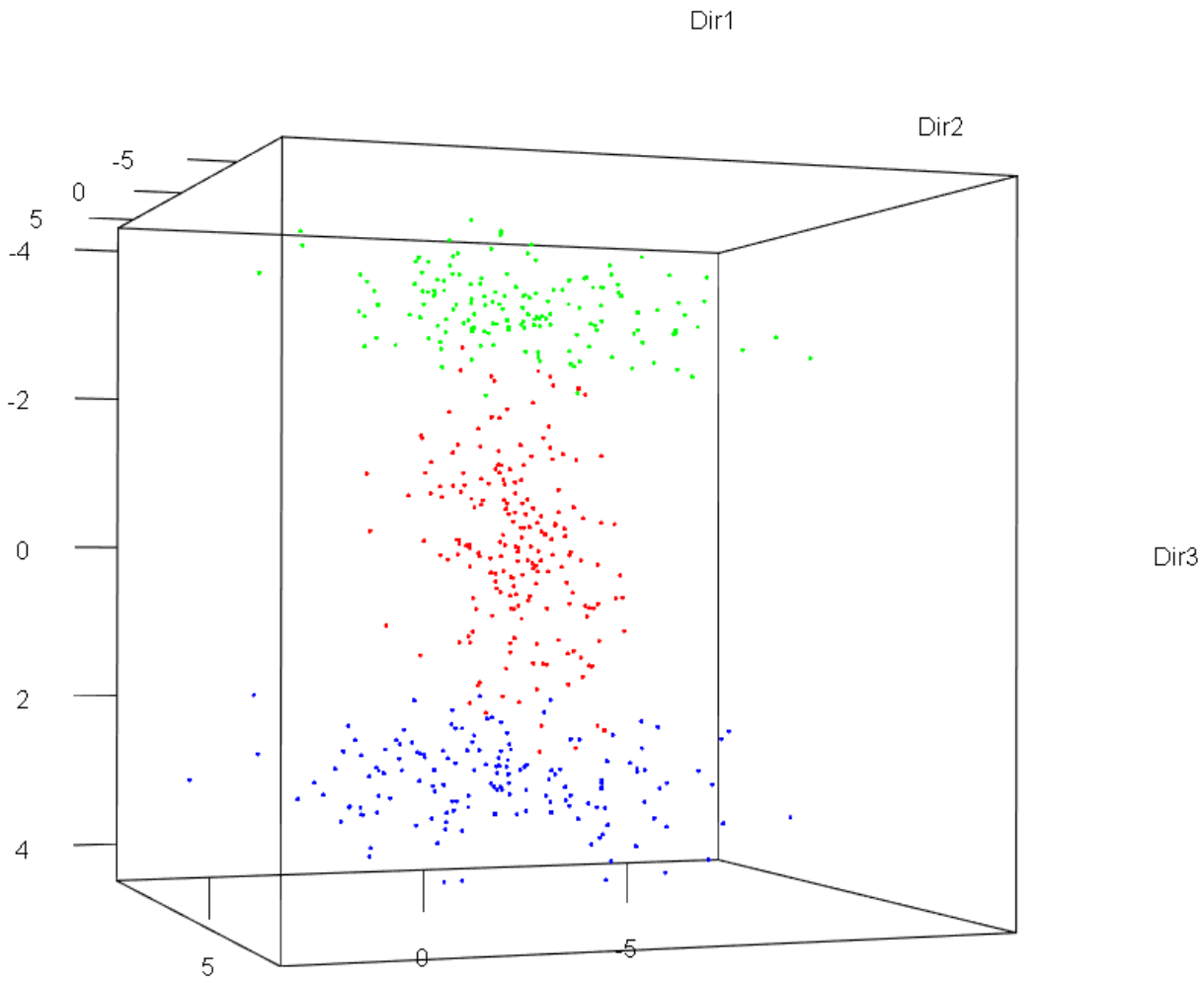


Легко бачити, як чітко розташовуються кластери по третьому напрямку: червоні і зелені купки виступають «підлогою» і «стелею», а синя – «колоною» між ними.

Спробуємо вивести на тривимірному рисунку три які-небудь напрямки. З матричної діаграми розсіювання, вочевидь, в розгляд варто включити третій напрямок.

Розглянемо спершу 1-3 напрямки.

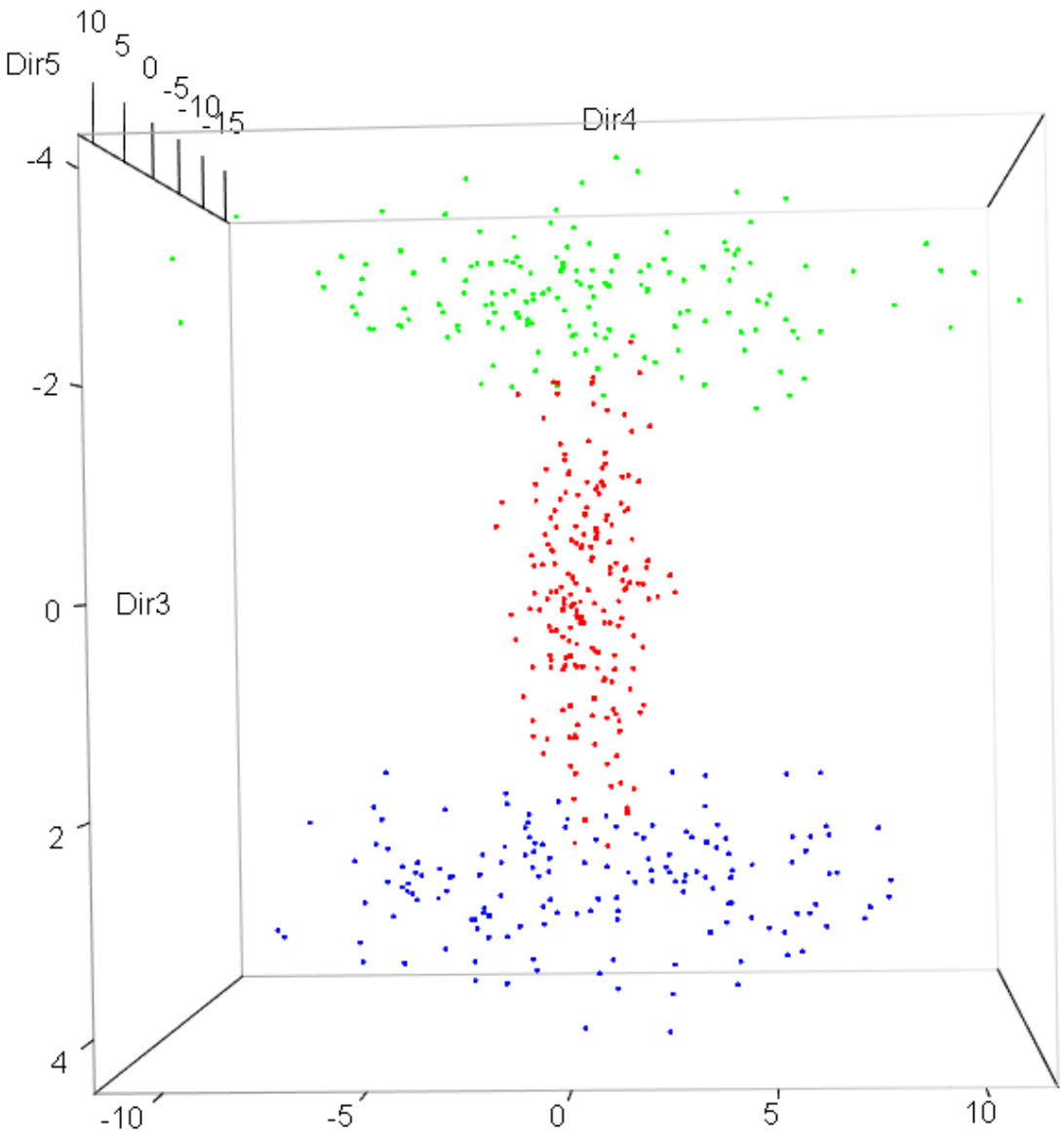
```
> plot3d((as.matrix(data) %*% datamod$basis)[,1:3],
+        col = c('red', 'green', 'blue')[mod$classification])
```



На тривимірній діаграмі наочно підтверджуються наші припущення про кластеризацію.

Спробуємо вивести 3-5 напрямки.

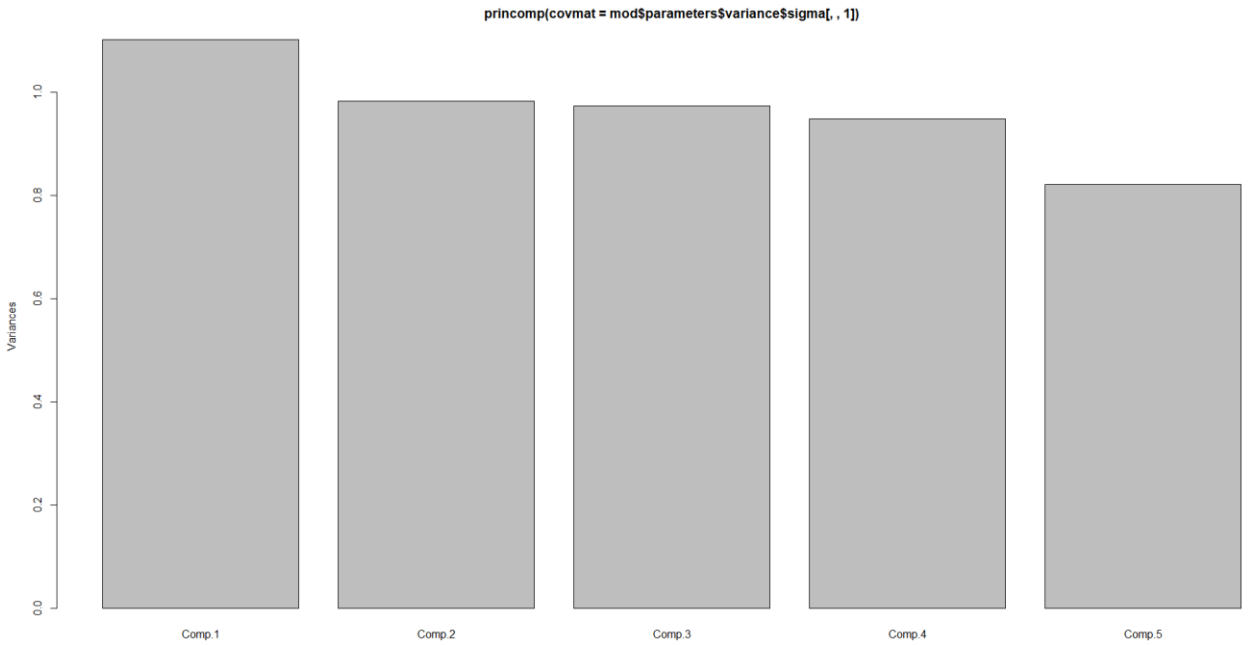
```
> plot3d((as.matrix(data) %*% datamod$basis)[,3:5],
+        col = c('red', 'green', 'blue')[mod$classification])
```



Майже не аналогічний рисунок. Але тут, помітимо, сині і зелені площини розташовуються в певній мірі навіть паралельно, що для кластеризації, на мою думку, виглядає ще краще.

Тепер подивимось на вимірність кластерів. Почнемо з першого.

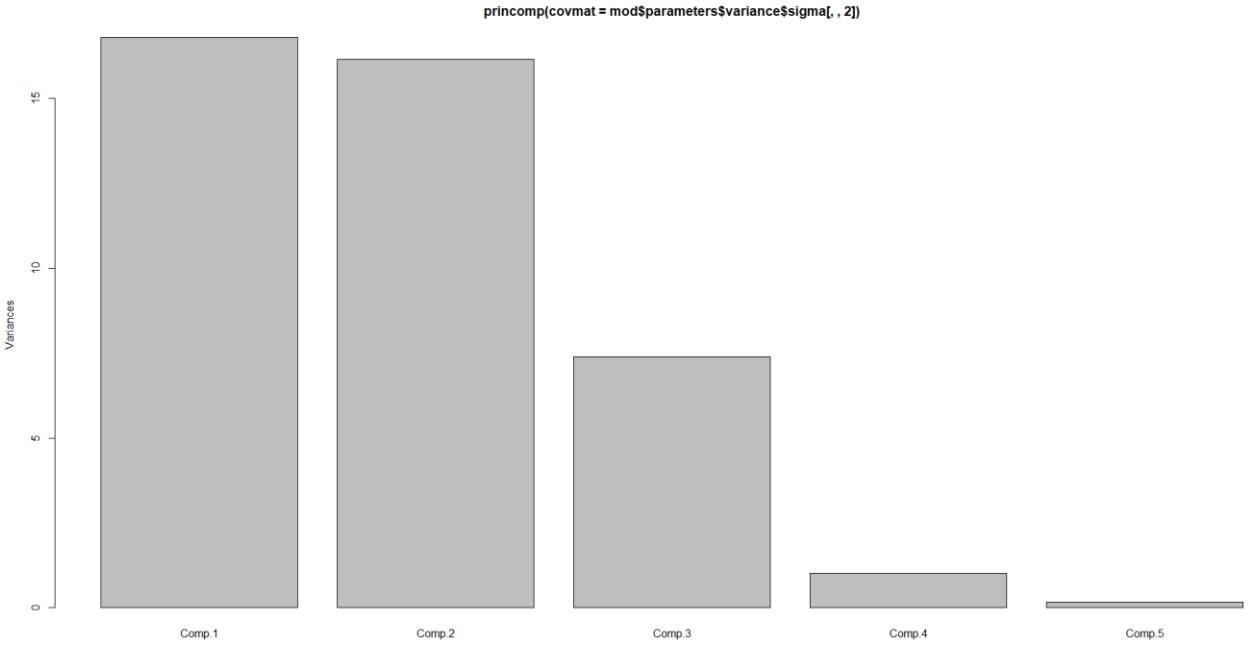
```
> plot(princomp(covmat = mod$parameters$variance$sigma[,,1]))
```



Можемо бачити, що коваріаційна матриця першої компоненти має всі п'ять власних чисел, що не сильно відрізняються за значенням (немає значних провалів). Тому є підстави стверджувати, що справжня вимірність простору, у якому розташована перша компонента, становить 5.

Проробимо те ж саме для другої компоненти.

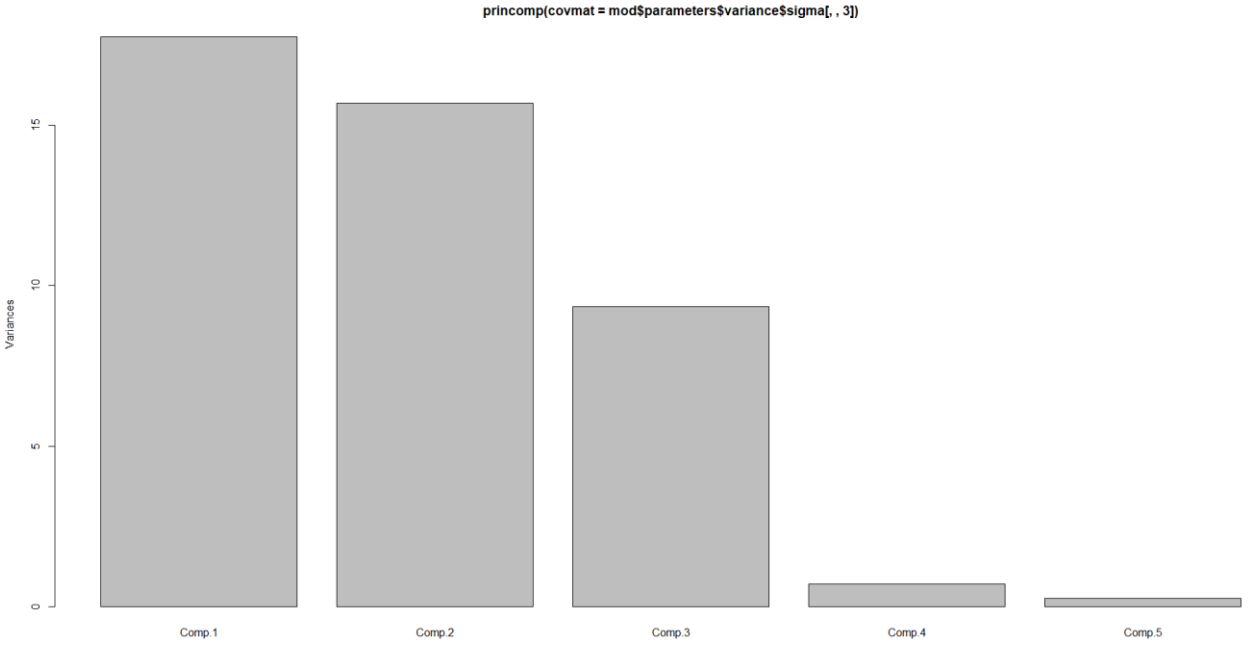
```
> plot(princomp(covmat = mod$parameters$variance$sigma[,,2]))
```



Тут вже картина дещо інакша: перші два власних числа є приблизно однаковими, в той час як третє вже точно вдвічі менше за друге, а четверте і взагалі наближається до нуля. Тому тут дещо важче однозначно вказати якою буде вимірність простору, у якому розташована друга компонента. На мою думку, це скоріше 3, аніж 2, але і це можливо.

І наостанок для третьої компоненти.

```
> plot(princomp(covmat = mod$parameters$variance$sigma[,,3]))
```



Схожа ситуація з минулим випадком, але тут третє число вже не так сильно «просідає» відносно другого, тому тут маю більшу впевненість, що вимірність цього простору буде 3.