Кластерный анализ

Анализ и визуализация многомерных данных с использованием R

Марина Варфоломеева, Вадим Хайтов

Кластерный анализ

- Методы построения деревьев
- Методы кластеризации на основании расстояний
- Примеры для демонстрации и для заданий
- Кластерный анализ в R
- Качество кластеризации:
  - кофенетическая корреляция
  - количество кластеров
  - поддержка ветвей
- Сопоставление деревьев: танглграммы

Вы сможете

- Выбирать подходящий метод аггрегации (алгоритм кластеризации)
- Строить дендрограммы
- Оценивать качество кластеризации (Кофенетическая корреляция, поддержка ветвей)
- Рассчитывать оптимальное число кластеров
- Сопоставлять дендрограммы, полученные разными способами, при помощи танглграмм

Кластерный анализ

Какие бывают методы построения деревьев?

Методы класстеризации на основании расстояний (о них сегодня)

- Метод ближайшего соседа
- Метод отдаленного соседа
- Метод среднегруппового расстояния
- Метод Варда
- и т.д. и т.п.

Методы кластеризации на основании признаков

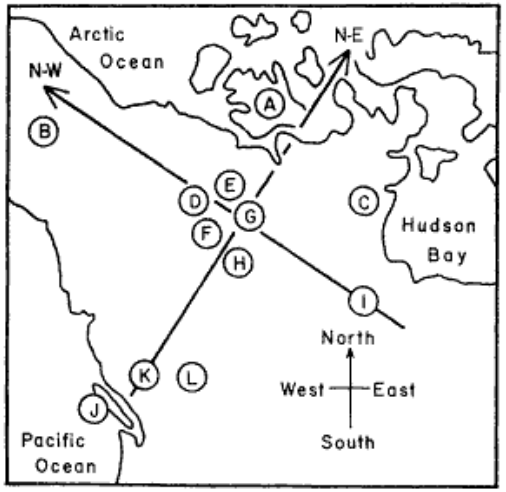
- Метод максимальной бережливости
- Метод максимального правдоподобия

И это еще далеко не все

Примеры

Пример: Волки

Морфометрия черепов у волков в Скалистых горах и в Арктике (Jolicoeur, 1959)



Данные взяты из работы Morrison (1990):

- A — волки из Арктики (10 самцов, 6 самок)
- L — волки из Скалистых гор (6 самцов, 3 самки)

```
library(candisc)
data("Wolves")
```

Знакомимся с данными

```
dim(Wolves)
```

```
## [1] 25 12
```

```
colnames(Wolves)
```

```
##  [1] "group"    "location" "sex"      "x1"       "x2"       "x3"
##  [7] "x4"       "x5"       "x6"       "x7"       "x8"       "x9"
```

```
head(rownames(Wolves))
```

```
## [1] "rmm1" "rmm2" "rmm3" "rmm4" "rmm5" "rmm6"
```

```
any(is.na(Wolves))
```

```
## [1] FALSE
```

```
table(Wolves$group)
```

```
##
## ar:f ar:m rm:f rm:m
##    6   10    3    6
```

Пример: Ирисы

```
data("iris")
```

Знакомимся с данными

```
dim(iris)
```

```
## [1] 150   5
```

```
colnames(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"
```

```
head(rownames(iris))
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

```
# Делаем осмысленные имена строк
Species <- substr(iris$Species, 0, 2)
rownames(iris) <- make.unique(Species)
# Делаем случайную выборку для этой демонстрации
set.seed(191231)
ids <- sample(nrow(iris), 50)
siris <- iris[ids, ]
```

Задание:

- Постройте ординацию nMDS данных о морфометрии волков и ирисов
- Оцените качество ординации
- Обоснуйте выбор коэффициента
- Раскрасьте точки на ординации волков в зависимости от географического происхождения (`group`), а на ординации ирисов — от вида (`Species`)

Решение: Волки

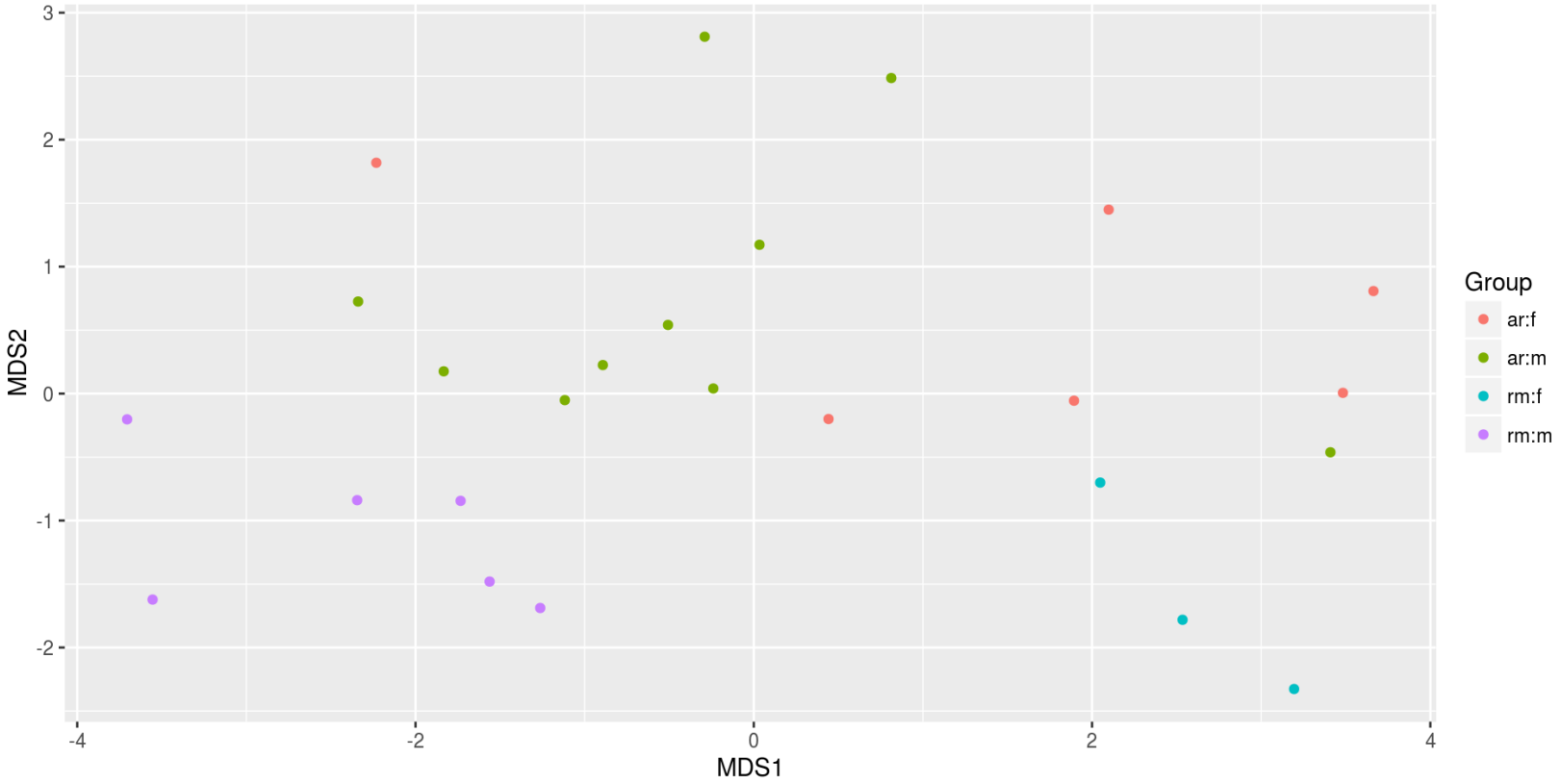
```
library(vegan)
library(ggplot2)
s_w <- scale(Wolves[, 4:ncol(Wolves)]) ## стандартизируем
ord_w <- metaMDS(comm = s_w, distance = "euclidean", autotransform = FALSE)
```

```
## Warning in metaMDS(comm = s_w, distance = "euclidean", autotransform
## = FALSE): 'comm' has negative data: 'autotransform', 'noshare' and
## 'wascores' set to FALSE
```

```
## Run 0 stress 0.101
## Run 1 stress 0.143
## Run 2 stress 0.101
## ... procrustes: rmse 0.0000114  max resid 0.0000343
## *** Solution reached
```

```
dfr_w <- data.frame(ord_w$points, Group = Wolves$group)
gg_w <- ggplot(dfr_w, aes(x = MDS1, y = MDS2)) + geom_point(aes(colour = Group))
```

Решение: Волки



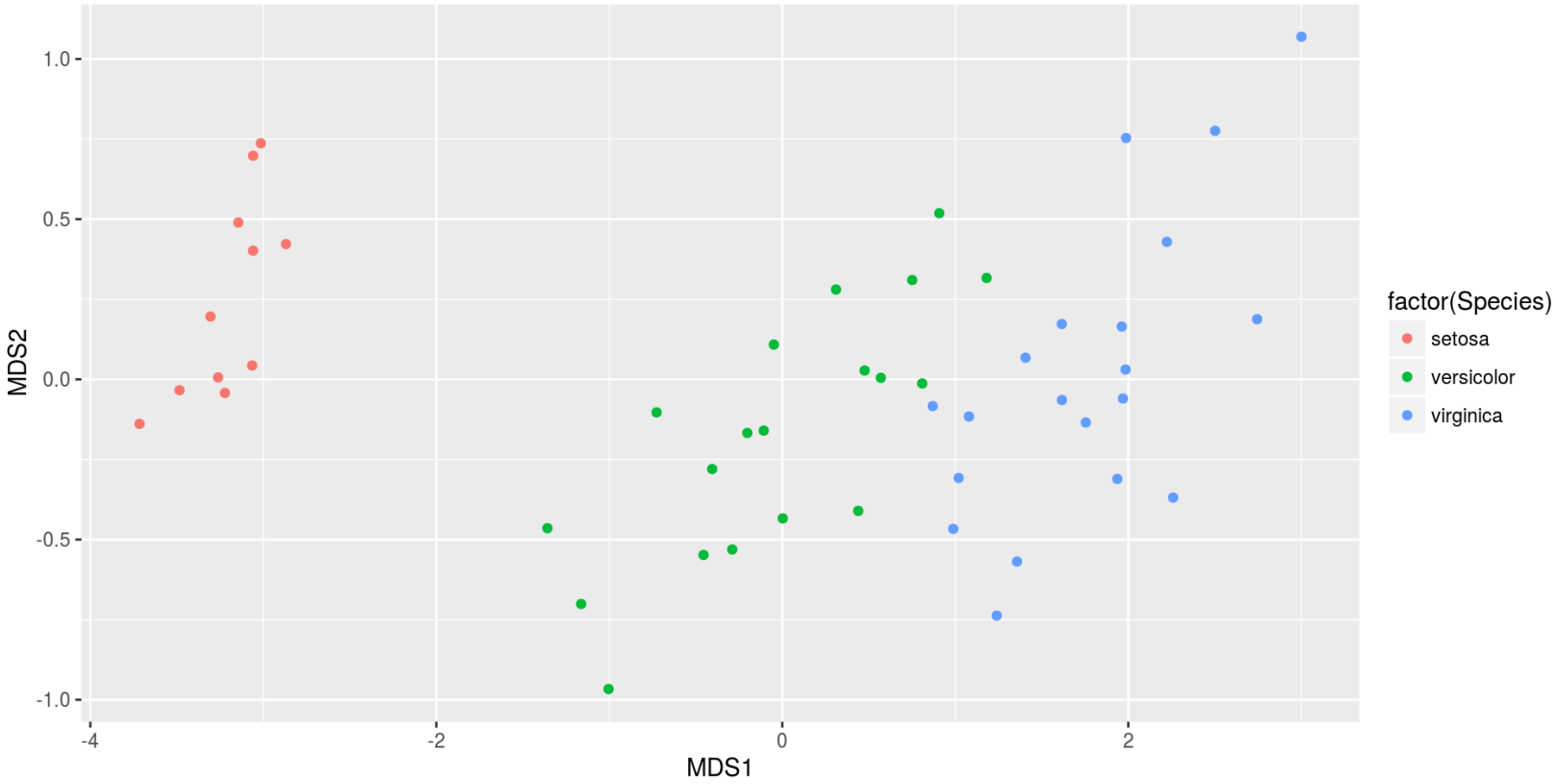
Решение: Ирисы

```
ord_i <- metaMDS(comm = siris[, -5], distance = "euclidean", autotransform = FALSE)
```

```
## Run 0 stress 0.0272
## Run 1 stress 0.0345
## Run 2 stress 0.0315
## Run 3 stress 0.0314
## Run 4 stress 0.0321
## Run 5 stress 0.0356
## Run 6 stress 0.039
## Run 7 stress 0.0314
## Run 8 stress 0.0272
## ... New best solution
## ... procrustes: rmse 0.000195  max resid 0.000609
## *** Solution reached
```

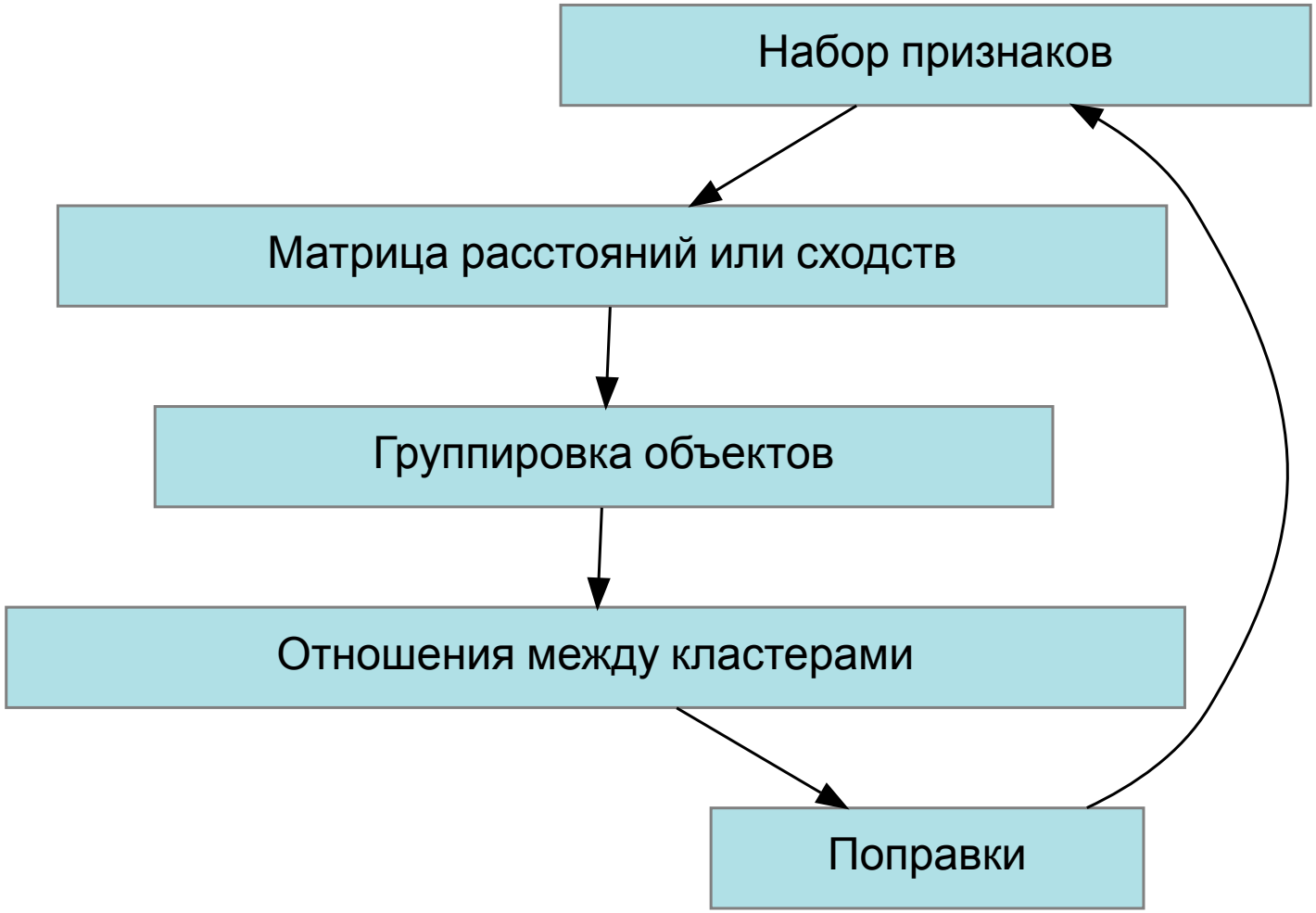
```
dfr_i <- data.frame(ord_i$points, Species = siris$Species)
gg_i <- ggplot(dfr_i, aes(x = MDS1, y = MDS2)) + geom_point(aes(colour = factor(Species)))
```

Решение: Ирисы



Методы класстеризации на основании расстояний

Этапы кластеризации



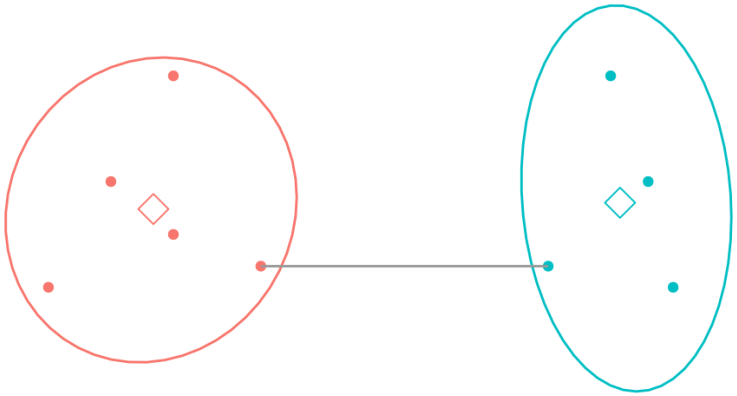
От чего зависит результат кластеризации

Результат кластеризации зависит от

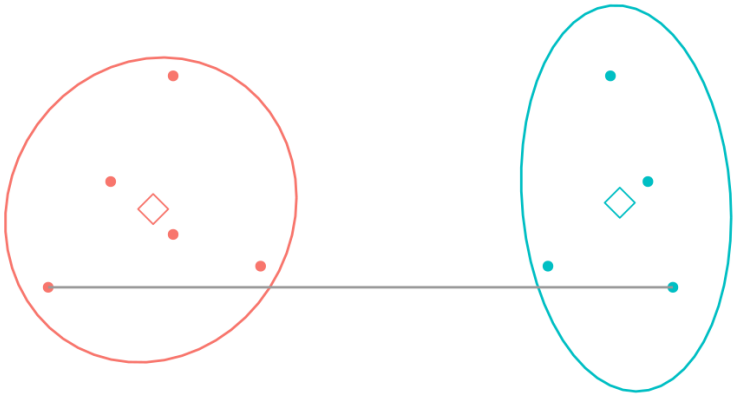
- коэффициента сходства-различия
- от алгоритма кластеризации

Методы кластеризации

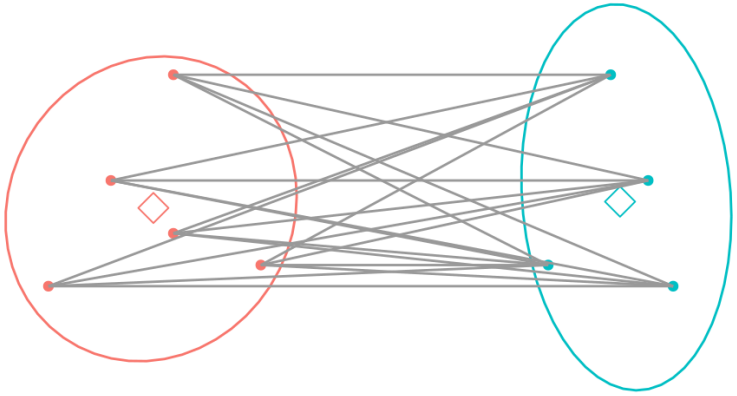
Метод ближайшего соседа



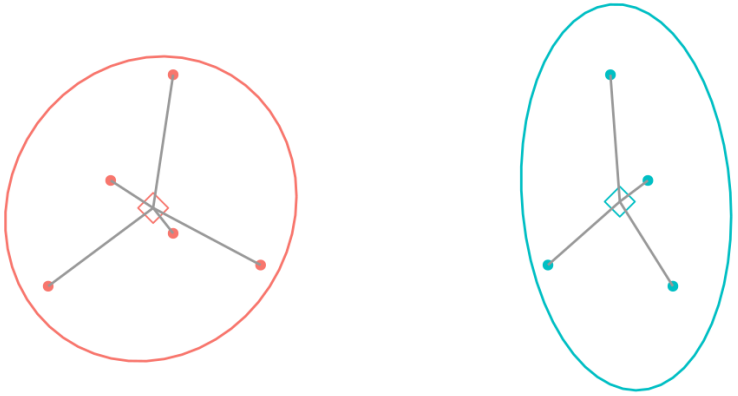
Метод отдаленного соседа



Метод среднегруппового расстояния

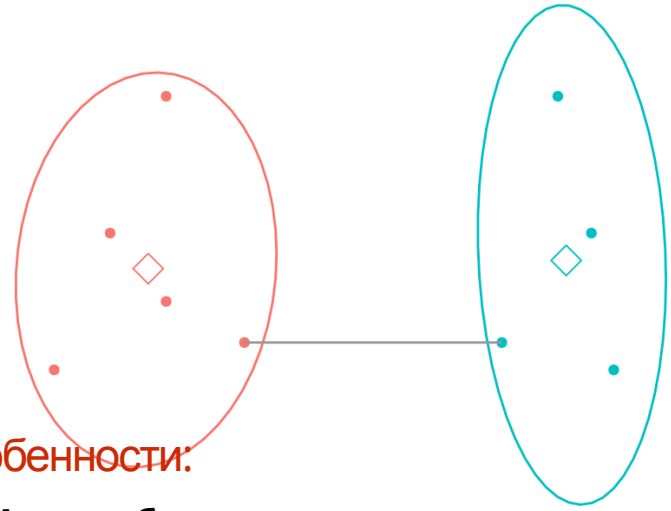


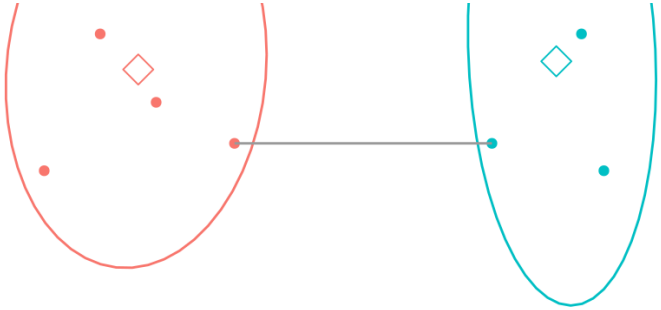
Метод Варда



Метод ближайшего соседа

= nearest neighbour = single linkage



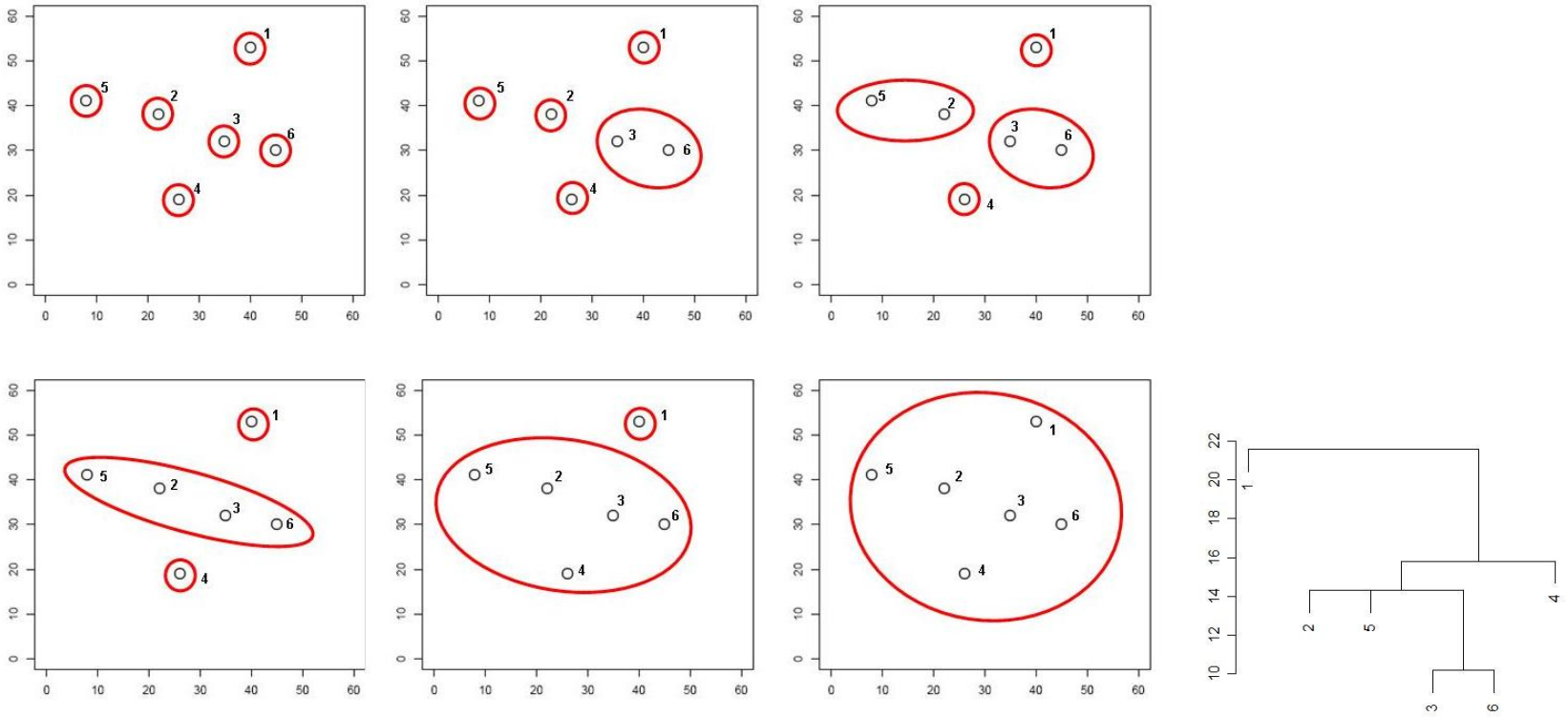


- к кластеру присоединяется ближайший к нему кластер/объект
- кластеры объединяются в один на расстоянии, которое равно расстоянию между ближайшими объектами этих кластеров

Особенности:

- Может быть сложно интерпретировать, если нужны группы
- объекты на дендрограмме часто не образуют четко разделенных групп
- часто получаются цепочки кластеров (объекты присоединяются как бы по-одному)
- Хорош для выявления градиентов

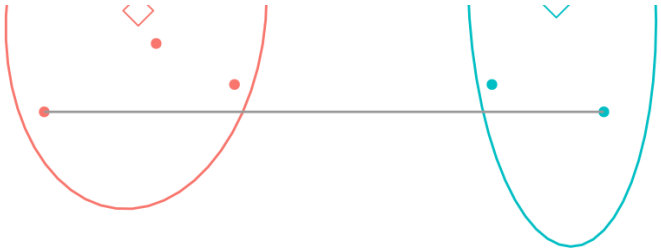
Как работает метод ближайшего соседа

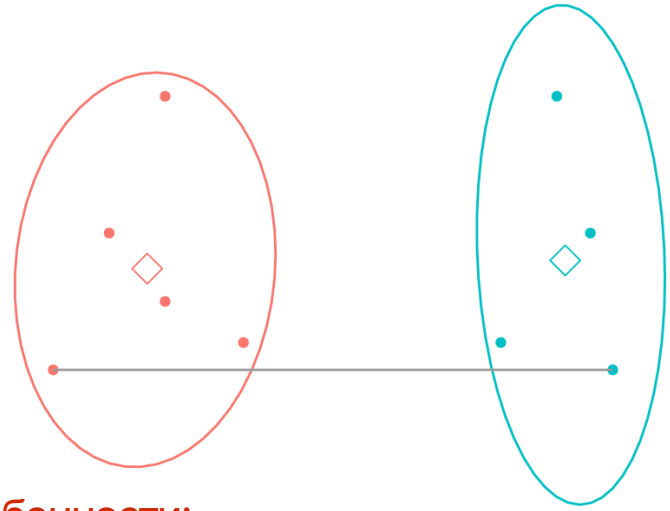


http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/Hierarchisches_Clustern_Beispiel.php

Метод отдаленного соседа

= furthest neighbour = complete linkage



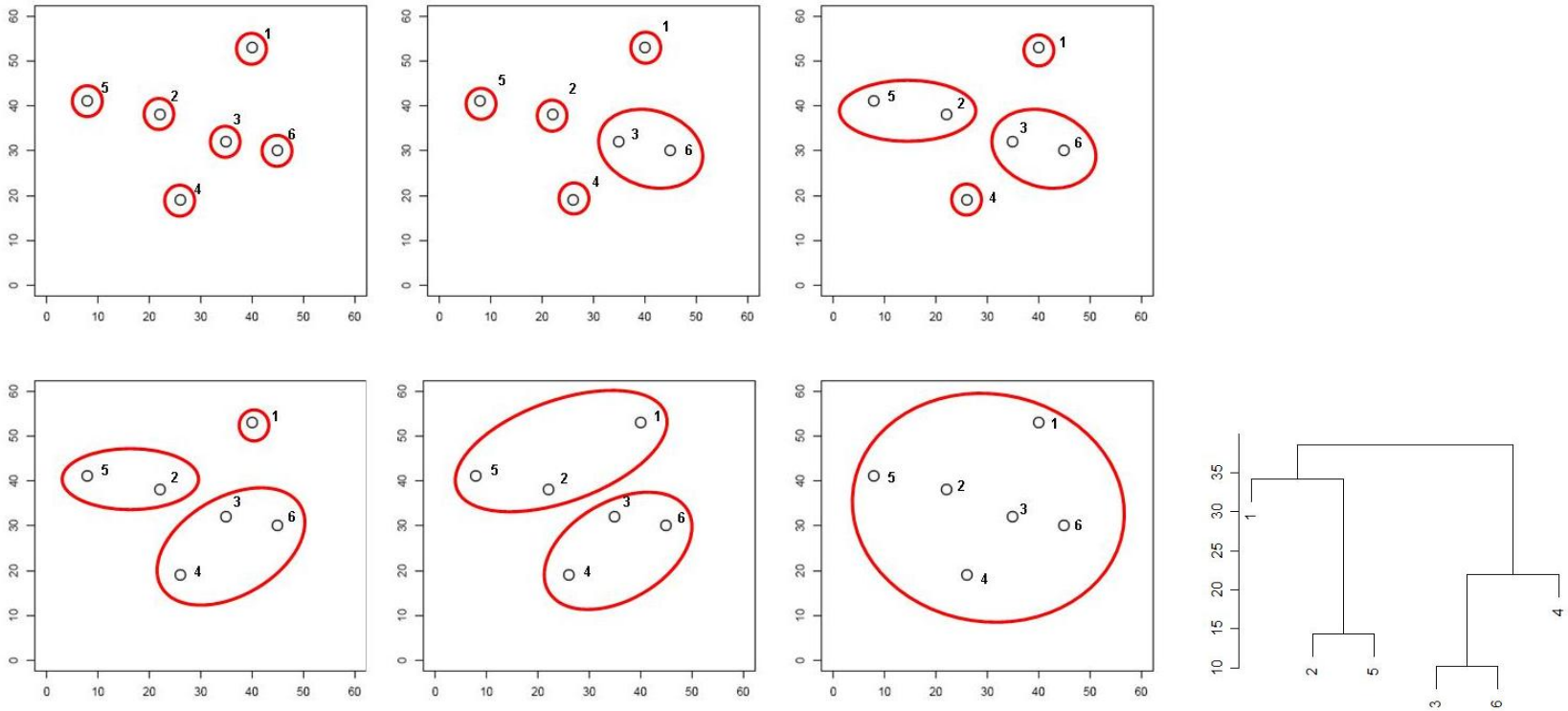


- к кластеру присоединяется отдаленный кластер/объект
- кластеры объединяются в один на расстоянии, которое равно расстоянию между самыми отдаленными объектами этих кластеров (следствие - чем более крупная группа, тем сложнее к ней присоединиться)

Особенности:

- На дендрограмме образуется много отдельных некрупных групп
- Хорош для поиска дискретных групп в данных

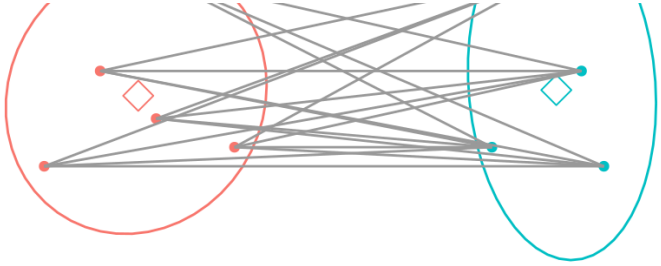
Как работает метод отдаленного соседа

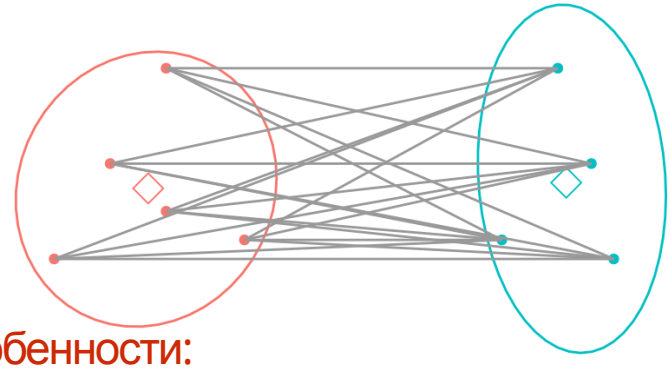


http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/Hierarchisches_Clustern_Beispiel.php

Метод невзвешенного попарно

= UPGMA = Unweighted Pair Group Method wit

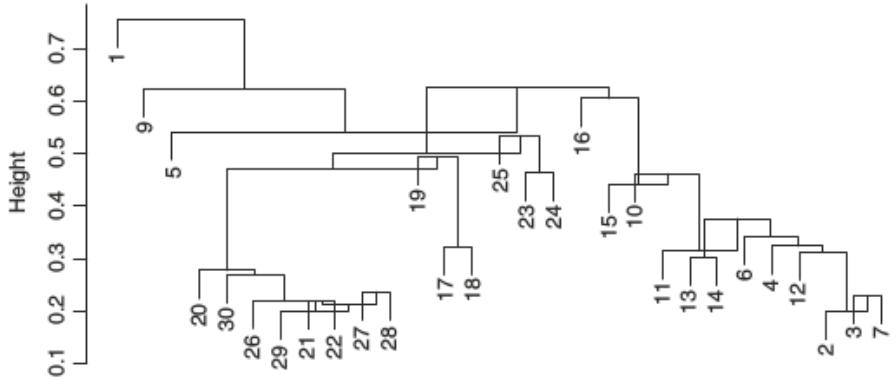




- кластеры объединяются в один на расстоянии, которое равно среднему значению всех возможных расстояний между объектами из разных кластеров.

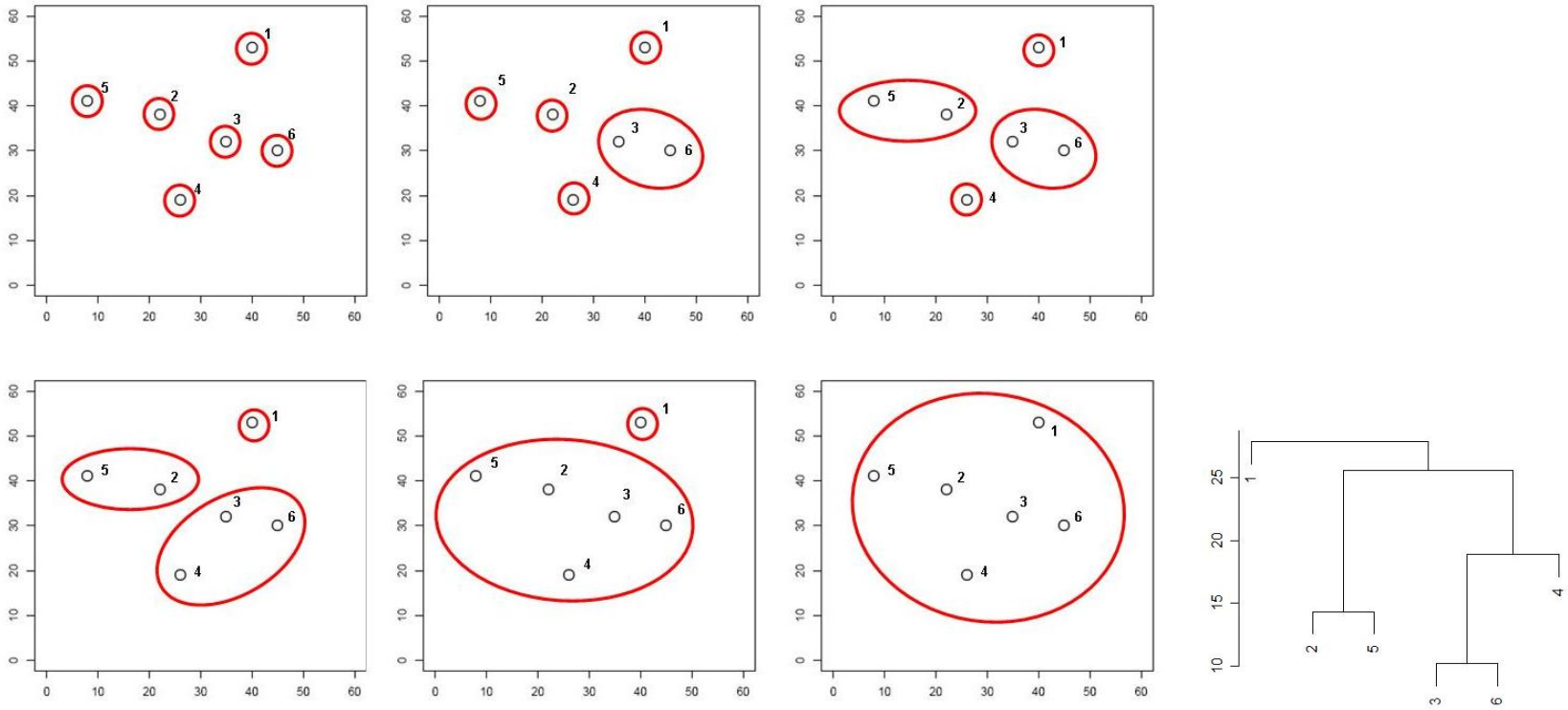
Особенности:

- UPGMA и WUPGMC иногда могут приводить к инверсиям на дендрограммах



из Borcard et al., 2011

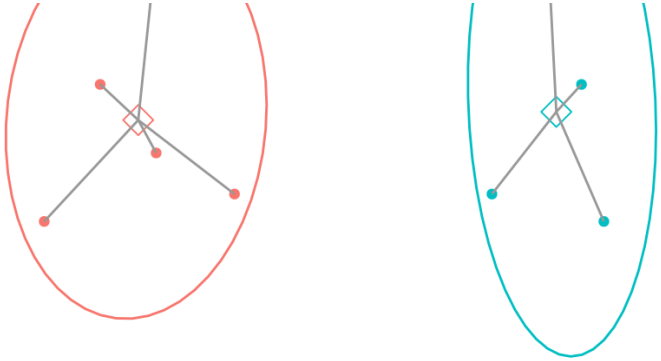
Как работает метод среднегруппового расстояния

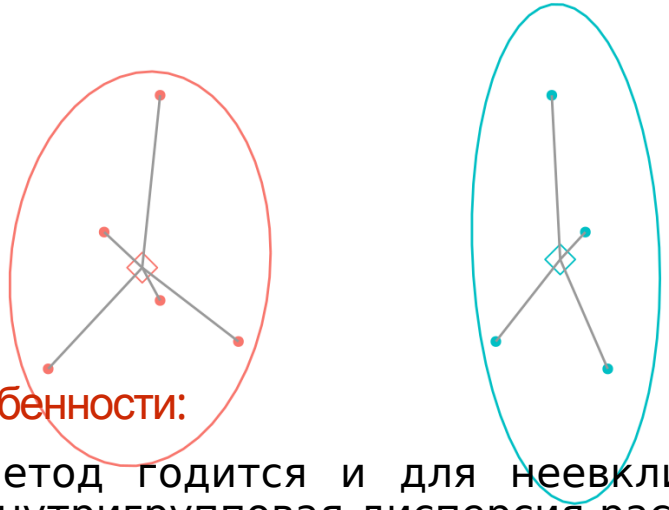


http://www-m9.ma.tum.de/material/felix-klein/clustering/Methoden/Hierarchisches_Clustern_Beispiel.php

Метод Варда

= Ward's Minimum Variance Clustering





- объекты объединяются в кластеры так, чтобы внутригрупповая дисперсия расстояний была минимальной

Особенности:

- метод годится и для неевклидовых расстояний несмотря на то, что внутригрупповая дисперсия расстояний рассчитывается так, как будто это евклидовы расстояния

Кластерный анализ в R

Кластеризация

Давайте построим деревья при помощи нескольких алгоритмов кластеризации (по стандартизованным данным, с использованием Евклидова расстояния) и сравним их.

```
# Нам понадобится матрица расстояний
d <- dist(x = s_w, method = "euclidean")
# Пакеты для визуализации кластеризации
library(ape)
library(dendextend)
```

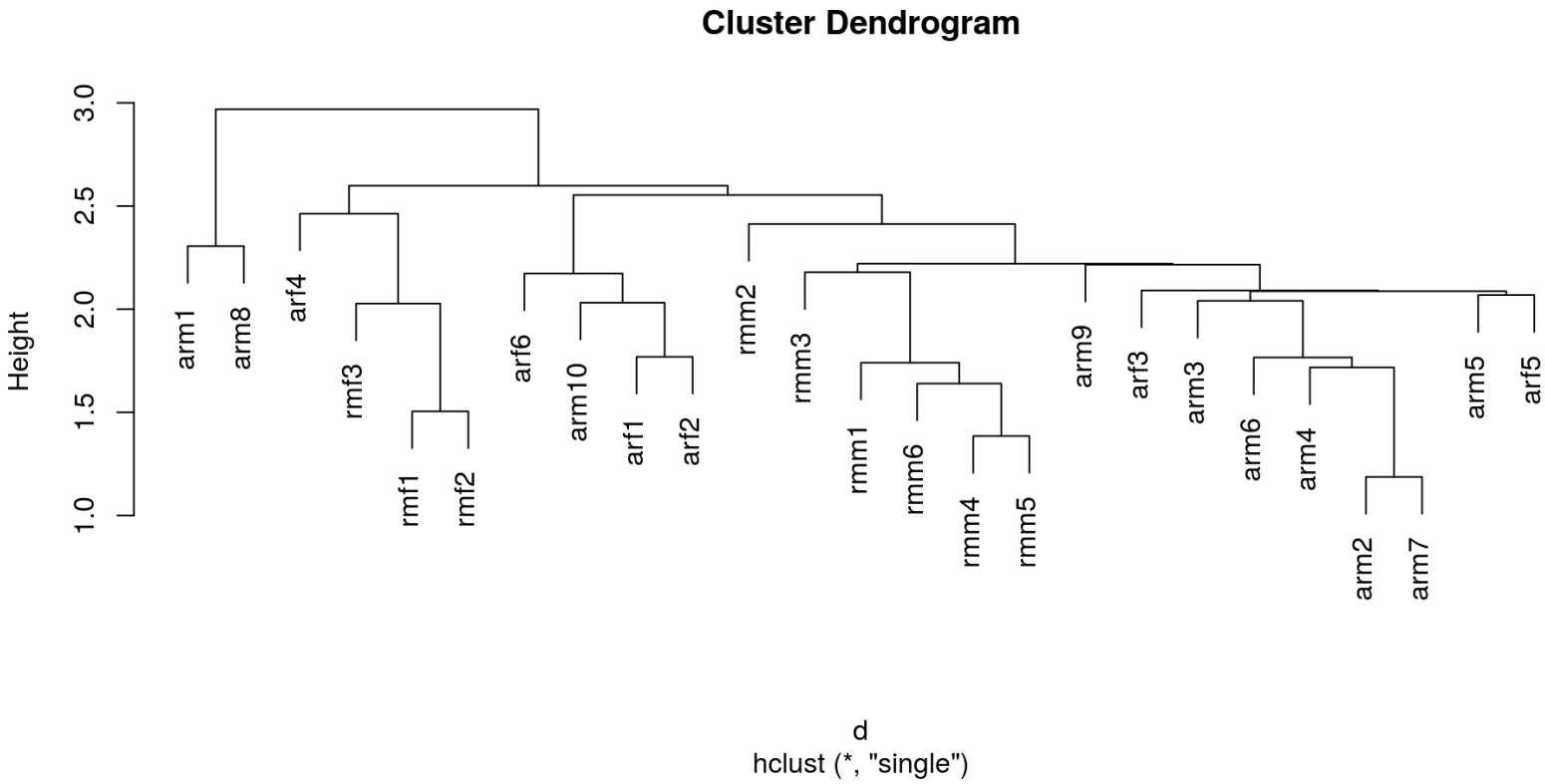
Метод ближайшего соседа в R

```
hc_single <- hclust(d, method = "single")
```

- И это все?
- Нет!

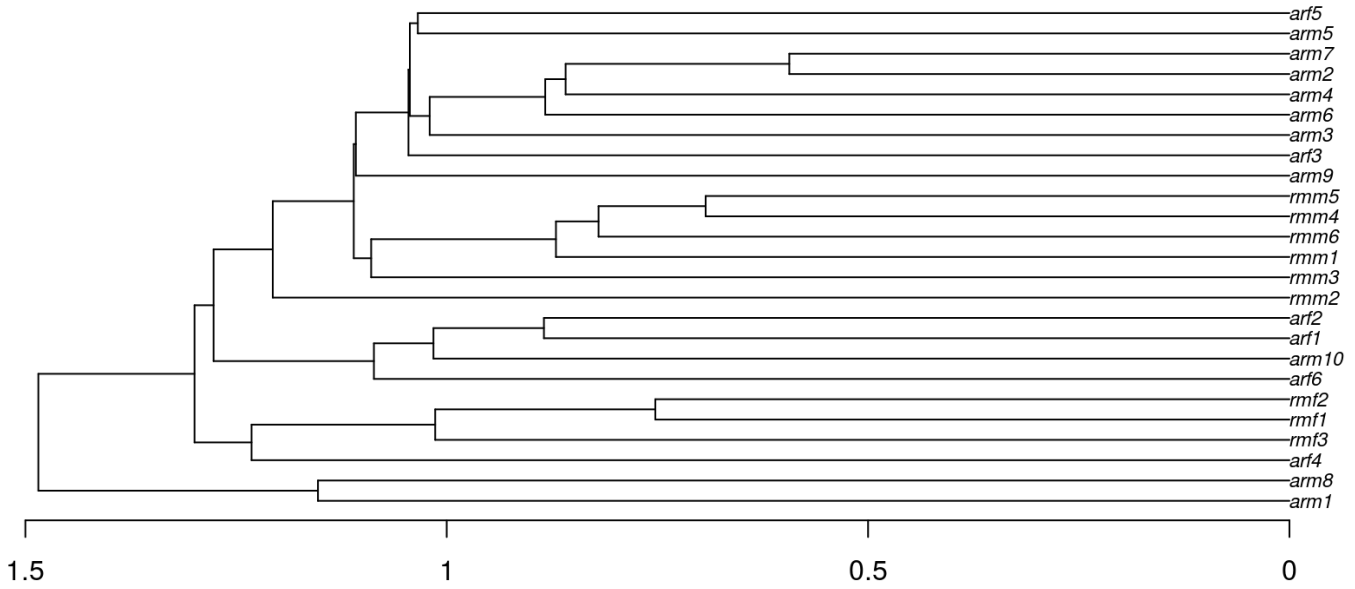
Визуализируем при помощи базовой графики

```
plot(hc_single)
```



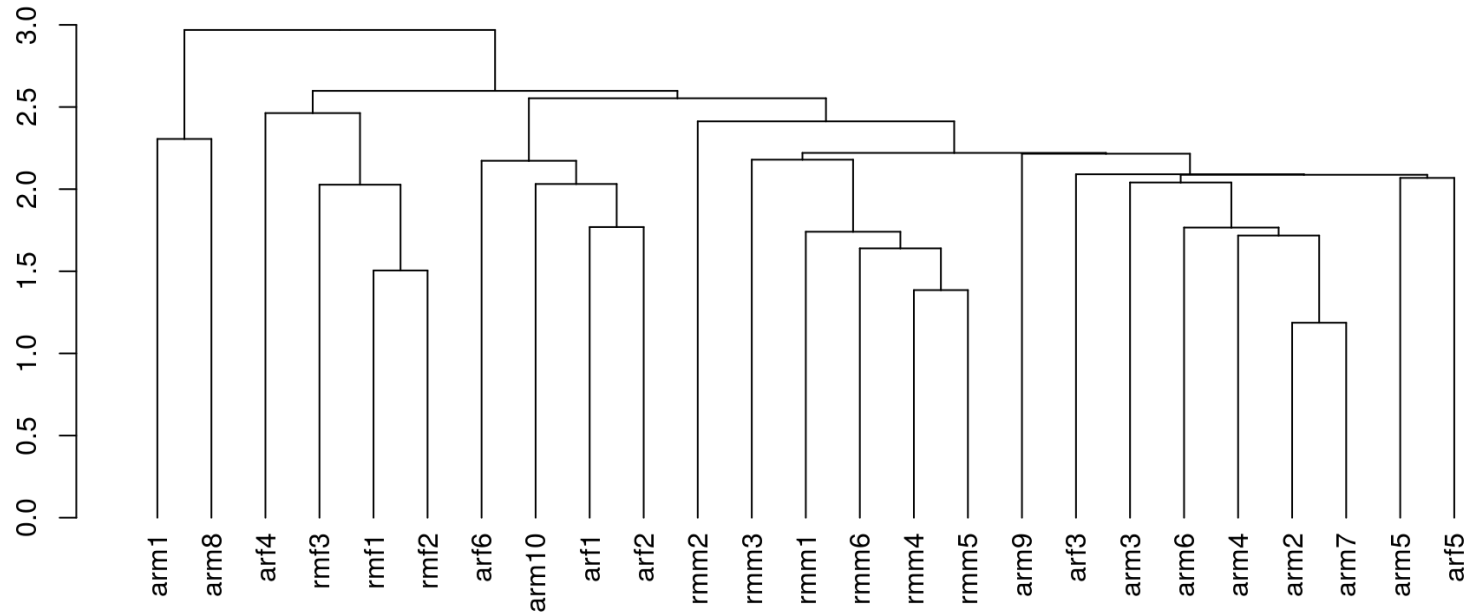
Визуализируем средствами ape

```
ph_single <- as.phylo(hc_single)
plot(ph_single, type = "phylogram", cex = 0.7)
axisPhylo()
```



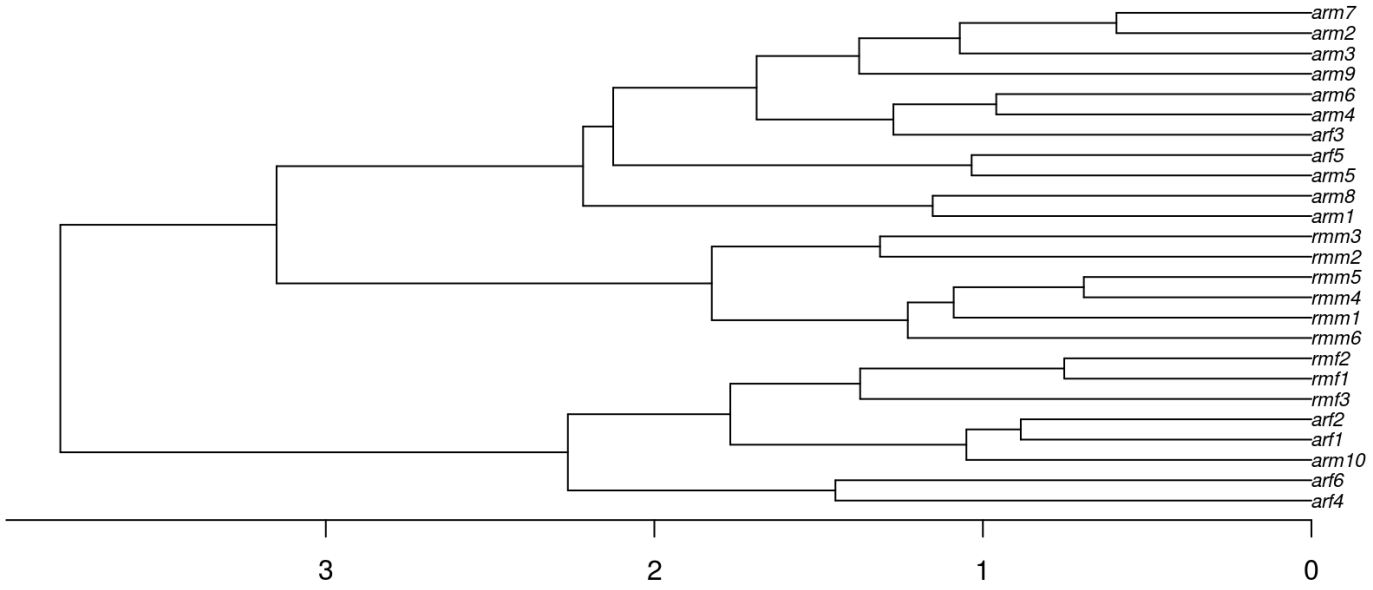
Визуализируем средствами dendextend

```
den_single <- as.dendrogram(hc_single)
plot(den_single)
```



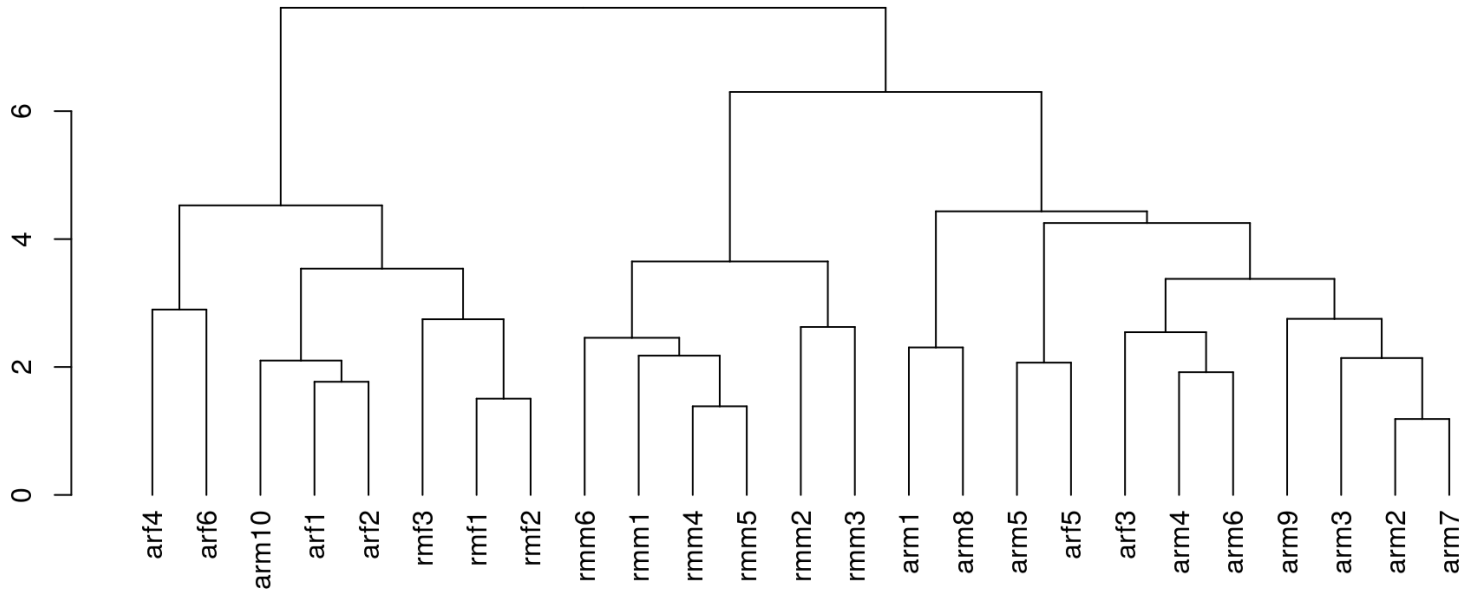
Метод отдаленного соседа в R

```
hc_compl <- hclust(d, method = "complete")
ph_compl <- as.phylo(hc_compl)
plot(ph_compl, type = "phylogram", cex = 0.7)
axisPhylo()
```



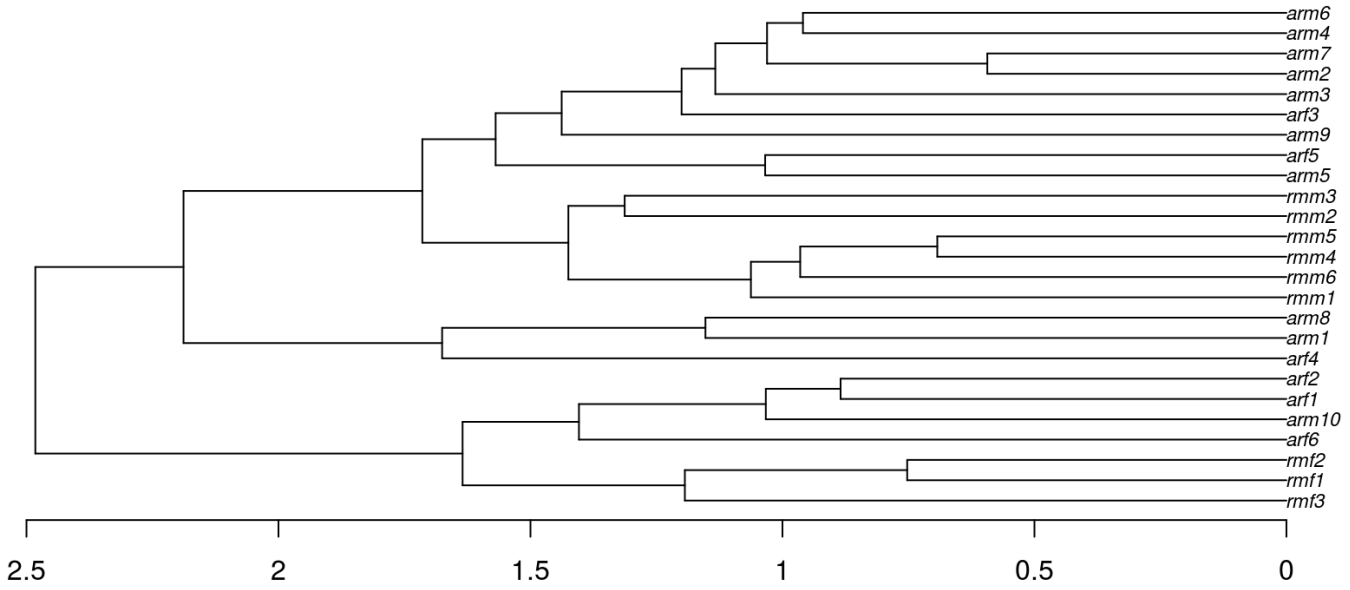
Визуализируем дерево, полученное методом отдаленного соседа, средствами dendextend

```
den_compl <- as.dendrogram(hc_compl)
plot(den_compl)
```



Метод невзвешенного попарного среднего в R

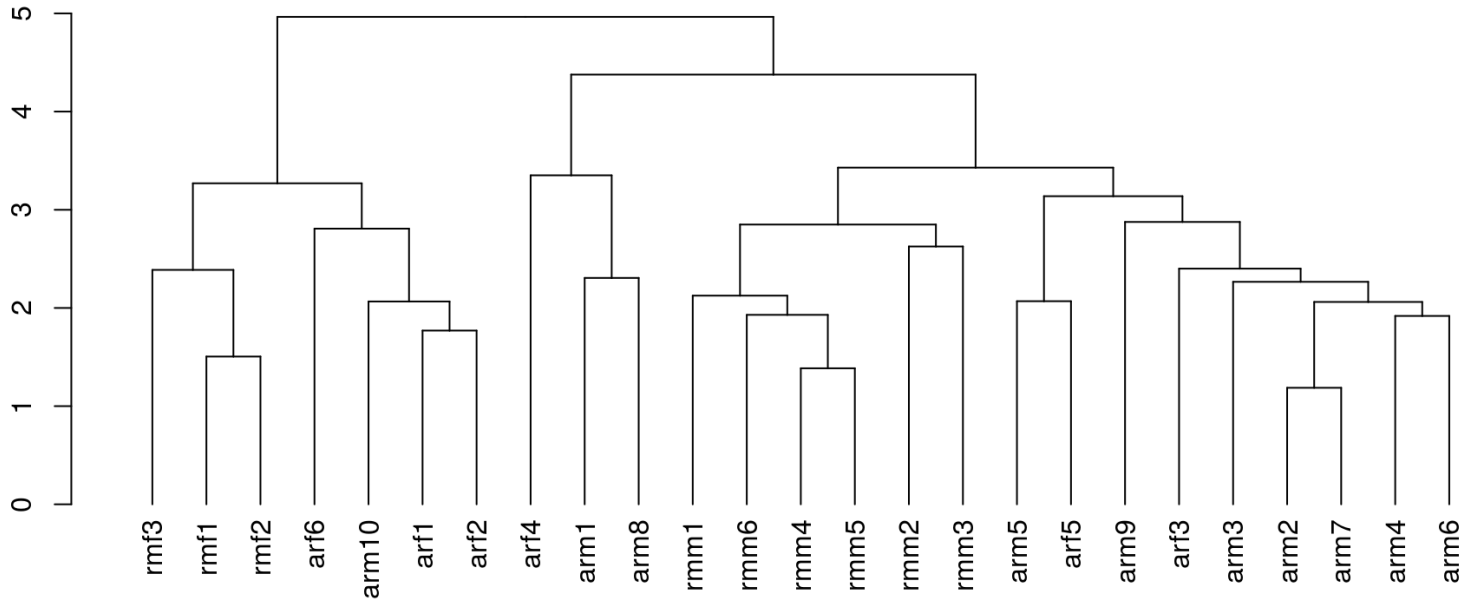
```
hc_avg <- hclust(d, method = "average")
ph_avg <- as.phylo(hc_avg)
plot(ph_avg, type = "phylogram", cex = 0.7)
axisPhylo()
```



Визуализируем дерево, полученное методом невзвешенного попарного среднего, средствами

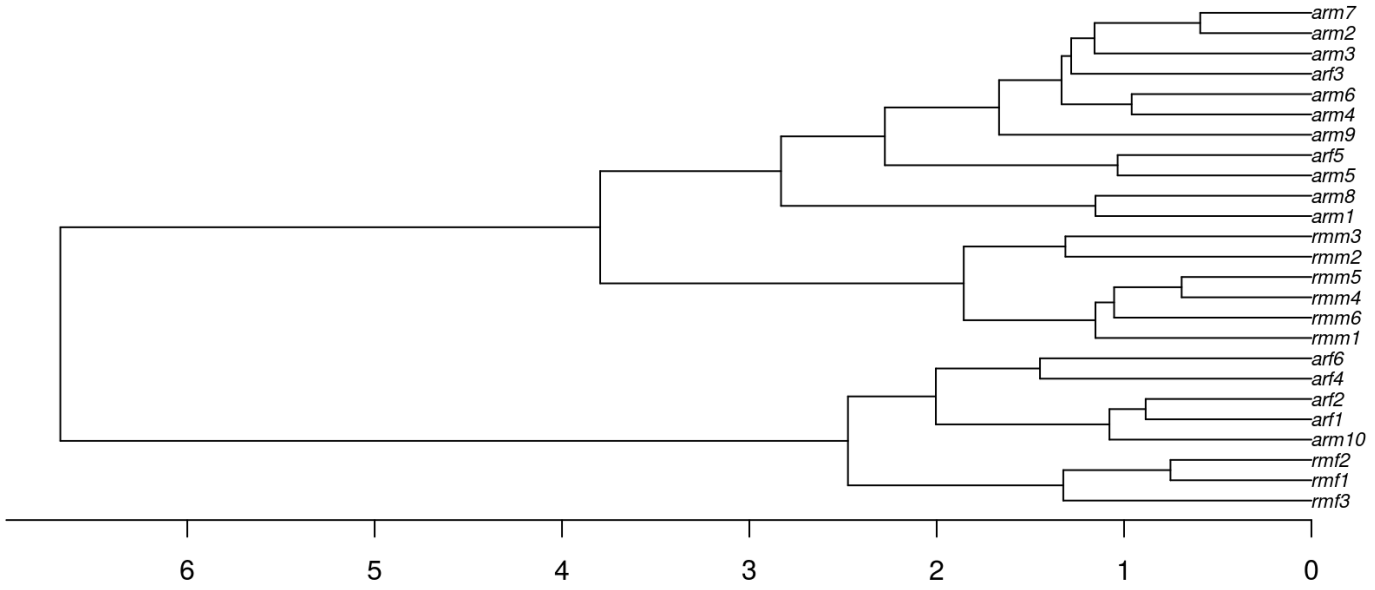
dendextend

```
den_avg <- as.dendrogram(hc_avg)
plot(den_avg)
```



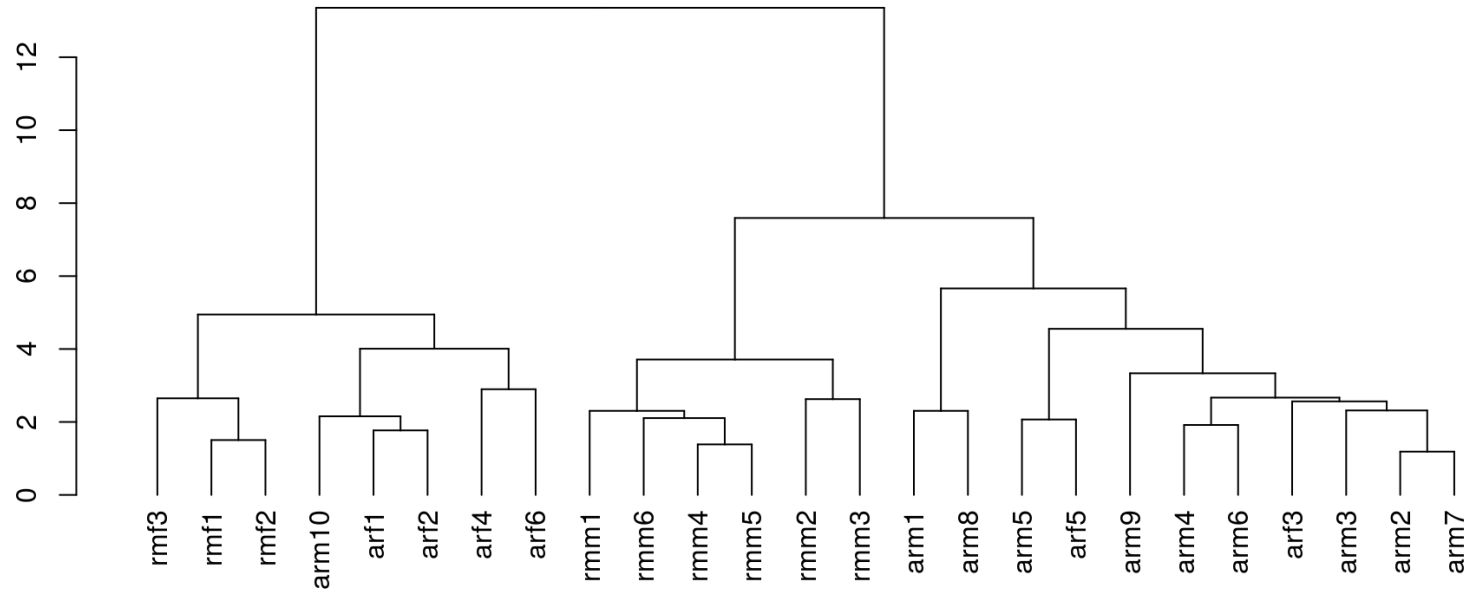
Метод Варда в R

```
hc_w2 <-hclust(d, method = "ward.D2")
ph_w2 <- as.phylo(hc_w2)
plot(ph_w2, type = "phylogram", cex = 0.7)
axisPhylo()
```



Визуализируем дерево, полученное методом Варда, средствами dendextend

```
den_w2 <- as.dendrogram(hc_w2)
plot(den_w2)
```



Качество кластеризации

Кофенетическая корреляция

Кофенетическая корреляция

Кофенетическое расстояние - расстояние между объектами на дендрограмме

Кофенетическую корреляцию можно рассчитать как пирсоновскую корреляцию (обычную) между матрицами исходных и кофенетических расстояний между всеми парами объектов

Метод, который дает наибольшую кофенетическую корреляцию дает кластеры лучше всего отражающие исходные данные

Можно рассчитать при помощи функции из пакета `ape`

Кофенетическая корреляция в R

```
c_single <- cophenetic(ph_single)
c_compl <- cophenetic(ph_compl)
c_avg <- cophenetic(ph_avg)
c_w2 <- cophenetic(ph_w2)

cor(d, as.dist(c_single))

## [1] 0.565

cor(d, as.dist(c_compl))

## [1] 0.706

cor(d, as.dist(c_avg))

## [1] 0.745

cor(d, as.dist(c_w2))

## [1] 0.726
```

Задание:

Оцените для данных об ирисах при помощи кофенетической корреляции качество кластеризаций, полученных разными методами.

Какой метод дает лучший результат?

Качество и количество кластеров

На каком уровне нужно делить дендрограмму на кластеры?

- Можно субъективно, на любом выбранном уровне. Главное, чтобы кластеры были осмысленными и интерпретируемыми.
- Можно выбрать, глядя на распределение расстояний ветвления
- Можно оценить стабильность кластеризации при помощи бутстрепа и выбрать оптимальное число кластеров.

Стабильность кластеров

На хорошей кластеризации кластеры должны воспроизводиться.

Оптимальное число кластеров можно определить рассчитывая меру нестабильности для каждой из выборок бутстрепа (Fang and Wang (2012))

Много раз берем по 2 выборки бутстрепом, и оцениваем стабильность.

```
library(fpc)
nsel <- nselectboot(d, B = 1000, clustermethod = hclustCBI, seed = 9646, method =
"average", krange=3:11)
```

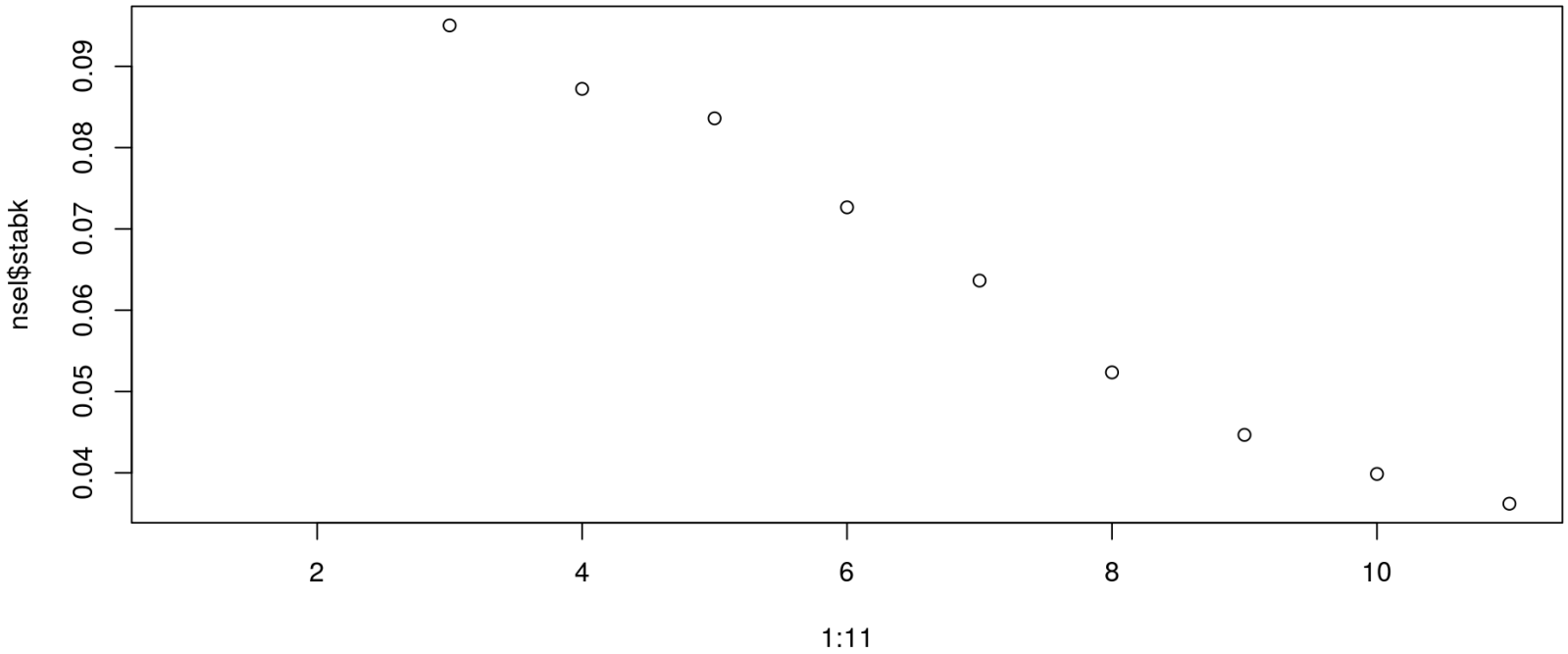
Оптимальное число кластеров — с минимальным значением нестабильности

```
nsel$kopt # оптимальное число кластеров

## [1] 11

nsel$stabk # средние значения нестабильности

##  [1]     NA     NA 0.0950 0.0872 0.0836 0.0727 0.0636 0.0524 0.0447
## [10] 0.0399 0.0362
```

Визуализируем значения нестабильности

Чтобы легче было выбирать, и чтобы понять, что происходит, изобразим значения нестабильности на графике

```
plot(1:11, nsel$stabk)
```



Ширина силуэта

Ширина силуэта — мера степени принадлежности объекта к кластеру. Это среднее расстояние от данного объекта до других объектов из того же кластера, в сравнении с аналогичной величиной для ближайшего кластера.

Оценим ширину силуэта для 3 или 6 кластеров

```
complete3 <- cutree(hclust(d), 3)
qual3<- cluster.stats(d, complete3)
qual3$clus.avg.silwidths
```

```
##     1     2     3
## 0.334 0.340 0.149
```

```
complete6 <- cutree(hclust(d), 6)
qual6<- cluster.stats(d, complete6)
qual6$clus.avg.silwidths
```

```
##     1     2     3     4     5     6
## 0.220 0.142 0.372 0.205 0.335 0.095
```

```
mean(qual3$clus.avg.silwidths); mean(qual6$clus.avg.silwidths)
```

```
## [1] 0.274
```

```
## [1] 0.228
```

Бутстреп поддержка ветвей

"An approximately unbiased test of phylogenetic tree selection" (Shimodaria, 2002)

Этот тест использует специальный вариант бутстрепа — multiscale bootstrap. Мы не просто многократно берем бутстреп-выборки и оцениваем для них вероятность получения топологий (BP p-value), эти выборки еще и будут с разным числом объектов. По изменению BP при разных объемах выборки можно вычислить AU (approximately unbiased p-value).

```
library(pvclust)
# итераций должно быть 1000 и больше, здесь мало для скорости
set.seed(42)
cl_boot <- pvclust(t(s_w), method.hclust = "average", nboot = 100, method.dist =
"euclidean")
```

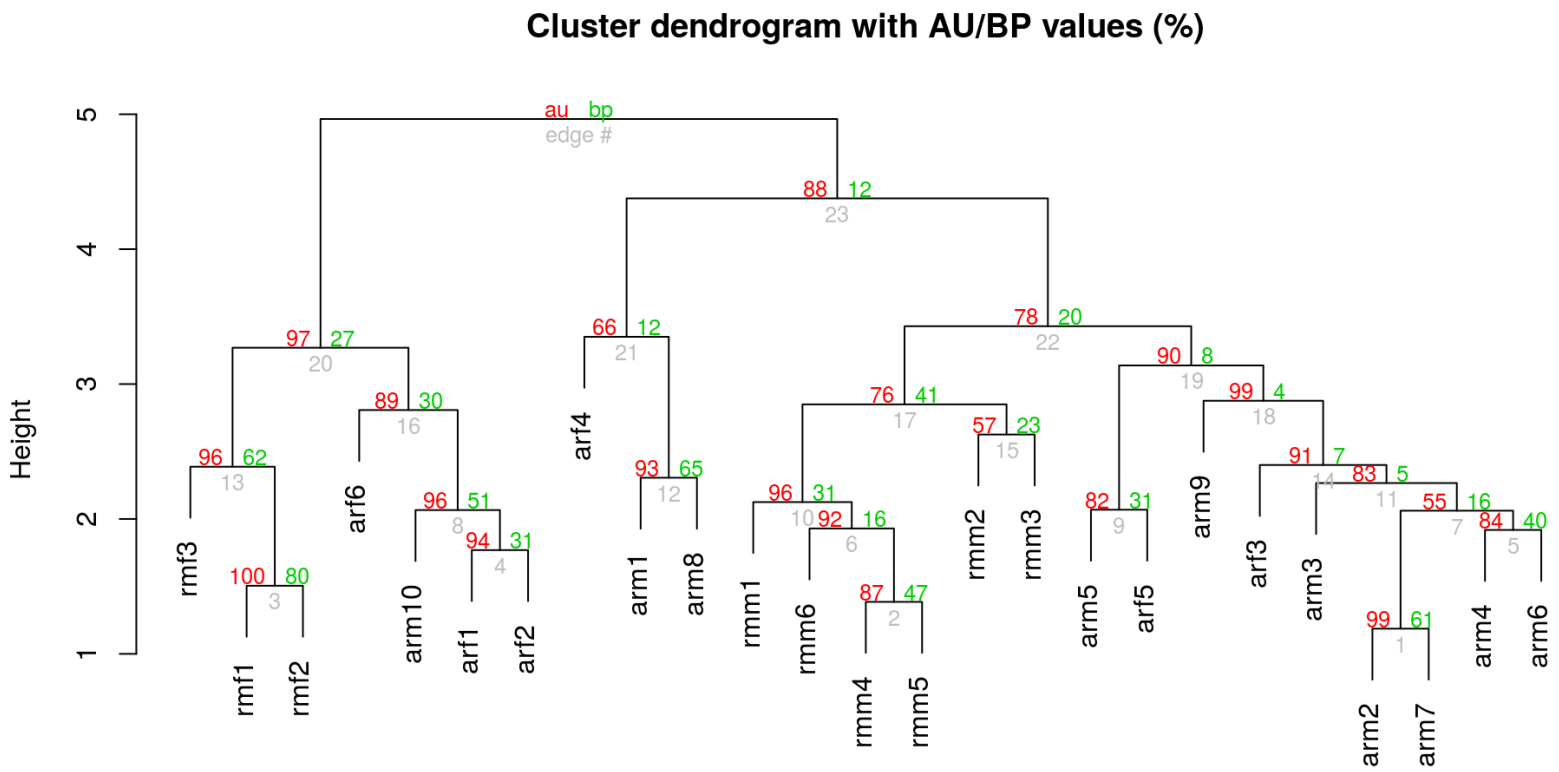
```
## Bootstrap (r = 0.44)... Done.
## Bootstrap (r = 0.56)... Done.
## Bootstrap (r = 0.67)... Done.
## Bootstrap (r = 0.78)... Done.
## Bootstrap (r = 0.89)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.0)... Done.
## Bootstrap (r = 1.11)... Done.
## Bootstrap (r = 1.22)... Done.
## Bootstrap (r = 1.33)... Done.
```

```
## Warning in a$p[] <- c(1, bp[r == 1]): number of items to replace is
## not a multiple of replacement length
```

Дерево с величинами поддержки

AU — approximately unbiased p-values (красный),BP — bootstrap p-values (зеленый)

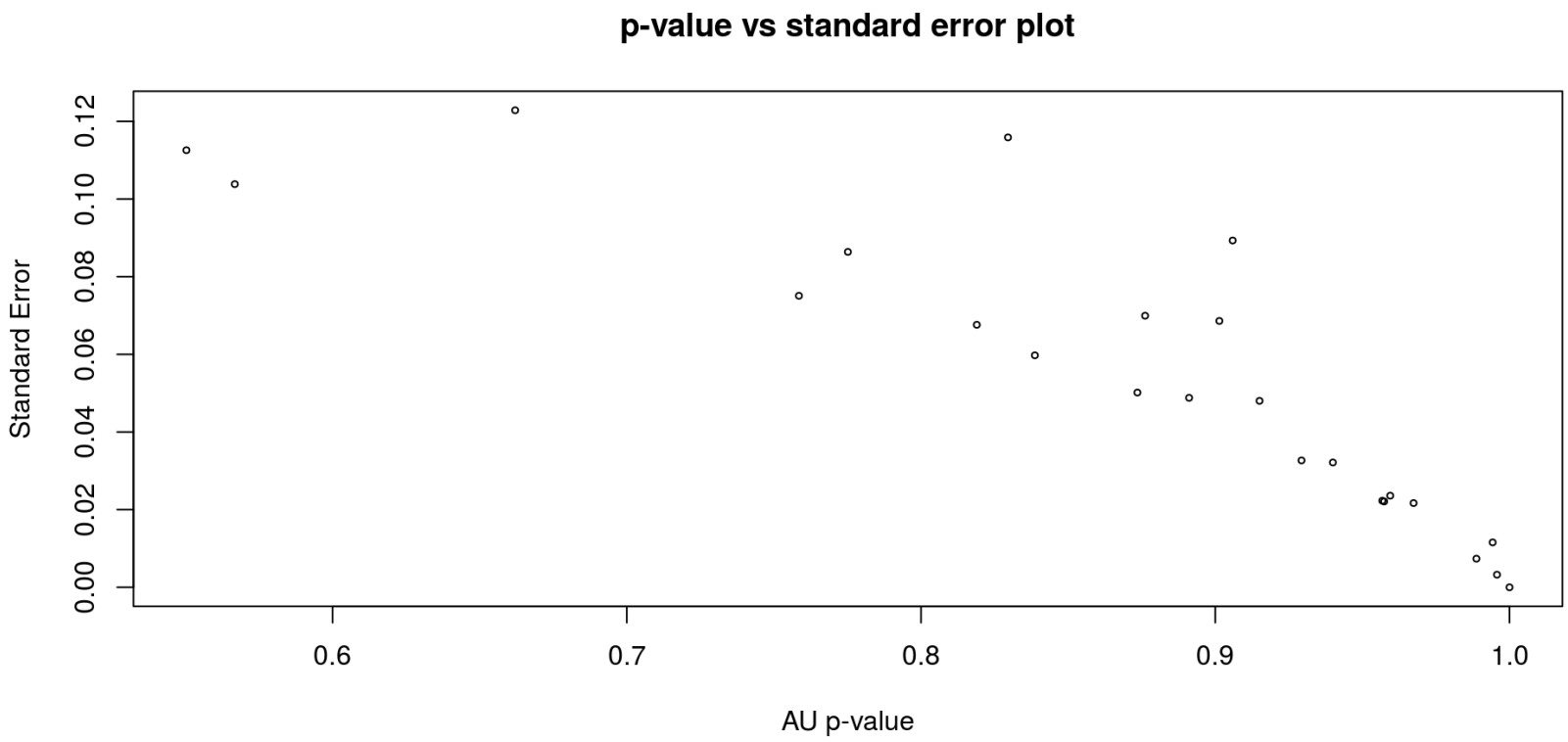
```
plot(cl_boot)
```



Для диагностики качества оценок AU

График стандартных ошибок для AU p-value нужен, чтобы оценить точность оценки самих AU. Чем больше было бутстреп-итераций, тем точнее будет оценка.

```
seplot(cl_boot, cex = 0.5)
```



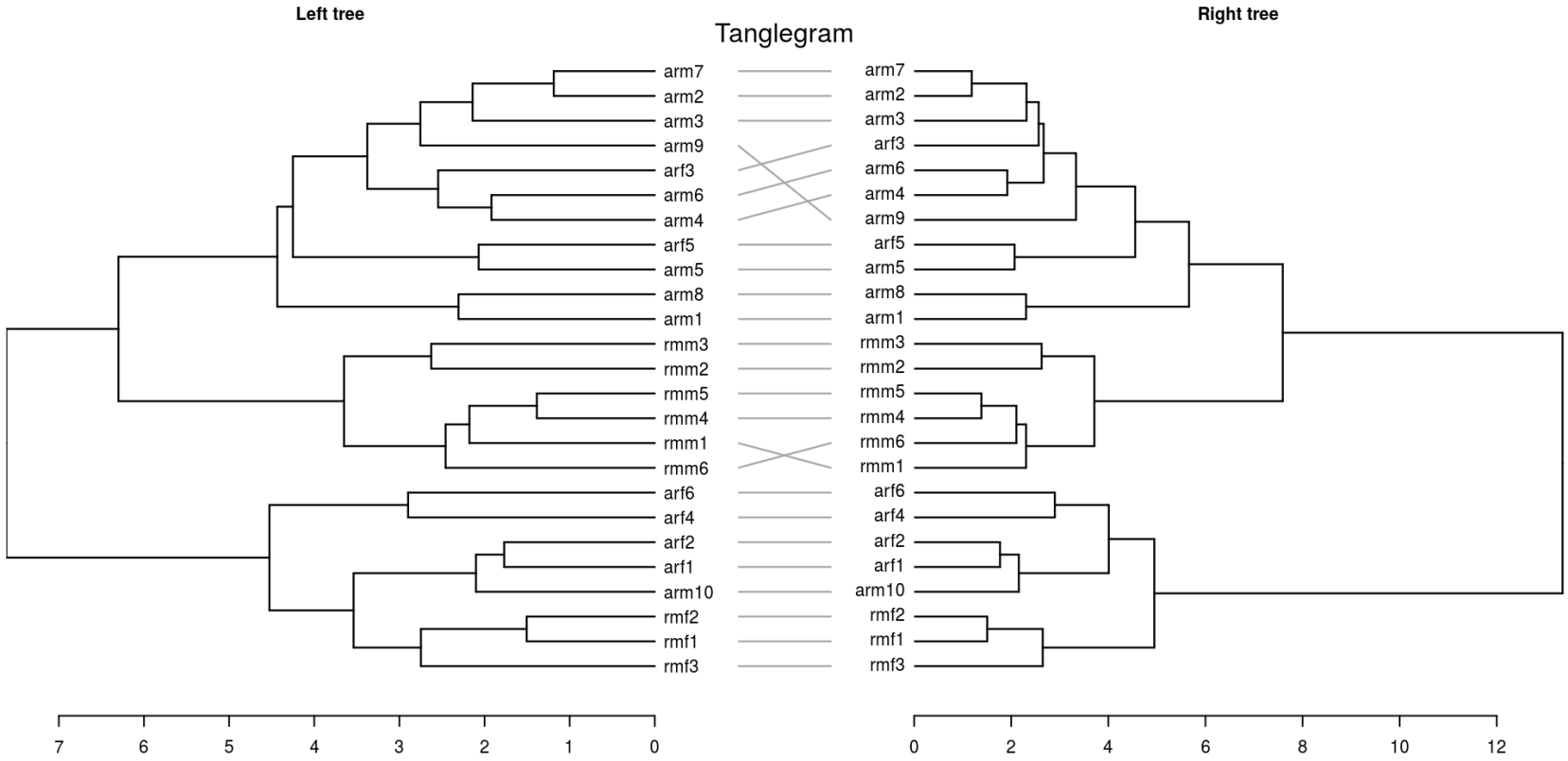
Сопоставление деревьев: Танглграммы

Танглграмма

Два дерева (с непохожим ветвлением) выравнивают, вращая случайным образом ветви вокруг оснований. Итеративный алгоритм. Картина каждый раз разная.

```
set.seed(395)
untang_w <- untangle_step_rotate_2side(den_compl, den_w2, print_times = F)

# танглграмма
tanglegram(untang_w[[1]], untang_w[[2]],
           highlight_distinct_edges = FALSE,
           common_subtrees_color_lines = F,
           main = "Tanglegram",
           main_left = "Left tree",
           main_right = "Right tree",
           columns_width = c(8, 1, 8),
           margin_top = 3.2, margin_bottom = 2.5,
           margin_inner = 4, margin_outer = 0.5,
           lwd = 1.2, edge.lwd = 1.2,
           lab.cex = 1, cex_main = 1)
```

Танглграмма



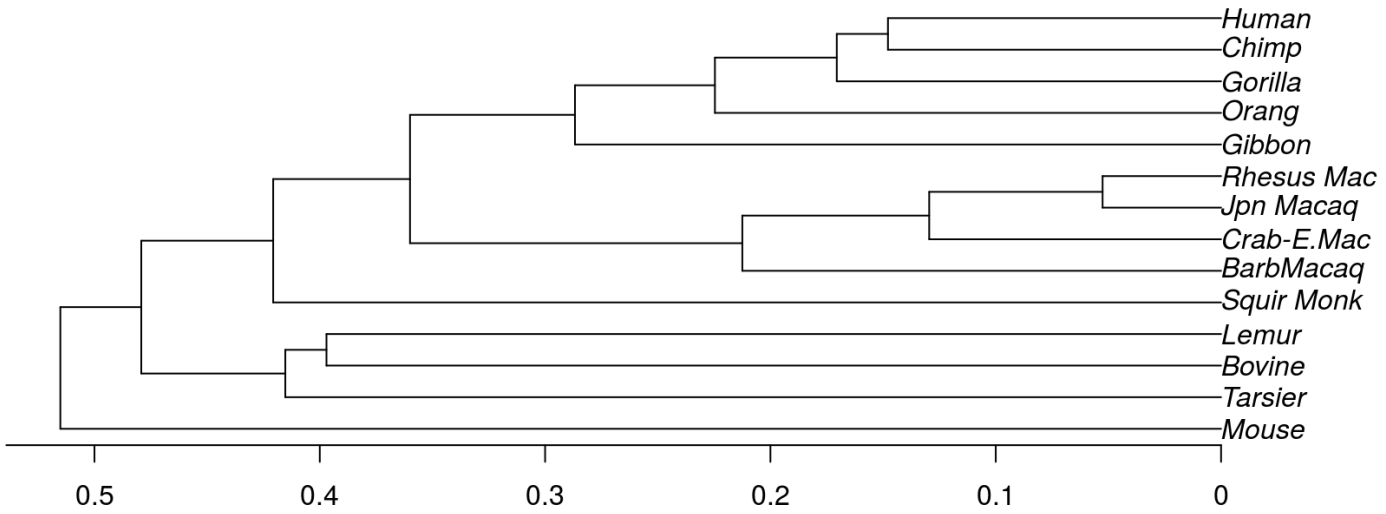
Задание

Постройте танглграмму для данных о морфометрии ирисов из дендрограмм, полученных методом ближайшего соседа и методом Варда.

Деревья по генетическим данным

И небольшая демонстрация - дерево по генетическим данным

```
webpage <-"http://evolution.genetics.washington.edu/book/primates.dna"
primates.dna <- read.dna(webpage)
d_pri <- dist.dna(primates.dna)
hc_pri <- hclust(d_pri, method = "average")
ph_pri <- as.phylo(hc_pri)
plot(ph_pri)
axisPhylo()
```



Take home messages

- Результат кластеризации зависит не только от выбора коэффициента, но и от выбора алгоритма.
- Качество кластеризации можно оценить разными способами.
- Кластеризации, полученные разными методами, можно сравнить на танглграммах.

Дополнительные ресурсы

- Borcard, D., Gillet, F., Legendre, P., 2011. Numerical ecology with R. Springer.
- Legendre, P., Legendre, L., 2012. Numerical ecology. Elsevier.
- Quinn, G.G.P., Keough, M.J., 2002. Experimental design and data analysis for biologists. Cambridge University Press.

И еще ресурсы

Как работает UPGMA можно посмотреть здесь:

- http://www.southampton.ac.uk/~re1u06/teaching/upgma/

Как считать поддержку ветвей (пакет + статья):

- pvclust: An R package for hierarchical clustering with p-values [WWW Document], n.d. URL http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/ (accessed 11.7.14).

Для анализа молекулярных данных:

- Paradis, E., 2011. Analysis of Phylogenetics and Evolution with R. Springer.