Специальные случаи применения анализа главных компонент

Анализ и визуализация многомерных данных с использованием R

Марина Варфоломеева, Вадим Хайтов

Анализ морфометрических данных при помощи анализа главных компонент

- Классический подход к морфометрии
- Геометрическая морфометрия
- Эволюция формы

Вы сможете

- Проанализировать морфометрические данные корректно удалив влияние абсолютного размера
- Рассказать, что происходит во время обобщенного прокрустова анализа
- Проанализировать данные о координатах меток используя методы геометрической морфометрии
- Понимать, каким образом происходит отображение филогенетического древа в пространство форм

Классический подход к морфометрии

Классический подход к морфометрии

Для анализа формы различных структур анализируются расстояния между метками, а не их координаты.

Признаки сильно интегрированных структур, например частей скелета, лучше анализировать совместно друг с другом. Один из вариантов анализа - анализ главных компонент.

Пример: морфометрия черепах

Черепахи - единственные живые представители анапсид (череп не имеет височных окон). Морфология черепа важна для их систематики (Claude et al., 2004).

Данные - 24 разных измерения черепов черепах 122 ныне живущих пресноводных, морских и наземных видов и одного ископаемого.

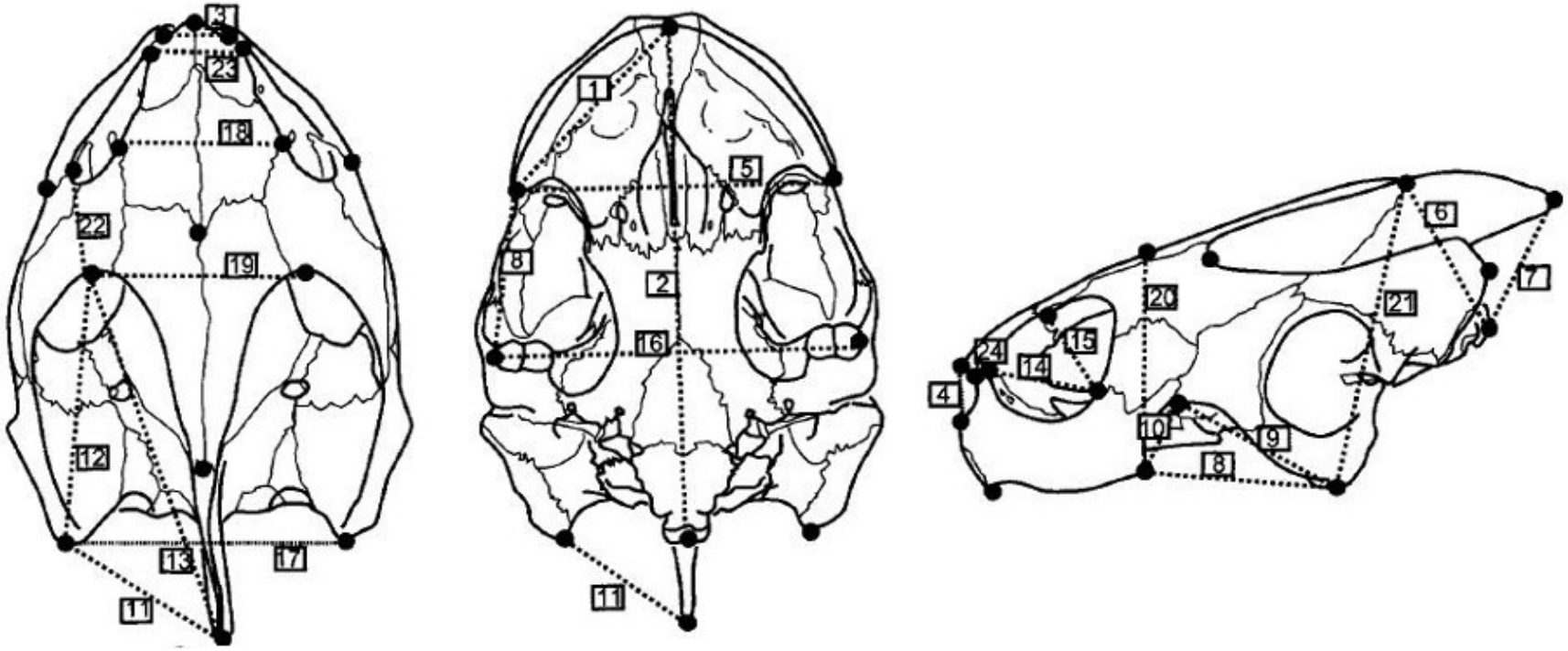


Рис. 30.1 из Zuur et al. 2007

Читаем данные

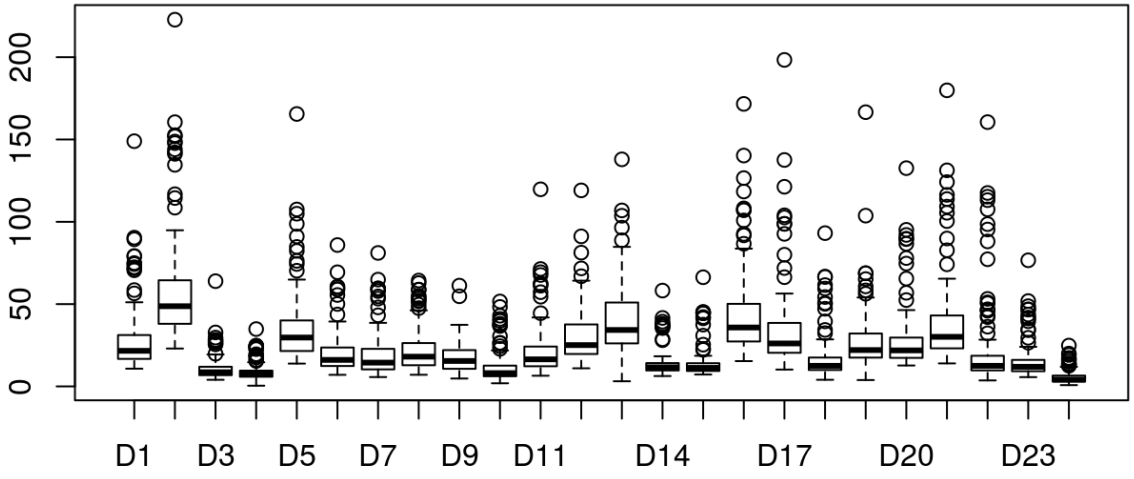
```
turt <- read.table("data/turtles.txt", header = TRUE)
turt$Environment3 <- factor(turt$Environment3, levels = c(0, 1, 2, 9), labels =
c("Freshwater", "Terrestrial", "Marine", "Fossil"))
colnames(turt)
```

```
##  [1] "nspecies"     "species_name" "Family"       "SuperFamily"
##  [5] "Order"        "Environment"  "Environment3" "D1"
##  [9] "D2"           "D3"           "D4"           "D5"
## [13] "D6"           "D7"           "D8"           "D9"
## [17] "D10"          "D11"          "D12"          "D13"
## [21] "D14"          "D15"          "D16"          "D17"
## [25] "D18"          "D19"          "D20"          "D21"
## [29] "D22"          "D23"          "D24"
```

Данные из Zuur et al. 2007

Чтобы понять, нужно ли стандартизовать исходные данные, построим боксплот

```
boxplot(x = turt[8:31])
```



- Наверное, лучше стандартизовать

Задание: Проведите анализ главных компонент

- Сколько изменчивости объясняют компоненты?
- Сколько компонент достаточно для описания данных?

Решение: Делаем анализ главных компонент по стандартизованным данным

```
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.3-3
```

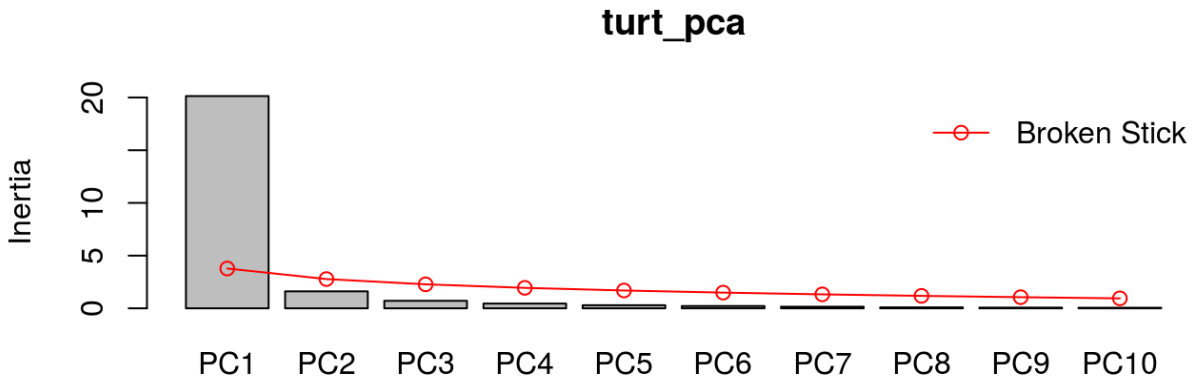
```
turt_pca <- rda(turt[, 8:31], scale = TRUE)
```

Решение: Сколько компонент достаточно для описания данных?

```
eig <- eigenvals(turt_pca)[1:5]
eig*100/sum(eig) # доля объясненной изменчивости
```

```
##   PC1   PC2   PC3   PC4   PC5
## 86.76  6.94  3.07  1.96  1.27
```

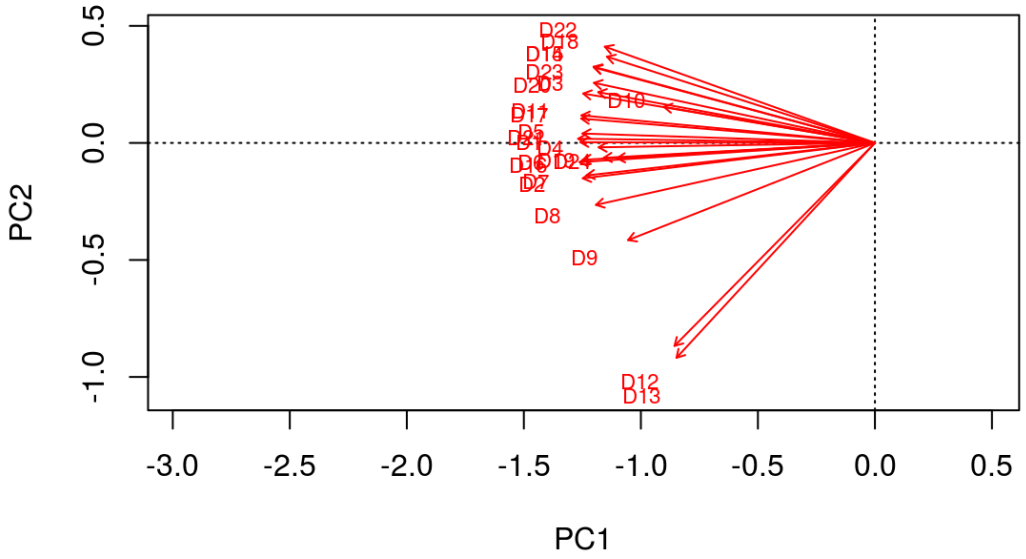
```
screeplot(turt_pca, bstick = TRUE)
```



- Первая компонента объясняет очень много, остальные - почти ничего. Одной компоненты достаточно?
- Нет! Не все так просто.

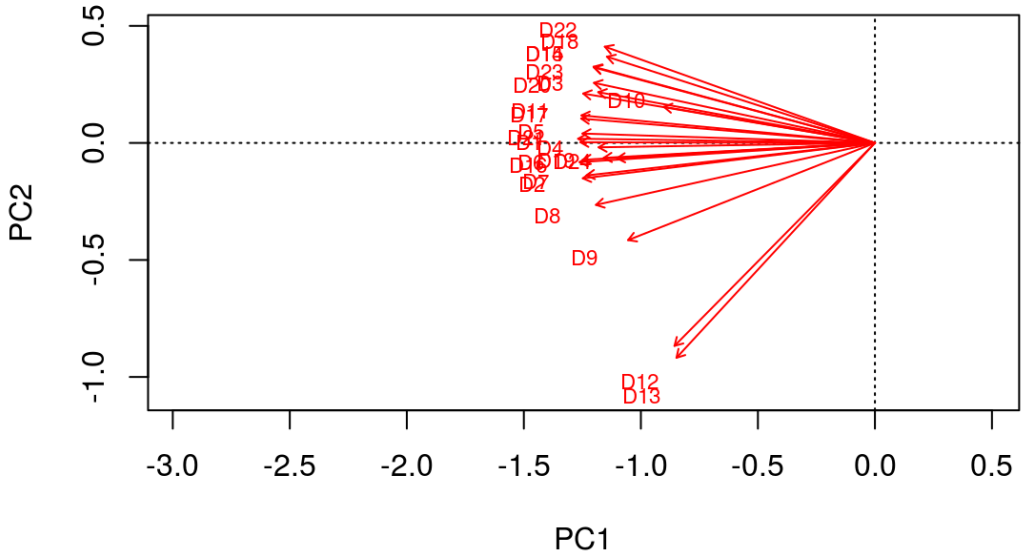
Что странного в этой картинке?

```
biplot(turt_pca, display = "species", scaling = 2)
```



- Как вы думаете, почему у всех переменных большие нагрузки по первой компоненте?
- Первая компонента отражает размеры особей.

При анализе сырых морфометрических данных первая компонента отражает размер объектов и, возможно, немножко - их форму



Задание:

Придумайте способ избавится от влияния размера

Классические способы избавится от влияния размера:

- использовать одну из исходных переменных как оценку "размера": использовать в PCA остатки от регрессий исходных признаков от "размера"
- стандартизация исходных данных при помощи деления на величину "размера" для каждого образца (корень из суммы квадратов измерений)
- сделать двойное центрирование (логарифмированных) исходных данных
- и т.д. и т.п.

Двойное центрирование

Нам достаточно центрировать строки, т.к. столбцы будут центрированы автоматически в процессе анализа главных компонент.

```
# Функция, которая может центрировать вектор
center <- function(x){
  x - mean(x, na.rm = TRUE)
}
# применяем эту функцию к каждой строке
dbcent <- t(apply(turt[, 8:31], 1, center))
# получившийся датафрейм пришлось транспонировать,
# поскольку apply() результаты от каждой строки
# возвращает в виде столбцов
```

Задание:

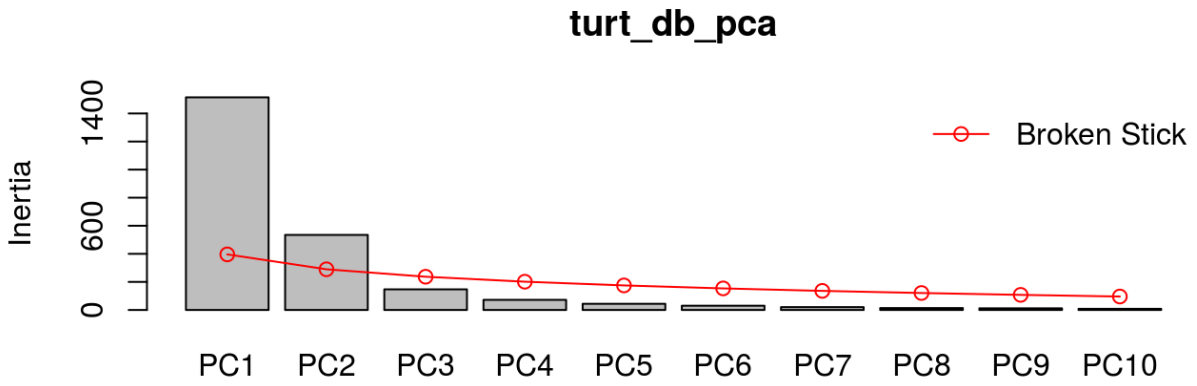
- Проведите анализ главных компонент по центрированным данным (`dbcent`). При помощи скольки компонент можно адекватно описать данные?
- Постройте график факторных нагрузок. Изменилась ли интерпретация компонент?

Решение: После двойного центрирования большие собственные числа у нескольких компонент

```
turt_db_pca <- rda(dbcent)
eig_db <- eigenvals(turt_db_pca)[1:5]
eig_db*100/sum(eig_db)
```

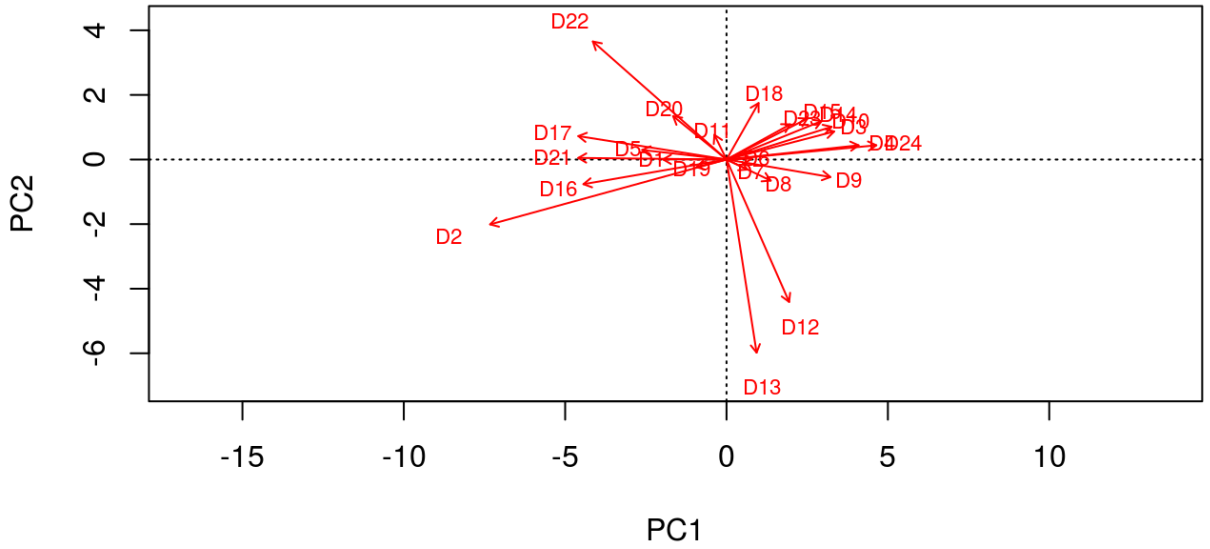
```
##   PC1   PC2   PC3   PC4   PC5
## 65.48 23.12  6.36  3.13  1.91
```

```
screeplot(turt_db_pca, bstick = TRUE)
```



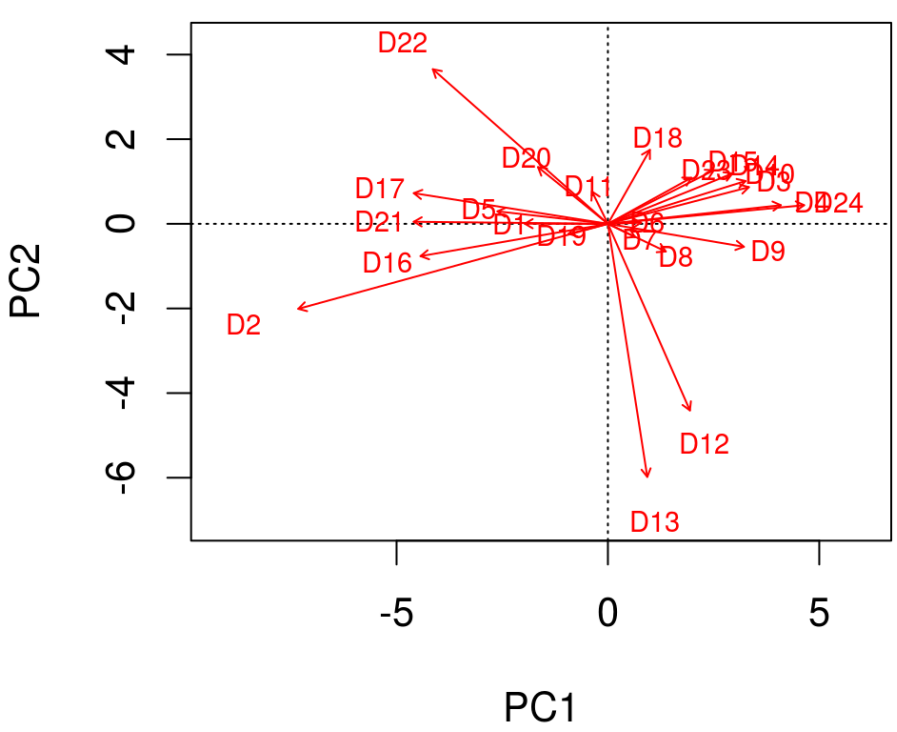
Решение: После двойного центрирования у переменных высокие нагрузки на несколько компонент, влияние размера удалено

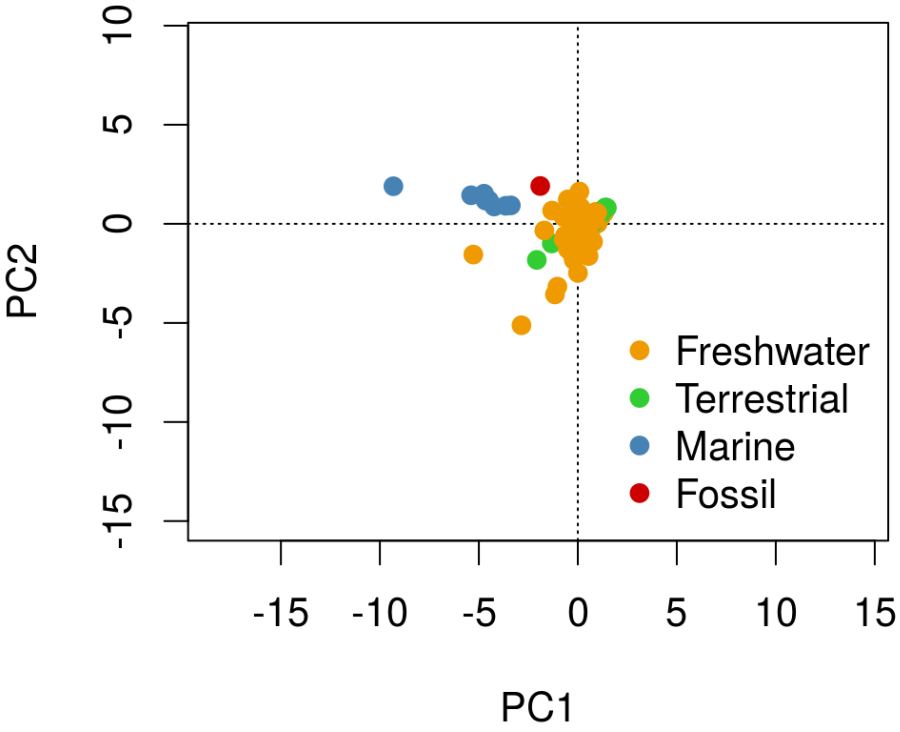
```
biplot(turt_db_pca, display = "species", scaling = 2)
```



Интерпретируем как обычно: компонента отражает несколько признаков

Ординация черепах по морфометрии черепов (двойное центрирование данных)





- У пресноводных большие D12 и D13, и маленькая D2. У морских наоборот
- Ископаемая черепаха похожа на нынешних морских

Код для графика ординации черепах по морфометрии черепов

```
op <- par(mfrow = c(1, 2), mar = c(4, 4, 0.5, 0.5), cex = 1.3)
biplot(turt_db_pca, display = "species", scaling = 2)
# цвета для графика факторных координат
colvec <- c("orange2", "limegreen", "steelblue", "red3")
# пустой график
plot(turt_db_pca, type = "n", scaling = 1)
# точки, раскрашенные по уровням фактора turt$Environment3
points(turt_db_pca, display = "sites", scaling = 1, pch = 21,
       col = colvec[turt$Environment3], bg = colvec[turt$Environment3])
# легенда
legend("bottomright", legend = levels(turt$Environment3), bty = "n", pch = 21,
       col = colvec, pt.bg = colvec)
par(op)
```

Но настоящие джедаи теперь анализируют координаты меток, а не расстояния между ними!

Геометрическая морфометрия

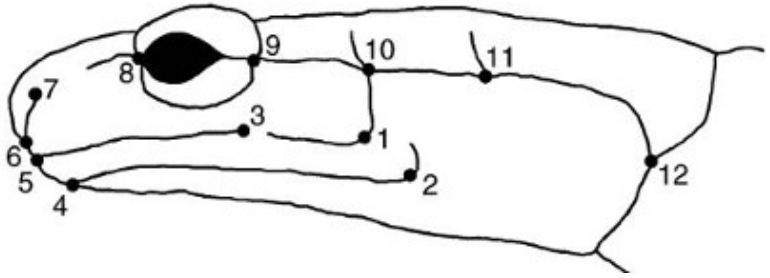
Пример: Форма головы Апалачских саламандр рода *Plethodon*

Plethodon jordani и *P.teyahalee* встречаются вместе и раздельно.В совместно обитающих популяциях меняется форма головы обоих видов. В разных группах популяций этот процесс параллельно приводит к одинаковым результатам. По-видимому, одной из причин параллельной эволюции может быть межвидовая конкуренция (Adams, 2004, 2010).

Plethodon jordani - Jordan's Salamander by John P Clare on Flickr

Plethodon cf. *teyahalee* by squamatologist on Flickr

Морфометрия головы саламандр




```
# install.packages("geomorph", dependencies = TRUE)
library(geomorph)
```

```
## Loading required package: rgl
```

```
## Loading required package: ape
```

```
data(plethodon)
str(plethodon, vec.len = 2, give.attr = F)
```

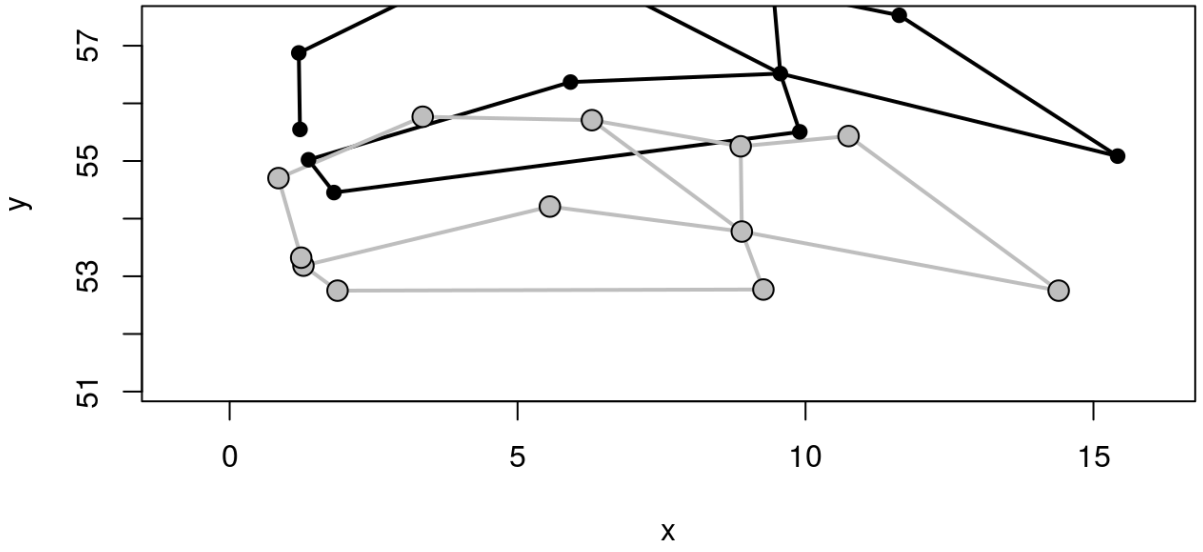
```
## List of 5
##  $ land   : num [1:12, 1:2, 1:40] 8.89 9.27 ...
##  $ links  : num [1:14, 1:2] 4 3 2 1 1 ...
##  $ species: Factor w/ 2 levels "Jord","Teyah": 1 1 1 1 1 ...
##  $ site   : Factor w/ 2 levels "Allo","Symp": 2 2 2 2 2 ...
##  $ outline: num [1:3631, 1:2] 0.399 0.4 ...
```

рис. из Adams, 2004, 2010

Сырые морфометрические данные еще не выровнены

Все образцы разного размера и разной ориентации в пространстве. На этом графике — два образца для примера.

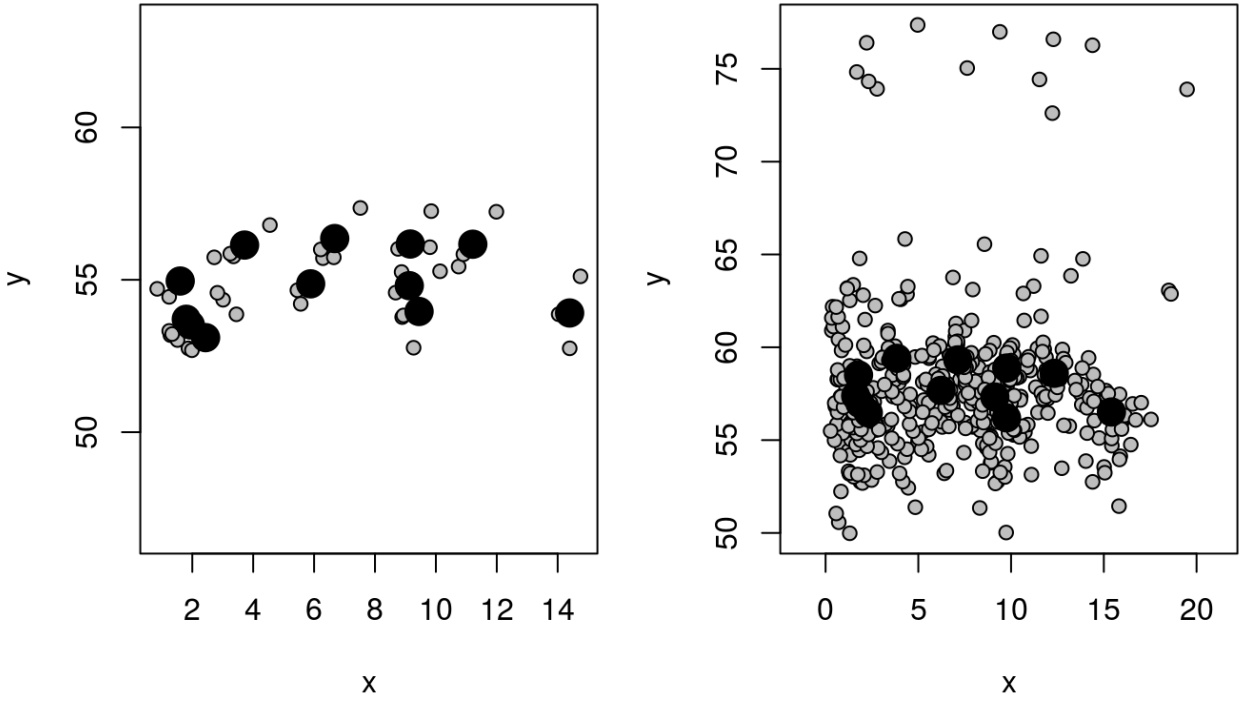
```
plotRefToTarget(plethodon$land[, , 1], plethodon$land[, ,10],
                method = "points", mag = 1, links = plethodon$links)
```



Если нарисовать невыровненные образцы, получится полная каша. Что делать?

Слева - три образца, справа - все. Жирные точки - центроиды соответствующих меток

```
op <- par(mfrow = c(1, 2), mar = c(4, 4, 1, 1))
plotAllSpecimens(plethodon$land[, , 1:3], links=plethodon$links)
plotAllSpecimens(plethodon$land,links=plethodon$links)
par(op)
```



Геометрическая морфометрия

1. Влияние размера удаляется при помощи обобщенного прокрустова анализа (масштабирование, поворот и сдвиг координат)
2. Преобразованные координанты меток используются как признаки объектов (конкретных особей) в анализе главных компонент. Получается морфопространство. Главные компоненты отражают изменения формы.

- можно получить усредненную форму для любой группы выровненных координат
- можно сравнить форму любой особи со средней формой
- можно проследить изменение формы вдоль осей главных компонент

Прокрустов анализ

Тезей убивает разбойника Прокруста (источник https://mrpsmythopedia.wikispaces.com/Procrustes)

Шаг 1. Выравниваем данные при помощи обобщенного прокрустова анализа

Generalized Procrustes Analysis (GPA)

Минимизируем сумму квадратов расстояний между одноименными метками, меняя масштаб, поворачивая и сдвигая координаты. Вот как это выглядит на данных про черепах:

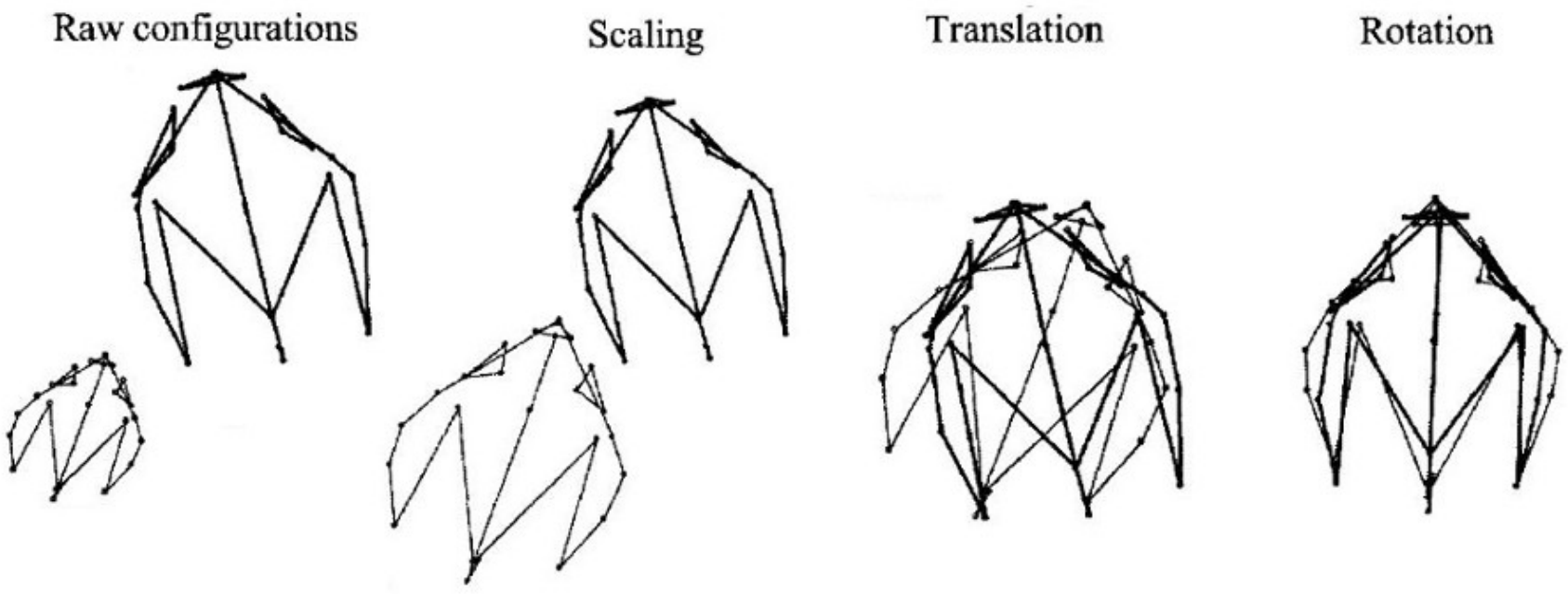
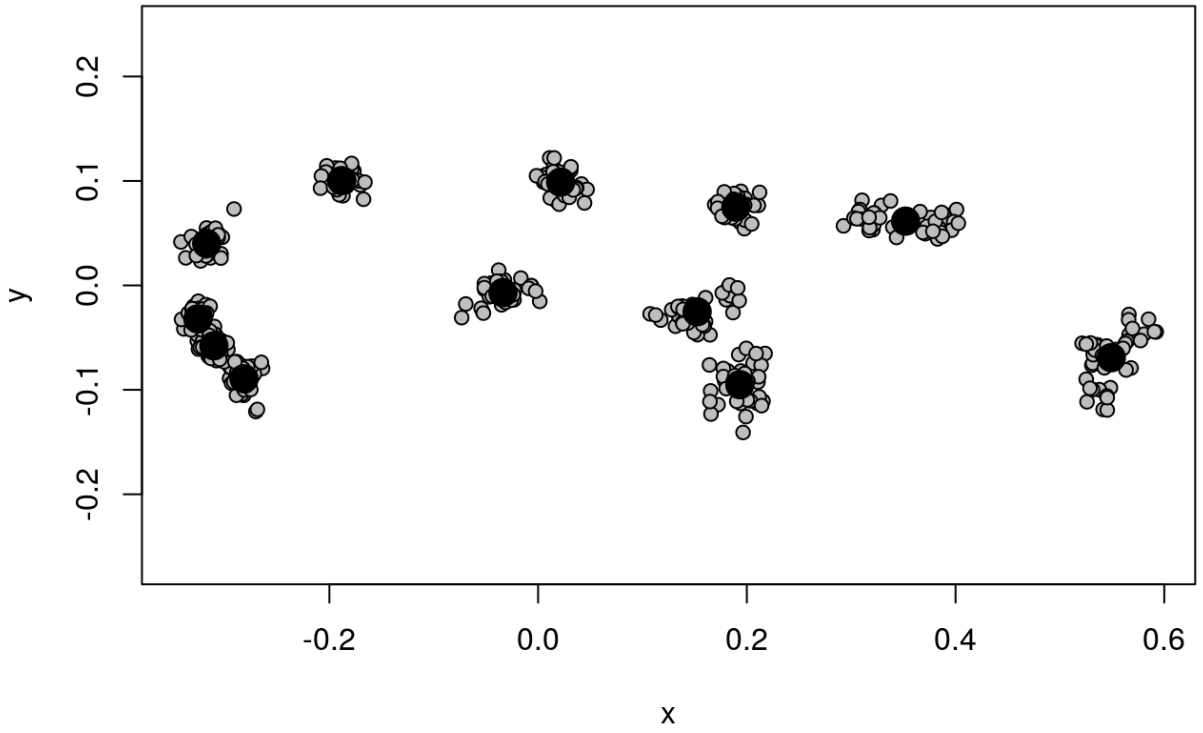


Рис. 30.8 из Zuur et al. 2007 с изменениями

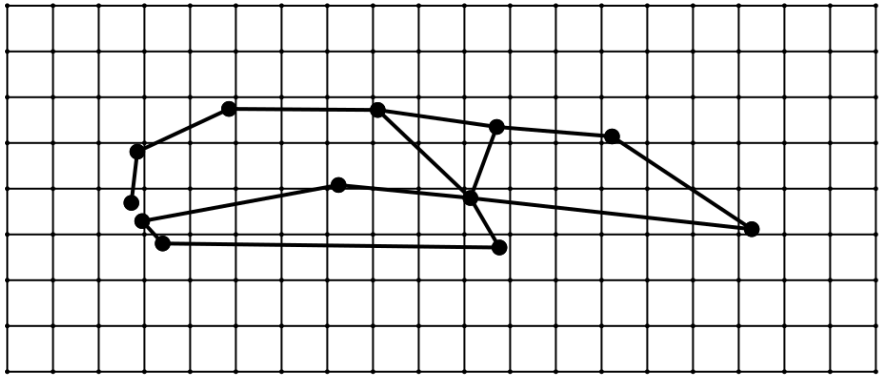
Выравниваем головы саламандр

```
Y.gpa <- gpagen(plethodon$land)
plotAllSpecimens(Y.gpa$coords,links=plethodon$links)
```



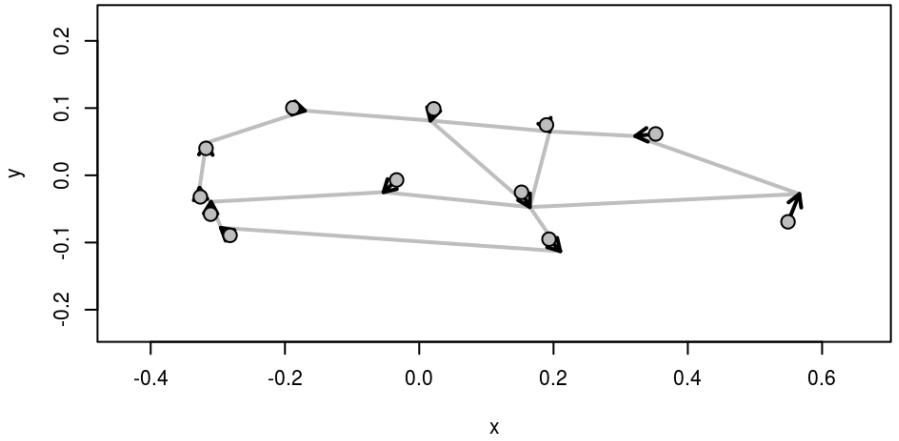
Усредненная форма

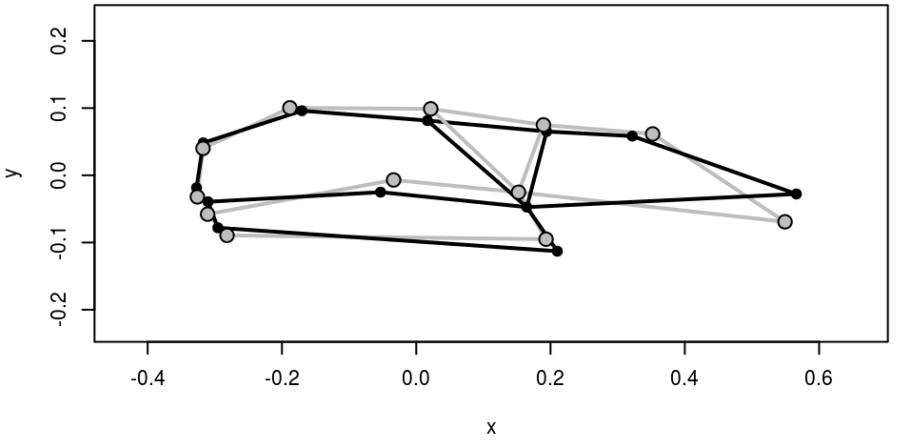
```
ref <- mshape(Y.gpa$coords)
plotRefToTarget(ref, ref, method = "TPS", links = plethodon$links)
```

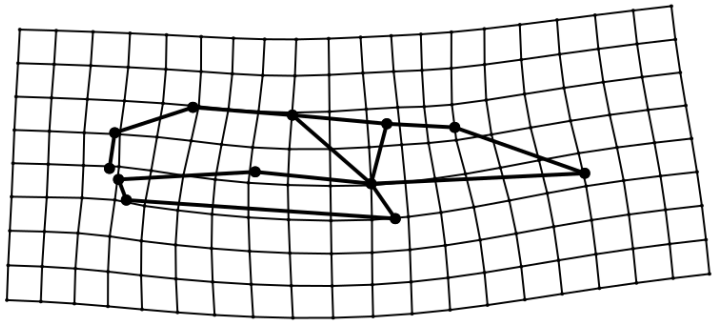


Можем посмотреть, как отличается любой из образцов от усредненной формы

Изменение формы можно представить графически несколькими способами







Код для графиков сравнения образцов с усредненной формой

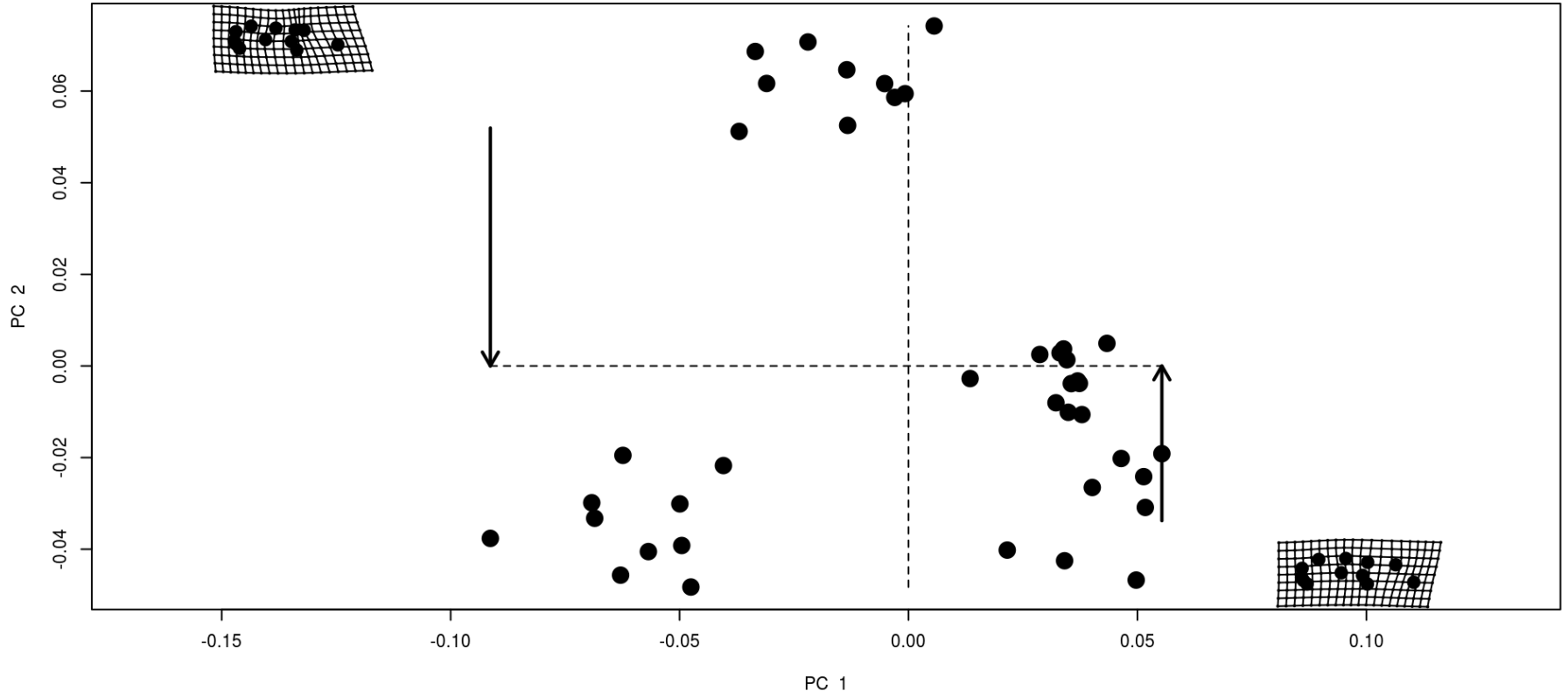
```
# матрица, в которой хранится разметка общего графика
m <- matrix(data = c(1, 1, 2, 2,
                     3, 3, 3, 3,
                     3, 3, 3, 3),
            nrow = 3,
            ncol = 4,
            byrow = TRUE)
l <- layout(m, heights = c(3, 2))
# layout.show(l) # можно просмотреть разметку

# Графики
op <- par( mar = c(4, 4, 1, 1))
# 1) изменение конфигурации обозначено векторами
plotRefToTarget(ref, Y.gpa$coords[, , 11],
                method = "vector", mag = 1,
                links = plethodon$links)
# 2) формы обозначены точками
plotRefToTarget(ref, Y.gpa$coords[, , 11],
                method = "points", mag = 1,
                links = plethodon$links)
# 3) сплайн
plotRefToTarget(ref, Y.gpa$coords[, , 11],
                method = "TPS", mag = 1,
                links = plethodon$links)
par(op)
```

Шаг 2. Создаем морфопространство

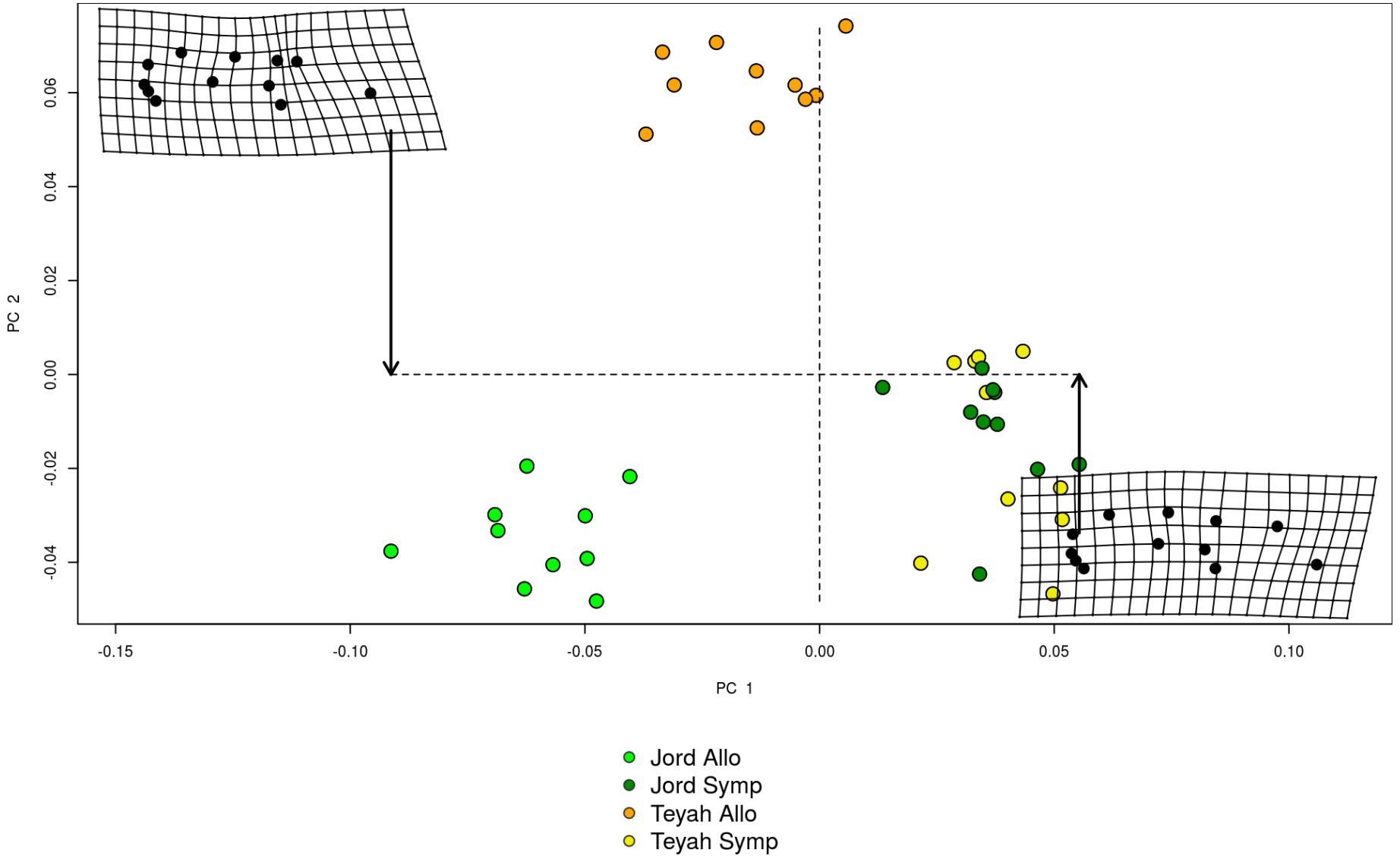
Анализ главных компонент по координатам меток для выровненных образцов. Главные компоненты отражают изменения формы.

```
plotTangentSpace(Y.gpa$coords)
```




```
## Importance of components:
```

Можно раскрасить по группам



Код для графика ординации и для легенды

```
op <- par(mar = c(4, 4, 0, 0))
gp <- as.factor(paste(plethodon$species, plethodon$site)) # группа должна быть фактором
# задаем соответствие цветов уровням фактора
colvec <- c("Jord Allo" = "yellow2",
            "Jord Symp" = "orange",
            "Teyah Allo" = "green4",
            "Teyah Symp" = "green1")
# вектор цветов в порядке заданном фактором gp
colvec <- colvec[match(gp, names(colvec))]
res <- plotTangentSpace(Y.gpa$coords, groups = colvec, verbose = TRUE)
par(op)


# легенда
op <- par(mar = c(0, 0, 0, 0))
plot.new(); legend("center", legend = levels(gp),
                   bty = "n", pch = 21,
                   col = "grey20",
                   pt.bg = levels(as.factor(colvec)))
par(op)
```

Задание:

Исследуйте структуру объекта результатов

- Сколько процентов изменчивости объясняют первые 2 или 3 компоненты?
- Как изменяется форма вдоль 2 компоненты в отрицательном и положительном направлении относительно средней формы? Постройте график

Решение: Структура результатов

- `$pc.summary` - результаты анализа главных компонент
- `$pc.scores` - факторные координаты образцов
- `$pc.shapes` - формы на противоположных концах главных компонент

```
str(res, max.level = 2, vec.len = 2, give.attr = FALSE)
```

```
## List of 3
##  $ pc.summary:List of 6
##   ..$ sdev      : num [1:24] 0.0431 0.0396 ...
##   ..$ rotation  : num [1:24, 1:24] -0.185 0.054 ...
##   ..$ center    : num [1:24] 0.1523 -0.0252 ...
##   ..$ scale     : logi FALSE
##   ..$ x         : num [1:40, 1:24] -0.036993 -0.000749 ...
##   ..$ importance: num [1:3, 1:24] 0.0431 0.3674 ...
##  $ pc.scores : num [1:40, 1:24] -0.036993 -0.000749 ...
##  $ pc.shapes :List of 4
##   ..$ PC1min: num [1:12, 1:2] 0.169 0.217 ...
##   ..$ PC1max: num [1:12, 1:2] 0.142 0.179 ...
##   ..$ PC2min: num [1:12, 1:2] 0.131 0.186 ...
##   ..$ PC2max: num [1:12, 1:2] 0.185 0.204 ...
```

Решение: Доля объясненной изменчивости

```
res$pc.summary$importance[, 1:5] # Доля изменчивости объясненной 1-5 компонентами
```

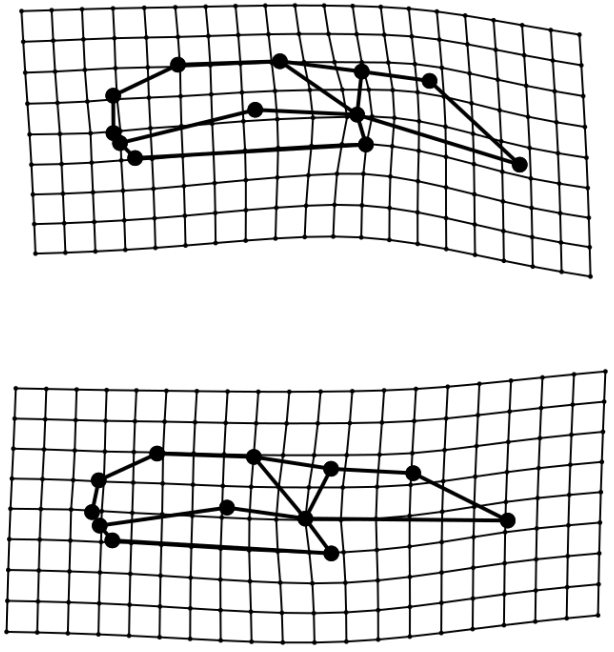
```
##                            PC1    PC2    PC3    PC4    PC5
## Standard deviation     0.0431 0.0396 0.0203 0.0151 0.0131
## Proportion of Variance 0.3674 0.3102 0.0820 0.0451 0.0342
## Cumulative Proportion  0.3674 0.6777 0.7597 0.8048 0.8390
```

```
head(res$pc.scores[, 1:5]) # Факторные координаты по 1-5 компонентам
```

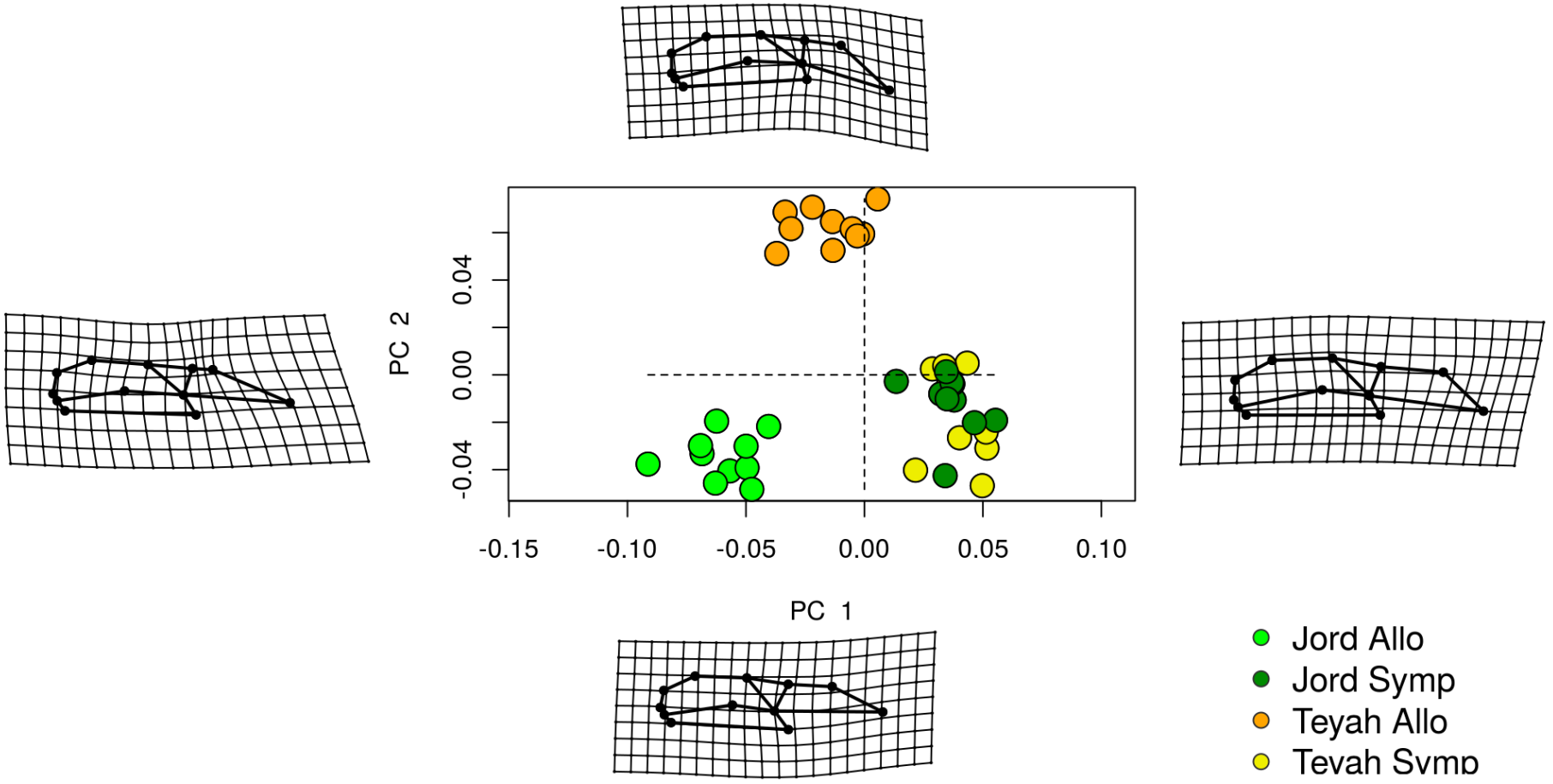
```
##             PC1    PC2       PC3      PC4      PC5
## [1,] -0.036993 0.0512 -0.001697 -0.00313 -0.01094
## [2,] -0.000749 0.0594  0.000137 -0.00277 -0.00812
## [3,]  0.005600 0.0742 -0.005261 -0.00503 -0.00275
## [4,] -0.013481 0.0646 -0.045844 -0.00789  0.00982
## [5,] -0.033470 0.0686  0.013629  0.00736  0.02235
## [6,] -0.005215 0.0616 -0.029933 -0.00575 -0.02406
```

Решение: Изменение формы вдоль 2 компоненты относительно средней формы

```
op <- par(mfrow = c(2, 1), mar = c(1, 1 , 1, 1))
# изменение формы вдоль 2 компоненты в положительном направлении
plotRefToTarget(M1=ref, M2=res$pc.shapes$PC2max, method="TPS", links = plethodon$links)
# изменение формы вдоль 2 компоненты в отрицательном направлении
plotRefToTarget(M1=ref, M2=res$pc.shapes$PC2min, method="TPS", links = plethodon$links)
par(op)
```



Можно нарисовать одновременно изменение формы вдоль обеих компонент и ординацию



Код для графика изменения форм вдоль первых двух главных компонент

```
mat <- matrix(c(0, 1, 0,
                2, 6, 3,
                0, 4, 5),
              nrow = 3, ncol = 3, byrow = TRUE)
l <- layout(mat, widths = c(1, 2, 1), heights = c(1, 3, 1))
# layout.show(l)
op <- par(mar = c(0, 0 , 0, 0)) # параметры для 1-4 графиков
# графики форм (слева, справа, снизу, сверху)
plotRefToTarget(M1 = ref, M2 = res$pc.shapes$PC2max, method = "TPS", links =
plethodon$links)
plotRefToTarget(M1 = ref, M2 = res$pc.shapes$PC1min, method = "TPS", links =
plethodon$links)
plotRefToTarget(M1 = ref, M2 = res$pc.shapes$PC1max, method = "TPS", links =
plethodon$links)
plotRefToTarget(M1 = ref, M2 = res$pc.shapes$PC2min, method = "TPS", links =
plethodon$links)
# легенда снизу слева
par(mar = c(0.5, 0, 0, 0)) # параметры для легенды
plot.new(); legend("center", legend = levels(gp), bty = "n", pch = 21,
                   col = "grey20",pt.bg = levels(as.factor(colvec)), cex = 2)
# в центре
par(mar = c(4, 4 , 1, 1), cex = 1) # параметры для последнего графика
plotTangentSpace(Y.gpa$coords, warpgrids = FALSE, groups = colvec)
par(op)
```

Эволюционные изменения формы

Фило-морфо пространство

Если у вас есть данные о средних формах для каждого вида и данные о филогении (из любого источника), то можно изобразить эволюционные изменения формы

Этапы:

1. Выравнивание средних форм для таксонов при помощи обобщенного прокрустова анализа
2. Ординация таксонов при помощи анализа главных компонент
3. Поиск анцестральных состояний количественных признаков (форм) методом максимального правдоподобия
4. Наложение филогенетического дерева и анцестральных форм на график ординации

Фило-морфопространство саламандр рода Plethodon

P. serratus, P. cinereus, P. shenandoah, P. hoffmani, P. virginia, P. nettingi, P. hubrichti, P. electromorphus, P. richmondi

```
data(plethspecies)
str(plethspecies, vec.len = 2, give.attr = F)
```

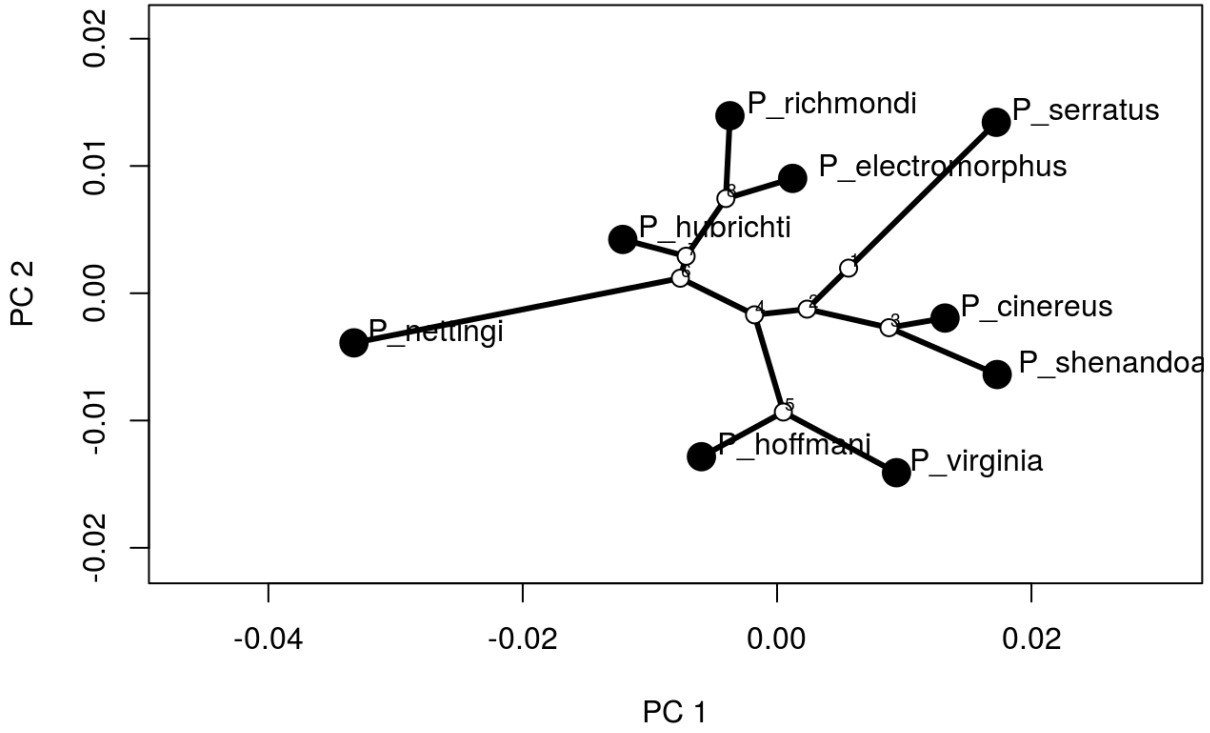
```
## List of 2
##  $ land: num [1:11, 1:2, 1:9] 0.217 0.259 ...
##  $ phy :List of 4
##   ..$ edge       : int [1:16, 1:2] 10 10 11 12 12 ...
##   ..$ Nnode      : int 8
##   ..$ tip.label  : chr [1:9] "P_serratus" "P_cinereus" ...
##   ..$ edge.length: num [1:16] 15.17 3.84 ...
```

Выравниваем средние формы для видов

```
Yphyl.gpa <- gpagen(plethspecies$land) #GPA-alignment
```

Наложение филогенетического дерева и анцестральных форм на график PCA ординации

```
plotGMPhyloMorphoSpace(plethspecies$phy, Yphyl.gpa$coords)
```




```
##    [,1]    [,2]  [,3]    [,4]    [,5]    [,6]   [,7]    [,8]   [,9]
## 1 0.213 -0.0192 0.247 -0.0789 -0.0252 -0.0150 -0.263 -0.0952 -0.289
## 2 0.214 -0.0206 0.249 -0.0795 -0.0258 -0.0155 -0.263 -0.0947 -0.290
```

Take home messages

- Классический подход к морфометрии
  - анализируют расстояния между метками
  - для корректного анализа необходимо удалить влияние размера и оставить форму, но сделать это корректно почти невозможно
- Геометрическая морфометрия
  - анализируют координаты меток
  - различные конфигурации выравнивают при помощи обобщенного прокрустова анализа
  - преобразованные координаты точек используют в анализе главных компонент
  - чтобы визуализировать эволюцию форм, можно наложить филогенетическое древо на ординацию

Дополнительные ресурсы

- Bookstein, F.L., 2003. Morphometric Tools for Landmark Data Geometry and Biology. Cambridge University Press.
- Claude, J., 2008. Morphometrics With R. Springer.
- GEOL G562 - Geometric Morphometrics [WWW Document], n.d. URL http://www.indiana.edu/~g562/PBDB2013/ (accessed 4.1.15).
- Zelditch, M., Swiderski, D.L., Sheets, D.H., Fink, W.L., 2004. Geometric Morphometrics for Biologists. Academic Press.
- Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analysing ecological data. Springer.