Корреспондентный анализ и анализ главных компонент

Анализ и визуализация многомерных данных с использованием R

Вадим Хайтов, Марина Варфоломеева

Корреспондентрый анализ и анализ главных компонент

- Сложности при анализе видового состава сообществ при помощи анализа главных компонент
  - Анализ сырых данных
  - Трансформация Хеллингера
  - Расстояние хорды
- Корреспондентный анализ
  - Анализ таблиц сопряженности, хи-квадрат
  - Оси в корреспондентном анализе
  - Интерпретация графиков в корреспондентном анализе

Вы сможете

- Избавляться от "эффекта подковы" в анализе главных компонент при помощи трансформаций данных
- Проводить корреспондентный анализ таблиц сопряженности
- Объяснить, что именно означает взаиморасположение точек объектов и переменных на графиках результатов корреспондентного анализа
- Интерпретировать графики результатов корреспондентного анализа

Анализ видового состава сообществ. Трансформации данных

Пример: Птицы в лесах Австралии

Обилие 102 видов птиц в 37 сайтах в юго-восточной Австралии (Mac Nally, 1989; данные из Quinn, Keough, 2002). Можно ли описать отношения между сайтами небольшим числом главных компонент?

```
library(readxl)
birds <- read_excel(path = "data/macnally.xlsx")
str(birds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    37 obs. of  104 variables:
##  $ SITE    : chr  "Reedy Lake" "Pearcedale" "Warneet" "Cranbourne" ...
##  $ HABITAT : chr  "Mixed" "Gippsland Ma" "Gippsland Ma" "Gippsland Ma" ...
##  $ V1GST   : num  3.4 3.4 8.4 3 5.6 8.1 8.3 4.6 3.2 4.6 ...
##  $ V2EYR   : num  0 9.2 3.8 5 5.6 4.1 7.1 5.3 5.2 1.2 ...
##  $ V3GF    : num  0 0 0.7 0 12.9 10.9 6.9 11.1 8.3 4.6 ...
##  $ V4BTH   : num  0 0 2.8 5 12.2 24.5 29.1 28.2 18.2 6.5 ...
##  $ V5GWH   : num  0 0 0 2 9.5 5.6 4.2 3.9 3.8 2.3 ...
##  $ V6WTTR  : num  0 0 0 0 2.1 6.7 2 6.5 4.2 5.2 ...
##  $ V7WEHE  : num  0 0 10.7 3 7.9 9.4 7.1 2.6 2.8 0.6 ...
##  $ V8WNHE  : num  11.9 11.5 12.3 10 28.6 6.7 27.4 10.9 9 3.6 ...
##  $ V9SFW   : num  0.4 8.3 4.9 6.9 9.2 0 13.1 3.1 3.8 3.8 ...
##  $ V10WBSW : num  0 12.6 10.7 12 5 8.9 2.8 8.6 5.6 3 ...
##  $ V11CR   : num  1.1 0 0 0 19.1 12.1 0 9.3 14.1 7.5 ...
##  $ V12LK   : num  3.8 0.5 1.9 2 3.6 6.7 2.8 3.8 3.2 2.4 ...
##  $ V13RWB  : num  9.7 11.6 16.6 11 5.7 2.7 2.4 0.6 0 0.6 ...
##  $ V14AUR  : num  0 0 2.3 1.5 8.8 0 2.8 1.3 0 0 ...
##  $ V15STTH : num  0 0 2.8 0 7 16.8 13.9 10.2 12.2 11.3 ...
##  $ V16LR   : num  4.8 3.7 5.5 11 1.6 3.4 0 0 0.6 5.8 ...
##  $ V17WPHE : num  27.3 27.6 27.5 20 0 0 16.7 0 0 0 ...
##  $ V18YTH  : num  0 0 0 0 0 0 0 0 0 9.6 ...
##  $ V19ER   : num  5.1 2.7 5.3 2.1 1.4 2.2 0 1.2 1.3 2.3 ...
##  $ V20PCU  : num  0 0 0 0 0 0 0 0 2.8 2.9 ...
##  $ V21ESP  : num  0 3.7 0 2 3.5 3.4 5.5 5.1 7.1 0.6 ...
```

```r
# названия переменных
colnames(birds)
```

```
##   [1] "SITE"     "HABITAT"  "V1GST"    "V2EYR"    "V3GF"
##   [6] "V4BTH"    "V5GWH"    "V6WTTR"   "V7WEHE"   "V8WNHE"
##  [11] "V9SFW"    "V10WBSW"  "V11CR"    "V12LK"    "V13RWB"
##  [16] "V14AUR"   "V15STTH"  "V16LR"    "V17WPHE"  "V18YTH"
##  [21] "V19ER"    "V20PCU"   "V21ESP"   "V22SCR"   "V23RBFT"
##  [26] "V24BFCS"  "V25WAG"   "V26WWCH"  "V27NHHE"  "V28VS"
##  [31] "V29CST"   "V30BTR"   "V31AMAG"  "V32SCC"   "V33RWH"
##  [36] "V34WSW"   "V35STP"   "V36YFHE"  "V37WHIP"  "V38GAL"
##  [41] "V39FHE"   "V40BRTH"  "V41SPP"   "V42SIL"   "V43GCU"
##  [46] "V44MUSK"  "V45MGLK"  "V46BHHE"  "V47RFC"   "V48YTBC"
##  [51] "V49LYRE"  "V50CHE"   "V51OWH"   "V52TRM"   "V53MB"
##  [56] "V54STHR"  "V55LHE"   "V56FTC"   "V57PINK"  "V58OBO"
##  [61] "V59YR"    "V60LFB"   "V61SPW"   "V62RBTR"  "V63DWS"
##  [66] "V64BELL"  "V65LWB"   "V66CBW"   "V67GGC"   "V68PIL"
##  [71] "V69SKF"   "V70RSL"   "V71PDOV"  "V72CRP"   "V73JW"
##  [76] "V74BCHE"  "V75RCR"   "V76GBB"   "V77RRP"   "V78LLOR"
##  [81] "V79YTHE"  "V80RF"    "V81SHBC"  "V82AZKF"  "V83SFC"
##  [86] "V84YRTH"  "V85ROSE"  "V86BCOO"  "V87LFC"   "V88WG"
##  [91] "V89PCOO"  "V90WTG"   "V91NMIN"  "V92NFB"   "V93DB"
##  [96] "V94RBEE"  "V95HBC"   "V96DF"    "V97PCL"   "V98FLAME"
## [101] "V99WWT"   "V100WBWS" "V101LCOR" "V102KING"
```

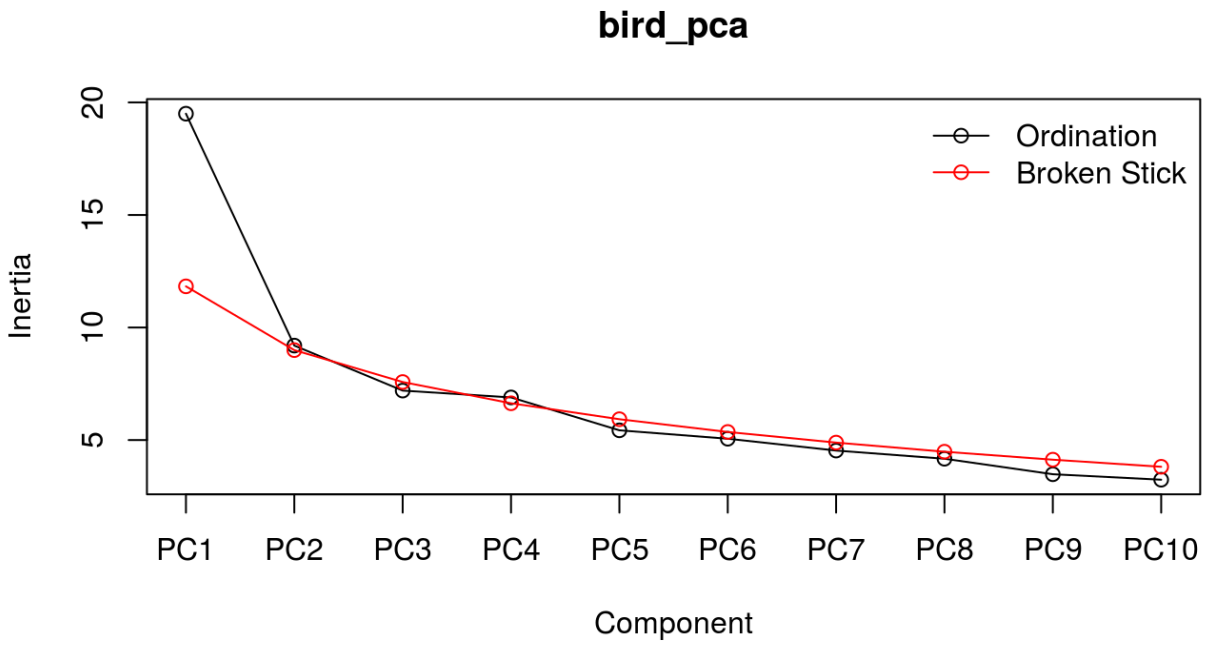
```r
# имена переводим в нижний регистр
colnames(birds) <- tolower(colnames(birds))
# есть ли пропущенные значения
any(!complete.cases(birds))
```

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

Задание: Проведите анализ главных компонент

Результаты анализа главных компонент

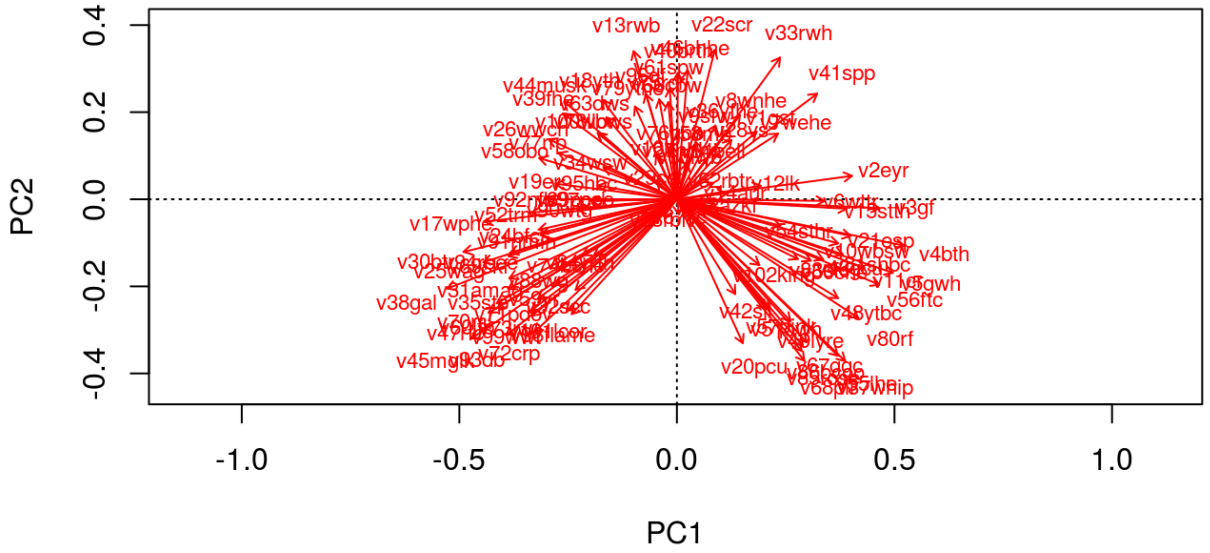
```
library(vegan)
bird_pca <- rda(birds[ , -c(1, 2)], scale = TRUE)
# summary(bird_pca)
screeplot(bird_pca, type = "lines", bstick = TRUE) # график собственных чисел
```



- Первые две компоненты объясняют умеренное количество изменчивости

Факторные нагрузки

```
biplot(bird_pca, display = "species", scaling = 2, type = "t")
```

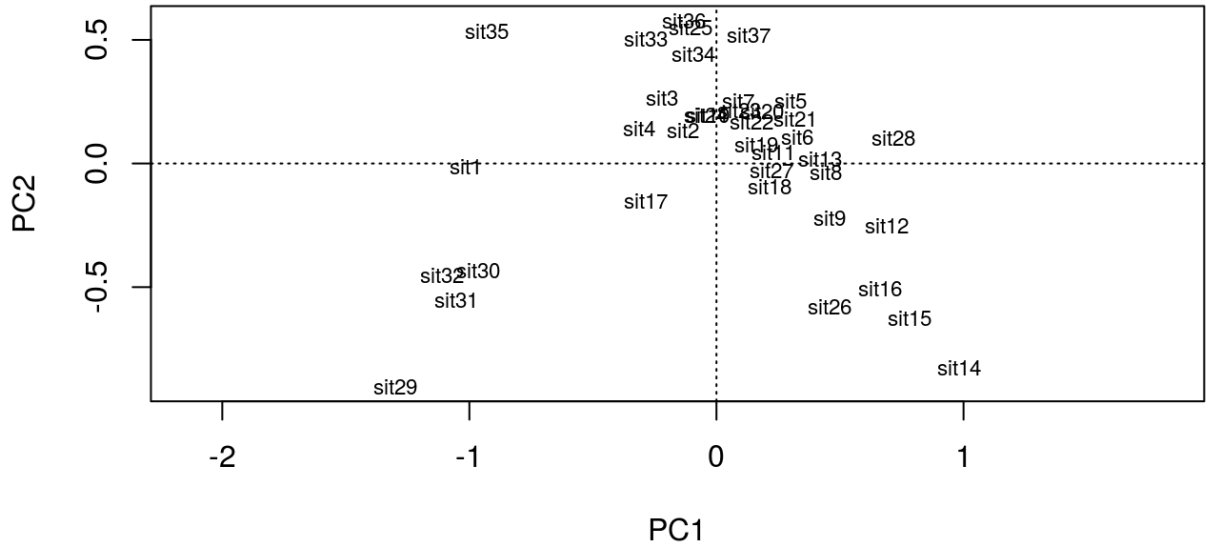


- У многих переменных факторные нагрузки велики сразу на две оси. Это может быть неудобно.

Обратите внимание, график в виде подковы!

Сайты 29 и 14 на самом деле расположены далеко друг от друга и мало похожи. Почему же они сближены на графике?

```
biplot(bird_pca, display = "sites", scaling = 1) # биплот расстояний
```



- Так происходит от того, что завышены корреляции между переменными из-за большого числа нулей

Чтобы исчез "эффект подковы" нужна трансформация исходных данных

- расстояние Хеллингера (Hellinger distance)
- хордальное расстояние (chord distance)

```
birds_h <- decostand(birds[ , -c(1, 2)], "hellinger")
birds_ch <- decostand(birds[ , -c(1, 2)], "norm") # chord distance
```

Задание: Проведите анализ главных компонент по трансформированным данным

Сравните долю дисперсии, объясненной первыми двумя компонентами с результатами анализа нетрансформированных данных.

- В каком случае объясненная дисперсия больше?

Сравните получившиеся ординации объектов.

- Исчез ли "эффект подковы" после трансформации?
- Изменилась ли группировка объектов?

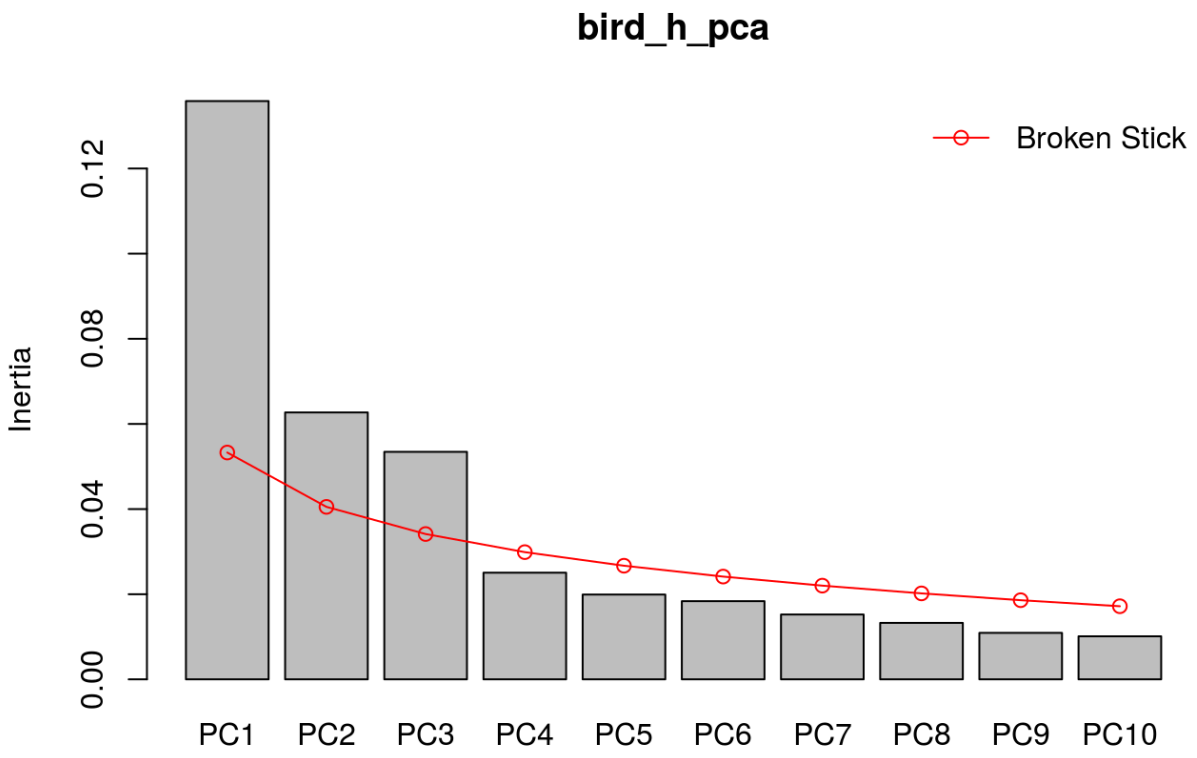
Анализ главных компонент

```
bird_h_pca <- rda(birds_h)
summary(bird_h_pca)
```

```
##
## Call:
## rda(X = birds_h)
##
## Partitioning of variance:
##               Inertia Proportion
## Total           0.459          1
## Unconstrained   0.459          1
##
## Eigenvalues, and their contribution to the variance
##
## Importance of components:
##                            PC1    PC2    PC3    PC4    PC5    PC6    PC7
## Eigenvalue              0.136 0.0627 0.0534 0.0250 0.0199 0.0184 0.0152
## Proportion Explained    0.296 0.1365 0.1163 0.0545 0.0433 0.0400 0.0331
## Cumulative Proportion   0.296 0.4321 0.5484 0.6029 0.6463 0.6863 0.7194
##                            PC8    PC9   PC10    PC11    PC12    PC13
## Eigenvalue              0.0133 0.0109 0.0101 0.00893 0.00821 0.00812
## Proportion Explained    0.0289 0.0238 0.0220 0.01943 0.01787 0.01767
## Cumulative Proportion   0.7482 0.7720 0.7940 0.81343 0.83129 0.84896
##                           PC14    PC15    PC16    PC17   PC18    PC19
## Eigenvalue              0.00678 0.00653 0.00582 0.00547 0.0051 0.00479
## Proportion Explained    0.01476 0.01421 0.01268 0.01190 0.0111 0.01043
## Cumulative Proportion   0.86373 0.87794 0.89062 0.90252 0.9136 0.92405
##                           PC20    PC21    PC22    PC23    PC24    PC25
## Eigenvalue              0.00416 0.00362 0.00331 0.00311 0.00305 0.00241
## Proportion Explained    0.00906 0.00787 0.00720 0.00676 0.00664 0.00524
## Cumulative Proportion   0.93311 0.94098 0.94818 0.95495 0.96158 0.96682
##                           PC26    PC27    PC28    PC29    PC30    PC31
## Eigenvalue              0.00225 0.00199 0.00195 0.00161 0.00156 0.00135
```

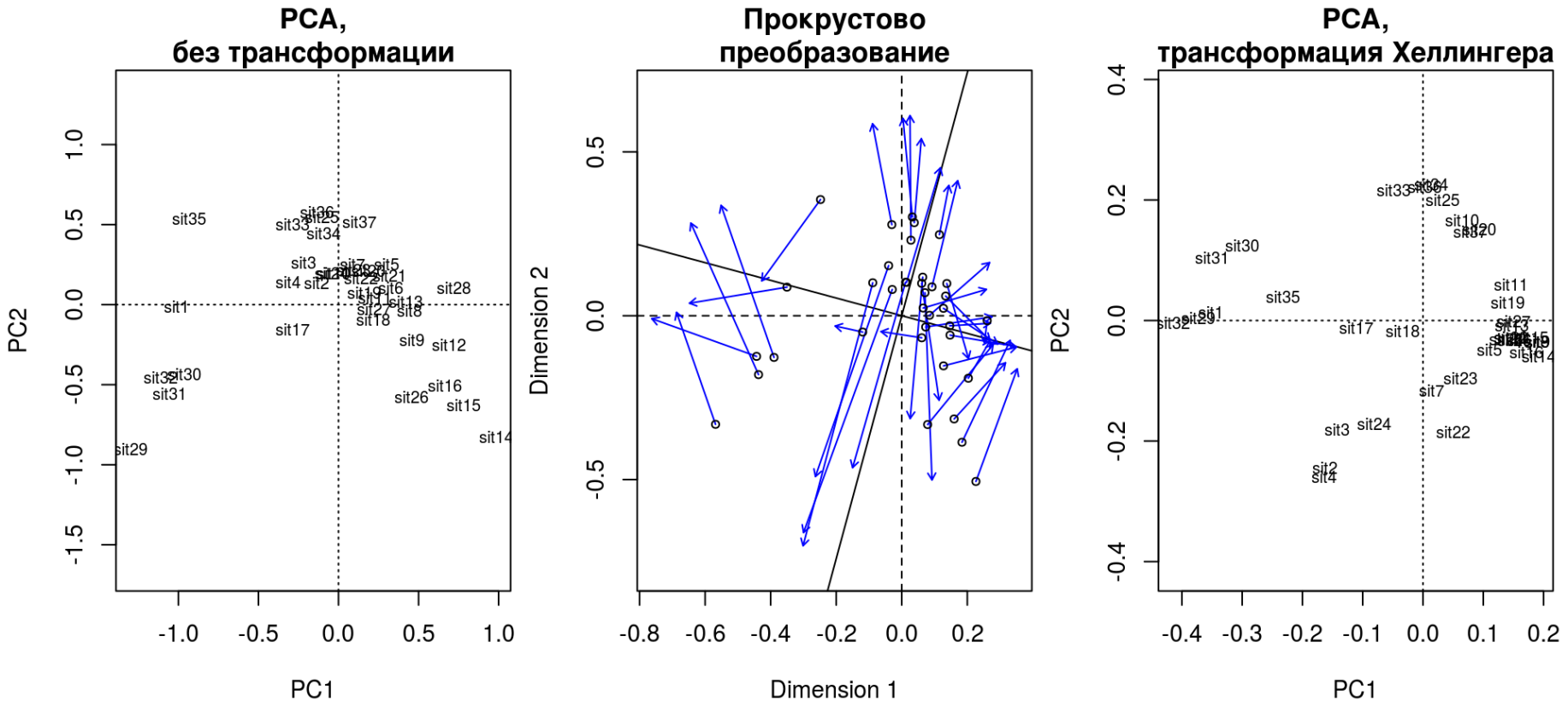
Собственные числа

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



Ординация до и после трансформации данных

```
op <- par(mfrow = c(1, 3), cex = 0.9, mar = c(4, 4, 2.5, 0.5))
biplot(bird_pca, display = "sites", scaling = 1, main = "PCA,\nбез трансформации")
plot(procrustes(bird_h_pca, bird_pca), main = "Прокрустово\nпреобразование")
biplot(bird_h_pca, display = "sites", scaling = 1, main = "PCA,\nтрансформация Хеллингера")
par(op)
```



Для успешного применения анализа главных компонент нужно:

- Линейные связи между переменными (т.к. матрица корреляций или ковариаций)
- Исключить наблюдения, в которых есть пропущенные значения
- Если много нулей - трансформация данных
- Если очень много нулей - удалить такие переменные из анализа

Корреспондентный анализ

Пример: Крысы

Число грызунов разных видов в нескольких сайтах в южной Калифорнии (Bolger et al. 1997). Некоторые из этих сайтов оказались изолированы из-за урбанизации. Кроме того, несколько сайтов в исследовании из нефрагментированной местности.

```
rats <- read.csv("data/bolger1.csv")
head(rats, 2)
```

```
##       SITE TYPE RRATTUS MMUSCUL PCALIFO PEREMIC RMEGALO NFUSCIP
## 1  Florida FRAG       0      13       3       1       1       2
## 2 Sandmark FRAG       0       1      57      65       9      16
##   NLEPIDA PFALLAX MCALIFO
## 1       0       0       0
## 2       8       2       3
```

```
# имена переводим в нижний регистр
colnames(rats) <- tolower(colnames(rats))
# есть ли пропущенные значения
any(!complete.cases(rats))
```

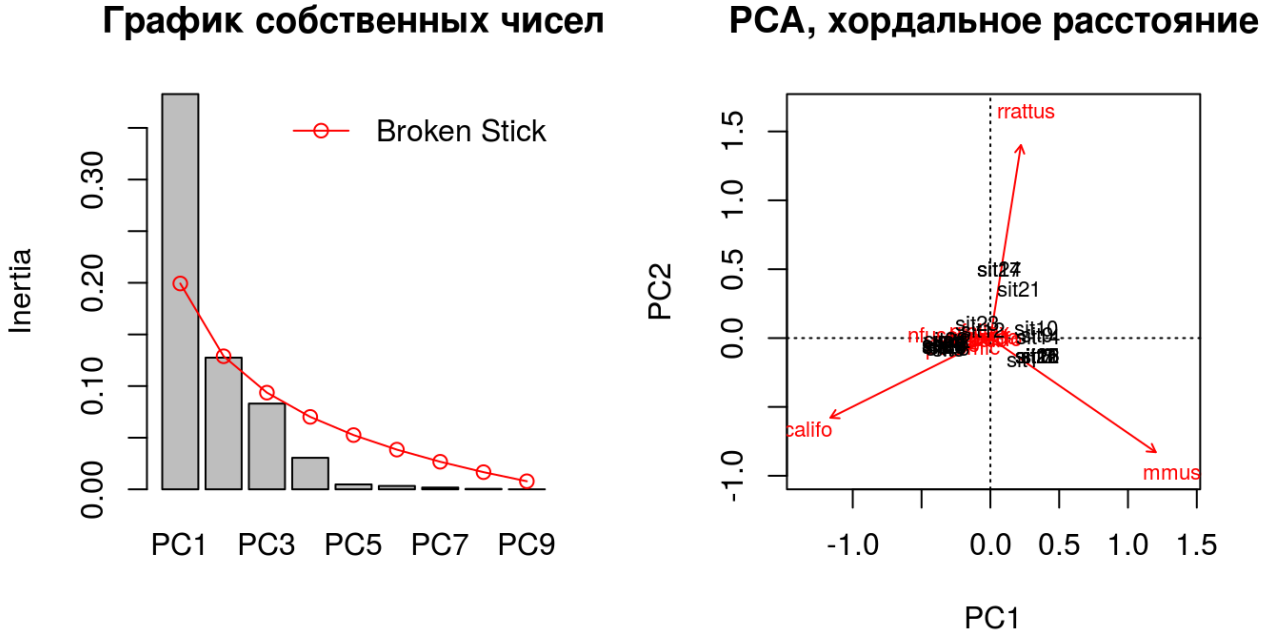
```
## [1] FALSE
```

Данные из Quinn, Keough, 2002

Задание: проведите анализ главных компонент

- Используйте хордальное расстояние
- Нарисуйте биплот расстояний

```r
rats_ch <- decostand(rats[ , -c(1, 2)], "norm") # chord distance
rats_ch_pca <- rda(rats_ch)
# summary(rats_ch_pca)
op <- par(mfrow = c(1, 2))
screeplot(rats_ch_pca, bstick = TRUE, main = "График собственных чисел")
biplot(rats_ch_pca, scaling = 1, main = "PCA, хордальное расстояние")
par(op)
```



Анализ таблиц сопряженности

Корреспондентный анализ был придуман для анализа сводных таблиц вроде этой:

Горох	Желтый	Зеленый
Гладкий	99	42
Морщинистый	29	13

Ожидаемые частоты

Таблица наблюдаемых частот

Горох	Желтый	Зеленый	Сумма
Гладкий	99	42	141
Морщинистый	29	13	42
Сумма	128	55	183

Ожидаемая частота желтого и гладкого, если эти признаки независимы:

вероятность быть желтым * вероятность быть гладким * общее число горошин

$$\frac{141}{183} \times \frac{128}{183} \times 183 = \frac{141 \times 128}{183} = 98.6$$

Можно посчитать ожидаемые частоты в каждой ячейке

Таблица наблюдаемых частот

Горох	Желтый	Зеленый	Сумма
Гладкий	99	42	141
Морщинистый	29	13	42
Сумма	128	55	183

Ожидаемая частота желтого и гладкого, если эти признаки независимы:

вероятность быть желтым * вероятность быть гладким * общее число горошин

$$\frac{141}{183} \times \frac{128}{183} \times 183 = \frac{141 \times 128}{183} = 98.6$$

Таблица ожидаемых частот

Горох	Желтый	Зеленый
Гладкий	98.6	42.4
Морщинистый	29.4	12.6

Проверяем гипотезу о независимости столбцов и строк

Таблица наблюдаемых частот

Горох	Желтый	Зеленый	Сумма
Гладкий	99	42	141
Морщинистый	29	13	42
Сумма	128	55	183

Таблица ожидаемых частот

Горох	Желтый	Зеленый
Гладкий	98.6	42.4
Морщинистый	29.4	12.6

- Если столбцы и строки независимы, то наблюдаемые частоты не будут отличаться от ожидаемых
- Для каждой ячейки: $\chi^2 = \sum \frac{(\text{наблюд.частота} - \text{ожид.частота})^2}{\text{ожид.частота}}$
- Общий хи-квадрат - это сумма по таблице $\chi^2 = \sum \chi^2_{ij}$

Корреспондентный анализ

хи-квадрат - это мера независимости переменных (строк и столбцов)

Корреспондентный анализ помогает визуализировать матрицу хи-квадратов, если переменных очень много

Механика похожа на анализ главных компонент (вернее, SVD - singular value decomposition)

Вместо столбцов и строк исходных данных получатся новые переменные - главные оси (principal axes)

Свойства главных осей

- Главные оси независимы друг от друга (перпендикулярны)
- Каждая последующая объясняет меньше общей инерции (общего хи-квадрат, а не изменчивости!!!)
- Всего осей может быть не больше чем минимальное из этих значений: (число строк - 1), (число столбцов - 1)
- Первая ось - переменные, которые объясняют максимум зависимости строк от столбцов (значения которых сильнее всего отличаются от ожидаемых для данных объектов)
- Результаты изображаются в виде точечных графиков, похожих на биплоты (осторожно, scaling!)

Корреспондентный анализ данных про крыс

```
rats_ca <- cca(rats[ , -c(1, 2)])
summary(rats_ca)
```

```
##
## Call:
## cca(X = rats[, -c(1, 2)])
##
## Partitioning of mean squared contingency coefficient:
##               Inertia Proportion
## Total            1.72          1
## Unconstrained    1.72          1
##
## Eigenvalues, and their contribution to the mean squared contingency coefficient
##
## Importance of components:
##                            CA1   CA2   CA3    CA4    CA5    CA6    CA7
## Eigenvalue              0.746 0.459 0.288 0.1528 0.0357 0.0246 0.0113
## Proportion Explained    0.434 0.267 0.167 0.0889 0.0207 0.0143 0.0066
## Cumulative Proportion   0.434 0.701 0.868 0.9572 0.9779 0.9922 0.9989
##                            CA8
## Eigenvalue             0.00198
## Proportion Explained   0.00115
## Cumulative Proportion  1.00000
##
## Scaling 2 for species and site scores
## * Species are scaled proportional to eigenvalues
## * Sites are unscaled: weighted dispersion equal on all dimensions
##
##
## Species scores
##
##             CA1     CA2     CA3    CA4      CA5      CA6
## rrattus   2.606  5.2987 -0.0486 -0.205  0.03006 -0.00376
```

Сколько общей инерции объясняют первые две главных оси?

```
eig <- eigenvals(rats_ca)
eig/sum(eig)*100
```

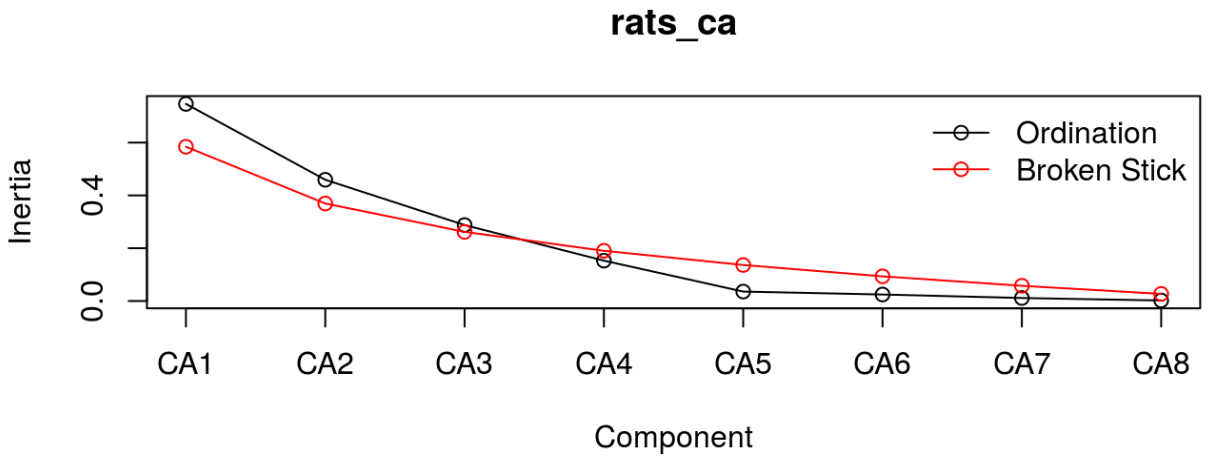
```
##  CA1  CA2  CA3  CA4  CA5  CA6  CA7  CA8
## 43.4 26.7 16.7  8.9  2.1  1.4  0.7  0.1
```

```
cumsum(eig)/sum(eig)*100
```

```
##   CA1   CA2   CA3   CA4   CA5   CA6   CA7   CA8
##  43.4  70.1  86.8  95.7  97.8  99.2  99.9 100.0
```

Сколько главных осей достаточно?

```
screeplot(rats_ca, type = "lines", bstick = TRUE)
```



Scaling

Разные варианты scaling не меняют порядок расположения, но меняют расстояния между объектами.

scaling 1 (для графиков "расстояний")- расстояния между объектами пропорциональны хи-квадрату

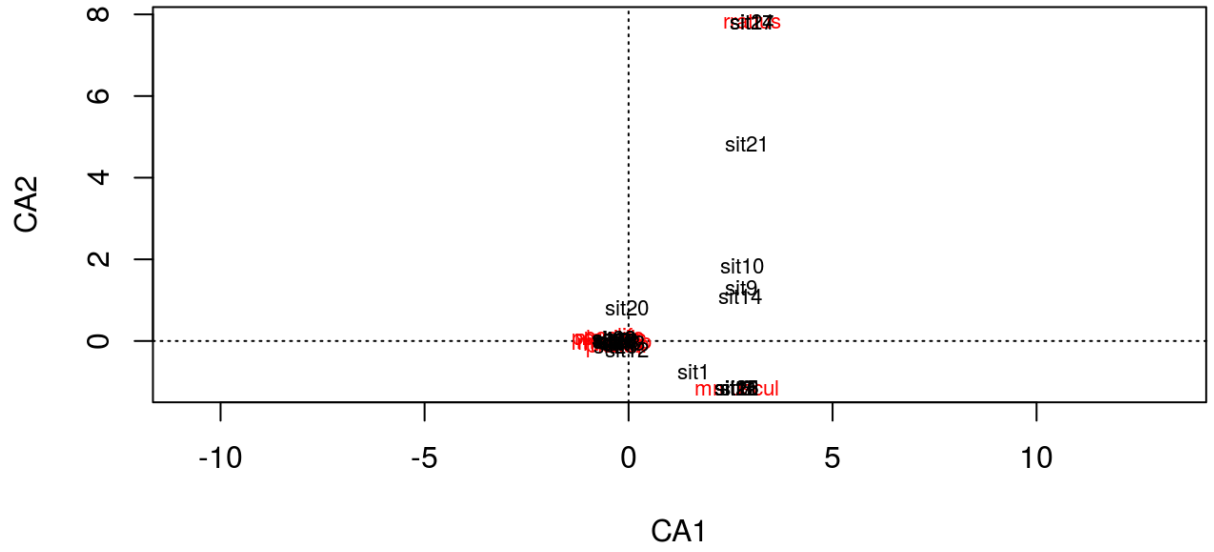
scaling 2 (для графиков "переменных")- расстояния между переменными пропорциональны хи-квадрату

scaling 3 (компромиссный вариант)- нечто среднее между 1 и 2

Биплот расстояний

- Если объект и переменная расположены рядом, то у этого объекта значение переменной выше ожидаемого (ожидаемого при условии независимости объектов и переменных).
- Если объект и переменная расположены далеко, то у этого объекта значение переменной ниже ожидаемого.

```
plot(rats_ca, scaling = 1)
```



Создаем функцию, чтобы быстрее рисовать цветные графики

```
col_ord_plot <- function(ord, scaling = 1, colvec = NULL, colfac, pch = 21,
    lab.cex = 1, leg.cex = 0.9, leg.pos = "bottom", ncol = 1, display.labs = TRUE,
    display.legend = TRUE, ...) {
    if (is.null(colvec)) {
        # создаем вектор цветов
        ncolours <- length(levels(colfac))
        colvec <- rainbow(ncolours, s = 0.8, v = 0.9)
    }
    plot(ord, type = "n", scaling = scaling, ...)  # пустой график
    # точки, раскрашенные по уровням фактора
    points(ord, display = "sites", scaling = scaling, pch = pch, col = colvec[colfac],
        bg = colvec[colfac], ...)
    if (display.labs == TRUE) {
        # подписи переменных
        text(ord, display = "species", scaling = scaling, cex = lab.cex)
    }
    if (display.legend == TRUE) {
        # легенда
        legend(x = leg.pos, legend = levels(colfac), bty = "n", pch = pch,
            col = colvec, pt.bg = colvec, cex = leg.cex, ncol = ncol)
    }
}
```

Графики PCA и CA

```
op <- par(mfrow = c(1, 2), mar = c(4, 4, 2, 1), cex = 0.9)
col_ord_plot(ord = rats_ch_pca, colvec = c("steelblue", "red2"),
             colfac = rats$type, leg.pos = "topleft", main = "PCA, хордальное расстояние")
col_ord_plot(ord = rats_ca, colvec = c("steelblue", "red2"),
             colfac = rats$type, leg.pos = "topleft", main = "CA")
par(op)
```

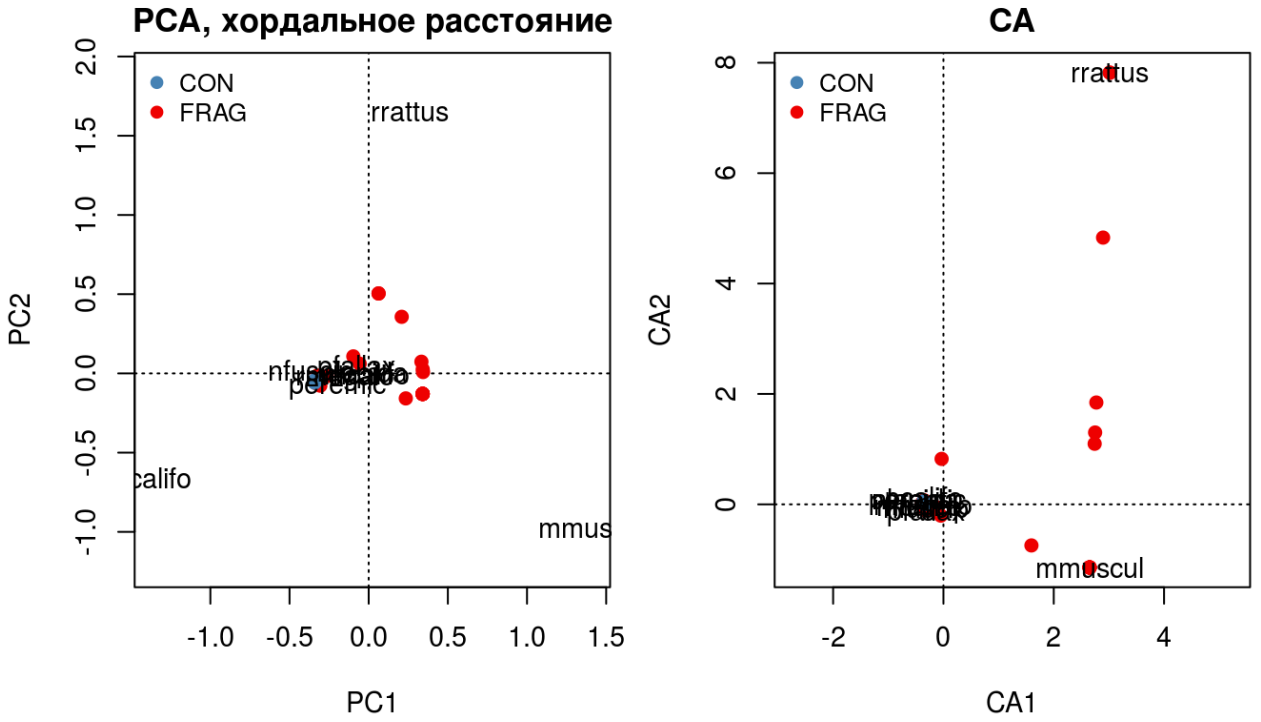
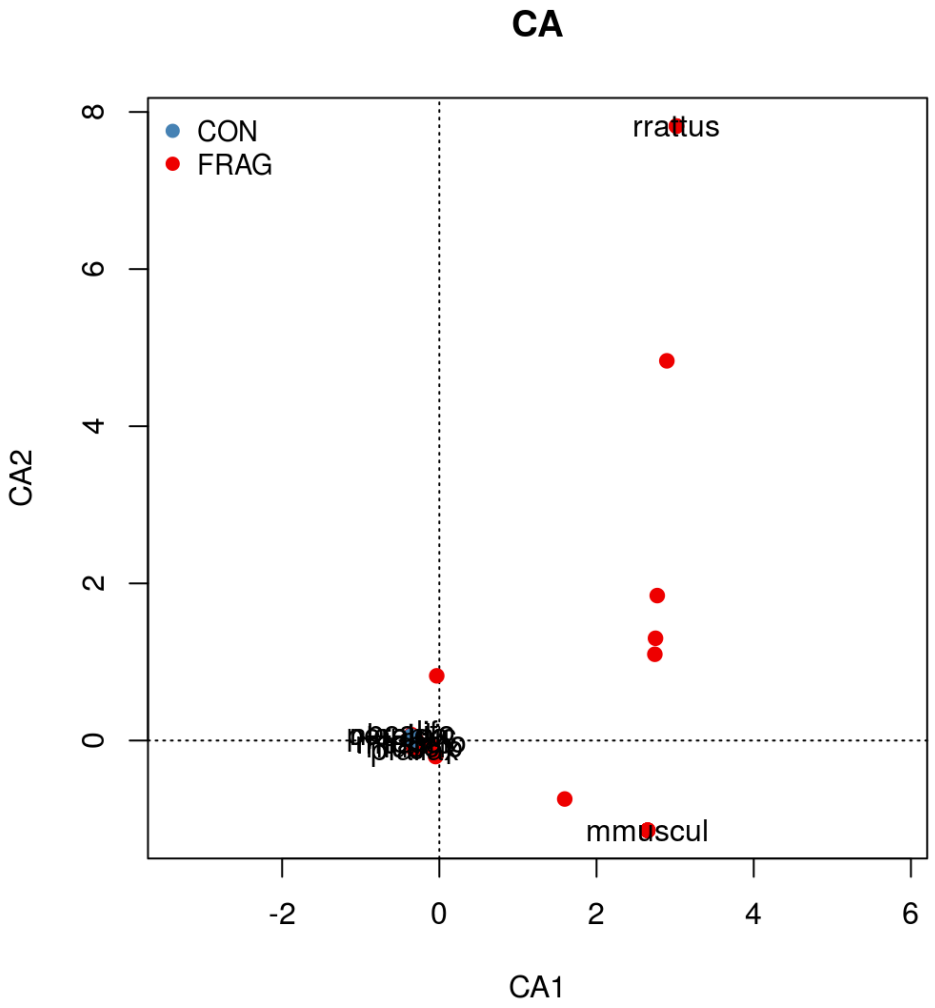


График CA



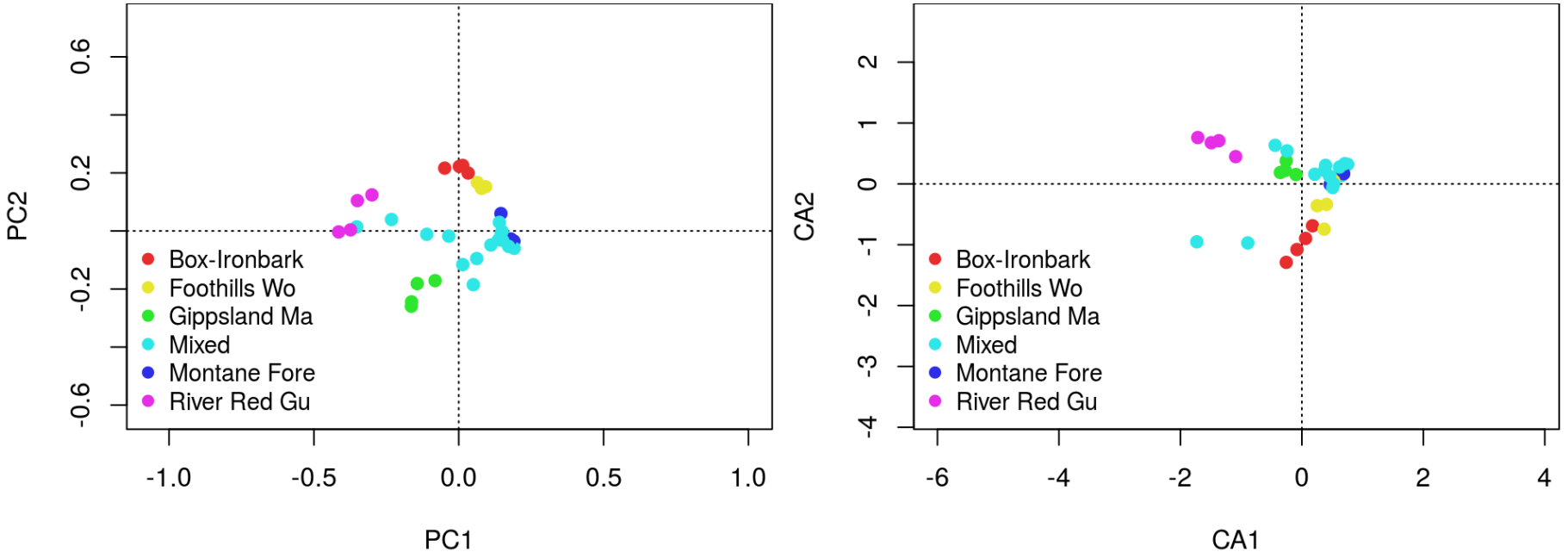
- На графике CA видно, что в нескольких местах больше *R.rattus* и *M.musculus*, чем ожидается (это интродуценты)

Задание: Проведите корреспондентный анализ данных про птиц

- исчез ли эффект подковы?

Решение

```
bird_ca <- cca(birds[, -c(1, 2)])
op <- par(mfrow = c(1, 2), mar = c(4, 4, 0, 0.5))
col_ord_plot(ord = bird_h_pca, colfac = factor(birds$habitat), leg.pos = "bottomleft",
    ncol = 1, display.labs = FALSE)
col_ord_plot(ord = bird_ca, colfac = factor(birds$habitat), leg.pos = "bottomleft",
    ncol = 1, display.labs = FALSE)
par(op)
```



- Остался "эффект дуги" (так называется "эффект подковы" для корреспондентного анализа)

Take home messages

- Анализ главных компонент
  - При анализе счетных признаков трансформация данных нужна, чтобы избежать "эффекта подковы"
- Корреспондентный анализ
  - придуман для анализа таблиц сопряженности
  - использует расстояние хи-квадрат
  - собственные числа отражают хи-квадрат, объясненный осями (степень зависимости столбцов и строк)

Дополнительные ресурсы

- Borcard, D., Gillet, F., Legendre, P., 2011. Numerical ecology with R. Springer.
- Legendre, P., Legendre, L., 2012. Numerical ecology. Elsevier.
- Oksanen, J., 2011. Multivariate analysis of ecological communities in R: vegan tutorial. R package version 2–0.
- The Ordination Web Page URL http://ordination.okstate.edu/ (accessed 10.21.13).
- Quinn, G.G.P., Keough, M.J., 2002. Experimental design and data analysis for biologists. Cambridge University Press.
- Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analysing ecological data. Springer.