Анализ избыточности (Redundancy analysis, RDA)

Многомерные методы на R, весна 2015

Марина Варфоломеева
Каф. Зоологии беспозвоночных, СПбГУ

Анализ избыточности (Redundancy analysis, RDA)

- Связь нескольких наборов переменных
- Анализ избыточности, теория и практика
- Проверка значимости ординации
- Выбор оптимальной модели
- Частный анализ избыточности и компоненты объясненной инерции
- Компоненты объясненной изменчивости

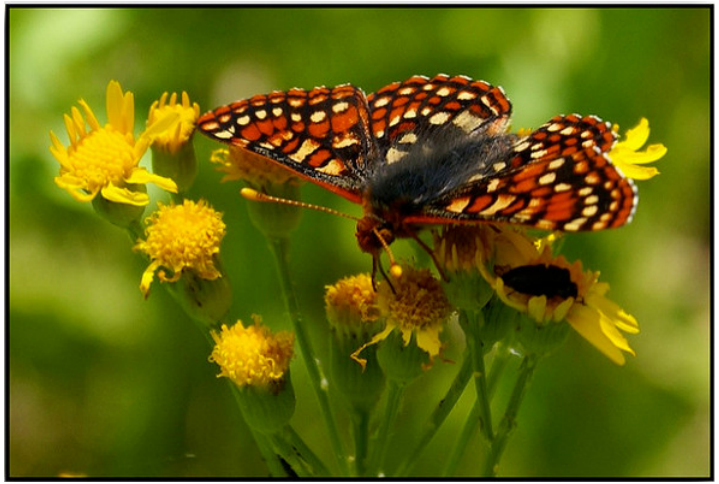
Вы сможете

- Проводить анализ избыточности
- Оценивать долю объясненной инерции
- Интерпретировать компоненты по нагрузкам переменных
- Строить ординацию объектов в пространстве компонент
- Проверять значимость модели ординации при помощи пермутационного теста
- Разделять объясненную инерцию на компоненты, связанные с разными наборами переменных, при помощи частного анализа избыточности

Связь нескольких наборов переменных

Пример: генетика бабочек Euphydryas editha

Частоты разных аллелей фосфоглюкоизомеразы и данные о факторах среды для 16 колоний бабочек **Euphydryas editha** в Калифорнии и Орегоне

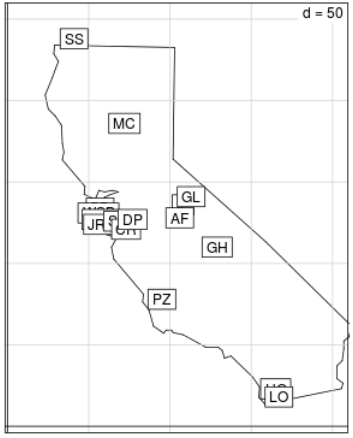


Winged Wonder by Roger Lynn on Flickr

данные McKechnie et al., 1975

Как вы думаете, что будет определять генетическую структуру в колониях бабочек?

```
library(ade4)
data(butterfly)
# расположение сайтов
s.label(butterfly$xy, contour = butterfly$contour, inc = FALSE)
```



Структура данных

- $xy - координаты колоний
- $envir - 4 фактора среды для колоний
- $genet - частоты 6 аллелей в колониях
- $contour - карта Калифорнии

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

```
# List of 5
#  $ xy     :'data.frame':  16 obs. of  2 variables:
#   ..$ x: num [1:16] 41 57 56 57 58 ...
#   ..$ y: num [1:16] 238 134 131 127 124 ...
#  $ envir  :'data.frame':  16 obs. of  4 variables:
#   ..$ Altitude     : num [1:16] 500 800 570 550 550 ...
#   ..$ Precipitation: num [1:16] 43 20 28 28 28 ...
#   ..$ Temp_Max     : num [1:16] 98 92 98 98 98 ...
#   ..$ Temp_Min     : num [1:16] 17 32 26 26 26 ...
#  $ genet  :'data.frame':  16 obs. of  6 variables:
#   ..$ 0.4 : num [1:16] 0 0 0 0 0 ...
#   ..$ 0.6 : num [1:16] 3 16 6 4 1 ...
#   ..$ 0.8 : num [1:16] 22 20 28 19 8 ...
#   ..$ 1   : num [1:16] 57 38 46 47 50 ...
#   ..$ 1.16: num [1:16] 17 13 17 27 35 ...
#   ..$ 1.2 : num [1:16] 1 12 2 3 6
```

Создадим переменные с более короткими названиями для удобства

```
# частоты аллелей
gen <- butterfly$genet
head(gen, 3)
```

```
#     0.4 0.6 0.8  1 1.16 1.3
# SS    0   3  22 57   17   1
# SB    0  16  20 38   13  13
# WSB   0   6  28 46   17   3
```

```
# переменные среды и географические координаты
env_geo <- cbind(butterfly$envir, butterfly$xy)
head(env_geo, 3)
```

```
#     Altitude Precipitation Temp_Max Temp_Min  x   y
# SS       500            43       98       17 41 238
# SB       800            20       92       32 57 134
# WSB      570            28       98       26 56 131
```

Анализ избыточности, теория и практика

Анализ избыточности (RDA)

- Основан на анализе главных компонент
- Нужно две матрицы данных: матрица зависимых переменных и матрица предикторов
- Нужно найти такие компоненты матрицы зависимых переменных, которые являются линейными комбинациями предикторов и отражают максимум изменчивости.

Условие применимости такое же как у PCA:

- Линейная зависимость переменных-откликов от предикторов

Не будем вдаваться в математику RDA - перейдем к практике

Собственно, RDA (Redundancy analysis, анализ избыточности) в vegan

- Зависимые переменные (отклики) - генетические данные
- Независимые переменные (предикторы) - переменные среды

```
library(vegan)
bf_rda <- rda(gen ~ Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
summary(bf_rda)
```

```
#
# Call:
# rda(formula = gen ~ Altitude + Precipitation + Temp_Max + Temp_Min,      data = env_geo)
#
# Partitioning of variance:
#               Inertia Proportion
# Total             730      1.000
# Constrained       380      0.521
# Unconstrained     350      0.479
#
# Eigenvalues, and their contribution to the variance
#
# Importance of components:
#                          RDA1    RDA2    RDA3    RDA4     PC1     PC2
# Eigenvalue            357.20 19.6391 2.30457 0.75999 211.979 101.540
# Proportion Explained    0.49  0.0269 0.00316 0.00104   0.291   0.139
# Cumulative Proportion   0.49  0.5165 0.51966 0.52070   0.811   0.950
```

Структура общей изменчивости

О структуре изменчивости можно судить **по суммам собственных чисел ординационных осей** (ограниченных и неограниченных)

```
Partitioning of variance:
             Inertia Proportion
Total          729.6     1.0000
Constrained    379.9     0.5207
Unconstrained  349.7     0.4793
```

- Total - всех осей - общая изменчивость исходной матрицы откликов (генетич. структуры в разных сайтах)
- Constrained - осей, кот. являются комбинациями факторов среды - изменчивость объясненная средой
- Unconstrained - необъясненная изменчивость

Важность различных компонент

Можно более подробно оценить, как распределяется изменчивость между осями

```
Eigenvalues, and their contribution to the variance

Importance of components:
                          RDA1     RDA2    RDA3    RDA4
Eigenvalue            357.2026 19.63905 2.30457 0.75999
Proportion Explained    0.4896  0.02692 0.00316 0.00104
Cumulative Proportion   0.4896  0.51650 0.51966 0.52070
                           PC1      PC2      PC3     PC4     PC5
Eigenvalue            211.9795 101.5400 27.13436 7.91921 1.12065
Proportion Explained    0.2905   0.1392  0.03719 0.01085 0.00154
Cumulative Proportion   0.8113   0.9504  0.98761 0.99846 1.00000
```

- Много изменчивости объяснено, но много осталось необъясненной. Первые две ограниченных оси объясняют 51% изменчивости, но первые две неограниченных объясняют еще 43%

Распределение изменчивости, потенциально объяснимой факторами

```
Accumulated constrained eigenvalues
Importance of components:
                          RDA1     RDA2    RDA3  RDA4
Eigenvalue            357.2026 19.63905 2.30457 0.760
Proportion Explained    0.9402  0.05169 0.00607 0.002
Cumulative Proportion   0.9402  0.99193 0.99800 1.000
```

- Первая ограниченная ось объясняет большую часть потенциально объяснимой изменчивости. Остальные оси почти ничего не объясняют.

Собственные векторы, нагрузки переменных = “species scores”

```
scores(bf_rda, display = "species", choices = 1:5)
```

```
Species scores

        RDA1     RDA2      RDA3     RDA4     PC1     PC2
0.4   0.6371 -0.66036  0.197281 -0.24232  0.9681 -0.4640
0.6   0.8943 -0.75253 -0.006618  0.12069  2.2443 -0.7172
0.8   2.6182 -0.09492  0.141566  0.16990  2.7677  0.0607
1    -6.3502  0.12080  0.054622  0.04704 -3.3365 -2.4763
1.16  1.5535  1.33696  0.117722 -0.04195 -2.3548  2.5813
1.3   0.6470  0.05005 -0.504573 -0.05336 -0.2887  1.0155
```

Корреляции между откликами и предикторами

- Сильная корреляция между генетической структурой и средой только для первой ограниченной оси. Для других - умеренные или слабые.

```
spenvcor(bf_rda)
```

```
#  RDA1  RDA2  RDA3  RDA4
# 0.836 0.353 0.332 0.183
```

Визуализация ординации

- Какие предикторы важнее всего?
- Какими факторами определяется значение зависимых переменных?

Триплоты:

- переменные-отклики ("species"),
- объекты ("sites")
- переменные-предикторы (непрерывные в виде векторов, дискретные в виде центроидов)

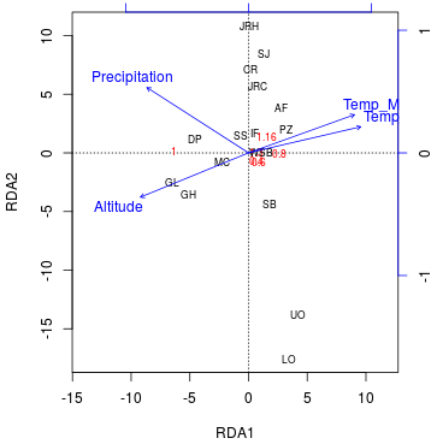
Биплоты:

- отклики + предикторы
- объекты + предикторы

Триплот корреляций (scaling = 2): Какие переменные среды важнее всего?

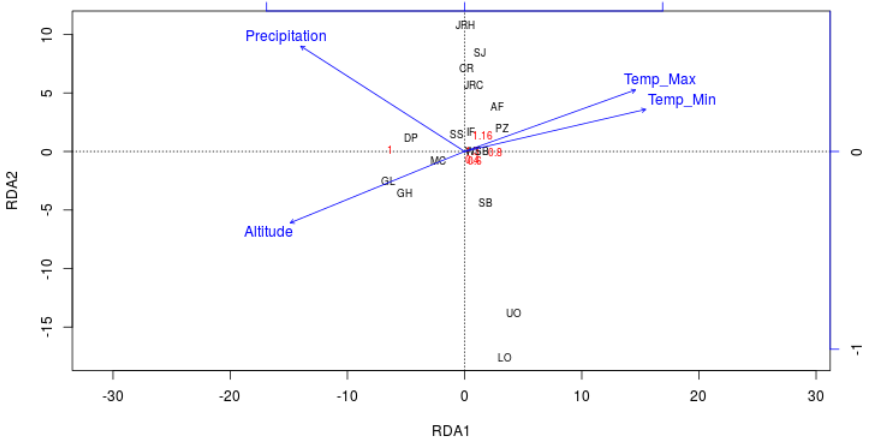
- Векторы - независимые переменные, факторы среды
- Надписи - объекты (сайты, особи, популяции и пр.)
- Красные надписи - зависимые переменные
- Косинусы углов между векторами - корреляции между соотв. переменными
- Расстояния не имеют смысла
- Проекция объекта на линию-вектор - значение переменной для данного объекта

```
plot(bf_rda, scaling = 2)
```



Пример интерпретации триплота корреляций

```
plot(bf_rda, scaling = 2)
```

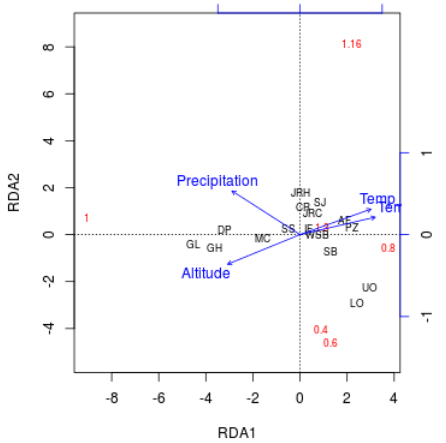


- Вдоль первой оси - температура, высота и осадки
- Вдоль второй оси - немного осадки

Триплот расстояний (scaling = 1):

- Надписи - объекты (сайты, особи, популяции и пр.)
- Красные надписи - зависимые переменные
- Векторы - независимые переменные, факторы среды
- Расстояния между точками - расстояния между наблюдениями
- Углы между векторами предикторов и откликов - корреляции между соотв. переменными, другие углы не имеют смысла
- Проекция объекта на линию-вектор отражает примерное положение данного объекта вдоль соотв. переменной (но не значение)
- Отношения между дискретными и

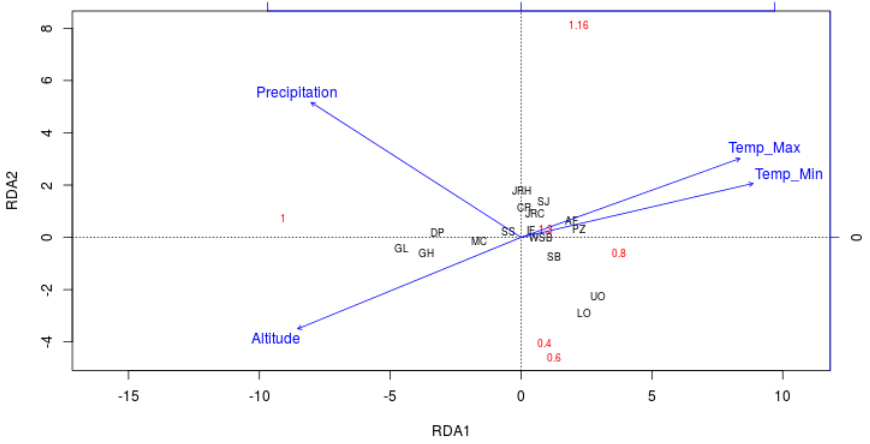
```
plot(bf_rda, scaling = 1)
```




```
# s.label(butterfly$xy, contour = butterfly$contour
```

Пример интерпреатции триплота расстояний

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



- Генетическая структура в LO и UO похожа, но не похожа на остальные места
- GL и GH - более высокогорные сайты, чем LO и UO

Проверка значимости ординации

Общий тест на значимость ординации

- тестируем гипотезу о том, что отношения между генотипом и средой значимы. H_0: значения предикторов в пробах не зависят от переменных среды (генетическая структура не зависит от среды)
- основан на пермутациях: проверяем, насколько наблюдаемая связь сильнее, чем если случайно переставить данные
- статистика - сумма всех соб. чисел ограниченных осей

Общий тест: Влияют ли факторы на зависимые переменные?

Есть ли связь генетики со средой?

```
anova(bf_rda)
```

```
# Permutation test for rda under reduced model
# Permutation: free
# Number of permutations: 999
#
# Model: rda(formula = gen ~ Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
#          Df Variance    F Pr(>F)
# Model     4      380 2.99  0.015 *
# Residual 11      350
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- связь генетической структуры и среды значима

Тест факторов, type I эффекты: Какие факторы влияют на зависимые переменные?

- Генетическая структура популяций бабочек достоверно зависит от высоты, если в модель включены др. факторы.
- Но это Type I эффекты - они зависят от порядка включения факторов в модель. Т.е. после включения высоты в модель другие факторы уже не влияют.

```
anova(bf_rda, by = "term")
```

```
# Permutation test for rda under reduced model
# Terms added sequentially (first to last)
# Permutation: free
# Number of permutations: 999
#
# Model: rda(formula = gen ~ Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
#              Df Variance    F Pr(>F)
# Altitude      1      279 8.79  0.004 **
# Precipitation  1       72 2.28  0.119
# Temp_Max      1       20 0.62  0.571
# Temp_Min      1        8 0.26  0.796
# Residual     11      350
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Тест факторов, type III эффекты: Какие факторы влияют на зависимые переменные?

- Если протестировать каждый из факторов отдельно, при у словии, что все остальные включены в модель, то получится, что ни один из них не влияет.

```
anova(bf_rda, by = "mar")
```

```
# Permutation test for rda under reduced model
# Marginal effects of terms
# Permutation: free
# Number of permutations: 999
#
# Model: rda(formula = gen ~ Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
#              Df Variance    F Pr(>F)
# Altitude      1        3 0.08   0.95
# Precipitation  1       34 1.08   0.34
# Temp_Max      1       25 0.79   0.47
# Temp_Min      1        8 0.26   0.79
# Residual     11      350
```

Тест значимости осей, ограниченных факторами:

- H_0: значения переменных-откликов для объектов не зависят от переменных-предикторов
- пермутационный: выбирает оси, которые объясняют больше изменчивости, чем из др. матриц, полученных путем перестановок
- Генетическая структура значимо меняется вдоль первой главной оси

```
anova(bf_rda, by = "axis")
```

```
# Permutation test for rda under reduced model
# Marginal tests for axes
# Permutation: free
# Number of permutations: 999
#
# Model: rda(formula = gen ~ Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
#          Df Variance     F Pr(>F)
# RDA1      1      357 11.24  0.001 ***
# RDA2      1       20  0.62  0.530
# RDA3      1        2  0.07  0.985
# RDA4      1        1  0.02  0.996
# Residual 11      350
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

У нас проблема. Если мы тестируем любой из факторов, после включения остальных в модель - он не влияет. Это значит, что модель не оптимальна.

Можно использовать пошаговый выбор модели: добавляем в модель лучшие переменные и снова исключаем те, что потеряли значимость.

Модели с разным числом предикторов можно сравнить при помощи пермутационного теста (AIC для ограниченных ординаций не существует!)

Осторожно! В vegan факторы включенные в модель обозначаются "-", а факторы, исключенные из модели - "+"

Выбор оптимальной модели

Для пошагового выбора нам понадобятся полная и нулевая модели

```
m1 <-rda(gen ~ Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
m0 <- rda(gen ~ 1, data = env_geo)
m <- ordistep(m0, scope = formula(m1))
```

```
#
# Start: gen ~ 1
#
#                 Df AIC    F Pr(>F)
# + Temp_Min       1 100 9.75  0.005 **
# + Altitude       1 101 8.69  0.005 **
# + Temp_Max       1 101 8.12  0.005 **
# + Precipitation  1 102 7.30  0.010 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Step: gen ~ Temp_Min
#
#            Df AIC    F Pr(>F)
# - Temp_Min  1 106 9.75  0.005 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#                 Df AIC    F Pr(>F)
```

Оптимальная модель, отобранная при помощи пошагового алгоритма

```
m$anova
```

```
#           Df AIC    F Pr(>F)
# + Temp_Min  1 100 9.75  0.005 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Оптимальная модель содержит только один предиктор - минимальную температуру

Частный анализ избыточности и компоненты объясненной инерции

Зачем нужен частный анализ избыточности?

Мы уже обнаружили тесную связь генотипов со средой, и даже знаем, с какими переменными.

Но генотипы в близких местах могут быть похожи от того, что

- там сходный климат
- поток генов между близкими колониями облегчен

Нужно удалить влияние географического положения.

Частный анализ избыточности

- зависимость от одного набора переменных (предикторов), когда влияние другого (ковариат) исключено.

Техника:

1. Множественная регрессия предикторов от ковариат.
2. Остатки от этой регрессии (то, что от ковариат не зависит) - в PCA в качестве переменных среды.

Legendre & Legendre 1998

Делаем частный RDA: зависимость генетической структуры от среды с учетом географического положения

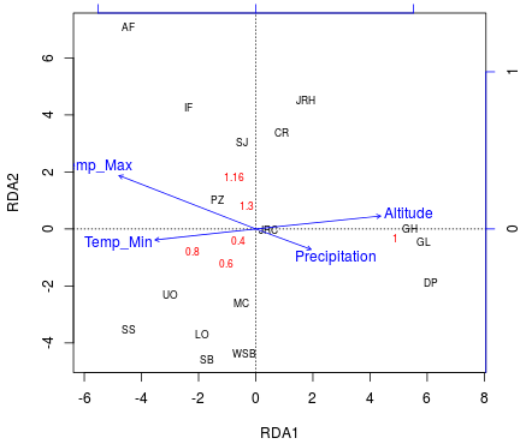
```
bf_prda_1 <- rda(gen ~ Altitude + Precipitation + Temp_Max + Temp_Min + Condition(x + y), data = env_ge
anova(bf_prda_1) ## Пермутационный тест
```

```
# Permutation test for rda under reduced model
# Permutation: free
# Number of permutations: 999
#
# Model: rda(formula = gen ~ Altitude + Precipitation + Temp_Max + Temp_Min + Condition(x +
#          Df Variance  F Pr(>F)
# Model     4      268  3  0.023 *
# Residual  9      201
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Климат объясняет генетическую изменчивость, даже после удаления влияния географических координат

График ординации

```
plot(bf_prda_1)
```



- Смысл остался прежним, изменились нюансы
- Первая ось - высота и мин температура (кот. определяется высотой). Макс. температура с высокими нагрузками по обеим осям. Осадки почти ничего не объясняют.

Задание: Проверьте значимость частного RDA, описывающего зависимость генетической структуры от среды с учетом географического положения

```
anova(bf_prda_1)
anova(bf_prda_1, by = "term")
anova(bf_prda_1, by = "mar")
anova(bf_prda_1, by = "axis")
```

Компоненты объясненной изменчивости

Компоненты изменчивости

К этому моменту мы знаем, что климат объясняет генетическую изменчивость, даже после удаления влияния географических координат.

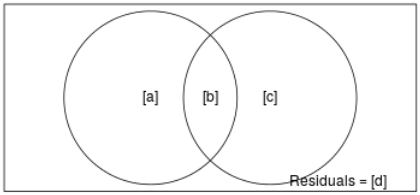
Но сколько именно обясняется одной лишь географической близостью, а сколько - общим действием климата и географии?

Общую изменчивость делим на части:

- связана с переменными среды,
- с пространственными координатами,
- может быть объяснена тем и другим вместе,
- не объяснена ни тем не другим.

Нужно для рассчета компонент изменчивости

```
showvarparts(2)
```



1. a + b + c - вся потенциально объяснимая средой и географией изменчивость
2. a - изменчивость, объясненная климатом
3. c - изменчивость, объясненная географией
4. b - изменчивость, совместно объясненная средой и географией

Чтобы выделить компоненты изменчивости нам нужно несколько элементов

У нас уже есть **частный RDA №1: зависимость генетики от среды с учетом географии** (для a)

Нам нужен **частный RDA №2: генетика от географии с учетом свойств среды** (для c)

```
bf_prda_2 <- rda(gen ~ x + y + Condition(Altitude + Precipitation + Temp_Max + Temp_Min), data = env_ge
```

И **полная модель RDA генетика от среды и географического положения** (для a + b + c)

```
bf_rda_full <- rda(gen ~ x + y + Altitude + Precipitation + Temp_Max + Temp_Min, data = env_geo)
```

Задание: Найдите компоненты инерции

1. изменчивость, потенциально объяснимую средой и географией
2. изменчивость, связанную только со средой, но не с географией
3. изменчивость, связанную только с географией, но не со средой
4. изменчивость, объясненную одновременно средой и географией

Подсказка

Смотрите на результаты разных RDA

1) Сколько изменчивости потенциально объясняется средой и географией?

```
sum_full <- summary(bf_rda_full)
# sum_full
```

```
Partitioning of variance:
  Inertia Proportion
Total          729.6     1.0000
Constrained    =528.2=   0.7239
Unconstrained   201.4     0.2761
```

Изменчивость, объясненная вместе средой и географией, здесь достаточно велика

```
It <- sum_full$constr.chi
```

В отличие от нее, доля изменчивости, объясненной ограниченной матрицей, может быть довольно малой по оношении к общей изменчивости. Некоторые советуют сосредоточиться на доле от потенциально объяснимой изменчивости (от sum_full$constr.chi)

2) Изменчивость, объясненная климатом

```
sum_prda_1 <- summary(bf_prda_1)
# sum_prda_1
```

```
Partitioning of variance:
  Inertia Proportion
Total           729.6     1.0000
Conditioned     259.7     0.3560
Constrained    =268.4=    0.3679
Unconstrained   201.4     0.2761
```

```
Ie <- sum_prda_1$constr.chi
```

3) Изменчивость, объясненная географией

```
sum_prda_2 <- summary(bf_prda_2)
# sum_prda_2
```

```
Partitioning of variance:
  Inertia Proportion
Total           729.6     1.0000
Conditioned     379.9     0.5207
Constrained    =148.3=    0.2032
Unconstrained   201.4     0.2761
```

```
Ig <- sum_prda_2$constr.chi
```

4) Изменчивость, совместно объясненная средой и географией

Iab = It - Ia - Ib

```
Ieg <- sum_full$constr.chi - sum_prda_1$constr.chi -sum_prda_2$constr.chi
```

Компоненты изменчивости - сводим результаты вместе

```
bf_results <- data.frame(Inertia = c(Ie, Ig, Ieg, It),
           row.names = c('Только среда',
                         'Только география',
                         'Среда и география вместе',
                         'Общая объяснимая инерция'))
bf_results$Proportion <- bf_results$Inertia/sum(bf_results$Inertia[1:3])
colnames(bf_results) <- c('Инерция', 'Доля')
bf_results
```

```
#                           Инерция  Доля
# Только среда                  268 0.508
# Только география              148 0.281
# Среда и география вместе      111 0.211
# Общая объяснимая инерция      528 1.000
```

Среда объясняет 50% общей изменчивости генетической структуры - очень много, но и география объясняет 30%. И только 21% объясняется совместным влиянием среды и географии

Take home messages

- Анализ избыточности помогает установить связь между несколькими наборами переменных. Один из наборов считается зависимым, другой считается объясняющим
- Для анализа необходимо, чтобы зависимости переменных-откликов от предикторов были линейными
- В ходе анализа выделяют два типа осей - ограниченные (объясненные) переменными-предикторами, и неограниченные (необъясненные) ими
- Частный анализ избыточности позволяет описать зависимость двух наборов переменных с поправкой на влияние дополнительных переменных (ковариат)
- При помощи частного анализа избыточности можно выделить компоненты изменчивости связанные с несколькими (2-4) наборами переменных-предикторов

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.