Канонический корреспондентный анализ

Анализ и визуализация многомерных данных с использованием R

Вадим Хайтов, Марина Варфоломеева

Вы сможете

- Проводить канонический корреспондентный анализ
- Оценивать долю объясненной инерции
- Интерпретировать канонические оси по координатам переменных
- Строить ординацию объектов в пространстве канонических осей
- Проверять значимость модели ординации
- Разделять общую изменчивость на компоненты при помощи частного канонического корреспондентного анализа

Связь нескольких наборов переменных: виды и среда

Пример: клещи-орибатиды на сплавине одного из канадских озер

На площадке 10 x 2.6м взяли стратифицированную случайную выборку на 7 типах субстратов (70 проб). Исследователи хотели разделить влияние среды и положения в пространстве на структуру сообщества клещей-орибатид (Borcard & Legendre 1994).

Oribatid mite with a visiting friendly springtail by Andy Murray on Flickr

Tourbière/Peat bog* by peupleloup on Flickr

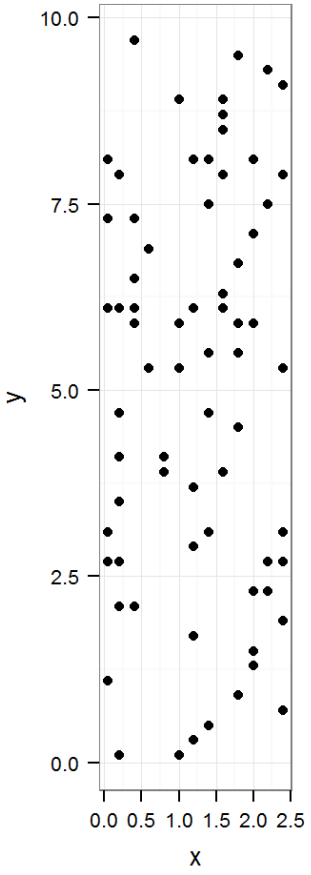
Читаем данные

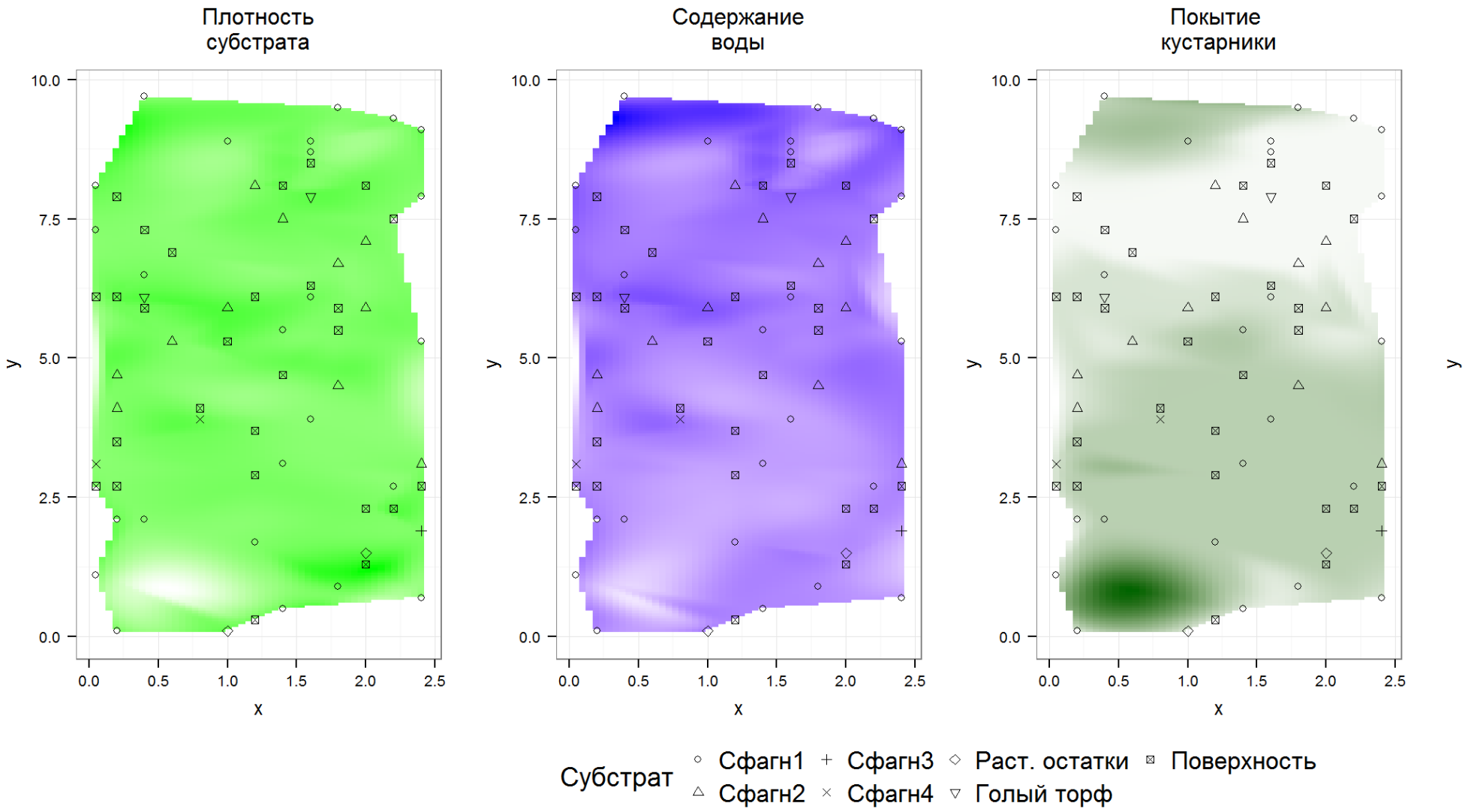
```
##   Brachy PHTH HPAV RARD SSTR Protopl
## 1     17    5    5    3    2       1
## 2      2    7   16    0    6       0
```

```
## 'data.frame':    70 obs. of  2 variables:
##  $ x: num  0.2 1 1.2 1.4 2.4 1.8 0.05 2 2 1.2 ...
##  $ y: num  0.1 0.1 0.3 0.5 0.7 0.9 1.1 1.3 1.5 1.7 ...
```

```
## 'data.frame':    70 obs. of  5 variables:
##  $ SubsDens : num  39.2 55 46.1 48.2 23.6 ...
##  $ WatrCont : num  350 435 372 360 204 ...
##  $ Substrate: Factor w/ 7 levels "Sphagn1","Sphagn2",..: 1 5 7 1 1 1 1 7 5 1 ...
##  $ Shrub    : Ord.factor w/ 3 levels "None"<"Few"<"Many": 2 2 2 2 2 2 2 3 3 3 ...
##  $ Topo     : Factor w/ 2 levels "Blanket","Hummock": 2 2 2 2 2 2 2 1 1 2 ...
```

Схема асположения проб




Плотность субстрата
Содержание воды
Покытие кустарники
y
x
10.0
7.5
5.0
2.5
0.0
0.5
1.0
1.5
2.0
2.5
Субстрат
Сфагн1
Сфагн2
Сфагн3
Сфагн4
Раст. остатки
Голый торф
Поверхность

Косвенный градиентный анализ

- ищем ось макс. варьирования (анализ главных компонент (PCA), корреспондентный анализ (CA))
- затем регрессионый анализ в завсимости от предикторов

- Опасности:
- теряем связи переменных среды и отброшенных компонент высоких порядков.
  - **Выход** использовать прямой градиентный анализ.
- PCA нужны линейные зависимости (т.к. матрица ковариаций или корреляций). Если связи нелинейны - искривление градиентов.
  - **Выход** использовать другие расстояния (CA использует хи-квадрат).

Прямой градиентный анализ

Прямой градиентный анализ еще называют **каноническим** анализом.

Канонической формой функции или выражения в математике называют такую упрощенную форму, к которой можно свести функцию или выражение без потери самой важной информации.

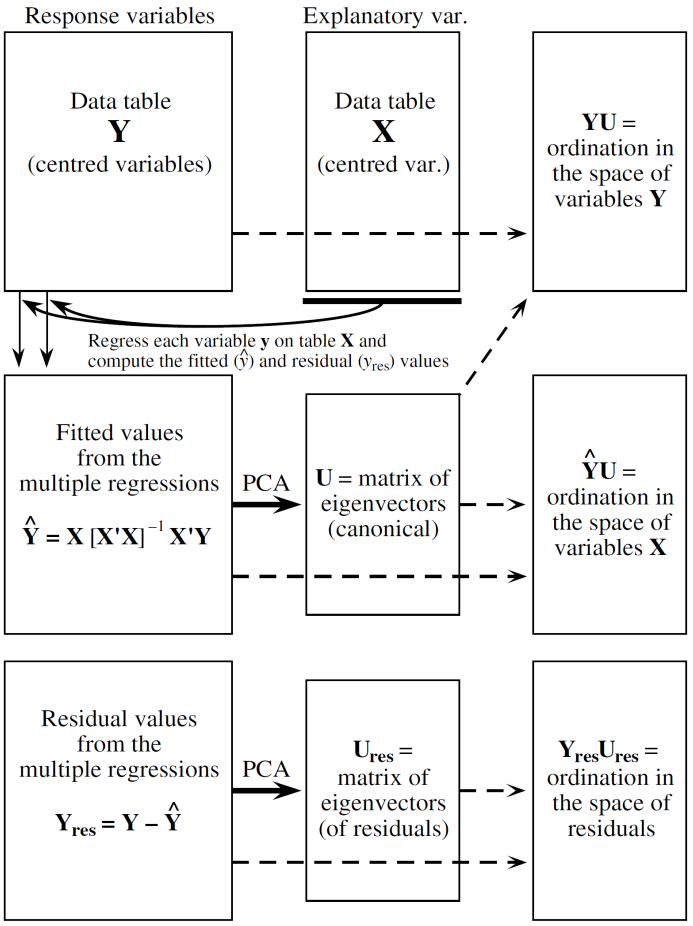
Например, канонической формой матрицы ковариации является матрица собственных значений

Канонический анализ

В общем виде каноничесий анализ сводится к тому же, что мы видели в CA и PCA, то есть к поиску собственных значений (задающих оси максимального варьирования) и собственных векторов (коэффициентов, определяющих координаты в новом пространстве).

НО! Оси ордиации определяются переменными среды.

Принципиальная схема канонического анализа



из Legendre & Legendre, 2012

Вспомним регрессионный анализ

$$\hat{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Здесь $\hat{y}$ - предсказанные значения (fitted values)

$x_1, x_2 \ldots x_p$ - предикторы

$\beta_0, \beta_1, \beta_2 \cdots + \beta_p$ - коэффициенты

Регрессионный анализ в матричном виде

Как всегда, независимая чсть модели - это модельная матрица **X**
Зависимая часть в анализе - это не вектор **y**, как мы привыкли, а матрица **Y**
Коэффициенты, будут представлены тоже в виде матрицы β.

Тогда матрица предсказанных значений

$$\hat{\mathbf{Y}} = \beta \times X$$

При этом матрица коэффициентов - это

$$\beta = [X'X]^{-1}X'Y$$

Матрица остатков - это

$$Y_{resid} = Y - \hat{Y}$$

SVD и канонический анализ

Основан на SVD трех матриц:

1. SVD матрицы исходных значений (Y) - дает информацию об общей изменчивости в системе - это просто PCA или CA.
2. SVD матрицы предсказнных значений ($\hat{Y}$) - дает информацию системе, ограниченной (Constrained) предкторами (X).
3. SVD матрицы остатков (Y_{resid}) - дает информацию о системе за вычитом влияния предикторов.

Канонический корреспондентный анализ (CCA)

- Основан на корреспондентном анализе, то есть вместо исходных данных используется матрица **Q** (см. дополнительные материалы "За кулисами CA и CCA").
- Мера изменчивости в этом анализе - *Inertia*.
- 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 инерции

Условия применимости такие же как у CA

Новые оси

- **Канонические** - Корреспондентный анализ на основе матрицы предсказанных значений.
- **Неканонические** - корреспондентный анализ таблицы остатков от регрессии координат по предикторам (каноническим осям) от переменных среды.

Два подхода к анализу

Рассмотрение ординации в канонических осях дает информацию о связи между видами, сайтами и средой.

Анализ неканонических осей - отношения между видами или сайтами после того, как учтен эффект среды.

CCA - канонический кореспондентный анализ в vegan

- Зависимые переменные (отклики) - обилие видов
- Независимые переменные (предикторы) - переменные среды

Задание

Используя представления об ограниченной ординации в форме RDA, постройте, по аналогии, ограниченную ординацию в форме CCA. Объект, содержащий результаты должен называться `mite_cca`

В качестве предикторов в модели возьмите следующие факторы из датафрейма `mite.env`: SubsDens, WatrCont, Substrate, Topo

Решение

```
mite_cca <- cca(mite ~ SubsDens + WatrCont + Substrate + Topo, data = mite.env)
```

Общая инерция и ее разложение

О ней можно судить **по суммам собственных чисел** (ограниченных и неограниченных осей)

```
Partitioning of mean squared contingency coefficient:
              Inertia Proportion
Total           1.696      1.000
Constrained     0.735      0.433
Unconstrained   0.961      0.567
```

Важность различных компонент

Можно более подробно оценить, как распределяется изменчивость между осями

В общем наборе осей (их 34 + 9 = 43!)

```
Eigenvalues, and their contribution to the mean squared contingency coefficient

Importance of components:
                        CCA1  CCA2   CCA3   CCA4   CCA5    CCA6 ...
Eigenvalue             0.426 0.129 0.0667 0.0443 0.0321 0.01447 ...
Proportion Explained   0.251 0.076 0.0393 0.0261 0.0189 0.00853 ...
Cumulative Proportion 0.251 0.327 0.3666 0.3927 0.4117 0.42020 ...
```

Только в ограниченных (канонических) осях (их всего 9!)

```
Accumulated constrained eigenvalues
Importance of components:
                        CCA1  CCA2   CCA3   CCA4   CCA5   CCA6 ...
Eigenvalue             0.426 0.129 0.0667 0.0443 0.0321 0.0145 ...
Proportion Explained   0.580 0.175 0.0907 0.0603 0.0437 0.0197 ...
Cumulative Proportion 0.580 0.755 0.8458 0.9061 0.9498 0.9695 ...
```

- Часть изменчивости структуры сообществ орибатид действительно удается объяснить изменениями среды. Канонические оси объясняют 43% общей инерции если рассматривать всю изменчивость (т.е. 43 оси). Из них первые две оси объясняют 33 % общей изменчивости.

Но, если рассматривать первые две канонические оси в пределах только ограниченных осей (т.е. 9 осей) - то среда объясняет целых 75%

Распределение изменчивости, потенциально объяснимой факторами

```
Accumulated constrained eigenvalues
Importance of components:
                          CCA1  CCA2   CCA3   CCA4   CCA5   CCA6   CCA7 ...
Eigenvalue              0.426 0.129 0.0667 0.0443 0.0321 0.0145 0.0107 ...
Proportion Explained    0.580 0.175 0.0907 0.0603 0.0437 0.0197 0.0145 ...
Cumulative Proportion 0.580 0.755 0.8458 0.9061 0.9498 0.9695 0.9840 ...
```

- Первая ограниченная ось объясняет 58%, вторая 17% того, что можно потенциально связать с воздействием среды, остальные оси почти ничего не объясняют. Значит, характеристики среды, в принципе, можно свести к двум комплексным независимым переменным.

Собственные векторы, нагрузки переменных = “species scores”

- обилие каких видов сильнее варьирует вдоль оси?

Другие части результатов

- Собственные векторы (“species scores”) координаты видов
- Координаты объектов (“site scores”)
- Координаты объектов в канонических осях (“Site constraints”)
- (“Biplot scores”) координаты для биплотов
- Центроиды сайтов для бинарных переменных среды на диаграмме ординации (“Centroids for factor constraints”)

Корреляции между откликами (обилиями видов) и предикторами (средой)

```
##  CCA1  CCA2  CCA3  CCA4  CCA5  CCA6  CCA7  CCA8  CCA9
## 0.905 0.768 0.639 0.695 0.606 0.526 0.521 0.448 0.269
```

- Несмотря на то, что канонические оси объясняют не очень большое количество изменчивости, почти все они сильно или умеренно коррелируют с характеристиками среды.

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)

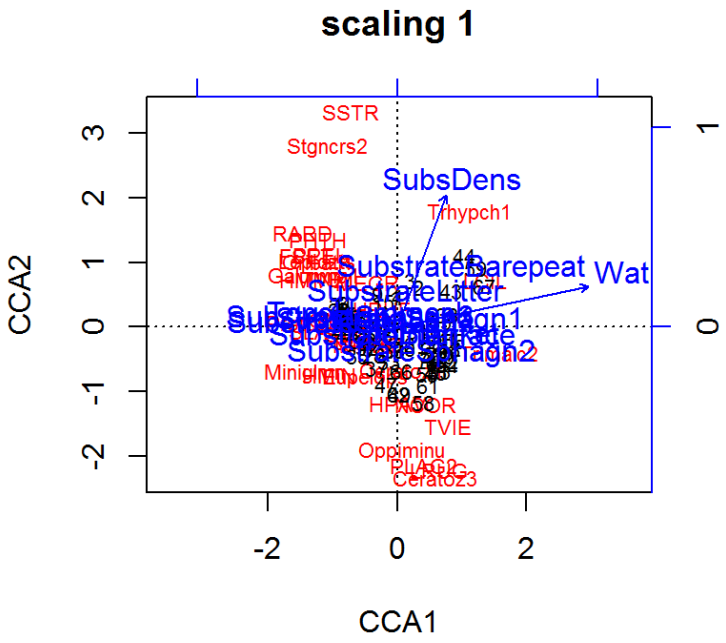
Биплоты:

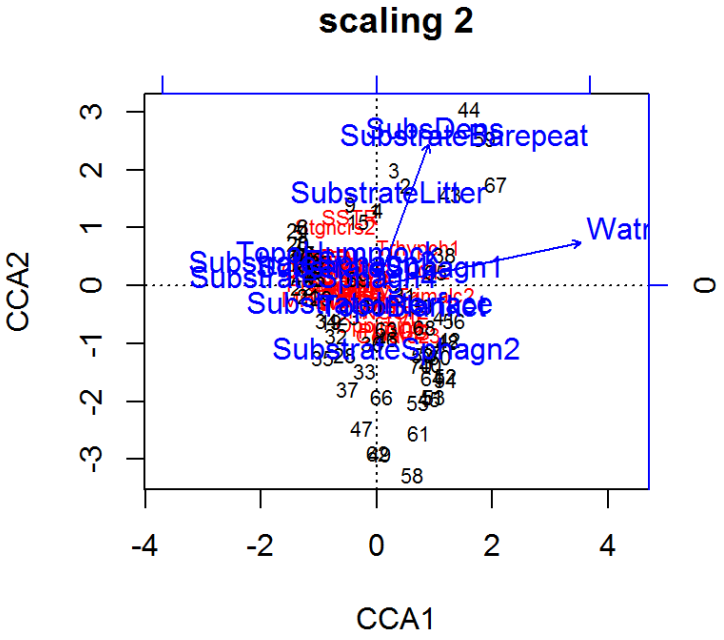
- отклики + предикторы
- объекты + предикторы

Примеры триплотов

Осторожно, в CCA есть два типа координат объектов:

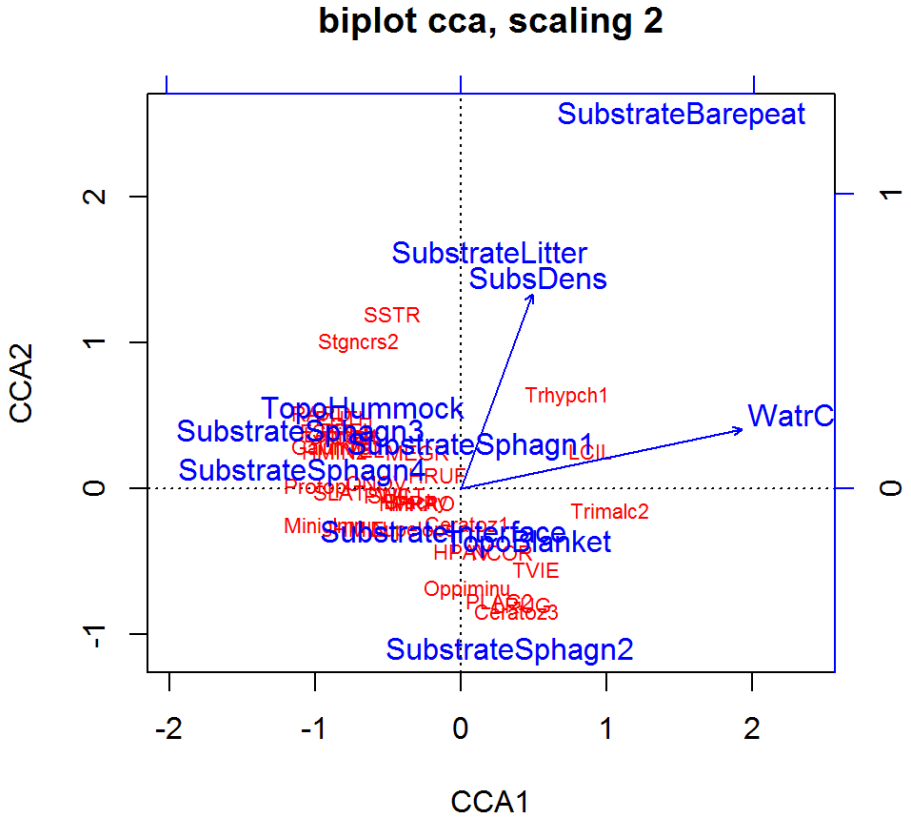
- WA - "weighted average scores" - без ограничений переменными среды, но при этом они отличаются от координат полученных в CA
- LC - "linear combination scores" - результат множественной регрессии WA-координат по линейным комбинациям переменных среды. Palmer (1993) рекомендовал графики с ними, потому что WA непонятно как интерпретировать - и это по-умолчанию используется в vegan



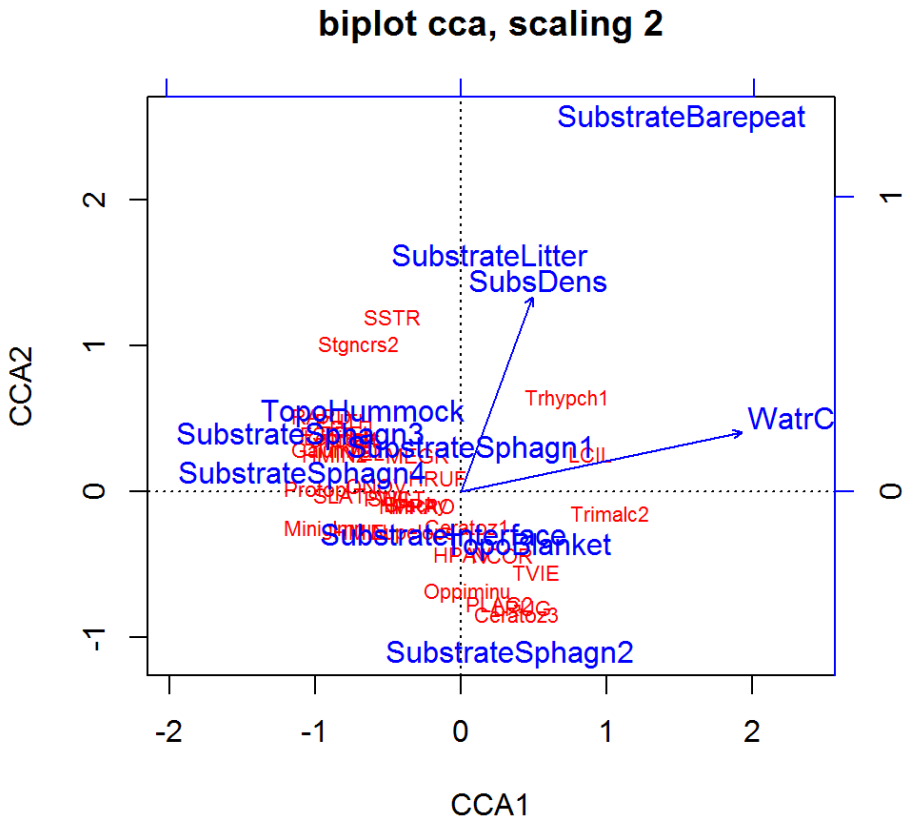


Отношения между видами и средой (scaling = 2)

- Проекция вида под прямым углом на переменную - оптимум вида по этой переменной
- Вид вблизи центроида качественной переменной - скорее всего часто встречается на сайтах с такими качествами
- Расстояния между центроидами и центроидами и объектами **НЕ равны** χ^2



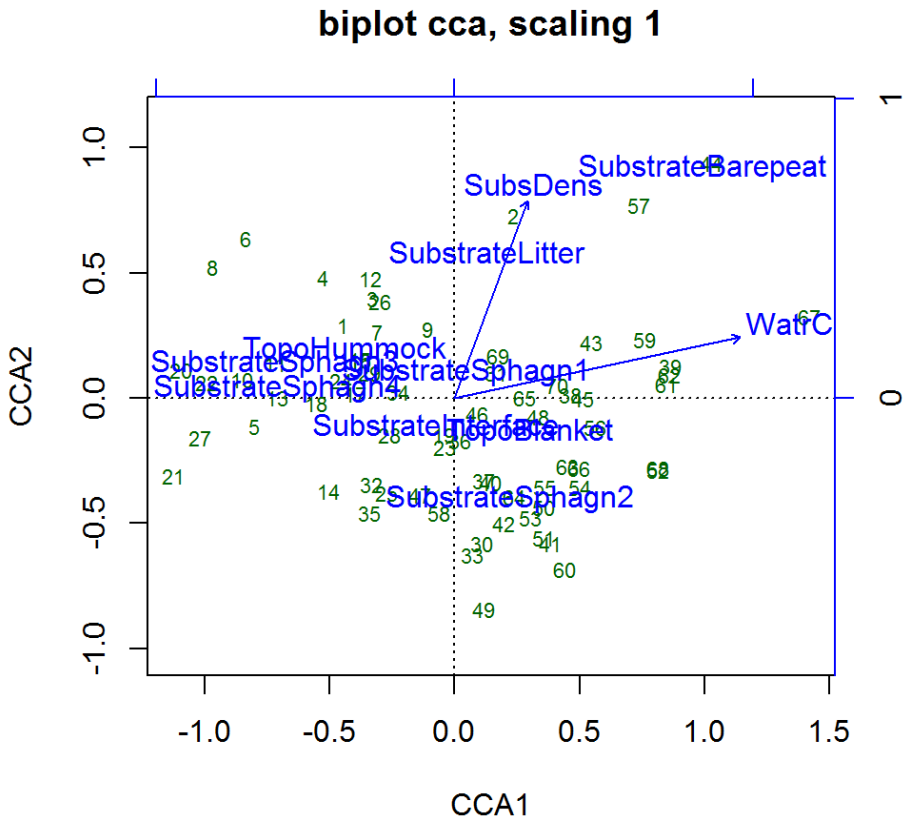
Пример интерпретации графика в scaling = 2



Отношения между объектами и средой (scaling = 1)

- Проекция объекта под прямым углом на колич. переменную - приблизительная позиция объекта вдоль этой переменной
- Объект вблизи центроида качественной переменной скорее всего обладает данным качеством
- Расстояние между центроидами качественных переменных и между центроидами объектов - χ^2

Пример интерпретации графика в scaling 1



Проверка значимости ординации

Общий тест на значимость ординации

- Тестируем гипотезу о том, что отношения между структурой сообщества и средой значимы - H_0: обилие видов в пробах не зависит от значений переменных среды
- основан на пермутациях: случайно перемешиваем данные и проверяем, насколько редко будет наблюдаться связь сильнее, чем данная
- статистика - сумма всех канонических соб. чисел

Общий тест: влияют ли факторы на зависимые переменные?

Есть ли связь структуры сообщества со средой?

```
anova(mite_cca)
```

```
## Permutation test for cca under reduced model
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = mite ~ SubsDens + WatrCont + Substrate + Topo, data = mite.env)
##          Df ChiSquare   F Pr(>F)
## Model     9     0.735 5.1  0.001 ***
## Residual 60     0.961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Структура сообщества связана с условиями среды

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

- Структура сообщества клещей зависит от плотности и типа субстрата, от содержания воды и топографии
- Осторожно, это Type I эффекты. Они зависят от порядка включения факторов в модель!

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

```
## Permutation test for cca under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = mite ~ SubsDens + WatrCont + Substrate + Topo, data = mite.env)
##           Df ChiSquare     F Pr(>F)
## SubsDens   1     0.100  6.24  0.005 **
## WatrCont   1     0.377 23.55  0.001 ***
## Substrate  6     0.199  2.07  0.057 .
## Topo       1     0.059  3.69  0.020 *
## Residual  60     0.961
## ---
## 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 условии, что остальные уже в модели, получится, что тип субстрата не влияет, а влияют остальные факторы.

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

```
## Permutation test for cca under reduced model
## Marginal effects of terms
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = mite ~ SubsDens + WatrCont + Substrate + Topo, data = mite.env)
##           Df ChiSquare     F Pr(>F)
## SubsDens   1     0.073  4.56  0.017 *
## WatrCont   1     0.239 14.93  0.001 ***
## Substrate  6     0.180  1.88  0.067 .
## Topo       1     0.059  3.69  0.020 *
## Residual  60     0.961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Тест значимости осей, ограниченных факторами:

- H_0: изменение состава сообщества в пробах вдоль данной оси не зависит от переменных среды
- пермутационный тест: случайно переставляем данные и проверяем, насколько редко данная ось объясняет больше изменчивости чем в исходном анализе. Если редко, то варьирование вдоль оси значимо
- Структура сообщества значимо меняется вдоль первых 5 канонических осей

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

```
## Permutation test for cca under reduced model
## Marginal tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: cca(formula = mite ~ SubsDens + WatrCont + Substrate + Topo, data = mite.env)
##          Df ChiSquare     F Pr(>F)
## CCA1      1     0.426 26.61  0.001 ***
## CCA2      1     0.129  8.04  0.001 ***
## CCA3      1     0.067  4.17  0.001 ***
## CCA4      1     0.044  2.77  0.006 **
## CCA5      1     0.032  2.01  0.040 *
## CCA6      1     0.014  0.90  0.541
## CCA7      1     0.011  0.67  0.819
## CCA8      1     0.008  0.51  0.941
## CCA9      1     0.004  0.22  1.000
## Residual 60     0.961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ограничения CCA

- Редкие виды - использовать когда репрезентативная выборка или удалить из анализа
- Чтобы снизить вес больших численностей видов - логарифмировать
- Доля объясненной изменчивости (% of total inertia) - аналогично R^2 - смещенная оценка. Простых поправок не придумано.
- Симметричные распределения переменных среды (преобразования)
- Нельзя включать шумные переменные среды
- Нельзя трансформацию Хелингера - будет уже не χ^2

Частная ординация

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

Техника: множественная регрессия предикторов от ковариат. Остатки от этой регрессии (то, что от ковариат не зависит) - в CCA в качестве переменных среды.

Legendre & Legendre 1998

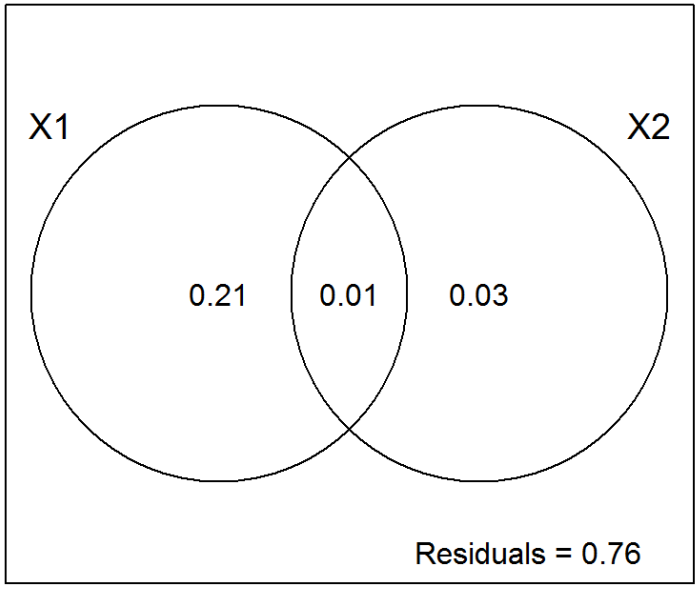
Попробуем частную ординацию и пространственный анализ

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.

```
mod <- varpart(mite, ~ SubsDens + WatrCont + Substrate + Topo, ~ x + y, data =
cbind(mite.env, mite.xy))
```

Выделяем компоненты изменчивости

```
##
## Partition of variation in RDA
##
## Call: varpart(Y = mite, X = ~SubsDens +
WatrCont + Substrate +
## Topo, ~x + y, data = cbind(mite.env, mite.xy))
##
## Explanatory tables:
## X1:  ~SubsDens + WatrCont + Substrate + Topo
## X2:  ~x + y
##
## No. of explanatory tables: 2
## Total variation (SS): 627803
##              Variance: 9098.6
## No. of observations: 70
##
## Partition table:
##                       Df R.squared
Adj.R.squared Testable
## [a+b] = X1             9   0.31605
0.21346     TRUE
## [b+c] = X2             2   0.06584
0.03796     TRUE
## [a+b+c] = X1+X2       11   0.36374
0.24307     TRUE
## Individual fractions
## [a] = X1|X2            9
0.20511     TRUE
## [b]                    0
0.00835    FALSE
## [c] = X2|X1            2
0.02961     TRUE
## [d] = Residuals
0.75693    FALSE
## ---
## Use function 'rda' to test significance of
fractions of interest
```



Borcard et al., 1992

Take home messages

- Канонический корреспондентный анализ позволяет описать, как стуктура изменчивости определяется внешними переменными.
- При помощи частного канонического корреспондентного анализа можно разделять изменчивость на несколько компонент.

Дополнительные ресурсы

- Borcard, D., Gillet, F., Legendre, P., 2011. Numerical ecology with R. Springer.
- Jongman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R. (eds.), 1995. Data analysis in community and landscape ecology. Cambridge University Press
- 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.