Frequently Asked Questions

Jari Oksanen

June 13, 2006

Abstract

The document answers some frequently asked questions and explains some design decisions I have made in vegan.

Contents

1 Scaling in redundancy analysis

This chapter discusses the scaling of scores (results) in redundancy analysis and principal component analysis performed by function `rda` in the `vegan` library. Principal component analysis, and hence redundancy analysis, is a variant of singular value decomposition (SVD). Functions `rda` and `prcomp` (library `mva`) even use SVD internally in their algorithm. In SVD a centred data matrix is decomposed into orthogonal components so that $x_{ij} = \sum_k \sigma_k u_{ik} v_{jk}$, where u_{ik} and v_{jk} are orthonormal coefficient matrices and σ_k are singular values. Orthonormality means that sum of squared columns is one and their cross-product is zero, or $\sum_i u_{ik}^2 = \sum_j v_{jk}^2 = 1$, and $\sum_i u_{ik} u_{il} = \sum_j v_{jk} v_{jl} = 0$ for $k \neq l$. This is a decomposition, and the original matrix is found exactly from the singular vectors and corresponding singular values, and first two singular components give the best rank = 2 least squares estimate of the original matrix.

Principal component analysis is often presented (and performed in legacy software) as an eigenanalysis of covariance matrices. Instead of data matrix, we analyse a matrix of covariances and variances $\mathbf{S}$. The result will be orthonormal coefficient matrix $\mathbf{U}$ and eigenvalues $\mathbf{\Lambda}$. The coefficients u_{ik} ares identical to SVD (except for possible sign changes), and eigenvalues λ_k are related to the corresponding singular values by $\lambda_k = \sigma_k^2/(n-1)$. With classical definitions, the sum of all eigenvalues equals the sum of variances of species, or $\sum_k \lambda_k = \sum_j s_j^2$, and it is often said that first axes explain a certain maximized proportion of total

Table 1: Alternative scalings for RDA used in the functions `prcomp` and `princomp` (package `mva`), and the one used in the `vegan` function `rda` and the proprietary software `Canoco` scores in terms of orthonormal species (u_{ik}) and site scores (v_{jk}), eigenvalues (λ_k), number of sites (n) and species standard deviations (s_j). In `rda`, const $= \sqrt[4]{(n-1)\sum \lambda_k}$. Corresponding negative scaling in `vegan` and corresponding positive scaling in `Canoco` is derived dividing each species by its standard deviation s_j (possibly with some additional constant multiplier).

	Site scores u^*_{ik}	**Species scores** v^*_{jk}
`prcomp, princomp`	$u_{ik}\sqrt{n-1}\sqrt{\lambda_k}$	v_{jk}
`rda, scaling=1`	$u_{ik}\sqrt{\lambda_k/\sum \lambda_k} \times \text{const}$	$v_{jk} \times \text{const}$
`rda, scaling=2`	$u_{ik} \times \text{const}$	$v_{jk}\sqrt{\lambda_k/\sum \lambda_k} \times \text{const}$
`rda, scaling=3`	$u_{ik}\sqrt[4]{\lambda_k/\sum \lambda_k} \times \text{const}$	$v_{jk}\sqrt[4]{\lambda_k/\sum \lambda_k} \times \text{const}$
`rda, scaling < 0`	u^*_{ik}	$\sqrt{\sum \lambda_k/(n-1)}s_j^{-1}v^*_{jk}$
`Canoco, scaling=-1`	$u_{ik}\sqrt{n}\sqrt{\lambda_k/\sum \lambda_k}$	$v_{jk}\sqrt{n}$
`Canoco, scaling=-2`	$u_{ik}\sqrt{n}$	$v_{jk}\sqrt{n}\sqrt{\lambda_k/\sum \lambda_k}$
`Canoco, scaling=-3`	$u_{ik}\sqrt{n}\sqrt[4]{\lambda_k/\sum \lambda_k}$	$v_{jk}\sqrt{n}\sqrt[4]{\lambda_k/\sum \lambda_k}$

variance in the data. The other orthonormal matrix $\mathbf{V}$ can be found indirectly as well, so that we have the same components in both methods.

The coefficients u_{ik} and v_{jk} are of the same (unit) length for all axes k, but singular values σ_k or eigenvalues λ_k give the information of the importance of axes, or the 'axis lengths.' Instead of the orthonormal coefficients, or equal length axes, it is customary to use eigenvalues to scale at least one of the alternative scores to reflect the importance of axes or describe the true configuration of points. Table 1 shows some alternative scalings used in various software. These alternatives apply to principal components analysis in all cases, and in redundancy analysis, they apply to species scores and constraints or linear combination scores; weighted averaging scores have somewhat wider dispersion.

In community ecology, it is common to plot both species and sites in the same graph. If this graph is a graphical display of SVD, or a graphical, low-dimensional approximation of the data, the graph is called a biplot. The graph is a biplot if the transformed scores satisfy $x_{ij} = c\sum_k u^*_{ij}v^*_{jk}$ where c is a scaling constant. In functions `princomp`, `prcomp` and `rda`, $c = 1$ or the plotting scores are the straight biplot scores so that the singular values (or eigenvalues) are expressed for sites, and species are left unscaled. For `Canoco` $c = n^{-1}\sqrt{n-1}\sqrt{\sum \lambda_k}$ with positive `Canoco` scaling values. All these c are constants for a matrix, so these are all biplots with different internal scaling of species and site scores with respect to each other. For `Canoco` with postive scaling values and `vegan` with negative scaling values, no constant c can be found, but the correction is dependent on species standard deviations s_j, so this alternative does not define a biplot.

There is no natural way of scaling species and site scores to each other, but all functions and programs above selected different strategies. The eigenvalues in redundancy and principal components analysis are scale dependent and change when the the data are multiplied by a constant. If we have percent cover data, the eigenvalues are typically very high, and the scores scaled by eigenvalues will have much wider dispersion than the orthonormal set. If we express the

percentages as proportions, or divide the matrix by 100, the eigenvalues will be reduced by factor 100^2, and the scores scaled by eigenvalues will have much narrower dispersion than the orthonormal set. For graphical biplots we should be able to fix the relation and make it invariant for scale changes. The solution adoption in the R standard function `biplot.princomp` is to scale site and species scores independently, and typically very differently, but plot each with separate scales so that both sets fill the graph area. The solution in `Canoco` and `rda` is to use proportional eigenvalues $\lambda_k / \sum \lambda_k$ instead of original eigenvalues. These proportions are invariant with scale changes, and typically they have a nice range for plotting two data sets in the same graph.

In this chapter, I used always centred data matrices. In principle SVD could be done with original, non-centred data, but there is no option for this in `rda`, because I think that non-centred analysis is dubious and I do not want to encourage its use (if you think you need it, you are certainly so good in programming that you can change that one line in `rda.default`). I do think that the arguments for non-centred analysis are often twisted, and the method is not very good for its intended purpose, but there are better methods for finding fuzzy classes. Normal, centred analysis moves the origin to the average of all species, and the dimensions describe differences from this average. Non-centred analysis leaves the origin in the empty site with no species, and the first axis usually runs from the empty site to the average site. Second and third non-centred components are often very similar to first and second (etc.) centred components, and the best way to use non-centred analysis is to discard the first component and use only the rest. This is better done with directly centred analysis.

2 Why to use weighted averages scores instead of linear combinations in constrained ordination

Constrained ordination methods such as Constrained Correspondence Analysis (CCA) and Redundancy Analysis (RDA) produce two kind of site scores [3, 4]:

- LC or Linear Combination Scores which are linear combinations of constraining variables.
- WA or Weighted Averages Scores which are such weighted averages of species scores that are as similar to LC scores as possible.

Many computer programs for constrained ordinations give only or primarily LC scores, following Mike Palmer's recommendation [3]. However, functions `cca` and `rda` in the `vegan` package use primarily WA scores. This chapter explains the reasons for this choice.

Briefly, the main reasons are that

- LC scores *are* linear combinations, so they give us only the (scaled) environmental variables. This means that they are independent of vegetation and cannot be found from the species composition. Moreover, identical combinations of environmental variables give identical LC scores irrespective of vegetation.

- Bruce McCune has demonstrated that noisy environmental variables result in deteriorated LC scores whereas WA scores tolerate some errors in environmental variables [2]. All environmental measurements contain some errors, and therefore it is safer to use WA scores.

This articles studies mainly the first point. The users of `vegan` have a choice of either LC or WA (default) scores, but after reading this article, I believe that most of them do not want to use LC scores, because they are not what they were looking for in ordination.

2.1 LC Scores are Linear Combinations

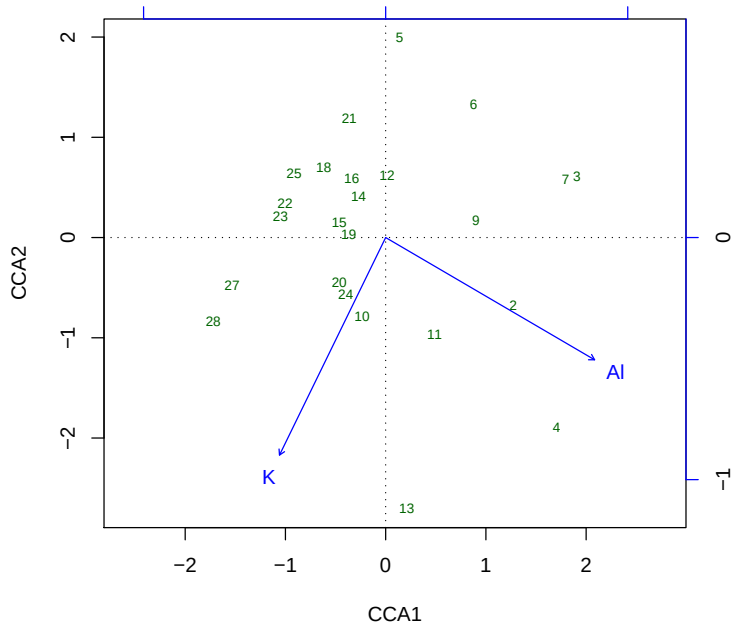
Let us perform a simple CCA analysis using only two environmental variables so that we can see the constrained solution completely in two dimensions:

```
> library(vegan)
> data(varespec)
> data(varechem)
> orig <- cca(varespec ~ Al + K, varechem)
```

Function `cca` in `vegan` uses WA scores as default. So we must specifically ask for LC scores (Fig. 1).

Figure 1 LC scores in CCA of the original data.

```
> plot(orig, dis = c("lc", "bp"))
```

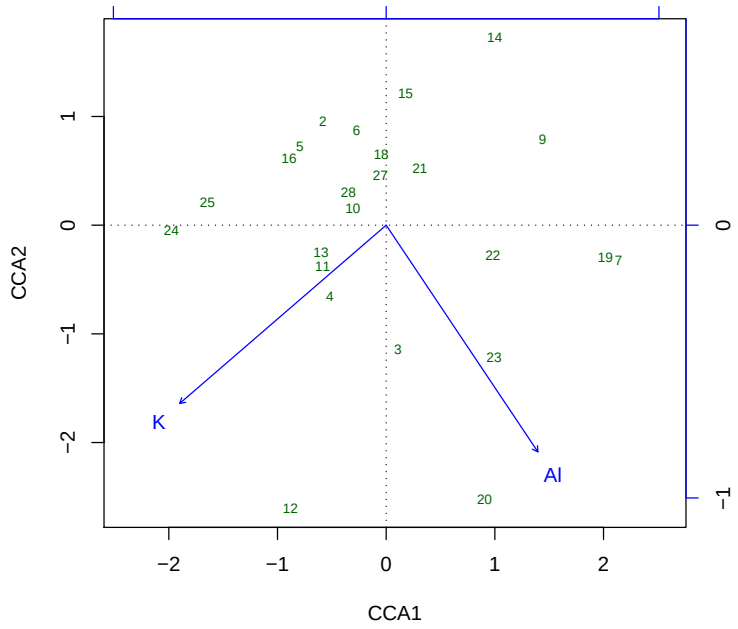


What would happen to linear combinations of LC scores if we shuffle the ordering of sites in species data? Function `sample()` below shuffles the indices.

```
> i <- sample(nrow(varespec))
> shuff <- cca(varespec[i, ] ~ Al + K, varechem)
```

Figure 2 LC scores of shuffled species data.

```
> plot(shuff, dis = c("lc", "bp"))
```



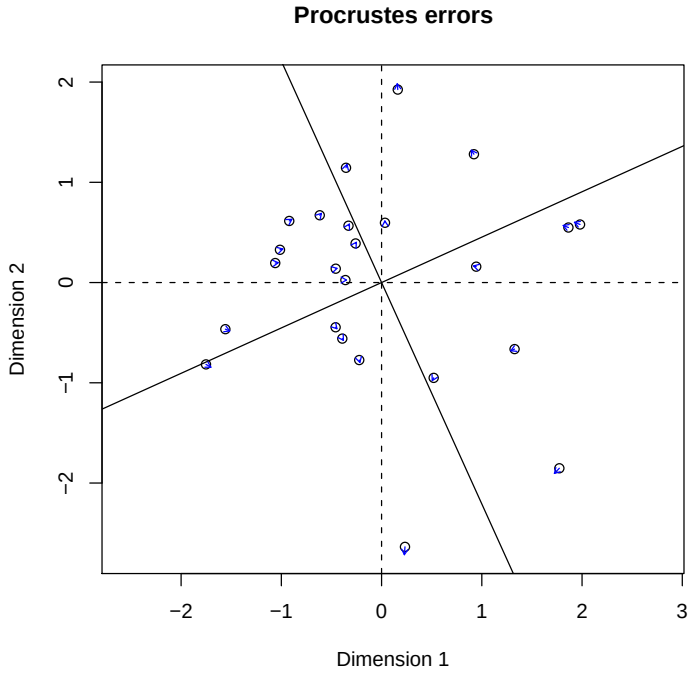
It seems that site scores are fairly similar, but oriented differently (Fig. 2). We can use Procrustes rotation to see how similar the site scores indeed are (Fig. 3). There is a small difference, but this will disappear if we use Redundancy Analysis (RDA) instead of CCA (Fig. 4). Here we use a new shuffling as well.

```
> tmp1 <- rda(varespec ~ Al + K, varechem)
> i <- sample(nrow(varespec))
> tmp2 <- rda(varespec[i, ] ~ Al + K, varechem)
```

LC scores indeed are linear combinations of constraints (environmental variables) and *independent of species data*: You can shuffle your species data, or change the data completely, but the LC scores will be unchanged in RDA. In CCA the LC scores are *weighted* linear combinations with site totals of species data as weights. Shuffling species data in CCA changes the weights, and this can cause changes in LC scores. The magnitude of changes depends on the variability of site totals.

Figure 3 Procrustes rotation of LC scores from CCA of original and shuffled data.

```
> plot(procrustes(scores(orig, dis = "lc"), scores(shuff, dis = "lc")))
```



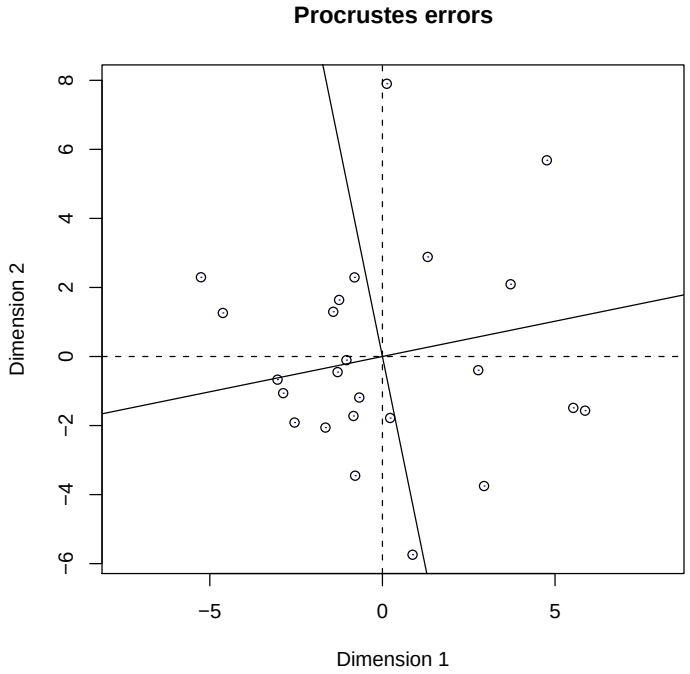
The original data and shuffled data differ in their goodness of fit[1].

```
> orig

Call:
cca(formula = varespec ~ Al + K, data = varechem)

              Inertia Rank
Total           2.083
Constrained     0.476    2
Unconstrained   1.607   21
Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:
  CCA1   CCA2
0.3608 0.1152

Eigenvalues for unconstrained axes:
    CA1     CA2     CA3     CA4     CA5     CA6     CA7     CA8
```

[1]Or probably differ: The randomization is done while generating this article, and different versions may have different randomizations.

Figure 4 Procrustes rotation of LC scores in RDA of the original and shuffled data.

```
> plot(procrustes(scores(tmp1, dis = "lc"), scores(tmp2, dis = "lc")))
```



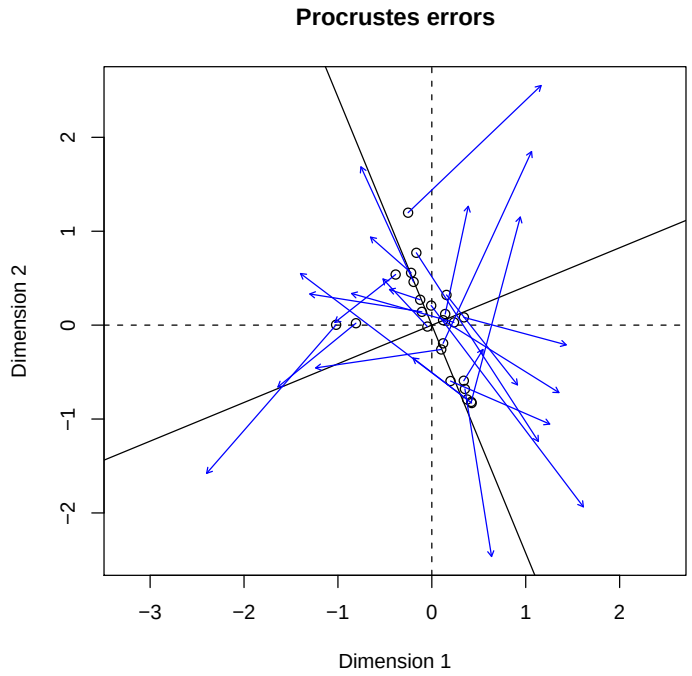

```
0.37476 0.24036 0.19696 0.17818 0.15209 0.11840 0.08364 0.07567
(Showed only 8 of all 21 unconstrained eigenvalues)

> shuff

Call:
cca(formula = varespec[i, ] ~ Al + K, data = varechem)

              Inertia Rank
Total          2.0832
Constrained    0.1174    2
Unconstrained  1.9658   21
Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:
   CCA1    CCA2
0.08581 0.03156

Eigenvalues for unconstrained axes:
   CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8
0.5062 0.3530 0.2163 0.1836 0.1726 0.1210 0.0942 0.0882
```

```
(Showed only 8 of all 21 unconstrained eigenvalues)
```

Similarly their WA scores will be (probably) very different (Fig. 5).

Figure 5 Procrustes rotation of WA scores of CCA with the original and shuffled data.

```
> plot(procrustes(orig, shuff))
```



The example used only two environmental variables so that we can easily plot all constrained axes. With a larger number of environmental variables the full configuration remains similarly unchanged, but its orientation may change, so that two-dimensional projections look different. In the full space, the differences should remain within numerical precision:

```
> tmp1 <- rda(varespec ~ ., varechem)
> tmp2 <- rda(varespec[i, ] ~ ., varechem)
> tmp1

Call:
rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe +      Mn + Zn + Mo + Baresoil

              Inertia Rank
Total          1825.7
Constrained    1459.9   14
Unconstrained   365.8    9
Inertia is variance
```

```
Eigenvalues for constrained axes:
    RDA1     RDA2     RDA3     RDA4     RDA5     RDA6     RDA7     RDA8
820.1042 399.2847 102.5617  47.6317  26.8382  24.0481  19.0644  10.1670
    RDA9    RDA10    RDA11    RDA12    RDA13    RDA14
  4.4288   2.2720   1.5353   0.9255   0.7155   0.3119

Eigenvalues for unconstrained axes:
    PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
186.192  88.464  38.188  18.402  12.839  10.552   5.519   4.521   1.092
```

```
> proc <- procrustes(scores(tmp1, dis = "lc", choi = 1:14), scores(tmp2,
+     dis = "lc", choi = 1:14))
> max(residuals(proc))

[1] 2.458252e-14
```

In `cca` the difference would be somewhat larger than now observed 2.4583e-14 because site weights used for environmental variables are shuffled with the species data.

2.2 Factor constraints

It seems that users often get confused when they perform constrained analysis using only one factor (class variable) as constraint. The following example uses the classical dune meadow data [1]:

```
> data(dune)
> data(dune.env)
> summary(dune.env)

       A1         Moisture Management       Use    Manure
 Min.   : 2.800   1:7      BF:3       Hayfield:7   0:6
 1st Qu.: 3.500   2:4      HF:5       Haypastu:8   1:3
 Median : 4.200   4:2      NM:6       Pasture :5   2:4
 Mean   : 4.850   5:7      SF:6                    3:4
 3rd Qu.: 5.725                                    4:3
 Max.   :11.500
```

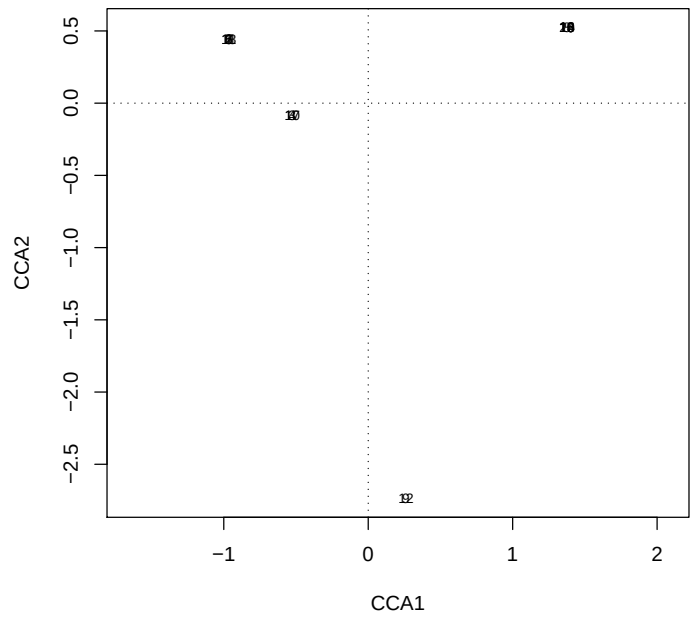
```
> orig <- cca(dune ~ Moisture, dune.env)
> orig

Call:
cca(formula = dune ~ Moisture, data = dune.env)

              Inertia Rank
Total          2.1153
Constrained    0.6283    3
Unconstrained  1.4870   16
Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:
```

```
  CCA1   CCA2   CCA3
0.4187 0.1330 0.0766

Eigenvalues for unconstrained axes:
     CA1      CA2      CA3      CA4      CA5      CA6      CA7      CA8
0.409782 0.225913 0.176062 0.123389 0.108171 0.090751 0.085878 0.060894
     CA9     CA10     CA11     CA12     CA13     CA14     CA15     CA16
0.056606 0.046688 0.041926 0.020103 0.014335 0.009917 0.008505 0.008033
```

When the results are plotted using LC scores, sample plots fall only in four alternative positions (Fig. 6). In the previous chapter we saw that this happens

Figure 6 LC scores of the dune meadow data using only one factor as a constraint.

```
> plot(orig, dis = "lc")
```

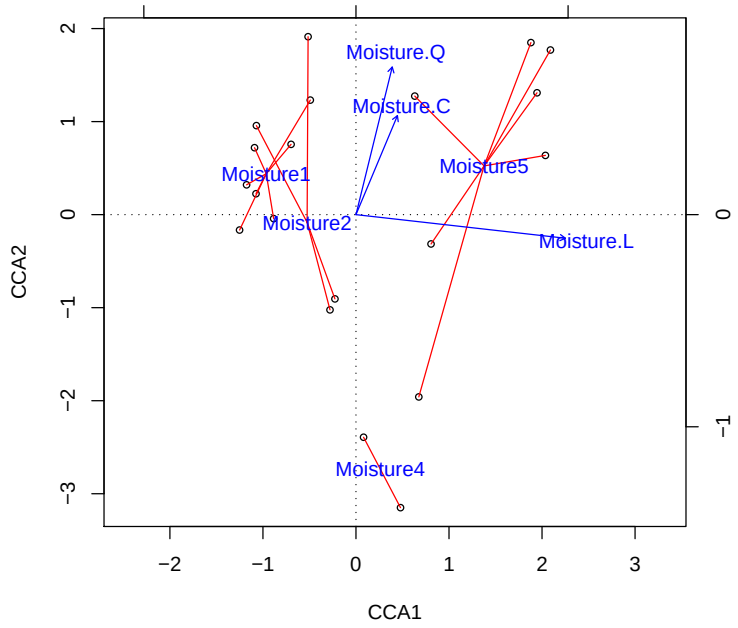


because LC scores *are* the environmental variables, and they can be distinct only if the environmental variables are distinct. However, normally the user would like to see how well the environmental variables separate the vegetation, or inversely, how we could use the vegetation to discriminate the environmental conditions. For this purpose we should plot WA scores, or LC scores and WA scores together: The LC scores show where the site *should* be, the WA scores shows where the site *is*.

Function `ordispider` adds line segments to connect each WA score with the corresponding LC (Fig. 7). This is the standard way of displaying results of

Figure 7 A "spider plot" connecting WA scores to corresponding LC scores. The shorter the web segments, the better the ordination.

```
> plot(orig, display = "wa", type = "points")
> ordispider(orig, col = "red")
> text(orig, dis = "cn", col = "blue")
```



discriminant analysis, too. Moisture classes `1` and `2` seem to be overlapping, and cannot be completely separated by their vegetation. Other classes are more distinct, but there seems to be a clear arc effect or a "horseshoe" despite using CCA.

2.3 Conclusion

LC scores are only the (weighted and scaled) constraints and independent of vegetation. If you plot them, you plot only your environmental variables. WA scores are based on vegetation data but are constrained to be as similar to the LC scores as only possible. Therefore `vegan` calls LC scores as `constraints` and WA scores as `site scores`, and uses primarily WA scores in plotting. However, the user makes the ultimate choice, since both scores are available.

3 t-values in constrained ordination

`Vegan` does not report t-values for constrained ordination methods. The reason for this omission is that the suggested t-values are wrong and misleading. This section first explains the arguments against t-values and then shows how you can get them if you really want to do so.

The t-values are based on the multivariate regression model where the weighted averages (WA) scores are dependent variables and environmental variables (constraints) are independent variables. That is, we try to explain the ordination scores by constraints. The t-values are based on the standard errors of the coefficients of this regression model. The regression model only uses the constrained ordination and ignores the residual variation in the data, and gives much too narrow confidence intervals.

The easiest way of obtaining the t-values is to refit the constrained ordination model as a multivariate linear model (`mlm`) which automatically is fitted with the standard `lm` function when the response is multivariate. The following implements a similar model as in CANOCO[2] [5]:

```
> as.mlm.cca <- function(x) {
+     w <- weights(x)
+     wa <- sweep(x$CCA$wa, 1, sqrt(w), "*")
+     X <- qr.X(x$CCA$QR)
+     lm(wa ~ ., data = as.data.frame(X))
+ }
```

Essentially, this is nothing but a multiple linear model weighted with row totals w_i. The weights appear to be different for WA scores and constraints, but the latter are extracted by `qr.X` function and already were weighted by $\sqrt{w_i}$, but WA scores are unweighted. The same function also works with `rda` and `capscale`, but these have constant weights $w_i = 1$ and the function can be simplified.

The t-values can be found in the `summary` of this model:

```
> mod <- cca(varespec ~ pH + Ca, varechem)
> mlmod <- as.mlm.cca(mod)
> summary(mlmod)

Response CCA1 :

Call:
lm(formula = CCA1 ~ ., data = as.data.frame(X))

Residuals:
      Min        1Q    Median        3Q       Max
-0.312972 -0.143757  0.004526  0.156529  0.248044

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0007043  0.0372082  -0.019 0.985077
```

[2]It seems that in more correctly weighted model you should multiply WA scores with row weights w_i and constraints with $\sqrt{w_i}$.

```
pH           3.4344602  0.8582352   4.002 0.000647 ***
Ca          -0.0031683  0.0007552  -4.195 0.000407 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1822 on 21 degrees of freedom
Multiple R-Squared: 0.5891,        Adjusted R-squared:  0.55
F-statistic: 15.05 on 2 and 21 DF,  p-value: 8.797e-05


Response CCA2 :

Call:
lm(formula = CCA2 ~ ., data = as.data.frame(X))

Residuals:
     Min       1Q   Median       3Q      Max
-0.63461 -0.32304 -0.02968  0.26146  0.77031

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.011424   0.088281  -0.129    0.898
pH           3.213052   2.036263   1.578    0.130
Ca           0.002672   0.001792   1.491    0.151

Residual standard error: 0.4323 on 21 degrees of freedom
Multiple R-Squared: 0.2026,        Adjusted R-squared: 0.1266
F-statistic: 2.668 on 2 and 21 DF,  p-value: 0.09282
```

The weighted data are centred, and the intercept should be zero (and it would be with alternative weighting scheme) Would these t-values be correct, we also could use the standard `anova` of the refitted model to test the significance of the constraints:

```
> anova(mlmod)

Analysis of Variance Table

            Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)  1  0.0019   0.0193      2     20 0.9808700
pH           1  0.4062   6.8420      2     20 0.0054459 **
Ca           1  0.5065  10.2649      2     20 0.0008562 ***
Residuals   21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It is easy to see that P-values of the anova are much better (smaller) than from the permutation tests:

```
> anova(mod, by = "terms", permut = 200)
```

```
Permutation test for cca under direct model
Terms added sequentially (first to last)

Model: cca(formula = varespec ~ pH + Ca, data = varechem)
         Df  Chisq      F N.Perm Pr(>F)
pH        1 0.1458 1.7453    200   0.07 .
Ca        1 0.1827 2.1864    200   0.02 *
Residual 21 1.7547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This also means that the t-values have the same bias.

References

[1] R. H. Jongman, C. J. F. ter Braak, and O. F. R. van Tongeren. *Data analysis in community and landscape ecology.* Pudoc, Wageningen, 1987.

[2] B. McCune. Influence of noisy environmental data on canonical correspondence analysis. *Ecology*, 78:2617–2623, 1997.

[3] M. W. Palmer. Putting things in even better order: The advantages of canonical correspondence analysis. *Ecology*, 74:2215–2230, 1993.

[4] C. J. F. ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67:1167–1179, 1986.

[5] C. J. F ter Braak and P. Šmilauer. *CANOCO reference manual and user's guide to Canoco for Windows: Software for canonical community ordination (version 4).* Microcomputer Power, Ithaca, NY, 1998.