R包vegan的典范对应分析（CCA）

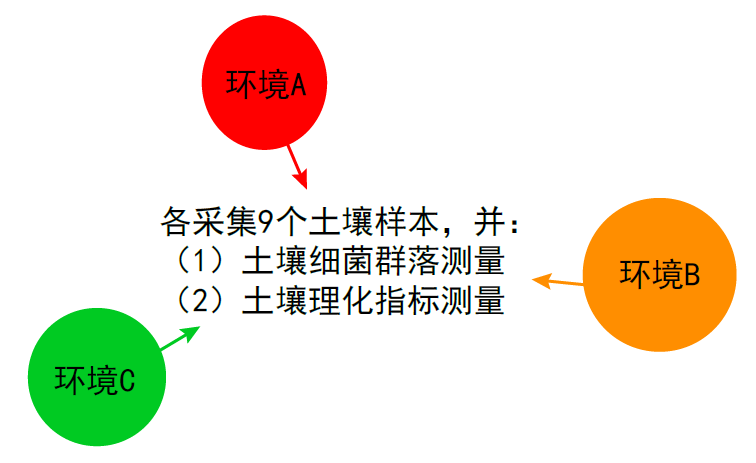
典范对应分析（canonical correspondence analysis，CCA）是单峰约束排序方法，是[对应分析](http://blog.sciencenet.cn/blog-3406804-1206643.html)（CA）与多元回归的结合，其算法源自[冗余分析](http://blog.sciencenet.cn/blog-3406804-1182489.html)（RDA）。与仅使用物种多度数据的CA相比，CCA可以结合多种环境变量一起分析，从而更好地反映群落与环境的关系。

类似CA，原始的响应变量矩阵（物种多度数据集）首先被转化为一个描述样方对对Pearson χ2统计量的贡献率的Ǭ矩阵；之后通过Ǭ矩阵完成回归集，与RDA相比，使用加权多元回归代替简单多元回归。

本篇以某16S扩增子测序所得细菌群落数据，简介R包vegan的CCA过程。

示例文件、R脚本等的百度盘链接：

<https://pan.baidu.com/s/1UWlvIUWrnl4gyoEhpCGVtw>



某研究对三种环境中的土壤进行了采样（环境A、B、C，各环境下采集9个土壤样本），结合二代测序观测了土壤细菌群落物种组成，并对多种土壤理化指标进行了测量。获得两个数据集。

（1）文件“otu\_table.txt”为通过16S测序所得的细菌OTU丰度表格，下文统称为“物种”，尽管OTU并不完全等于物种。

（2）文件“env\_table.txt”记录了多种土壤理化指标信息，即环境数据。

接下来期望通过CCA分析这些数据，用以分析群落物种组成与环境的关系（作为示例，暂且忽略本数据更适用线性模型还是单峰模型）。

## vegan包执行CCA

vegan包中，通过cca()执行CCA，该函数的输入格式大致如下（与[RDA](http://blog.sciencenet.cn/blog-3406804-1182489.html)的rda()函数输入格式相似）。

library(vegan)

#详情 ?cca

#调用格式 1

cca(Y, X, W)

#或者格式 2

cca(Y~var1+var2+var3+factorA+var2\*var3+Condition(var4))

在第1种调用格式中，Y为响应变量矩阵，X为解释变量矩阵，W为偏CCA的协变量矩阵。当不考虑协变量时，可直接输入响应变量矩阵Y和解释变量矩阵X，即“cca(Y, X)”。这种写法虽然简单，但要求解释变量矩阵或协变量矩阵中不能含有因子变量（定性变量）。（当仅存在响应变量矩阵Y时，则执行CA排序）

因此，通常建议使用第2种写法。其中Y为响应变量矩阵，var1、var2、var3为定量解释变量，factorA为因子变量，var2\*var3为变量2和变量3的交互作用项，var4为协变量。与上述第1种写法相比，不仅能够识别因子变量，还能够对变量间的交互作用加以考虑。

本示例中，环境变量全部为定量解释变量，且暂不考虑解释变量间的交互作用以及协变量的情况。因此两种写法分别为：

#读入物种数据，以细菌 OTU 水平丰度表为例

otu <- read.delim('otu\_table.txt', row.names = 1, sep = '\t', stringsAsFactors = FALSE, check.names = FALSE)

otu <- data.frame(t(otu))

#读取环境数据

env <- read.delim('env\_table.txt', row.names = 1, sep = '\t', stringsAsFactors = FALSE, check.names = FALSE)

#CCA，详情 ?cca

#调用格式 1；这里无协变量矩阵，所以直接输入响应变量矩阵和解释变量矩阵

otu\_cca <- cca(otu, env)

#或者格式 2；Y~. 是 Y~var1+var2+...+varn 的简写，不涉及交互作用及协变量

otu\_cca <- cca(otu~., env)

## CCA结果的初步解读

CCA可获得的约束轴数为min[p–1, m, n–1]。其中，p为响应变量（物种）数量；m为定量解释变量数量以及定性解释变量（因子变量）的因子水平的自由度（即该变量因子水平数减1）；n为排序对象（样方）数量。

CCA中，排序轴承载的总变差≠总方差，而是通过一个叫均方列联系数（mean squared contingency coefficient）或称总惯量（total inertia）的指标表征，它代表了Ǭ矩阵所有值的平方和。CCA中R2即代表了总惯量（而非总方差）被环境变量所解释的程度，约束轴承载了被成功解释的惯量部分。由于CCA与CA共享一套基础算法，CCA是在CA的基础上添加约束过程发展而来，其约束算法源自RDA中使用的多元回归。因此其很多特征与CA（体现在样方与物种的关系）或RDA（体现在环境变量的解释规则）相似，可分别参考前文[CA](http://blog.sciencenet.cn/blog-3406804-1206630.html)或[RDA](http://blog.sciencenet.cn/blog-3406804-1182485.html)。

**CCA结果总概**

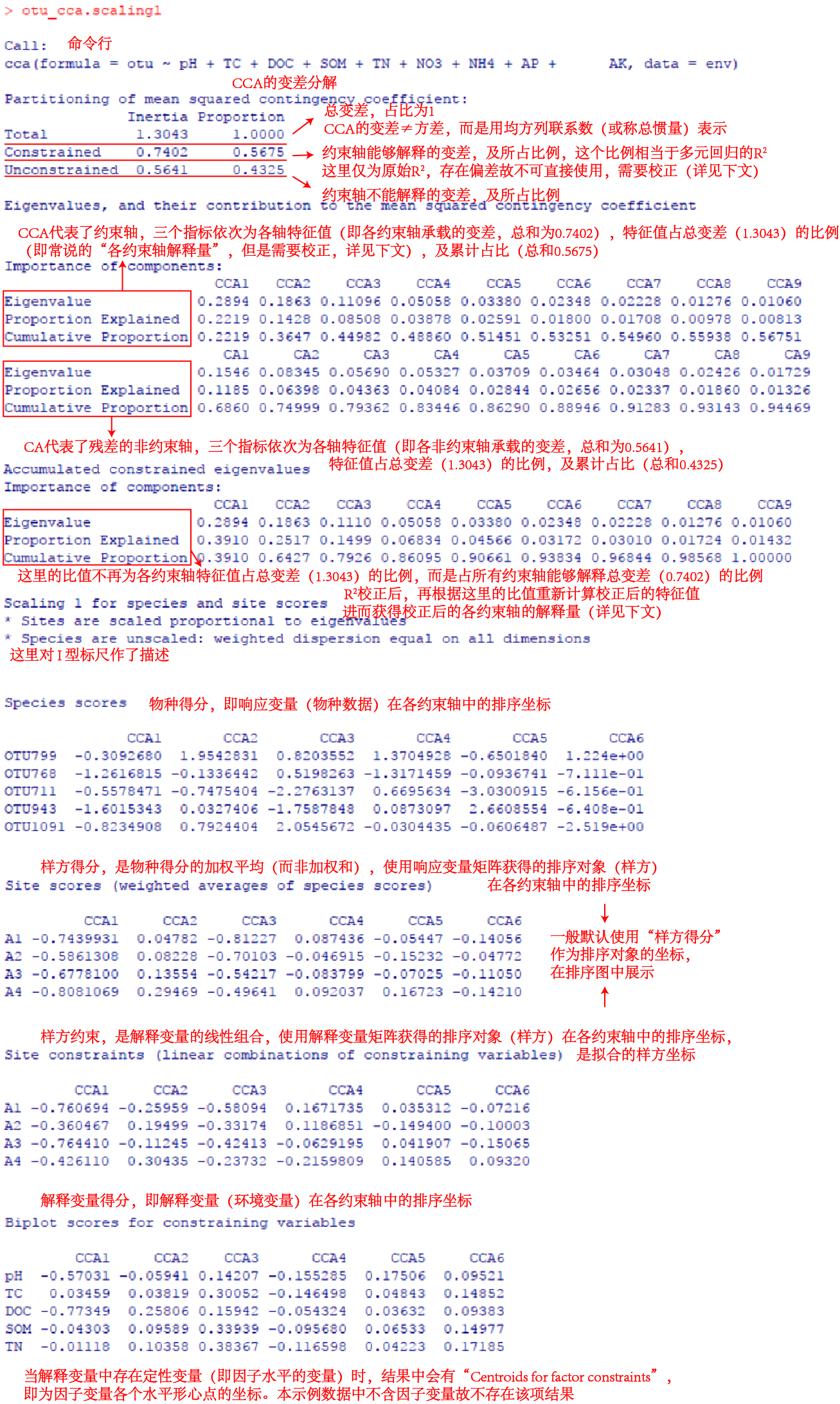
通过summary()查看CCA结果概要。

#查看统计结果信息，以 I 型标尺为例

otu\_cca.scaling1 <- summary(otu\_cca, scaling = 1)

otu\_cca.scaling1

#关于 I、II 型标尺的解读，见下文排序图



**CCA排序图**

CCA三序图中一个非常瞩目的特征是物种在约束轴的排序位置反映其生态梯度最适点，这个特征使物种组成的生物学解释更加直观和容易。

CCA排序图中，样方和物种常用点表示，定性解释变量同样以点表示（质心位置），定量解释变量以向量表示。对于CCA中两种主要标尺的解读方式如下所述。

对于样方和物种的关系，其解释与[CA](http://blog.sciencenet.cn/blog-3406804-1206643.html)中的解释相同。

I型标尺图中，（1）排序图内样方之间的距离近似于它们的χ2距离，排序图中两个样方点越近，代表这些样方内的物种组成越相似；（2）一个样方点靠近一个物种点，表示该物种对于该样方的贡献比较大。

II型标尺图中，（1）排序图内物种之间的距离近似于它们的χ2距离，排序图中两个物种点越近，代表它们的相对多度沿样方分布越相似；（2）一个物种点靠近一个样方点，表示该物种在该样方内存在的可能性很大，或在该样方内的多度比在其它样方内大。

对于解释变量（环境变量）与样方或物种的关系。

I型标尺图中，（1）将对象点垂直投影到变量向量或延长线上，投影点位置接近该样方内该解释变量数值的位置。（2）定性解释变量质心的点靠近某一样方，表明其在该样方中的状态更可能为“1”。

II型标尺图中，（1）将物种点垂直投影到变量向量或延长线上，投影点的位置表示该物种在该环境变量梯度上的最适区域。（2）定性解释变量质心的点靠近某一物种，表明该物种更可能出现在（或大量存在）该变量的状态为“1”的样方中。

对于本示例的CCA结果，如下为一些简略的排序图展示，便于观测数据。关于更美观的可视化方案，可参考本文末。

#作图查看排序结果，详情 ?plot.cca

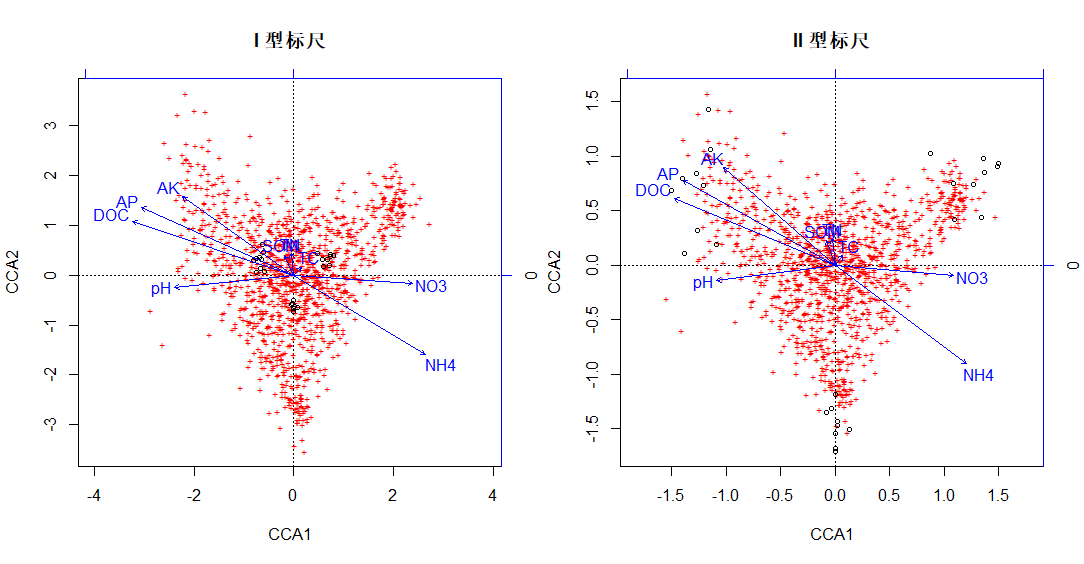
#三序图，包含 I 型标尺和 II 型标尺，样方坐标展示为使用物种加权计算的样方得分

par(mfrow = c(1, 2))

plot(otu\_cca, scaling = 1, main = 'I 型标尺', display = c('wa', 'sp', 'cn'))

plot(otu\_cca, scaling = 2, main = 'II 型标尺', display = c('wa', 'sp', 'cn'))

在这里，I型标尺中，样方得分是物种得分加权平均，故排序图中物种通常分布在样方范围之外；II型标尺中，物种得分是样方得分的平均值，并按出现该物种的所有样方中该物种丰度加权，故排序图中样方通常分布在物种范围之外。



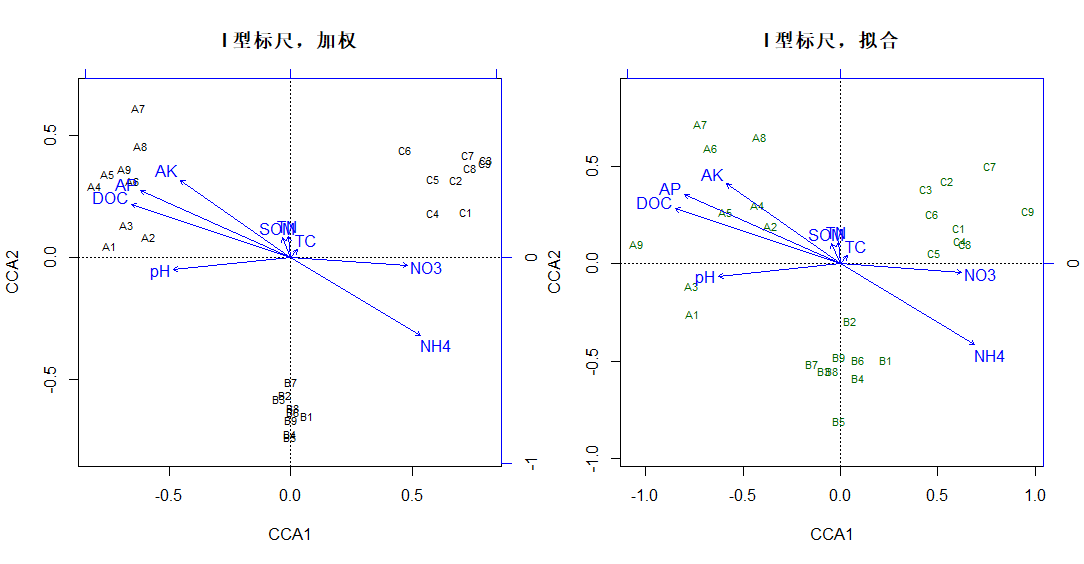
#比较分别使用物种加权计算的样方坐标以及拟合的样方坐标的差异

#隐藏物种，以 I 型标尺为例展示双序图

par(mfrow = c(1, 2))

plot(otu\_cca, scaling = 1, main = 'I 型标尺，加权', display = c('wa', 'cn'))

plot(otu\_cca, scaling = 1, main = 'I 型标尺，拟合', display = c('lc', 'cn'))



**主要信息提取**

若有需要，考虑在结果中将需要的信息提取出来。

#scores() 提取排序得分（坐标），以 I 型标尺为例，前四轴为例

#使用物种加权和计算的样方得分

otu\_cca\_site.scaling1 <- scores(otu\_cca, choices = 1:4, scaling = 1, display = 'wa')

#物种变量（响应变量）得分

otu\_cca\_sp.scaling1 <- scores(otu\_cca, choices = 1:4, scaling = 1, display = 'sp')

#环境变量（解释变量）得分

otu\_cca\_env.scaling1 <- scores(otu\_cca, choices = 1:4, scaling = 1, display = 'bp')

#或者在 summary() 后提取，以 I 型标尺为例，前四轴为例

otu\_cca.scaling1 <- summary(otu\_cca, scaling = 1)

#使用物种加权和计算的样方得分

otu\_cca\_site.scaling1 <- otu\_cca.scaling1$site[ ,1:4]

#物种

otu\_cca\_sp.scaling1 <- otu\_cca.scaling1$species[ ,1:4]

#环境

otu\_cca\_env.scaling1 <- otu\_cca.scaling1$biplot[ ,1:4]

#若需要输出在本地

#样方

write.table(data.frame(otu\_cca\_site.scaling1), 'otu\_cca\_site.scaling1.txt', sep = '\t', col.names = NA, quote = FALSE)

#物种

write.table(data.frame(otu\_cca\_sp.scaling1), 'otu\_cca\_sp.scaling1.txt', sep = '\t', col.names = NA, quote = FALSE)

#环境

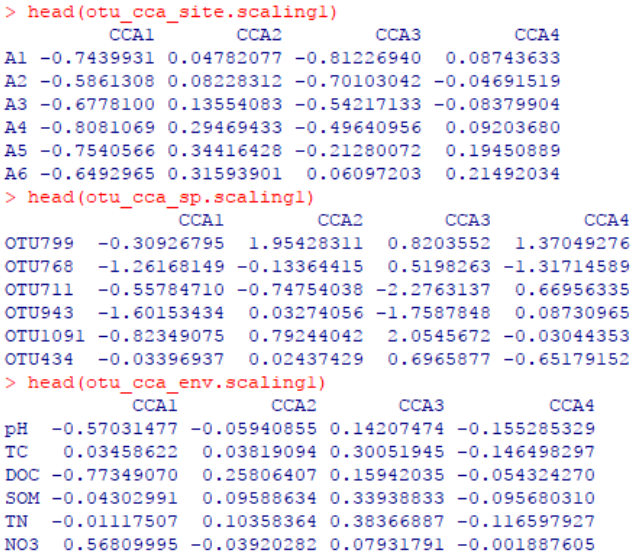
write.table(data.frame(otu\_cca\_env.scaling1), 'otu\_cca\_env.scaling1.txt', sep = '\t', col.names = NA, quote = FALSE)

#不建议直接在原始数据集中提取，因为这样提取的坐标数值未经标尺缩放处理，不利于反映生物学问题

#otu\_cca$CCA$u[ ,1:4]

#otu\_cca$CCA$v[ ,1:4]

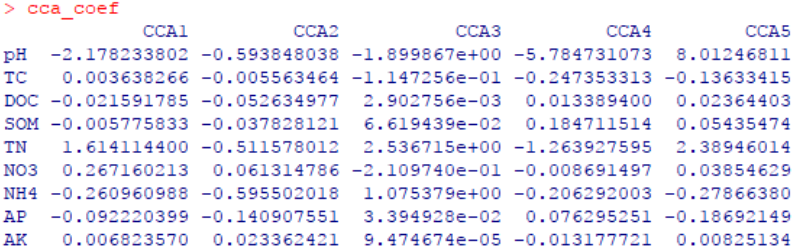
#otu\_cca$CCA$biplot[ ,1:4]



#若对典范特征系数（即每个解释变量与每个约束轴之间的回归系数）感兴趣

#coef() 提取 CCA 典范系数

cca\_coef <- coef(otu\_cca)



## CCA的R2校正和约束轴的置换检验

和多元回归的R2一样，CCA约束轴承载的总变差（响应变量矩阵的总变差能够被解释变量解释的部分，如果用比例表示，即等同于多元回归的R2，也就是约束轴的总解释量）以及各约束轴解释量等，有待校正。初始所得的CCA模型需要经过校正和检验后才可使用，否则可能会带来较大的偏差，原因可[参考前文](http://blog.sciencenet.cn/blog-3406804-1182486.html)。

**校正R2及约束轴特征值**

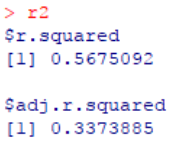
对于校正前后R2的提取，可使用RsquareAdj()完成。R2经过校正后通常会降低，这是必然的。

#RsquareAdj() 提取 R2，详情 ?RsquareAdj()

r2 <- RsquareAdj(otu\_cca)

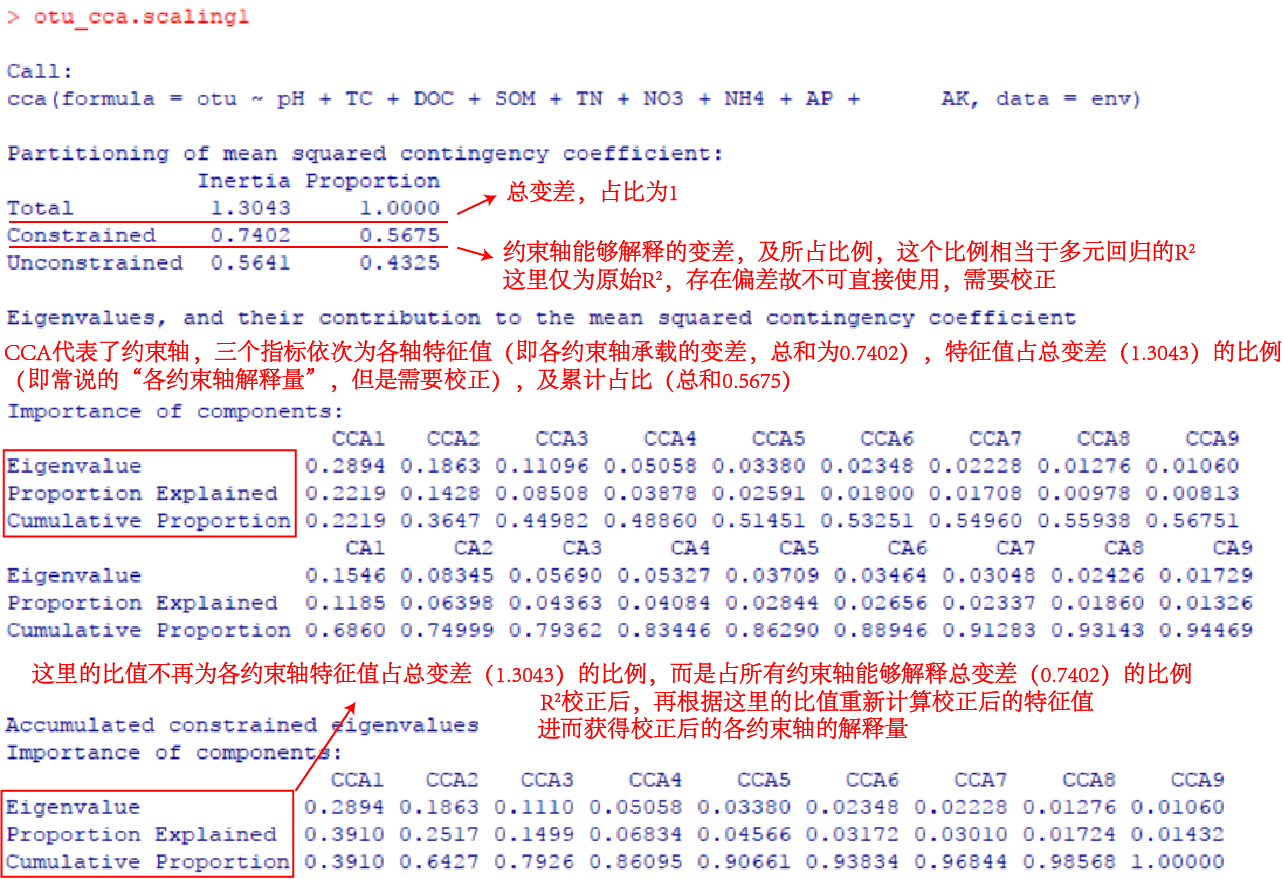
otu\_cca\_noadj <- r2$r.squared #原始 R2

otu\_cca\_adj <- r2$adj.r.squared #校正后的 R2



对于示例数据的CCA排序结果，通过上文，我们已经得知了响应变量矩阵的总变差为1.3043，未校正前的约束轴解释的变差总和为0.7402（占比0.5675，即未校正前的R2）；那么根据校正后的R2（0.33739），可还原校正R2后的约束轴解释的变差总和约为“1.3043×0.33739=0.4401”。

对于各约束轴校正后的解释量。例如，已知CCA1的解释变差占约束轴总解释变差的0.3910，由于校正后的R2为0.33739，那么可推断校正R2后的CCA1解释率为“0.3910×0.33739=0.1319”；同理，校正R2后CCA2解释率“0.2517×0.33739=0.0849”；其余约束轴的解释量以此类推。

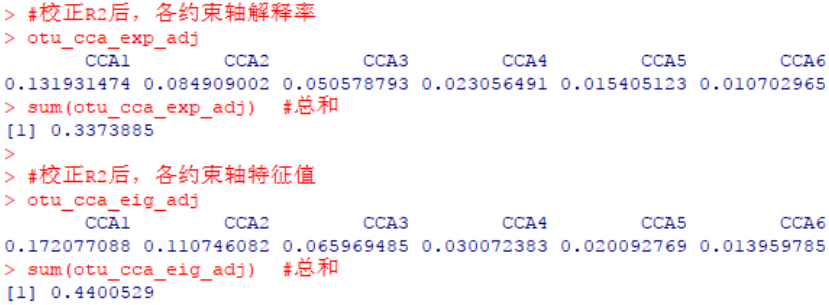


#概念如上所述，如下为代码实现

#关于约束轴承载的特征值或解释率，应当在 R2 校正后重新计算

otu\_cca\_exp\_adj <- otu\_cca\_adj \* otu\_cca$CCA$eig/sum(otu\_cca$CCA$eig)

otu\_cca\_eig\_adj <- otu\_cca\_exp\_adj \* otu\_cca$tot.chi



**约束轴的置换检验及p值校正**

我们看到本示例的约束轴解释率还算可以（以前两轴为例，解释量0.1319+0.0849=0.2168，不算很低），那么，这些约束轴是否有效呢？

对于置换检验，首先检验全模型，通过后再检验各约束轴。

#置换检验

#所有约束轴的置换检验，即全局检验，基于 999 次置换，详情 ?anova.cca

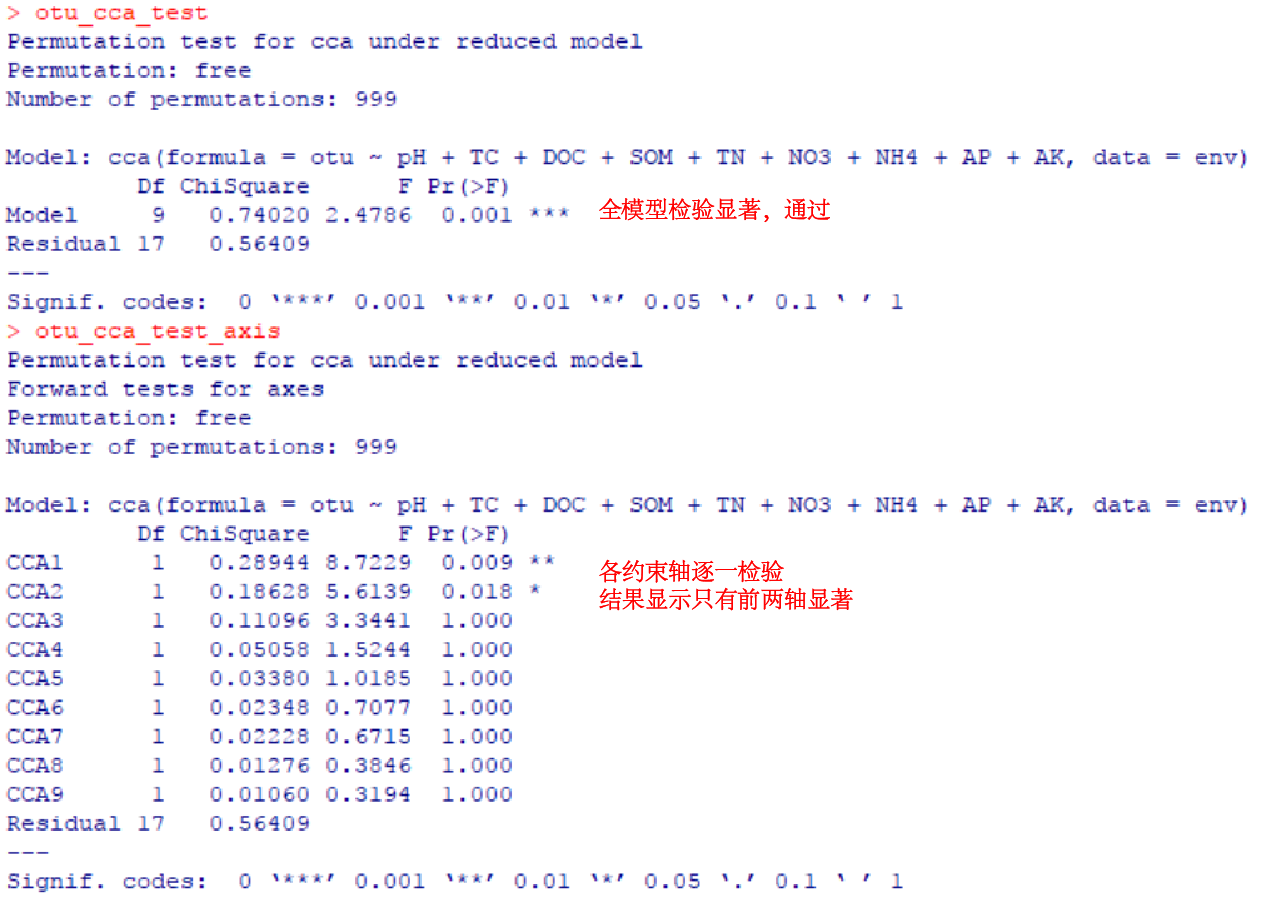
otu\_cca\_test <- anova.cca(otu\_cca, permutations = 999)

#各约束轴逐一检验，基于 999 次置换

otu\_cca\_test\_axis <- anova.cca(otu\_cca, by = 'axis', permutations = 999)

#p 值校正（Bonferroni 为例）

otu\_cca\_test\_axis$`Pr(>F)` <- p.adjust(otu\_cca\_test\_axis$`Pr(>F)`, method = 'bonferroni')



本示例中仅前两轴显著，也就是说，第三及之后的约束轴理论上不能再用作分析了。

注：如果检验未通过，在考虑更换其它方法之前，不妨先试试剔除一些环境变量，因为也有可能是某些环境变量与群落特征之间缺乏较好的线性关系所致。

## CCA的变量选择（前向选择为例）

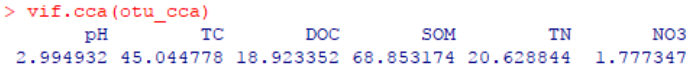
有时解释变量太多，需要想办法减少解释变量的数量。减少解释变量的数量可能有两个并不冲突的原因：第一是寻求简约的模型，利于我们对模型的解读；第二是当解释变量过多时会导致模型混乱，例如有些解释变量之间可能存在较强的线性相关，即共线性问题，可能会造成回归系数不稳定。

下文以前向选择为例展示CCA的解释变量选择过程，前向选择概念可[参考前文](http://blog.sciencenet.cn/blog-3406804-1182487.html)。

方差膨胀因子（Variance Inflation Factor，VIF）显示环境变量间共线性很严重。有几个变量的VIF值特别大（超过10甚至20），所以有必要剔除一些变量。

#计算方差膨胀因子，详情 ?vif.cca

vif.cca(otu\_cca)



vegan中前向选择过程可通过ordistep()、ordiR2step()等函数完成，它们所依据的原理是不同的。详情可见[前文RDA](http://blog.sciencenet.cn/blog-3406804-1182489.html)。

#前向选择，这里以 ordiR2step() 的方法为例，基于 999 次置换检验，详情 ?ordiR2step

otu\_cca\_forward\_pr <- ordiR2step(cca(otu~1, env), scope = formula(otu\_cca), R2scope = TRUE, direction = 'forward', permutations = 999)

#以 otu\_cca 和 otu\_cca\_forward\_pr 为例，简要绘制双序图比较变量选择前后结果

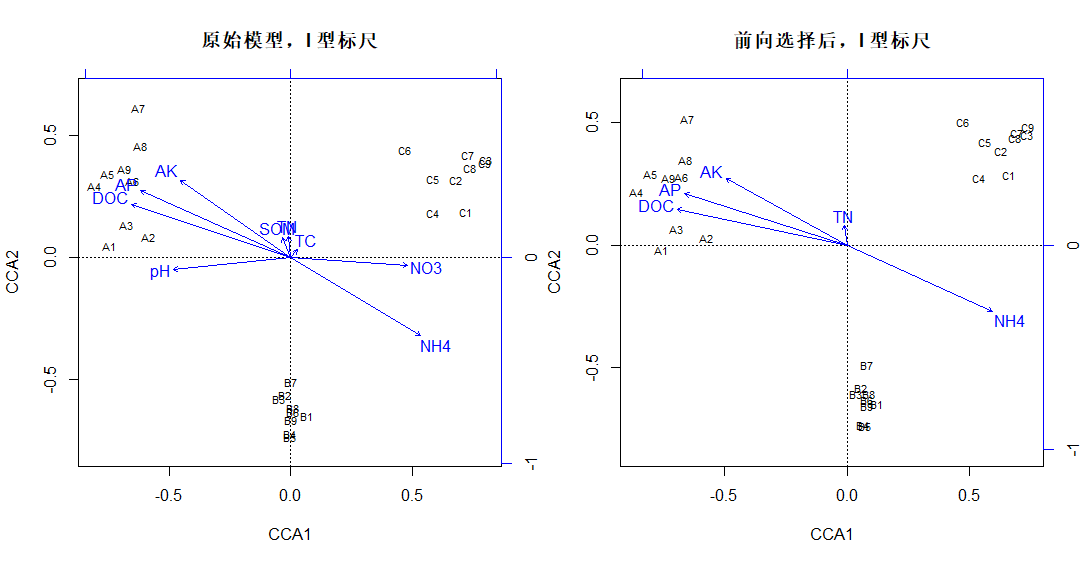
par(mfrow = c(1, 2))

plot(otu\_cca, scaling = 1, main = '原始模型，I 型标尺', display = c('wa', 'cn'))

plot(otu\_cca\_forward\_pr, scaling = 1, main = '前向选择后，I 型标尺', display = c('wa', 'cn'))

#细节部分查看

summary(otu\_cca\_forward\_pr, scaling = 1)



#比较选择前后校正后 R2 的差异，详情 ?RsquareAdj

#可以看到变量选择后，尽管去除了很多环境变量，但总 R2 并未损失很多

RsquareAdj(otu\_cca)$adj.r.squared

RsquareAdj(otu\_cca\_forward\_pr)$adj.r.squared

#所有约束轴的全局检验，999 次置换，详情 ?anova.cca

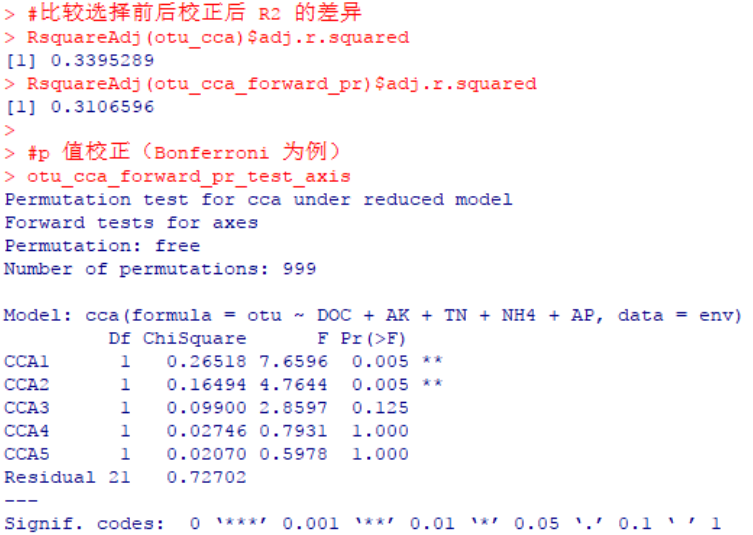
otu\_cca\_forward\_pr\_test <- anova.cca(otu\_cca\_forward\_pr, permutations = 999)

#各约束轴逐一检验，999 次置换

otu\_cca\_forward\_pr\_test\_axis <- anova.cca(otu\_cca\_forward\_pr, by = 'axis', permutations = 999)

#p 值校正（Bonferroni 为例）

otu\_cca\_forward\_pr\_test\_axis$`Pr(>F)` <- p.adjust(otu\_cca\_forward\_pr\_test\_axis$`Pr(>F)`, method = 'bonferroni')



结果显示前两轴显著，我们考虑将简约CCA模型的前两轴坐标提取出，以便用于后续分析或可视化。

##提取或输出变量选择后的排序坐标

#提取方式和上文一致，这里通过 scores() 提取，以 I 型标尺为例，前两轴为例

#使用物种加权和计算的样方得分

otu\_cca\_forward\_pr\_site.scaling1 <- scores(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'wa')

write.table(data.frame(otu\_cca\_forward\_pr\_site.scaling1), 'otu\_cca\_forward\_pr\_site.scaling1.txt', sep = '\t', col.names = NA, quote = FALSE)

#物种变量（响应变量）得分

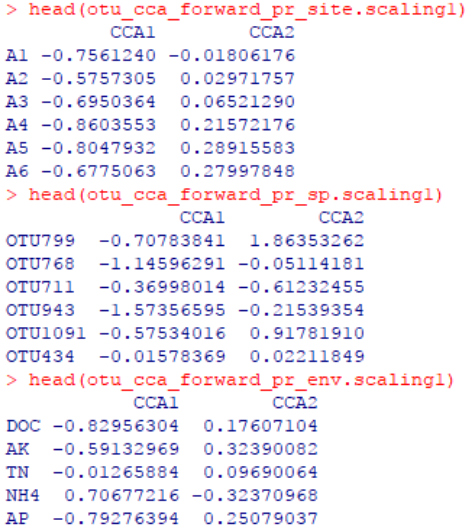
otu\_cca\_forward\_pr\_sp.scaling1 <- scores(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'sp')

write.table(data.frame(otu\_cca\_forward\_pr\_sp.scaling1), 'otu\_cca\_forward\_pr\_sp.scaling1.txt', sep = '\t', col.names = NA, quote = FALSE)

#环境变量（解释变量）得分

otu\_cca\_forward\_pr\_env.scaling1 <- scores(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'bp')

write.table(data.frame(otu\_cca\_forward\_pr\_env.scaling1), 'otu\_cca\_forward\_pr\_env.scaling1.txt', sep = '\t', col.names = NA, quote = FALSE)



此外，[如前所述](http://blog.sciencenet.cn/blog-3406804-1182487.html)，尽管变量选择策略具有明显的优势，但一定切记不应盲目依赖自动选择程序在回归模型中选择相关的环境变量，即仅通过统计学手段得到的最优变量集合可能并非具有很好的生物学意义。

## CCA的变差分解

属于不同类别的两组或两组以上解释变量共同解释一组响应变量的现象非常常见，当我们对它们中的每一个所单独能解释的变差（或者两组或多组变量共同解释的变差）产生兴趣时，就会用到变差分解（variation partitioning）。

vegan中，CCA的变差分解通过varpart()执行。早期版本的vegan无法对CCA执行变差分解，但是现在的版本支持了。

#以两组环境变量（DOC、AP+AK）为例，运行变差分解，详情 ?varpart

#参数 chisquare = TRUE，执行 CCA 的变差分解；chisquare = FALSE，执行 RDA 的变差分解

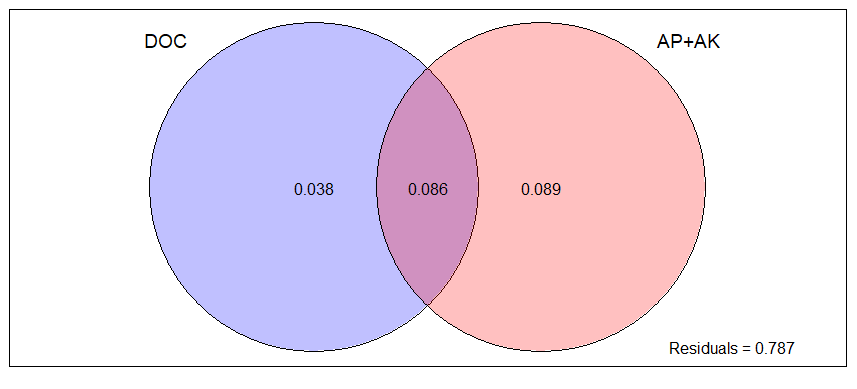
#注：如果 varpart() 不支持 CCA，请更新 R 版本（如 R3.6 的 vegan）

otu\_cca\_vp <- varpart(otu, env['DOC'], env[c('AP', 'AK')], chisquare = TRUE)

otu\_cca\_vp

plot(otu\_cca\_vp, digits = 2, Xnames = c('DOC', 'AP+AK'), bg = c('blue', 'red'))

对于该示例，DOC解释了总变差的12.4%，其中单独解释的比例为3.8%；AP+AK单独解释了8.9%，和DOC共同解释了8.6%；DOC+AP+AK未能解释的变差占比78.7%。

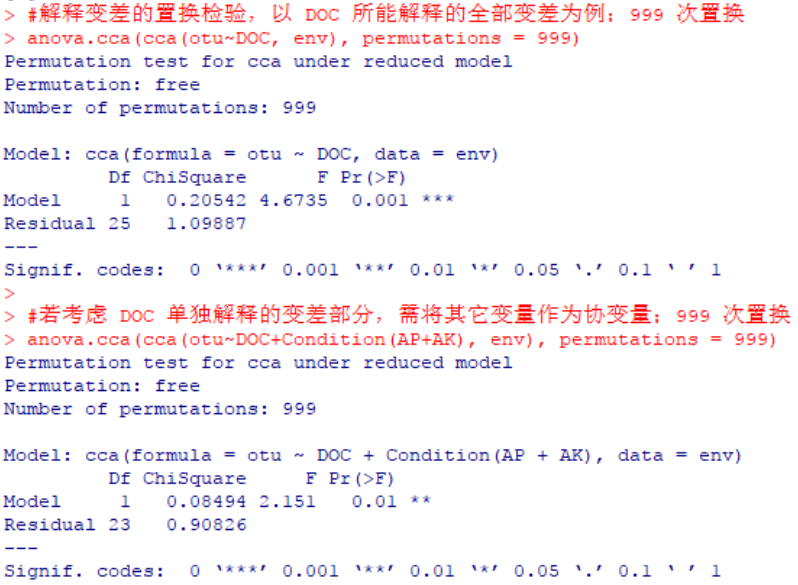


#解释变差的置换检验，以 DOC 所能解释的全部变差为例；999 次置换

anova.cca(cca(otu~DOC, env), permutations = 999)

#若考虑 DOC 单独解释的变差部分，需将其它变量作为协变量；999 次置换

anova.cca(cca(otu~DOC+Condition(AP+AK), env), permutations = 999)



## 作图示例

上文已经展示了通过内置的plot()函数，绘制简单的排序图用于观测数据。现在添加些参数在里面，让它美观一些。

#以上述前向选择后的简约模型 otu\_cca\_forward\_pr 为例作图展示前两轴

#计算校正 R2 后的约束轴解释率

exp\_adj <- RsquareAdj(otu\_cca\_forward\_pr)$adj.r.squared \* otu\_cca\_forward\_pr$CCA$eig/sum(otu\_cca\_forward\_pr$CCA$eig)

cca1\_exp <- paste('CCA1:', round(exp\_adj[1]\*100, 2), '%')

cca2\_exp <- paste('CCA2:', round(exp\_adj[2]\*100, 2), '%')

#plot() 作图，详情 ?plot.cca

#样方展示为点，物种暂且展示为“+”，环境变量为向量

par(mfrow = c(1, 2))

plot(otu\_cca\_forward\_pr, type = 'n', display = c('wa', 'cn'), choices = 1:2, scaling = 1, main = 'I型标尺，双序图', xlab = cca1\_exp, ylab = cca2\_exp)

points(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'wa', pch = 19, col = c(rep('red', 9), rep('orange', 9), rep('green3', 9)), cex = 1)

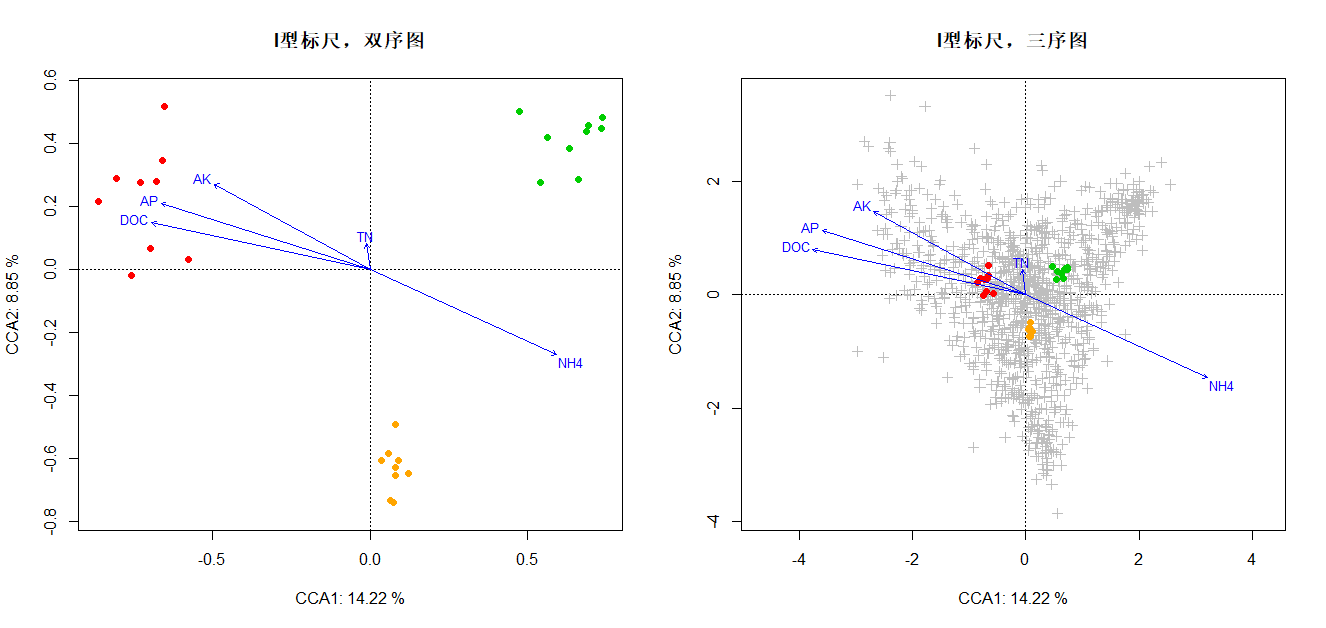
text(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'cn', col = 'blue', cex = 0.8)

plot(otu\_cca\_forward\_pr, type = 'n', display = c('wa', 'sp', 'cn'), choices = 1:2, scaling = 1, main = 'I型标尺，三序图', xlab = cca1\_exp, ylab = cca2\_exp)

points(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'sp', pch = 3, col = 'gray', cex = 1)

points(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'wa', pch = 19, col = c(rep('red', 9), rep('orange', 9), rep('green3', 9)), cex = 1)

text(otu\_cca\_forward\_pr, choices = 1:2, scaling = 1, display = 'cn', col = 'blue', cex = 0.8)



或者可在导出坐标后，通过其它作图包（如ggplot2）绘制。

#提取样方和环境因子排序坐标，前两轴，I 型标尺

otu\_cca\_forward\_pr.scaling1 <- summary(otu\_cca\_forward\_pr, scaling = 1)

otu\_cca\_forward\_pr.site <- data.frame(otu\_cca\_forward\_pr.scaling1$sites)[1:2]

otu\_cca\_forward\_pr.env <- data.frame(otu\_cca\_forward\_pr.scaling1$biplot)[1:2]

#手动添加分组

otu\_cca\_forward\_pr.env$name <- rownames(otu\_cca\_forward\_pr.env)

otu\_cca\_forward\_pr.site$name <- rownames(otu\_cca\_forward\_pr.site)

otu\_cca\_forward\_pr.site$group <- c(rep('A', 9), rep('B', 9), rep('C', 9))

#ggplot2 作图

library(ggplot2)

library(ggrepel) #用于 geom\_label\_repel() 添加标签

p <- ggplot(otu\_cca\_forward\_pr.site, aes(CCA1, CCA2)) +

geom\_point(aes(color = group)) +

stat\_ellipse(aes(color = group), level = 0.95, show.legend = FALSE, linetype = 2) +

scale\_color\_manual(values = c('red', 'orange', 'green3')) +

theme(panel.grid.major = element\_line(color = 'gray', size = 0.1), panel.background = element\_rect(color = 'black', fill = 'transparent'),

legend.title = element\_blank(), legend.key = element\_rect(fill = 'transparent'), plot.title = element\_text(hjust = 0.5)) +

labs(x = cca1\_exp, y = cca2\_exp, title = 'I型标尺，双序图') +

geom\_vline(xintercept = 0, color = 'gray', size = 0.5) +

geom\_hline(yintercept = 0, color = 'gray', size = 0.5) +

geom\_segment(data = otu\_cca\_forward\_pr.env, aes(x = 0, y = 0, xend = CCA1, yend = CCA2), arrow = arrow(length = unit(0.2, 'cm')), size = 0.3, color = 'blue') +

geom\_text(data = otu\_cca\_forward\_pr.env, aes(CCA1 \* 1.2, CCA2 \* 1.2, label = name), color = 'blue', size = 3) +

geom\_label\_repel(aes(label = name, color = group), size = 3, box.padding = unit(0.5, 'lines'), show.legend = FALSE)

p

