---

title: "Beta多样性限制性排序"

author: "文涛 南京农业大学"

date: "`r Sys.Date()`"

output:

word\_document:

toc: yes

toc\_depth: 5

---

```{r setup, include=F, message=F}

knitr::opts\_chunk$set(

echo=T, comment=NA, message=F, warning=F,

fig.align="center", fig.width=5, fig.height=3, dpi=300)

```

### Beta多样性限制性排序CCA/RDA/CPCoA/LDA

> 本节作者：文涛 南京农业大学

>

> 版本1.0.1，更新日期：2020年7月13日

> 本项目永久地址： https://github.com/YongxinLiu/MicrobiomeStatPlot ，本节目录 222CPCoA，包含R markdown(\\*.Rmd)、Word(\\*.docx)文档、测试数据和结果图表，欢迎广大同行帮忙审核校对、并提修改意见。提交反馈的三种方式：1. 公众号文章下方留言；2. 下载Word文档使用审阅模式修改和批注后，发送至微信(meta-genomics)或邮件(metagenome@126.com)；3. 在Github中的Rmd文档直接修改并提交Issue。审稿人请在创作者登记表 https://www.kdocs.cn/l/c7CGfv9Xc 中记录个人信息、时间和贡献，以免专著发表时遗漏。

#### 基本概念

在本章节中以下概念可以通用：

限制性=约束性=有监督=典范

非限制性=非约束性=无监督

\*\*限制性排序\*\*，又称为约束性排序，不同于非约束排序。因为约束性排序从一开始就考虑到了解释变量，并进行运算，然后提取出来和解释变量相关的数据结构。其实非限制性排序也可以加入解释变量，只是在计算之后进行的被动添加，也就是说非限制性排序并不受外部变量的影响。

从表达的意义上来讲约束性排序可以探索解释变量矩阵和响应变量矩阵之间的关系，而非限制性排序只是一种描述响应变量矩阵的，探索性的分析方法。

限制性排序是两个矩阵之间关系的寻找，这就牵扯到不同的限制性排序方法。如果仅仅关注于微生物生态学领域，经常用的有两种，分别是冗余分析（RDA）和典范对应分析（CCA）。这两种方法作为典型的生态方向的应用都是使用多元回归寻找解释变量矩阵和响应变量矩阵之间的关系，分别结合PCA排序或者CA排序展示结果。两个矩阵之间的关系往往可以做显著性分析，这里经常使用置换检验。

这里额外提及LDA和CPCoA，它们也是约束性排序，只是将连续型变量替换为分类型变量(分组)，目的是寻找能够解释已知分组对象的定量解释变量的组合，也就是分组可以解释响应变量矩阵的程度。

##### 冗余分析(RDA）

主成分分析（Principal Component Analysis，PCA）结合多元回归分析，即为冗余分析（Redundancy analysis，RDA）。作为限制性排序的代表方法之一，在微生物生态学领域经常用于寻找环境变量和微生物群落之间的关系等。首先计算解释变量同微生物群落之间的多元多重线性回归，其次使用回归拟合值矩阵进行PCA排序。如果回归拟合值矩阵使用对应分析（Correspondence Analysis，CA）排序拟合，那么就是典范对应分析（canonical correspondence analysis，CCA）。

如果有解释变量作为协变量，从此衍生出来偏RDA分析，表示在控制协变量后解释变量可以解释的响应变量矩阵。

从上一节《[221.Beta多样性PCoA和NMDS排序](https://mp.weixin.qq.com/s/roqJmna0ihVhskCs5vElEw)》我们也逐渐认识到，随着生态学的发展，以欧氏距离为代表的传统分析方法并不能很好解决微生物组数据，由此衍生出来的Bray-Curtis distance等诸多适合于生态学领域的距离。这些距离是没有办法使用RDA分析的，随之而来就有了基于距离的冗余分析（Distance-based redundancy analysis，db-RDA）。当然不仅仅是生态学的距离，任何距离都可以使用db-RDA。随着研究进一步发展，我们了解到微生物和解释变量之间会有大量的非线性关系，比如单峰响应，此时可以使用非线性关系的CCA分析。

无论是哪种方法，发展到如今，都可以通过R包Vegan进行很好的处理生态学数据，下面我们就一个公开发表文章中的案例进行结果解读，而后进行简单的示范性演示。

RDA分析结果得到排序，总结了响应矩阵中主要的变化模式，可以通过解释变量矩阵来解释。可以用于讨论选择适当的可视化尺度并解释此排序。在RDA的结果中经常会显示有关多个约束轴（RDA轴）和非约束轴（PCA轴）的信息。每个RDA轴都有一个与之关联的特征值。由于总方差等于所有特征值的总和（受约束的），因此每个轴解释的方差比例只是给定特征值与解的总方差的商。有时，残差的排序和/或残差之间的相关性可能比特征明确的残差更具有生态学意义。通过排序和相关性检查RDA解决方案的非规范（不受约束）向量，可以部分解释残差。或者，可以在对一组响应变量执行多元线性回归（multiple linear regression，MLR）之后，对残差矩阵执行PCA分析，探索生态学意义。 RDA可以将PCA轴与RDA轴一起显示。 PCA轴汇总了不受约束的（残差）方差。

##### 阅读RDA排序的双序图和三序图

RDA结果可以表示为双序图或三序图（图1）。 这些图的解释取决于所选择的标尺。 通常，如果对象之间的距离具有特定值，或者大多数解释性变量是逻辑值，则考虑I型标尺（type I scaling）。 如果对变量之间的相关关系更感兴趣，请考虑使用II型标尺（ type II scaling ）。 进一步的解释在下面讨论。

![](./priciple\_1.jpg)

![](./priciple\_2.jpg)

图1. RDA双序图(a)和RDA三序图(b)的示意图。

a）RDA双序图将对象作为点，将响应或解释变量作为向量（红色箭头）。 标称变量(Levels of nominal variables)的级别绘制为点（红色）。

b）在一个三序图中，对象被标为点（蓝色），而响应变量和解释变量（分别为红色和绿色箭头）被绘制为矢量。物种绘制为点（绿色）。

- I型缩放-距离图（以对象为中心）

目标点之间的距离近似于欧几里得距离。因此，可以期望将靠近在一起的对象具有相似的变量值。这并非总是如此，因为RDA仅展示了数据集中的部分变化。

代表响应变量的矢量之间的角度是没有意义的。

代表响应变量的点和代表解释变量的矢量之间的角度反映了它们的（线性）相关性。

II型缩放-相关图（关注响应变量）

对象点之间的距离不是近似欧几里得距离。

对象在表示响应变量的点上的直角投影近似于给定对象的变量值。

所有矢量之间的角度反映了它们的（线性）相关性。相关性等于矢量之间的角度的余弦值

![演示图1](./priciple\_3.jpg)

![演示图2](./priciple\_4.jpg)

图2. 坐标对象在矢量上的投影(a)和矢量之间的角度(b)的示意图。

坐标点在变量向量上的投影（如面板a中的点i所示）近似于为该对象实现的变量值。因此，目测，与大多数其他对象相比，可以预期对象i受变量1影响较大。但是，可以预期对象ii相对于其他对象受的变量1影响较小。请注意，虚线通常不会在双序图中显示，此处为演示而显示。使用II型尺度缩放时，向量之间的夹角余弦近似于它们表示的变量之间的相关性。在这种情况下，∠a接近90，这表明变量“ 1”和“ 2”显示出很小的相关性（即它们几乎是正交的，就像独立的轴一样）。 ∠b小于90°，  
  
表明变量“ 2”和“ 3”之间呈正相关，而∠c接近180°，表明变量“ 2”和“ 4”之间具有强烈的负相关（即变量“增加”的方向） 2”和“ 4”彼此相对）。变量5是非数值解释变量，由点表示。在变量4上的直角投影表明两者是正相关的。

##### 典范对应分析(CCA）

同RDA有许多相似之处，作为非限制性排序典型算法的CCA，是基于单峰模型，同CA分析一样的基础算法，但在排序迭代过程中将解释变量同响应变量矩阵之间做了加权回归分析。这里重点关注CCA使用的条件：1. 物种在整个生态梯度环境中都有采样；2. 物种在生态梯度上呈现单峰分布响应。

##### 限制性主坐标分析(CPCoA）

##### 线性判别分析(LDA）

#### 实例解读

#### 示例一. 以取材部位和基因型为条件的主坐标轴分析(CPCoA/CCA)

[Zgadzaj, R., et.al., 2016 .PNAS ](http://www.pnas.org/content/113/49/E7996.short)

这篇文章分析了百脉根根瘤的微生物组成，同时在根瘤缺失突变体条件下发现根和根际微生物均有较大差异的变化。

![image](http://bailab.genetics.ac.cn/markdown/scatterplot2.png)

图2. 散点图展示限制性主坐标轴分析(Constrained PCoA/ CCA)取材部位和基因型间的差异。

1. (A) 采用CCA方法结合Bray-Curtis距离，分析以取样部分(compartment)条件下可显示各组最大差异投影平面；图顶部19.97% of variance (P<0.001)表示当前所展示的平面坐标系，可解释所有样品间总差异的19.97%的，并且各组间存在显著差异(P<0.001)；

2. (B) 以基因型为条件分析最大解释基因型组间差异的空间平面，可解释9.82%的变异，并且有显著差异，其中作者按形状标出了各基因型；同时作者还按compartment进行着色，在这一平面上，compartment仍能很好的分开。

3. 图表结果：Compartment可解释19.97%差异，且区分明显；突变体与WT(gifu)可以区分，区分不大(占9.82%变异中的17.75%的纵轴上可区分)；各突变体间很难区分，完全混在一起；在基因型最大解释平面上，compartment仍能非常好的在第一轴上区分。

> Fig. 2. (A) Constrained PCoA plot of Bray–Curtis distances between samples including only the WT constrained by compartment (19.97% of variance, P > 0.001; n = 94). (B) Constrained PCoA plot of Bray–Curtis distances constrained by genotype (9.82% of variance explained, P < 0.001; n = 164). Each point corresponds to a different sample colored by compartment, and each host genotype is represented by a different shape. The percentage of variation indicated in each axis corresponds to the fraction of the total variance explained by the projection.

#### 示例二 Nature Communication：根系代谢物塑造根际微生物群来驱动土壤对植物生长和防御的反馈

这篇文章由Matthias Erb 等人在2018年发表在Nature Communications期刊上。题目为：Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota。揭示了植物根际微生物区系组成、植物性能和植物-草食相互作用的机制。

![示例2](./example\_2.jpg)

图7 BXs对第二代植物根和根际的土壤微生物区系有一定的影响。野生型（WT）植物根际和根际细菌（a）和真菌（b）群落的主坐标的部分正则分析（n=10）。利用细菌操作分类学单元（bOTU, c）和真菌OTU（fOTU, d）的双标图，显示单个OTU对不同处理的分离的贡献。[点击查看中文解读](https://mp.weixin.qq.com/s/d80uZRffhhIsA-gOOxcKmA)

> Fig. 7 Benzoxazinoid-conditioning shapes soil microbiota in the roots and rhizosphere of the second plant generation. a, b Partial Canonical analysis of

Principal Coordinates (CAP) of rhizosphere and root bacterial (a) and fungal (b) communities of wild type (WT) plants grown in soils previously

conditioned by WT (BX+) or benzoxazinoid (BX)-deficient bx1 mutant plants (BX−) (n = 10). CAP ordinations using Bray–Curtis distance were

constrained for the factors sample type and soil conditioning. Model, explained fraction of totasl variance and model significance are indicated above the

plots. Axes report the proportions of total variation explained by the constrained axes. c, d Biplot with the bacterial Operational Taxonomic Unit (bOTU, c)

and fungal OTU (fOTU, d) scores of the CAP showing the contribution of individual OTUs to the separation of the different treatments. For unconstrained

ordination and individual OTUs, refer to Supplementary Fig. 9

\*\*result\*\*：

为了了解BX土壤处理是否改变了响应植物的根相关菌群，研究人员分析了生长在BX+和BX−土壤中野生植物的根和根际样品中的真菌和细菌群落。无限制和受限制的纵坐标显示，细菌和真菌在根际和根之间以及在BX+和BX−土壤之间有明显的分离（图7a, b，补充图9a, b）。多变量统计证实了这些效应的显著性（补充表7和8；参考补充数据3来深入分析BX土壤处理对微生物多样性的影响）。放线菌门OTUs和属于子囊菌纲和球囊菌门的OTUs的亚群对根和根际样品的分离贡献最大（图7c, d）

> To understand whether BX soil conditioning changes the rootassociated

microbiota of the responding plants, we profiled fungal

and bacterial communities in root and rhizosphere samples of

WT plants grown in BX+ and BX− soils. Both unconstrained

and constrained ordinations revealed a clear separation of bacteria

and fungi between rhizosphere and roots and between BX+

and BX− soils (Fig. 7a, b, Supplementary Fig. 9a, b). Multivariate statistics confirmed the significance of these effects (Supplementary

Tables 7 and 8; we refer to the Supplementary Data 3 for in

depth analysis of BX soil conditioning effects on microbial

diversity). Actinobacteria OTUs and a subset of OTUs belonging

to Ascomycota and Glomeromycota contributed most strongly to

the separation of root and rhizosphere samples (Fig. 7c, d).

#### 示例三 GCB | 高强度人类活动背景下微生物对环境变化的抗性和恢复力

本文由Laibin Huang等人于2020年发表在Global Change Biology上，即使在人类活动强度高于全球沿海生态系统平均水平的情况下，分析了在大尺度生态系统解除胁迫因子后，微生物群落整体也能在短时间内恢复到与原始状态相似的状态，这意味着即使在强烈的人为干扰下，土壤微生物群落也具有很强的恢复能力。[点击访问中文解读](https://mp.weixin.qq.com/s/\_w07hBrDtCzh6Vnj6QpPgg)

![示例3](example\_3.png)

图6 RDA分析对珠江三角洲沿岸微生物群落变化的环境解释。（a） 前20门变化的环境解释；（b）与硝化过程有关的属变化的环境解释。USF\_2014（粉色圆圈）和USF\_2015（蓝色圆圈）分别代表2014年和2015年上游的原位土壤；DSF\_（粉色钻石）和DSF\_（蓝色钻石）分别代表2014年和2015年下游的原位土壤。（无限制置换试验，n=999）

> FIGURE7 Environmental explanation of the changes in microbial community along the Pearl River Estuary by RDA analysis. (a)

Environmental explanation of the changes in top 20 phylum; (b) environmental explanation of the changes in genus related to nitrification

process. USF\_2014 (pink circle) and USF\_2015 (blue circle) represent field soils from upstream in 2014 and 2015, respectively; DSF\_2014

(pink diamond) and DSF\_2015 (blue diamond) represent field soils from downstream in 2014 and 2015, respectively. \*.05; \*\*.01; \*\*\*<.001

(Unrestricted permutation test, n = 999)

\*\*result\*\*：

在RDA分析中，2014年采集的土壤样品中20个最高门中的16个相对丰度较高，而2015年采集的土壤样品中只有3个相对较高，这与之前的结果一致。

如图7b所示，第一轴解释方差为82.1％（p = .001；置换检验）和

第二轴解释方差为0.9％（p = .992;置换检验）。RDA的第一轴和第二轴分别解释了这种关系，这表明第一轴基于硝化微生物群落将所有环境变量和土壤样品显着分离。

> in the RDA analysis, the relative abundance of 16 out of the 20 top phyla were higher in the soil samples collected in 2014, whereas only three of these phyla were higher in the soil samples collected in 2015, which was consistent with the previous results.

As shown in Figure 7b, 82.1% (p = .001; permutation test) and

0.9% (p = .992; permutation test) of the cumulative variance of the

relationship were explained by the first and second axes of RDA, respectively, which indicated that the first axis significantly separated all of the environmental variables and soil samples based on nitrifying microbial communities.

#### 绘图实战

### 安装专题R包：amplicon

在github上安装即可；

```{R}

# 基于github安装

# library(devtools)

# install\_github("microbiota/amplicon")

#基于bioconductor安装，暂不可用

# library(BiocManager)

# BiocManager::install("amplicon")

```

## 载入程序包

```{R}

suppressWarnings(suppressMessages(library(amplicon)))

```

## 实验设计 data/Metadata

设置基本参数：读取实验设计、设置分组列名、图片宽高

Setting the basic parameter

```{r parameter, warning=FALSE}

# Data reading

metadata = read.table("../data/metadata.tsv", header=T, row.names=1, sep="\t", comment.char="", stringsAsFactors = F)

# colnames of group ID in metadata

# 设置实验分组列名

group = "Group"

# Output figure width and height

# Letter纸图片尺寸为单栏89 mm，双栏183 mm，页面最宽为247 mm

# 推荐比例16：10，即半版89 mm x 56 mm; 183 mm x 114 mm

width = 89

height = 59

#-创建保存文件夹

dir.create("./beta")

# 手动指定分组列和顺序，默认为字母顺序

# metadata[[group]] = factor(metadata[[group]], levels = c("WT","KO","OE"))

# 按实验设计中分组出现顺序

# metadata[[group]] = factor(metadata[[group]], levels = unique(metadata[[group]]))

```

## 有监督PCoA/CCA

Constrained PCoA

要求至少分组数量 >= 3时才可用，否则请跳过此步分析

```{r}

# 设置距离矩阵类似，常用bray\_curtis或unifrac

# Data reading

distance\_mat = read.table("../data/beta/bray\_curtis.txt", header=T, row.names=1, sep="\t", comment.char="")

distance\_mat[1:3, 1:3]

# Plotting Constrained PCoA

(p = beta\_cpcoa\_dis(distance\_mat, metadata, groupID = group))

# Saving figure

ggsave("../data/beta/CPCoA.pdf", p, width = width, height = height, units = "mm")

```

### RDA 或者 CCA

```{R}

source("G:\\Shared\_Folder\\Function\_local\\R\_function\\Liu\_project\\amplicon-master\\R\\rda-cca.R")

```

#### 导入数据

```{R}

#导入otu表格

otu = read.delim("../data/otutab.txt",row.names = 1)

#导入注释文件

tax = read.delim("../data/taxonomy.txt",row.names = 1)

#导入分组文件

map = read.delim("../data/metadata.tsv",row.names = 1)

head(map)

# 导入环境因子文件

env = read.delim("../data/env.txt",row.names = 1)

# head(map)

```

#### 运行RDA\_CCA

RDA 或CCA 模型的选择原则:RDA是基于线性模型，CCA是基于单峰模型。一般会选择CCA来做直接梯度分析。首先函数先做DCA(detrendedcorrespondence analysis)分析，并提取结果中Lengths ofgradient 的第一轴的大小。如果大于4.0，就应该选CCA;如果3.0-4.0 之间，选RDA 和CCA均可;如果小于3.0，RDA 的结果要好于CCA。

所以这里会有同时做RDA和CCA的情况。

```{R}

result =RDA\_CCA(otu = otu,map = map,ps = NULL,env = env,group = "Group")

#=--提取图

p = result[[1]]

p

```

```{R}

#--提取作图数据

plotdata = result[[2]]

head(plotdata)

```

```{R}

# 提取带有样本标签的图形

p2 = result[[3]]

p2

```

检测解释变量对响应变量的解释度和显著性。

```{R}

# 提取环境因子同群落的差异检测

table = result[[4]]

head(table)

```

Lengths ofgradient 的第一轴的大小在3.0-4.0 之间，会同时计算RDA和CCA，供选择。

```{R}

# 如果两种模型都选择，则列表中5，6就被激活了#

result[[5]]

result[[6]]

```

#### 参考文献

- Zgadzaj, R., Garrido-Oter, R., Jensen, D.B., Koprivova, A., Schulze-Lefert, P. and Radutoiu, S., 2016. Root nodule symbiosis in Lotus japonicus drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. Proceedings of the National Academy of Sciences, 113(49), pp.E7996-E8005.

- Hu, Lingfei, et al. "Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota." Nature Communications 9.1 (2018).

- Huang, Laibin, et al. "Microbial resistance and resilience in response to environmental changes under the higher intensity of human activities than global average level." Global Change Biology 26.4 (2020): 2377-2389.

- Spatial and Temporal Variation in a Caribbean Herbivorous Fish Assemblage

- Diversity and dynamics of rare and of resident bacterial populations in coastal sands

- The energy–diversity relationship of complex bacterial communities in Arctic deep-sea sediments

- GUSTA ME http://mb3is.megx.net/gustame/constrained-analyses/rda

> 责编：刘永鑫 中科院遗传发育所

> 版本更新历史

> 1.0.0，2020/6/27，文涛，初稿

> 1.0.1，2020/7/13，刘永鑫，LDA/CCA/CPCoA背景介绍

> 1.0.1，2020/7/17，席娇，结构调整，文字修改