R包mixOmics进行微生物群落偏最小二乘判别分析（PLS-DA）

在代谢组学分析中经常可以见到主成分分析（PCA）、偏最小二乘判别分析（PLS-DA）等分析手段，目的为区分样本差异，或在海量数据中挖掘潜在标志物。PCA与PLS-DA的不同之处在于，PCA是一种无监督的模式，而偏PLS-DA则是一种有监督的模式。作为延伸，我们可考虑将其应用在生态群落分析中，以区分环境样本差异。此处结合微生物群落研究中的16S扩增子分析数据，给大家分享怎样在R中进行PLS-DA分析。

首先介绍示例数据。我们此处共有30个16S测序样本，均来自土壤。因试验需求，在土壤中添加了某化学物质，目的为探究该化学物质对土壤微生物群落的影响。这30个样本中，10个为不添加化学物质的对照组（control组），10个为添加低浓度化学物质的处理组（treat-low组），10个为添加高浓度化学物质的处理组（treat-higt组）。

此处我们希望通过PLS-DA分析，判断该化学物质是否显著影响了土壤细菌群落的组成。

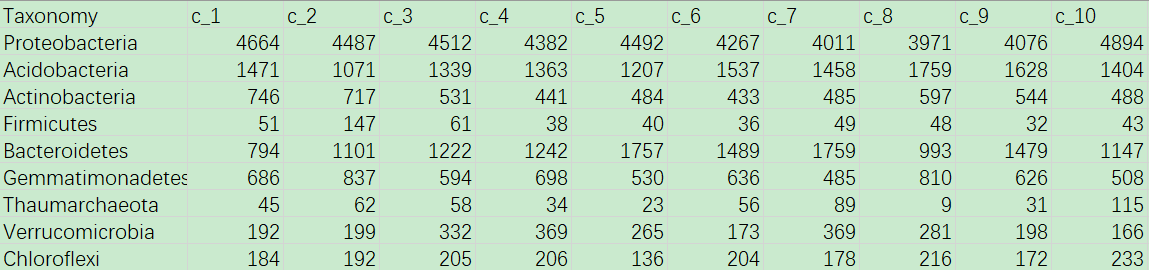
作图示例文件、R脚本等，已上传至百度盘，无提取码

<https://pan.baidu.com/s/1E9ZQbLch0AA4XIOq2PJa2Q>

## 示例文件简要

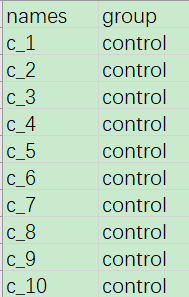
文件“phylum\_table.txt”为16S测序所得的细菌类群丰度表格，此处统计至细菌“门水平”（即界门纲目科属种的“门”这一水平），其内容展示如下。

每一列为一个样本，每一行为一类门，交叉区域为每类门在各样本中的丰度。



文件“group.txt”为样本分组信息，其内容展示如下。

第一列（names）为各样本名称；第二列（group）为各样本的分组信息，即这些样本分别属于未添加化学物质的对照组（control组）或分别添加了低/高化学物质的处理组（treat-low/treat-high组）。



## 使用mixOmics包对细菌群落数据进行偏最小二乘判别分析（PLS-DA）

首先读入细菌群落数据，以及样本分组文件。

##读入文件

#门水平丰度表

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

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

#样本分组文件

group <- read.delim('group.txt', sep = '\t', stringsAsFactors = FALSE)

读入细菌门水平丰度表格，赋值给新数据框“phylum”，并进行转置（计算要求，每一行为一个样本，每一列为细菌门类群信息）。

之后导入mixOmics包，并使用mixOmics包中的plsda()命令进行PLS-DA分析。plsda()的详细参数可使用?plsda查看。

library(mixOmics)

##PLS-DA 分析

#基于门水平丰度表，只展示前 3 个排序轴

phylum <- phylum[group$names, ]

plsda\_result <-plsda(phylum, group$group, ncomp = 3)

phylum，先前读入的细菌门水平丰度表，注意此处需要首先确认丰度表中的样本顺序；group$group ，各样本分组，分组顺序与phylum中的样本顺序需要一一对应；nf =3，在排序结果中显示3个排序轴，若想展示更多的排序轴（如第三轴、第四轴等），可更改此参数。此处示例只展示前三个轴。plsda()的其余参数详见帮助。分析结果赋值给“plsda\_result”。

PLS-DA分析完毕，我们可简要查看结果。

#简要查看结果

plsda\_result

#或

names(plsda\_result)

#查看排序轴解释量

plsda\_result$explained\_variance$X

#查看样本排序坐标

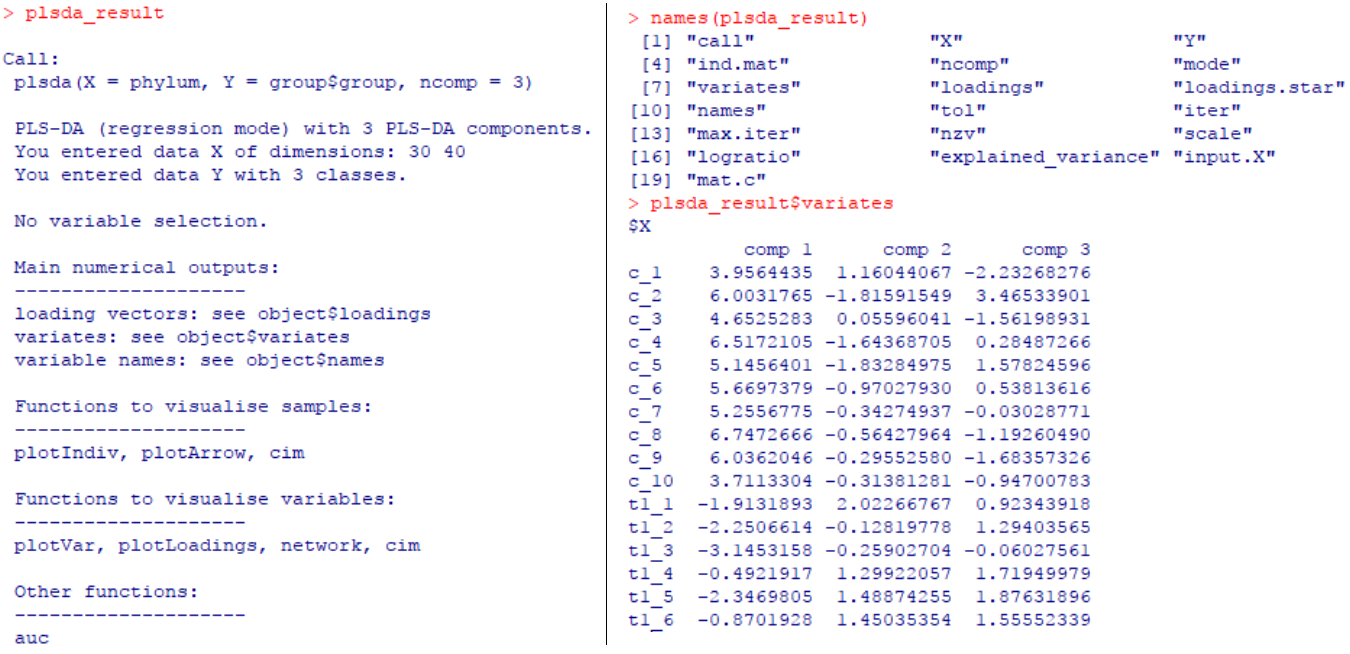
plsda\_result$variates$X

#查看细菌门类群排序坐标

plsda\_result$loadings$X

直接输入排序结果“plsda\_result”回车后，我们可看到排序结果简要，包括了我们的输入命令、主要结果信息的查看方式（例如，“variates: see object$variates”，即我们若输入“plsda$variates”则可直接查看各样本在各排序轴中的坐标）等。

使用names(plsda\_result)之后可看到结果“plsda\_result”中所包含的所有信息类型，共计19项内容，此处不再一一说明，可使用“pca$variates”等命令查看结果中每部分的更详细内容。



使用mixOmics包内置命令plotIndiv()绘制PLS-DA分析结果图。plotIndiv()的详细参数可使用?plotIndiv查看。

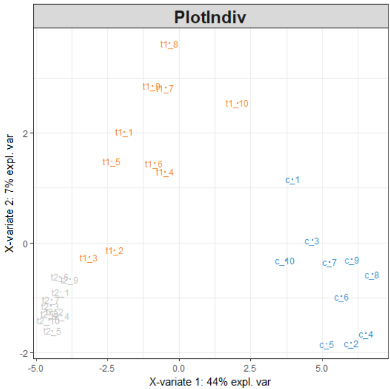
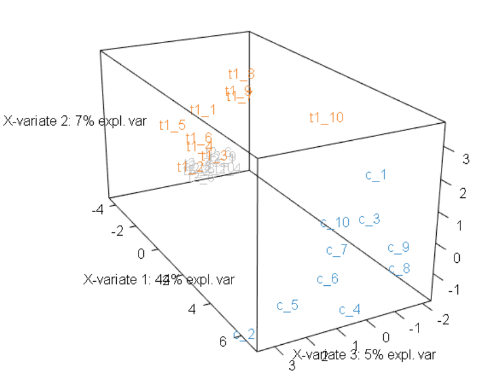
#使用 plotIndiv() 绘制 PLS-DA 分析结果

plotIndiv(plsda\_result, ind.names = TRUE, style = 'ggplot2')

plotIndiv(plsda\_result, ind.names = TRUE, style = '3d')

plotIndiv()是一个非常便捷的命令，可以快速绘制PLS-DA分析结果。plotIndiv()内嵌了ggplot2、lattice、graphics 、3d等做图样式，提供了较为灵活的二维、三维展示图作图风格样式选择。我们这里使用plotIndiv()分别对PLS-DA分析结果中前2、3个最主要的排序轴进行简要展示，未作详细的参数调整，若需要可参见命令帮助。

PLS-DA分析作图结果如下，可以见到3个分组（未添加化学物质的对control组以及分别添加了低/高化学物质的treat-low/treat-high组）的样本在图中区分明显。结合本文一开始所提及的分析目的，可判断在土壤中添加了某化学物质后，会显著影响土壤细菌群落的组成。

若有其它需要，可在PLS-DA分析结果“plsda\_result”中提取相关内容。如下示例所示，提取各样本的排序坐标，并将各样本所对应的分组加入其中。

#提取坐标轴解释量（前两轴）

plsda\_result\_eig <- {plsda\_result$explained\_variance$X}[1:2]

#提取样本点坐标（前两轴）

sample\_site <- data.frame(plsda\_result$variates)[1:2]

#为样本点坐标添加分组信息

sample\_site$names <- rownames(sample\_site)

names(sample\_site)[1:2] <- c('plsda1', 'plsda2')

sample\_site <- merge(sample\_site, group, by = 'names', all.x = TRUE)

#可选输出各样本的 PLS-DA 分析结果

write.table(sample\_site, 'plsda\_sample.txt', row.names = FALSE, sep = '\t', quote = FALSE)

示例输出结果“plsda\_sample.txt”如下所示。包含4列内容，分别为各样本名称、各样本在PLS-DA分析结果第一主轴中的坐标、各样本在PLS-DA分析结果第二主轴中的坐标、各样本的分组信息。



## 使用ggplot2包进行PLS-DA作图

mixOmics包中内置命令plotIndiv()的功能强大，一般情况下即可满足作图需求。

但有时，对于已经习惯了使用常规作图R包（如ggplot2、lattice等）作图的我们来讲，可能还是使用这些作图R包去完成会更加顺手一些。因此在这里继续基于已经计算得到的PLS-DA分析结果，展示一个使用ggplot2进行作图的示例。

ggplot2作图示例命令如下，作图文件为上文中的结果“sample\_site”（包含了各样本的排序坐标及分组信息），绘制散点图，并添加各分组的95%置信椭圆。

library(ggplot2)

#使用 ggplot2 简单绘制 PLS-DA 结果图

plsda\_plot <- ggplot(sample\_site, aes(plsda1, plsda2, color = group, label = names)) +

geom\_point(size = 1.5, alpha = 0.6) +

stat\_ellipse(show.legend = FALSE) + #添加 95% 置信椭圆

scale\_color\_manual(values = c('#1D7ACC', '#F67433', '#00815F')) +

theme(panel.grid = element\_line(color = 'grey50'), panel.background = element\_rect(color = 'black', fill = 'transparent')) +

theme(legend.title = element\_blank(), legend.key = element\_rect(fill = 'transparent')) +

labs(x = paste('PLS-DA axis1 ( explained variance ', round(100 \* plsda\_result\_eig[1], 2), '% )', sep = ''), y = paste('PLS-DA axis2 ( explained variance ', round(100 \* plsda\_result\_eig[2], 2), '% )', sep = ''))

#ggsave('plsda\_plot.pdf', plsda\_plot, width = 6, height = 5)

ggsave('plsda\_plot.png', plsda\_plot, width = 6, height = 5)

PLS-DA作图示例图输出如下。可以见到3个分组（未添加化学物质的对control组以及分别添加了低/高化学物质的treat-low/treat-high组）的样本在图中区分明显。结合本文一开始所提及的分析目的，可判断在土壤中添加了某化学物质后，会显著影响土壤细菌群落的组成。

