Basic Analysis of Simplex output

Ken Locey

October 2015

OVERVEW

This R Markdown document is designed to be opened and ran in the RStudio. The chunks of code below allow for basic analysis of **simplex** output. This includes univariate and multivariate relationships, and graphical exploration.

For an R-based introduction to multivariate analysis: https://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html

SETUP

A. Clear and set the working directory

```
rm(list=ls())
getwd()
setwd("~/GitHub/simplex")
```

B. Import packages; install if needed

```
#install.packages("vegan") # Example of how an install can be done
require("vegan")
require("car")
```

C. Import simulated data generated by simplex models

As it is iterating through randomly assembled models, simplex writes its output to six .csv files. For each file, each row corresponds to a single model. Consequently, the ith row in each file corresponds to ith model that was assembled and run by simplex. Let's import the simulated data files.

```
# A table where each columns corresponds to a state variable or model output.
simplex.dat <- read.csv("~/GitHub/simplex/results/simulated_data/examples/SimData.csv")
simplex.dat$h.tau <- log((simplex.dat$width * simplex.dat$height)/simplex.dat$flow.rate)

# Replacing 0's with 1's to allow log-transforms below
simplex.dat$total.abundance[simplex.dat$total.abundance<=0] <- 1
simplex.dat$N.max[simplex.dat$N.max<=0] <- 1
simplex.dat$species.richness[simplex.dat$species.richness<=0] <- 1
simplex.dat$resource.particles[simplex.dat$resource.particles<=0] <- 1
simplex.dat$resource.concentration[simplex.dat$resource.concentration<=0] <- 1

simplex.dat$total.abundance <- log(simplex.dat$total.abundance)
simplex.dat$N.max <- log(simplex.dat$N.max)
```

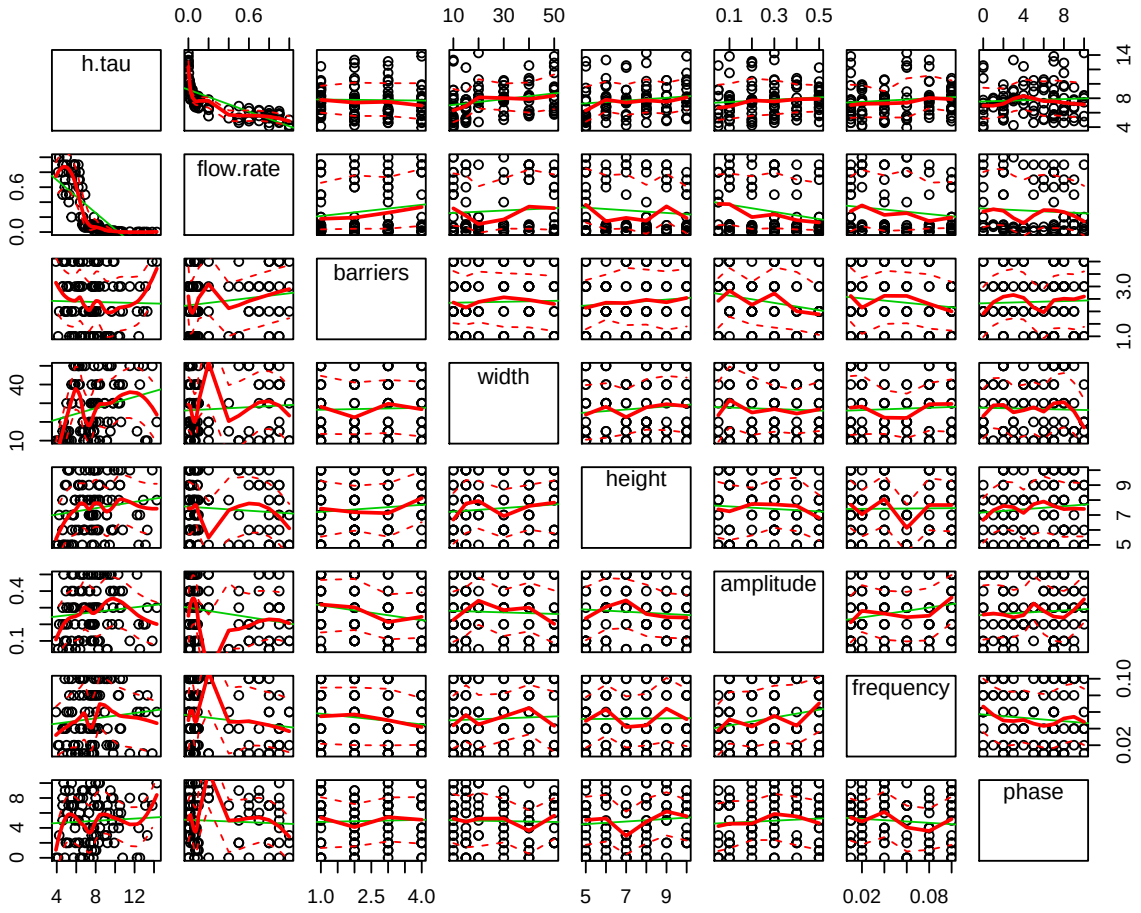
```
simplex.dat$species.richness <- log(simplex.dat$species.richness)
simplex.dat$resource.particles <- log(simplex.dat$resource.particles)
simplex.dat$resource.concentration <- log(simplex.dat$resource.concentration)
simplex.dat[is.na(simplex.dat)] <- 0
```

UNIVARIATE ANALYSES

Perhaps we want to ask whether **simplex** produces (i) relationships that are well-known to occur in ecological systems (ii) auto-correlated relationships between independent variables (iii) or novel relationships of use to a specific study or question. Each of these questions has its particular use, but rather than generate one x-y relationship after another, we generate an entire field of relationships. That is, given x variables, we can explore x*(x-1) x-y relationships.

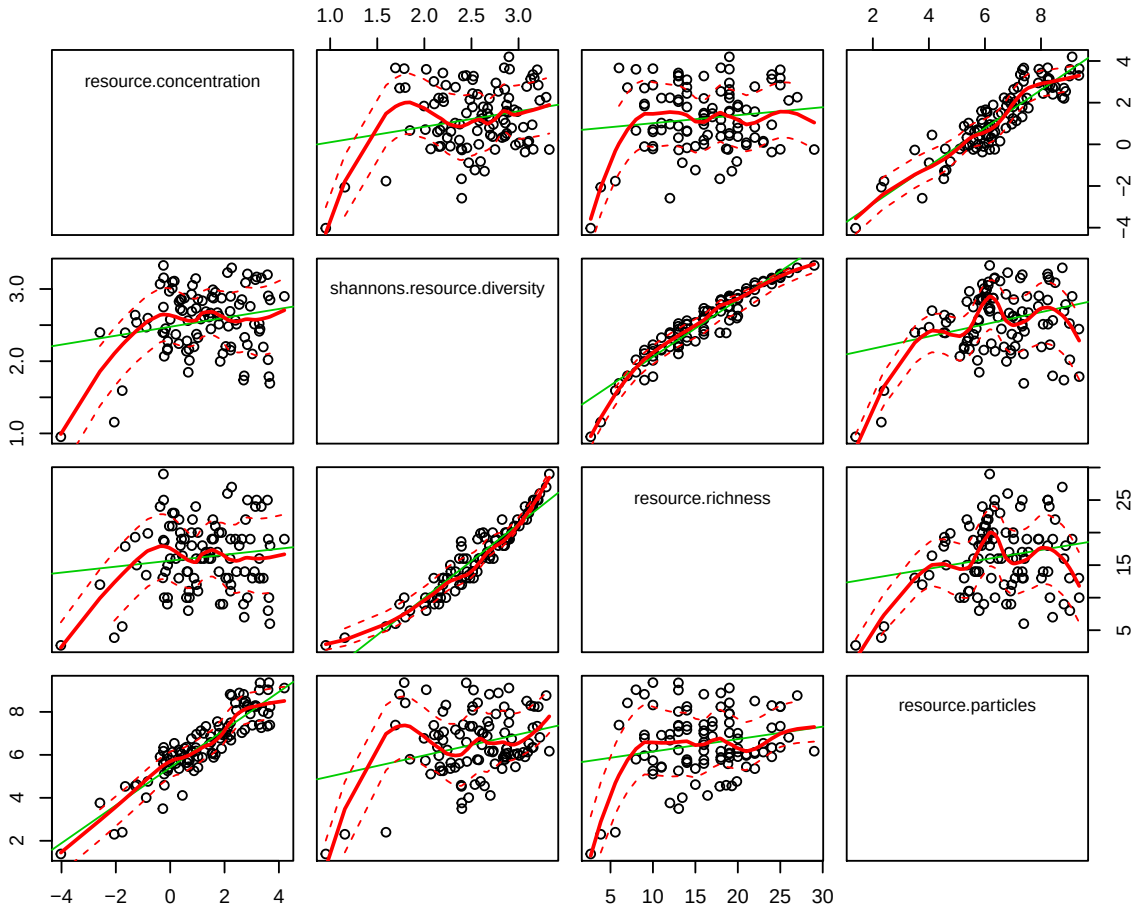
Let's begin by examining relationships between some physical variables. As we'll see, the only physical variables that appear to be correlated are ecosystem residence time and the rate of flow.

```
# Physical and metacommunity variables
phys.dat <- as.matrix(subset(simplex.dat,
                      select = c(h.tau,
                                 flow.rate,
                                 barriers,
                                 width,
                                 height,
                                 amplitude,
                                 frequency,
                                 phase)))

scatterplotMatrix(phys.dat, main="Physical data", diagonal = "none")
```



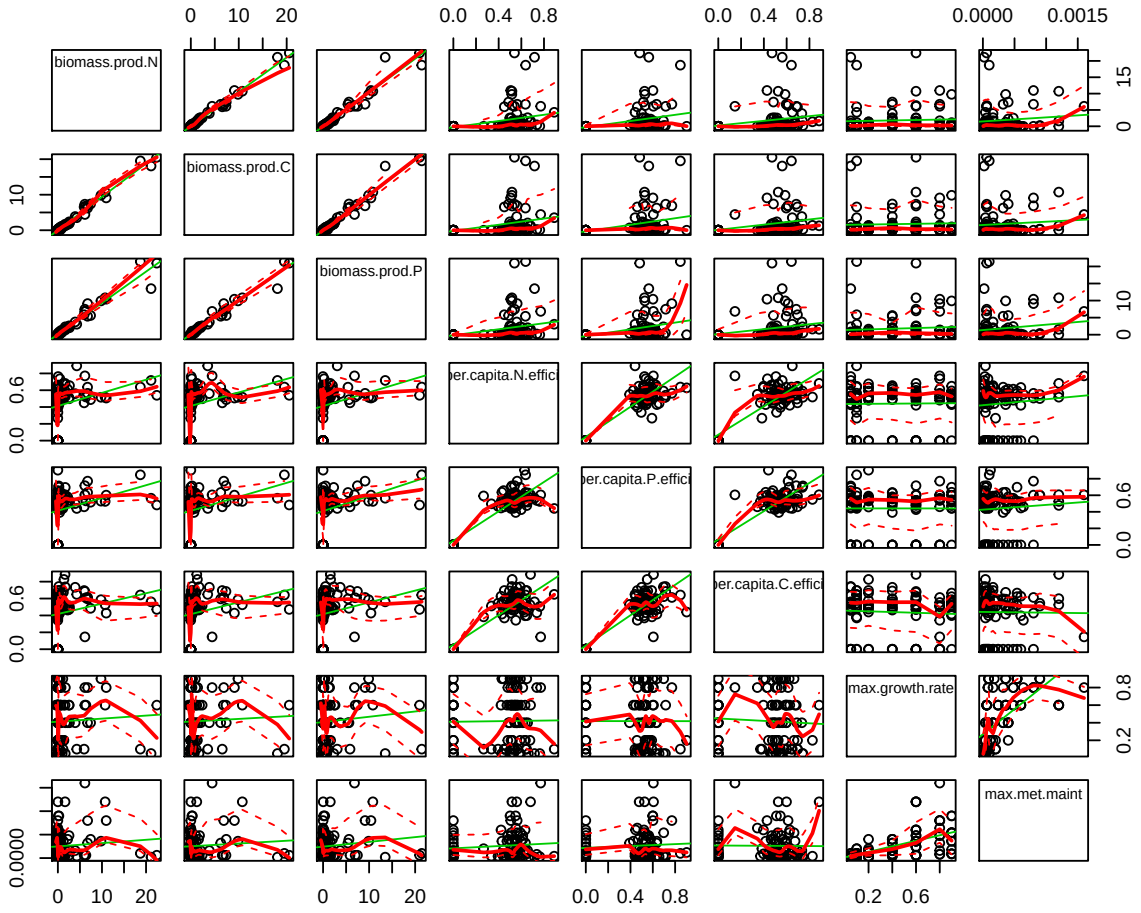
Next, let's explore relationships between some resource-related variables. As we'll see, the number of resource particles, their concentration, resource richness or the number of resource types, and resource diversity (a combination of resource richness and the variance in abundance among resources) are all highly and positively correlated.

```
# Resource variables
res.dat <- as.matrix(subset(simplex.dat,
                      select = c(resource.concentration,
                      shannons.resource.diversity,
                      resource.richness,
                      resource.particles)))

scatterplotMatrix(res.dat, main="Resource data", diagonal = "none")
```



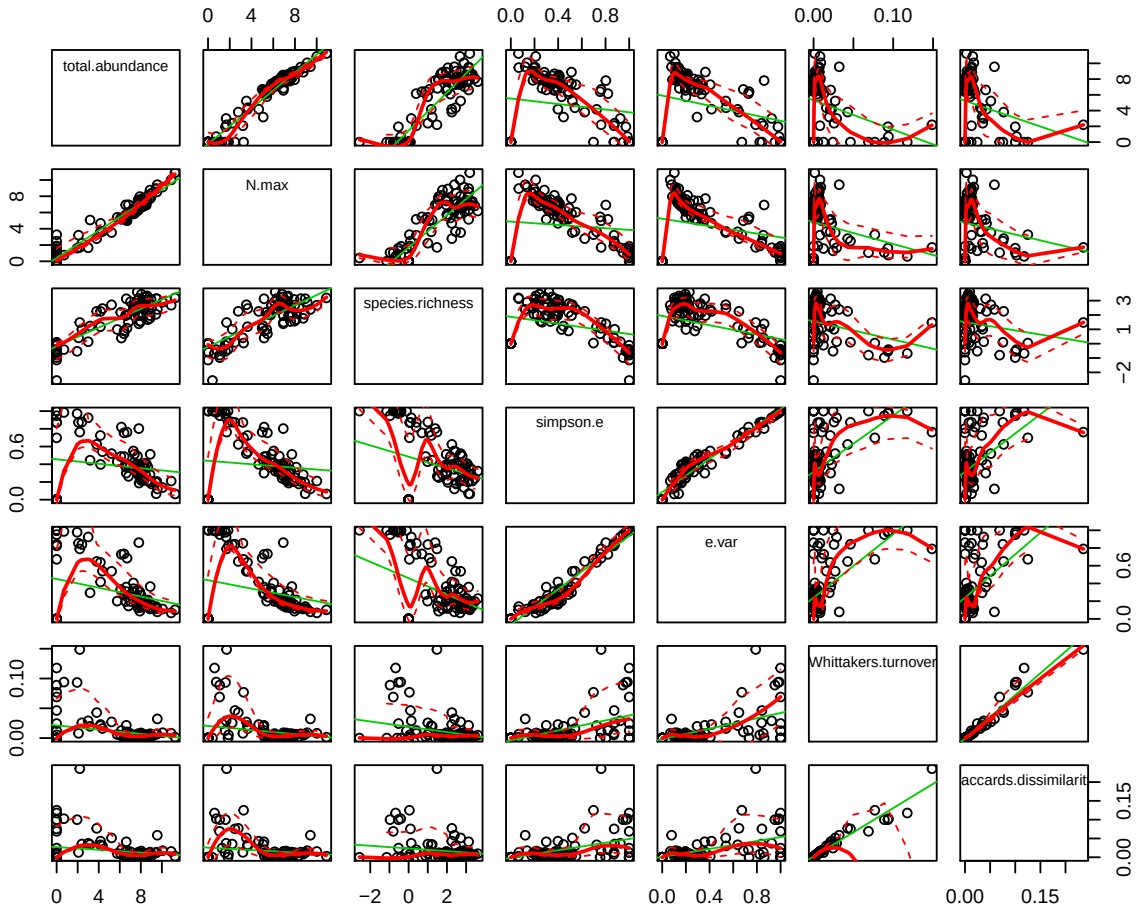
We can also explore relationships between physiological variables, noticing that, biomass production in Carbon, Nitrogen, and Phosphorus are all positively and strongly correlated. This is largely because **simplex** (as of 26 October 2015) does not include any explicit stoichiometry. Also notice that an interesting life history trade-off arises in **simplex** however, that is, increasing growth rate leads to increasing metabolic maintenance.

```
# Physiological variables
physio.dat <- as.matrix(subset(simplex.dat,
                      select = c(biomass.prod.N,
                      biomass.prod.C,
                      biomass.prod.P,
                      avg.per.capita.N.efficiency,
                      avg.per.capita.P.efficiency,
                      avg.per.capita.C.efficiency,
                      max.growth.rate,
                      max.met.maint)))

scatterplotMatrix(physio.dat, main="Physiological data", diagonal = "none")
```



Finally, let's explore relationships between some diversity related variables. As we can see different evenness measure basically reflect each other, as do different measures of species turnover. Likewise, we see a strong postivie relationships between total abundance and species richness and between total abundance and the abundance of the most abundant species; which we should expect.

```
diversity.dat <- as.matrix(subset(simplex.dat,
                      select = c(total.abundance,
                      N.max,
                      species.richness,
                      simpson.e,
                      e.var,
                      Whittakers.turnover,
                      Jaccards.dissimilarity)))

scatterplotMatrix(diversity.dat, main="Diversity data", diagonal = "none")
```



Summary statistics

We can use the 'sapply' function to generate summary statistics (mean, variance, etc.) column by column for any of our **simplex** output. For example, analyzing phys.dat as a data.frame:

```
sapply(as.data.frame(phys.dat), mean) # sample mean
```

```
##     h.tau flow.rate  barriers     width    height amplitude frequency
##  7.760074  0.282918  2.380000 27.000000  7.430000  0.273000  0.052000
##     phase
##  4.930000
```

```
sapply(as.data.frame(phys.dat), var) # sample variance
```

```
##        h.tau    flow.rate     barriers        width       height
## 5.466889e+00 1.183039e-01 1.147071e+00 2.065657e+02 2.974848e+00
##    amplitude    frequency        phase
## 2.335455e-02 1.105051e-03 9.843535e+00
```

```
sapply(as.data.frame(phys.dat), sd) # sample standard deviation
```

```
##      h.tau  flow.rate   barriers      width     height  amplitude
##  2.3381379  0.3439534  1.0710139 14.3723922  1.7247749  0.1528219
##  frequency      phase
##  0.0332423  3.1374409
```

```
sapply(as.data.frame(phys.dat), median) # sample median
```

```
##     h.tau flow.rate  barriers     width    height amplitude frequency
##  7.600902  0.090000  2.000000 25.000000  7.000000  0.300000  0.040000
##     phase
##  5.000000
```

HIGHLY CORRELATED VARIABLES

Suppose we want to find the most highly correlated variables in our data. First, lets define a function to return the x most highly correlated variables in a dataframe.

Then, let's call our function to find the x most highly correlated variables.

```
dat <- cbind(phys.dat, physio.dat, res.dat, diversity.dat)
mosthighlycorrelated(dat, 15)
```

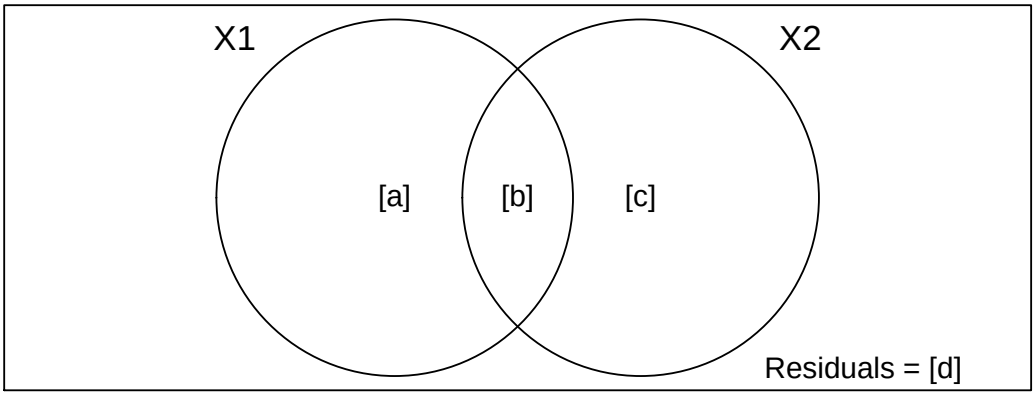
```
##                  First.Variable             Second.Variable Correlation
## 252              biomass.prod.N              biomass.prod.C   0.9938858
## 280              biomass.prod.C              biomass.prod.P   0.9850824
## 588             total.abundance                       N.max   0.9766499
## 279              biomass.prod.N              biomass.prod.P   0.9748475
## 672                   simpson.e                       e.var   0.9718603
## 728         Whittakers.turnover      Jaccards.dissimilarity   0.9698928
## 504 shannons.resource.diversity           resource.richness   0.9476535
## 530      resource.concentration          resource.particles   0.8963326
## 336 avg.per.capita.N.efficiency avg.per.capita.P.efficiency   0.8928350
## 615             total.abundance            species.richness   0.8884601
## 433                       h.tau      resource.concentration  -0.8880980
## 364 avg.per.capita.P.efficiency avg.per.capita.C.efficiency   0.8766532
## 363 avg.per.capita.N.efficiency avg.per.capita.C.efficiency   0.8592703
## 616                       N.max            species.richness   0.8467161
## 28                        h.tau                  flow.rate  -0.7467032
```

Variance partitioning.

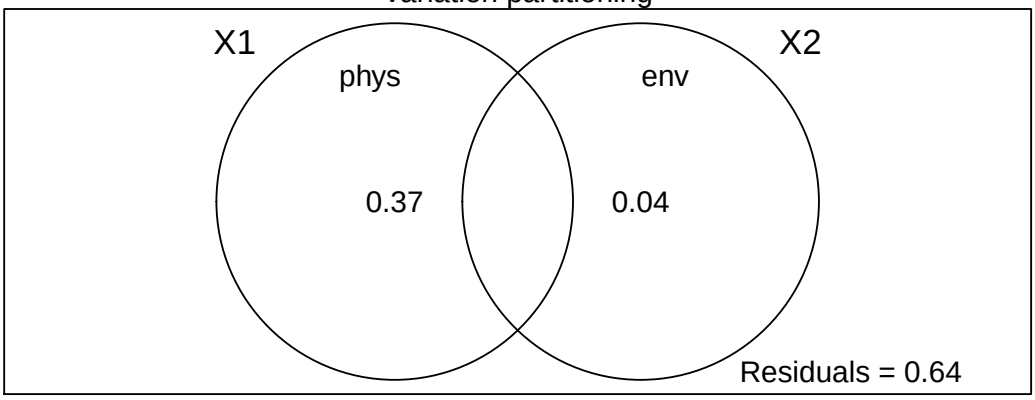
A common question to ask is whether physical variables (geograhic distance, area, volume, flow rate, etc.) has a larger influence on ecological diversity than, say, resource-related variables.

One way to address this question is to use variance partitioning. A note of caution, typing 'help(varpart)' reveals that R only uses (simple) linear regression when there is one response variable and uses redundancy analysis ordination (RDA) for two or more response variables.

Executing the code below shows that, in **simplex** models, the phys variables explain more variation in total abundance, and species richness, evenness, and turnover than do resource related variables.



Variation partitioning



Values <0 not shown