Multivariate Analysis Module III: Deep Dive

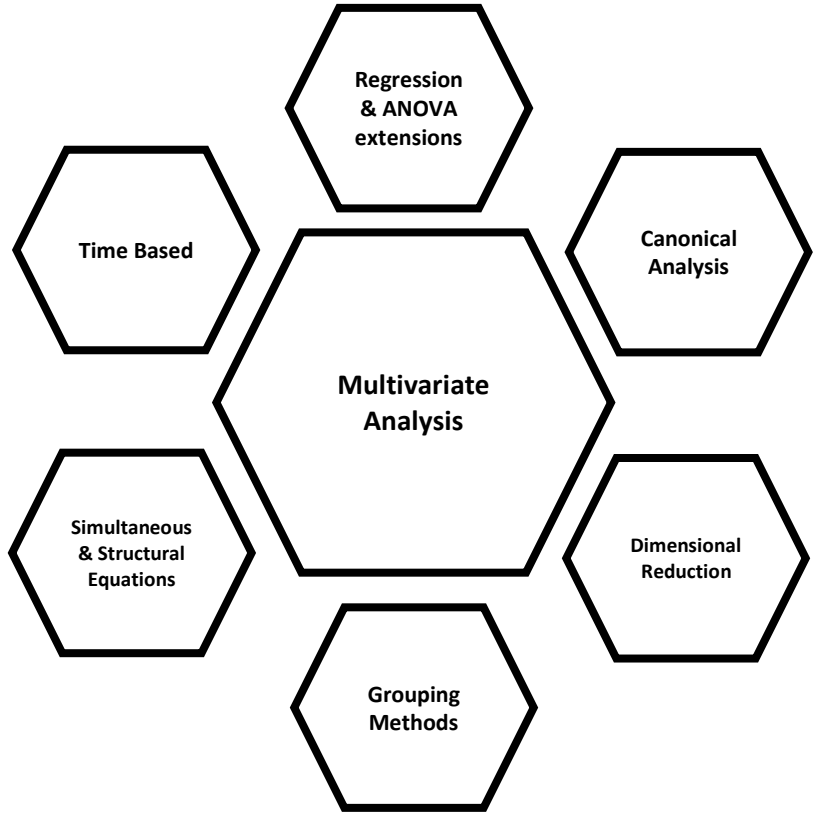
Dr. Mark Williamson

DaCCoTA

University of North Dakota

Introduction

- Last time, we covered MANOVA, MANCOVA, clustering, and recursive partitioning in more detail
- Today, we'll cover Canonical Analysis and Time Based methods



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA.

Reviewing the Basics

Canonical Correlation Analysis: comparing U matrix with V matrix (symmetric)

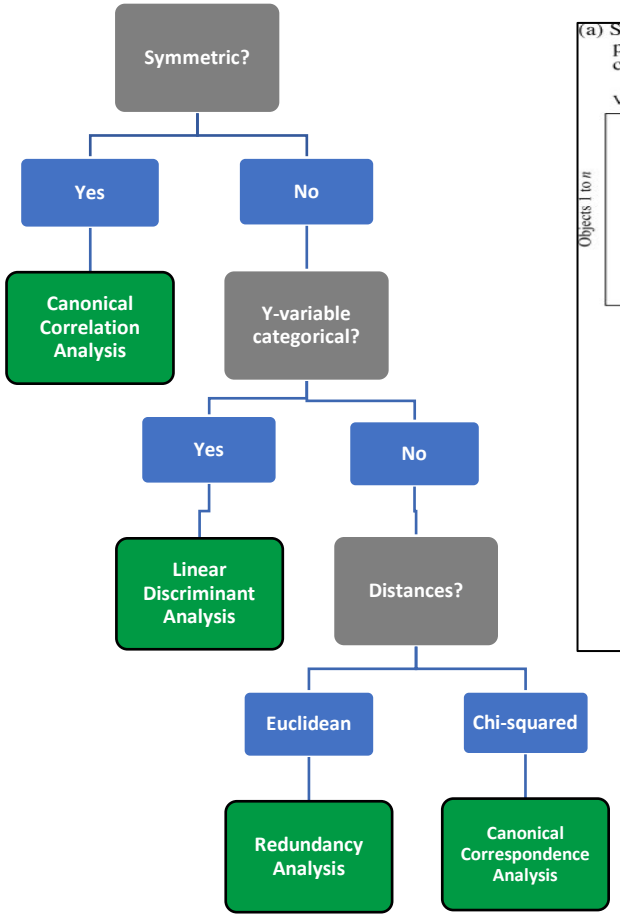
Canonical Correspondence Analysis: comparing Y matrix with X matrix (asymmetric) with Chi-Square distances

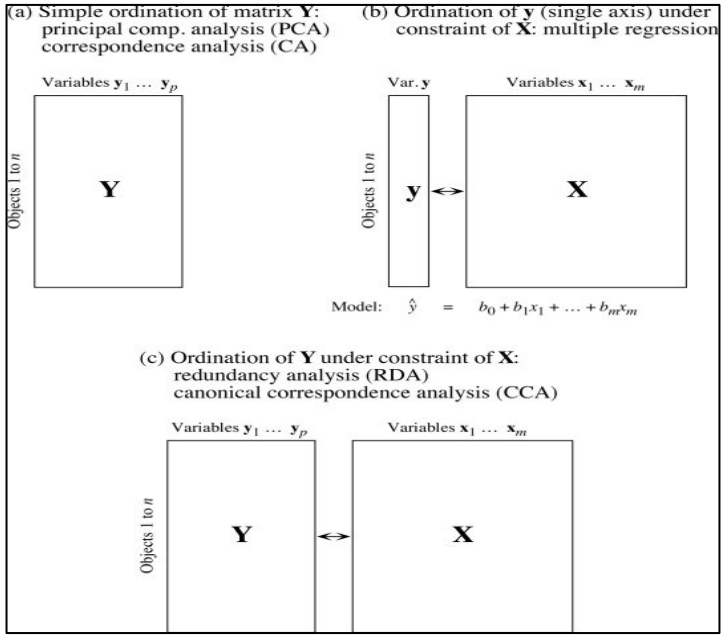
Redundancy Analysis: comparing Y matrix with X matrix (asymmetric) with Euclidean distances

Linear Discriminant Analysis: predicting Y (class of objects) like in multinomial regression, but including dimensional reduction

Vector Autoregression: comparing multiple time series that are related to one another

Principal Response Curves: comparing treatment effects in repeated measures for multiple Y variables





DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Reviewing the Basics

Canonical Correlation Analysis: comparing U matrix with V matrix (symmetric)

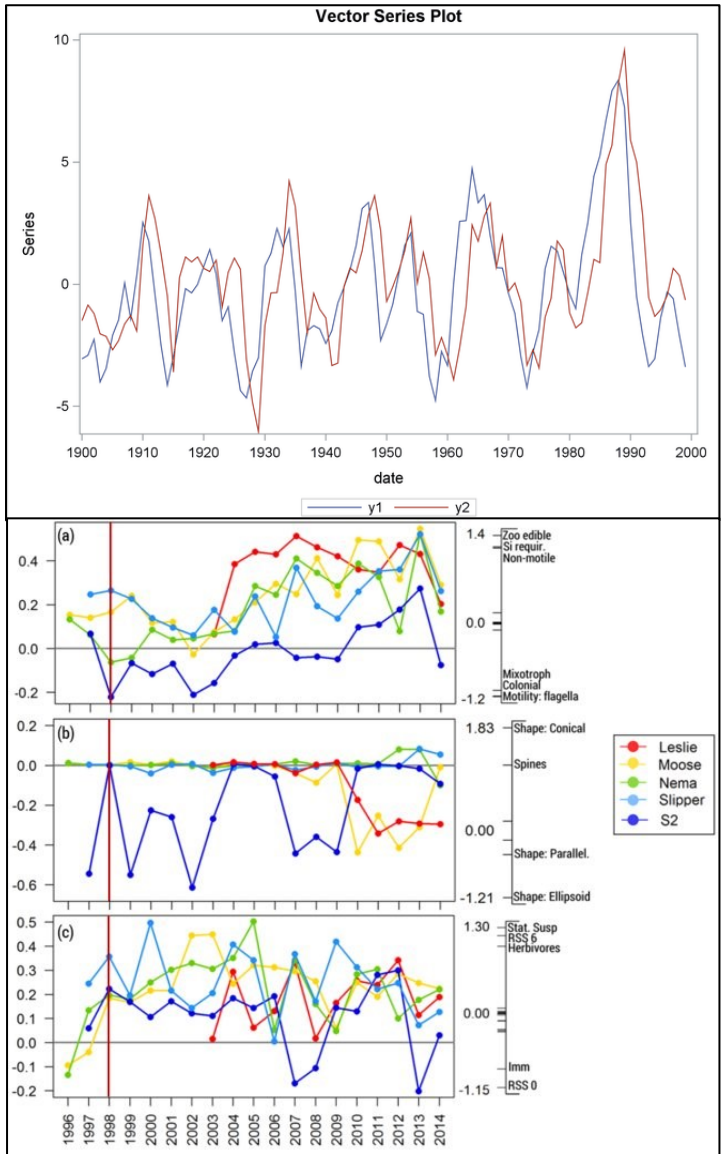
Canonical Correspondence Analysis: comparing Y matrix with X matrix (asymmetric) with Chi-Square distances

Redundancy Analysis: comparing Y matrix with X matrix (asymmetric) with Euclidean distances

Linear Discriminant Analysis: predicting Y (class of objects) like in multinomial regression, but including dimensional reduction

Vector Autoregression: comparing multiple time series that are related to one another

Principal Response Curves: comparing treatment effects in repeated measures for multiple Y variables



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Details

Canonical Analysis

Canonical: *conforming to a general rule or acceptable procedure* [1]

Canonical analysis is a multivariate technique which is concerned with determining the relationships between groups of variables in a data set. The data set is split into two groups X and Y, based on some common characteristics. The purpose of canonical analysis is then to find the relationship between X and Y [2]

[1] https://www.merriam-webster.com/dictionary/canonical
[2] https://en.wikipedia.org/wiki/Canonical_analysis

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Details 2

Canonical Correspondence Analysis vs. Redundancy Analysis

- Both stem from ecological research
- *'According to folklore, rda should be used with "short gradients" rather than cca. However, this is not based on research which finds methods based on Euclidean metric as uniformly weaker than those based on Chi-squared metric. However, standardized Euclidean distance may be an appropriate measures'* [3]
- *'CCA is a good choice if the user has clear and strong a priori hypotheses on constraints and is not interested in the major structure in the data set'* [3]
- *'CCA focuses more on species composition, i.e. relative abundance, if you have a gradient along which all species are positively correlated, RDA will detect such a gradient while CCA will not'* [4]

[3] https://mran.microsoft.com/snapshot/2015-11-17/web/packages/vegan/vegan.pdf
[4] https://sites.ualberta.ca/~lkgray/uploads/7/3/6/2/7362679/slides-gradientanalysis.pdf

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Details 3

Vector Autoregression vs. Principal Response Curves

- A Vector autoregressive (VAR) model is useful when one is interested in predicting multiple time series variables using a single model. At its core, the VAR model is an extension of the univariate autoregressive model [5]

- PRC is a special case of rda (redundancy analysis) with a single factor for treatment and a single factor for time points in repeated observations. In vegan, the corresponding rda model is defined as rda(response ~ treatment * time + Condition(time)) [6]

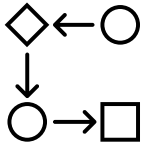
- Vector autoregression is more common in economics while Principal Response Curves are more common in ecology

[5] https://www.econometrics-with-r.org/16-1-vector-autoregressions.html
[6] https://rdrr.io/rforge/vegan/man/prc.html

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

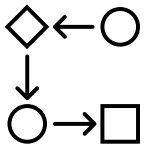
Step-by-step Example 1



Canonical Analysis using the Cars93 dataset

A. How do car fuel variables and engine variables compare to one another (Canonical Correlation Analysis)?
B. Can car variables predict the type of drive train a car has (Linear Discriminant Analysis)?
C. How do car count variables and metric variables compare to one another (Canonical Correspondence and Redundancy Analysis)?

Step-by-step Example 1



Setup

```
#librairies
library(dplyr)
library(ggplot2)
library(GGally)
library(CCA)
library(CCP)
library(MASS)
library(heplots)
library(candisc)
library(vegan)
library(vars)

#dataset
names(Cars93)
head(Cars93)
```

```
[1] "Manufacturer"      "Model"           "Type"            "Min.Price"       "Price"           "Max.Price"
 [7] "MPG.city"         "MPG.highway"     "AirBags"         "DriveTrain"      "Cylinders"       "EngineSize"
[13] "Horsepower"       "RPM"             "Rev.per.mile"    "Man.trans.avail" "Fuel.tank.capacity" "Passengers"
[19] "Length"           "Wheelbase"       "Width"           "Turn.circle"     "Rear.seat.room"  "Luggage.room"
[25] "Weight"           "Origin"          "Make"
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

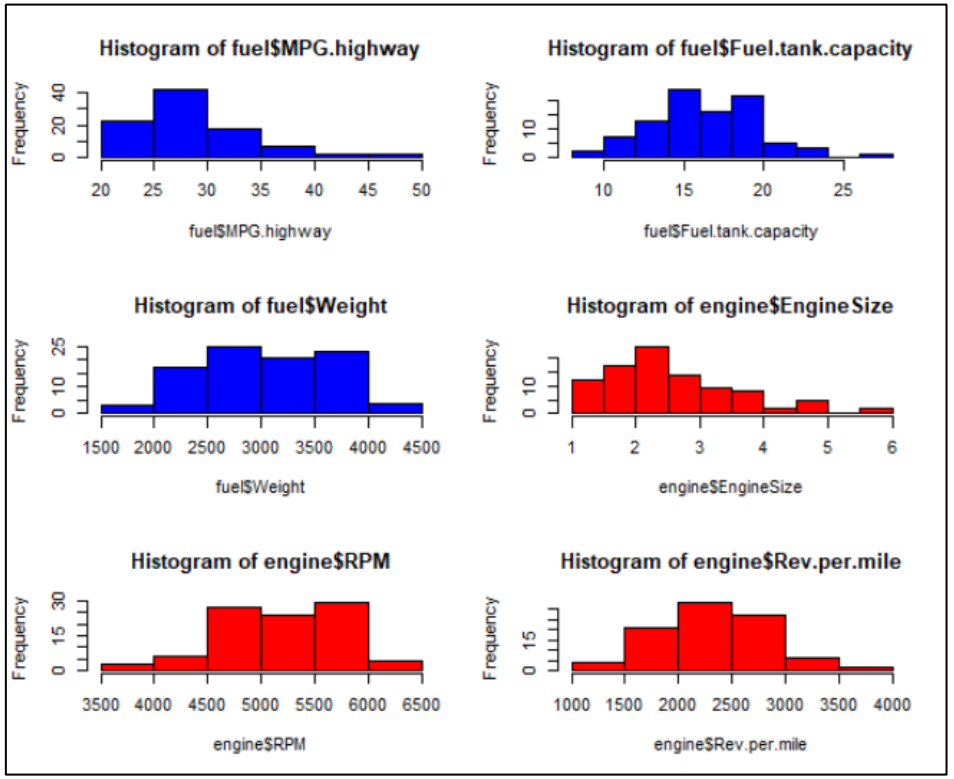
Step-by-step Example 1

Canonical Correlation Analysis

```
#set up variable sets
fuel <- data.frame(MPG.highway=Cars93$MPG.highway,
            Fuel.tank.capacity=Cars93$Fuel.tank.capacity,
            Weight=Cars93$Weight)
engine <-data.frame(EngineSize=Cars93$EngineSize,
            RPM=Cars93$RPM, Rev.per.mile=Cars93$Rev.per.mile)
```

```
#test normality
par(mfrow=c(3,2))
hist(fuel$MPG.highway, col='blue'); hist(fuel$Fuel.tank.capacity,
            col='blue'); hist(fuel$Weight, col='blue')
hist(engine$EngineSize, col="red"); hist(engine$RPM, col="red");
            hist(engine$Rev.per.mile, col="red")
par(mfrow=c(1,1))
```

```
#looking at variable sets
ggpairs(fuel)
ggpairs(engine)
```



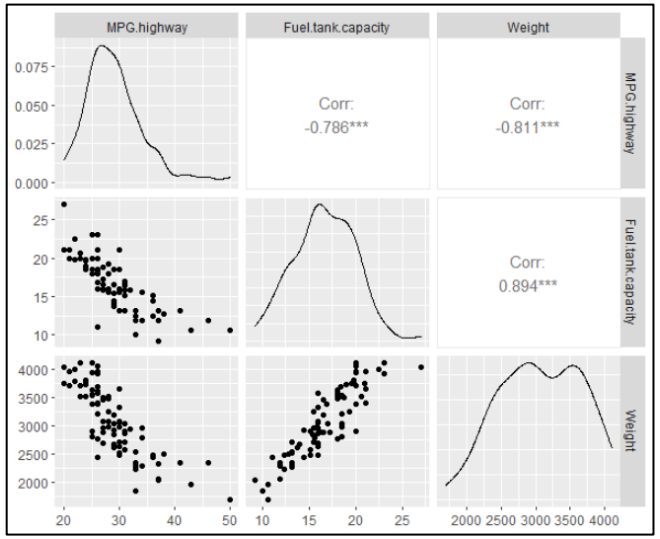
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

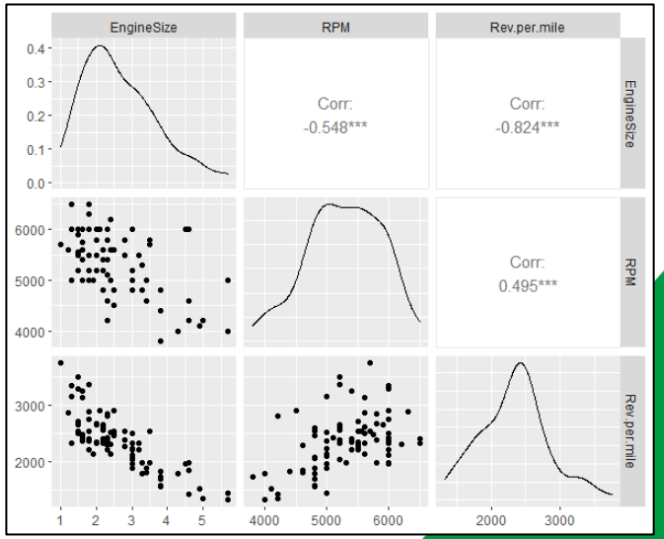
UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1

Canonical Correlation Analysis

```
#set up variable sets
fuel <- data.frame(MPG.highway=Cars93$MPG.highway,
            Fuel.tank.capacity=Cars93$Fuel.tank.capacity,
            Weight=Cars93$Weight)
engine <-data.frame(EngineSize=Cars93$EngineSize,
            RPM=Cars93$RPM, Rev.per.mile=Cars93$Rev.per.mile)

#test normality
par(mfrow=c(3,2))
hist(fuel$MPG.highway, col='blue'); hist(fuel$Fuel.tank.capacity,
            col='blue'); hist(fuel$Weight, col='blue')
hist(engine$EngineSize, col="red"); hist(engine$RPM, col="red");
            hist(engine$Rev.per.mile, col="red")
par(mfrow=c(1,1))

#looking at variable sets
ggpairs(fuel)
ggpairs(engine)
```





DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1

Canonical Correlation Analysis Cont.

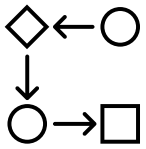
```
#correlations at and between variable sets
matcor(fuel, engine)

#canonical correlation analysis
cc1<-cc(fuel, engine)

#canonical correlations
cc1$cor

#raw canonical coefficients
cc1[3:4]

#canonical loadings (type of latent variable)
cc2 <-comput(fuel, engine, cc1)
cc2[3:6]
```

```
$Xcor
                MPG.highway Fuel.tank.capacity     Weight
MPG.highway       1.0000000       -0.7860386 -0.8106581
Fuel.tank.capacity -0.7860386       1.0000000  0.8940181
Weight           -0.8106581       0.8940181  1.0000000

$Ycor
            EngineSize       RPM Rev.per.mile
EngineSize   1.0000000 -0.5478978   -0.8240086
RPM         -0.5478978  1.0000000    0.4947642
Rev.per.mile -0.8240086  0.4947642    1.0000000

$XYcor
                MPG.highway Fuel.tank.capacity     Weight EngineSize       RPM Rev.per.mile
MPG.highway       1.0000000       -0.7860386 -0.8106581 -0.6267946 0.3134687    0.5874968
Fuel.tank.capacity -0.7860386       1.0000000  0.8940181  0.7593062 -0.3333452   -0.6097098
Weight           -0.8106581       0.8940181  1.0000000  0.8450753 -0.4279315   -0.7352642
EngineSize       -0.6267946       0.7593062  0.8450753  1.0000000 -0.5478978   -0.8240086
RPM               0.3134687      -0.3333452 -0.4279315 -0.5478978  1.0000000    0.4947642
Rev.per.mile      0.5874968      -0.6097098 -0.7352642 -0.8240086  0.4947642    1.0000000
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 1



Canonical Correlation Analysis Cont.

#correlations at and between variable sets
matcor(fuel, engine)

#canonical correlation analysis
cc1<-cc(fuel, engine)

#canonical correlations
cc1$cor

```
[1] 0.85432506 0.28478630 0.06798047
```

#raw canonical coefficients
cc1[3:4]

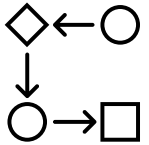
```
$xcoef
                    [,1]        [,2]        [,3]
MPG.highway      0.036729812 -0.193923071  0.263776837
Fuel.tank.capacity 0.021540947 -0.644418794 -0.272485909
Weight           0.001846686  0.001900533  0.003139611

$ycoef
                  [,1]         [,2]        [,3]
EngineSize    0.9112733314 -1.5202454004  0.093249971
RPM           0.0001183839 -0.0007129807 -0.001877324
Rev.per.mile -0.0002201029 -0.0032065224  0.001553295
```

#canonical loadings (type of latent variable)
cc2 <-comput(fuel, engine, cc1)
cc2[3:6]

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Canonical Correlation Analysis Cont.

#correlations at and between variable sets
matcor(fuel, engine)

#canonical correlation analysis
cc1<-cc(fuel, engine)

#canonical correlations
cc1$cor

#raw canonical coefficients
cc1[3:4]

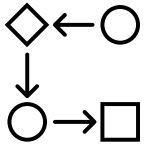
#canonical loadings (type of latent variable)
cc2 <-comput(fuel, engine, cc1)
cc2[3:6]

```
$corr.X.xscores
                  [,1]       [,2]       [,3]
MPG.highway      -0.7427862 -0.28166210  0.6074003
Fuel.tank.capacity  0.8906098 -0.29826770 -0.3432937
Weight           0.9937537  0.06997607 -0.0869311

$corr.Y.xscores
                [,1]       [,2]       [,3]
EngineSize    0.8514764 -0.009132577  0.005100486
RPM          -0.4283287 -0.099416597 -0.053819240
Rev.per.mile -0.7289815 -0.143264891  0.009330138

$corr.X.yscores
                  [,1]       [,2]       [,3]
MPG.highway      -0.6345808 -0.08021351  0.041291363
Fuel.tank.capacity  0.7608703 -0.08494255 -0.023337266
Weight           0.8489887  0.01992823 -0.005909617

$corr.Y.yscores
                [,1]       [,2]       [,3]
EngineSize    0.9966656 -0.03206817  0.0750287
RPM          -0.5013651 -0.34909192 -0.7916867
Rev.per.mile -0.8532835 -0.50306103  0.1372473
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Canonical Correlation Analysis Cont.

```
#significance test for canonical dimensions
rho <-cc1$cor
n <-dim(fuel)[1]
p <-length(fuel)
q <-length(engine)

p.asym(rho, n, p, q, tstat = "Wilks") #1st significant only
p.asym(rho, n, p, q, tstat = "Hotelling") #1st significant only
p.asym(rho, n, p, q, tstat = "Pillai") #1-3 significant

#standardized canonical coefficients
s1 <- diag(sqrt(diag(cov(fuel)))) #for fuel
s2 <- diag(sqrt(diag(cov(engine)))) #for engine

s1 %*% cc1$xcoef
s2 %*% cc1$ycoef
```

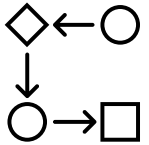
```
Wilks' Lambda, using F-approximation (Rao's F):
          stat     approx df1      df2    p.value
1 to 3:  0.2470733 18.2730423   9 211.8857 0.00000000
2 to 3:  0.9146502  2.0071313   4 176.0000 0.09549698
3 to 3:  0.9953787  0.4132093   1  89.0000 0.52199772

Hotelling-Lawley Trace, using F-approximation:
          stat     approx df1 df2    p.value
1 to 3:  2.794843515 26.6027698   9 257 0.00000000
2 to 3:  0.092904336  2.0361534   4 263 0.08971305
3 to 3:  0.004642801  0.4163045   1 269 0.51933675

Pillai-Bartlett Trace, using F-approximation:
          stat     approx df1 df2      p.value
1 to 3:  0.815595888 11.0767102   9 267 1.132427e-14
2 to 3:  0.085724584  2.0076012   4 273 9.367253e-02
3 to 3:  0.004621345  0.4304482   1 279 5.123098e-01
```

```
 [,1]     [,2]      [,3]
[1,] 0.19583330 -1.033945  1.4063858
[2,] 0.07064075 -2.113288 -0.8935822
[3,] 1.08935340  1.121118  1.8520456


          [,1]      [,2]       [,3]
[1,]  0.94532125 -1.5770463  0.09673407
[2,]  0.07064345 -0.4254582 -1.12025893
[3,] -0.10928254 -1.5920593  0.77122105
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Canonical Correlation Analysis Summary

Tests of Canonical Dimensions

Dimension	Canonical Corr.	F	df1	df2	p
1	0.85	18.27	9	211.9	0.0000
2	0.28	2.01	4	176.0	0.0955
3	0.07	0.41	1	89.0	0.5220

Table 2: Standardized Canonical Coefficients

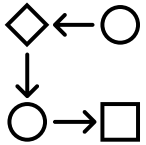
	Dimension 1
Fuel Variables	
MPG	0.20
Tank Capacity	0.07
Weight	1.09
Engine Variables	
Engine size	0.95
RPM	0.07
Rev per mile	-0.11

https://stats.idre.ucla.edu/r/dae/canonical-correlation-analysis/

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Linear Discriminant Analysis

```
#set up dataset
CarsLD <-data.frame(DriveTrain=Cars93$DriveTrain, MPG.highway=Cars93$MPG.highway,
            Fuel.tank.capacity=Cars93$Fuel.tank.capacity, Weight=Cars93$Weight, EngineSize=Cars93$EngineSize,
            RPM=Cars93$RPM, Rev.per.mile=Cars93$Rev.per.mile)
```

```
#scale data
CarsLD[2:7]<-scale(CarsLD[2:7])
apply(CarsLD[2:7], 2, mean)
apply(CarsLD[2:7], 2, sd)
```

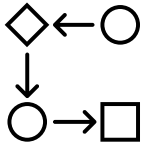
MPG.highway	Fuel.tank.capacity	Weight	EngineSize	RPM	Rev.per.mile
2.644200e-17	1.254095e-16	-1.512637e-16	-2.160780e-16	-5.311705e-16	3.743823e-16
MPG.highway	Fuel.tank.capacity	Weight	EngineSize	RPM	Rev.per.mile
1	1	1	1	1	1

```
#training and test samples
set.seed(1)

sample <-sample(c(TRUE, FALSE), nrow(CarsLD), replace=TRUE, prob=c(0.7,0.3))
train <-CarsLD[sample, ] #70% as training set
test <-CarsLD[!sample, ] #30% as testing set
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Linear Discriminant Analysis Cont.

```
#fit LDA model
LDA1 <-lda(DriveTrain~., data=CarsLD)
LDA1

#make predictions
predicted <-predict(LDA1, test)

names(predicted)
head(predicted$class)
head(predicted$posterior)
head(predicted$x)

#find accuracy of model
mean(predicted$class==test$DriveTrain) #80% accuracy
```

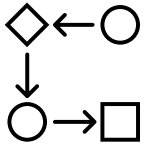
0.8

```
Call:
lda(DriveTrain ~ ., data = CarsLD)

Prior probabilities of groups:
     4WD     Front      Rear
0.1075269 0.7204301 0.1720430

Group means:
      MPG.highway Fuel.tank.capacity     Weight  EngineSize       RPM Rev.per.mile
4WD    -0.6163148         0.5322619  0.3934534 -0.09422154 -0.2692084    0.06705994
Front   0.2162122        -0.2467828 -0.2270703 -0.21493519  0.1574941    0.17397110
Rear   -0.5201921         0.7007393  0.7049487  0.95892957 -0.4912512   -0.77041646

Coefficients of linear discriminants:
                    LD1       LD2
MPG.highway      0.02070679 -0.4583823
Fuel.tank.capacity 0.31032866  0.9522342
Weight          -0.76900966  0.2888419
EngineSize       1.49478625 -1.2719076
RPM              0.08492831 -0.5829880
Rev.per.mile    -0.05483095  0.1531040

Proportion of trace:
  LD1   LD2
0.584 0.416
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Linear Discriminant Analysis Cont.

```
#fit LDA model
LDA1 <-lda(DriveTrain~., data=CarsLD)
LDA1

#make predictions
predicted <-predict(LDA1, test)

names(predicted)
head(predicted$class)
head(predicted$posterior)
head(predicted$x)

#find accuracy of model
mean(predicted$class==test$DriveTrain) #80% accuracy
```

0.8

```
[1] "class"     "posterior" "x"


[1] Front Front Front Front Rear  Front
Levels: 4WD Front Rear


        4WD     Front      Rear
2  0.06442994 0.7795587 0.15601140
6  0.10805248 0.8258750 0.06607248
12 0.02658847 0.8772618 0.09614969
13 0.05677410 0.8772765 0.06594944
18 0.01344075 0.1400639 0.84649534
23 0.02823698 0.9410790 0.03068401


         LD1        LD2
2   0.27337901  0.1115372
6  -0.47730054  0.3882579
12 -0.04259095 -0.6380602
13 -0.42855194 -0.1176370
18  2.81160767  0.4081441
23 -0.95145150 -0.7128459
```

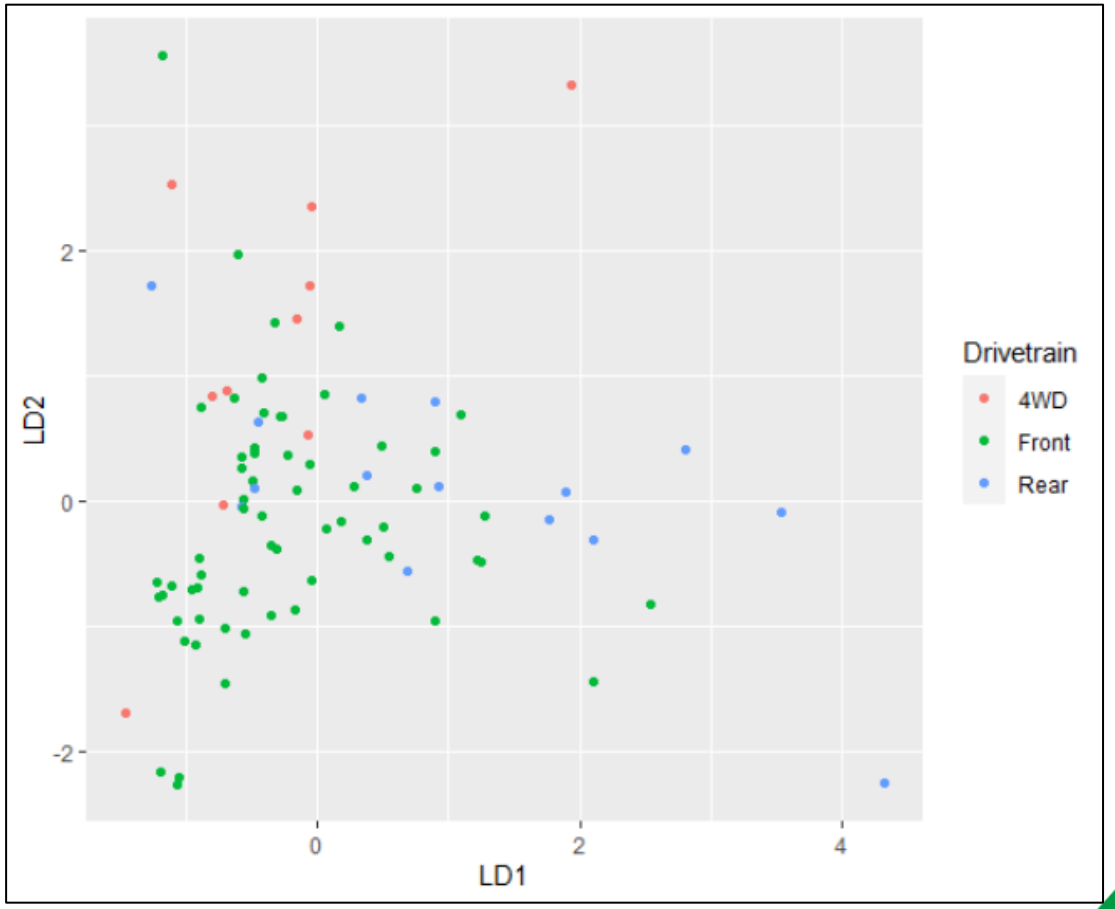
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 1

Linear Discriminant Analysis Cont.

```
#visualize the results

DriveTrain2 <-data.frame(Drivetrain=CarsLD$DriveTrain)
LDA_plot <-cbind(DriveTrain2, predict(LDA1)$x)
ggplot(LDA_plot, aes(LD1, LD2)) +
  geom_point(aes(color=Drivetrain))
```

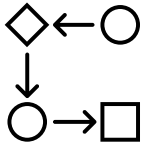


https://www.statology.org/linear-discriminant-analysis-in-r/

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1



Canonical Correspondence and Redundancy Analysis

```
#(X, Y, Z matrices) where Z is conditioning matrix
#X, Y, Z -> partial correspondence/redundancy analysis
#X, Y    -> canonical correspondence/redundancy analysis
#X       -> principal components analysis

CarCounts <-data.frame(Cylinders=Cars93$Cylinders,
                       Passengers=Cars93$Passengers,
                       Fuel.tank.capacity=Cars93$Fuel.tank.capacity)
CarMetrics<-data.frame(EngineSize=Cars93$EngineSize,
                       Horsepower=Cars93$Horsepower, RPM=Cars93$RPM,
                      Rev.per.mile=Cars93$Rev.per.mile, Price=Cars93$Price,
                       MPG.highway=Cars93$MPG.highway)

CarCounts$Cylinders <-as.integer(CarCounts$Cylinders)
str(CarCounts); str(CarMetrics)
```

```
'data.frame': 93 obs. of  3 variables:
 $ Cylinders       : int  2 4 4 4 2 2 4 4 4 5 ...
 $ Passengers      : int  5 5 5 6 4 6 6 6 5 6 ...
 $ Fuel.tank.capacity: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
```

```
'data.frame': 93 obs. of  6 variables:
 $ EngineSize  : num  1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...
 $ Horsepower  : int  140 200 172 172 208 110 170 180 170 200 ...
 $ RPM         : int  6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...
 $ Rev.per.mile: int  2890 2335 2280 2535 2545 2565 1570 1320 1690 1510 ...
 $ Price       : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
 $ MPG.highway : int  31 25 26 26 30 31 28 25 27 25 ...
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

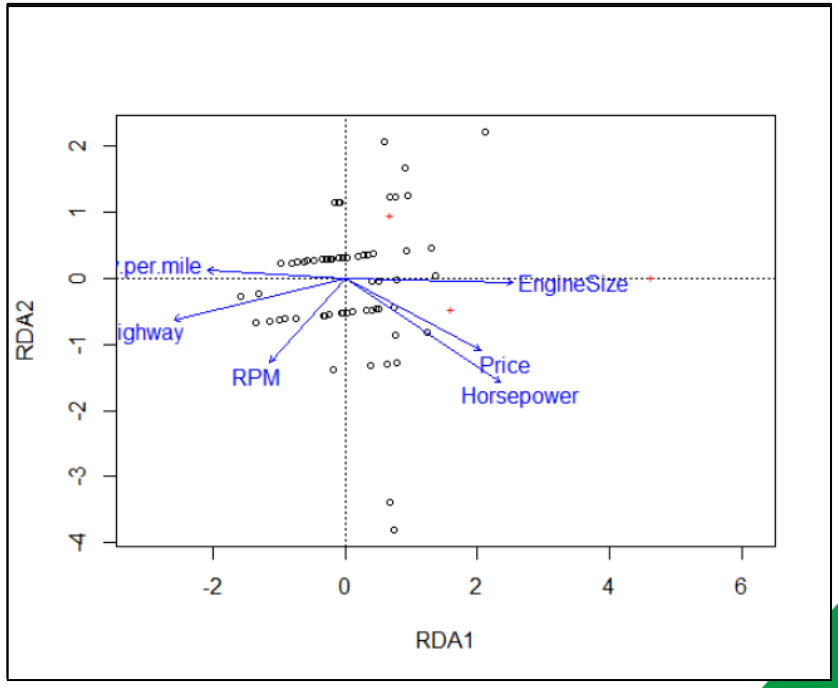
UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1

Canonical Correspondence and Redundancy Analysis Cont.

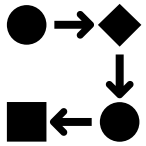
```
#basic cca with all variables
Car_CCA1 <-cca(CarMetrics, CarCounts)
Car_CCA1
plot(Car_CCA1)

#more nuanced formula
Car_CCA2 <-cca(CarCounts ~ EngineSize*(RPM
+ Rev.per.mile) + Price, data=CarMetrics)
Car_CCA2
plot(Car_CCA2)

#redundancy analysis
Car_RDA <-rda(CarMetrics, CarCounts)
Car_RDA
plot(Car_RDA)
```

```
Call: rda(X = CarCounts, Y = CarMetrics)

              Inertia Proportion Rank
Total         13.2302     1.0000
Constrained    9.7567     0.7375    3
Unconstrained  3.4735     0.2625    3
Inertia is variance

Eigenvalues for constrained axes:
 RDA1  RDA2  RDA3
9.254 0.430 0.073

Eigenvalues for unconstrained axes:
   PC1    PC2    PC3
2.7100 0.4647 0.2989
```



https://mran.microsoft.com/snapshot/2015-11-17/web/packages/vegan/vegan.pdf

Step-by-step Example 2



Time-based multivariate methods using simulated data and pyrifos data

A. How do two time series relate to one another (Vector Autoregression?
B. How do doses of pesticide affect invertebrate abundances across time (Principal Response Curves)?

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

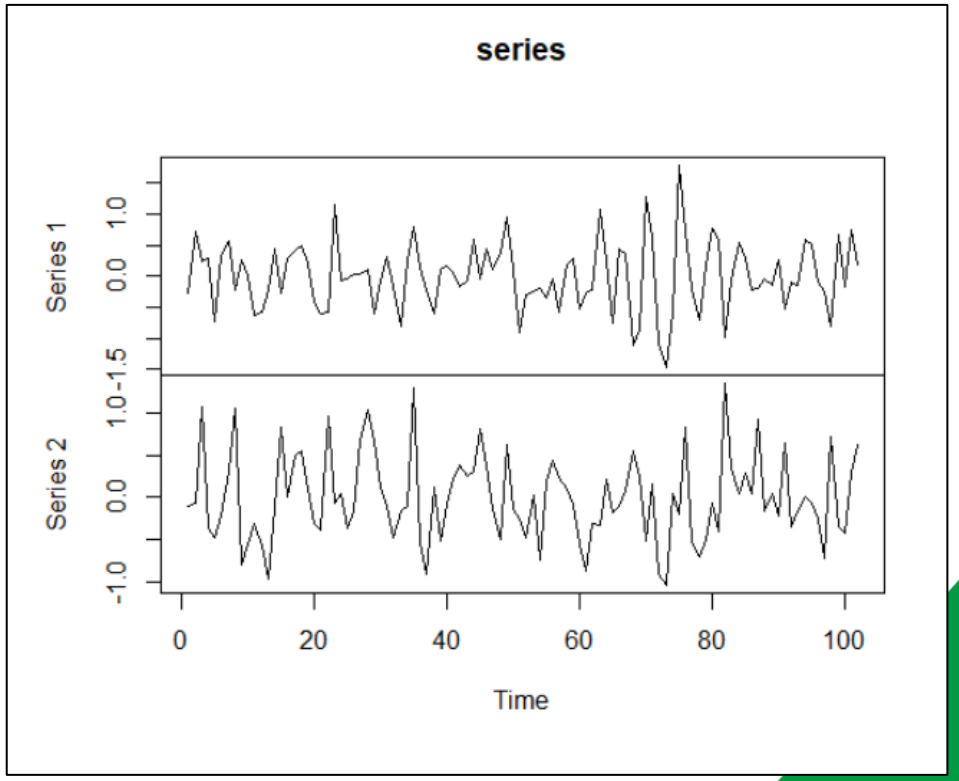
UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2

Vector Autoregression

```
set.seed(123) # For reproducibility

# Generate sample
t <- 100 # Number of time series observations
k <- 2 # Number of endogenous variables
p <- 2 # Number of lags

# Generate coefficient matrices
A.1 <- matrix(c(.1, .2, .3, .4), k) # Coefficient matrix of lag 1
A.2 <- matrix(c(-.4, -.3, -.2, -.1), k) # Coefficient matrix of lag 2
A <- cbind(A.1, A.2) # Companion form of the coefficient matrices

# Generate series
series <- matrix(0, k, t + 2*p) # Raw series with zeros
for (i in (p + 1):(t + 2*p)){ # Generate series with e ~ N(0,0.5)
  series[, i] <- A.1%*%series[, i-1] + A.2%*%series[, i-2] + rnorm(k, 0, .5)
}
```

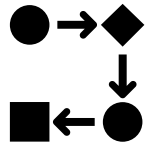
```
series <- ts(t(series[, -(1:p)])) # Convert to time series format
names <- c("V1", "V2") # Rename variables
plot.ts(series) # Plot the series
```



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2



Vector Autoregression Cont.

```
var.1 <- VAR(series, 2, type = "none") # Estimate the model

var.aic <- VAR(series, type = "none", lag.max = 5, ic = "AIC")

summary(var.aic)
```

```
VAR Estimation Results:
========================
Endogenous variables: Series.1, Series.2
Deterministic variables: none
Sample size: 100
Log Likelihood: -129.082
Roots of the characteristic polynomial:
0.7472 0.7472 0.2515 0.2515
Call:
VAR(y = series, type = "none", lag.max = 5, ic = "AIC")
```

```
Estimation results for equation Series.1:
=========================================
Series.1 = Series.1.l1 + Series.2.l1 + Series.1.l2 + Series.2.l2

           Estimate Std. Error t value Pr(>|t|)
Series.1.l1  0.13820    0.08363   1.652 0.101709
Series.2.l1  0.35797    0.09118   3.926 0.000163 ***
Series.1.l2 -0.54641    0.08561  -6.383 6.15e-09 ***
Series.2.l2 -0.15098    0.08983  -1.681 0.096061 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.4548 on 96 degrees of freedom
Multiple R-Squared: 0.3468,   Adjusted R-squared: 0.3196
F-statistic: 12.74 on 4 and 96 DF,  p-value: 2.339e-08
```

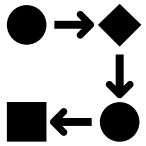
```
Estimation results for equation Series.2:
=========================================
Series.2 = Series.1.l1 + Series.2.l1 + Series.1.l2 + Series.2.l2

           Estimate Std. Error t value Pr(>|t|)
Series.1.l1  0.27986    0.09040   3.096  0.00257 **
Series.2.l1  0.16766    0.09855   1.701  0.09214 .
Series.1.l2 -0.22586    0.09253  -2.441  0.01648 *
Series.2.l2 -0.12703    0.09709  -1.308  0.19386
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.4915 on 96 degrees of freedom
Multiple R-Squared: 0.1511,   Adjusted R-squared: 0.1158
F-statistic: 4.273 on 4 and 96 DF,  p-value: 0.00317
```

```
Covariance matrix of residuals:
          Series.1 Series.2
Series.1  0.20673 -0.02813
Series.2 -0.02813  0.24158

Correlation matrix of residuals:
          Series.1 Series.2
Series.1   1.0000  -0.1259
Series.2  -0.1259   1.0000
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 2



Vector Autoregression Cont.

```
#True values
A

# Extract coefficients, standard errors etc. from the object
# produced by the VAR function
est_coefs <- coef(var.aic)

# Extract only the coefficients for both dependend variables
# and combine them to a single matrix
est_coefs <- rbind(est_coefs[[1]][, 1], est_coefs[[2]][, 1])

# Print the rounded estimates in the console
round(est_coefs, 2)
```

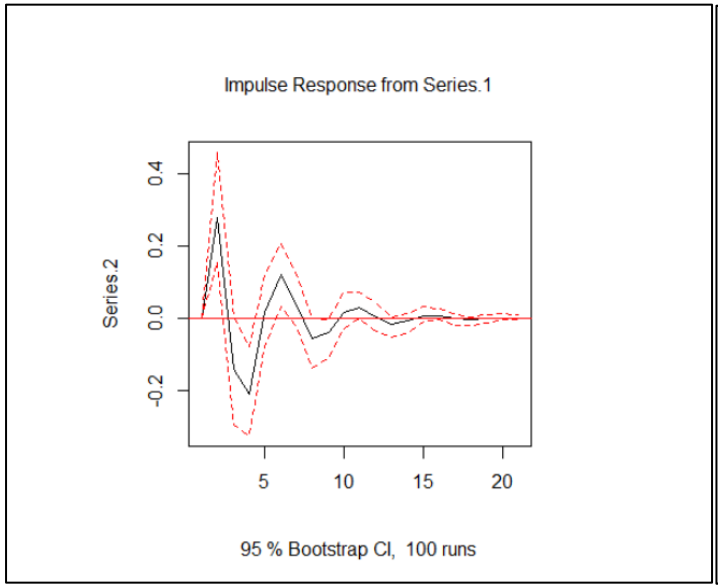
```
     [,1] [,2] [,3] [,4]
[1,]  0.1  0.3 -0.4 -0.2
[2,]  0.2  0.4 -0.3 -0.1
```

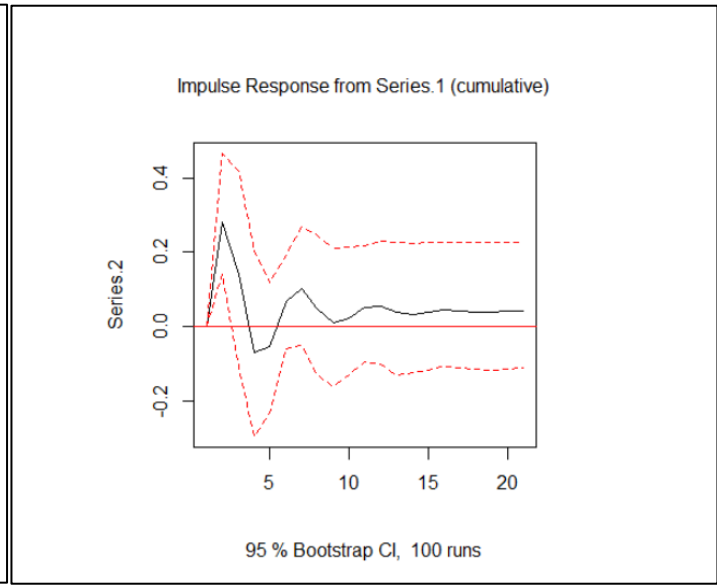
```
 Series.1.l1 Series.2.l1 Series.1.l2 Series.2.l2
[1,]        0.14        0.36       -0.55       -0.15
[2,]        0.28        0.17       -0.23       -0.13
```

Step-by-step Example 2

Vector Autoregression Cont.

```
#Impulse response
# Calculate the IRF
ir.1 <- irf(var.1, impulse = "Series.1",
            response = "Series.2", n.ahead = 20,
            ortho = FALSE)

# Plot the IRF
plot(ir.1)

# Calculate impulse response
ir.2 <- irf(var.1,impulse="Series.1",
            response="Series.2",n.ahead = 20,
            ortho = FALSE,
            cumulative = TRUE)

# Plot
plot(ir.2)
```





https://www.r-econometrics.com/timeseries/varintro/

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 2

Principal Response Curves

```
## Chlorpyrifos experiment and experimental design: Pesticide
## treatment in ditches (replicated) and followed over from 4
weeks
## before to 24 weeks after exposure

data(pyrifos)
week <- gl(11, 12, labels=c(-4, -1, 0.1, 1, 2, 4, 8, 12, 15, 19, 24))
dose <- factor(rep(c(0.1, 0, 0, 0.9, 0, 44, 6, 0.1, 44, 0.9, 0, 6), 11))
ditch <- gl(12, 1, length=132)

# PRC
mod <- prc(pyrifos, dose, week) #response, treatment, time
mod          # RDA

summary(mod)   # PRC
```

```
Call:
prc(response = pyrifos, treatment = dose, time = week)
Species scores:
   Simve    Daplo    Cerpu    Alogu    Aloco    Alore    Aloaf    Copsp    Ostsp    Slyla    Acrha    Aloex    Chysp
-1.461934 -0.796510 -0.295171 -0.152301 -0.096273 -0.171335 -0.231967 -0.635966 -1.257492  0.302872 -0.057396 -0.124049 -0.051689
   Alona    Plead    Oxyte    Grate    Copdi    NauLa    CilHa    Strvi    amosp    Ascmo    Synsp    Squro    Squmu
-0.034638 -0.075268 -0.013815 -0.052667 -0.777088 -2.636100 -0.486881 -1.669474  0.738371 -0.037926  0.014409 -0.143790  0.246185
   Polar    Kerqu    Anufi    Mytve    Mytvi    Mytmu    Lepsp    Leppa    Colob    Colbi    Colun    Lecsp    Lecqu
-0.251255 -0.269397 -0.235362 -0.040448 -0.049452 -0.057589 -0.542922 -0.046123  0.393234  0.075905  0.450495 -0.257170  0.048327
   Lecco    Leclu    Lecfl    Tripo    Cepsp    Monlo    Monae    Scalo    Trilo  Tripo.1    Tricy    Trisp    Tepat
 0.158007 -0.027077  0.221912 -0.117056  0.440303  0.287108  0.048918  0.041981  0.021257 -0.134025 -0.182409 -0.042651 -0.004198
   Rotne    Notla    Filsp    Lopox  hydrspec  bothrosp  olchaeta  erpoocto  glsicomp  alglhete  hebdstag  sphidae  ansuvote
-0.078168  0.062163  0.091561  0.016854  0.026485 -0.216815  0.633674  0.490029  0.078527  0.039728 -0.490678 -0.796015 -0.076512
 armicris  bathcont  binitent  gyraalbu  hippcomp  lymnstag  lymnaes7  physfont  plbacorn  popyanti  radiovat  radipere  valvcris
-0.913681 -0.039855  1.060788 -0.017975 -0.219974  0.143403 -0.073502  0.014349 -0.046098 -0.691904  0.010777  0.340163 -0.005753
 valvpisc  hycarina  gammpule  aselaqua  proameri  collembo  caenhora  caenluct  caenrobu  cloedipt  cloesimi  aeshniae  libellae
 0.145523 -0.567803 -0.830166 -0.858606 -0.063401 -0.016264 -3.136867 -1.292300 -0.068624 -2.574625 -0.675580 -0.115677  0.044524
 conagrae  corident  coripanz  coripunc  cymabons  hesplinn  hespsahl  notoglau  notomacu  notoobli  notoviri  pacoconc  pleaminu
-0.886795 -0.007484 -0.065501  0.096124 -0.025213  0.037779 -0.018078 -0.301948 -0.027226 -0.044790 -0.117398 -0.009087 -0.038705
 sigadist  sigafall  sigastri  sigarasp  colyfusc  donacis6  gyrimari  haliconf  haliflav  haligruf  haliobli  herubrev  hya_herm
-0.041594  0.009987 -0.032743  0.150817 -0.019847  0.042751 -0.005753 -0.243054 -0.024222 -0.244814 -0.071501  0.069877  0.175327
 hyglpusi  hyhyovat  hypoplan  hyporusp  hytuinae  hytuvers  laphminu  noteclav  rhantusp  sialluta  ablalong  ablaphmo  cltanerv
 0.006404 -0.013159 -0.008145 -0.126197 -1.259664 -0.963901 -0.344204 -0.004303 -0.036774 -0.603320 -0.008145 -1.627590 -0.041132
 malopisp  mopetenu  prdiussp  pstavari  chironsp  crchirsp  crclglat  ditendsp  mitegchl  pachgarc  pachgvit  popegnub  popedisp
-0.025919 -0.004740 -0.301790 -0.045054 -1.027839 -0.009087 -0.015746 -0.045402 -0.125429  0.006628  0.016265 -0.121971  0.037879
 acriluce  chclpige  conescut  cricotsp  liesspec  psclbarb  psclgsli  psclobvi  psclplat  psclpsil  pscladsp  cladotsp  laa_spec
 0.004324  0.004756 -0.446524 -0.066095 -0.058403  0.015576 -0.327165  0.197081  0.028310 -0.003991 -0.003086 -0.293623 -0.018548
 patanysp  tatarssp  zaa_spec  anopmacu  cepogoae  chaoobsc  cucidae4  tabanusp  agdasphr  athrater  cyrncren  holodubi  holopici
-0.079841 -0.364072 -0.027165  0.089046 -1.389767 -1.328261 -0.018078  0.006309 -0.147828 -0.036774 -0.038705 -0.051532 -0.332631
 leceriae  lilurhom  monaangu  mystazur  mystloni  oecefurv  oecelacu  triabico  paponysp
-0.162413 -0.004929 -0.350402 -0.018078 -1.630725 -0.291847 -0.140893 -0.048357 -0.053182

Coefficients for dose + week:dose interaction
which are contrasts to dose 0
rows are dose, columns are week
      -4     -1    0.1      1      2      4      8     12     15     19     24
0.1 0.01335 0.02543 0.01887 0.007525 0.03886 0.02524 0.01494 0.02841 0.02077 0.04002 0.01449
0.9 0.01500 0.03580 0.03582 0.088245 0.09207 0.07966 0.02503 0.06563 0.04455 0.03249 0.02857
6   0.03074 0.02280 0.08397 0.213933 0.20045 0.21296 0.10381 0.08691 0.05695 0.06091 0.03372
44  0.02586 0.03623 0.13520 0.233266 0.24009 0.27059 0.23891 0.18650 0.14465 0.10672 0.05739
```

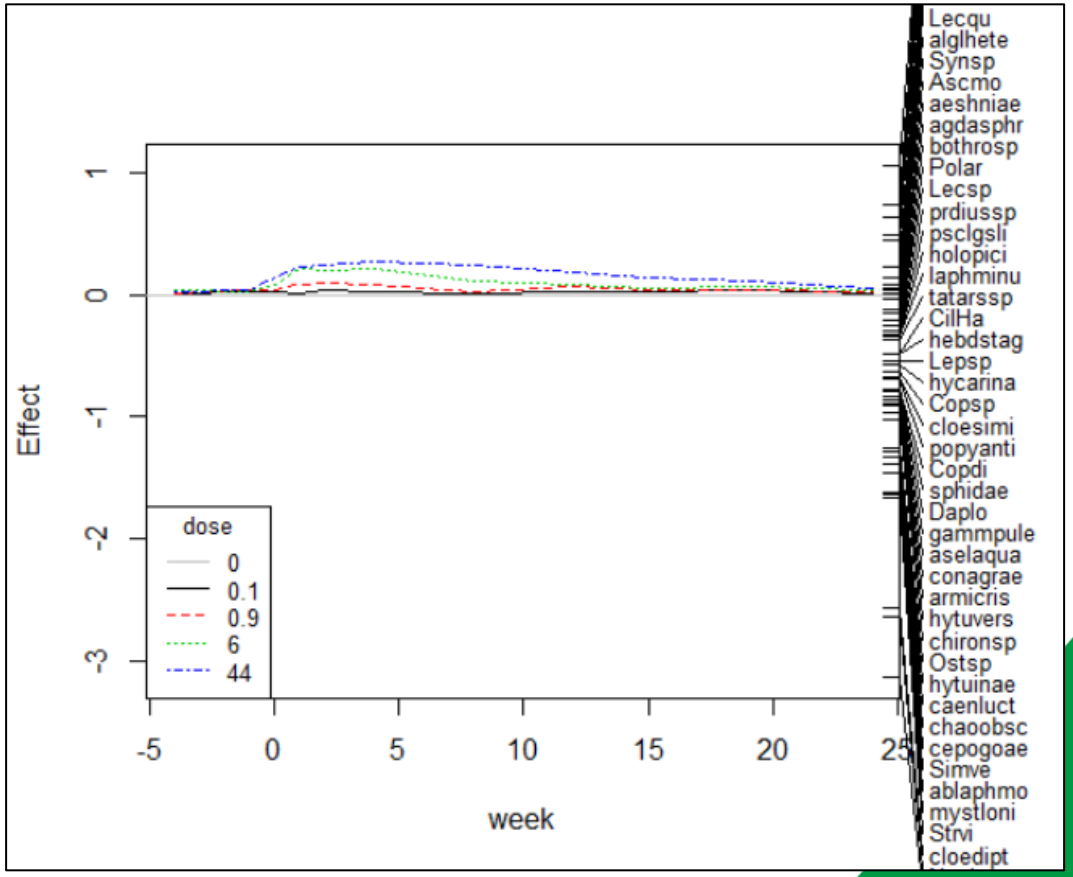
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Step-by-step Example 2

Principal Response Curves

```
logabu <- colSums(pyrifos)
plot(mod, select = logabu > 100)
## Ditches are randomized, we have a time series, and are only
## interested in the first axis
ctrl <- how(plots = Plots(strata = ditch,type = "free"),
          within = Within(type = "series"), nperm = 99)
anova(mod, permutations = ctrl, first=TRUE)


#reduced
pyrifos_red <-pyrifos[,1:10]

mod2 <- prc(pyrifos_red, dose, week) #response, treatment, time
mod2          # RDA

summary(mod2)   # PRC
plot(mod2)
```



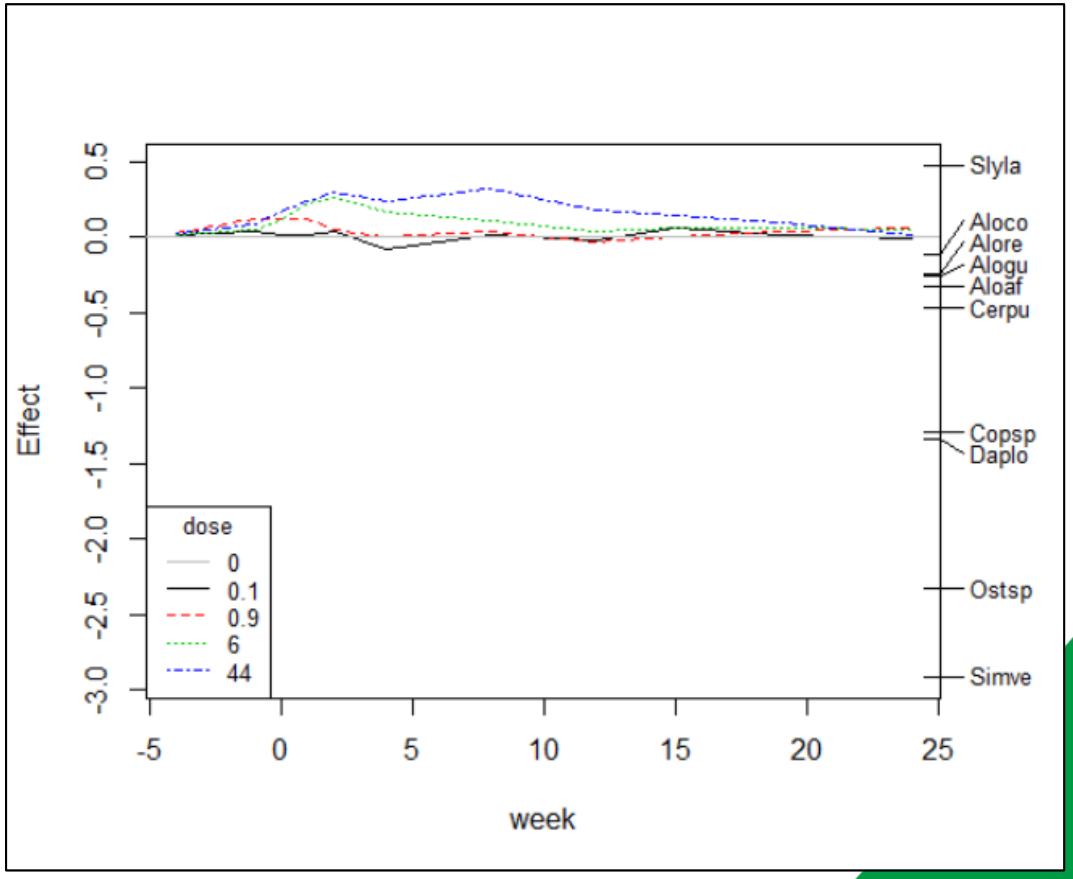
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 2

Principal Response Curves

```
logabu <- colSums(pyrifos)
plot(mod, select = logabu > 100)
## Ditches are randomized, we have a time series, and are only
## interested in the first axis
ctrl <- how(plots = Plots(strata = ditch,type = "free"),
        within = Within(type = "series"), nperm = 99)
anova(mod, permutations = ctrl, first=TRUE)


#reduced
pyrifos_red <-pyrifos[,1:10]

mod2 <- prc(pyrifos_red, dose, week) #response, treatment, time
mod2          # RDA

summary(mod2)   # PRC
plot(mod2)
```



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

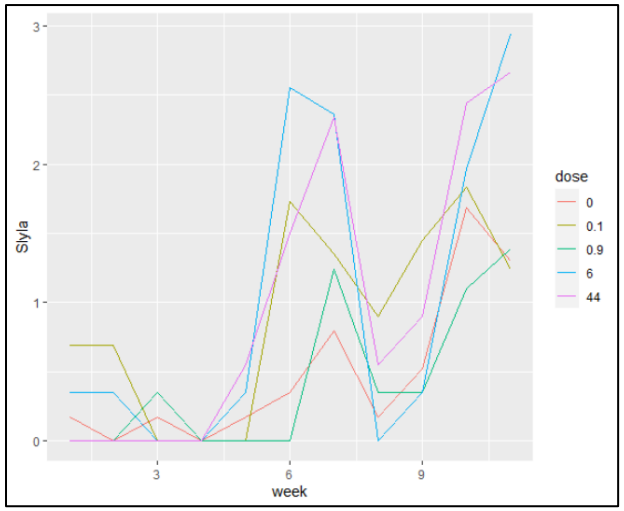
UNIVERSITY OF NORTH DAKOTA

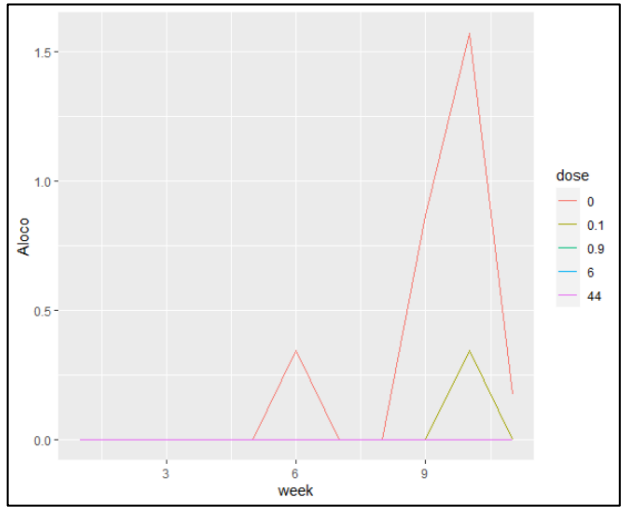
Step-by-step Example 2

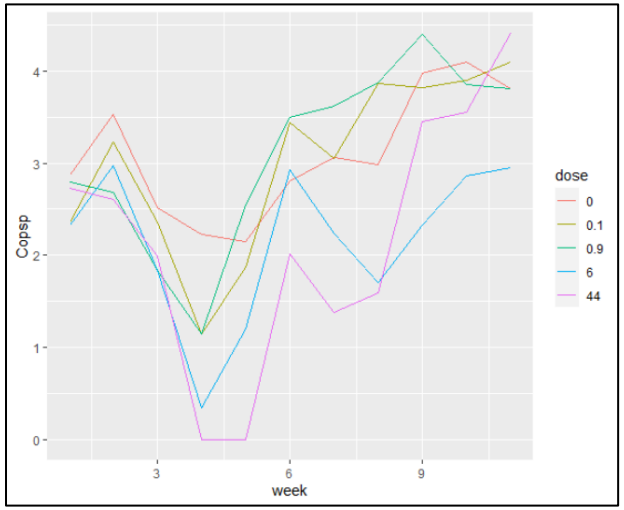
Principal Response Curves

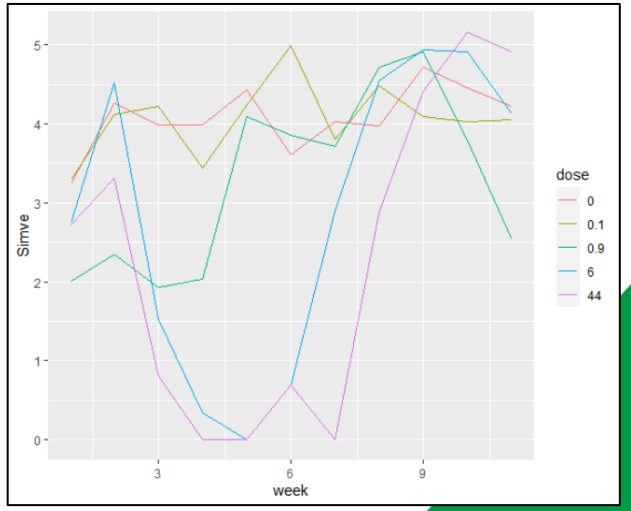
```
#individual abundance

pyrifos_combo <-cbind(pyrifos, dose, week)
pyrifos_combo$week <- as.integer(pyrifos_combo$week)

Slyla_m <-aggregate(Slyla~ dose*week, FUN=mean, data=pyrifos_combo)
Aloco_m <-aggregate(Aloco~ dose*week, FUN=mean, data=pyrifos_combo)
Copsp_m <-aggregate(Copsp~ dose*week, FUN=mean, data=pyrifos_combo)
Simve_m <-aggregate(Simve~ dose*week, FUN=mean, data=pyrifos_combo)

g1 <- ggplot()+geom_line(data=Slyla_m, aes(x=week, y=Slyla, col=dose))
g2 <- ggplot()+geom_line(data=Aloco_m, aes(x=week, y=Aloco, col=dose))
g3 <- ggplot()+geom_line(data=Copsp_m, aes(x=week, y=Copsp, col=dose))
g4 <- ggplot()+geom_line(data=Simve_m, aes(x=week, y=Simve, col=dose))

g1;g2;g3;g4
```

https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/prc









DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Assessment

qualtrics XM

UNIVERSITY OF NORTH DAKOTA UND

https://und.qualtrics.com/jfe/form/SV_bsc4eev0HZbxYua

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Caveats and Concerns 

- General
  - Need to interpret firehose of outputs
  - Simpler analyses may be better
- Canonical Correlation Analysis [7]
  - Assumption of normality for both sets of variables
  - Bad for small sample sizes
- Canonical Correspondence Analysis [8]
  - Overfitting is a danger
  - Interpretation depends on good explanatory variable selection
  - Measurable X-variables and Y-variable composition is usually desirable over unmeasurable factors
- Vector Autoregression [9]
  - Different numbers of lags can be set
- Principal Response Curves
  - Graphs can be hard to interpret with too many variables

[7] https://stats.idre.ucla.edu/r/dae/canonical-correlation-analysis/
[8] http://ordination.okstate.edu/overview.htm
[9] http://www.phdeconomics.sssup.it/documents/Lesson17.pdf

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

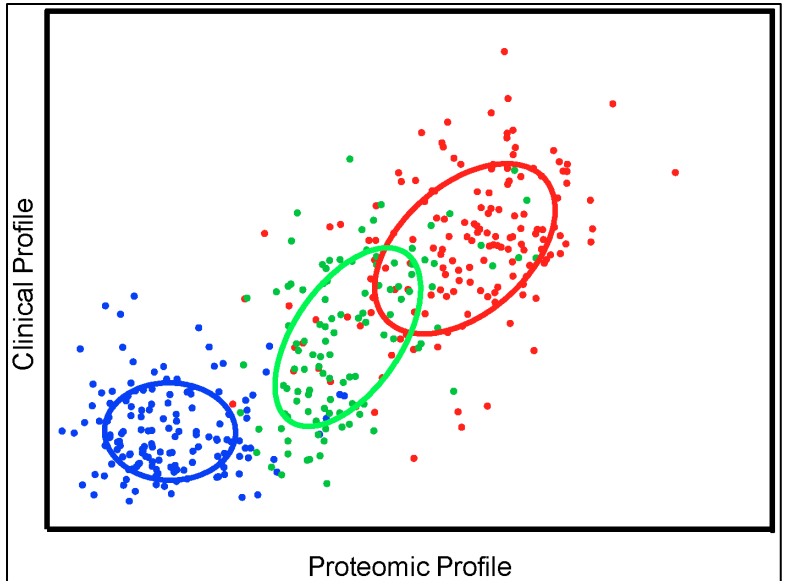
UND
UNIVERSITY OF NORTH DAKOTA.

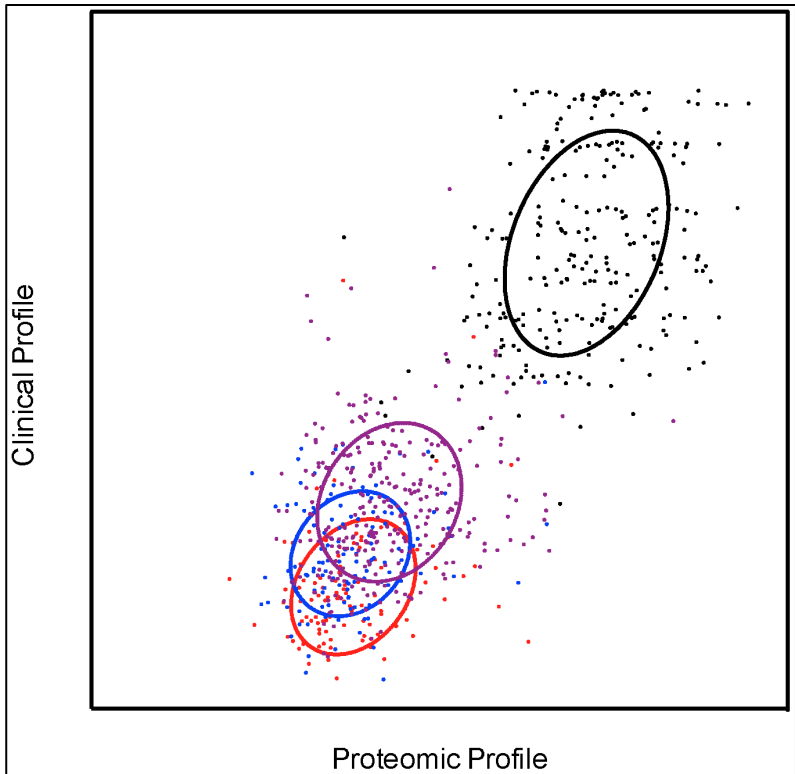
Real World Examples

Canonical Correlation Analysis

We propose an approach based on asymmetrical sparse canonical correlation analysis (SCCA) that finds multivariate correlations between the 'omics measurements and the complex clinical phenotypes. We correlated plasma proteomics data to multivariate overlapping complex clinical phenotypes from tuberculosis and malaria datasets. We discovered relevant 'omic biomarkers that have a high correlation to profiles of clinical measurements and are remarkably sparse, containing 1.5–3% of all 'omic variables.

Rousu, J., Agranoff, D. D., Sodeinde, O., Shawe-Taylor, J., & Fernandez-Reyes, D. (2013). Biomarker Discovery by Sparse Canonical Correlation Analysis of Complex Clinical Phenotypes of Tuberculosis and Malaria. PLoS Computational Biology, 9(4), e1003018, doi:10.1371/journal.pcbi.1003018.





Real World Examples

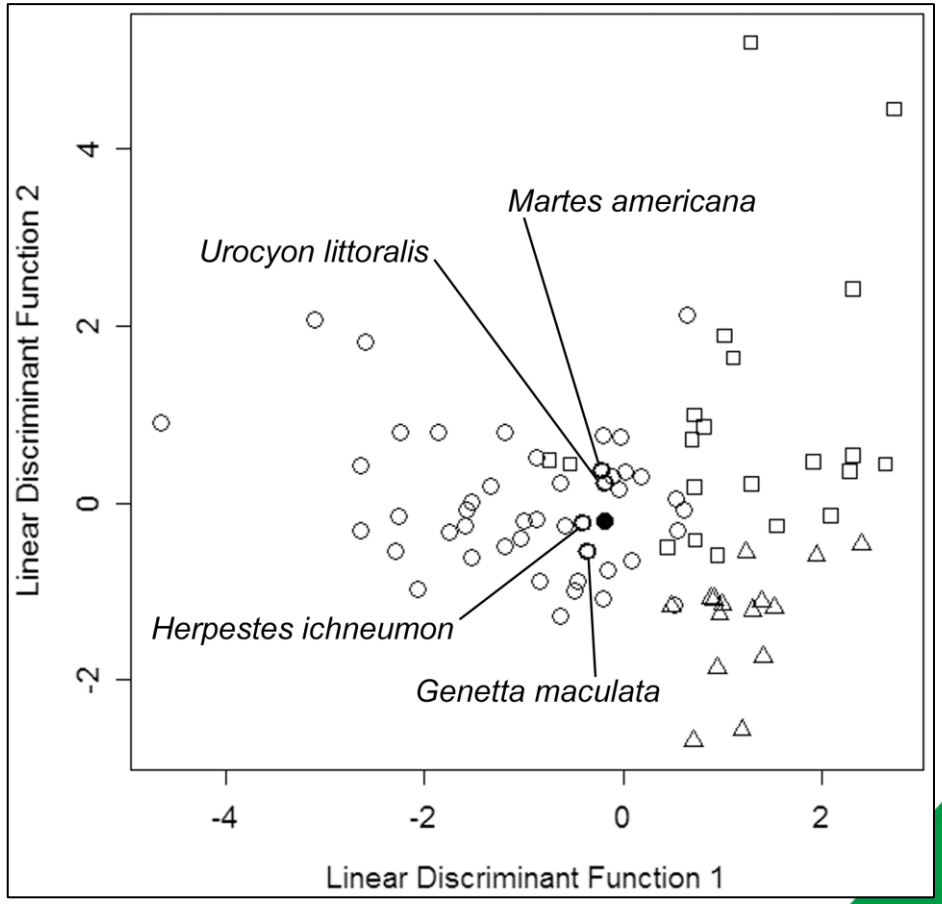
Linear Discriminant Analysis

Tomiya, S. (2011). A New Basal Caniform (Mammalia: Carnivora) from the Middle Eocene of North America and Remarks on the Phylogeny of Early Carnivorans. PloS One, 6(9), e24146, doi:10.1371/journal.pone.0024146.

A new carnivoramorphan from the middle Eocene of southern California, Lycophocyon hutchisoni, is described.

A discriminant analysis of the estimated body weight and dental ecomorphology predicted a mesocarnivorous diet for L. hutchisoni, and the postcranial morphology suggests a scansorial habit.

Six ecomorphological variables were used to maximally separate three dietary groups: carnivores (open circles), omnivores/hard-object feeders (open squares), and insectivores (open triangles). Data for 82 extant taxa are from Friscia et al. [71] and those for L. hutchisoni (filled circle) are based on holotype UCMP 85202. Four labeled taxa are the closest to L. hutchisoni in their posterior probabilities of dietary-group affiliations.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

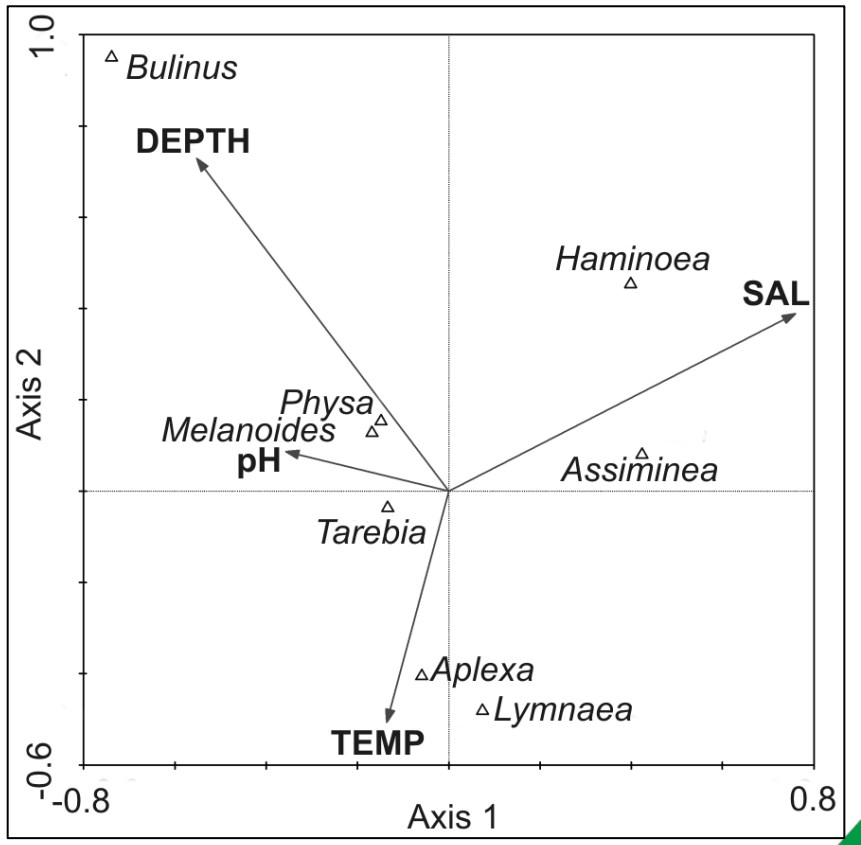
UND
UNIVERSITY OF NORTH DAKOTA.

Real World Examples

Canonical Correspondence Analysis

Miranda, N. A. F., Perissinotto, R., & Appleton, C. C. (2011). Population Structure of an Invasive Parthenogenetic Gastropod in Coastal Lakes and Estuaries of Northern KwaZulu-Natal, South Africa. PloS One, 6(8), e24337, doi:10.1371/journal.pone.0024337.

A canonical correspondence analysis was conducted between log-transformed gastropod (8 species, Fig. 3) abundance data and standardized environmental data (depth, salinity, temperature, dissolved oxygen, pH, median sediment particle size, turbidity, nitrates and phosphates) collected at the study site between 2007 and 2010. A forward selection model was used to determine the four environmental variables which best explained the variation in density data of gastropod species.



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA.

Real World Examples

Redundancy Analysis

Dong, C., Gao, M., Guo, C., Lin, L., Wu, D., & Zhang, L. (2017). The underlying processes of a soil mite metacommunity on a small scale. *PloS One, 12*(5), e0176828, doi:10.1371/journal.pone.0176828.

The relative importance of spatial (including trend variables, i.e., geographical coordinates, and broad- and fine-scale spatial variables) and environmental factors in driving the soil mite metacommunity was determined by variation partitioning. Mantel and partial Mantel tests and a redundancy analysis (RDA) were also used to identify the relative contributions of spatial and environmental variables.

Then, a redundancy analysis (RDA) was used to further assess which environmental variables (PCs) contributed most of the variation in the soil mite metacommunity composition.

Furthermore, the results of the RDA illustrated that the composition of the soil mite metacommunity in 2012 was influenced by PC1 and PC3 and in 2013 was influenced by PC1, PC3 and PC4 (Table 3).

Factor	2012		2013	
	R^2	*P*	R^2	*P*
PC1 [a]	0.14	<0.001***	0.11	<0.001***
PC2	0.01	0.61	0.01	0.43
PC3	0.06	0.03*	0.14	<0.001***
PC4	0.04	0.07	0.07	0.02*

[a] PC indicates each of the factors that were obtained from the PCA for each of the data sets.

* $P<0.05$.

*** $P<0.001$.

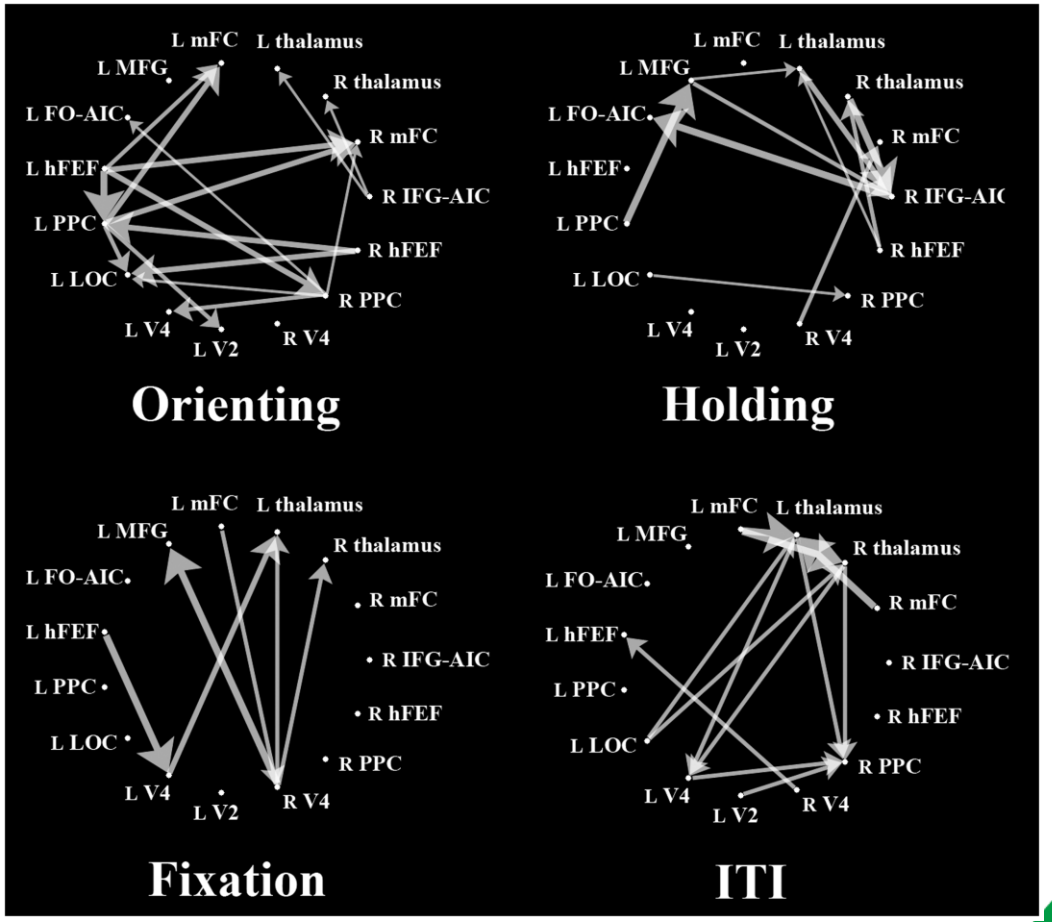
https://doi.org/10.1371/journal.pone.0176828.t003

Real World Examples

Vector Autoregression

Ozaki, T. J. (2011). Frontal-to-Parietal Top-Down Causal Streams along the Dorsal Attention Network Exclusively Mediate Voluntary Orienting of Attention. *PloS One, 6*(5), e20079, doi:10.1371/journal.pone.0020079.

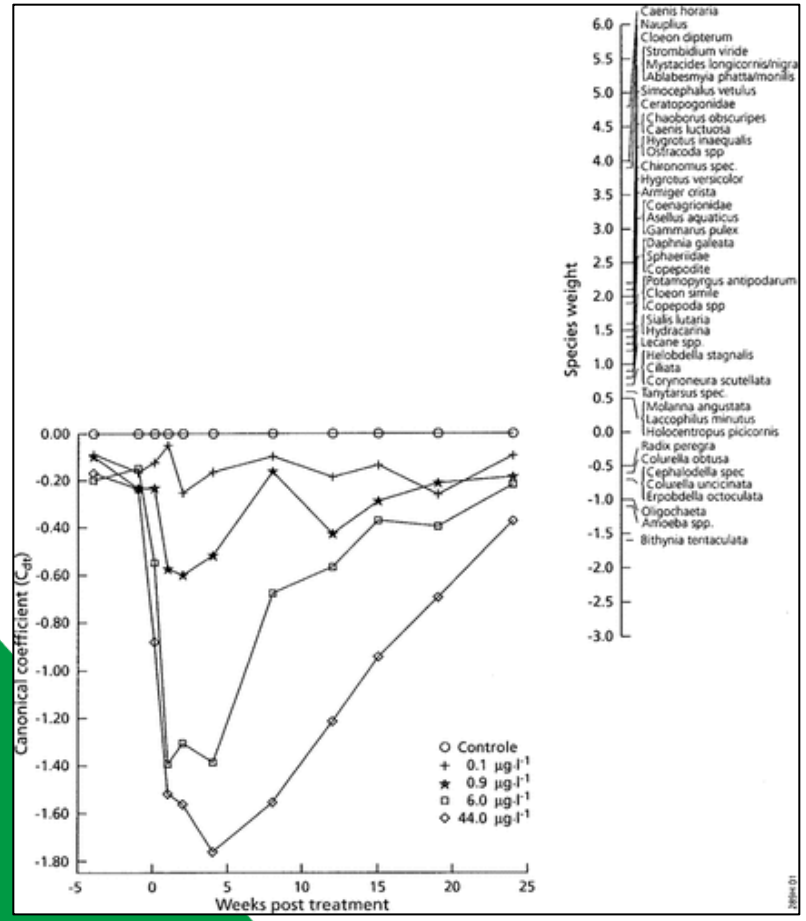
To evaluate causal flows between ROIs (regions of interest), we computed pGC (partial Granger causality) using Seth's Granger Causal Connectivity Analysis toolbox, based on multivariate vector autoregressive (MVAR) models including lags of multiple time-series [41], as described [42].

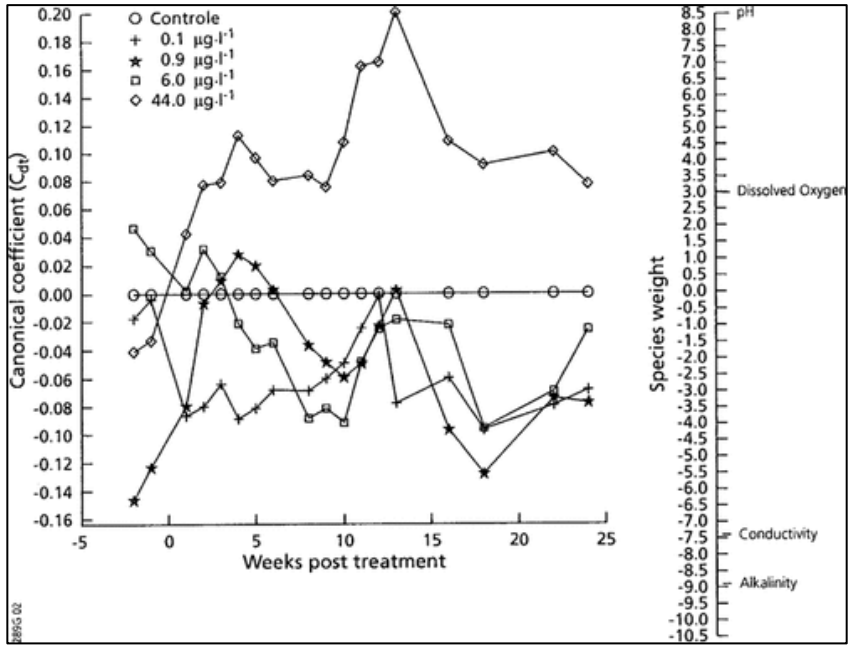


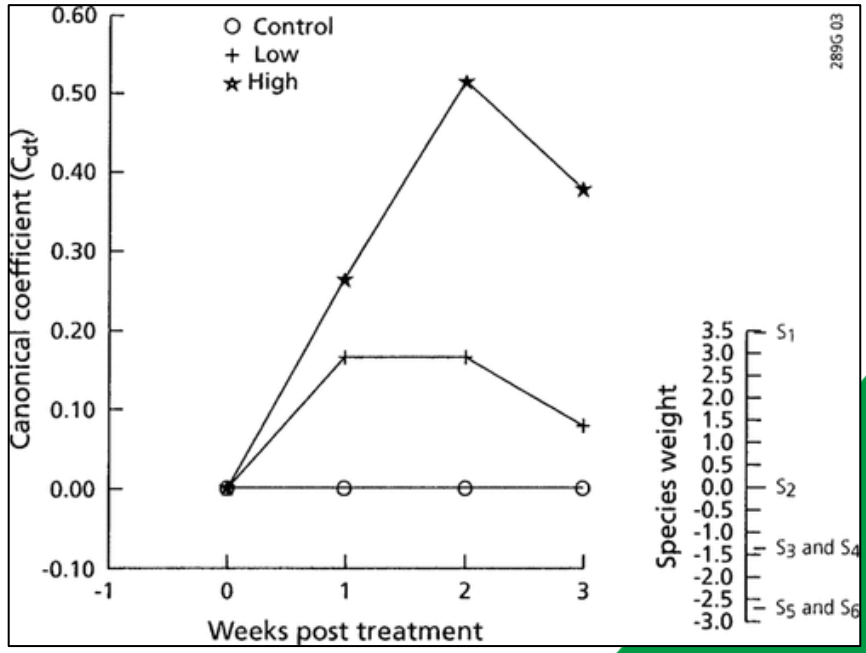
Real World Examples

Principal Response Curves

Van den Brink, P.J. and Braak, C.J.F.T. (1999), Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress. Environmental Toxicology and Chemistry, 18: 138-148. https://doi.org/10.1002/etc.5620180207



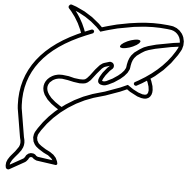




DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA.

Summary and Conclusion



- Canonical Analysis is the merger of ordination techniques and multiple regression
- Vector Autoregression is the extension of univariate regression and compares multiple time series
- Principal Response Curves is the extension of repeated measures and compares and is a special kind of redundancy analysis
- All are more advanced techniques that require certain types of data, careful construction, and concentrated interpretation

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF
NORTH DAKOTA.

Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY