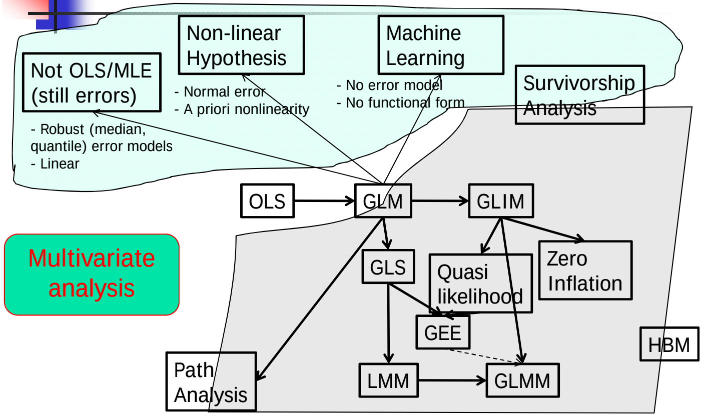
Multivariate statistics

1. Linear
2. Error in Y only
3. Errors normal
4. Errors independent
5. Errors homoscedastic

Road map



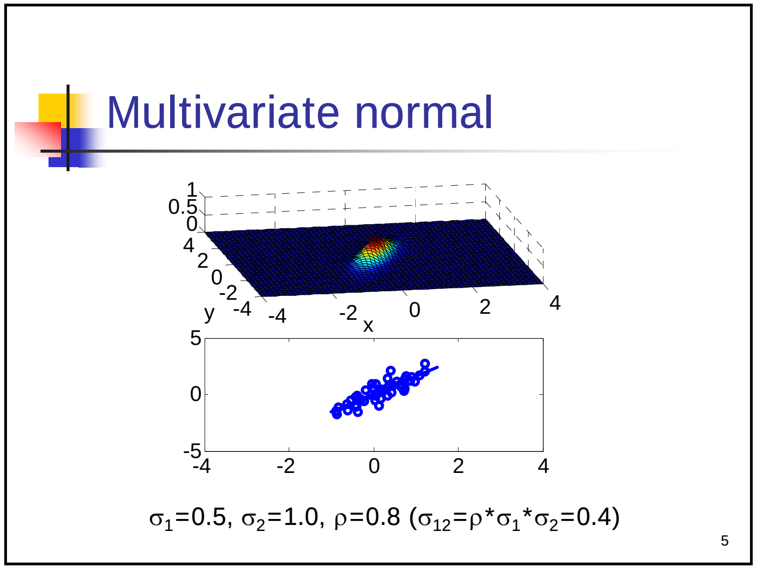
Multivariate statistics

- The statistical study of data that has many different variables (measurements)
  - All variables (except one in some cases) are continuous
  - GLM technically multivariate but usually excluded
  - Very heavy use of linear algebra

Leave behind

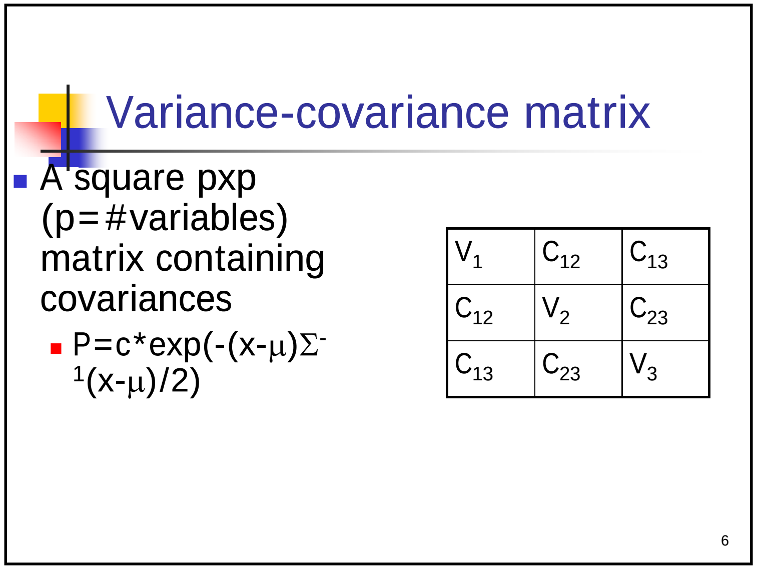
- Significance tests
- y=f(x) (modelling/prediction)

- Purely descriptive
  - Basically exploring covariance structure



Multivariate normal

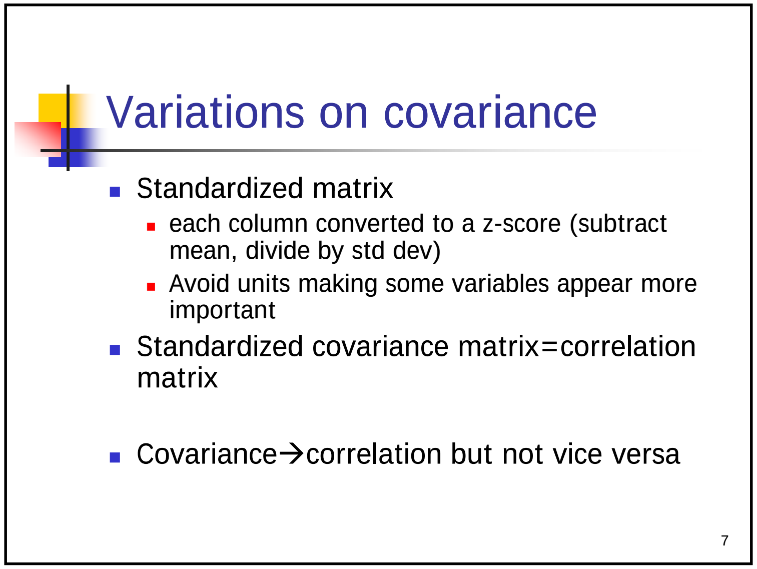
σ_1=0.5, σ_2=1.0, ρ=0.8 ($\sigma_{12}=\rho*\sigma_1*\sigma_2$=0.4)



Variance-covariance matrix

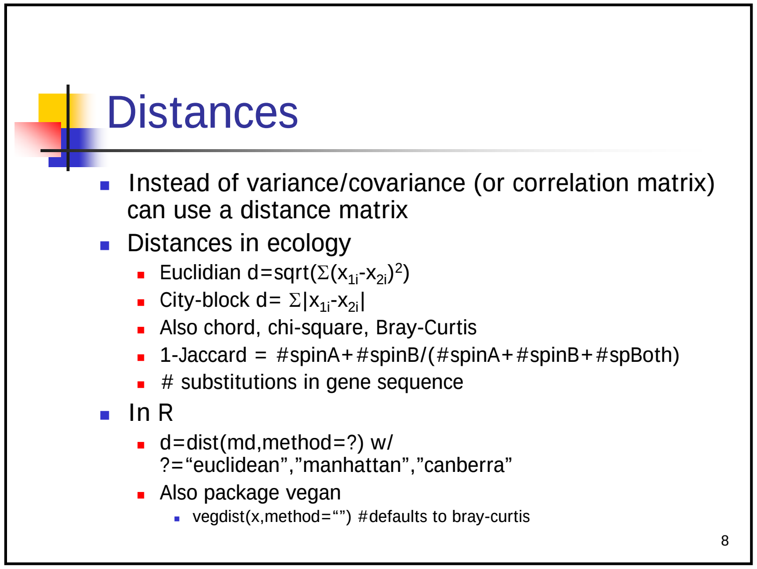
- A square pxp (p=#variables) matrix containing covariances
  - $P=c*exp(-(x-\mu)\Sigma^{-1}(x-\mu)/2)$

V_1	C_{12}	C_{13}
C_{12}	V_2	C_{23}
C_{13}	C_{23}	V_3



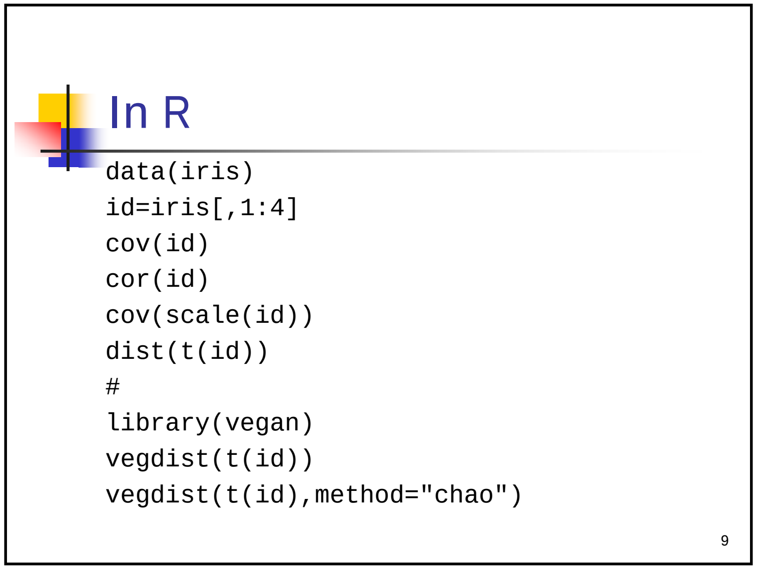
Variations on covariance

- Standardized matrix
  - each column converted to a z-score (subtract mean, divide by std dev)
  - Avoid units making some variables appear more important
- Standardized covariance matrix=correlation matrix
- Covariance→correlation but not vice versa



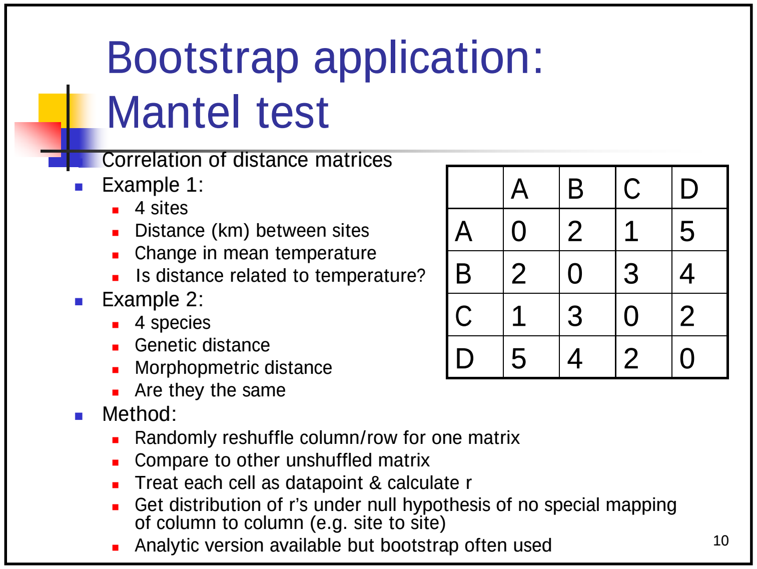
Distances

- Instead of variance/covariance (or correlation matrix) can use a distance matrix
- Distances in ecology
  - Euclidian d=sqrt($\Sigma(x_{1i}-x_{2i})^2$)
  - City-block d= $\Sigma|x_{1i}-x_{2i}|$
  - Also chord, chi-square, Bray-Curtis
  - 1-Jaccard = #spinA+#spinB/(#spinA+#spinB+#spBoth)
  - # substitutions in gene sequence
- In R
  - d=dist(md,method=?) w/ ?="euclidean","manhattan","canberra"
  - Also package vegan
    - vegdist(x,method="") #defaults to bray-curtis



In R

```
data(iris)
id=iris[,1:4]
cov(id)
cor(id)
cov(scale(id))
dist(t(id))
#
library(vegan)
vegdist(t(id))
vegdist(t(id),method="chao")
```

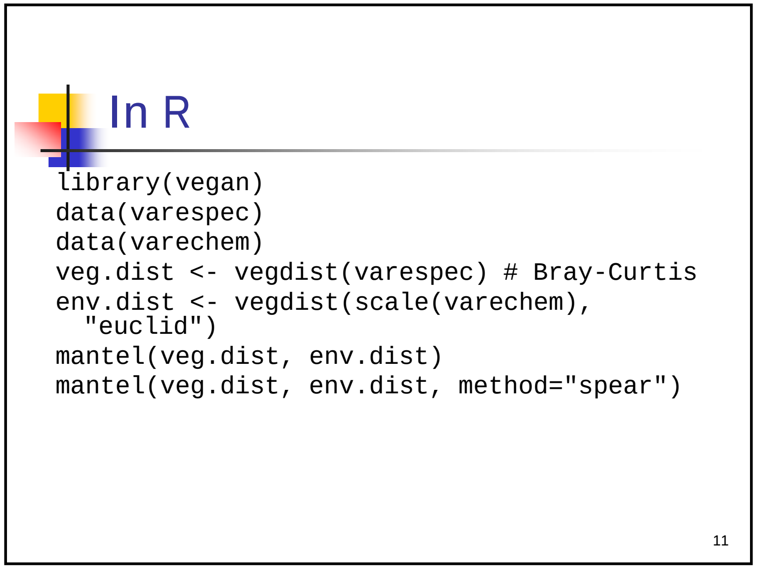


Bootstrap application: Mantel test

Correlation of distance matrices

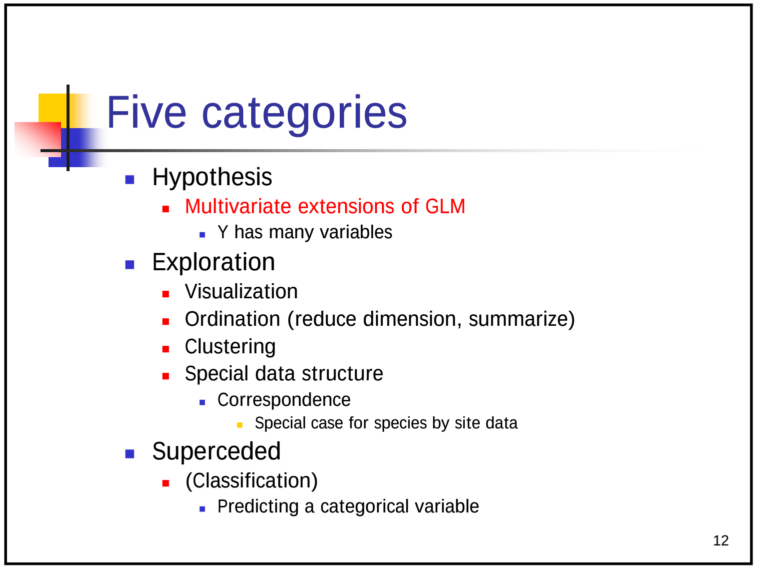
- Example 1:
  - 4 sites
  - Distance (km) between sites
  - Change in mean temperature
  - Is distance related to temperature?
- Example 2:
  - 4 species
  - Genetic distance
  - Morphopmetric distance
  - Are they the same
- Method:
  - Randomly reshuffle column/row for one matrix
  - Compare to other unshuffled matrix
  - Treat each cell as datapoint & calculate r
  - Get distribution of r's under null hypothesis of no special mapping of column to column (e.g. site to site)
  - Analytic version available but bootstrap often used

	A	B	C	D
A	0	2	1	5
B	2	0	3	4
C	1	3	0	2
D	5	4	2	0



In R

```
library(vegan)
data(varespec)
data(varechem)
veg.dist <- vegdist(varespec) # Bray-Curtis
env.dist <- vegdist(scale(varechem),
  "euclid")
mantel(veg.dist, env.dist)
mantel(veg.dist, env.dist, method="spear")
```



Five categories

- Hypothesis
  - Multivariate extensions of GLM
    - Y has many variables
- Exploration
  - Visualization
  - Ordination (reduce dimension, summarize)
  - Clustering
  - Special data structure
    - Correspondence
      - Special case for species by site data
- Superceded
  - (Classification)
    - Predicting a categorical variable

Sample multivariate data in R

```
data(iris)
names(iris)
sp=iris[,5]
sp
id=iris[,1:4]
id
#let=ifelse(sp=="setosa","s",ifelse(sp=="versicolor","
  v","g"))
let<-substr(iris$Species,1,2)
let
mns=aggregate(id,by=list(Species=sp),mean)
mns
mnsp=mns[,1]
mns=mns[,:2:5]
```

Multivariate extensions (hypothesis testing)

- Earlier example of correction for multiple tests (Bonferonni)
  - Two populations, 35 morphometric measures (e.g. corolla length)
  - Are the populations distinct (different means)
  - Previously 35 t-tests w/ correction
- Better way
  - Assume data in a 35-dimensional normal distribution
  - Use normal machinery
- Three analogues (for continuous y)
  - T-test (2 discrete)→Hotelling's T^2
  - ANOVA (n discrete)→MANOVA
  - Regression (n continuous)→RDA

In R

```
#MANOVA
data(iris)
m=manova(cbind(Sepal.Length,Sepal.Width,Petal.Wid
  th,Petal.Length)~Species,data=iris)
m
summary(m)
```

R output

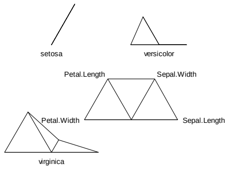
```
Call:
   manova(y ~ iris$Species)

Terms:
                iris$Species Residuals
Sepal.Length         63.2121   38.9562
Sepal.Width          11.3449   16.9620
Petal.Length        437.1028   27.2226
Petal.Width          80.4133    6.1566
Deg. of Freedom            2       147

Residual standard error: 0.5147894 0.3396877 0.4303345 0.2046500
Estimated effects may be unbalanced
> summary(m)
              Df Pillai approx F num Df den Df    Pr(>F)
iris$Species   2  1.192   53.466      8    290 < 2.2e-16 ***
Residuals    147
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Five categories

- Hypothesis
  - Multivariate extensions of GLM
    - Y has many variables
- Exploration
  - Visualization
  - Ordination (reduce dimension, summarize)
  - Clustering
  - Special data structure
    - Correspondence
      - Special case for species by site data
- Superceded
  - (Classification)
    - Predicting a categorical variable

Visualization (Exploration 1)

- Many ways to visual multivariate data
  - Human eye very good at finding patterns



```
library(MASS)
parcoord(id)
parcoord(mns)
stars(mns,full=F,key.loc=c(4,3),
  labels=mnsp)
```

Five categories

- Hypothesis
  - Multivariate extensions of GLM
    - Y has many variables
- Exploration
  - Visualization
  - Ordination (reduce dimension, summarize)
  - Clustering
  - Special data structure
    - Correspondence
      - Special case for species by site data
- Superceded
  - (Classification)
    - Predicting a categorical variable

Ordination (Exploration 2)

- A collapsing of dimensionality=simplification
  - E.g. have 20 variables, simplify to two (can plot)
- Many methods
  - Principle Component Analysis (PCA)
  - Principle Coordinate Analysis (PCoA)
  - Multidimensional Scaling (MDS)

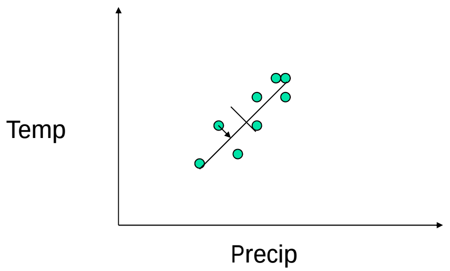
Two goals

- Reduce # of measurements
  - 20 measurements of body size
- Ordinate (order) the data
  - Species abundance data at 20 sites

Principal Component Analysis

- Identify the major axes of variation
  - Literally axis 1 is the axis of greatest varition
- Calculation: the eigenvectors/eigenvalues of the covariance matrix
  - Often use the standardized matrix
- Eigenvalues (λ) proportional to amount of variation explained
- Eigenvectors give directions
- Get coordinates of original data on new axes – can keep subset

Visual PCA



In R

```
data(iris)
id=iris[,1:4]
pc=princomp(log(id),cor=T) #log normalize, cor scales
pc
summary(pc)
plot(pc) #scree tells relative variance of axes
loadings(pc) #how variables map into axes
pcid=predict(pc) #get projected coordinates for plots
#classic plot
plot(pcid[,2]~pcid[,1],xlab="PC1",ylab="PC2")
#also library(MASS), eqscplot(x,y,…)
#classic w/ letters
plot(pcid[,2]~pcid[,1],xlab="PC1",ylab="PC2“,type=n)
text(pcid[,1:2],labels=let)
#biplot
biplot(pc,xlabs=let)
```

R output

```
pc = princomp(log(id), cor = T)
> pc
Call:
princomp(x = log(id), cor = T)
Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4 
1.7124583 0.9523797 0.3647029 0.1656840 
 4  variables and  150 observations.
> summary(pc)
Importance of components:
                          Comp.1    Comp.2     Comp.3    Comp.4
Standard deviation     1.7124583 0.9523797 0.36470294 0.1656840
Proportion of Variance 0.7331284 0.2267568 0.03325206 0.0068628
Cumulative Proportion  0.7331284 0.9598851 0.99313720 1.0000000
> loadings(pc)
Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.504 -0.455  0.709  0.191
Sepal.Width  -0.302 -0.889 -0.331       
Petal.Length  0.577        -0.219 -0.786
Petal.Width   0.567        -0.583  0.580

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
```

Summary of outputs

- Loadings – relative importance of different variables (x_i) in a PC axis
  - The arrows in a biplot
- Scores – position (coordinates) of an observation/row on PCA axes
  - The points in a biplot
- Variance – proportion of variance for axis is $\lambda_i/\Sigma\lambda_i$
  - Scree plot is just bars of λ_i - looks like talus slope
  - Cumulative or pareto gives cumulative variance

Another morphological

```
d<-read.table("sparrowda.txt",h=T)
d2<-d[d$observer==3 & d$Month==6,]
str(d2)
m<-princomp(d2[,2:7],cor=T)
biplot(m)
summary(m)
loadings(m)
```

And an ordination

```
#and a ordination
library(vegan)
data(varespec)
m<-
  princomp(varespec[,1:20],cor=T)
summary(m)
plot(m)
biplot(m)
loadings(m)
scores(m)
```

And in climate

- Take pressure at grid on globe at one point in time as one row (1 observation)
  - Time in rows, points on globe in columns
  - Do PCA
  - Get main axes of variability over time
  - → PNA, NAO, many other teleconnections

PCA & Collinearity

- If you have highly collinear data messing up a multivariate regression
- Do:
  - PCA (or PCoA)
  - Determine # of axes to keep
  - Interpret the axes
  - Do regression vs. the transformed coordinates

In R

```
d<-read.csv("dickcissel.csv",h=T)
str(d)
 m<-princomp(d[,3:15],cor=T)
summary(m)
biplot(m)
m2<-lm(d$abund~m$scores[,1]+
  m$scores[,2])
plot(d$abund~m$scores[,1])
```

Factor analysis

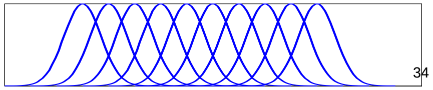
- Line is very blurry with PCA
- PCA has goal of finding axis of maximum varition
- Factor has goal of finding p<n factors that explain (predict) the multivariate cloud with error
  - Original derivation
    - $x = \Lambda f + \varepsilon$
    - $\Sigma = \mathrm{cov}(x) = \mathrm{cov}(\Lambda f + \varepsilon) = \mathrm{cov}(\Lambda f) + \mathrm{cov}(\varepsilon) = \Lambda \Lambda^T + \Psi$
    - Non-unique upto Qf where Q orthogonal
    - PCA is now alternative
  - IQ was the original motivation (test scores in different subjects all correlated – 1 underlying factor?)
- Big difference – Factor axes are not unique
  - Can rotate axes to make loadings high or 0 – i.e. improve interpretation

Factor analysis in R

```
#d<-
  read.table("sparrowda.txt",h=T)
#d2<-d[d$observer==3 &
  d$Month==6,]
#str(d2)
fact<-
  factanal(d2[,2:7],3,scores=c("reg
  ression"),rotation="varimax")
loadings(m)
```

Correspondence analysis

- Works on a chi-squre table (counts)
  - Classically species in columns, sites in rows
  - Not for morphology!
- Treats rows & columns equally (PCA & Factor do not)
  - Order sites based on species while ordering species based on sites found in
- Two different interpretations
  - Reciprocal averaging (RA)
  - Gaussian ordination
    - Find the unknown (abstract) axis such that species abundances vs this axis are:

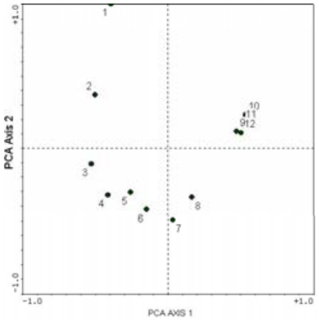


Correspondence Analysis (CA) in R)

```
#and a ordination
library(vegan)
data(varespec)
m<-cca(varespec) # no [,1:20]!!
summary(m)
plot(m)
```

Horseshoe effects

- Sometimes get a plot that looks like a horseshoe
- Most common in CA
- Two ends of gradient have middle plants with zeros
  - These zeros cause the ends to look similar
  - In fact one zero is due to too hot, other is too cold
- Can correct with detrending (e.g. DECORANA) but current recommendation is to not do this



PCA on distances instead of covariances

- Variety of methods analogous to PCA for distance matrix
  - Allows use of preferred distance measure (e.g use Jaccard when only have presence-absence)
- Many names
  - Principle Coordinate Analysis (PCoA) – just PCA with distances
    - PCA is a special case with Euclidean distance and k=# principal components
  - Metric Multidimensional Scaling
  - Classical Scaling
- Works by:
  - Center and scale distances, then apply PCA

In R

```
dat=varespec[,1:20]
mms<-cmdscale(vegdist(dat>0),k=2)
mms #just scores (coordinates)
library(MASS)
eqscplot(mms,type="n") #library(MASS)
text(mms,rownames(dat))
```

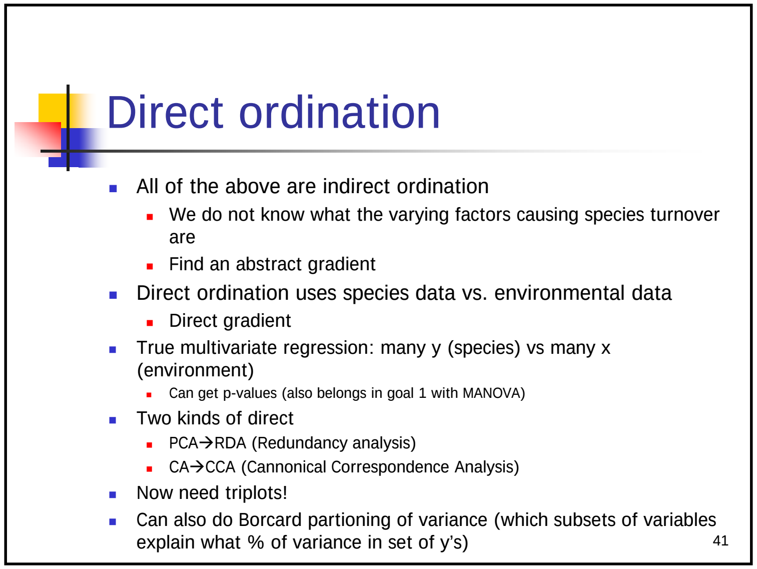
Non-metric MDS

- NMDS
  - Non-metric MDS – only rank order of distances
  - Can use when faith in distances is weak
- Works by:
  - Attempts to spread points out in 2 (+) dimensions to keep distances proportional
    - Requires squishing/stretching
  - Builds n-dimensional ball-stick model, tries to find rotation that requires least squishing to flatten
  - Stress=measure of how much squishing occurred

```
library(vegan)
m<-metaMDS(dat,distance="jaccard",k=2)
```

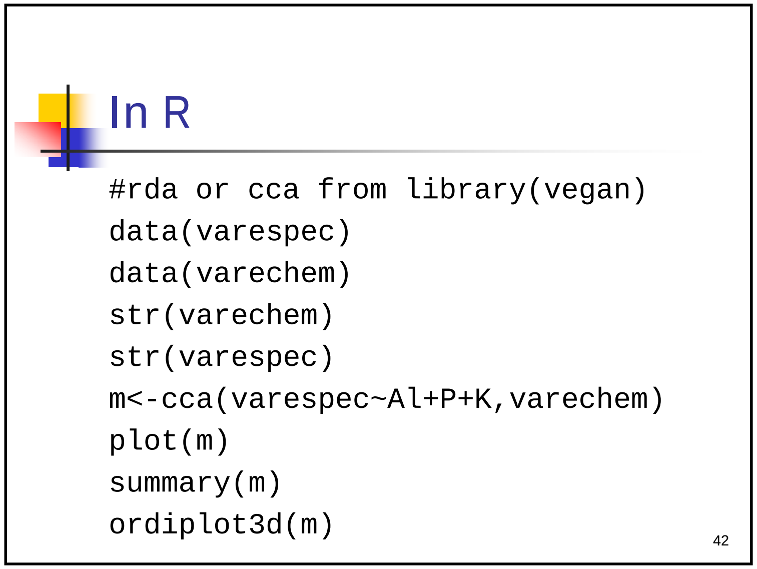
Which to use?

- If we do not have a special idea of distance
  - PCA if goal is description
  - PCA or factor if goal is explanation
- If we have a metric of distance
  - PCoA/MDS
  - NMDS
- If ordination (speciesXsite w/ abundances)
  - If gradient is small (all species on all gradient, linear responses)→PCA/PCoA
  - If gradient is large (species come in/out, Gaussian responses)→CA



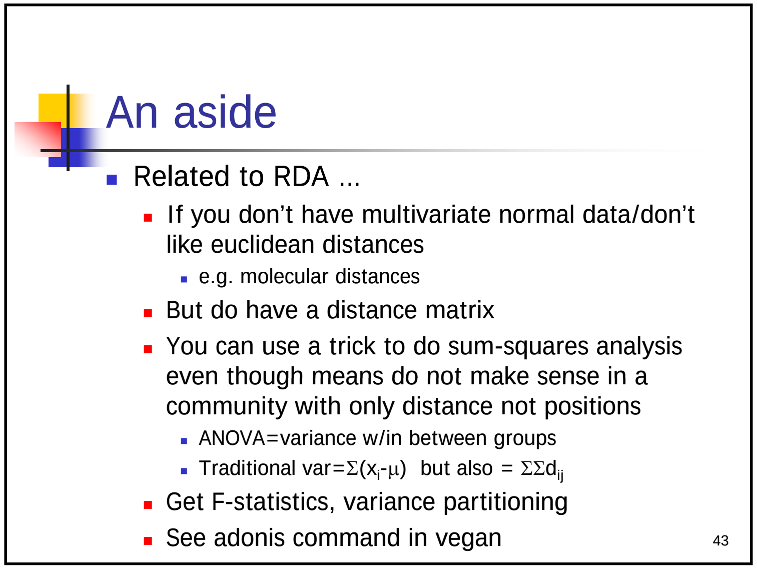
Direct ordination

- All of the above are indirect ordination
  - We do not know what the varying factors causing species turnover are
  - Find an abstract gradient
- Direct ordination uses species data vs. environmental data
  - Direct gradient
- True multivariate regression: many y (species) vs many x (environment)
  - Can get p-values (also belongs in goal 1 with MANOVA)
- Two kinds of direct
  - PCA→RDA (Redundancy analysis)
  - CA→CCA (Cannonical Correspondence Analysis)
- Now need triplots!
- Can also do Borcard partioning of variance (which subsets of variables explain what % of variance in set of y's)



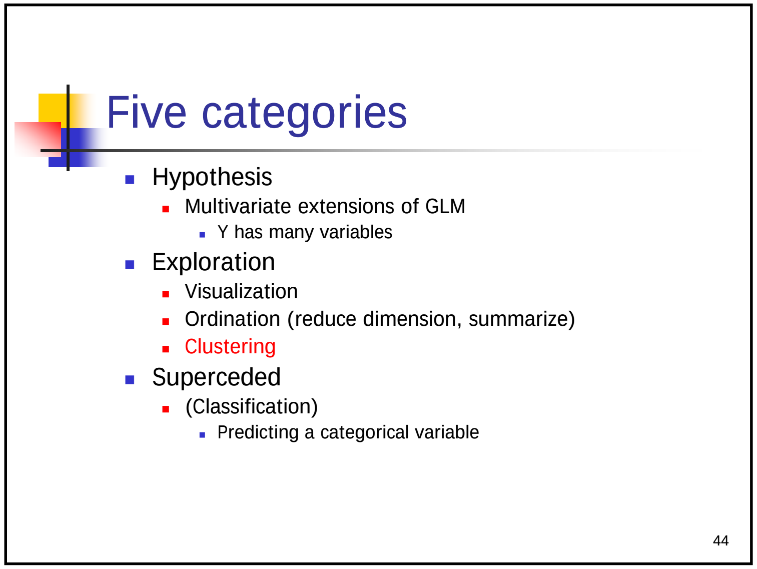
In R

```
#rda or cca from library(vegan)
data(varespec)
data(varechem)
str(varechem)
str(varespec)
m<-cca(varespec~Al+P+K,varechem)
plot(m)
summary(m)
ordiplot3d(m)
```



An aside

- Related to RDA …
  - If you don't have multivariate normal data/don't like euclidean distances
    - e.g. molecular distances
  - But do have a distance matrix
  - You can use a trick to do sum-squares analysis even though means do not make sense in a community with only distance not positions
    - ANOVA=variance w/in between groups
    - Traditional var=$\Sigma(x_i-\mu)$ but also = $\Sigma\Sigma d_{ij}$
  - Get F-statistics, variance partitioning
  - See adonis command in vegan



Five categories

- Hypothesis
  - Multivariate extensions of GLM
    - Y has many variables
- Exploration
  - Visualization
  - Ordination (reduce dimension, summarize)
  - Clustering
- Superceded
  - (Classification)
    - Predicting a categorical variable

Clustering (Exploration 3)

- Are there natural “clumps” or clusters in n-dimensional space?
- Takes a distance matrix as input
  - may be Euclidean or not
- Several types
  - Hierarchical (builds a tree of points based on similarity)
    - Aggregative (builds leafs up)
      - `In R: hclust, agnes, mclust`
    - Divisive (builds trunk down)
      - `In R: diana, mona`
  - Non-hierchical
    - K-means (user inputs K, uses gravity-like method)
      - `In R: kmeans, pam, clara, fanny`

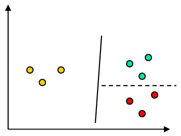
In R

```
cl=hclust(dist(id),method="single")
cl
summary(cl)
plot(cl)
cutree(cl,h=3)
cutree(cl,h=0.75)
cutree(cl,h=0.5)
cutree(cl,k=3)
cutree(cl,k=4)
library(cluster)
cl=diana(dist(id))
plot(cl)
cutree(cl,k=3)
plot(cl,lab=let,w=2,cex=0.8)
```

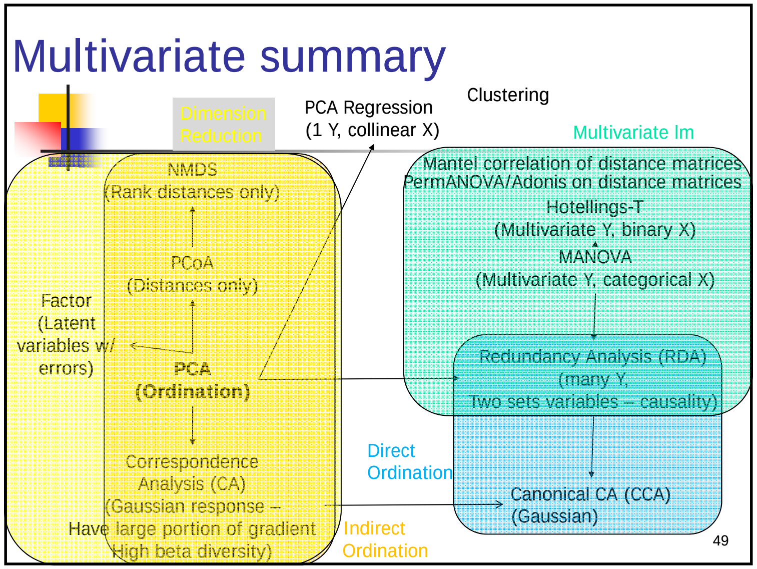
Five categories

- Hypothesis
  - Multivariate extensions of GLM
    - Y has many variables
- Exploration
  - Visualization
  - Ordination (reduce dimension, summarize)
  - Clustering
- Superceded
  - (Classification)
    - Predicting a categorical variable

(Classification) (Superceded)



- Linear discriminant analysis
  - Given a training set of data with a discrete y variable, and multivariate continuous x
  - Find hyperplanes (n-dimensional lines) that divide the groups
- Simple to calculate but now superceded by other techniques



Multivariate summary

Dimension Reduction

PCA Regression
(1 Y, collinear X)

Clustering

Multivariate lm

NMDS
(Rank distances only)

PCoA
(Distances only)

Factor
(Latent
variables w/
errors)

PCA
(Ordination)

Correspondence
Analysis (CA)
(Gaussian response –
Have large portion of gradient
High beta diversity)

Mantel correlation of distance matrices
PermANOVA/Adonis on distance matrices

Hotellings-T
(Multivariate Y, binary X)

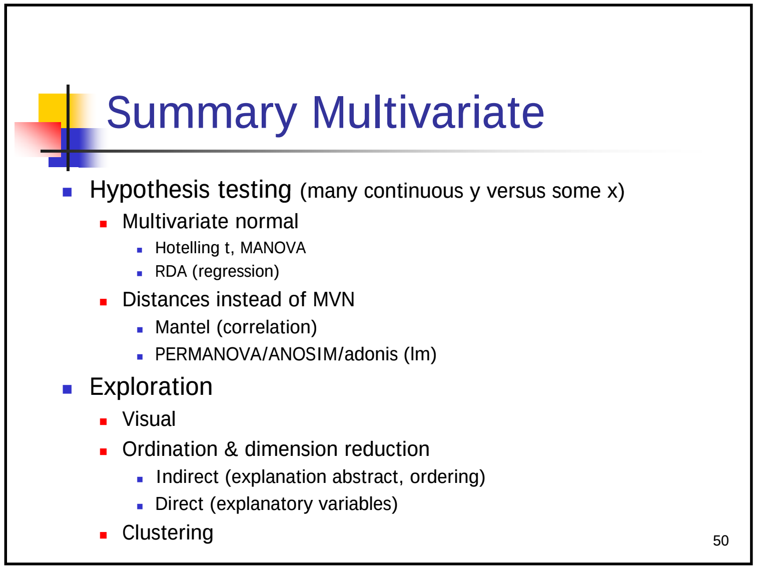
MANOVA
(Multivariate Y, categorical X)

Redundancy Analysis (RDA)
(many Y,
Two sets variables – causality)

Direct
Ordination

Canonical CA (CCA)
(Gaussian)

Indirect
Ordination



Summary Multivariate

- Hypothesis testing (many continuous y versus some x)
  - Multivariate normal
    - Hotelling t, MANOVA
    - RDA (regression)
  - Distances instead of MVN
    - Mantel (correlation)
    - PERMANOVA/ANOSIM/adonis (lm)
- Exploration
  - Visual
  - Ordination & dimension reduction
    - Indirect (explanation abstract, ordering)
    - Direct (explanatory variables)
  - Clustering