

A (brief) introduction to ordination and the vegan package

Eduard Szöcs

Institute for Environmental Sciences - University of Koblenz-Landau



SEFS9, July 5th 2015

Datasets
oooooooo

Indirect Gradient Analysis
oooooooooooooooooooo

Direct Gradient Analysis
oooo

Permutation Tests
o

End

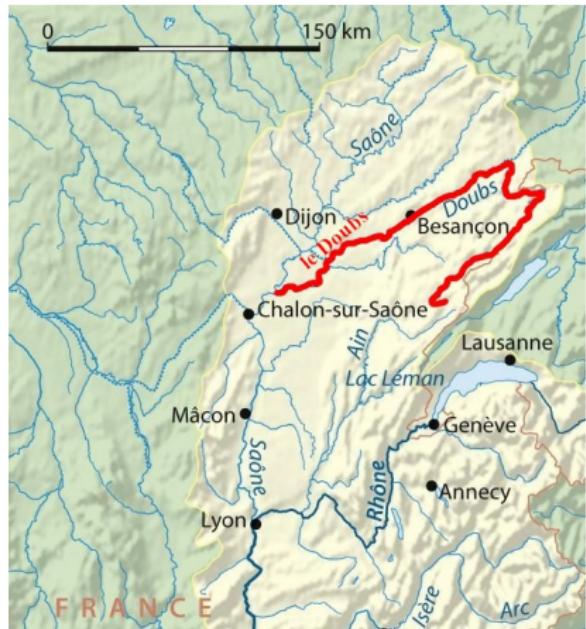
	Raw data	Transformed data	Unimodal	Distance-based	Model-based
Unconstrained	PCA		CA, DCA	PCoA, NMDS	MM, LVM
Constrained	RDA	tb-PCA tb-RDA	CCA	db-RDA	CAO, CQO
Other				Permanova, Dispersion	manyglm

(Nearly) no maths today ;)

Datasets

Demonstration: Doubs river fish communities

5 / 42



- ▶ Fish
- ▶ 30 sites along the Doubs River

Verneaux, J. (1973) Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. These d'état, Besançon. 1–257.

Datasets



Indirect Gradient Analysis



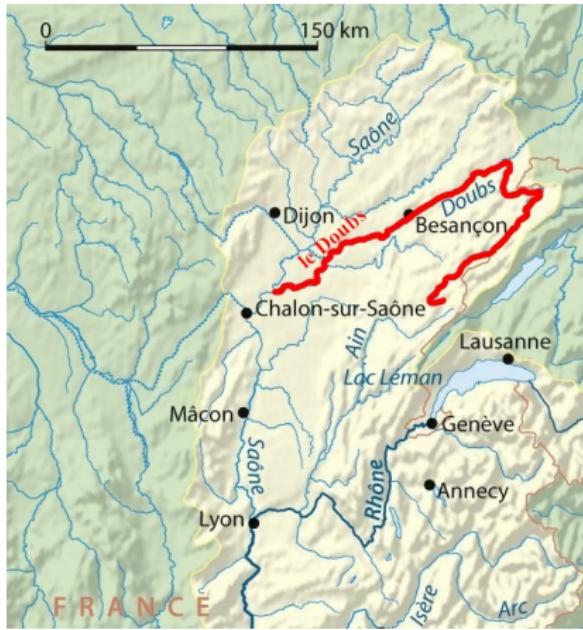
Direct Gradient Analysis



Permutation Tests



End



- ▶ Fish
 - ▶ 30 sites along the Doubs River

Questions

- ▶ How does fish composition change downstream?
 - ▶ Environmental drivers?

Verneaux, J. (1973) Cours d'eau de Franche-Comte (Massif du Jura). Recherches ecologiques sur le reseau hydrographique du Doubs. Essai de biotypologie. These d'etat, Besancon. 1-257.

```
Dabu <- read.table('doubtsAbu.csv', sep = '|', header = TRUE)
Denv <- read.table('doubtsEnv.csv', sep = '|', header = TRUE)
Dspa <- read.table('doubtsSpa.csv', sep = '|', header = TRUE)
```

dim(Dabu)

[1] 30 27

30 sites, 27 taxa

```
head(Dabu[, 1:18])
```

	CHA	TRU	VAI	LOC	OMB	BLA	HOT	TOX	VAN	CHE	BAR	SPI	GOU	BRO	PER	BOU	PSO	ROT
1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	5	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	5	5	5	0	0	0	0	0	0	0	0	0	1	0	0	0	
4	0	4	5	5	0	0	0	0	0	1	0	0	1	2	2	0	0	
5	0	2	3	2	0	0	0	0	5	2	0	0	2	4	4	0	0	
6	0	3	4	5	0	0	0	0	1	2	0	0	1	1	1	0	0	

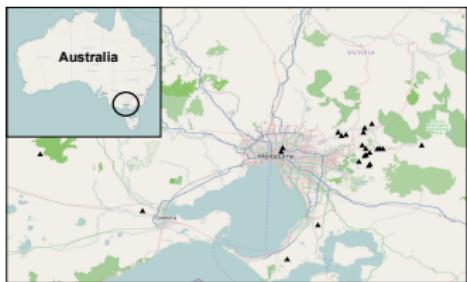
```
# Dimension and first rows of Environmental data
dim(Denv)

[1] 30 11
```

30 sites, 11 variables

```
head(Denv)

  das alt pen deb pH dur pho nit amm oxy dbo
1 0.3 934 48.0 0.84 7.9 45 0.01 0.20 0.00 12.2 2.7
2 2.2 932 3.0 1.00 8.0 40 0.02 0.20 0.10 10.3 1.9
3 10.2 914 3.7 1.80 8.3 52 0.05 0.22 0.05 10.5 3.5
4 18.5 854 3.2 2.53 8.0 72 0.10 0.21 0.00 11.0 1.3
5 21.5 849 2.3 2.64 8.1 84 0.38 0.52 0.20 8.0 6.2
6 32.4 846 3.2 2.86 7.9 60 0.20 0.15 0.00 10.2 5.3
```



- ▶ Macroinvertebrates
 - ▶ 24 sites
 - ▶ covering a salinity and toxicity gradient

The dataset is published in: Szöcs, E., Kefford, B.J., Schäfer, R.B., 2012. Is there an interaction of the effects of salinity and pesticides on the community structure of macroinvertebrates? *Science of the Total Environment* 437, 121–126.



- ▶ Macroinvertebrates
 - ▶ 24 sites
 - ▶ covering a salinity and toxicity gradient

Questions:

- ▶ Interaction between salinization and pesticides?
 - ▶ Which species are affected?
 - ▶ Other influences?

The dataset is published in: Szöcs, E., Kefford, B.J., Schäfer, R.B., 2012. Is there an interaction of the effects of salinity and pesticides on the community structure of macroinvertebrates? *Science of the Total Environment* 437, 121–126.

Exercise: Salinization and Pesticides

```
# setwd('3-Ordination/data/')
abu <- read.table('melbourneAbu.csv', sep = ';', header = TRUE)
env <- read.table('melbourneEnv.csv', sep = ';', header = TRUE)
```

```
# dimensions of data.frame  
dim(env)
```

[1] 24 23

dim(abu)

[1] 24 76

24 sites, 22 environmental variables, 75 taxa

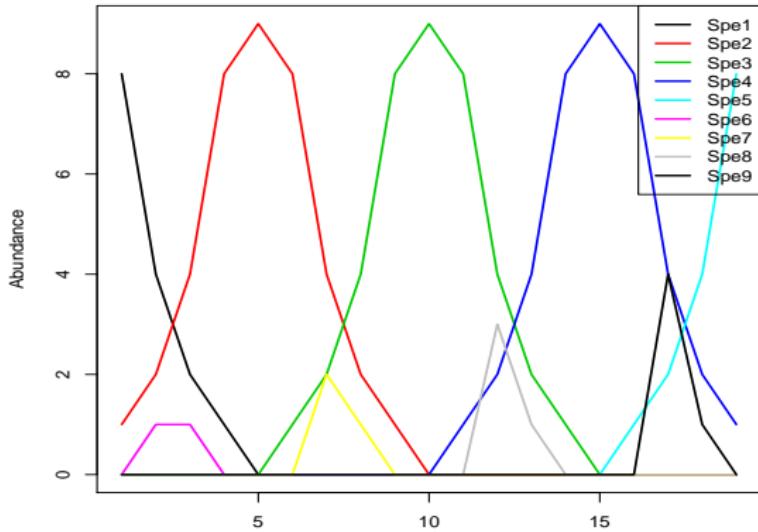
```
head(env[ , 1:10])
```

ID	T	pH	oxygen	Depth	maxwidth	minwidth	rifperc	poolperc	Bedrock	
1	1-11	16.8	7.67	80.1	0.9	15	12.0	0	100	0
2	2-11	16.5	7.29	83.0	0.9	30	15.0	0	100	0
3	3-11	17.3	7.20	77.9	0.4	4	2.5	0	100	0
4	4-11	15.6	7.84	72.0	0.7	8	2.5	0	100	0
5	5-11	17.2	6.97	69.9	0.9	7	4.0	0	100	0
6	6-11	15.5	7.26	80.0	0.2	3	2.0	5	95	0

Exercise: Dummy abundances

10 / 42

```
# Load dummy data
dummy <- read.table('dummydata.csv', header = TRUE, sep = ';')
# plot dummy data
matplot(dummy[ , -1], type = 'l', xlab = 'Site', ylab = 'Abundance',
        lty = 'solid', lwd = 2, col = 1:9)
legend('topright', legend = colnames(dummy)[-1],
       col = 1:9, lty = 'solid', lwd = 2)
```



Indirect Gradient Analysis

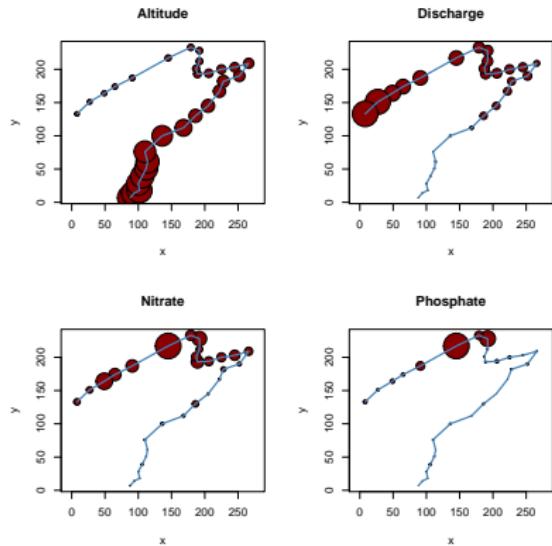
Principal Components Analysis (PCA)

Principal coordinates analysis (PCoA)

Nonmetric Multidimensional Scaling (NMDS)

Principal Components Analysis (PCA) — Why?

12 / 42

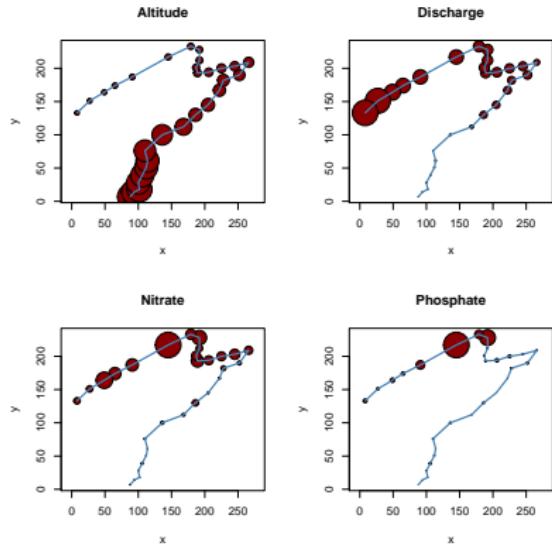


- ▶ 11 variables

Questions:

- ▶ Which variables are correlated?
- ▶ Which sites have similar conditions?
- ▶ How do conditions change downstream?

Solutions?



- ▶ 11 variables

Questions:

- ▶ Which variables are correlated?
- ▶ Which sites have similar conditions?
- ▶ How do conditions change downstream?

Solutions?

- ▶ pairwise comparisons
- ▶ 3D possible
- ▶ more than 3 dimensions?

- ▶ *"Look from another angle on the data"*
- ▶ PCA is just a rotation of the coordinate system
- ▶ The rotation is done so that the first axis contains as much variation as possible
- ▶ Second axis than most of remaining variation

Short Demo.

- ▶ *"Look from another angle on the data"*
- ▶ PCA is just a rotation of the coordinate system
- ▶ The rotation is done so that the first axis contains as much variation as possible
- ▶ Second axis than most of remaining variation

Short Demo.

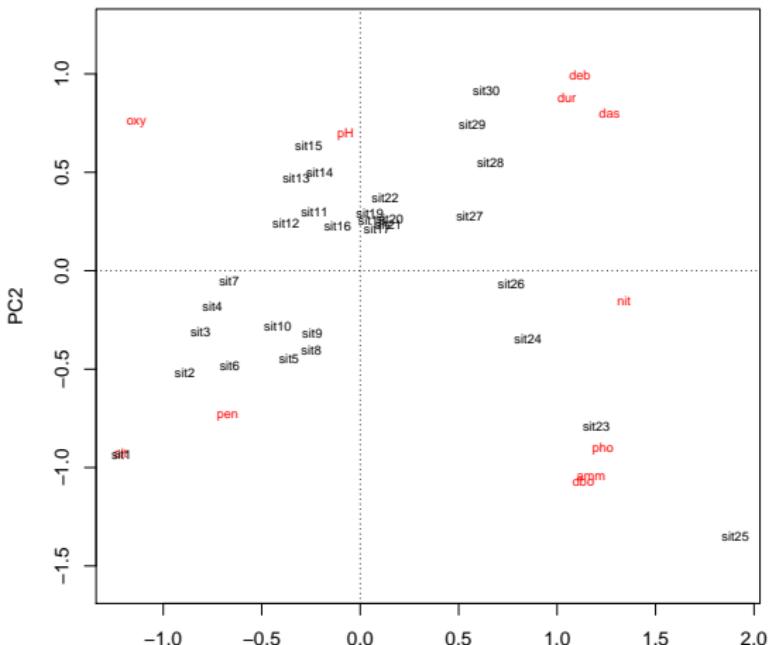
Maths:

- ▶ The covariance (or correlation) matrix is decomposed into its Eigenvectors and Eigenvalues.
- ▶ The Eigenvectors give the rotation needed
- ▶ The Eigenvalues stretch the axes

Principal Components Analysis (PCA) — How?

14 / 42

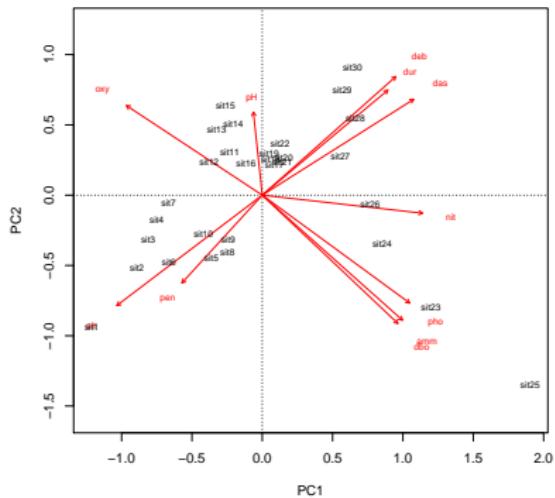
```
require(vegan)  
PCA <- rda(Denv, scale = TRUE)  
plot(PCA, scaling = 3)
```



Principal Components Analysis (PCA) — Interpretation? (I)

15 / 42

```
biplot(PCA, cex = 5, scaling = 3)
```

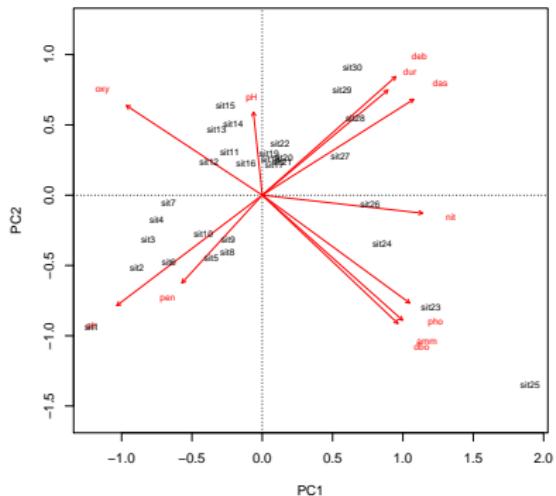


- ▶ angle between variables **approx.** their correlation
- ▶ distance between sites **approx.** their euclidean distance
- ▶ projecting a site on a variable **approx.** the relative value

Principal Components Analysis (PCA) — Interpretation? (I)

15 / 42

```
biplot(PCA, cex = 5, scaling = 3)
```

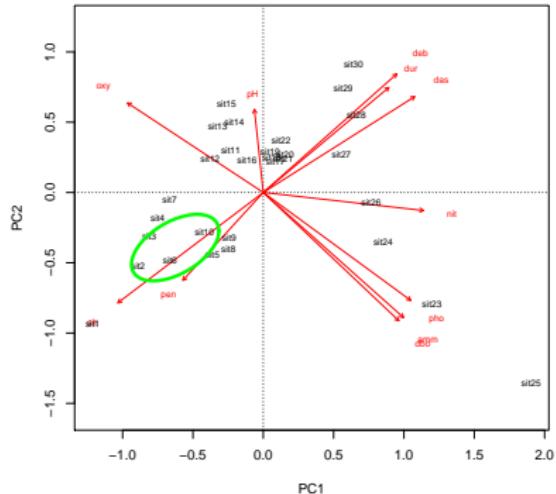


- ▶ angle between variables **approx.** their correlation
 - ▶ distance between sites **approx.** their euclidean distance
 - ▶ projecting a site on a variable **approx.** the relative value
-
- ▶ scaling = 1 - to interpret (only) distances between sites
 - ▶ scaling = 2 - to interpret (only) correlations between variables

Principal Components Analysis (PCA) — Interpretation? (II)

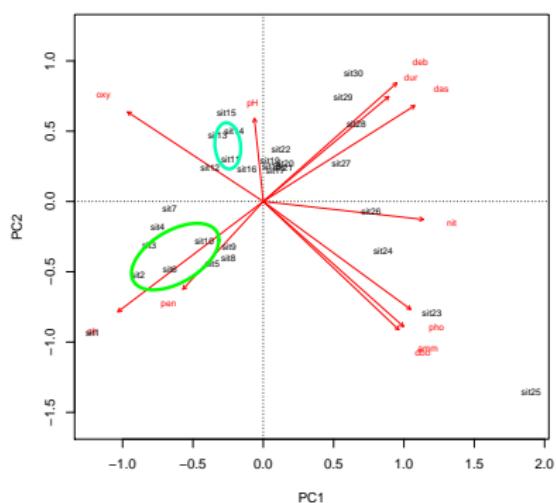
16 / 42

- ▶ high altitude + slope, low discharge



Principal Components Analysis (PCA) — Interpretation? (II)

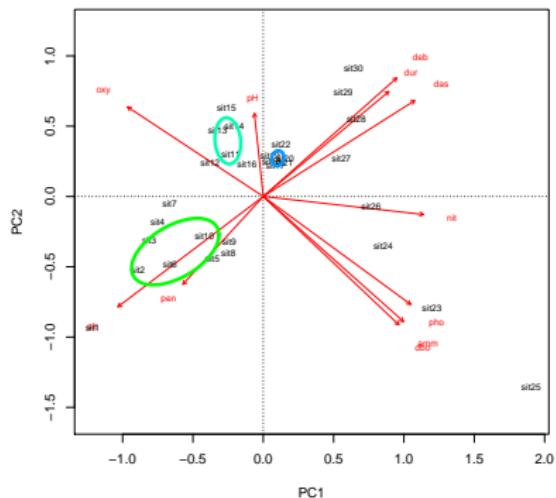
16 / 42



- ▶ high altitude + slope, low discharge
- ▶ high oxygen, low nutrient

Principal Components Analysis (PCA) — Interpretation? (II)

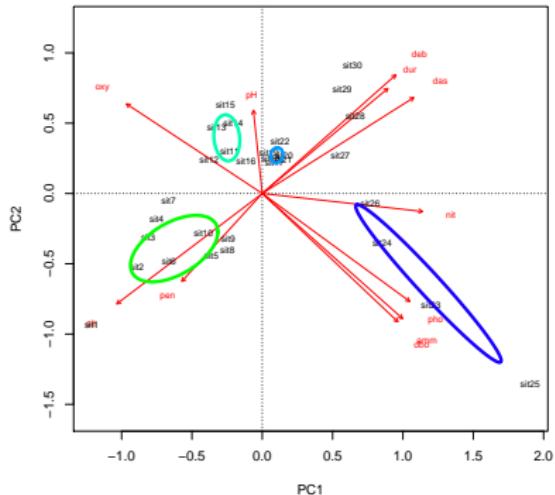
16 / 42



- ▶ high altitude + slope, low discharge
- ▶ high oxygen, low nutrient
- ▶ intermediate

Principal Components Analysis (PCA) — Interpretation? (II)

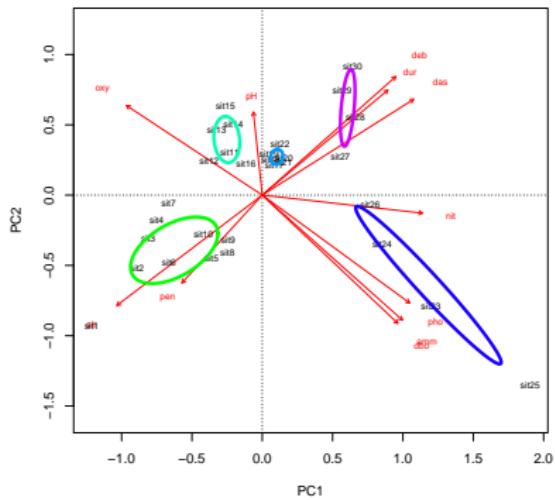
16 / 42



- ▶ high altitude + slope, low discharge
- ▶ high oxygen, low nutrient
- ▶ intermediate
- ▶ nutrient rich

Principal Components Analysis (PCA) — Interpretation? (II)

16 / 42



- ▶ high altitude + slope, low discharge
- ▶ high oxygen, low nutrient
- ▶ intermediate
- ▶ nutrient rich
- ▶ high discharge, low altitude, medium nutrient

Principal Components Analysis (PCA) — Interpretation? (III)

```
summary(PCA, display = NULL, scaling = 3)
```

Call:

```
rda(X = Denv, scale = TRUE)
```

Partitioning of correlations:

	Inertia	Proportion
Total	11	1
Unconstrained	11	1

Eigenvalues, and their contribution to the correlations

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Eigenvalue	5.9687	2.1638	1.06516	0.73873	0.40027	0.33565	0.1727
Proportion Explained	0.5426	0.1967	0.09683	0.06716	0.03639	0.03051	0.0157
Cumulative Proportion	0.5426	0.7393	0.83616	0.90331	0.93970	0.97022	0.9859
	PC8	PC9	PC10	PC11			
Eigenvalue	0.10821	0.02368	0.01707	0.005993			
Proportion Explained	0.00984	0.00215	0.00155	0.000540			
Cumulative Proportion	0.99575	0.99790	0.99946	1.000000			

Scaling 3 for species and site scores

- * Both sites and species are scaled proportional to eigenvalues on all dimensions
- * General scaling constant of scores:

Your turn!

Load the Melbourne dataset (only environmental variables).

Exclude the variables ID, logCond and logmaxTU.

Perform a PCA.

Which variables are correlated?

How much variance is explained by the first 2 axes?

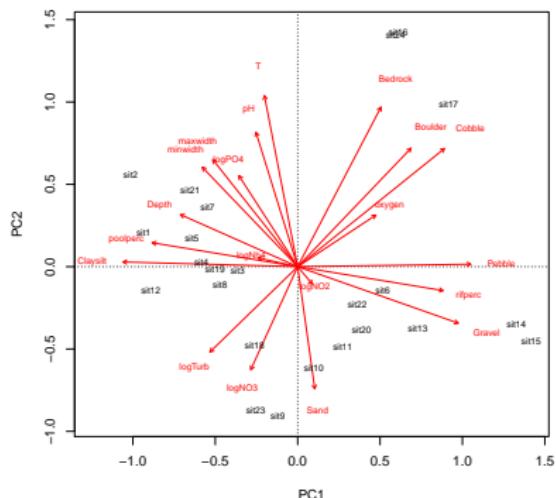
How could the two PCA axes be interpreted?

Exercise

```
take <- env[ , !names(env) %in% c('ID', 'logCond', 'logmaxTU')]
PCA <- rda(take, scale = TRUE)
cumsum(PCA$CA$eig / PCA$tot.chi)[1:2]
```

PC1	PC2
0.2839873	0.4537452

```
biplot(PCA, scaling = 3)
```



- ▶ multiple variables interrelated
- ▶ 1st axis can be interpreted as *hydrological gradient*
- ▶ 2nd axis can be interpreted as *chemistry gradient*

Question:

- ▶ How is diversity related to salinity, pesticides and other variables?

Question:

- ▶ How is diversity related to salinity, pesticides and other variables?

Problem:

- ▶ Only 24 sites
 - ▶ but 22 (potentially correlated) explanatory variables
 - ▶ strong hypotheses about salinity and pesticides

A Solution:

Question:

- ▶ How is diversity related to salinity, pesticides and other variables?

Problem:

- ▶ Only 24 sites
- ▶ but 22 (potentially correlated) explanatory variables
- ▶ strong hypotheses about salinity and pesticides

A Solution:

- ▶ Reduce number of variables to *Principal Components*
- ▶ regress these

Excursus — principal component regression (PCR)

```
# calculate shannon diversity index
div <- diversity(abu[, -1], index = 'shannon')
pc <- scores(PCA, choices = c(1, 2), scaling = 1, display = 'sites')
model_data <- data.frame(div, pc, logCond = env$logCond, logmaxTU = env$logmaxTU)
model <- lm(div ~ PC1 + PC2 + logCond + logmaxTU, data = model_data)
summary(model)
```

Call:

```
lm(formula = div ~ PC1 + PC2 + logCond + logmaxTU, data = model_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.64415	-0.15688	0.02063	0.18219	0.57929

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.83079	0.43429	4.216	0.000468 ***
PC1	0.01971	0.16691	0.118	0.907262
PC2	0.02192	0.19570	0.112	0.911996
logCond	-0.20942	0.13050	-1.605	0.125049
logmaxTU	-0.12572	0.07316	-1.718	0.101994

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

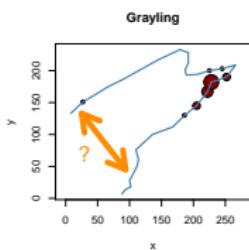
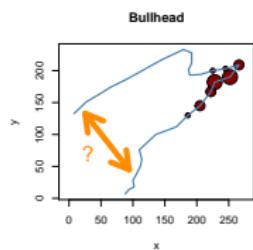
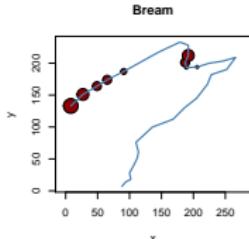
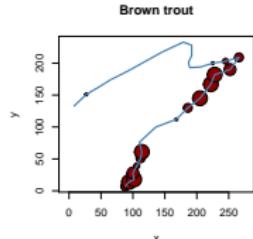
Residual standard error: 0.3645 on 19 degrees of freedom

Multiple R-squared: 0.2682, Adjusted R-squared: 0.1141

F-statistic: 1.741 on 4 and 19 DF, p-value: 0.1827

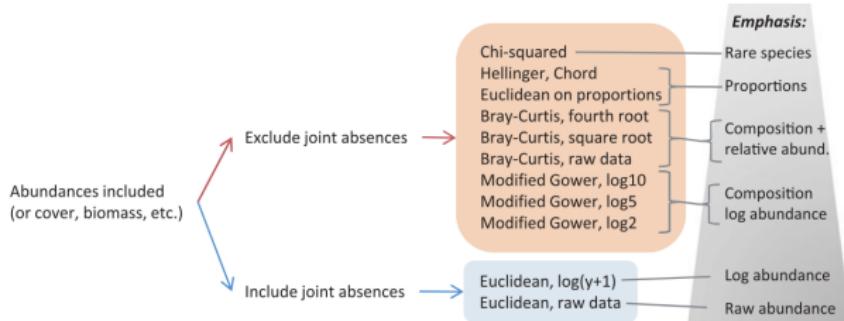
Abundances — The Problem with zeros

22 / 42



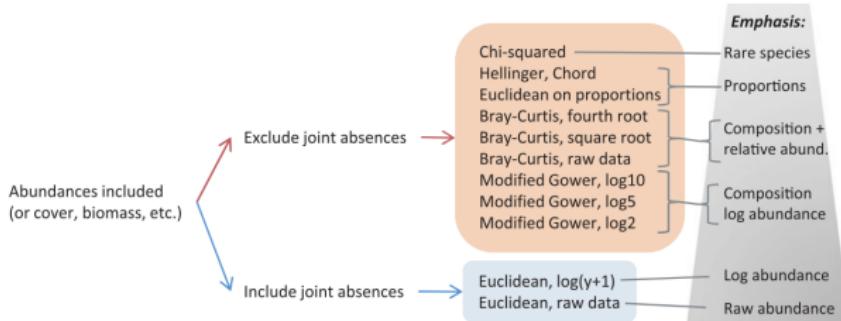
- ▶ Species may be absent due to different factors (too high flow, too saline, etc.)
- ▶ *Absence* contains less information than *Presence*
- ▶ PCA preserves the euclidean distance between sites
- ▶ Need another measure of similarity for (raw) abundances

Dissimilarity measures



from: Anderson, M.J., Crist, T.O., et al. , 2011. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. *Ecology Letters* 14, 19–28.

Dissimilarity measures



	Spe1	Spe2	Spe3
sit1	0	4	8
sit2	0	1	1
sit3	1	0	0

```
vegdist(mat, method = 'euclidean')
```

	sit1	sit2
sit2	7.615773	
sit3	9.000000	1.732051

```
vegdist(mat, method = 'bray')
```

	sit1	sit2
sit2	0.7142857	
sit3	1.0000000	1.0000000

from: Anderson, M.J., Crist, T.O., et al. , 2011. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. Ecology Letters 14, 19–28.

- ▶ Works on distance matrices
- ▶ Species can be added as *weighted averages*
- ▶ Eigenvalue based
- ▶ PCoA with euclidean distance == PCA

Principal coordinates analysis (PCoA)

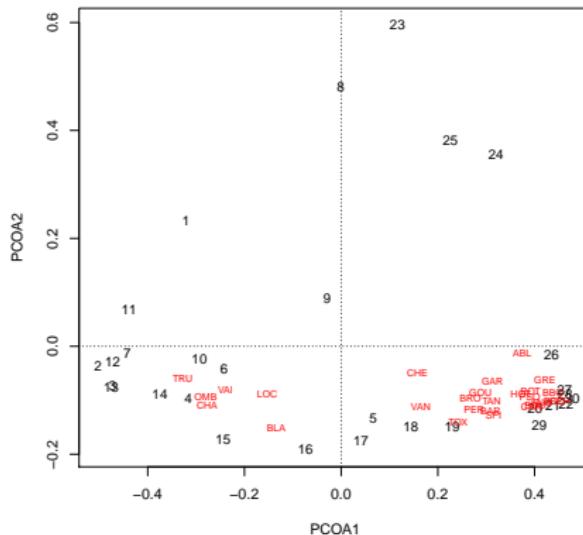
25 / 42

```
# Distance matrix
Dabu_dist <- vegdist(Dabu, method = 'bray')

# PCoA
PCOA <- cmdscale(Dabu_dist, eig = TRUE)

# Create plot
plot(PCOA$points, type = 'n',
      xlab = 'PCOA1', ylab = 'PCOA2')
text(PCOA$points,
     labels = rownames(Dabu), cex = 0.9)
abline(h = 0, lty = 'dotted')
abline(v = 0, lty = 'dotted')
# Add species as weighted averages
wa <- wascores(PCOA$points, Dabu)
text(wa, labels = colnames(Dabu),
     col = 'red', cex = 0.7)
```

```
# explained variance
(PCOA$eig / sum(PCOA$eig))[1:2] * 100
[1] 49.24914 15.95758
```



- ▶ Similar to PCoA
- ▶ Does not preserve exact distances between objects
- ▶ Possibly better representation in low dimensions
- ▶ **Not** eigenvalue based, iterative algorithm
- ▶ Axes have no meaning, just the relative distances

Nonmetric Multidimensional Scaling (NMDS)

27 / 42

```
# Distance matrix
Dabu_0 <- Dabu[!rowSums(Dabu) == 0, ]
Dabu_dist <- vegdist(Dabu_0, method = 'bray')

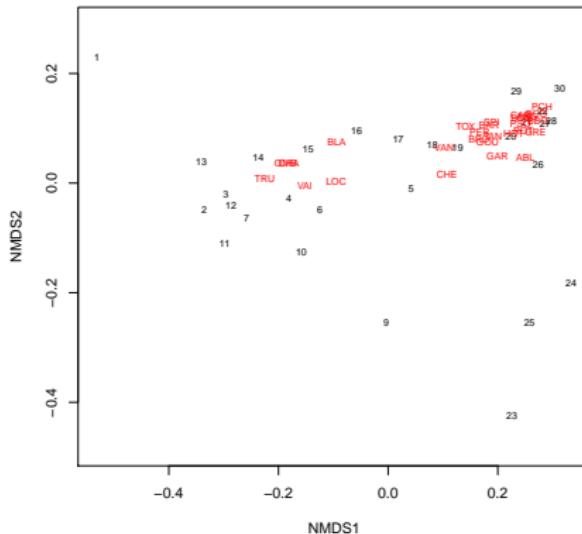
# NMDS
NMDS <- metaMDS(Dabu_dist, k = 2, trace = 0)

# Plot
plot(NMDS, type = 't')

# Add species as weighted averages
wa <- wascores(NMDS$points, Dabu_0)
text(wa, labels = colnames(Dabu),
     col = 'red', cex = 0.7)
```

```
# Stress value
NMDS$stress

[1] 0.07429467
```



Your turn!

Using the artificial dummy dataset.

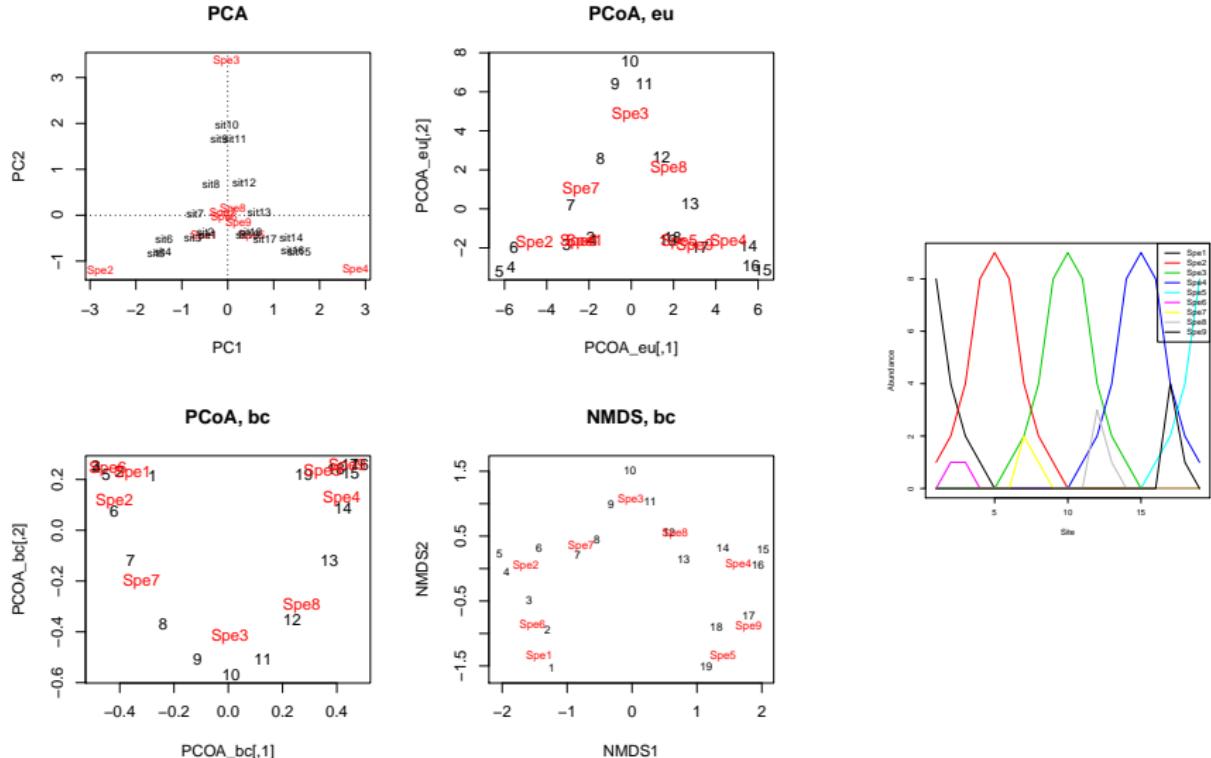
Run:

1. PCA
2. PCoA with euclidean distance
3. PCoA with Bray-Curtis dissimilarity
4. NMDS with Bray-Curtis dissimilarity

What are the differences between ordinations?
Which represent better the underlying gradient?

Exercise

29 / 42



Datasets
○○○○○○○

Indirect Gradient Analysis
○○○○○○○○○○○○○○○●○○○○

Direct Gradient Analysis
○○○○○○○

Permutation Tests
○

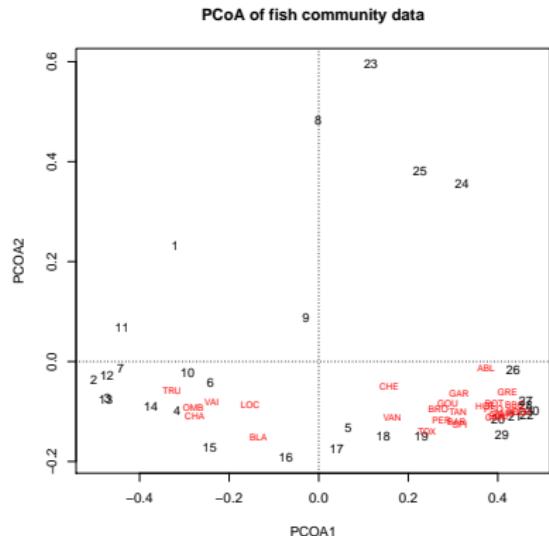
End

- This ordination is **only** driven by fish community data

Question:

- How can we interpret the gradients in community composition?

A solution:



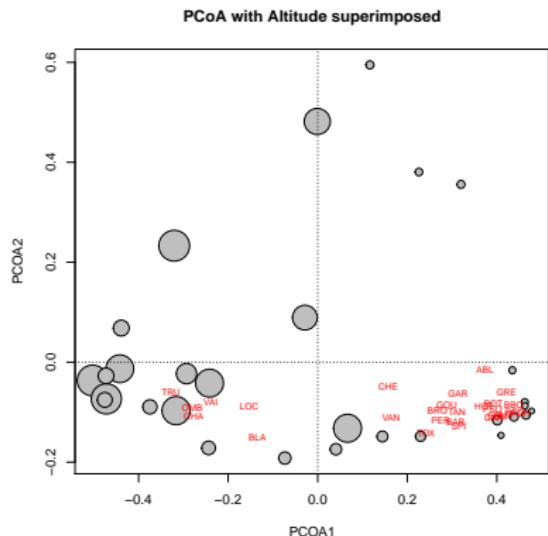
- ▶ This ordination is **only** driven by fish community data

Question:

- ▶ How can we interpret the gradients in community composition?

A solution:

- ▶ Superimpose environmental variables

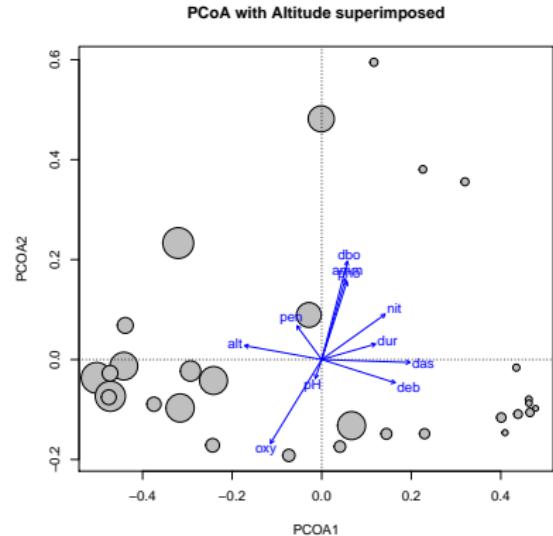


Fit environmental variables to ordination (II)

```
# PCoA of fish community data
plot(PCOA$points,
     xlab = 'PCOA1', ylab = 'PCOA2',
     cex = 5*Denv$alt / max(Denv$alt),
     main = 'PCoA with Altitude',
     bg = 'grey75', pch = 21)
abline(h = 0 , lty = 'dotted')
abline(v = 0 , lty = 'dotted')

# Fit Altitude to site-scores
ef <- envfit(PCOA, Denv)
plot(ef)
ef # summary

# Fit GAM
ordisurf(PCOA, Denv$alt, add = TRUE)
```



- ▶ Post hoc method
- ▶ non-linearity?

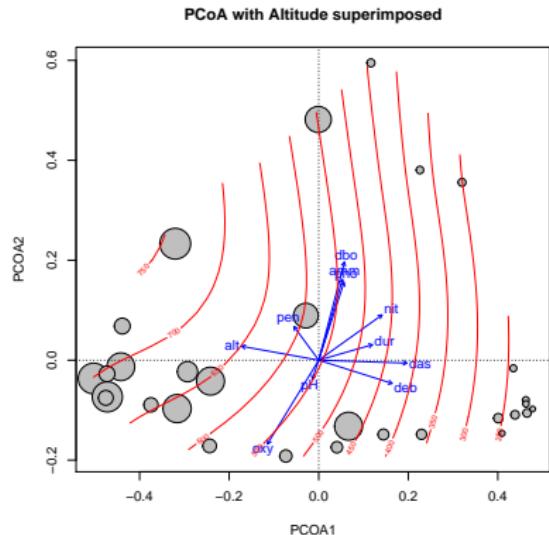
Fit environmental variables to ordination (II)

31 / 42

```
# PCoA of fish community data
plot(PCOA$points,
     xlab = 'PCOA1', ylab = 'PCOA2',
     cex = 5*Denv$alt / max(Denv$alt),
     main = 'PCoA with Altitude',
     bg = 'grey75', pch = 21)
abline(h = 0 , lty = 'dotted')
abline(v = 0 , lty = 'dotted')

# Fit Altitude to site-scores
ef <- envfit(PCOA, Denv)
plot(ef)
ef # summary

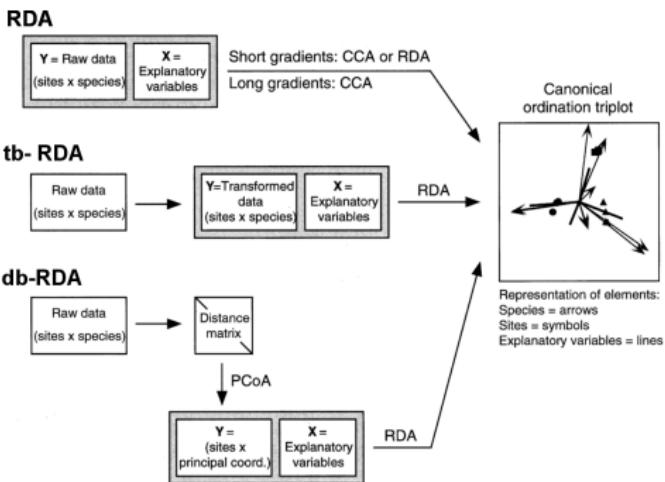
# Fit GAM
ordisurf(PCOA, Denv$alt, add = TRUE)
```



- ▶ Post hoc method
- ▶ non-linearity?
- ▶ be careful with summary
- ▶ Constrained ordination a better alternative

Direct Gradient Analysis

- ▶ Redundancy analysis (RDA)
- ▶ Transformation-based RDA (tb-RDA)
- ▶ Distance-based RDA (db-RDA)



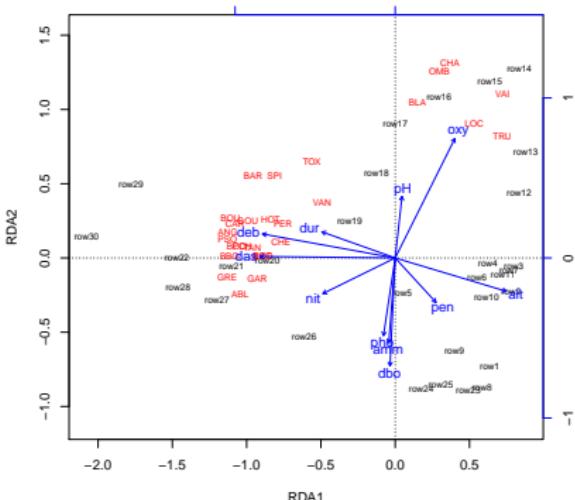
Adapted from: Legendre, P., Gallagher, E.D., 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129, 271–280.

- ▶ Associates both environmental and community data at once
- ▶ Combination of regression and PCA:
 1. Regress explanatory variables on community data
 2. Run PCA on fitted values
- ▶ Can test hypothesis about relationships

```
RDA <- rda(Dabu ~ ., data = Denv,  
            scale = TRUE)  
plot(RDA, scaling = 3)
```

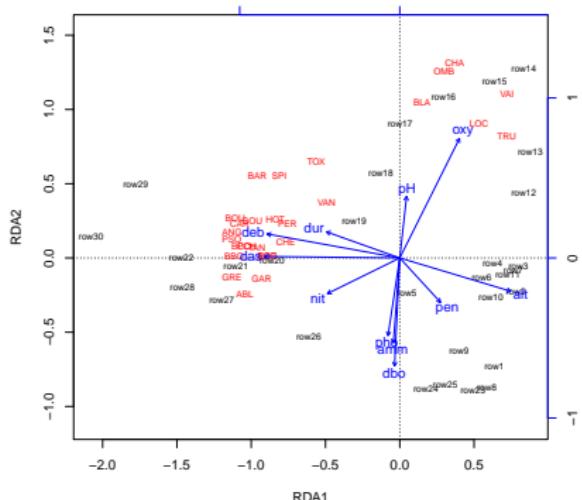
► Formula interface

- Left side: Response matrix
- Right side: Response variables from Denv



Redundancy analysis (RDA) — Interpretation? (I)

36 / 42



- ▶ Similar to PCA
- ▶ Projecting a site on response or explanatory variable approx. the value
- ▶ angles between response and expl. variables approx. their correlations

Redundancy analysis (RDA) — Interpretation? (II)

Call:

```
rda(formula = Dabu ~ das + alt + pen + deb + pH + dur + pho + nit + amm + oxy + dbo, c
```

Partitioning of correlations:

	Inertia	Proportion
Total	27.000	1.0000
Constrained	20.177	0.7473
Unconstrained	6.823	0.2527

Eigenvalues, and their contribution to the correlations

Importance of components:

	RDA1	RDA2	RDA3	RDA4	RDA5	RDA6	RDA7	RDA8	RDA9	RDA10	RDA11	PC1	PC2
Eigenvalue	14.714	2.6433	1.1341	0.76821	0.33807	0.28135	0.09356						
Proportion Explained	0.545	0.0979	0.0420	0.02845	0.01252	0.01042	0.00347						
Cumulative Proportion	0.545	0.6429	0.6849	0.71333	0.72585	0.73627	0.73974						
	PC3	PC4	PC5	PC6	PC7	PC8		PC9	PC10	PC11	PC12	PC13	PC14
Eigenvalue	0.76862	0.5399	0.38271	0.30039	0.25244	0.19487							
Proportion Explained	0.02847	0.0200	0.01417	0.01113	0.00935	0.00722							
Cumulative Proportion	0.91861	0.9386	0.95278	0.96391	0.97326	0.98047							
	PC15	PC16	PC17	PC18									
Eigenvalue	0.12593	0.10979	0.09775	0.06904	0.05043	0.04074							
Proportion Explained	0.00466	0.00407	0.00362	0.00256	0.00187	0.00151							
Cumulative Proportion	0.98514	0.98920	0.99282	0.99538	0.99725	0.99876							

Datasets

Indirect Gradient Analysis
Eigenvalue 0.01813 0.007958 0.005045 0.002412

Permutation Tests

o

Direct Gradient Analysis
Proportion Explained 0.00067 0.000280 0.000180 0.000090

End

Datasets

Indirect Gradient Analysis

Direct Gradient Analysis

Permutation Tests

End

Your turn!

Using the melbourne data.

Permutation Tests

Your turn!

Using the melbourne data.

- ▶ Distance-based hypothesis testing ((PER-)MANOVA, SIMPER, ANOSIM)
- ▶ Dispersion measures (β -Diversity, Functional diversity)
- ▶ Model-based ordination / hypothesis testing (See work of David Warton et al.)