EXAM KEY

RAW RESULTS:

Mean = 6.69 +/- 1.62 (Median = 7.20)

Maximum = 8.85

CORRECTED RESULTS:

Mean = 7.55 +/- 1.84 (Median = 8.13)

Maximum = 10.0

Section 1: Statistical Background

Co-variance:

Definition:
Amount of variability shared by two variables

Range of Values:
-infinite to +infinite

Formula:

$$Covariance = \frac{\sum (Xi - \overline{X})(Yi - \overline{Y})}{(n-1)}$$

Section 1: Statistical Background

Pearson Correlation:

Definition:
Quantifies intensity of association between two variables.
Covariance of X and Y divided by SD in X and SD in Y

Range of Values:
-1 to +1

$$r = \frac{Covariance}{\sqrt{(Variance\ X)(Variance\ Y)}}$$

Formula:

$$r = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum (Xi - \bar{X})^2 \sum (Yi - \bar{Y})^2}}$$

Section 1: Statistical Background

Orthogonality:

Definition:
Degree of independence between two variables.

Range of Values:
0 to 100%

Formula:

```
Orthogonality,% = 100(1-r^2)
```

Section 1: Statistical Background

Coefficient of determination:

Definition:
Amount of variance shared by two variables.

Range of Values:
0 to 100%

Formula:

$$r = \frac{Covariance}{\sqrt{(Variance\ X)(Variance\ Y)}}$$

r squared, where:

$$r = \frac{\sum (Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum (Xi - \bar{X})^2 \sum (Yi - \bar{Y})^2}}$$

Section 1: Statistical Background

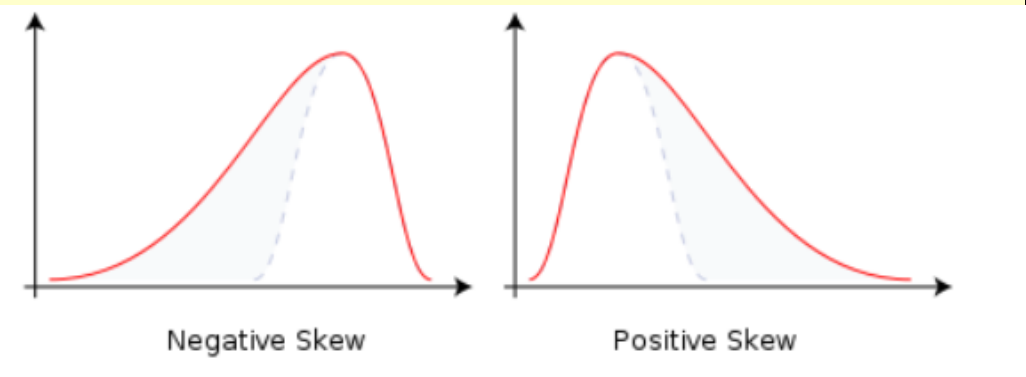
Skewness:

Definition:
Measures asymmetry of probability distributions.

Range of Values:
- infinite to + infinite

Formula:

$$skewness = \frac{\Sigma(y_i - \bar{y})^3}{(n-1)s^3}$$



Section 1: Statistical Background

More Skewness:

	y	(y - M)	$(y - M)^3$	
	8.04	0.54	0.16	
	6.95	-0.55	-0.17	
	7.58	0.08	0.00	
	8.81	1.31	2.24	
	8.33	0.83	0.57	
	9.96	2.46	14.87	
	7.24	-0.26	-0.02	
	4.26	-3.24	-34.04	
	10.84	3.34	37.23	
	4.82	-2.68	-19.27	
	5.68	-1.82	-6.04	
sum = $\sum y$ =	82.51	0.00	-4.46	sum = $\sum$deviations^3
mean = $(\sum y)/n$ = M	7.50		83.65	= (n-1) stdev^3
st dev = $\sqrt{}$var	2.03		-0.0533	= skewness

Section 1: Statistical Background

B) Spell out these acronyms (+0.1 each):

NMS:

EOF:

MRPP:

PO:

CCA:

Section 1: Statistical Background

B) Spell out these acronyms (+0.1 each):

EOF: Empirical Orthogonal Functions

Parallel EOF/PC analyses of the monthly SST and SLP anomaly fields, carried out independently by two of the present authors, were based on the temporal covariance matrix from the 1900–93 period of record.

A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production*

Nathan J. Mantua,+ Steven R. Hare,# Yuan Zhang,+
John M. Wallace,+ and Robert C. Francis@

Section 2: Interpretation

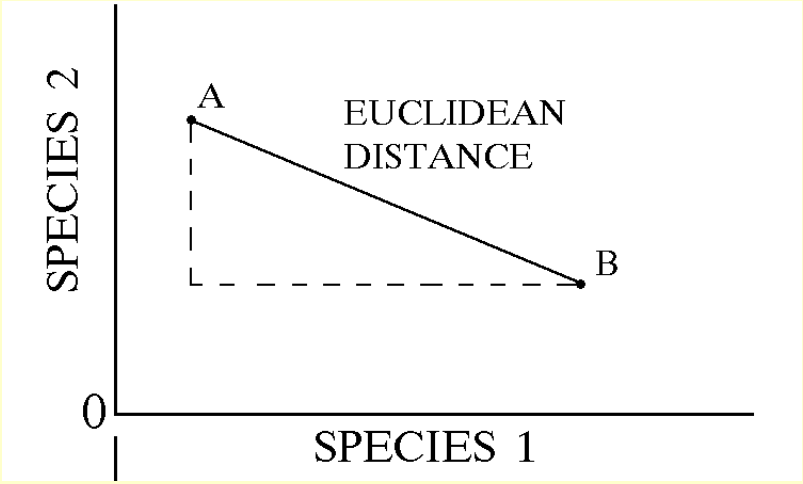
A) Clustering: You have four samples with two species each. (this question is worth 2 points):

Calculate and report the pair-wise Euclidean distances below (show calculations-leave square roots if necessary) (+0.10 each)

$$D = \sqrt[k]{x^k + y^k}$$

$k = 2$ gives Euclidean distance

$k = 1$ gives city-block distance



Section 2: Interpretation

	Sp1	Sp2
Plot 1	1	0
Plot 2	1	1
Plot 3	10	0
Plot 4	10	10

D (plot1, plot2) =

- Euclidean: sqrt [(1-1)^2 + (1-0) ^2]
  sqrt [0+1] = 1

		Euclidean			
		P1	**P2**	**P3**	**P4**
City-block	**P1**	-	**1**	Sqt (81 + 0)	Sqt (81+ 100)
	P2	**1**	-	Sqt (81 + 1)	Sqt (81+ 81)
	P3	(9 + 0)	(9+ 1)	-	Sqt (100 + 0)
	P4	(9 + 10)	(9 + 9)	(0 + 10)	-

Section 2: Interpretation

Which two plots would be linked first by the clustering algorithm? Explain (+0.25)

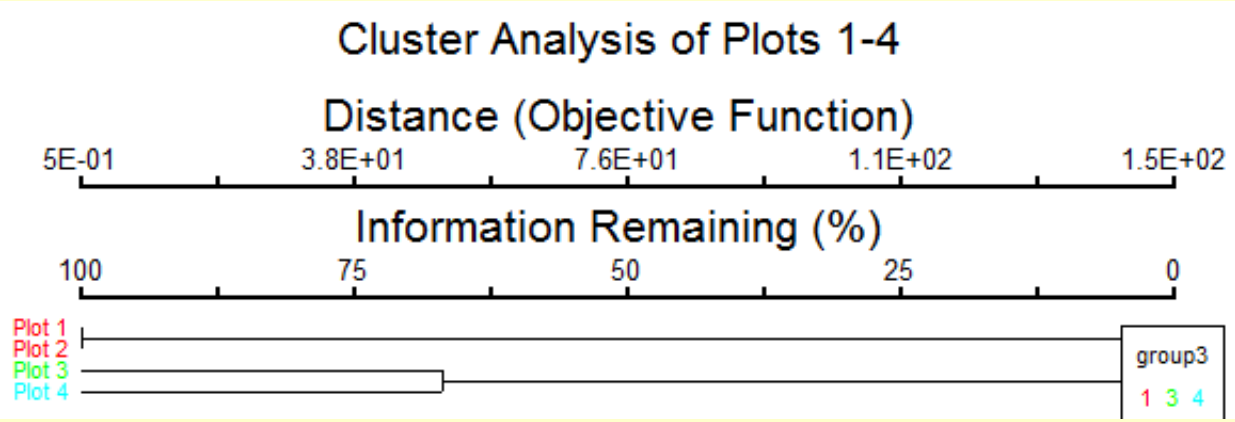
Plots 1 and 2 – because their distances are 1

Draw the four samples in “species-space” (+0.25):

	Sp1	Sp2
Plot 1	1	0
Plot 2	1	1
Plot 3	10	0
Plot 4	10	10

Section 2: Interpretation

Draw a generic dendogram, showing the relative distances between these four plots using “Euclidean distance” (+0.25):



Explain how the “information remaining (%)” and the “distance (objective function)” change as you combine these four samples into groups. Explicitly consider how many groups will you end up, and what are the maximum / minimum values of the explained variance (+0.25):

Information remaining is 100% before any groupings are made, and 0% when all plots are grouped. The objective distance will vary inversely, increasing as we create larger groups. The clustering ends with a single group and 0% of the information remaining.

Section 2: Interpretation

B) PCA (this question is worth 2 points): You performed a PCA using 999 randomization runs and got the following results:

VARIANCE EXTRACTED, FIRST 5 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	1304619.625	81.808	81.808	728259.750
2	279978.156	17.556	99.365	409313.875
3	7583.378	0.476	99.840	249840.922
4	2100.831	0.132	99.972	143525.641
5	447.436	0.028	100.000	63789.176

Based on this table, which PCA axes explains more variability than would be expected by chance? Explain your rationale? (+0.125)

Compare Observed Eigenvalue versus Broken-Stick Eigenvalue (OE > BS)

Section 2: Interpretation

These are the “loadings of the variables” in the axes:

FIRST 5 EIGENVECTORS, scaled to unit length.
These can be used as coordinates in a distance-based biplot,
where the distances among objects approximate their Euclidean distances.

	Eigenvector				
vars	**1**	**2**	**3**	**4**	**5**
year	0.0171	-0.0134	-0.9987	0.0435	0.0121
MEI	0.0040	0.0100	-0.0353	-0.9175	0.3960
PDO	0.0007	0.0038	-0.0284	-0.3951	-0.9182
upwell36	0.5460	-0.8375	0.0203	-0.0077	-0.0004
upwell39	0.8376	0.5462	0.0074	0.0088	-0.0011

Why do the loadings of the upwelling variables in axis 1 and 2 differ? (+0.125):

Because these variables do not co-vary perfectly. Axis 1 captures the positive co-variation and axis 2 captures the negative co-variation.

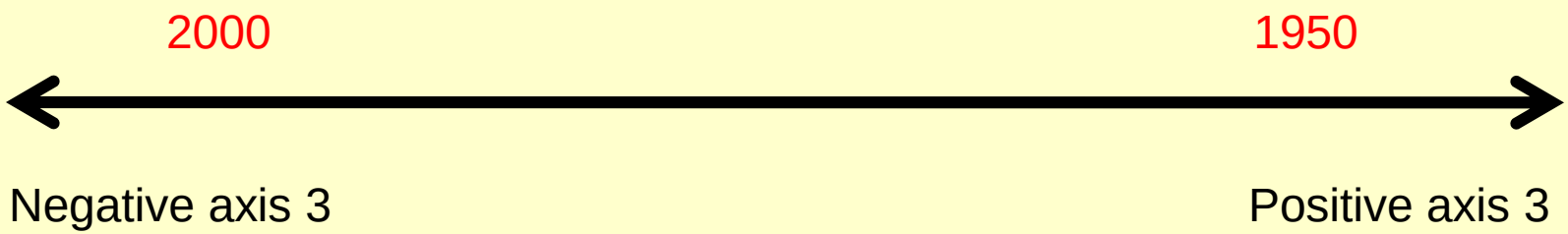
Section 2: Interpretation

Which axis indicates the shared in-phase variability of the two upwelling indices (+0.125)? Axis 1

Which axis indicates the shared out-of-phase variability of the two upwelling indices (+0.125)? Axis 2

Draw axis 3, (as a line), and plot the two following years along that axis: 1950 and 2000.
(Hint: I care about the relative location) (+0.125):

vars	**3**
year	-0.9987



Section 2: Interpretation

Check the randomization results below, and calculate p values for the two axes (+0.125 each):

RANDOMIZATION RESULTS 999 = number of randomizations

Axis	Eigenvalue from real data	Eigenvalues from randomizations: Minimum	Average	Maximum	p *
1	0.13046E+07	0.99879E+06	0.10048E+07	0.10738E+07	1 / 1000
2	0.27998E+06	0.51035E+06	0.57937E+06	0.58544E+06	1000 / 1000

Section 2: Interpretation

APPLICATION OF STOPPING RULES

Last useful Rule
axis acronym Explanation (see Peres-Neto, Jackson & Somers 2005)

1 Rnd-Lambda Observed eigenvalue as compared to randomization p values

1 Avg-Rnd Observed eigenvalue compared to average eigenvalue from randomizations

1 BS Observed eigenvalue compared to broken-stick eigenvalue

	Eigenvalue from	Eigenvalues from randomizations			
Axis	real data	Minimum	Average	Maximum	p *
1	0.13046E+07	0.99879E+06	0.10048E+07	0.10738E+07	0.001
2	0.27998E+06	0.51035E+06	0.57937E+06	0.58544E+06	1

Section 2: Interpretation

What is the inherent assumption of PCA? (+0.125):
Data normality (skewness / kurtosis) and “zeroes” under control)

What rule of thumb have we used to determine whether data distributions are skewed? (+0.125):
-1 < skewness < 1

What rule of thumb have we used to determine if there are too many zeroes in the species data for performing a PCA? (+0.125):
< 20% “zeroes” in the dataset

List two “criteria” have we seen used in the literature to determine which PCA axes are meaningful (+0.125 each):
Scaled Eigenvalue > 1, Eigenvectors explain > 60% variation, p values

What is the difference between EOF and PCA (+0.125):
None: Empirical Orthogonal Functions are PCAs in Earth Science / Oceanography

Section 2: Interpretation

C) (NMDS results: (this question is worth 2 points):
-Explain: Why do we need to perform many replicate runs when we calculate the observed axes with NMDS, unlike with PCA (+0.250):

NMDS relies on computing power to figure out the best relationships…
by exploring the variable landscape. Replicate runs are needed to ensure the different random starting points (configurations) do not drive the resulting patterns. This entails finding the best "global" solution, not a "local" stress minimum.

- Explain: Why do the NMDS axes do not necessarily decrease in the percent of variance explained, like happens with PCA (+0.250):

PCA attempts to explain as most variance as possible, in a hierarchical way.

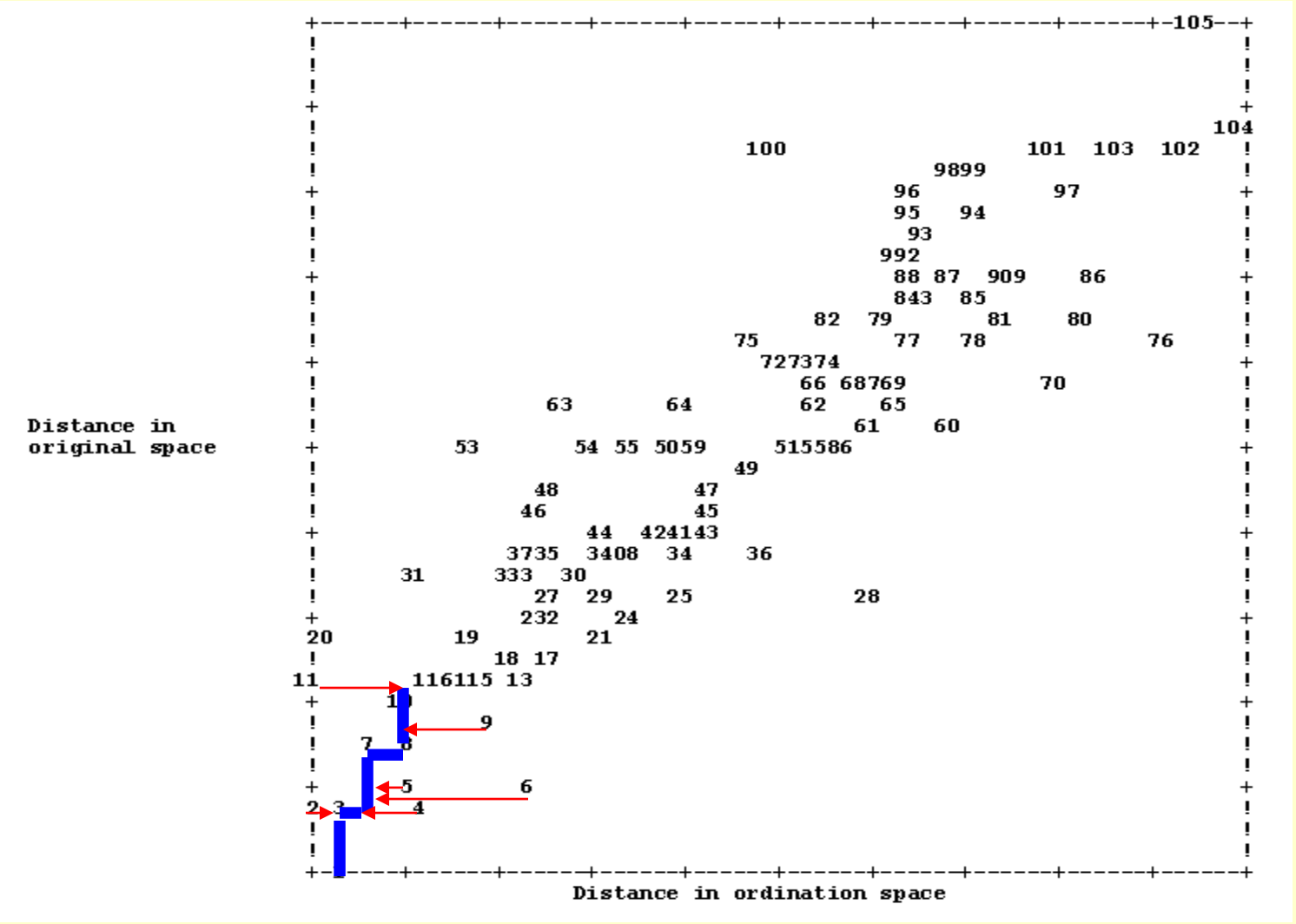
NMDS starts building the "best solution" wherever the robots land…
and the axes are thus not built hierarchically, but depend on the starting (random) configuration. Thus, subsequent axes may explain more variance.

Section 2: Interpretation

- Define stress, as used in the NMDS (+0.250):

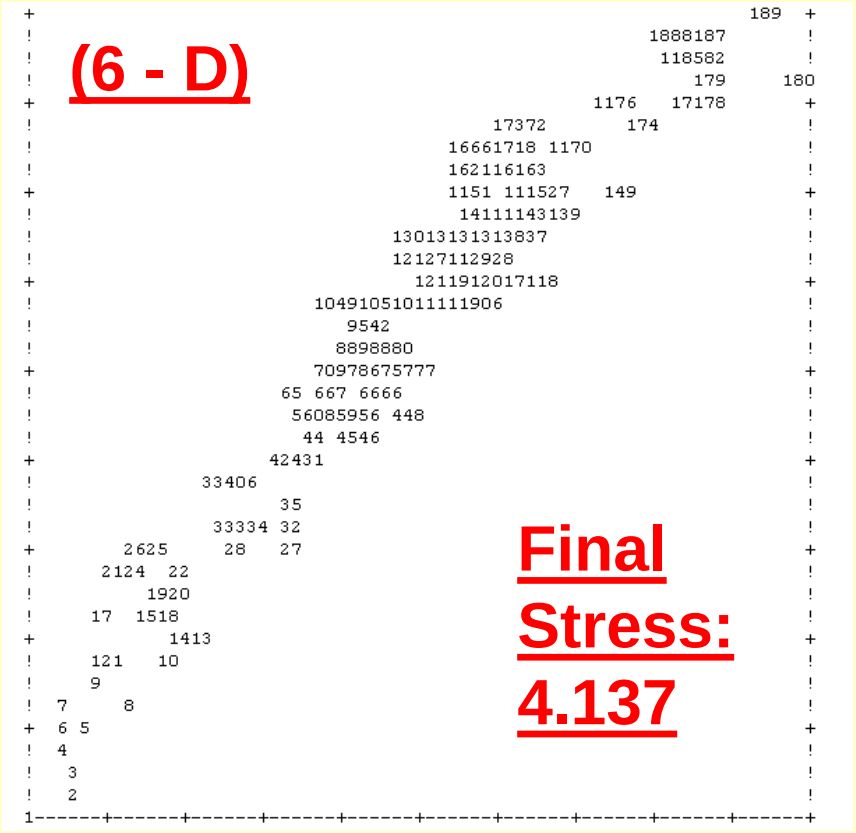
Stress:

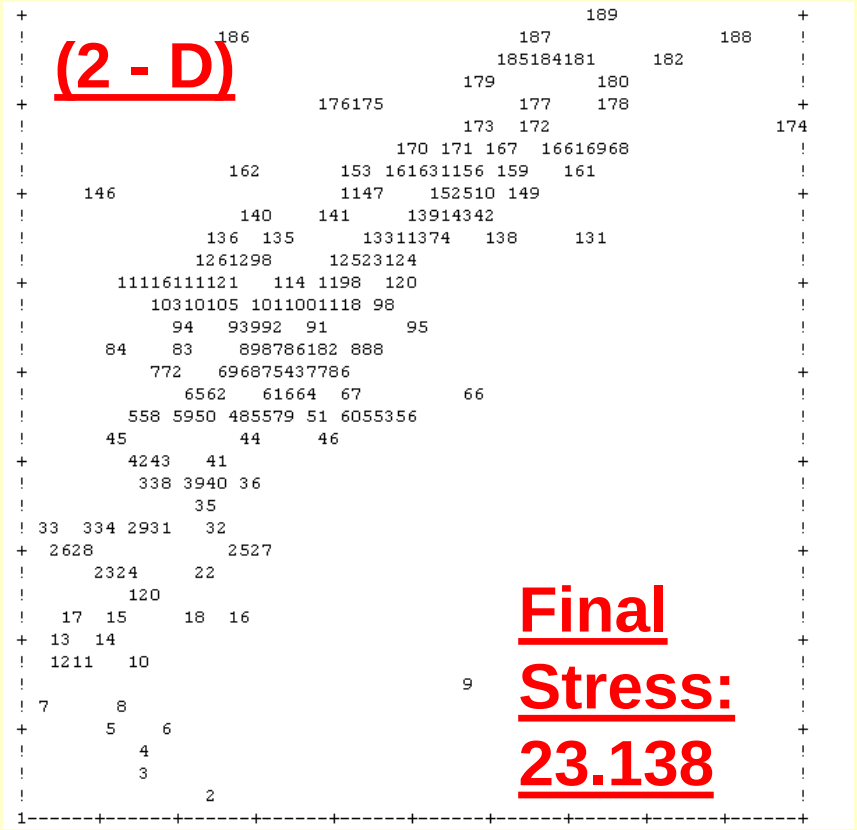
degree of departure from the 1:1 line in the relationship between k-dimensional distances (x axis) and distances in original data space (y axis)



Section 2: Interpretation

To illustrate how stress is calculated, make two plots comparing NMDS results: one with high stress and one with low stress
(label the plots accordingly: “high” vs “low” stress (+0.250).





Section 2: Interpretation

Explain how many axes are meaningful, in this NMDS analysis (+0.125 each):

Rule 1: Explain the rule and report the result:
P value < 0.05 (answer: 3 axes)

Rule 2: Explain the rule and report the result:
Stress decline of 5 units (answer: 3 axes)

STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)

	Stress in real data 10 run(s)			Stress in randomized data Monte Carlo test, 20 runs			
Axes	Minimum	Mean	Maximum	Minimum	Mean	Maximum	p
1	38.376	46.541	54.222	41.561	48.626	54.483	0.0476
2	20.366	22.469	25.766	21.752	24.574	28.997	0.0476
3	13.418	13.670	14.855	13.809	15.954	17.877	0.0476
4	8.919	8.954	9.268	8.579	10.807	12.085	0.0952
5	6.078	6.288	6.587	6.662	7.863	9.987	0.0476
6	4.138	4.217	4.499	4.635	5.716	7.708	0.0476

Section 2: Interpretation

Explain the Clarke's rule of thumb for assessing the meaning of the calculated stress (+0.1 each):

Stress < 0: Mistake? Trick?
(Stress can NEVER be < 0)

	Clarke's rules of thumb
< 5	An excellent representation with no prospect of misinterpretation. This is, however, rarely achieved.
5-10	A good ordination with no real risk of drawing false inferences
10-20	Can still correspond to a usable picture, although values at the upper end suggest a potential to mislead. Too much reliance should not be placed on the details of the plot.
> 20	Likely to yield a plot that is relatively dangerous to interpret. By the time stress is 35-40 the samples are placed essentially at random, with little relation to the original ranked distances.

Section 2: Interpretation

Finally, interpret this figure (+0.125 each):

What axis defines
the distributional differences
between Skipjack / Mahimahi:

Axis 1

What seabird species is associated
with all three predators:
Mahimahi, Skipjack , Odontocetes?

SOTE (in the origin)

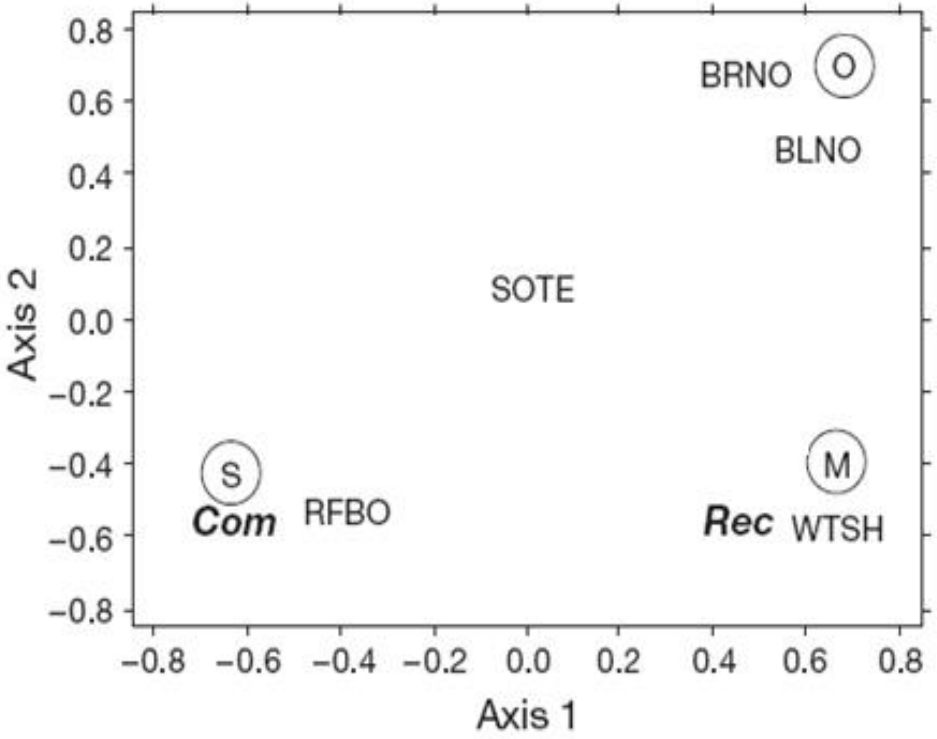


Fig. 7. NMDS plot showing 2-dimensional distances among the 5 most abundant seabirds (WTSH = wedge-tailed shearwater, SOTE = sooty tern, RFBO = red-footed booby, BLNO = black noddy, BRNO = brown noddy), 2 fishery types (commercial [Com] vs. recreational [Rec]), and 3 subsurface-predator types (M = mahimahi, O = odontocete, S = skipjack)

Section 3: Short Essays

Short Essays(1/2 page maximum): (1 point each)

A) Note: this is a philosophical question:
Briefly explain hypothesis testing (Hint: describe the null / alternate hypothesis) and explain how well this scientific approach jives with the advent of multivariate statistical methods.

Explain whether and how we should adapt hypothesis testing when using multi-variate methods?

Make the alternate hypotheses more information-rich
(rather than expecting the patterns; predict the patterns)

Use hypothesis-testing methods, rather than merely exploratory methods

Section 3: Short Essays

Short Essays(1/2 page maximum): (1 point each)

B) Note: this is a practical question:
Briefly explain how we can make multi-variate analyses more hypotheses based. Explicitly describe two approaches we can follow to do so: one occurs before data analysis takes place and one occurs during data analysis.

BEFORE: Only explore specific hypotheses, design study to sample across gradients / groups, and make sure you have the right number of samples.

DURING: Select a stringent alpha level and clear guidelines for data analysis; do not continue performing tests until you get significance; test for type-I errors