Multivariate Statistics (**MARS 6300**) - Homework 3 **Name:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Distributed:** Friday, February 16, 2018 **Due:** Thursday, March 1, 2018

**Instructions:** Copy and paste your answers below and turn in a word file and two excel files by the end of due day via email to khyrenba@gmail.com. Please use email title “MARS 6300 hw#3” and label all files with you’re a suffix including your name (e.g., MARS6300\_hw3\_hyrenbach). Unlabeled emails / files will be penalized 10% of points.

You are free to use any reference materials of your choice. While you are encouraged to work together, make sure you turn your own assignment. This homework is worth 5 points.

**The objectives of this homework are:**

A) To review and practice data relativizations.

B) To investigate the effects of relativizations on PCA.

C) To perform and interpret PCA analyses.

D) To critically evaluate the PCA literature.

E) To evaluate the reporting of PCA analyses.

To complete this homework, you will need:

* Instruction file: “MARS6300\_hw3.doc” (open with word file) – turn in
* “upwell\_PCA.xls” data file: (open with excel) – do not turn in
1. **Relativizations:**

You have three samples with three variable measurements.

Use the column summary tool in PC\_ORD and calculate column totals shown below:

****

Because you want each variable to have an equivalent weight in the analysis, you will relativize the data on the basis of the columns.

First, relativize the data by each column maxima – fill in the cells below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable 1** | **Variable 2** | **Variable 3** |
| **Sample 1** |  |  |  |
| **Sample 2** |  |  |  |
| **Sample 3** |  |  |  |
| **TOTAL** |  |  |  |

Did this relativization work well? Why / why not? – Explain:

Next, relativize the data by each column totals (in PC-ORD, this is the general relativization approach, with p = 1) – fill in the cells below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable 1** | **Variable 2** | **Variable 3** |
| **Sample 1** |  |  |  |
| **Sample 2** |  |  |  |
| **Sample 3** |  |  |  |
| **TOTAL** |  |  |  |

Did this relativization work well? Why / why not? – Explain:

C) Load the file “relativizations.wk1” into PC-ORD and perform the following relativizations by columns. For each instance, report two pieces of information:

i) paste resulting matrix. (Note: you can copy and paste the data from the matrix – select all and copy – or you can get a screen shot. If you do the latter, please crop image before pasting)

ii) summarize the rows / columns using PC-ORD tools and report the sums of columns / rows.

(Note: copy table from “results.txt” file and paste)

* Relativize by adjusting by the mean: What did PC-ORD do?
* Relativize by adjusting by rank: What did PC-ORD do?
* Relativize by adjusting by standard deviate: What did PC-ORD do?

**2) PCA Analysis – Interpretation:**

Load the “upwell\_PCA.xls” file and check the “summary of the column data (the variables)”.

 Report the skewness for the five variables:

 Skewness

**----------------------------------**

 **1 time \_\_\_\_\_\_\_\_\_**

 **2 MEI \_\_\_\_\_\_\_\_\_**

 **3 PDO \_\_\_\_\_\_\_\_\_**

 **4 upwell36 \_\_\_\_\_\_\_\_\_**

 **5 upwell39 \_\_\_\_\_\_\_\_\_**

Report the percent of “empty” cells: \_\_\_\_\_\_\_\_\_\_\_\_

Based on these results, would you proceed with the PCA analysis? Why / Why not?

Also report the “sum” of each of these five variables:

 Sums

**----------------------------------**

 **1 time \_\_\_\_\_\_\_\_\_**

 **2 MEI \_\_\_\_\_\_\_\_\_**

 **3 PDO \_\_\_\_\_\_\_\_\_**

 **4 upwell36 \_\_\_\_\_\_\_\_\_**

 **5 upwell39 \_\_\_\_\_\_\_\_\_**

On the basis of these totals, which variable do you think will have the largest effect determining the Euclidean distances between samples? Explain why:



Perform a PCA using the recommended settings presented in lecture:

Note: Set up the randomization tests for 999 runs

Save the “results.txt” file and look for the following results:

* Find the amount of variance explained by the first 5 axes:

---------------------------------------------------------------

 Broken-stick

AXIS Eigenvalue % of Variance Cum.% of Var. Eigenvalue

---------------------------------------------------------------

 1 5458.660 81.808 81.808 3047.112

 2 1171.457 17.556 99.365 1712.610

 3 31.730 0.476 99.840 1045.359

 4 8.790 0.132 99.972 600.526

 5 1.872 0.028 100.000 266.900

---------------------------------------------------------------

Based on this table, which PCA axis explains more variability than would be expected by chance? Explain Why?

How come we explain 100% of the variance with 5 axes?

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**

time 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

-----------------------------------------------------------------------------

Which are the two variables with the strongest loadings (coefficients) loading in axis 1.

Interpret what this means (how are these variables influencing axis1?):

Which are the two variables with the strongest loadings (coefficients) loading in axis 2.

Interpret what this means (how are these variables influencing axis2?):

Based on this interpretation, where would you expect the sample from June 1999 (s99.458) - the maximum upwelling recorded during this time series - to fall on the plot of axis 1 and axis 2 (Hint: Sample lies on positive / negative end of axis 1? On positive / negative end of axis 2?)

Explain:

You can check the scores of this sample (#162 in the list of COORDINATES OF samples) in axis 1 and axis 2. The scores are: for axis 1: \_\_\_\_\_\_\_\_ and for axis 2: \_\_\_\_\_\_\_\_

Check the randomization results below, and calculate the p values, for the two axes, using the equation: p-value for an axis is = (n+1)/(N+1),

where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

RANDOMIZATION RESULTS

 999 = number of randomizations

-------------------------------------------------------------------------

 Eigenvalue Eigenvalues from randomizations

 from ----------------------------------------

Axis real data Minimum Average Maximum p \*

 1 5458.7 4179.0 4201.9 4401.5 \_\_\_\_\_\_\_\_

 2 1171.5 2227.8 2426.6 2450.2 \_\_\_\_\_\_\_\_

-------------------------------------------------------------------------

\* p-value for an axis is (n+1)/(N+1), where n is the number of randomizations

with an eigenvalue for that axis that is equal to or larger than the observed

eigenvalue for that axis. N is the total number of randomizations.

Finally, use three “stopping rules”, to determine how many PCA axes to use. For each method, listed below, enter in the spaces provided the number of axes (starting with 1) that meet the rule. For instance, if axis 1and axis 2 meet the criterion Rnd-Lambda, write down “2”.

APPLICATION OF STOPPING RULES

-----------------------------------------------------------------------------

Last useful Rule

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

\_\_\_\_ Rnd-Lambda Observed eigenvalue as compared to randomization p values

\_\_\_\_ Avg-Rnd Observed eigenvalue compared to average eigenvalue

 from randomizations

\_\_\_\_ BS Observed eigenvalue compared to broken-stick eigenvalue

To check for the outliers, you can use “summery > outlier analysis”. Make sure you select Euclidean Distance as you distance measure. Ask for those outlier samples that are over 2 SD units away and request for all outputs (ranks / graph). Report the following information for the top two outliers:

------------------------------------------

 ENTITY AVERAGE STANDARD

 RANK NAME DISTANCE DEVIATIONS

------------------------------------------

 1

 2

Finally, look at the graph using the “graph > graph ordination > 2d” menu. View “simple scatterplot” and look for these outliers? Label these two samples to identify them. Copy and paste the image here:

View “main plot” and look for the correlations of “time” with both axes:

r (with axis 1): \_\_\_\_\_\_\_\_\_\_\_

r (with axis 2): \_\_\_\_\_\_\_\_\_\_\_

Using “Statistics > Correlations with Main Matrix”, calculate the correlations of each variable with each axis. What is the correlation of “time” with axis 3? Report the r here: \_\_\_\_\_\_\_\_\_\_\_\_\_

Note that the third axis is not significant (based on the Eigenvector), but “time” is very strongly correlated with it. Can you explain this result? (Hint: is time correlated with any other four oceanographic variables – check correlation matrix scatterplots).

**3) PCA Analysis – the influence of data relativization: relativization by maximum**

Load the “upwell\_PCA.xls” file and perform a relativization by maximum.

Report the missing values in the eigenvector table below:

VARIANCE EXTRACTED, FIRST 5 AXES

---------------------------------------------------------------

 Broken-stick

AXIS Eigenvalue % of Variance Cum.% of Var. Eigenvalue

---------------------------------------------------------------

 1

 2

 3

 4

 5

---------------------------------------------------------------

Why are the eigenvalues from this table much lower than the eigenvalues from the previous example? Explain.

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**

time

MEI

PDO

upwell36

upwell39

-----------------------------------------------------------------------------

Which are the two variables with the strongest loadings (coefficients) loading in axis 1.

Interpret what this means (how are these variables influencing axis1?):

Which are the two variables with the strongest loadings (coefficients) loading in axis 2.

Interpret what this means (how are these variables influencing axis2?):

Check the randomization results below, and calculate the p values, for the two axes, using the equation: p-value for an axis is = (n+1)/(N+1),

where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

RANDOMIZATION RESULTS

 999 = number of randomizations

-------------------------------------------------------------------------

 Eigenvalue Eigenvalues from randomizations

 from ----------------------------------------

Axis real data Minimum Average Maximum p \*

 1

 2

 3

Finally, use three “stopping rules”, to determine how many PCA axes to use. For each method, listed below, enter in the spaces provided the number of axes (starting with 1) that meet the rule. For instance, if axis 1and axis 2 meet the criterion Rnd-Lambda, write down “2”.

APPLICATION OF STOPPING RULES

-----------------------------------------------------------------------------

Last useful Rule

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

\_\_\_\_ Rnd-Lambda Observed eigenvalue as compared to randomization p values

\_\_\_\_ Avg-Rnd Observed eigenvalue compared to average eigenvalue

 from randomizations

\_\_\_\_ BS Observed eigenvalue compared to broken-stick eigenvalue

Finally, briefly describe how axis 1 and 2 have changed, as you performed the data relativization (Compare results from questions 2 and 3).

**4) PCA Analysis – the influence of data relativization: general relativization**

Load the upwell\_PCA.xls file and perform a general relativization by columns using p = 1.

Report the missing values in the eigenvector table below:

VARIANCE EXTRACTED, FIRST 5 AXES

---------------------------------------------------------------

 Broken-stick

AXIS Eigenvalue % of Variance Cum.% of Var. Eigenvalue

---------------------------------------------------------------

 1

 2

 3

 4

 5

---------------------------------------------------------------

Why are the eigenvalues from this table much lower than the eigenvalues from the previous example? Explain.

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**

time

MEI

PDO

upwell36

upwell39

-----------------------------------------------------------------------------

Which are the two variables with the strongest loadings (coefficients) loading in axis 1.

Interpret what this means (how are these variables influencing axis1?):

Which are the two variables with the strongest loadings (coefficients) loading in axis 2.

Interpret what this means (how are these variables influencing axis2?):

Check the randomization results below, and calculate the p values, for the two axes, using the equation: p-value for an axis is = (n+1)/(N+1),

where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

RANDOMIZATION RESULTS

 999 = number of randomizations

-------------------------------------------------------------------------

 Eigenvalue Eigenvalues from randomizations

 from ----------------------------------------

Axis real data Minimum Average Maximum p \*

 1

 2

 3

Finally, use three “stopping rules”, to determine how many PCA axes to use. For each method, listed below, enter in the spaces provided the number of axes (starting with 1) that meet the rule. For instance, if axis 1and axis 2 meet the criterion Rnd-Lambda, write down “2”.

APPLICATION OF STOPPING RULES

-----------------------------------------------------------------------------

Last useful Rule

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

\_\_\_\_ Rnd-Lambda Observed eigenvalue as compared to randomization p values

\_\_\_\_ Avg-Rnd Observed eigenvalue compared to average eigenvalue

 from randomizations

\_\_\_\_ BS Observed eigenvalue compared to broken-stick eigenvalue

Finally, briefly describe how axis 1 and 2 have changed, as you performed the data relativization (Compare results from questions 2, 3 and 4).

**5a) Critical reading of the literature:** Read the three posted PCA papers (Mantua et al. 1997, Nelson et al. 2004, Bonaiuto et al. 2003) and report the following:

\* Mantua et al. 1997:

- Did paper report how many samples / variables were measured? How many?

- Did the paper report the number of “empty” cells? How many?

- Did the paper quantify normality of the variables? How?

- Did the paper quantify cross-correlations between variables?

- Did the paper use any data relativizations? If yes, briefly explain:

\* Nelson et al. 2004:

- Did paper report how many samples / variables were measured? How many?

- Did the paper report the number of “empty” cells? How many?

- Did the paper quantify normality of the variables? How?

- Did the paper quantify cross-correlations between variables?

- Did the paper use any data relativizations? If yes, briefly explain:

\* Bonaiuto et al. 2003 (focus on scale 1 only):

- Did paper report how many samples / variables were measured? How many?

- Did the paper report the number of “empty” cells? How many?

- Did the paper quantify normality of the variables? How?

- Did the paper quantify cross-correlations between variables?

- Did the paper use any data relativizations? If yes, briefly explain:

**5b) Reporting of results in the literature:** Read the three posted PCA papers (Mantua et al. 1997, Nelson et al. 2004, Bonaiuto et al. 2003) and report the following:

\* Mantua et al. 1997:

- Did paper report PC axes loadings of variables? Were “important” variables identified? How?

- Did the paper report the % of explained variance by each axis?

- How many axes were used to explain the pattern?

- Did the paper discuss how the number of axes was selected? If yes, explain:

- Did the paper report p values? If yes, report how many permutations were used.

\* Nelson et al. 2004:

- Did paper report PC axes loadings of variables? Were “important” variables identified? How?

- Did the paper report the % of explained variance by each axis?

- How many axes were used to explain the pattern?

- Did the paper discuss how the number of axes was selected? If yes, explain:

- Did the paper report p values? If yes, report how many permutations were used.

\* Bonaiuto et al. 2003 (focus on scale 1 only):

- Did paper report PC axes loadings of variables? Were “important” variables identified? How?

- Did the paper report the % of explained variance by each axis?

- How many axes were used to explain the pattern?

- Did the paper discuss how the number of axes was selected? If yes, explain:

- Did the paper report p values? If yes, report how many permutations were used.