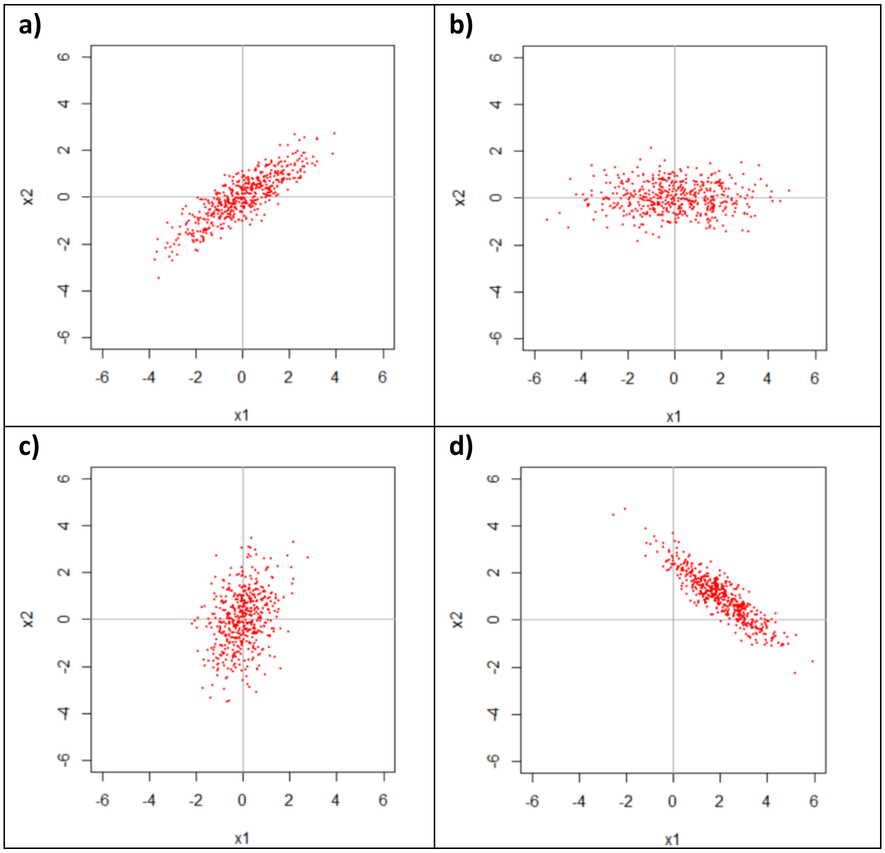
DSC 424
Assignment #3
Dr. John McDonald

Deliverables: Turn in your answers in a single PDF file. Copy any R output relevant to your answer into your document and explain your answer thoroughly and include a copy of the full analysis in your report along with your conclusions.

Note that this homework spans two modules. You should not wait until the second module to get going on it because many of the beginnings of these problems will be doable after the first week. It will most likely take you more than one week to complete.

1) **Remember to submit the milestones for the final project during these two weeks.**

2) (10 points) For each of the following datasets, draw the principal component vectors for the dataset, and for their **lengths,** estimate the size of the eigenvalue (i.e. the variance in the direction of the principal component). Note, you do note need to do this precisely, but you should be able to get a rough estimate from the graph. One question to think about is where should the principal component vectors be based (i.e. where should the arrow's tail be?)



3) **(10 points)** Answer each of the following by hand for the following matrices/vectors, and then verify your answers with R code:

$$M = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}, N = \begin{bmatrix} 21 & -2 & 1 \\ -3 & 10 & -11 \\ 3 & -22 & -1 \end{bmatrix}, v = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$$

a) Compute the eigenvalues and eigenvectors of *M*

b) Verify that *v* is an eigenvector of *N*. Note that you do not have to solve for the eigenvectors and also that the result may be somewhat approximate. Remember, what does it mean for *v* to be an eigenvector?

c) For the eigenvector in b), what is the corresponding eigenvalue (note, you do not need to solve for it. Hint: what does it mean that *v* is an eigenvector? Also note that it will be approximate (there may be a difference in the second or third decimal place.)

4) **(Principal Component Analysis, 20 points)** Begin with the “census2.csv” datafile, which contains census data on various tracts in a district. The fields in the data are

1. Total Population (thousands)
2. Professional degree (percent)
3. Employed age over 16 (percent)
4. Government employed (percent)
5. Median home value (dollars)

a) Conduct a principal component analysis using the covariance matrix (the default for prcomp and many routines in other software) and interpret the results. How much of the variance is accounted for in the first component and why is this?

b) Try dividing the MedianHomeValue field by 100,000 so that the median home value in the dataset is measured in $100,000’s rather than in dollars. How does this change the analysis?

c) Are there any other fields that are in particular need of scaling? Explain why or why not.

d) Compute the PCA with the correlation matrix instead. How does this change the result and how does your answer compare with your answer in b)? How does the meaning of the first component change?

e) Analyze the correlation matrix for this dataset for entries that are significant (i.e. different from zero) at a 95% confidence level. Are there any variables that are correlated with **most** of the other variables or are uncorrelated with **all** of the other variables? What is the importance of doing this and what might you consider doing with such variables?

f) Discuss what using the correlation matrix for PCA does compared to the covariance matrix and why it may or may not be appropriate in this case.

5) **(Principal Component Analysis, 20 points):** The data given in the file 'Employment.txt' is the percentage employed in different industries in Europe countries during 1979. Techniques such as Principal Component Analysis (PCA) can be used to examine which countries have similar employment patterns. There are 26 countries in the file and 10 variables as follows:

Variable Names:

1. Country: Name of country
2. Agr: Percentage employed in agriculture
3. Min: Percentage employed in mining
4. Man: Percentage employed in manufacturing
5. PS: Percentage employed in power supply industries
6. Con: Percentage employed in construction
7. SI: Percentage employed in service industries
8. Fin: Percentage employed in finance
9. SPS: Percentage employed in social and personal services
10. TC: Percentage employed in transport and communications.

a. Is scaling appropriate for this data? Explain why or why not.

b. Note that whatever your answer for c) the "principal" function will scale your data. This is because scaling is the default behavior in factor analysis. So, compute an initial principal component analysis using "prcomp" **with scaling** and apply the knee and var=1 criteria. How many components does each method suggest? Explain how confident you are in this result and if there are any ambiguities, why you made the choice you did.

c. Is VARIMAX factor rotation being applied in your computation in b)? Explain.

d. Print the component coefficients for the number of components you chose in b). For each component, write out the formula and give a brief interpretation. How easy are they to separate in-terms of meaning?

e. Run a parallel analysis for this dataset to compute a suggested number of components. Use the results of this and what you got with the knee and var=1 to choose a number of components. Explain your choice in detail.

f. Use "principal" to compute the Principal Factor Analysis with this number of components, and with VARIMAX factor rotation. **Give the formula for each component and a brief interpretation. Has rotating improved the ability to interpret the components?**

g. What countries have the highest and lowest values for each factor (only include the number of components specified in part e). For each of those countries, give the principal component scores (again only for the number of components specified in part a).

h. Consider the loadings matrix in e, how appropriate is the number of components you selected? Try running the analysis with one more and one fewer component. What do the results suggest for the number of components to finally select?

6) **(20 points, Common Factor Analysis)**
For this problem, you will analyze partial from intelligence tests given to children. Each child was given 11 tests on which they were rated. These were:

```
info     = 'Information'
comp     = 'Comprehension'
arith    = 'Arithmetic'
simil    = 'Similarities'
vocab    = 'Vocabulary'
digit    = 'Digit Span'
pictcomp = 'Picture Completion'
parang   = 'Paragraph Arrangement'
block    = 'Block Design'
object   = 'Object Assembly'
coding   = 'Coding';
```

Download the datafile, load it and complete the following analysis steps

a) Should the data be scaled or not for running PCA? Explain why/why not in detail.
b) Run an initial corrplot and an initial unrotated PCA (i.e. no VARIMAX). Use the corrplot and the techniques from the lecture to determine the appropriate number of factors to extract. Are there any variables that will likely be single-variable factors? Explain.
c) Run a Principal Factor Analysis with VARIMAX rotation and report the loadings with a cutoff of .4, and also plot the contributions to the components using either a biplot or PCA_Plot_Psych. Analyze the loadings and the plot. How clean and useful are the variable separations? Give a name to each component.
d) Then sort the scores by the first and second component (you will have to do this separately for each). Consider the cases (children) that score highly or extremely low on each. What do the scores mean for each of these cases? Are there any surprises?
e) Run a Common Factor Analysis (exploratory) and compare the loadings to those of the principal factor analysis. Note any significant differences and explain how they affect the factors practically.

7) **(Paper review)** An academic paper on principal component analysis in genetics research is posted in the "Supplimental Reading List" and also included in the homework materials. Read the paper and review the paper's use of PCA. In your analysis, you should address the following:

a) How suitable is their data for PCA? How are they applying PCA? Are they trying to extract interpretable underlying variables, or is their goal dimensionality reduction?
b) Are they scaling or do they have another method for equalizing variance for PCA?
c) Are they using factor rotation, and if so, what kind of factor rotation do they use?
d) How many components do they concentrate on in their analysis and how much variance does it capture? What criteria did they use for their choice?
e) Do they evaluate, and how do they evaluate the stability of the components?
f) What conclusions does PCA allow them to draw?

8) (**Reflection)** Post a comment on the lectures 3 & 4 forum regarding some topic covered during these lectures.