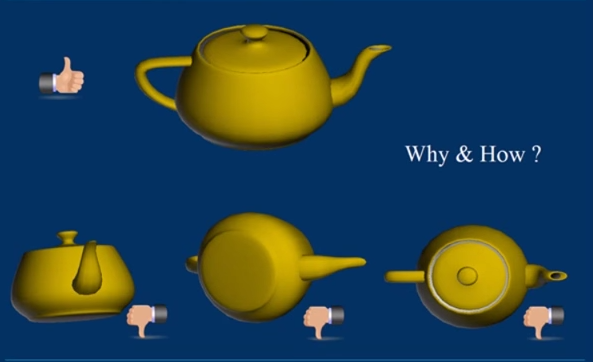
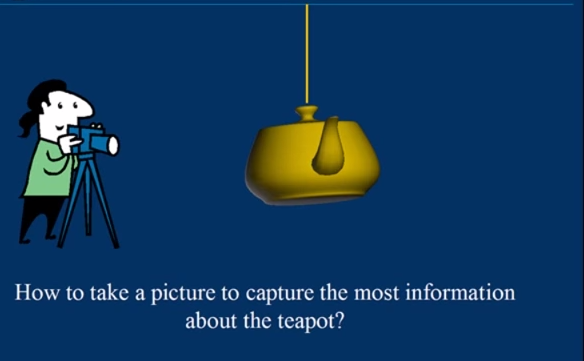
Principal Component Analysis (PCA) Assignment:

**Introduction / Background Material:**

Principal Component Analysis (PCA) is a data reduction technique that extracts what is most important and reduces the data to a smaller set with minimal loss of information. PCA reveals and minimizes undesirable redundancy in datasets. Essentially, it identifies patterns in datasets that would be hard to detect otherwise, and expresses the dataset in a way that shows the similarities and differences.

These two pictures illustrate PCA:



If a data set contained the three bottom photos of the teapot, a PCA could transform that data into the top photo.

A PCA results in “principal components”, which are filtered, compressed, and non-redundant transformations of the dataset that can be used in place of the raw data in further analyses. For instance, it is used in face recognition algorithms, genomics, and reducing climatic data in ecological niche modelling (there are links to literature at the end).

**Objectives:**

* To be able to recognize when a PCA is appropriate for a given dataset
* Understand how PCA works, its assumptions, and how to interpret the results
* Gain familiarity with the PCA syntax in R

**Data**: Evan’s Hive Data

Source: McGill Apicultural Association Beehives

Four beehives were equipped with in-hive sensors that detected temperature, relative humidity, barometric pressure, and decibels. A month long experiment was conducted: 2 hives were exposed to electromagnetic radiation from a Wi-Fi router and 2 control hives were outside of the Wi-Fi signal. Briefly, the context of this research involves developing hive sensors for future applications in honeybee research and apiary management. Ideally, it would be really sweet to have sensors that use wireless technologies (like Wi-Fi) that would send the hive data to a website or phone app, but that might interfere with the bees’ navigation capabilities and affect pollination rates and yields. Sensors logged data 18 times a minute, but here the data was filtered and averaged per hour. This data set is of the pilot project.

How the variables are coded:

In this study, there were two temperature and humidity sensors with different accuracies:

* The DHT22 sensor: ±0.1% RH and ±0.1°C resolution
  + “dht22\_t” and “dht22\_h” = temperature and humidity respectively
* The DHT11 sensor: ±2% RH and ±0.5°C resolution
  + “dht11\_t” and “dht11\_h”

Barometric Pressure (kPa)

* “pa”

Decibels (dB)

* “db”

**Question:** Can you see similarities or overlapping aspects in any of these variables?

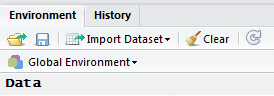
**Answer**: two temperature data columns and two humidity columns – this will show up in a PCA

PCA Principle 1: In general high correlation between variables is a telltale sign of high redundancy in the data.

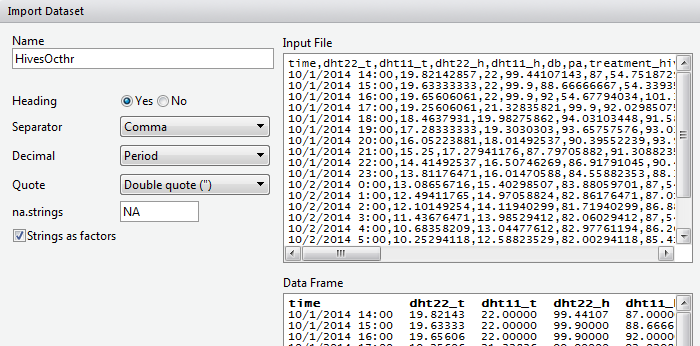
**How to download:**

Download the “HivesOcthr.csv” from myCourses (along with this .doc and the .R script)

In the top right of R Studio, click the import dataset, and select the “HivesOcthr.csv”



When asked for “heading”: click “Yes”



**Exercise:**

Load the data, subset\*, and plot.

\*you’ll notice that the raw data contains more columns than of interest. The subset function is used and can be pretty useful in data manipulations

The temperature and humidity variables are noticeably the most correlated. But what about the other variables?

* Evan intuitively thought that there might be a relationship between temperature and humidity inside the beehives, and also thought these in-hive climatic conditions could influence the loudness of the bees’ buzz (db).

Most of the step-by-step instructions from here onwards are in the .R script. This document elaborates on some of steps or math.

The underpinning mathematics involve variance, co-variance, and matrix algebra. This process is hidden in the principal component analysis function in R, but is described here.

Variance: How close the data is from the mean. Standard deviation is commonplace and calculated by taking the square root of the average of the squared differences of values from the mean value. It is a measure of one variable.

Covariance: measures the statistical relationship between multi-dimensional data sets. Unlike standard deviation, which asses a single dimension of a dataset independently, covariance measures how much dimensions of datasets vary with each other. Give a dataset with dimensions X and Y, the formula below is used to calculate the covariance. It can be used for > 2 dimensions however.

C:\Users\N\Desktop\covar.PNG

Run covariance test on the data (formula given on .R script). Notice that the result is the same 6x6 dimension as the first plot we made.

Interpreting the results lies in sign of the covariance values and how close the values are to 0 or 1. A positive value indicates that both dimensions increase together. If the value is negative, then as one dimension increases, the other decreases. A covariance is zero indicates that the two dimensions are independent of each other.

**Matrix Algebra: Eigenvectors and Eigenvalues**

The heart of PCA lies in concepts from linear algebra.

Matrix algebra will be mentioned because it is a process PCA uses. This will get more technical than what is needed to understand how PCA works.

Matrices of a square dimension, 2\*2, 3\*3, 4\*4, n\*n, etc. produce something called eigenvectors. There are n of them given an n\*n matrix. Eigenvectors represent the direction and distance of the variance. It is similar to vectors from introductory physics that represent x and y position of a slope, but applied to covariance.

<http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors#Real_matrices>

The data of the eigenvectors are referred as “loadings”, which indicate the strength or influence of each variable on the PC.

**Question:** what patterns can you see in the eigenvector columns?

**Answer:** the first PC is the most negative, and PC 2 is the second most negative or close to negative, and decreases to PC6.

This reorganization of the covariance is how PCAs are calculated. The actual mechanisms of eigenvalues and matrix math won’t be covered in more detail than this, but this is the step-by-step process of arriving at at PCA: covariance, compute eigenvalues and eigenvectors of the covariance matrix,

We will come back to this with the results of the PCA. It is ok if it is not super clear now.

**PCA: The Easiest Step!**

To repeat, PCA is a method of finding patterns in multidimensional data, and expressing the patterns. It is useful because these patterns can be used to compress the data without losing much information.

Comparing the pca$loadings to the covariance matrix eigen values, it can be seen that they are identical except the pca$loadings do not show all of the eigenvalues, just the ones of interest. This is the connecting piece between eigenvalues and PCA. It is ok to not understand the actual matrix math underneath this step, but just know that this is the footwork underneath a PCA.

**How to interpret summary** ( ): It is very straightforward

* Proportion of Variance: this row states the % of the total variance a principal component captures. The first PC has the highest, and decreases down to the last PC
* Cumulative variance: this is just the sum of proportion of variances for each additional PC
* Standard deviation= st.dev of the PC.

The summary always displays the PCs in decreasing order.

**Question** What information does a Scree plot show?

* It shows the proportion of variance each PC explains

There you have it! The rest of the assignment is visualizing how each data column influence the PCA. Below are the basic assumptions PCA follows and links to more tutorials and academic articles that use PCAs.

**Assumptions of PCA:**

Principal components analysis is a popular tool for studying high-dimensional data. It relies on four major assumptions (Shlens):

1. Linearity. This means that the only interaction among different signal sources is that they add together. If the strength of a combined signal were the product of the strengths of contributing signals, for instance, this would be a non-linear interaction and PCA would not work.

2. The interesting dynamics have the largest variances.

3. Mean and variance are sufficient statistics. Since PCA is designed around the covariance matrix of mean-centered data, the only statistics it considers are the mean and variance. If the data cannot adequately be described by its mean and variance (e.g. it is not Gaussian or exponentially distributed), PCA will be inadequate.

4. Orthogonal components. This is a particularly strong assumption made by PCA.

It is this assumption that allows PCA to be computationally

PCA Principle 1: In general high correlation between variables is a telltale sign of high redundancy in the data.

PCA Principle 2: The most important dynamics are the ones with the largest variance.

**Future reading**

**Tutorials**:

Shlens. A Tutorial on Principal Component Analysis. Copy retrieved [04-09-2008] from:

http://www.cs.cmu.edu/~elaw/papers/pca.pdf (2005)

https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition\_jp.pdf

<https://www.youtube.com/watch?v=BfTMmoDFXyE>

http://www.cs.otago.ac.nz/cosc453/student\_tutorials/principal\_components.pdf

**Academic articles:**

Dubois, Patrick CA, Gosia Trynka, Lude Franke, Karen A. Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova et al. "Multiple common variants for celiac disease influencing immune gene expression." *Nature genetics* 42, no. 4 (2010): 295-302.

Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2, no. 1 (1987): 37-52.