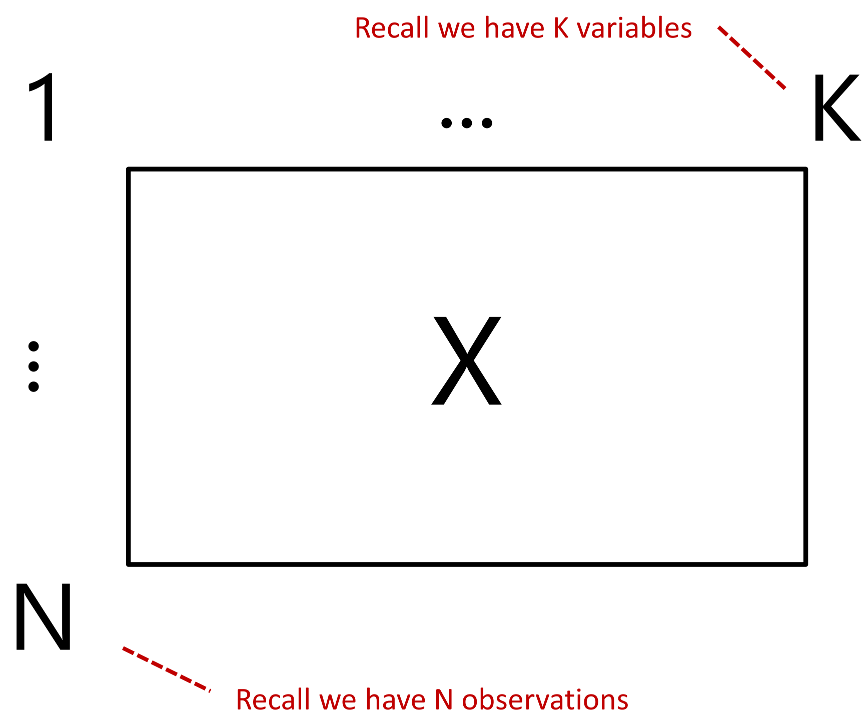
Chemical Engineering 4H03

Principal Component Analysis (PCA)

Jake Nease
McMaster University

Review: Data Sources

- PCA considers a singe data matrix (table) called **X**
  - What goes in the columns of **X**?
  - What go in the rows of **X**?

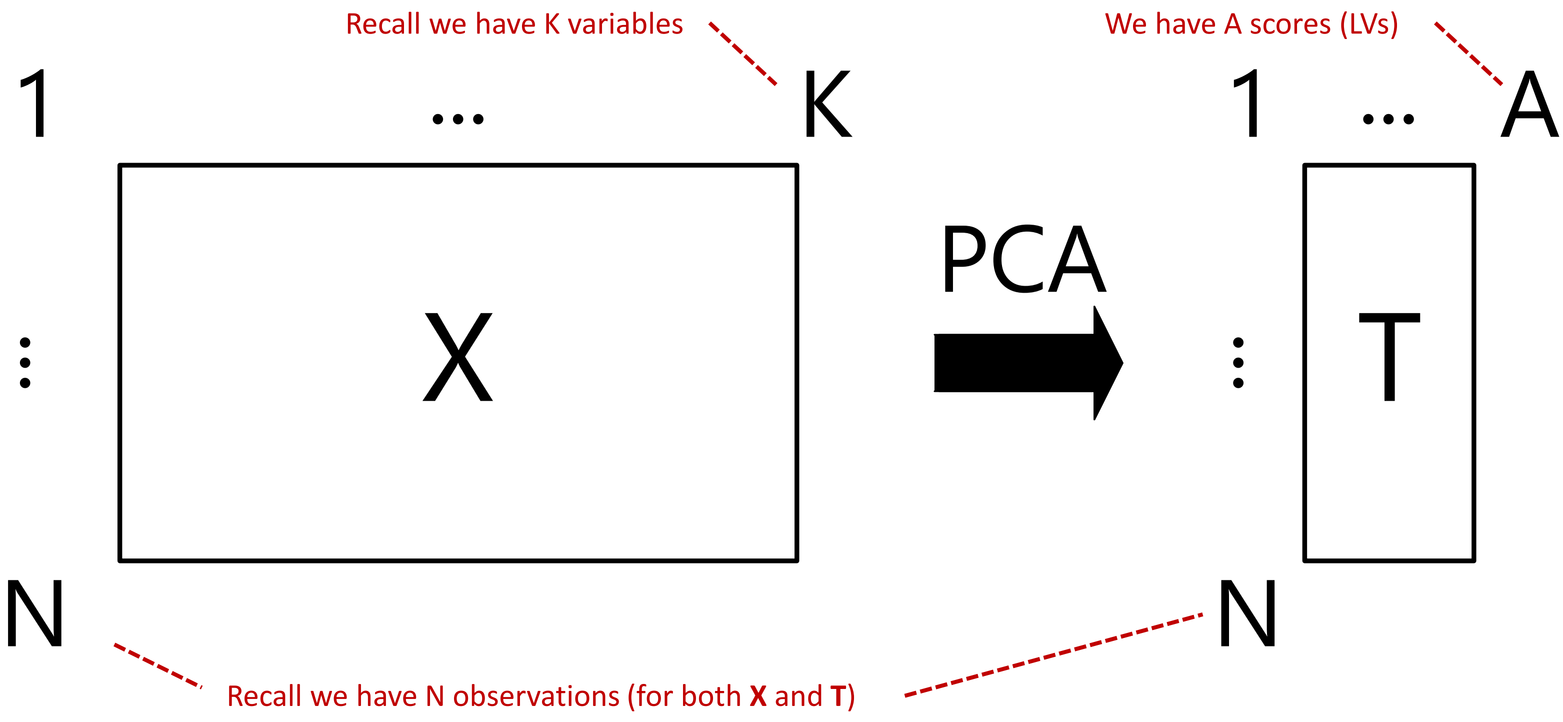


Review: Visualization

- How would you visualize this data?
- EXAMPLE: Assume N = 300 and K = 50
  - Not an uncommon shape for data
- IDEAS?
  - One column at a time (time series, histogram, box plot)
  - One row at a time (spectral data)
  - Multiple columns at a time (scatterplot matrix)
- Note: Scatterplot matrix requires $\frac{K(K-1)}{2}$ pairs!

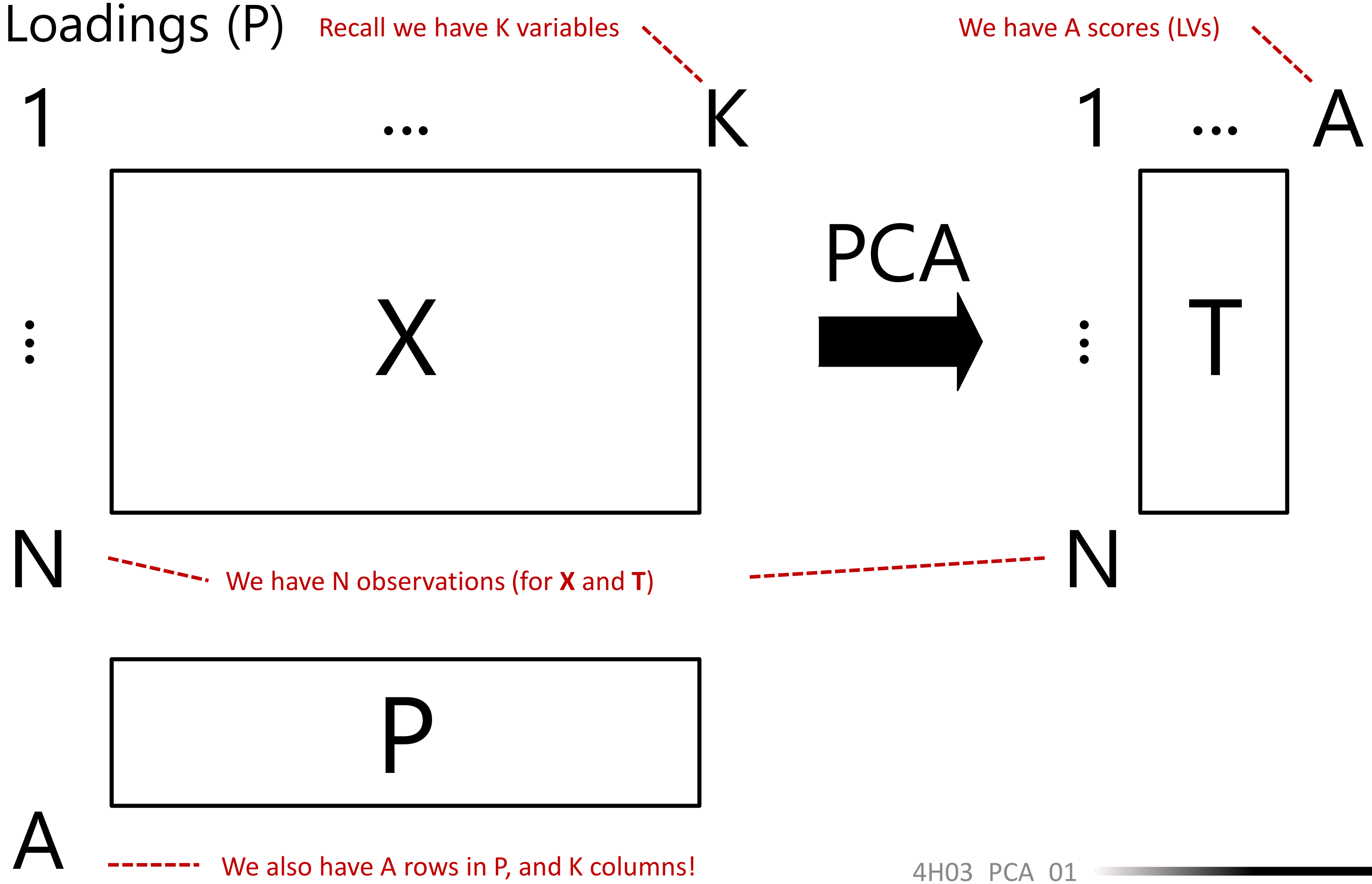
Review: What is PCA?

- **Mathematical Objective**
  - Find the best summary of data **X** using the **fewest** number of "summary variables"
  - These "summary variables" are known as the scores, **T**



Objectives for this Class

- To understand that the PCA model will compute
  - Scores (T)
  - Loadings (P)



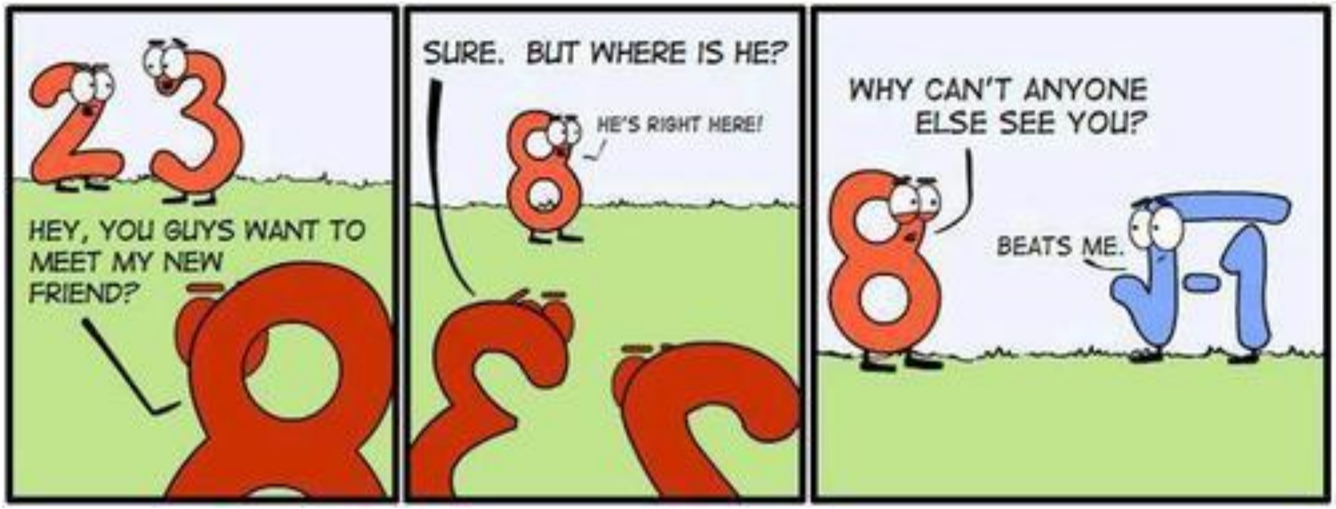
Objectives for this Class

- We desire to understand
  - Intuitive meaning of scores (**T**), loadings (**P**) and errors (**E**) in a PCA model
  - How to interpret these three things if presented with a model
  - How we can start building our own models with new data sets and learn from that data

- How?
  1. Data preprocessing
  2. Geometric interpretation of PCA (hand waving)
  3. Geometry (understand hand waving)
  4. Algebra (justify hand waving)

Data Preprocessing

Time to bust out the math



https://justinchristensen.weebly.com/uploads/1/6/8/0/16805852/7765276_orig.jpg

Some Math Notation For Ye'

- Our data is in matrix **X** which is $\in \mathbb{R}^{N,K}$
  - Each (row, column) in **X** is denoted as $x_{n,k}$
  - Ex: second data point in column 9 is $x_{2,9}$
  - Any **single entire column** of **X** is $\boldsymbol{x}_k$
  - Any **single entire row** of **X** is $\boldsymbol{x}_n$

- Our scores are in matrix **T** which is $\in \mathbb{R}^{N,A}$
  - Each (row, column) in **T** is denoted as $t_{n,a}$
  - Any **column** of **T** is $\boldsymbol{t}_a$

- Our loadings are in matrix **P** which is $\in \mathbb{R}^{K,A}$
  - Each (row, column) in **P** is denoted as $p_{k,a}$
  - Any **column** of **P** is $\boldsymbol{p}_a$

- Recall that a transpose swaps rows/columns
  - Denoted mathematically as $\boldsymbol{x}^T$ or sometimes $\boldsymbol{x}'$

Some Math Notation For Ye'

- Recall the **LENGTH** of any vector x is $||x||$
  - Can be considered the "Pythagorean" or Euclidian distance from the origin to a point at location x in N-space

Somewhere Aaron Childs is beaming proudly

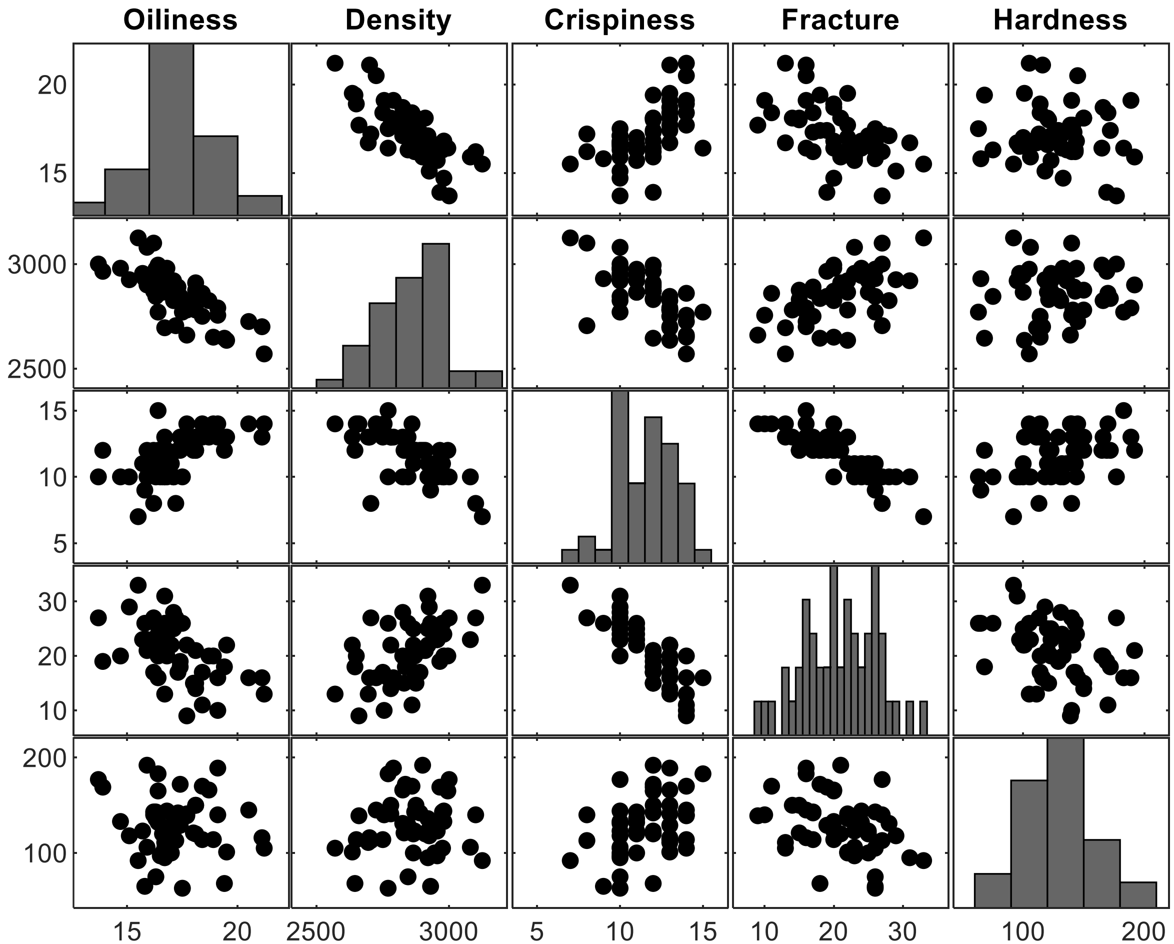
$$||x|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2 + \cdots + x_N^2}$$

$$||x|| = \sqrt{\sum_i x_i^2}$$

$$||x|| = \sqrt{x^T x}$$

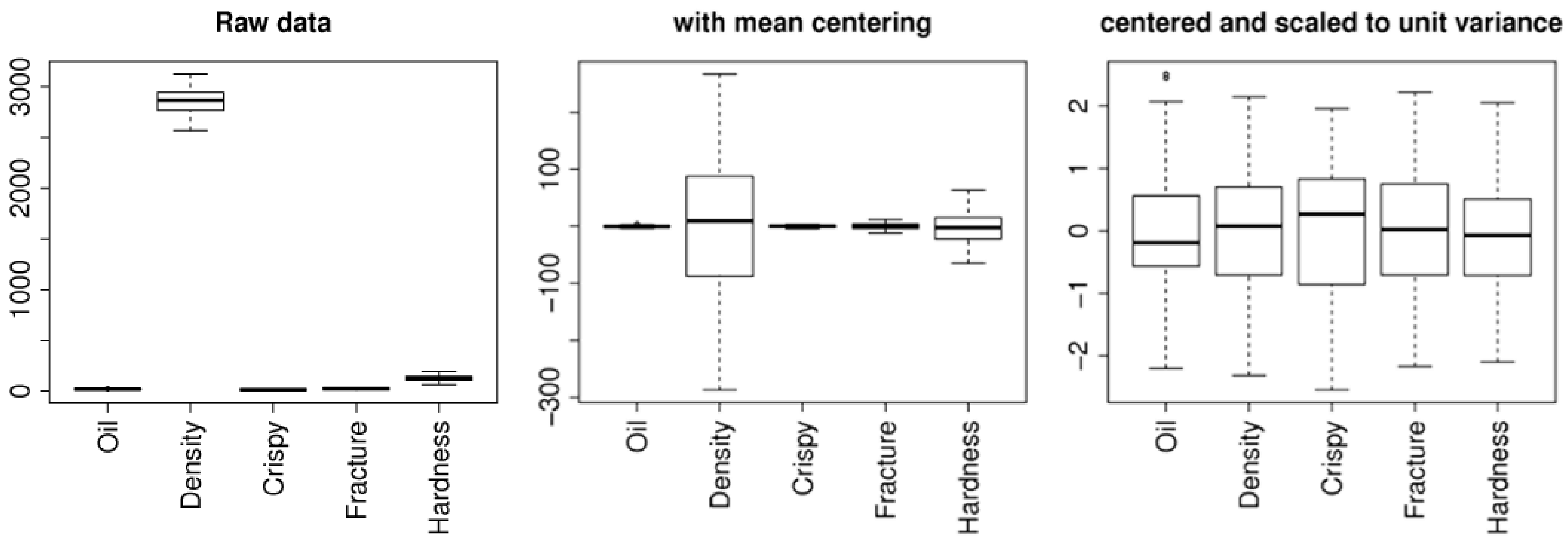
Preprocessing Example

- Consider the following data available for this course
  - Measures **five** properties of **50** pastry samples



Preprocessing Example

- Our first step is to **center** and **scale** the data
  - Discussion: Why?

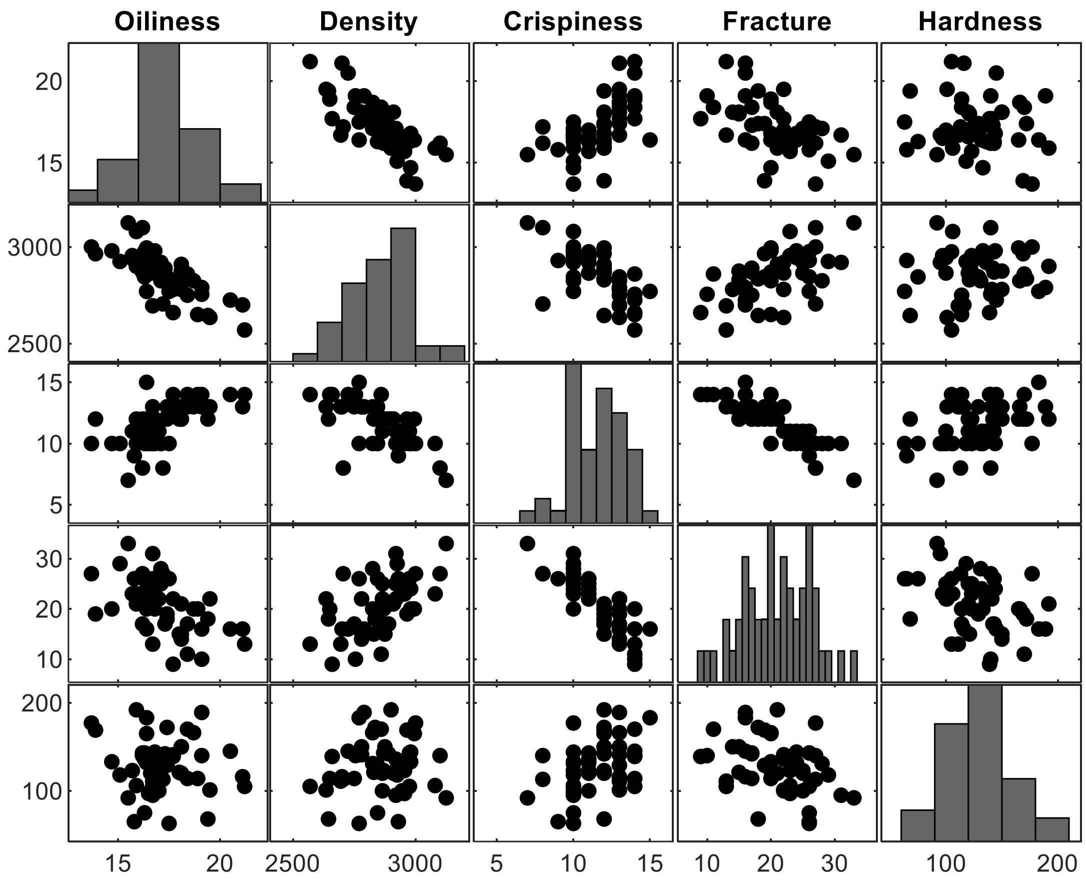


  - Centering: $\boldsymbol{x}_{k,center} = \boldsymbol{x}_{k,raw} - \bar{\boldsymbol{x}}_{k,raw}$ ($\bar{\boldsymbol{x}}_k$ is mean of $\boldsymbol{x}_k$)
  - Scaling: $\boldsymbol{x}_k = \dfrac{\boldsymbol{x}_{k,center}}{SD(\boldsymbol{x}_{k,center})}$
  - Does not change relationship between variables

Figures courtesy of ConnectMV

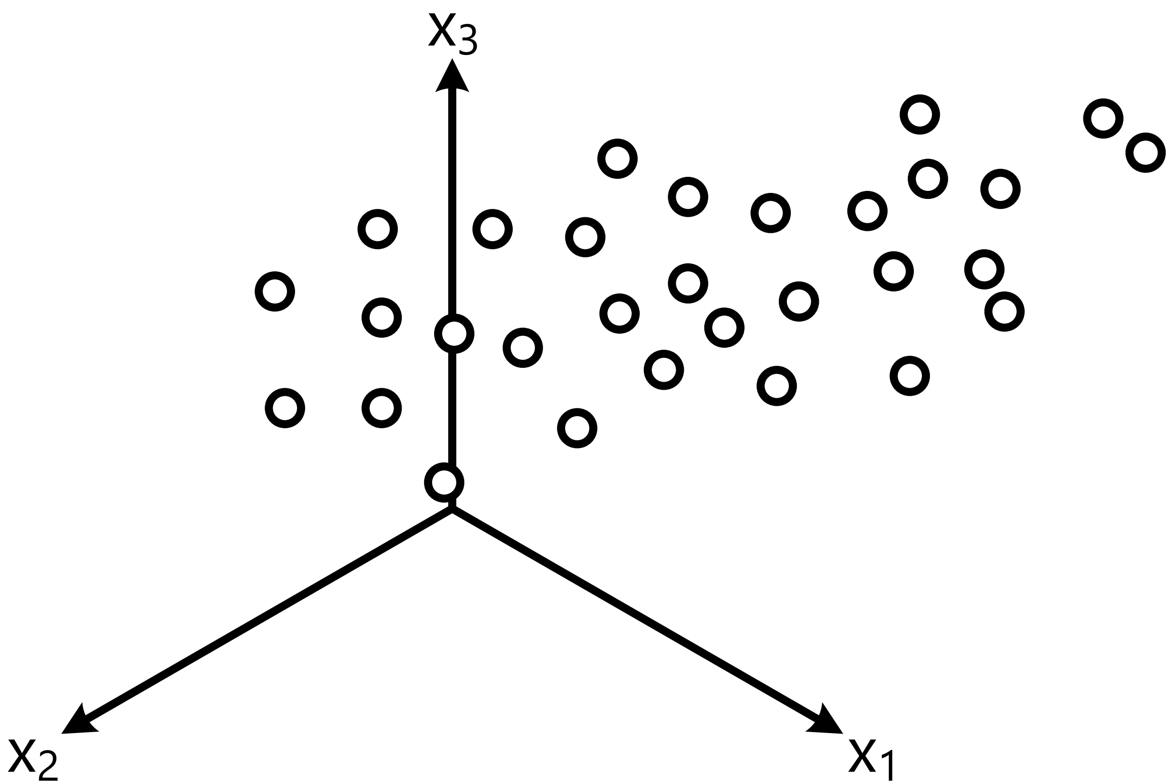
Preprocessing Example

- Does not change relationship between variables…
  - Only the **absolute scale** of them!



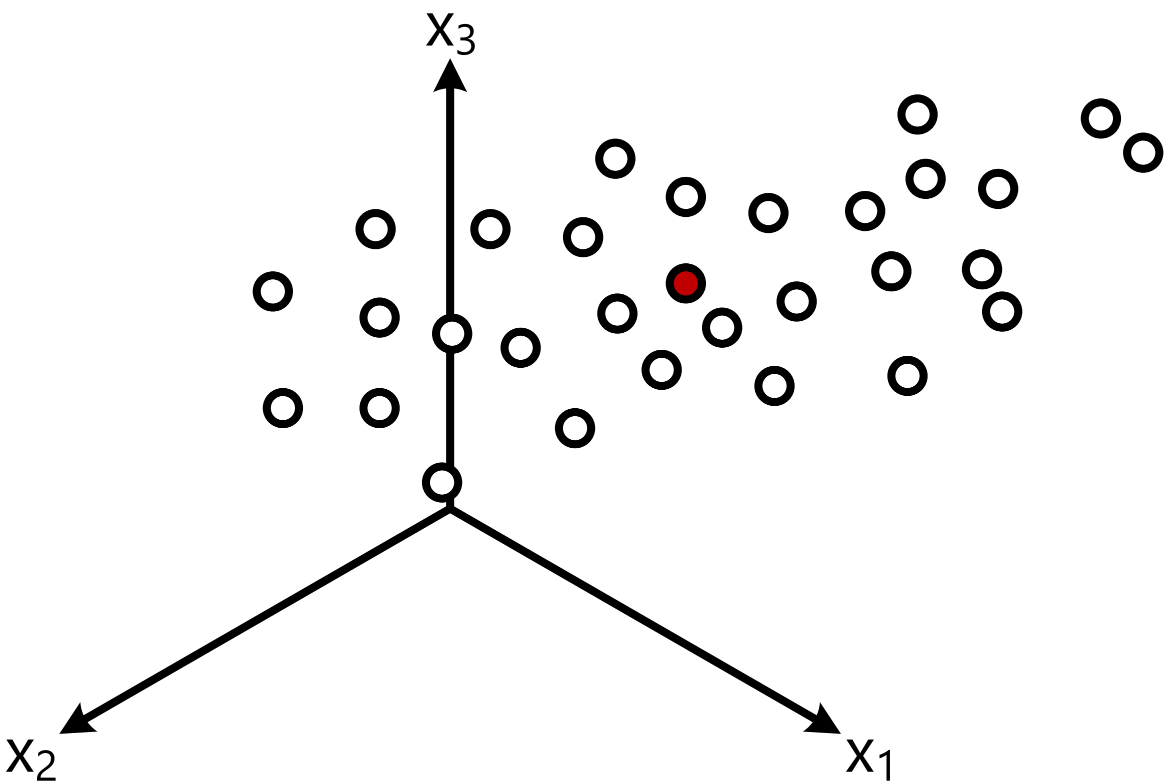
Geometric Interpretation

- Our original data cloud is "somewhere" in N-space
  - Really, this just means each column has its own units
  - Ex: Temperature, vibration, image, and concentration data



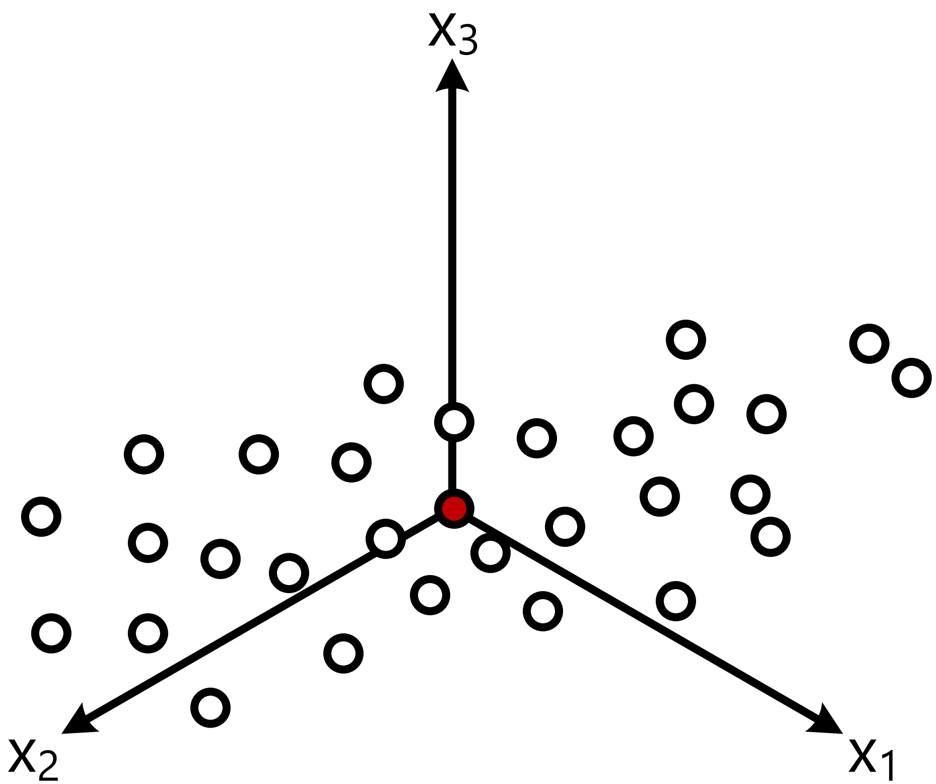
Geometric Interpretation

- We want to find the “center” of that data cloud
  - Represents the average of each dimension in N-space



Geometric Interpretation

- Then move the locus to the origin
  - Centering: move data cloud position (same locus or center)
  - Scaling: change axis proportions so all data treated equally



Geometric Interpretation of PCA

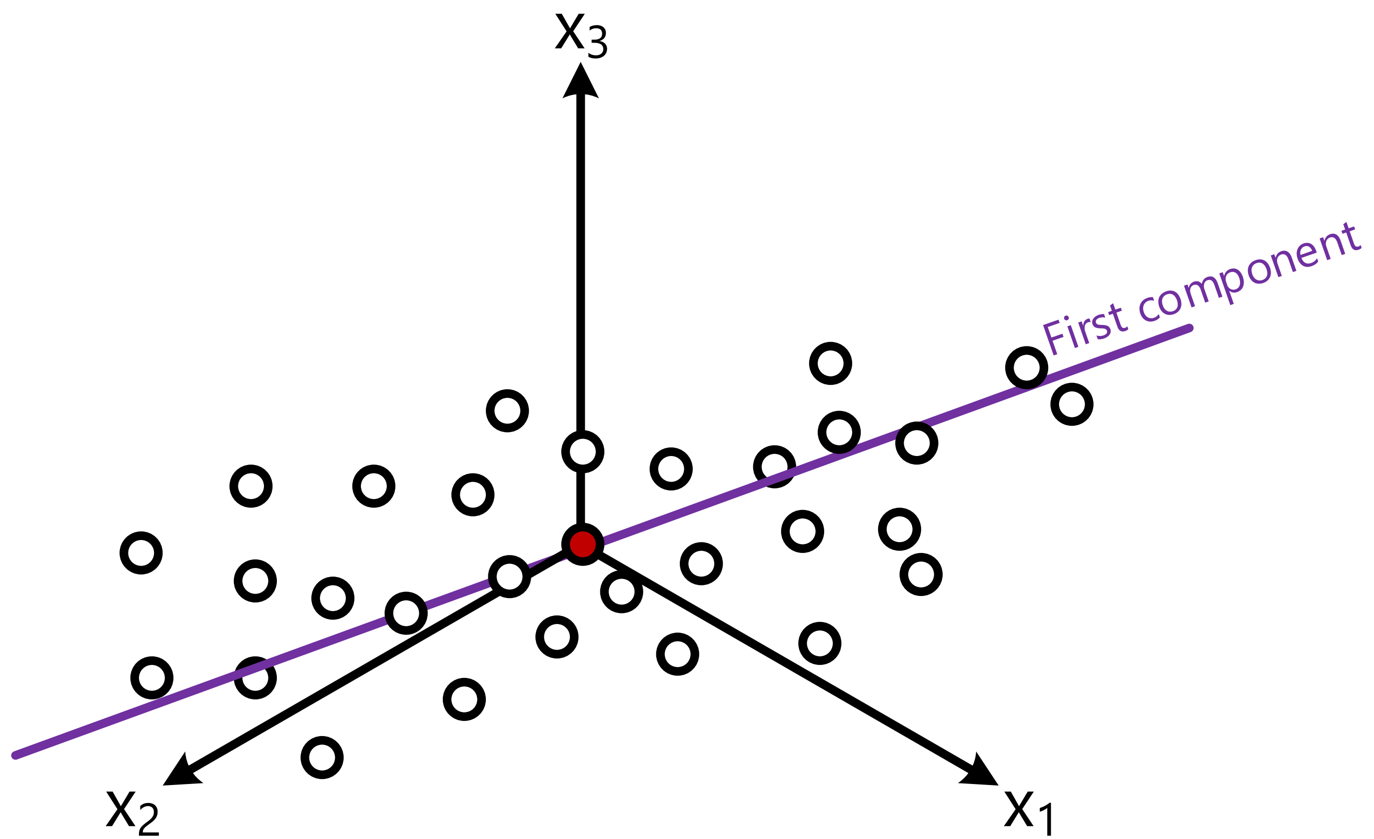
A Picture is Worth a Thousand Words*

*Step aside, *Ulysses*... Comic books are the new epic

https://en.wikipedia.org/wiki/The_School_of_Athens

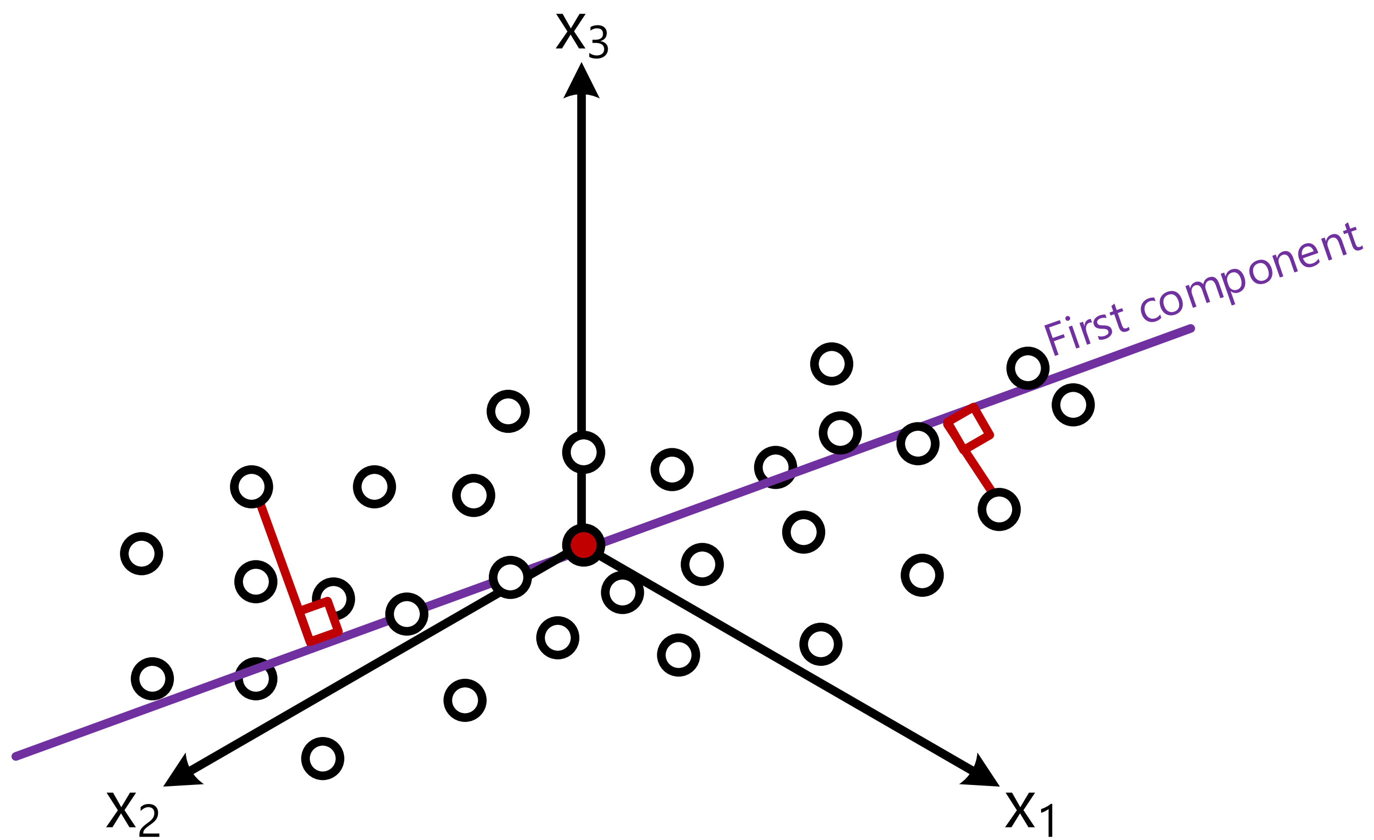
Geometric Interpretation of PCA

- Fit a line (straight) through the points in direction of greatest variance
  - In other words, minimizing errors. How to compute those…?



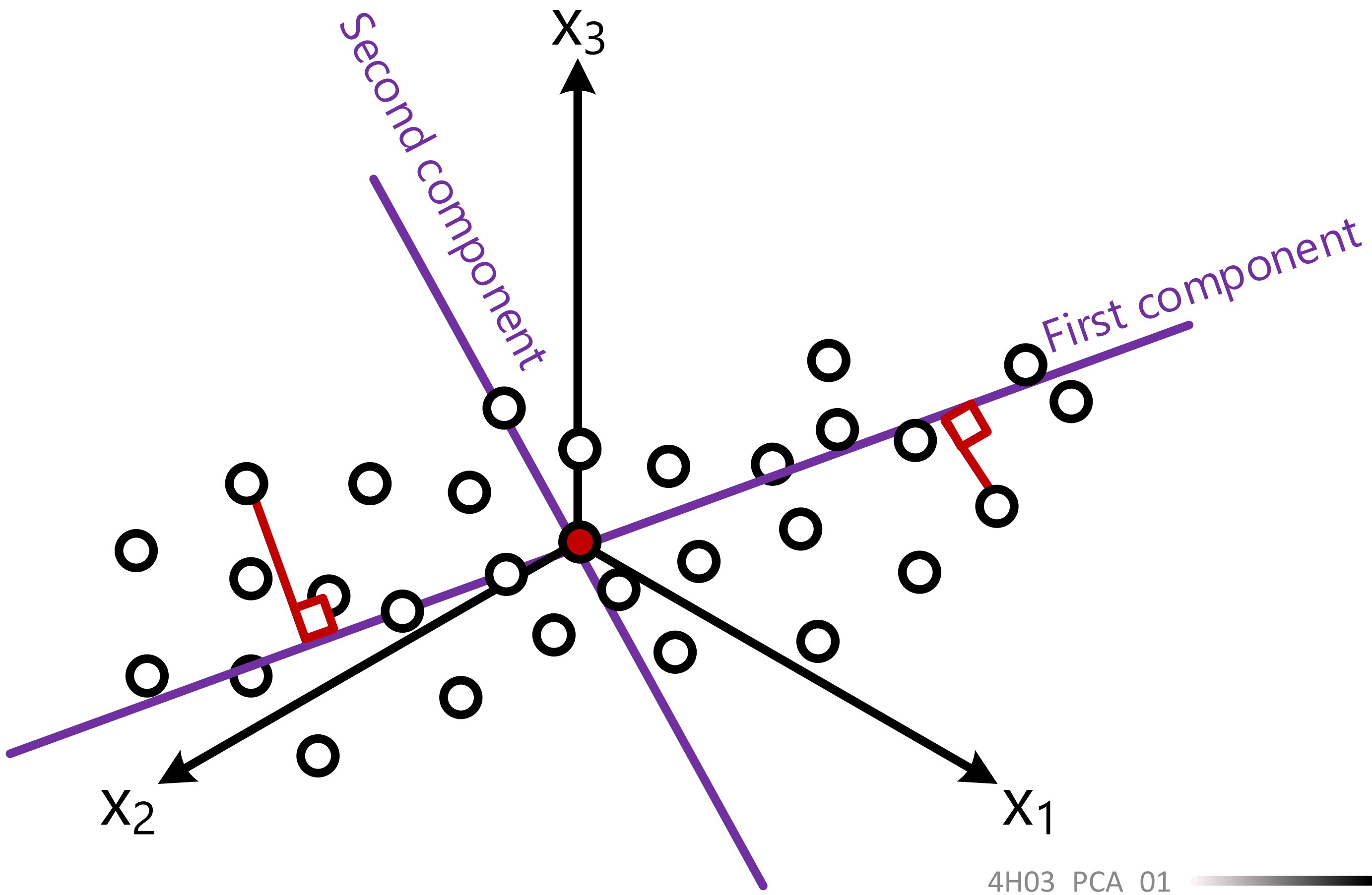
Geometric Interpretation of PCA

- Project each point onto the model component (90°)
  - Unsurprisingly, the distance from the model to each point will be classified later as the **error** of each point on the model



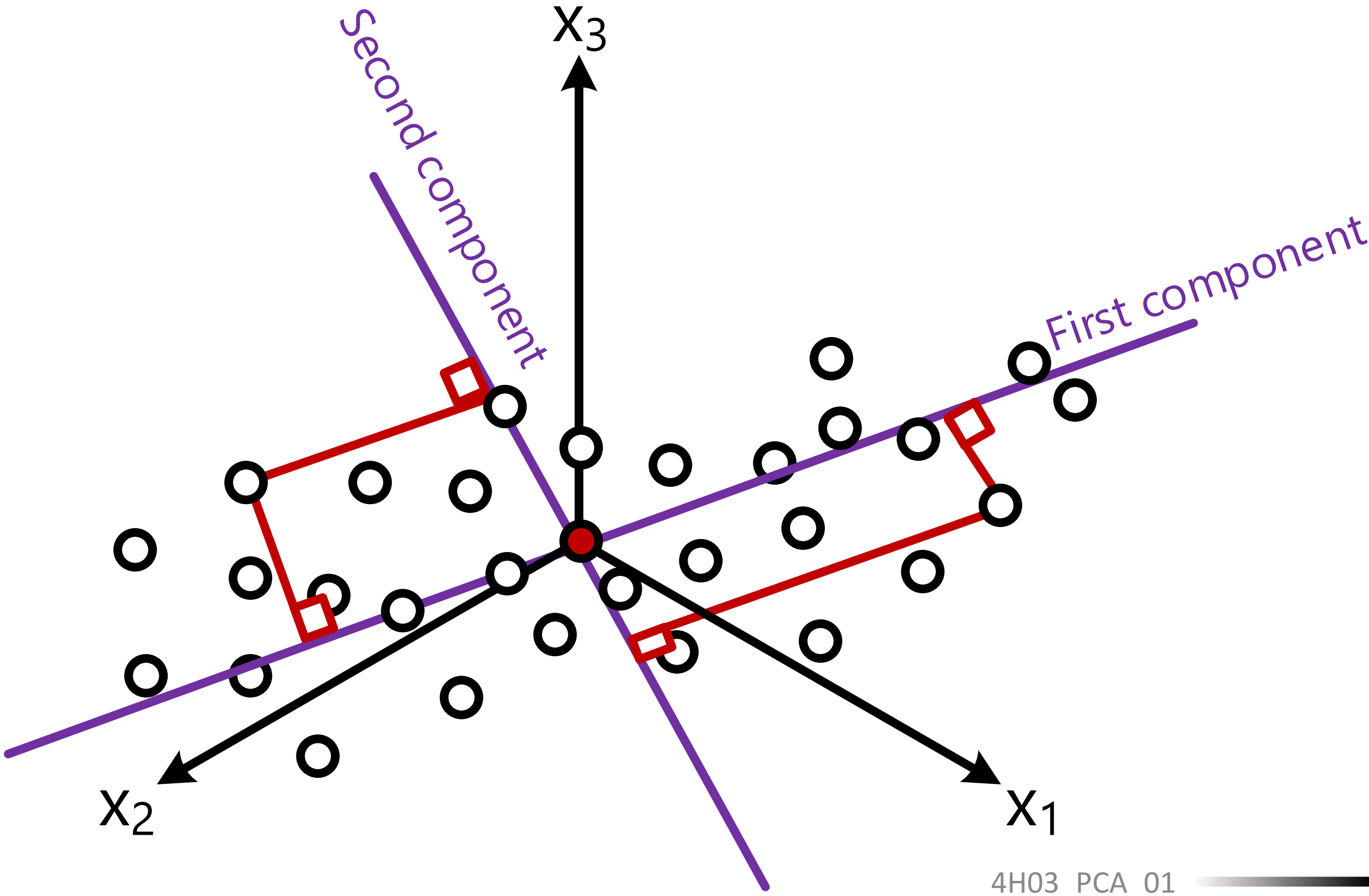
Geometric Interpretation of PCA

- Fit a line as the second component that best fits the data AND is orthogonal (perpendicular) to first



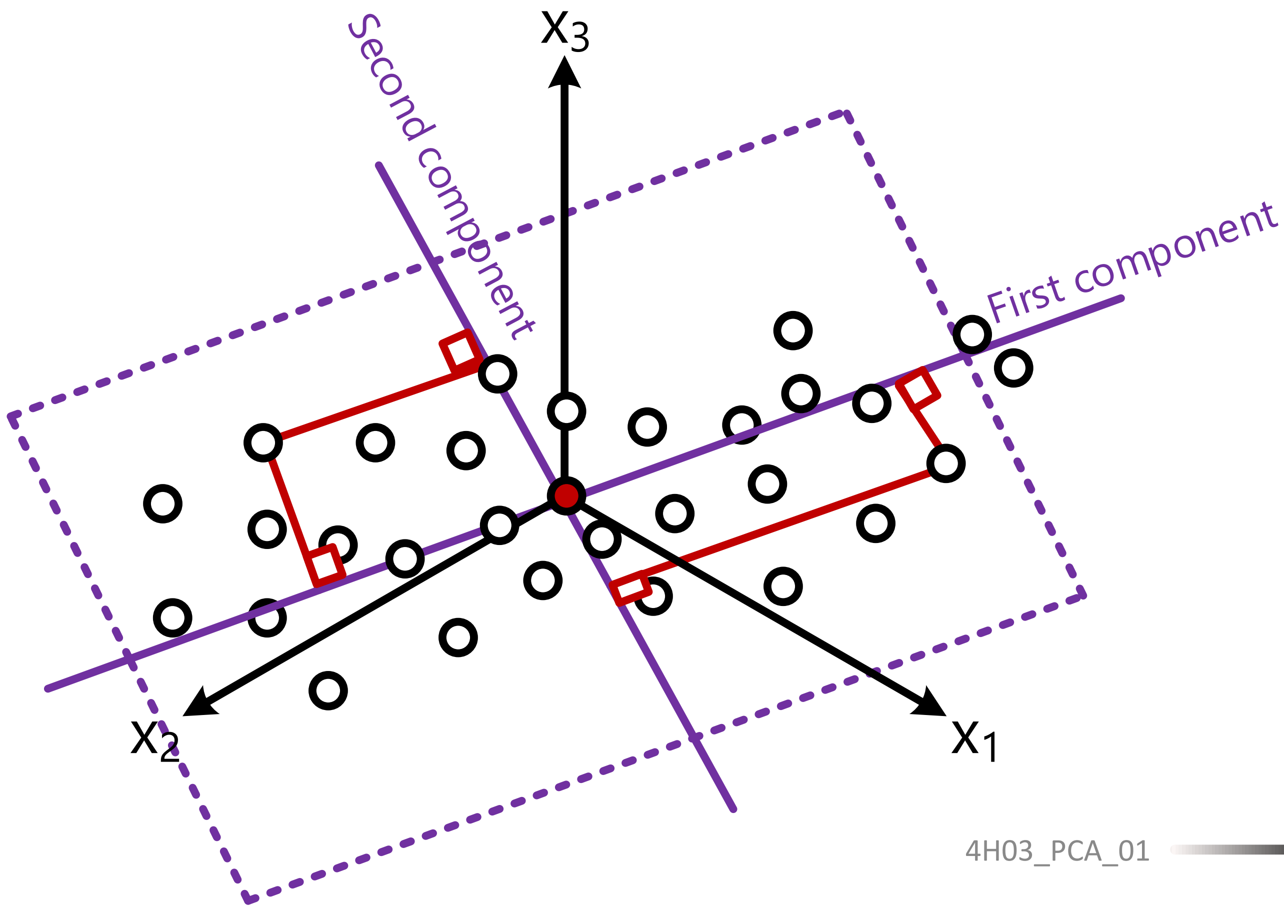
Geometric Interpretation of PCA

- Project points onto the second component
  - You might imagine that the errors of this component should be higher (less explained variance)



Geometric Interpretation of PCA

- The two components make a 2D plane
  - This is called a 2D "subspace" of the 3D space
  - I'll gently point out here that this works in N dimensions



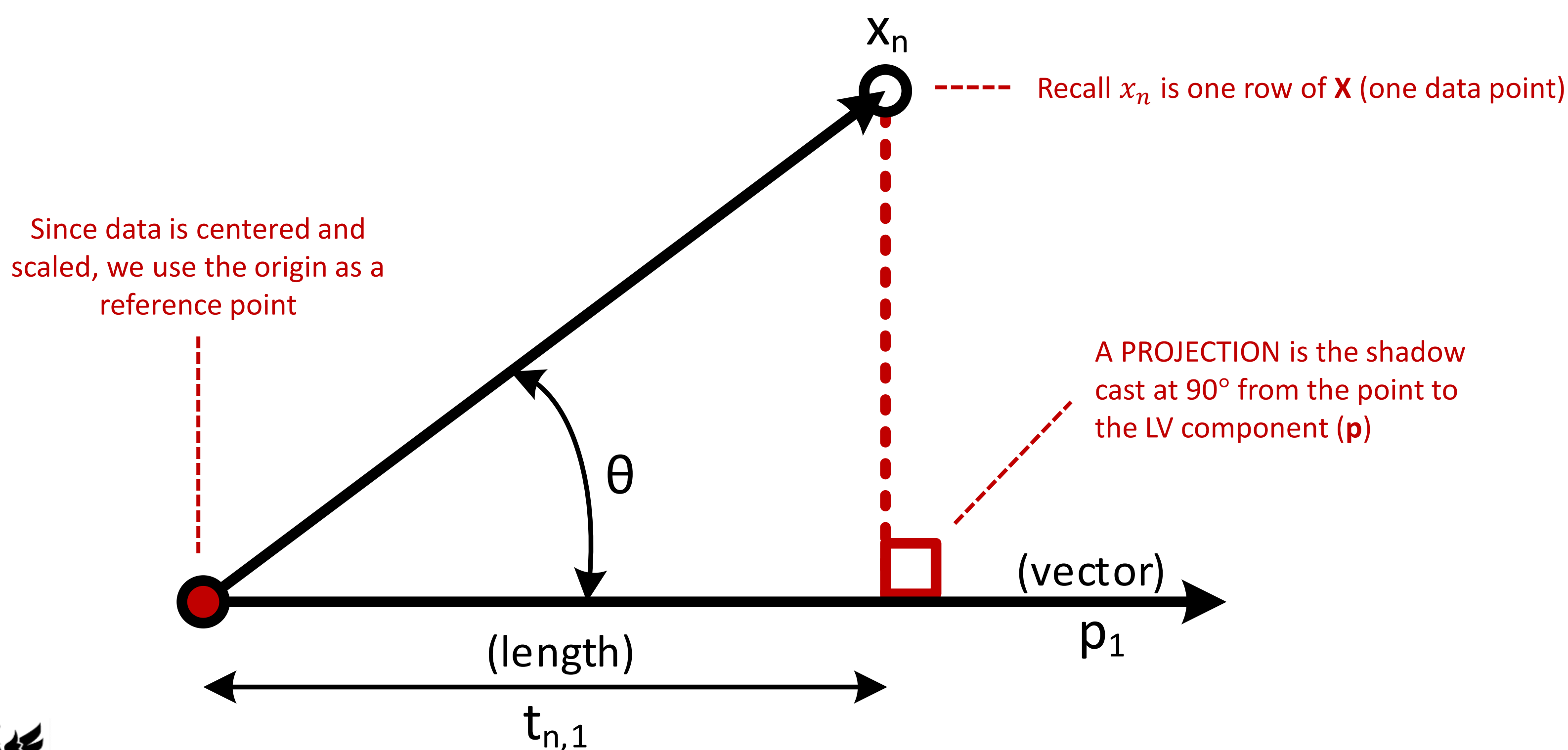
Mathematical Derivation of PCA

What, 3D is not good enough for you?

https://www.models-resource.com/resources/big_icons/10/9447.png

Mathematical Derivation of PCA

- So what have we done?
- We have broken **X** down into two parts:
  - Model predictions (projected points **on the plane**)
  - Residual distance (distance from true point to the plane)



Mathematical Derivation of PCA

- Remember SOHCAHTOA?

$$\cos(\theta) = \frac{ADJ}{HYP} = \frac{t_{n,1}}{\|\boldsymbol{x}_n\|}$$

OK, NOW Aaron Childs is proud of me

Also note here that for consistency, each vector is assumed to be a column

- Remember the definition of a dot product?

$$\boldsymbol{x}_n \cdot \boldsymbol{p}_1 \equiv \boldsymbol{x}_n^T \boldsymbol{p}_1 = \|\boldsymbol{x}_n\| \|\boldsymbol{p}_1\| \cos(\theta) \Rightarrow \cos(\theta) = \frac{\boldsymbol{x}_n^T \boldsymbol{p}_1}{\|\boldsymbol{x}_n\| \|\boldsymbol{p}_1\|}$$

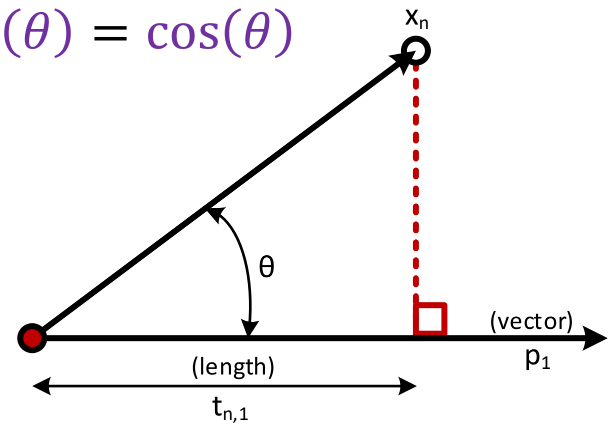
- If we combine these expressions together as $\cos(\theta) = \cos(\theta)$

$$\frac{t_{n,1}}{\|\boldsymbol{x}_n\|} = \frac{\boldsymbol{x}_n^T \boldsymbol{p}_1}{\|\boldsymbol{x}_n\| \|\boldsymbol{p}_1\|}$$

IMPORTANT – by definition, we assign $\boldsymbol{p}_a$ to have UNIT LENGTH. That is, $\|\boldsymbol{p}_a\| = 1$

$$\boldsymbol{t_{n,1} = x_n^T p_1}$$

$$(1 \times 1) = (1 \times k) \times (k \times 1)$$

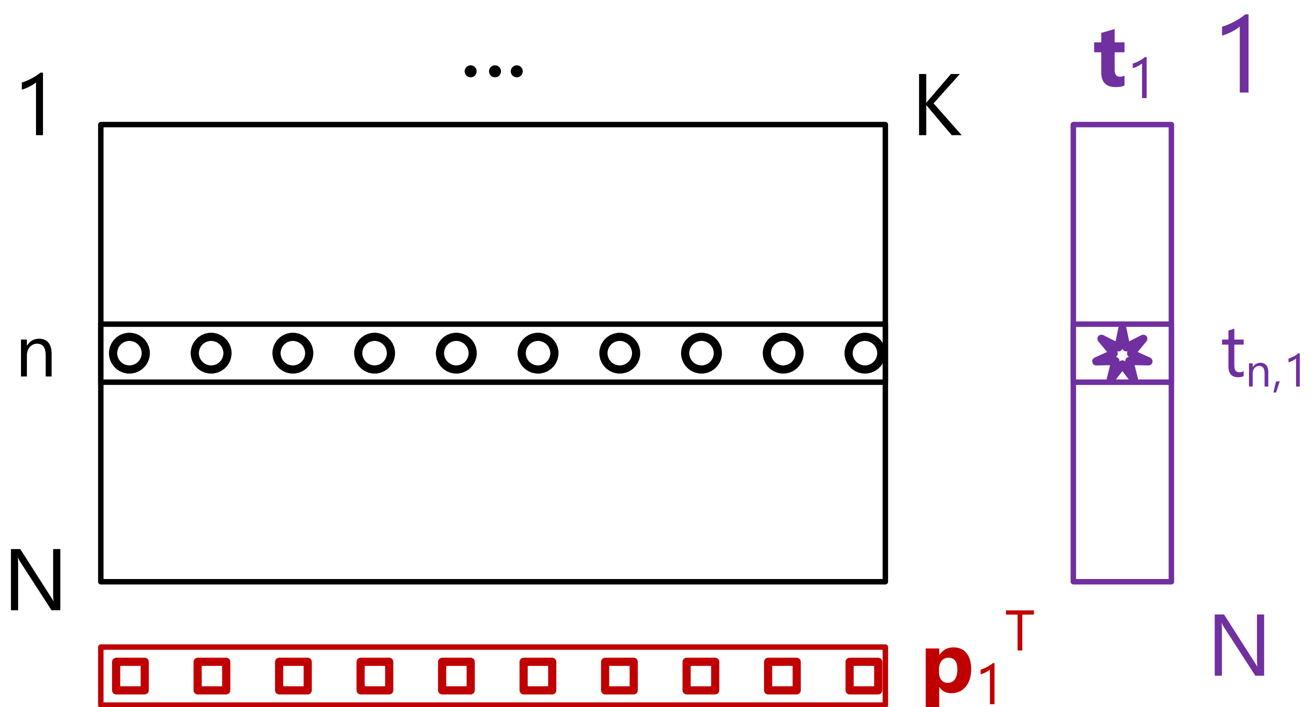


Mathematical Derivation of PCA

$$t_{n,1} = \boldsymbol{x}_n^T \boldsymbol{p}_1$$

$$t_{n,1} = x_{n,1}\, p_{1,1} + x_{n,2}\, p_{2,1} + \cdots + x_{n,k}\, p_{k,1} + \cdots + x_{n,K}\, p_{K,1}$$

- K terms add up as a **linear combination** to form $t_{n,1}$
- The entire first score vector is therefore $\boldsymbol{t}_1 = X\, p_1$



Workshop

$$t_{n,1} = x_{n,1}\, p_{1,1} + x_{n,2}\, p_{2,1} + \cdots + x_{n,k}\, p_{k,1} + \cdots + x_{n,K}\, p_{K,1}$$

- Given the following:
  - Values in $\boldsymbol{x}_n^T$ are centered and scaled
  - Entries in $\boldsymbol{p}_1$ are between -1 and 1
- How would you…
  - Get a large positive value of $t_{n,1}$?
  - Get a large negative value of $t_{n,1}$?
- What can you say about…
  - Two observations (rows) of **X** $\{15{,}30\}$ if $t_{15,1} \approx t_{30,1}$?
  - An observation with $t_{n,1} \approx 0$?

Graphical Tools for PCA Analysis

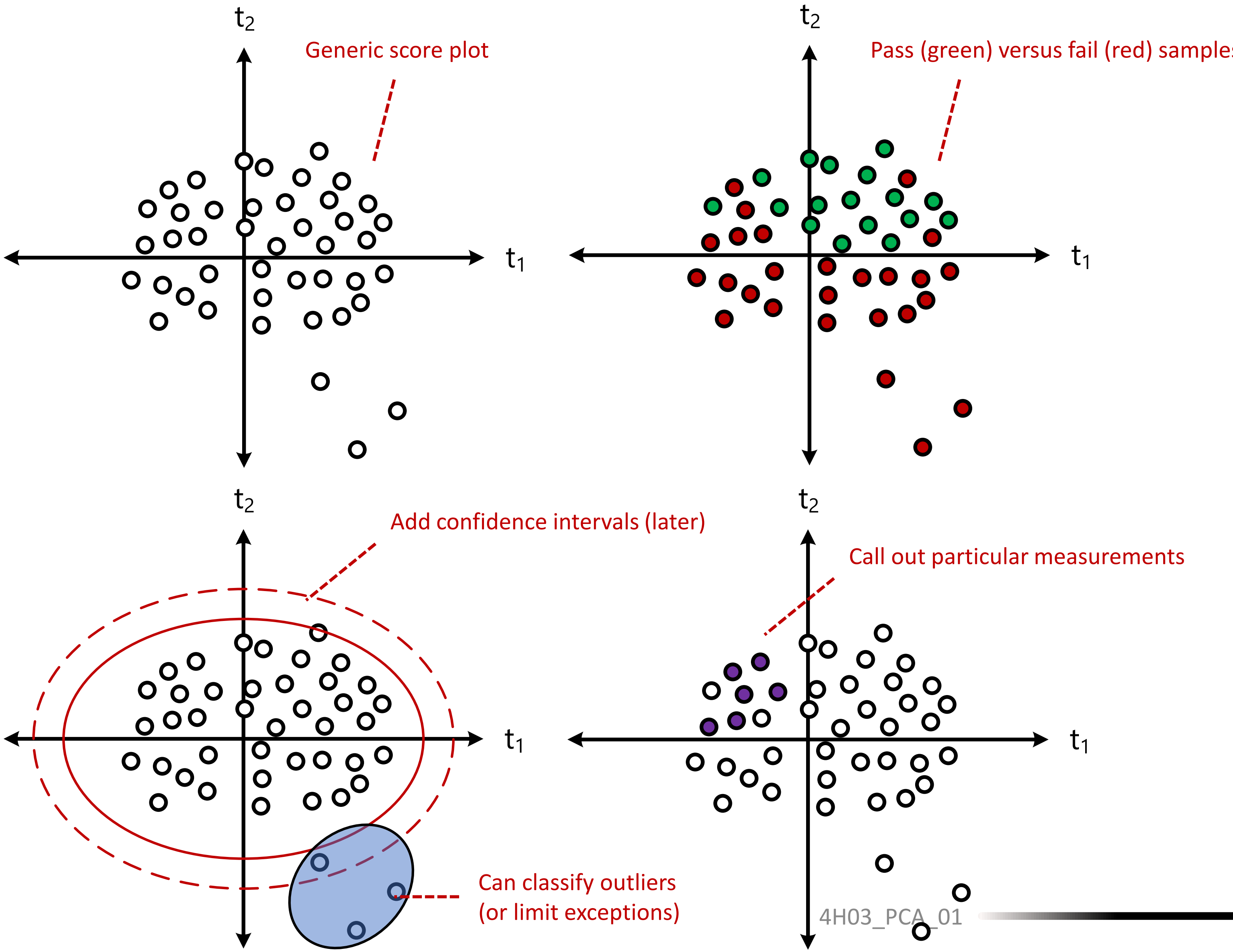
Know the Score

https://d2me2qg8dfiw8u.cloudfront.net/content/uploads/2019/07/16113817/Wimbledon-final-scoreboard-from-PA-752x428.jpg

Score Plots

- Represent a convenient visualization of two scores
  - 3D score plots are possible but a little harder to parse
- Plot any subset (or all) observations as a **scatterplot**
  - The scatterplot can plot t_1 versus t_2, t_2 versus t_3…
  - That is, **entire rows of $\boldsymbol{t}$**
- Can then perform some useful visualization methods
  - Tag outlier data
  - Tag as pass/fail
  - Identify competitor products
  - Can be combined with **loadings plots**

Score Plot Examples

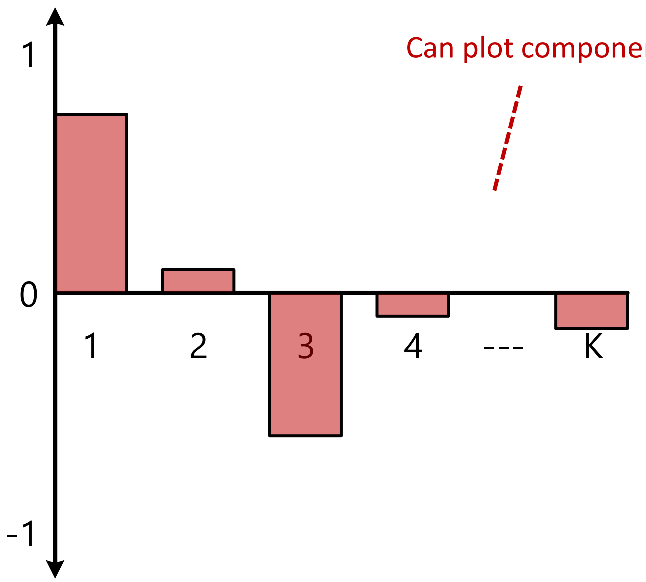


Loadings Plots

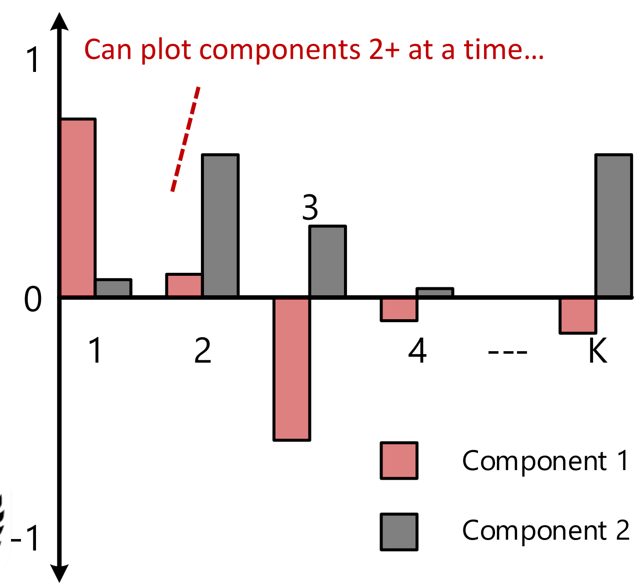
- Represent the loadings vectors in **P** ($\boldsymbol{p}_a$)
- Usually plotted one at a time
- Frequently visualized as a bar plot
- Offers convenient visualization of “important” columns in **X** for a given latent variable (component)
  - Recall, $\boldsymbol{p}_a$ is a vector that represents the “coefficients” of $t_{n,1} = x_{n,1}\, p_{1,1} + x_{n,2}\, p_{2,1} + \cdots + x_{n,k}\, p_{k,1} + \cdots + x_{n,K}\, p_{K,1}$
- May be combined visually with a score plot

Loadings Plot Examples

Can plot components one at a time…

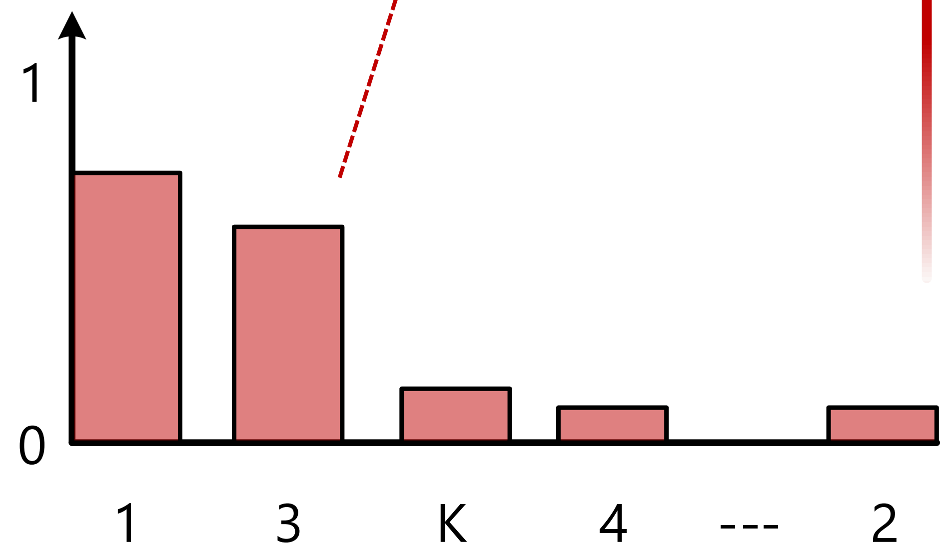


Can plot components 2+ at a time…



Can even plot a component in absolute-value descending (or ascending) order to emphasize what is important.

This is generally known as a **pareto plot**



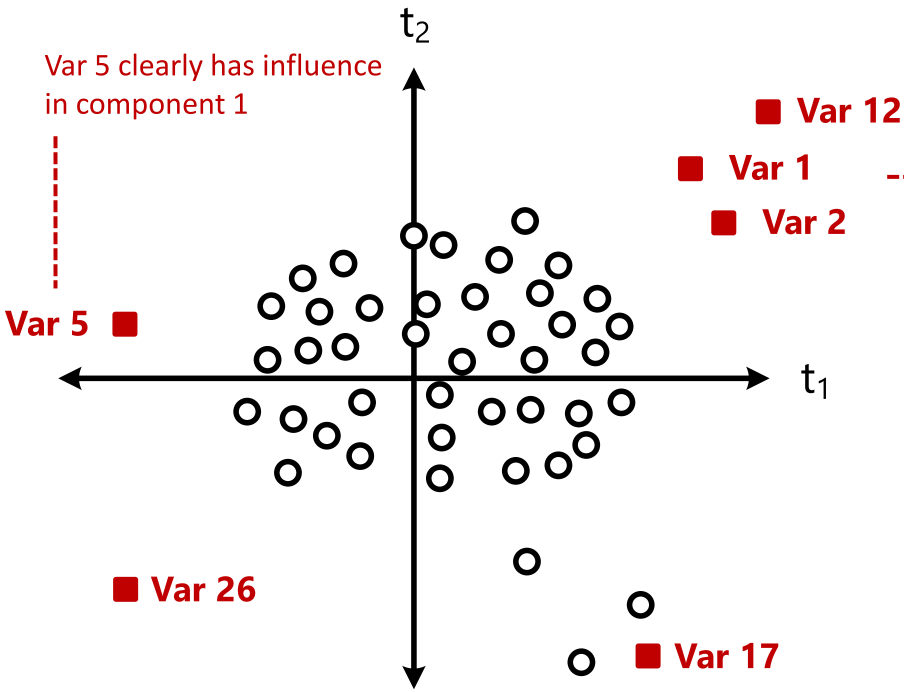
Wait… Why is the maximum value for these plots ± 1 again?

Monitoring and Combined Plots

- **Combined** plots attempt to display multiple inferences on the same axes
  - For example, you may plot the **loadings** as vectors in the **contribution plots** to help visualize what might lead to certain observations (see pastry examples)
  - Also may let you identify "problem" variables
- **Monitoring** plots (such as Dofasco *CasterSOS* or the latent temperature variables) display an easy-to-read score as a timeseries
  - If the loadings for that score are known (spoiler: they are), the monitoring plot can be used to quickly diagnose problems in the real variables

Combined Plot Examples

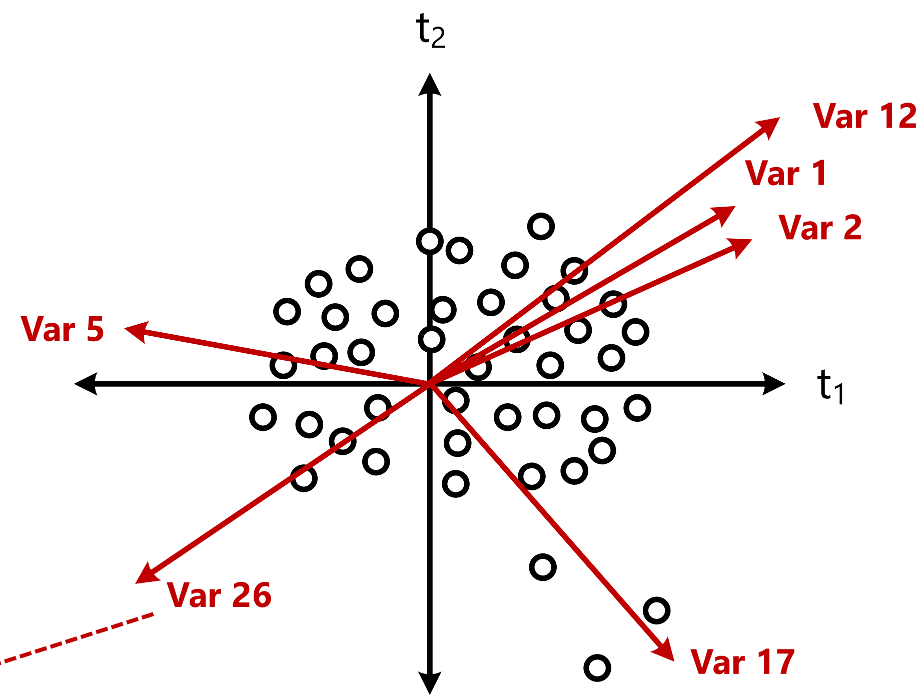
Var 5 clearly has influence in component 1



Vars 1, 2, and 12 all seem highly correlated (temperatures?)

Var 17 seems to be a measurement that contributes to outliers

These are the same plots, just with vectors or points



Var 26 seems to be NEGATIVELY correlated with vars 1, 2 and 12

Class Workshop

Interpreting the Pastry Data

https://d2me2qg8dfiw8u.cloudfront.net/content/uploads/2019/07/16113817/Wimbledon-final-scoreboard-from-PA-752x428.jpg

Pastry Data

- We will work with the pastry data set from the Avenue repository
  - I will post the code for this discussion for you to try after!

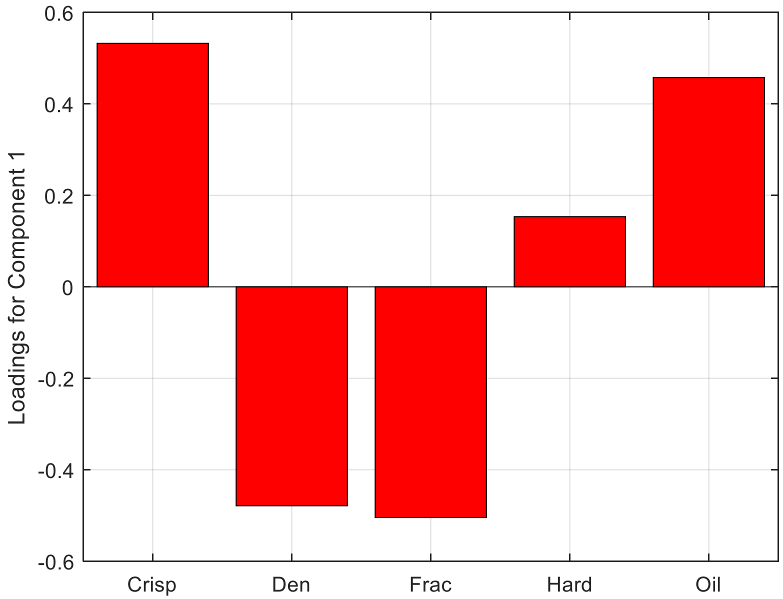
- I have made some plots for the slides but lets go through the MATLAB code together

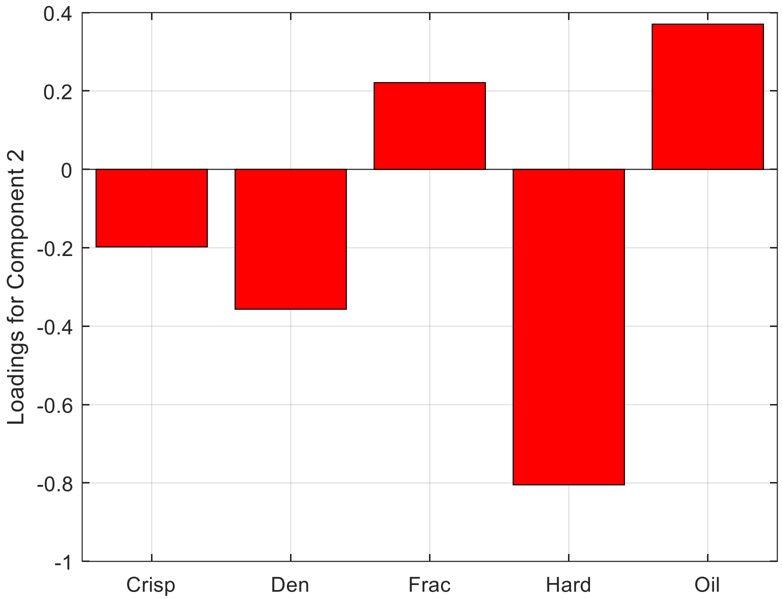
Producing a Score

- First, let's reproduce $t_{17,1}$ for pastry 17
  - x_{raw} = [18.9 2650   14   20   114]'
  - p_1 = [0.478 -0.479 0.532 -0.507 0.153]'
  - $\bar{x}_{raw}$ = [17.2   2857.6   11.5   20.9   128.2]'
  - $SD(x_{raw})$ = [1.59   124.5 1.78   5.47   31.1]'

Loadings Plot

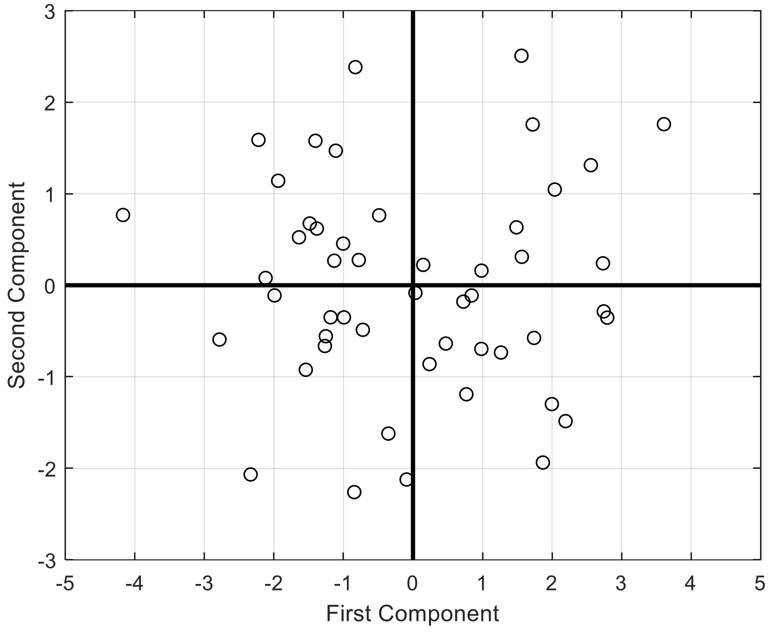
- What can we learn from these loadings plots?

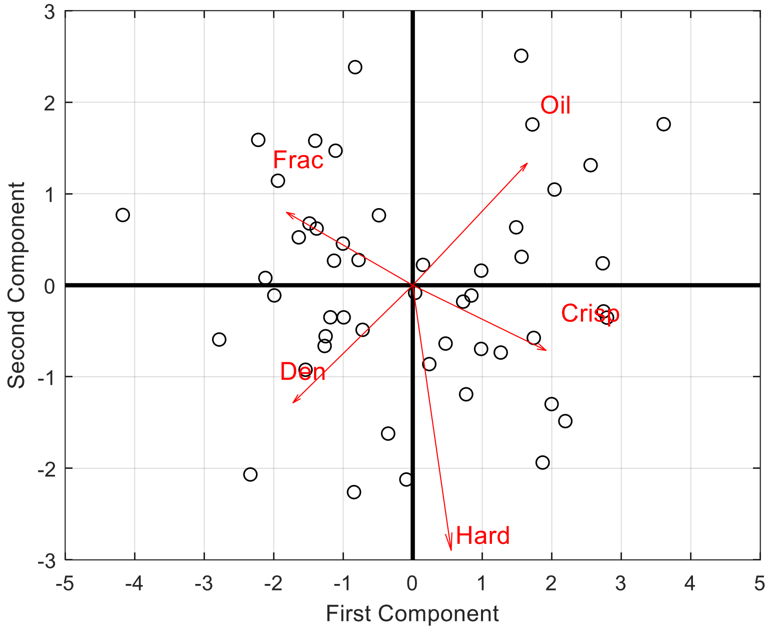




Contribution Plots (I)

- What can we learn from these contribution plots?





Contribution Plots (II)

- What can we learn from this contribution plot?
  - Green circles are those that customers rated ≥ 4/5
  - What kinds of pastries do customers tend to like?