Dimensionality Reduction

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

Dimensionality Reduction Outline

- Definition
- When to Use
- Mathematical Overview
- Key Concepts
- Examples
  - Diet in the UK
  - Genetics around the world

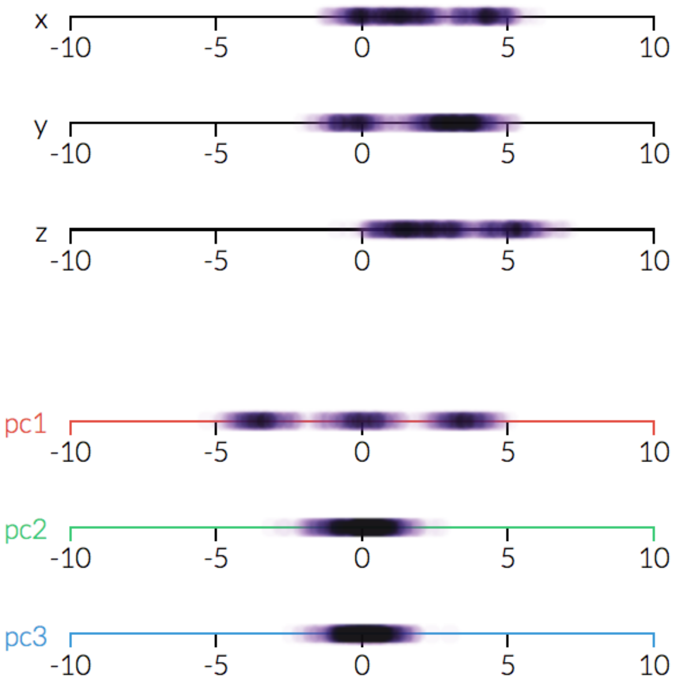
Dimensionality Reduction

A mathematical process to reduce the number of random variables to consider

Discuss: why may we want to do this?

Dimensionality Reduction

- Reduce the dimension of quantiative data to a more manageable set of variables
- Reduced set can then be input to reveal underlying patterns in the data and/or as inputs in a model (regression, classification, etc.)



Use Cases for Dimensionality Reduction

- Thousands of sensors used to monitor an industrial process
    - Reducing the data from these 1000s of sensors to a few features, we can then build an interpretable model
    - Goal : predict process failure from sensors

- Understanding diet around the world
    - Amount of foods eaten among populations across the world
    - Goal: identify diet similarity among populations

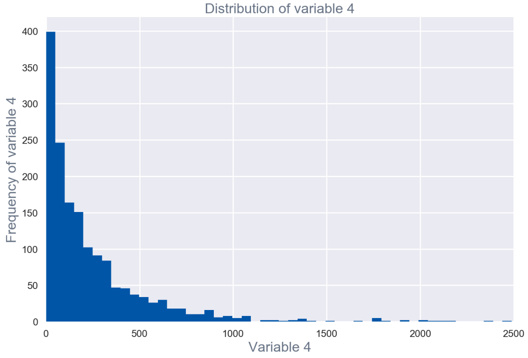
- Identify genetic ancestry
    - Determine ancestral origins based on genetic variation
    - Goal: Learn more about our genetic history

As an extension of EDA

- Gain insight into a set of data
- Understand how different variables relate to one another

EDA Approaches to “Get a Feel for the Data”

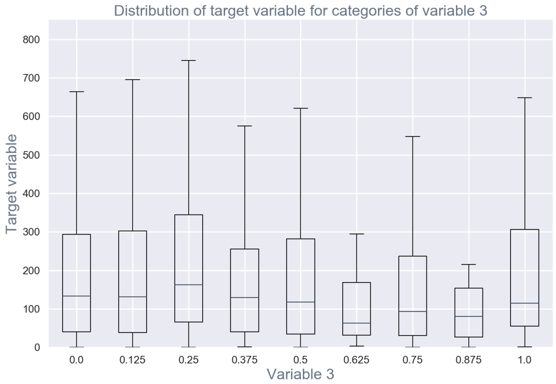
Understanding the relationship between variables in your dataset



Univariate

understanding a single variable

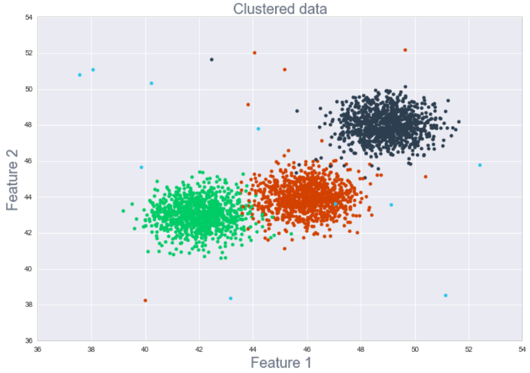
i.e.: histogram, densityplot, barplot



Bivariate

understanding relationship between 2 variables

i.e.: boxplot, scatterplot, grouped barplot, boxplot



Dimensionality Reduction

projecting high-D data into a lower-D space

i.e.: PCA, ICA, Clustering

Image source: https://www.svds.com/value-exploratory-data-analysis/

As an extension of EDA

- Gain insight into a set of data
- Understand how different variables relate to one another

Note: PCA/Dimensionality reduction can also be used for modeling & prediction

Key Terms:

- **Principal Component (PC)** - a linear combination of the predictor variables
- **Loadings** - the weights that transform the predictors into components (aka weights)
- **Screeplot** - Variance explained of each component

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Goal : combine multiple numeric predictor variables into a smaller set of variables. Each variable in this smaller set is a weighted linear combination of the original set.

This smaller set of variables -- the *principal components* (PCs) - “explain” most of the variability of the full set of variables….but uses many fewer dimensions to do so.

The weights (loadings) used to form the PCs explain the relative contributions of the original variables to the new PCs.

“Simple” PCA : Two predictor variables (X_1 and X_2)

For two variables, X_1 and X_2, there are two principal components Z_i ($i = 1$ or 2):

$$Z_i = w_{i,1}X_1 + w_{i,2}X_2$$

$w_{i,1}$ and $w_{i,2}$: weightings (*loadings*)

- Transform the original variables into principal components

Z_1 : the first principal component (PC1)

- The linear combination that best explains the total variance

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Stock Price returns for Chevron (CVX) and ExxonMobil (XOM)

PC1 and PC2 are the dotted lines on the plot

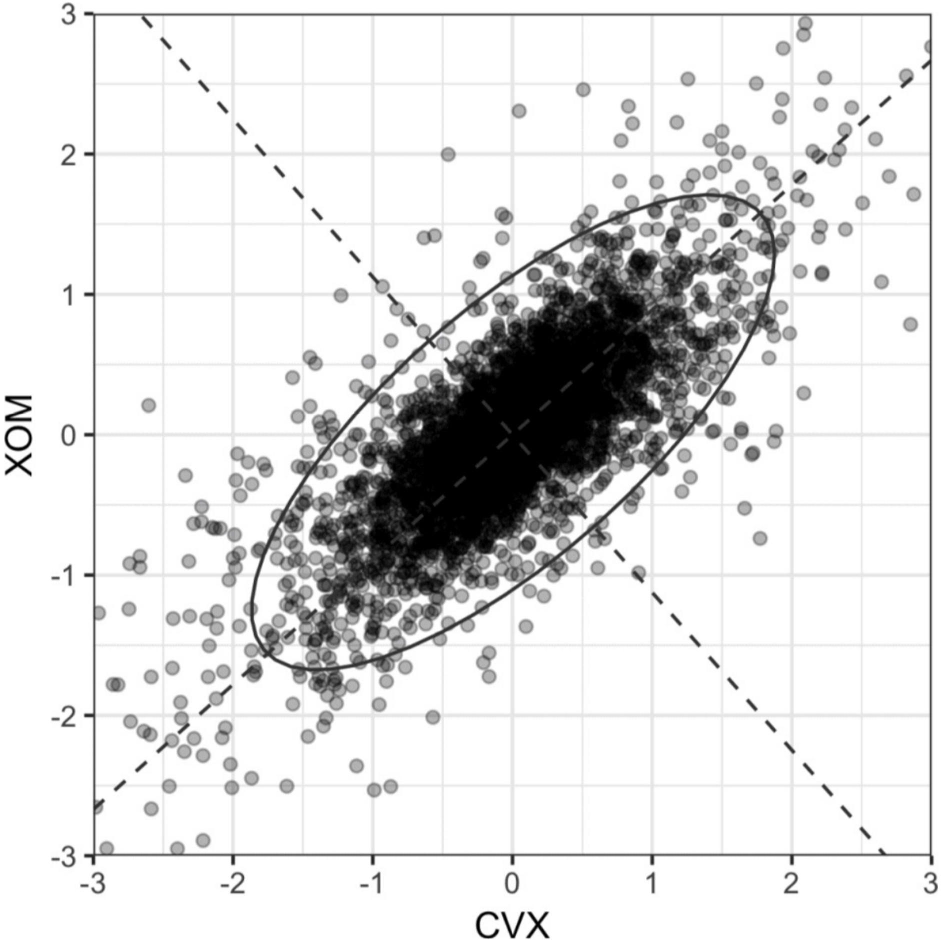
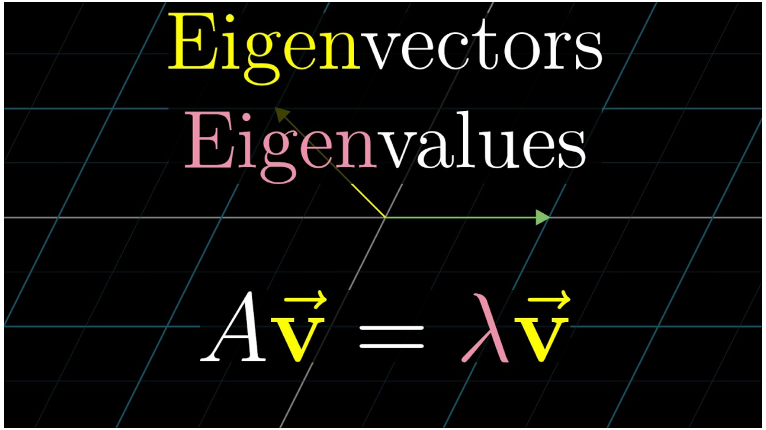


Figure 7-1. The principal components for the stock returns for Chevron and ExxonMobil



PCA

1. Center the data
2. Compute covariance matrix
3. Compute eigenvectors and eigenvalues of covariance matrix
4. (Optional) subset eigenvectors by picking the m largest eigenvalues
5. Project datapoints into PCs by multiplying with eigenvectors

But….PCA shines when you're dealing with high-dimensional data. So we have to move *beyond* two predictors to many predictors….

Step 1: Combine all predictors in linear combination

Step 2: Assign weights that optimize the collection of the covariation to the first PC (Z_1) (maximizes the % total variance explained)

Step 3: Repeat Step 2 to generate new predictor Z_2 (second PC) with different weights. By definition Z_1 and Z_2 are uncorrelated. Continue until you have as many new variables (PCs) as original predictors

Step 4: Retain as many components as are needed to account for *most* of the variance.

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

S&P 500 Data: 5648 days (1993-2015) x 517 stocks

	ADS	CA	MSFT	RHT	CTSH	CSC	EMC	IBM	XRX	ALTR	ADI	AVGO	BRCM	FSLR	INTC	LLTC	MCHP	MU	NVDA
1/29/93	0	0.06012444	-0.0220998	0	0	0.01889746	0.00736807	0.0921652	0.25914009	-0.0071053	-0.0157849	0	0	0	-0.0504878	-0.0898696	0	0.03702057	0
2/1/93	0	-0.180389	0.02762115	0	0	0.01888884	0.01842489	0.11520651	-0.1007745	0.06389288	-0.0157929	0	0	0	0.09536733	0.0449348	0	0.03702038	0
2/2/93	0	-0.1202566	0.03589987	0	0	-0.0755726	0.02948172	-0.0230413	0.02879553	-0.0141924	0.0473628	0	0	0	0	0.0674022	0	0.12340155	0
2/3/93	0	0.0601242	-0.024857	0	0	-0.151128	0.00368875	-0.2534543	-0.04319	-0.0071053	0.20523612	0	0	0	-0.050495	0.0224674	0	-0.0123403	0
2/4/93	0	-0.3607697	-0.0607567	0	0	0.11335029	-0.0221136	0.0698618	0	-0.0070962	-0.0315699	0	0	0	0	0.0224674	0	-0.0740409	0
2/5/93	0	0.03005777	0.09389247	0	0	0.09445283	-0.0479066	0.04657454	0.17276006	-0.0212976	-0.0631478	0	0	0	-0.0476873	-0.0674022	0	-0.0123403	0
2/8/93	0	0.03006643	-0.0607498	0	0	-0.1133503	-0.0110568	0.11643635	-0.04319	0.00709618	0	0	0	0	-0.0196321	-0.1235743	0	-0.0617008	0
2/9/93	0	-0.0901902	-0.063521	0	0	-0.1322391	-0.0147456	0.06986181	-0.115169	0.04969143	-0.0157929	0	0	0	-0.0112235	0.0224674	0	0	0
2/10/93	0	0.12025657	0.02209981	0	0	0.09445283	0.01474557	-0.2561599	0.01439448	0.02838473	0.01578495	0	0	0	0.04487956	0.11233699	0	0.07404095	0
2/11/93	0	0.03005825	-0.0220927	0	0	-0.0188975	0.01474556	-0.1397236	-0.04319	0.02129762	-0.0315699	0	0	0	-0.0532953	0.06740222	0	-0.0246804	0
2/12/93	0	-0.0901901	-0.0358999	0	0	-0.0377863	-0.0073681	-0.0698618	-0.1871546	0	-0.0473628	0	0	0	-0.0336561	-0.112337	0	-0.0370204	0
2/16/93	0	-0.6313411	-0.0607497	0	0	-0.0377863	-0.0479066	-0.0931491	-0.04319	-0.0283938	-0.1262955	0	0	0	-0.098175	-0.1460417	0	-0.0246803	0
2/17/93	0	0.12025657	-0.0165712	0	0	-0.1700254	-0.0110568	0.04657453	-0.08638	-0.0142015	0.03157785	0	0	0	0.04487955	0	0	-0.0123403	0
2/18/93	0	-0.1803808	0.00828562	0	0	-0.0566751	0.00368875	-0.0931491	-0.08638	0	-0.0157849	0	0	0	-0.0168315	0.0224674	0	0.03702056	0
2/19/93	0	0.03006595	-0.0469427	0	0	0	0.00736807	-0.0232873	0.115169	0.01419237	0.03157785	0	0	0	0.10378311	0.15727183	0	0.14808196	0
2/22/93	0	0.03005825	-0.0662782	0	0	-0.1322477	-0.0184249	0.13972361	0	0.02839382	-0.0631557	0	0	0	-0.0168317	-0.0674022	0	0	0
2/23/93	0	-0.0300583	0.03314266	0	0	0	-0.0479066	-0.0698618	-0.1439645	-0.0070962	0.03157785	0	0	0	-0.0336631	-0.0337047	0	-0.0493606	0
2/24/93	0	0.15031459	0.10769942	0	0	0.01888884	0.04421782	0.1397236	0.08638003	0.00709618	0	0	0	0	0.11781411	-0.0224674	0	0.09872137	0
2/25/93	0	0.15032277	0.04142827	0	0	0.01888884	-0.0110568	0.37259628	0	0.02839382	0	0	0	0	0.0112163	0.15727183	0	-0.0370205	0
2/26/93	0	-0.0300659	-0.0193286	0	0	-0.0188888	0.01105682	0.06986181	0.05759106	0	0.01578495	0	0	0	-0.028055	-0.0224674	0	-0.074041	0
3/1/93	0	-0.180381	-0.0497068	0	0	-0.0944614	-0.0073681	0	0.04358505	0.00709618	0.09472561	0	0	0	-0.0336631	0.0224674	0	0	0
3/2/93	0	0	0.06351413	0	0	0.15113659	0.00368875	-0.0698618	0.116229	0	0.11051053	0	0	0	0.09537435	-0.0449348	0	0.12340155	0
3/3/93	0	0.12025658	0	0	0	-0.0566751	0.03684977	0.16301088	0.02905891	-0.0070962	0	0	0	0	0.01402383	-0.112337	0	-0.0370204	0
3/4/93	0	-0.1503146	-0.0220927	0	0	0.0377863	0.00367932	-0.0698618	-0.1452879	-0.0070962	-0.015785	0	0	0	-0.0252473	-0.0674022	0	0.01234016	0
3/5/93	0	0.03005825	-0.0165714	0	0	-0.0944614	0.00368875	-0.0232873	0	0.03549001	0	0	0	0	-0.0617041	0.0449348	0	0.02468042	0
3/8/93	0	0.06012444	0.02209275	0	0	0.01888884	-0.025793	0.11643634	0.21792524	0	0.04736279	0	0	0	0.06731932	0.13480441	0	0.09872114	0
3/9/93	0	0.09019015	0.00552151	0	0	0.09446144	0.00736807	0.09314908	-0.0290523	-0.0070962	-0.0157849	0	0	0	0.0112163	0.0898696	0	0	0
3/10/93	0	0.03006595	0.01104991	0	0	0	0.01105681	-0.1862981	0.02905891	-0.0141924	-0.0157849	0	0	0	-0.0196321	-0.0449348	0	0	0
3/11/93	0	-0.0300583	0.02761408	0	0	0.22670058	0	-0.1862982	-0.0581112	0.00709618	0.06314774	0	0	0	-0.0196392	0.01123011	0	0	0
3/12/93	0	0	0.06627822	0	0	-0.0188975	0.01474556	0.30273448	-0.1452813	0.02839381	0.06314774	0	0	0	0.02524749	0.01123729	0	0.13574153	0

For this example: we'll focus on 16 top companies

Screeplot

The vernacular definition of "scree" is an accumulation of loose stones or rocky debris lying on a slope or at the base of a hill or cliff.

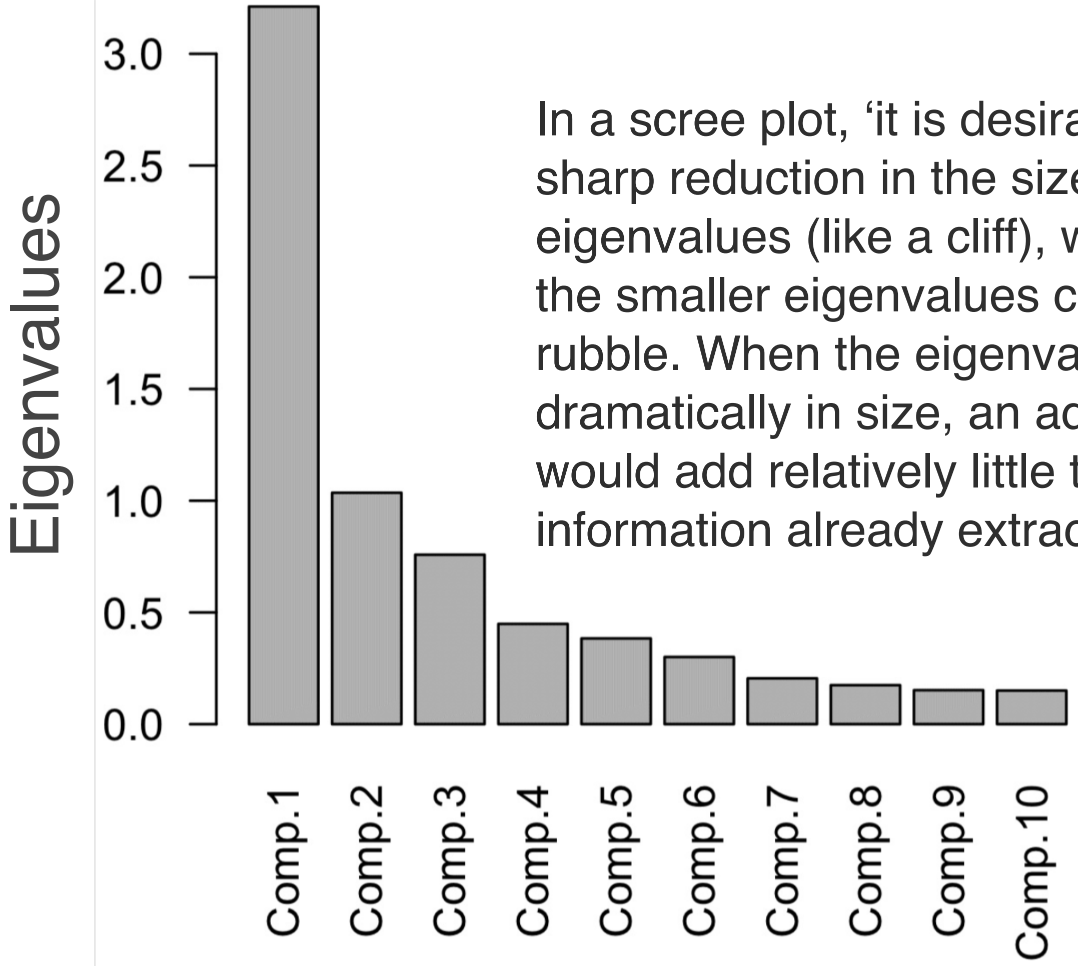


Figure 7-2. A screeplot for a PCA of top stocks from the S&P 500

In a scree plot, 'it is desirable to find a sharp reduction in the size of the eigenvalues (like a cliff), with the rest of the smaller eigenvalues constituting rubble. When the eigenvalues drop dramatically in size, an additional factor would add relatively little to the information already extracted.' (Source)

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Loading of PCs 1-5

PC1: Overall stock market trend

PC2: Price change of energy stocks

PC3: movements of Apple and CostCo.

PC4: movements of Schlumberger to other stocks

PC5: Financial companies

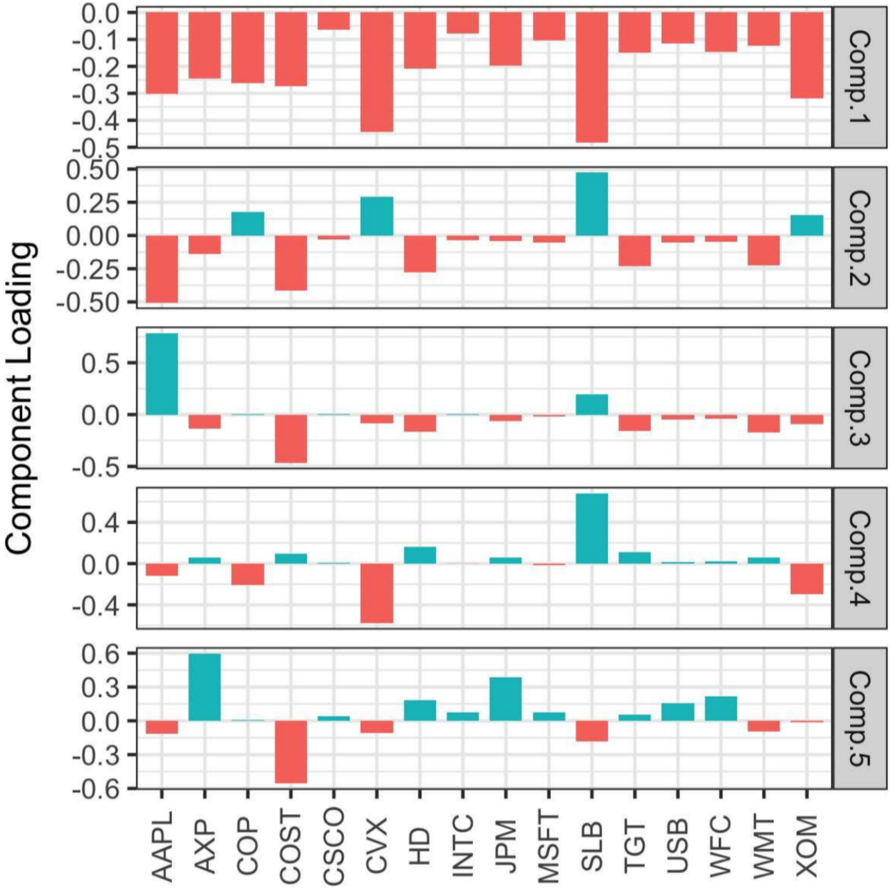


Figure 7-3. The loadings for the top five principal components of stock price returns

How many PCs to select?

Option 1: Visually through the screeplot: elbow method

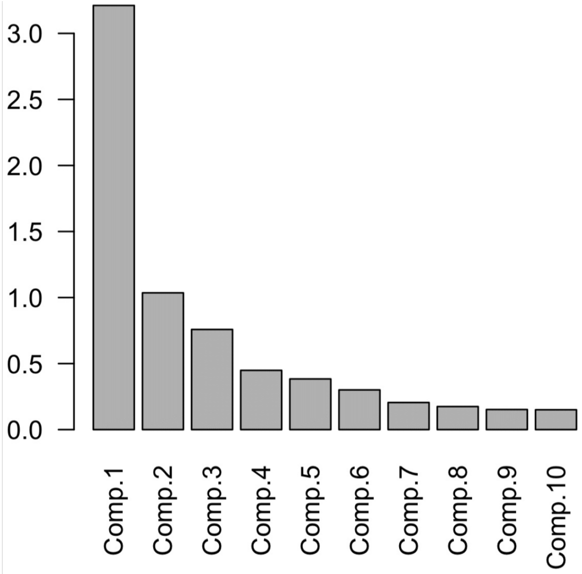


Figure 7-2. A screeplot for a PCA of top stocks from the S&P 500

Option 2: % Variance explained (i.e. 80% variance explained)

Option 3: Inspect loadings for an intuitive interpretation

Option 4: Cross-validation

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

https://forms.gle/HAc9XiVoV1gsAMZH8

PCA : Key Ideas

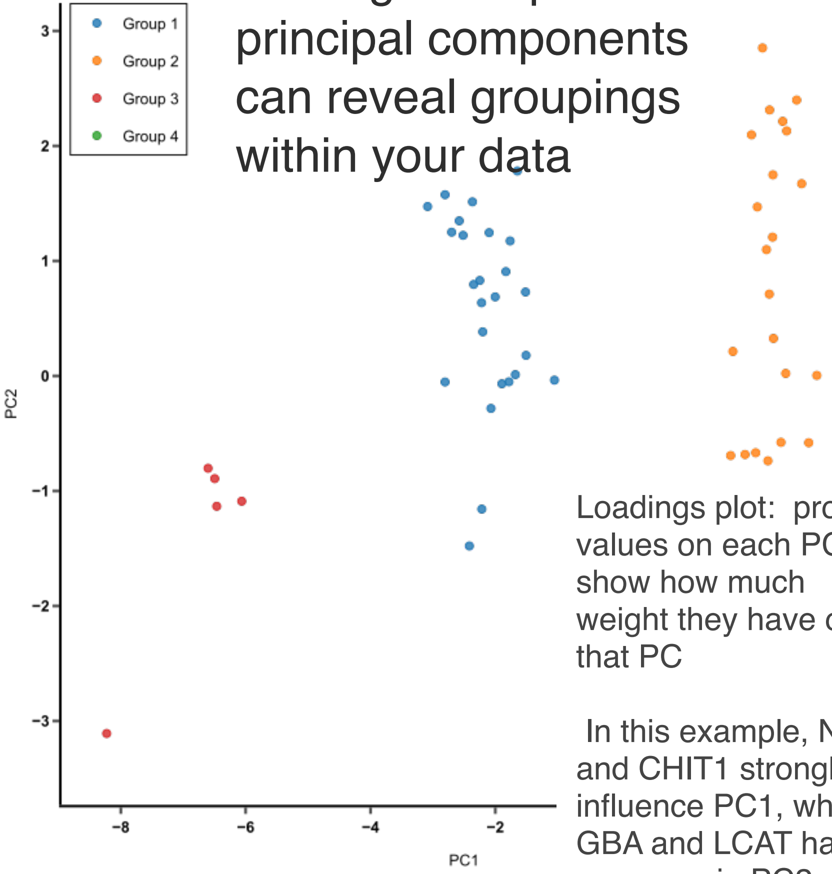
1. PCs are linear combinations of the predictor variables (numeric data only)
2. Calculated to minimize correlation between components (minimizes redundancy)
3. A limited number of components will typically explain most of the variance in the outcome variable
4. Limited set of PCs can be used in place of original predictors (dimensionality reduction)

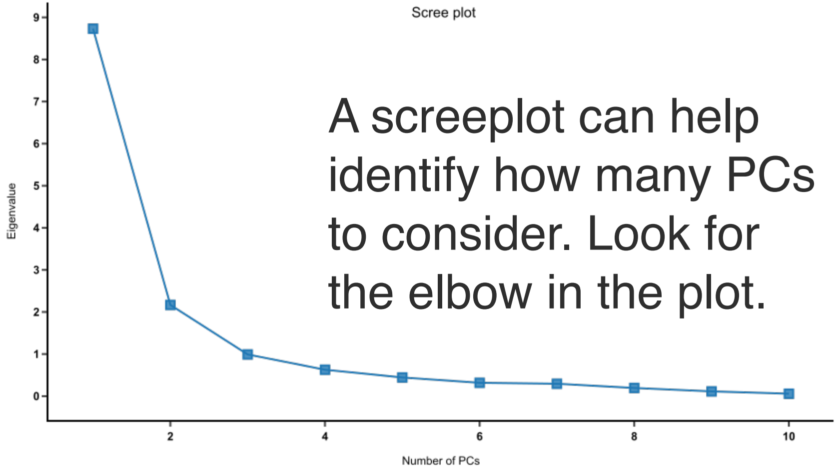
For more on PCA:

- https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/
- http://setosa.io/ev/principal-component-analysis/

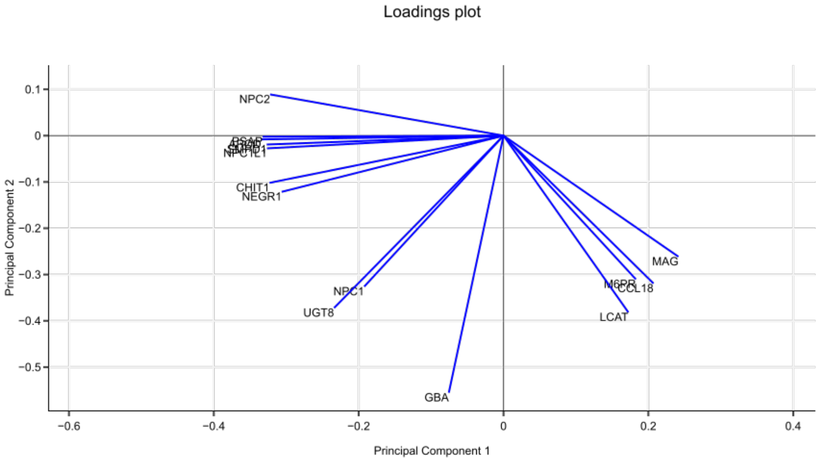
Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Plotting the top principal components can reveal groupings within your data





A screeplot can help identify how many PCs to consider. Look for the elbow in the plot.



Loadings plot: project values on each PC show how much weight they have on that PC

In this example, NPC2 and CHIT1 strongly influence PC1, while GBA and LCAT have more say in PC2

https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/

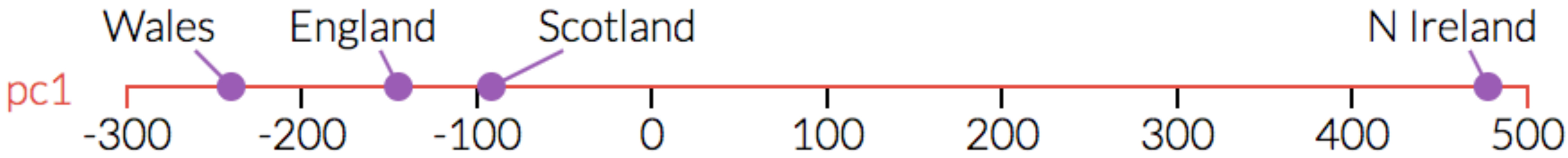
Case Study: Diet in the UK

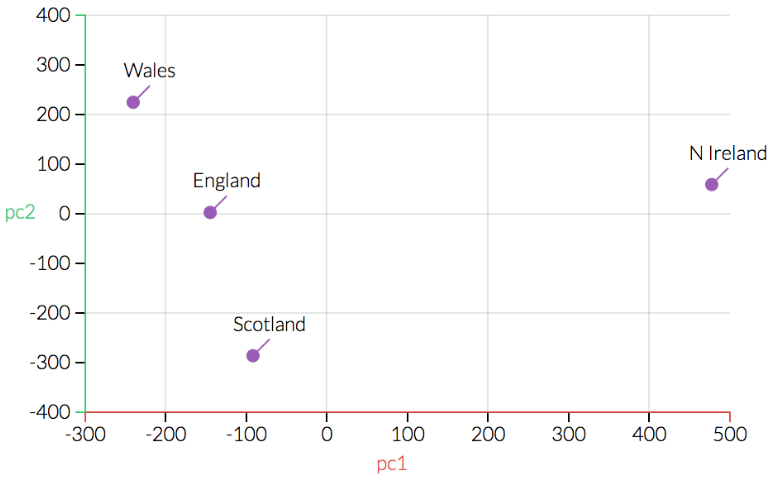
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

17 foods x 4 countries

http://setosa.io/ev/principal-component-analysis/

PCA: Diet in the K





Can already tell that N Ireland is distinct

If we look back at the raw data Northern Ireland, eats way more fresh potatoes and way fewer fresh fruits, cheese, & fish. This reflects real world geography...

http://setosa.io/ev/principal-component-analysis/

Source: Denver Public Library

Case Study: Genetics and Geography

Letter | Published: 31 August 2008

Genes mirror geography within Europe

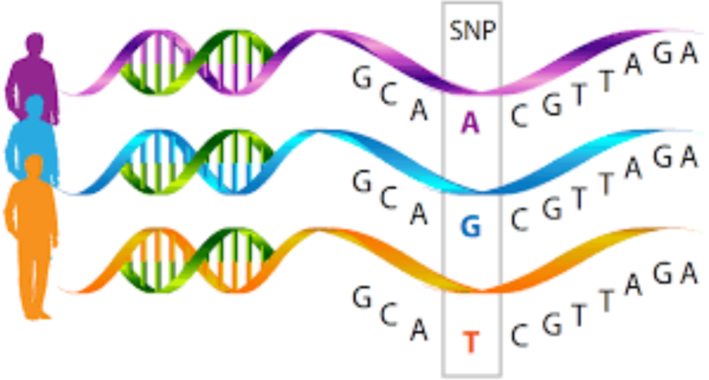
John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

Nature **456**, 98–101 (06 November 2008) | Download Citation

The Data: 1,387 Europeans x 500,000 SNPs

SNP (Single Nucleotide Polymorphism)

- Reminder: Your DNA is made up of four bases: G, C, T, & As
- A SNP is a position in one's DNA that varies between individuals (appears in at least 1% of the population)
  - This results from normal human variation
  - Some contribute to disease, but many are just differences between humans
  - These are used by companies like 23andMe and Ancestry.com



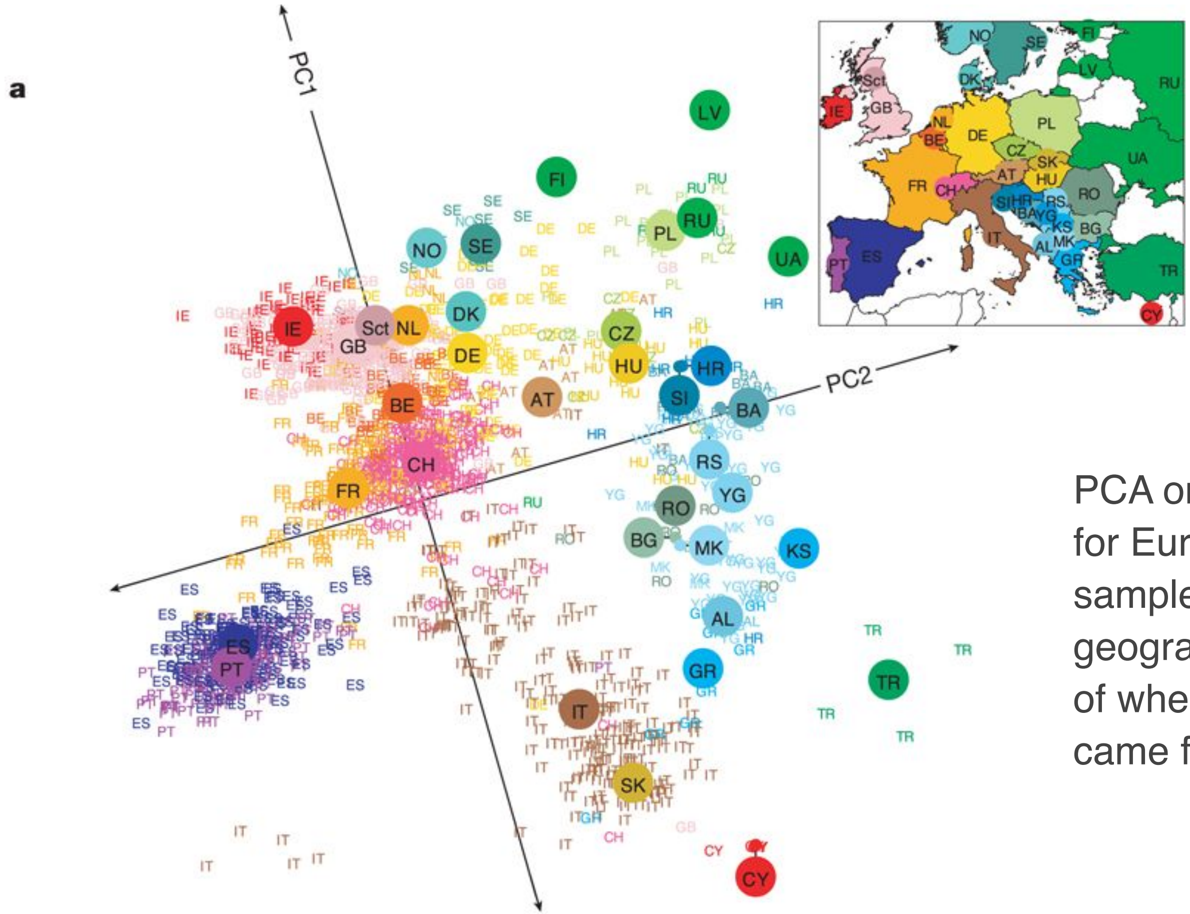
The Data: 1,387 Europeans x 500,000 SNPs

Step 1: Measure genotype (GCTA) at 500,000 positions (SNPs) along the genome in 1387 European individuals

Step 2: Calculate PCs from 500,000 SNPs

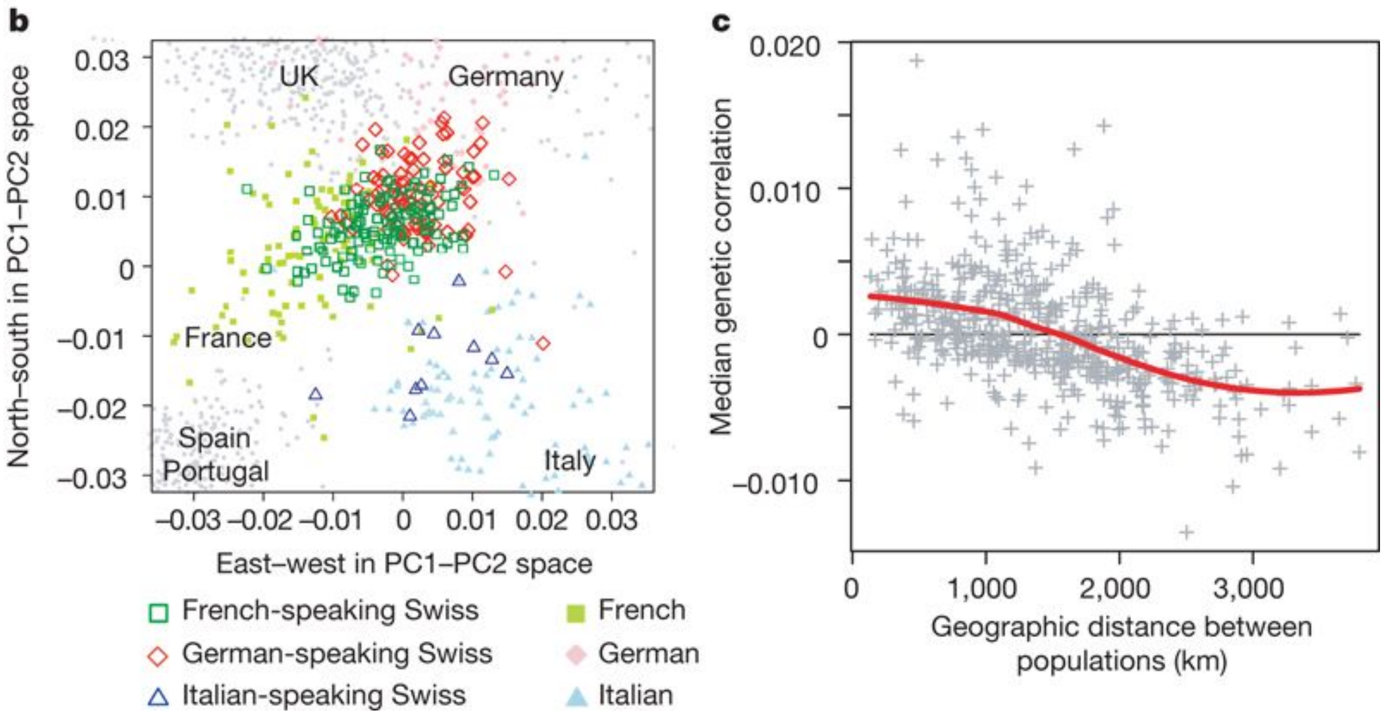
Step 3: Plot PC1 and PC2 (each point is an individual)

Step 4: Compare to the map of Europe



PCA on SNP data for European samples reflects geographic location of where samples came from

PC1 is East-West ; PC2 is North-South



https://www.nature.com/articles/nature07331

This analysis used 500,000 SNPs from 1,387 individuals.
How many PCs would have been calculated?

A
2

B
10

C
1,387

D
500,000

E
I’m lost

This analysis used 500,000 SNPs from 1387 individuals. How many PCs explain geographic differences across Europe by genetic ancestry?

A
2

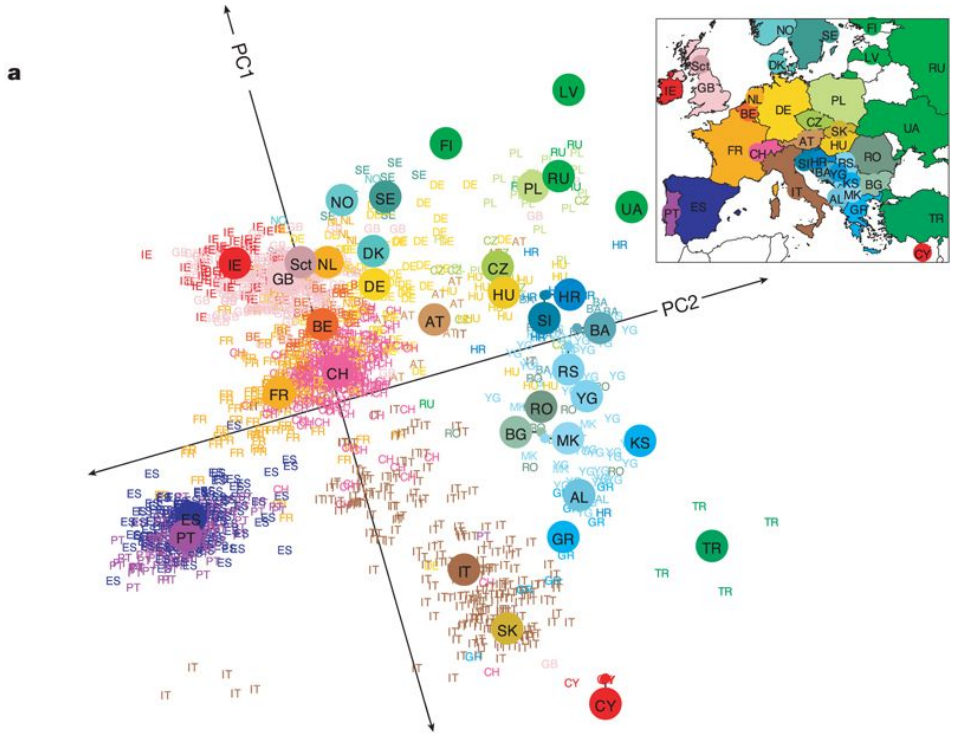
B
10

C
1,387

D
500,000

E
I'm lost

Which of the following is NOT true?



https://www.nature.com/articles/nature07331

A PC1 explains geographic differences from North to South

B PC2 explains geographic differences from East to West

C The French (FR) are not genetically related to the Scottish (Sct)

D The French are more closely related genetically to Germans (DE) than they are the Fins (GL)

E The Spanish (ES) and Portuguese (PT) are genetically similar

Dimensionality Reduction with PCA: Pros & Cons

Pros:

- Helps compress data; reduced storage space.
- reduces computation time.
- helps remove redundant features (if any)
- Identifies outliers in the data

Cons:

- may lead to some amount of data loss.
- tends to find linear correlations between variables, which is sometimes undesirable.
- fails in cases where mean and covariance are not enough to define datasets.
- may not know how many principal components to keep
- highly affected by outliers in the data

https://www.geeksforgeeks.org/dimensionality-reduction/