

CS/ENGR M148 L11: Principal Component Analysis

Sandra Batista

This week in discussion section:

Lab on PCA on single-cell data

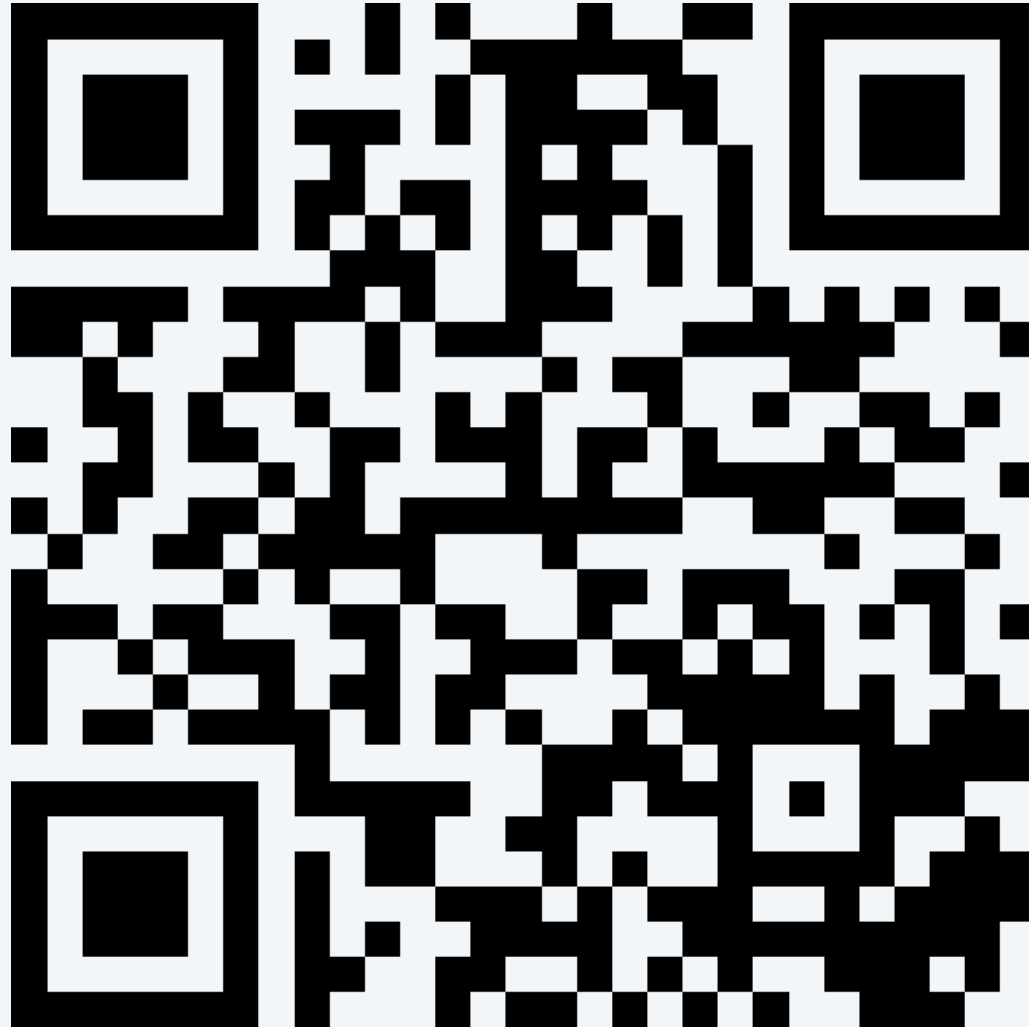
No project check-in this week.

Midterm grading underway. We hope to be done by Wednesday next week.

We'll be sharing a mid-quarter survey for extra credit...

Join our slido for the week...

<https://app.sli.do/event/8heJrvRsridgaJyz1qgUYv>

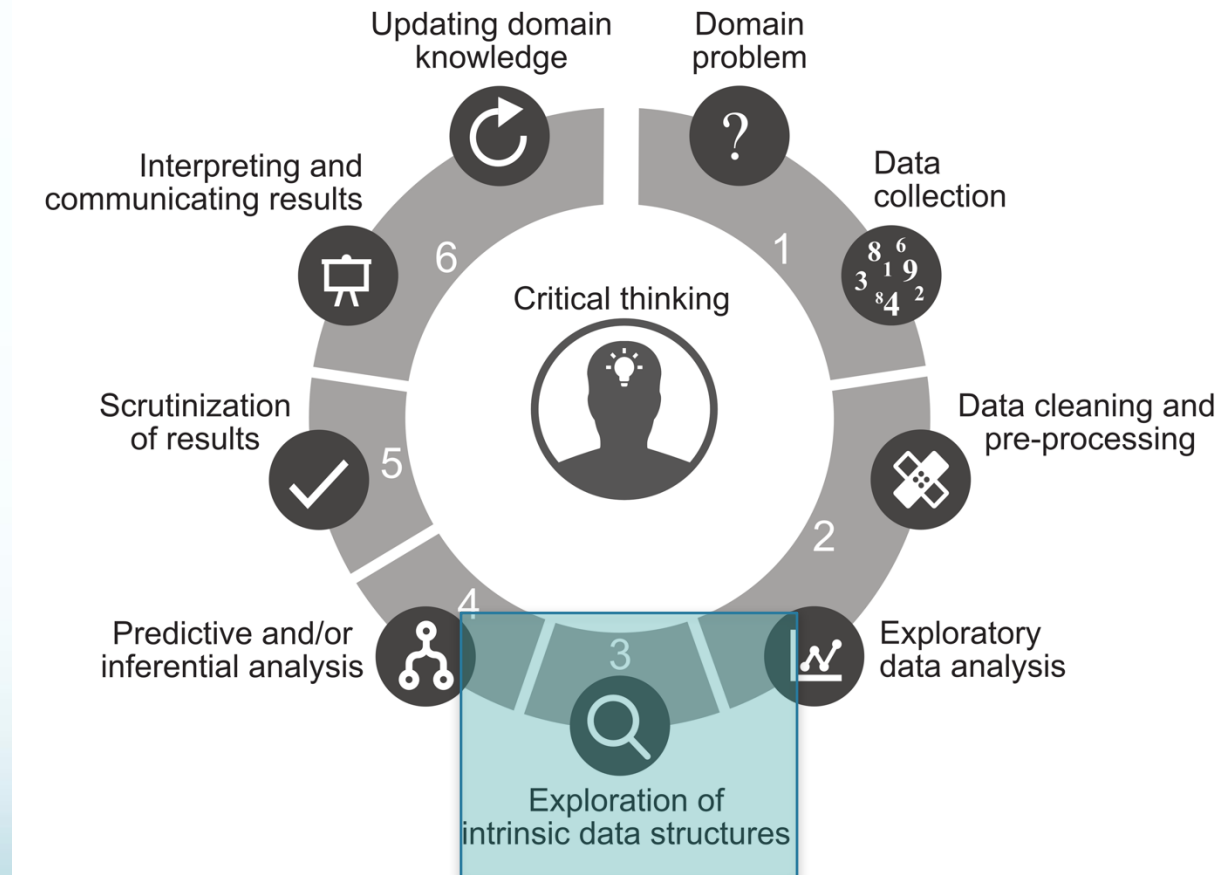


Today's Learning Objectives

Students will be able to:

- Explain the goal of principal component analysis (PCA)
- Understand what SVD is
- Use SVD to perform PCA
- Apply PCA to a real nutrition data set

Data Science Life Cycle (DSLCL)



[Yu, Barter 2024]

Dimensionality reduction

High-dimensional data is data with many features (such as thousands of variables) e.g. gene expression data, nutrition data

Dimensionality Reduction is the process of creating a lower-dimensional representation of a dataset.

1. Summarize the strongest patterns and relationships between the *variables* in the data (
2. Make computation on the data easier by reducing its size.

E.g.: Principal component analysis summarizes low-dimensional linear relationships in high-dimensional datasets [Arter 2024]

Principal Component Analysis (PCA)

Principal component analysis is an algorithm that computes a series of “orthogonal” linear combinations that have the maximum possible variance relative to the origin

By default, the origin is the data point whose measurements across all variables equal zero

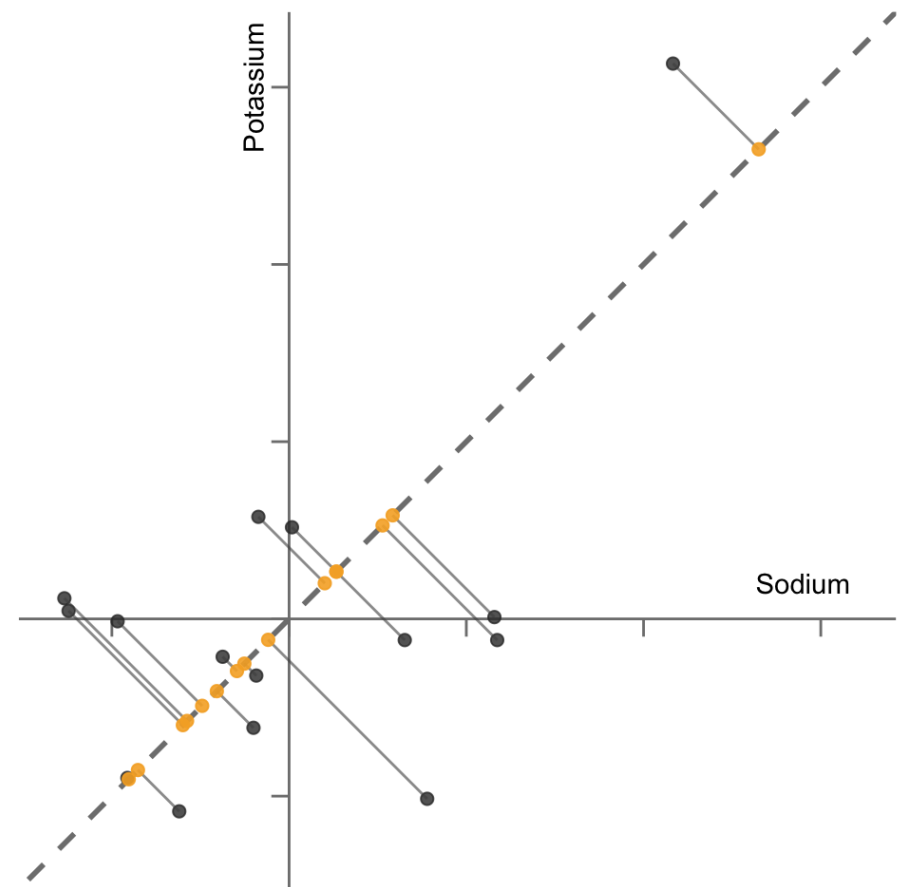
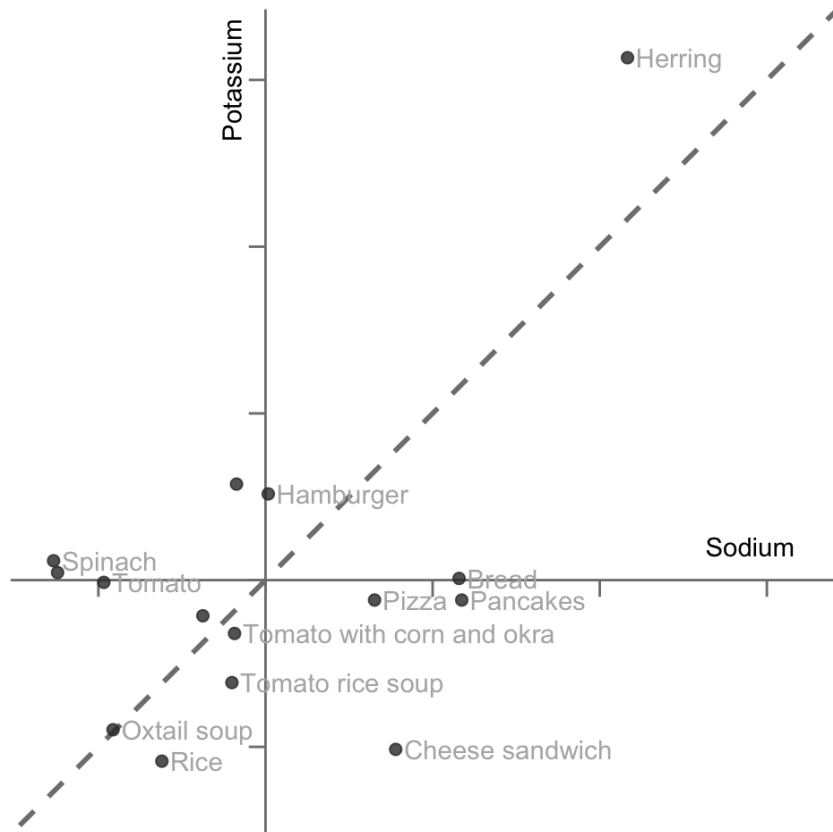
If data has been mean-centered, the origin is the mean of all the measurements

Principal Component Analysis (PCA)

Principal component analysis is an algorithm that computes a series of “orthogonal” linear combinations that have the maximum possible variance relative to the origin

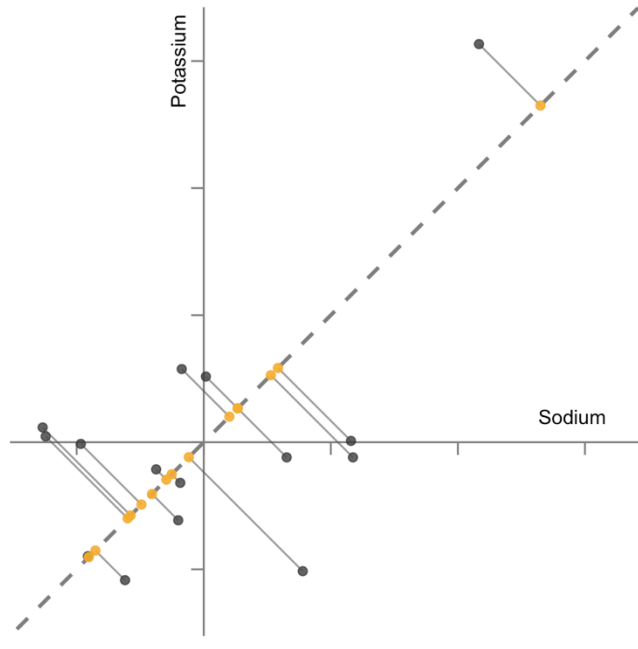
Variance measures of how well the variable’s measurements can distinguish between the observations.

First Principal Component(PC)

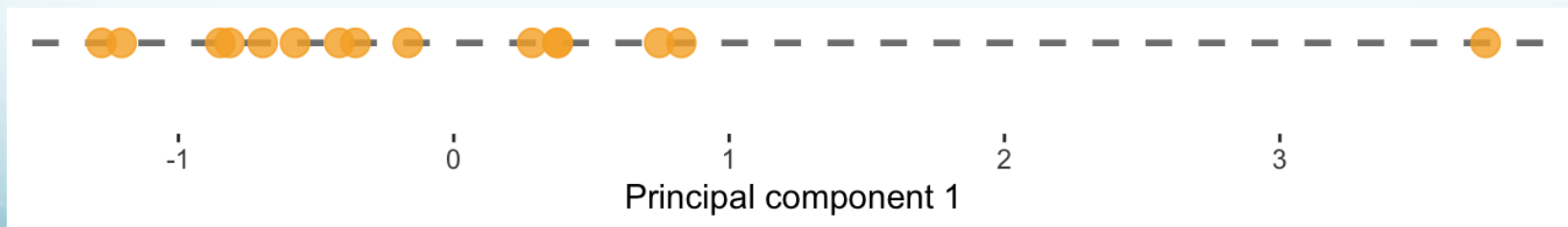


[Yu, Barter 2024]

First Principal Component(PC)



The first PC is the linear combination of the original variables corresponding to the direction along which the projected data points exhibit maximum variation.

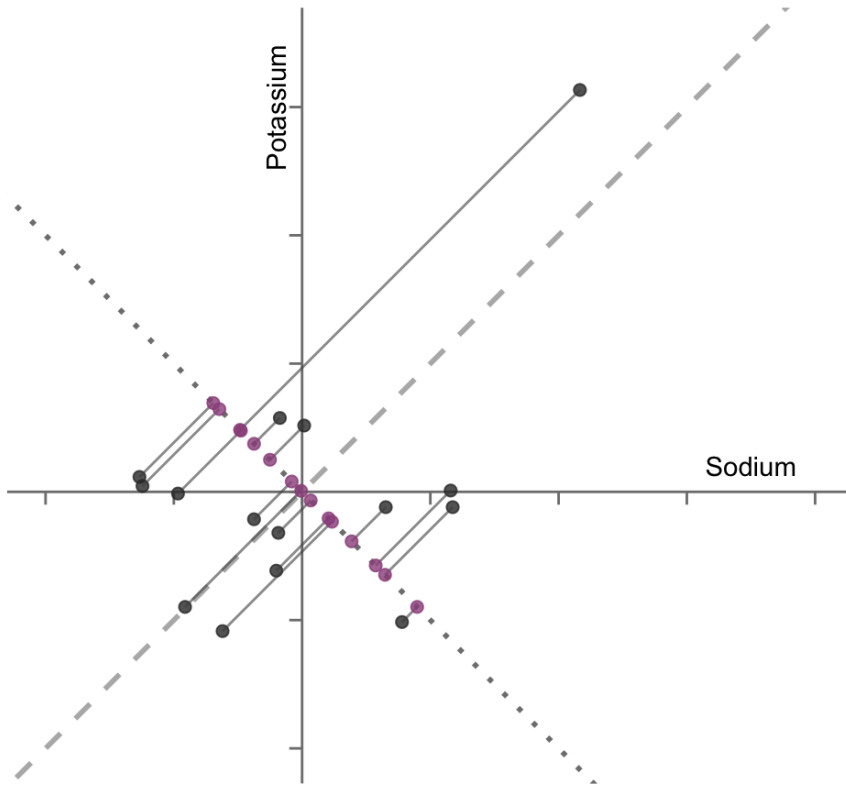


[Yu, Barter 2024]

Second PC and so forth...

Second PC is the linear combination of the original variables corresponding to the direction along which the projected data points exhibit the next highest variation.

Can have a PC for up to the number of variables.



Preprocessing for PCA

Standardization involves both **mean-centering** (subtracting the mean from each column) and **SD-scaling** (dividing each column by its standard deviation (SD)).

(SD-scaling is sometimes called normalizing.)

Common practice but not absolutely necessary.

Consider if

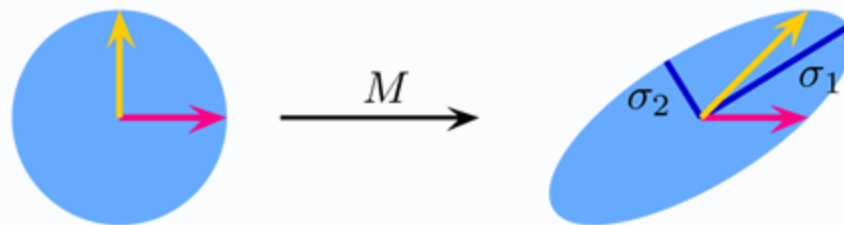
1. Means or 0 more important
2. If common scale needed or will it remove meaning of scale in data

Today's Learning Objectives

Students will be able to:

- ✓ Explain the goal of principal component analysis (PCA)
 - Understand what SVD is
 - Use SVD to perform PCA
 - Apply PCA to a real nutrition data set

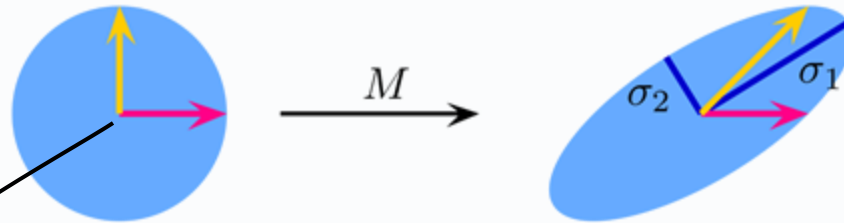
How a matrix describes (is code for) a function



$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \end{bmatrix}$$

$$M \quad (\mathbf{x})$$

The code depends on the choice of *basis*



A basis consists of a set of independent vectors that span a vector space.

$$\begin{array}{|c|c|} \hline 1 & 1 \\ \hline 0 & 1 \\ \hline \end{array} \cdot \begin{array}{|c|} \hline 4 \\ \hline 2 \\ \hline \end{array} = \begin{array}{|c|} \hline 6 \\ \hline 2 \\ \hline \end{array}$$

$M \quad (\mathbf{x})$

The code depends on the choice of *basis*



A basis consists of a set of independent vectors that span a vector space.

0	1
1	-1

Bases of a vector space

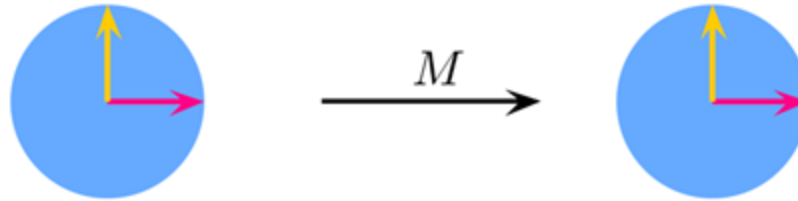
- Vector spaces may have many bases.
- The number of elements in a basis of a vector space is called the ***dimension*** of the vector space.
- The *standard basis* for R^n consists of the n vectors $(1,0,0,\dots,0)$, $(0,1,0,0,\dots,0)$, \dots , $(0,0,\dots,1)$.



- An **orthogonal basis** consists of vectors that are mutually orthogonal. The standard basis is orthogonal, and additionally **orthonormal**, i.e. all vectors in the basis have norm 1.

A simple matrix (linear transformation)

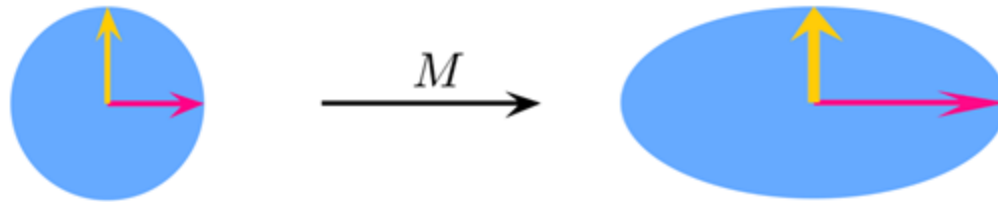
- The identity matrix



1	0
0	1

Another simple matrix (linear transformation)

- The diagonal matrix = scaling transformation



2	0
0	1

What is singular value decomposition about?

- The singular value decomposition (SVD) is a way that every matrix is essentially diagonal given appropriate matrices for the domain and range spaces.
- SVD clarifies linear transformations and their accompanying matrix representations.

Singular value decomposition

- **Input:** an $m \times n$ matrix
- **Output:** a set of numbers called *singular values* and two collections of vectors: a set of *right singular vectors* and another set of *left singular vectors*.

The SVD theorem

Theorem : Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored into a singular value decomposition (SVD),

$$A = USV^T,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e., $UU^T = VV^T = I$) and $S \in \mathbb{R}^{m \times n}$ is diagonal with $r = \text{rank}(A)$ leading positive diagonal entries. The p diagonal entries of S are usually denoted by σ_i for $i = 1, \dots, p$, where $p = \min\{m, n\}$, and σ_i are called the singular values of A .

Singular value decomposition

- **Input:** an $m \times n$ matrix
- **Output:** a set of numbers called *singular values* and two collections of vectors: a set of *right singular vectors* and another set of *left singular vectors*.

$$X = UDV^T.$$

Singular value decomposition

$$X = UDV^T.$$

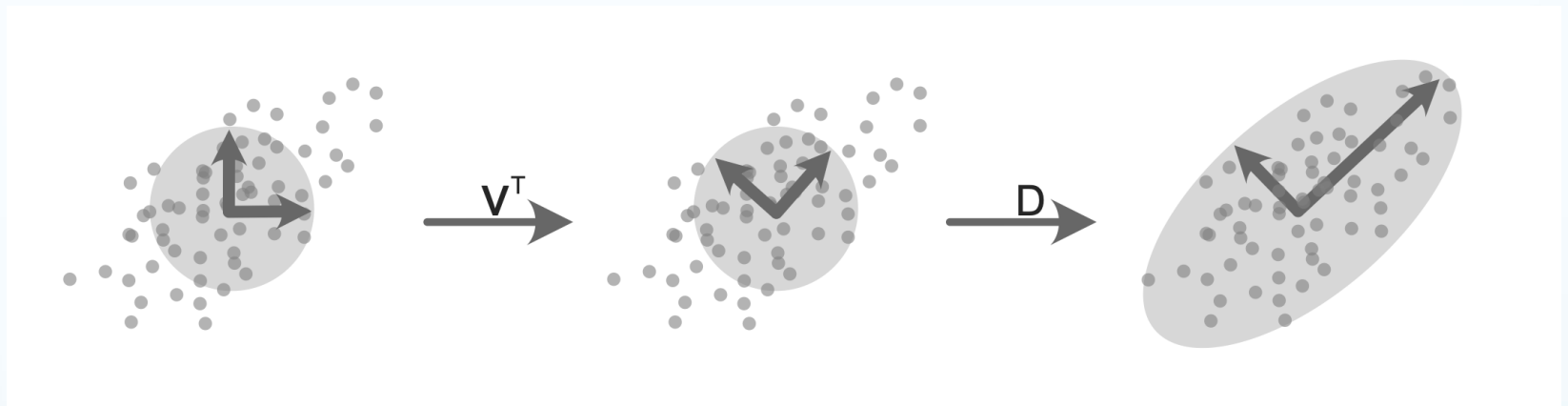
The left matrix, U, contains the ***left-singular vectors***

The matrix D is a **diagonal matrix** whose entries correspond to the magnitude or strength of the corresponding principal component directions. **Singular values are diagonal entries of D.**

V contains the **right-singular vectors** of the data matrix and each right-singular vector corresponds to a principal component.

Each column of V contains the coefficients of the corresponding principal component linear combination.

Rotating and scaling matrices in SVD



Variable Loadings

Variable loadings corresponds to the coefficient (or weight) of the variable in the linear combination that defines the principal component.

Variable Loadings can be used for feature importance if variables on the same scale.

The variable loadings for each PC are extracted from the relevant column of the right-singular vector matrix, V .

Variable Loadings

	(PC1)	(PC2)	(PC3)	(PC4)	(PC5)	(PC6)
(Sodium)	0.42	0.63	0.23	-0.58	0.15	-0.1
(Potassium)	0.48	-0.28	-0.43	-0.3	-0.64	0.1
(Calcium)	0.29	-0.24	0.79	0.22	-0.35	-0.27
(Phosphorus)	0.49	-0.06	0.12	0.28	0.3	0.76
(Magnesium)	0.38	-0.51	-0.16	-0.11	0.59	-0.46
(Total choline)	0.35	0.45	-0.33	0.66	-0.08	-0.35

PC1 = 0.42 sodium + 0.48 potassium + 0.29 calcium
+ 0.49 phosphorus + 0.38 magnesium
+ 0.35 total choline.

PC1 for a food data point

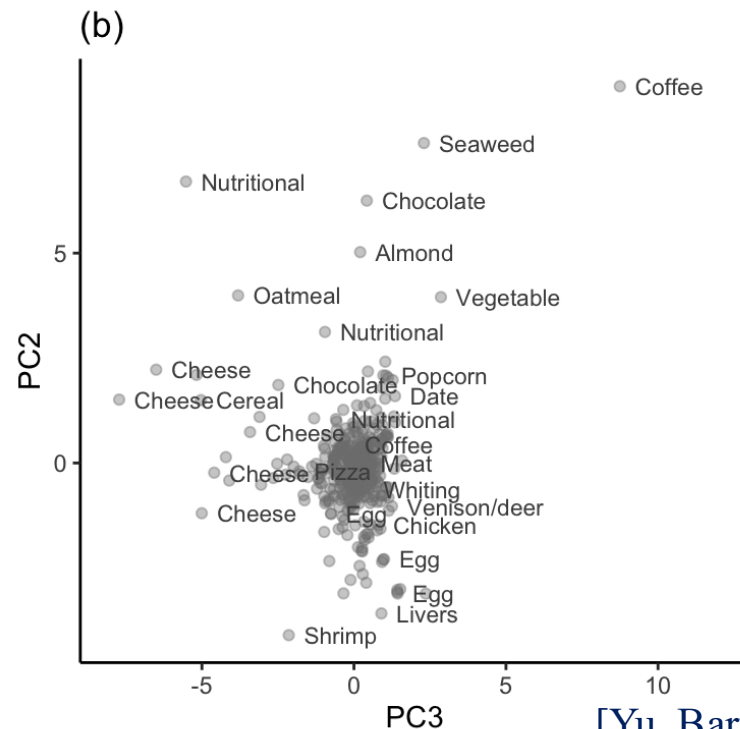
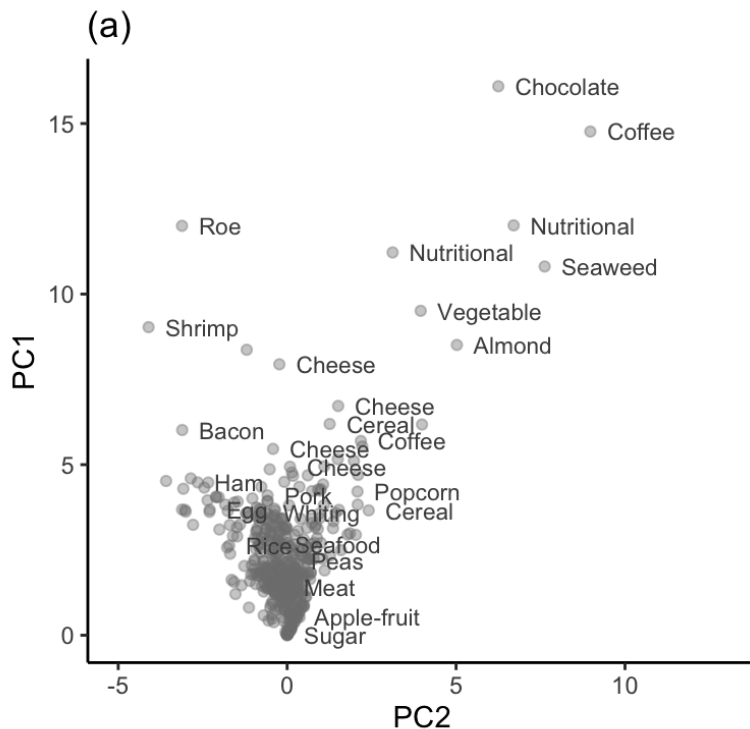
Sodium	Potass.	Calc.	Phosp.	Magn.	Chol.
2.87	2.3	0.75	2.71	1.29	2.34

$$\begin{aligned}\text{PC1}(\text{herring}) = & 0.42 \times 2.87 + 0.48 \times 2.3 + 0.29 \times 0.75 \\ & + 0.49 \times 2.71 + 0.38 \times 1.29 + 0.35 \times 2.34\end{aligned}$$

PC Data set

*To calculate the PC data set, multiply the original data set and the **right-singular vectors**, V*

$$X^{PC} = XV.$$



Variance Explained

D contains the *singular values* on its diagonal

The magnitude of the singular values measures the variability in the original data that is being captured by each principal component.

$$D = \begin{bmatrix} 252.84 & 0 & 0 & 0 & 0 & 0 \\ 0 & 97.36 & 0 & 0 & 0 & 0 \\ 0 & 0 & 91.13 & 0 & 0 & 0 \\ 0 & 0 & 0 & 80.76 & 0 & 0 \\ 0 & 0 & 0 & 0 & 62.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 45.01 \end{bmatrix}.$$

Variance Explained

The sum of the variance of the columns in the original dataset equals the sum of the squared singular values:

$$\sum_j \text{Var}(X_j) = \sum_j d_j^2,$$

$$\text{prop of variability explained by PC}_j = \frac{d_j^2}{\sum_i d_i^2}.$$

31

Variance Explained

The sum of the variance of the columns in the original dataset equals the sum of the squared singular values:

$$\sum_j \text{Var}(X_j) = \sum_j d_j^2,$$

$$\text{prop of variability explained by PC}_j = \frac{d_j^2}{\sum_i d_i^2}.$$

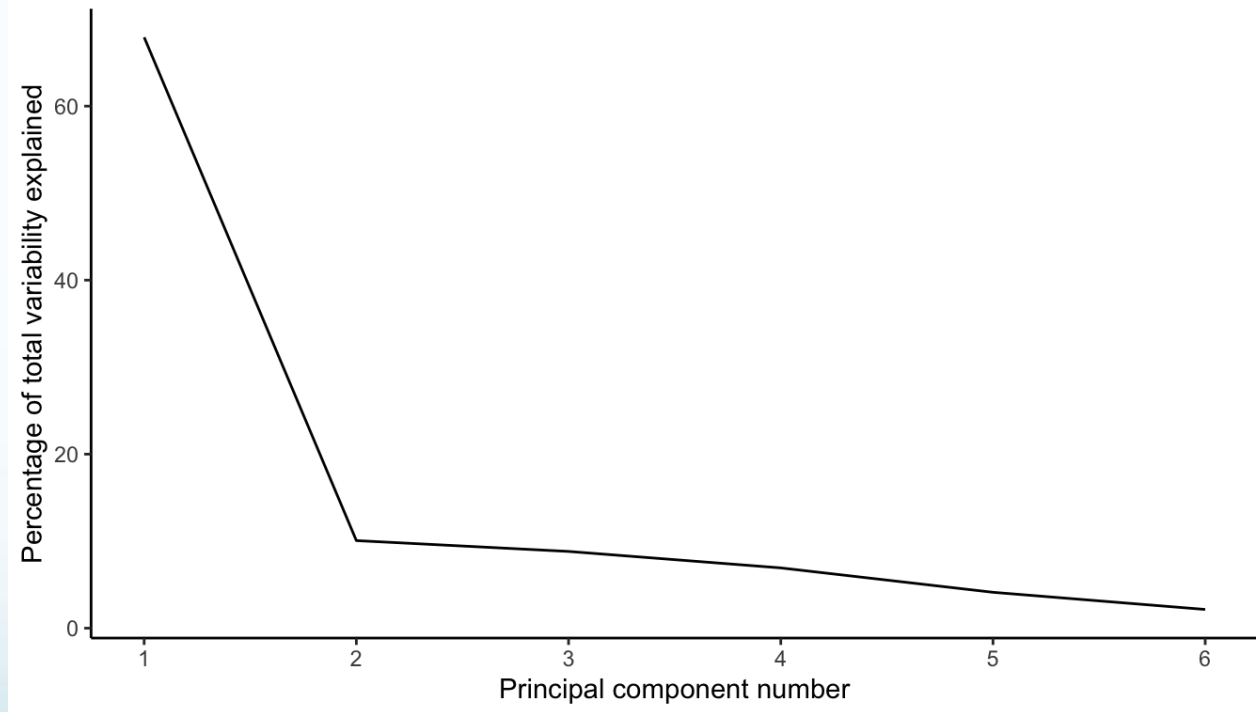
32

Variance Explained

Principal component	Percentage of variability explained	Cumulative percentage of variability explained
1	67.9	67.9
2	10.1	78.0
3	8.8	86.8
4	6.9	93.7
5	4.1	97.8
6	2.2	100.0

Scree plot

A **scree plot** shows the proportion of variability explained by each principal component either in a bar chart or as a line plot.



Common practice: use 'elbow' method to determine PCs to use

[Yu, Barter 2024]

Relevance of orthonormality of U and V

- Recall that two vectors x, y are orthogonal if their dot product $x \cdot y = 0$. The norm of a vector is the square of its dot product with itself.
- If U is an orthonormal matrix, then it follows that $U^T U = I$. In other words, U^T is the inverse of U , and similarly V^T is the inverse of V .

Relevance of orthonormality of U and V

- If the SVD of M is given what is $M^T M$?

Algorithms for singular value decomposition

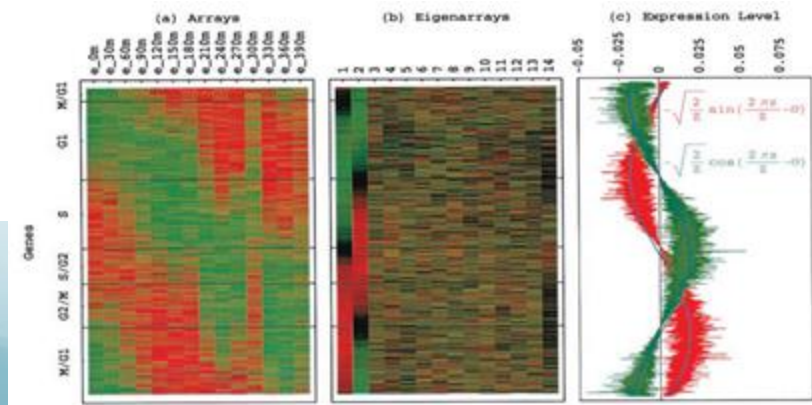
- There are efficient algorithms for computing singular values without computing singular vectors.
- Singular values are stable under perturbation. Singular vectors are sensitive to perturbation, but there are stability results for the subspaces they span. **This makes SVD suitable for application to noisy data**

Applications of singular value decomposition

- Singular value decomposition has numerous applications in mathematics, science and engineering. These include:
 - Determining the rank of a matrix
 - Computing the pseudoinverse of a matrix
 - Low rank matrix approximation
 - Imputation
 - Principal component analysis
 - Clustering
 - Visualization of high-dimensional data in two dimensions
 - Removal of noise from data
 - ...

Low rank approximation for gene expression analysis

- The SVD was first applied to gene expression analysis on data produced with microarrays ([Rachaudhuri, Stuart and Altman, 2000](#), [Alter et al., 2000](#)).



Orly Alter

SVD and decomposition of the covariance matrix.

- Recall that the covariance matrix for a data matrix M with n observations (i.e. n cells) is a matrix C given by

$$c_{jk} = \frac{1}{n-1} \sum_{i=1}^n (m_{ij} - \bar{m}_j)(m_{ik} - \bar{m}_k).$$

- If the matrix M is *centered*, i.e. for each gene the expression averaged over cells is zero, then this reduces to $C = \frac{1}{n-1} M^T M$.

- Eigendecomposition of the covariance matrix C can be performed by SVD of M .**
- Eigendecomposition decomposes the *covariance* matrix of a dataset. The eigenvalues are square of the singular values.

Today's Learning Objectives

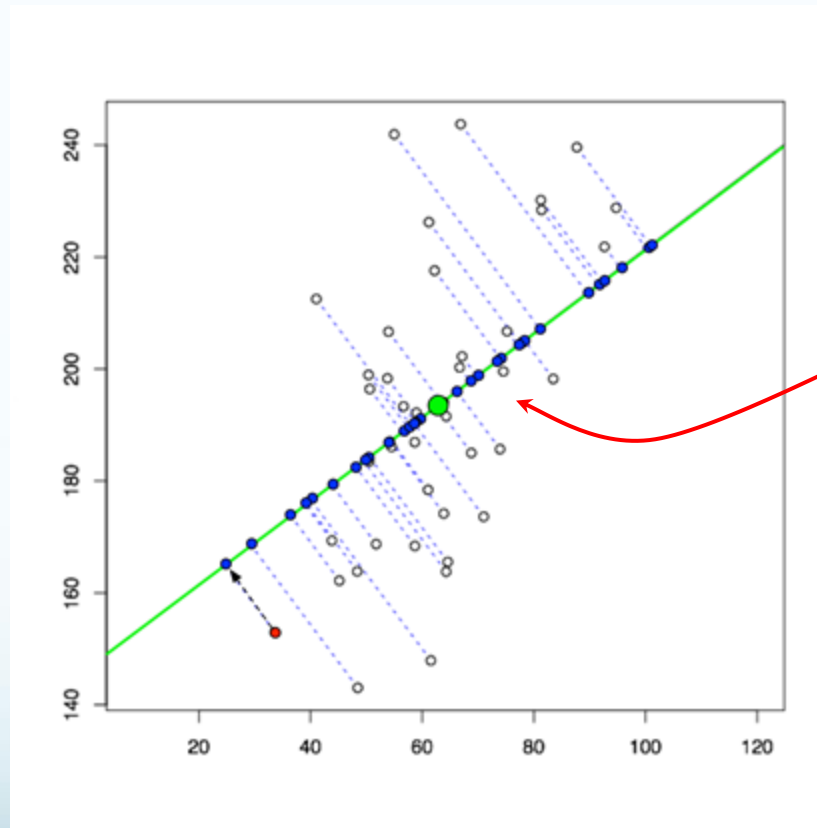
Students will be able to:

- ✓ Explain the goal of principal component analysis (PCA)
- ✓ Understand what SVD is
 - Use SVD to perform PCA
 - Apply PCA to a real nutrition data set



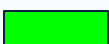
Principal component analysis

- The SVD of a (centered) M , yields a decomposition of $M^T M$ as i.e., eigendecomposition of the covariance matrix C can be performed by SVD of M .
- The low rank approximation of using first k columns of V .
- Set V_k to be the first k columns of V , i.e., $V_k = [v_1, v_2, \dots, v_k]$. Then the projection of the points in M by V_k , i.e., $P = MV_k$ has **numerous useful (and beautiful) properties.**

An example of a PCA projection



*maximizes the variance
of the projected points*

	<i>math</i>
	<i>statistics</i>
	<i>computer science</i>

- Start with a **data matrix** M .
- **Center** M .
- M has a singular value decomposition that is derived from viewing M as a **linear transformation**. The matrix V consists of the eigenvectors which diagonalize the **covariance matrix** $M^T M$.
- **Compute** V .
- Let V_k be the truncation of V to its first k columns. We **know from linear algebra** that this is a meaningful restriction because M_k is a good low-rank approximation to M .
- **Project** the **data matrix** M with V_k to obtain a new set of points: MV_k .
- The projection has the property that it will **maximize the variance** of the projected points.

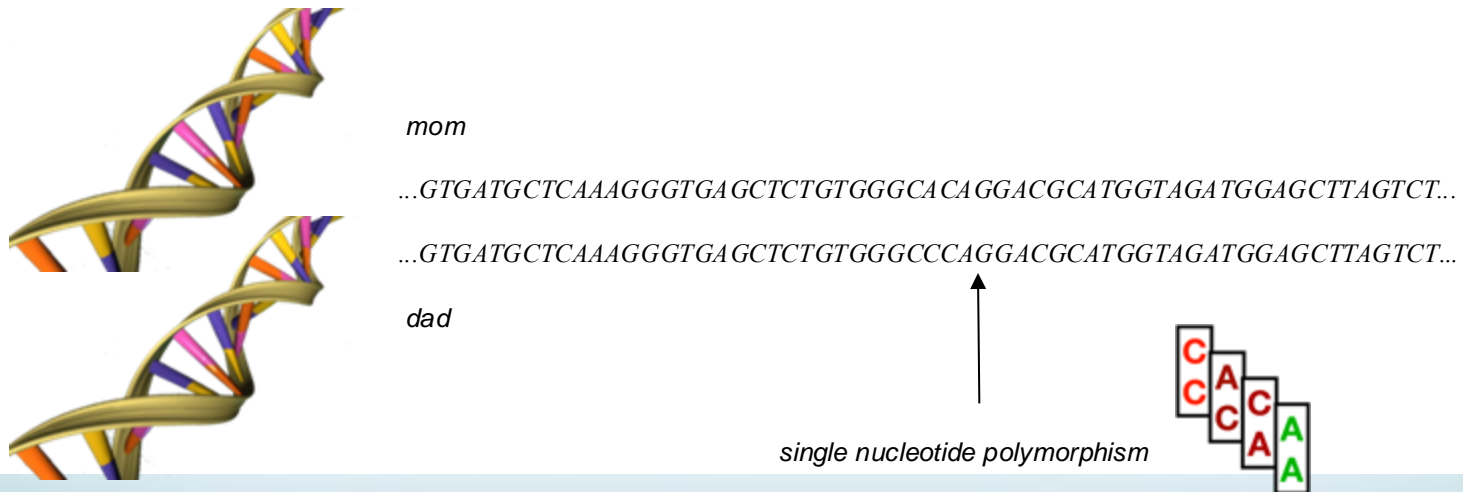
PCA Step by Step

1. (Preprocessing) Consider whether it makes sense to mean-center and/or SD-scale the variables (the answer will *usually*—but not always—be yes to both). If the answer is unclear, you will need to make a judgment call.
2. (Preprocessing) Visualize the distribution of each variable to determine whether a log-transformation (or any other kind of transformation) may increase the symmetry of the distributions. If so, you may want to apply a log-transformation to the relevant variables.

3. Apply SVD to your numeric data matrix to conduct principal component analysis (i.e., to compute the matrices V and D).
4. Based on the right-singular vector matrix, V , identify which variables are contributing most to each principal component. Based on the diagonal singular value matrix, D , compute the proportion of variability explained by each principal component.
5. Compute the principal component transformation of the original data matrix, X , by multiplying the original data by the right-singular vector matrix, V , as in $X^{PC} = XV$.

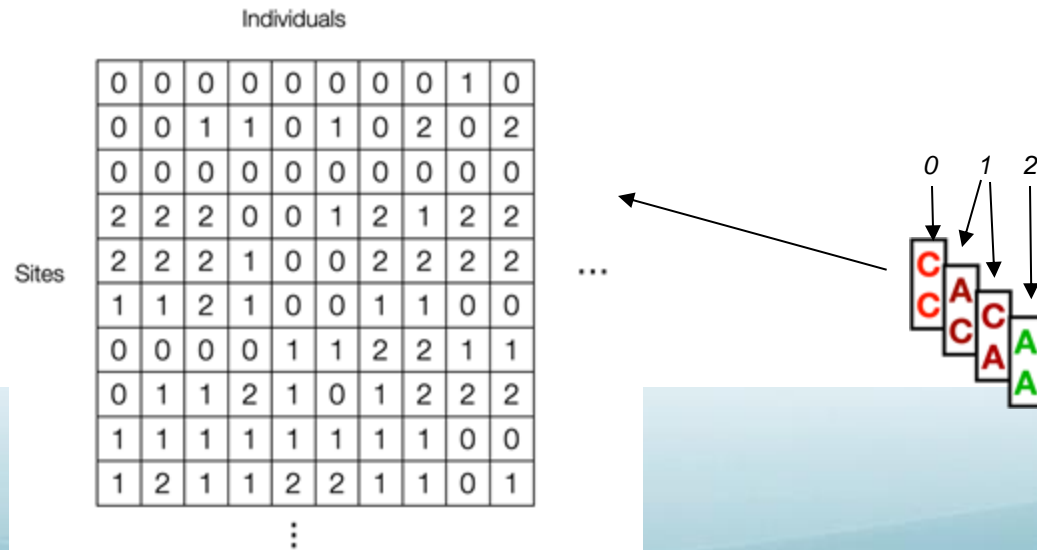
An application of PCA: the human genotype matrix

- Differences between any pair of human genomes are largely in the same sites, and consist of single nucleotide polymorphisms (SNPs).
- Most human SNPs are biallelic.

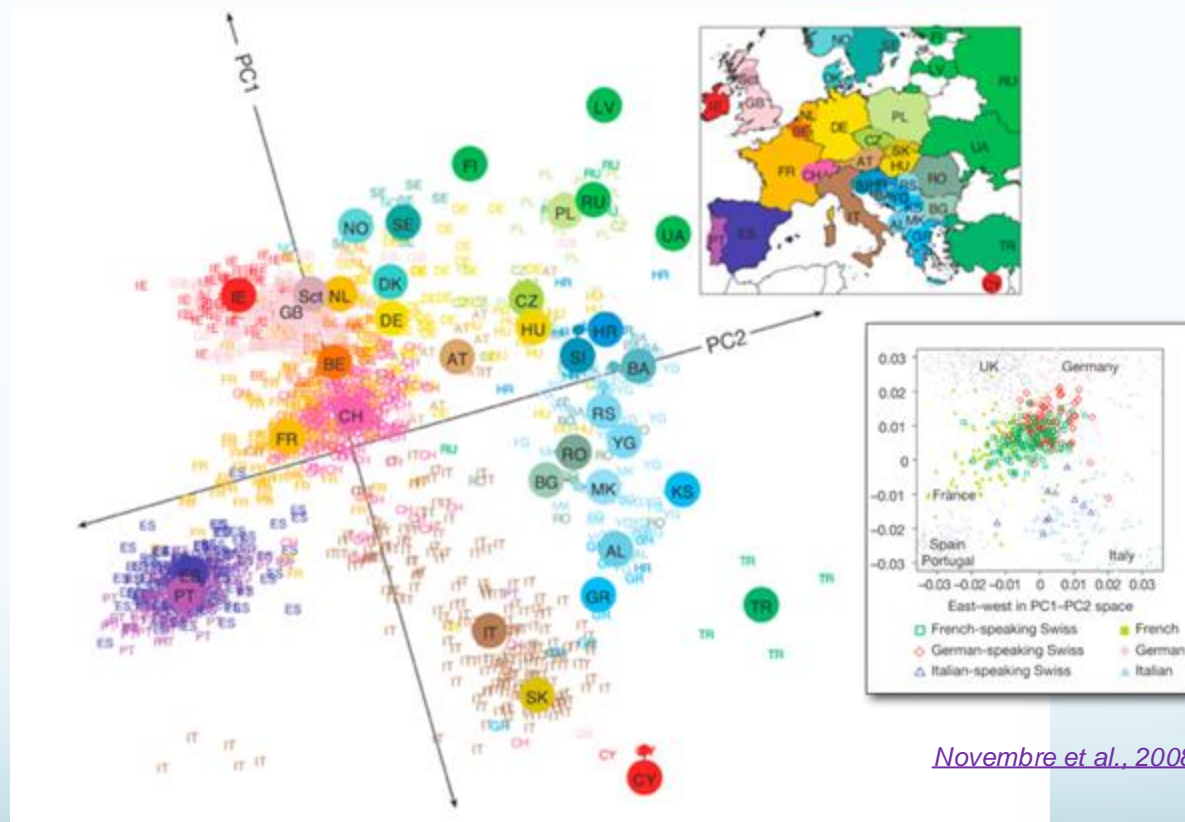


An application of PCA: the human genotype matrix

- Differences between any pair of human genomes are largely in the same sites, and consist of single nucleotide polymorphisms (SNPs).
- Most human SNPs are biallelic.



“Genes mirror geography within Europe”



Novembre et al., 2008

Today's Learning Objectives

Students will be able to:

- ✓ Explain the goal of principal component analysis (PCA)
- ✓ Understand what SVD is
- ✓ Use SVD to perform PCA
 - Apply PCA to a real nutrition data set

Your turn:

PCA on Nutrition Data

Please get the Jupyter notebook for PCA on US Department of Agriculture's (USDA) **Food and Nutrient Database for Dietary Studies (FNDDS)** ([Fukagawa et al. 2022](#)) Data:

Go to:

[https://colab.research.google.com/drive/1UwcuL31OrIRV
KFIRPyDcAM1tdG73j8ko?usp=sharing](https://colab.research.google.com/drive/1UwcuL31OrIRV KFIRPyDcAM1tdG73j8ko?usp=sharing)

Save a copy to your Google Drive and keep notes there...

Today's Learning Objectives

Students will be able to:

- ✓ Explain the goal of principal component analysis (PCA)
- ✓ Understand what SVD is
 - Use SVD to perform PCA
- ✓ Apply PCA to a real nutrition data set
- ✓

Citations:

Yu, B., & Barter, R. L. (2024). Veridical data science: The practice of responsible data analysis and decision making. The MIT Press.

Shah, C. (2020) A hands-on introduction to data science. Cambridge University Press.

Some slides adapted from CalTech CS183 Spring 2021 Lior Pachter Lab: These slides are distributed under the [CC BY 4.0 license](#)

Thank You
