Module 5AB: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

1. Review of Hypothesis testing

- Review χ^2 goodness of fit tests: assumptions, how it works, demonstrate use with any distribution (it is a non-parametric method)
  - χ^2 contingency test: a specialized type of χ^2 goodness of fit test with the Ho: two variables are **independent** (tested using probability multiplication rule)
- Reviewed Z scores: $\frac{Signal}{Noise} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

2. Student's t tests:

- Different versions (one sample, paired, two sample), with different assumptions
- Same principle as Z score but replace **σ** with **s**, adding uncertainty, the t-distribution is wider than Z
- $\frac{Signal}{Noise} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$
- Nonparametric versions that can be used when assumptions are not met
- Welch's approximate t test (still parametric – assumes normal distribution) when variances between the two populations are wildly different; adjusts degrees of freedom downward

3. ANOVA & Correlation

4. Regression

1. **Review of Hypothesis testing**
2. **Student's t tests**
3. **ANOVA & Correlation**
   - Extension of the two-sample t test when >2 populations are compared
   - Allows accurate **α** (no inflation)
   - How well does the model of "belonging to a particular treatment group" explain the total variation?
   - You are allocating variation: **between group variation,** and **within group (stochastic) variation**
   - $\frac{Signal}{Noise} = \frac{MS_{groups}}{MS_{error}}$
   - Post-hoc testing to determine **WHICH group(s)** has/have significantly different population means (**μ**)
   - (Pearson) Linear Correlation:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Covariance(X,Y)}{s_x s_y}$$

4. **Regression**

1. **Review of Hypothesis testing**
2. **Student's t tests**
3. **ANOVA & Correlation**
   - (Pearson) Linear Correlation:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2}\sqrt{\sum (Y - \bar{Y})^2}} = \frac{Co\,\mathrm{var}\,iance(X,Y)}{s_x s_y}$$

4. **Regression**
   - Independent (explanatory) variable impacts dependent (response) variable
   - Many different types that have slightly different assumptions
   - Typical: Homoscedasticity and normally distributed Y around each X_i
   - Note structural similarities between slope and correlation

$$\bar{Y} = a + \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}\bar{X}$$

   - Brief overview of General Linear models to emphasize the similarity: stepwise comparison of the full model to a model with one variable (or interaction) removed to see if there is a statistically significant improvement in fit.

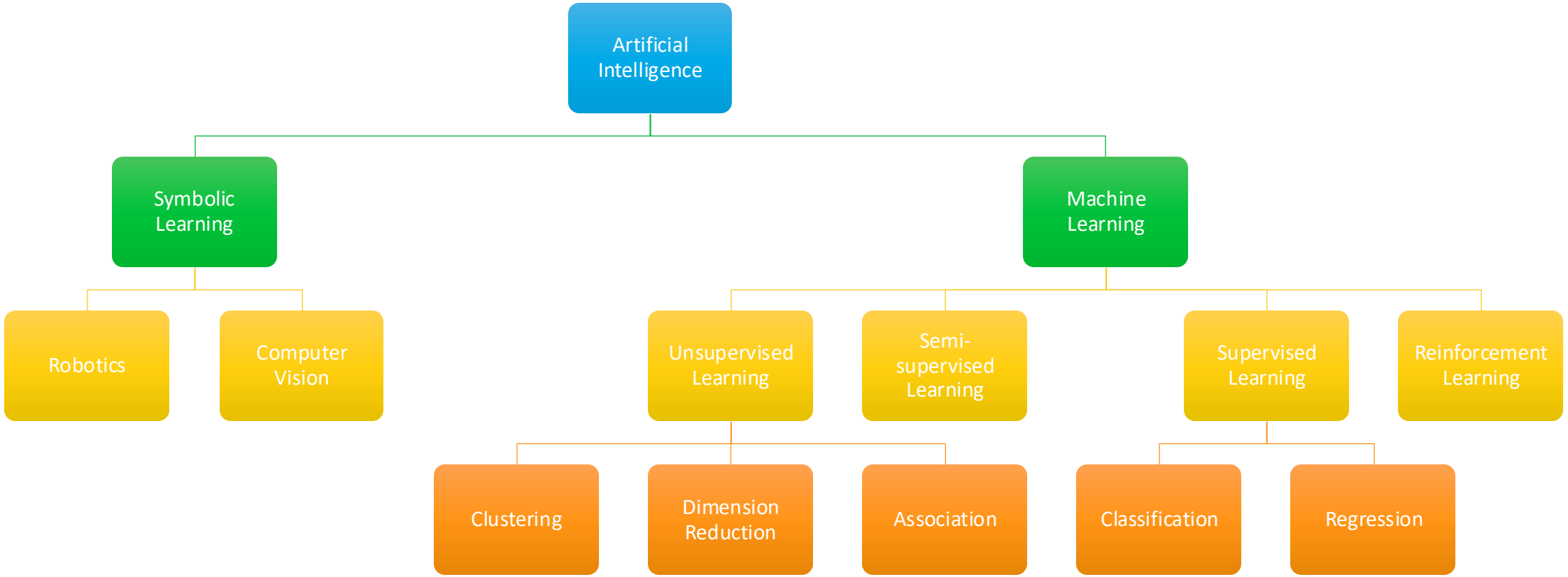
Agenda:

- Sometimes parametric methods are not powerful enough
- What is **unsupervised learning?**

Finding hidden patterns in data without labels

1. Clustering: K-means
   - Uses: candidate miRNA targets, gene expression data
2. Dimension Reduction: PCA, Discriminant Analysis
   - single-cell analysis, gene expression; population structure

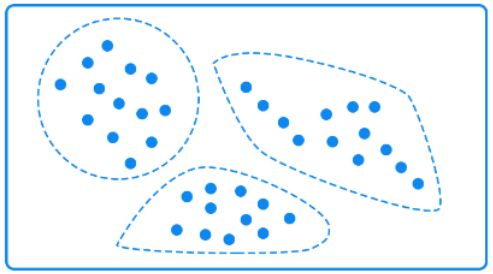
- Survey of computational methods: bootstrapping, permutation, and simulation


Artificial Intelligence
Symbolic Learning
Machine Learning
Robotics
Computer Vision
Unsupervised Learning
Semi-supervised Learning
Supervised Learning
Reinforcement Learning
Clustering
Dimension Reduction
Association
Classification
Regression

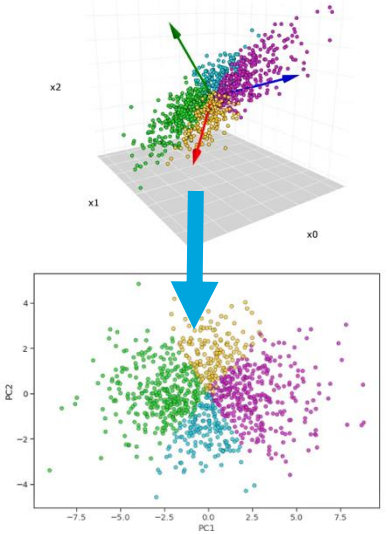
Machine Learning

Unsupervised Learning

Clustering



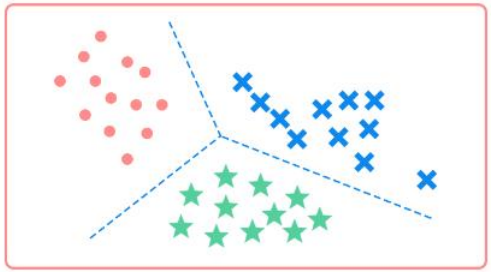
Dimension Reduction



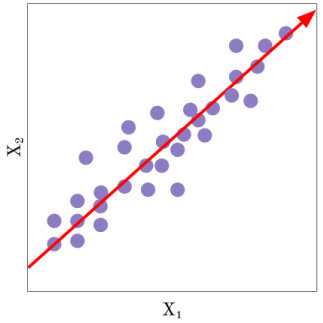
Association

Supervised Learning

Classification



Regression



K-means clustering algorithm

Clusters group data points together that share similarities

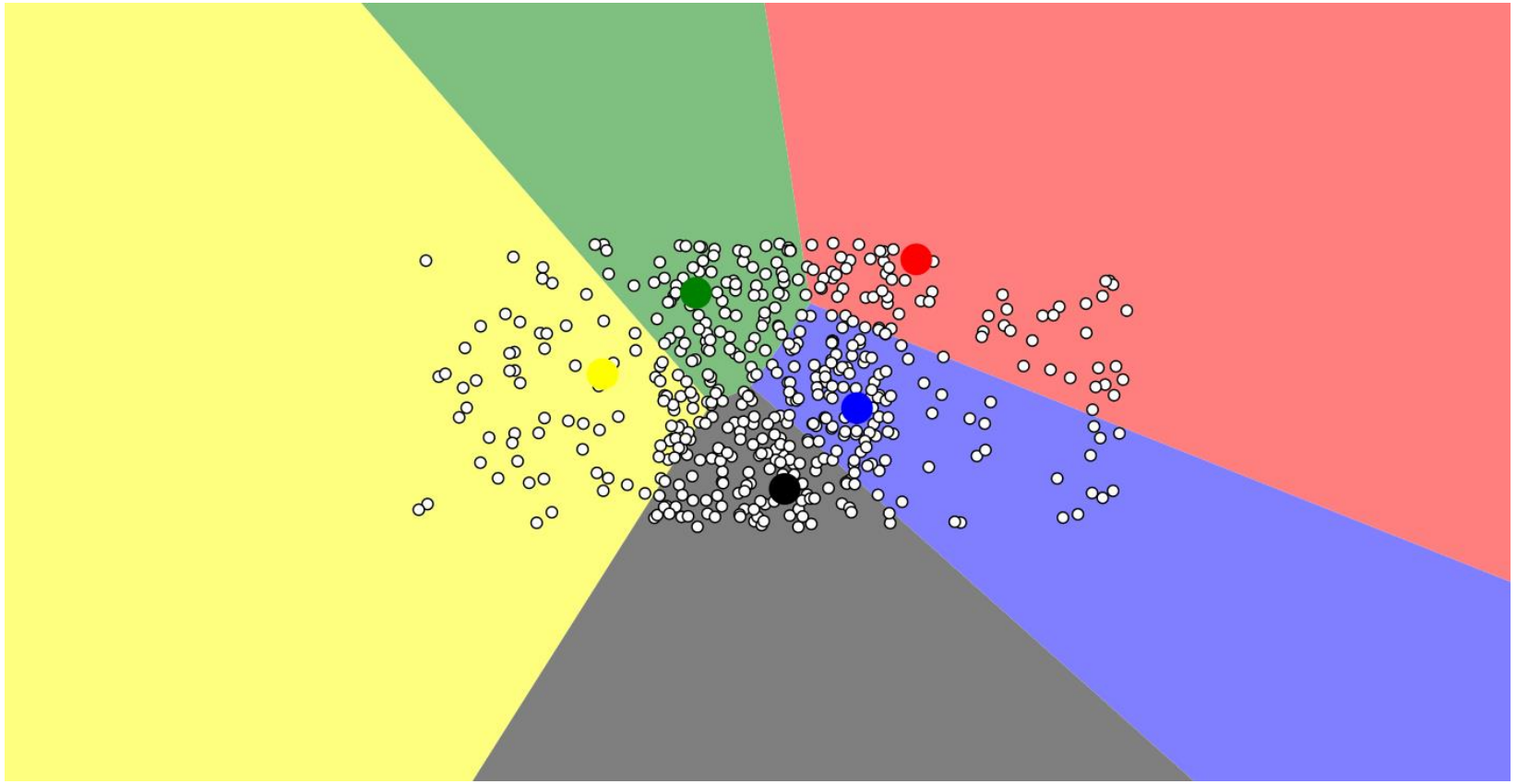
- N data points; k clusters → for each data point, assign it to one of the k groups
- Criteria: distance of data point from the center value (mean) of the kth group
- Method:

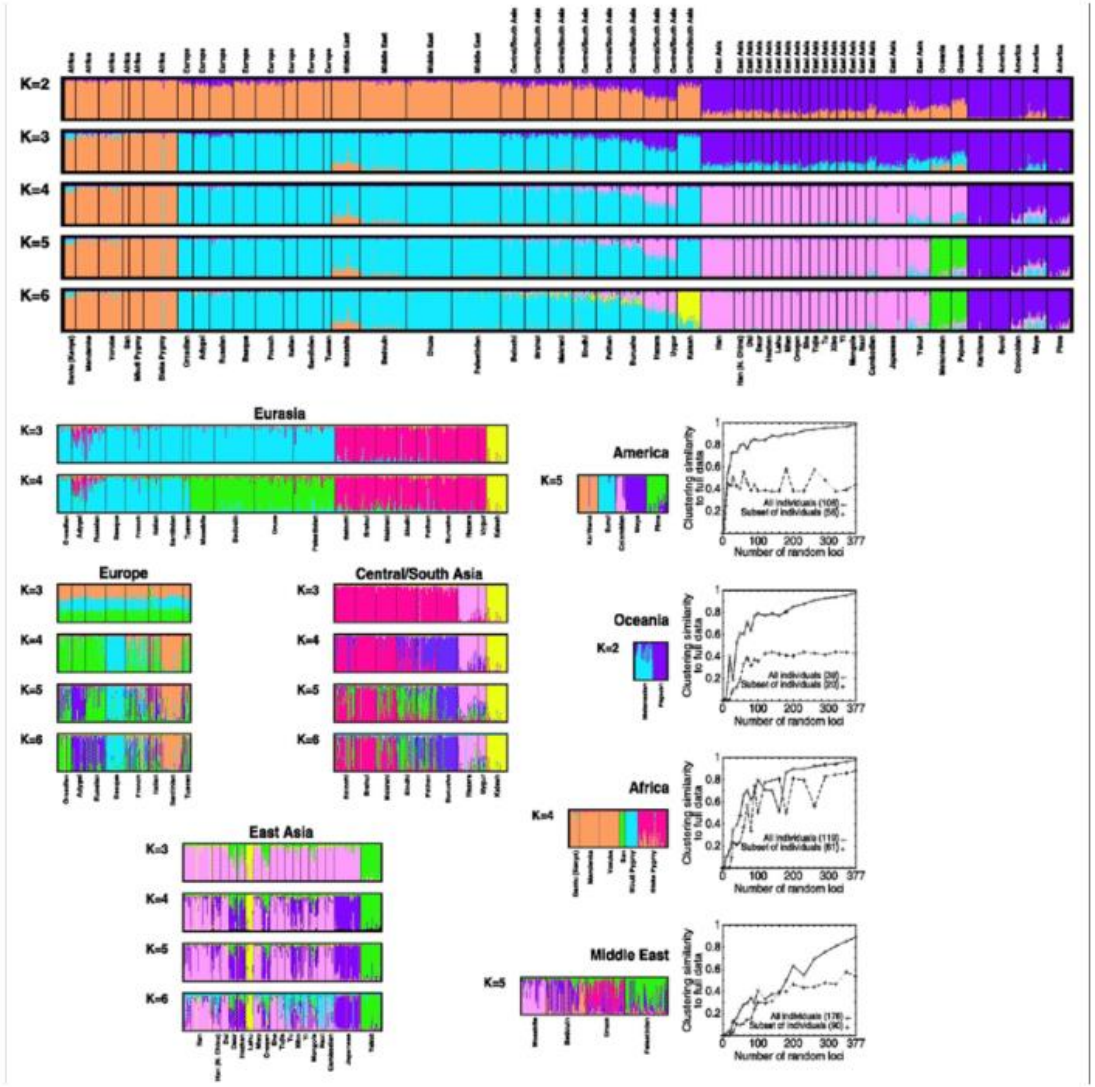
  1. **Random creation of k clusters** with centroids
  2. **Assignment** of each data point to a cluster based on shortest Euclidean distance to centroid
  3. **Updated centroids** (updating mean to include newly assigned point values), and the process repeats until a 'good' cluster is found (the value of the centroids stop changing between iterations)

- Sensitive to noise (data needs to be highly separated); highly dependent on initial assignment of centroids and k; can get stuck in local minima
- Pros: fast!

Code free way of visualizing k means!

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/





From Rosenberg et al. (2002) estimated population structure for the 52 sampled populations of the HGDP-CEPH panel for pre-chosen values of K = 2 through K = 6. Each cluster (K) is represented by a different color. Each individual is a vertical line, which depicts an estimate of that individual's membership in each cluster (multiple colors indicate membership in more than one cluster). Thin black lines denote individual populations. Population labels are shown at the bottom of the figure, while broad regional labels are listed at the top of the figure. While broad geographic clustering occurs, note that many individuals share genetic similarities with more than one cluster. This is particularly true within continents and for individuals from populations at the borders of continents

Principal **C**omponent **A**nalysis (PCA)

Identifies the major drivers of variation

Why PCA:

- Very few assumptions
- Non-parametric
- It **reduces** the dimensionality of your data
  - It may be surprising to you that you can reduce the dimensionality of your data without losing much information.
  - This occurs when the **variables are highly correlated**.
    - If you have included the following variables in your data set: arm length, leg length, height, you probably don't need them all – a linear combination of the three of them would capture the variation.
    - You can then use a smaller dataset of uncorrelated characteristics (or a smaller set of linear combinations of characteristics)
- Pearson, 1901 (yes, it is > 100 years old).

Principal **C**omponent **A**nalysis (PCA)

Identifies the major drivers of variation

Why PCA:

- Very few assumptions
- Non-parametric
- It **reduces** the dimensionality of your data

How PCA:

- It shifts the axes of your data from standard Cartesian axes (x, y, z) to a new set of axes, the PC axes, calculated from the largest Eigenvalue, and the second Eigenvalue (which is uncorrelated to the linear combination of the first Eigenvalue via….algebra. Magic, too, but mostly algebra).
- The largest Eigenvalue is a linear combination of the variables in your dataset.
- Therefore, these new axes demonstrate the linear characteristics that are most important in driving the variation of the characteristic under investigation.

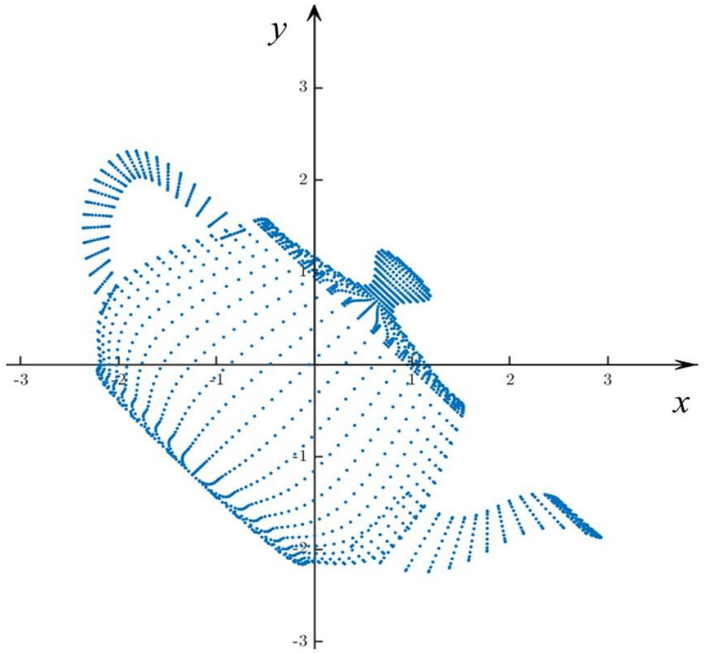
- This is A LOT, but I found some material that helps walk us through PCA principles without algebra

Principal **C**omponent **A**nalysis (PCA)

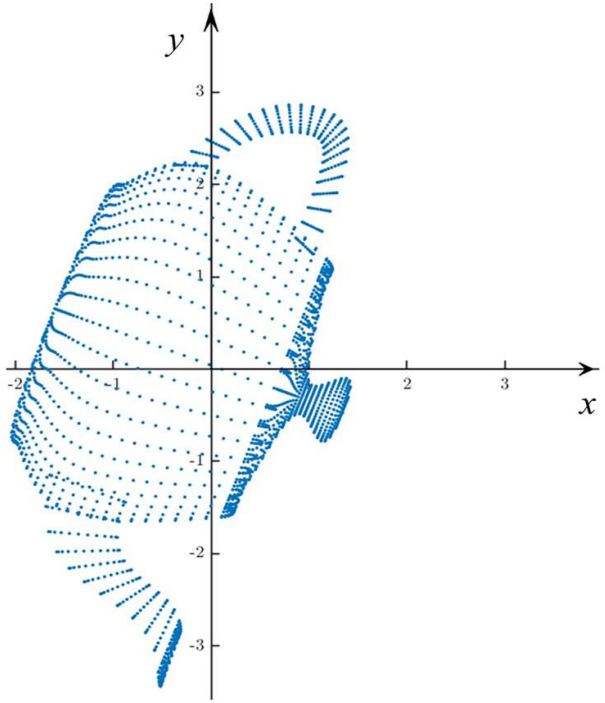
(A)

	A	B	C
1		VARIABLES	
2	Observation	x	y
3	1	0.64	-0.21
4	2	0.67	-0.22
5	3	0.70	-0.23
6	4	0.72	-0.23
7	5	0.73	-0.24
8	6	0.73	-0.24
9	7	0.73	-0.24
10	8	0.73	-0.24
11	9	0.72	-0.23
12	10	0.70	-0.23
13	11	0.67	-0.22
14	12	0.64	-0.21
15	13	0.56	-0.42
16	14	0.59	-0.43
17	15	0.62	-0.43
18	16	0.64	-0.44
19	17	0.65	-0.44
20	18	0.66	-0.44
21	19	0.66	-0.44
22	20	0.65	-0.43
23	...	...	...
4610	4608	-1.60	0.65
4611			

(B)

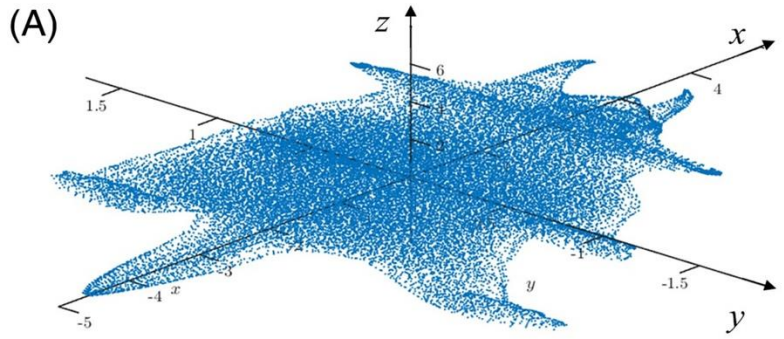


(C)



Principle **C**omponent **A**nalysis (PCA)

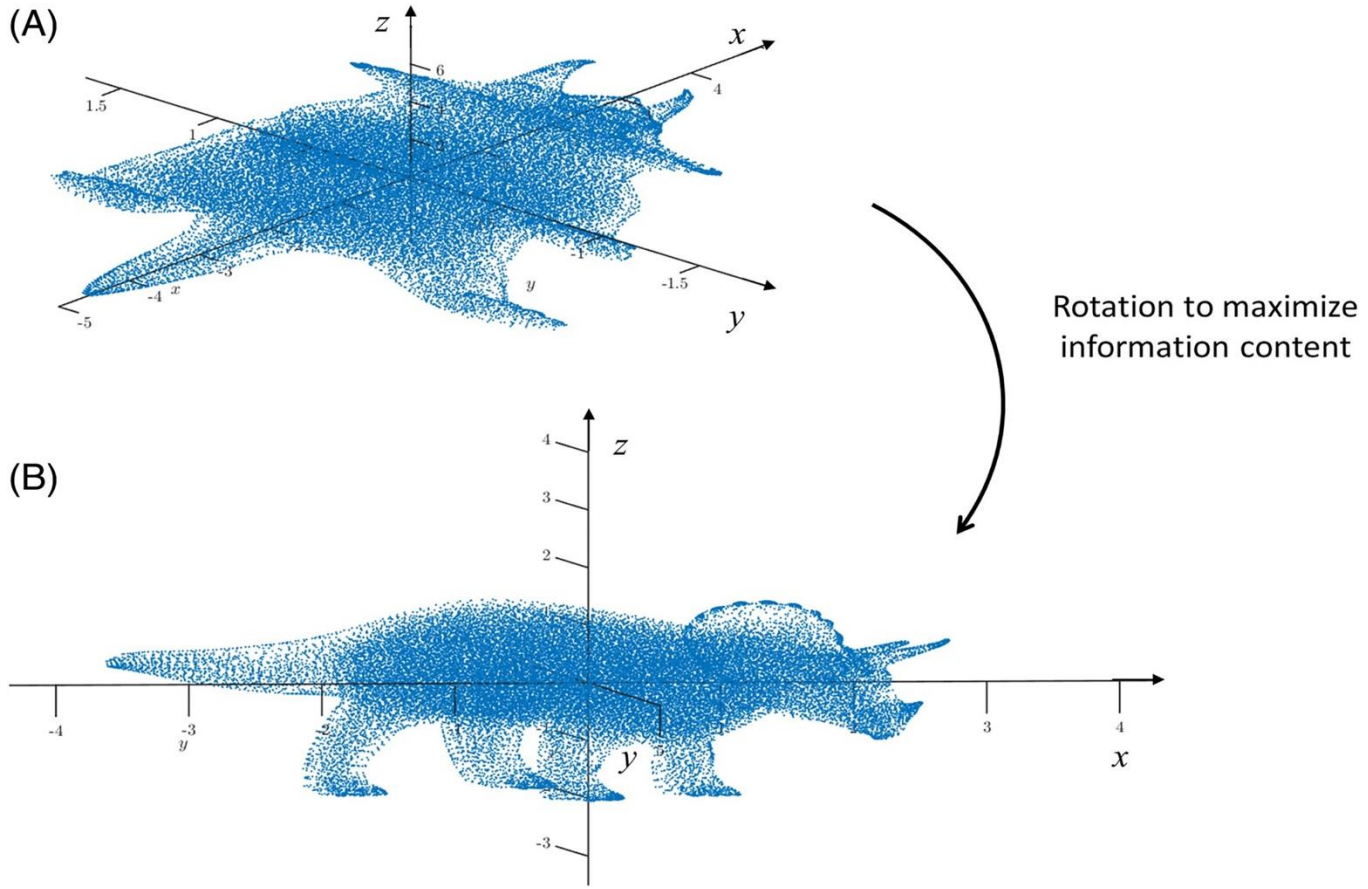
- Now we increase data points
- 3-D (x, y, z) for 36,876 points



Rotation to maximize information content

Principle **C**omponent **A**nalysis (PCA)

- Now we increase data points
- 3-D (x, y, z) for 36,876 points

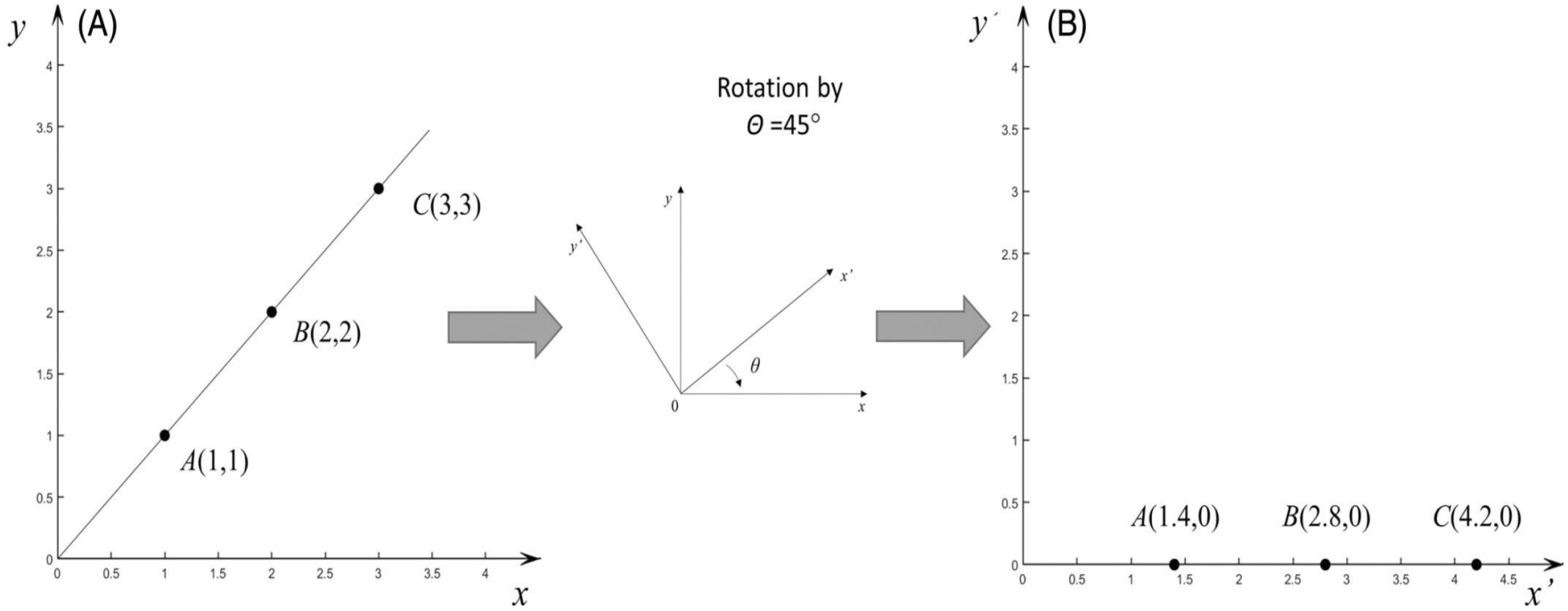


Teaching Statistics, Volume: 46, Issue: 1, Pages: 38-52, First published: 03 January 2024, DOI: (10.1111/test.12363)

Principle **C**omponent **A**nalysis (PCA)

- What happens if there are ≥ 4 variables? How do we uncover the structure of a high-dimensional data set?
- **Variation is information**; the more dispersed along an axis, the greater the information content is along that axis.
- You can use matrix decomposition:
  - N predictor variables → n x n covariance (or dispersion) matrix
  - Eigenvectors from this matrix gives us the direction of maximum variation
  - Eigenvalues weights the importance of the new axes

$$\mathbf{X} = \begin{bmatrix} & x & y \\ A| & 1 & 1 \\ B| & 2 & 2 \\ C| & 3 & 3 \end{bmatrix}.$$



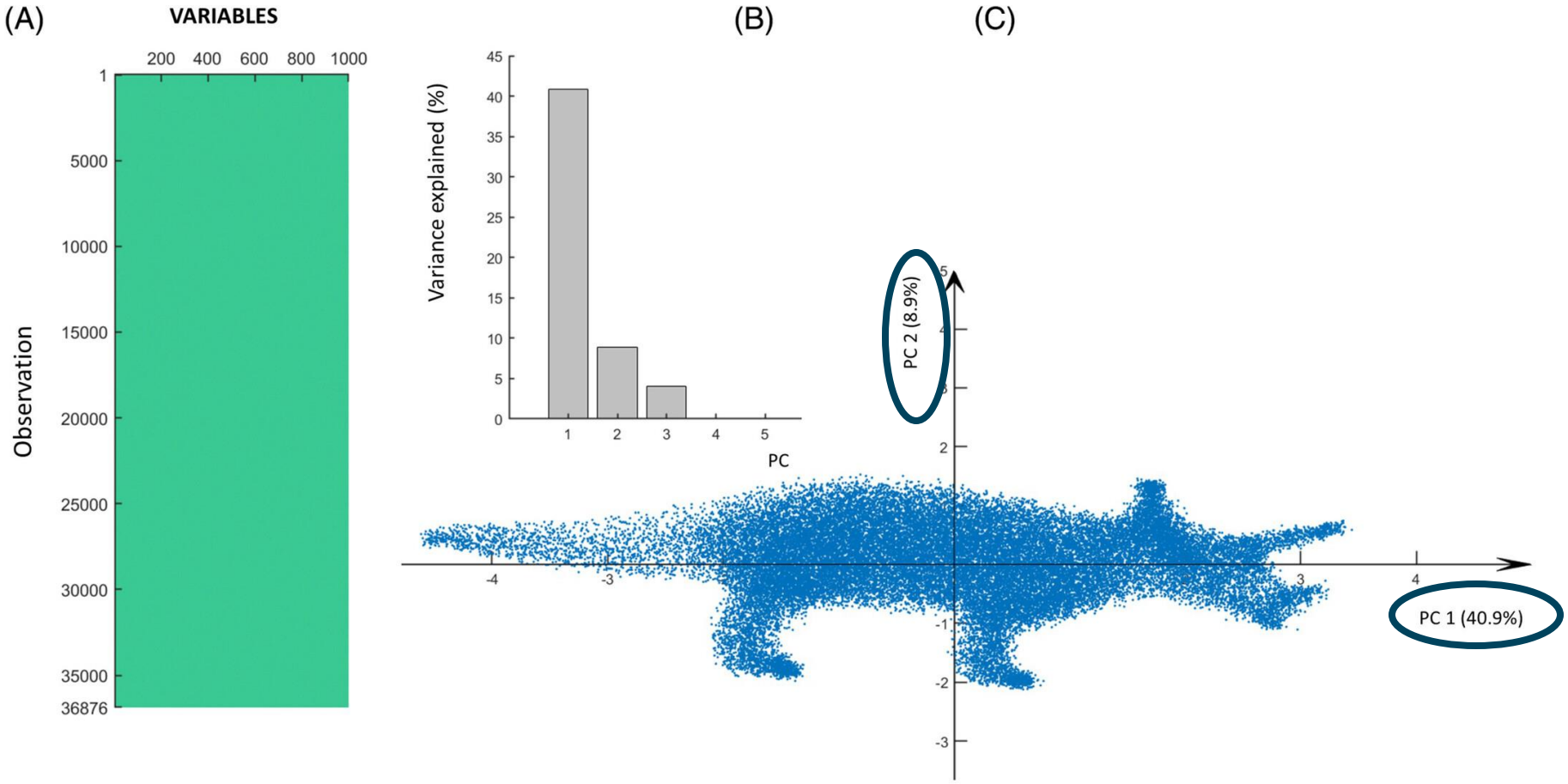
This becomes too complex as data become high dimensional.

Use Principal Components which are linear combinations of the original variables:

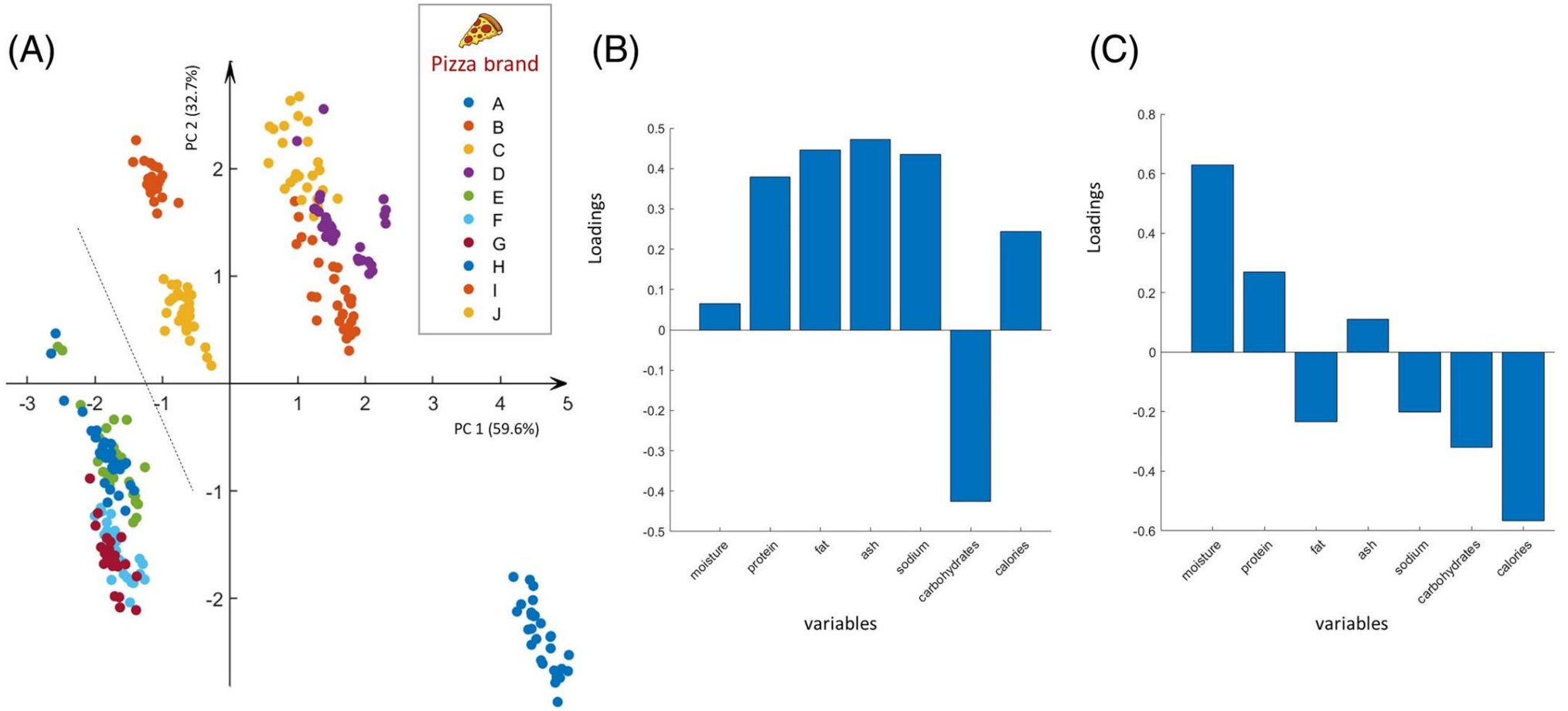
1. linear combination of data points are ordered by their variance
2. linear combination of data points are uncorrelated

Principal **C**omponent **A**nalysis (PCA)

Revisiting example, but with PCA:



Principal **C**omponent **A**nalysis (PCA)

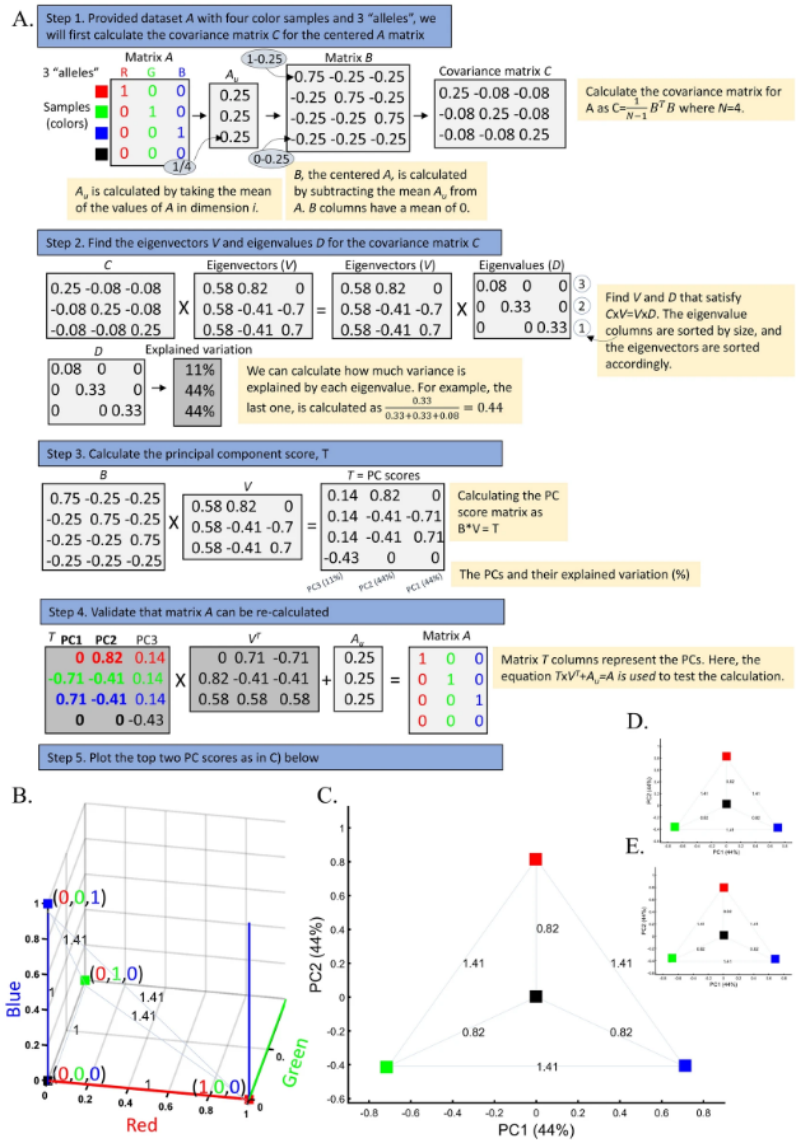


Teaching Statistics, Volume: 46, Issue: 1, Pages: 38-52, First published: 03 January 2024, DOI: (10.1111/test.12363)

Elhaik E. doi: 10.1038/s41598-022-14395-4. PMID: 36038559; PMCID: PMC9424212.

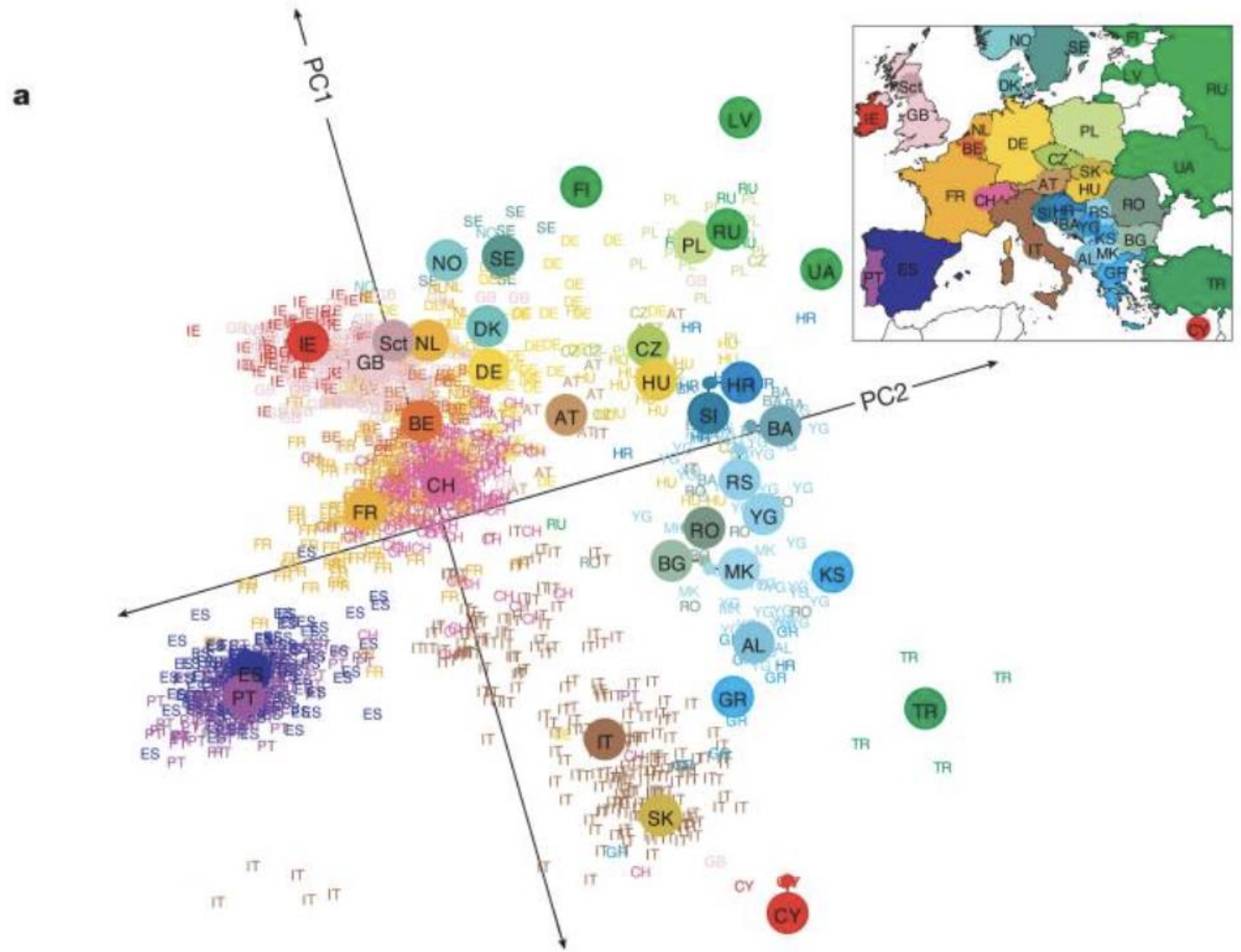
Figure 1

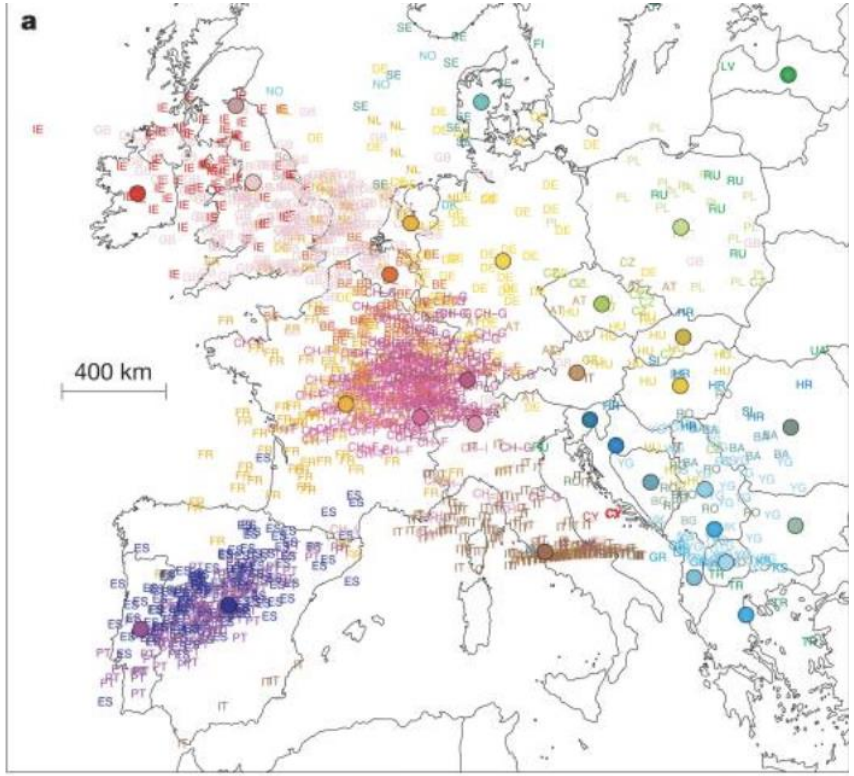
From: Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated

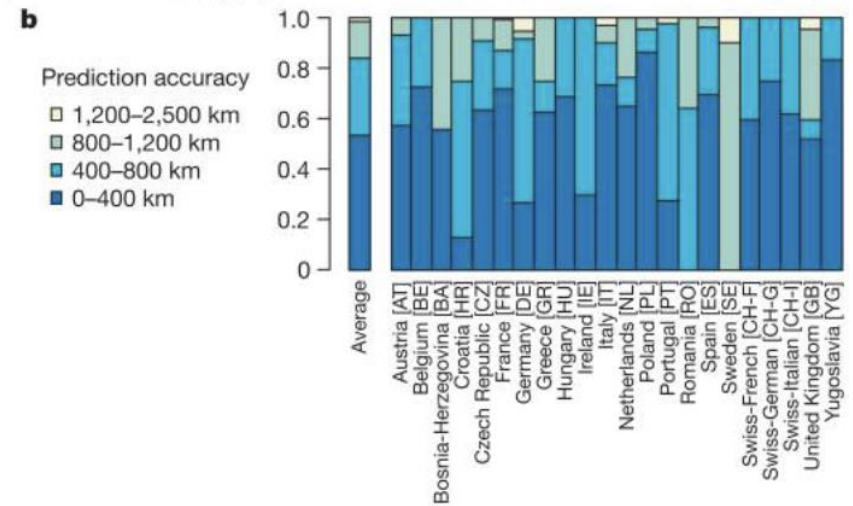


Applying PCA to four color populations. (**A**) An illustration of the PCA procedure (using the singular value decomposition (SVD) approach) applied to a color dataset consisting of four colors (n_{All}=1). (**B**) A 3D plot of the original color dataset with the axes representing the primary colors, each color is represented by three numbers ("SNPs"). After PCA is applied to this dataset, the projections of color samples or populations (in their original color) are plotted along their first two eigenvectors (or principal components [PCs]) with (**C**) n_{All}=1, (**D**) n_{All}=100, and (**E**) n_{All}=10,000. The latter two results are identical to those of (**C**). Grey lines and labels mark the Euclidean distances between the color populations calculated across all three PCs.

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686





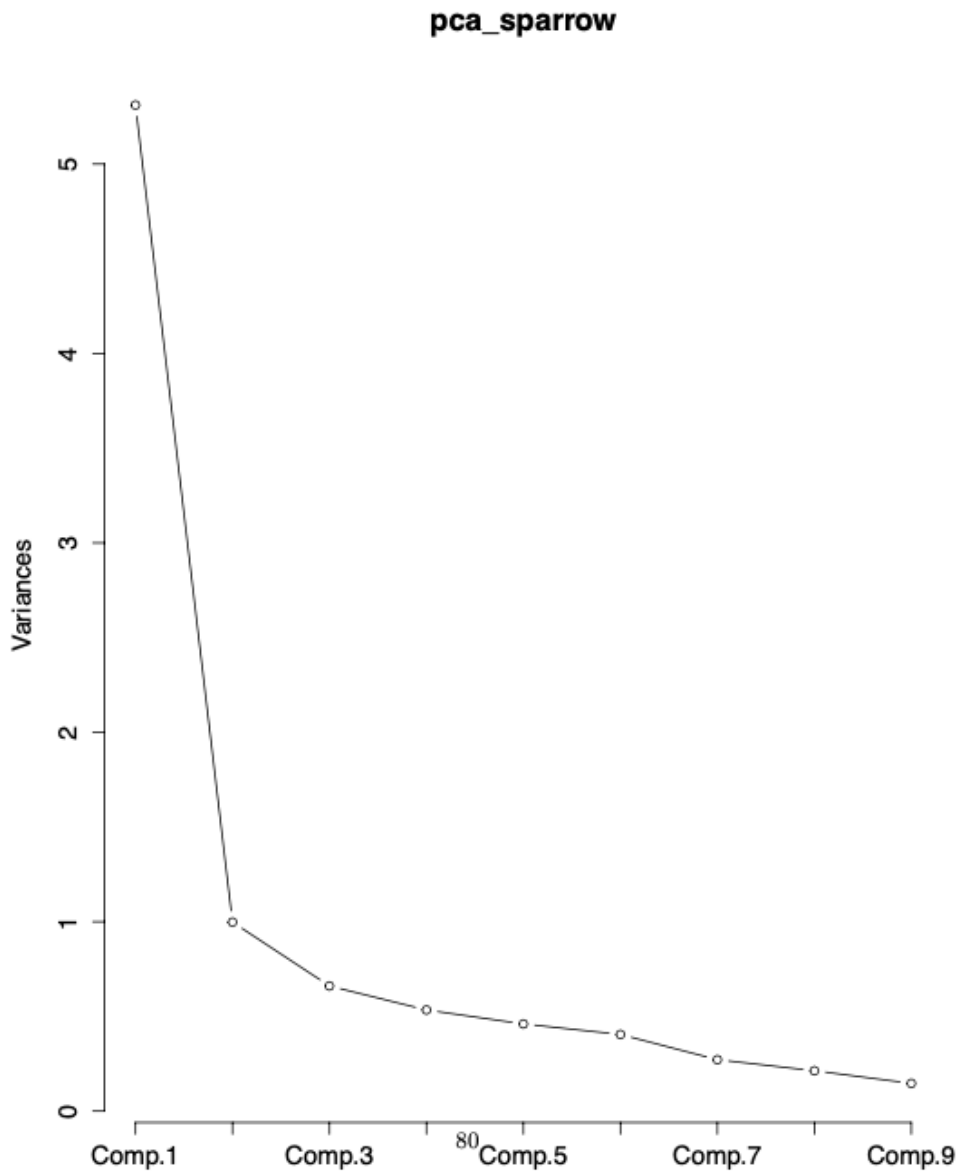


Novembre et al (2008). Genes mirror geography within Europe.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2735096/

- By removing the redundant variables, PCA captures which variables are actually foundational and are responsible for the phenomenon that we are measuring
- Bumpus Dataset.

Figure 11.1: This graph shows an "elbow" which gives a graphic reminder of how many variables you need to include to account for most of the variation in your data



Review of traditional Methods

Hypothesis testing

Possible Null distributions:

- Binomial
- X^2
- Normal
- Poisson
- F
- student's t

t-test
- One sample
- Paired
- Two Sample

ANOVA

Regression

Correlation

X^2 GOF

X^2 Contingency

Sign test

Mann-Whitney U

Kruskal-Wallis test

Spearman

Why use nonparametric tests at all?

Nonparametric tests **always** have less power than their parametric counterpart because you **always throw out information** by using only rank (and not magnitude): **type II error > β**

Why use nonparametric tests at all then?

- When used correctly, a nonparametric test should give a real Type I error rate = **α**

This seems kinda lame, right?

- But if you used a parametric test in its place (which would be using the parametric inappropriately since it doesn't meet the requirements), the parametric test will give a **type I error > α**

	Parametric	Nonparametric
Assumptions not met	**Type I > α**	**Type I = α**
Assumptions met	**Type II = β**	**Type II > β**

ACTUAL: indicated by Type I, Type II
STATED: indicated by α, β

Other “Modern” Statistics Methods

But there are many biologically interesting phenomenon that are not easily described by the tools we have examined so far….

Sometimes there is no standard method

Computers have dramatically expanded the toolkit of statistics/research

Computational methods:

When assumptions of best method available can’t be met

Random sampling is still assumed

No standard method exists

Massive amount of calculations

When we don’t know the null distribution

Two major categories of computational methods

Null sampling distributions:

1. Simulation – hypothesis testing

2. Randomization/Permutation

Precision of estimates:

3. Bootstrapping – sampling distribution of estimate; the values for the parameter estimates that we might obtain and their probabilities.

Two major categories of computational methods

Null sampling distributions:

1. Simulation – hypothesis testing

Determine the null distribution (from the parameters expected under the null hypothesis) by simulation of the sampling process

5 main steps

1. **Create and sample imaginary population**
   -parameters specified by null hypothesis
   -Same protocol that was used to collect real data
2. **Calculate test statistic on simulated sample**
3. **Repeat many times**
4. **Form the null distribution**
   - Gather simulated values for the test statistic
5. **Compare test statistic from the actual data to the null distribution**

This is a BROAD topic. We can't cover everything, but – since the first half of this module is on ML, these simulations will be relevant: https://chi-feng.github.io/mcmc-demo/

Randomization/Permutation (a resampling method):

- Asks: **are two variables independent?**
- **Assumptions:** random sampling, distribution of variables have approximately same shape

- Versatile
  - Variables can be any combination of numerical or categorical
  - We don't need a null hypothesis *because we build it ourselves*. A randomization test generates a ***null distribution*** for the association between two variables.
  - **MWU test is a type of permutation tests** – but you lose power when you use ranks instead of the actual data

- Basis: **Permutation**
  - Sampling without replacement
  - Method:
    1. Create data set
       - Response variable of a test statistic measuring association **randomly assigned to Explanatory variable**
         - **You are effectively exchanging labels**
       - **All data points are used exactly once**
    2. Calculate measure of association for randomized sample
    3. Repeat randomization many times
       - A NULL distribution

Pretty much gives you a p-value and not much else!

Randomization example:

The following is a very small data set of birth weights (in kg) of either singleton or individuals who were born with a twin. Create a legitimate randomized data set:

Singleton: 3.5, 2.7, 2.6, 4.4

Twin: 3.4, 4.2, 1.7

Two major categories of computational methods

Null sampling distributions:

1. Simulation – hypothesis testing

2. Randomization/Permutation

Precision of estimates:

3. Bootstrapping – sampling distribution of estimate; the values for the parameter estimates that we might obtain and their probabilities.

- You don't use the estimated sampling distribution to test a hypothesis; you only use it to get the SE!
- Usually rely on normal approximation for the sampling distribution.

Bootstrapping:

- ‘re-sampling’ the actual data
  - **Sampling with replacement**
  - Pick the original number of points for each group

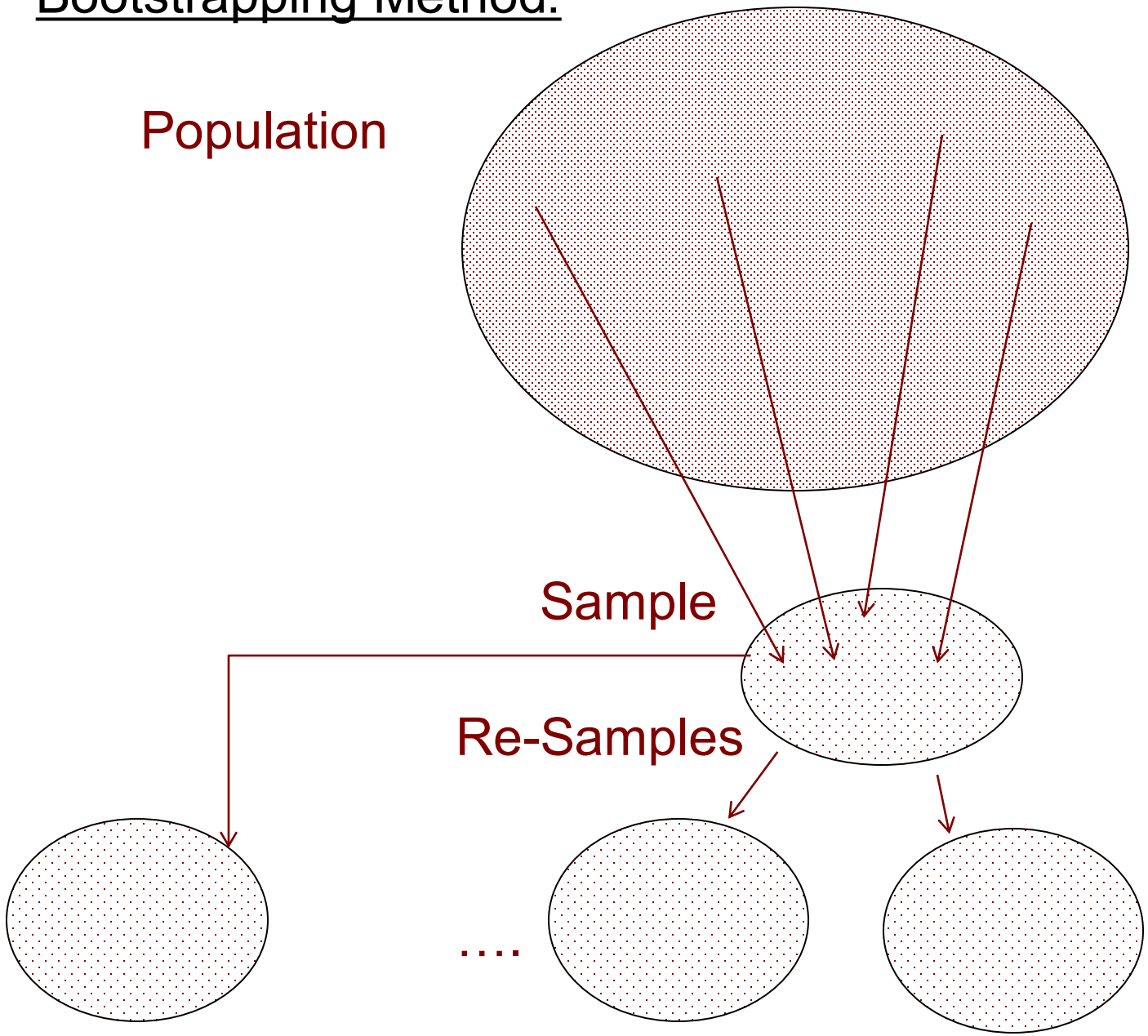
- Approximates the *sampling distribution* of an estimate
  - ***But NOT*** *the null (sampling) distribution as with simulation and randomization*

- Nonparametric and be applied to virtually any parameter – including means, proportions, correlations, linear model coefficients

- Used to find confidence interval and the bootstrap standard error
  - Precision method
  - Particularly useful when there is no ready formula for standard error (median, eigenvalue)

- Estimate uncertainty in phylogenies

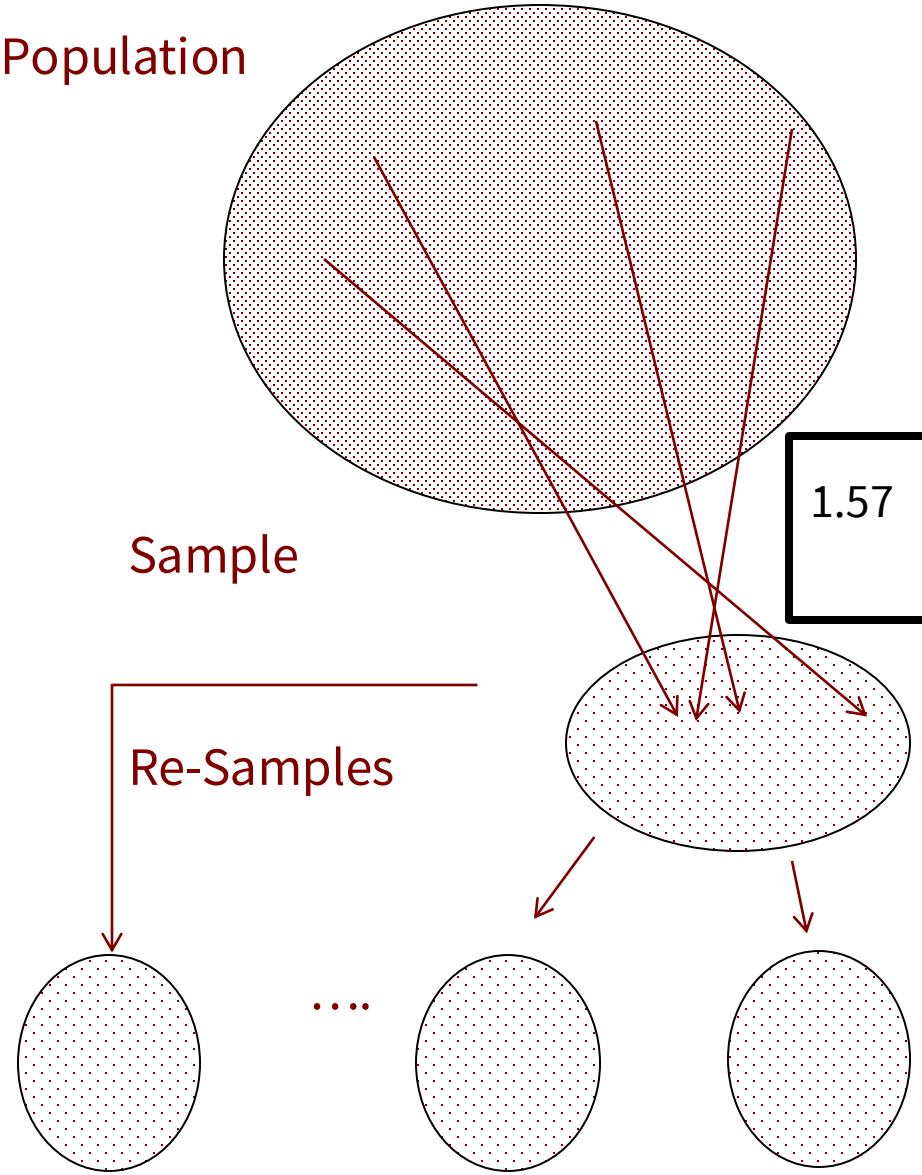
Bootstrapping Method:



Sample size: Large enough so that frequency distribution of sample is reasonable approximation of frequency distribution of population

Too small samples, result in standard errors that are too small and confidence errors are that are too narrow --> overestimate precision

Bootstrapping Method:

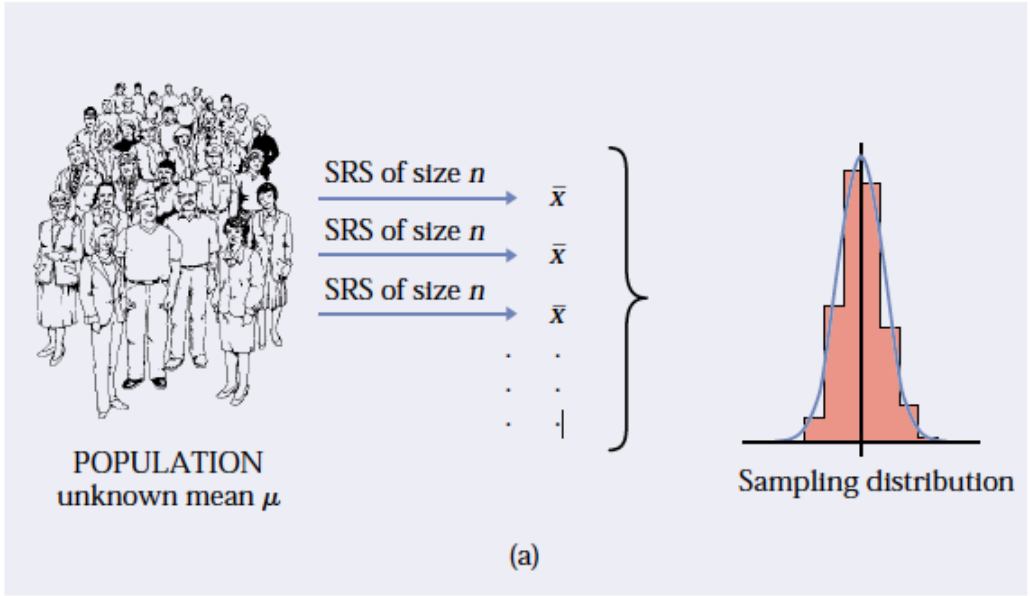


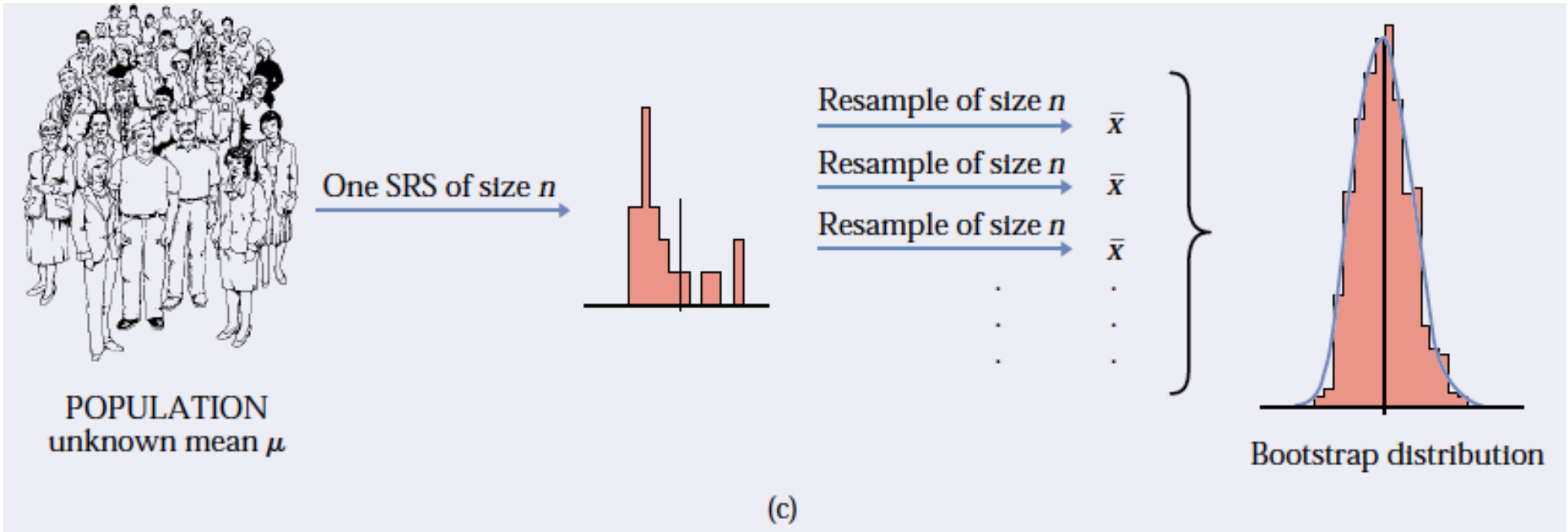
3.12 0.00 1.57 19.67 0.22 2.20
Mean = 4.46

1.57 0.22 19.67 0.00 0.22 3.12
Mean = 4.13

0.22 3.12 1.57 3.12 2.20 0.22
Mean = 1.74

0.00 2.20 2.20 2.20 19.67 1.57
Mean = 4.64





http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf

A bootstrap analysis with a small sample will cause:

---

a) a large standard error and a wide confidence interval

b) a small standard error and a narrow confidence interval

c) a large standard error and a narrow confidence interval

d) a small standard error and a wide confidence interval