PCA and DNA: Using Math to Unlock Genes

Tom Alberts

University of Utah
Department of Mathematics

U Masters of Statistics Program

- Campus wide statistics program with specializations in mathematics, biostatistics, econometrics, educational psychology, sociology
- Takes 1.5-2 years to finish after the undergrad degree
- Working students take it part time (evening classes) and finish in 3 years
- Classes in statistical inference, regression analysis, multilinear models, mathematical statistics, and probability theory
- Also a data science track available, which is more computational
- Graduates get local jobs in banking, insurance, data science/software, medical research groups, etc

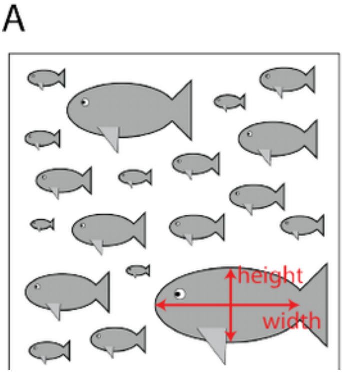
U Masters of Statistics Program

- Campus program: mstat.utah.edu
- Mathematics specialization: www.math.utah.edu/graduate/programs.php

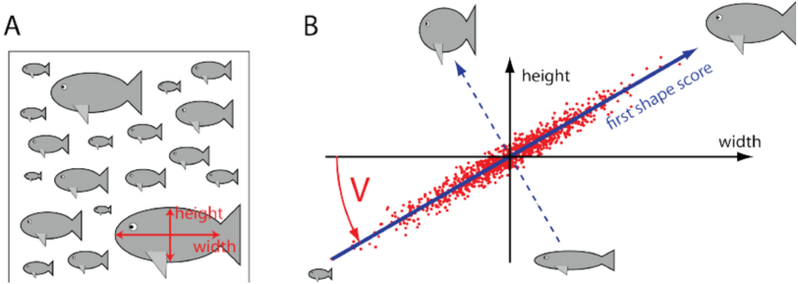
Principal Components Analysis

- Statistical method for determining the structure of large data sets
- If each point in the data set contains multiple measurements, it can be used to determine which combinations of measurements are the most important ones and which can be ignored (**dimension reduction**)

Example Data Set: Fish Measurements



Example Data Set: Fish Measurements



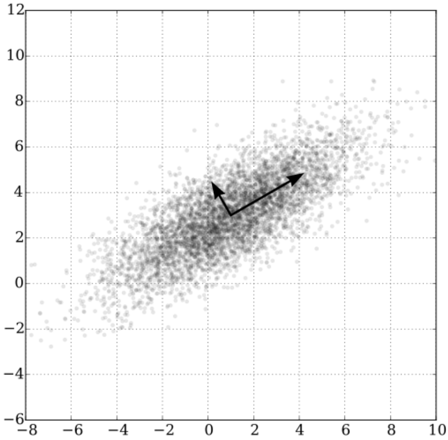
Applications to DNA and Population Modelling

- In a 2007 study, 3192 individuals had their DNA sequenced
- Each person's DNA was genotyped at 500,568 locations
- Which combination of these 500,000+ DNA measurements are most explanatory? Do they mean anything?

Best Fitting Ellipses

- Scatterplots of data sets often have an elliptical shape
- With higher dimensional data the points contained in an ellipsoid
- Given a data set there is a notion of an “best fitting ellipsoid”
- Principal components analysis finds it for you

Fitting an Ellipse



Geometry of Ellipsoids

- An ellipsoid in d dimensions is determined by d **principal axes**
- Principal axes are all perpendicular to each other
- Also need the length of the ellipsoid along each principal axis
- Finally, need the center of the ellipse (mean point of the data cloud)

Equation of An Ellipse

- Center our ellipses at zero
- In 2-dimensions:

$$ax^2 + 2bxy + cy^2 = r^2$$

where $ac - 4b^2 > 0$

$$\text{ax}^2 + 2bxy + cy^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Equation of An Ellipse

- Center our ellipses at zero
- In higher-dimensions:

$$\sum_{i=1}^{d} a_i x_i^2 + \sum_{i \neq j} b_{ij} x_i x_j = r^2$$

Represent as a quadratic form by a $d \times d$ symmetric matrix. Diagonalization gives principal axes and lengths.

Working with the Data Points

- What matrix to diagonalize for a cloud of data points?
- **Covariance matrix**

Setup:

- For n different individuals, collect d different numerical measurements
- Represent the measurements of each individual by a vector in $\mathbb{R}^d$
- Thus the data set becomes n points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$

Covariance Matrix from Data Points

$$\textbf{Average vector: } \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \in \mathbb{R}^d$$

- Thus $\bar{\mathbf{x}}(j) =$ average value of jth measurement, $j = 1, \ldots, d$, across all n individuals
- We center our ellipse at the point $\bar{\mathbf{x}}$

Covariance Matrix from Data Points

Sample variance of jth measurement:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))^2$$

Sample covariance between jth and kth measurement:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))(\mathbf{x}_i(k) - \bar{\mathbf{x}}(k))$$

Note if $j = k$, then sample covariance = sample variance

Covariance Matrix from Data Points

Sample variance of jth measurement:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))^2$$

Sample covariance between jth and kth measurement:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))(\mathbf{x}_i(k) - \bar{\mathbf{x}}(k))$$

Sample variance measures the **spread** of the jth measurement (how much it concentrates near its mean)

Covariance Matrix from Data Points

Sample variance of jth measurement:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))^2$$

Sample covariance between jth and kth measurement:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))(\mathbf{x}_i(k) - \bar{\mathbf{x}}(k))$$

Sample covariance measures the **strength of the linear relationship** between measurements j and k

Covariance Matrix from Data Points

Sample covariance measures the **strength of the linear relationship** between measurements j and k

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i(j)-\bar{\mathbf{x}}(j))(\mathbf{x}_i(k)-\bar{\mathbf{x}}(k)) = \frac{1}{n}\mathbf{z}_j^T\mathbf{z}_k = \frac{1}{n}||\mathbf{z}_j||\,||\mathbf{z}_k||\cos\theta$$

where $\mathbf{z}_j(i) = \mathbf{x}_i(j) - \bar{\mathbf{x}}(j)$ is the n-dimensional vector of the jth measurements, recentered about their means. Analogous for $\mathbf{z}_k$. Then θ is the angle between the vectors.

Covariance Matrix from Data Points

Let $\mathbf{X}$ be the $d \times n$ matrix with columns being the measurements for different individuals:

$$\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_n]$$

Let $\bar{\mathbf{X}}$ be the $d \times n$ matrix with rows being the averages of the different measurements:

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{\mathbf{x}}(1) \rightarrow \\ \bar{\mathbf{x}}(2) \rightarrow \\ \vdots \\ \bar{\mathbf{x}}(d) \rightarrow \end{bmatrix}$$

Covariance Matrix from Data Points

Then $\hat{\mathbf{\Sigma}} := \frac{1}{n}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T$ is a $d \times d$ matrix with entries

$$\hat{\mathbf{\Sigma}}_{jk} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i(j) - \bar{\mathbf{x}}(j))(\mathbf{x}_i(k) - \bar{\mathbf{x}}(k))$$

$\hat{\mathbf{\Sigma}}$ is called the **sample covariance matrix.** It is computed *from the data*.

Properties of Covariance Matrix

- $\hat{\boldsymbol{\Sigma}}$ is $d \times d$
- $\hat{\boldsymbol{\Sigma}}$ is symmetric
- $\hat{\boldsymbol{\Sigma}}$ is positive definite
- $\hat{\boldsymbol{\Sigma}}$ is a *Gram matrix*

Thus can diagonalize:

$$\hat{\boldsymbol{\Sigma}} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \ldots + \lambda_d \mathbf{v}_d \mathbf{v}_d^T$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$, and $\hat{\boldsymbol{\Sigma}} \mathbf{v}_i = \lambda_i \mathbf{v}_i$

Dimension Reduction

$$\widehat{\Sigma} = \widehat{\Sigma}_d = \sum_{i=1}^{d} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

with $\lambda_1 > \lambda_2 > \ldots > \lambda_d$. Idea of dimension reduction is to approximate by

$$\widehat{\Sigma}_r \approx \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

with $r \ll d$.

Dimension Reduction

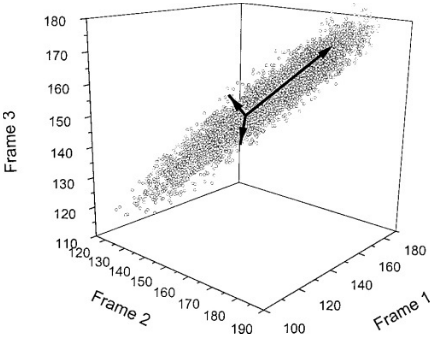
$$\widehat{\Sigma}_r \approx \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

Can use this to consider the reduced dataset

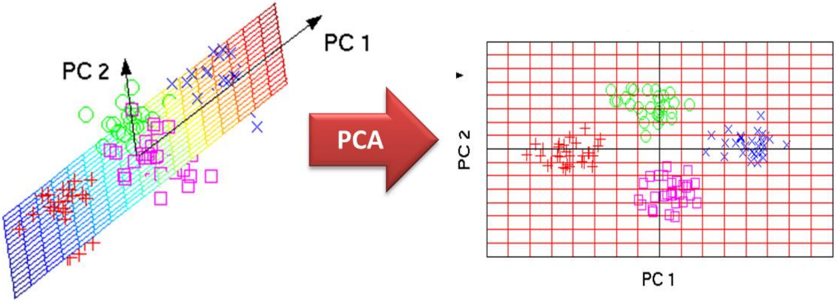
$$\left\{ \widehat{\Sigma}_r \mathbf{x}_i \right\}_{i=1}^{N}$$

which lives in the r-dimensional space $\mathrm{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$

Dimension Reduction



Dimension Reduction



Dimension Reduction

- There is a **variational characterization** of PCA that explains why the method is useful
- Suppose you had to reduce the dataset from d dimensions to 1 dimension, by **projecting** the data points onto a line. Which line would you choose?
- It turns out that the "best" line is exactly in the direction of $\mathbf{v}_1$, the eigenvector corresponding to the largest eigenvalue
- When you project onto this line, the new one-dimensional dataset has the **largest** variance possible among all possible choices of the line. Thus you preserve as much of the data's "character" as is possible.

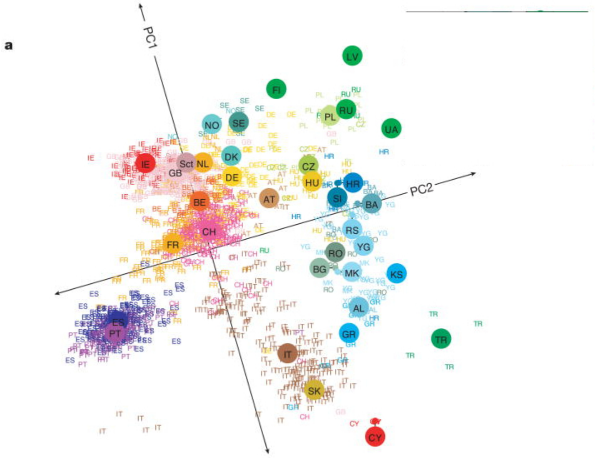
Dimension Reduction

- If instead you could reduce the data from d dimensions to 2 dimensions, then the “best” choice is to project onto $\mathrm{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$
- Contains as much of the variance as is possible among all two-dimensional planes to project onto

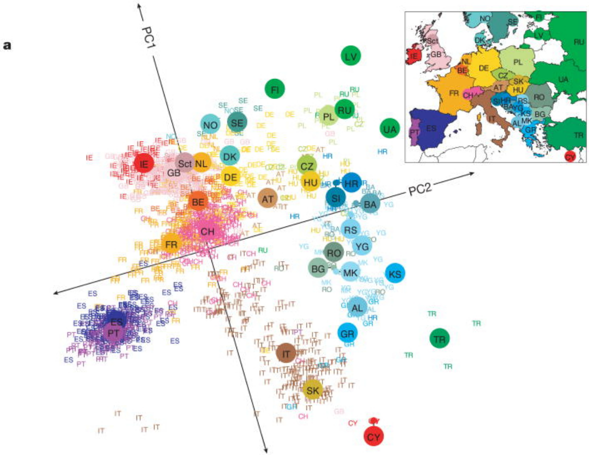
Applications to DNA and Population Modelling

- In a 2007 study, 3192 individuals had their DNA sequenced
- Each person's DNA was genotyped at 500,568 locations
- Which combination of these 500,000+ DNA measurements are most explanatory? Do they mean anything?

First Two Principal Components of DNA Data



First Two Principal Components of DNA Data



First Two Principal Components of DNA Data