# Statistically Speaking: Dimensionality reduction with PCA

DAVID L. TABB, PH.D.

NOVEMBER 9, 2017

# Overview

- Why we need dimensionality reduction
- The value of principal component analysis
- The story of Harold Hotelling

# Dimensionality Reduction

- We often think of biological data as a table; each column is a sample, and each row represents some property or entity we measure for each sample.

*Example: columns are patients and rows are transcripts for which we measure expression*

- The set of measurements for each sample can be thought of as coordinates in *n* dimensions, where *n* is the number of measurements.

# Too many measurements?

- Different measurements may be redundant. They may contain *mutual information* or be *correlated* with each other.
  - Genes may be expressed in response to the same transcription factor.
  - Length and mass measurements scale with physical size.

# Why is that a problem?

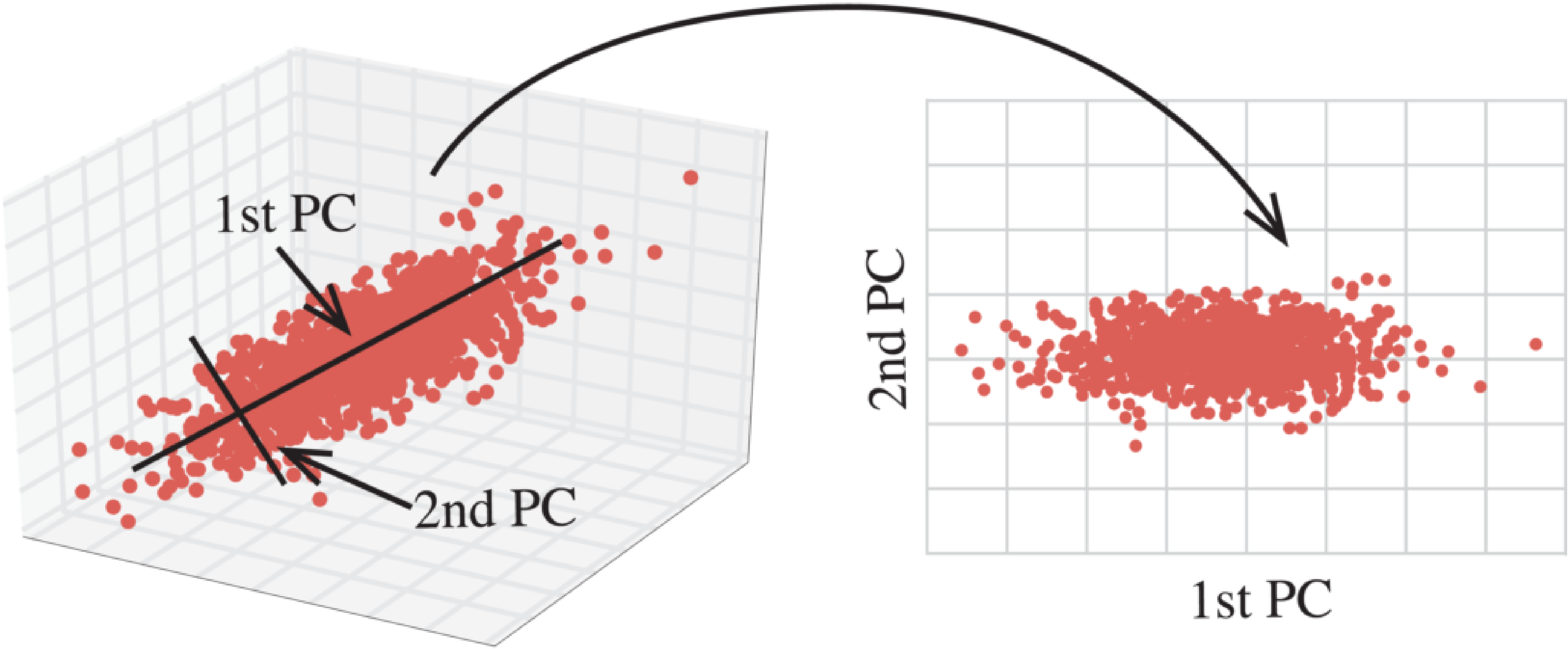
- You are making a classifier. You seek the five genes changing the most between cohorts. When you combine the five genes to one classifier, discrimination doesn't improve.
- Any time you have far more measurements than you have samples, multiple testing correction can limit sensitivity.

# Principal Component Analysis

- “Component:” PCA accepts a table of of *n* metrics for each sample and returns *n* component values for each sample.
- “Principal:” PCA prioritizes the components by the sample variability they contain.
- Input metrics may be correlated, but output components minimize this relationship.

http://www.theanalysisfactor.com/tips-principal-component-analysis/

# PCA rotates data in *n*-D space



http://blog.kaggle.com/2017/04/10/exploring-the-structure-of-high-dimensional-data-with-hypertools-in-kaggle-kernels/

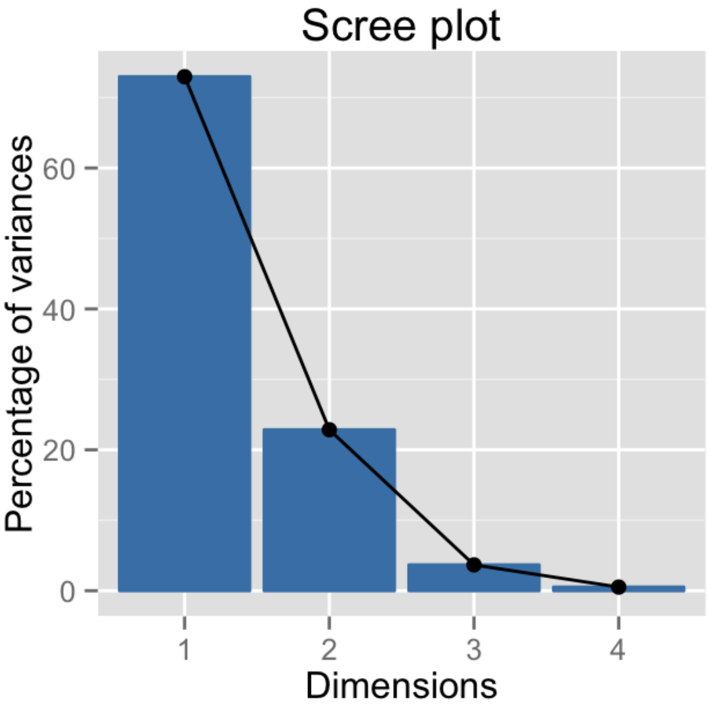
# Linear combinations

- Each component is formed by adding together the original variables, each multiplied by a weight for that row.

$$C_1 = v_1w_1 + v_2w_2 + v_3w_3 + v_4w_4 \ldots$$

# Screeplots and variance

- You must keep enough components to account for substantial variability.
- How many components are necessary for first ½? First ¾? First 7/8?

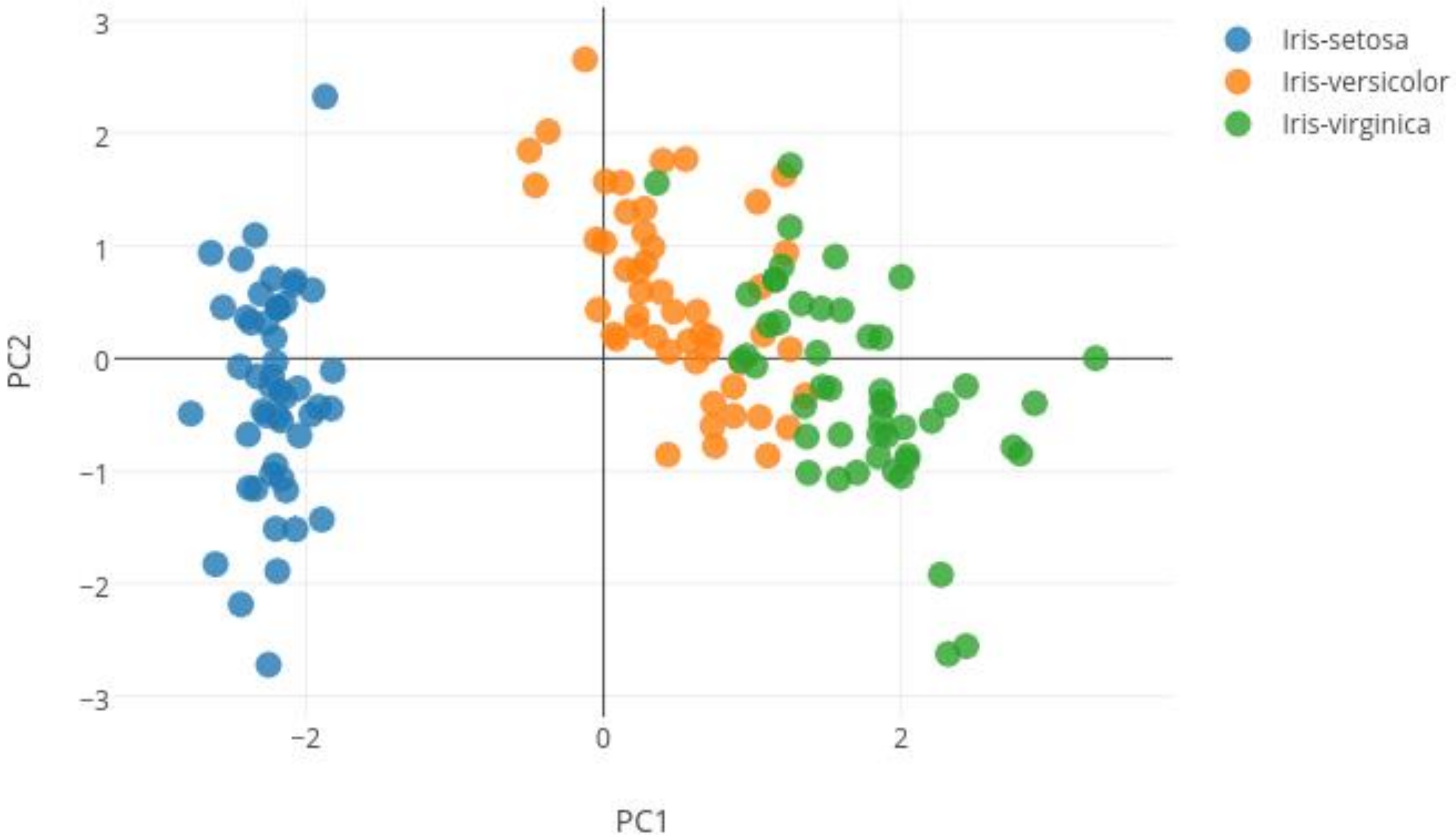


See also "eigenvalues"

http://www.sthda.com/english/wiki/

R. Cangelosi and A. Goriely, *Biology Direct* (2006) 2:2

# PCA plot typically shows first two components



https://plot.ly/ipython-notebooks/principal-component-analysis/

# Harold Hotelling:statistician, economist, and traveler

stat-or.unc.edu/support

- Created Hotelling $T^2$ for multivariate hypothesis testing (key to QC)
- Extended RA Fisher’s methods to create principal component analysis
- Headed U-North Carolina Institute of Mathematical Statistics (1946)
- Pioneered econometrics of shared resources
- Lived in USA, UK, India, Argentina

https://www.ncpedia.org/biography/hotelling-addyson-harold

# Takeaways

- When dozens, hundreds, or thousands of measurements are made per sample, dimensionality reduction is a worthy goal.
- Principal component analysis is classic, and it appears in most statistical environments.
- PCA is most useful in visualizing overall scatter of samples through measured space.