

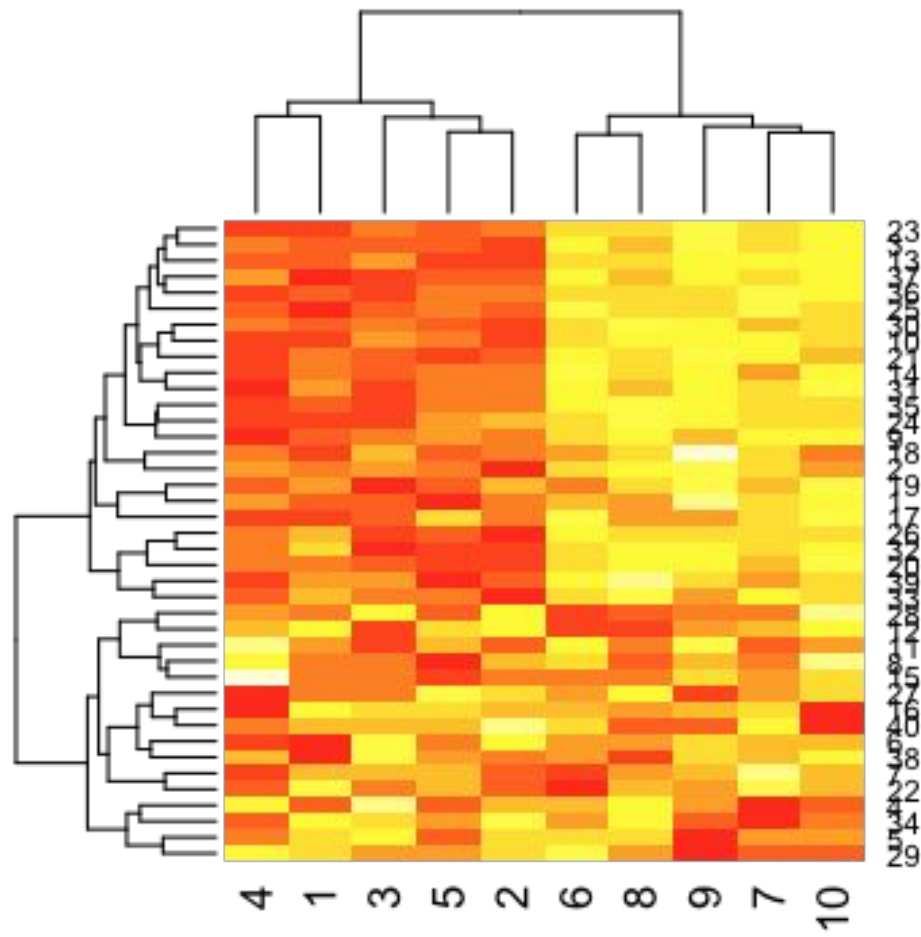
# Dimension reduction

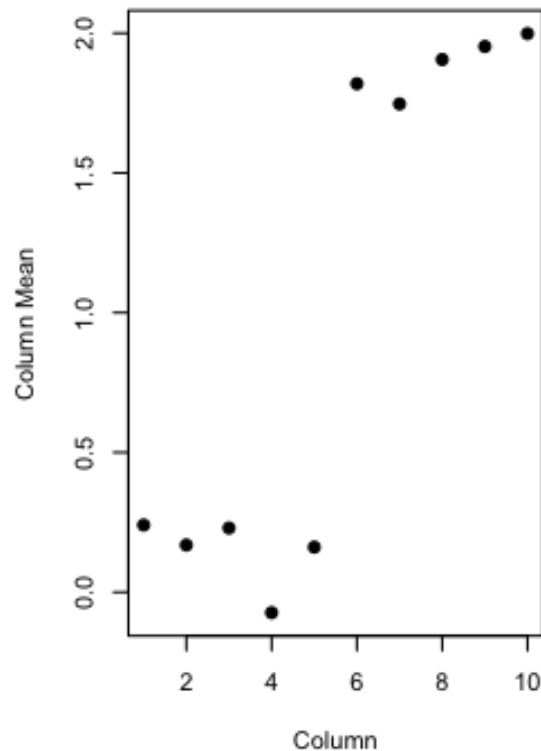
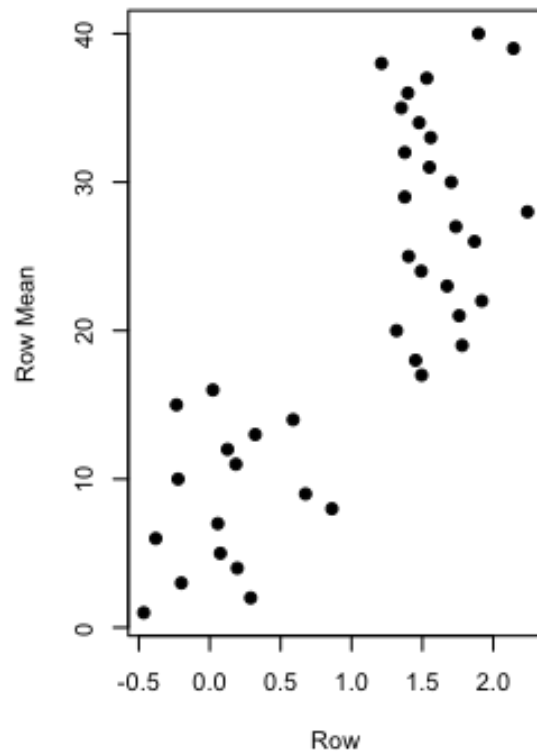
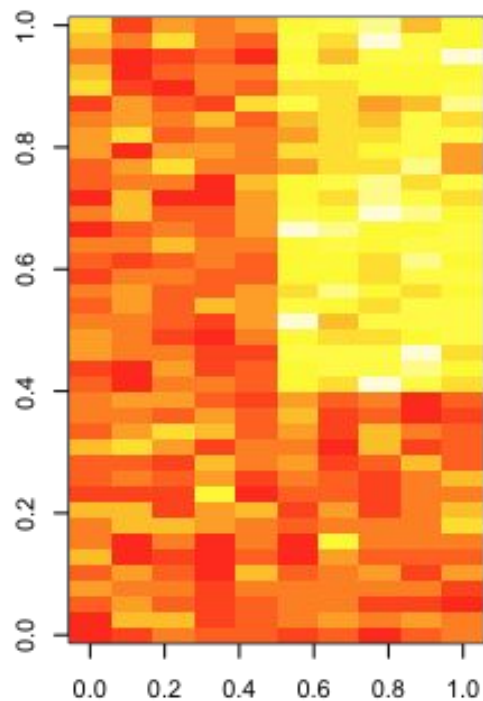
---

Jeff Leek

@jtleek

[www.jtleek.com](http://www.jtleek.com)





# PCA and SVD

PCA & SVD have different math goals

SVD can be used to estimate PCs

First proposed in genomics by

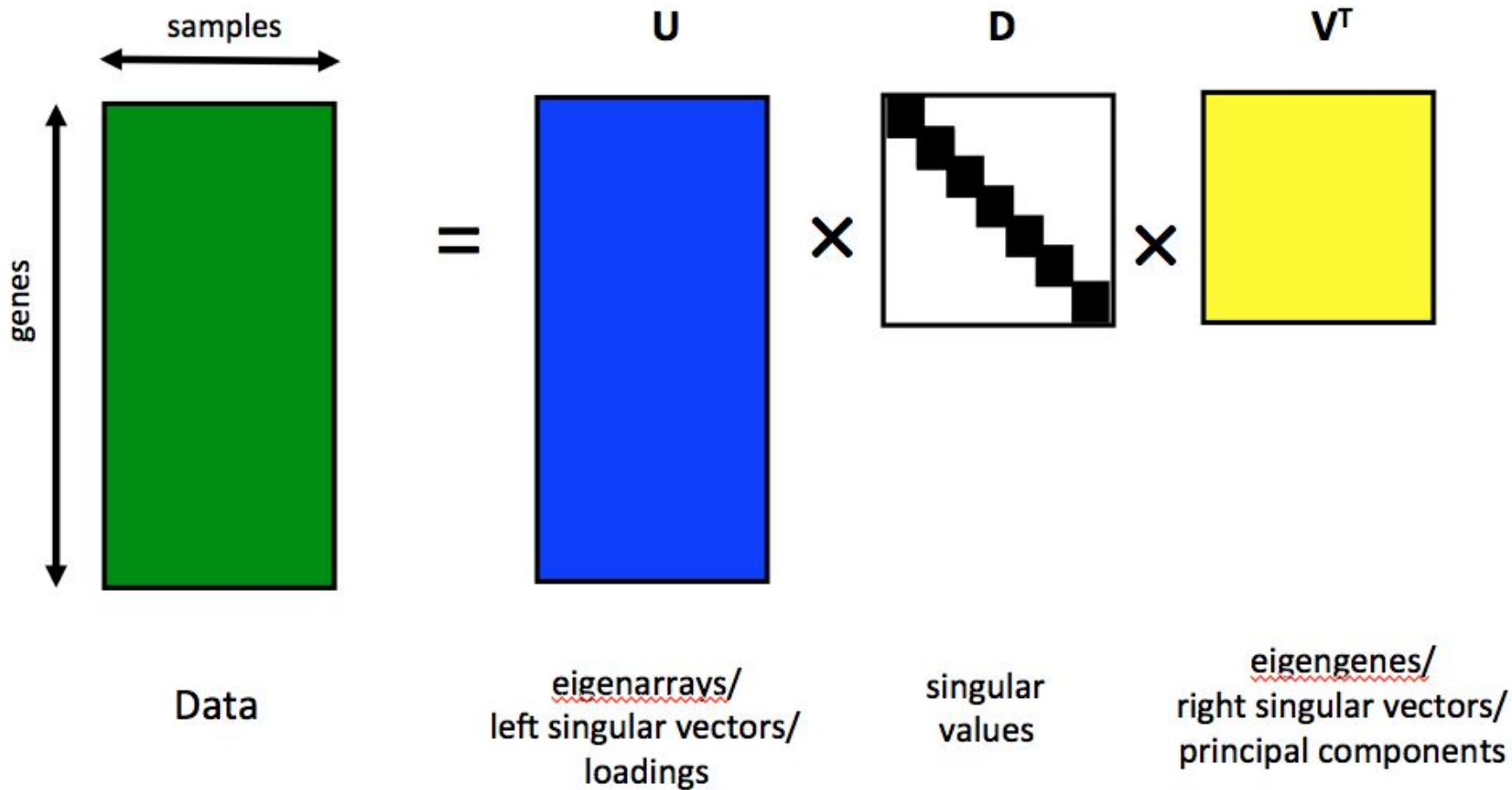
Alter et al. 2000 PNAS

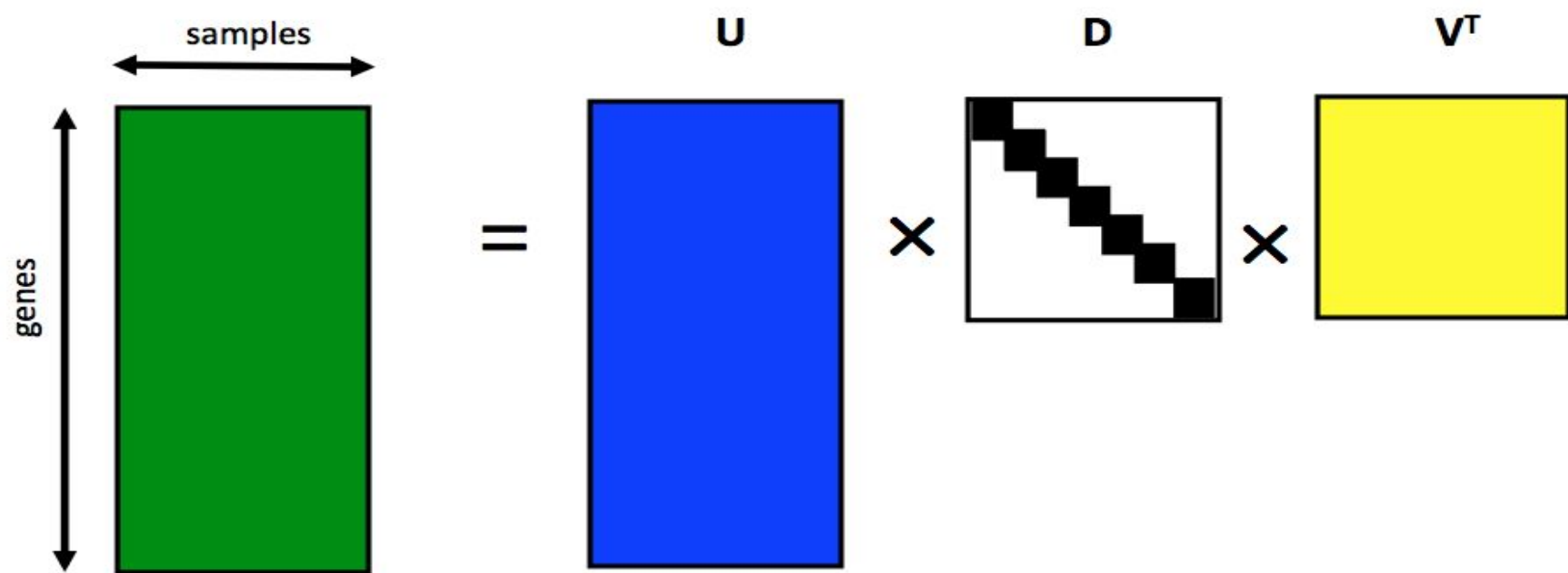
# Related Problems

You have multivariate matrix of data  $\mathbf{X}$

- Find a new set of multivariate variables that are uncorrelated and explain as much variance across rows as possible.
- Find the best matrix created with fewer variables (lower rank) that explains the original data.

The first goal is statistical and the second goal is data compression.



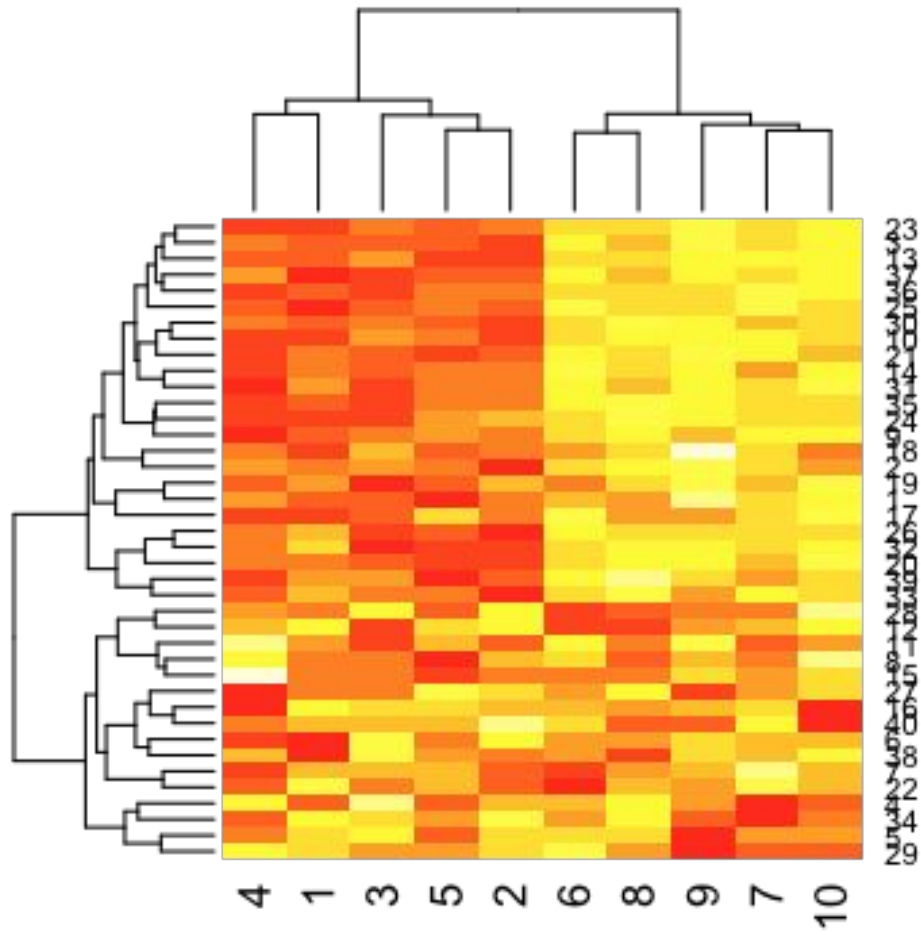


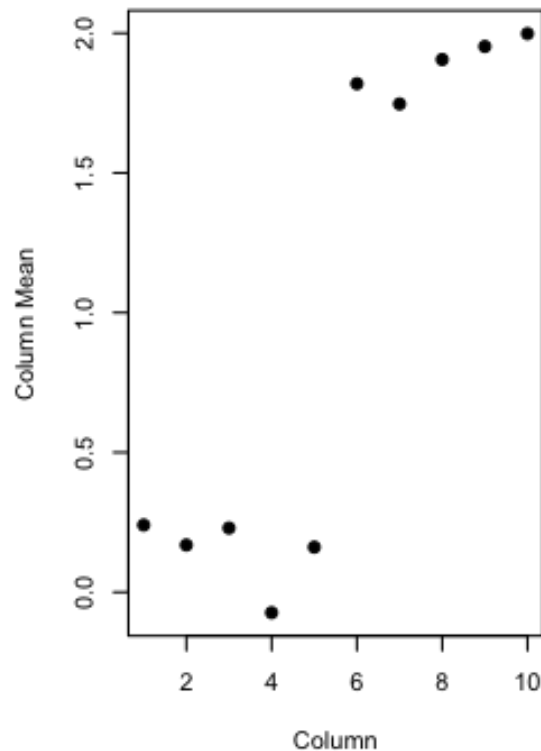
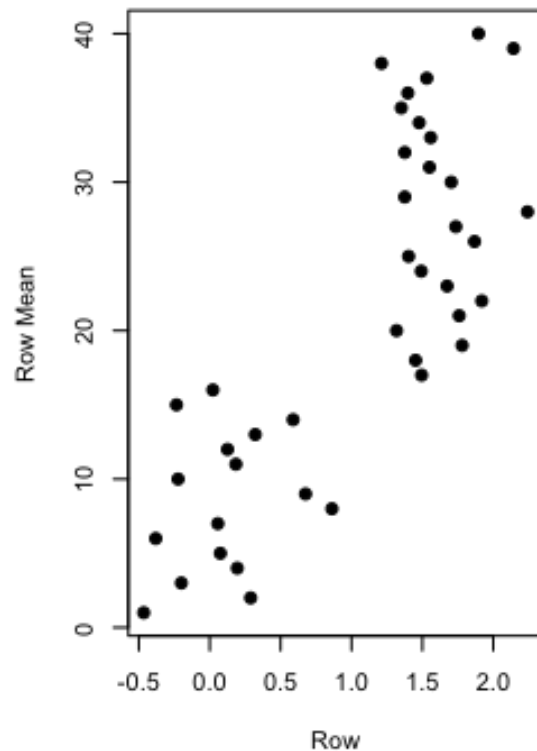
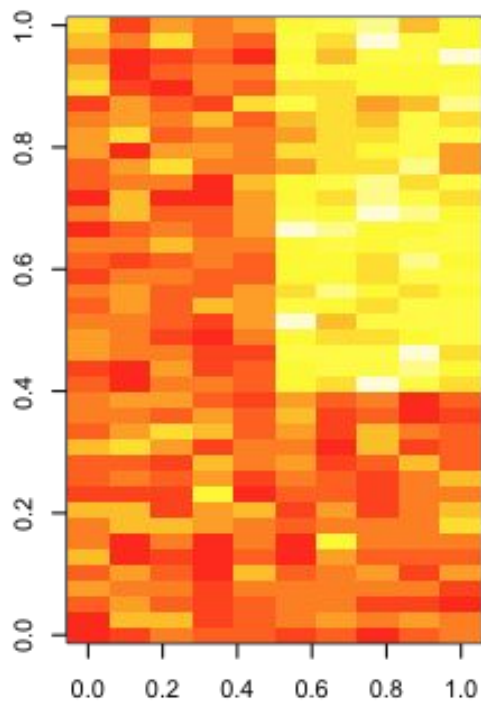
Columns of  $V^T$ /rows of  $U$  are orthogonal and calculated one at a time  
 Columns of  $V^T$  describe patterns across genes  
 Columns of  $U$  describe patterns across arrays

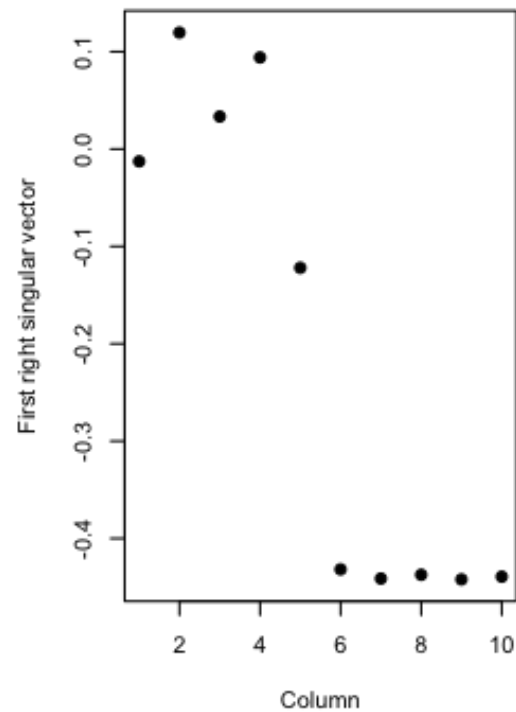
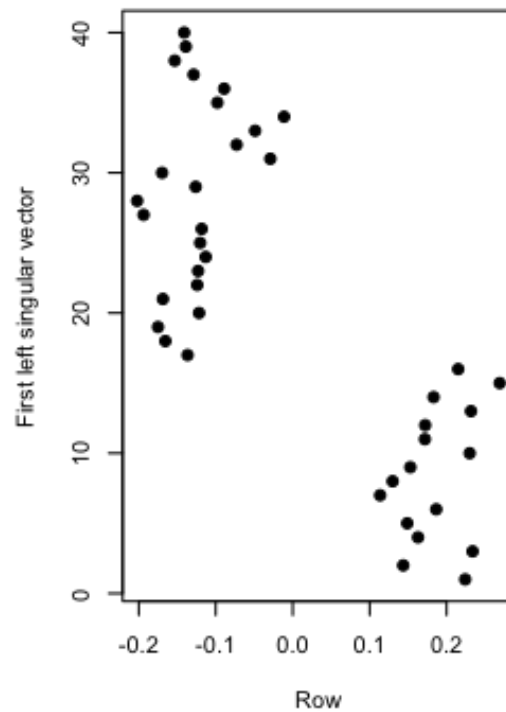
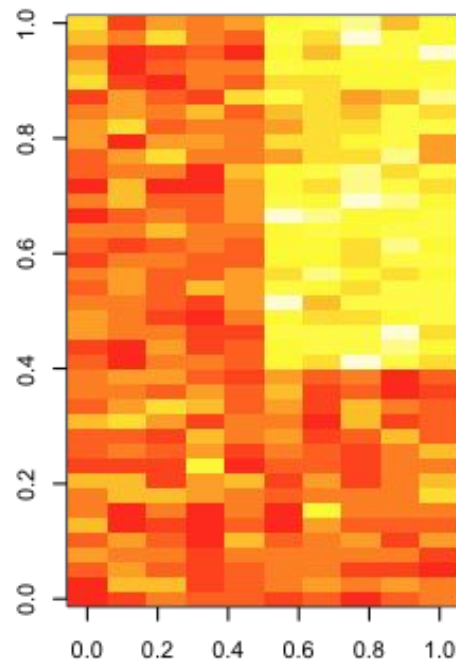
$d_i^2 / \sum_{i=1}^n d_i^2$  is the percent of variation explained by the  $i$ th column of  $V$

**Singular vectors/principal components**  
**Method to identify patterns in the data**









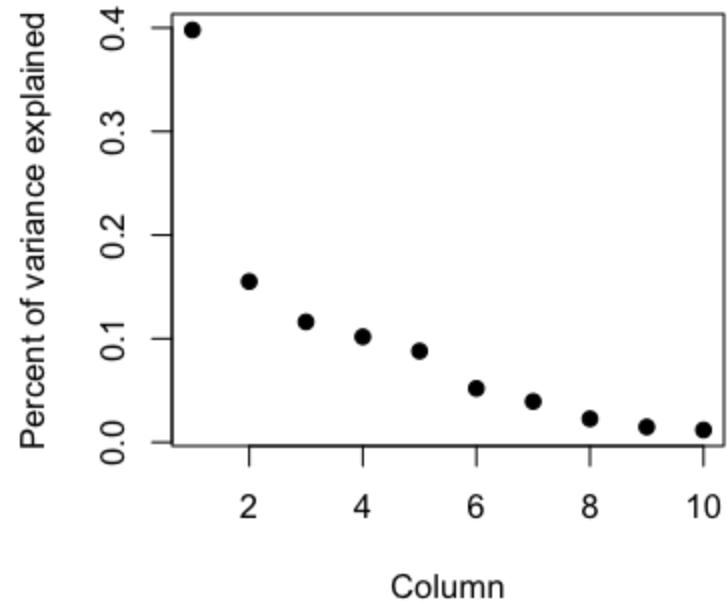
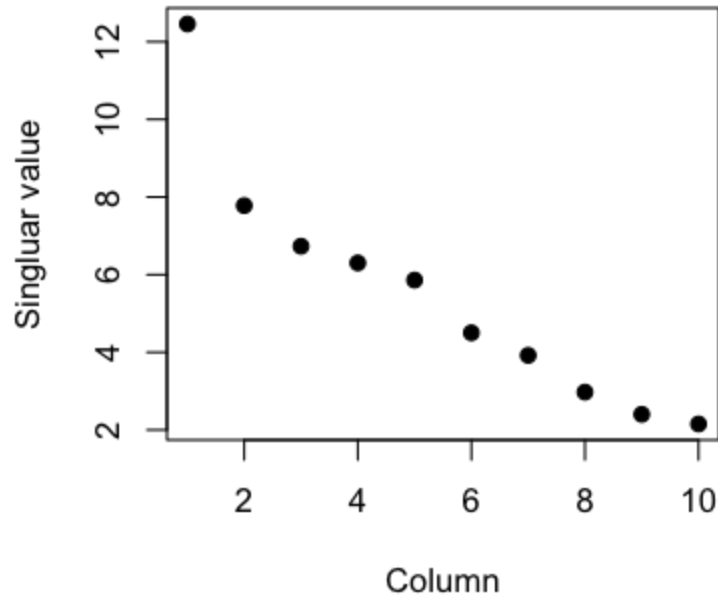
# Singular values

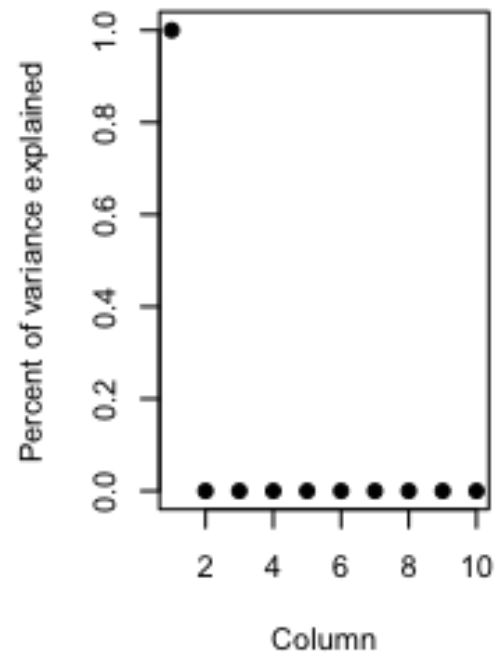
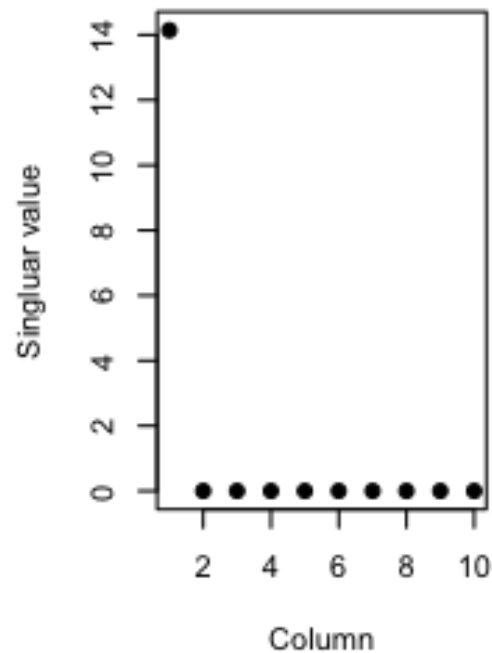
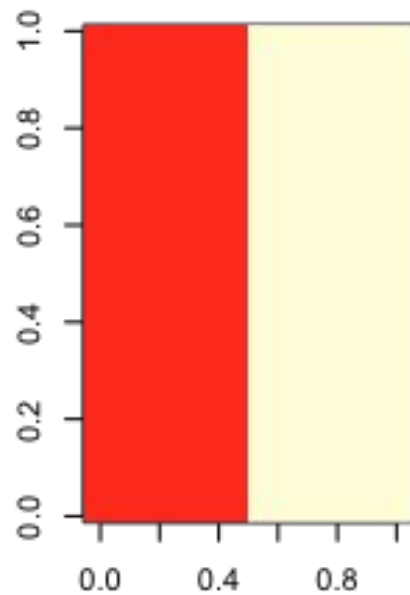
D is a diagonal matrix

$d_{ii}$  = ith singular value

$d_{ii}^2 / \sum d_{jj}^2$  = percent variance

explained by ith singular vectors

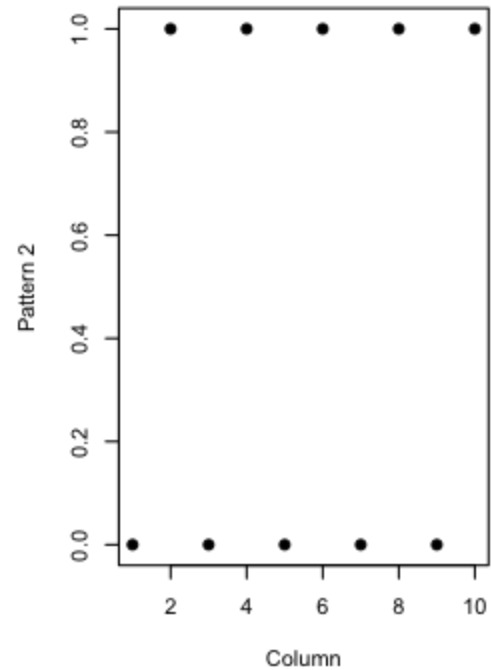
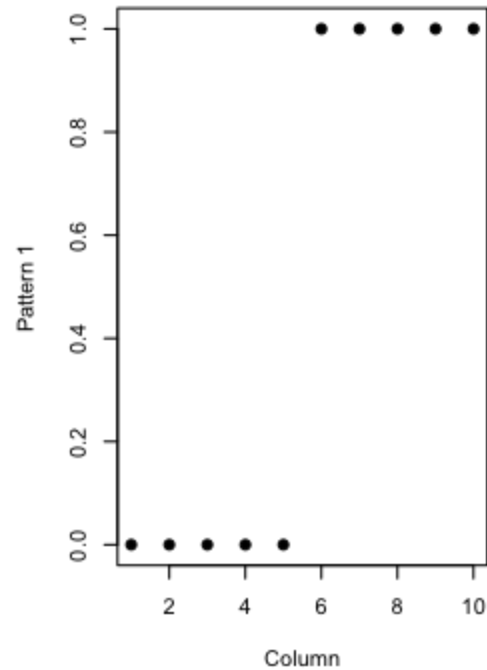
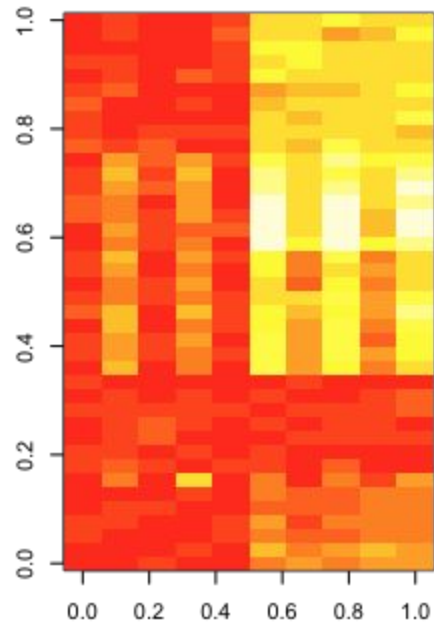




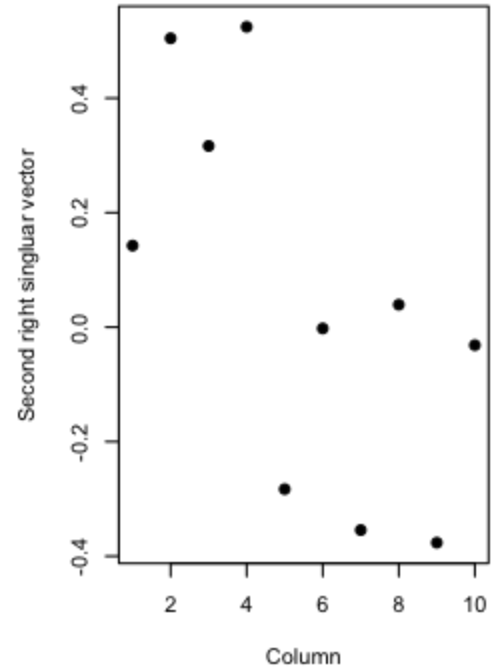
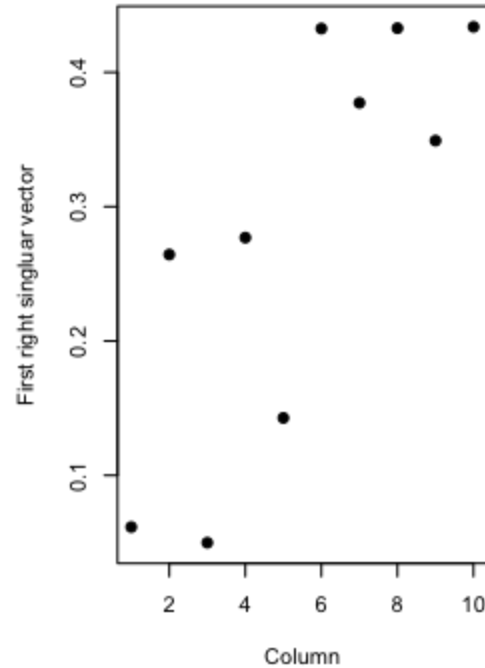
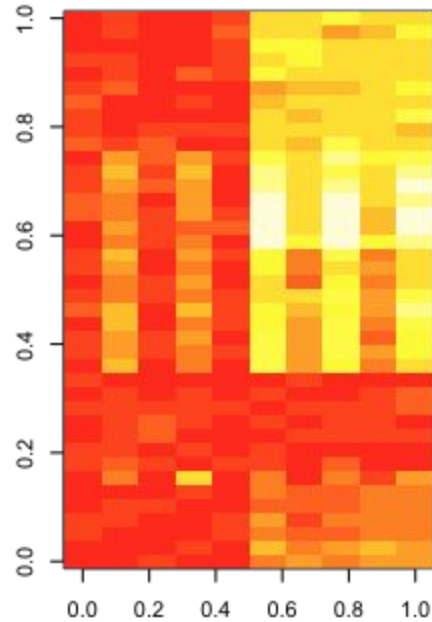
More than one pattern

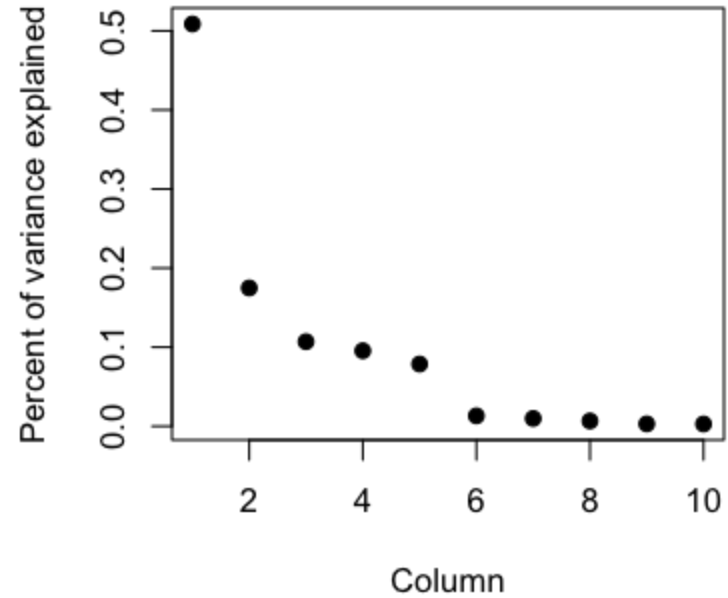
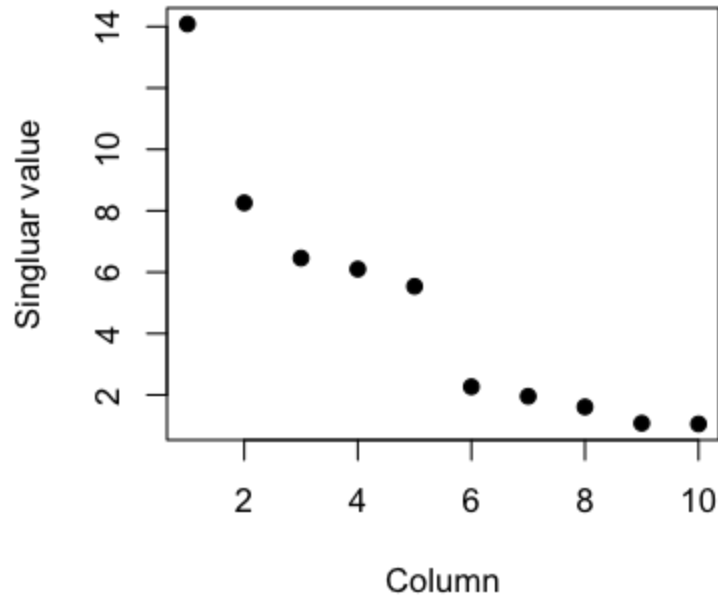
Patterns are orthogonal

One PC/SV may not equal one “variable”







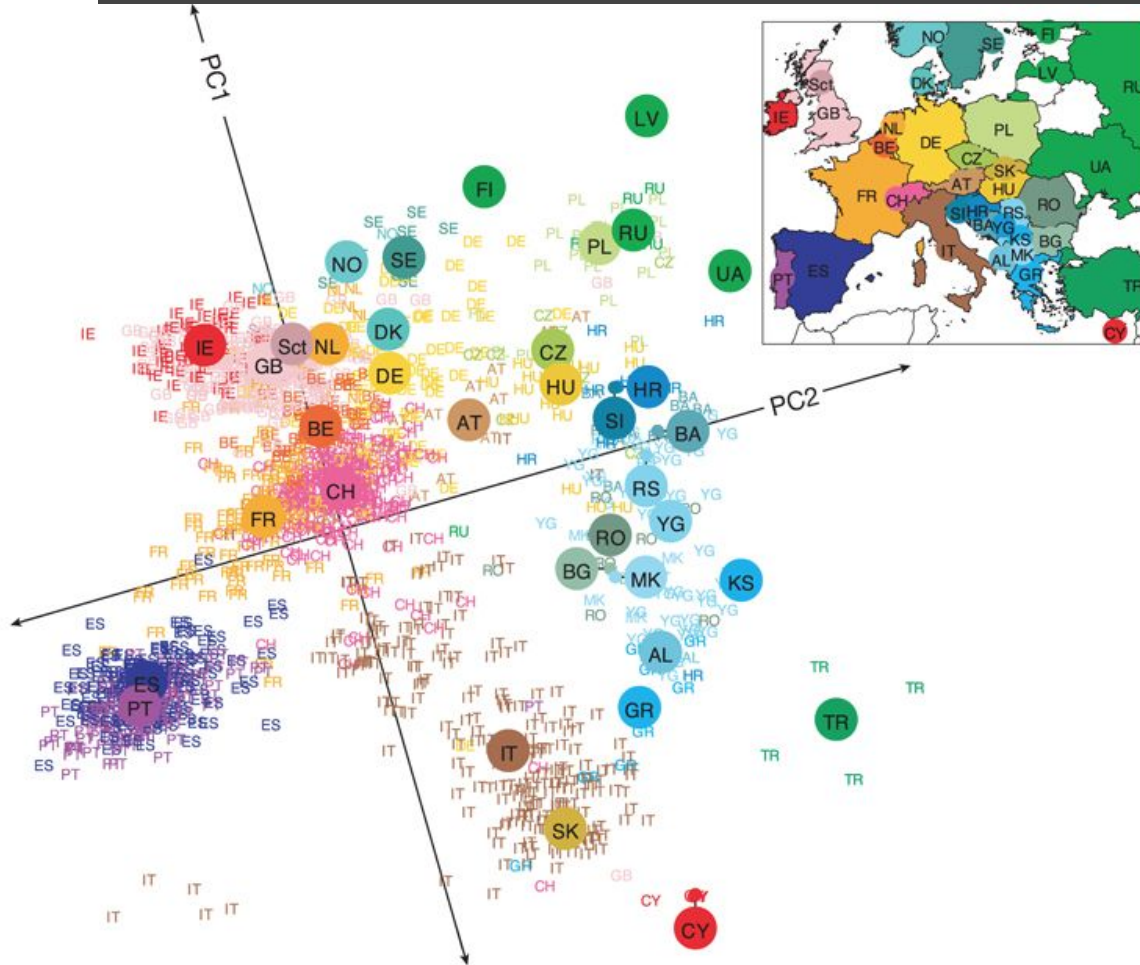


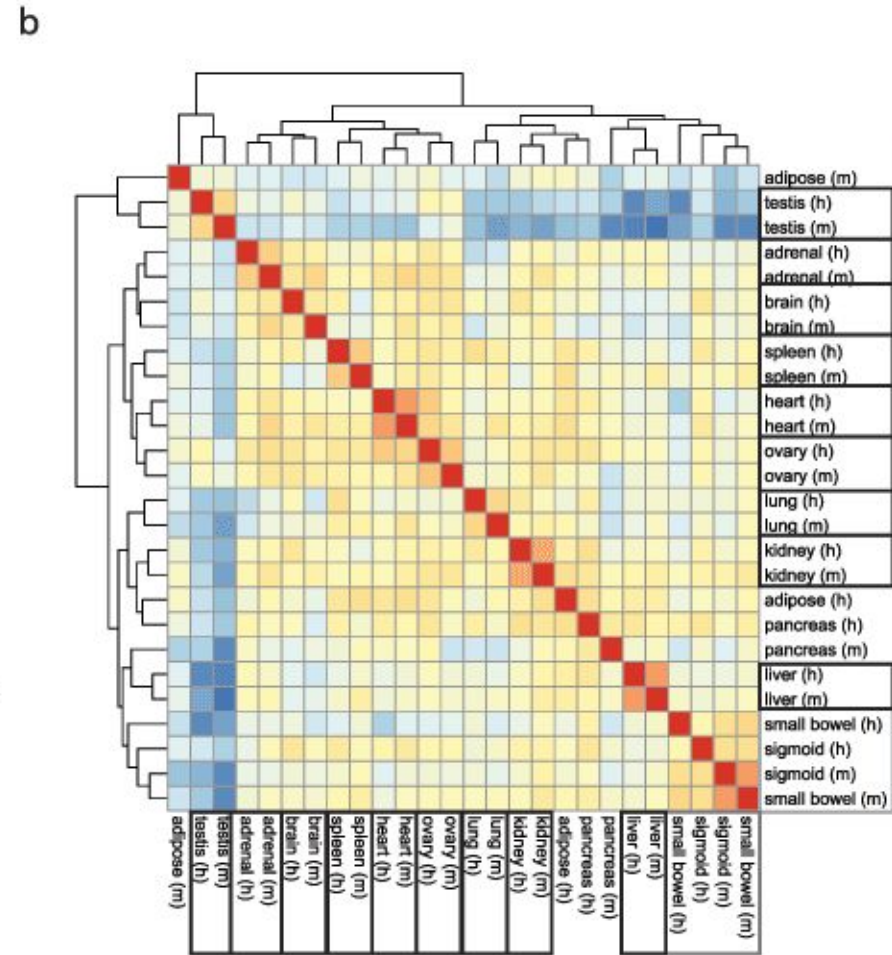
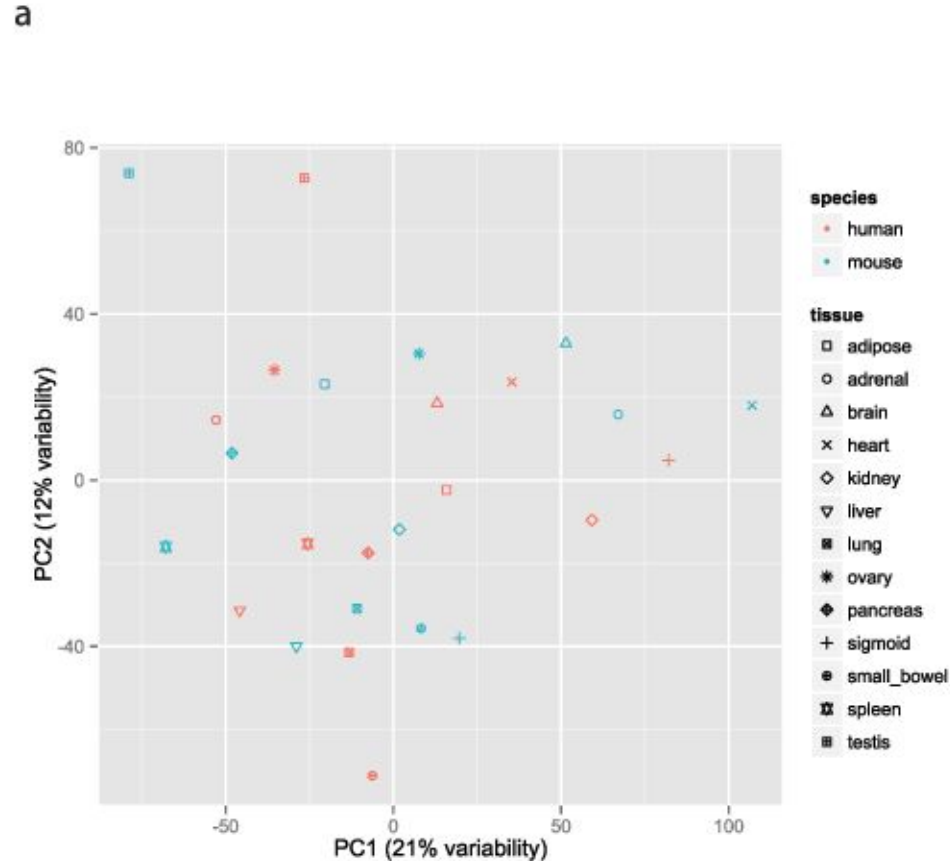
How this is used

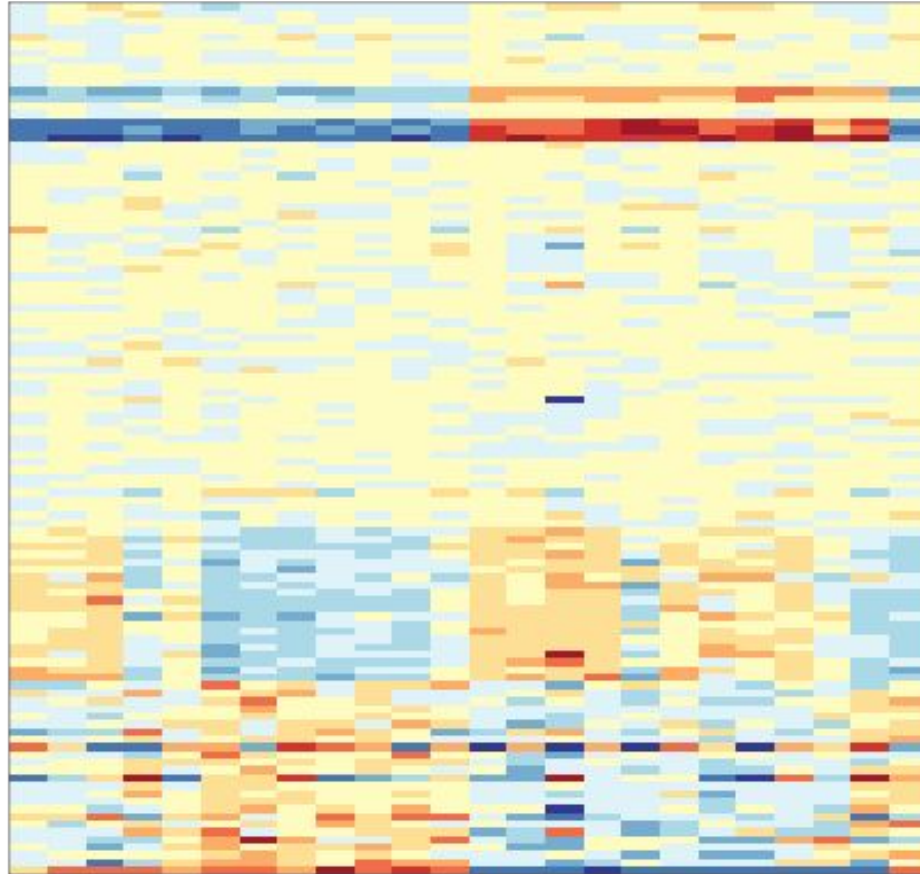
Identify meaningful patterns

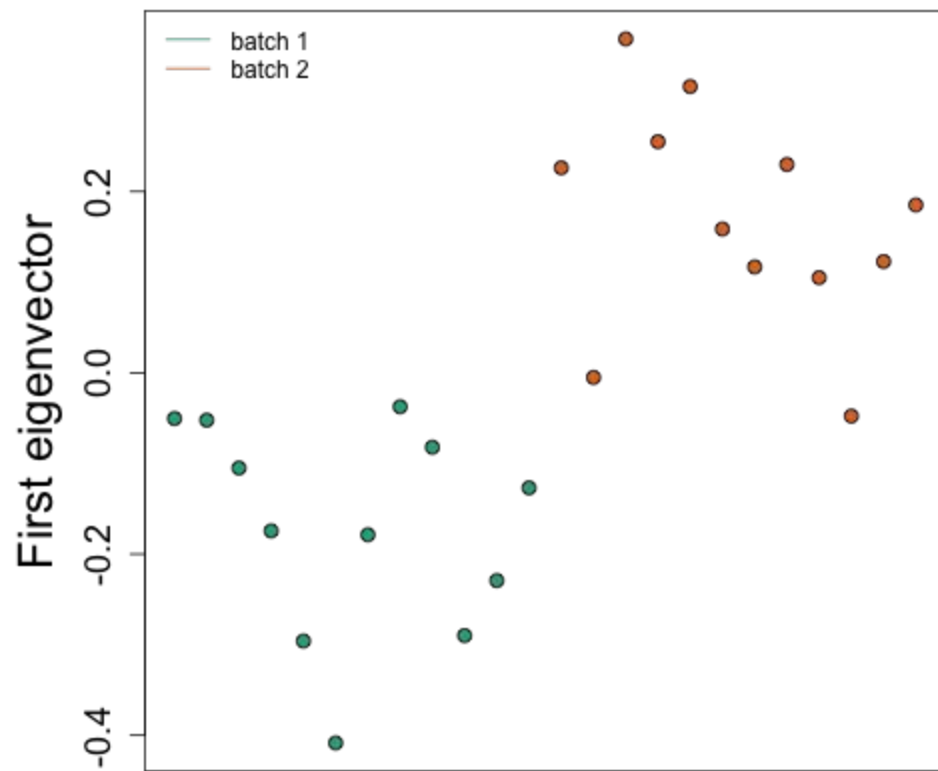
Find batch effects

**a**









# Notes and further resources

- Widely used for batch effects
  - <http://www.nature.com/nrg/journal/v11/n10/full/nrg2825.html>
- There are many more decompositions people use
  - multidimensional scaling, independent components analysis, non-negative matrix factorization
- More discussion in this course
  - <https://www.edx.org/course/advanced-statistics-life-sciences-harvardx-ph525-3x>