B4M36SAN
Dimensionality reduction

Anh Vu Le & Jan Blaha

Outline

- PCA
  - motivational example (*BreastCancer* dataset)
  - PCA principles with an artificial dataset
  - *BreastCancer* dataset revisited

- PCA vs LDA

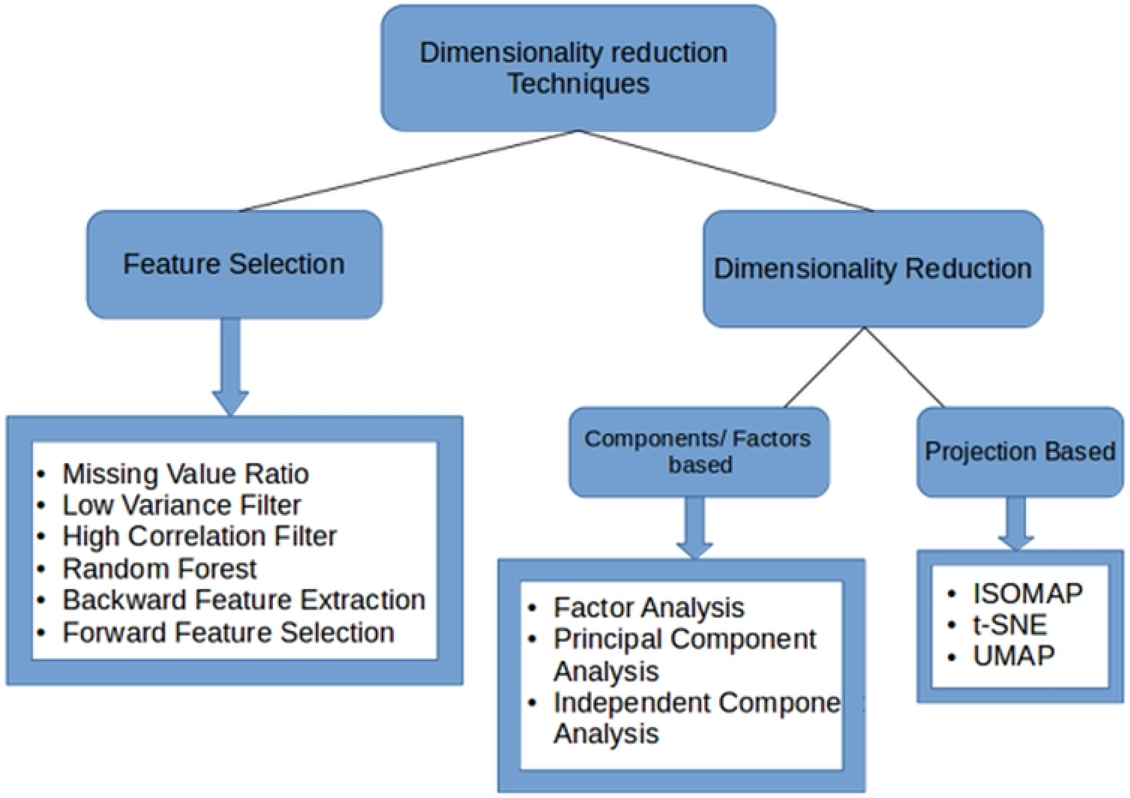
- tSNE
  - gentle introduction

Dimensionality reduction principles

- Benefits of reducing dimensionality
  - faster and often more accurate learning of classifiers
  - removing redundancy
  - visualization

Dimensionality reduction principles

- Benefits of reducing dimensionality
  - faster and often more accurate learning of classifiers
  - removing redundancy
  - visualization



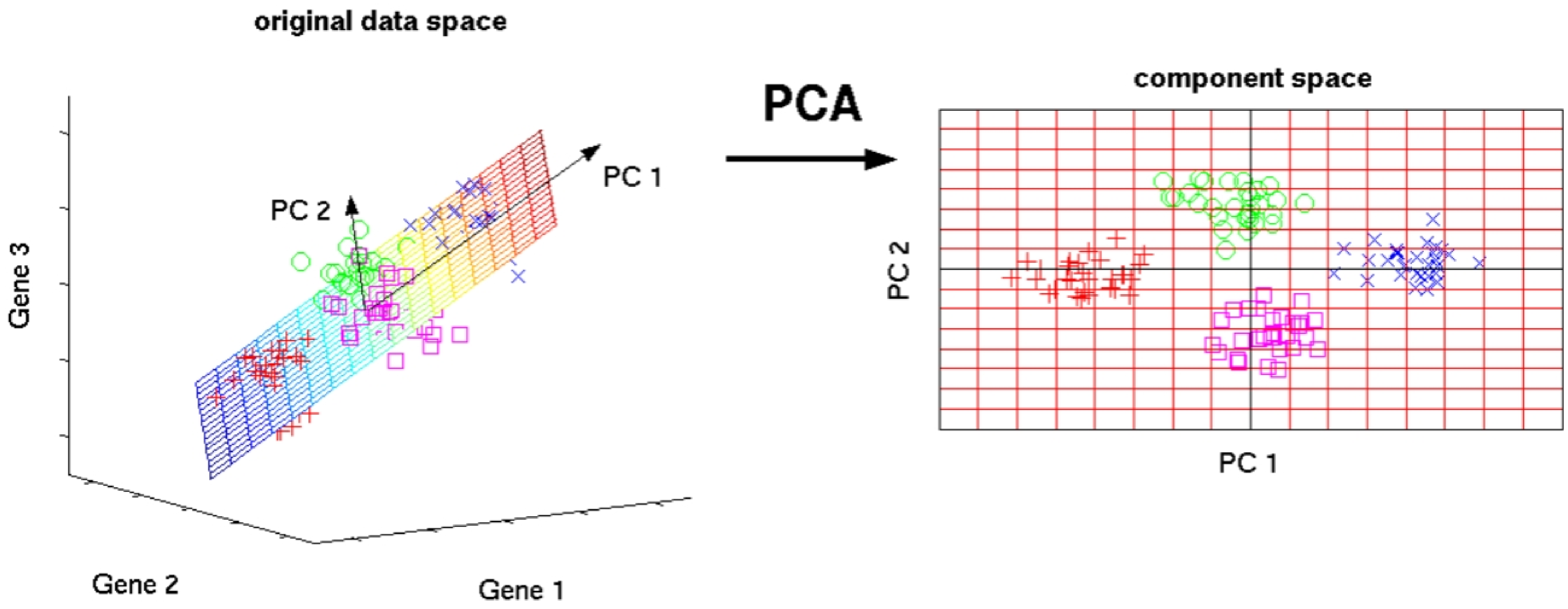
[Dimensionality Reduction Techniques | Python (analyticsvidhya.com)]

Dimensionality reduction principles

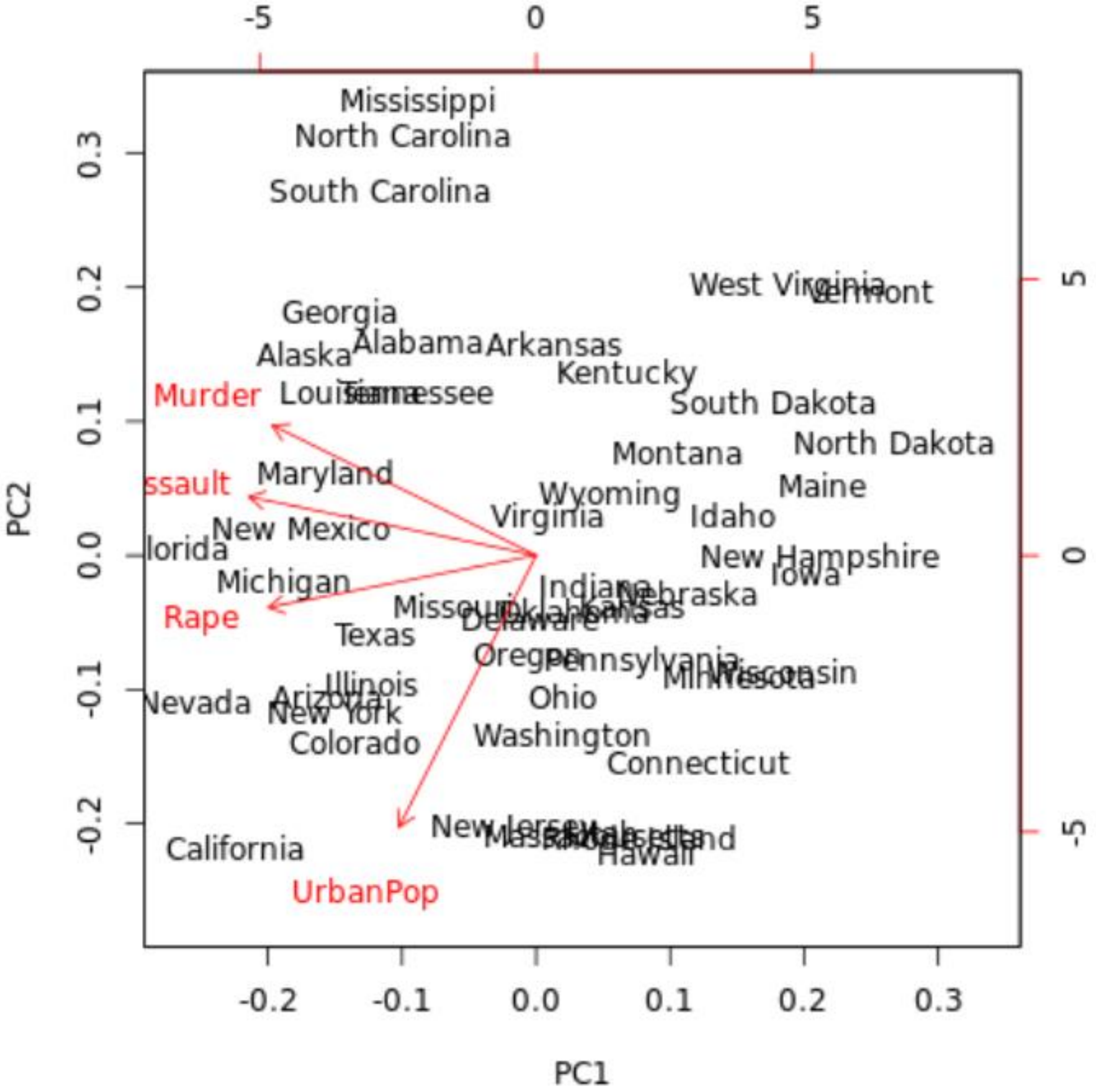
- Benefits of reducing dimensionality
  - faster and often more accurate learning of classifiers
  - removing redundancy
  - visualization

- Cost of reducing dimensionality
  - information loss
  - new axes may be difficult to interpret

Math score	English score	Age
3.5	3.7	17
4.0	3.2	18
2.3	2.1	18
2.0	3.9	17
1.0	2.9	18



Ph.D. thesis - Matthias Scholz - Max Planck Institute of Molecular Plant Physiology (matthias-scholz.de)



-5
0
5
Mississippi
North Carolina
South Carolina
West Virginia
Vermont
Georgia
Alabama
Arkansas
Alaska
Kentucky
Murder
Tennessee
South Dakota
North Dakota
Montana
Maryland
ssault
Wyoming
Maine
Virginia
Idaho
New Mexico
lorida
New Hampshire
Iowa
Michigan
Indiana
Nebraska
Rape
Texas
Oregon
Pennsylvania
Wisconsin
Illinois
Ohio
Nevada
New York
Colorado
Washington
Connecticut
New Jersey
Hawaii
California
UrbanPop
0.3
0.2
0.1
0.0
-0.1
-0.2
PC2
-0.2
-0.1
0.0
0.1
0.2
0.3
PC1

PCA

LDA

Finds axes/directions of:

max scatter
of the **entire data set**

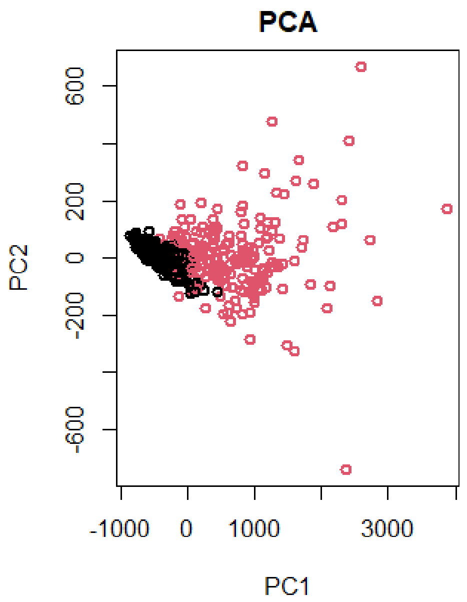
max scatter **between**
AND
min scatter **within** classes

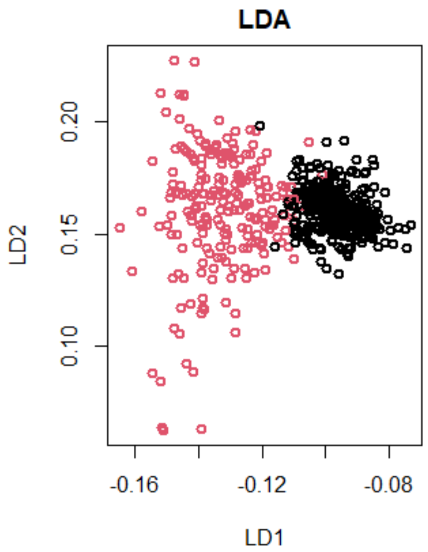
Eigenproblem leading to the new axes

$Cov(X)$

$S_W^{-1} S_B$

Projection (*BreastCancer*)

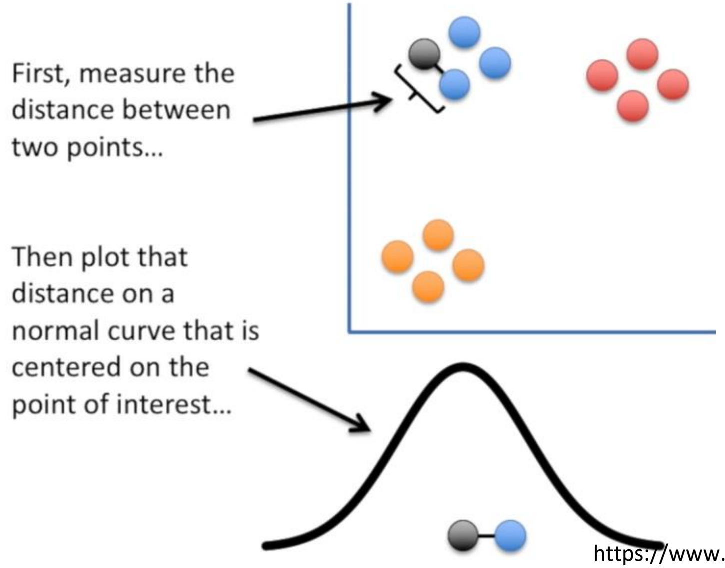




t-SNE

- PCA focuses on data as whole
  - “Makro” method
  - cannot capture finer details of the topology of the data

- Project into a lower dimension while preserving neighborhood relationships
  - t-SNE, ISOMAP



https://www.youtube.com/watch?v=NEaUSP4YerM