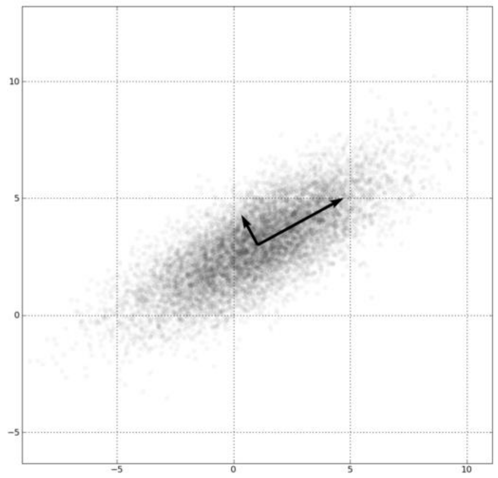
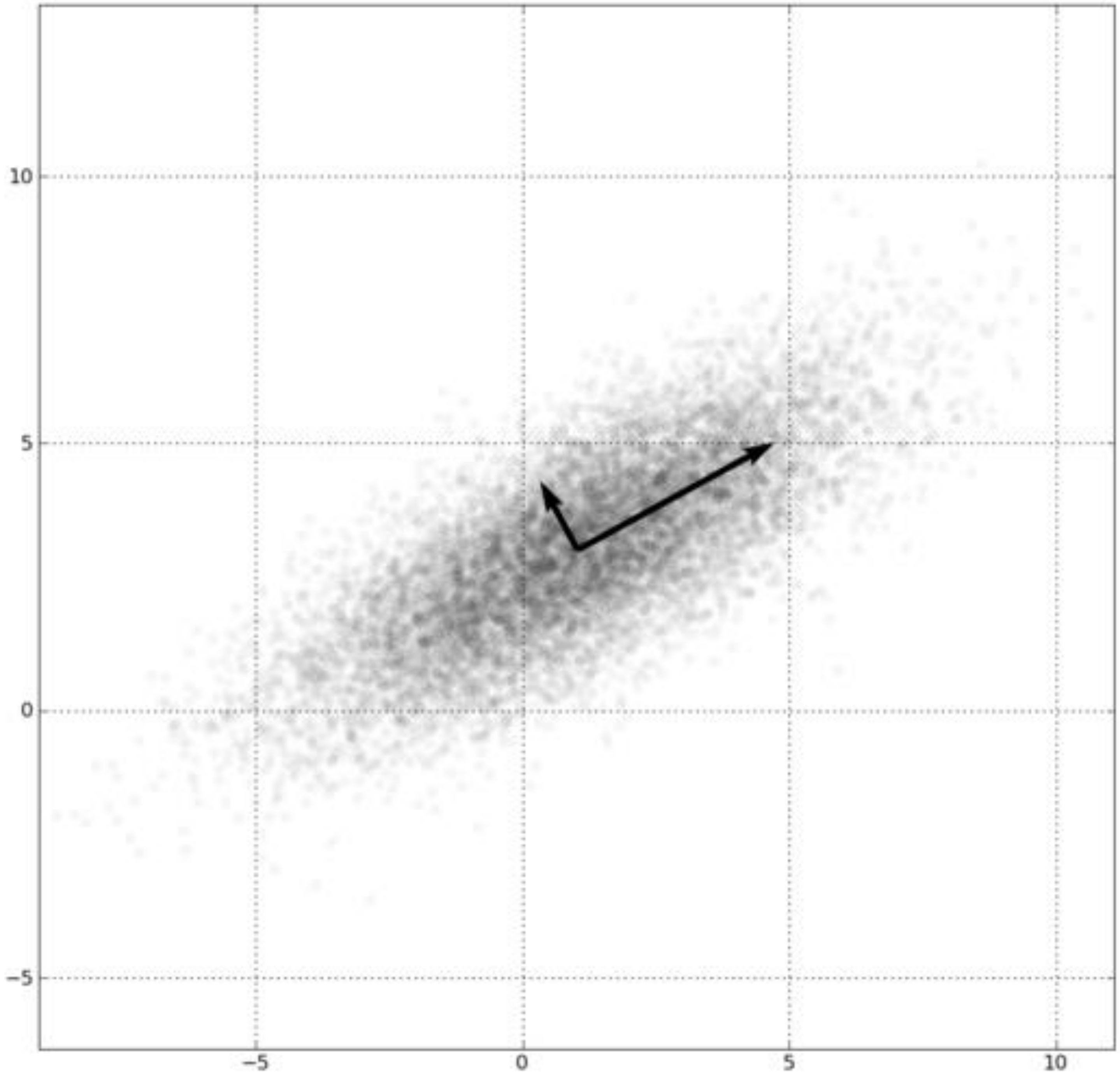
Principal Component Analysis

by Ahmet Sacan

PCA: Change of basis

- Is there another basis, which is a <u>linear</u> combination of the original basis, that <u>best</u> re-expresses our data set?
- Linearity simplifies the problem by restricting the set of potential bases





10
5
0
−5
−5
0
5
10

Linear Transformation

- Let X and Y be m×n matrices related by a linear transformation P
  - Y = XP
  - X: original dataset
  - Y: transformed/re-represented dataset
  - P is a n×n matrix
- Geometrically, P is a rotation and a stretch which transforms X into Y.
- The columns of P, {p1, . . . , pn}, are a set of new basis vectors for expressing the rows of X.
  - These basis vectors will become the "principal components of X"

“Best” Transformation

- To reduce redundancy:
  - each variable should co-vary as little as possible with other variables.
- i.e., we would like the co-variances between separate measurements/components to be zero.

Covariance/Scatter Matrix

- Assume X is shifted to its mean

$$S_X = \frac{1}{m-1} X^T X$$

$$S_Y = \frac{1}{m-1} Y^T Y$$

- What property should S_Y have? (specifically, consider the off-diagonals)
- Goal: Find an orthogonal matrix P where Y=XP, such that S_Y is diagonalized.

Best P

- The coefficient 1/(m-1) is removed for convenience

$$S_Y = Y^T Y = (XP)^T (XP) = P^T (X^T X) P = P^T S_X P$$

- Every real symmetric matrix A (e.g., S_X,S_Y) can be diagonalized, moreover, the eigen decomposition is:

$$A = EDE^T$$

  – Where E is an orthogonal matrix (eigenvectors of A) and D is real and diagonal (eigenvalues of A)

Best P

$$S_Y = P^T S_X P = P^T EDE^T P$$

- If we select P=E,

$$S_Y = E^T EDE^T E = D$$

- So, the <u>best transformation P is simply the eigenvectors of S_X</u>
- The eigenvalues give the amount of variation along each eigenvector

PCA solution: summary

- Given n dimensional matrix X with m rows
- Mean-shift X
- Calculate scatter matrix:

$$S_x = \frac{X^T X}{m-1}$$

- Find the eigen decomposition of S_x:

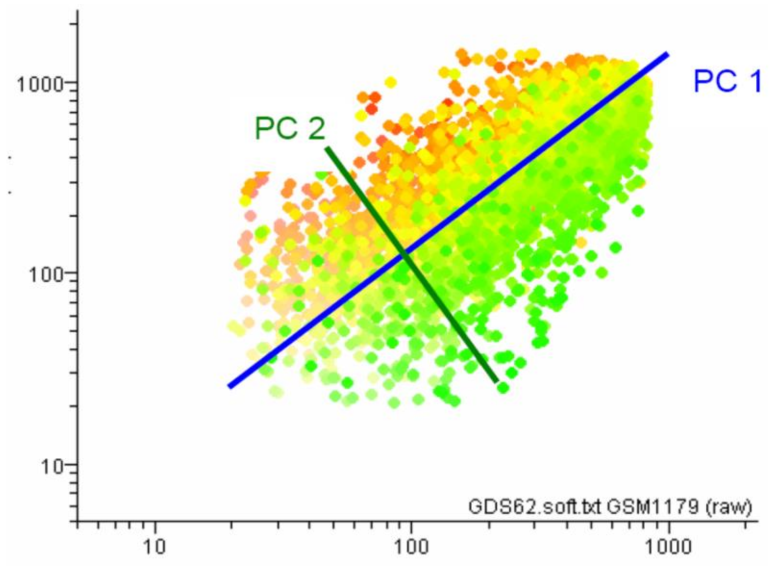
$$S_x = PDP^T$$

- Use P as the transformation matrix to produce Y:

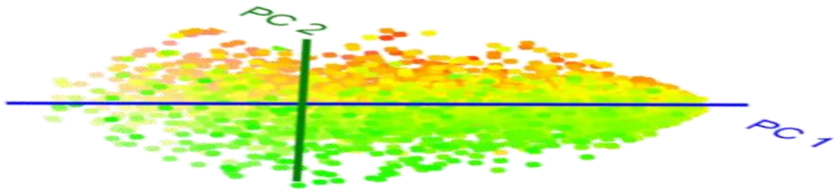
$$Y = XP$$

Another solution using SVD

- $X=U\Sigma V^*$
  - Σ is a diagonal matrix of "singular values". (which are the square roots of the eigenvalues of X^TX)
  - Columns of U are orthonormal output basis vector directions of X (eigenvectors of X^TX)
  - Columns of V are orthonormal input basis vector directions of X
- SVD of data matrix X vs. eigenvalue decomposition of covariance matrix X^TX



1000
100
10
PC 1
PC 2
GDS62.soft.txt GSM1179 (raw)
10
100
1000



PC 2
PC 1

Reconstruction from Y to X

- Remember: $Y = XP$
- Multiply both sides with P^{-1}:

$$YP^{-1} = XPP^{-1}$$
$$X = YP^{-1}$$

- Because P is an orthogonal matrix, $P^T = P^{-1}$

$$X = YP^T$$

Reconstruction from [[partial Y]] to X

- Let X have n columns
  - Y would have n columns too.
  - P would be n-by-n matrix
- Let $Y_{partial}$ be the first k columns of Y
- Let $P_{partial}$ be the first k columns of P
- We can reconstruct an approximation of X:

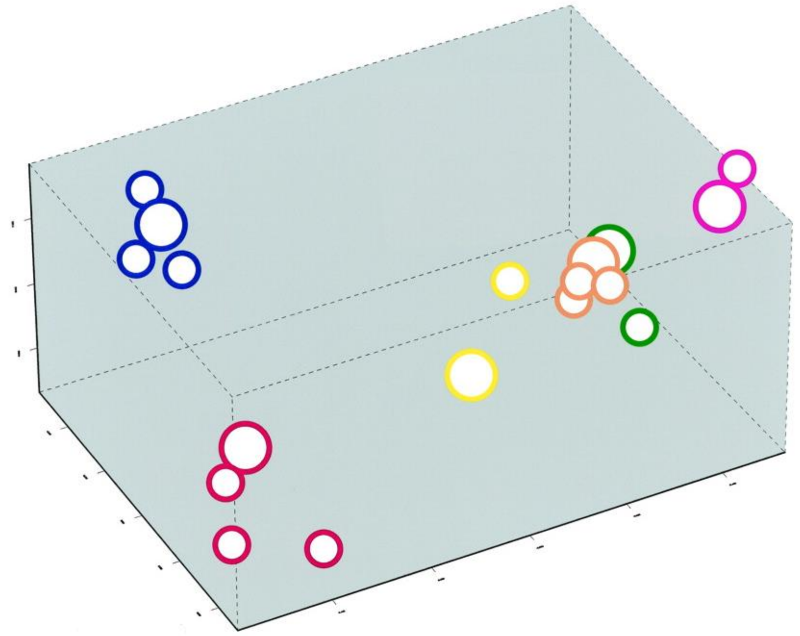
$$X_{approx} = Y_{partial} P_{partial}^T$$

- Note that X_{approx} has n columns, **not k**.
- One can compare X_{approx} and X (e.g., using mean-squared-difference of all of their values) to find out the "reconstruction error"

Multidimensional Scaling (MDS)

- Assersohn, 2002
  - Samples from fine needles aspirates (FNA) and from tumors in breast cancer.
  - Color: patient, Large circle: tumor, Small circle: FNA

B



A