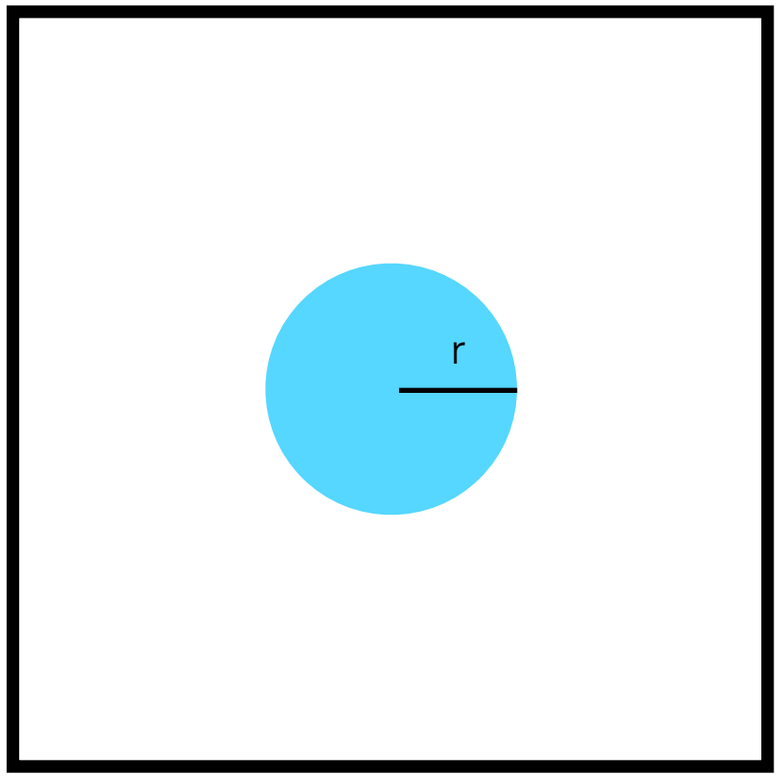
DATA MINING & MACHINE LEARNING:

PRINCIPAL COMPONENT ANALYSIS

THE CURSE OF DIMENSIONALITY

- Consider a uniformly-populated sample in 2 parameters (dimensions). The fraction of the sample found within a radius r of the data's center is:

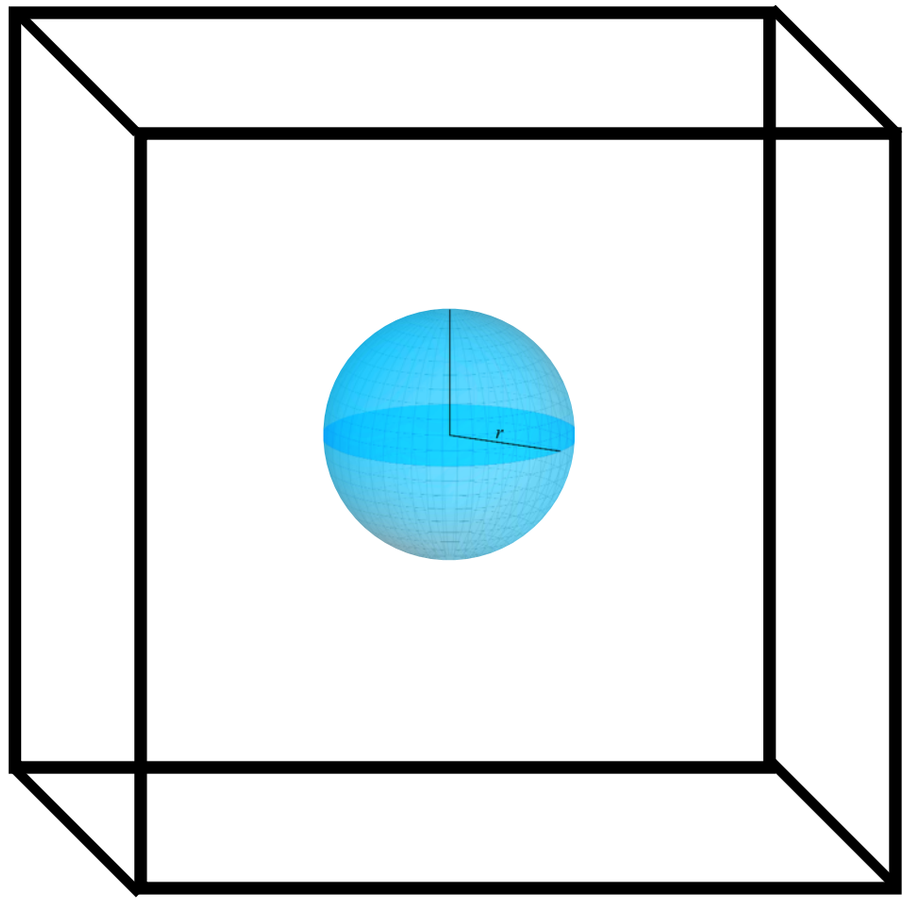
$$f_2 = \frac{V_{param}}{V_{total}} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4} \approx 78.5\%$$



THE CURSE OF DIMENSIONALITY

- As a model becomes more complex (of higher dimensionality), more data is required to constrain it.

$$f_3 = \frac{V_{param}}{V_{total}} = \frac{(4/3)\pi r^3}{(2r)^3} = \frac{\pi}{6} \approx 52.3\%$$



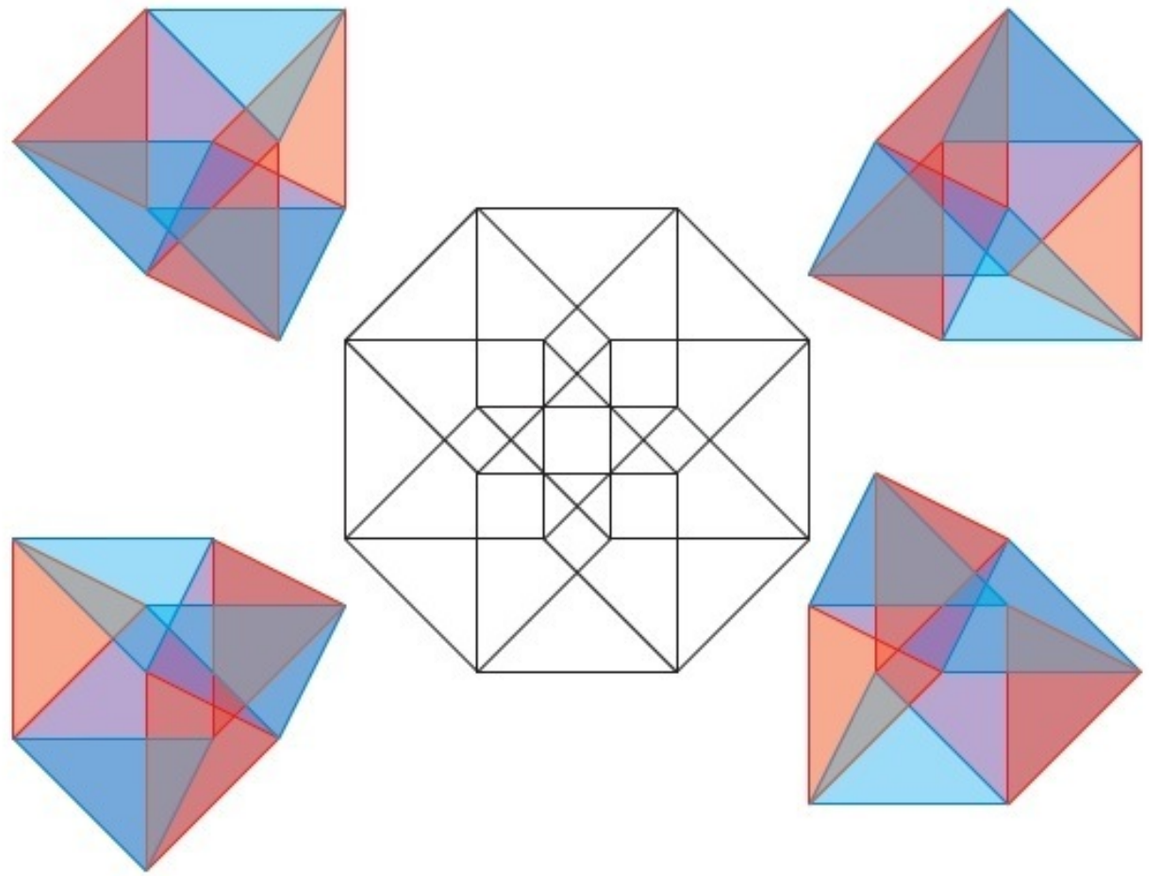
THE CURSE OF DIMENSIONALITY

- for the case of D dimensions,

$$V_D(r) = \frac{2r^D \pi^{D/2}}{D\Gamma(D/2)}$$

$$f_D = \frac{V_D(r)}{(2r)^D} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}$$

$$\lim_{D\to\infty} f_D = 0$$



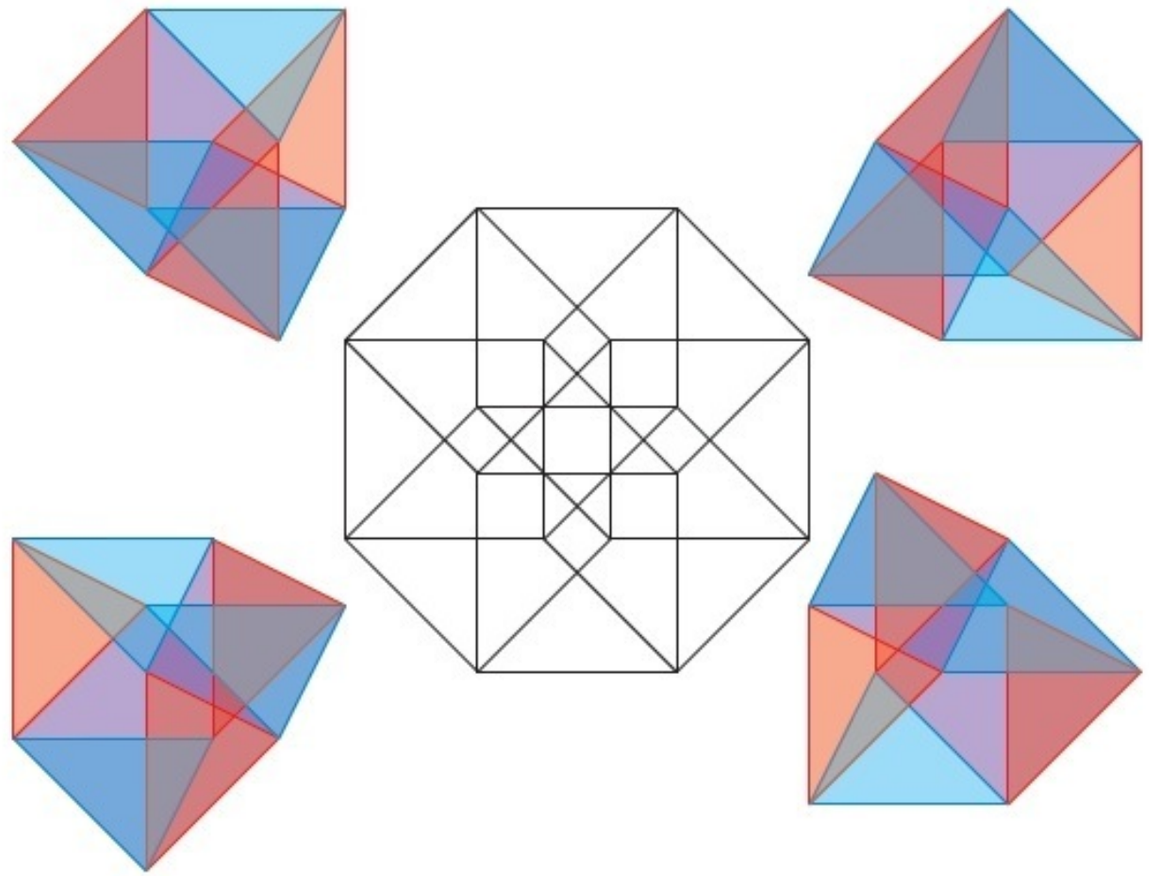
THE CURSE OF DIMENSIONALITY

- for the case of D dimensions,

$$V_D(r) = \frac{2r^D \pi^{D/2}}{D\Gamma(D/2)}$$

$$f_D = \frac{V_D(r)}{(2r)^D} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}$$

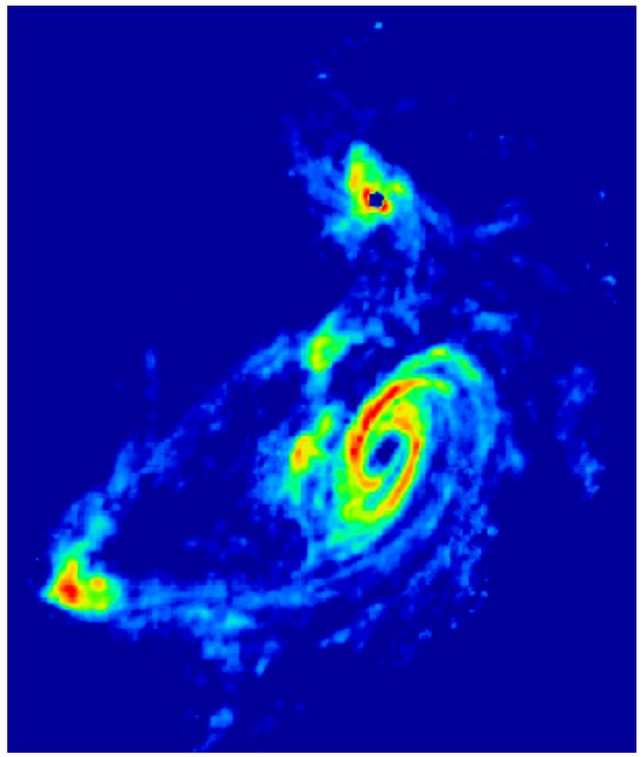
$$\lim_{D\to\infty} f_D = 0$$

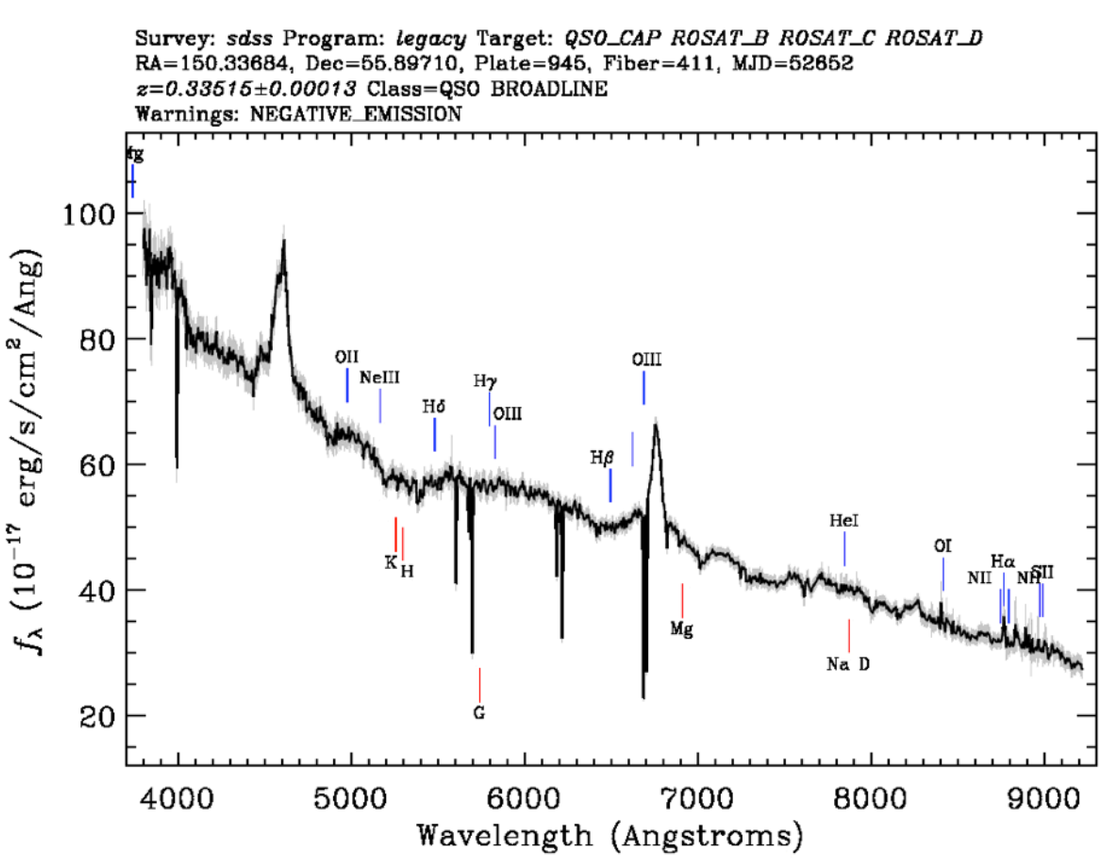


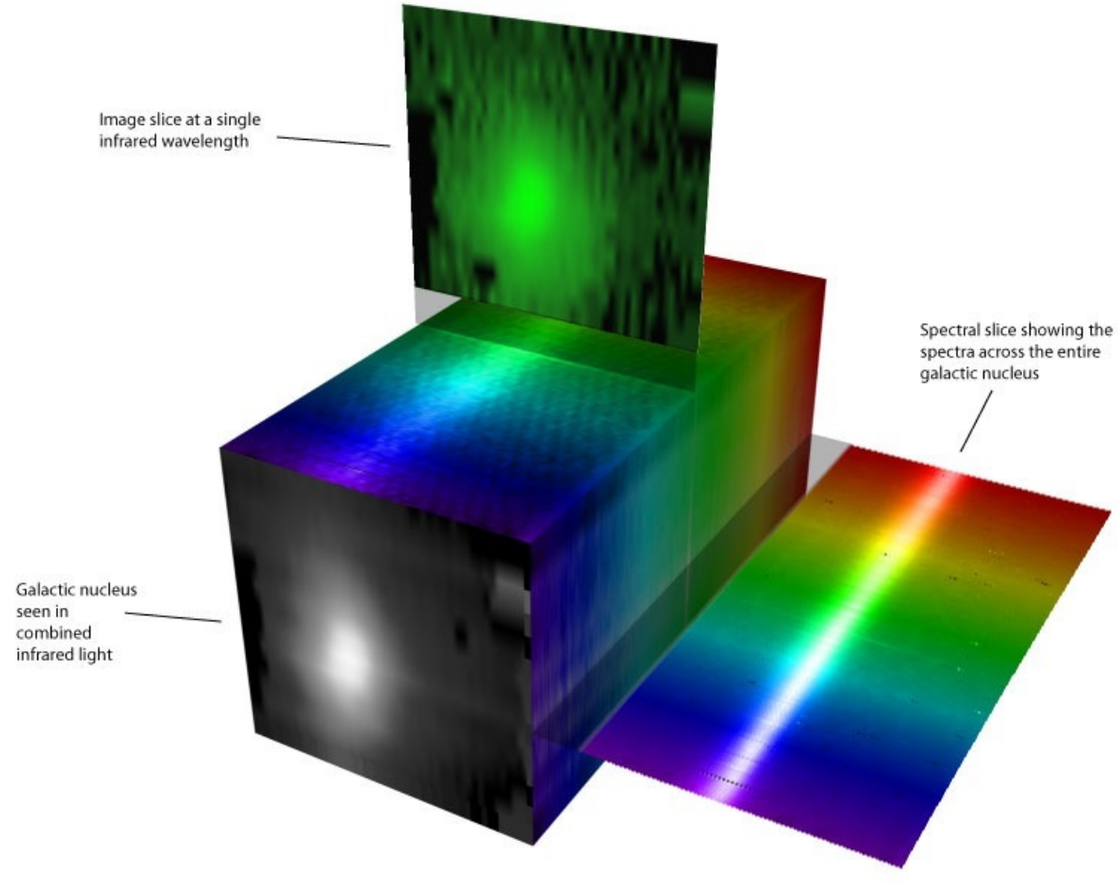
- **The number of data points required to evenly sample a hypervolume grows exponentially with dimension.**

THE CURSE OF DIMENSIONALITY

- more familiar “hyperspheres”







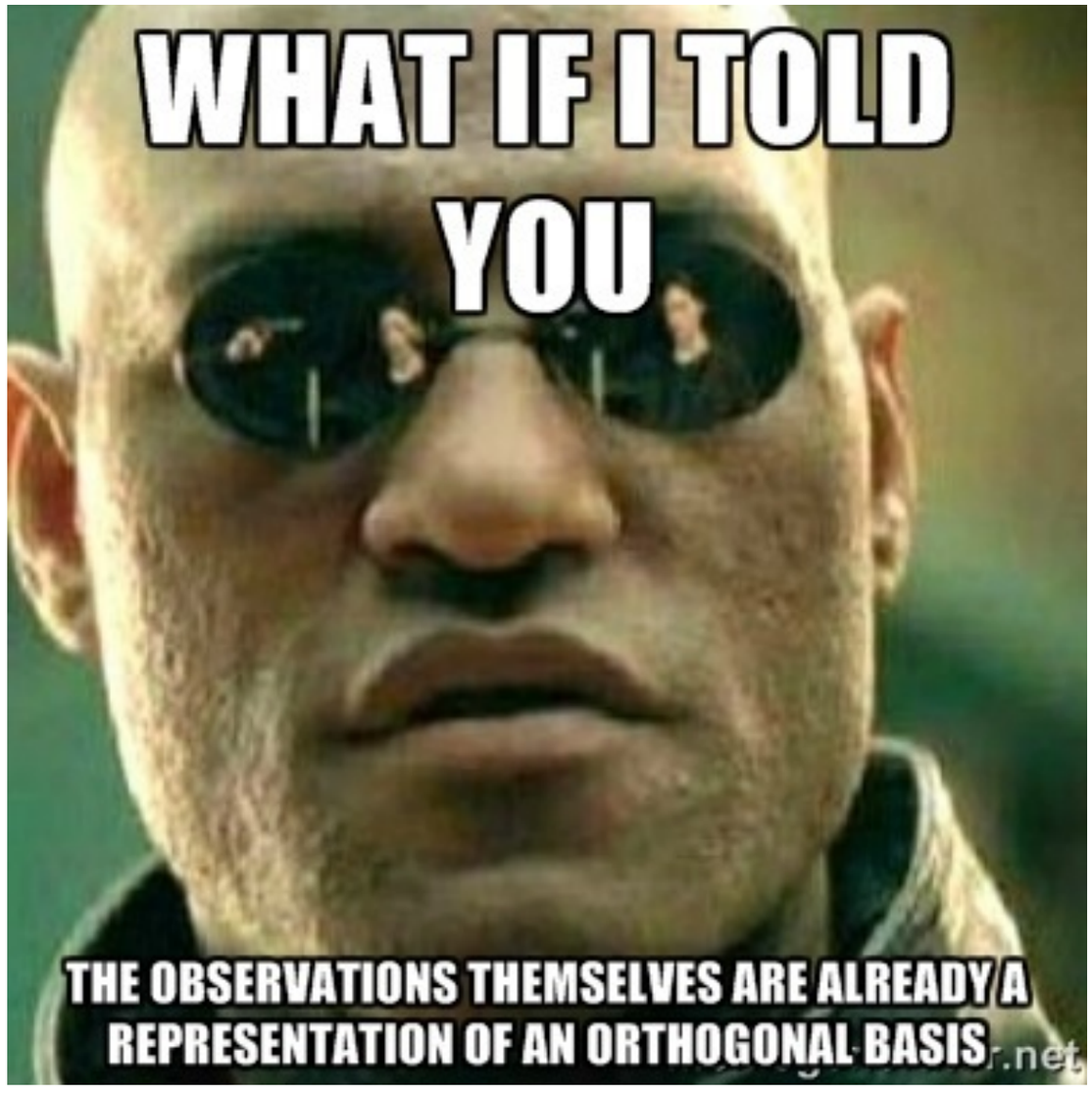
INTRINSIC DIMENSIONALITY

- certain projections within the data capture the principal physical and statistical correlations between measured quantities

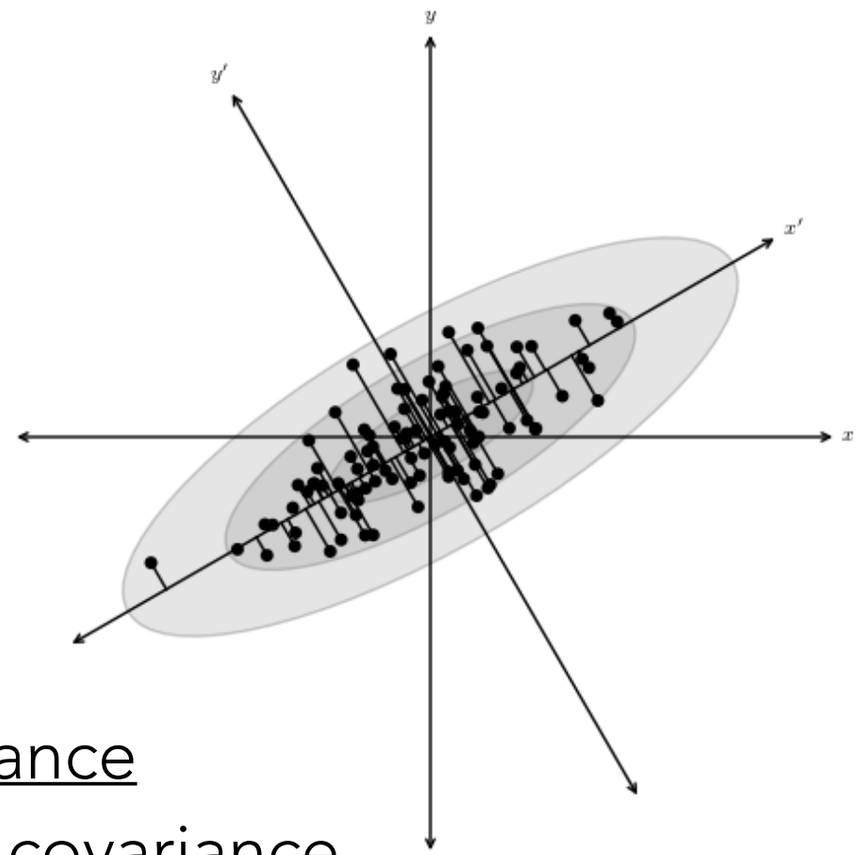
- if we can find these efficiently, we can use them to:

    - reduce the dimensionality of data

    - more simply visualize, classify data



INTRINSIC DIMENSIONALITY



- Principal Component Analysis (PCA)
    - identifies the axes with maximal variance (principal components) and minimal covariance
    - effectively a least squared minimization of the points with respect to the principal components

COVARIANCE MATRICES

- Variance

$$Var(x) = \frac{\sum \left(x_i - \bar{x} \right)^2}{N}$$

- Covariance

$$COV(x,y) = \frac{\sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)}{N}$$

- Covariance Matrix

$$C_X = \frac{1}{N-1} X^T X$$

COVARIANCE MATRICES

- Calculating a Covariance Matrix:

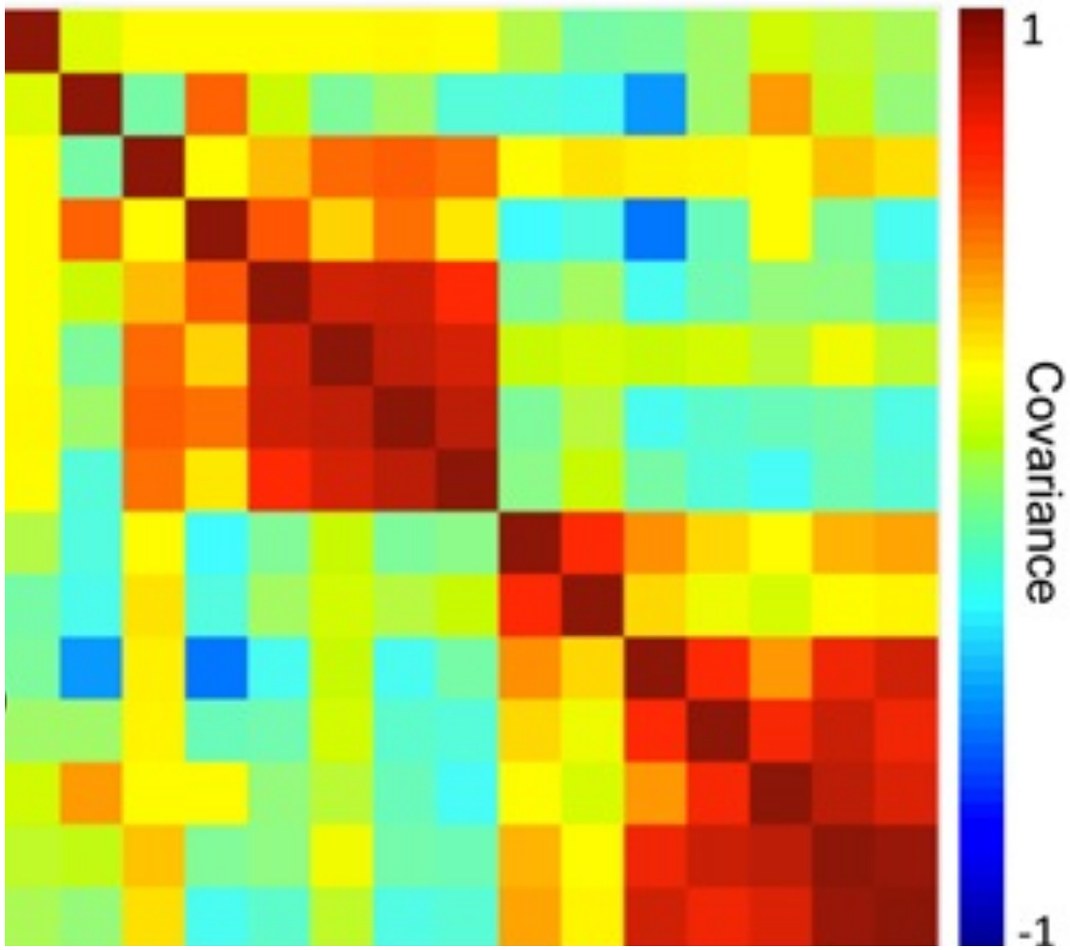
  - Assume data $\{x_i\}$ with N observations of K parameters
  - Center data by subtracting the mean of each feature in $\{x_i\}$ and write this N x K matrix as X.
  - The centered data is given by the covariance matrix:

$$C_X = \frac{1}{N-1} X^T X$$

COVARIANCE MATRICES

- Covariance Matrix

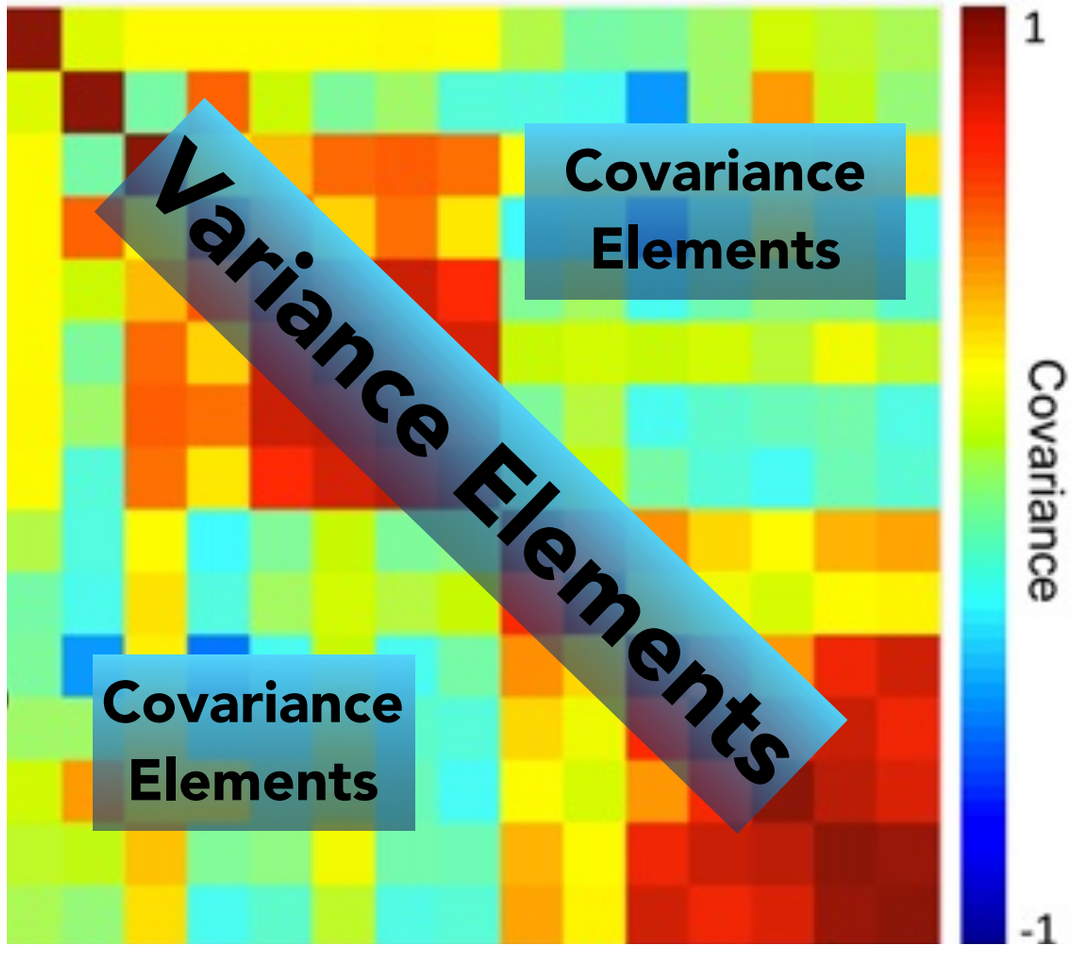
$$COV(x,y) = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{N}$$



$$C_X = \frac{1}{N-1} X^T X$$

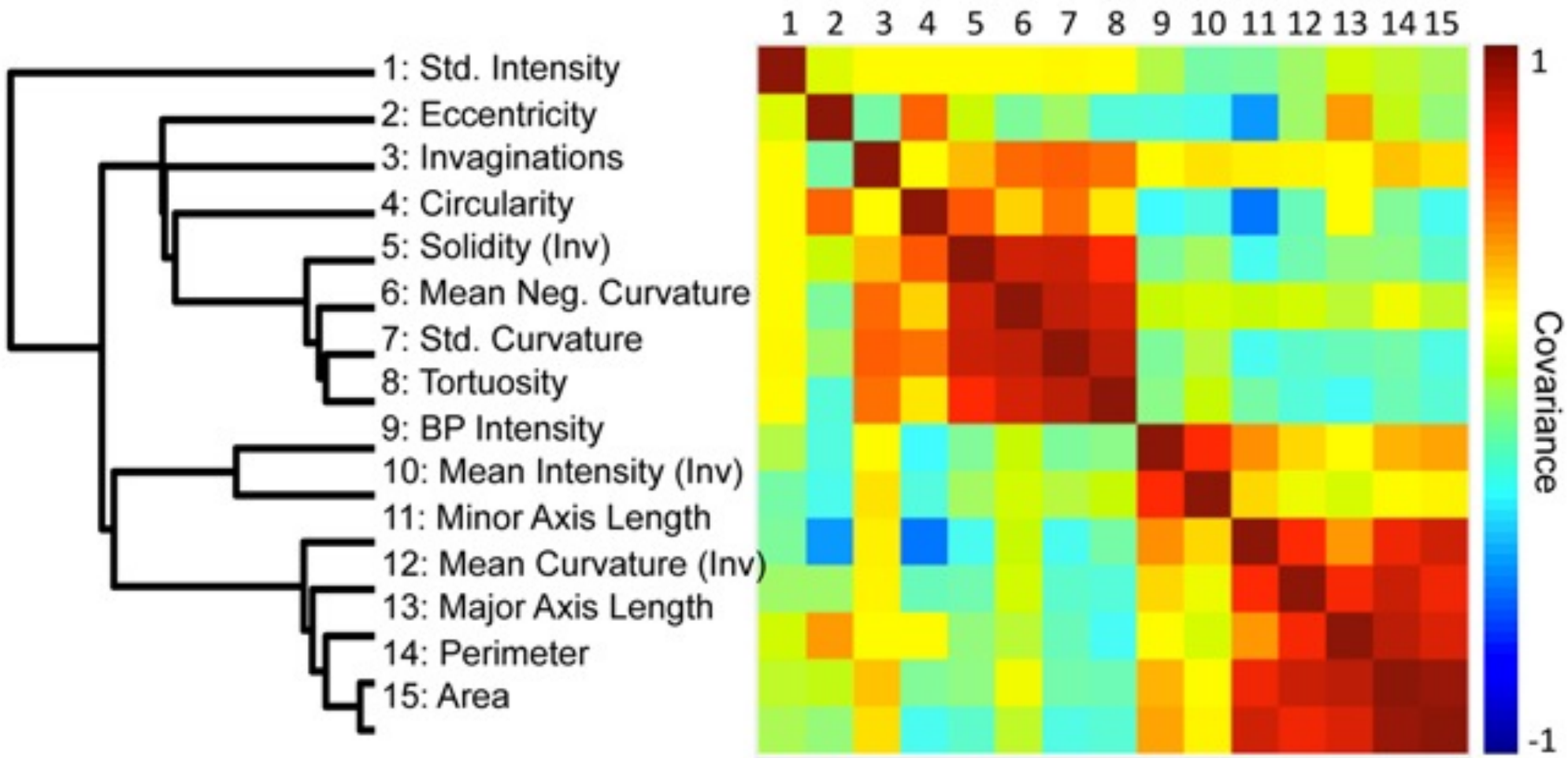
COVARIANCE MATRICES

- Covariance Matrix



COVARIANCE MATRICES

- Covariance Matrix



COVARIANCE MATRICES

- An example:

The table below displays scores on math, English, and art tests for 5 students. Note that data from the table is represented in matrix **A**, where each column in the matrix shows scores on a test and each row shows scores for a student.

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

A

Given the data represented in matrix **A**, compute the variance of each test and the covariance between the tests.

COVARIANCE MATRICES

- Calculating a Covariance Matrix:

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$\Rightarrow$

$$\begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

A

  - Assume data $\{x_i\}$ with N observations of K parameters
  - Center data by subtracting the mean of each feature in $\{x_i\}$ and write this N x K matrix as X.
  - The centered data can be presented as the covariance matrix:

$$C_X = \frac{1}{N-1} X^T X$$

COVARIANCE MATRICES

- <u>Step 1:</u> Transform the raw scores from matrix **A** into deviation scores, **a**

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

A

a = A - 1 1'A (1/n)

where **1** is an n x 1 column vector of ones,

a is an n x k matrix of deviation scores,

A is an n x k matrix of raw scores

COVARIANCE MATRICES

- Step 1: Transform the raw scores from matrix **A** into deviation scores, **a**

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

A

a = A - 1 1'A (1/n)

$$\mathbf{a} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} (1/5)$$

$$\mathbf{a} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix} - \begin{bmatrix} 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \\ 66 & 60 & 60 \end{bmatrix} = \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix}$$

COVARIANCE MATRICES

- Step 2: Compute **a'a**, the n x n deviation sums of squares and cross products matrix for **A**

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

A

$$\mathbf{a}'\mathbf{a} = \begin{bmatrix} 24 & 24 & -6 & -6 & -36 \\ 0 & 30 & 0 & 0 & -30 \\ 30 & -30 & 0 & 30 & -30 \end{bmatrix} \begin{bmatrix} 24 & 0 & 30 \\ 24 & 30 & -30 \\ -6 & 0 & 0 \\ -6 & 0 & 30 \\ -36 & -30 & -30 \end{bmatrix} = \begin{bmatrix} 2520 & 1800 & 900 \\ 1800 & 1800 & 0 \\ 900 & 0 & 3600 \end{bmatrix}$$

COVARIANCE MATRICES

- Step 3: Divide each element in the deviation sum of squares matrix by n

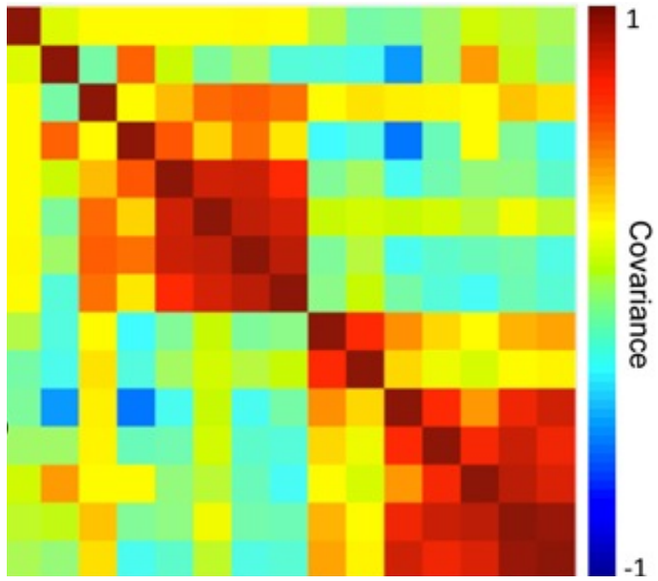
Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

$$\Rightarrow \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

A

$$\mathbf{a}'\,\mathbf{a} / n = \begin{bmatrix} 2520/5 & 1800/5 & 900/5 \\ 1800/5 & 1800/5 & 0/5 \\ 900/5 & 0/5 & 3600/5 \end{bmatrix} = \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

PRINCIPAL COMPONENT ANALYSIS



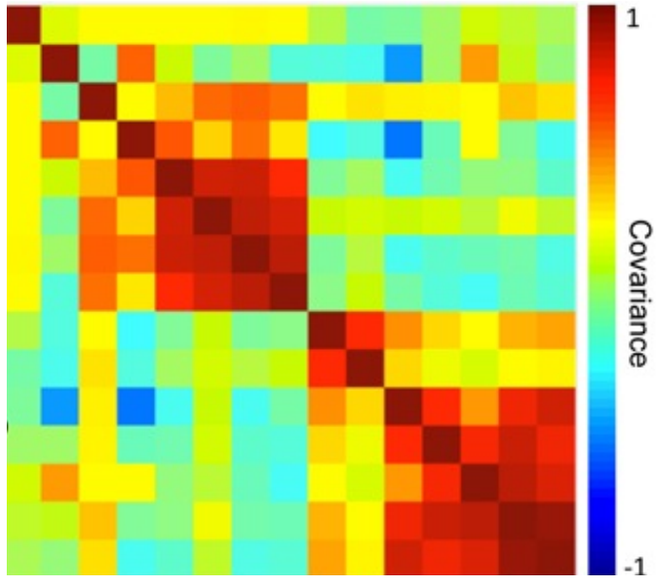
- Theoretical outline:
    - Identify a projection of $\{x_i\}$, say, $\boldsymbol{R}$, that is aligned with the directions of maximal variance.
    - We call this projection $Y = XR$ and its covariance is:

$$C_Y = R^T X^T XR = R^T C_X R$$

with

$$C_X = \frac{1}{N-1} X^T X$$

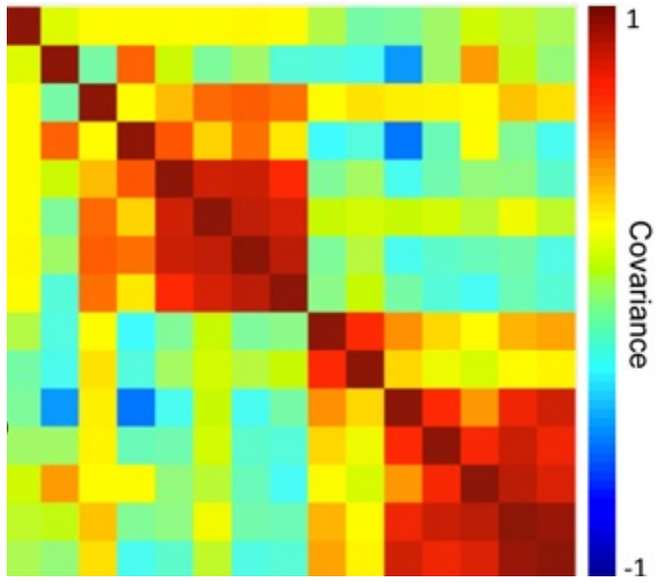
PRINCIPAL COMPONENT ANALYSIS



- Theoretical outline:
    - The first principle component, r_1 of $\boldsymbol{R}$, is defined as the projection with the maximal variance, derived using Lagrange multipliers and defining the cost function as:

$$\phi(r_1, \lambda_1) = r_1^T C_X r_1 - \lambda_1 (r_1^T r_1 - 1)$$

PRINCIPAL COMPONENT ANALYSIS



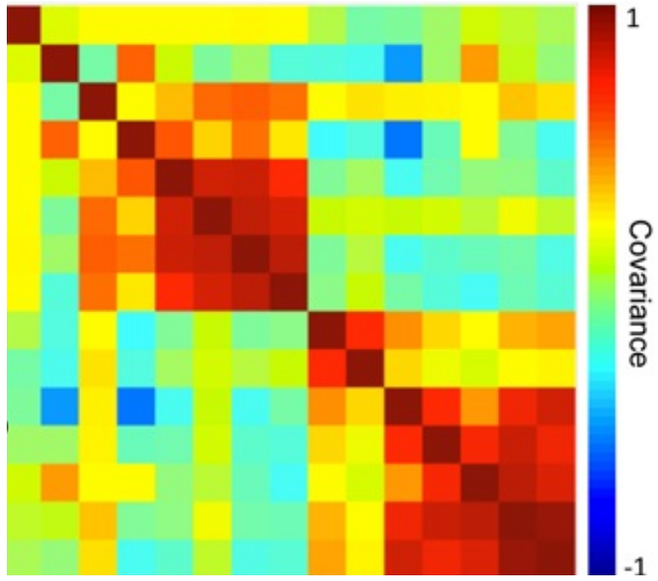
$$\phi(r_1, \lambda_1) = r_1^T C_X r_1 - \lambda_1 (r_1^T r_1 - 1)$$

- Setting the derivative of $\phi(r_1, \lambda_1)$ with respect to r_1 equal to zero gives:

$$C_X r_1 - \lambda_1 r_1 = 0$$

- So λ_1 is the root of the equation $\det(C_X - \lambda_1 I) = 0$ and a principal component of the covariance matrix.
- The variance for the first principal component is maximized when $\lambda_1 = r_1^T C_X r_1$ is the largest eigenvalue of the covariance matrix

PRINCIPAL COMPONENT ANALYSIS



$$\phi(r_1, \lambda_1) = r_1^T C_X r_1 - \lambda_1 (r_1^T r_1 - 1)$$

- The second (and further principal components can be similarly derived after requiring that the principal components of the cost function are uncorrelated.

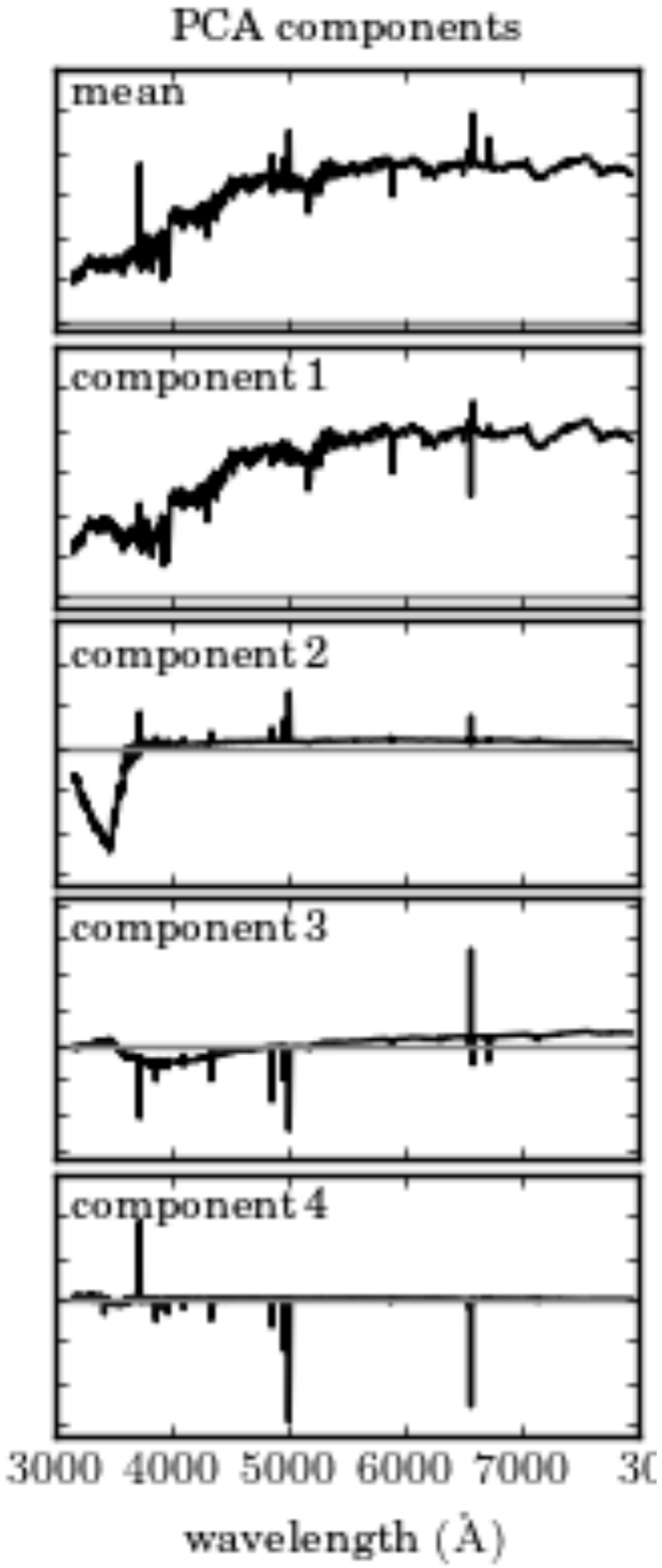
- Ordering the eigenvectors by their eigenvalue defines the set principal components for X

PRINCIPAL COMPONENT ANALYSIS

- A large, diverse dataset can be encoded in a small number of eigenvectors (94% of SDSS galaxies can be constructed with 10 eigenvectors)

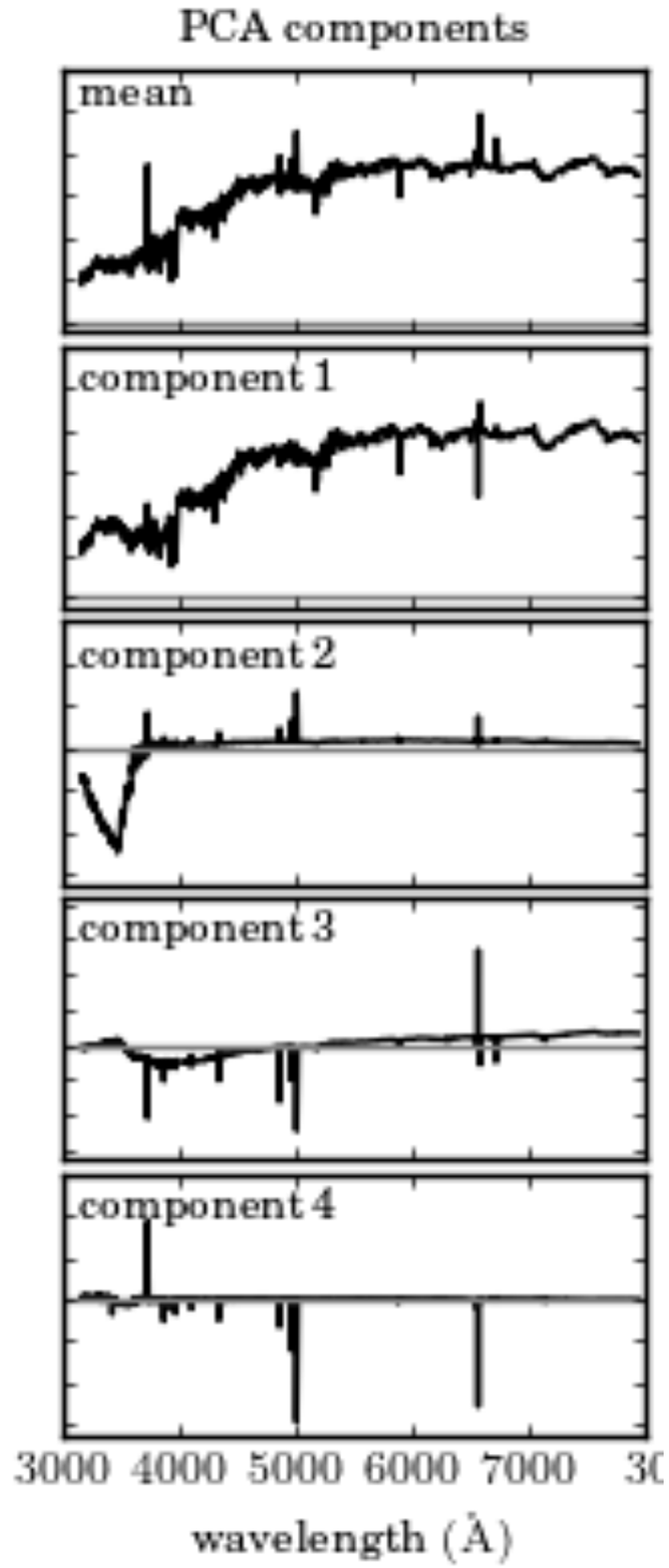
- High eigenvector numbers = low-order components (e.g., continuum)

- Low eigenvector numbers = higher-order features (e.g., emission lines)



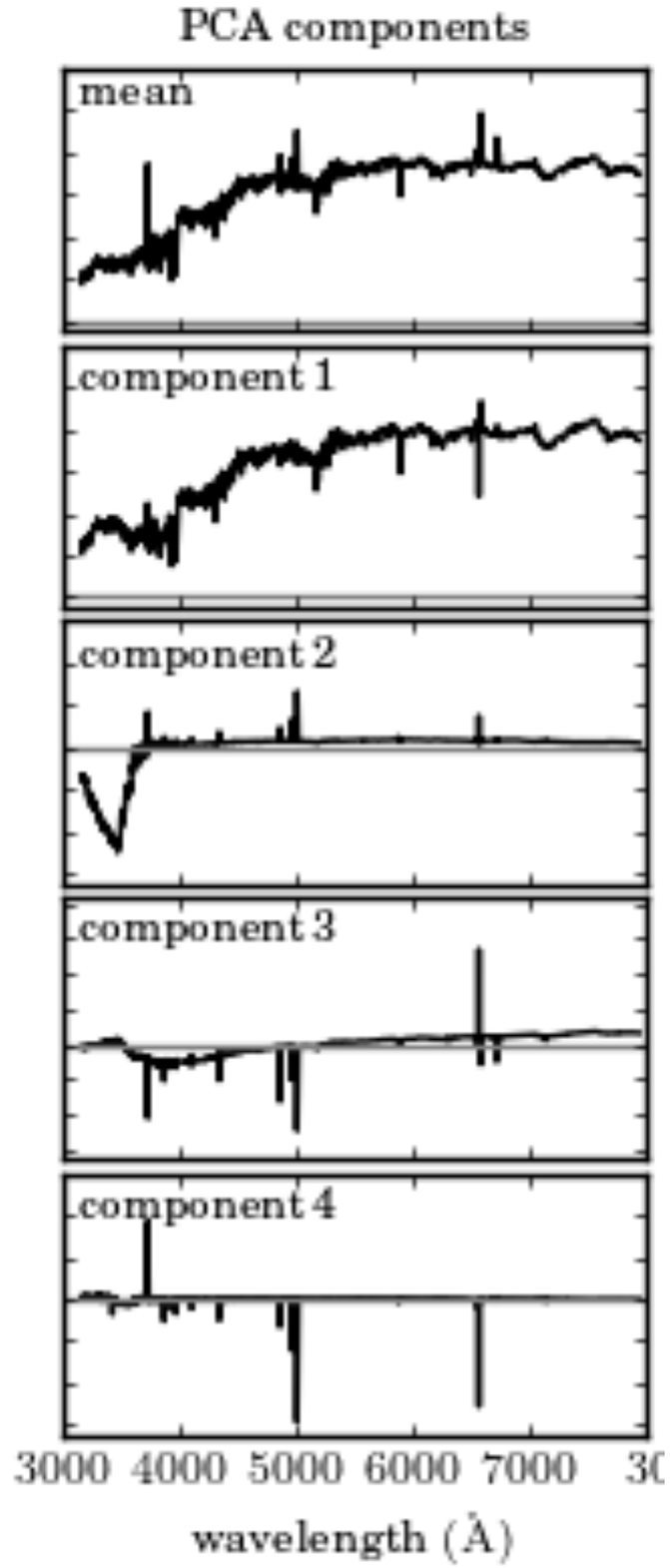
PRINCIPAL COMPONENT ANALYSIS

- Efficient computational methods:
  - Singular Value Decomposition (SVD) of X
  - eigenvalue decomposition of C_X (use when N>>K)
  - eigenvalue decomposition of M_X (use when K>>N)



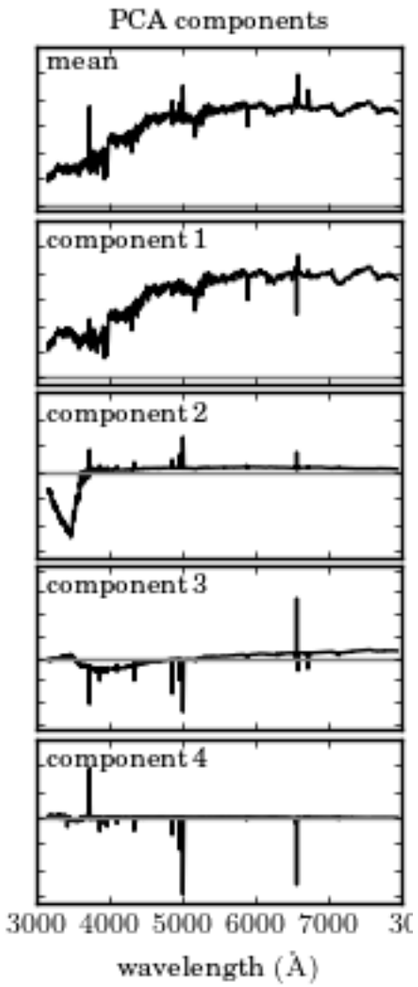
PRINCIPAL COMPONENT ANALYSIS

- Examples of effective pre-processing steps for greater PCA efficiency:
  - flux normalization
  - background subtraction

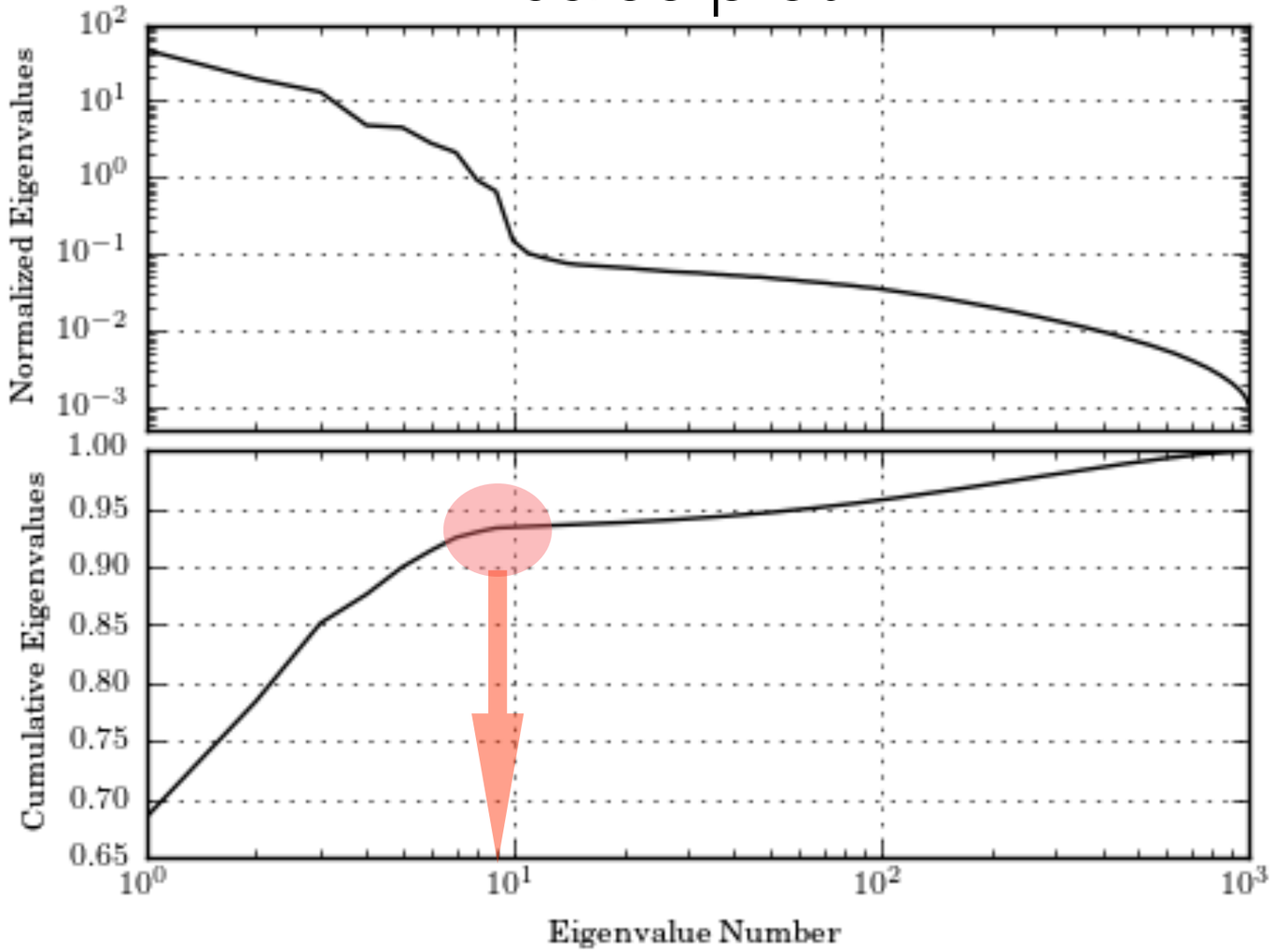


PRINCIPAL COMPONENT ANALYSIS

- How many principle components should be fit?

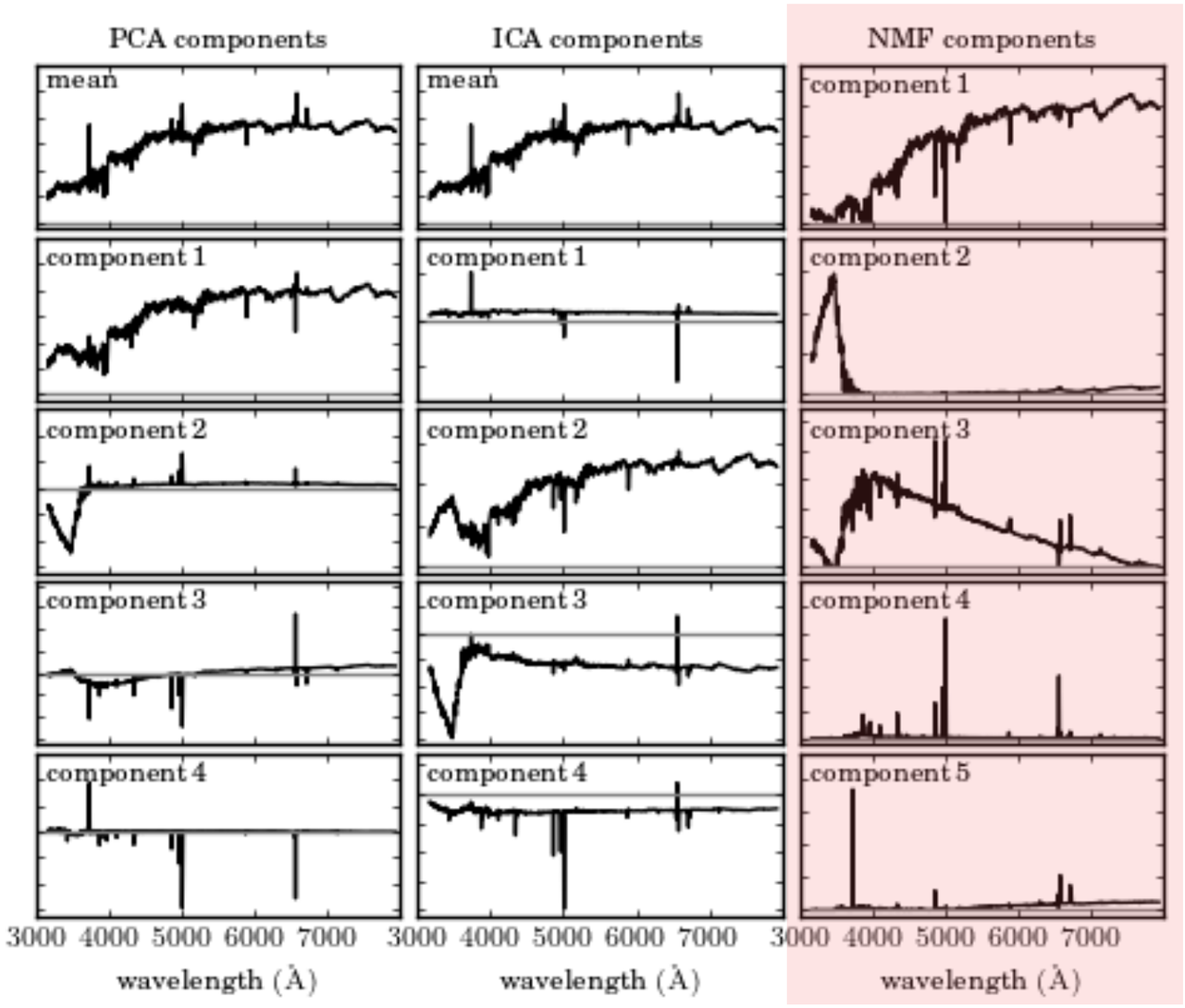


"scree plot"



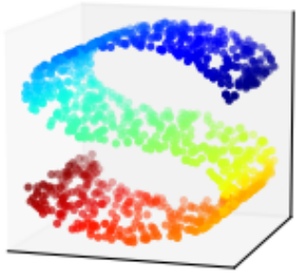
PRINCIPAL COMPONENT ANALYSIS

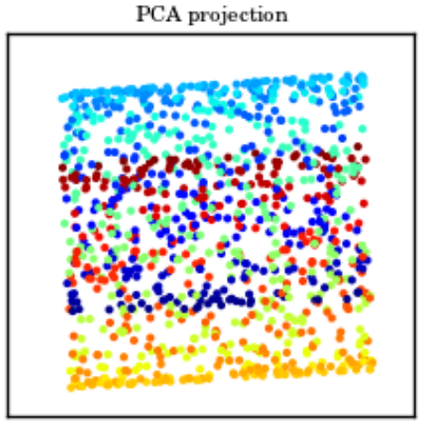
- Preferred methods?



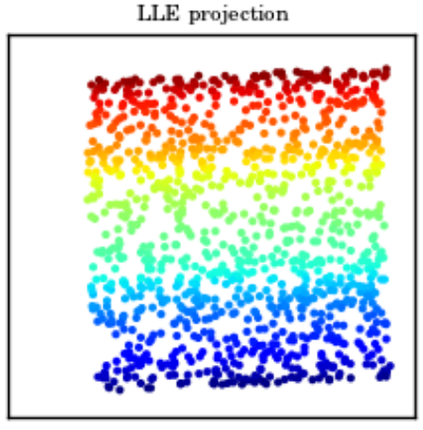
- e.g., Nonnegative Matrix Factorization (NMF), for cases when we know that a data vector can be represented by a linear sum of positive components (e.g., galaxy spectra)

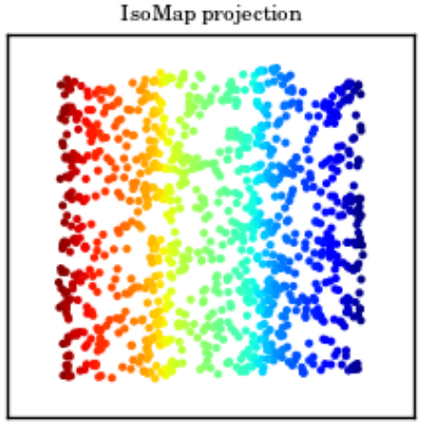
MANIFOLD LEARNING





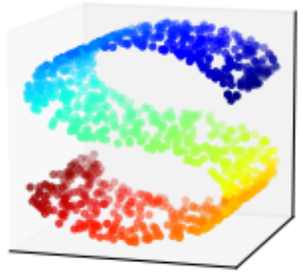
- Non-linear dimensionality reduction:
  - Locally Linear Embedding (LLE)
  - Isometric Mapping (IsoMap)

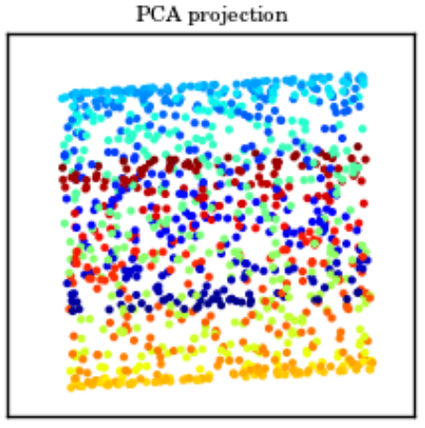


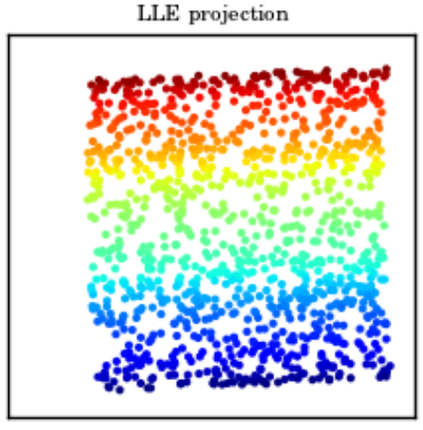


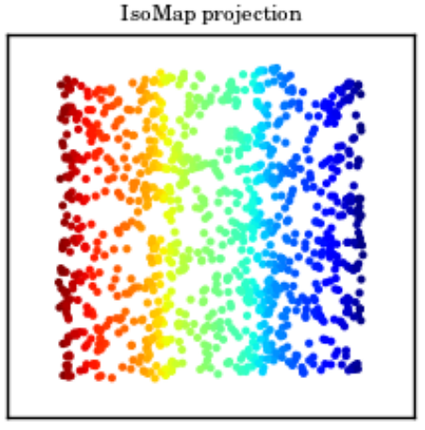
- the preservation of locality enables non-linear features to be captured with fewer components

MANIFOLD LEARNING



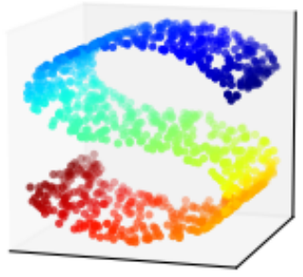


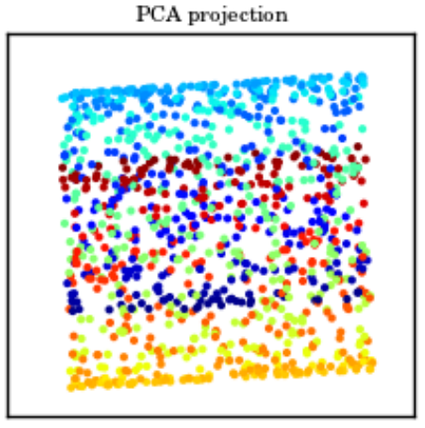


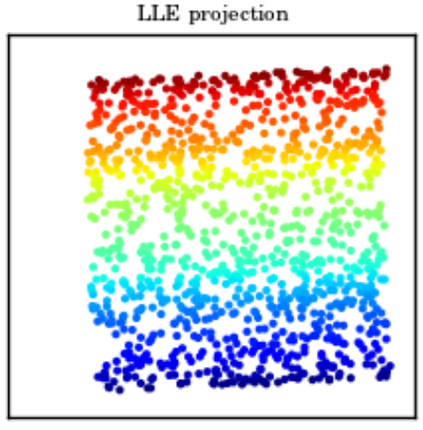


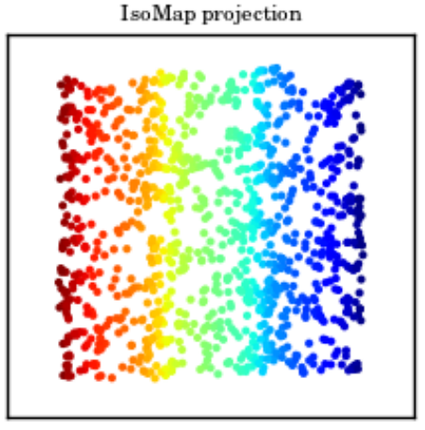
- <u>Weaknesses:</u>
  - noisy / gap-ridden data
  - tuning the right number of nearest neighbors
  - no clean mapping onto a set number of dimensions
  - outlier sensitivity
  - reconstruction from the manifold

CHOOSING A DIMENSIONALITY REDUCTION TECHNIQUE FOR YOUR PROBLEM









- Accuracy
- Interpretability
- Scalability
- Simplicity