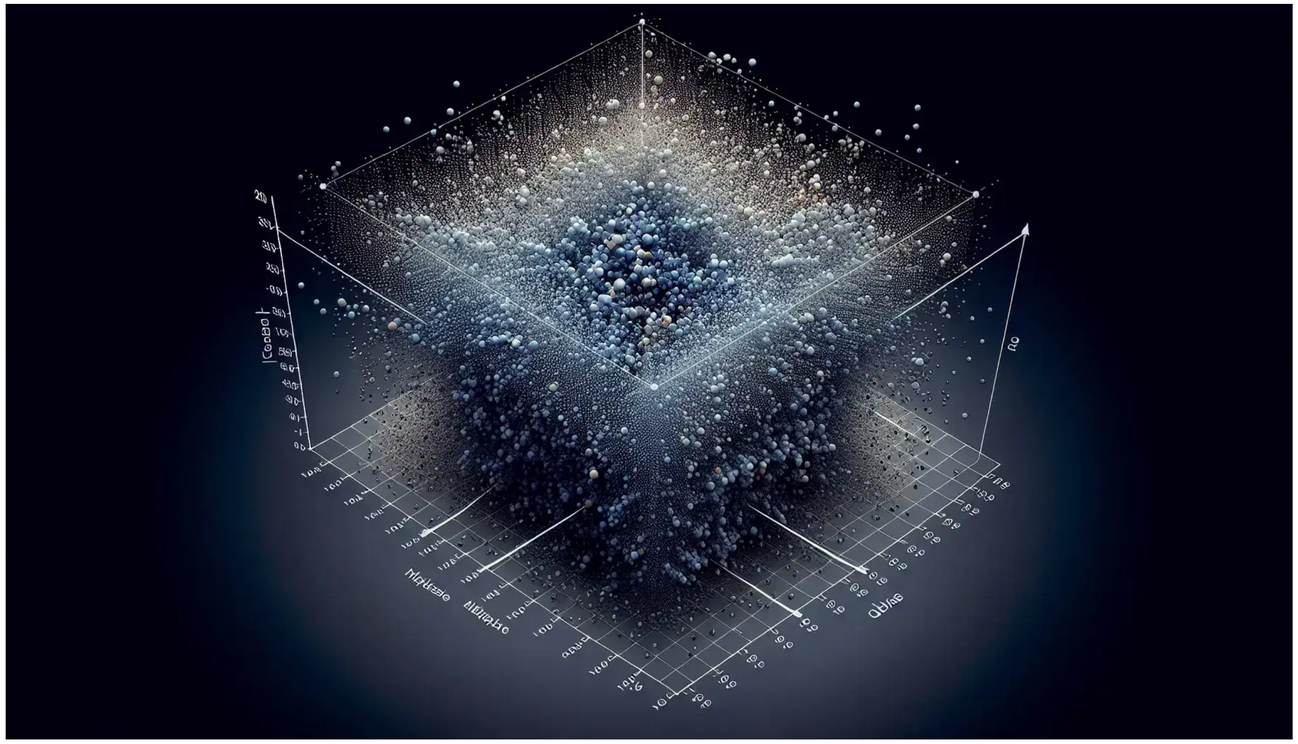
Principal Component Analysis (PCA), k-means clustering, Cross validation techniques & Applications



Dr. Veselina Kalinova

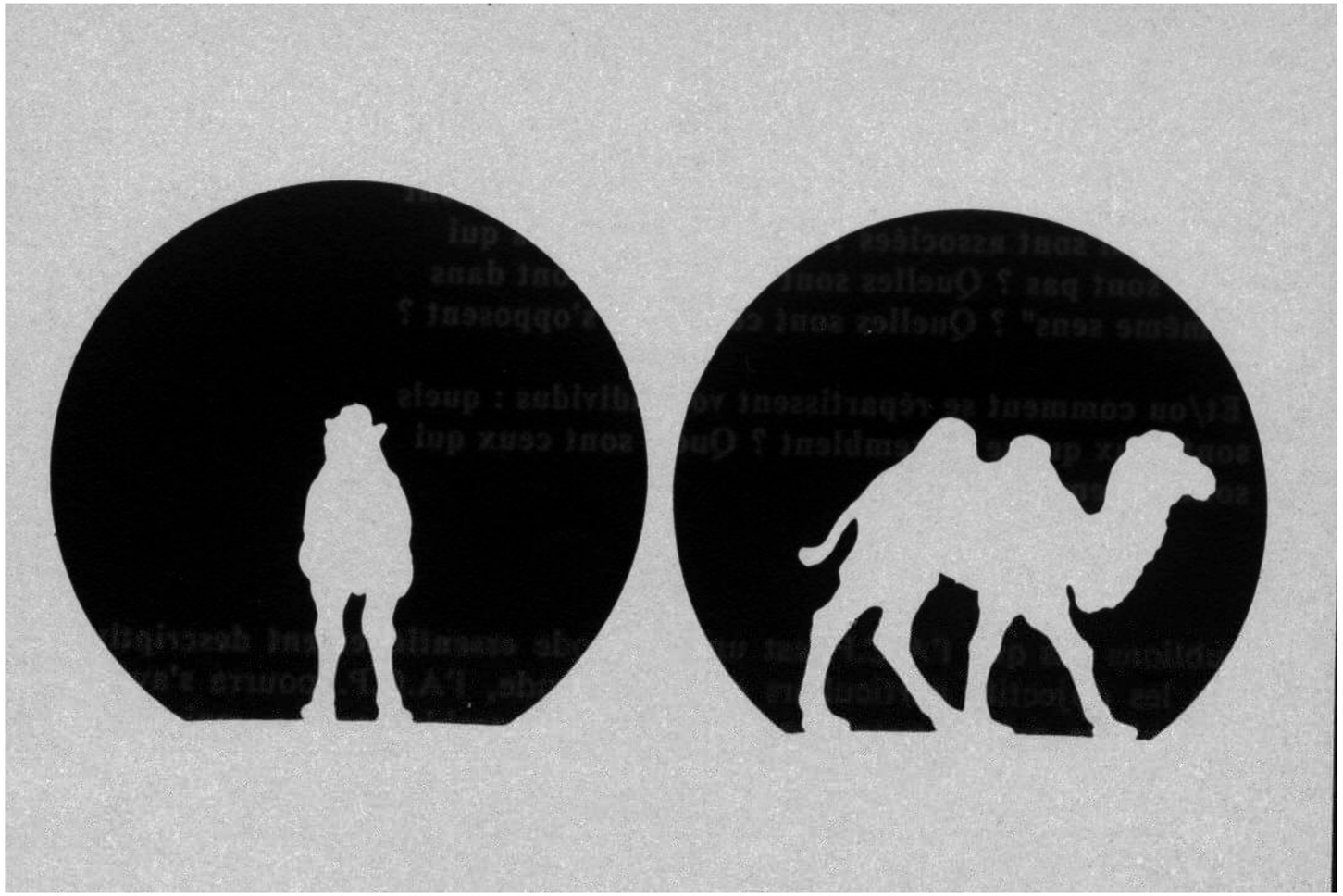
Institute for Sustainable Hydrogen Economy
Forschungszentrum Jülich

Max-Planck-Institut für Radioastronomie

OPINCHARGE

HC·H2 Helmholtz-Cluster Wasserstoff

Motivation



credit: http://web.stanford.edu/class/bios221/PCA_Slides.html

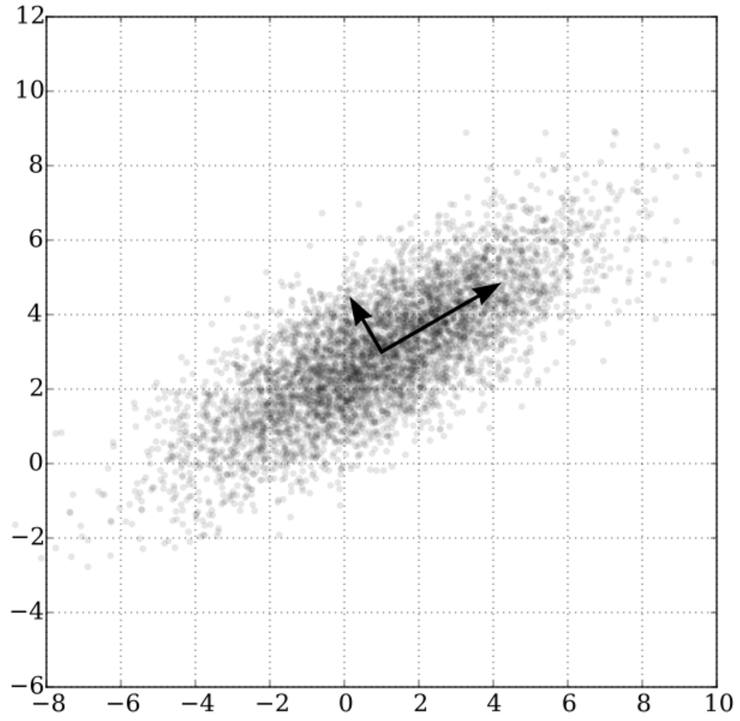
Which projection of the image will provide more information for the object?

It's the projection that maximises the area of the shadow and an equivalent measurement is the sums of squares of the distances between points in the projection, we want to see as much of the variation as possible, that's what PCA does.

Principal Component Analysis (PCA): definition

Karl Pearson (1857 - 1936), English mathematician, biostatistician, who developed PCA in 1901 year.

- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- The number of principal components is less than or equal to the number of original variables.
- This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- The resulting vectors are an uncorrelated orthogonal basis set

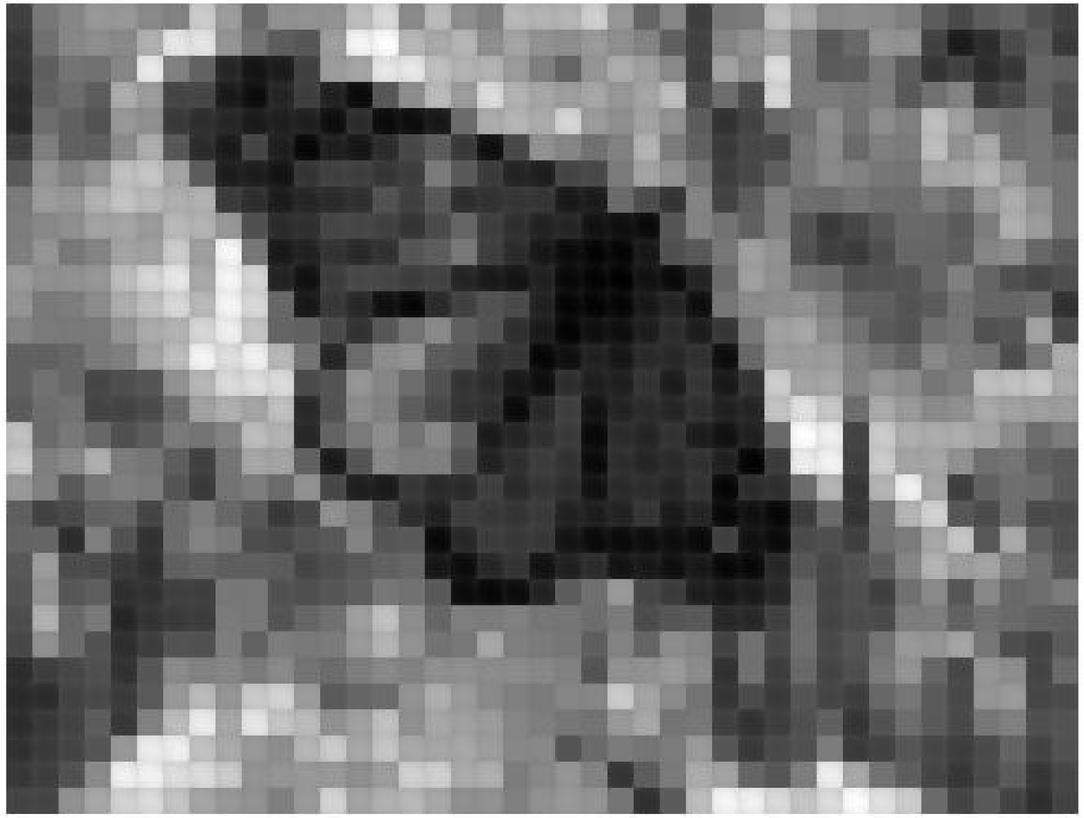


PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.

Reduced size of images via PCA

Original Image

PCA compression: 144D ⇒ 1D

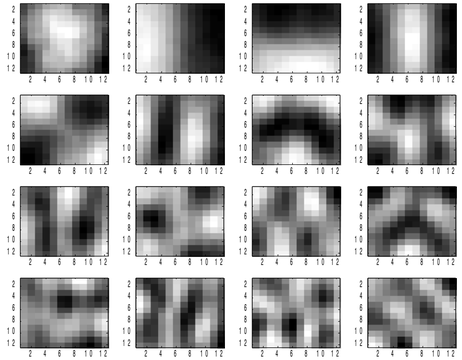


- Divide the original 372x492 image into patches:
  - Each patch is an instance that contains 12x12 pixels on a grid
- View each as a 144-D vector

Reduced size of images via PCA: more components

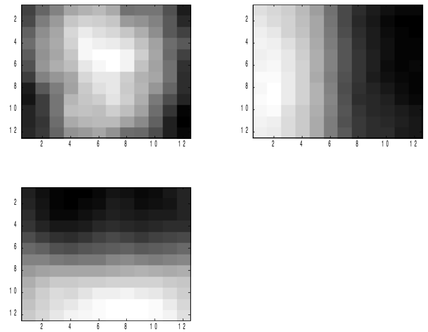
PCA compression: 144D ⇒ 16D

16 most important eigenvectors



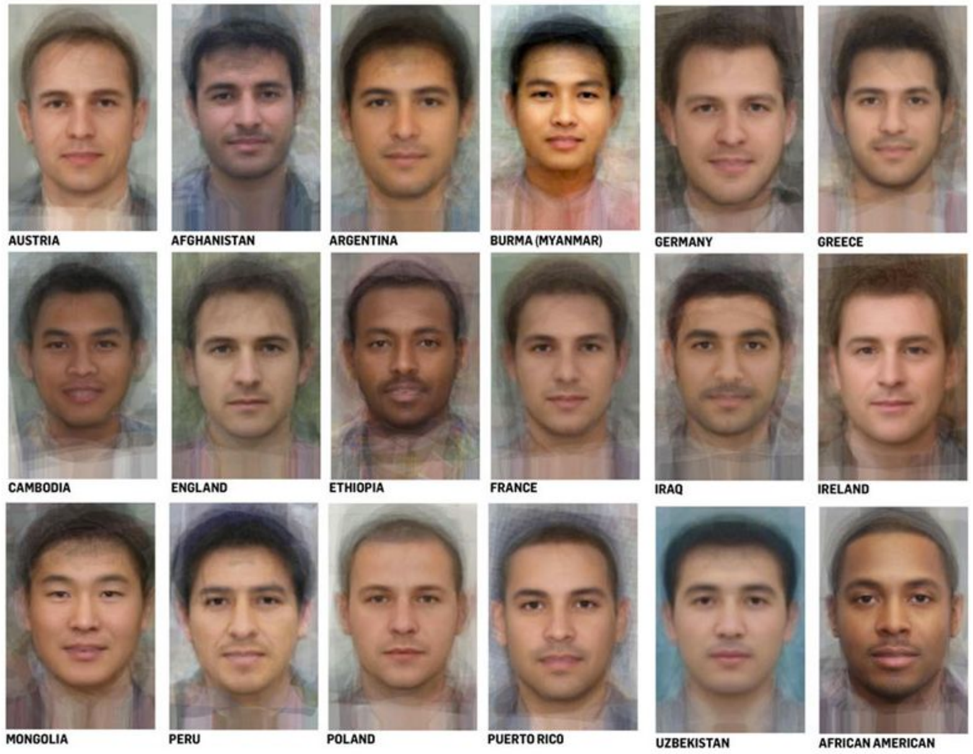
PCA compression: 144D ⇒ 3D

3 most important eigenvectors

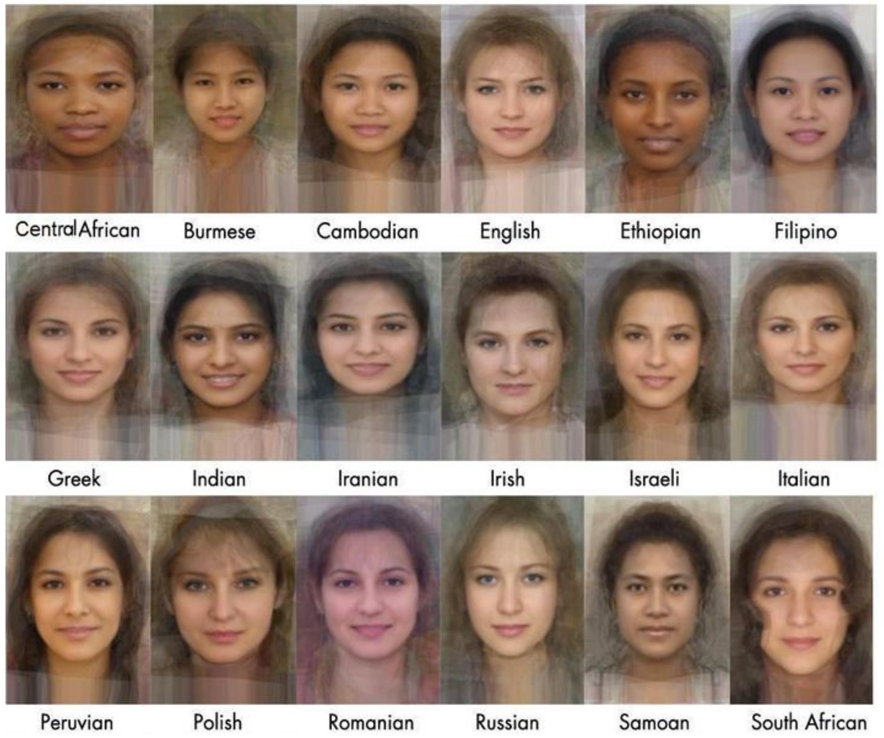


PCA eigenfaces: common features

Average Men of the world



Average Women of the world



source: https://leadingpersonality.wordpress.com/2013/09/30/average-faces-of-men-and-women-around-the-world/

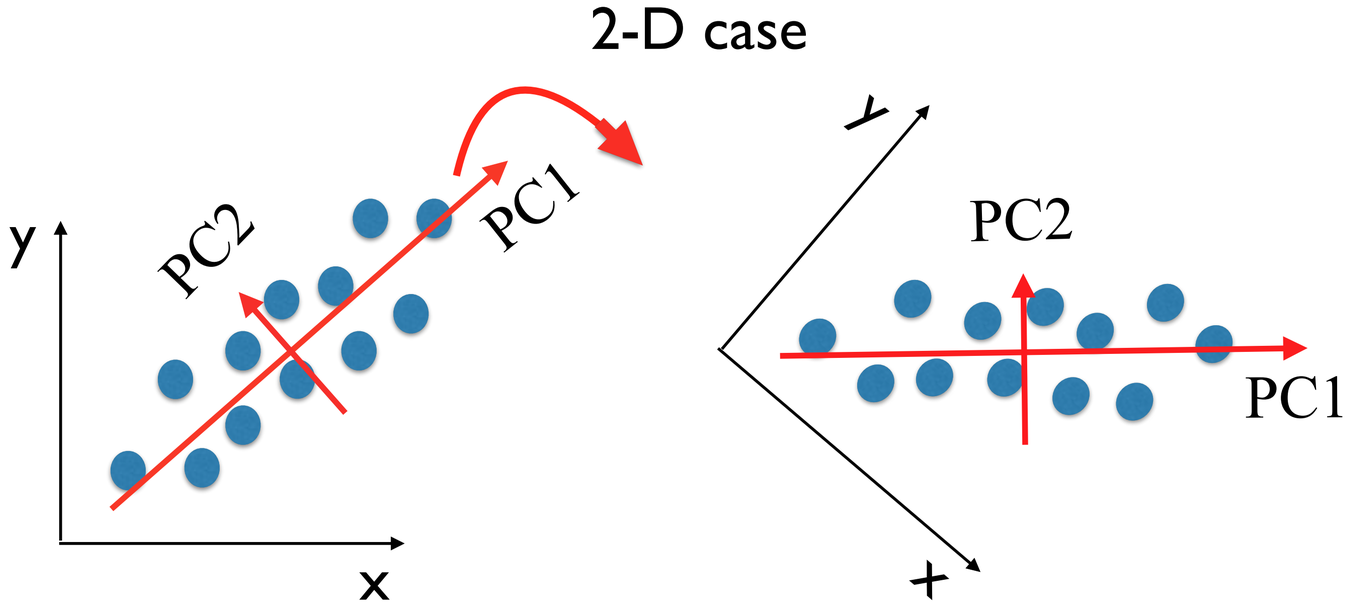
PCA eigenfaces: emotions

happy face

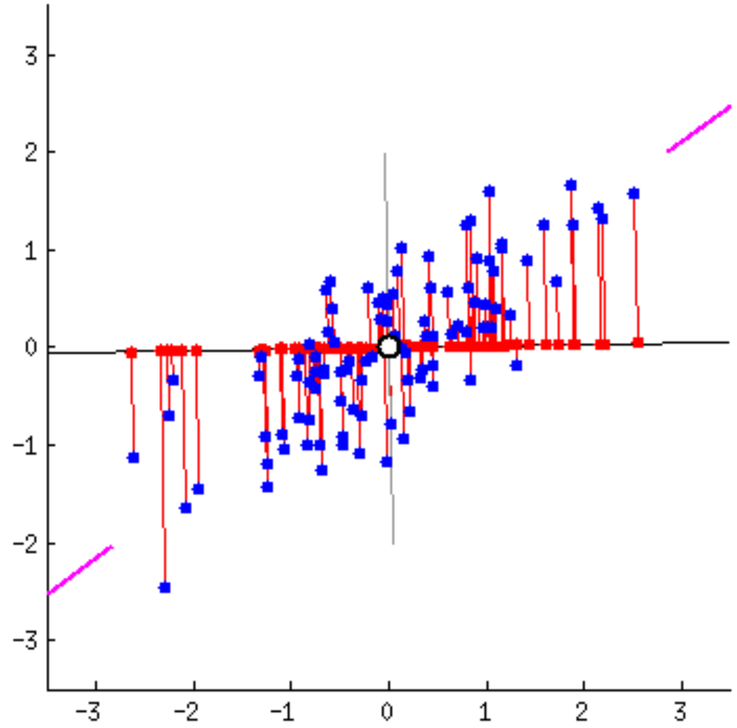
disgust face

credit: Barnaba Poszos

Principal Component Analysis: projection



PC1 captures the direction of the most variation
PC2 captures the direction of the **2nd** most variation
…
PCn captures the direction of the **nnd** most variation



red lines represent the eigenvectors' axes, i.e. PC axes: orthogonal base of eigenvectors

source: http://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Eigenvalue quation

- The **eigenvalue equation** is a fundamental concept in linear algebra, especially in the context of matrix analysis, and is central to methods like PCA.
- It defines the relationship between a matrix, its eigenvalues, and eigenvectors.

The Eigenvalue Equation

The eigenvalue equation for a square matrix $\mathbf{A}$ is given by:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

Where:

- $\mathbf{A}$ is a square matrix (e.g., a covariance matrix in PCA),
- $\mathbf{v}$ is a non-zero vector called an **eigenvector**,
- λ is a scalar (a number) called an **eigenvalue**.

Eigenvalue quation

What the Eigenvalue Equation Means

The eigenvalue equation essentially states that when the matrix $\mathbf{A}$ multiplies the vector $\mathbf{v}$, the result is a scaled version of $\mathbf{v}$, and the scalar factor is λ. In other words:

- $\mathbf{A}\mathbf{v}$ produces a new vector that is just the original eigenvector $\mathbf{v}$ multiplied by the scalar λ.
- The eigenvector direction does not change, only its magnitude gets scaled by the eigenvalue λ.

Solving the eigenvalue quation

1. **Rearrange the equation**:

$$\mathbf{A}\mathbf{v} - \lambda\mathbf{v} = 0$$

This can be written as:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0$$

where $\mathbf{I}$ is the identity matrix.

2. **Non-trivial solutions**: For a non-trivial solution (i.e., $\mathbf{v} \neq 0$), the matrix $(\mathbf{A} - \lambda\mathbf{I})$ must be singular (i.e., its determinant must be zero). This gives the **characteristic equation**:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

This is a polynomial equation in λ, and solving it gives the eigenvalues.

3. **Find eigenvectors**: Once the eigenvalues λ are found, substitute each λ back into the equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0$ to find the corresponding eigenvectors $\mathbf{v}$.

Example of eigenvalue quation

Example of the Eigenvalue Equation

Suppose $\mathbf{A}$ is a 2x2 matrix:

$$\mathbf{A} = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix}$$

To find the eigenvalues λ, solve the characteristic equation:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

The identity matrix $\mathbf{I}$ is:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So $\mathbf{A} - \lambda \mathbf{I}$ is:

$$\mathbf{A} - \lambda \mathbf{I} = \begin{pmatrix} 4 - \lambda & 1 \\ 2 & 3 - \lambda \end{pmatrix}$$

Now, compute the determinant:

$$\det \begin{pmatrix} 4 - \lambda & 1 \\ 2 & 3 - \lambda \end{pmatrix} = (4 - \lambda)(3 - \lambda) - 2 \times 1 = 0$$

Expanding this:

$$(4 - \lambda)(3 - \lambda) = 12 - 7\lambda + \lambda^2$$

$$12 - 7\lambda + \lambda^2 - 2 = 0 \quad \Rightarrow \quad \lambda^2 - 7\lambda + 10 = 0$$

Solving this quadratic equation:

$$\lambda = \frac{7 \pm \sqrt{(-7)^2 - 4 \cdot 1 \cdot 10}}{2 \cdot 1} = \frac{7 \pm \sqrt{49 - 40}}{2} = \frac{7 \pm \sqrt{9}}{2}$$

$$\lambda_1 = 5, \quad \lambda_2 = 2$$

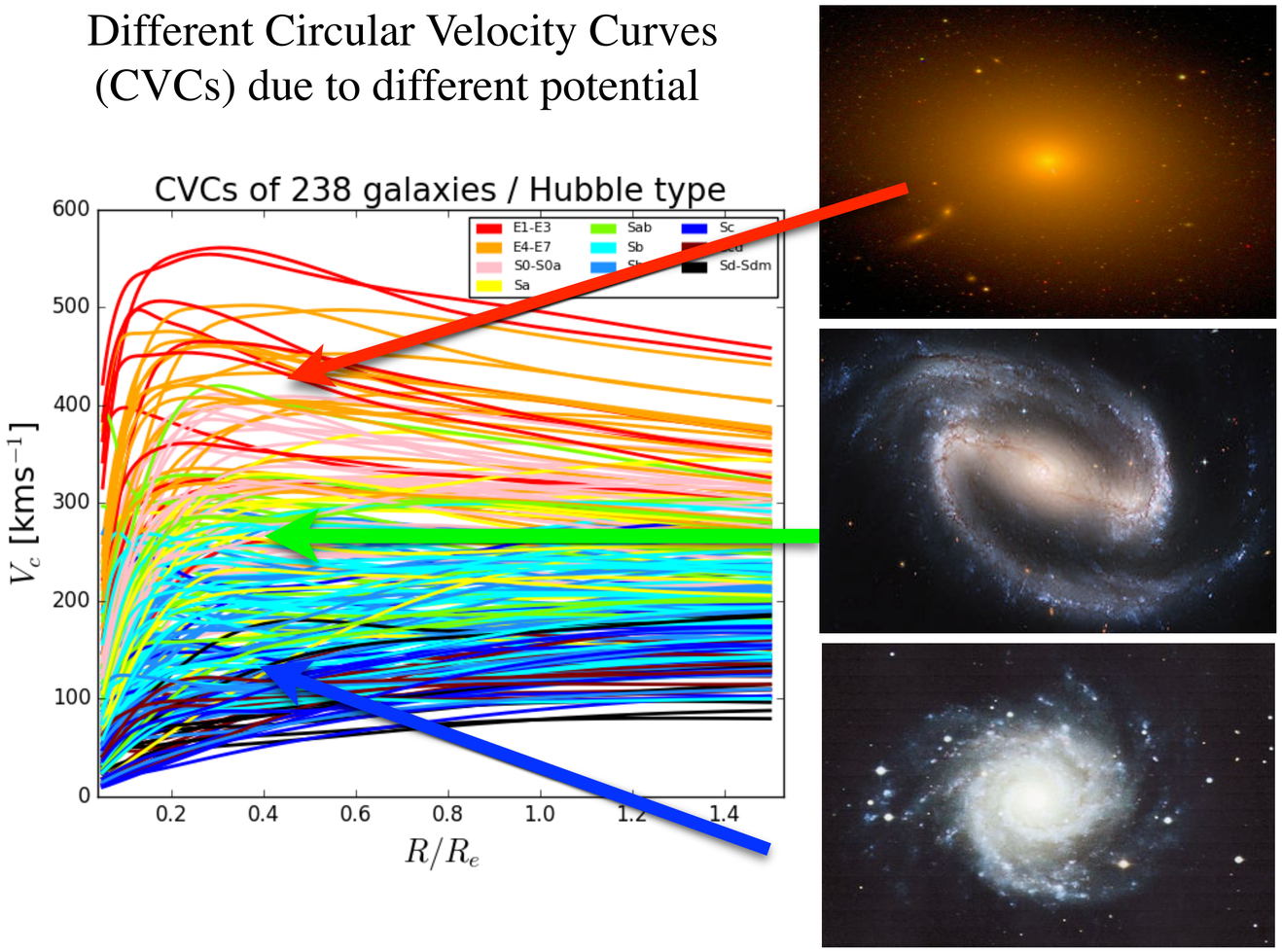
These are the eigenvalues. To find the eigenvectors, we substitute $\lambda_1 = 5$ and $\lambda_2 = 2$ back into the eigenvalue equation to solve for $\mathbf{v}$.

Key Takeaways

- The eigenvalue equation $\mathbf{Av}=\lambda\mathbf{v}$ shows that multiplying a matrix by an eigenvector results in the same eigenvector scaled by its eigenvalue
- Eigenvalues tell us how much the corresponding eigenvectors are scaled
- In PCA, the **eigenvalues** represent the amount of variance captured by each principal component, while **eigenvectors** provide the directions (principal components)

Application of PCA: classifying galaxy velocity curves

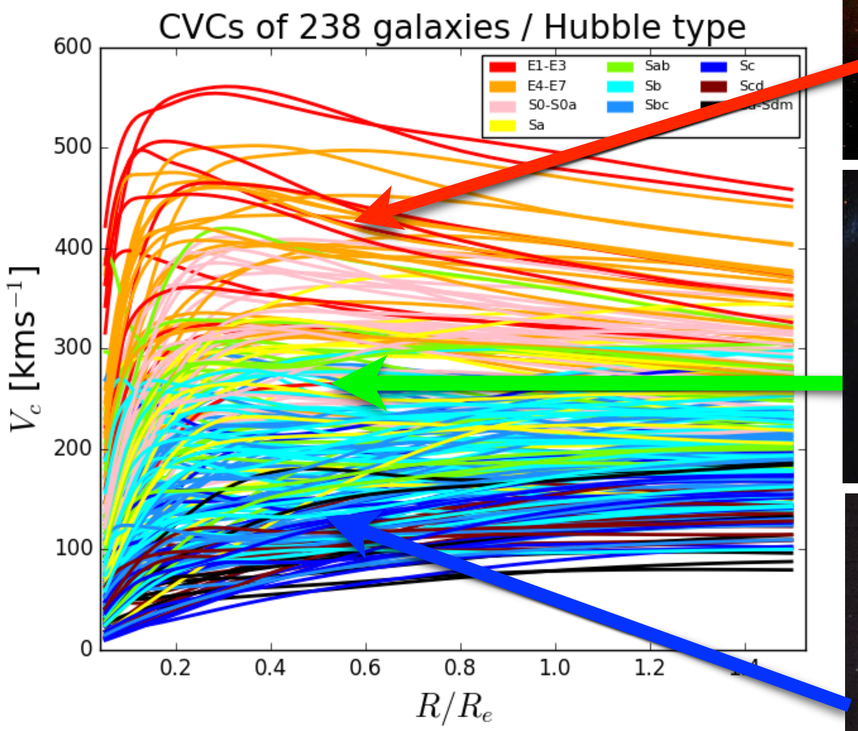
Different Circular Velocity Curves (CVCs) due to different potential



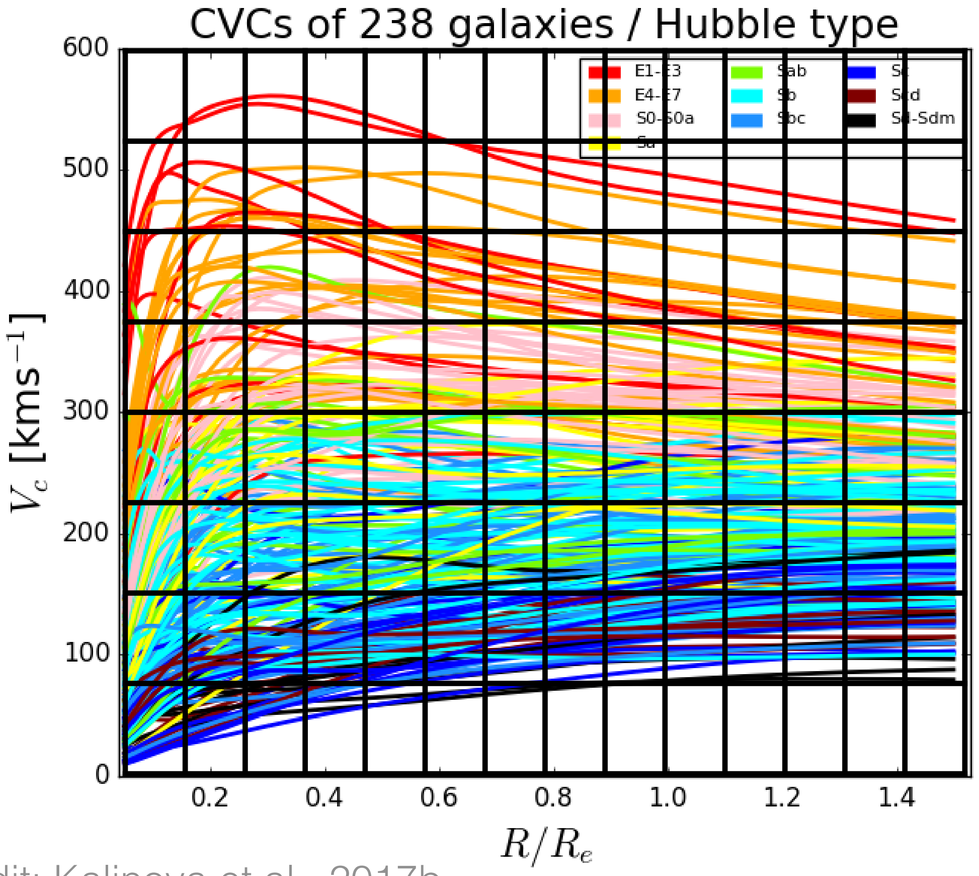
credit: Kalinova et al., 2017b

Application of PCA: classifying galaxy velocity curves

Different Circular Velocity Curves (CVCs) due to different potential

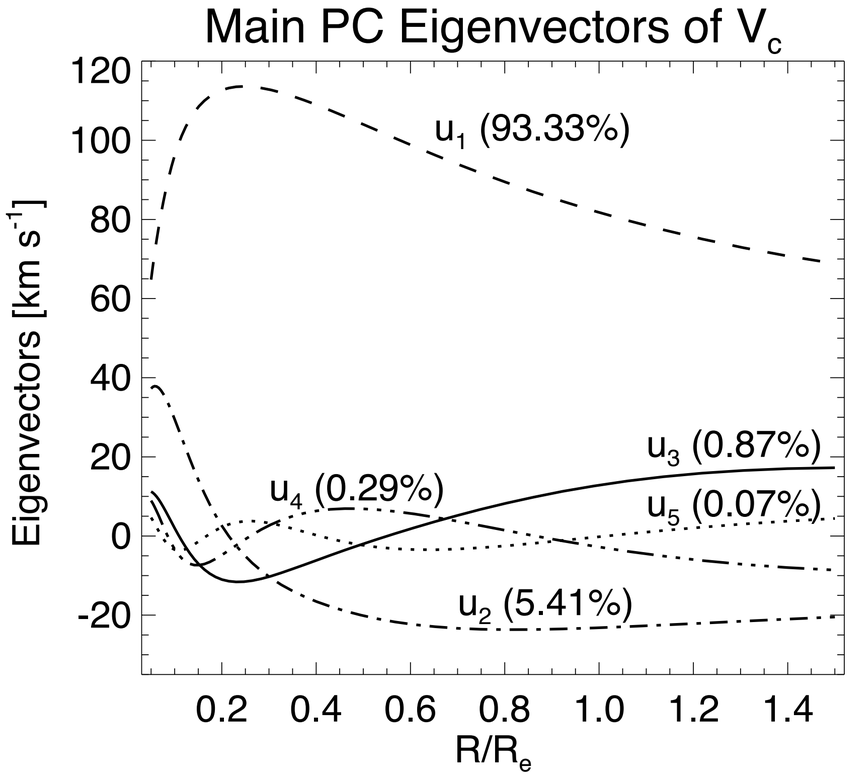


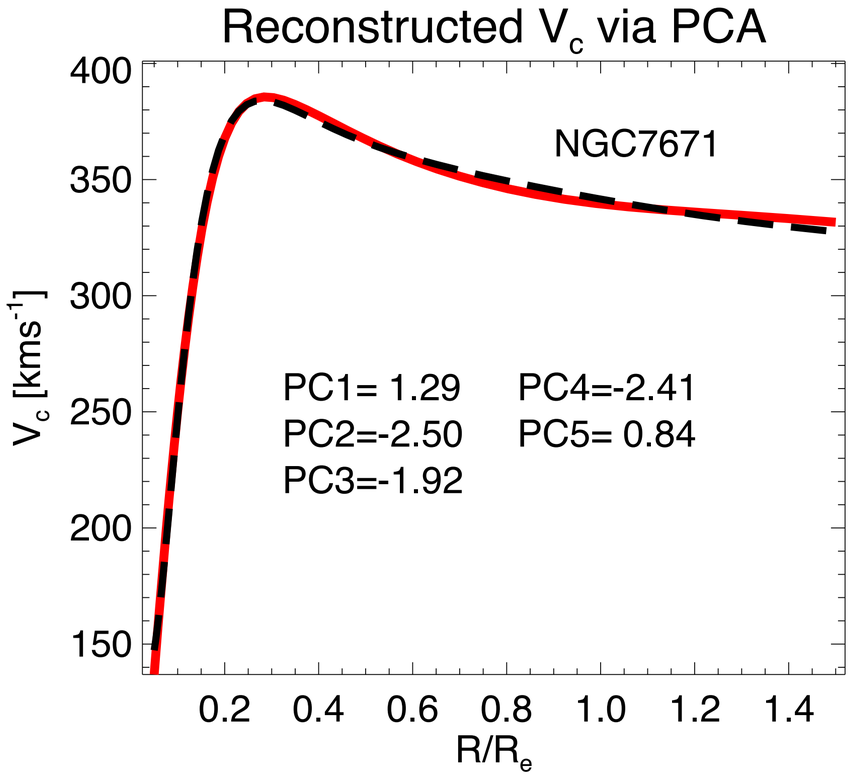
- We compare the shapes of the rotation curves in each cell of the radius (x-axis)
- all curves are combined in one data structure



credit: Kalinova et al., 2017b

Reconstructed velocity curve





$$V_{c,\mathrm{rec}} = (PC_1\mathbf{u_1} + PC_2\mathbf{u_2} + PC_3\mathbf{u_3} + PC_4\mathbf{u_4} + PC_5\mathbf{u_5}) + \overline{V_c}$$

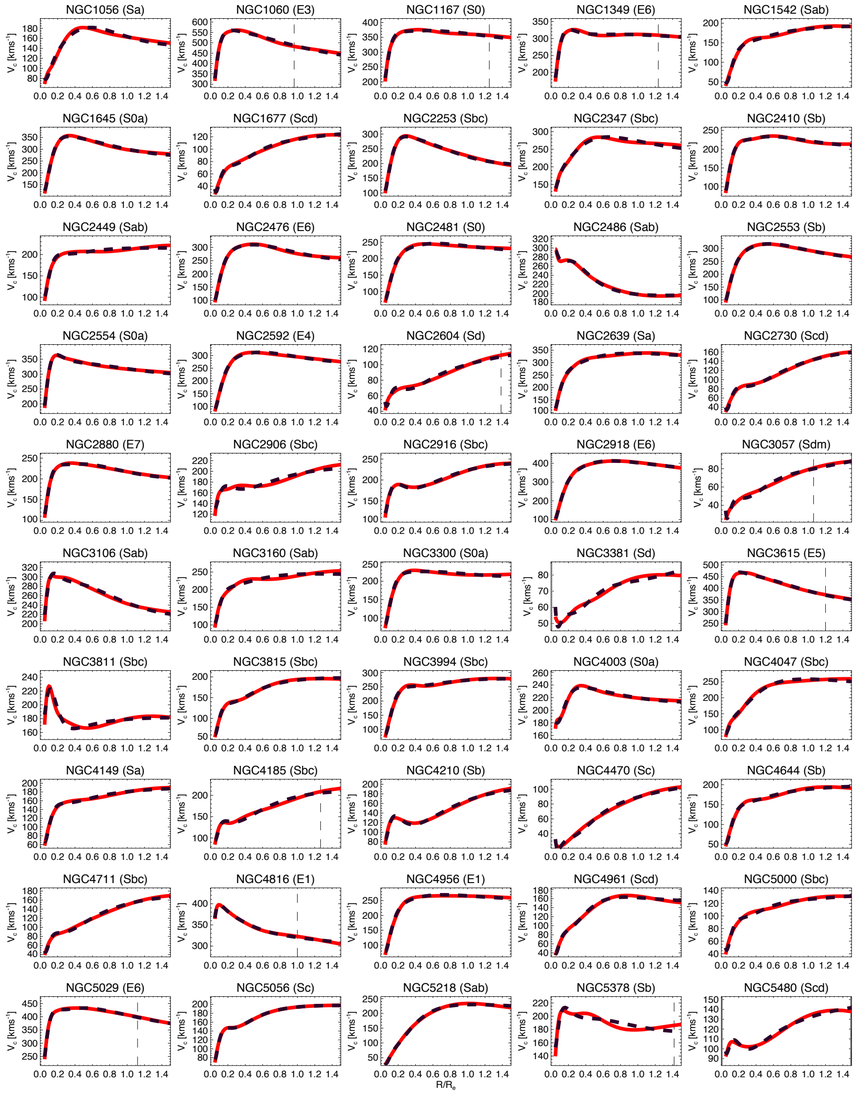
reconstructed

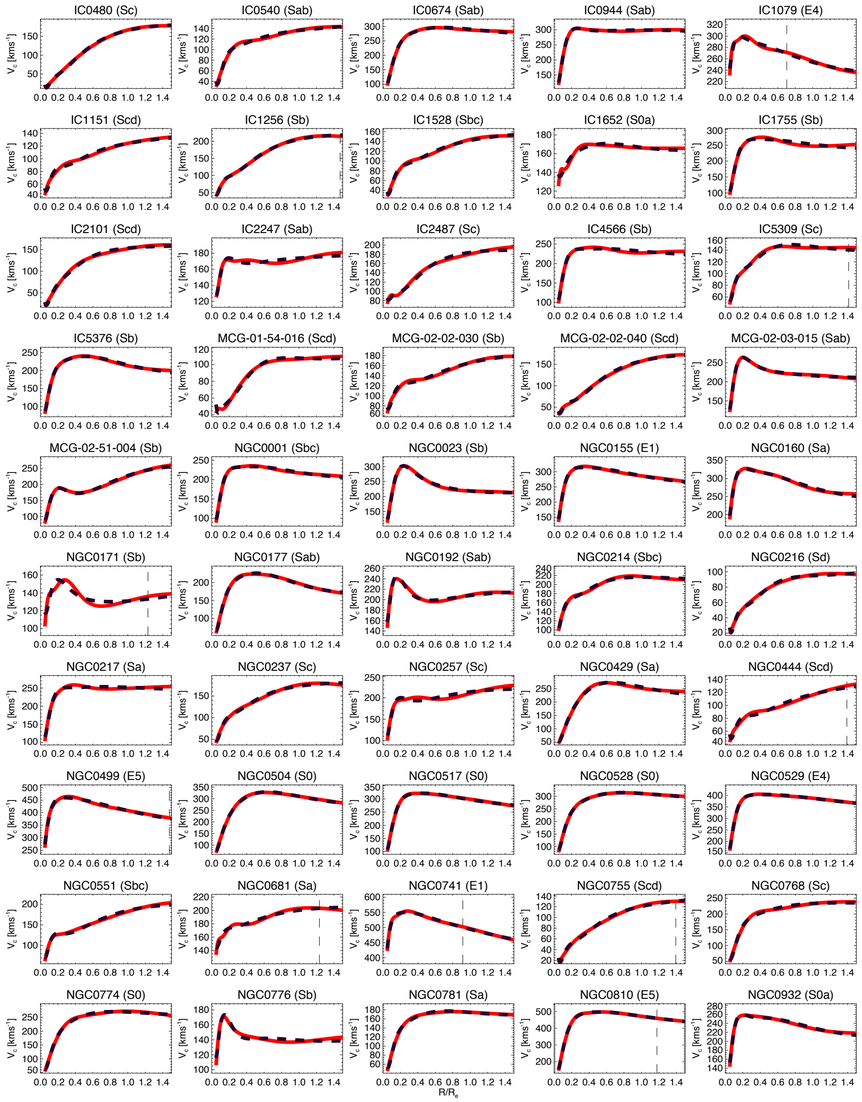
P C1 = +0.79, P C2 = −1.86, PC3 = −1.98, PC4 = −1.90, PC5 = +1.82.

mean velocity of the sample

credit: Kalinova et al., 2017b

Reconstructed velocity curves

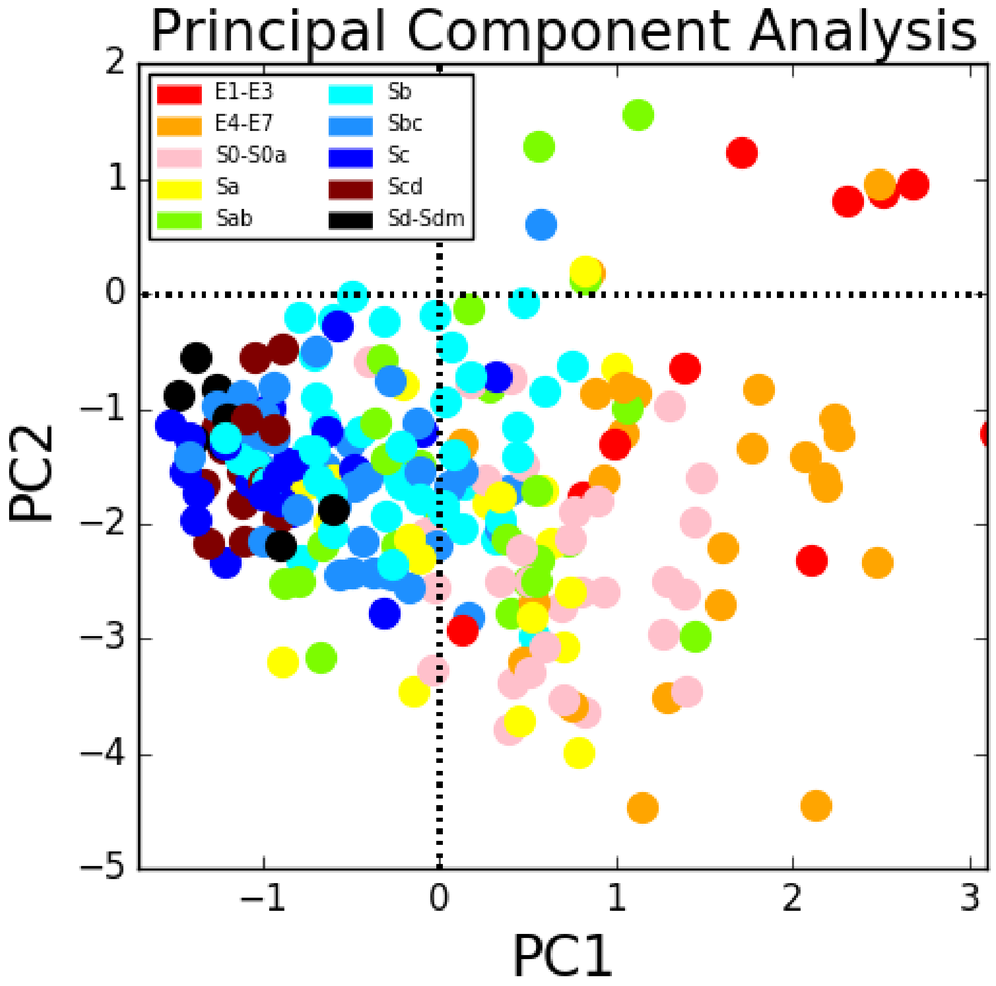




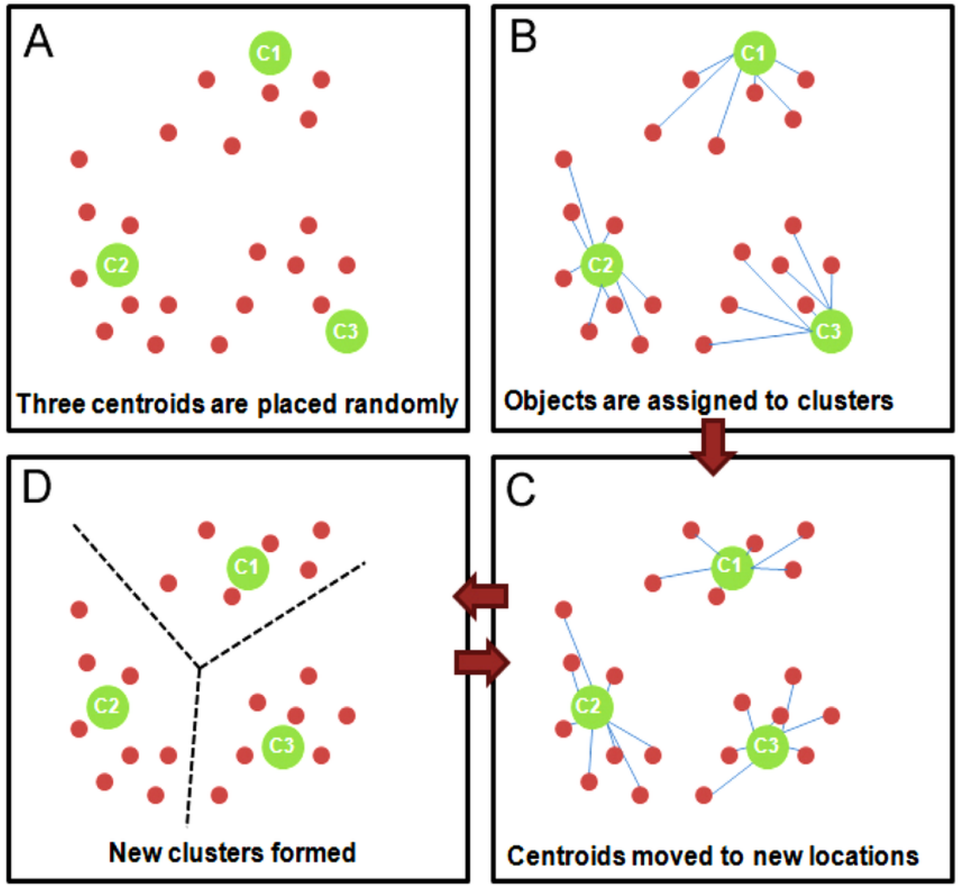
credit:
Kalinova et al.,
2017b

JÜLICH Forschungszentrum

Applying k-means clustering on the PC plane

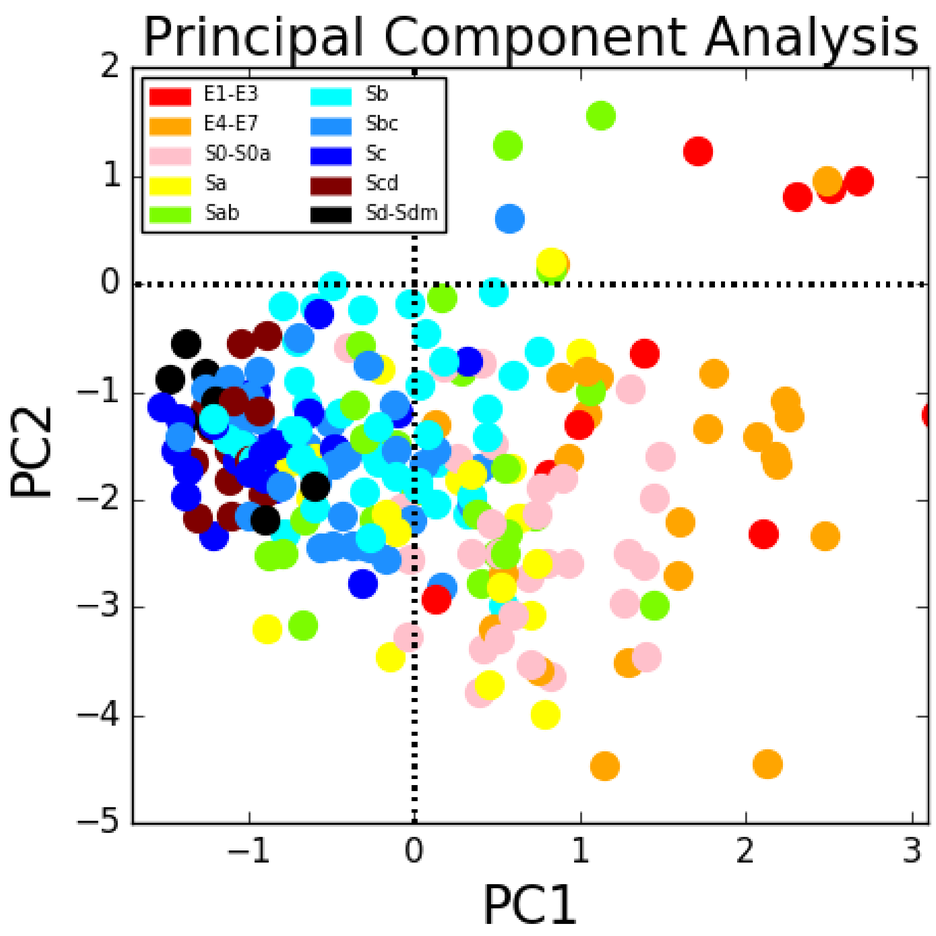


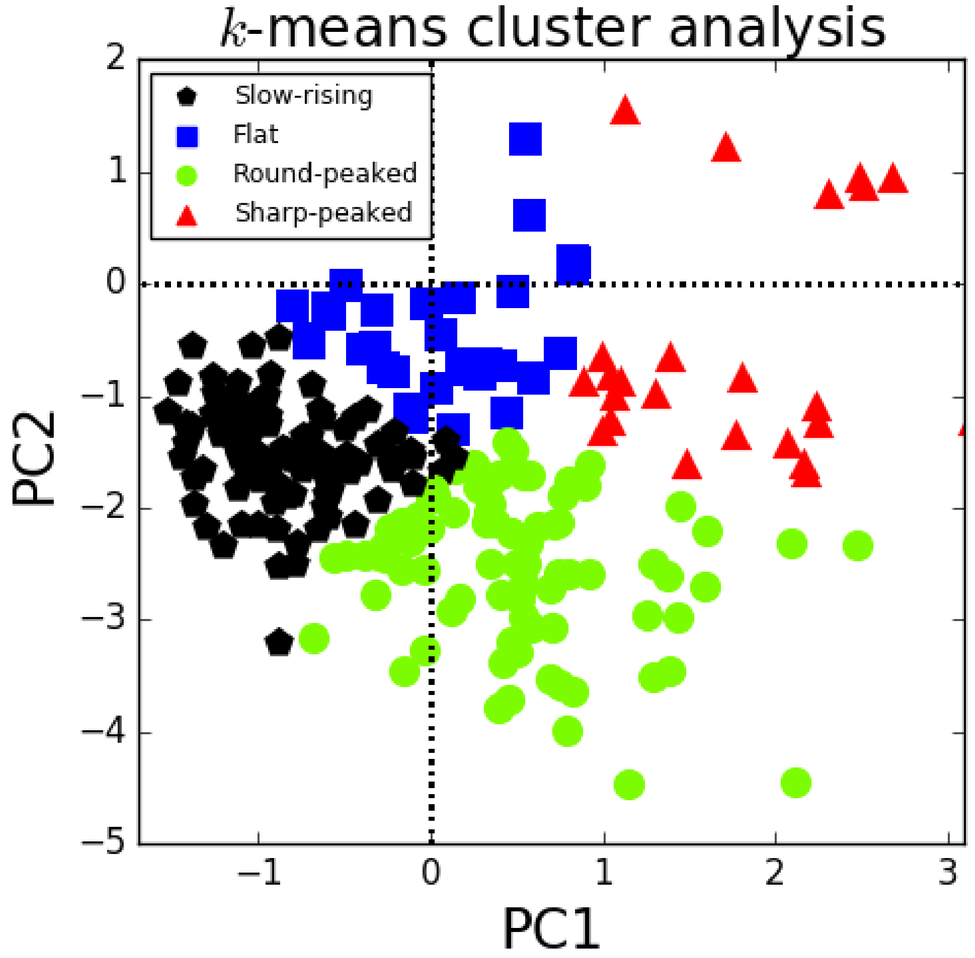
credit: Kalinova et al., 2017b



credit: Rumdeep K. Grewal

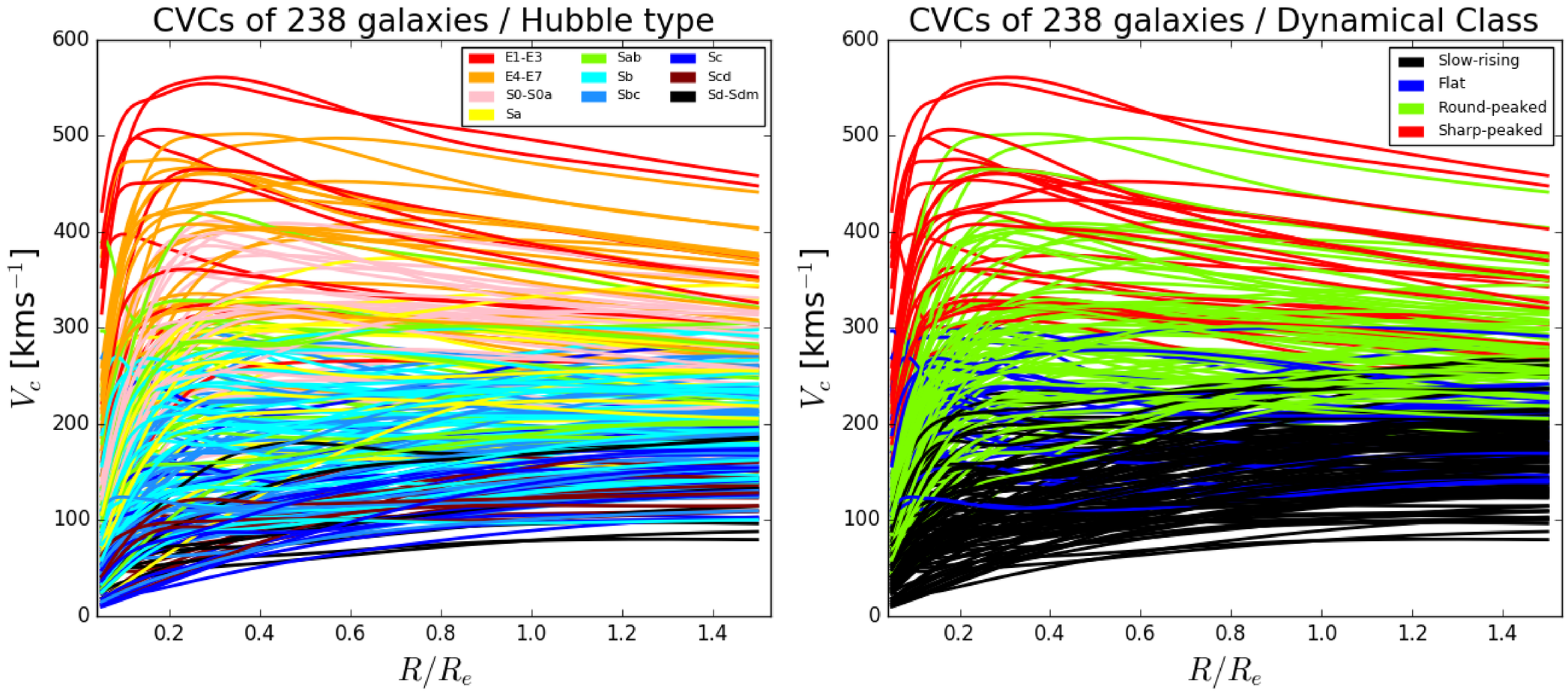
k-means clustering analysis





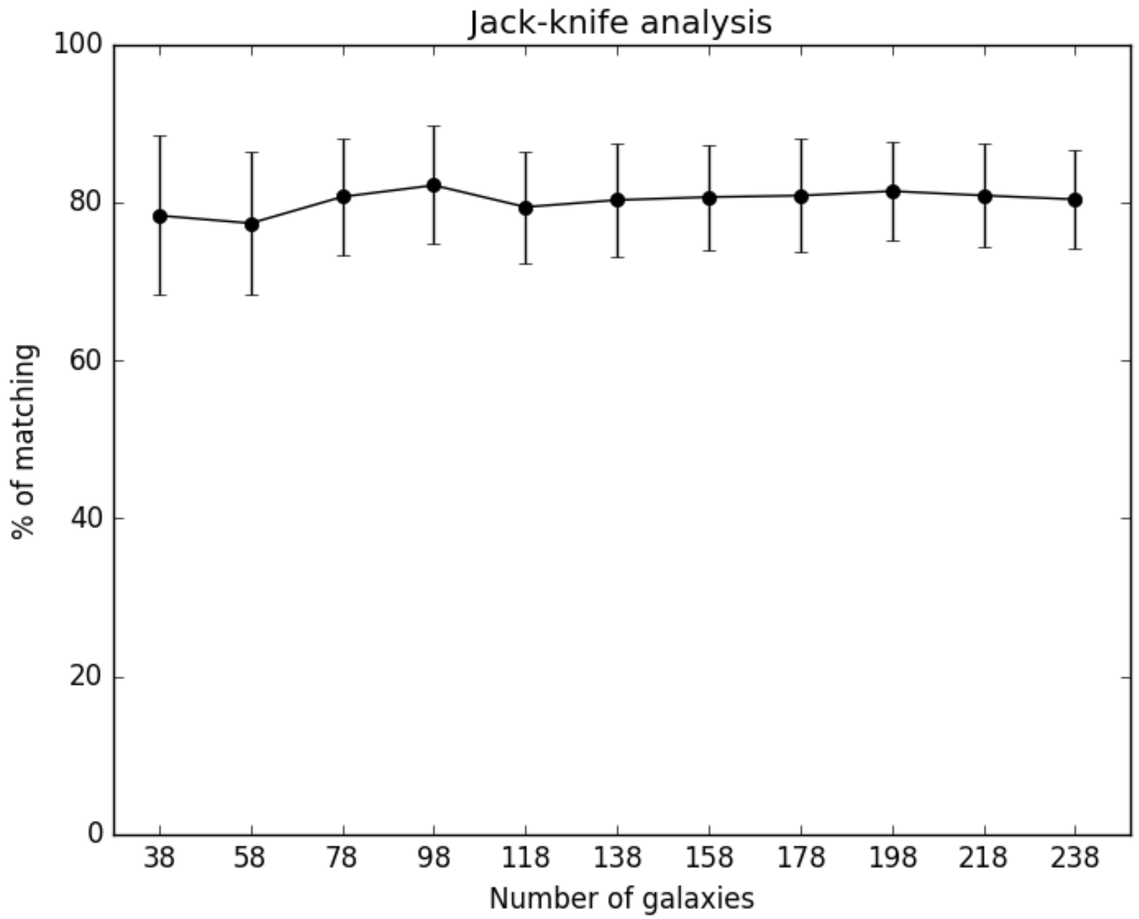
credit: Kalinova et al., 2017b

Classification of velocity curves: cluster category



credit: Kalinova et al., 2017b

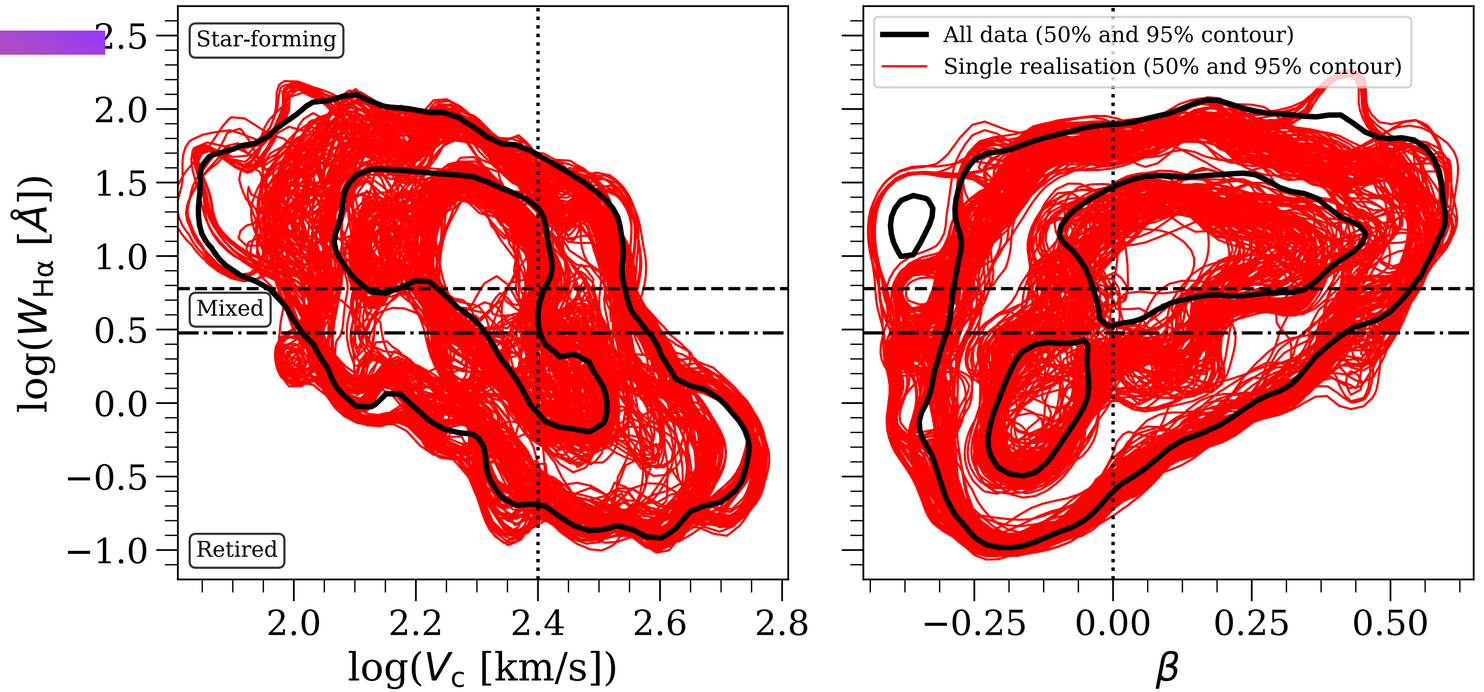
Cross-validation techniques: Jack-knife



credit: Kalinova et al., 2017b

- important to verify the ML results through modifying our sample
- Jack-knife: change the sample (e.g. 20 galaxies for each iteration), but not the sample distribution

Cross-validation techniques: Boot



credit: Kalinova et al., 2022

- It is important to verify the ML results through modification of the sample
- Bootstrap: modify the distribution of the sample

Conclusion

- Principal Component Analysis (PCA) is powerful machine learning technique for dimensionality reduction tests of large datasets
- Can be use for classification, regressions, decomposing, and hierarchical structure tasks
- Limitations: needs representative samples to find common orthogonal base (otherwise, it might lead to biases); uses linear base

credit: Kalinova et al., 2017b

Conclusion

- Principal Component Analysis (PCA) is a powerful machine learning technique for dimensionality reduction tasks of large datasets
- Can be use for classification, regression, decomposing, and hierarchical structure tasks
- Limitations: needs representative samples to find common orthogonal base (otherwise, it might lead to biases); uses linear base

THANK YOU!

credit: Kalinova et al., 2017b