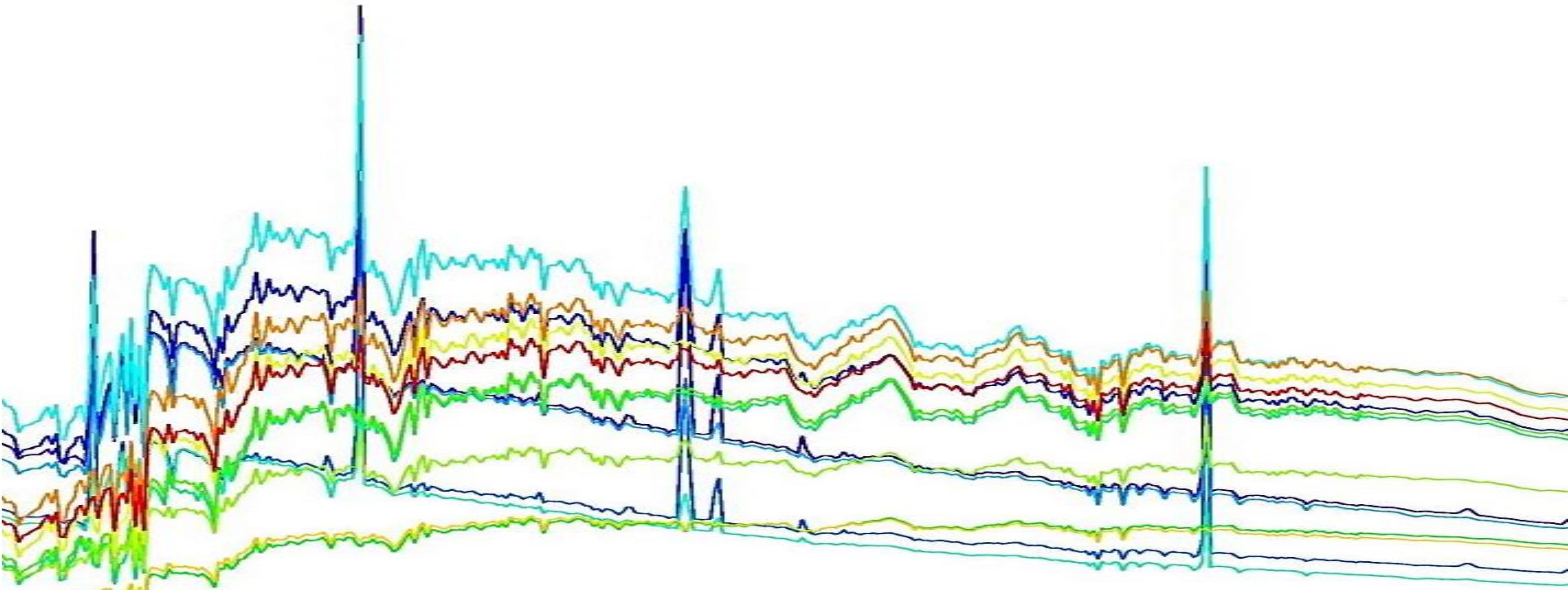
A. Karampelas, E. Kontizas, M. Kontizas

Unsupervised spectral classification of synthetic galaxies using Principal Components Analysis

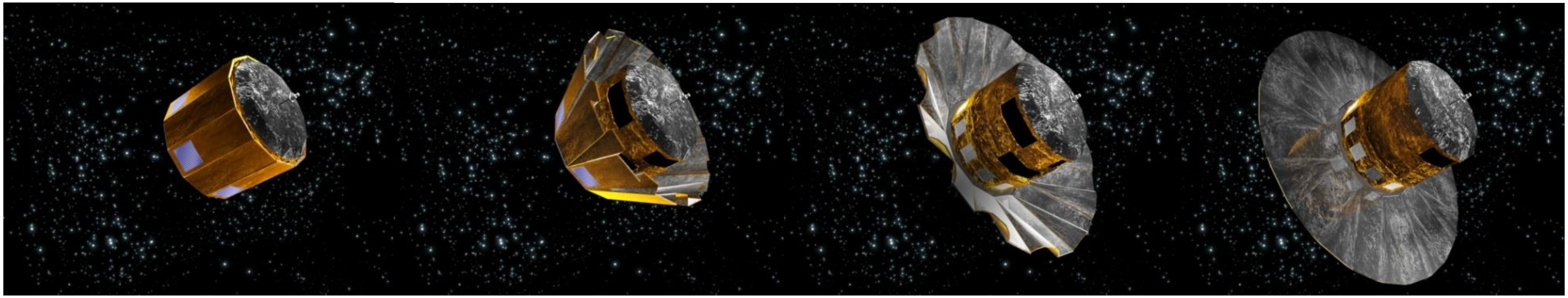


The 11th Hellenic Astronomical Conference, Athens, September 11, 2013

ESA's Gaia Mission (2013-2018)

Launch: Nov 17 – Dec 5 2013

<u>in less than 3 months!</u>



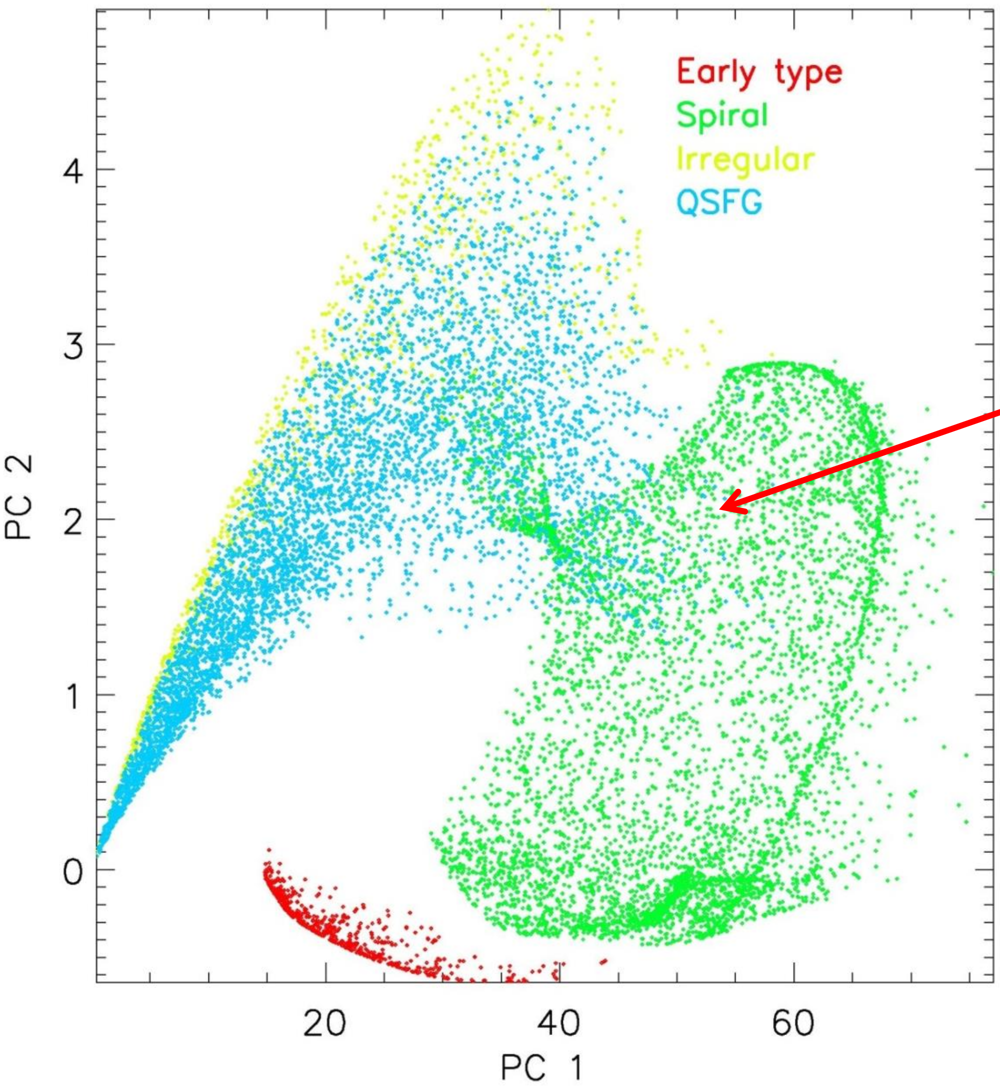
<u>Main Goal</u>: Construct the biggest and the most precise 3D stellar map of the Milky Way ever made

<u>Observables:</u> 1 billion galactic and extragalactic objects up to V = 20-25 mag (Jordi et al. 2006)

Stars / **Galaxies** / Quasars / Solar system objects / Exoplanets / SN

<u>Our task</u>: Classify and parameterize galaxy spectra (UGC)

Galaxy classification using **PCA**



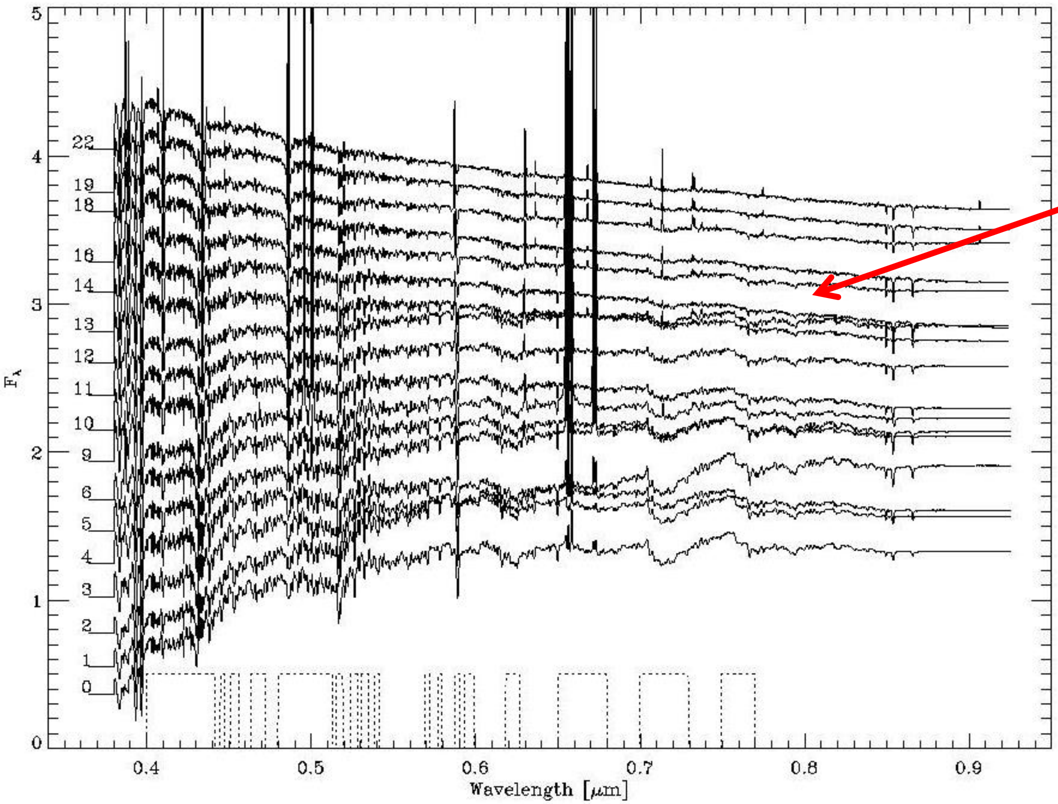
Optimization of synthetic galaxy spectra for the ESA's Gaia Mission
(Karampelas et al. 2012, A&A)

<u>Similar work</u>

Folkes et al. 1999, MNRAS
Madgwick et al. 2003, ApJ
Yip et al. 2004, AJ
Steiner et al. 2009, MNRAS
Wang et al. 2011, MNRAS

Galaxy classification using **k-means clustering**

ASK (Automatic Spectroscopic K-means-based) classification



99% of the SDSS/DR7 galaxy spectra can be assigned to only 17 major classes

(Sánchez-Almeida et al. 2010, AJ)

<u>Similar work</u>

Sánchez Almeida et al. 2011, AJ

Ascasibar & Sánchez Almeida et al. 2011, MNRAS

Sánchez Almeida et al. 2012, AJ

Aguerri et al. 2012, A&A

Galaxy classification

PCA is a very successful classification method, but it does not provide prototypical spectra, like **ASK** does.

Possible solution

Combination of **PCA** and **DT** (Decision Tree)

(Karampelas et al. in prep.)

A similar approach: Wang et al. 2011, MNRAS

Data: synthetic galaxy spectra

(Livanou et al. in prep.)

- ✓ Produced with PÈGASE.2
  (Fioc & Rocca-Volmerange 1997, 1999; Le Borgne & Rocca-Volmerange 2002)
- ✓ Four spectral types – Early-type / Spiral / Irregular / QSFG
- ✓ Various redshifts (z: 0.0 – 0.6)

Used here

A subset of 7,160 optimal (Karampelas et al. 2012) $z = 0$ synthetic spectra

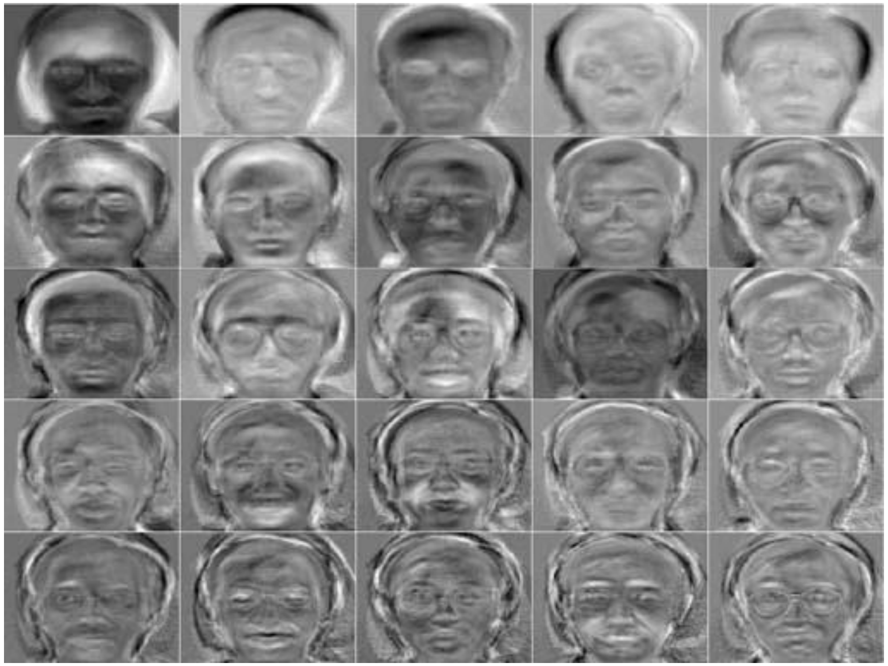
Principal Components Analysis (PCA)
(Karhounen-Loeve transformation)

- Linear orthogonal transformation in a new base, in which the data variance is highlighted.

- New axes = Principal Components (PCs)

- Very effective in: **Data compression, Dimensionality reduction, Noise extraction**

- Applications: Astronomy, Biology, Graphology, Face and Fingerprint Recognition



Principal Components Analysis (PCA)

Implementation

1. Construction of the variance-covariance matrix
2. Determination of eigenvalues and eigenvectors of the matrix.
3. Eigenvectors = new axes = PCs
4. Eigenvalue sorting in descending order λ_1, λ_2, …
5. PC1 corresponds to λ_1, PC2 to λ_2 etc.
    - PC1: Summarizes the majority of the data variance
    - PC2: Summarizes the majority of the rest of the data variance etc.

Full spectrum reconstruction

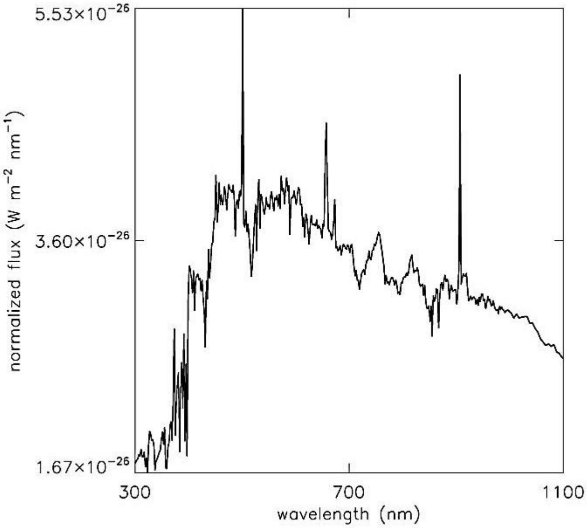
$$\text{Spectrum} = \alpha_1 PC1 + \alpha_2 PC2 + \ldots + \alpha_k PCk$$

Partial reconstruction is usually sufficient:

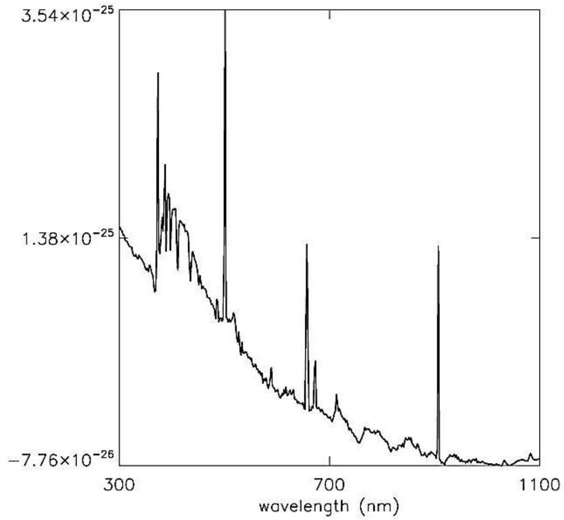
$$\text{Spectrum} \approx \alpha_1 PC1 + \alpha_2 PC2 + \ldots + \alpha_5 PC5$$

Principal Components (PCs) and % of the total variance

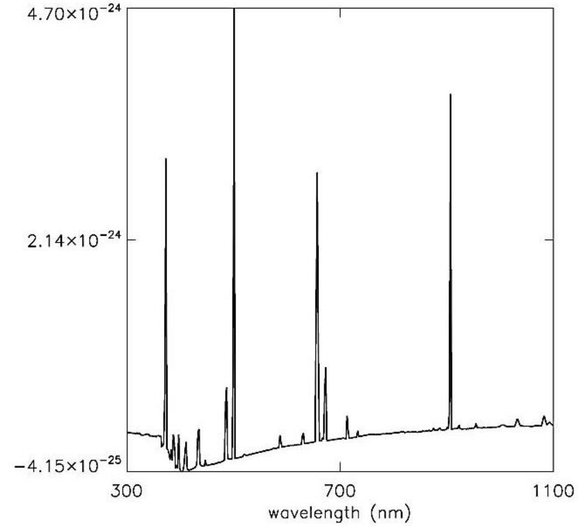
PC1: 84.993%



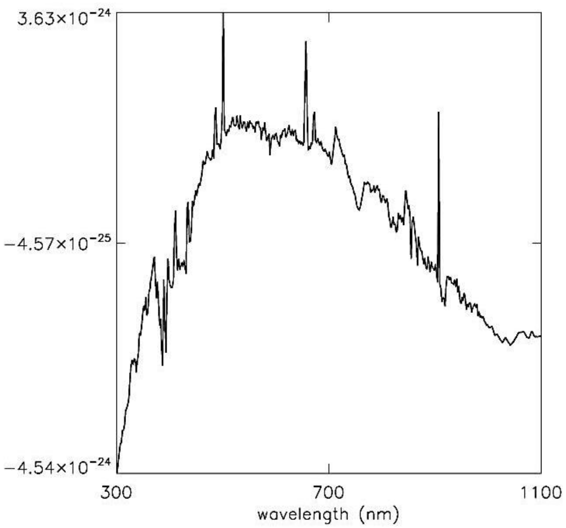
PC2: 14.277%



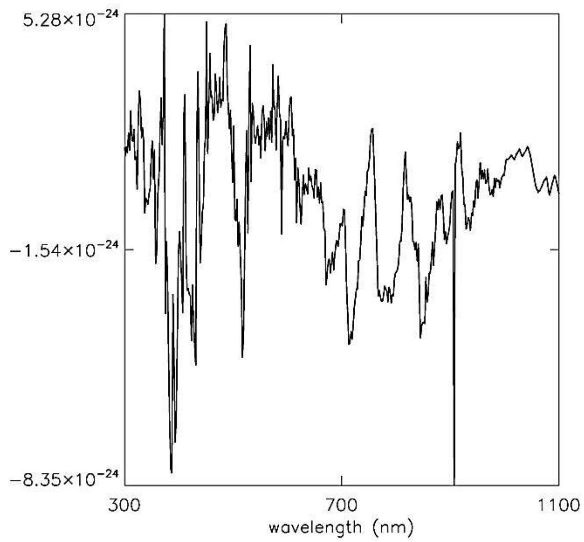
PC3: 0.664%



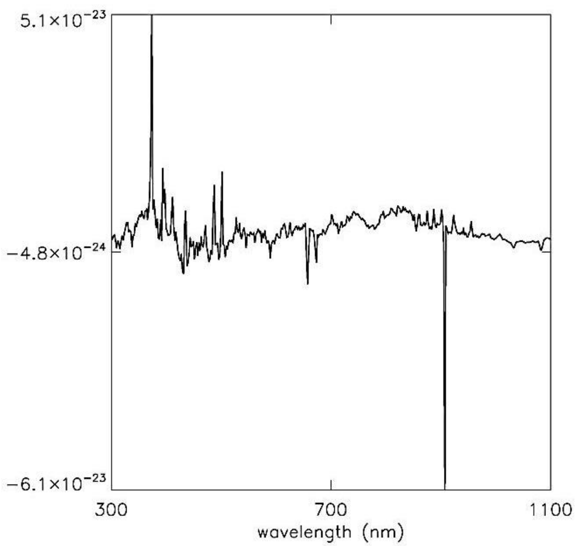
PC4: 0.038%



PC5: 0.021%



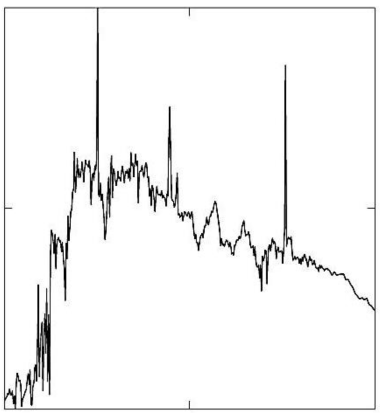
PC6: 0.004%



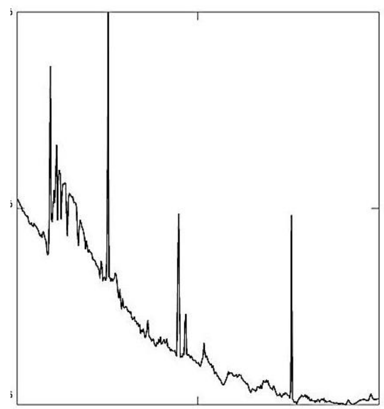
Reconstruction

$$\text{Spectrum} = \alpha_1 \text{PC1} + \alpha_2 \text{PC2} + \alpha_3 \text{PC3} + \ldots =$$

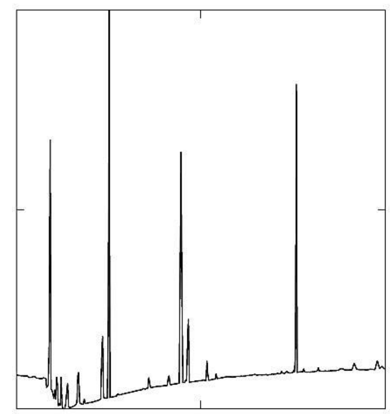
$= \alpha_1$



$+ \alpha_2$



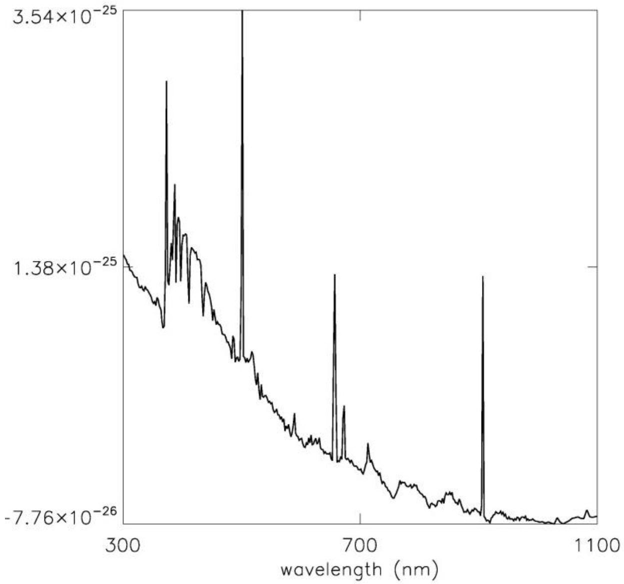
$+ \alpha_3$



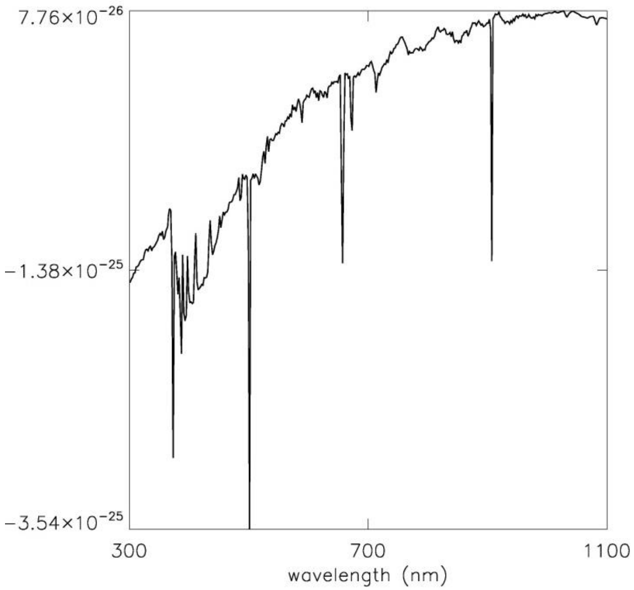
$+ \ldots$

Sign of the admixture coefficients

$1 \cdot PC2 =$

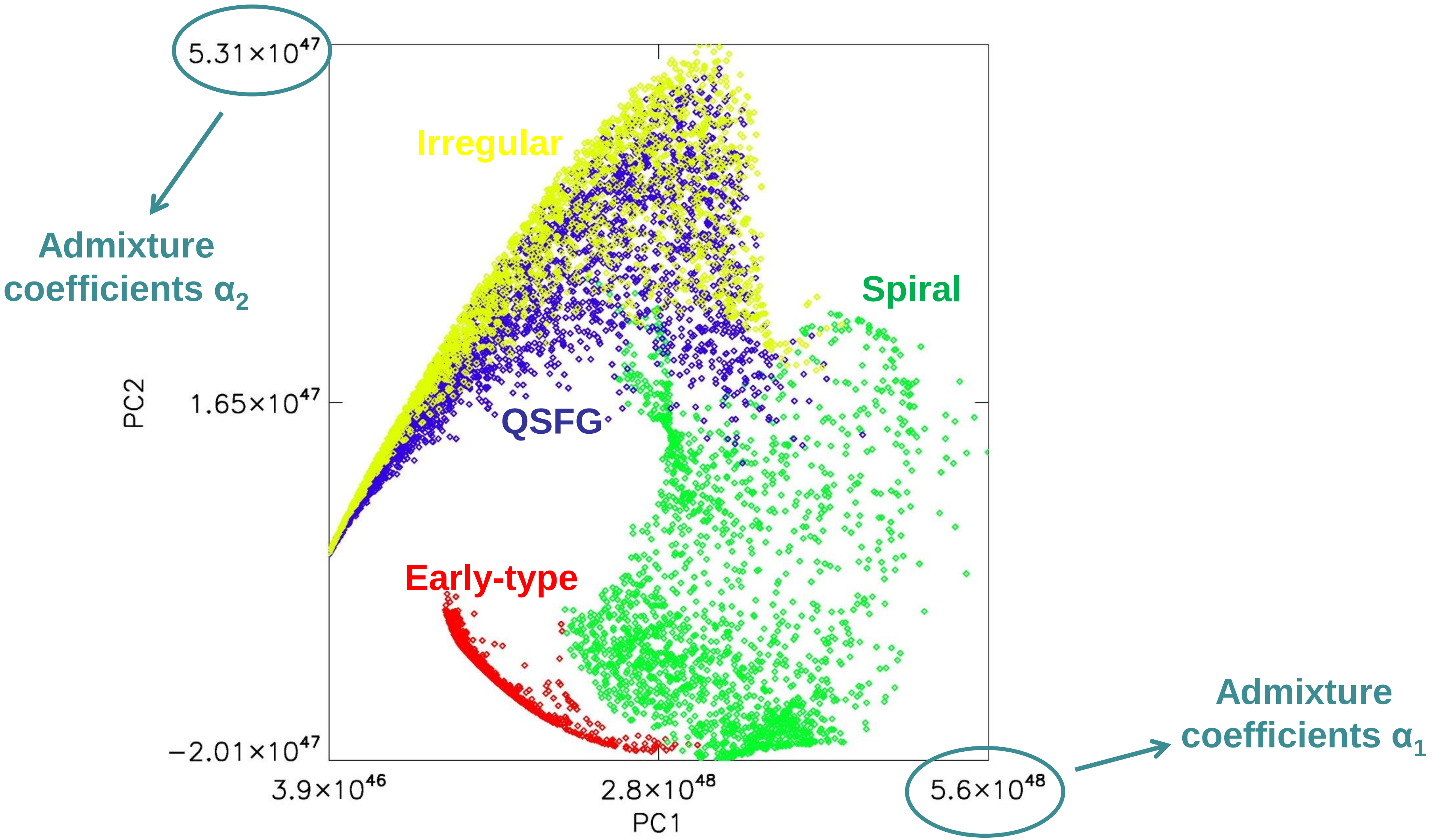


$-1 \cdot PC2 =$

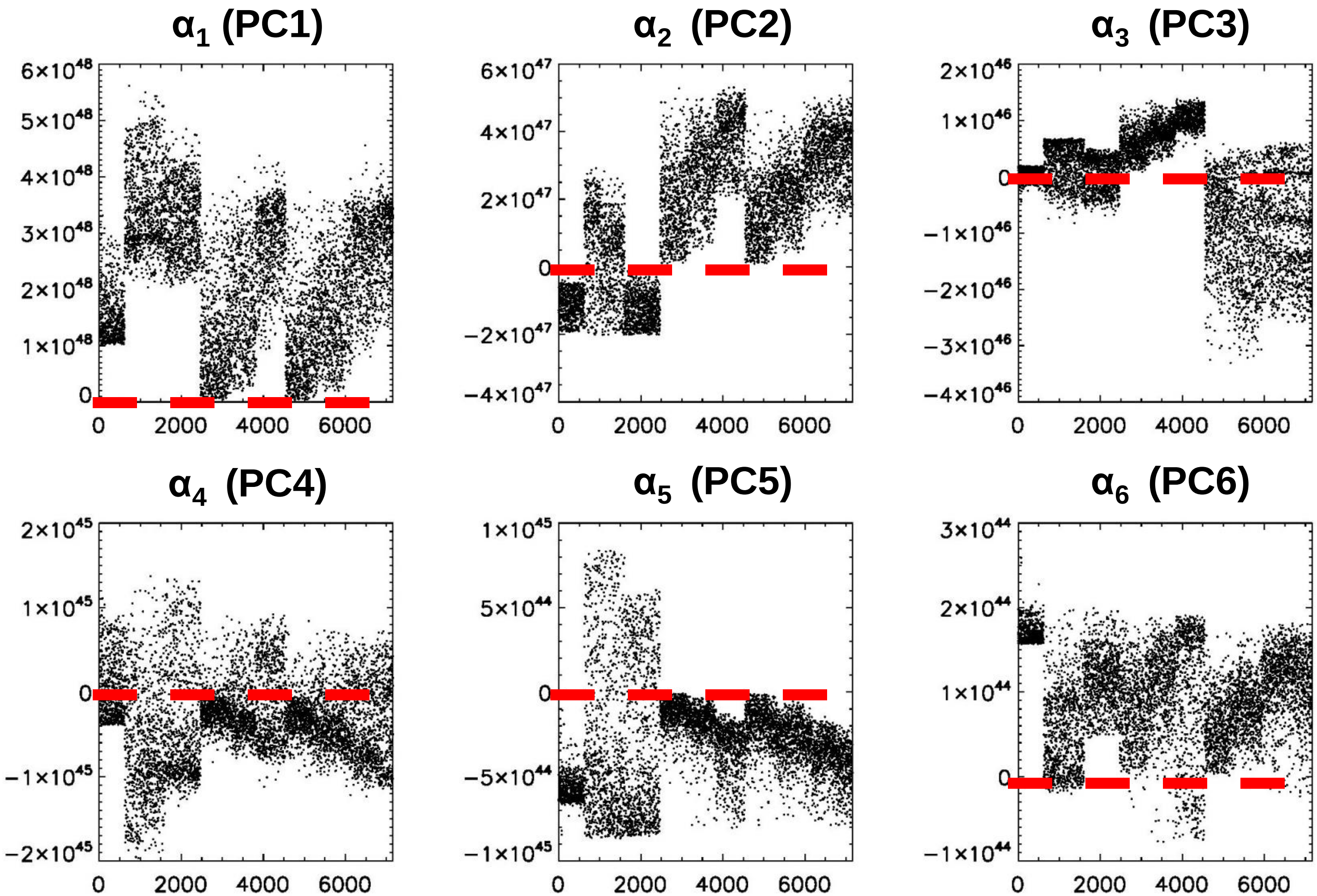


Different sign = different kind of contribution of each PC to the galaxy spectrum

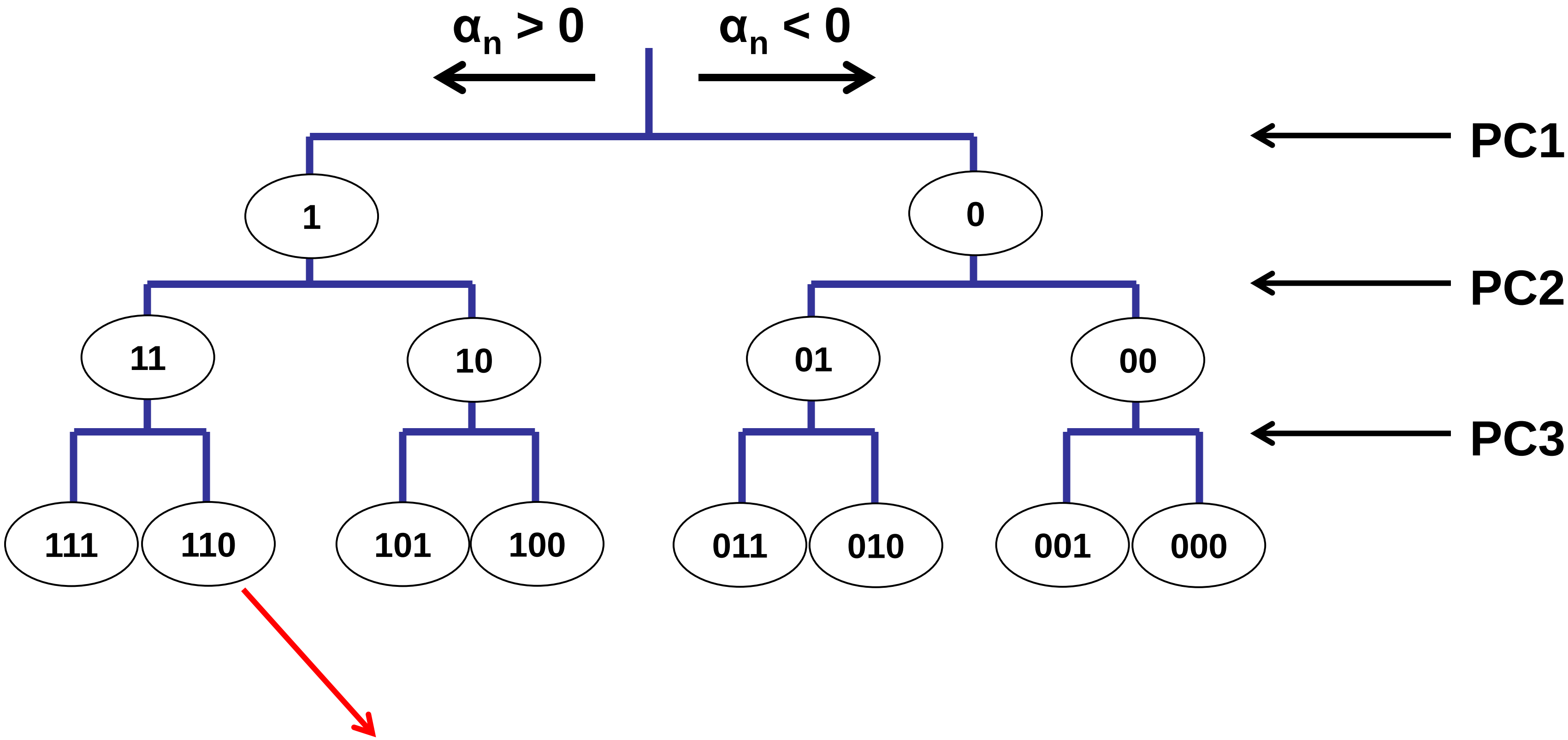
Projection of spectra to PC1/PC2 axes



Admixture coefficients

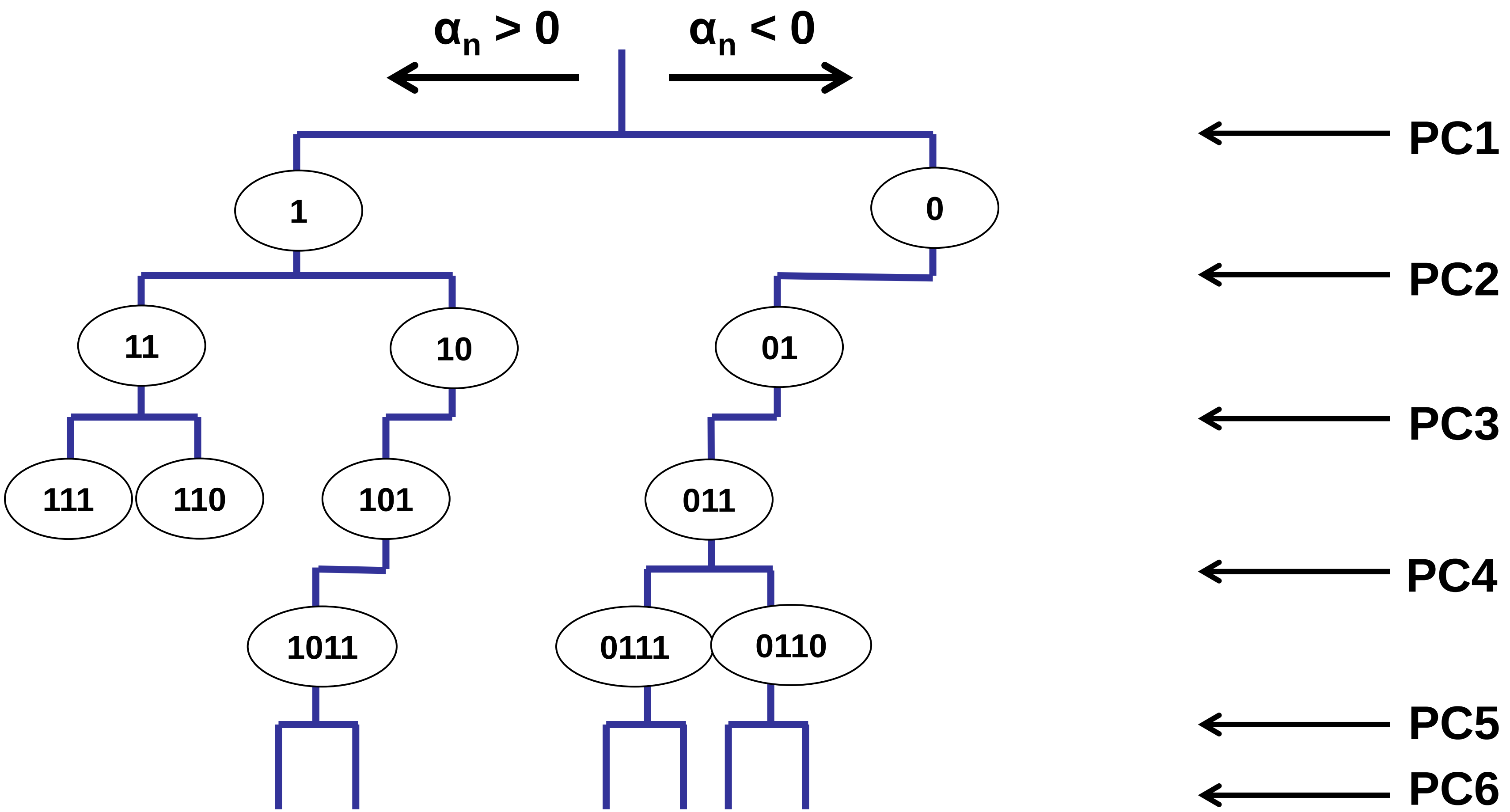


Decision Tree



110: spectra with PC1 > 0, PC2 > 0, PC3 < 0

Decision Tree (an example)



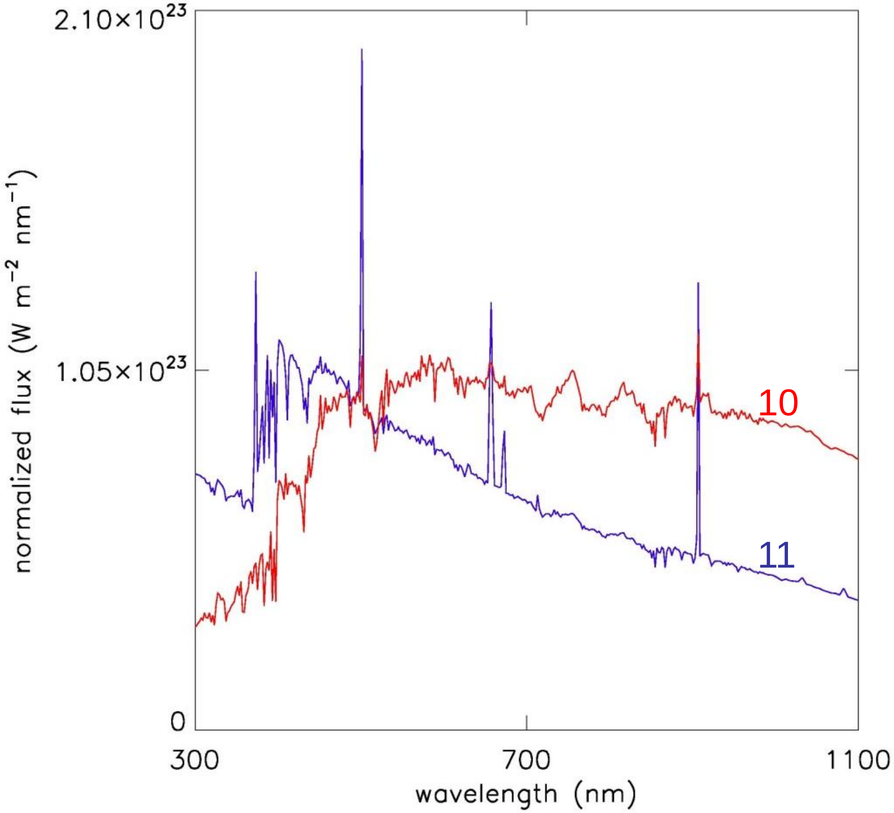
Finalizing the Decision Tree

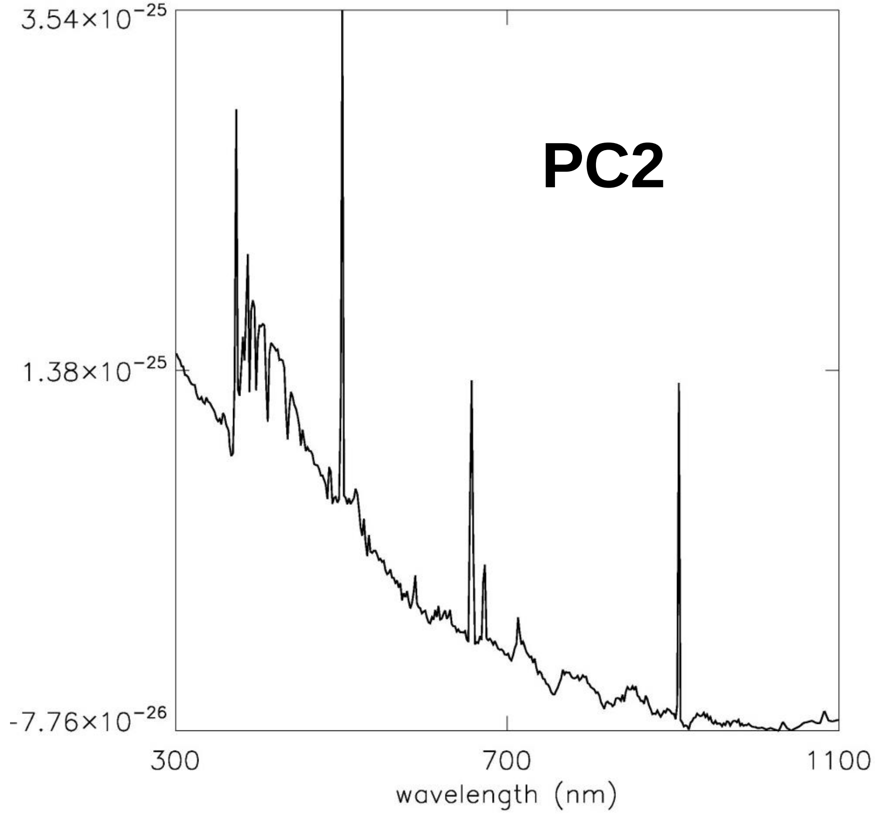
Possible ways to finalize a part of the tree

- When a class is poorly divided after 2-3 partitions
- When the spectra of a class are similar (PCA, χ^2, K-S test)
- When the corresponding PC_n represents noise

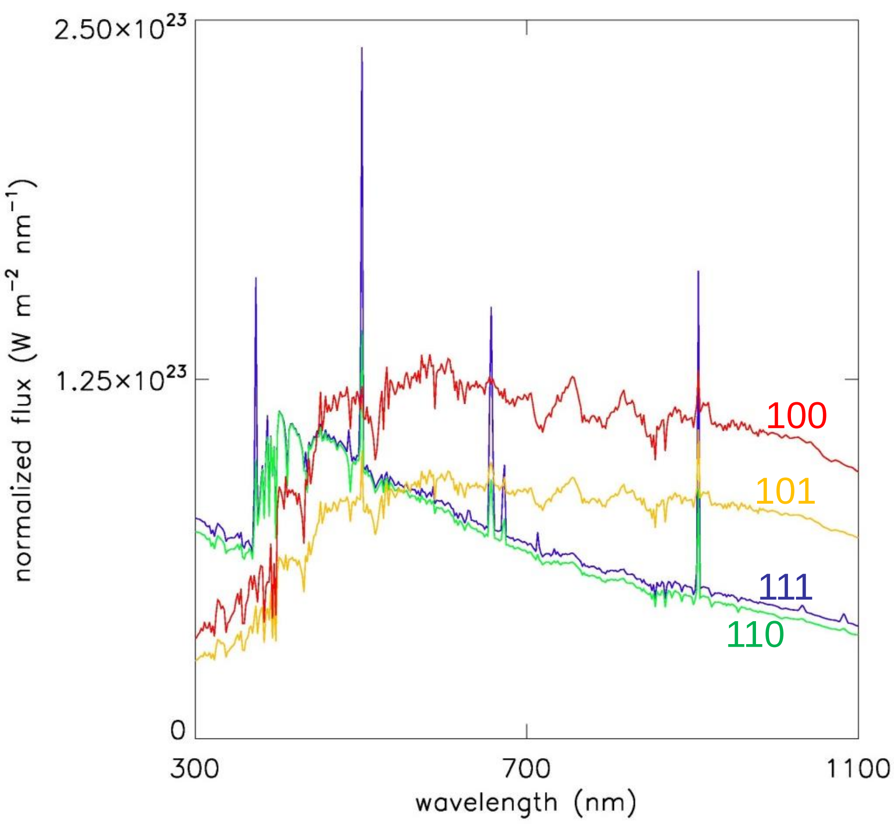
Large tree: may overfit the data
Small tree: may not capture the important structure

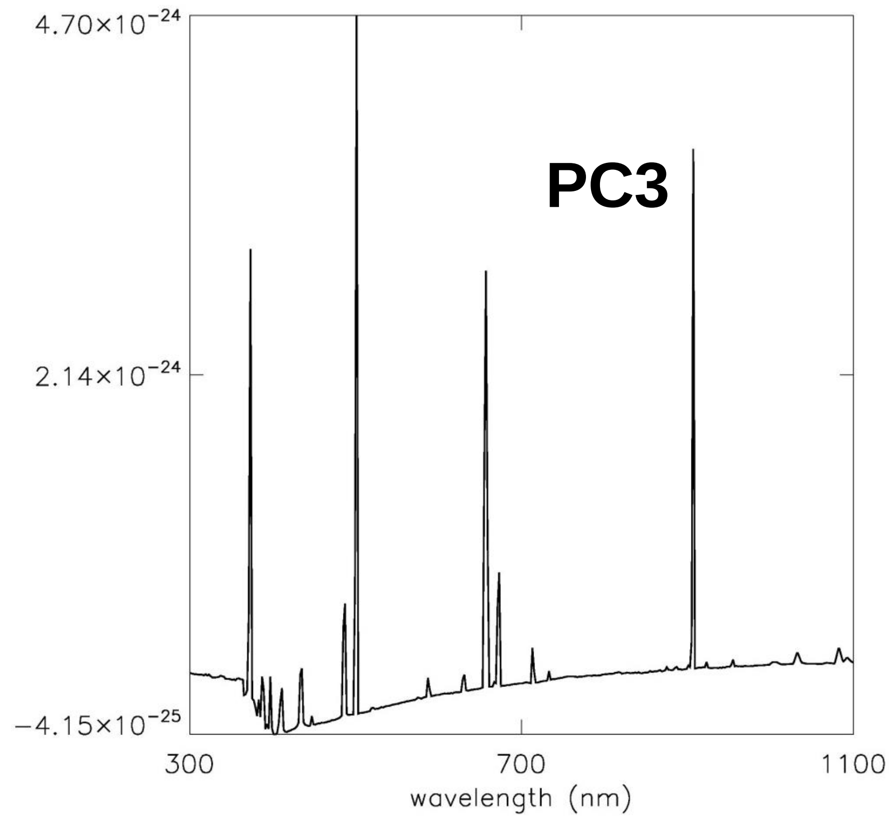
Mean spectra of subtypes determined from the sign of PC2 coefficients



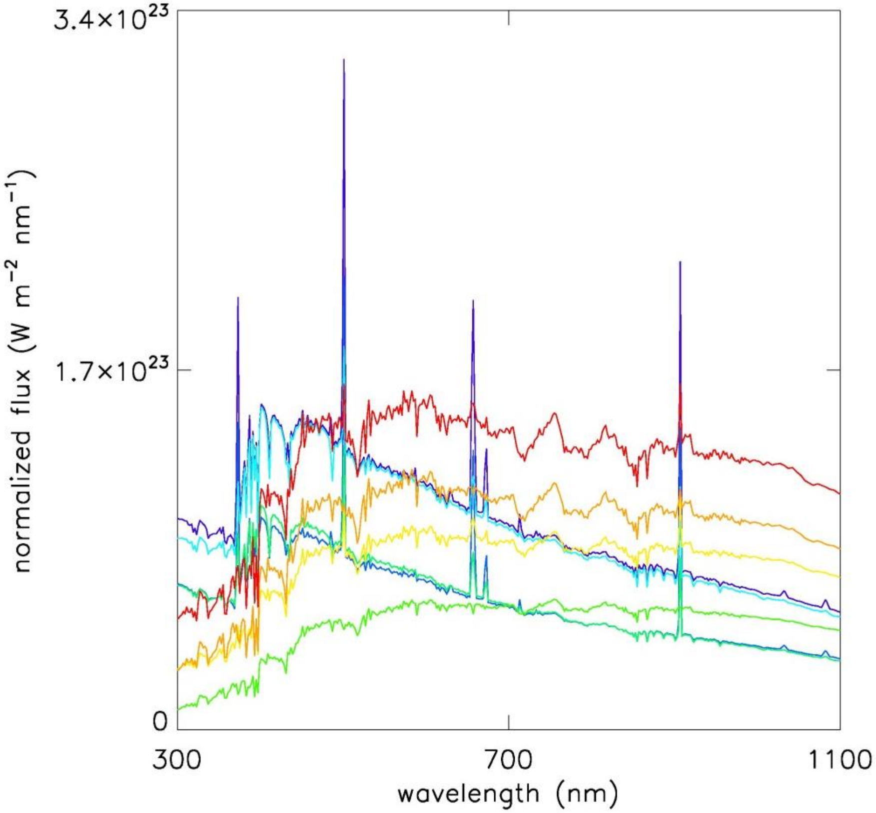


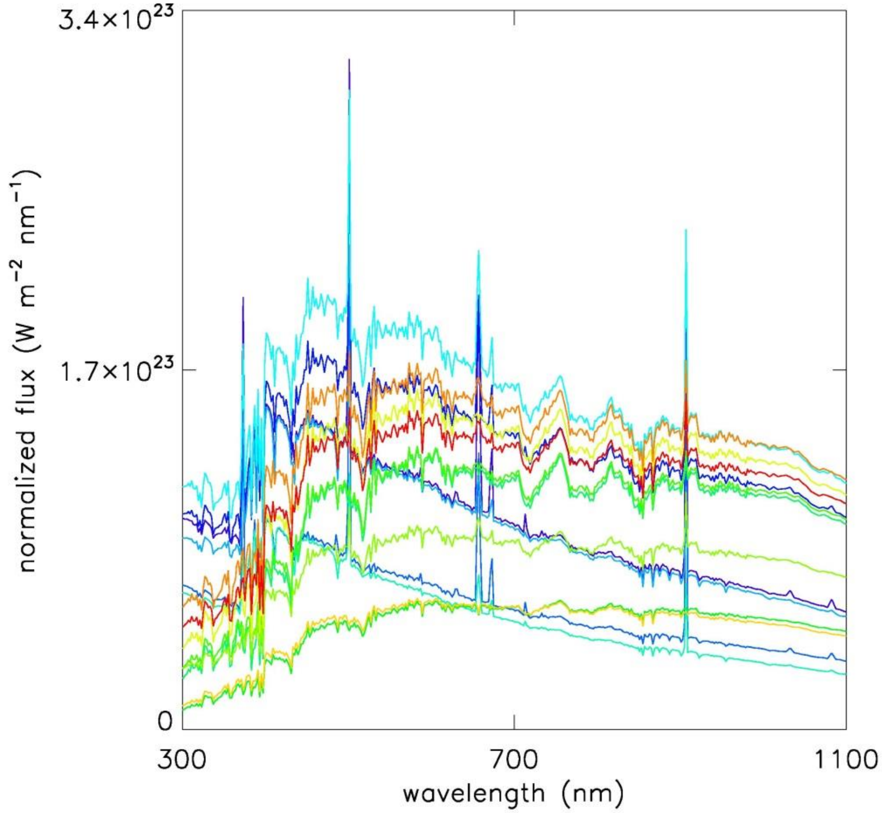
Mean spectra of subtypes determined from the sign of PC3 coefficients



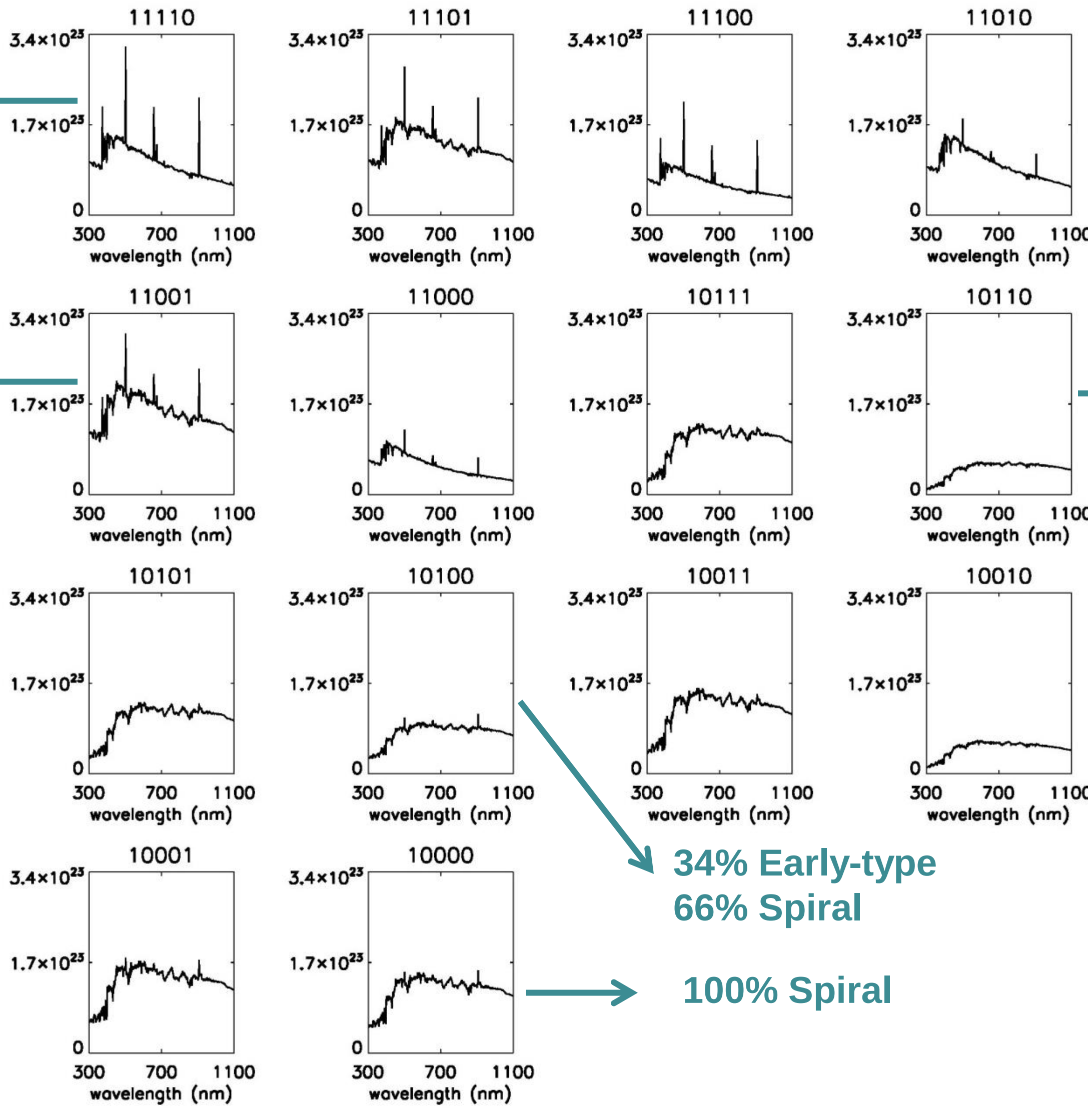


Mean spectra of subtypes determined from the sign of PC4 and PC5 coefficients





Mean spectra of subtypes determined from the sign of PC5 coefficients



17% Spiral
67% Irregular
16% QSFG

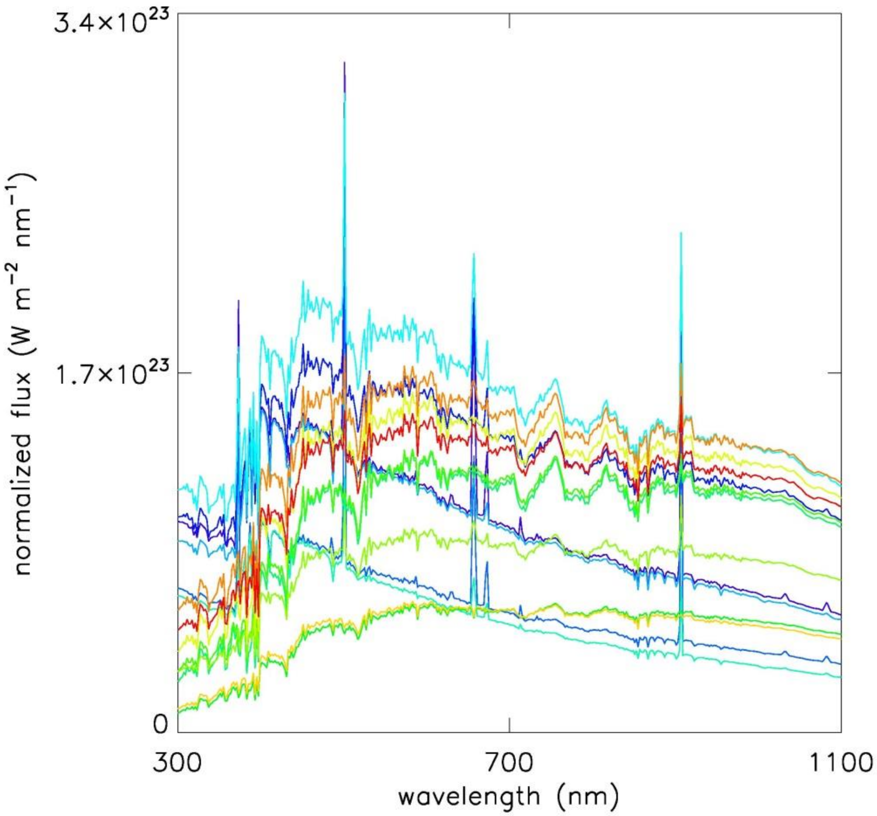
100% Spiral

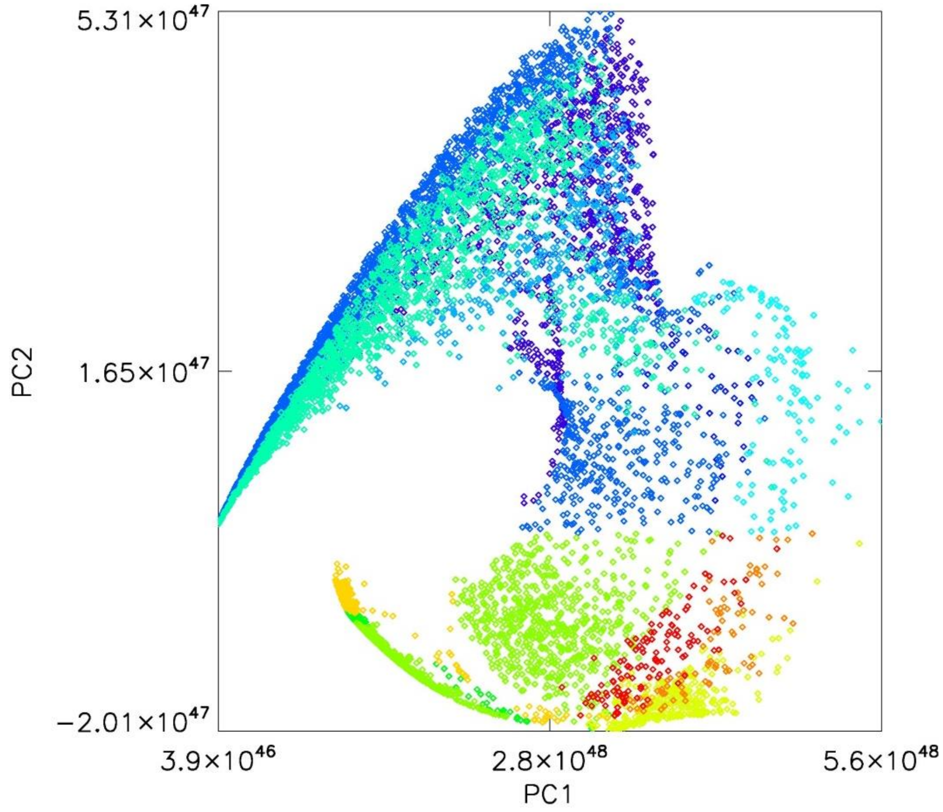
100% Early-type

34% Early-type
66% Spiral

100% Spiral

Projection of subtypes to PC1/PC2 axes





Summary

- We propose a PCA/DT spectral classification method.
- The method is currently implemented on synthetic galaxy spectra for ESA's Gaia Mission.
- The spectra seem to be divided into unique subtypes.

Future Plans

- Fine-tune the method and compare with ASK.
- Implementation to the major space and ground-based surveys.