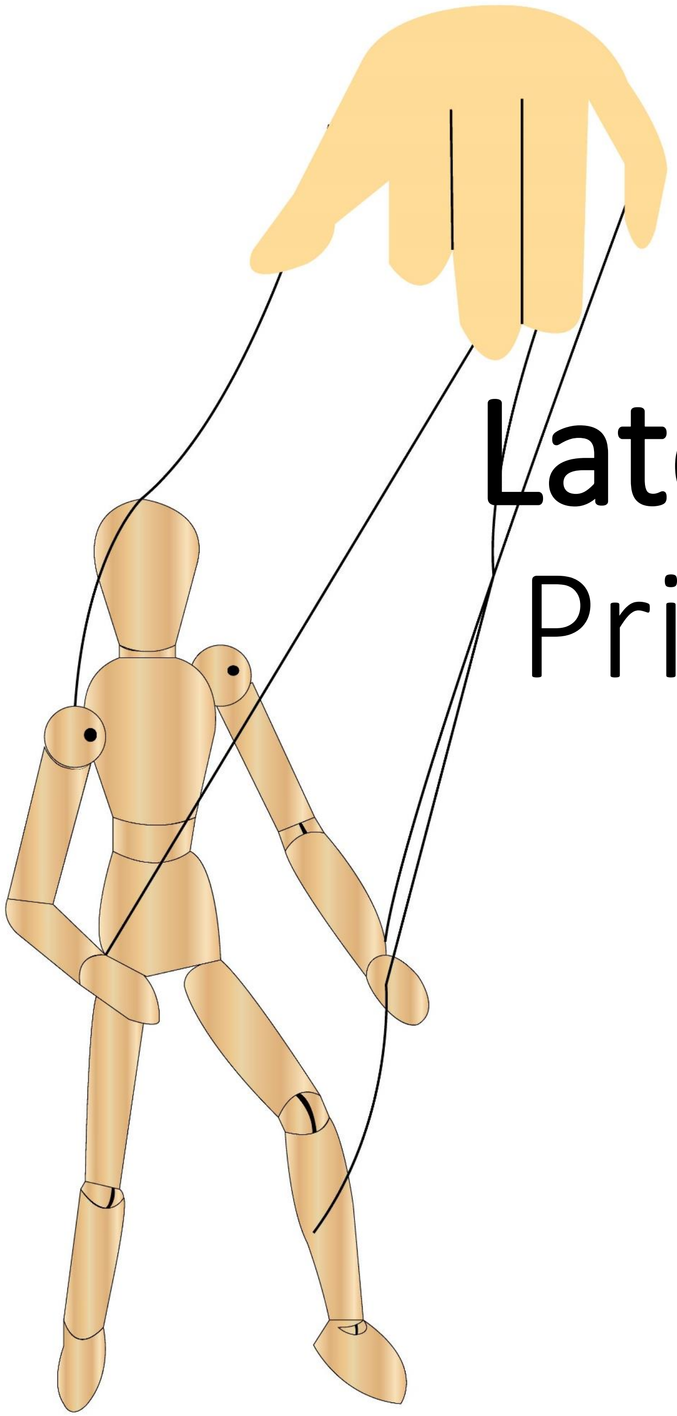


Latent Variable Models:
Principal Components

Biprateep Dey

University of Pittsburgh
Physics & Astronomy

Motivation: How can we describe wine?

Features:

State: Liquid → Useless!

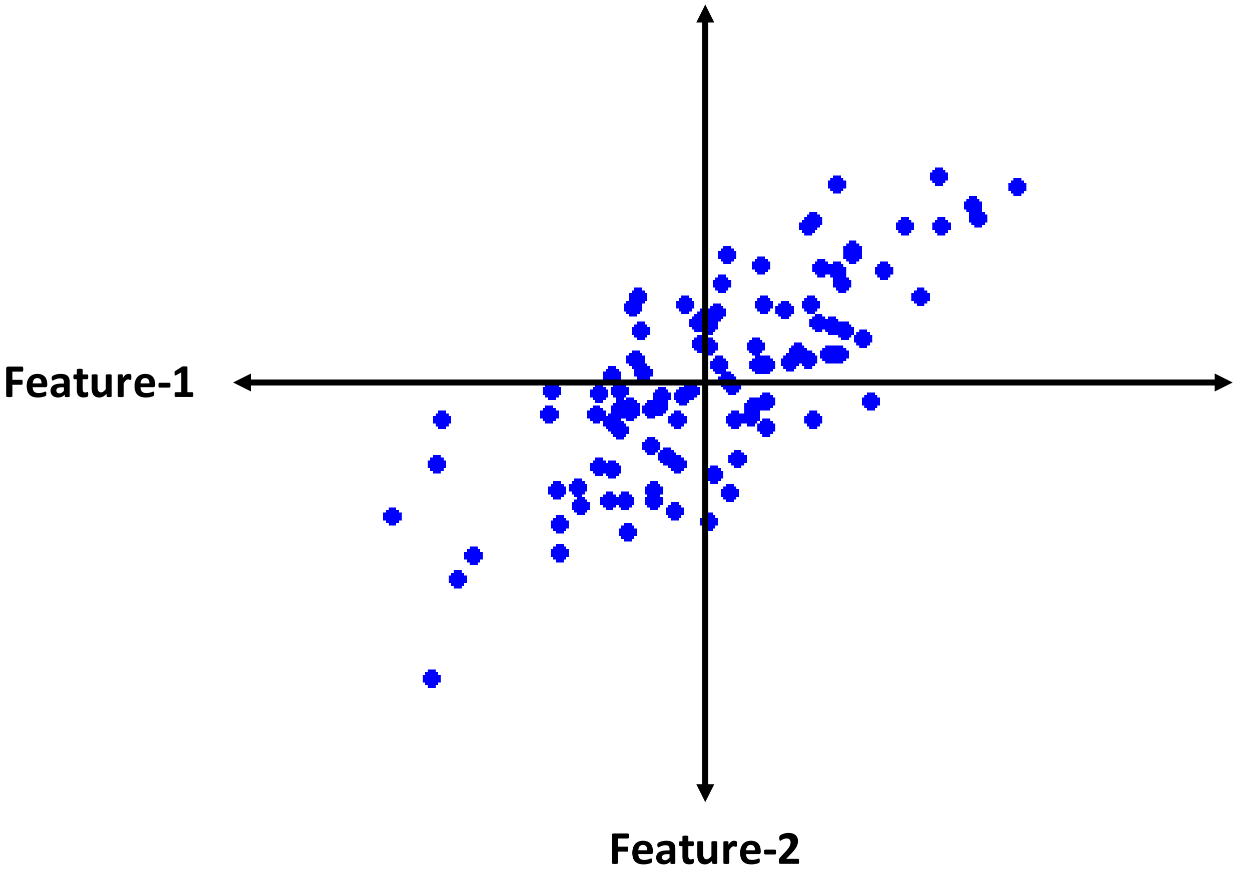
Color: Red vs White → Does not summarize well all the variations

Hue: Describes the variation → Not enough to reconstruct the original

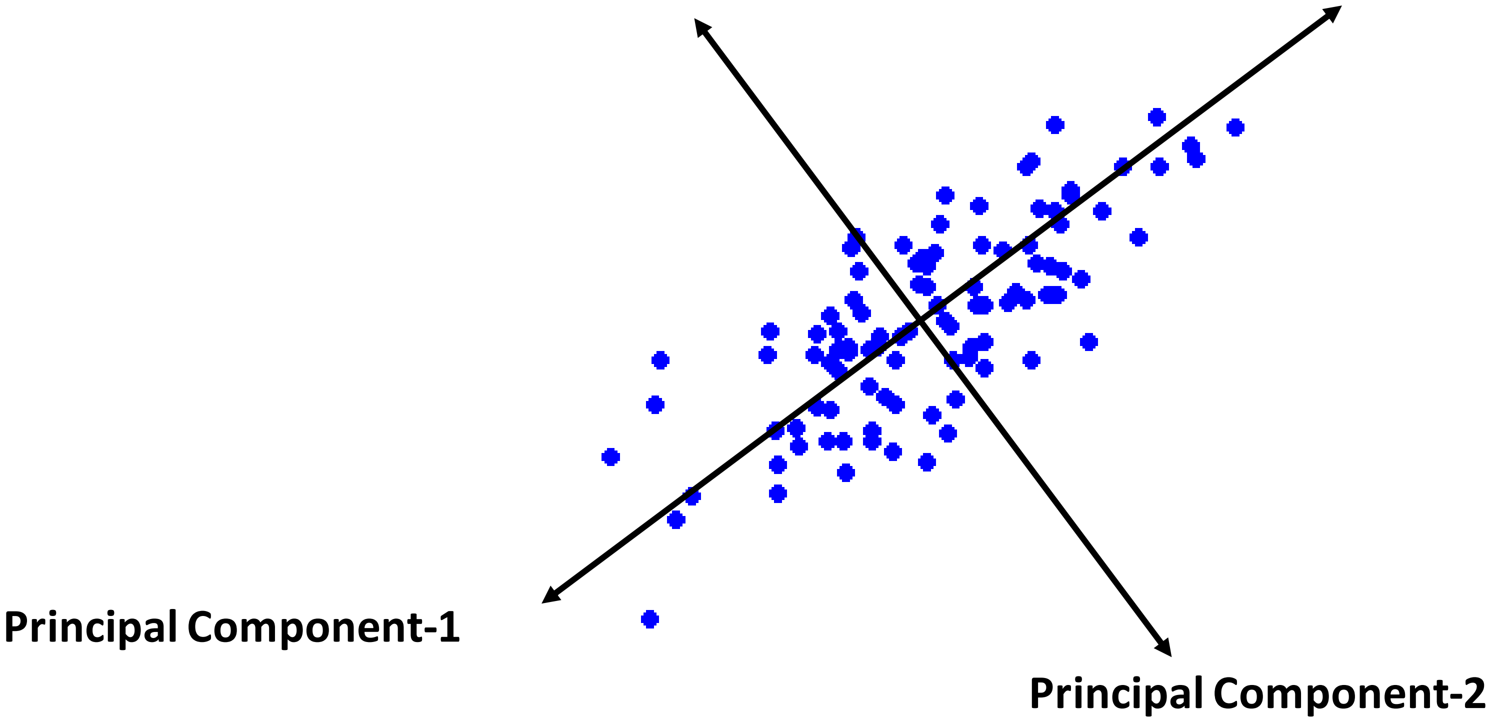
Alcohol Content, Grape Variety, …

A “Good” summary should
use feature that **represent most variation** &
efficiently reconstruct the input

Applying the same reasoning on data



Applying the same reasoning on data



Mathematically

$$\sigma^2(x,x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ a.k.a variance

Eigen vectors≡Principal Component Axes

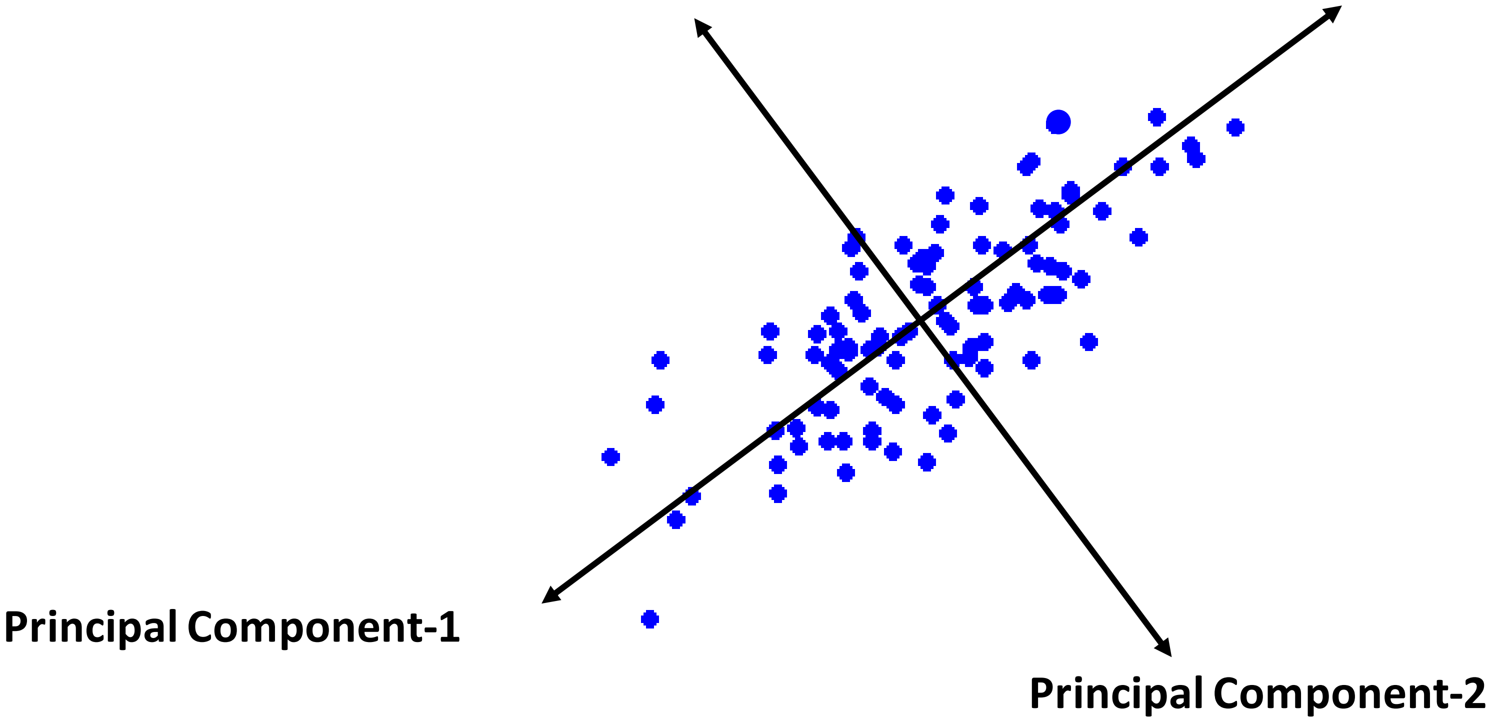
$$\begin{pmatrix} \sigma^2(x,x) & \sigma^2(x,y) \\ \sigma^2(y,x) & \sigma^2(y,y) \end{pmatrix} \xrightarrow{\text{Diagonalize}} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

$$\sigma^2(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$ a.k.a covariance

Eigen values≡Principal Components
(Arranged in decreasing order)

A “Good” summary should
use feature that represent most variation &
efficiently reconstruct the input

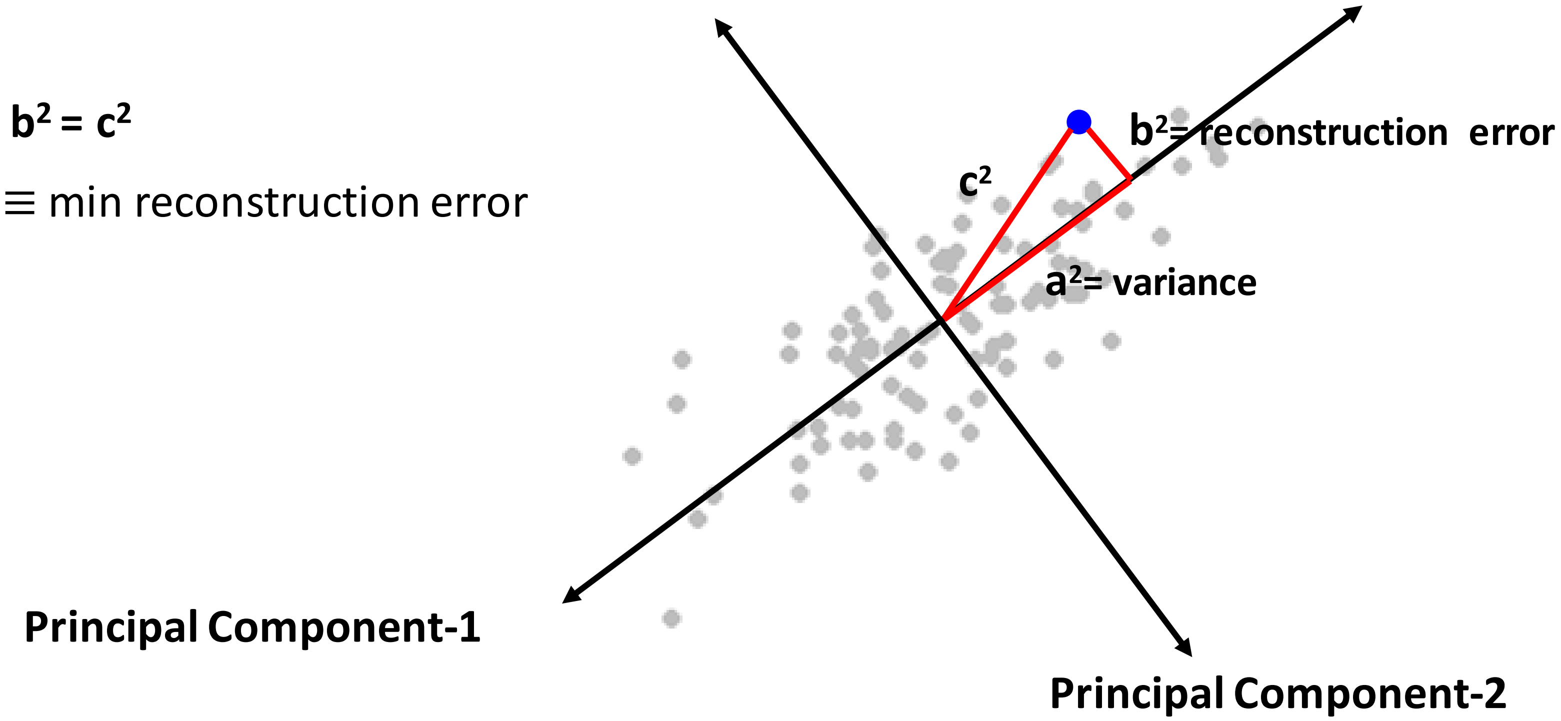
Principal Component Axes Minimize Reconstruction Error



Principal Component Axes Minimize Reconstruction Error

$a^2 + b^2 = c^2$

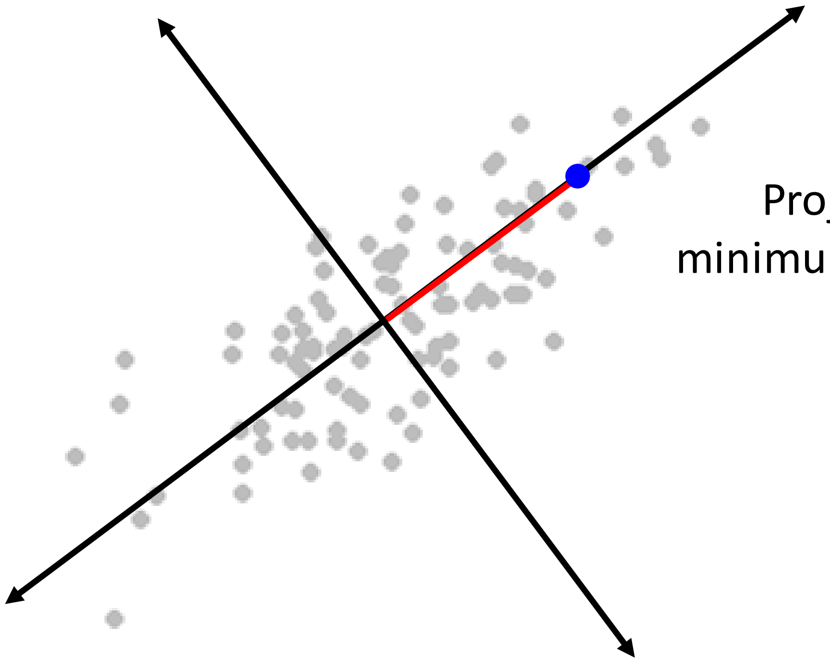
max variance ≡ min reconstruction error



Principal Component Axes Minimize Reconstruction Error

$\mathbf{a^2 + b^2 = c^2}$

max variance ≡ min reconstruction error



Projection on axis with minimum reconstruction error*

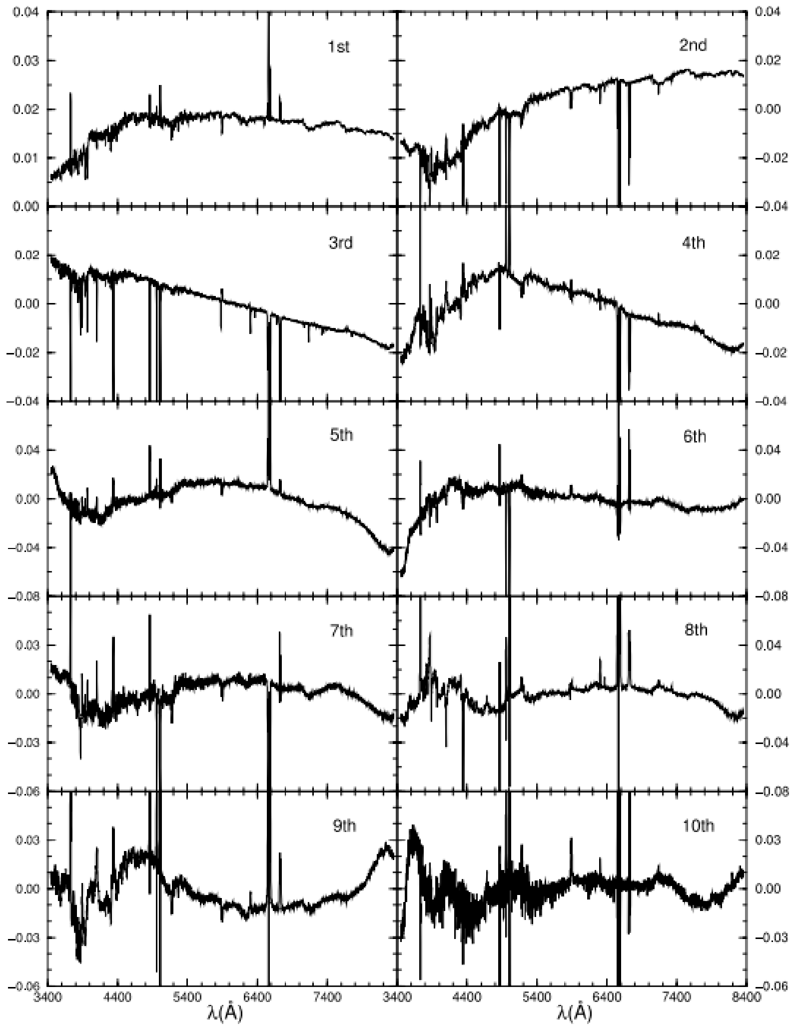
*In general we select the first *n* components which account for most of the variance

Choosing a subset of Principal Components allow us to **reduce dimensionality** (i.e. data compression)

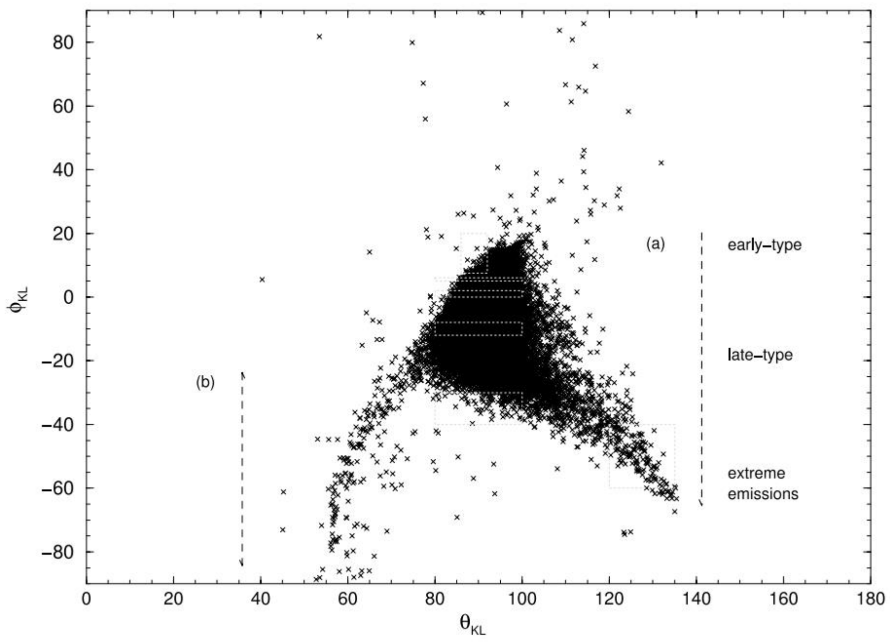
Principal Component Analysis (PCA)* in Astronomy

*a.k.a. Karhunen-Loève (KL) transform

Dimensionality Reduction of SDSS Galaxy Spectra (Yip et al. 2004)



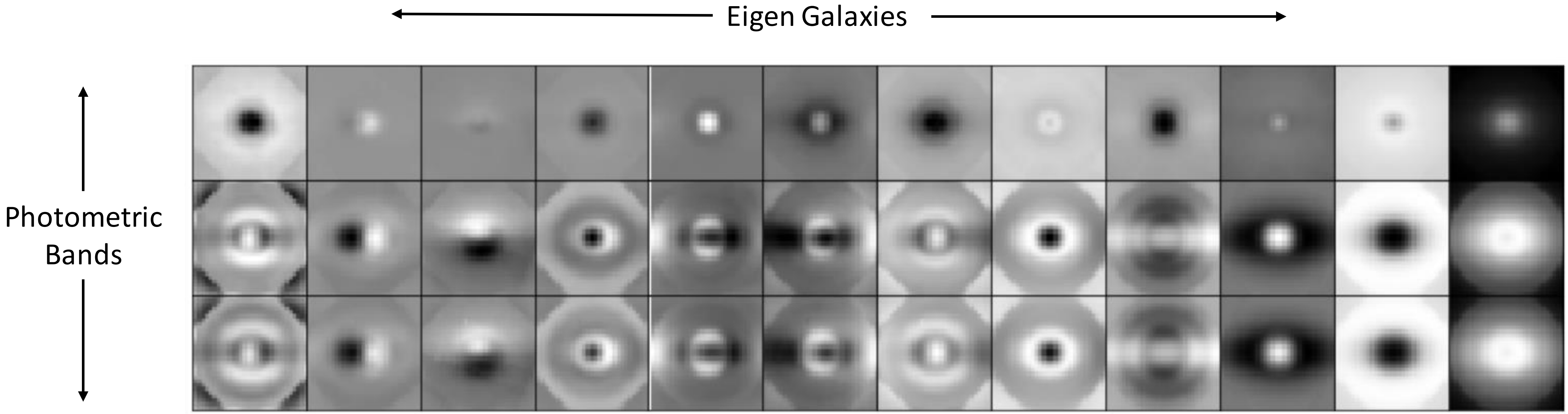
First 10
Eigen-spectra



2D plot separates galaxies
based on physical properties

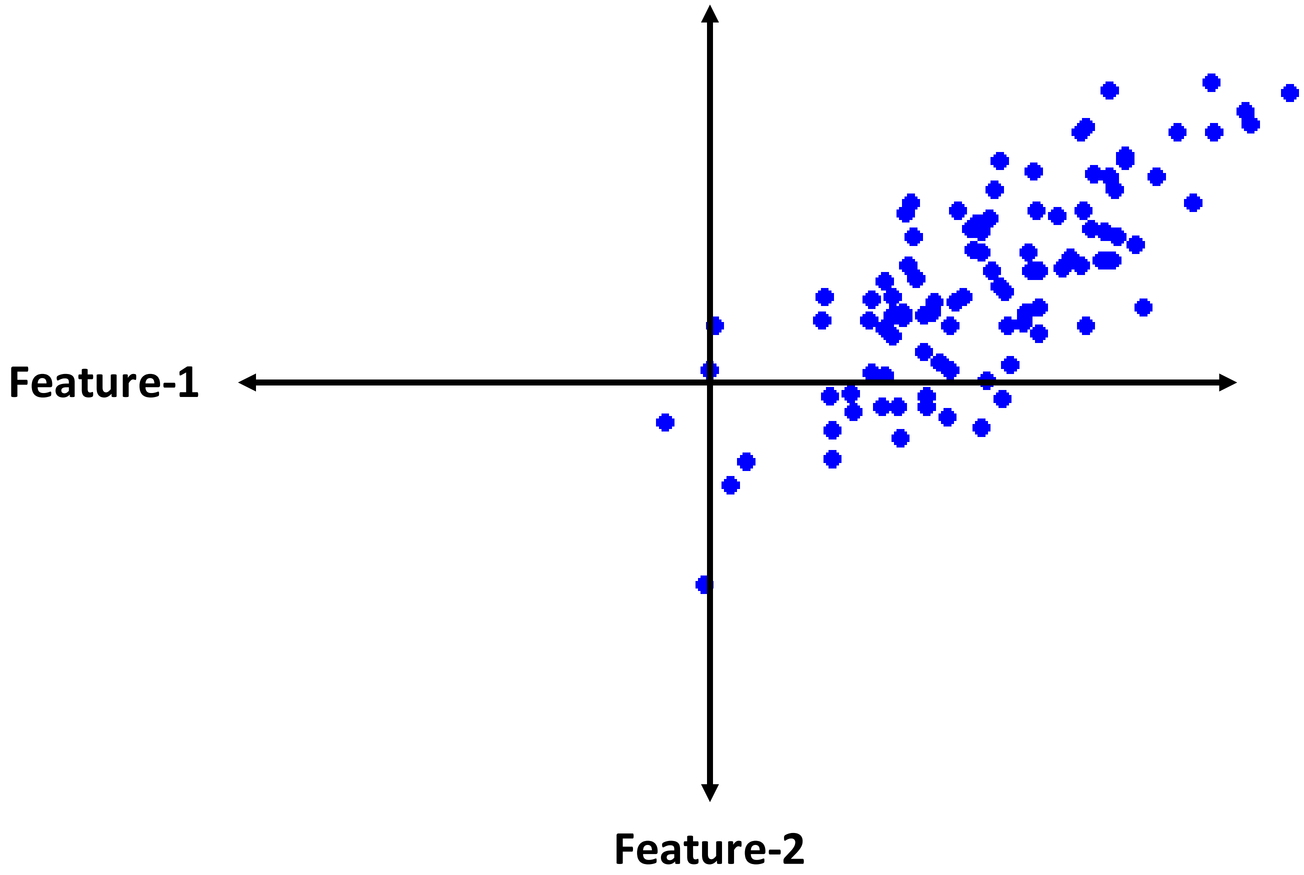
Applications: Templates for redshift fitting, explore distribution of objects, outliers detection, …

Principal Components of images (Uzeirbegovic et al. 2020)

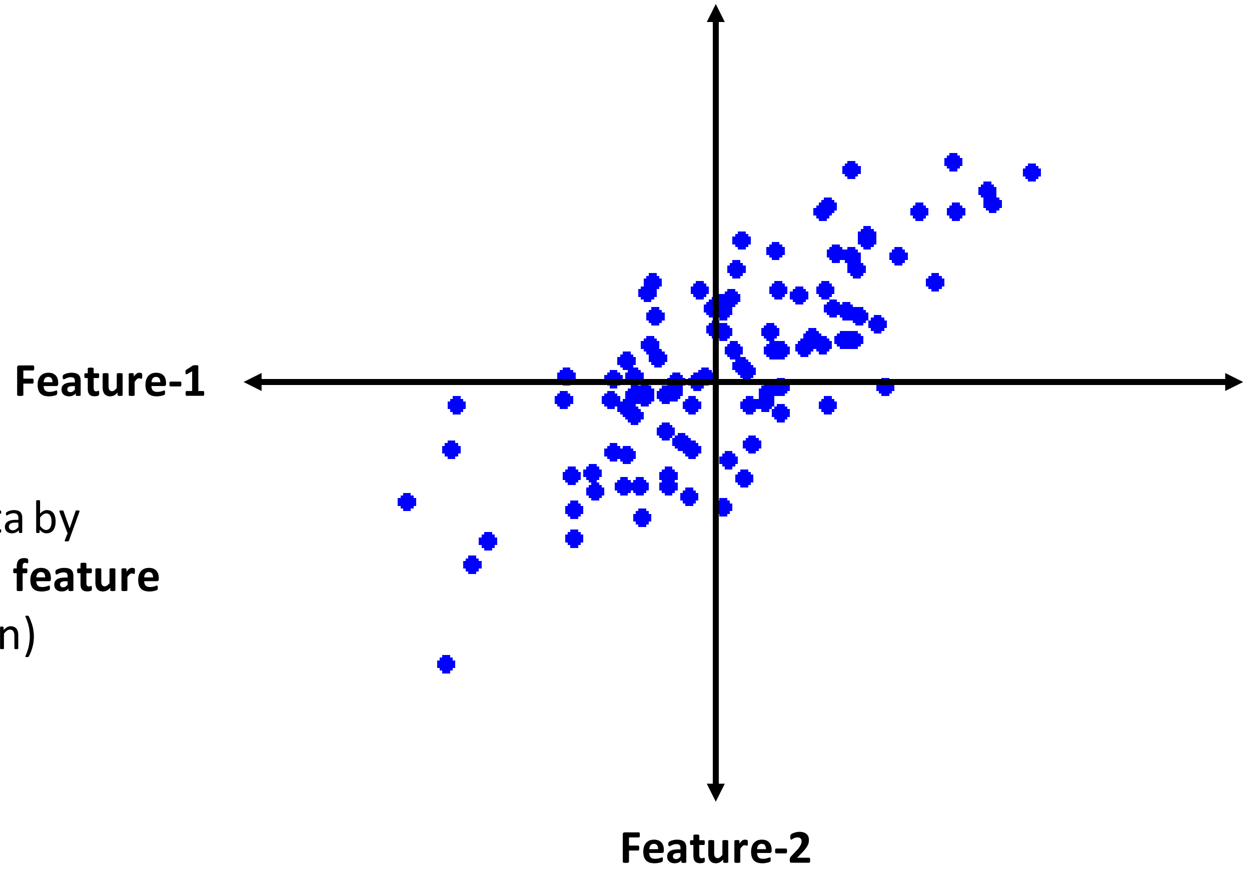


Caveat: PCA is only a rotation

Centering Data

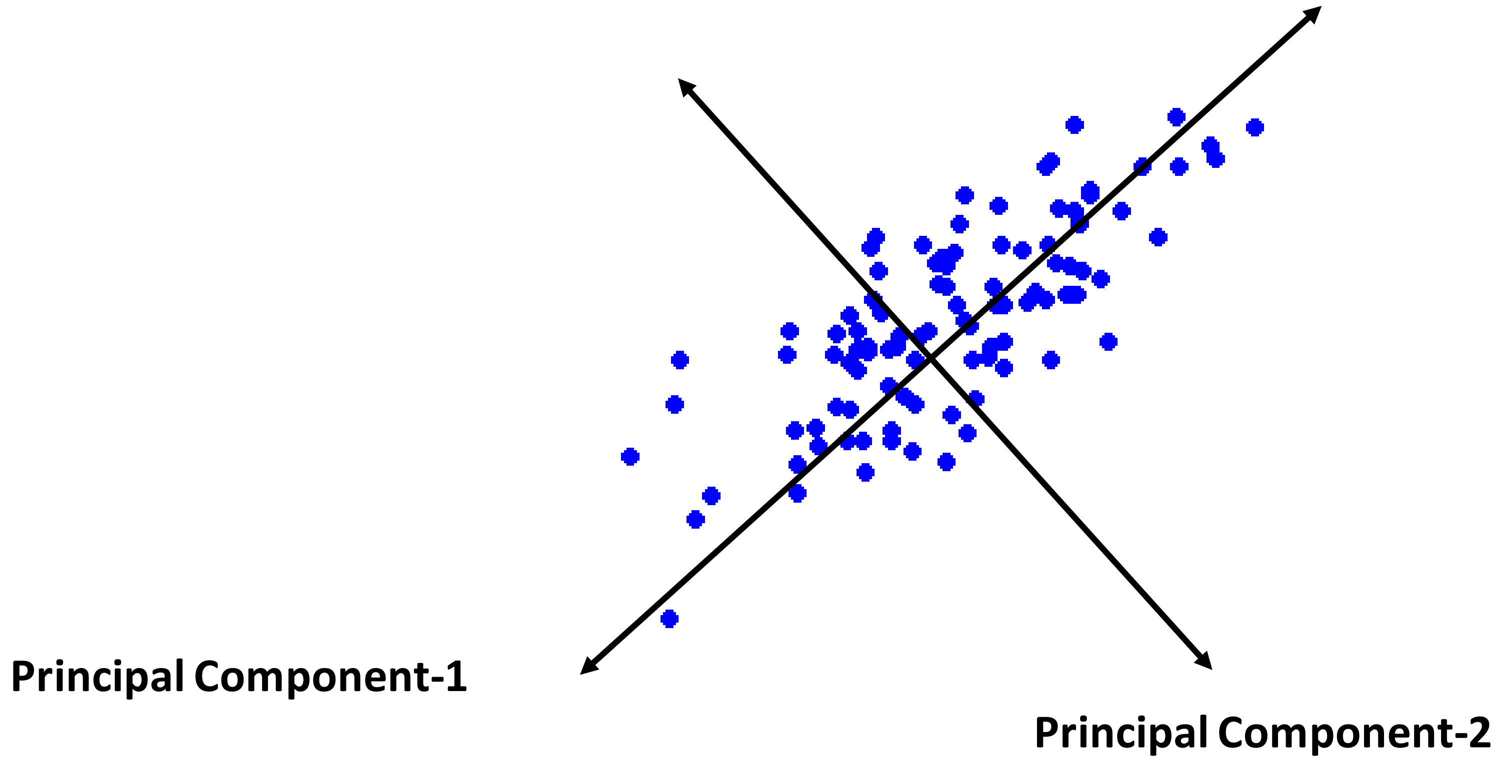


Centering Data



Center (and optionally scale) data by **subtracting the mean from each feature** (and divide by standard deviation)

Centering Data



Caveat: Linear Independence of Basis Vectors

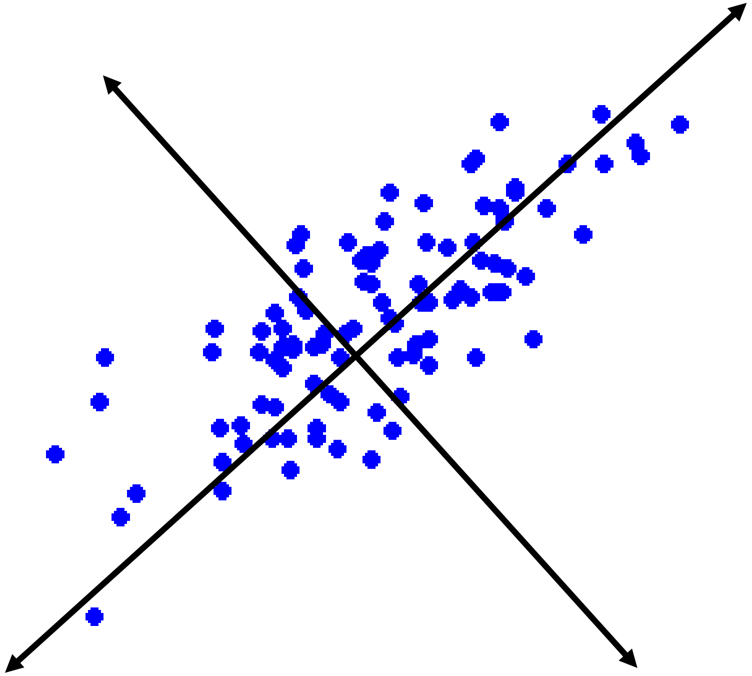
$$\neq$$

Statistical Independence of Random Variables

Orthogonal basis vectors are **linearly independent**, diagonal covariance matrix ensures new features are **uncorrelated**

Does NOT mean

New features are **statistically independent**



Principal Component-1

Principal Component-2

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

Caveat: Feature with most variance

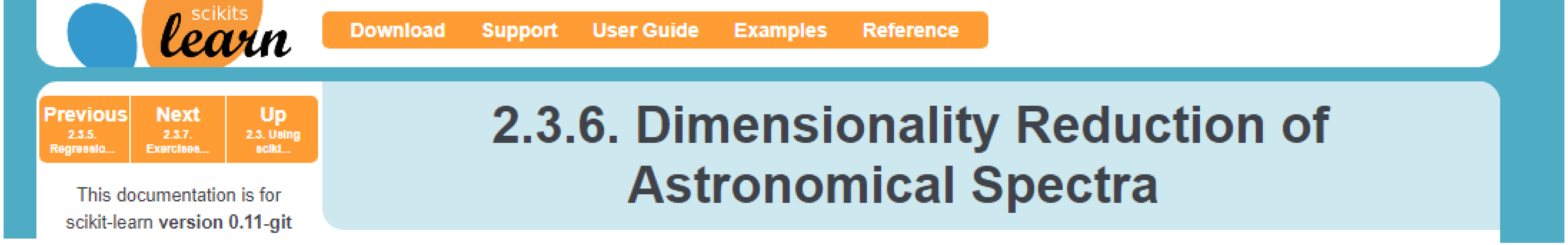
$\neq$

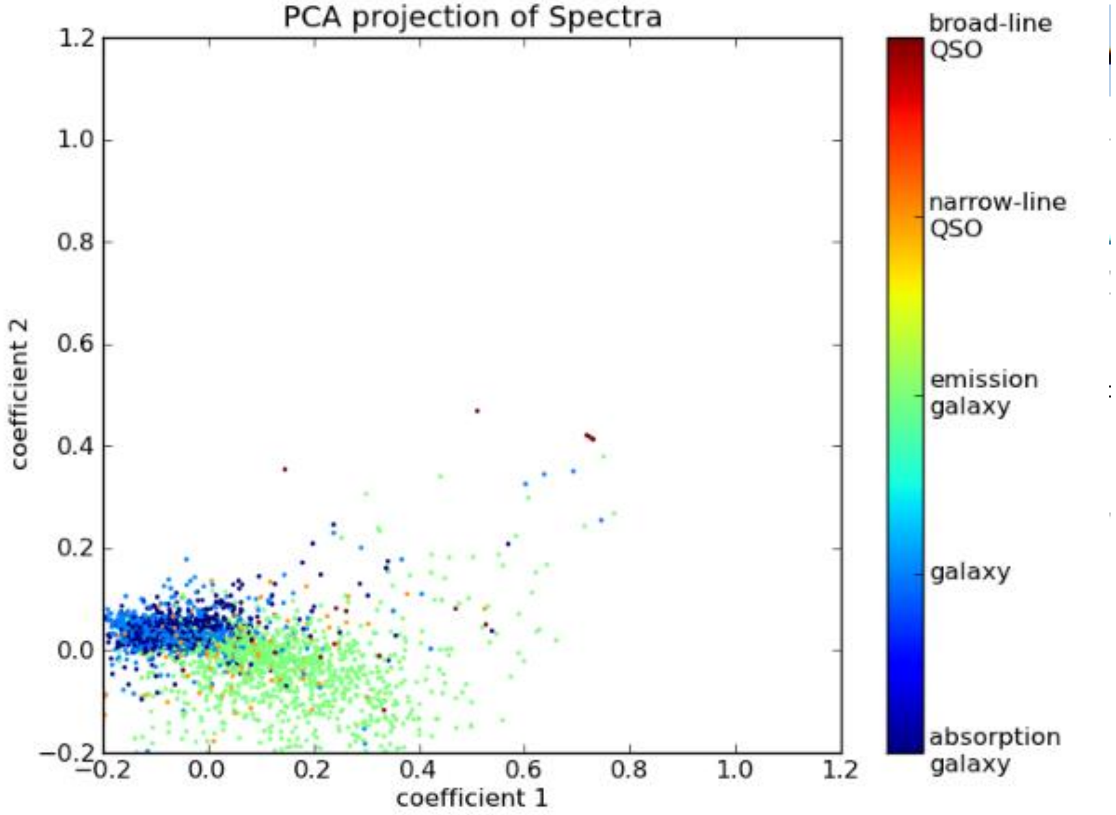
The most important feature

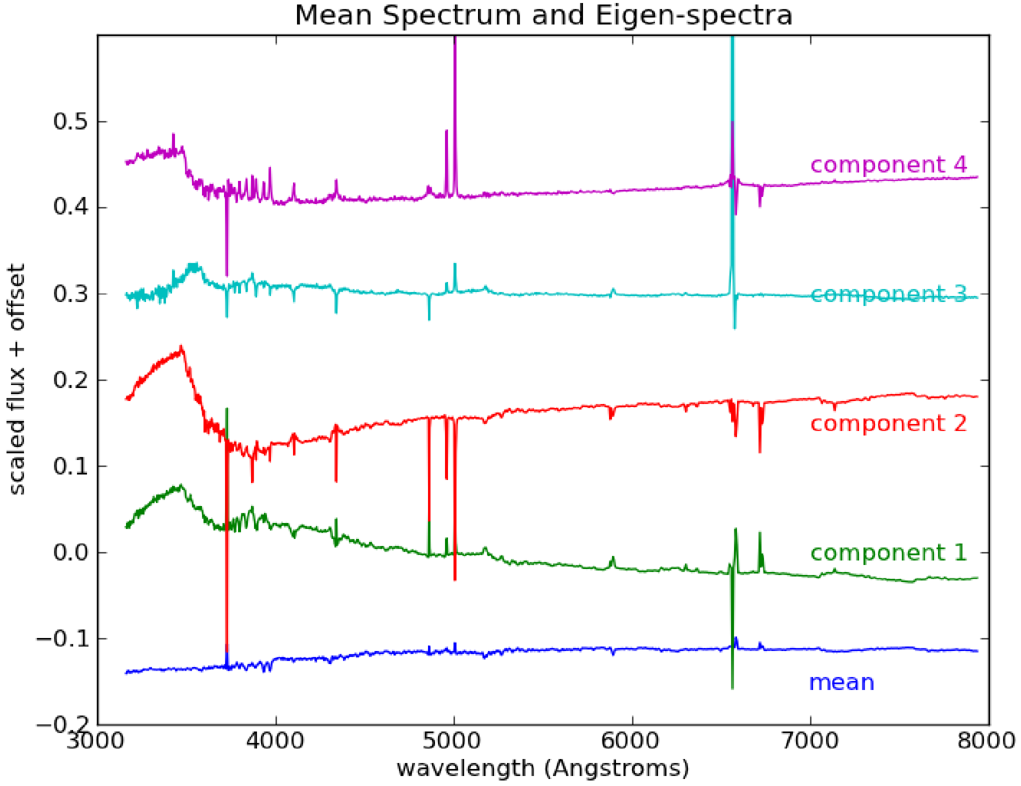
Neil Lawrence
@lawrennd

"Have you run PCA on it?" is the data scientist's equivalent of "Have you switched it off and on again?"

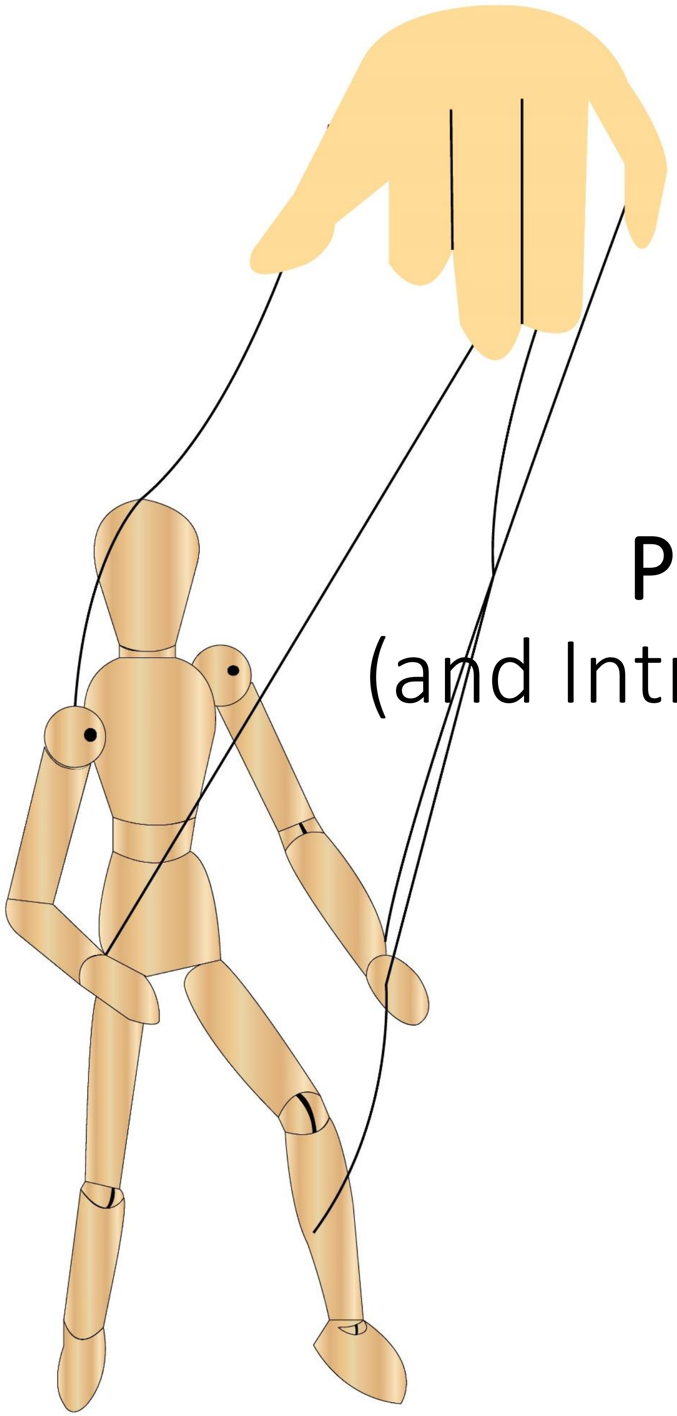
1:44 PM · Jun 12, 2020 · Twitter Web App







[2.3.6. Dimensionality Reduction of Astronomical Spectra — scikit-learn 0.11-git documentation]



PCA: A Probabilistic Perspective
(and Introduction to Latent Variable Models)

Latent Variable Models

Low Dimensional variable you believe created the data (Mass, Age, Metallicity)

$$Y = f(X)$$

Latent Variable Model

High Dimensional data you have observed (Spectra, Images)

Assumption-1: Linear Model

Transformation Matrix
(unknown)

Random Noise
(unknown)

$$Y = \mathbf{W}\, X + \boldsymbol{\varepsilon}$$

Observed Vector

Latent Variable Vector
(unknown)

Assumption-2: Gaussian Noise

$$Y = \mathbf{W}\, X + \boldsymbol{\varepsilon}$$

$$\varepsilon \sim \mathcal{N}(0, \Psi)$$

Likelihood: $p(y_i | x_i, W) = \mathcal{N}(W x_i, \Psi)$

* I assumed that data is centered, this does not lose generalizability

Assumption-3: Gaussian Priors on Latent Variables

$$\mathrm{Y} = \mathbf{W}\,\mathrm{X} + \boldsymbol{\varepsilon}$$

$$\mathrm{X} \sim \mathcal{N}(0, \mathbf{I})$$

Likelihood: $p(y_i|W) = \mathcal{N}(0, WW^T +\ \Psi)$

Assumption-4: Isotropic Noise

$$Y = \mathbf{W}\,X + \boldsymbol{\varepsilon}$$

$$\Psi = \sigma^2\,\mathbf{I}$$

Likelihood: $p(y_i|W) = \mathcal{N}(0, WW^T + \sigma^2\,\mathbf{I})$

Maximum Likelihood Estimator (MLE)

Likelihood: $p(y_i|W) = \mathcal{N}(0, WW^T + \sigma^2\,\mathbf{I})$

Maximize Log Likelihood

$\equiv$

Maximize: $\underbrace{Tr(W^T W \times S)}_{\text{Projection of Sample covariance along new axes (under the assumption } \sigma^2 \to 0)}$ S =Sample Covariance

Projection of Sample covariance along new axes
(under the assumption $\sigma^2 \to 0$)

Maximum Likelihood Estimator (MLE)

Maximize Log Likelihood

$$\equiv$$

$$\text{Maximize: } \underbrace{Tr(W^T W \times S)}_{\text{Projection of Sample covariance along new axes (under the assumption } \sigma^2 \to 0)}$$

Which is exactly what PCA is!

You DO NOT need to use these assumptions for your latent variable model !

Latent Variable Models

$$Y = f(X)$$

- **f(), linear:** Assume any general prior on X and use Maximum Likelihood Estimation/ Maximum A Posteriori Estimation
- **f(), linear:** Assume non isotropic noise → Factor Analysis
- **f(), non-linear:** Use a neural network to model → Autoencoder
- **f(), non-linear:** X has noise, Use neural network → Variational Autoencoder
- **f(), non-linear (aka linear with infinite dimensions) and with Gaussian Priors** → Gaussian Process Latent Variable Model

Summary

- Principal Components rotate your axes towards maximum variance

- Principal Components have the lowest reconstruction error →
Good for dimensionality reduction

- Principal Component Analysis is just one specific kind of Latent Variable Model