TEXAS A&M UNIVERSITY STATISTICS

Principal Components Analysis

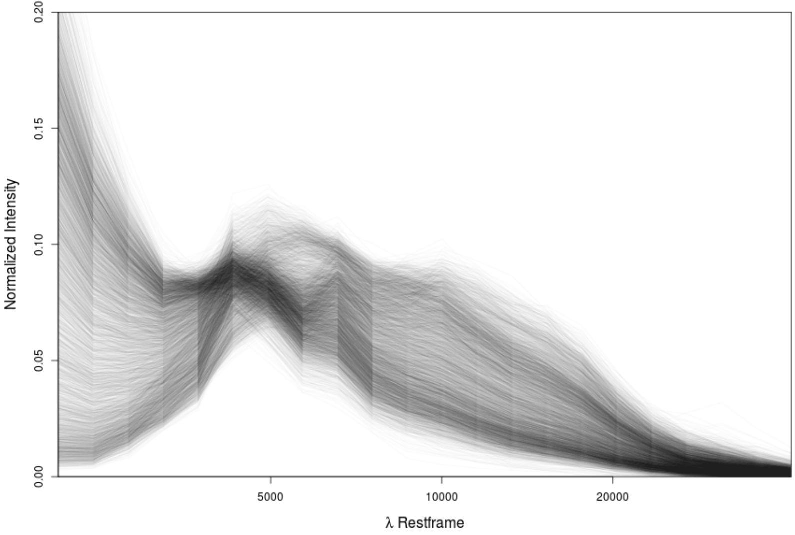
James Long

November 17, 2015

References

- **Elements of Statistical Learning** (Tibshirani, Hastie, Friedman)
  - Chapter 14.5
  - `http://statweb.stanford.edu/~tibs/ElemStatLearn/`
- **Functional Data Analysis** Ramsay and Silverman
  - Chapters 8 and 9
- **Statistics, Data Mining, and Machine Learning in Astronomy** (Ivezic, et al)
  - Section 7.3
- **Modern Statistical Methods for Astronomy** (Feigelson, Babu)
  - Section 8.4.2

Synthetic Photometry



Dimension Reduction

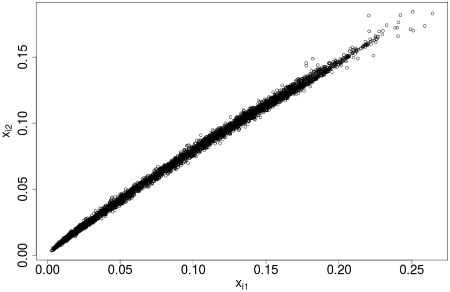
- $X \in \mathbb{R}^{n \times p}$ is synthetic photometry
  - $n = 3984$ is number of galaxies
  - $p = 22$ is number of synthetic filters
  - $x_i \in \mathbb{R}^p$ is i^{th} row of X
- p is the "dimension" of the data
- Sometimes the vectors x_i are all (approximately) in some lower dimensional subspace of $\mathbb{R}^p$
- Finding and characterizing this subspace is called "dimension reduction"

Dimension Reduction Example

Consider

$$\{(x_{i1}, x_{i2})\}_{i=1}^{n}$$

the first two dimensions of synthetic photometry for each observation.



Message:

- The <u>intrinsic</u> dimension is 1.
- We can compress the two dimensional data into 1 dimension.

Principal Components Analysis (PCA) Idea

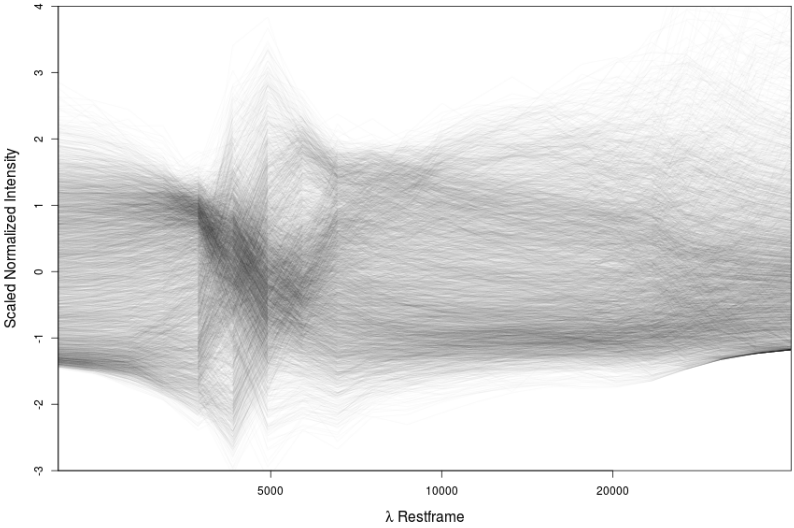
- Realign axes so
  - Most variation on first axis
  - Second most variation on second axis
  - . . .
- Ignore higher axes because minimal variation in these directions.
- The principal components describe how the new axes map to the old axis.

PCA is typically applied to a scaled version of X.

$$X' = (X - 1\mu^T)S^{-1}$$

- Remove column means (μ)
- Scale column variances to 1.
  - S is diagonal with S_{jj} = standard deviation column j of X

Scaled Data Matrix X'



PCA Math – Singular Value Decomposition

The singular value decomposition of X' (assuming $n > p$) is

$$X' = U\Sigma V^T$$

where

- U is $n \times p$ with $U^T U = I$[1]
  - The data in the new coordinate system.
- V is $p \times p$ with $V^T V = I$
  - V rotates the new coordinates to the old coordinates.
- Σ is $p \times p$ diagonal with $\Sigma_{jj} > \Sigma_{ii}$ for $j < i$ [2]
  - Σ scales the new coordinates to the old coordinates.

---

[1] U is $n \times p$ in R and $n \times n$ in theory.

[2] Σ is $p \times p$ in R and $n \times p$ in theory.

Reconstructing the data

- A $q \leq p$ dimensional reconstruction of X' (in R notation) is

$$X'_q = U[,1{:}q]\Sigma[1{:}q,1{:}q]V[,1{:}q]^T$$

- If the data lies (approximately) on a q dimensional subspace then

$$X'_q \approx X'$$

- Obtain an approximation of the original data

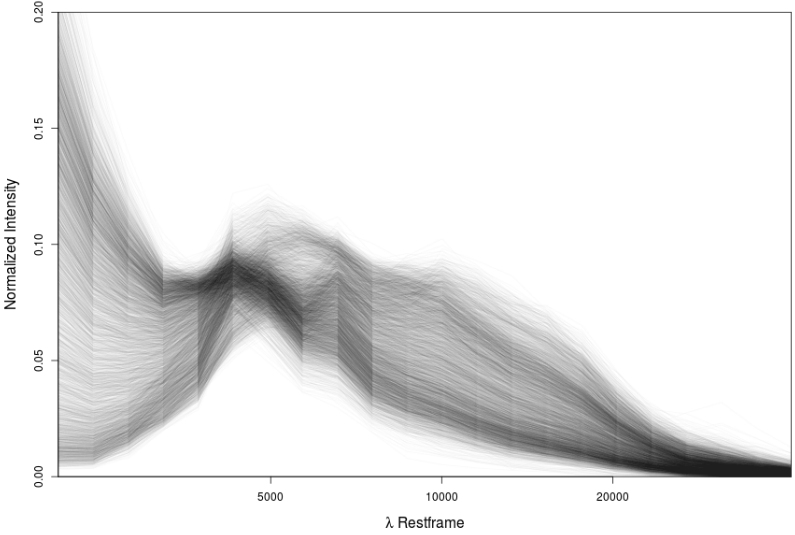
$$X_q = X'_q S + 1\mu^T$$

and

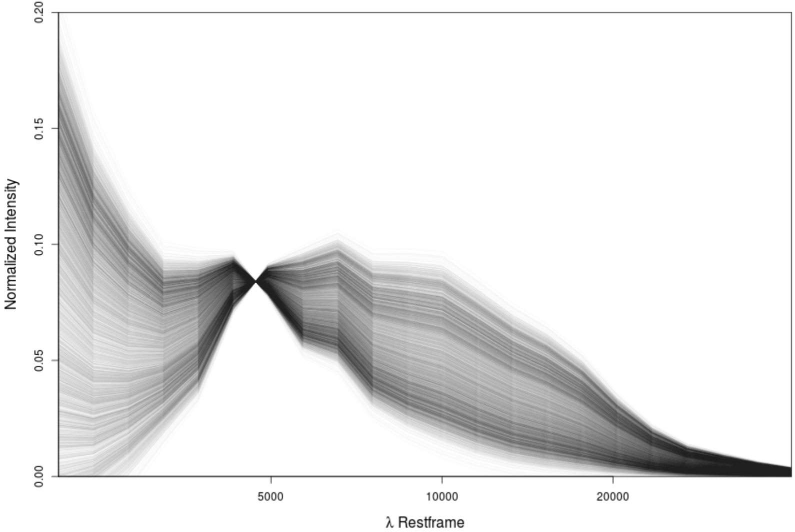
$$X_q \approx X$$

For functional data we can do a visual check.

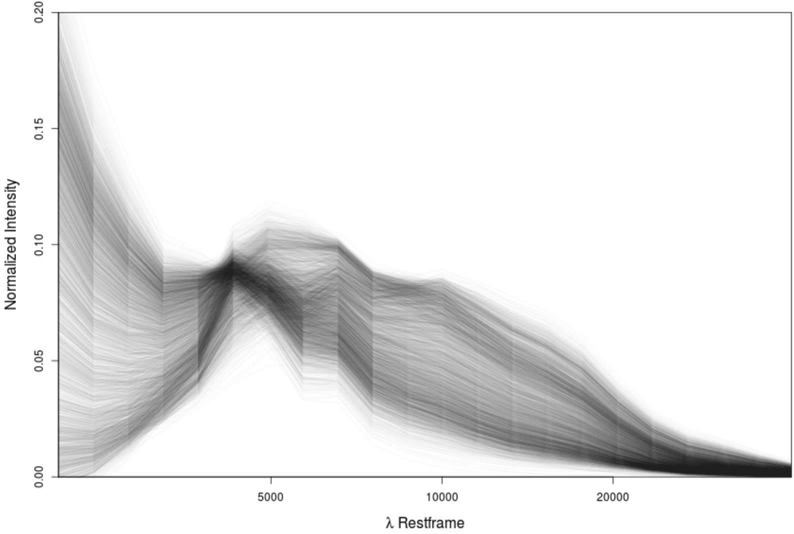
Synthetic Photometry



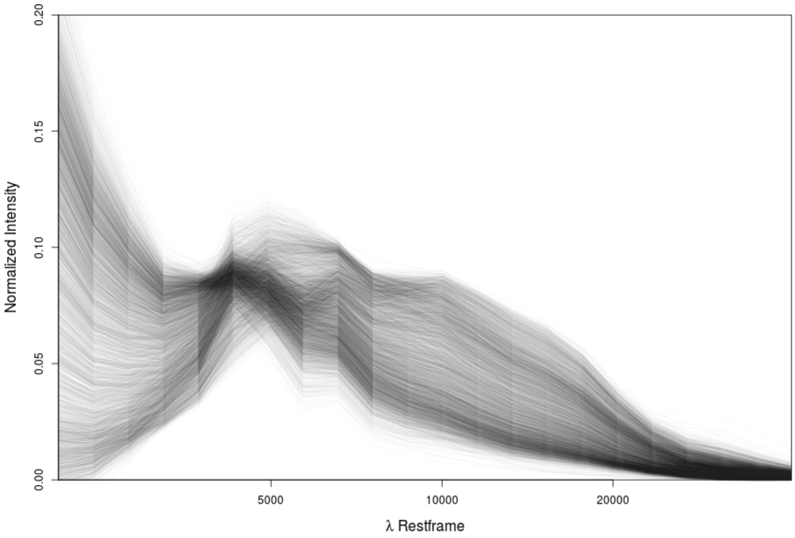
Reconstruction with $q = 1$ Principal Component



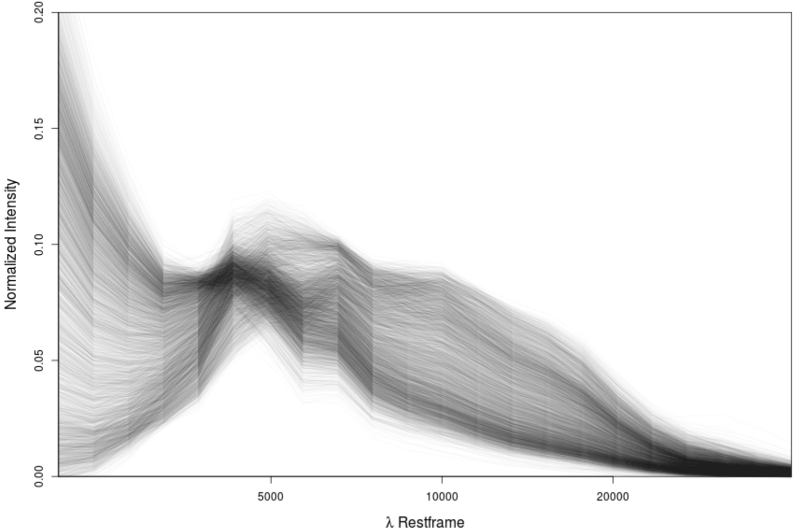
Reconstruction with $q = 2$ Principal Components



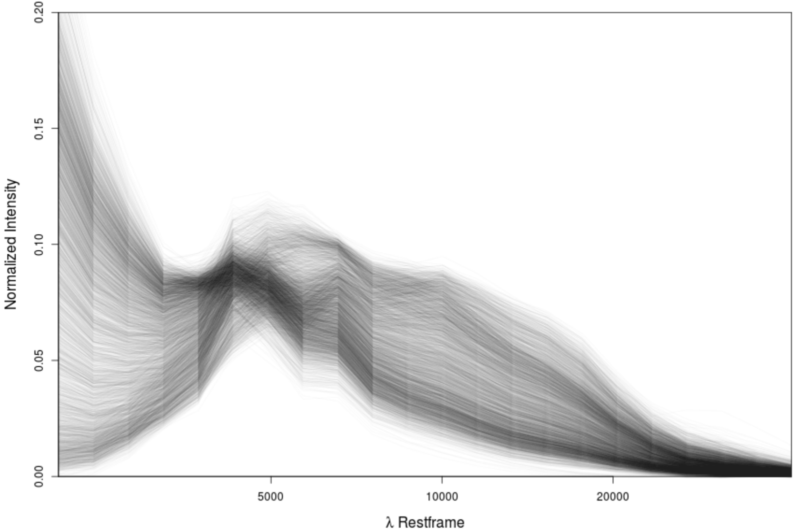
Reconstruction with $q = 3$ Principal Components



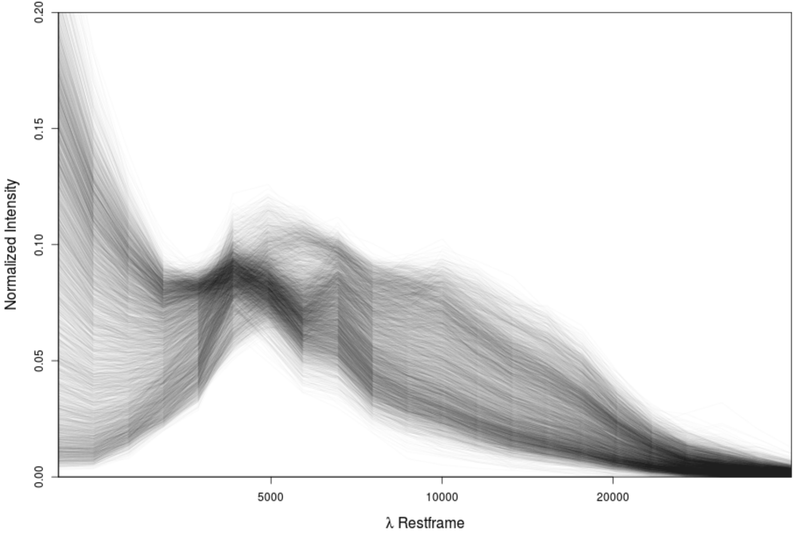
Reconstruction with $q = 4$ Principal Components



Reconstruction with $q = 5$ Principal Components



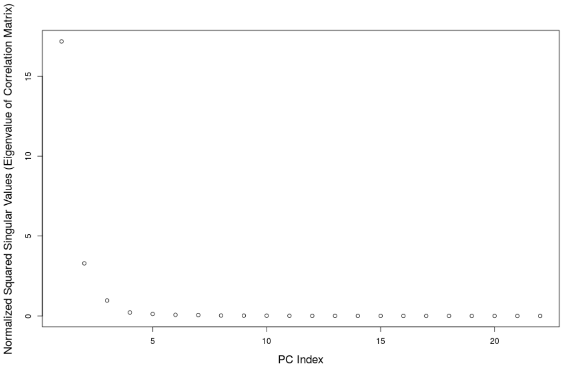
Synthetic Photometry



What's Happening in the SVD Formula

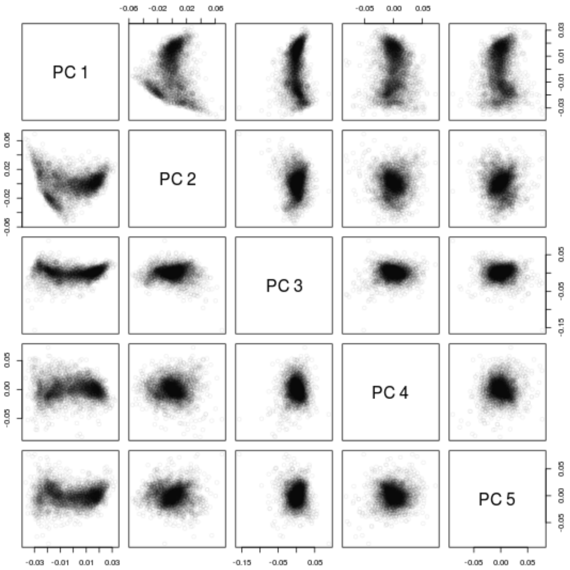
- We see
$$X_q \approx X$$
for $q = 2$.
- So
$$X'_q = U[,1{:}q]\Sigma[1{:}q,1{:}q]V[,1{:}q]^T \approx U\Sigma V^T = X'$$
- So Σ_{jj} for $j > 2$ are small.

Scree Plot

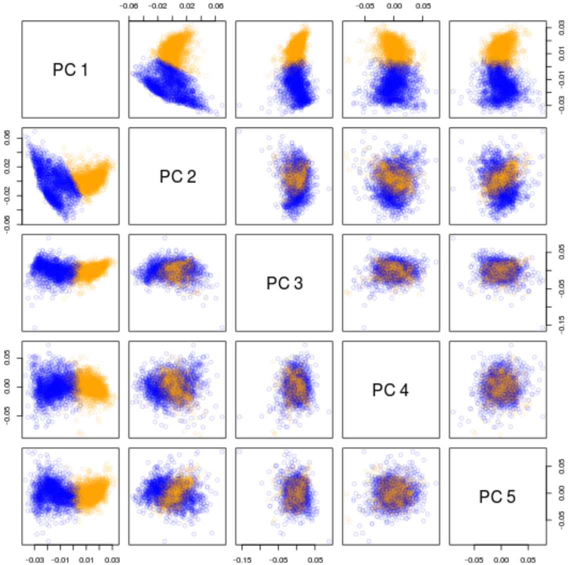


- y–axis values Σ_{jj}^2/n (sum to $p = 22$)
- Most variation can be explained by a small number of principal components.
- This plot is helpful for deciding how many PCs to use (choose q).

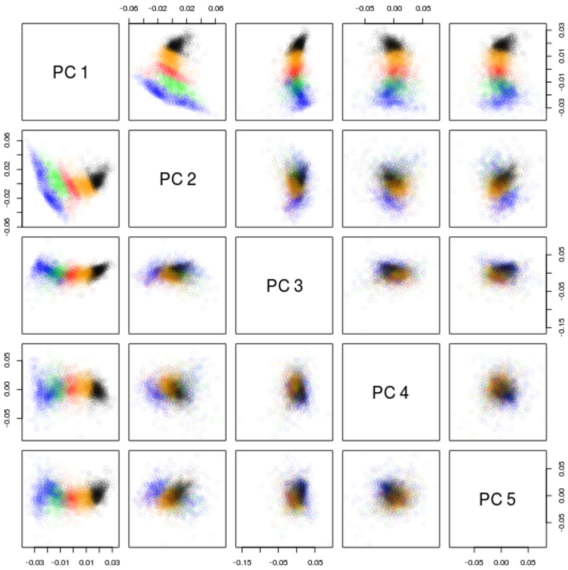
Principal Components ($U[,1{:}5]$)



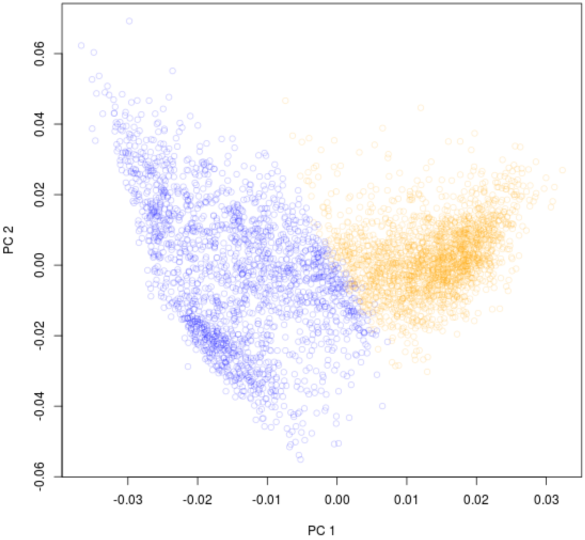
Principal Components with Hierarchical Clustering



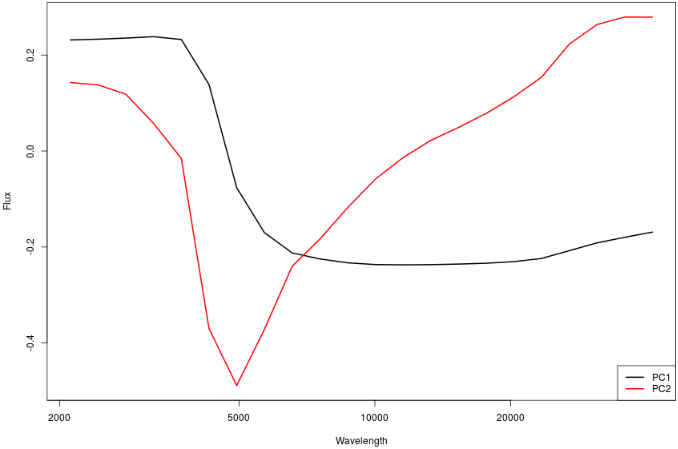
Principal Components with Hierarchical Clustering



Principal Components ($U[,1{:}2]$)



Two Principal Components $V[, 1{:}2]$



First $q = 2$ columns of U:

$$X_q = (U[, 1{:}q]\Sigma[1{:}q, 1{:}q]V[, 1{:}q]^T)S + 1\mu^T$$

Uses of Result

- Are there actually clusters in the data or a continuous set of shapes that can be characterized by 2 or 3 values (principal components)?
- Continuous composities: For any galaxy we can calculate "neighbors" in PC space and make composites based on neighbors.

Related Methods

- Non–negative matrix factorization
  - Chapter 14.6 of Hastie, Tibshirani, Friedman
- Functional principal components analysis (FPCA)
  - PCA here required synthetic photometry at same wavelengths. FPCA could be applied to restframe actual photometry.