Some Statistical Aspects of Photometric Redshift Estimation

James Long

November 9, 2016

Photometric Redshift Estimation Background

Non–Negative Least Squares

Non–Negative Matrix Factorization

EAZY Results

Outline

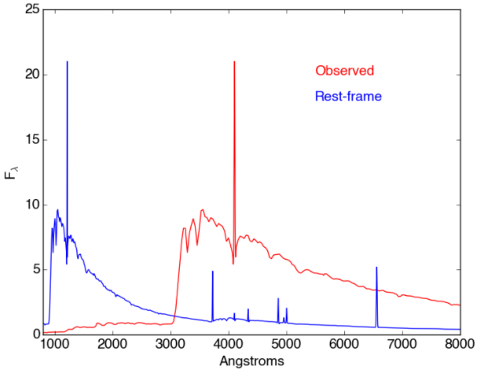
Photometric Redshift Estimation Background

Non–Negative Least Squares

Non–Negative Matrix Factorization

EAZY Results

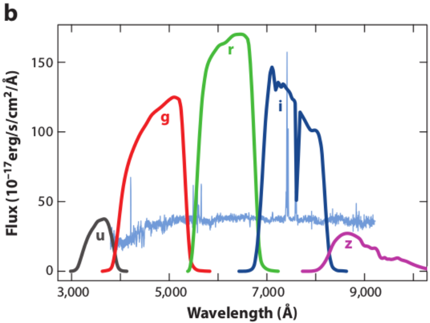
Redshift



- f_{obs} (red line) is the observed spectrum
- f_{rf} (blue line) is the true, unobserved spectrum
- $f_{rf}(\lambda) = f_{obs}(\lambda(1 + z))$
- goal: estimate z
- relatively easy if we observe f_{obs}

Photometric Data

- spectral data is very expensive to collect
- much more photometric data available
- photometry is the spectra observed at small set of wavelengths



Photometric redshift estimation (photoz) is the process of estimating redshift from photometric data.

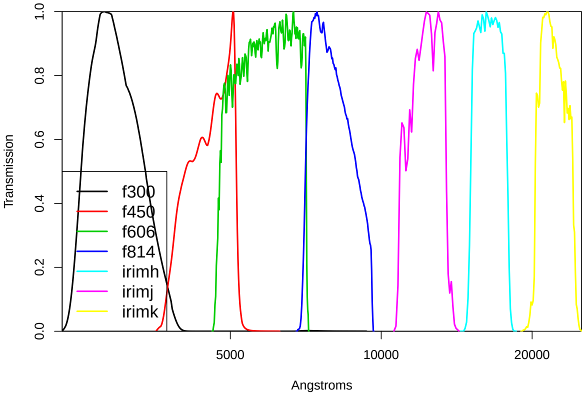
Source: Schafer "A Framework for Statistical Inference in Astrophysics."
`http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-022513-115538`

Machine Learning Approach to Photoz

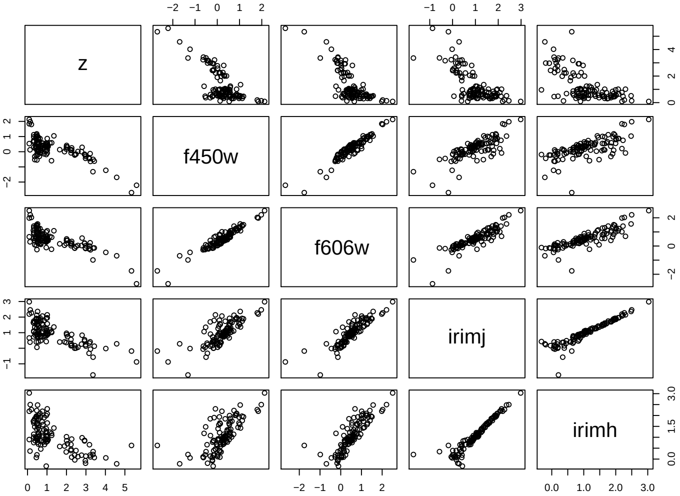
Idea:

- create training set by taking spectroscopy on subset of data
- construct classifier on training data, apply to unlabeled data

Example: Collected photometry in 7 filters for 114 galaxies.

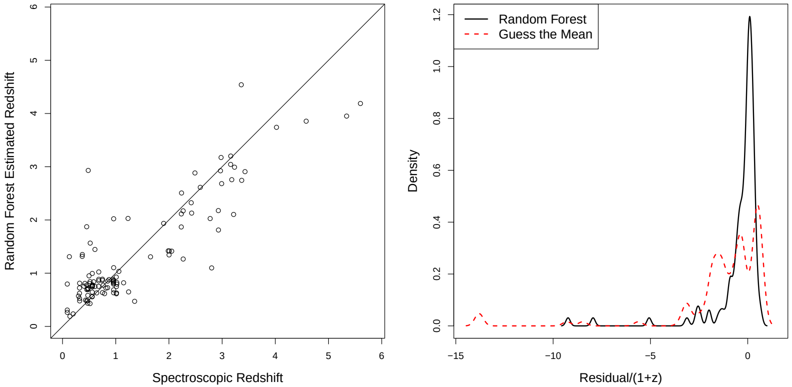


Spectroscopic Redshift and Photometry



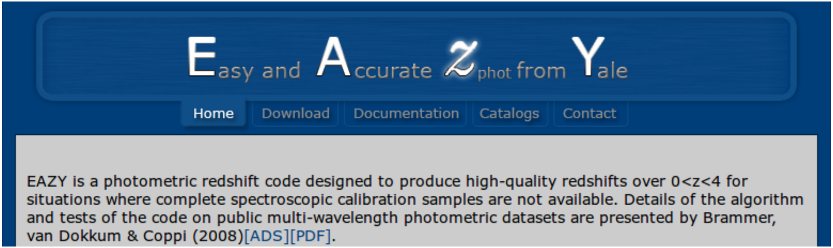
Only showing 4 of 7 filters.

Apply Random Forest



- ignored uncertainty in photometry (features)
- training sets are often from nearby objects, unlabeled data far away objects
- difficult to compute uncertainty in estimate

Easy Accurate Zphot from Yale (EAZY)



- developed by Brammer, van Dokkum, Coppi
- synthesizes ideas from several earlier works

Paper: http://adsabs.harvard.edu/abs/2008ApJ...686.1503B. Code: https://github.com/gbrammer/eazy-photoz/

Simple Idea

Notation:

- determine set of model spectra T_i for $i = 1, \ldots, n_T$
  - actual spectra
  - spectra generated by theoretical models
- let $T_{z,i}$ be spectra i redshifted to z
- let $T_{j,z,i}$ be spectra i redshifted to z, convolved with filter j
- F_j and σ_j are the flux and flux error in band j

Optimization Function:

$$\widehat{z} = \underset{z}{\operatorname{argmin}} \min_{1 \leq i \leq n_T} \sum_{j=1}^{J} \left(\frac{T_{j,z,i} - F_j}{\sigma_j} \right)^2$$

Outline

Photometric Redshift Estimation Background

Non–Negative Least Squares

Non–Negative Matrix Factorization

EAZY Results

Generalizing the Template Set

Consider expanding templates through linear combinations

$$T_z = \sum_{i=1}^{n_T} \alpha_i T_{z,i} = \boldsymbol{\alpha}^T T_z$$

Astronomical theory says that $\alpha_i \geq 0 \ \ \forall i$. So optimization problem becomes

$$\widehat{z} = \underset{z}{\operatorname{argmin}} \ \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}>0} \sum_{j=1}^{J} \left(\frac{\boldsymbol{\alpha}^T T_z - F_j}{\sigma_j} \right)^2$$

This is known as non–negative least squares in statistics.
Solve on a grid of z.

More Familiar Statistical Notation

The non-negative least squares estimate is

$$\widehat{\beta} = \underset{\beta:\beta\geq 0}{\text{argmin}}\ ||Y - X\beta||_2^2$$

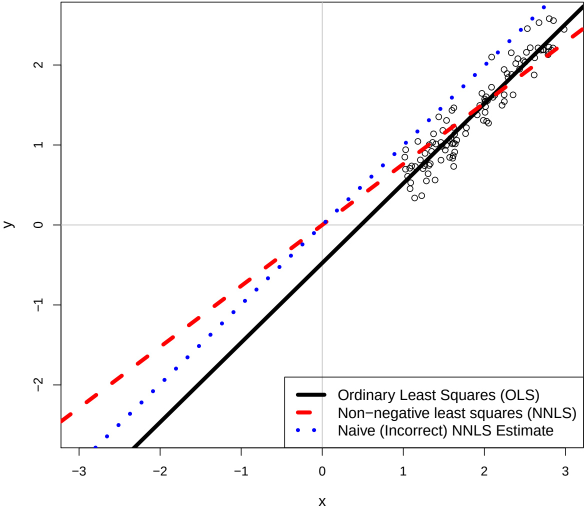
where

- $X \in \mathbb{R}^{n \times p}$ is design matrix
- $Y \in \mathbb{R}^{n}$ is response

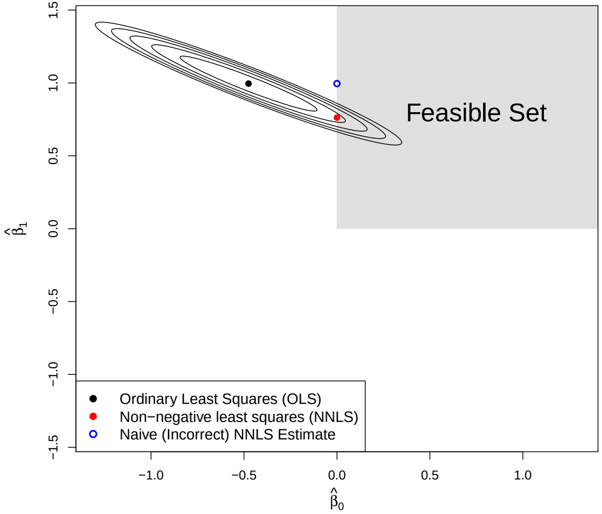
without the $\beta \geq 0$ constraint the problem has the familiar LS form

$$\underset{\beta}{\text{argmin}}\ ||Y - X\beta||_2^2 = (X^T X)^{-1} X^T Y.$$

Example Data



Example Likelihood Surface



Non–Negative Least Squares Algorithms

- `R`–package `nnls` and `scipy.optimize.nnls` use active set method (Lawson and Hanson 1974 book "Solving Least Squares Problems")
- EAZY uses "Multiplicative Updates for Nonnegative Quadratic Programming" by Sha 2007 *Neural Computation*

Outline

Constructing Template Set

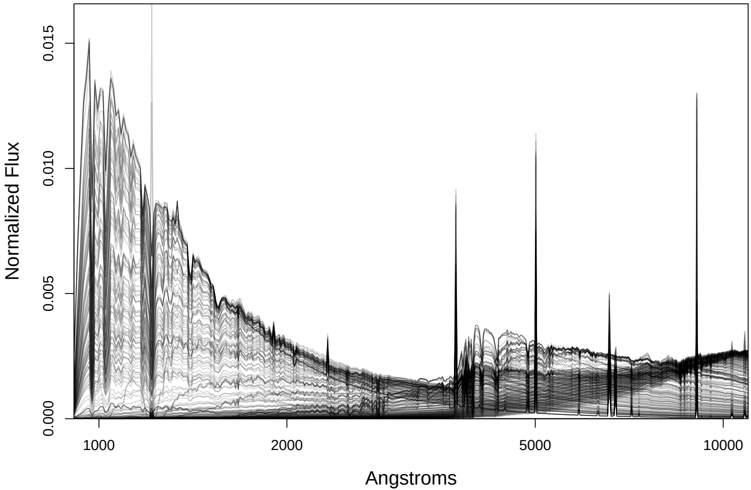
T_i are the templates. They may be

- observed data
  - advantages: "real" data that does not make physical assumptions
  - disadvantages: sometimes expensive to collect, little data available at high redshifts
- output from physical simulations
  - advantages / disadvantages reversed from observed data

Having a small set of T_i is convenient because

- interpretation is easier
- computation is faster

Constructing Template Set



T_i for $i = 1, \ldots, 259$ filters, lots of redundancy, like to reduce set

Dimension Reduction for Template Construction

- $X \in \mathbb{R}^{n \times p}$ are templates
  - $n =$ number of templates
  - $p =$ number of bins for each tempate
- p is the "dimension" of the data
- assume: the row vectors x_i are (approximately) in some lower dimensional subspace of $\mathbb{R}^p$
- finding and characterizing this subspace is called "dimension reduction"

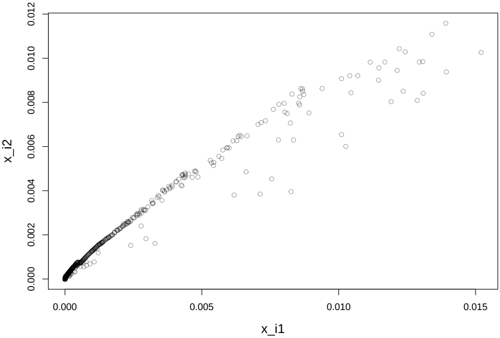
We would like this subspace to be characterized as a linear combination of <u>positive</u> bases.

Dimension Reduction Example

consider

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

the first two dimensions of filters



Message:

- intrinsic dimension is near 1
- 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
- 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

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]
  - 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

$$X^* \approx X_q^*$$

- obtain an approximation of the original data

$$X_q = X_q^* S + 1\mu^T$$

and

$$X \approx X_q$$

- reduced template set is $V[,1{:}q]$

Non–negative Matrix Factorization

Decomposition: When the data matrix X is positive we can decompose

$$X_{n \times p} \approx W_{n \times r} H_{r \times p}$$

where rows of H are basis.

Algorithm: Maximize

$$L(W, H) = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} \log(WH)_{ij} - (WH)_{ij})$$

ie maximum likelihood under the model that x_{ij} is Poisson$((WH)_{ij})$

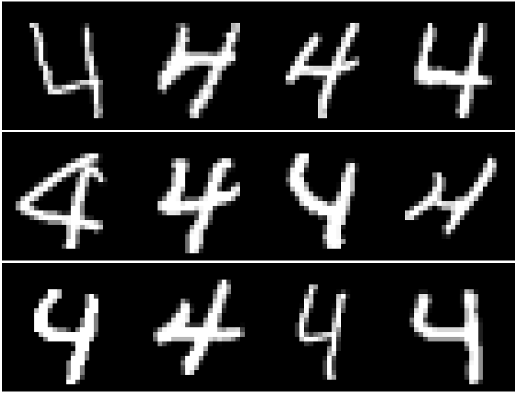
Source: Elements of Statistical Learning. Haste, Tibshirani, Friedman. Chapter 14.6. Available: `http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Non–negative Matrix (NMF) Factorization

- the NMF basis vectors often have more physical interpretation than PCA basis
  - eg NMF spectral basis elements look like spectra
- optimizing log likelihood difficult for NMF
- identifiability issues with model*

Source: Donoho and Stodden. "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?"
https://web.stanford.edu/~vcs/papers/NMFCDP.pdf

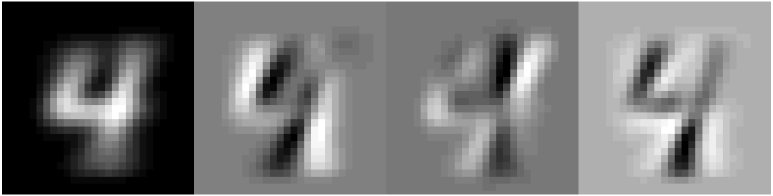
MNIST Data Set – 12 Fours Out of 4072



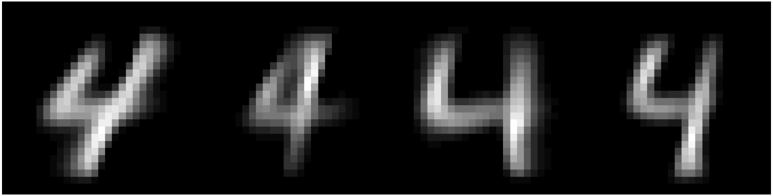
- each image 28×28 pixels
- vectorize image i to $x_i \in \mathbb{R}^{784}$
- $X \in \mathbb{R}^{4072 \times 784}$
- apply PCA and NMF

MNIST Results PCA versus NMF

PCA (Mean + 3 principal components)

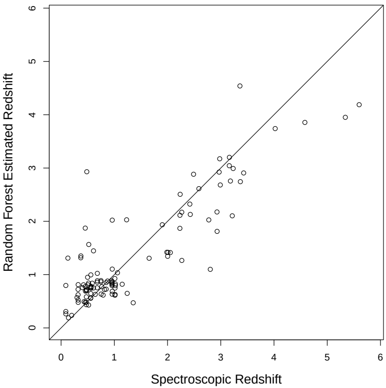


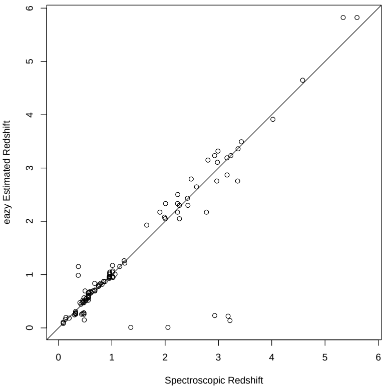
NMF (4 Basis Vectors)



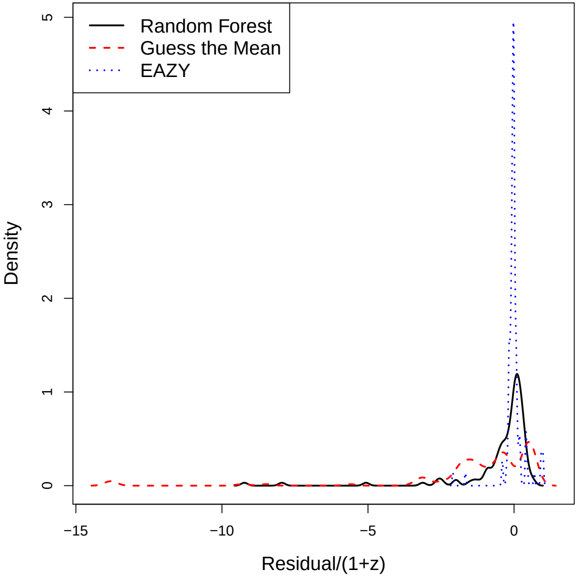
Outline

Random Forest – EAZY Comparison





Random Forest – EAZY Comparison



Other Challenges / Opportunities / Approaches

- accounting for template error (model misspecification)
- estimating templates from photometric data
- propagating uncertainty on redshift to the next stage of analysis
- other work on photoz:
  - "Robust machine learning applied to astronomical data sets. III. Probabilistic photometric redshifts for galaxies and quasars in the SDSS and GALEX." Ball ApJ 2008
  - "Bayesian photometric redshift estimation." Benitez ApJ 2000
  - "Random forests for photometric redshifts." Carliles ApJ 2010