Rau, Seitz, Brimioulle, Frank, Friedrich, Gruen, Hoyle in prep.

Machine Learning for PhotoZ PDFs

Markus Michael Rau, SPV Stella Seitz

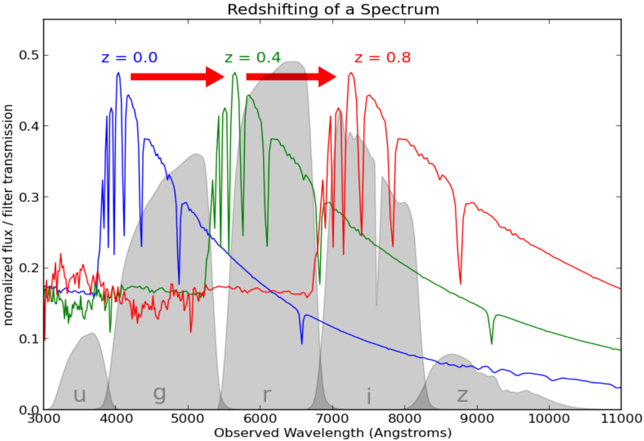
USM Munich

January 19, 2015

Outline:

- What are Photometric Redshifts and why do we need them?
- What are Conditional Probability Density Functions (PDFs) and why do we need them?
- How do we know how accurate they are?
- How do we estimate them efficiently?
- How well do our algorithms perform on data?

Photometric Redshifts



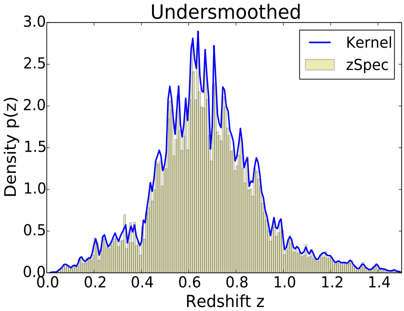
Source: http://www.astroml.org/sklearn_tutorial/_images/plot sdss_filters_2.png

Why Photometric Redshifts?

- Spectroscopic Redshifts expensive (long exposure times especially for faint objects)
- Small datasets
  $\rightarrow$ insufficient for many cosmological applications (cosmic shear, large scale structure, etc.)
- Solution: Photometric Surveys with spectroscopic overlap

Learn the mapping between the photometry of objects $\mathbf{f}$ and their spectroscopic redshifts z_{spec} and apply this model to objects without spectroscopy. Machine Learning often more accurate than traditional Template Fitting.
(Sánchez et al. 2014 (DES) arXiv:1406.4407)

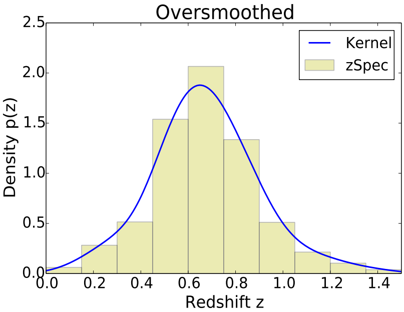
Kernel Density Estimation



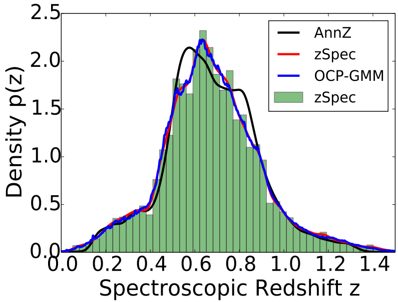
Kernel Density Estimate

$$\hat{p}(z) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}\left(z, z_i^{\text{spec}}, h\right) \tag{1}$$

- h: Bandwidth (i.e. standard deviation of Gaussian)
- z_i^{spec}: Kernel center
- h small → Undersmoothed
- h big → Oversmoothed

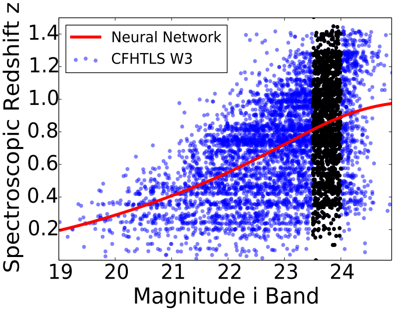


Why PhotoZ PDFs?



- Point Predictions from regression ML algorithms are not able to estimate stacked redshift PDFs
- Unsuited for many applications in cosmology (i.e. Cosmic Shear, Lensing Cluster Mass measurement)

Example in 2D



$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{N} w_i(\mathbf{f}) \mathcal{N}\left(z, z_i^{\text{spec}}, h\right) \tag{2}$$

- 2D example with one filter (Data from CFHTLS W3)
- Estimate conditional PDF in black region

The conditional PDF

- The conditional PDF $p(z|\mathbf{f})$ of the objects redshift z given its photometry $\mathbf{f}$ is defined as:

$$p(z|\mathbf{f}) = p(z, \mathbf{f})/p(\mathbf{f}) \tag{3}$$

- Weighted Kernel Density Estimate:

$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{N} w_i(\mathbf{f})\,\mathcal{N}\left(z, \mu = z_i^{\mathrm{spec}}, \sigma = h\right) \tag{4}$$

- Conditional Mean (Output of ANNz):

$$\hat{z}_{\mathrm{phot}}(\mathbf{f}) = \sum_{i=1}^{N} w_i(\mathbf{f})\, z_i^{\mathrm{spec}} \tag{5}$$

- Conditional Variance:

$$\hat{\sigma}^2(\mathbf{f}) = \sum_{i=1}^{N} w_i(\mathbf{f})\,\left(z_i^{\mathrm{spec}} - \hat{z}_{\mathrm{phot}}(\mathbf{f})\right)^2 \tag{6}$$

- Conditional PDF Estimation can be used to obtain arbitrary point predictions (Mean, Mode, Median)
- Incorporate redshift uncertainty in follow up analysis
- Novel algorithm: parametrizes conditional pdf highly efficient (5 numbers/object)
  $\rightarrow$ Scales well to large datasets (e. g. Euclid, DES)
- Allows the accurate reconstruction of the sample redshift pdf of a photometric sample

$$p(z) = \frac{1}{N} \sum_{i=1}^{N} p(z|\mathbf{f}_i) \tag{7}$$

Machine Learning Methodology

- Split available data into three datasets (60%, 20%, 20%) training set, validation set, test set
- Estimate the conditional redshift PDF on the training set
- Tune the estimate on the validation set
- Predict the performance on unseen data using the test set

Evaluation Metrics

- Kullback-Leibler Divergence

$$D(p||\hat{p}) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{\hat{p}(\mathbf{x})}\right) d\mathbf{x} \tag{8}$$

$$D(p||\hat{p}) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log\left(p(\mathbf{x})\right) d\mathbf{x} - \int_{-\infty}^{\infty} p(\mathbf{x}) \log\left(\hat{p}(\mathbf{x})\right) d\mathbf{x} \tag{9}$$

- Minimize $-\int_{-\infty}^{\infty} p(\mathbf{x}) \log\left(\hat{p}(\mathbf{x})\right)$
- Minimize mean negative log-likelihood loss (MNLL)

$$\mathrm{MNLL} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\hat{p}(\mathbf{z}_i|\mathbf{f}_i)\right) \tag{10}$$

Basic Concept

Remember:

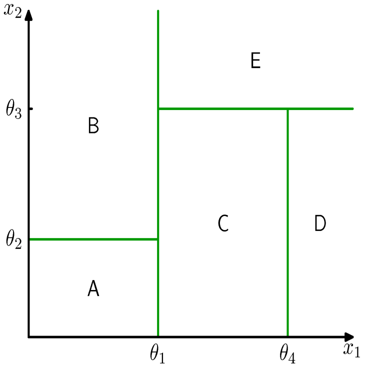
$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{N_{\rm tr}} w(\mathbf{f}_i)\,\mathcal{N}\left(z, \mu = z_i^{\rm spec}, \sigma = h\right) \qquad (11)$$

$N_{\rm tr}$: Number of objects in the training set

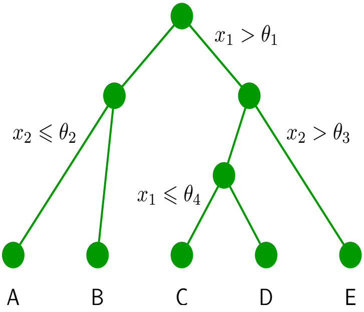
Questions:

- How to estimate the weights $w_i(\mathbf{f})$?
  - Using a Quantile Regression Forest
  - Using an Ordinal Classification Approach
- Given $\left\{w_i(\mathbf{f}), z_i^{\rm spec}\right\}$ how do we estimate the objects PDF?
  - Weighted Kernel Density Estimate (Eqn. 11)
  - Linear combination of normal densities (Gaussian Mixture Model)

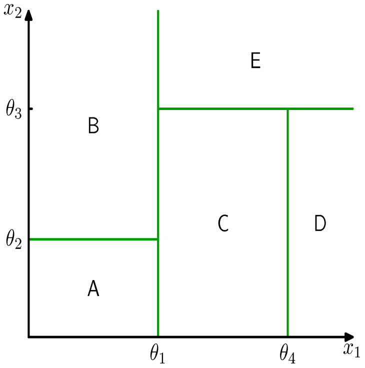
Regression Tree



research.microsoft.com/en-us/um/people/cmbishop/prml/index.htm



Quantile Regression Forest (Meinshausen 2006)



Single Tree:

$$w_i(\mathbf{f}) = \frac{I\left(\mathbf{f}_i^{\mathrm{tr}} \in \mathcal{R}_{l(\mathbf{f},\theta)}\right)}{\sum_{j=1}^{N_{\mathrm{tr}}} I\left(\mathbf{f}_j^{\mathrm{tr}} \in \mathcal{R}_{l(\mathbf{f},\theta)}\right)} \quad (12)$$

Tree Ensemble:

$$w_i(\mathbf{f}) = \frac{1}{k}\sum_{b=1}^{k} w_i(\mathbf{f}, \theta_b) \quad (13)$$

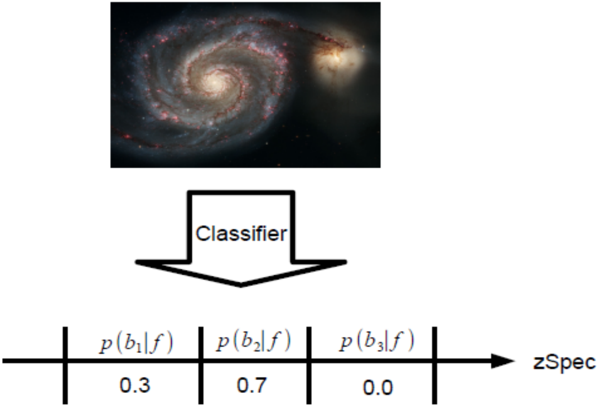
θ: parametrizes how the tree was grown

The Highest Weight Element

- Useful "by-product" from Quantile Regression Forest
  - Use weights from Quantile Regression Forest
  - Select the spectroscopic redshift with the highest weight
  - z_i^{spec} for $\max\left(w_i(\mathbf{f})\right)$
- Similar to Nearest Neighbour estimator
- Single floating point number per object

$\rightarrow$ Very efficient estimator for sample redshift PDF

Classification for PhotoZ PDFs



How do we estimate the weights $w_i(\mathbf{f})$?

- Idea: Bin the redshift range and use a probabilistic classifier to reconstruct the PDF. (Schapire et al. 2002, Frank et al. 2009)

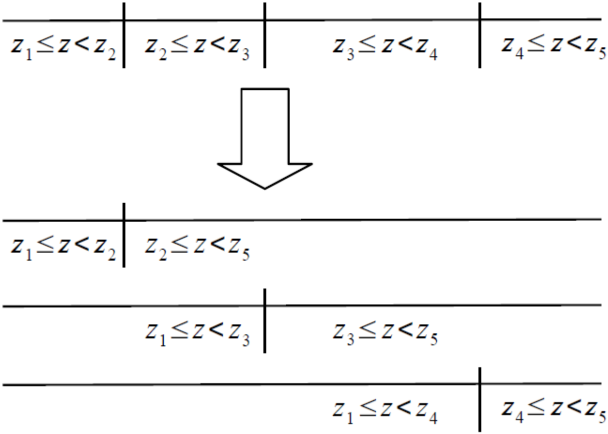
$$w_i(\mathbf{f}) = \frac{\hat{p}(b_i|\mathbf{f})}{n_{b_i}} \tag{14}$$

  - b_i: Index denoting the bin
  - $\hat{p}(b_i|\mathbf{f})$: probability that the redshift of an object with photometry $\mathbf{f}$ falls into bin b_i.
  - n_{b_i}: number of training set objects in bin b_i

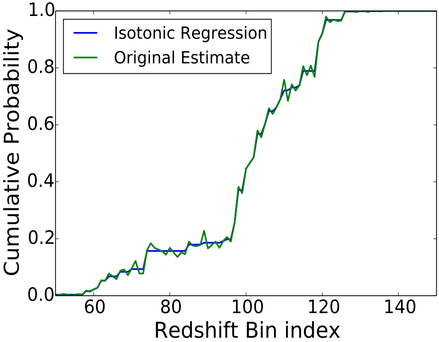
Ordinal Classification

- Idea: Treat the binned redshift as an ordinal scale variable to improve classification. (Frank et al. 2001)
- Nominal Classes:
  $p(\mathrm{Temp} = \mathrm{Cool}|\mathbf{x})$, $p(\mathrm{Temp} = \mathrm{Mild}|\mathbf{x})$, $p(\mathrm{Temp} = \mathrm{Hot}|\mathbf{x})$
- Ordinal Classes:
  $p(\mathrm{Temp} > \mathrm{Cool}|\mathbf{x})$, $p(\mathrm{Temp} > \mathrm{Mild}|\mathbf{x})$
- Recover Class probabilities:
  $p(\mathrm{Temp} = \mathrm{Cool}|\mathbf{x}) = 1 - p(\mathrm{Temp} > \mathrm{Cool}|\mathbf{x})$,
  $p(\mathrm{Temp} = \mathrm{Hot}|\mathbf{x}) = p(\mathrm{Temp} > \mathrm{Mild}|\mathbf{x})$,
  $p(\mathrm{Temp} = \mathrm{Mild}|\mathbf{x}) = p(\mathrm{Temp} > \mathrm{Cool}|\mathbf{x}) - p(\mathrm{Temp} > \mathrm{Mild}|\mathbf{x})$

Application to binned redshift



Calibrating Class Probabilities



- Errors in classification
- Monotonicity of cumulative probability not guaranteed
- Calibrate using Isotonic (Monotonic) Regression

Recapitulation

- The photometric redshift PDF for a new object is estimated from the training set

$$\left\{ w_i(\mathbf{f}), z_i^{\mathrm{spec}} \right\} \quad (15)$$

- The weights $\{w_i(\mathbf{f})\}$ are estimated using:
  - Quantile Regression Forest (QRF)
  - Not Ordinal (nominal) Classification PDF estimate (NOCP)
  - Ordinal Classification PDF estimate (OCP)
- The Highest Weight Element (HWE) is a single floating point estimate for the stacked redshift PDF
- Find density estimate for the weighted spectroscopic redshifts in the training set

Density Estimation

- Kernel Density Estimation

$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{N_{\mathrm{tr}}} w_i(\mathbf{f}) \mathcal{N}\left(z, \mu = z_i^{\mathrm{spec}}, \sigma = h\right) \qquad (16)$$

  - Select bandwidth h

- Density Estimation using Gaussian Mixture Models

$$\hat{p}(z|\mathbf{f}) = \sum_{i=1}^{K} \alpha_i(\mathbf{f}) \mathcal{N}(z, \mu_i(\mathbf{f}), \sigma_i(\mathbf{f})) \qquad (17)$$

  - Select number of mixture components K
  - Fit mixture components to weighted data

Bandwidth Selection

- 'Scott' Bandwidth:

$$\hat{\sigma}_{\mathrm{Scott}} = a \frac{\hat{\sigma}}{N_{\mathrm{tr}}^{1/5}} \tag{18}$$

- 'Hjort' Bandwidth:

$$\hat{\sigma}_{\mathrm{Hjort}} = a \frac{\hat{\sigma}}{N_{\mathrm{tr}}^{1/4}} \tag{19}$$

- Standard Deviation

$$\hat{\sigma}^2(\mathbf{f}) = \sum_{i=1}^{N_{\mathrm{tr}}} w_i(\mathbf{f}) \left(z_i^{\mathrm{spec}} - \hat{z}_{\mathrm{phot}}(\mathbf{f})\right)^2 \tag{20}$$

- Select the factor a using the validation set

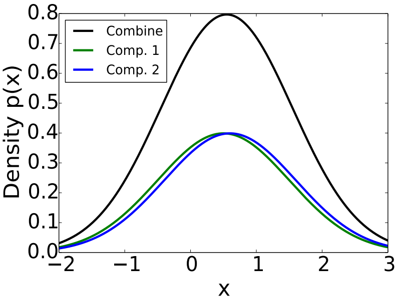
Gaussian Mixture Model

Motivation: Sparse parametrization

- More efficient (i.e. 5 floating point numbers per object)
- Easier to interpret

- Fit the parameters $\alpha_i(\mathbf{f})$, $\mu_i(\mathbf{f})$ and $\sigma_i(\mathbf{f})$ to the weighted data $\left\{w_i(\mathbf{f}), z_i^{\mathrm{spec}}\right\}$
- Fix a maximum number of mixture components K_{max} using the validation set
- Select the number of components $0 < K \leq K_{\mathrm{max}}$ on a per-object basis that minimizes the normalized entropy criterion

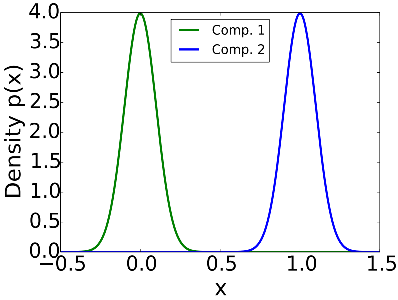
Normalized Entropy Criterion (Celeux & Soromenho 1996)



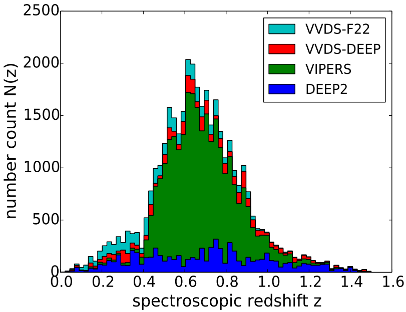
Minimize

$$NEC(K) = \frac{E(K)}{L(K) - L(1)} \quad (21)$$

- $E(K)$: Entropy measures "overlap" between components
- $L(K)$: maximum weighted log-likelihood for the model
- $L(K)$ (increasing in K) balanced by $E(K)$ (favours less overlap between components)

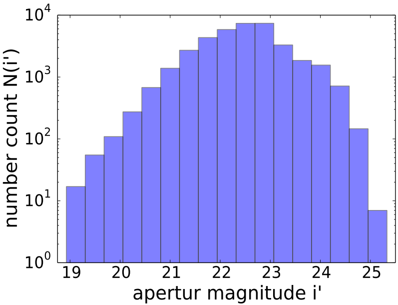


Dataset

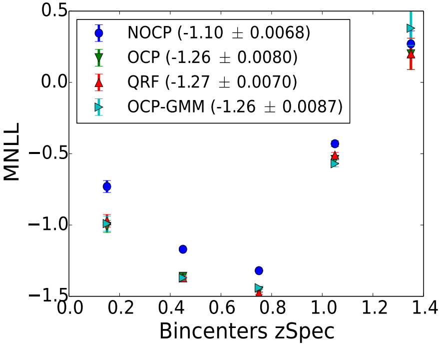


- Subsample from CFHTLS Wide (Brimioulle et. al. 2013)
- 5 band photometry (u^*, g', r', i', z')

- 31183 objects with $i' \leq 22.5$
- 6561 objects with $22.5 < i' \leq 24.5$



Redshift Conditional PDF

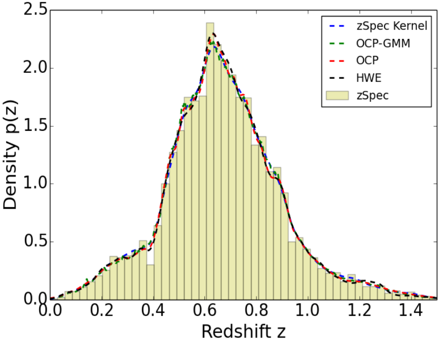


- Minimize mean negative log-likelihood loss (MNLL)

$$\mathrm{MNLL} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\hat{p}(\mathbf{z}_i|\mathbf{f}_i)\right) \tag{22}$$

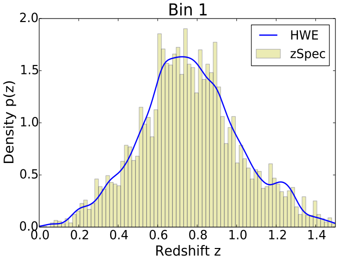
- Ordinal Classification improves performance
- Gaussian Mixture density estimate competitive with kernel and more efficient
  → 5 floating point numbers per object

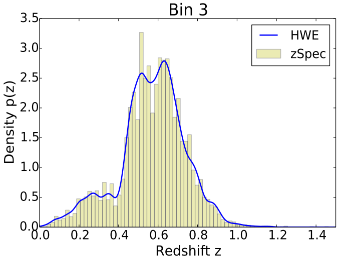
Stacked Redshift PDF

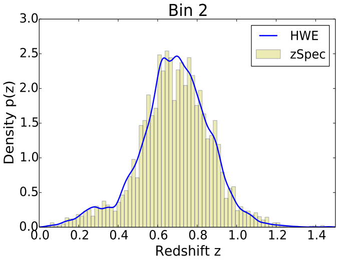


- Highest Weight Element accurate estimator of the redshift sample PDF
- z_i^{spec} associated with $\max(w_i(\mathbf{f}))$
- Efficient: 1 floating point number per object

Magnitude Selected Samples







- Selected on scaled i' band flux by equal frequency binning

$$f_{\mathrm{scaled},i'} = \frac{f_{i'} - \mu_{i'}}{\sigma_{i'}} \qquad (23)$$

  - $\mu_{i'}$: average flux
  - $\sigma_{i'}$: standard deviation of all fluxes (NOT flux error)

Conclusions

- Highest Weight Element accurately estimates the redshift sample PDF (1 floating point number)
- Ordinal classification improves classification accuracy
- Gaussian Mixtures very efficient (5 floating point numbers) for PDF estimation
- Point predictions (i.e. conditional mean) don't provide enough information for many applications in cosmology
  $\rightarrow$ Currently actively explored by us