Probabilistic photometric redshifts in the era of Petascale Astronomy

Matías Carrasco Kind
NCSA/Department of Astronomy
University of Illinois at Urbana-Champaign

Tools for Astronomical Big Data
March 9th - 11th, 2015
The need of distances in cosmology

3D Clustering of galaxies as a probe in cosmology, e.g., 2 point correlation function, power spectrum of the galaxy distribution, etc.
Photometry vs. Spectroscopy

- **Photometry** focuses on measuring the brightness of objects at different wavelengths.
- **Spectroscopy** involves analyzing the spectrum of light to determine the composition and properties of objects.

**Graphical Representation:**
- Photometry data is shown as a set of images with varying colors.
- Spectroscopy data is illustrated with a prism dispersing light into its constituent wavelengths, highlighting specific absorption lines for Hydrogen (H), Sodium (Na), and Calcium (Ca).
Big data problem

It’s happening! 😊

∼ 300 millions galaxies up to $z = 1.5$

5,000 squares degrees (1/8 sky)

Data management at NCSA

DES specially designed to probe the origin of dark energy

S/G class and photo-z needed

1 TB of data per day

2 years completed, 3 more to go
Big data problem

It’s happening! 😊

~ 300 millions galaxies up to $z = 1.5$
5,000 squares degrees (1/8 sky)

Data management at NCSA

DES specially designed to probe the origin of dark energy

S/G class and photo-z needed

1 TB of data per day

2 years completed, 3 more to go
Data science challenge

Determine redshift using limited information. 8 points instead of thousands!
Motivation

- Photo-$\sim$ Probability Density Functions needed
- Several methods/codes to compute photo-$\sim$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good but for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools
Motivation

- Photo-$z$ Probability Density Functions needed
- Several methods/codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good but for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools
Motivation

- Photo-$z$ Probability Density Functions needed

- Several methods/codes to compute photo-$z$

- Need for a meta-algorithm that combines multiple techniques

- PDF are good but for large datasets, storage and I/O will be an issue

- Machine Learning and statistical tools
Motivation

- Photo-$z$ Probability Density Functions needed
- Several methods/codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good but for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools
Motivation

- Photo-$\sim$ Probability Density Functions needed
- Several methods/codes to compute photo-$\sim$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good but for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools
Photo-$z$ PDF estimation (in 5 min.)
Photo-$z$ PDF estimation: TPZ

- TPZ (Trees for Photo-Z) is a supervised machine learning code
- Prediction trees and random forest
- Incorporate measurements errors and deals with missing values
- Ancillary information: expected errors, attribute ranking and others
- Application to the S/G

http://lcdm.astro.illinois.edu/code/mlz.html

Carrasco Kind & Brunner 2013a (MNRAS, 432, 1483)
Photo-z PDF estimation: TPZ example

Tree from DES training data with more than 7000 nodes

colors represent depth only
Photo-$z$ PDF estimation: Random forest

Combine predictions from trees

Trees are ideally uncorrelated and strong
Bootstrapping and error sampling
Random features at each node
TPZ has been tested in several databases with remarkable results
Photo-$z$ PDF estimation: SOM

- SOM (Self Organized Map) is an unsupervised machine learning algorithm

- Competitive learning to represent data conserving topology

- 2D maps and Random Atlas

- Framework inherited from TPZ

- Application to the S/G

Carrasco Kind & Brunner 2014a (MNRAS, 438, 3409)
SOM topologies

Different topologies can be used with or without periodic boundary conditions

Carrasco Kind & Brunner 2014a (MNRAS, 438, 3409)
Photo-$z$ PDF estimation: BPZ

- BPZ (Benitez, 2000) is a Bayesian template fitting method to obtain PDFs
  - Set of calibrated SED and filters
  - Doesn’t need training data
  - Priors can be included

Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)
Photo-$z$ PDF combination
Bayesian framework

Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)
Bayesian framework

Our approach

Supervised method

Unsupervised method

Template fitting

Weighting scheme

$\Downarrow$

photo-$z$ PDF

Outliers

Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)
Photo-$z$ PDF combination: Results

- Several combination methods
- Bayesian model averaging (BMA) and combination (BMC) are the best
- Same applies to S/G (Kim, Brunner & CK in prep.)
Naïve Bayes Classifier (same used for spam emails) to identify "spam" galaxies using information from multiple techniques

Each feature provides information about these two classes, and can be combined to make a stronger classifier
Photo-$\sim$ PDF representation and storage
Photo-$z$ PDF storage: Strategies

Single Gaussian fit

Multi-Gaussian fit

Monte Carlo sampling

Sparse representation techniques

Reduce number of points while increasing accuracy
Photo-$z$ PDF storage: Strategies

- Single Gaussian fit
- Multi-Gaussian fit
- Monte Carlo sampling
- Sparse representation techniques

Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)
Photo-$z$ PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N_{\text{original}}^2$ bases

With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs

Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)
Use Gaussian and Voigt profiles as bases, need $N_{\text{original}}^2$ bases.

With only 10-20 bases achieve 99.9 % accuracy.

Use 32-bits integer per basis, compression.

Store Multiple PDFs.

---

Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)
Photo-$\sim$ PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N_{\text{original}}^2$ bases

With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs

Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)
Use Gaussian and Voigt profiles as bases, need $N_{\text{original}}^2$ bases

With only 10-20 bases achieve 99.9% accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs

Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)
$N(z)$ and sparse representation

By definition:

$$N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) \, dz$$
$N(z)$ and sparse representation

By definition:

$$N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz$$

Using sparse representation, we represent each PDF $p_{z_k}$ as:

$$p_{z_k} \approx D \cdot \delta_k$$

$D$ is the dictionary, $\delta_k$ is the sparse vector, then

$$N(z) = \sum_{k=1}^{N} \delta_k \cdot \int_{z-\Delta z/2}^{z+\Delta z/2} D dz$$

by precomputing:

$$\delta_N = \sum_{k=1}^{N} \delta_k$$

$$I_D(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} d_j dz \quad j = 1, 2, \ldots, m$$

Only bases are integrated
**N(z) and sparse representation**

By definition:

\[
N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz
\]

Using sparse representation, we represent each PDF \( p_{z_k} \) as:

\[
p_{z_k} \approx D \cdot \delta_k
\]

\( D \) is the dictionary, \( \delta_k \) is the sparse vector, then

\[
N(z) = \sum_{k=1}^{N} \delta_k \cdot \int_{z-\Delta z/2}^{z+\Delta z/2} D dz
\]

by precomputing:

\[
\delta_N = \sum_{k=1}^{N} \delta_k
\]

\[
I_D(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} d_j dz \quad j = 1, 2, \ldots, m
\]

\( N(z) \) is reduced to a simple dot product

\[
N(z) = I_D(z) \cdot \delta_N
\]
Talk on Friday (ad)

- Machine Learning in DES
- Photo-$z$ in DES early data
- Photo-$z$ PDF in DESDM
- New tools to access these from DB
Conclusions

✔ Compute photo-z PDF

✔ Combine PDFs efficiently
Better than individual, outliers identification (arXiv:1403.0044)

✔ PDF Sparse Representation
99.9% accuracy in P(z) and N(z) with 15 points (arXiv:1404.6442)

✔ Uses of these tools!
Clustering, weak lensing, DES, DESDM, etc...
Questions?

Matias Carrasco Kind
NCSA/UIUC
mcarras2@ncsa.illinois.edu
http://matias-ck.com/
https://github.com/mgckind
For PDFs with less than 4 peaks 5-10 points should be sufficient.

Sparse representation gives more accurate and more compressed representation for \( N(z) \), 99.9\% accuracy with 15 points (200 points originally).
Combination of Gaussian and Voigt profiles

Covering the whole redshift space, at each location we have several bases
Out of Bag (cross-validation) data used to validate trees/maps. Changes for every tree/map and is not used during training. We can learn from the cross-validation data!
Photo-$z$ PDF estimation: Error and validation

Out of Bag (cross-validation) data used to validate trees/maps
Changes for every tree/map and is not used during training
We can learn from the cross-validation data!

Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)
Photo-$\sim$ PDF estimation: SOM 2D toy example

Suppose 2D data distributed in a given space.

De-project the data in a 2D map.

Each cell will contain objects with similar properties.
Photo-$\tilde{z}$ PDF estimation: SOM 2D toy example

Suppose 2D data distributed in a given space.

De-project the data in a 2D map.

Each cell will contain objects with similar properties.