ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

How to produce, combine, store and use photo-z PDFs

Matías Carrasco Kind

Department of Astronomy
University of Illinois

DES meeting at UIUC May 19-23, 2014

Motivation

- Photo-z PDF are important in cosmology
- 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 PDF are important in cosmology
- 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 PDF are important in cosmology
- 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 PDF are important in cosmology
- 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

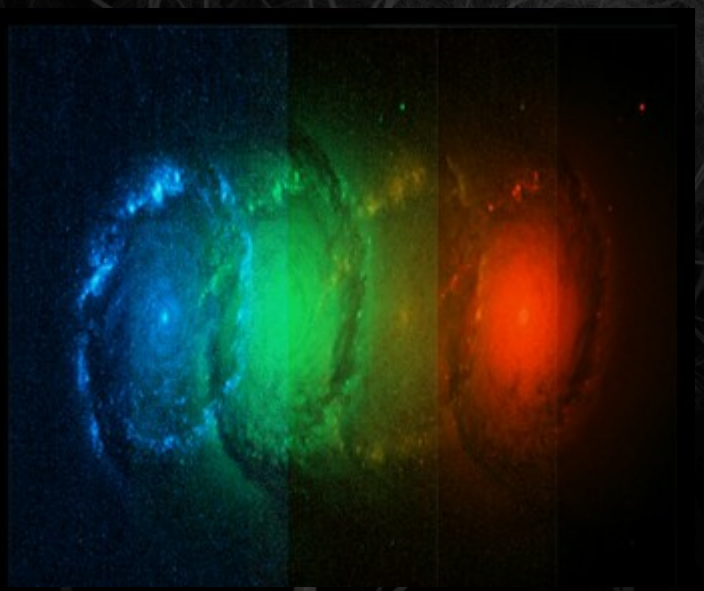
- Photo-z PDF are important in cosmology
- Several! methods/codes to compute photo-z
- Need for a meta-algorithm that combines multiple techniques
- PDF are good **but** for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools

Motivation

- Photo-z PDF are important in cosmology
- 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

How to produce, combine, store and use P(z)

Photo-z PDF estimation



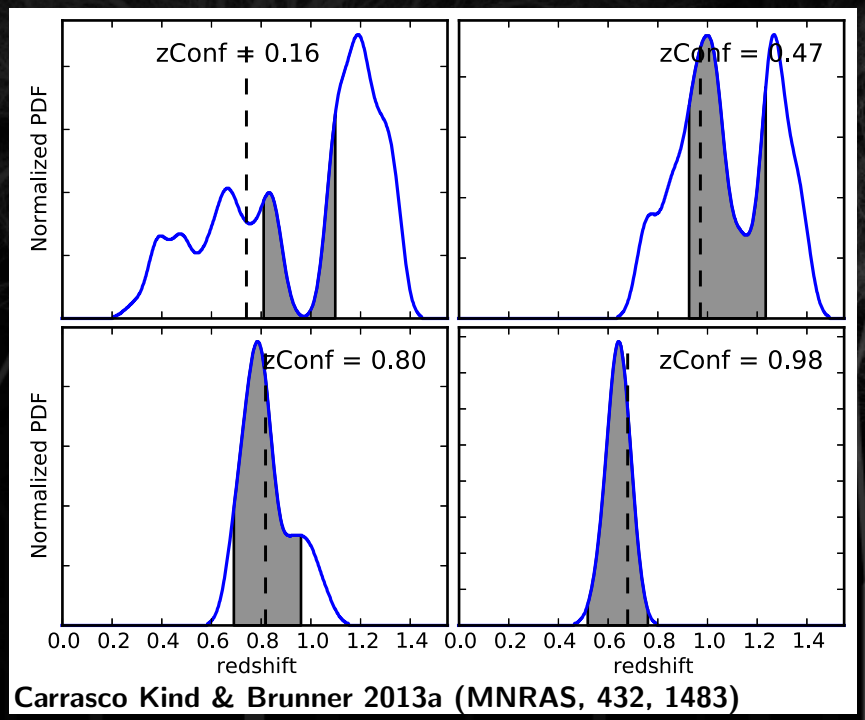
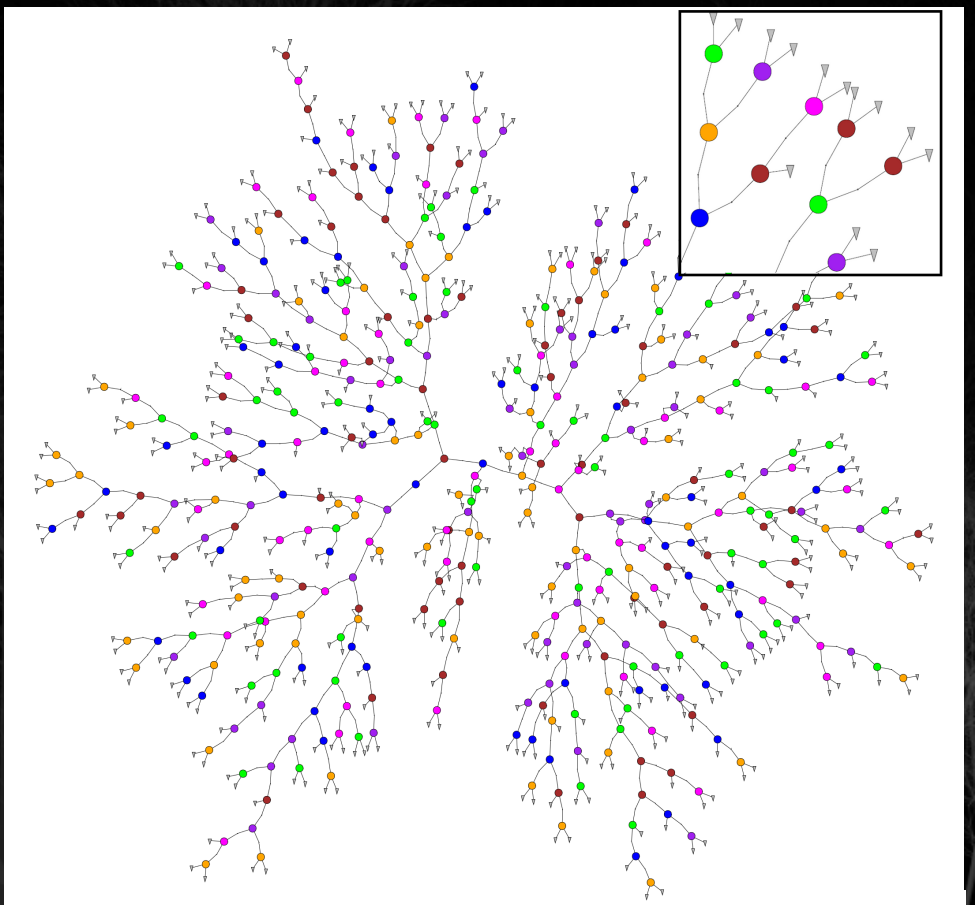


Photo-z PDF estimation: TPZ

- 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

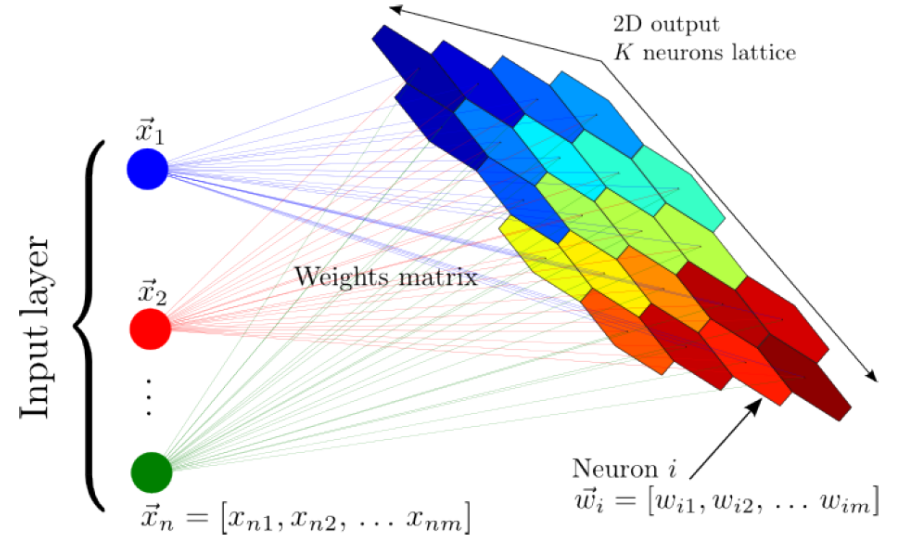


Carrasco Kind & Brunner 2013a (MNRAS, 432, 1483)

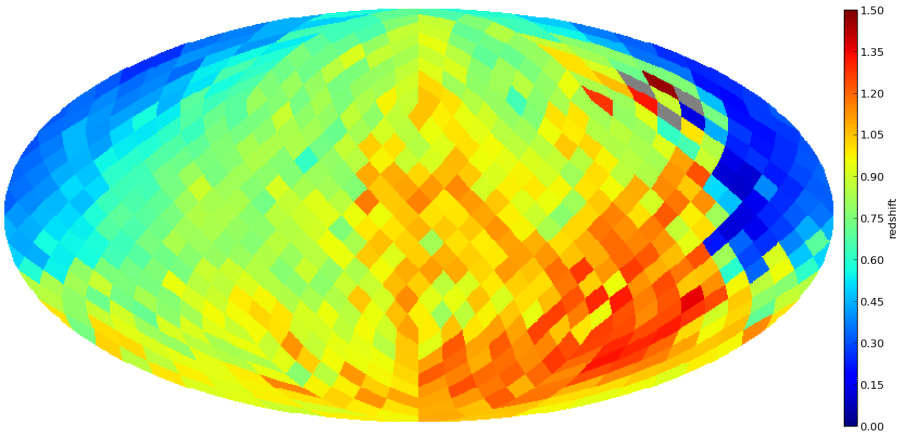
http://lcdm.astro.illinois.edu/code/mlz.html

Photo-z PDF estimation: SOM

- SOM(Self Organized Map) is a 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)



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

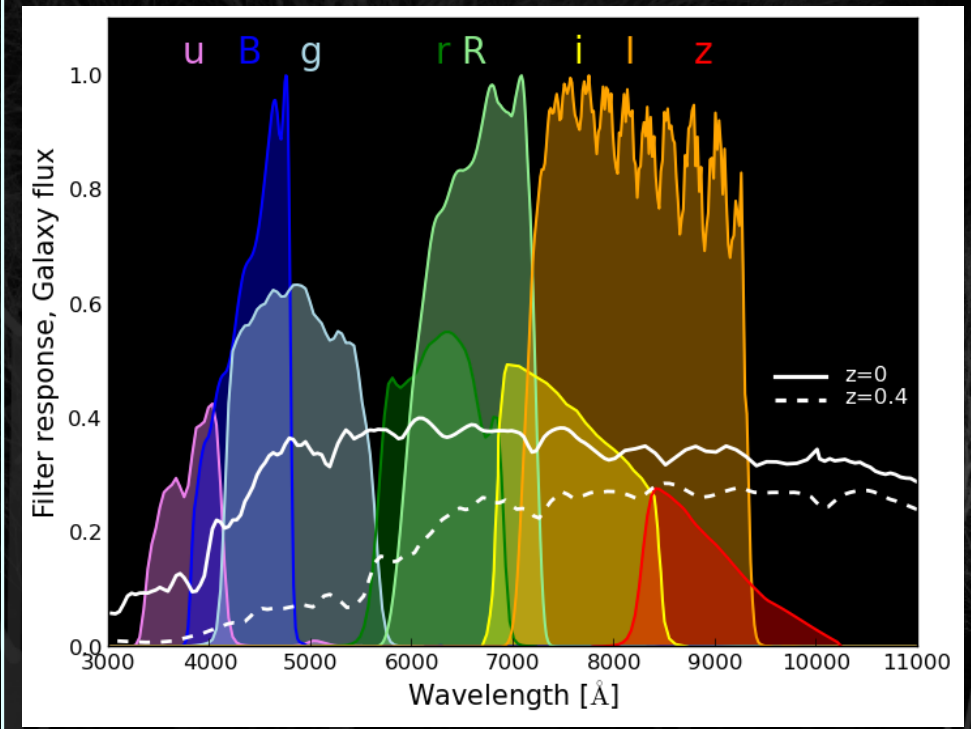


Photo-z PDF estimation: MLZ

MLZ :Machine Learning for photo-Z

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

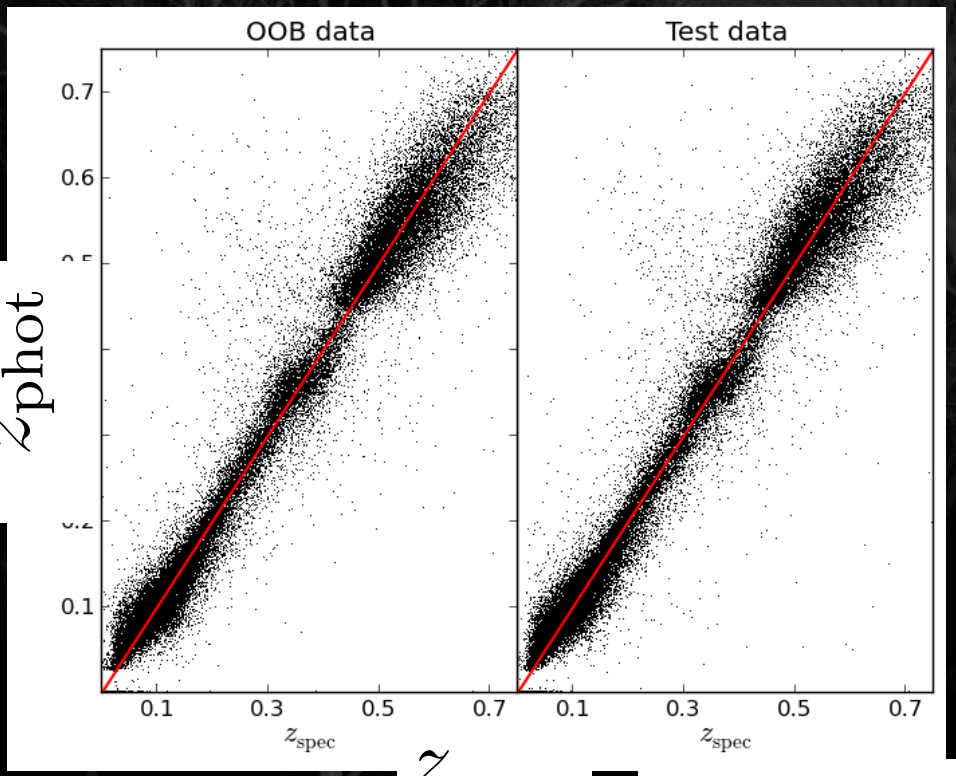
- TPZ, SOM and BPZ incorporated in one python framework, more can be added
- Public, parallel and easy to use
- PDF Sparse representation included
- Current version 1.1, GitHub repository (https://github.com/mgckind/MLZ)
- pycuda and numba in folder

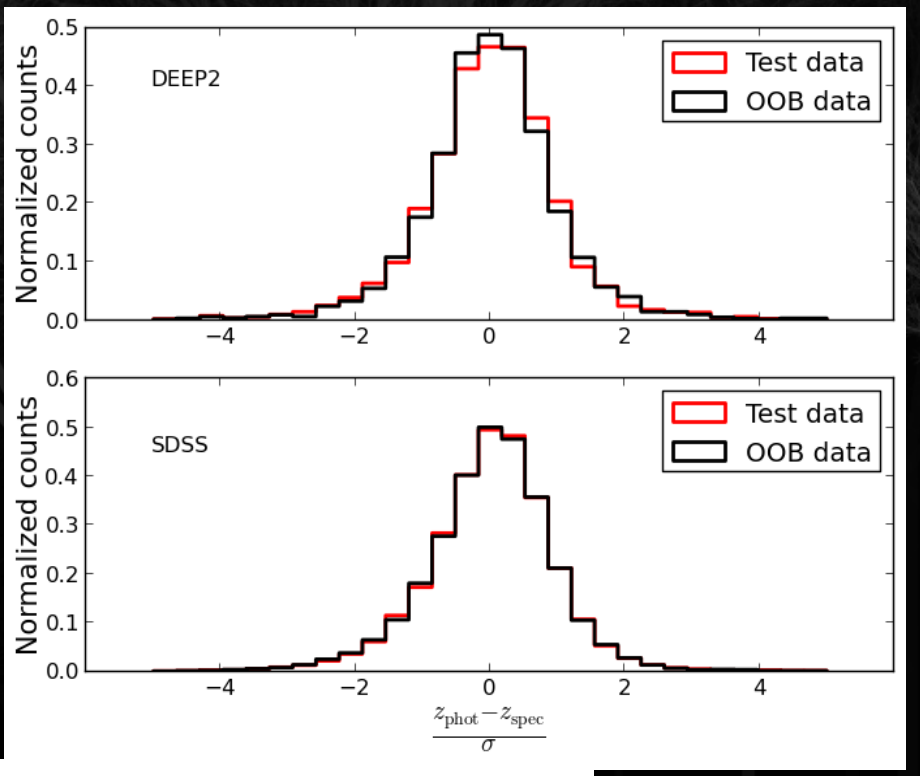
Photo-z PDF estimation: Error and validation

Out of Bag 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!

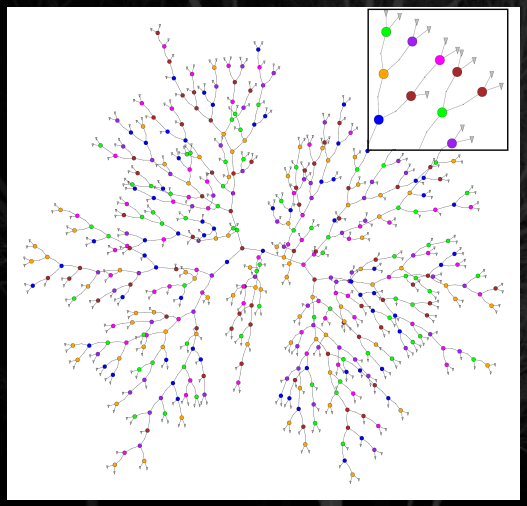




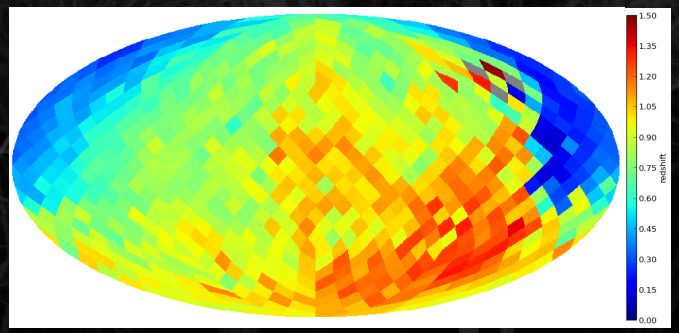
Carrasco Kind & Brunner 2014c (MNRAS submitted)

How to produce, combine, store and use P(z)

Photo-z PDF combination



+



+

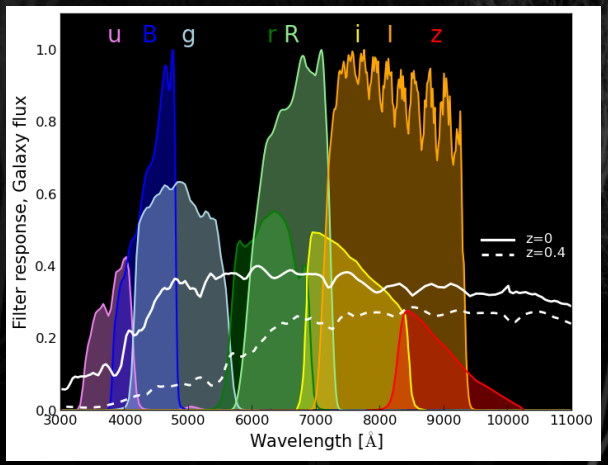
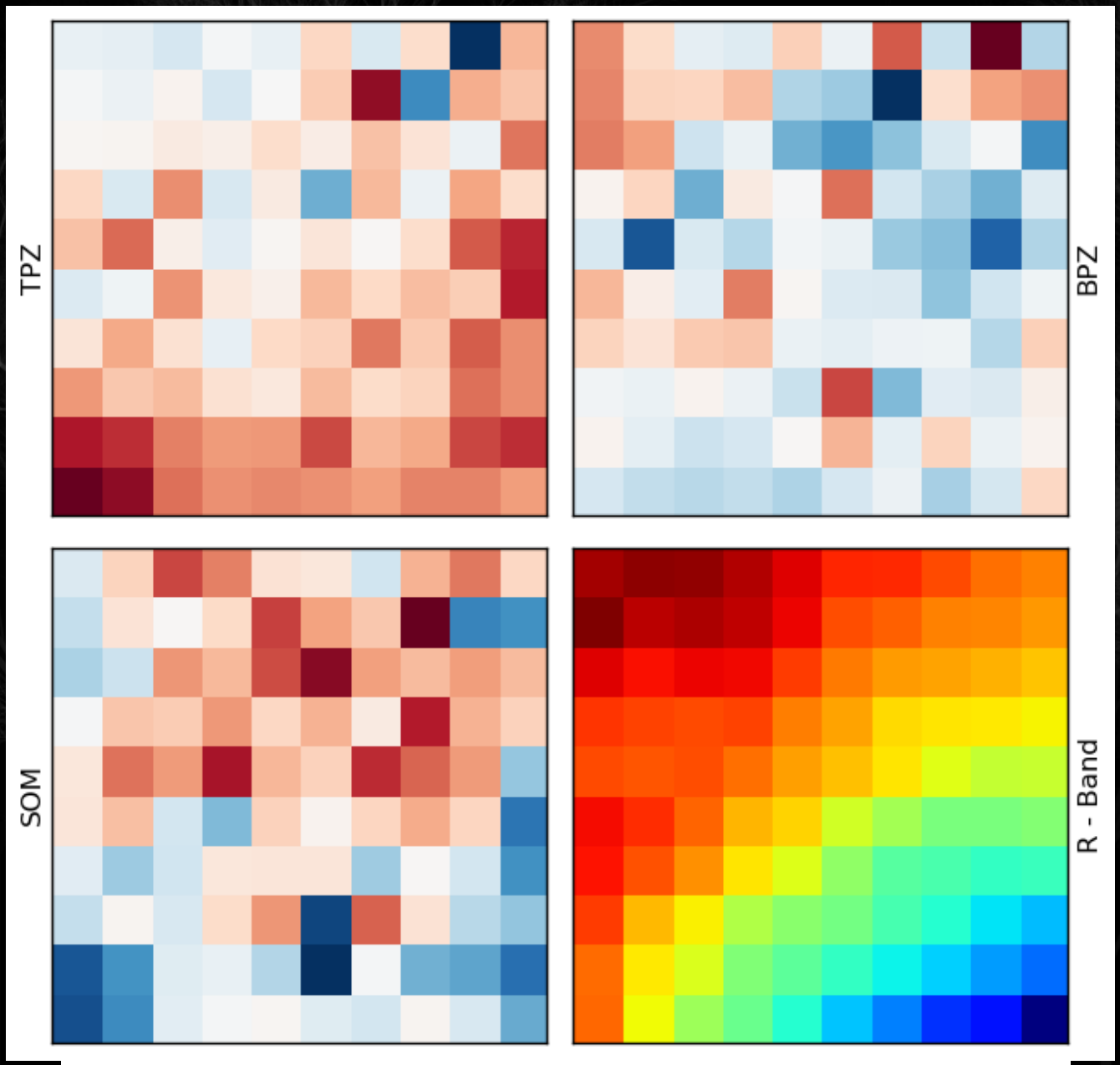


Photo-z PDF combination: Bayesian framework



Carrasco Kind & Brunner 2014c (MNRAS submitted)

This approach

Supervised method

+

Unsupervised method

+

Template fitting

+

Weigthing scheme

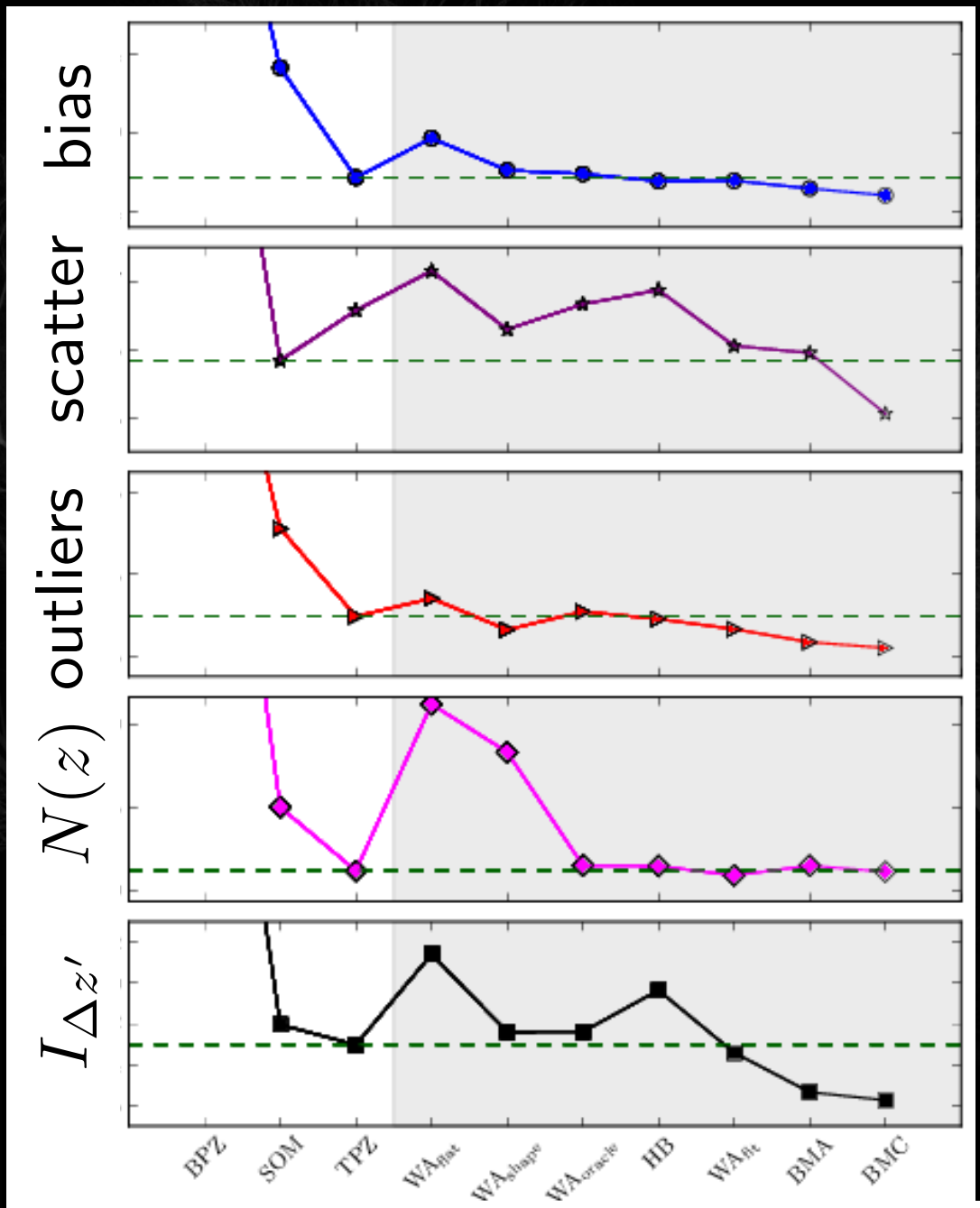
⇓

photo-z PDF

+

Outliers

Photo-z PDF combination: Results



Carrasco Kind & Brunner 2014c (MNRAS submitted)

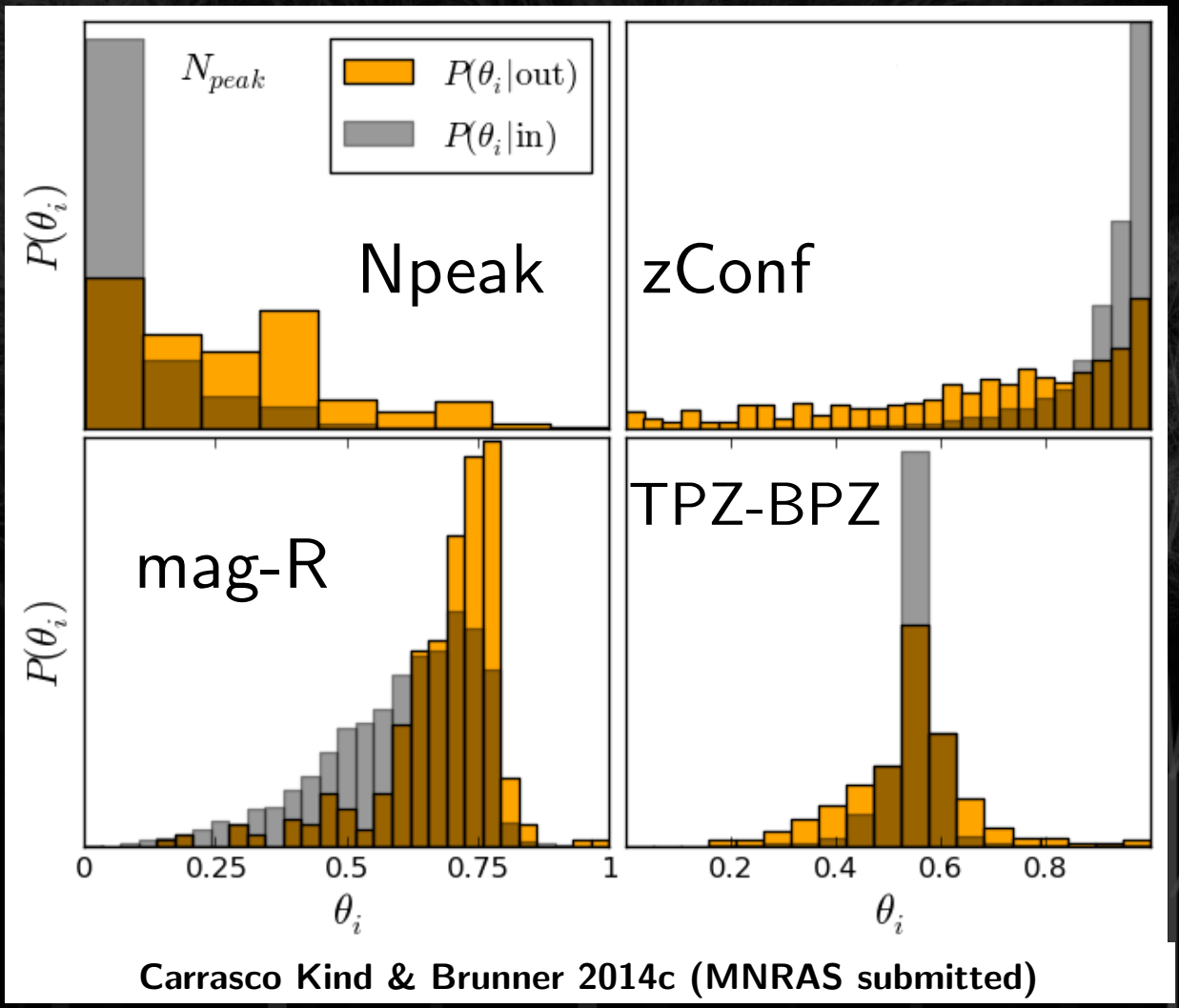
- Several combination methods
- Bayesian model averaging (BMA) and combination (BMC) are the best
- We introduce the I-score which combine multiple metics after being rescaled to compare different methods and/or codes

$$I_{\Delta z'} = \sum w_i M_i$$

Photo-z PDF combination: Outliers

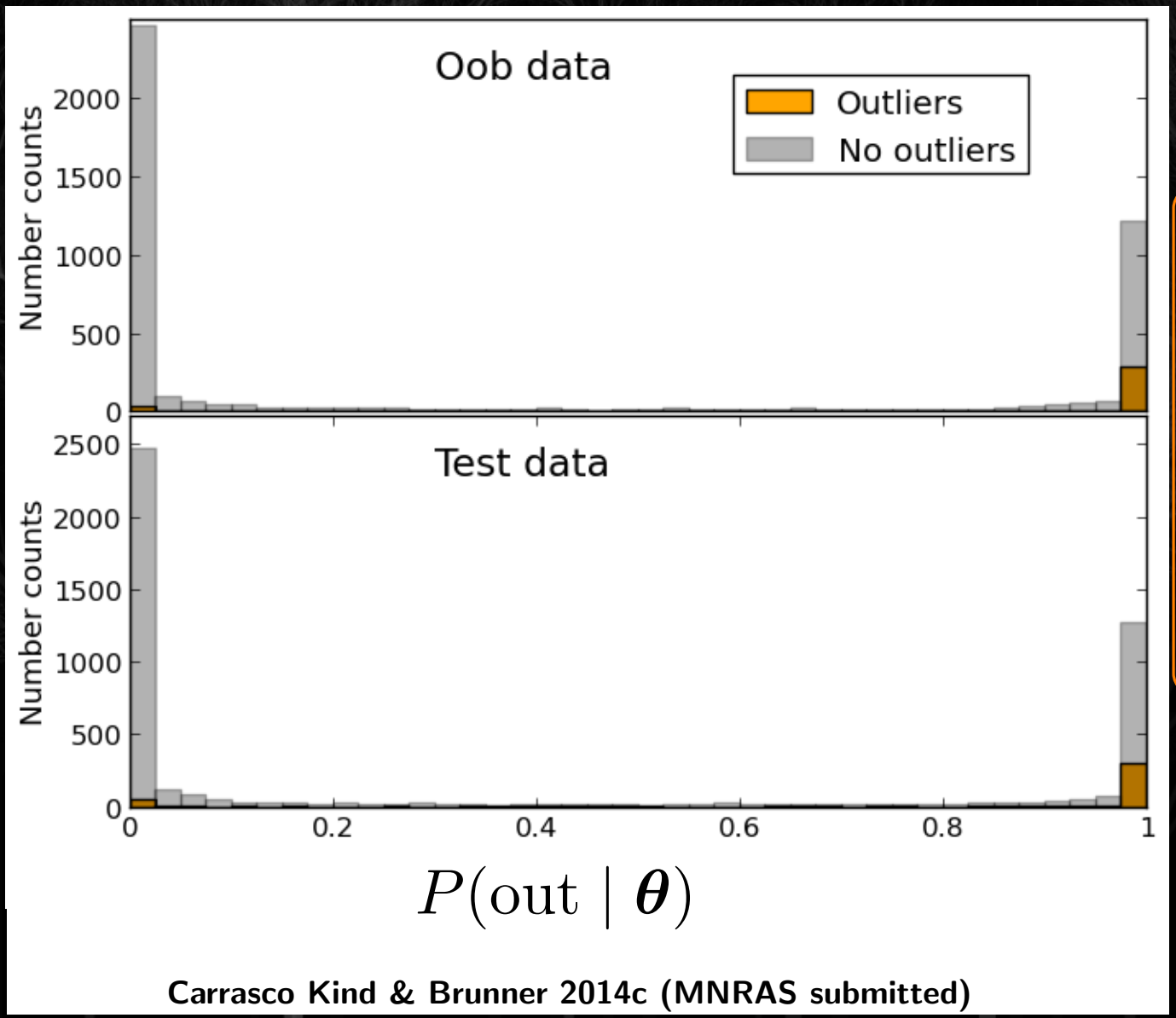
Naïve Bayes Classifier (same used for spam emails) to identify "spam" galaxies using information from multiple techniques

Each feature provides information about these two classes, and can be combined to make a stronger classifier



Carrasco Kind & Brunner 2014c (MNRAS submitted)

Photo-z PDF combination: Outliers



Carrasco Kind & Brunner 2014c (MNRAS submitted)

- Highly bimodal
- Little contamination
- Good discriminant
- Consistent between samples

Photo-z PDF storage

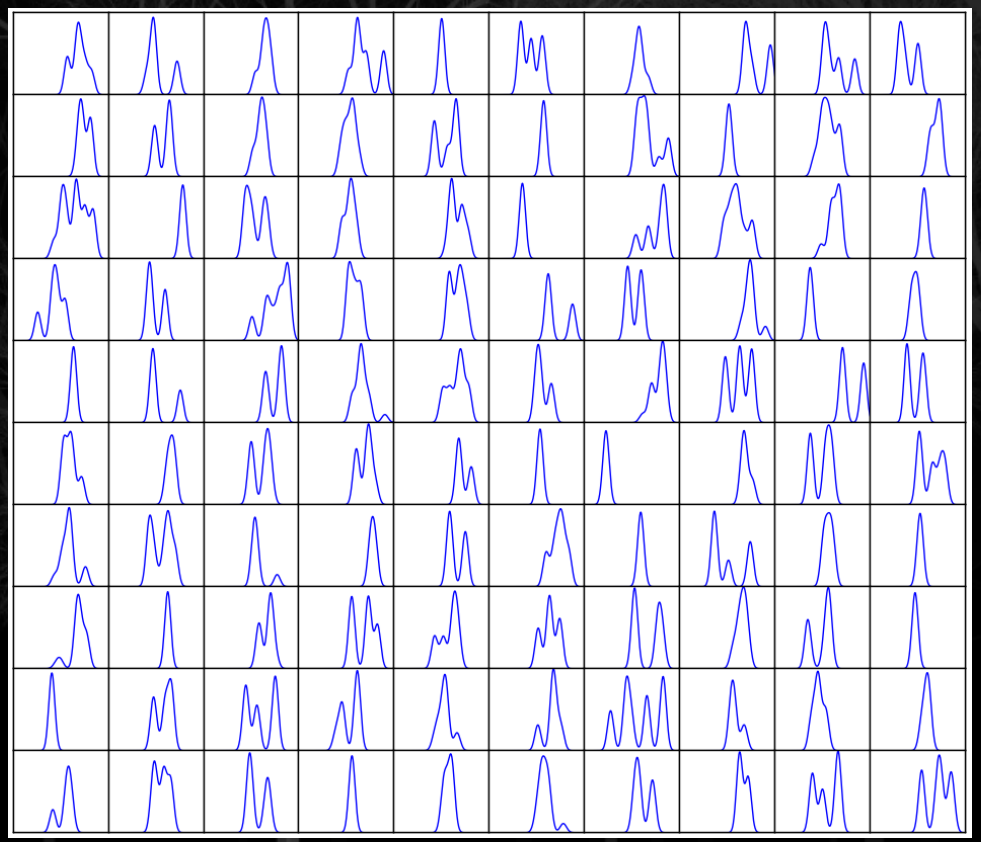


Photo-z PDF storage: Strategies

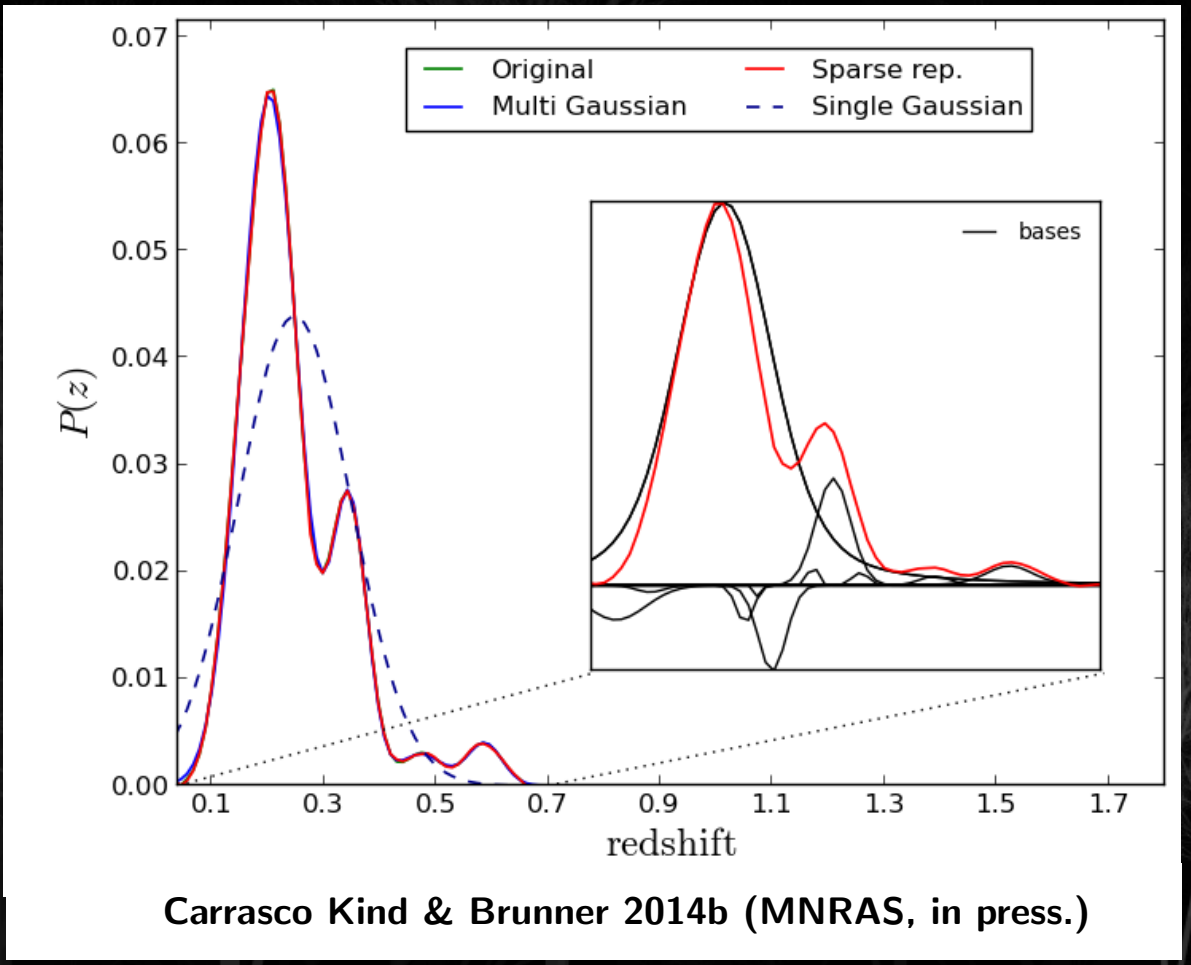
Single Gaussian fit

Multi-Gaussan fit

Monte Carlo sampling

Sparse representation techniques

Carrasco Kind & Brunner 2014b, MNRAS in press, arxiv: 1404.6442



Carrasco Kind & Brunner 2014b (MNRAS, in press.)

Photo-z PDF storage: Sparse representation

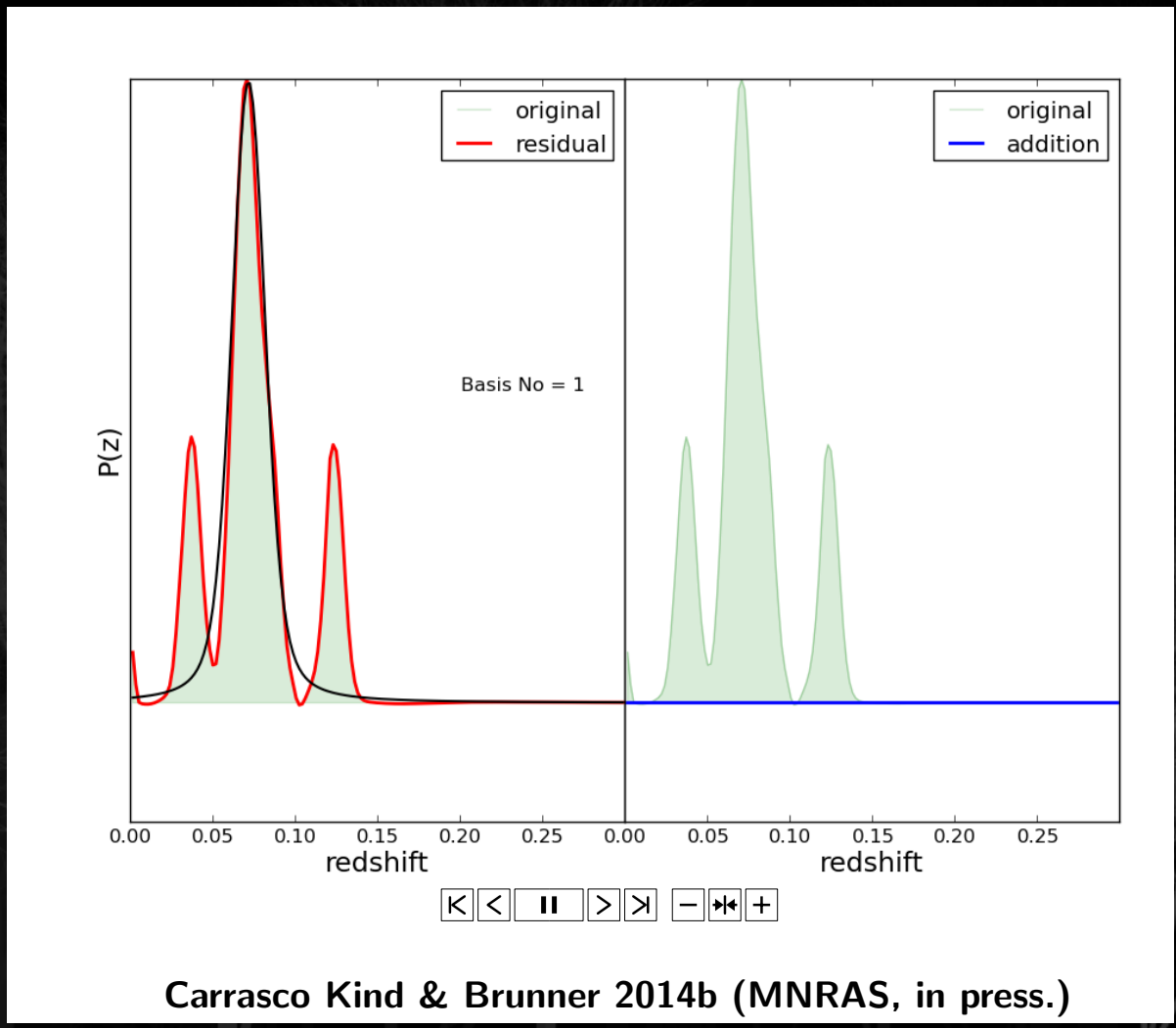
Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

Find basis and amplitud to reduce residual on each step

With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs



Carrasco Kind & Brunner 2014b (MNRAS, in press.)

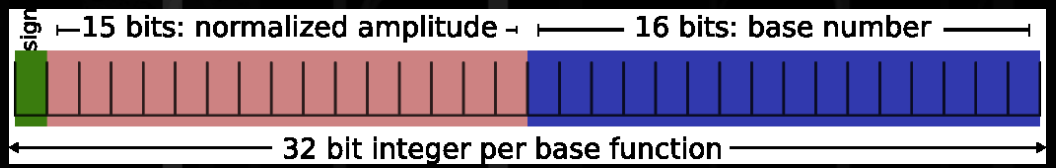
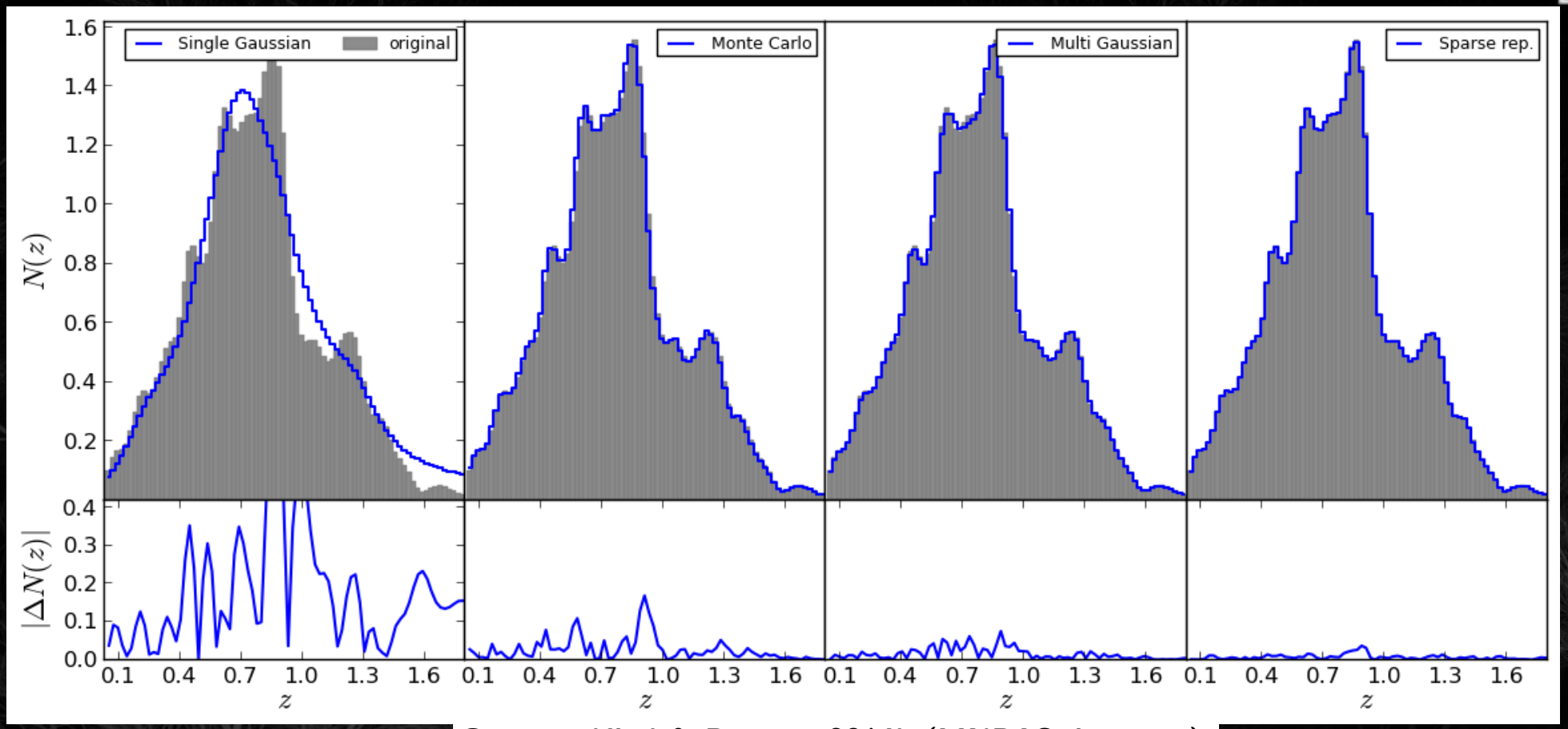


Photo-z PDF storage: Results



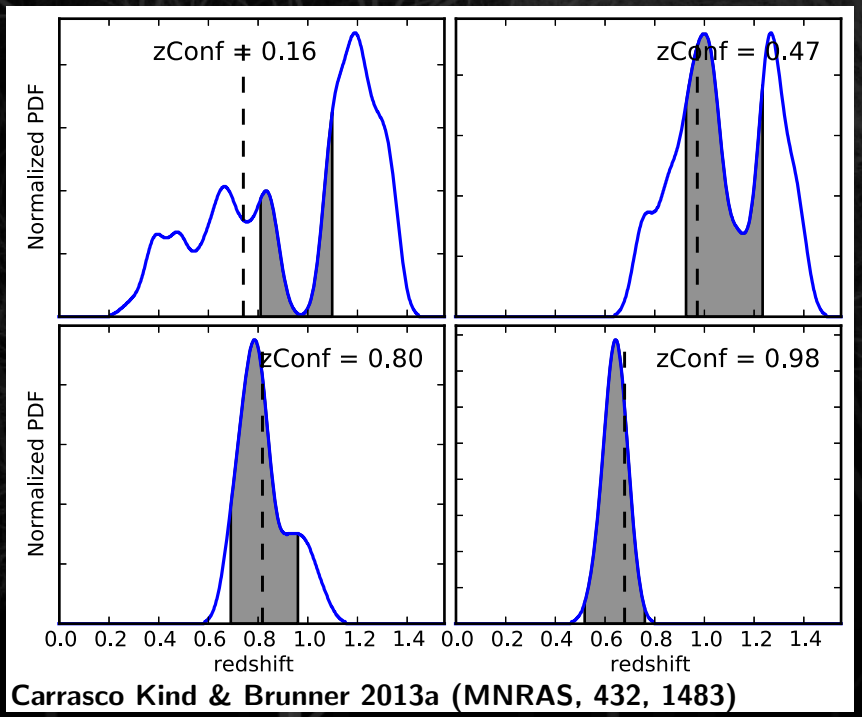
Carrasco Kind & Brunner 2014b (MNRAS, in press.)

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)

How to produce, combine, store and use P(z)

Photo-z PDF applications



Carrasco Kind & Brunner 2013a (MNRAS, 432, 1483)

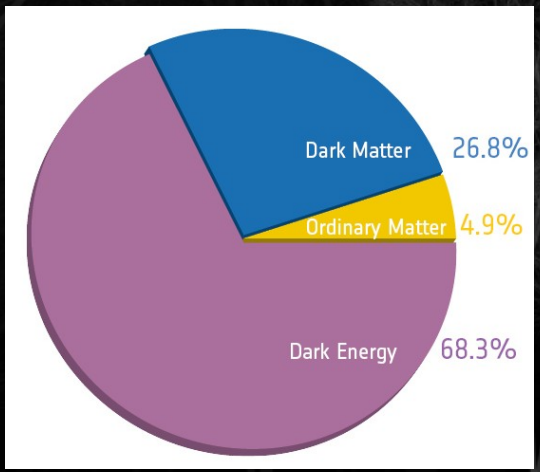


Photo-z PDF application: $N(z)$

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 pz_k as:

$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k$ $\mathbf{D}$ is the dictionary, $\boldsymbol{\delta}_k$ is the sparse vector, then

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

Only bases are integrated

by precomputing:

$$\boldsymbol{\delta}_N = \sum_{k=1}^{N} \boldsymbol{\delta}_k \qquad \mathbf{I_D}(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{d}_j dz \qquad j = 1, 2, \ldots, m$$

$N(z)$ is reduce to a simple dot product

$$\boxed{N(z) = \mathbf{I_D}(z) \cdot \boldsymbol{\delta}_N}$$

Photo-z PDF application: $N(z)$

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 pz_k as:

$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k$ $\mathbf{D}$ is the dictionary, $\boldsymbol{\delta}_k$ is the sparse vector, then

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

Only bases are integrated

by precomputing:

$$\boldsymbol{\delta}_N = \sum_{k=1}^{N} \boldsymbol{\delta}_k \qquad \mathbf{I_D}(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{d}_j dz \qquad j = 1, 2, \ldots, n$$

$N(z)$ is reduce to a simple dot product

$$\boxed{N(z) = \mathbf{I_D}(z) \cdot \boldsymbol{\delta}_N}$$

Conclusions

✓ Compute photo-z PDF

Individual techniques (MLZ; arXiv:1303.7269, arXiv:1312.5753)

✓ 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)

✓ Sparse rep. for cosmology

Use bases framework to speed things up (in prep.)

THANKS!

Questions?

Matias Carrasco Kind
University of Illinois
mcarras2@illinois.edu
https://sites.google.com/site/mgckind/