Photometric Redshifts with Random Forests

Alex Szalay

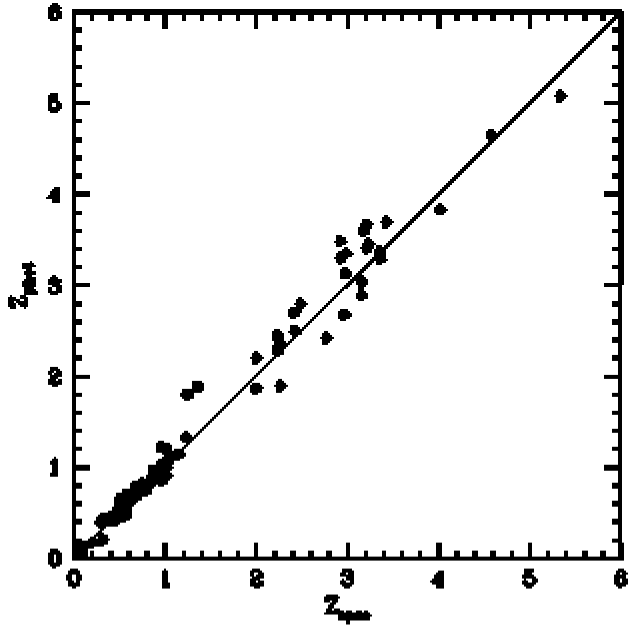
JHU / MPA

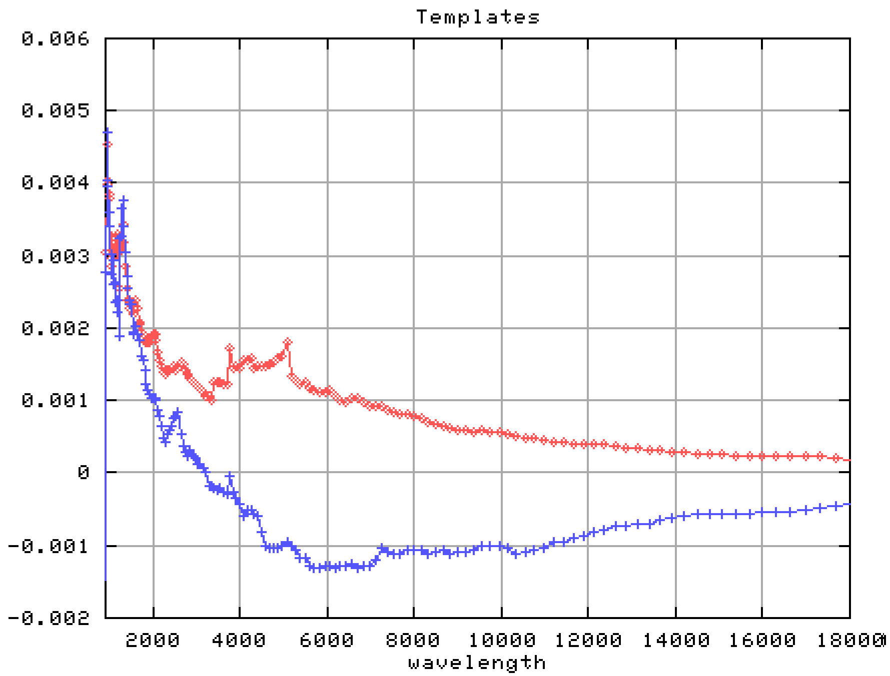
Photometric Redshift Techniques

Techniques

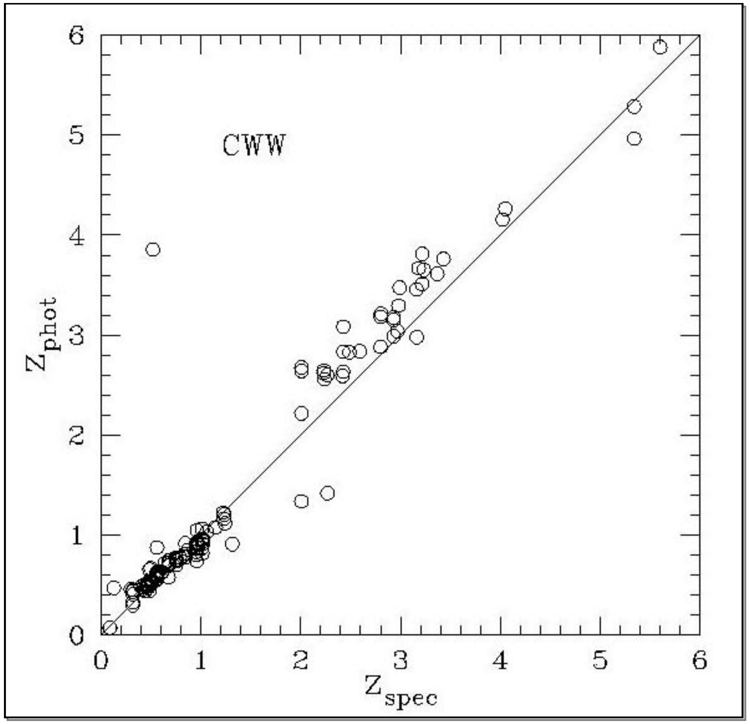
- Phenomenological (PolyFit, ANNz, kNN, RF)
  - *Simple, quite accurate, fairly robust*
  - *Little physical insight, difficult to extrapolate, M- bias*
- Template-based (KL, HyperZ…)
  - *Simple, physical model*
  - *Calibrations, templates, issues with accuracy*
- Hybrid (‘base learner’)
  - *Physical basis, adaptive*
  - *Complicated, compute intensive*

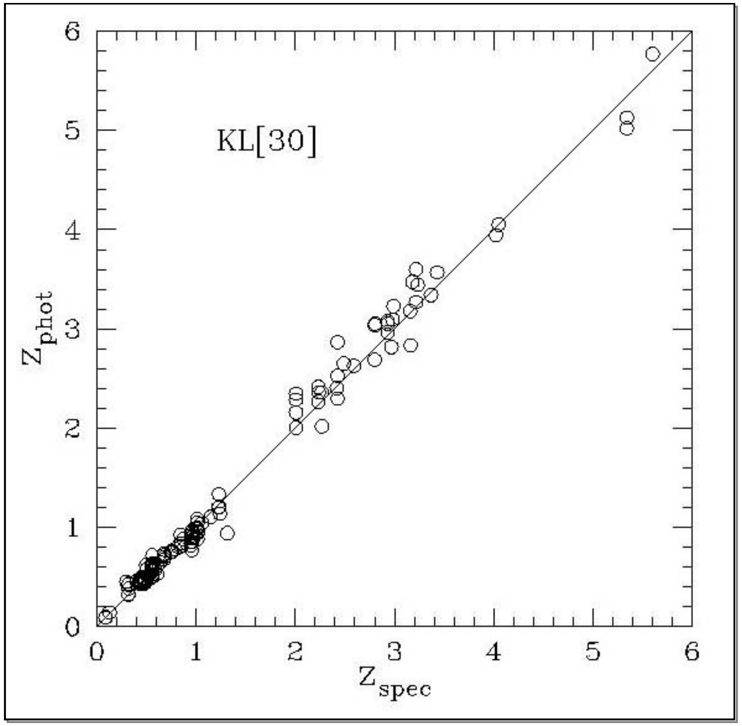
Training the Bases





Hubble Deep Field





initial

Hybrid + 30 iterations

Accuracy of SDSS PhotoZ

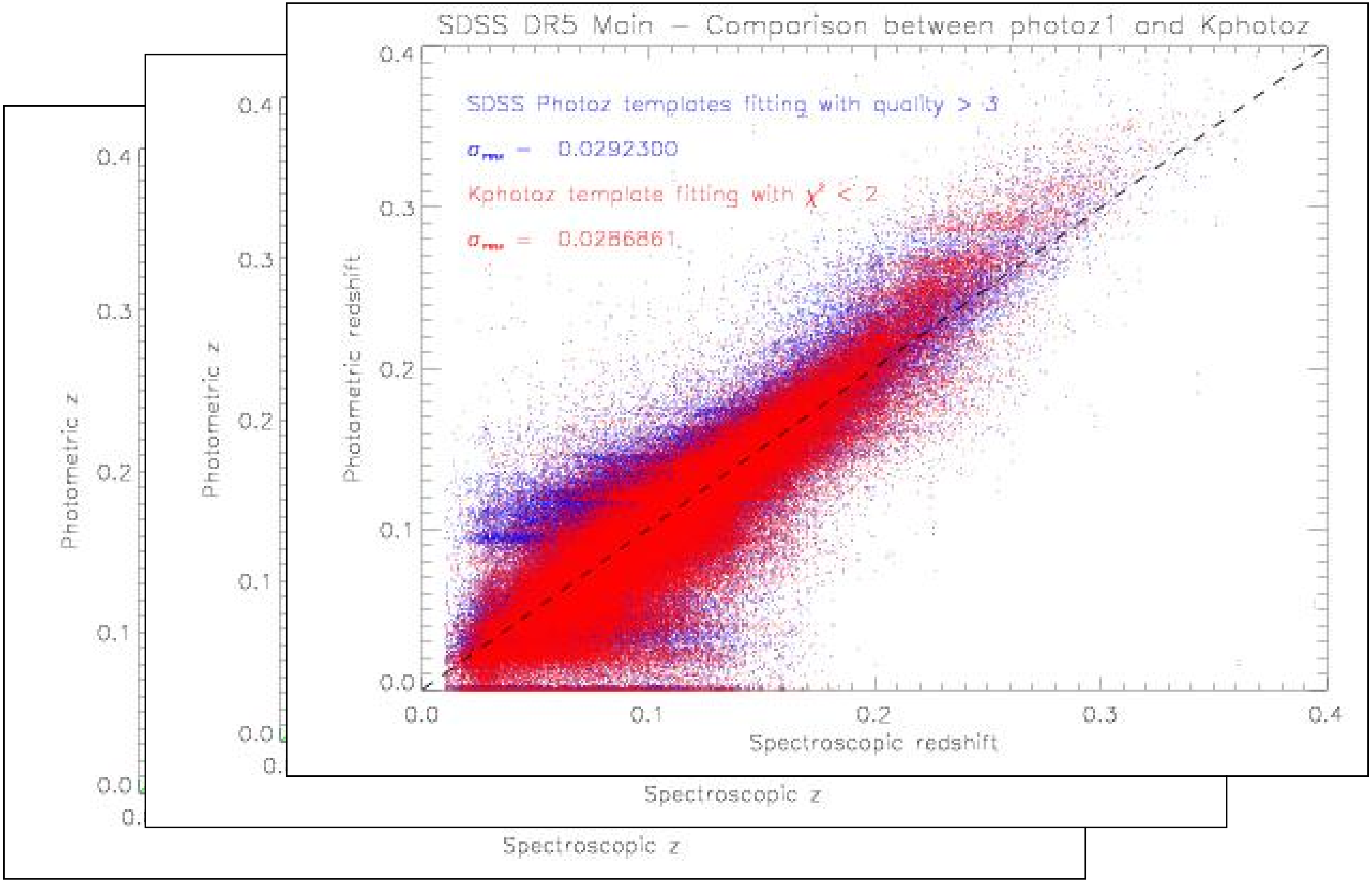
- At least 5 groups computed SDSS photoz
  - *JHU/Hungary, Fermilab, NYU, Lahav, Sussex*
- Comparison by Celine Eminian (Sussex)
- Most techniques perform at about the same level
  - *Getting to 0.025 easy, beyond it is getting hard*

	Main	***LRG***
Kphotoz(*)	0.028	0.022
ANNz	0.019	0.022
photoz1	0.029	0.025
photoz2	0.023	0.026

SDSS PhotoZ

- Spectro sample (670K unique galaxies in DR5):
  - *Main r_{pet}<17.77*
  - *LRG color cut, about 1 mag fainter, 5% of total*
- Photometry (132M primary galaxies)
  - *Out of these 21M is r_{pet}<20.77*
- Photoz for LRG is much better
- Currently two different versions stored in the DB

SDSS Main Sample



Recent Developments

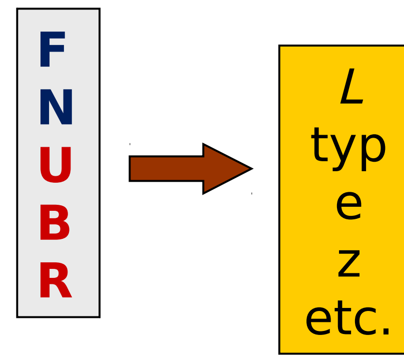
- “Unified theory” of photometric redshifts (Budavari 2010)
  - *Not a regression problem*
  - *Kernel density estimators, constrained by model priors*
- Random Forests at JHU
  - ***S. Carliles**, C. Priebe, A. Szalay, T. Budavari, S. Heinis (2009)*
  - *Slightly better than other estimators*
  - *Estimated errors close to Gaussian, and accurate*
- Physically motivated removal of various systematics
  - *Inclination ⇔ Self Absorption in a galaxy (Yip et al 2011)*
  - *Effect of emission lines*

Unified Theory of Photoz

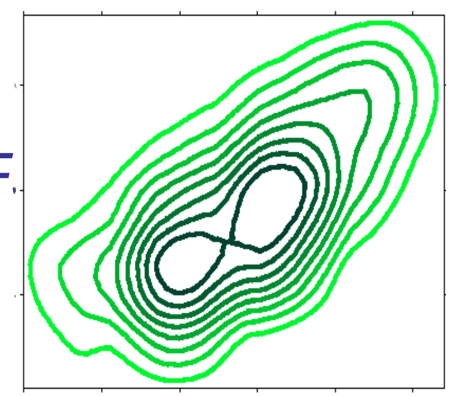
- Tamas Budavari, *Ap.J.*, **695**, 747 (2009)
- Bayesian approach to photo-z
- Essentially all existing techniques are a limiting case

Photometric Inversion

- The general inversion problem
  - *Constrain various properties consistently*
  - *Propagate uncertainties and correlations*



- Estimates are secondary
  - *Probability density functions instead*
  - *Scientific analyses to use the full PDF*



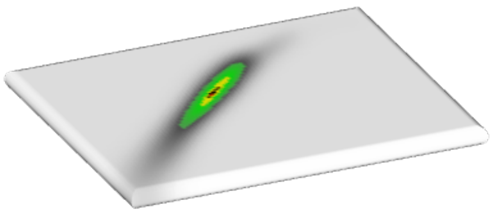
A Unified Framework

- Training and Query sets with different observables

$$T: \quad \{\boldsymbol{x}_t, \boldsymbol{\xi}_t\}_{t \in T}$$
$$Q: \quad \{\boldsymbol{y}_q\}_{q \in Q}$$
$$M: \quad \boldsymbol{\theta}$$

- Model yields observables for given parameter
  - *Prediction via* $p(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{\theta}, M)$ *and has prior* $p(\boldsymbol{\theta} | M)$
  - *Also folds in the photometric accuracy*

- We are after $p(\boldsymbol{\xi} | \boldsymbol{y}_q, M)$



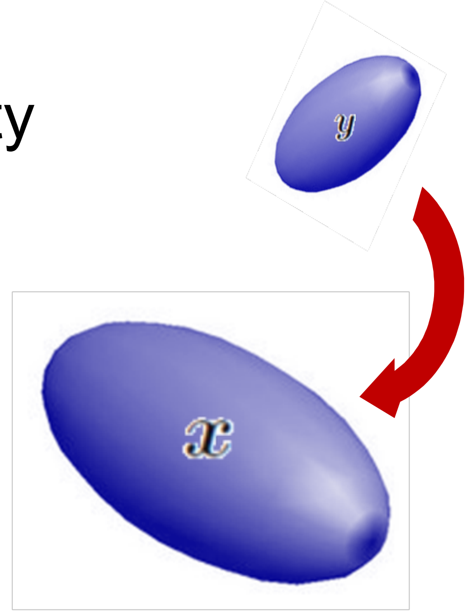
Connecting the Observables

- The model provides the probability density

$$p(\boldsymbol{x}|\boldsymbol{y}_q, M) = \int d\boldsymbol{\theta}\; p(\boldsymbol{x}|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|\boldsymbol{y}_q, M)$$

with

$$p(\boldsymbol{\theta}|\boldsymbol{y}_q, M) = \frac{p(\boldsymbol{\theta}|M)\, p(\boldsymbol{y}_q|\boldsymbol{\theta}, M)}{p(\boldsymbol{y}_q|M)}$$

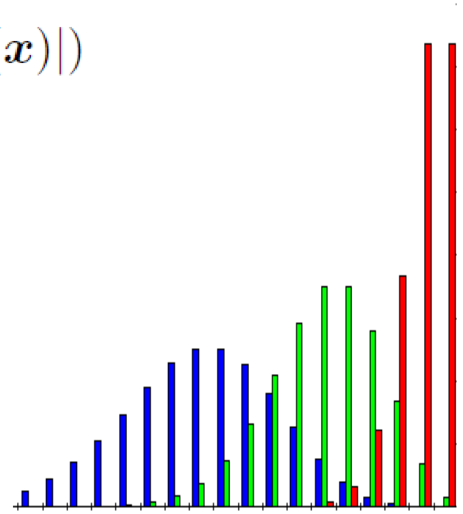


- Think empirical conversion formulas but better
  - *For example, from UJFN to ugriz with errors*

Empirical Relation

- Usually just assume a function $\xi = \hat{\xi}(x)$
  - *Wrong! We know there are degeneracies…*
- There is a more general relation $p(\xi|x)$
  - *Usual restriction is* $p(\xi|x) = \delta(|\xi - \hat{\xi}(x)|)$
  - *Correct estimation*

$$p(\xi|x) = \frac{p(\xi, x)}{p(x)}$$

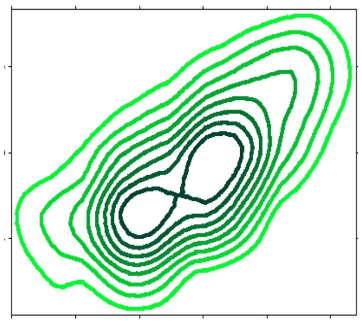


T. Budavari

Properties of Interest

- The final constraint is

$$p(\boldsymbol{\xi}|\boldsymbol{y}_q, M) = \int d\boldsymbol{x}\; p(\boldsymbol{\xi}|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{y}_q, M)$$



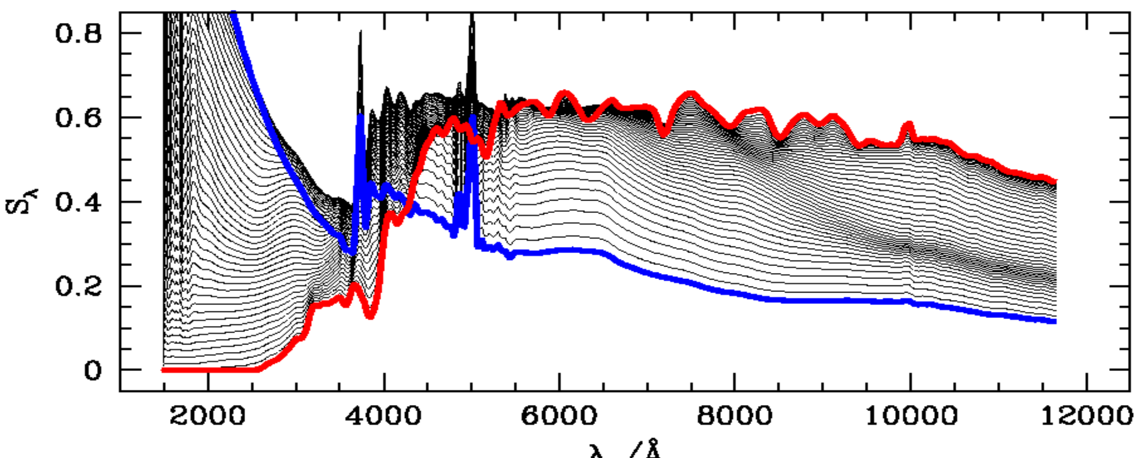
- Estimate by the mean
  - *If the result is unimodal (no guarantee)*

$$\bar{\boldsymbol{\xi}}(\boldsymbol{y}_q) = \int d\boldsymbol{\xi}\, \boldsymbol{\xi}\, p(\boldsymbol{\xi}|\boldsymbol{y}_q, M)$$

Template Fitting

- Artificial training set $\{x_t, \xi_t\} = \{\bar{x}(\theta_t), \bar{\xi}(\theta_t)\}$
  - *From a grid of model points*
  - *No errors*
- Analytic result $p(x|\theta, M) = \delta(|x - \bar{x}(\theta)|)$

$$p(\xi|y_q, M) \propto \sum_{t \in T} \delta(|\xi - \xi_t|)\, p(\theta_t|M)\, N(y_q|\bar{y}(\theta_t), C_q)$$



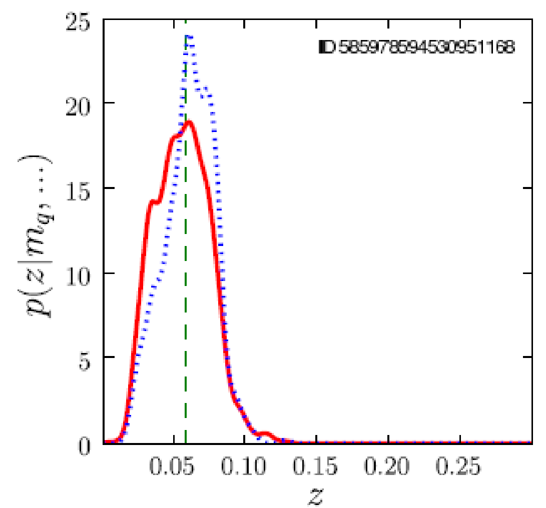
T. Budavari

Improved Empirics

- Minimalist model
  - *Normal distributions, same quantities:* $\bar{x}(\theta) = \theta$ *and* $\bar{y}(\theta) = \theta$
  - *With simple prior, the mapping is analytic , e.g., for flat*

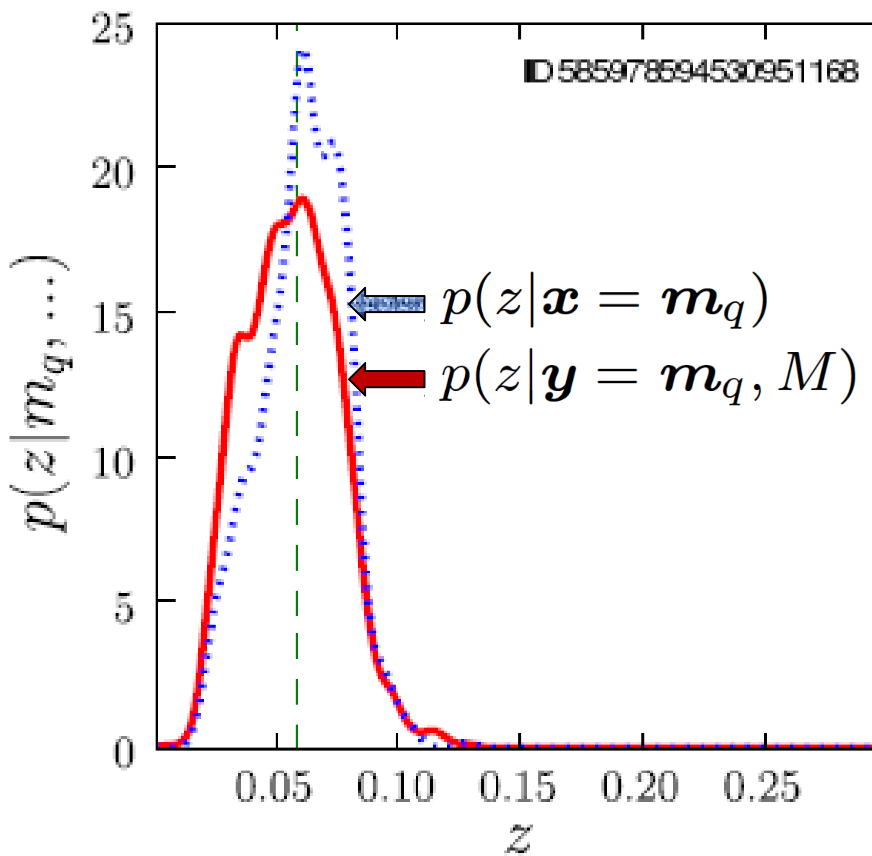
$$p(\boldsymbol{x}_t|\boldsymbol{y}_q, M) = \int d\boldsymbol{\theta}\ N(\boldsymbol{x}_t|\boldsymbol{\theta}, \mathrm{C}_t)\, N(\boldsymbol{\theta}|\boldsymbol{y}_q, \mathrm{C}_q)$$

- Empirical relation
  - *Fitting function as before or rather*
  - *General relation from densities*
- Numerical summation over neighbors



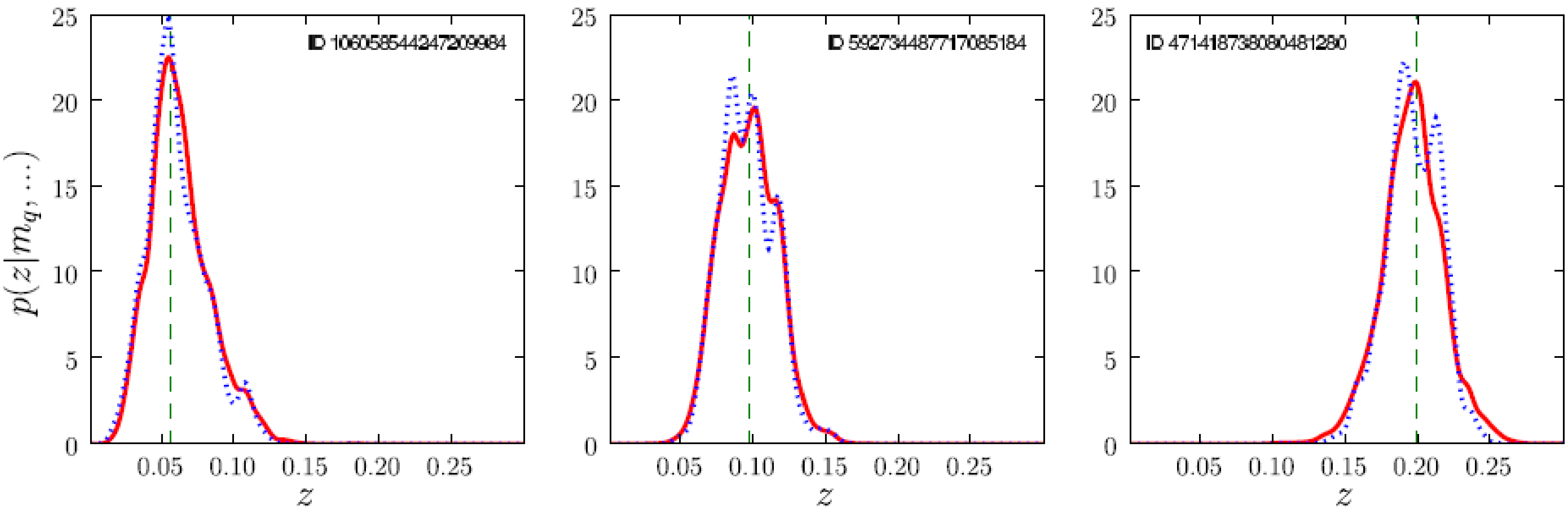
T. Budavari

It works!

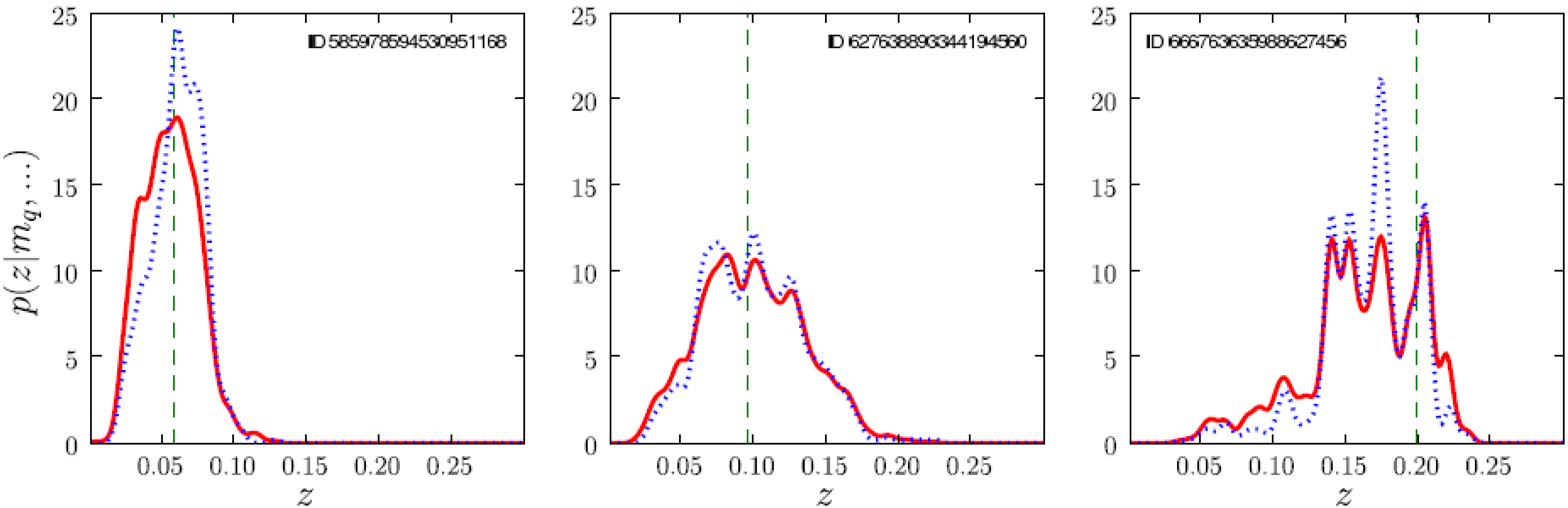


ID 585978594530951168
$p(z|\boldsymbol{x} = \boldsymbol{m}_q)$
$p(z|\boldsymbol{y} = \boldsymbol{m}_q, M)$
$p(z|m_q, \ldots)$
z
0
5
10
15
20
25
0.05
0.10
0.15
0.20
0.25

Red Galaxies



Blue Galaxies



Advanced Methods

- Mapping observables via models
  - *Any complete basis on wavelength range*
  - *Physics in the prior*
- Relation of properties
  - *Conditional densities*
- Empirical but with templates
  - *Unified framework at its best*

$$T: \quad \{\boldsymbol{x}_t, \boldsymbol{\xi}_t\}_{t \in T}$$
$$Q: \quad \{\boldsymbol{y}_q\}_{q \in Q}$$
$$M: \quad \boldsymbol{\theta}$$
$$p(\boldsymbol{\theta}|M)$$
$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, M)$$

Summary

- Upcoming photometric surveys = tons of data
  - *Have to make best use of them: Bayesian inference*
- Objective evidence for associations
  - *Probabilities from ensemble statistics*
- Photometric inversion from first principles
  - *Old methods in the limits*
  - *Suggests new techniques*

Random Forest

- Recent effort at JHU
  - *S. Carliles, C. Priebe, A. Szalay, T. Budavari, S. Heinis*
- RF: Leo Berman and Adele Cutler
- Create many (~500) random subsamples of training set (about 2/3 each)
- Build a piecewise linear regression *Tree* for each
- These Trees make up the *Forest*: each provides an estimated parameter value → *probability distribution*
- Their mean and sigma is the value and error of the final estimate → *robust!*
- Why does it work?

Very promising

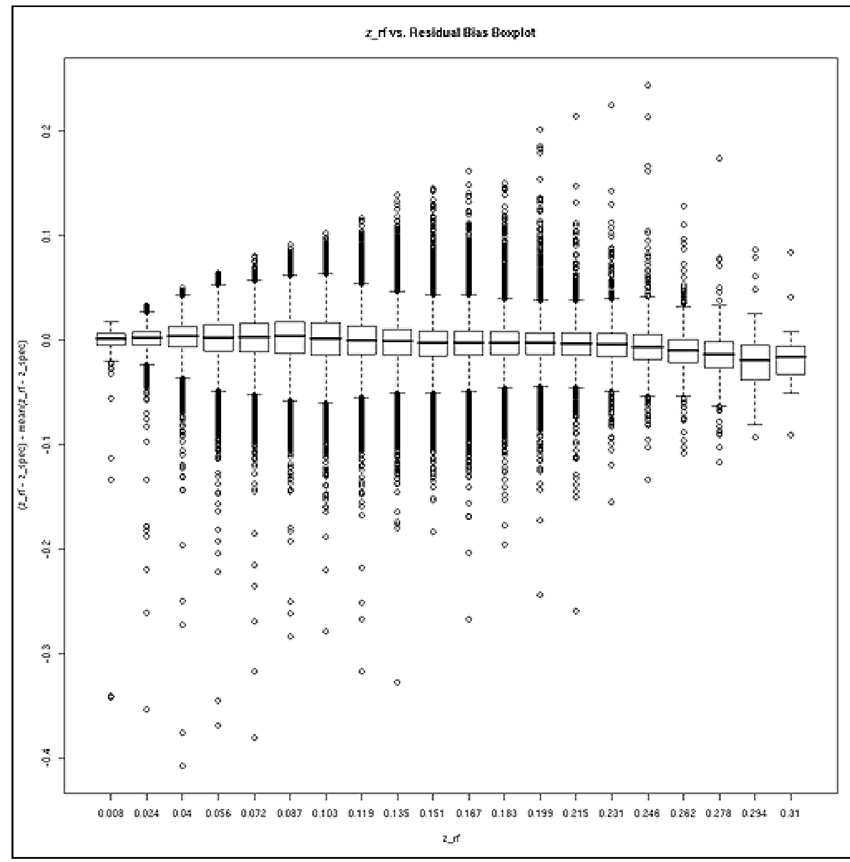
- Consistent estimation of value and its error
- Good scatter vs training set size
- Very few outliers
- Mix of MAIN and LRG
- No χ2<2 clipping
- 100k training set:

  MSE=0.023 MAE=0.017

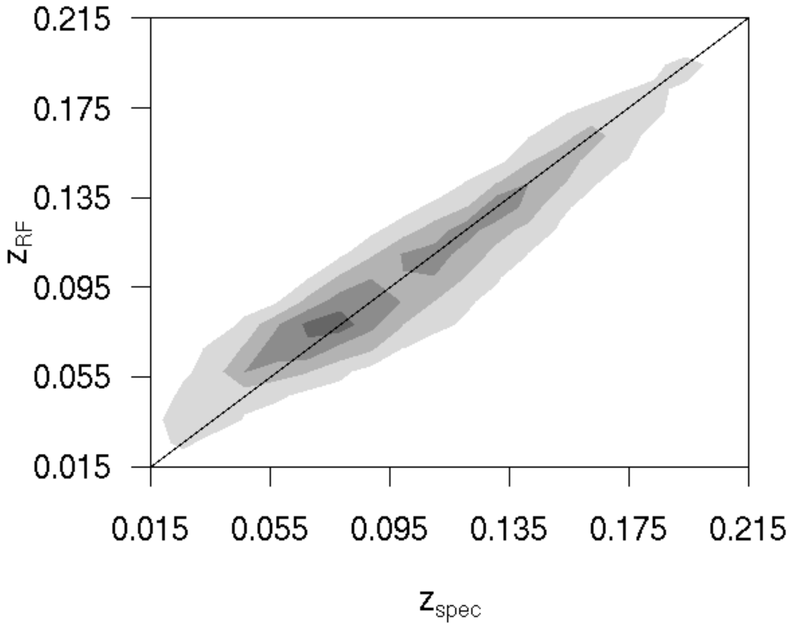
  -> 0.015 with clipping
- 10k training set

  MSE=0.026 MAE=0.019

  deltaZ vs zPred =>



Zspec vs Zrf

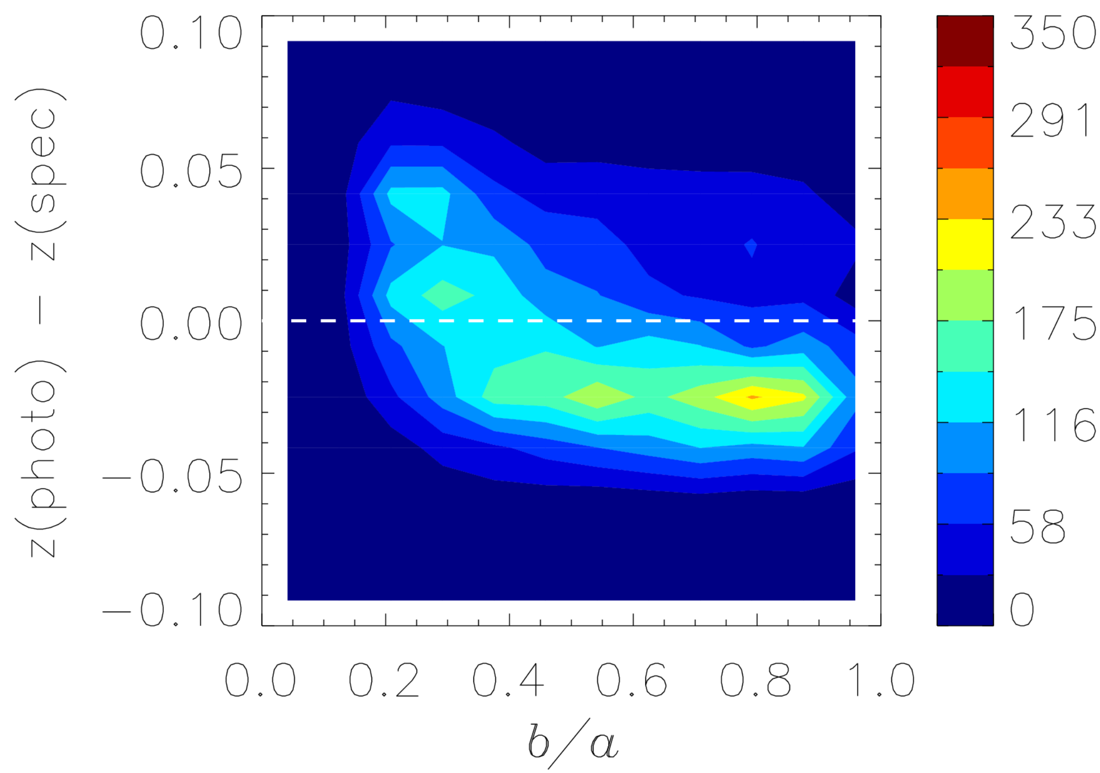


Carliles et al 2009

$$\frac{z_{pred}(i) - z_{spec}(i)}{\sigma(i)}$$

Photo-z Bias vs. Galaxy Inclination

- Edge-on galaxies are redder, mimic higher redshift galaxies
- Photo-z bias is -0.02 for face-on galaxies
- SDSS disk galaxiess, Spec-z = 0.065-0.075, a 30% effect!
- Once axial ratio is included in RF training, bias goes away



C-W Yip et al. 2011

Cyberbricks

- 36-node Amdahl cluster using 1200W total
- Zotac Atom/ION motherboards
  - *4GB of memory, N330 dual core Atom, 16 GPU cores*
- Aggregate disk space 43.6TB
  - *63 x 120GB SSD = 7.7 TB*
  - *27x 1TB Samsung F1 = 27.0 TB*
  - *18x.5TB Samsung M1= 9.0 TB*
- Blazing I/O Performance: 18GB/s
- Amdahl number = 1 for under $30K
- Using the GPUs for data mining:
  - *6.4B multidimensional regressions (photo-z) in 5 minutes over 1.2TB of data*
  - *Running the Random Forest algorithm inside the DB*

Why Does it Work?

- Robustness:
  - *There are always bad points in the training set*
  - *Through the random sampling (~50%) these only make it into half of the neighborhoods*
  - *Whenever a bad point is there, estimator is on the tail*
  - *Whenever bad point is missing, Gaussian*
- Gaussianity:
  - *Through the sampling and averaging, we are creating a new random variable with much better statistical properties than the original estimates with a high skewness and kurtosis*
  - *Central Limit Theorem at work*
  - *The main question is, in which dimension are we approaching the asymptotic limit?*

Simple Analytic Model of RF

Definitions

- Training data with smooth trends removed, $i=1..N$
- Residuals x_i, with zero mean and second moment
- Sampling rate f
- Regression trees $t=1..T$
- Leaf nodes have exactly M points

Estimator for a Query Point

- Consider a single query point
- In each tree there will be a single leaf node containing it
- The estimator from a given tree is calculated as the mean of its M neighbors

$$y_t = \frac{1}{M} \sum_{n=1}^{N} w_{ti} x_i$$

- w_{ti} are the weights (0,1), adding up to M, marking the members of the particular leaf node

Many Trees: Forest

- The ensemble average over many trees gives

$$\langle y_t \rangle = \frac{1}{M} \sum_i \langle w_{ti} \rangle_t \langle x_i \rangle_e = 0$$

(since x has zero mean)

$$\langle y_t^{\ 2} \rangle = \frac{1}{M^2} \sum_{i,j} \langle w_{ti} w_{tj} \rangle_t \langle x_i x_j \rangle_e$$

- The x_i are independent random variates, thus

$$\langle x_i x_j \rangle = \delta_{ij} \sigma_i^2 = \delta_{ij} \sigma^2$$

$$\langle y_t^2 \rangle = \frac{\sigma^2}{M^2} \sum_i \langle w_{ti}^2 \rangle$$

Averaging the Weights

- Once we consider a large number of trees, each point has a probability p_i that it participates in a leaf node for our query point
- The weights will have a multinomial distribution (we draw M points out of N with p_i probability), thus

$$E(w_{ti}) = M\ p_i$$

$$Var(w_{ti}) = M\ p_i(1 - p_i)$$

$$\left\langle w_{ti}^2 \right\rangle = M\ p_i(1 - p_i) + M^2 p_i^2$$

- Summing over all the points

$$\sum_i \left\langle w_{ti}^2 \right\rangle = M + (M^2 - M)\sum_i p_i^2 = M + (M^2 - M)\rho^2$$

The Effective Bandwidth

- Here $\rho^2 = 1/v$ is the “effective bandwidth of the kernel arising from the local neighborhoods
- v is the effective degrees of freedom
- The variance of the estimator is

$$\langle y_t^2 \rangle = \sigma^2 \left[\frac{1}{M} + \left(1 - \frac{1}{M} \right) \frac{1}{v} \right]$$

- The effective degrees of freedom will depend on the sampling rate
- For this toy model there is no bias error, as we assumed a zero mean. For a real use case there will be an optimum bandwidth, like for an adaptive kernel

The Forest Estimator

- The different trees are obviously correlated

$$\langle y_t y_r \rangle = \frac{\sigma^2}{M^2} \sum_i \langle w_{ti} w_{ri} \rangle = \frac{\sigma^2}{M^2} \sum_i \langle w_{ti} \rangle \langle w_{ri} \rangle = \sigma^2 \sum_i p_i^2 = \frac{\sigma^2}{v}$$

- The forest estimator and its variance

$$Y = \frac{1}{T} \sum_t y_t$$

$$\langle Y^2 \rangle = \frac{1}{T^2} \sum_{t,r} \langle y_t y_r \rangle = \frac{1}{T^2} \sum_t \langle y_t^2 \rangle + \frac{1}{T^2} \sum_{t \neq r} \langle y_t y_r \rangle.$$

The Variance

- Using the tree estimator variance and covariance

$$\left\langle Y^2 \right\rangle = \frac{1}{T^2}\left[T\sigma^2 \left(\frac{1}{M} + \left(1 - \frac{1}{M}\right)\frac{1}{v} + T(T-1)\frac{\sigma^2}{v} \right)\right]$$

$$\left\langle Y^2 \right\rangle = \sigma^2 \left[\frac{1}{v} + \frac{1}{TM}\left(1 - \frac{1}{v}\right)\right]$$

- The variance mostly depends on v, and only weakly on the forest size T, as seen in our experiments

Summary

- A simple analytic toy model shows how the Central Limit Theorem creates an asymptotically Gaussian estimator for the RF
- The Random Forest technique approximates a kernel density estimator based integration over the training set
- The convergence primarily depends on the size of the kernel, i.e. the sampling rate
- There has to be an optimum bandwidth, possibly variable over our photo-z domain
- The RF photo-z very closely resembles the Budavari implementation for the Bayesian photo-z