SPIDERz - A SUPPORT VECTOR MACHINE FOR PHOTOMETRIC REDSHIFT ESTIMATION

Orientation

Galaxy redshifts are important

- Many reasons!

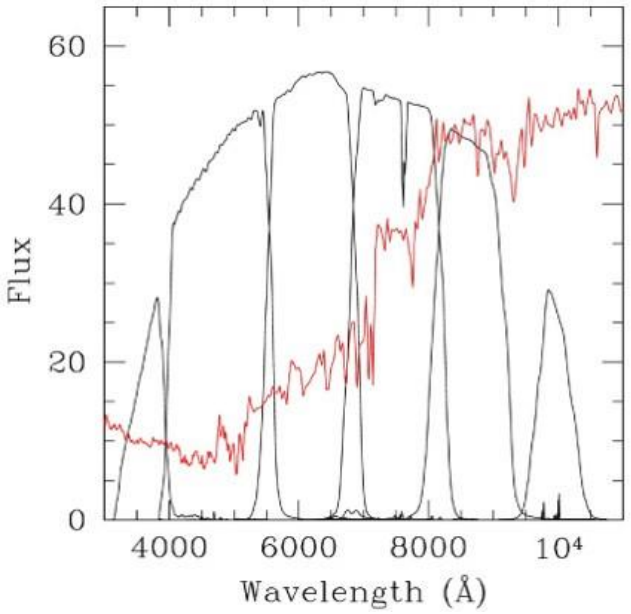
But

Measuring galaxy spectra is too slow for large scale surveys

The (potential) solution:

Photo-z estimation

- Estimate redshift from flux in a limited number of filter bands
- Doing so accurately and with well understood errors is an important data challenge for current and future large multi-band extragalactic surveys



Why make a SVM for photo-z estimation?

SVMs have been successfully applied in other areas of astrophysics

- classification of objects into stellar, galactic, or active galaxy categories — Marton et al. 2016; Malek et al. 2013; Hassan et al. 2013; Solarz et al 2013; Klement et al. 2011; Peng et al. 2002
- classification of structures in interstellar medium — e.g Beaumont et al. 2011
- galaxy morphological classification — e.g Huertas-Company et al. 2007

Past SVM attempts for photo-zs were intriguing but limited

- low redshifts ($z < 1$) or simulated data — Wadadekar 2004; Wang et al. 2007

SVMs are useful for exploring inclusion of parameters beyond photometry

- learning algorithm can treat input parameters symmetrically

In contrast with some other empirical methods

- computational time for training is roughly linear in the number of input parameters
- Our custom SVM method naturally outputs 'effective' redshift probability distribution (PDF)

Supervised learning with SVM

TRAINING

Training galaxies contain photometry and are labeled with known spectroscopic redshifts:

$\overrightarrow{x_i} = [\mathrm{u}, \mathrm{b}, \mathrm{g}, \mathrm{r}, \mathrm{i}]$

$y_i = z_{spec}$

$\vec{x}_i, z_{spec}$

SVM 'learns' from galaxies in the training set and builds a predictive model

$\mathrm{M} = f(\vec{x}_i, z_{spec})$

EVALUATION

Evaluation galaxies contain only photometry:

$\overrightarrow{x_j} = [\mathrm{u}, \mathrm{b}, \mathrm{g}, \mathrm{r}, \mathrm{i}]$

The predictive model is applied to galaxies in the evaluation set to obtain photo-z estimations

$\mathrm{M}(\vec{x}_j) = z_{photo}$

We can compare photo-z estimations for the evaluation set to known spectroscopic redshifts to assess the performance of model.

SPIDERz: **SuP**port vector classification for **IDE**ntifying **R**edshifts

Reported in

- E. Jones & J. Singal, 2017, *A&A*, “Analysis of a Custom Support Vector Machine for Photometric Redshift Estimation and the Inclusion of Galaxy Shape Information.” in press (arXiv:1607.00044)

Available from

- spiderz.sourceforge.net

SPIDERz: SuPport vector classification for IDEntifying Redshifts

Implements Support Vector Classification (SVC) in IDL

- galaxy vectors are assigned class labels according to redshift
- each bin represents a different class in the multi-class system
  - i.e. dataset ranging from z = 0 to 5 and with bins of size 0.1 forms a 51 class system

Training

- Multi-class solutions can be approximated with a series of binary class solutions
- We use a one vs. one or 'pairwise coupling' approach that constructs and solves a binary class system for every possible pairing of classes:

m classes → $\frac{m(m-1)}{2}$ binary class problems with $\frac{m(m-1)}{2}$ unique optimal hyperplane solutions

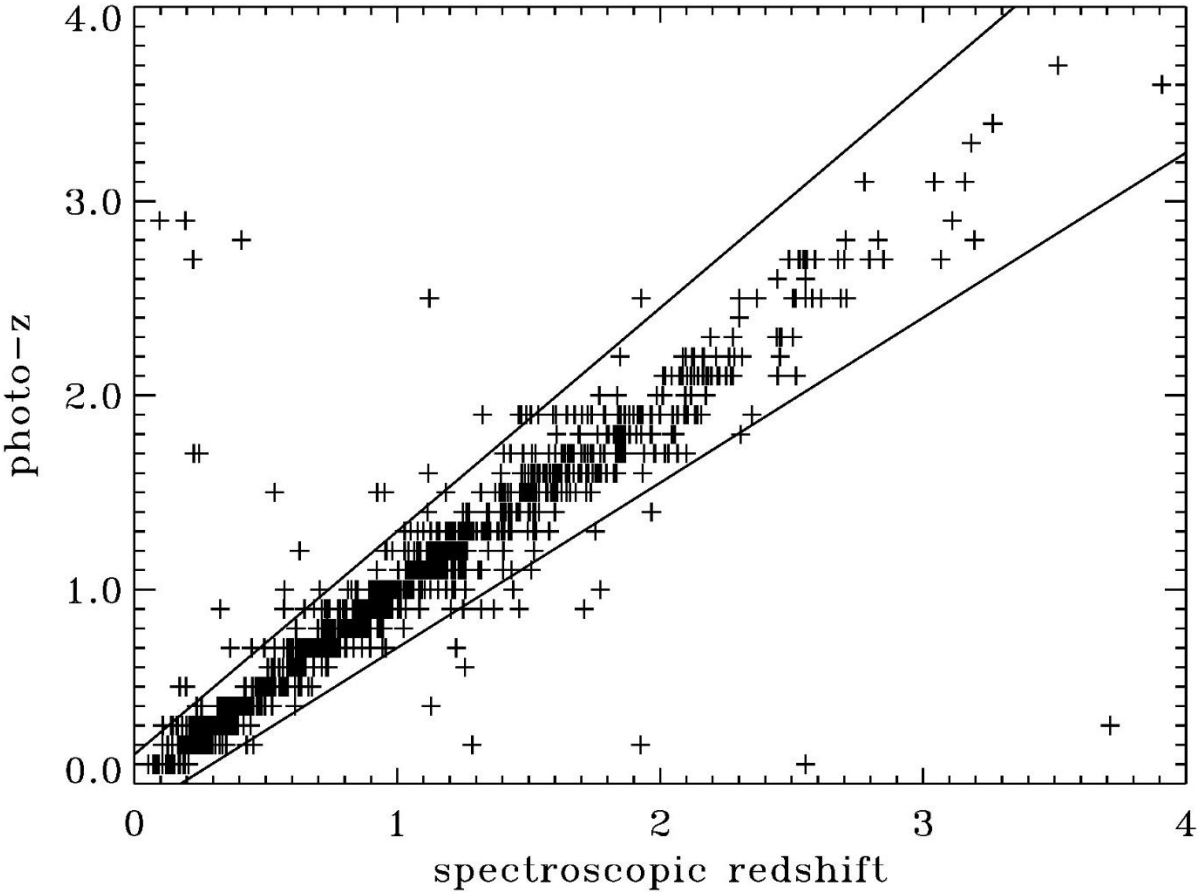
Evaluation

Predictive model consisting of $\frac{m(m-1)}{2}$ binary classifiers is applied to evaluation set of galaxies

- The class (or redshift bin) to which a galaxy is most assigned becomes its final discrete predicted redshift value
- The distribution of binary classification results resembles a probability distribution

COSMOSxHST Data Set

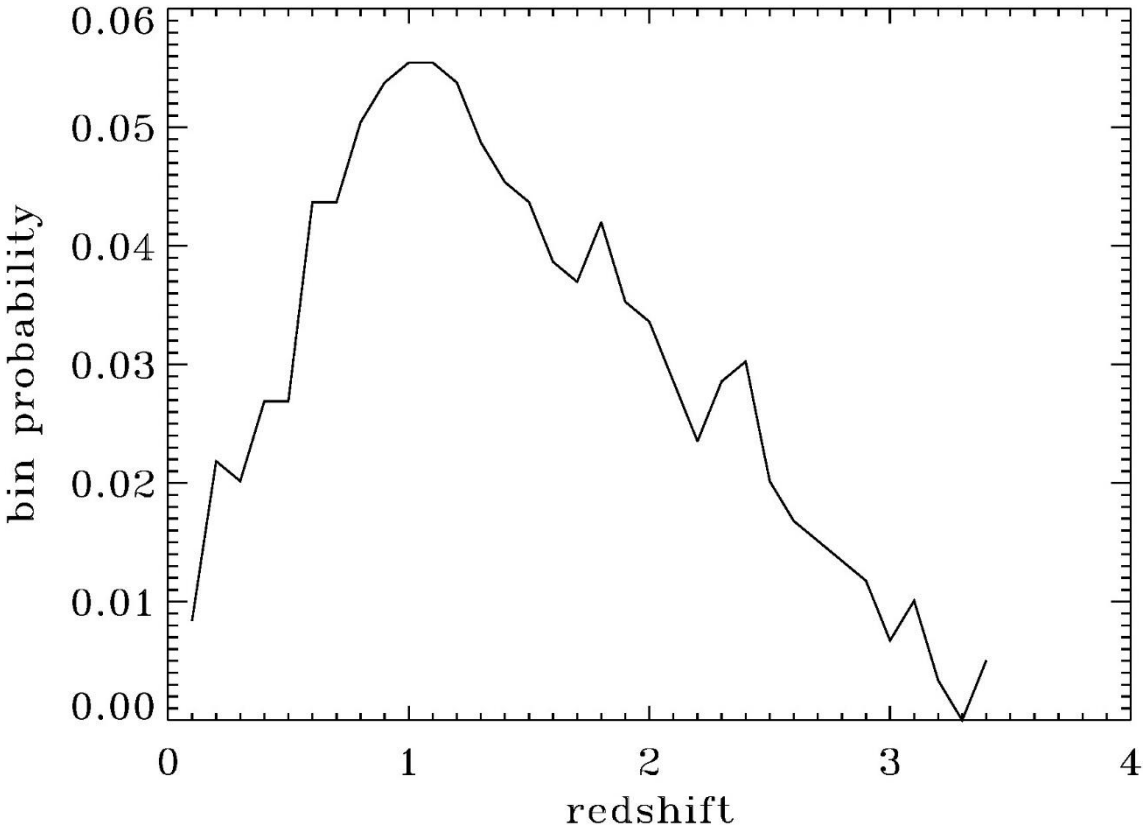
- Same COSMOS photometry and morphology as previous but with available spectro-zs from HST (Momcheva et al., 2016)
- Makes set with 3048 galaxies (6.8% z>2)



2.6% outliers
RMS = .056
R-RMS = 0.04

10 band COSMOSxHST SPIDERz results, binsize 0.01, 1200 training

SPIDERz 'effective PDF' options

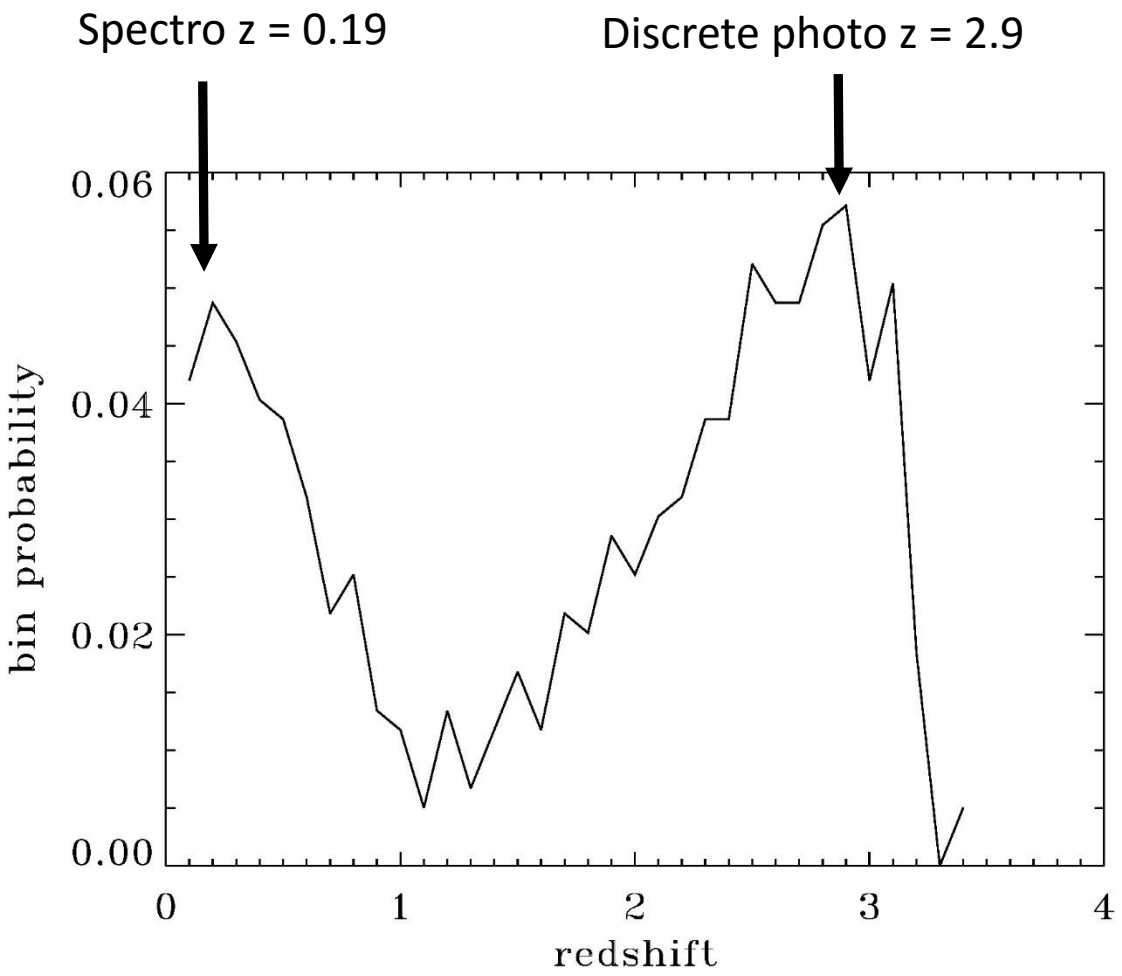


- Because of the $\frac{m(m-1)}{2}$ binary class solutions we actually have a distribution of photo-z results
- Could preserve all $\frac{m(m-1)}{2}$ results as a photo-z PDF of sorts
- More later…

SPIDERz PDF options

PDFs can reveal potential “catastrophic outliers”

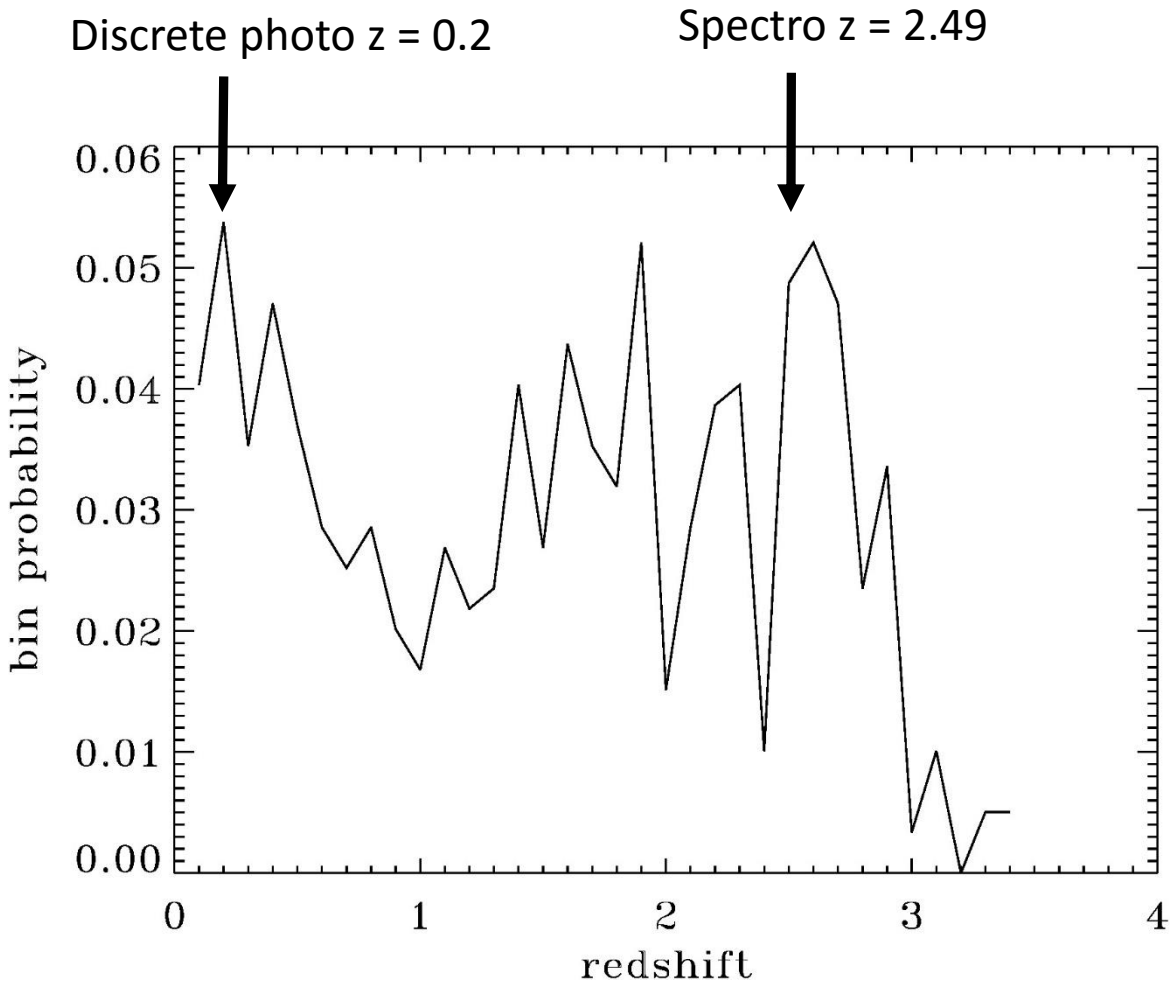
Double peaks - (Very photogenic example from COSMOSxHST 10 band)



SPIDERz PDF options

PDFs can reveal potential “catastrophic outliers”

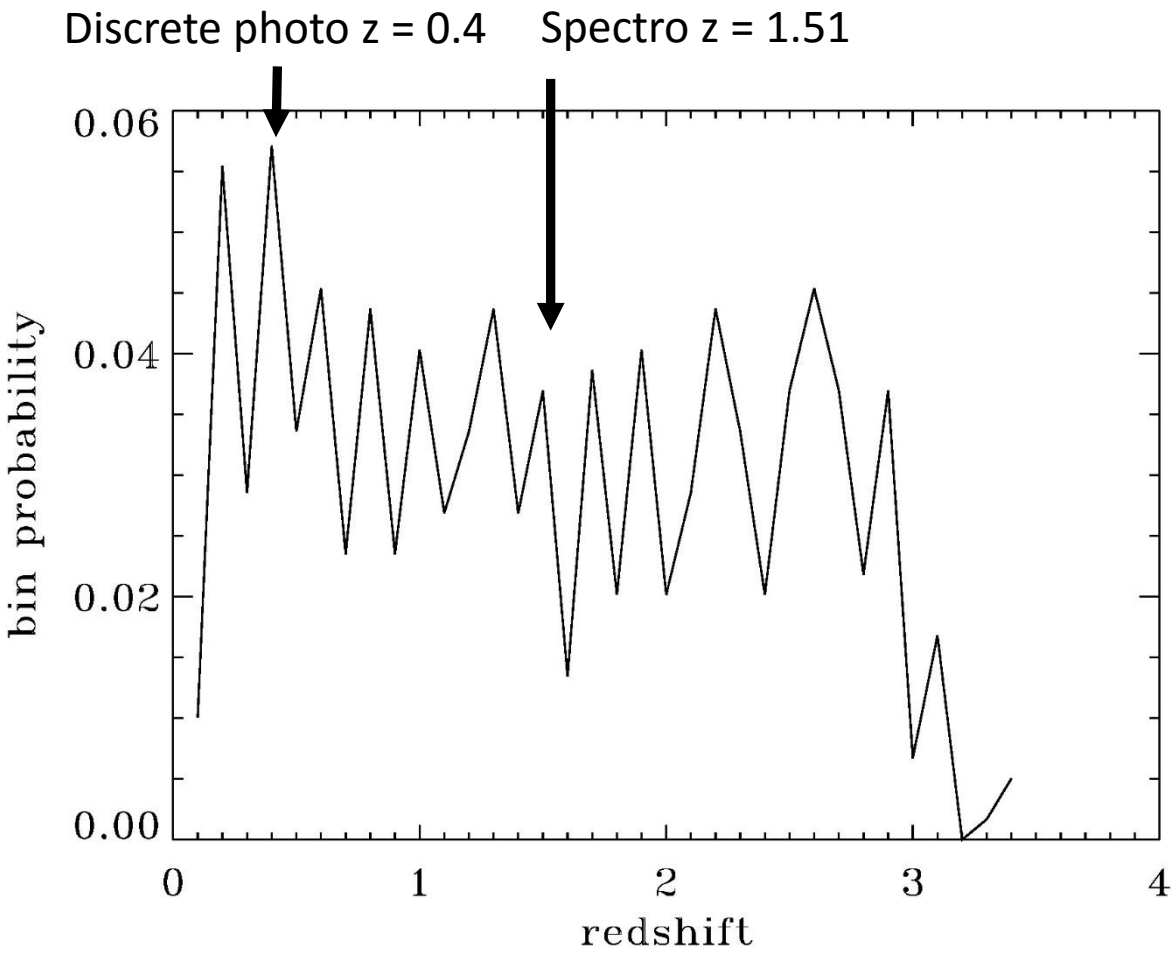
Double peaks - (Another example from COSMOSxHST 10 band)



SPIDERz PDF options

PDFs can reveal potential “catastrophic outliers”

Weak peak - (Another example from COSMOSxHST 10 band)



Identifying potential catastrophic outliers with EPDFs

- Want to use characteristic features present in EPDFs to flag potential outlier or catastrophic outlier galaxy estimates

- We focus on identifying distributions with multiple peaks

Flagging criteria for identifying multiply peaked EPDFs

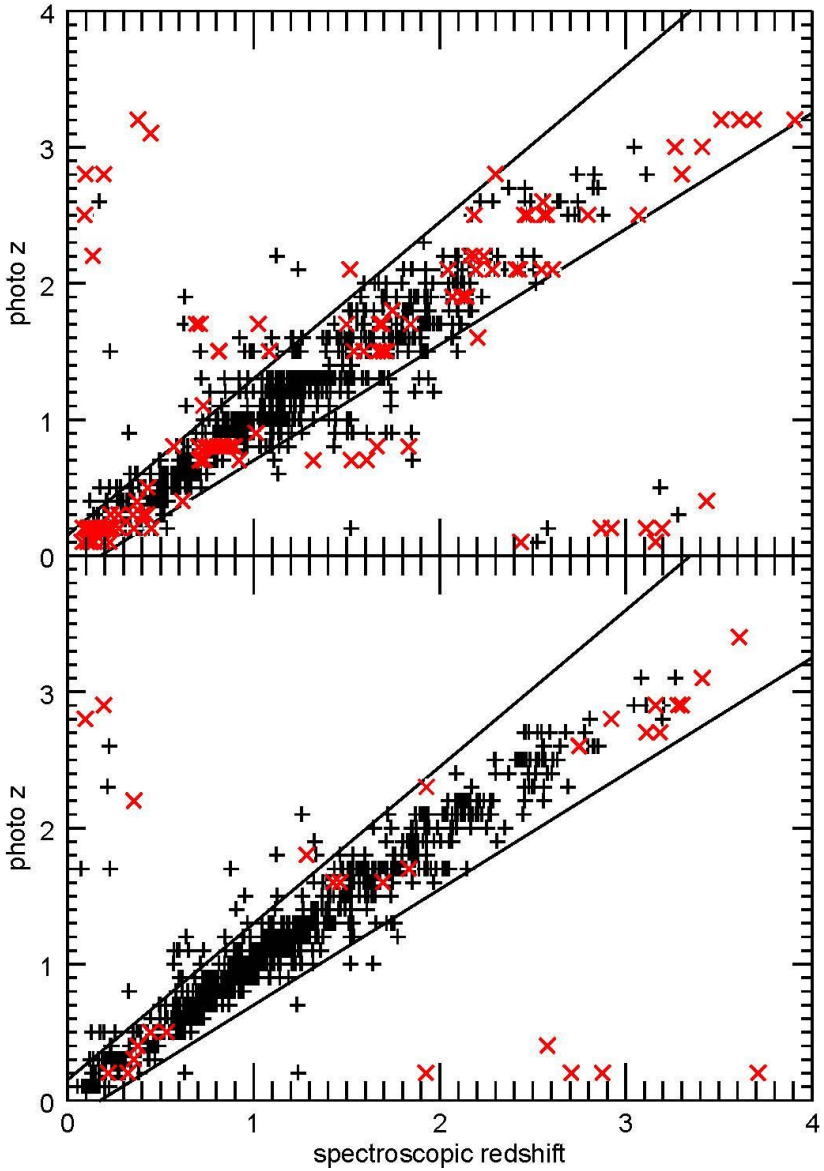
1. redshift distance between candidate peak and primary peak:

$$\Delta z_{peak} = \left| z_i - z_{primary} \right|$$

2. relative probability compared to primary peak:

$$p_f = \frac{p_i}{p_{primary}}$$

Flagged galaxies shown in red for test determinations performed with SPIDERz and using test data comprised of 5 optical bands (top) and 10 optical and infrared bands (bottom)

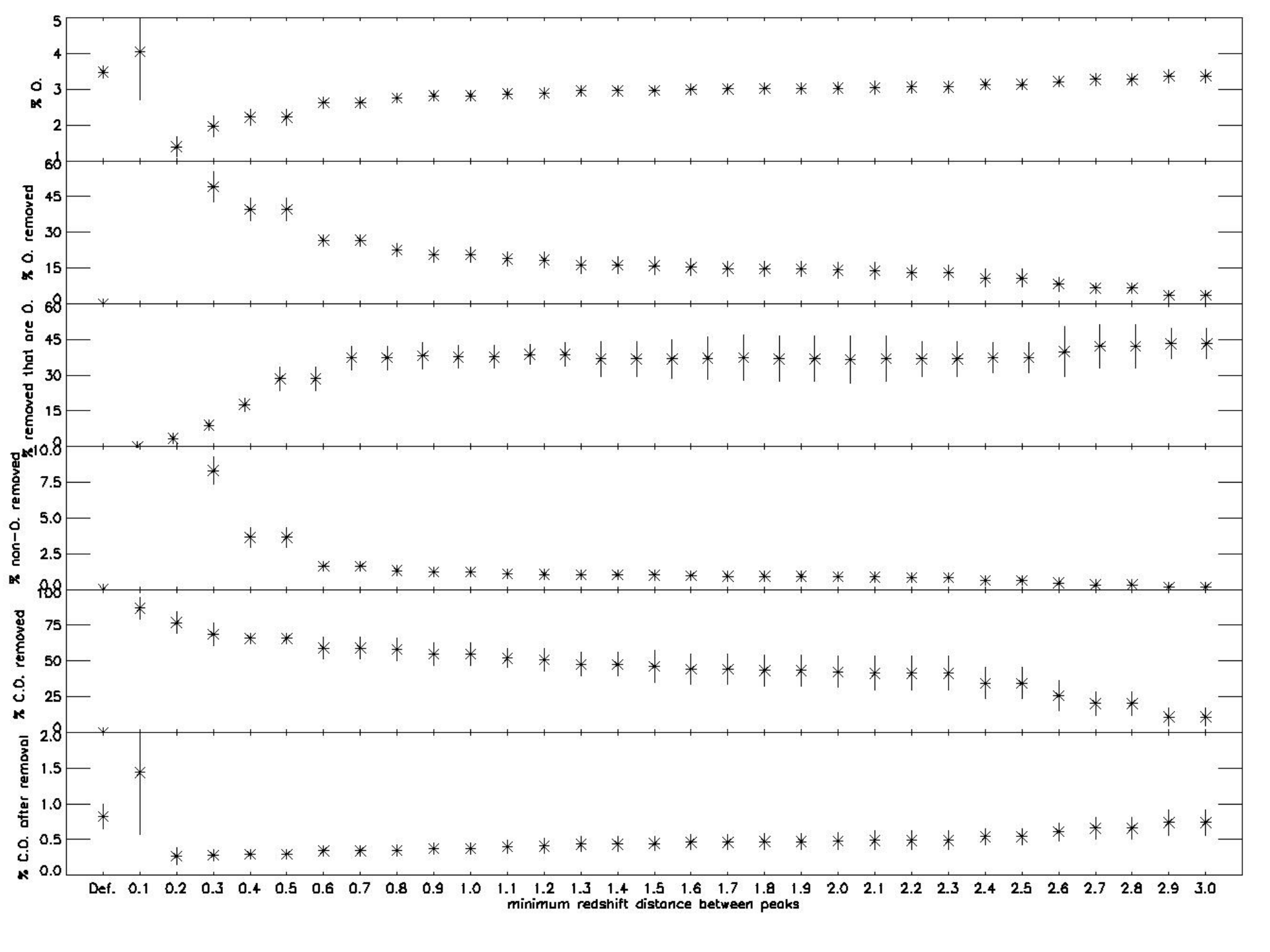


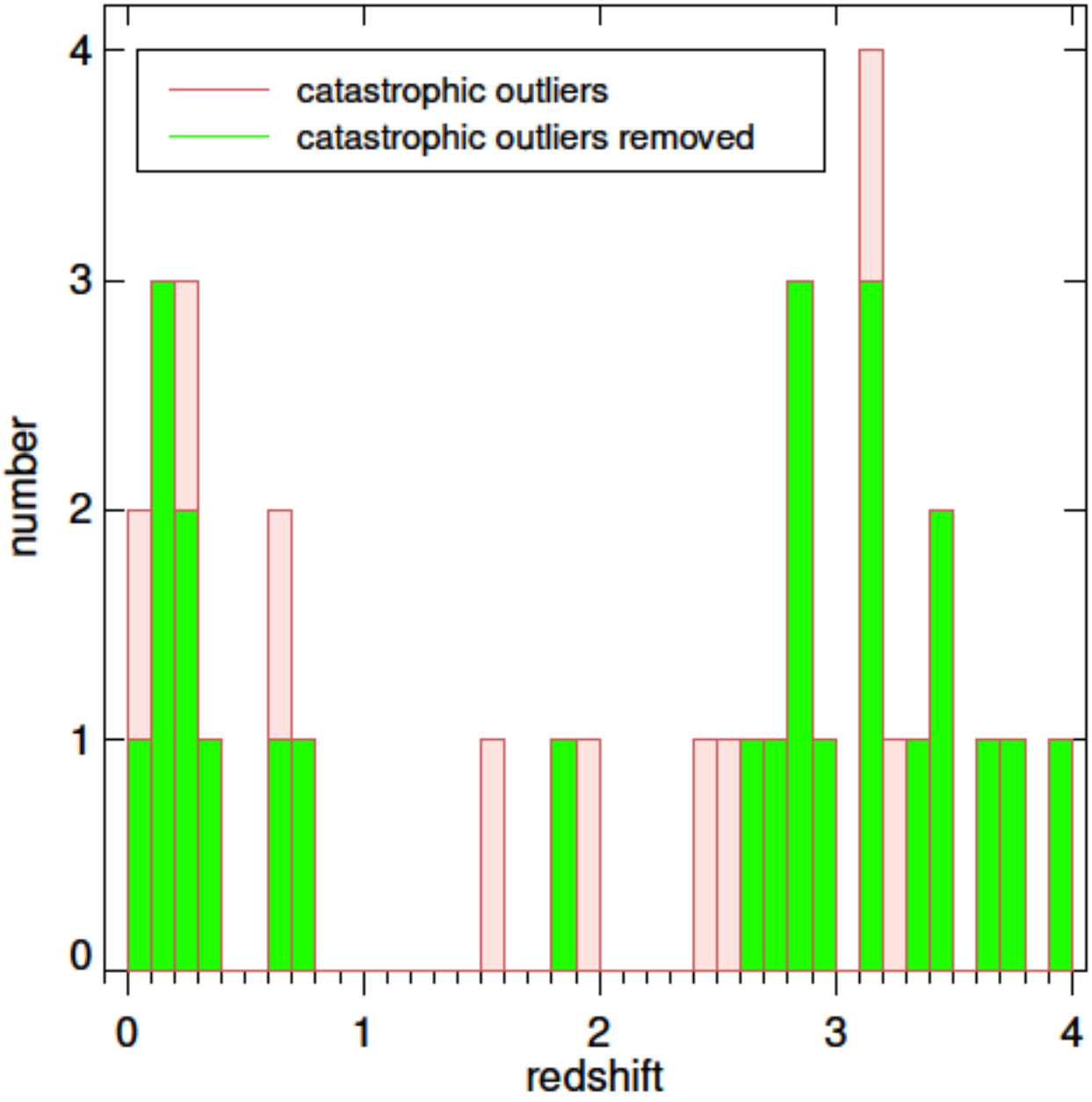
5-bands (u, V, r, i, z+)

- Outliers reduced by ~28%
- Catastrophic outliers reduced by ~77%
- Incorrectly removed 5.0 % of non-outliers
- RMS reduced by ~ 60%

10-bands (u, B, V, r, i, z+, Y, H, J, Ks)

- Outliers reduced by ~37%
- Catastrophic outliers reduced by ~60%
- Incorrectly removed only 3.4% of non-outliers
- RMS reduced by ~ 63%


% O.
% O. removed
% removed that are O.
% non-O. removed
% C.O. removed
% C.O. after removal
Def.
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
minimum redshift distance between peaks


catastrophic outliers
catastrophic outliers removed
number
redshift
0
1
2
3
4
0
1
2
3
4