Joint Stellar Mass - Redshift PDFs using Random Forest

SUNIL MUCESH

SUPERVISORS: PROF. OFER LAHAV, DR. TIM SCANLON & DR. WILL HARTLEY

Project Goal and Motivation

- Machine learning (ML) can be used to predict properties of galaxies. However, a point estimate is not very useful.
- To characterise the uncertainties, PDFs are required.
- It has been shown that PDFs can be generated using ML.
- Taking a step further, can ML be used to build joint pdfs?
- In particular, stellar mass – redshift joint pdfs are important.

Random Forest: Introduction

- Random forest is an ensemble learning technique based on using many decision trees.
- It can be used for classification and regression.
- Easy to implement and understand (not a black box).
- It has been used to predict redshifts, stellar masses and star formation rates of galaxies.

Random Forest: Algorithm

A random forest consists of many decision trees with a few tweaks.

1. Sample randomly from data with replacement.
2. Choose only a subset of features.
3. Create a decision tree from the bootstrapped sample.
4. Repeat to create a random forest.

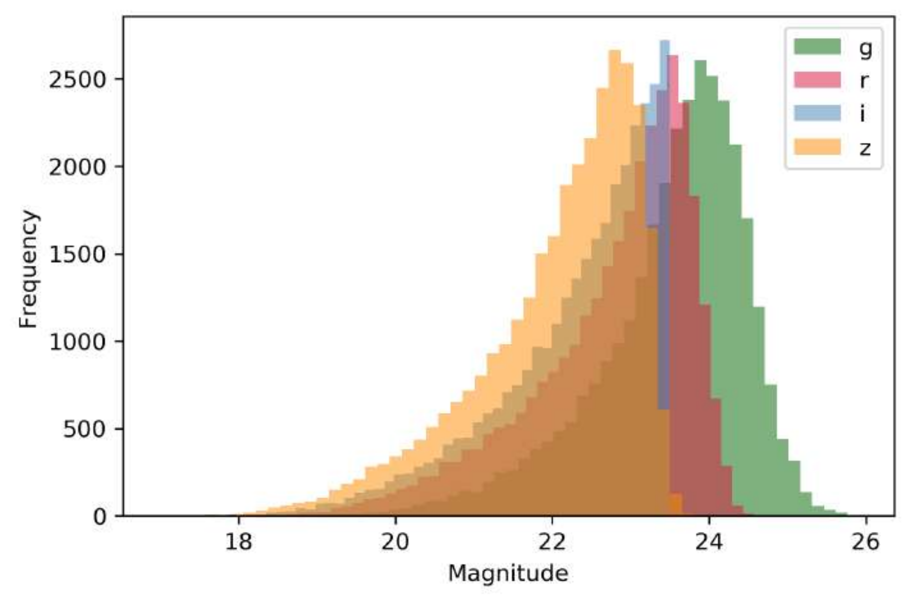
To make a prediction:

- Classification - Majority Vote
- Regression - Average

Data: COSMOS-DECam & DES Y3 Gold

Pre-processing:

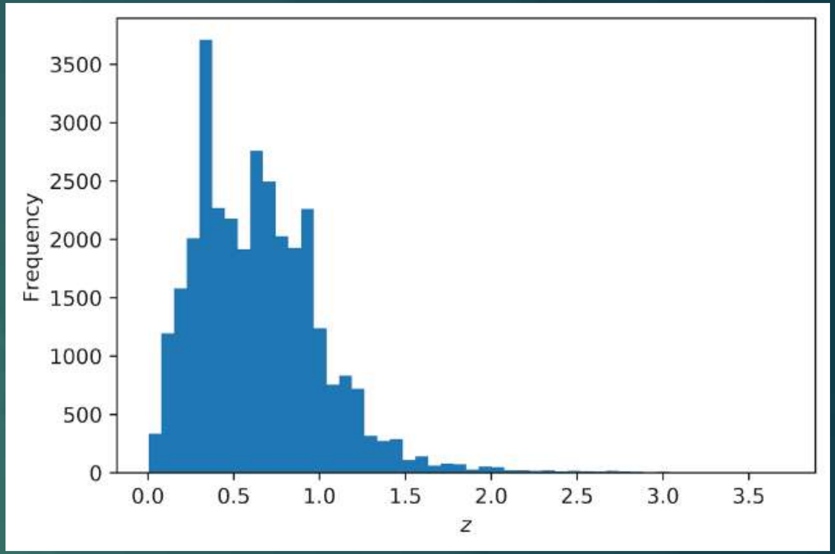
- COSMOS-DECam dataset contains approximately 240,000 galaxies.
- Cuts: *griz* magnitudes < 30, *i*-band < 23.5 and magnitude errors < 0.1
- Reduced dataset matched to DES Y3 Gold using magnitudes and errors drawn randomly for each galaxy. This is done to mimic data in a relatively deep and wide field.
- Final dataset contains approximately 31,000 galaxies.

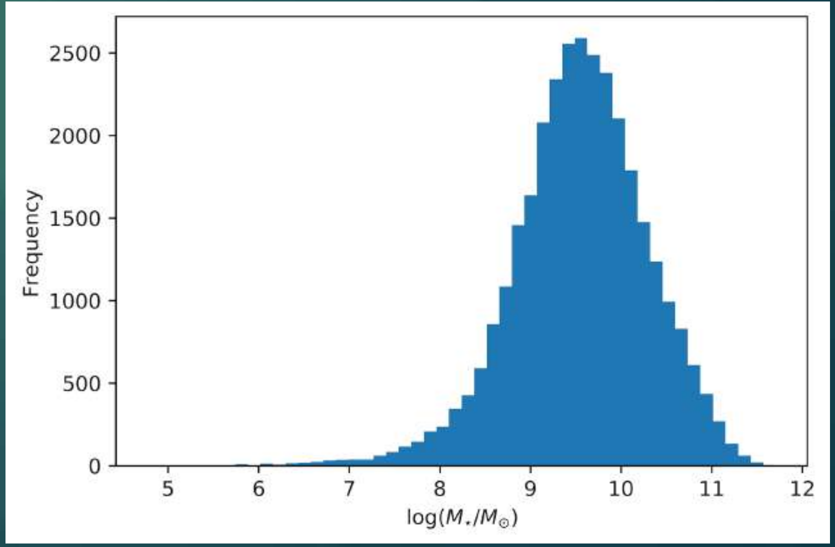


Data: COSMOS-DECam & DES Y3 Gold

Pre-processing:

- COSMOS-DECam dataset contains approximately 240,000 galaxies.
- Cuts: *griz m*agnitudes < 30, *i*-band < 23.5 and magnitude errors < 0.1
- Reduced dataset matched to DES Y3 Gold using magnitudes and errors drawn randomly for each galaxy. This is done to mimic data in a relatively deep and wide field.
- Final dataset contains approximately 31,000 galaxies.



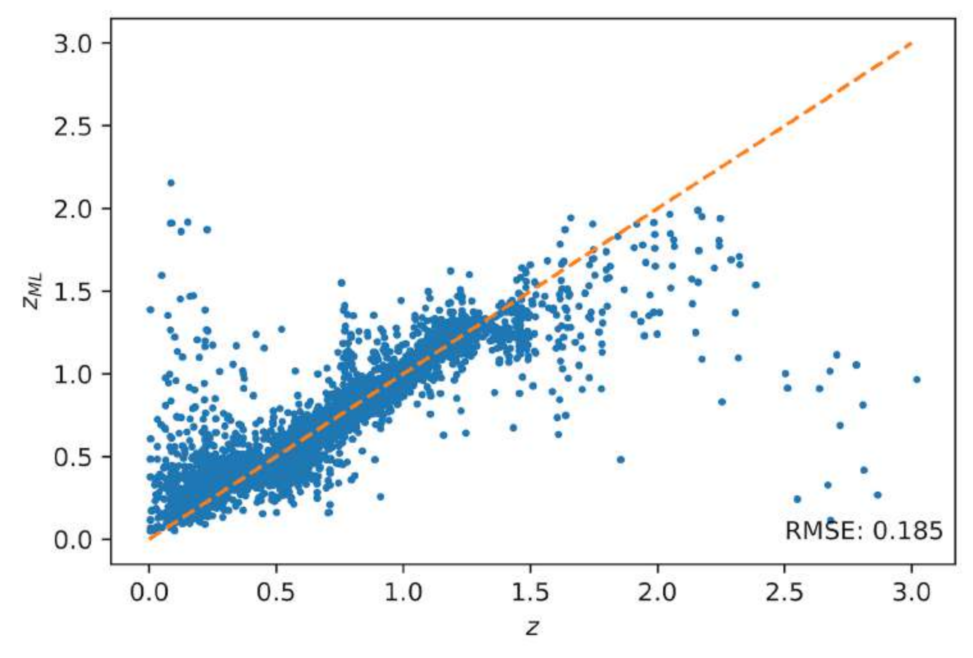


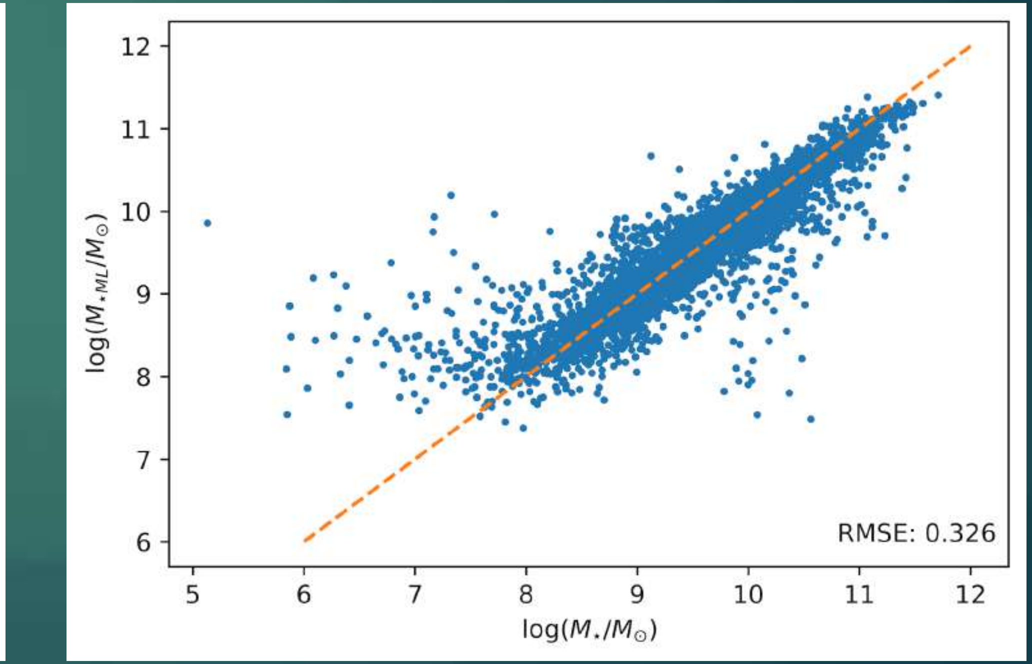
Incorporating Errors

- We can incorporate errors in the data into the algorithm.
- One way of doing this is to scatter the magnitudes of a galaxy according to the errors multiple times.
- Draw errors randomly from a gaussian distribution centred about the magnitude and with standard deviation given by the error.

Results: Point Estimates

- Trained two RFs to predict redshift and stellar mass with 80% data.
- Input features: magnitudes in *griz* bands + colours (*g-r*, *r-i*, *i-z*) and errors.





Probability distributions

- Random forest can be described as a clustering algorithm.
- It aims to group together similar galaxies and these end up in the same leaf nodes of the decision trees.
- For a point estimate, we averaged the redshift or stellar mass values of the galaxies in leaf nodes.
- To extract a probability distribution, we can simply gather all the values in the leaf nodes in all the decision trees.

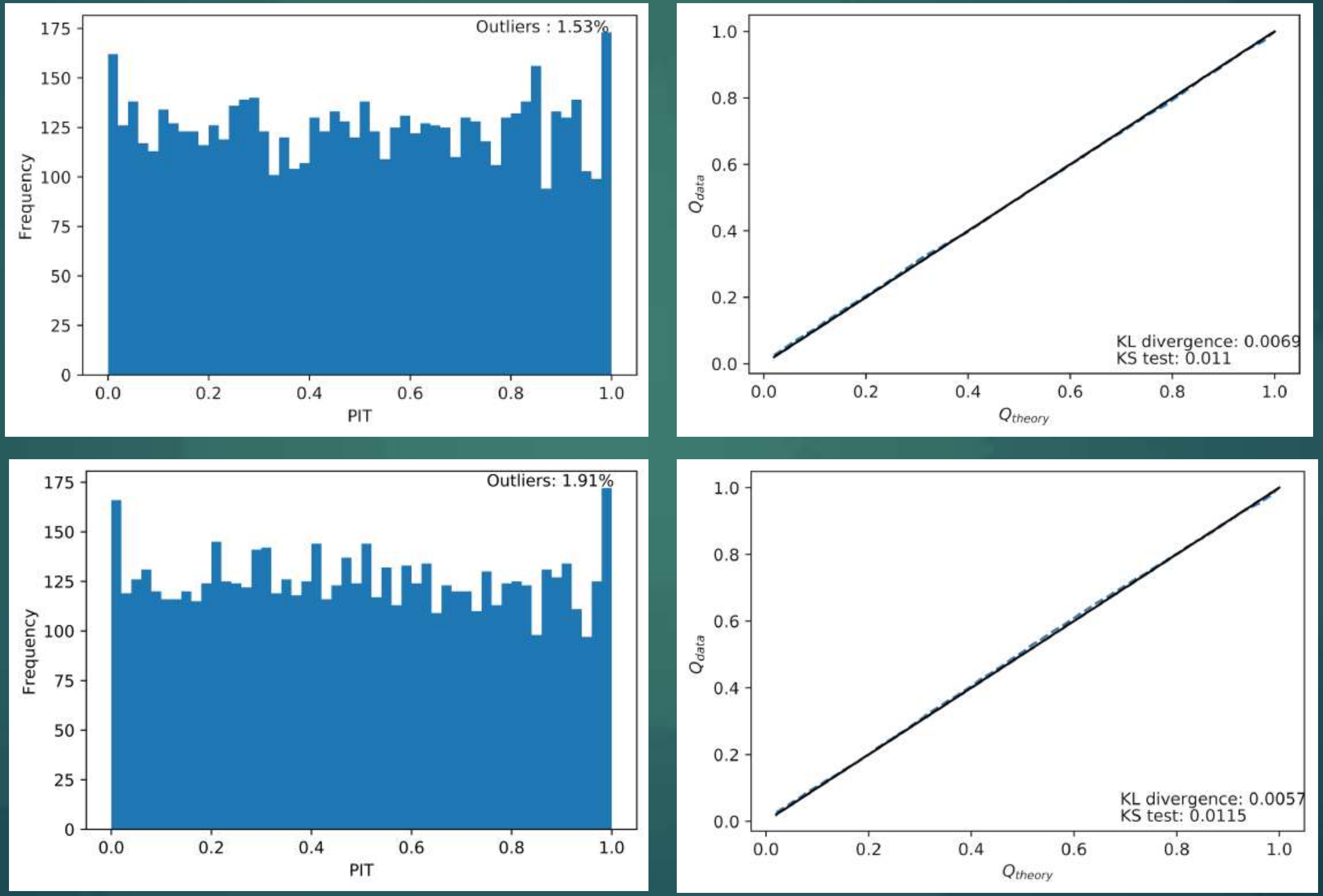
How accurate are the extracted PDFs?

- Unlike point estimates, the ‘true’ PDFs are not available for comparison.
- To get started, we can compare the true value to the extracted probability distribution using the probability integral transform defined by:

$$PIT = \int_{-\infty}^{z_{true}} p(z)\, dz$$

- If accurate, the true value should be a random draw from its respective probability distribution, and as a result, the distribution of PIT values should be uniform for an ensemble of galaxies.
- To test uniformity of the PIT distribution, we can use metrics such as KL divergence and KS Test.

Results: Redshift & Stellar Mass PIT distributions



Theory: Joint PDFs

Simultaneous Method:

- Build one model which predicts redshift and stellar mass simultaneously.
- Extract the joint distribution.

Separate Method:

- Build two separate models, one which predicts redshift and another which predicts stellar mass given a redshift.
- To get the joint pdf.

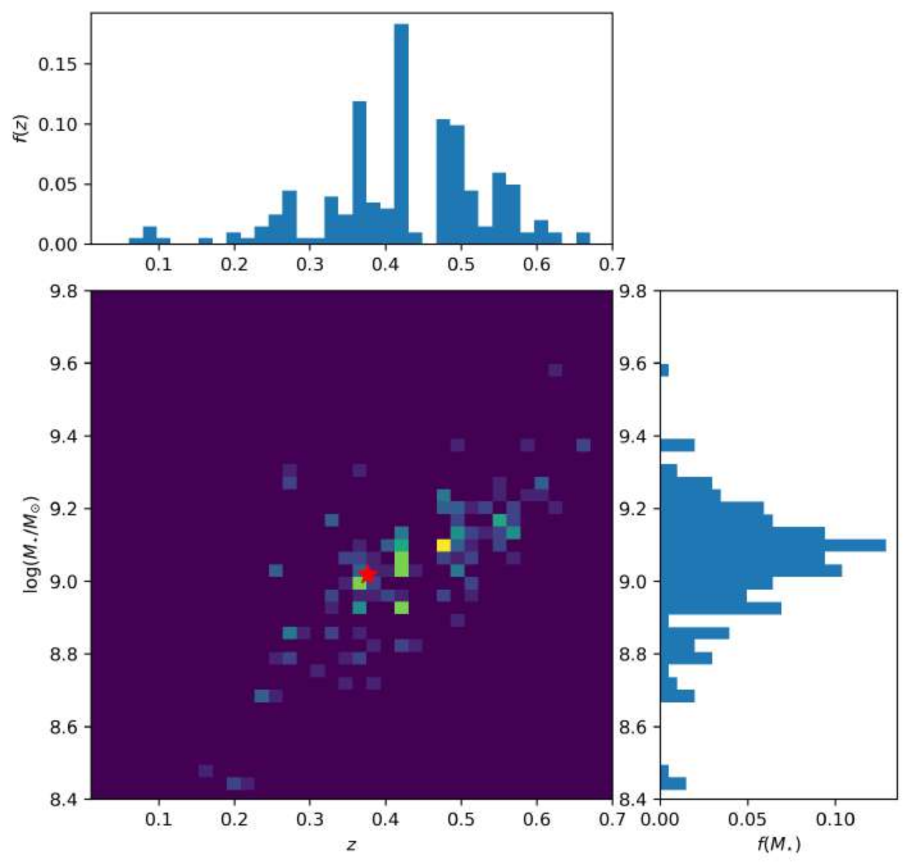
$$f(M,z) = f(M|z) * f(z)$$

Steps: Separate Method

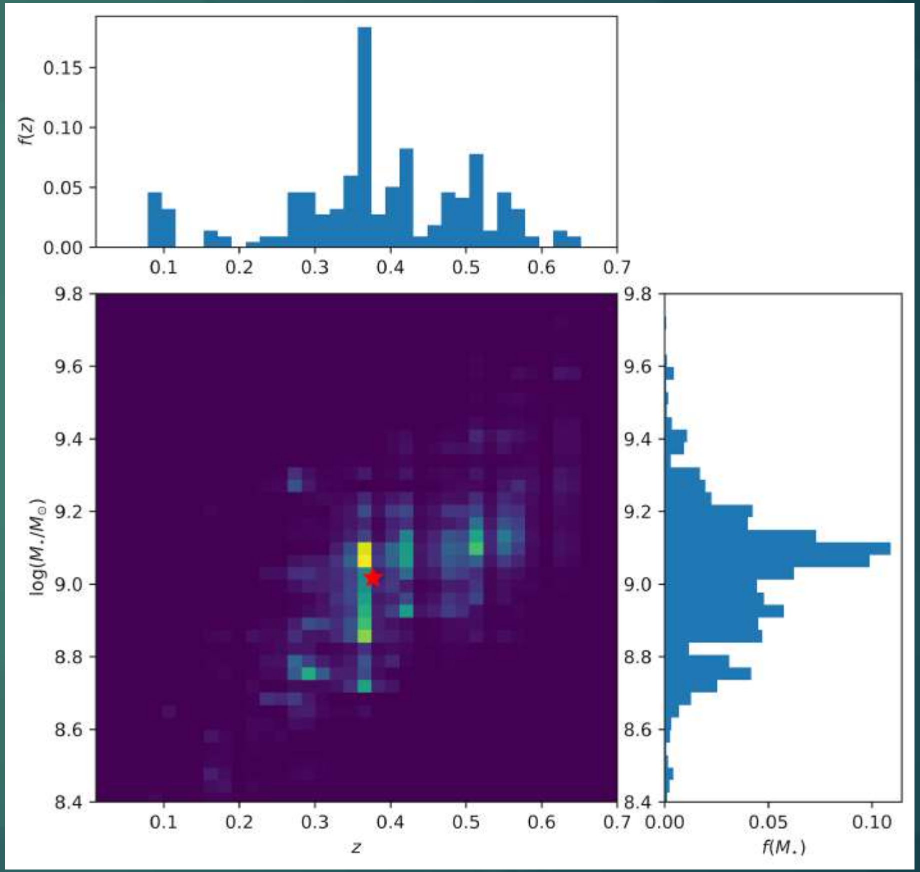
1. Train the first model to predict redshift.
2. Train the second model to predict stellar mass but include redshift + all features used in the first model.
3. For a test galaxy, extract the marginal pdf of redshift from the first model.
4. For each value of redshift, run the second model to extract conditional pdf of stellar mass | redshift. All the other features are kept the same.
5. Bin each conditional probability distribution into fixed redshift and stellar mass bins.
6. Finally, multiply the binned conditional probability distributions by the marginal pdf of redshift to get the joint pdf.

$$f(M, z) = f(M|z) * f(z)$$

Results: Joint PDFs

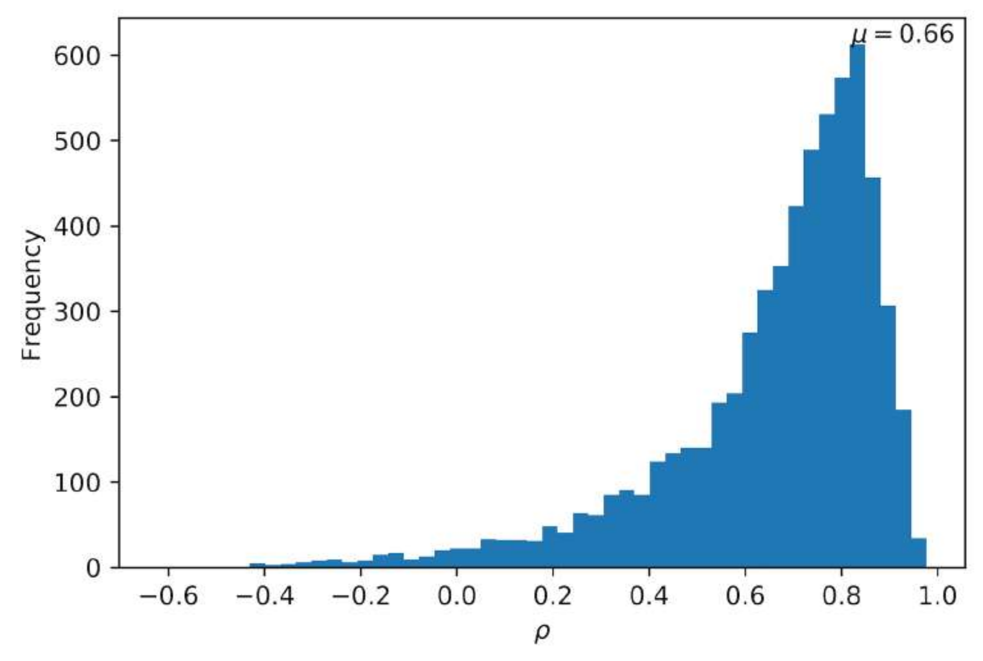


Simultaneous

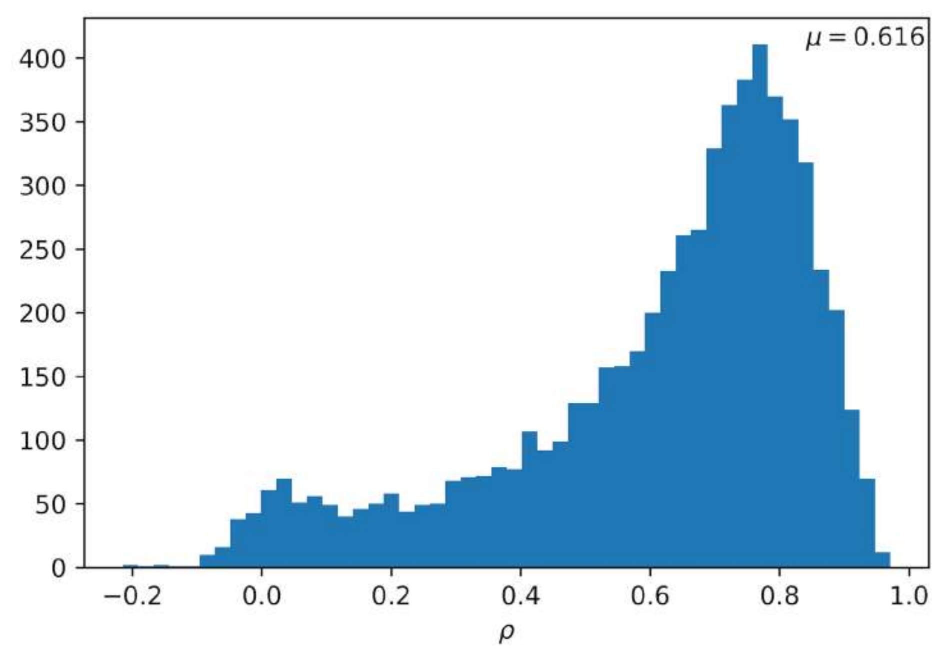


Separate

Results: Correlation Coefficient Distributions



Simultaneous



Separate

Summary and Next Steps…

- Used random forest to predict point estimates of redshift and stellar mass.
- Extracted pdfs of redshift and stellar mass.
- Tested their validity using PIT distributions and metrics. They appear to be valid.
- Used two different methods to build joint pdfs of redshift and stellar mass. Both produce similar results.

- Validate the joint pdfs using synthetic data.
- Goal is to build a python package which can produce joint pdfs on the fly. Future applications to LSST and Euclid.
- Method can be used to extract joint pdfs for any other galaxy properties.

Thank you. Any questions?