Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Improving photo-z estimation under covariate shift with StratLearn

Chiara Moretti, R. Trotta, M. Autenrieth,
D. van Dyk, R. Serra, +

Spoke 3 General Meeting, Elba 5-9 / 05, 2024

SISSA DATASCIENCE
Machine Learning for the Natural Sciences

SISSA

Scientific Rationale

Covariate shift

Different distributions in source (training) set and target dataset

$$p_S(x) \neq p_T(x) \qquad \text{but} \qquad p_S(y|x) = p_T(y|x)$$

→ can be due to **selection effects** (e.g. brighter/low redshift objects more likely to be observed)

Ubiquitous in astronomy!

→ ML algorithms show **poor generalisation** properties

Photometric redshift estimation

- obtain redshifts of several objects at once from imaging (vs spectroscopy, more accurate but more expensive)
- Key in ongoing/future cosmological surveys like Euclid, LSST
- Typically estimated with template fitting or **ML based methods**

Technical Objectives, Methodologies and Solutions

→ Our proposed solution: StratLearn

Code declined for photo-z estimation (applied to lensing in arXiv:2401.04687)

- Data partitioned in strata, based on **propensity scores**

$$e(x_i) = P(s_i = 1|x_i)$$

→ Estimated via binary classification, via logistic regression

- Conditional density estimators (Series, ker-NN) trained within each stratum, then combined with weighted average

→ Approach is **general and multi-purpose**

→ Can be combined with other estimators/models

Timescale, Milestones and KPIs

MILESTONE 7:

Target: Porting code from R to julia → 50x faster
KPI: code ported and available in public repo (StratLearn for photo-z)

Target: Assess performance in context of photo-z estimation on data that feature covariate shift
KPI: Application to simulated data produced for LSST (from Stylianou+2022)

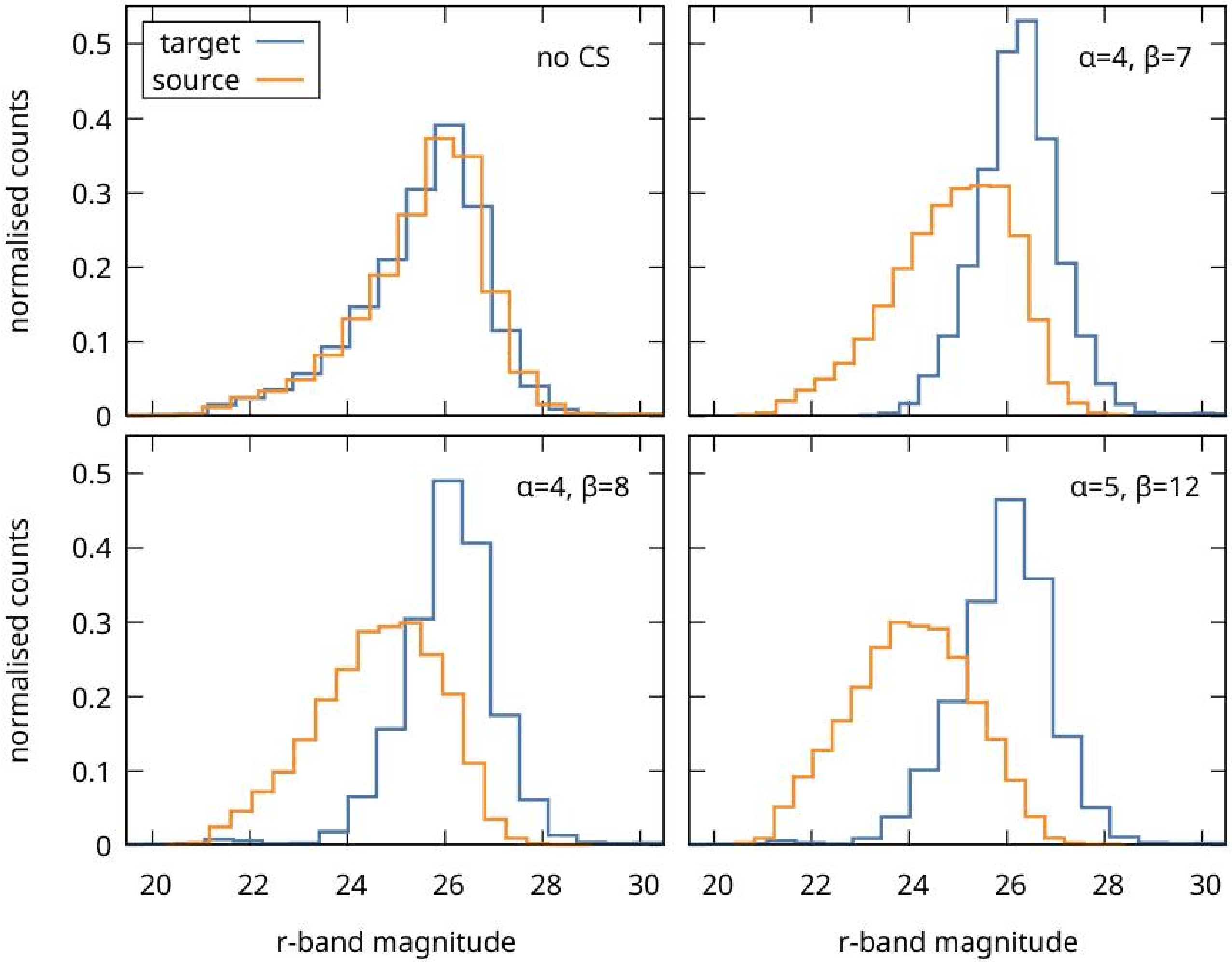
Paper almost ready!

Accomplished Work, Results

Assess performance on **simulated data** (Buzzard flock simulations, produced for LSST)

→ 100k simulated galaxies, spectroscopic (true) redshift + photometry in 6 bands (*ugrizy*)

- Introduce **CS with rejection sampling on r-band**, using Beta distribution (same approach as Izbicki+16)
- Partition data based on propensity scores + training
- Cond. density estimation (redshift pdf)

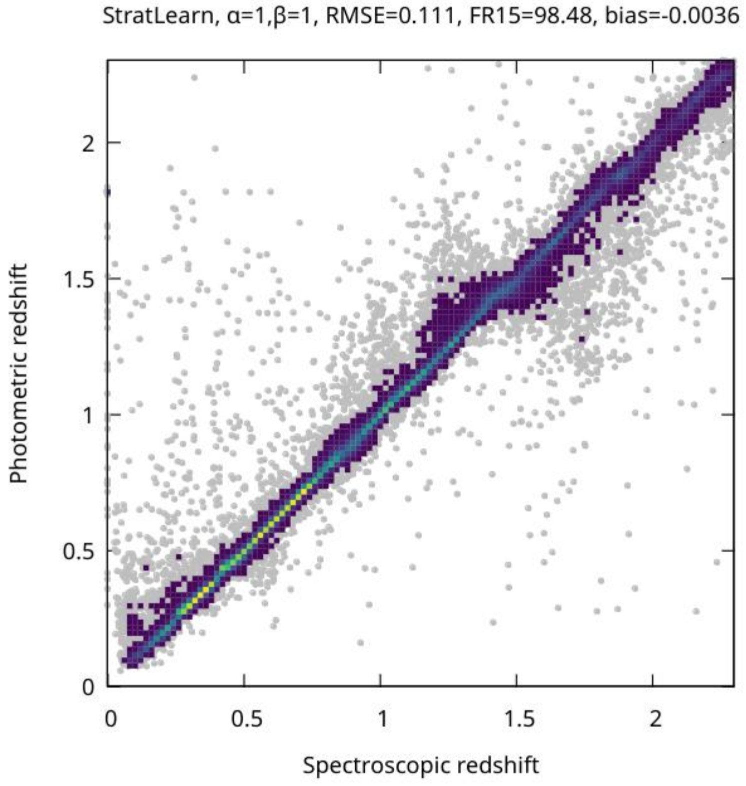


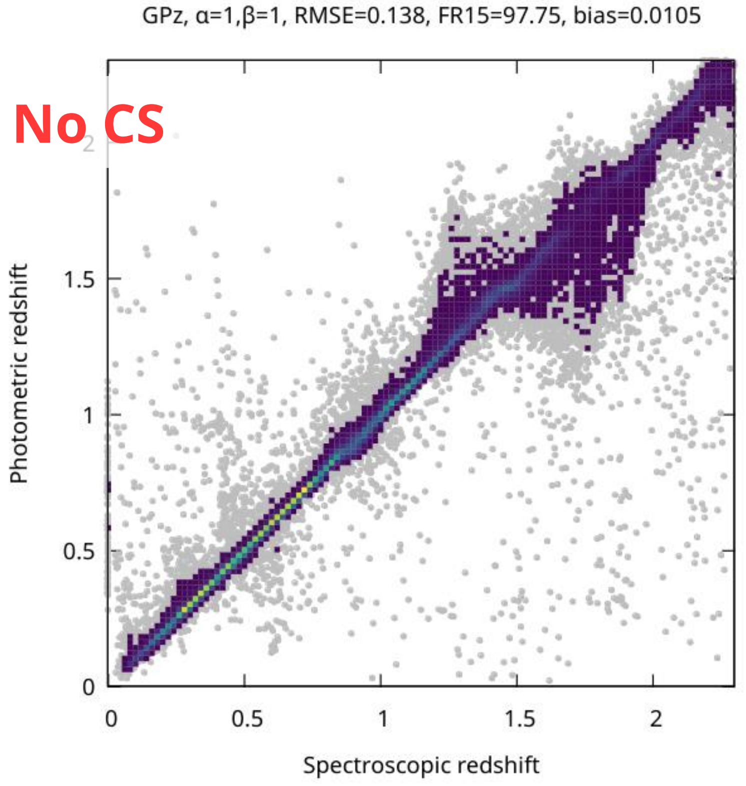
Accomplished Work, Results

StratLearn performance:

Several metrics to assess redshift point estimates, PIT for redshift pdf

Benchmark against **GPz** code





Improved performance in all scenarios explored

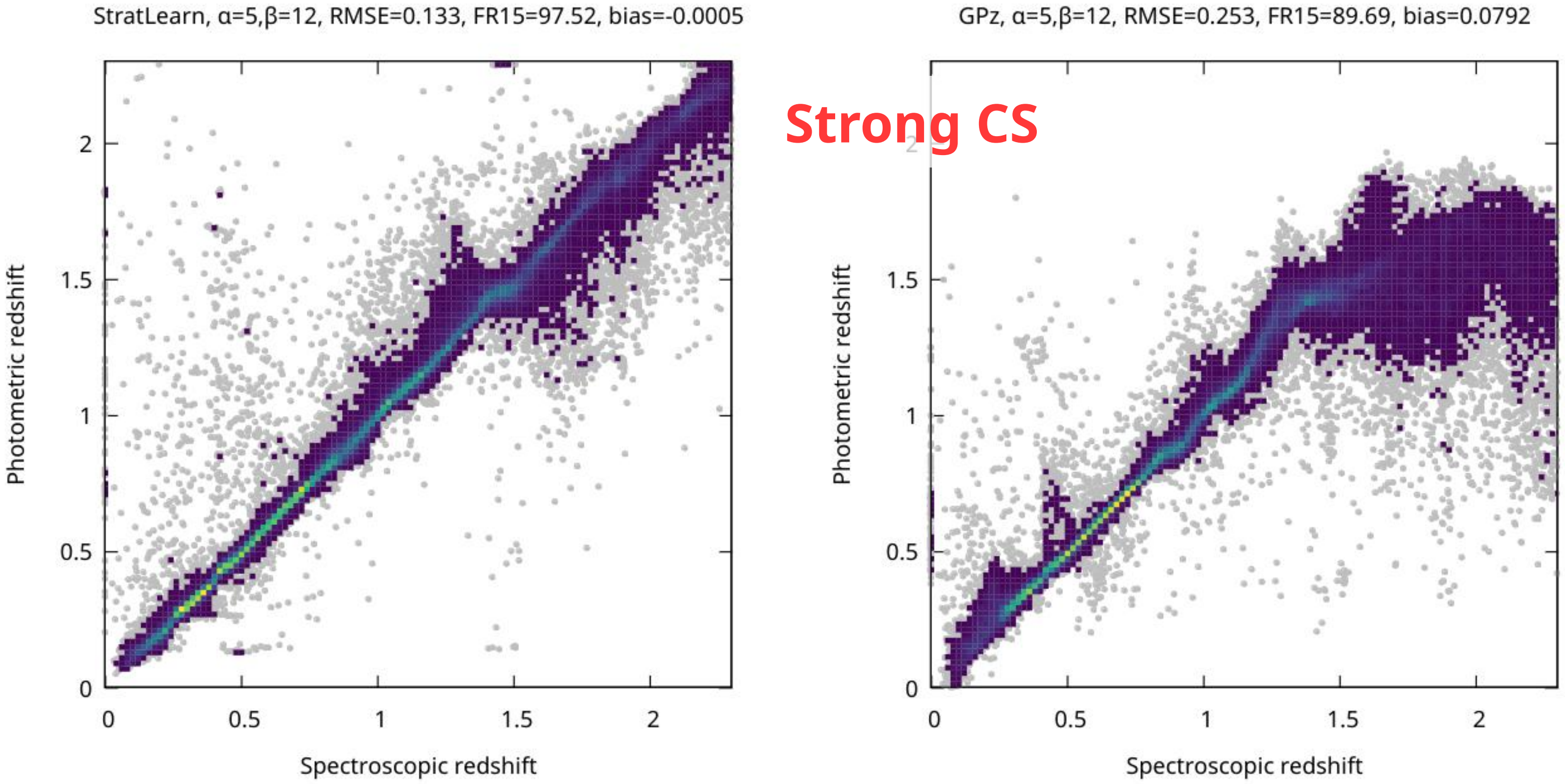
- Reduced bias
- Reduced error
- Less catastrophic errors

Accomplished Work, Results

StratLearn performance:

Several metrics to assess redshift point estimates, PIT to estimate redshift pdf

Benchmark against GPz code



Improved performance in all scenarios explored

- Reduced bias
- Reduced error
- Less catastrophic errors

Next Steps and Expected Results

Work in progress:

- Application to simulations: **paper** submission (May 2024)
- Code optimisation + some restructuring for easy usage
  - Julia offers more flexibility and easier to maintain!
- **First steps toward parallelisation**
- Apply to simulated (but realistic) data with Euclid-like properties

- Final goal is application to **real Euclid data**!