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, Roberto Trotta, M. Autenrieth, D. van Dyk, A. Mesinger

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024

Scientific Rationale

Covariate shift

Unrepresentative training datasets → $p_S(x) \neq p_T(x)$ but $p_S(y|x) = p_T(y|x)$

→ ML algorithms show **poor generalisation**

Ubiquitous problem in astronomy! Due to **selection effects** (brighter/low redshift objects more likely to be observed)

GOAL: improve generalisation properties of ML algorithms in presence of covariate shift

Scientific application:

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

→ Proposed solution: StratLearn

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

- Data partitioned in strata, based onquantiles of **propensity scores**

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

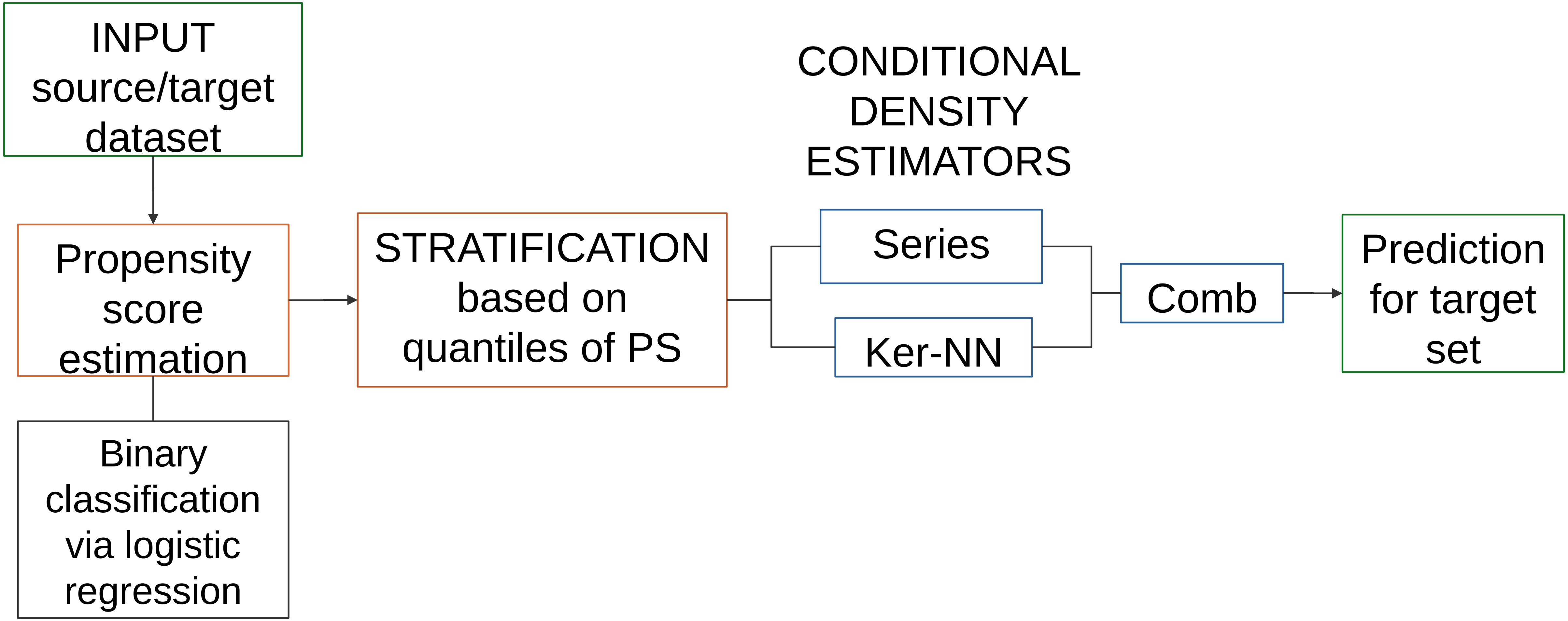
→ Estimated via binary classification with logistic regression

- 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

Technical Objectives, Methodologies and Solutions



INPUT source/target dataset

Propensity score estimation

Binary classification via logistic regression

STRATIFICATION based on quantiles of PS

CONDITIONAL DENSITY ESTIMATORS

Series

Ker-NN

Comb

Prediction for target set

Main Results

- Previous milestones
- Original code ported from R to julia → 50x faster KPI
- Code optimisation → 10x faster KPI
- Introduction of yaml parameterfile for easy usage
- Public github repository available at github.com/chiaramoretti/StratLearn-z KPI

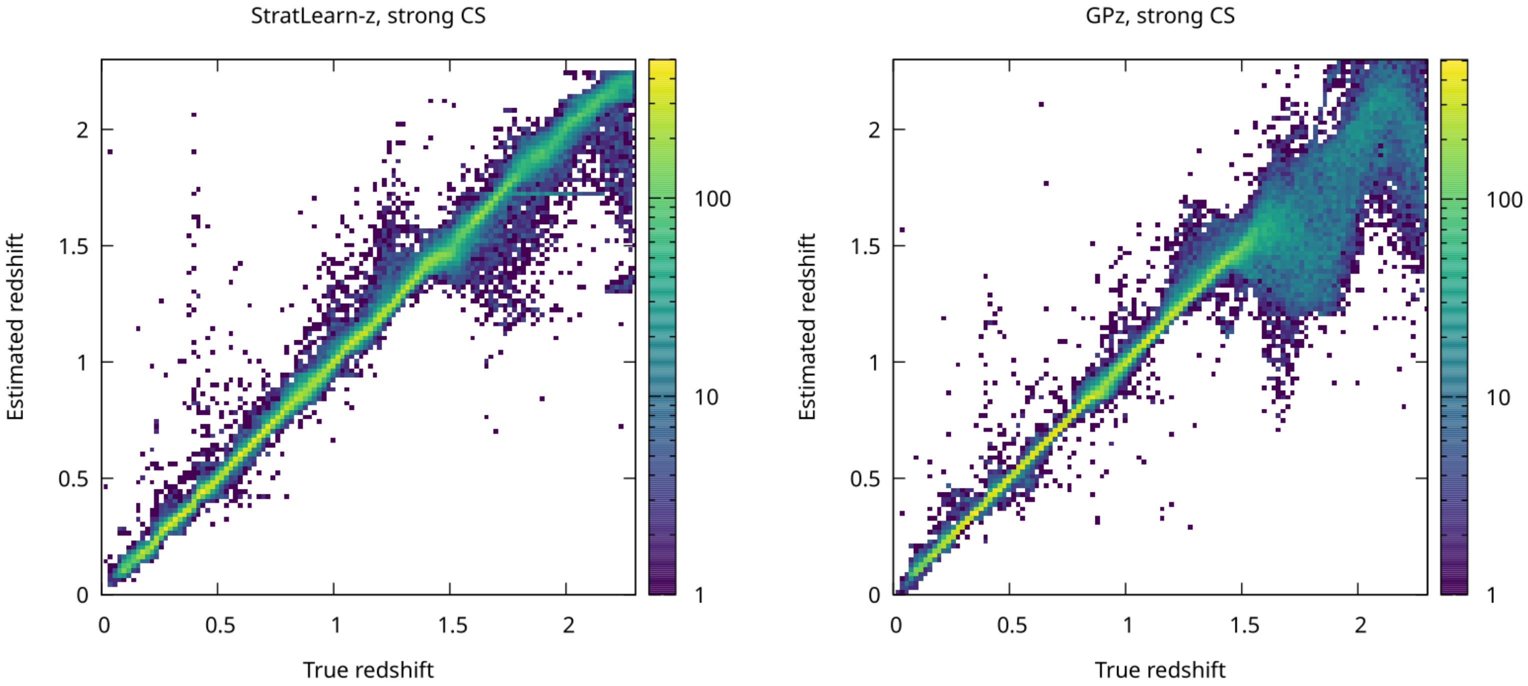
What's new?

- Generalised to read covariates from input datafile
- Additional script that only performs stratification → **easy combination with external photo-z codes**

Main Results

Application to simulated dataset (Buzzard flock simulations produced for DES, LSST) with introduced covariate shift

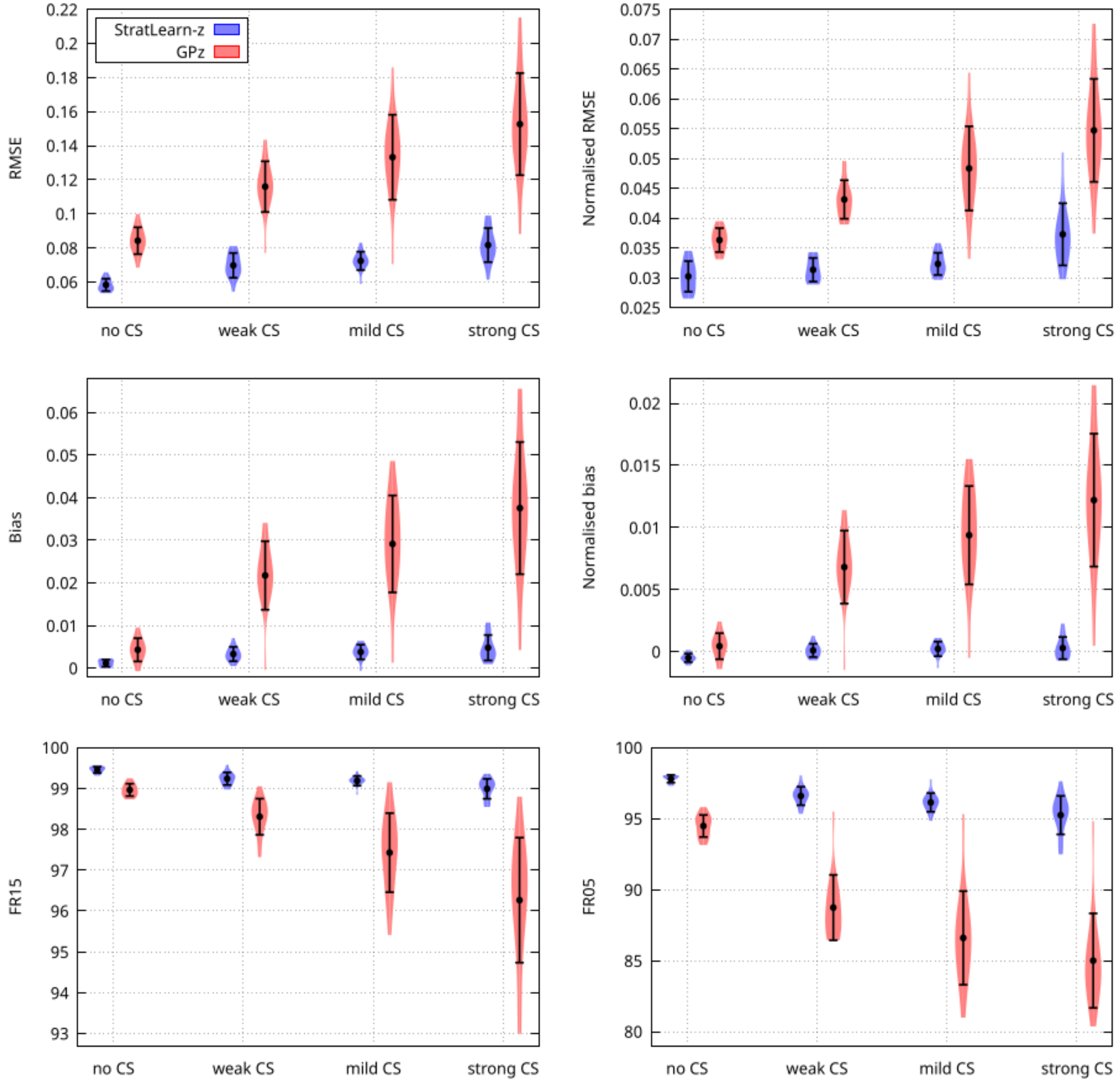
→ 100k objects with *ugrizy* photometry + redshifts
→ CS introduced by performing rejection sampling on the r-band



Main Results

Application to simulated dataset (Buzzard flock simulations produced for DES, LSST) with introduced covariate shift

Comparison with GPz code: **improved results** on all point estimate metrics considered



Main Results

Application to simulated dataset (Buzzard flock simulations produced for DES, LSST) with introduced covariate shift

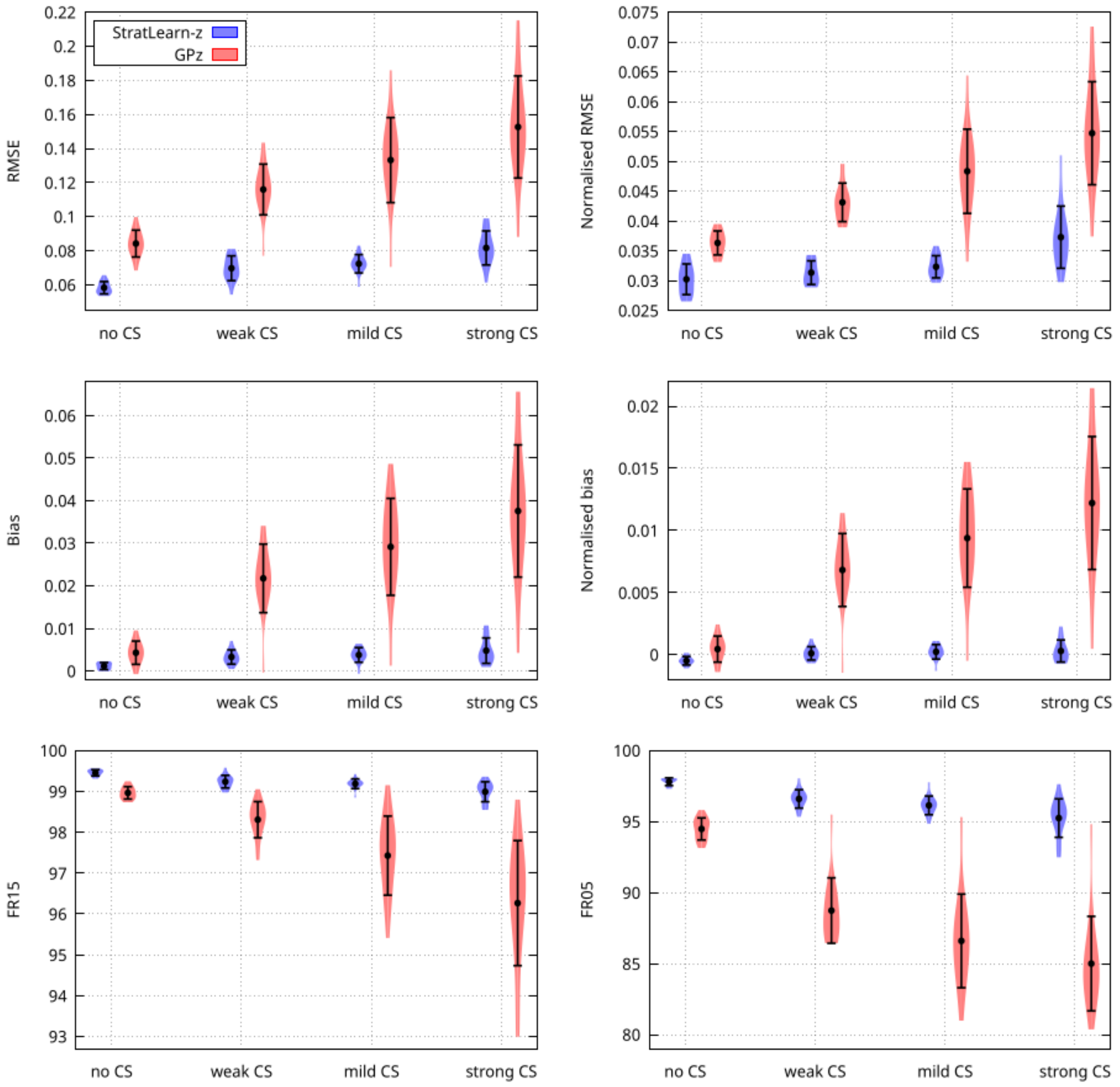
Comparison with GPz code: **improved results** on all point estimate metrics considered

Paper submitted (first review round completed) arXiv:2409.20379

KPI

Poster presentation @ COSMO

KPI



Final Steps

Ongoing work:

- Application to Euclid-like dataset based con COSMOS field
- → more realistic, used in Euclid photo-z challenge

KPI

25% completed
Expected to be done by April

- First step towards parallelisation: first target identified, currently ongoing

KPI

20% completed
Expected to be done by March

- Further optimisation of conditional density estimators

- Looking into combination with further models

Feasibility by end of contract still TBD