Letter of Recommendation for the DNF photo-z estimator

Title: Proposal for Directional Neighborhood Fitting as a general purpose photo-z estimator
Contributors to this letter: Ignacio Sevilla, Juan de Vicente, Laura Toribio (CIEMAT - Spain)
Endorsers: Jacobo Asorey (CIEMAT), Santiago Ávila (IFT), Francisco Castander (ICE-CSIC, IEEC), Juan García-Bellido (IFT), Juan Mena (CIEMAT), Ramon Miquel (IFAE), Anna Porredon (OSU), David Sánchez (CIEMAT), Eusebio Sánchez (CIEMAT)

0. Summary statement

Directional Neighborhood Fitting (DNF) is a nearest neighbor algorithm for photometric redshifts described in De Vicente, Sánchez & Sevilla-Noarbe (2016). It works on color and magnitude/flux space relying on a 'training' or reference data set to provide an estimate of the target galaxy photo-z and N(z) distribution through a fit to a hyperplane in this multivariate space, using a variable number of neighbors. It has been used in a variety of science cases successfully with the Dark Energy Survey (DES) data, and is one of its official photo-z estimators, included in the Gold catalogs.

We think this is a quick, well maintained, tested and performant photo-z estimator, with a good value for cross-validation with legacy data.

1. Scientific utility

DNF has been used in many science cases in the context of the Dark Energy Survey project, both using its point and N(z) estimates. The former quantities have been commonly used as part of selection functions for objects or extragalactic astrophysical studies. The latter as an ingredient for cosmological model fits. It works using either fluxes or magnitudes, and uses their estimates of the errors for improving its outputs.

A few recent examples in scientific literature follow:

- Vega-Ferrero et al. 2021 (a morphological catalog for DES Y3 data)
- The DES Collaboration 2021 (DES Y3 BAO measurement)
- Porredon et al. 2021 (estimation of galaxy-galaxy lensing constraints from DES Y3 data)
- Pocino et al. 2021 (optimization in simulations of the Euclid sample for clustering)
- Tanoglidis et al. 2021 (low surface brightness galaxies discovered in the Dark Energy Survey)
- Pieres et al. 2020 (Milky Way modelling using Dark Energy Survey data, DNF was used to aid in star-galaxy separation)
- Palmese et al. 2020 (statistical standard siren measurement for gravitational wave events)

- 44 other citations as of September 2021

It is also one of the official Dark Energy Survey photo-z estimates to be released by January 2022 as part of the DES Y3 cosmology dataset public release (see Sevilla-Noarbe et al. 2021 for details on the DES Y3 Gold catalog). Currently, it is being used as one of the exploratory estimators to combine DES and radio sources from the EMU pilot survey and as one of PAUS photo-z algorithms.

The DNF photo-z estimates at the time that the Vera Rubin Observatory LSST starts will be available in public DES data for O(500 M) common sources, which can be of scientific interest in order to compare with DES scientific results and catalogs. This complies with the recommendation of having vetted and widely used photo-z catalogs over the LSST footprint.

Currently, a version based on DES Year 1 photometry is available publicly at https://des.ncsa.illinois.edu/releases/y1a1/photoz.

2. Outputs

The following outputs are available for every object:
- Point estimate from a hyperplane fit in color/magnitude space (most accurate);
- Nearest neighbor association (provides best N(z) approximation in our experience);
- PDF estimate per galaxy using fit residuals;
- Independent photo-z uncertainty estimates coming either from flux/mag errors and from the fit to training sample;
- Nearest neighbor metrics to evaluate training set completeness.

3. Performance

DNF has had an active development in recent years that has been accompanied by performance metric tests and validations:

- Sanchez et al. (2014) was the first real data test for a prototype version of DNF (NIP-kNNz), where its good error and N(z) reconstruction is highlighted.
- De Vicente, Sánchez & Sevilla-Noarbe (2016) provides the description of the algorithm, and includes checks versus SDSS DR10 and VVDS datasets, as well as a run over the PHAT common dataset described in Hildebrandt et al. 2010. DNF was one of the best performant in terms of bias and scatter, with intermediate performance for outliers.
- Desprez et al., Euclid Collaboration (2020) tested several estimators including DNF with a good point estimate performance and outlier flagging.
- Sevilla-Noarbe et al. 2021 shows that in the validation dataset built for DES and described in Gschwend et al. (2018), DNF is the best code available in the Gold catalog in terms of bias and scatter.

- It has also provided excellent results in non-published results on MICE and Euclid Flagship simulations.

It has also been run on the same data set as the DESC study described in Schmidt et al. 2020 providing a similar performance as other codes in this comparative analysis of the Buzzard simulation ($\sigma_{IQR}/(1+z)$ ~ 0.016, mean (not median) bias ~ 0.003, $3\sigma_{IQR}$ outlier rate = 6%).

As with every 'training' based code, it has an important reliance on the availability of spectroscopic or narrow band external data. We are working actively towards providing an estimate of the reliability of the estimates through flags using internal consistency checks, number and distance of available training objects in the directional color/magnitude space that the algorithm uses.

4. Technical Aspects

The code is written in Python 2 as of today. We are working on the update of the code and its publication as we progress towards the DES Y3 cosmology public data release in the next few months. It has been used successfully by external parties for execution with the MICE simulations (http://maia.ice.cat/mice/), Buzzard simulations (https://buzzardflock.github.io/), and DES Science Portal (Gschwend et al. (2018)) by the respective owners of these infrastructures.

Scalability -- will meet. Photo-z estimates are produced at <1 ms/galaxy in estimation, including PDF outputs using a 2.3 GHz Xeon Gold 5118 CPU core.

Inputs/Outputs -- will meet. Inputs fluxes/magnitudes and their errors, from catalogs, as specified in Appendix B of DMTN-049. Other inputs (e.g. shapes) are possible when relevant information can be added. Outputs are point estimates, uncertainties and individual pdfs. These are produced in chunks of FITS files which we have parcelled, uploaded and checksummed against an Oracle DB routinely, using DES specific software in python (https://anaconda.org/conda-forge/des-easyaccess) which is of general use.

Storage constraints -- will meet. Those necessary to store the external, vetted reference set (-1 GB per half million training galaxies in DES Y6). PDF storage space is already budgeted for as per DMTN-049 in our understanding.

External data sets -- will meet. The reference data set currently used is available here. It is expected to be broadened as the DES progresses towards completion in its analysis. Current results shown in papers above are based on mostly public data, which will be completely public by the time of commissioning (specific OzDES spectra pending as of this writing, O(50k spectra)). This public repository holds 1.5 million spectra, with 10% being over z > 1.

Estimator Training and Iterative Development -- will probably meet. The algorithm, as it is not a machine-learning one strictly speaking, does not require a pre-training phase. It is done 'on the fly' at minimum CPU cost but with some memory requirements (see below). As with all training set based codes however, it implies a large reliance on the representativeness of this reference data set for accuracy. We intend to support the creation of reliability flags (in development as of today, as part of a DES program) to palliate this issue.
On the other hand, we have the ability to fully support this effort with in-house staff personnel, ideally also as part of the BCNMAD international team in-kind contribution.

Computational Processing Constraints -- will probably meet. Depending on the size of the training set and the photometric sample partition, the memory footprint will vary. For DES, this amounts to up to 16 GB of RAM per core. These catalogs were produced for ~300M objects within two days using a modest cluster allocation (7 nodes with 2 x Xeon Gold 6148 2.4Ghz with 20 cores each), as part of the DES Science Release process. This was run after the catalog was internally released by DES, at CIEMAT's premises by the code developers. However, the code has been implemented in the photometric redshift pipelines from the DES Science Portal (Gschwend et al. (2018)) and by other external parties for adding value to simulations without further issues.

Implementation Language -- will meet: Python2. It has been certified to run on CentOS at CIEMAT's premises but also on other Linux systems by external parties at Observatorio Nacional in Brazil and SLAC.

The code has a long-term commitment to active maintenance by the main author, who is a staff scientist working on this subject. The BCNMAD international team can provide additional support for implementation/execution as an in-kind contribution, if this is deemed appropriate.