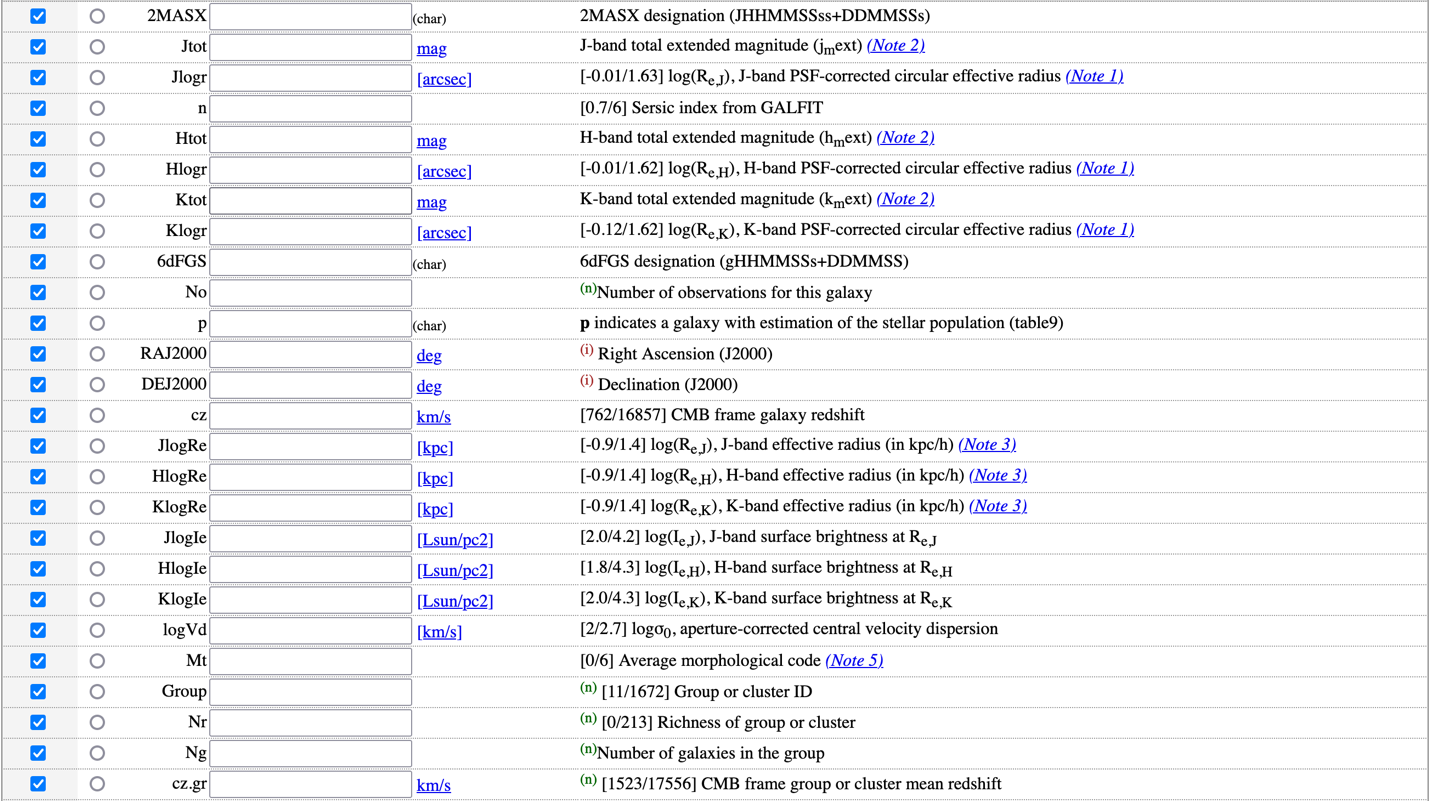
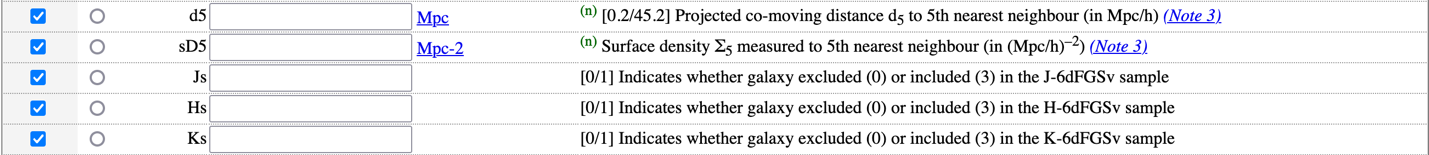
WHAT I HAVE BEEN DOING

1. **Obtaining the raw spectroscopy/velocity dispersions data**
2. 6dFGS (there are two data sources, I have not inspected whether they are different):
   * + 1. First source is what I actually used throughout my thesis, Campbell’s table on Vizier obtained [here](https://vizier.cds.unistra.fr/viz-bin/VizieR-3?-source=+J%2FMNRAS%2F443%2F1231%2Ftable2&-from=nav&-nav=cat%3AJ%2FMNRAS%2F443%2F1231%26tab%3A%7BJ%2FMNRAS%2F443%2F1231%2FFPsample%7D%26key%3Asource%3DJ%2FMNRAS%2F443%2F1231%2Ftable2%26HTTPPRM%3A%26) . I use two tables, first is FPsample table (FP quantities, 11102 galaxies) with the following schema:





second table is table2 (velocity dispersion, 11503 galaxies) with the following schema:

A screenshot of a computer

Description automatically generated

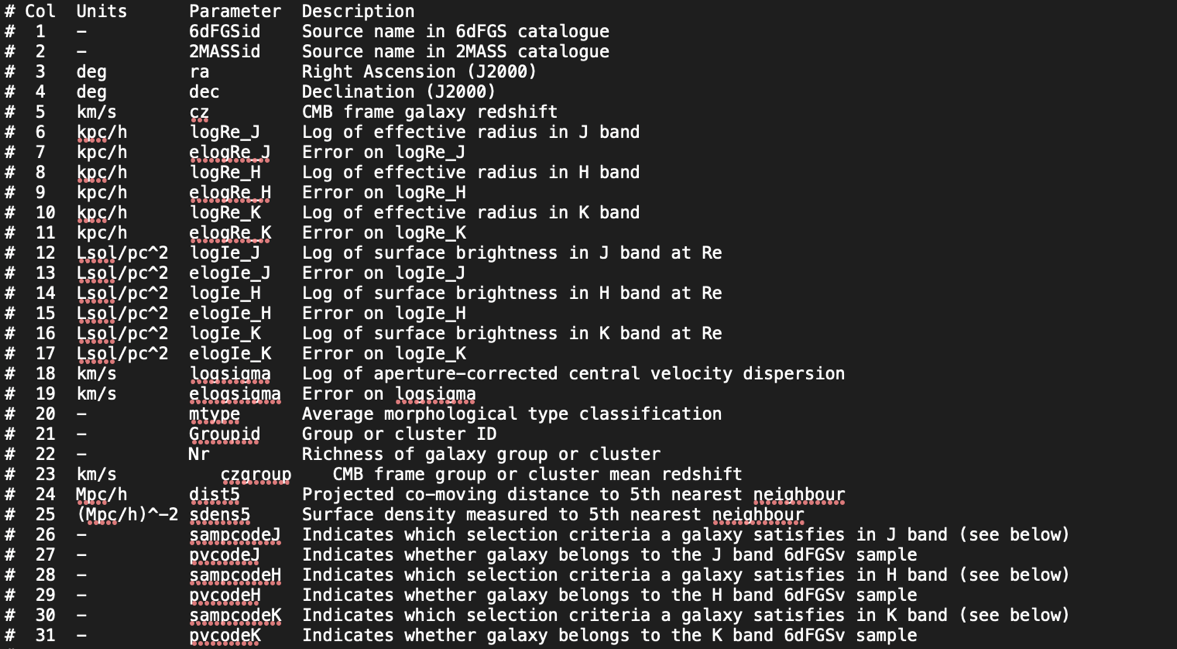
* + - 1. Second source is from Campbell et al. (2014) on MNRAS obtained [here](https://academic.oup.com/mnras/article/443/2/1231/1060810) (supplementary data section). The files are campbell\_table2.ascii (velocity dispersion data), campbell\_table4.ascii (NIR photometry) and campbell\_table8.ascii (derived FP data).
         1. table2 schema (unsure if the descriptions are true):

|  |  |  |  |
| --- | --- | --- | --- |
| column name | type | Unit | description |
| sdf | string |  | 6dFGS id |
| tmass | string |  | 2MASS id |
| mjd | float |  | Modified julian date of the observation |
| Z\_helio | float | Redshift | Heliocentric redshift |
| s2n | float |  | Signal to noise ratio of velocity dispersion measurement |
| Cor\_sigma | float |  | Cross-correlation R parameter? |
| Dex\_error | float | [km/s] | Error of log\_Vd? |

* + - * 1. table4 schema

|  |  |  |  |
| --- | --- | --- | --- |
| column name | type | Unit | description |
| MASS\_name | string |  | 2MASS id |
| J\_m\_ext | string | Mag | Extrapolated J magnitude |
| Fwhm\_j | float | Arcsecond? | Modified julian date of the observation |
| Delta\_r\_j | float | Arcsecond? | PSF correction? |
| Log\_re\_j | float | [arcsecond]? | Log PSF-corrected effective radius |
| nser | float |  | Sersic index |
| H\_m\_ext | float | Mag | Extrapolated H magnitude |
| Fwhm\_h | float | Arcsecond? |  |
| Delta\_r\_h | float | Arcsecond? |  |
| Log\_re\_h | float | [arcsecond]? |  |
| K\_m\_ext | float | Mag | Extrapolated K magnitude |
| Fwhm\_k | float | Arcsecond? |  |
| Delta\_r\_k | float | Arcsecond? |  |
| Log\_re\_k | float | [arcsecond]? |  |

* + - * 1. table8 schema



1. SDSS (34059 galaxies): obtained Cullan’s SDSS FP/PV sample.
2. LAMOST (85861 galaxies): the file lamost\_DR7\_VDcat\_20200825.fits obtained from Khaled. Schema:

A screenshot of a computer error

Description automatically generated

**Note:**

* I used vizier for 6dFGS because the tables on MNRAS do not have the raw velocity dispersions.

1. **Obtaining supplementary data**

* Two supplementary data required: John’s radii measurements and Tempel et al. SDSS DR8 groups and clusters data.
* John provided colours table (data/raw/r\_e\_jrl/colours.ascii), radii measurements (data/raw/r\_e\_jrl/jhk\_r\_e.csv), and LAMOST ETG list (data/raw/r\_e\_jrl/lamost\_good\_pv\_list.csv)
  + Colours table (I didn’t use these, so I don’t know what most of the columns are):

|  |  |  |  |
| --- | --- | --- | --- |
| column name | type | Unit | description |
| tmass | string |  | 2MASS ID |
| Ra | float | deg |  |
| Dec | float | deg |  |
| J\_ext | float |  |  |
| J\_ext\_error | float |  |  |
| G\_r\_ext | float |  |  |
| GAIA3\_B\_R | float |  |  |
| GAIA3\_B\_R\_error | float |  |  |
| PS1\_g\_r\_5 | float |  |  |
| PS1\_g\_r\_5\_error | float |  |  |
| SM3\_g\_r\_5 | float |  |  |
| SM3\_g\_r\_5\_error | float |  |  |
| J\_K\_5 | float |  |  |
| J\_K\_5\_err | float |  |  |
| W2\_W3 | float |  |  |
| W2\_W3\_err | float |  |  |

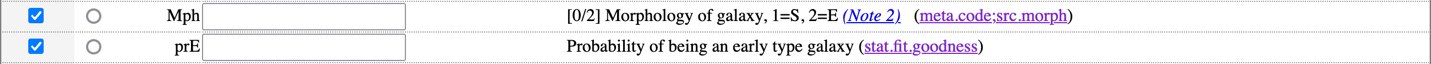
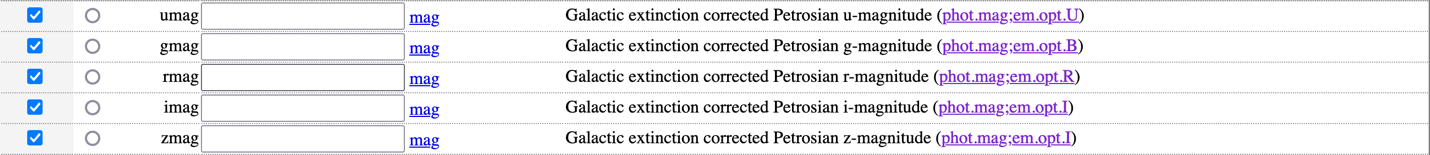
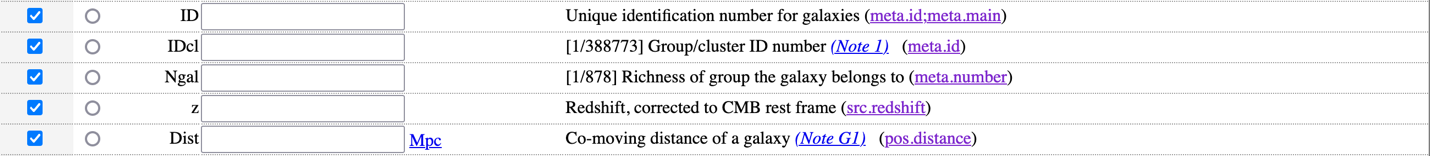
* + Radii measurements

|  |  |  |  |
| --- | --- | --- | --- |
| column name | type | Unit | description |
| tmass | string |  | 2MASS ID |
| Log\_r\_h\_app\_j |  |  |  |
| Log\_r\_h\_smodel\_j |  |  |  |
| Log\_r\_h\_model\_j |  |  |  |
| Red\_chi\_j |  |  |  |
| Galfit\_ser\_j |  |  |  |
| Fwhm\_j |  |  |  |
| Fit\_ok\_j |  |  |  |
| Log\_r\_h\_app\_h |  |  |  |
| Log\_r\_h\_smodel\_h |  |  |  |
| Log\_r\_h\_model\_h |  |  |  |
| Red\_chi\_h |  |  |  |
| Galfit\_ser\_h |  |  |  |
| Fwhm\_h |  |  |  |
| Fit\_ok\_h |  |  |  |
| Log\_r\_h\_app\_k |  |  |  |
| Log\_r\_h\_smodel\_k |  |  |  |
| Log\_r\_h\_model\_k |  |  |  |
| Red\_chi\_k |  |  |  |
| Galfit\_ser\_k |  |  |  |
| Fwhm\_k |  |  |  |
| Fit\_ok\_k |  |  |  |

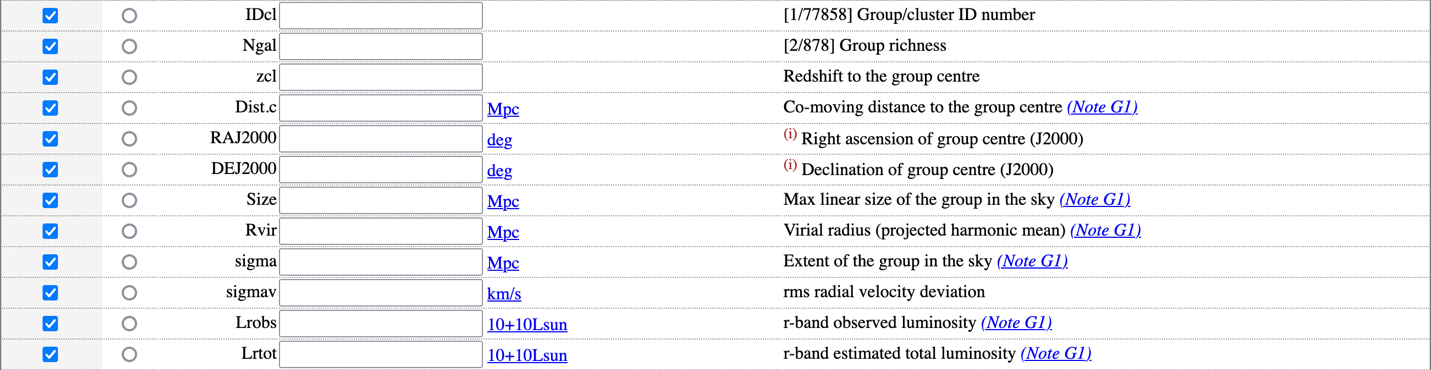
* + lamost\_good\_pv\_list.csv

|  |  |  |  |
| --- | --- | --- | --- |
| column name | type | Unit | description |
| Col1 | string |  | 2MASS ID |
| 2MASS | string |  | 2MASS ID |

* I downloaded Tempel et al. SDSS DR8 groups and clusters from vizier ([here](https://vizier.cds.unistra.fr/viz-bin/VizieR-3?-source=J/A%2bA/540/A106/dr8gal)). It contains two tables:
  + dr8gal 🡪 individual galaxy data. I filtered Ngal > 1 to exclude groups with only 1 member



* + dr8gr 🡪 groups and clusters data



**Note:** is vizier the correct place to look for astronomical data?

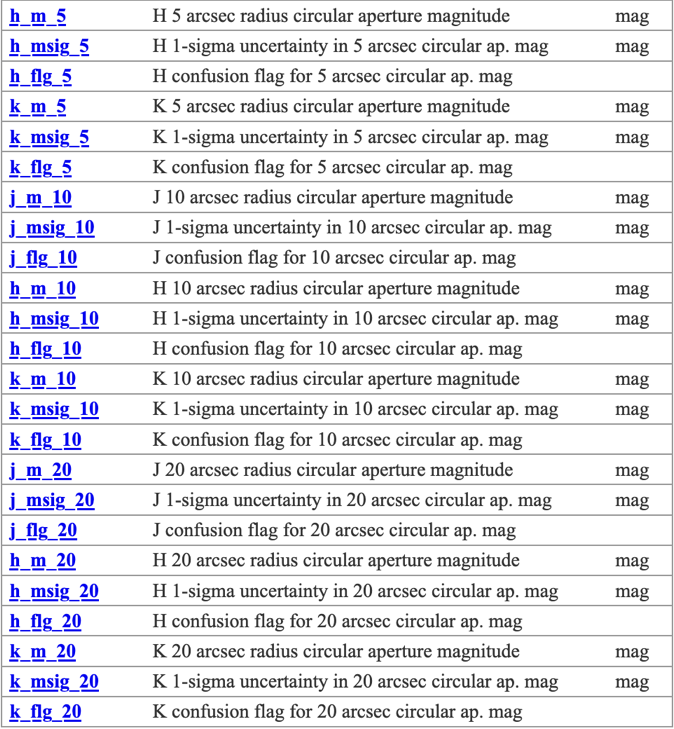
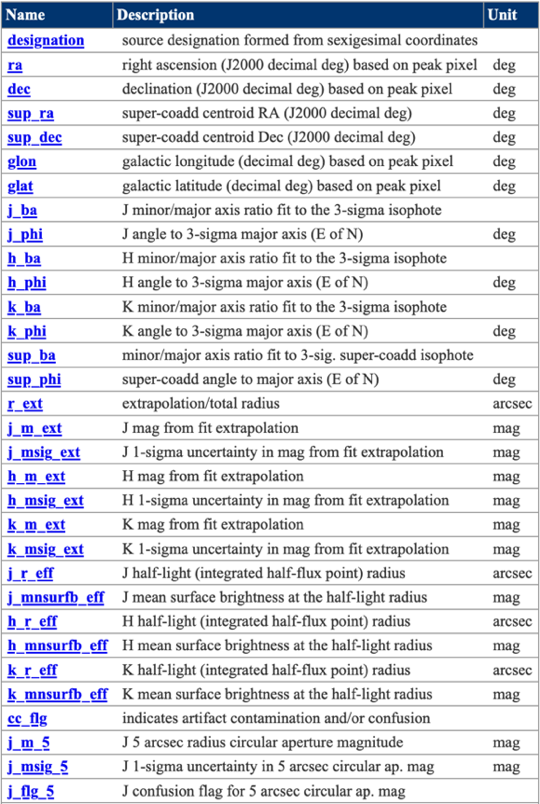
1. **Obtaining the sky coordinates from all the raw galaxies and obtaining 2MASS data**

* I use the get\_coordinates.py script to fetch the (ra, dec) and store them to data/preprocessed/sky\_coord/<survey>.ascii file (IPAC format).
* I had to multiply 6dFGS ra by 15 to get the ra in degrees (it was given in hour), while for SDSS and LAMOST I fetch the ra and dec directly.
* To set the cone search radius, I first queried 6dFGS galaxies to 2MASS XSC. I compared each input galaxy’s ra dec to its 2MASS counterpart’s ra dec. I found that the maximum separation was 2.22”.

A screenshot of a computer

Description automatically generated

* I set the cone search radius to be 2.25 arcsecond (based on the argument above but a bit more relaxed) and checked the ‘One to One Match’ box (basically doing a left join) so that I can simply merge the resulting dataframes later. These are the fields that I queried:



I saved the results at data/raw/2mass/<survey>\_tmass.csv. Also backed up the files in Google Drive (cannot push them to GitHub as they are too large).

* Note: is there a rule of thumb when doing sky cross-matching? How do we know we matched with the correct galaxies without visual inspection?

1. **Merge the original spectroscopy data with all others**
   * This step is done using combine\_spectrophoto.py
   * For 6dFGS (combine\_6df\_spectrophoto function):
     1. Open the 6dFGS FP sample (data/raw/6dfgs/sdfgs\_fp\_vizier.fits) containing 11102 galaxies.
     2. Open the 2MASS data containing 11102 photometry measurements (same number of rows and order as 6dFGS FP above).
     3. Join FP-2MASS by their dataframe indices.
     4. Check if the FP dataframe’s RAJ2000 and DEJ2000 are consistent with 2MASS dataframe’s ra\_01 and dec\_01. If consistent, I drop ra\_01 and dec\_01 from the joined dataframe.
     5. Open the 6dFGS velocity dispersions table (data/raw/6dfgs/sdfgs\_veldisp\_vizier.fits).
     6. 6dFGS veldisp table contains multiple measurements of the same galaxy. I picked the row with the best S\_N and drop the duplicates.
     7. Join 6dFGS FP-2MASS-veldisp.
     8. Save the table to data/preprocessed/spectrophoto.
   * For SDSS and LAMOST (I put them in one function combine\_sdss\_lamost\_spectrophoto as the steps are identical):
     1. Open the SDSS/LAMOST spectroscopy data (data/raw/sdss/SDSS\_spectro.csv and data/raw/lamost/lamost\_DR7\_VDcat\_20200825.fits).
     2. Open the 2MASS data.
     3. Join SDSS/LAMOST-2MASS by their dataframe indices.
     4. Check if SDSS/LAMOST dataframes’ ra and dec are consistent with 2MASS dataframe’s ra\_01 and dec\_01. If consistent, I drop ra\_01 and dec\_01 from the joined dataframe.
     5. Open John’s GALFIT measurements (radii, PSF corrections, other criteria).
     6. Join SDSS/LAMOST-2MASS-JRL on 2MASS id.
     7. Open Tempel’s individual galaxies and clusters data. Then I join them to get each galaxy’s cluster redshift.
     8. Join SDSS/LAMOST-2MASS-JRL-Tempel on galaxy’s ra and dec (Tempel does not provide SDSS objID).
     9. Save the table to data/preprocessed/spectrophoto.

* Note: I was not able to replicate the number of galaxies for SDSS. Last year, the SDSS+2MASS data was around 95000, but this time there are 158000 SDSS galaxies with 2MASS photometry. I don’t remember what CasJobs query I used to obtain SDSS previously.

1. **Derive r, s, i**
   * First, I installed dustmaps package using pip install dustmaps. Then, I run the src/utils/dustmaps\_downloader.py to download dustmaps\_sfd map to calculate Galactic extinction.
   * The r, s, i quantities are calculated using src/derive\_rsi.py script. It has the following steps:
     + 1. Extra step only for SDSS and LAMOST:
          1. append ‘SDSS’ and ‘LAMOST’ to their survey ID’s
          2. Perform the following selection criteria suggested by John:

Select fit\_ok\_j == ‘OK’

Select log\_r\_h\_model\_j > 0

Select red\_chi\_k <= 2

These removed 2212 galaxies from SDSS and 1237 galaxies from LAMOST.

* + - 1. Rename the columns for uniformity. Add ‘tmass’ column to 6dFGS and rename z\_lamost to z.
      2. Derive PSF-corrected radii. For 6dFGS, they are simply 10^Jlogr. For SDSS and LAMOST:
         1. Calculate the PSF correction
         2. Calculate the PSF-corrected radii
      3. Calculate CMB frame redshift for individual galaxies (also rederive for 6dFGS) using Khaled’s perform\_corr script from helio\_cmb.py script.



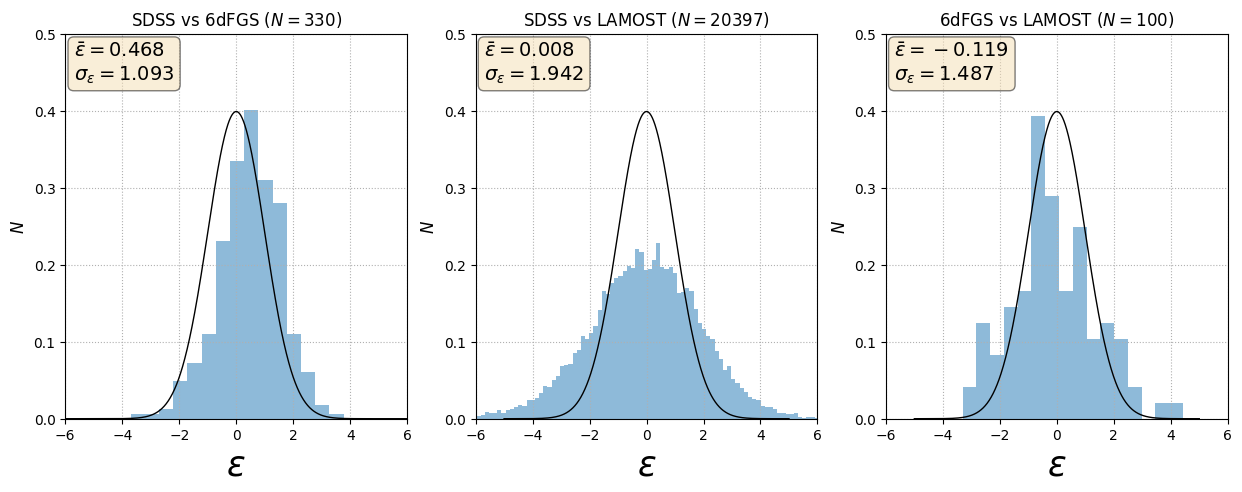
* + - 1. Create a new redshift column (z\_dist\_est) as distance estimator, which uses group or cluster mean CMB redshift if available, else using individual galaxy’s CMB redshift.
      2. Perform aperture size correction to the velocity dispersions.
         1. First, I calculated R-band radii from J-band radii (using circularized radii).
         2. Calculate aperture size-corrected velocity dispersion
      3. Calculate Galactic extinction in the JHK bands using SFD map.
      4. Calculate K-corrections using Chilingarian’s python script.
         1. I calculated the and colors first (corrected for Galactic extinction)
         2. Then I calculated the K-corrections using these colors
      5. Derive r and i
         1. First, I obtained redshift-distance table for to interpolate each galaxy’s estimated distance (dz\_cluster).
         2. Then I calculated circularized radii:
         3. Calculated r as follows:
         4. Calculated i as follows:
      6. Derive s
         1. Calculated s:
         2. Calculated es:
      7. Saved the files at data/processed/rsi\_derived/
  + Note: I compared the rsi results for 6dFGS with Campbell’s values. The veldisp (s) are pretty similar, but r and i are quite different. I think it’s because I used circularized radii, whereas Campbell did not.

A graph of a function

Description automatically generated with medium confidence

1. **Calibrating velocity dispersions**
   * Done in 4\_veldisp\_calibration.py.
   * Under the assumption of independence, the pairwise statistic

should follow a standard normal distribution (mean of 0 and variance of 1). But I got the following:



* I calibrated the velocity dispersions using two steps: first I scaled the error, and then I determined the offset.

1. **Finding common galaxies**

To perform the comparison, the repeat measurements are first obtained by matching the 2MASS id. In the script, this is done in the get\_common\_galaxies function. The galaxies with repeat measurements are stored at

data/processed/veldisp\_calibrated/repeat\_ori.csv

1. **Finding the error scaling**

Doing this with the method Matthew came up with. There are two approaches to calculate the error scalings. First is only inflating LAMOST errors (i.e. assuming k6df and ksdss to be 1), second is inflating two surveys (assuming only one k to be 1). I did the first approach, but only comparing the LAMOST errors with SDSS (because the scaling obtained from 6dFGS is so large, and the number of comparison is very small anyway). My implementation (get\_error\_scaling\_lamost\_only function in the script) for this approach:

* 1. First, I set the number of maximum iterations and the LAMOST scalings (all 1)
  2. At the beginning of the iteration, the SDSS and LAMOST errors are scaled (I scaled the log veldisp errors)
  3. Then I calculated the pairwise statistics for the LAMOST-SDSS or LAMOST-6dFGS common galaxies

Where ‘fiducial’ refers to either 6dFGS or SDSS

* 1. I applied sigma clipping: I only use the comparison if
  2. Then I store the remaining comparisons , the rms errors for LAMOST and the fiducial survey:

and the variance of the sigma-clipped pairwise statistics distributions (I simply find the squared of the standard deviation of the distribution, i.e. I did not fit a Gaussian and use the fitted std)

* 1. Then I calculated the new error scaling for LAMOST using the other two surveys as the fiducials:

The updated error scalings are then

I also checked the convergence. The iteration is stopped when

Usually the iteration already stops at the 3rd iteration.

* 1. I obtained: and
  2. The scaling with 6dFGS fiducial is too big, most likely because the common galaxies are few (69 galaxies). So I use the scaling with SDSS fiducial as LAMOST error.
  3. The scaling is applied as follows:

1. **Determining the veldisp offset**

I implemented sigcorr2.awk to Python (get\_offset function in the script). Steps:

* 1. First I defined the number of bootstraps. The first bootstrap, I directly use the measured velocity dispersions. For the other bootstraps, I use Monte Carlo sample

At the beginning of each bootstrap, I set the total offset for the three surveys to zero.

* 1. Then, iteratively, I calculate the offsets for the three surveys. At the beginning of the iteration, the offsets are subtracted from the non-subtracted (either measured values or MC values) velocity dispersions.
  2. For each iteration, I did the following for the three surveys (for loop):
     1. Start with 0 offset, offset error, significance, weight, and significance.
     2. Calculate the error-weighted veldisp in the target survey:

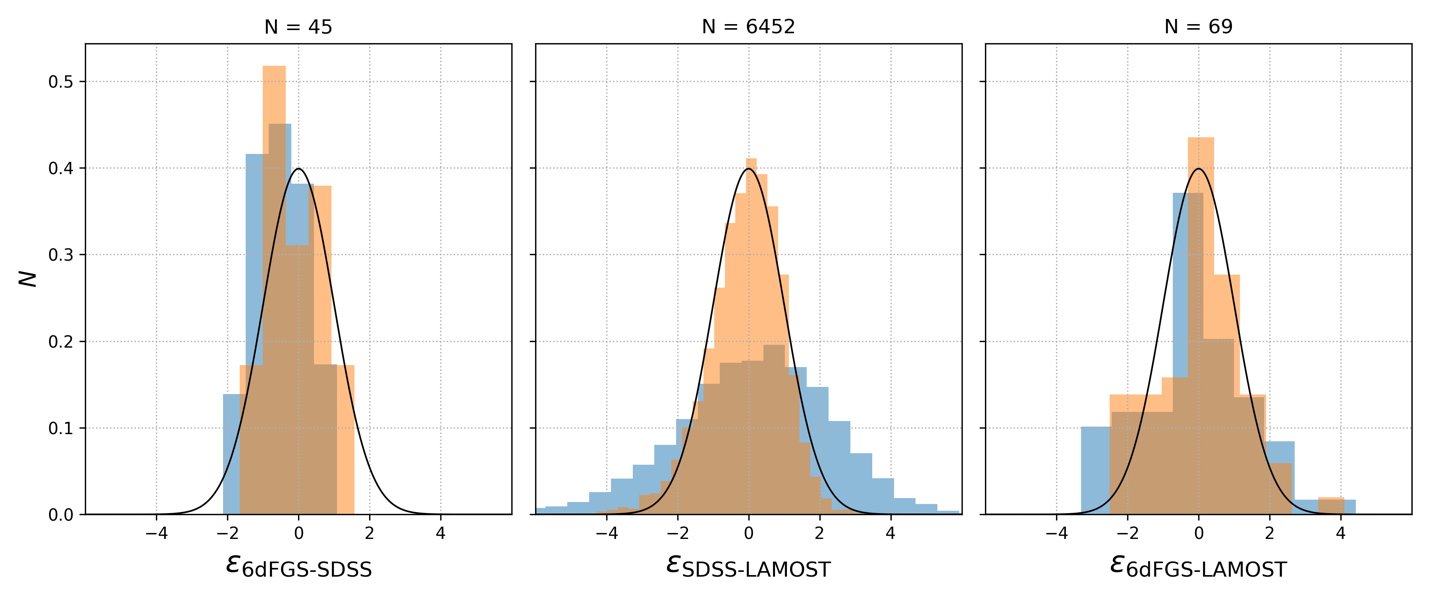
the (inverse) rms error

And the number of observations in the target survey (since always only 1 observation per survey).

And do the same for the observation in other surveys:

The sum is taken over the number of observation in other surveys, , which can be 1 or 2.

* + 1. For each galaxy, I calculated:
    2. Applied cut to galaxies where
    3. Determined the offset, error, and significance
    4. If , update the total offset (basically always update, but the threshold can be set):
    5. Iteration stopping criteria: maximum significance from the three surveys (offset/offset\_error) falls below 0.5 or the number of maximum iterations is reached.
    6. At the end of the iteration, I subtract all of the total offsets with SDSS offset (so SDSS’s totoff is 0).
    7. The offsets are estimated from the original data. The bootstrap samples are used to determine the robustness of the estimation. I obtained:
* Results:



* Notes:
  1. Should I have compared the linear velocity dispersions (Gaussian error?) instead of log\_Vd? ✅
  2. The order of finding the error scaling affects the results. If I try to determine LAMOST scaling first, I found and . But if I determine SDSS scaling first, I found and . Is this okay? ✅
  3. I think my implementation of Matthew’s sigcorr2.awk script was wrong. Basically we set N number of runs/bootstraps. But the total offset is only determined by the last run ✅
  4. Something is weird about the error scaling formula: in the first approach, if I want to find klamost, then the error rms for LAMOST is the denominator for the error rms ratio in the right-hand side. However, in the second approach, if I want to find klamost, the error rms for LAMOST appears in the numerator in the right-hand side 🡪 the second approach gives NaN and diverges

1. **Applying selection criteria (redshift, magnitude, and veldisp cut)**

* This step is carried out in 5\_apply\_selection.py
* Applied the following criteria with the number of galaxies remaining and excluded at each step:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | Selection limit | Ng | Nexc | Descriptions |
| 6dFGS | Full sample | 8803 |  | Christina’s slightly modified FP sample |
|  | Final sample | 8803 |  | I bypassed the selection for 6dFGS. The cz < 16120 selection actually removes 5 galaxies (max cz here is 16133) because of different way to obtain zcmb. |
| SDSS | Full sample | 29677 |  | Cullan’s full FP sample with 2MASS photometry (actually should be 31631 galaxies, the missing 2000 need to ask John to measure with GALFIT) |
|  | red\_chi\_j <= 2 | 29306 | 371 | Reduced J band chi square fit from John’s GALFIT |
|  | cz\_cmb\_group <= 16120 | 7783 | 21523 | Upper CMB frame group redshift limit |
|  | cz\_cmb\_group >= 3000 | 7762 | 21 | Lower CMB frame group redshift limit |
|  | m\_j - extinction\_j <= 13.65 | 4484 | 3278 | Extinction-corrected magnitude upper limit |
|  | s\_scaled > log10(112) | 4262 | 222 | veldisp lower limit |
|  | Final sample | 4262 |  | Final number of SDSS galaxies |
| LAMOST | Full sample | 62927 |  | Full LAMOST DR7 with 2MASS photometry |
|  | fit\_ok\_j == OK | 61922 | 1005 | J fit is labelled ‘OK’ from John’s GALFIT |
|  | red\_chi\_j <= 2 | 61693 | 229 | Reduced J band chi square fit from John’s GALFIT |
|  | cz\_cmb\_group <= 16120 | 12001 | 49692 | Upper CMB frame group redshift limit |
|  | cz\_cmb\_group >= 3000 | 11956 | 45 | Lower CMB frame group redshift limit |
|  | m\_j - extinction\_j <= 13.65 | 4281 | 7675 | Extinction-corrected magnitude upper limit |
|  | s\_scaled > log10(112) | 3399 | 882 | veldisp lower limit |
|  | Visual inspection is ETG | 2195 | 1204 | John’s visual inspection |
|  | Final sample | 2195 |  | Final number of LAMOST galaxies |

* Notes:
  1. Should I use z\_cmb\_group redshift (since this is volume limit, and I used that because that redshift is the distance estimate) ✅
  2. Should I use extinction-corrected magnitude for the limit? ✅
  3. Seems awkward that the veldisp limit is offseted by 6dFGS offset 🡪 just use log10(112) as the veldisp cut ✅
  4. Maybe I should just use ETG-selected galaxies from the beginning (the selection criteria are well-defined) 🡪 obtain 2MASS photometry and record how many remain 🡪 apply the volume, magnitude, and veldisp cut so the effect from each selection can be tracked clearly ✅
  5. May need to ask John to run GALFIT on the missing 2k galaxies…
  6. Last time I suggested to combine SDSS and LAMOST to a single FP, but can we even do that?

1. **Deriving photometric errors and selecting the final FP columns**

* This step is carried out in 6\_calculate\_phot\_errors.py.
* I followed the steps in Magoulas et al. 2012:
  1. I combine all FP samples and drop duplicated galaxies (since they have exactly the same photometry)
  2. For every galaxy, I calculate the extinction-corrected magnitude and colors: , , and .
  3. I group the galaxies into magnitude bins of width 0.2 mag.
  4. For each bin, I calculate the variance of each color:
  5. Then I calculate the error in J-band alone:
  6. Then I plot vs mean magnitude in each bin and fit it with a pairwise linear function, i.e. a function that approaches a constant at low threshold magnitude and increases linearly at larger magnitude. The result is as follows:

A graph with blue dots

Description automatically generated

* 1. I apply this formula for every galaxy, and then calculate the error in radii (assuming perfect anticorrelation) as:
  2. Finally, I pick the relevant columns for FP fitting and save them to data/foundation/fp\_sample folder.

1. **Fitting the FP and deriving the log-distance ratios**
2. **Mock galaxies simulation**
3. **Consistency checks**
4. **Smoothed velocity field map**
5. **Deriving cosmic bulk flow velocity**