This file contains a short description of all code in the ML\_variance\_decomposition repository. More elaborate comments can be found in the code itself. If upon reading the comments and this document you are still stuck don’t hesitate to send me an e-mail: [janwillemnijenhuis@gmail.com](mailto:janwillemnijenhuis@gmail.com).

**Part 1:** Folder structure and contents

The folder **ML\_variance\_decomposition** contains:

* **all\_code**: all files used in the paper.
* **Documentation:** this file.
* **LICENSE**: creative commons and no responsibility license.
* **Paper:** the paper.
* **README:** the readme file of GitHub.
* **Video:** for those within the bank it contains a video with an explanation of the file structure and the files themselves. If you don’t understand anything written in here please visit this explanation before you e-mail me, it might help.

The folder **all\_code** contains:

* The code used in the paper.
* **4\_mc1:** the plots, datafiles (.pickle) and tables (.xlsx) used in the first stage of section 4.
* **4\_mc2:** the plots, datafiles (.pickle) and tables (.xlsx) used in the second stage of section 4.
* **5\_application:** the plots, datafiles (.pickle) and tables (.xlsx) used in the application to the omh.

The latter 3 folders are explained in detail in the video.

The folder **4\_mc1** contains 6 folders with the results of the simulations for 3 facets, 250-1000 obs, for dichotomous and normal data. Each separate excel file contains one row of variance estimates. The summarizing statistics for both 2 and 3 facets for this experiment are stored in the folder **results**. The file **sim\_gtheory.sh** is the batch file used in the Lisa supercluster of the UvA.

The folder **4\_mc2** contains the folder **datafiles** in which all variance estimates are stored (.pickle extension). The folder **plots** contains all histograms used in the paper. The folder **results** contains all resulting (.xlsx) tables used in the paper of the summarizing statistics of the experiment. Here, the extension **\_total** refers to the overall results, the individual files don’t have this extension. There, the tables are not beautified.

The folder **5\_application** contains the folder **results\_OMH** which contains all folders of the 36 experiments that have been performed. Each of these folders contains 3 files, which cover the resulting estimates, variance components and generalizability coefficients. The file **omh.xlsx** contains all tables used in the paper. The other files contain the raw data for the overlap (with ‘cors’ in the filename) and variance components. These tables contain the same data as is summarized in **omh.xlsx**.

This folder also contains the folder **cors** which stores all results on the estimates of overlap. That is, plots of the heatmaps, and datafiles in (.pickle) and (.xlsx) format.

**Part 2:** Code documentation

The code documentation will be covered with descending importance. So the most important file is explained first, assisting scripts the last.

1. **experiment\_omh.py:** this is the experiment of the oversluitmodel hypotheken as performed in section 5 of the paper. You need access to the big data platform before you can use this, because it contains packages of dVb. You can change the parameters in the first section of the code (‘SETTINGS’) to perform either one of the 36 experiments or a different one.
2. **experiment\_example.**py: modified version of the former. It uses much simpler data and is used as illustration of the former. You can adjust the type of machine learning algorithm used, as well as the data. This is meant as a basis from which one can design his/her own experiment.
3. **init\_dist\_fns.py:** contains the code to run the second stage of mc experiments. The functions at the top support the code at the bottom. You can change the number of observations and facets there. It outputs a .pickle file with the results.
4. **Simulation\_g\_theory.py:** contains the code to run the first stage of mc experiments. You can alternate between the number of simulations, facets (2 or 3) and type (‘bi’ or ‘norm’). The script generates data components with fixed variance components, and then estimates these components for an increasing facet size. It outputs an excel file with the variance components.
5. **Sim\_g.py**: contains the code to run one single simulation of the above. This has been used in the lisa supercomputer, to run all these jobs in parallel for 3 facets.
6. **Calc\_res**.py: computes the mean,mae,std,rmse of the first stage mc experiments of the file above. It returns a table in excel with the summarizing statistics.
7. **Unpack\_results.py:** computes the overlap between the top 1000 customers of each iteration with the top 1000 customers of the ‘true’ model in the application. It returns an excel file with the overlap for each facet combination.
8. **Unpack\_results1.py:** computes the mean,std,var,min,max of the second stage mc experiments of the file **init\_dist\_fns.py**. It returns a table in excel with the summarizing statistics (type = ‘stats’) or histogram of resulting components (type=’plots’).
9. **Test\_rf.py:** file is used to test for the best data generating process. Not used in final analysis, however the resulting dgp is used.
10. **Plot\_graphs.py:** file is used to plot Fig. 1 of the paper.
11. **Open\_omh.py:** converts the variance components from the application to excel format, so it can be beautified in tables.
12. **Open\_cors.py:** plots heatmap of overlap calculated in **calc\_res.py**.
13. **Dist\_pds.py:** initial file used to perform experiment 2 of mc simulations, is replaced by **init\_dist\_fns.py**.