

#####

Supplementary Material 2 - Electronic Appendix to the paper
"Interaction Forests: Identifying and exploiting interpretable quantitative and
qualitative interaction effects" by Roman Hornung 1,* and Anne-Laure Boulesteix 1

1 Institute for Medical Information Processing, Biometry and Epidemiology,
University of Munich, Marchioninstr. 15, 81377 Munich, Germany

* For questions, please contact: hornung@ibe.med.uni-muenchen.de

#####

Program and Platform:

#####

- Program: R, versions 3.6.3 and 4.0.3

- Used R packages that are available from CRAN:

"diversityForest", version: 0.3.1
"fastDummies", version: 1.6.2
"ggplot2", version: 3.3.2
"ggpubr", version: 0.4.0
"iRF", version: 2.0.0
"mvtnorm", version: 1.1-1
"nnet", version: 7.3-14
"obliqueRF", version: 0.3
"OpenML", version: 1.10
"plyr", version: 1.8.6
"R.utils", version: 2.10.1
"randomForestSRC", version: 2.9.3
"ranger", version: 0.12.1
"rotationForest", version: 0.1.3
"rstatix", version: 0.6.0
"scales", version: 1.1.1

- Used R packages that are available from GitHub:

"ccf", version: 0.1.0

The above package can be installed in R via:

```
# install.packages("remotes")  
remotes::install_github("jandob/ccf")
```

(Date: 9th April 2021)

- Used platforms: Linux (x86-64) (for the conduction of the analyses)
Windows 7 64-bit (for the evaluation of the results)

General information and contents of this Electronic Appendix:

#####

- Paths of the form `"/InteractionForests/..."` are used in all R scripts, where `"/"` is the R working directory. Therefore, the folder "InteractionForests" this README file is contained in is to be put into the R working directory. Alternatively, the R working directory can be changed to an arbitrary directory that contains the folder "InteractionForests".
- The following subfolders are found in "InteractionForests":
 - "Data": This subfolder contains the subfolders "Datasets" and "ExampleDatasets" and the files `"datainfo.Rda"`, `"df_bmr.RData"`, `"DownloadData.R"`, and `"DownloadAndProcessExampleData.R"`.

The subfolder "Datasets" contains the processed data sets as used in the large-scale real data study in the form of Rda files.

The subfolder "ExampleDatasets" contains the processed example data sets in the form of Rda files. These example data sets were used in Section C "Real data based exemplary interaction forest analyses" of Supplementary Material 1.

The R script `"DownloadData.R"` stems from the electronic appendix of Hornung (2020) and was written to download and preprocess the data sets. This file also generated the Rda file `"datainfo.Rda"`, also contained in "Data", which contains a `data.frame` with meta information on the data sets such as sample sizes and numbers of covariates. Note that `"DownloadData.R"` downloads all 243 data sets used by Couronné et al. (2018), where for the analyses performed for Hornung (2020) and for the current paper that subset of the 220 data sets was used that contains less than 10000 observations and a maximum of 500 covariates.

References:

- Couronné, R., Probst, P. & Boulesteix, A.-L. (2018) Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 19, 270.
- Hornung, R. (2020) Diversity Forests: Using split sampling to allow for complex split procedures in random forest. Technical Report No. 234, Department of Statistics, University of Munich.

File RData file `"df_bmr.RData"` is used by `"DownloadData.R"` and contains a `data.frame` that provides the OpenML data set IDs necessary to download the data sets from OpenML.

The R script `"DownloadAndProcessExampleData.R"` was used to download and process the example data sets.

- "Evaluation": This subfolder contains four R scripts: `"Evaluation_LargeStudy.R"`, `"Evaluation_Simulation.R"`, `"ExampleAnalyses.R"`, and `"RuntimeAnalysis.R"`.

"Evaluation_LargeStudy.R" was used to produce all results of the large-scale real data study.

"Evaluation_Simulation.R" was used to produce all results of the simulation study.

"ExampleAnalyses.R" was used to produce the exemplary real data analyses.

"RuntimeAnalysis.R" was used to produce the runtime analysis presented in the discussion of the paper.

These R scripts also contain the R code used to produce the figures.

- "Functions": This subfolder contains the R scripts "Functions_LargeStudy.R" and "Functions_Simulation.R".

These two R scripts contain the functions used in the large-scale real data study ("Functions_LargeStudy.R") and in the simulation study ("Functions_Simulation.R").

They are called by the corresponding R scripts found in the folder "JobScripts" that perform these studies.

- "JobScripts": This subfolder contains the following R scripts: "LargeStudy.R", "LargeStudy_ObliqueForest.R", "Simulation_100.R", "Simulation_500.R", "Simulation_1000.R", and "Simulation_subtree_100_500.R".

"LargeStudy.R" and "LargeStudy_ObliqueForest.R" perform the large-scale real data study. "LargeStudy.R" produces the results for all considered methods except for Oblique Forests, which are produced by "LargeStudy_ObliqueForest.R". The reason for using a separate R script for Oblique Forests was that these were computationally more demanding than the remaining methods, which is why we needed to perform the calculations for Oblique Forests on a different cluster than those for the other methods.

"Simulation_100.R", "Simulation_500.R", and "Simulation_subtree_100_500.R" produce the results for all methods except for IMDMS for sample sizes 100, 500, and for both, 100 and 500, respectively.

"Simulation_1000.R" produces the results for all methods for sample size 1000.

"Simulation_subtree_100_500" and "Simulation_1000.R" were executed on a different cluster than "Simulation_100.R" and "Simulation_500.R", because the former two scripts were computationally more expensive than the latter two scripts.

- "Results": This folder contains the raw results of the large-scale real data study and the simulation study as well as the figures produces in the analyses. More precisely, the folder contains the subfolders "Figures" and "LargeStudy_Rda_files" and the files "Results_LargeStudy.Rda", "Results_Simulation.Rda", "scenariogrid_LargeStudy_done", and "scenariogrid_Simulation".

The subfolder "Figures" contains all figures (mostly in PDF format) produced by "Evaluation/Evaluation_LargeStudy.R", "Evaluation/Evaluation_Simulation.R", and "ExampleAnalyses.R".

All Figures shown in the main paper and in Supplementary Material 1 are among the figures in this subfolder.

The raw results of the large-scale real data study are contained in "Results_LargeStudy.Rda", where the Rda file "scenariogrid_LargeStudy_done.Rda" contains a data.frame providing information on the settings considered in the study. These files are used in the evaluation of the large-scale real data study.

The raw results of the simulation study are contained in "Results_Simulation.Rda", where the Rda file "scenariogrid_Simulation.Rda" contains a data.frame providing information on the settings considered in the study. These files are used in the evaluation of the simulation study.

The subfolder "LargeStudy_Rda_files" is empty. When reproducing the results of the large-scale real data study (see further down) this folder will contain the results of the iterations of this study in a separate Rda file for each iteration. We store these iterations separately, because some methods delivered errors for some iterations.

- "SimulationDesign": This subfolder contains only a single R script labeled "SimulationSetup.R". This script show how the simulation parameters for the simulation of the quantitative and qualitative interactions were determined when setting up the simulation study.

Evaluation of the results:

#####

- For the evaluation of the results it is not necessary to re-perform the analyses:

The R scripts "Evaluation_LargeStudy.R", "Evaluation_Simulation.R", "ExampleAnalyses.R", and "RuntimeAnalysis.R" contained in the subfolder "Evaluation" produce all results shown in the main paper and in Supplementary Material 1 without the need of re-performing the analyses. These R scripts read in Rda files (stored in the subfolder "Results") that contain the raw results.

Full reproduction of the results:

#####

- All R code needed to fully reproduce the analyses is available in this electronic appendix.
- As a first step, the folder "InteractionForests" this README is contained in has to be put into the home directory ("~/") of a Linux machine.
- An MPI environment is required.

- The R scripts in the subfolder "JobScripts" perform the large-scale real data study and the simulation study.

The R scripts "LargeStudy.R", "Simulation_100.R", and "Simulation_500.R" require the RMPISNOW shell script from the R package "snow".

Therefore, before executing these scripts you need to install the RMPISNOW shell script from the installed 'snow' R package or 'inst' directory of the package sources of the 'snow' R package in an appropriate location, preferably on your path.

See <http://homepage.divms.uiowa.edu/~luke/R/cluster/cluster.html> (last accessed: 09th April 2021) for more details.

Subsequently, you need to create sh files, each for a different of the above R scripts. The following is the content of an example sh file "LargeStudy.sh":

```
#!/bin/bash
#SBATCH -o /myoutfiledirectory/myjob.%j.%N.out
#SBATCH -D /myhomedirectory
#SBATCH -J LargeStudy
#SBATCH --get-user-env
#SBATCH --clusters=myclustername
#SBATCH --partition=mypartitionname
#SBATCH --qos=mypartitionname
#SBATCH --nodes=??
#SBATCH --tasks-per-node=??
#SBATCH --mail-type=end
#SBATCH --mail-user=my@mail.de
#SBATCH --time=?:?:??
```

```
mpirun RMPISNOW < ./InteractionForests/JobScripts/LargeStudy.R
```

The R scripts "LargeStudy_ObliqueForest.R", "Simulation_1000.R", and "Simulation_subtree_100_500.R" use a shared memory environment (with a lot of RAM).

The following is the content of an example sh file "LargeStudy_ObliqueForest.sh":

```
#!/bin/bash
#SBATCH -o /myoutfiledirectory/myjob.%j.%N.out
#SBATCH -D /myhomedirectory
#SBATCH -J LargeStudy_ObliqueForest
#SBATCH --get-user-env
#SBATCH --clusters=myclustername
#SBATCH --partition=mypartitionname
#SBATCH --mem=?????mb
#SBATCH --cpus-per-task=??
#SBATCH --mail-type=end
#SBATCH --mail-user=my@mail.de
#SBATCH --export=NONE
#SBATCH --time=?:?:??
```

```
R --no-save -f ./InteractionForests/JobScripts/LargeStudy_ObliqueForest.R
```

The above sh-files of course have to be adjusted to be useable (e.g., the "?"s have to be replaced by actual numbers, the directories have to be adjusted and

you need to specify your e-mail address; an e-mail will be sent to this address once the job is finished).