Simulation study: "Leveraging multimodal data to predict outcomes of antidepressant treatment"

```
path <- "~/Documents/GitHub/article-predictionNP1BD3/"
setwd(path)
source("./FCT/simData.R")
```

1 Introduction

We proposed a three step strategy to build and assess predictive models for antidepressant treatment response:

- **step 1**: a logistic regression with linear effects and no interest on a subset of 10 biomarkers
- **step 2**: a random forest approach on a subset of 10 biomarkers. Variable importance of each biomarker will be assessed and a more complex logistic regression (with non-linear effects and interaction) will be built using only important biomarkers.
- **step 3**: SuperLearner will be train on a large set of biomarkers.

5-fold cross validation will be used to assess the predictive performances in term of AUC, brier score, and calibration.

However, there are some difficulties:

- What is the power of the random forest test vs. logistic?
  Of variable importance to detect useful biomarkers?
  Of the complex logistic regression to identify complex patterns?
- there are several way to assess variable importance in random forest, e.g. `impurity` based on the Gini index (function `importance_pvalues` in ranger) or the method developped by [1] (implemented in the R package forestControl).
- is the 5-fold cross validation a good way to estimate predictive performances (bias, variance)[1]? Many other approaches exists so it would be nice to show that this one leads to reasonable results[2].

[1] Martin N. asked something similar at the Brain Drug annual meeting

[2] We could use repeated 10-fold CV where is fold has the same prevalence

- how to choose the hyperparameters of the machine learning approach? E.g. in random forest we need to choose:

  1. a number of trees (argument `num.trees` in ranger, by default 500)
  2. a number of features considered for splitting a node (argument `mtry` in ranger, by default square root of the number of predictors)
  3. minimal number of data points to split a node (argument `min.node.size` in ranger, by default 1),
  4. maximum tree depth (argument `max.depth` in ranger, by default unlimited),
  5. objective function (argument `splitrule` in ranger, by default `"gini"`)
  6. sampling of the observations (argument `replace` in ranger, by default `TRUE`)

  For superLearner we need to choose:

  1. the library of learner we want to consider (argument `SL.library` in SuperLearner).
  2. how do we want to combine the learners? Take the best or combine the prediction of each learner in the best way possible.
  3. plus the hyperparameters corresponding to each learner.

2 Objectives

Assess the validity of the predictive approach:

- in term of type 1 error control, i.e. conclude that there a predictive value when in fact there is none at most $\alpha\%$ of the time.
- in term of type 2 error control, i.e. conclude that we cannot identify a predictive value when in fact there is one at most $\beta\%$ of the time.

Estimate the hyperparameters of the machine learning approaches.
From the existing litterature on Random Forests, the main parameter to tune is `mtry` [2]. Other parameters such as `num.tree` can be set to arbitrary large value that is computationnally feasible [3].
The super learner library should at least contain Random Forests and elastic-net regularized logistic regression.

3 Generative model

To perform the simulation study, we first start by defining a number of scenario that should ressemble real data. As there are many parameters we can vary, we will fix:

- the sample size to $n = 80$.

- the number of predictors to $p = 10$ and the number of useful predictors 2 (unless when we are under the null where it is 0).
- the marginal distribution of each predictor to a multivariate normal distribution with mean 0 and variance 1. Can be changed to better reflect real data!

We will vary:

- the strenght of the predictors (strong, weak, or null)
- the type of interaction between the predictors considering 3 scenarios: either no interaction, or normal values vs. abnormal values, or no main effect and only an interaction.
- the correlation between the predictors. The predictors will be divided in two groups and the within group correlation will be vary. Each group contain a single useful predictor.

Visually this can is summarized in Figure 1 and Figure 2.

4 References

[1] Ender Konukoglu and Melanie Ganz. Approximate false positive rate control in selection frequency for random forest. *arXiv preprint arXiv:1410.2838*, 2014.

[2] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

[3] Philipp Probst and Anne-Laure Boulesteix. To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.*, 18(1):6673–6690, 2017.

5 Setting

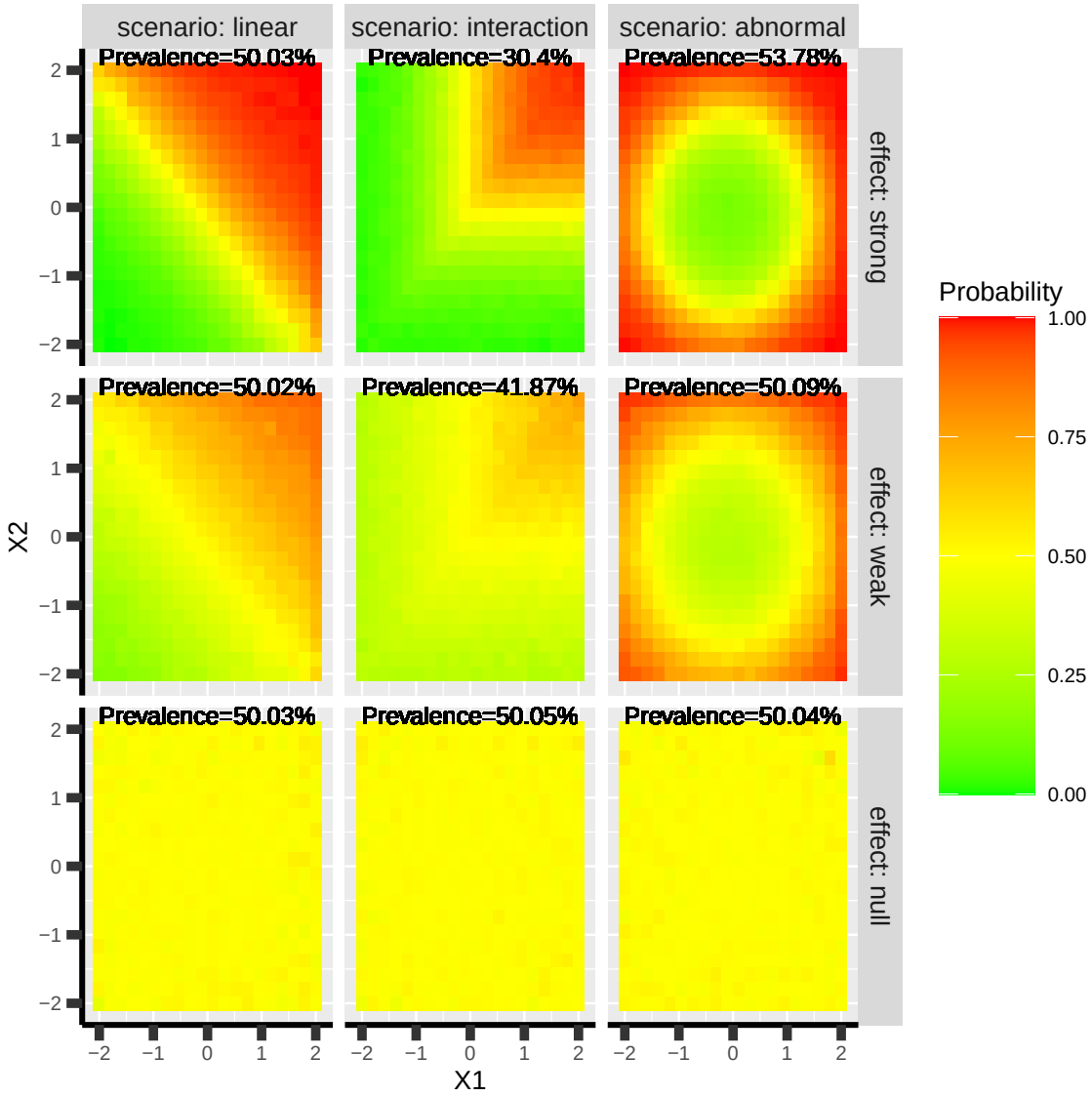

Figure 1: Probability of treatment response as a function of the useful biomarkers in each scenario.

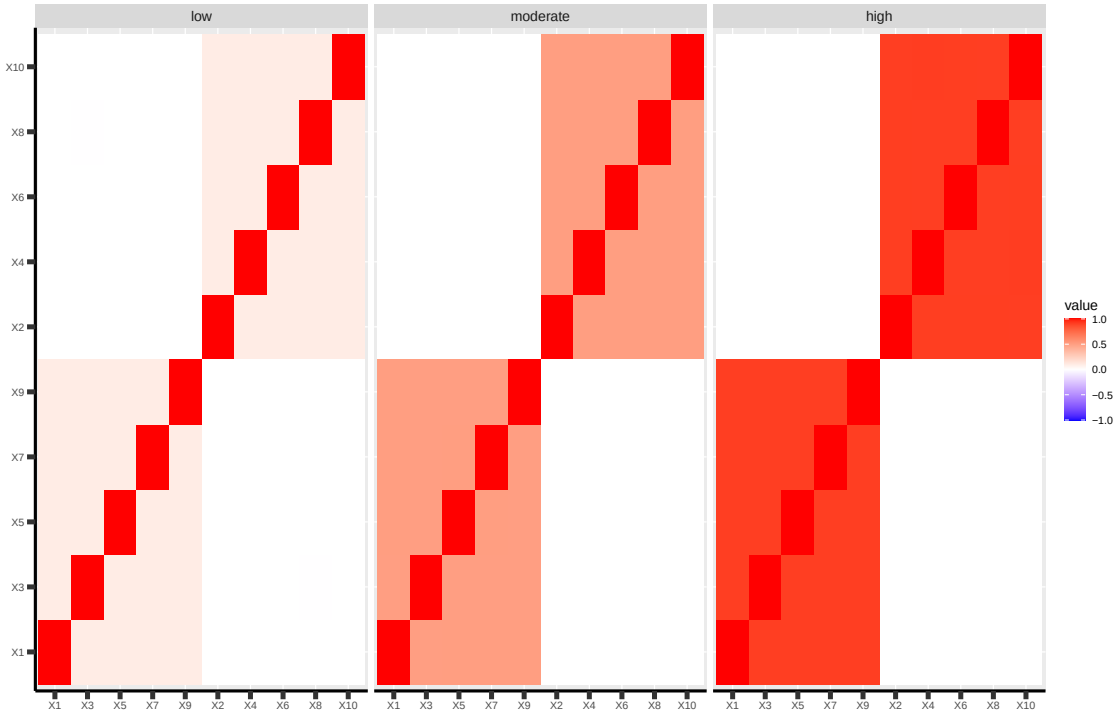

Figure 2: Correlation structure of the predictors.