# Group variable identification based on the random forests for grouped variables algorithm.

*6 month-intership proposal*

| | |
|---|---|
| **Institutions**: | Université Bretagne Sud, LMBA CNRS 6205, IRISA CNRS 6074. |
| **Lieu**: | Campus Tohannic, Vannes, France. |
| **Duration**: | 6 months. |
| **Supervisors**: | Audrey POTERIE & Charlotte PELLETIER. |
| **Contacts**: | `audrey.poterie@univ-ubs.fr`, `charlotte.pelletier@univ-ubs.fr`. |

**Keywords** : Random forests for grouped variables, group variable identification, reccursive feature selection, C++.

## Subject

Supervised learning consists in explaining and/or predicting an output variable by using some inputs. Here, we consider the context in which the inputs have a known and/or obvious group structure. In many supervised problems, inputs can have a group structure or groups of inputs can be defined to capture the underlying input associations. In these cases, the study of groups of variables can make more sense than the study of inputs taken individually.

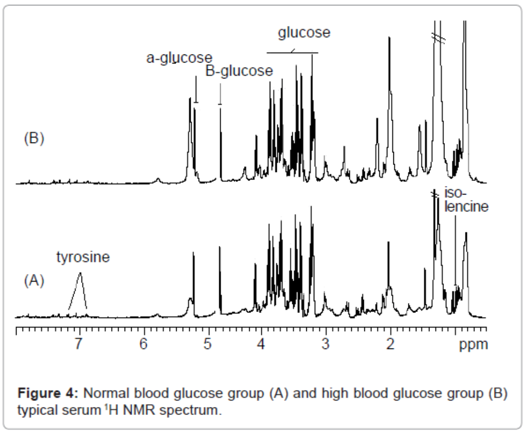

**Figure 4:** Normal blood glucose group (A) and high blood glucose group (B) typical serum $^1$H NMR spectrum.

Figure 1: Example of spectrometry data [1].

For example, in the analysis of gene expression data, datasets contain the expression levels of thousands of genes in a much smaller number of observations. Thus it has become frequent to use in the analysis only a small number of genes, which are clustered into several groups representing putative biological processes [2, 3]. Another example is functional data, like spectrometry data (see Figure 1), where researchers are often more interested by identifying discriminatory parts of the curve rather than individual wavelengths. Finally, in some supervised methods such as the CART algorithm [4], categorical inputs are converted into a group of dummy variables. Thus, it seems quite natural to treat the dummy variables related to a categorical variable as a group. In all these situations, elaborating a prediction rule based on groups of variables rather than on the individual variables can improve both interpretation and prediction accuracy [5]. Several methods have already been proposed to deal with this problem. For instance, the logistic regression regularized by the Group Lasso penalty (GL) enables to elaborate classification rules based on groups of input variables [6]. Recently, two decision tree approaches and one random forests algorithm *–called Random Forests for Grouped Variables–* have been developped to deal with groups of inputs [7, 8]. These methods build prediction rules based on groups of inputs. Furthermore in addition to the prediction purpose, these three new methods can also be used to perform group variable selection thanks to the introduction for each of them of some grouped variable importance scores.

## Objective

The internship will focus on the algorithm named Random Forests for Grouped Variables [8] (RFGV). The project will consist of three objectives.

The first objective will be to propose an efficient implementation of the RFGV algorithm (for both the training phase and the prediction phase). An implementation in R has already been proposed for binary classification problems. The first aim is twofold: (1) to improve this implementation by proposing a multi-threading version in C++/Java/Cython, and (2) to extend the implementation so as to deal with multi-class classification tasks and regression problems.

In supervised problems with grouped inputs, the group structure is often unknown. Then, the second objective will consist in developping an original and data-driven method to perform group variable identification. The proposed strategy will be based on the RFGV algorithm and its grouped importance score [7]. A wrapper strategy, such as the feature elimination algorithm proposed by [9] or the selection approach introduced by [5], could be tested. The method will be thoroughly assessed through experimental studies on several synthetic data sets.

Finally, the third objective will be to apply the proposed method on real data.

## Candidate profile

We are looking for a motivated and talented student in Master 2 (or equivalent) who:

- Has strong programming skills in at least one of the following languages: C++, Java, Cython. Skills in R and/or Python will be appreciate.
- Is keen to learn about random forests, feature selection and more generally supervised learning.
- Has good communication skills in French or in English (oral/reading/writing).

Being familliar random forests will be appreciate.

## Details

The 6-month internship will take place at the Université Bretagne Sud in Vannes. The salary will be around 500€ per month.

The student will be supervised by:

- Audrey Poterie: `audrey.poterie@univ-ubs.fr`,
- Charlotte Pelletier: `charlotte.pelletier@univ-ubs.fr`.

To apply for this position, the candidate is requested to firstly send a CV, a motivation letter and academic records to audrey.poterie@univ-ubs.fr. The application deadline is: 31th January 2021.

## References

[1] A. Y. Luaibi, A. J. Al-Ghusain, A. Rahman, M. H. Al-Sayah, and H. A. Al-Nashash, "Non-invasive blood glucose level measurement using nuclear magnetic resonance," in *2015 IEEE 8th GCC Conference & Exhibition*, pp. 1–4, IEEE, 2015.

[2] P. Tamayo, D. Scanfeld, B. L. Ebert, M. A. Gillette, C. W. Roberts, and J. P. Mesirov, "Metagene projection for cross-platform, cross-species characterization of global transcriptional states," *Proceedings of the National Academy of Sciences*, vol. 104, no. 14, pp. 5959–5964, 2007.

[3] S.-I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, no. 11, p. R76, 2003.

[4] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees.* CRC Press, 1984.

[5] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis," *Computational Statistics & Data Analysis*, vol. 90, pp. 15–35, 2015.

[6] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[7] A. Poterie, J.-F. Dupuy, V. Monbet, and L. Rouvière, "Classification tree algorithm for grouped variables," *Computational Statistics*, vol. 34, no. 4, pp. 1613–1648, 2019.

[8] A. Poterie, *Arbres de décision et forêts aléatoires pour variables groupées.* PhD thesis, INSA de Rennes, 2018.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.