Constructing bi-plots for random forest

Analytica Chimica Acta 1131 (2020) 146–155

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Tutorial

Constructing bi-plots for random forest: Tutorial

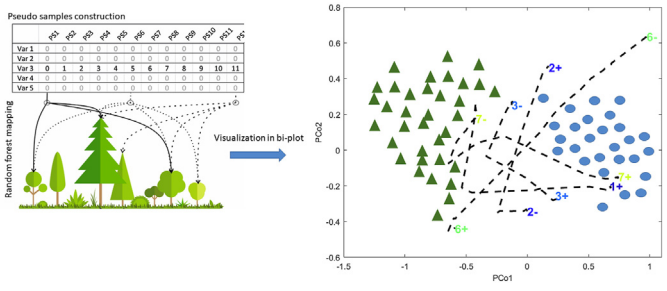
Lionel Blanchet [a, 1], Raffaele Vitale [b, c], Robert van Vorstenbosch [a], George Stavropoulos [a], John Pender [d, e], Daisy Jonkers [f], Frederik-Jan van Schooten [a], Agnieszka Smolinska [a, *]

[a] Department of Pharmacology and Toxicology, School of Nutrition, Toxicology and Translational Research in Metabolism (NUTRIM), Maastricht University Medical Center+, Maastricht, the Netherlands
[b] Laboratoire de Spectrochimie Infrarouge et Raman - LASIR CNRS - UMR 8516, Université de Lille, Bâtiment C5, F-59000, Lille, France
[c] Molecular Imaging and Photonics Unit, Department of Chemistry, Katholieke Universiteit Leuven, Celestijnenlaan 200F, B-3001, Leuven, Belgium
[d] Department of Medical Microbiology, School of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University Medical Centre+, Maastricht, the Netherlands
[e] Department of Medical Microbiology, School for Public Health and Primary Care (CAPHRI), Maastricht University Medical Centre+, Maastricht, the Netherlands
[f] Division Gastroenterology-Hepatology, Department of Internal Medicine, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, 6202 AZ, Maastricht, the Netherlands

HIGHLIGHTS

- Pseudo-samples enable visualization of the variable importance in random forest (RF).
- Interpretation of variable importance in RF and unsupervised random forest (URF).
- Possibility of obtaining so called bi-plot for RF and URF.
- Relation between variables are obtained using principal coordinates analysis.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 16 February 2020
Received in revised form
15 June 2020
Accepted 16 June 2020
Available online 11 July 2020

Keywords:
Random forest interpretation
Pseudo samples
Bi-plots
Proximity matrix
Principal coordinates analysis

ABSTRACT

Current technological developments have allowed for a significant increase and availability of data. Consequently, this has opened enormous opportunities for the machine learning and data science field, translating into the development of new algorithms in a wide range of applications in medical, biomedical, daily-life, and national security areas. Ensemble techniques are among the pillars of the machine learning field, and they can be defined as approaches in which multiple, complex, independent/uncorrelated, predictive models are subsequently combined by either averaging or voting to yield a higher model performance. Random forest (RF), a popular ensemble method, has been successfully applied in various domains due to its ability to build predictive models with high certainty and little necessity of model optimization. RF provides both a predictive model and an estimation of the variable importance. However, the estimation of the variable importance is based on thousands of trees, and therefore, it does not specify which variable is important for which sample group.

The present study demonstrates an approach based on the pseudo-sample principle that allows for construction of bi-plots (i.e. spin plots) associated with RF models. The pseudo-sample principle for RF. is explained and demonstrated by using two simulated datasets, and three different types of real data, which include political sciences, food chemistry and the human microbiome data. The pseudo-sample bi-

* Corresponding author.
E-mail address: a.smolinska@maastrichtuniversity.nl (A. Smolinska).
[1] Current address: Philips, Veenpluis 4-6, Best, the Netherlands.

https://doi.org/10.1016/j.aca.2020.06.043

plots, associated with RF and its unsupervised version, allow for a versatile visualization of multivariate models, and the variable importance and the relation among them.

Contents

1. Introduction

The influence of machine learning algorithms spreads throughout the scientific and business practices [1,2]. Credit risks are routinely evaluated by using logistic regression [3]. Faulty products are detected in industrial production by means of multivariate statistical process control [4]. Streaming services use neural networks to recommend movies related to the user's preferences. Among hundreds of approaches, the ensemble learning [5,6] methods have proved to be flexible and efficient. Ensemble learning has been particularly popular since the introduction of random forest (RF.) by L. Breiman in 2001 [7]. The underlying idea is that the combination of multiple weak and uncorrelated classifiers results in an improved ability to predict overall response. This approach has been immensely successful because of its capacity to construct a complex model with little to no parameter optimization involved. Once a decision model has been established, it can be used as a tool for prediction purposes. However, this approach is valid only if one is interested is the final choice made by the decision trees. This is because the model does not provide any information related to either biological or chemical meaning of the compounds responsible for the final choice made. Oftentimes, it is needed to understand what is driving the model towards a particular decision. For example, an investment banker (and their supervisors and regulators) may be interested in knowing why it is wise to invest millions in a specific company. RF can provide an estimation of the variable importance; although this quantitative information is often too crude to provide a clear explanation of the inner working of themodel. The model sees a variable as important, but it does not indicate changes in the variable values, i.e. if its amount increases or decreases, or whether it should be considered in interaction with another variable.

In ensemble methods, detailed information on variables is difficult to access because it is spread over the hundreds of weak classifiers that form the model. InRF, the variable importance is measured by calculating the increase in the model prediction error after permuting the values of a variable. A variable is considered as important if, after shuffling its original values, the prediction error increases. The opposite also holds; a variable is unimportant if permuting its value does not change the prediction error. Several variation-of-variable-importance approaches have been proposed, such as removing of a variable or retraining a model and in the following step, compare its error. Fisher et al. [8] suggested a model-agnostic form of variable importance, called model reliance, where the dataset is split in half, and the value of a variable is swap between the two halves. On the one hand, the classification trees that form an RF are easy to understand, similarly to simple classification and regression trees [9]. On the other hand, interpreting the decision rules present in an entire forest still remains a daunting task.

In the current study, a simple approach that allows for construction of bi-plots (i.e. spin plots) associated with RF models is demonstrated. Bi-plots are among the most popular and versatile visualization tool of multivariate models [10–13]. Bi-plots permit for the simultaneous display of information on the samples and the variables. The performance of the RF bi-plots is assessed by using two types of simulated, and various real datasets, i.e. political sciences, food chemistry and the human microbiome. These examples demonstrate the potential of this visualization tool on both categorical and continuous data, in both supervised and unsupervised analysis. The interpretation of the RF models employing bi-plots is shown for supervised and unsupervised version [14] of the tree-based techniques.

2. Material and methods

2.1. Random forest and unsupervised random forest

RF is an ensemble technique introduced by Leo Breiman, and it is based on the aggregation of a large number of uncorrelated and weak decision trees [7]. The idea behind it is to create a training set that consists of ~63% of samples (with replacement, i.e. bootstrap aggregating) from the original data; the remaining samples are used for internal validation of the RF model. Those samples are called out-of-bag (OOB) samples, and the corresponding error calculated for their prediction is called OOB error. The procedure is repeated *t* times, where t indicates the number of grown trees. A set of *t* single trees without pruning is created by using each of these bootstrap datasets. At each node, a subset of variables is randomly selected to choose the best binary split by implementing, e.g. Gini impurity index evaluated for OOB cases. Each of the *t* classification trees is next used to predict the OOB cases. The final decision is made by majority voting of all the *t* trees. Each RF model presented here is validated either internally by using OOB cases (microbiota data) or externally by using a test set (wine data and senate vote data) selected by employing the Kennard and Stone algorithm [15,16].

Similarly to RF, unsupervised random forest (URF) [14] uses a set of *t* weak decision trees. URF assumes that if the data embraces any possible trends (e.g. relevant class groupings), the data should be differentiated from a randomly generated version of itself [17]. The randomly generated data is, thus, created to perform a two-class RF classification model. The synthetic data within URF can be generated in various ways [18]. The most common and most straightforward approach used here is to permute the values of each predictor of the original data. This procedure leads to the creation of a synthetic dataset that has the same number of samples and variables as the original dataset. The meaningful classification results can be further visualized by using a proximity matrix, which can be considered as a similarity measure among samples. It is created by measuring the number of times that two samples end up in the same terminal node during the RF classification model. The final number is normalised by the number of trees used in the RF model, leading to a matrix *mxm*, where *m* is the number of samples in the data, with values varying between 0 and 1. The proximity matrix is then transformed into a dissimilarity matrix, and it can be used to perform any unsupervised analysis, such as principal component analysis, for identifying any relevant structure in the data.

2.2. Bi-plots for RF and URF

The within RF variable importance is obtained via variable importance measure (VIM), which is calculated by permuting the values of each variable in the OOB cases and then, predicting the values of these samples [7]. If a data matrix **X** with "p" number of variables is considered, VIM is obtained by first, randomly, permuting the values in the predictor variable "p", and thus, losing the association with the class vector. If the prediction accuracy for the cases in OOB decreases significantly in comparison with non-permuted variables values, it indicates a strong relation of the predictor variable "p" with the response (i.e. classes). The difference in prediction accuracy before and after permuting the values of variable "p", averaged over all trees, is a measure of variable importance; the higher the number is, the more important the predictor variable "p" is. VIM allows for selecting the most informative predictors; however, it does not provide information on the relation between predictors and the relative changes among the different groups explored in the RF classification model. Moreover, VIM values might be negatively affected by correlated variables [19].

The pseudo-sample principle [20,21] is used, which is based on the nonlinear plot idea described by Gower [22], to represent the variable importance in RF In that regard, a set of artificial samples, i.e. pseudo samples, is created to investigate the RF model. These samples are constructed to evaluate each variable independently. The graphical illustration of the pseudo sample approach is shown in Fig. 1A–B. The described procedure starts (Fig. 1A) from creating a proximity matrix (**P** within the RF model for data matrix **X** with "m" number of observation and "p" variables. The latter is changed into a dissimilarity matrix **DP**. After double-centering, the matrix can be imputed into PCA analysis, and the sample groupings can be visualized in a PCA score plot. In the further part of this tutorial, this type of plot and analysis will be referred to as principal coordinate analysis (PCoA) and its subsequent score plot. In Fig. 1B, a given pseudo sample is shown, where all the values are set to zero except for the variable being investigated. The latter is set to a particular value, e.g. the range between the minimum and maximum value observed for that variable in the real data. These values allow for describing a complete trajectory for every variable.

Consequently, for each variable, a matrix of size "l × p" is created, where "l" is the number of steps in the range of pseudo samples used to span the complete array of the original variable, and "p" is the number of original variables in data matrix **X**. Note that, overall, "p" pseudo-sample matrices are created, and it is possible to check the influence of these pseudo samples in the RF model. Those pseudo samples are then put into the RF model, where their proximity to the training samples can be estimated (pseudo-sample proximity matrix **PP** of size "lxm"). After changing the matrix **PP** into a dissimilarity matrix **DPP,** the position of the artificial samples regarding the real data can be visualized in a PCoA space (Fig. 2B). By changing the non-zero values of the artificial samples, the procedure progressively constructs a complete trajectory in the PCoA space, which allows for visualization of the importance and behavior of this variable in the RF model. An adequate metaphor of this procedure would be a chemical titration, where the titrating solution here is the variable that is progressively explored by increasing its value in the pseudo sample step by step and observing the corresponding result at each step. The exact procedure is applied to all variables. Ultimately, all variable trajectories can be displayed on top of the scatter plot obtained from the PCoA analysis of the proximity matrix, thus, leading to the RF bi-plot. The same procedure can be applied in URF.

The graphical representation of the projection of the pseudo samples proximities onto the PCoA space is shown in Fig. 2B. In this simple example, four different variables are shown. Note that in the case of linear techniques, such a plot is called a loading plot. Each of the demonstrated variables in Fig. 2B has different trajectories and influence in the model (taking into account principal coordinate 1). The trajectories of the variables one and three form a straight line; thus, it is expected that their influence in the RF model is also linear. Although both variables demonstrate a linear behavior, their importance is completely different. Variable one exhibits long trajectories, while variable three has a short trajectory fluctuating around zero. This indicates that variable one has relatively higher importance in the model than variable three. The trajectories of the variables two and four exhibit a nonlinear behavior over the entire range, and similarly to variables one and three, their importance in the model varies. Variable two has high importance in all of its range, while variable three reveals higher importance in the higher variable range. It is possible to calculate the overall importance of each variable by taking the absolute value of the difference between the maximum and minimum value of each pseudo sample. The results can be then, graphically, shown as any traditional bar plot used, for instance, in partial least square regression plot [23].

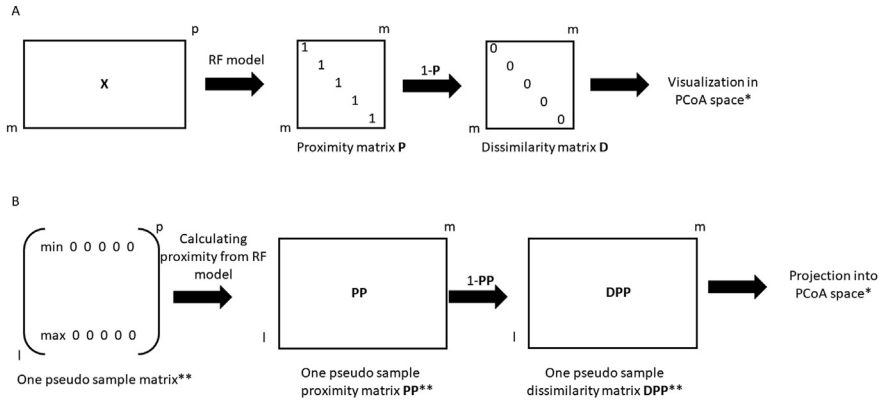


Fig. 1. Representations of the A) random forest proximity matrix of data matrix X; b) pseudo samples principle in the random forest model. The uniformly distributed range of pseudo sample values is indicated as "I"; *The samples of data matrix X are svisualised in PCoA score plot, and the pseudo samples are projected into the same PCoA space using the loading vector. **Note that there are "p" pseudo sample matrixes, "p" pseudo samples proximity matrixes and "p" pseudo sample dissimilarity.

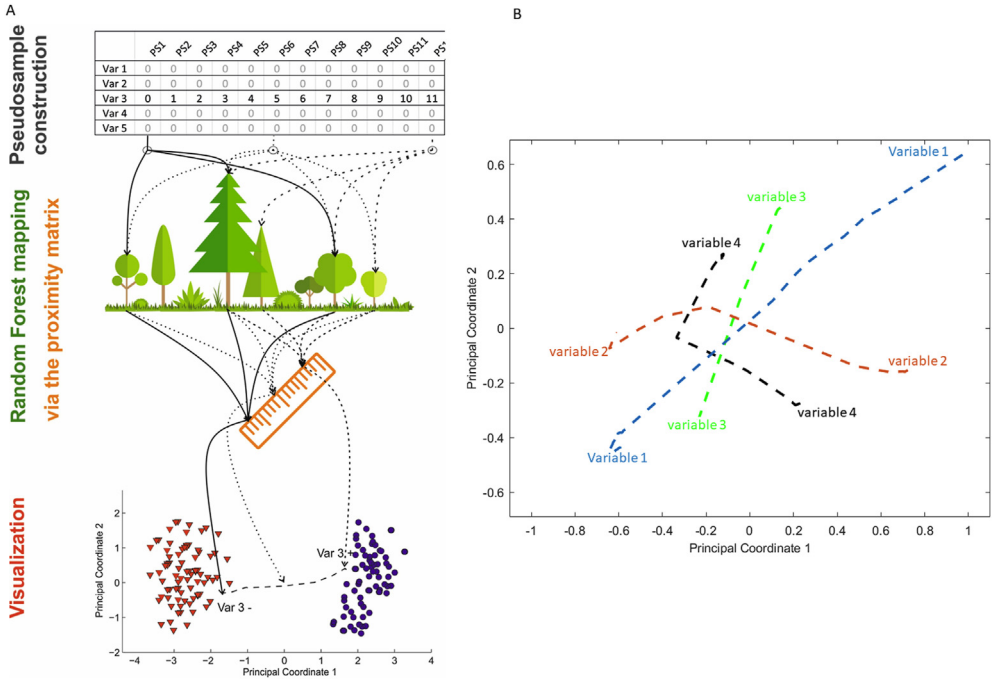


Fig. 2. A) A graphical representation of the pseudo samples approach. B) An example of four pseudo samples proximities for four variables projected into PCoA space.

The procedure used to obtain bi-plots for RF or URF is available in the following repository https://github.com/LioB-FRNL/RF_URF_biplots. The procedure is very fast, considering the fact that building an RF model of 1000 trees takes approximately a few minutes on a regular laptop.

3. Datasets

3.1. The synthetic benchmark datasets 1 and 2

Two simulated datasets were created to represent the pseudo-sample approach. The first simulated dataset consists of two classes, which are put in the shape of a chessboard. The second simulated dataset contains two classes with two circles: the inner circle refers to class one, and the outer circle refers to class two. Both cases are visualized in Fig. 3A–B, respectively, and they consist of 400 samples, 200 in each class, and 224 variables. To make the simulated data resembling the typical –omics data, the simulations are based on the covariance structured of the blood plasma metabolomics data [24]. The first dataset is simulated by creating two circles as sinus and cosine function and adding an array of random numbers from a normal distribution with specific sigma and mean values. A different level of noise was added as normally distributed pseudo-random numbers. In the final step, the subgroupings were created in each circle by taking the absolute values of the, randomly, selected 20% of all variables. This was done for 70% of the samples, randomly selected. The similar procedure was followed for the second dataset (Fig. 3B); however, instead of using sinus and cosine functions as the basis, the simulations created a set of random numbers from a normal distribution with different centroids to obtain the chessboard shape.

Both benchmark datasets demonstrated nonlinear relations among variables; thus, the linear techniques were not able to find the proper separation between the two classes. The most important set of variables was selected by taking 80% of the maximum value of the absolute value of the difference between the maximum and minimum value of each pseudo sample.

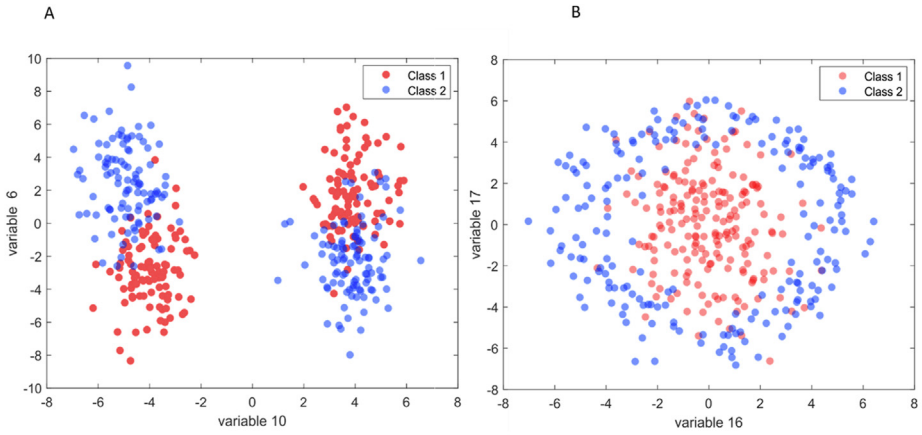


Fig. 3. The scatter score plot of synthetic data: A) dataset 1 representing two classes arranged in chessboard; B) dataset 2 consisting of two inner circles. Each dataset contains 400 samples and 224 variables.

3.2. Real datasets

3.2.1. Wine data

Three different real datasets were used to demonstrate the flexibility of the pseudo samples. The first data was a wine dataset, a benchmark dataset freely available from the UC machine learning repository [25,26]. The data consisted of 178 wine samples obtained from the same region of Italy (Piedmont), and using three grape varieties (59 Barolo, 71 Grignolino, 48 Barbera). A total of 13 constituents were analyzed in each sample: alcohol, malic acid, ash, the alkalinity of ash, magnesium, total phenols, flavonoids, non-flavonoid phenols, proanthocyanins, color intensity, hue, proline, and the ratio of the optical density at 280 over 315 nm of diluted wines.

3.2.2. US senate votes

The US Senate vote dataset compiled the votes of each US senators over 2015. All information was obtained by using the application programming interface (API) of GovTrack.us, which makes legislative data freely available. The data consisted of 100 samples, each being one senator, and 339 votes held in 2015. The votes were encoded either as 1 for yes, −1 for no and zero for abstention or missing vote.

3.2.3. Microbiota data

The last dataset used here is microbiota data obtained by amplicon sequencing using 454 pyrosequencing, as described before [27]. Shortly, metagenomic DNA obtained from fecal samples of 20 individuals with Crohn's disease (CD) and Ulcerative Colitis (UC) was used to sequence the V1−V3 hypervariable region of the 16S rRNA gene. Each individual delivered two fecal samples, one at disease remission state and a second one after subsequent development of exacerbation during a one-year follow-up period. The sequences were clustered into operational taxonomic units (OTUs) or phylotypes based on 97% sequence similarity (i.e. species level) against the Greengenes reference set by using the UCLUST algorithm leading to 2869 OTUs [28]. Inverse hyperbolic sine transformation and centering per individual were applied to the microbiota data before the actual analysis took place. After removing OTUs that were present in less than 20% of the samples, the final datasets consisted of 40 samples and 648 OTUs.

4. Results

4.1. The synthetic benchmark datasets 1 and 2

The simulated data were first used to show the feasibility of the RF bi-plot. The URF analysis was first performed on both datasets, and the final set of the most discriminatory variables were selected. In the case of the synthetic benchmark dataset one, the final number of the top discriminatory variables was 15, while for the synthetic benchmark data set two, it was 17. The top discriminatory variables were selected by taking the top 80% of the trajectories exhibiting the longest trajectories in the PCo1 and PCo2, i.e. directions which show the separation between the classes. Note that representing the trajectories for the whole set of variables can obscure the interpretation due to overcrowding of the plot. The corresponding bi-plots obtained from PCoA analysis performed on the proximity matrix obtained from URF are shown in Fig. 4A−B. Despite the nonlinear relation among variables, a separation between the two groups, class1 and class2, was achieved for both simulated datasets. However, the separation of the classes for chessboard dataset was obtained by taking the combination of the first two sPCo's, in case of the simulated dataset two, with two inner circles; PCo1 only suffices to get the clear distinction between class1 and class2.

The results obtained for the benchmark data set 1 demonstrate two subgroups of variables (Fig. 4A). The first group consisted of variables 4, 5, 6, 8, 10 and 11 exhibited longer trajectories than the remaining variables. This suggested higher importance of those variables in separating the two classes. The trajectories followed from the negative (i.e. lower) values for class1 and further to positive (i.e. higher) values for class2. It was also clearly seen that the trajectories of these two groups of variables followed very closely, indicating that a correlation might exist among the original variables.

The trajectories obtained for the benchmark dataset 2 (Fig. 4B) indicate that the majority of variables found as being important make a complete turn between the two groups, i.e. the variables have the positive and negative sides of the trajectories in the cloud of the points that belong to class1. Only a few variables, i.e. 2,4,6,14, and 17, changed the sign of their trajectory between class1 and class2.

4.2. Real datasets

4.2.1. Wine dataset

The wine dataset is a benchmark dataset, which allowed for validation of the RF bi-plots. In that regard, the RF bi-plot was compared with linear discriminant analysis bi-plot (Fig. 5A−B, respectively). Both models successfully discriminated the three classes of wines with a correct overall prediction for the test set of 100% for each class. As seen in Fig. 5B, the RF bi-plot displays nonlinear trajectories, which is an expected feature of an RF model. Each trajectory was labelled twice, once at each extremity to

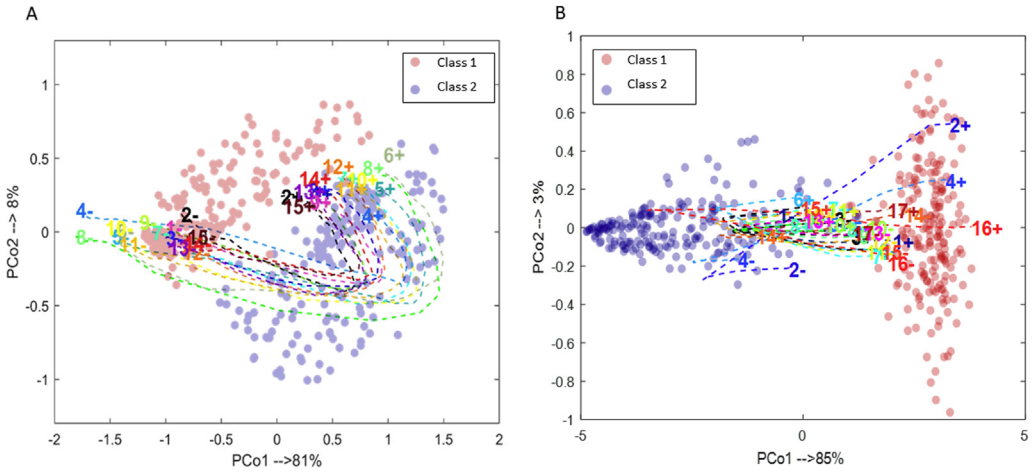


Fig. 4. PCoA bi-plot acquired from proximity obtained from the URF model using the most important variables for A) synthetic benchmark dataset one (chessboard shape); B) synthetic benchmark dataset two (two circles). Each variable is represented as one trajectory with positive (+) and negative (−) side.

picture the minimal and maximal values. Both approaches permited to conclude that the concentration of flavonoids was lower in the Grignolino group, or that alcohol level was higher in Barbera groups. As seen in Fig. 5B, the length of the trajectories of the wine data reflects the importance of each variable in the model. Similarly to the linear discriminant analysis (LDA) results, Barbera wines were characterized by higher amounts of proline and flavonoids. This can be observed since the sign of the trajectories points with its positive site to the direction of this class. The opposite trend was observed for Grignolino wine, which was characterized by lower concentrations of flavonoids, whose trajectory points with their negative values to that class of wine.

4.2.2. US senate votes

The variables in the previous datasets were all continuous. One attractive feature of RF is its ability to deal with either continuous or categorical data. The US Senate votes dataset allows one to demonstrate the use of RF bi-plots on categorical data. Moreover, the RF model was built here in an unsupervised fashion to show the potential of this approach as an exploratory data analysis method. In Fig. 6A, the URF score plot is shown with each senator denoted as a sample. A strong clustering was observed among the first principal coordinate which unsurprisingly corresponded to the Republican and Democrats parties. The moderates of both sides were relatively closer to each other at the bottom of the plot. In contrast, senators supported by the tea party (represented by Ted Cruz) clustered at the upper right of the figure, and the most liberals clustered on the top left (represented by Kirsten Gillibrand). The URF bi-plots providedus with additional information. We were able here to evaluate which votes were most significant for this clustering. In most cases, the clustering seemed to be more defined by the opposition to some votes rather than the support for others, as shown by the fact that the negative side of the trajectories was more clearly pointing towards each group. For example, voting against the vote 105 (*Senate Amendment. 817:To establish a deficit-neutral reserve* fund *to provide tax benefits to patriot employers that invest in American jobs and provide fair pay and benefits to workers and to eliminate tax benefits for corporations that ship jobs or profits overseas*) was an important feature for the Republicans.

The analysis of the US Senate votes can go further and look at additional information captured by the URF model. Fig. 6B–C presents the score plot and bi-plot, respectively, obtained on the plane formed by the third and fourth dimension of the PCoA of the URF proximity matrix. A different pattern is visible here; the two parties mostly overlap. The Republican Party seems to stretch along the third dimension, whereas the democrats are spread along the fourth one. The URF bi-plot shown next to the score plot associates the latter spread with, for instance, the votes 189 and 276, which concern, respectively, the need for approval by the congress of the addition of new countries to the Trans Pacific Partnership and a motion on budgetary discipline.

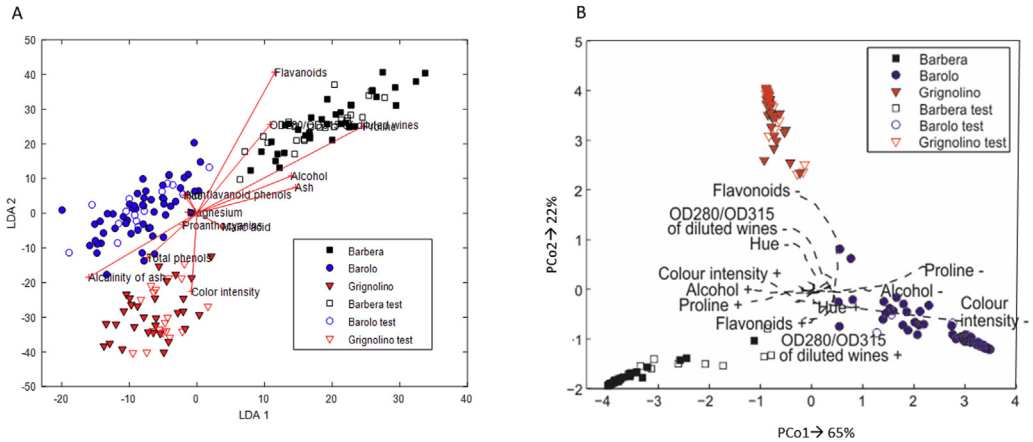


Fig. 5. A) Linear discriminant analysis bi-plot of wine data with clearly projected test samples. Each variable is indicated as a line originating from the middle of the plot B) PCoA bi-plot acquired from proximity obtained from RF model with projected test samples. Each variable is represented as a trajectory.

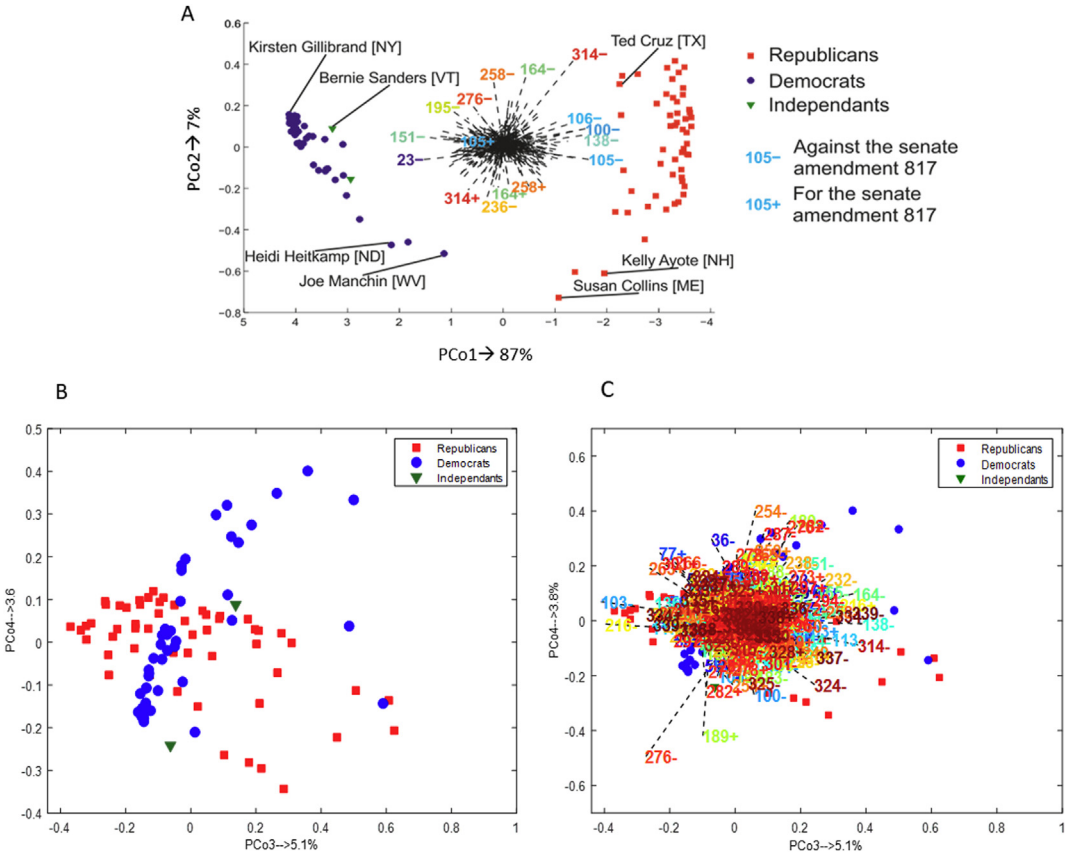


Fig. 6. A) PCoA bi-plot using the first two PCo's acquired from the proximity obtained from the URF model. B) PCoA score and; C) bi-plot obtained on the plane defined by the third and fourth PCo. Each variable is represented as a trajectory.

4.2.3. *Microbiota data*

The last example shown here is microbiota data, which consisted of 40 samples and 648 variables (OTUs). Fecal microbiota is used as a non-invasive biomarker particularly relevant for a number of diseases affecting the gastrointestinal tract [27,29]. The RF analysis led to the selection of the 12 most discriminatory variables with a correct overall prediction rate obtained from OOB cases of 88%. The names of the selected fecal bacterial taxa are reported in Table 1, while the corresponding RF bi-plot is shown in Fig. 7A. The individuals assigned as UC and CD created distinct groups, well separated along the first principal coordinate, explaining 72% of the variance.

As shown in Fig. 7A, the two groups demonstrate diverse variations. The CD group showed a more extensive spread than the UC group. Each group contained samples taken from the patient in both the active and inactive stage of the disease. The disease stages did not show any clustering withing the main classes, i.e. CD and UC (data not shown). The trajectories shown in the bi-plot exhibited various behaviours. Variable 6 shows a nonlinear trajectory, moving from the CD group (with a positive site) to the UC group (with the negative site). Moreover, this variable had the longest trajectory, indicating the relevance of this variable in discriminating CD and UC. It is relevant to notice that the large variation in the CD group seemed to be also expressed in the behavior of the trajectories. Variable 12 exhibits behavior seen in the simulation dataset 2, where its values make the complete turn (positive and negative sites end in the cloud of the CD group). This indicated that variable 12 is more informative about differences in variance between the two groups than the differences in the mean amount of that variable. PCoA bi-plot revealed that variables, 1,3,7, and 8 that are scattered in the middle of the plot, suggesting little importance of those variables (short trajectories). However, removing those variables led to a decreased prediction accuracy of the RF model. These two statements might seem contradictory; although upon further inspection, one can notice that these variable trajectories are very close to each other, and thus, suggesting a relation among them. The presence of this correlation implied that their overall importance is divided among them. All three variables belong to the same order-level Clostridiales; however, genus and species were not defined. For comparison purposes, the VIM obtained from the RF analysis for the most discriminatory fecal bacterial taxa (as indicated in Table 1) is shown in Fig. 7B. The VIM of fecal bacterial taxa was comparable to the importance demonstrated in the PCoA bi-plot (Fig. 7A). As can be seen, the most important bacteria taxa is variable number 6, while the variables 1,2,3, 7, and 8 exhibit the smallest importance in the classification model.

Table 1
Taxonomy of 12 fecal bacterial taxa that have the highest contribution to differentiating between UC and CD individuals.

Variable Numbe	Taxa names
1	Ruminococcaceae
2	Lachnospiraceae
3	Clostridiales
4	Roseburia
5	Holdemania
6	*Roseburia/Eubacteriumrectale*
7	Clostridiales
8	Clostridium
9	Roseburia
10	Roseburia
11	Ruminococcaceae
12	Roseburia

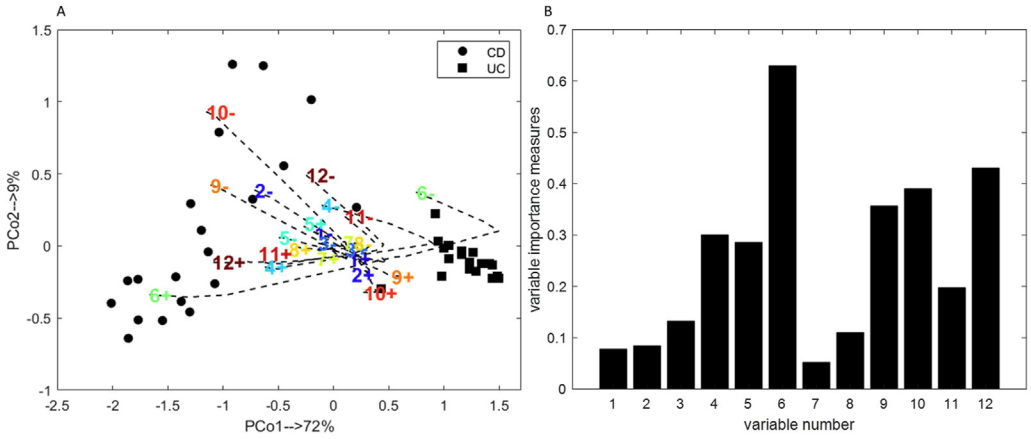


Fig. 7. A) PCoA bi-plot of microbiota data obtained from the proximity obtained from the RF model using the most discriminatory variables. Each variable is represented as a trajectory. B) The variable importance measures derived from the RF model using the most discriminatory variables of microbiota data.

5. Discussion and conclusions

The rapid development in many areas of science and technology has caused an enormous growth of data. Subsequently, this has become an essential opportunity in the machine learning field, where data is the core of it. RF, as an ensemble technique, has become frequently a method of choice for Kaggle competitions, and often, it has exceeded the popularity of another ensemble method, such as AdaBoost. Despite its popularity, the main disadvantage of RF, and also URF, is the lack of visualization of the variable importance. Information on the importance of the different variables is spread over several trees that are used in RF and URF. Both techniques allow for obtaining overall variable importance; however, quantitative information is missing. The possibility to obtain a so-called bi-plot, a typical representation of variable importance in various techniques including PCA or PLS, is lacking. The current tutorial demonstrates a novel approach to represent variables importance by using the pseudo-sample approach and the nonlinear bi-plot theory [11,20]. The utility of pseudo samples has been successfully applied to unravel the variable importance in various chemometrics techniques such as support vector machine [20,30], kernel-PLS (kPLS) [21,31], and dissimilarity-PLS [32]. Nonetheless, their application in ensemble techniques has not been studied thus far.

Here, pseudo-sample approach was used to visualize the variables importance and the relation among them for both RF and URF models. The results presented here demonstrate that this methodology can be applied to various data types, including continuous, count, and categorical. The main advantage of the proposed approach is its independency on the scaling of the data, and the presence of outliers or subgroups within the groups of interest. Moreover, the ability of RF and URF to model both linear and nonlinear relations is expressed in the shape of the individual trajectories. The pseudo samples shown here are particularly relevant for URF. As indicated already, the variable significance obtained in the standard URF represents the importance of differentiating the real data from a randomly generated version of itself [14]. Therefore, the utility of pseudo samples allows for defining which variables drive the differences in the PCoA score plot for the groups present in the dataset. In the datasets shown here, this was particularly visible for the simulated dataset one and dataset two, where several variables displayed linear trajectories; however, many showed a curved shape, suggesting a complex behaviour of that variables in the datasets. Interestingly, behavior was discovered for variables in the simulation dataset two, where several variables exhibited the complete turn of the trajectory. Those variables had positive and negative values for the same class. This behavior might be related to the structure of the data, i.e. the concentric spherical arrangement of the two classes. This suggests that many variables span a similar range for class 1 and class 2. This could be already seen in Fig. 3B, where the original values of variable 16 and 17 are shown. Asit is illustrated, both variables demonstrate positive and negative values for both classes. This behavior is remarkable since, in various metabolomics studies, it is usually seen that a compound (i.e. a variable) is relevant for group separation due to changes in their amount. Here, however, it is observed that those variables differ between classes not due to their relative difference in amount but due to changes in variance between the two classes. As indicated earlier, the variables can change their values between the classes. The same behavior was observed in the microbiota data, i.e. variable 12, representing Roseburia fecal bacterial taxa. A similar observation was found in the study by Vitale et al. [33,34], where pseudo samples were combined with k-PLS for regression analysis from mixture designs of experiments.

The results shown for the wine data using RF in combination with pseudo samples are in line with the outcome of LDA, i.e. the important parameters for each wine type and their changes in the amount. Moreover, if one looks at the relation among some wine parameters, for instance, the amount of flavonoids and the amount of ratio OD280/OD315 of diluted wine, they see that a high correlation is present in the LDA bi-plot. The same relation can be seen in PCoA bi-plot; however, in comparison to LDA, a more complex relation can be observed. The above-mentioned wine parameters exhibit a more substantial relation in the higher range of the parameters (i.e. on the positive sign of the trajectory) than in the lower range (i.e. on the negative sign of the trajectory). The useful element of pseudo samples is the possibility to investigate the behavior of each variable over its entire range. Therefore, it would be possible to define a cut-off for each variable to describe any groups or subgroups of samples seen in the bi-plots. Additionally, one could extend the values of the pseudo samples beyond the original values. The standard approach for creating pseudo samples is based on utilization of the original variables range. Nevertheless, it is possible to evaluate the behavior of the pseudo samples by using values that do not necessarily cover the values of the original variables, which is equivalent to extrapolation in the loading space for linear bi-plots. This property of the pseudo samples could be further examined, however, it is beyond the scope of this tutorial.

The results obtained from the US senate votes indicate that all variables used in the analysis reveal a linear trend. This is not surprising since the Senate votes are represented in values of 0 or 1. As seen in Fig. 6A, pseudo samples permitted for the identification

of variables that differentiate the two parties (variable 105 representing amendment 817). On the contrary, Fig. 6C enabled indentifing votes that were similar for both parties (variables 189 and 276 which correspond to the need for approval by the congress of the addition of new countries to the Trans Pacific Partnership and motion on budgetary discipline, respectively).

In the case of the microbiota data, the RF analysis combined with pseudo samples led to a set of 12 fecal taxa that differentiated UC and CD individuals. The importance of the fecal taxa that discriminated UC and CD shown in PCoA bi-plot is in line with VIM obtained from the RF classification model. The identified fecal bacterial taxa were, previously, identified as related to UC and CD [35–38]. Three relevant variables have been identified as *Roseburia,* one coming from the butyrate-producing bacteria, which has been previously reported as a primary candidate for microbial therapeutics in inflammatory bowel disease [35,38]. Interestingly, two of them exhibited very similar trajectories, indicating a positive correlation, and the possibility of belonging to the same bacteria species. Worth noting, variable 6, identified as *Roseburia Eubacteriumrectale*, has been found as the most important variable to distinguish UC and CD, and it displayed a complex trajectory when compared to variables 9 and 10. This might suggest that variables 9 and 10 belong to different bacteria species than *Eubacteriumrectale.*

The approach proposed here opens many possibilities for future investigations. The nonlinear bi-plots have various advantages, but due to the complexity of the procedure, they also bring multiple shortcomings. The most obvious one is the investigation of the correlation among variables. It has been indicated here that for variables whose trajectories follow a similar trend, a relation might be assumed. The direct investigation of the correlation among variables in nonlinear space is multifactorial since the variables might exhibit a relation only in a certain range and show very different behavior in another [34]. Furthermore, the approach demonstrated here consists of creating each variable on its own, while the other variables values are kept zeros. The choice of the “zero values” is intuitive since, for mean-centered data, one variable is varied while the rest of the variables are set to the mean value. As shown previously, this is particularly relevant for PLS and PCA models [20,21,31]. Nevertheless, it is worthwhile mentioning that due to the nature of RF any linear scaling of the data does not affect the model, thus, mean-centering of data is not necessary. Selecting another value is possible, and it would lead to the same trajectories, though, shifted to another position in the PCoA space. The inability of directly investigating the correlation among variables is a definite disadvantage when compared to standard linear bi-plots. If one wants to examine the direct relation among variables, it is necessary to vary two, or more, variables in pseudo samples simultaneously. This, however, is very challenging, mainly because the relation can change if one selects two variables or more to investigate.

The current study has demonstrated the possibility to visualize the variables importance and their changes in RF and URF models by using the pseudo-sample approach. This is the first study that shows the possibility of unrevealing the relation among important variables in RF and URF, applicable to various data types. The approach presented here is relevant for a possible definition of variable cut–off in biomarker discovery field.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

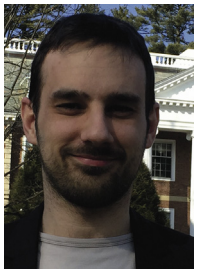
Acknowledgement

This work was supported by Netherlands Organisation for Scientific Research (NWO, the Netherlands) (grant number: 016.Veni.178.064).

References

[1] D. Boughaci, A.A.K. Alkhawaldeh, A new variable selection method applied to credit scoring, Algorithmic Finance 7 (1–2) (2018) 43–52.
[2] W.B. Xiao, Q. Zhao, Q. Fei, A comparative study of data mining methods in consumer loans credit scoring management, J. Syst. Sci. Syst. Eng. 15 (4) (2006) 419–435.
[3] M. Berlin, L.J. Mester, Retail credit risk management and measurement: an introduction to the special issue, J. Bank. Finance 28 (4) (2004) 721–725.
[4] M. Kharbach, Y. Cherrah, Y. Vander Heyden, A. Bouklouze, Multivariate statistical process control in product quality review assessment - a case study, Ann. Pharm. Fr. 75 (6) (2017) 446–454.
[5] D. Che, Q. Liu, K. Rasheed, X. Tao, Decision tree and ensemble learning algorithms with their applications in bioinformatics, Adv. Exp. Med. Biol. 696 (2011) 191–199.
[6] Z. Yin, H. Ai, L. Zhang, G. Ren, Y. Wang, Q. Zhao, H. Liu, Predicting the cytotoxicity of chemicals using ensemble learning methods and molecular fingerprints, J. Appl. Toxicol. 39 (10) (2019).
[7] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[8] A. Fisher, R. Cynthia, F. Dominici, All Models Are Wrong but Many Are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, Using Model Class Reliance, 2018 arXiv arXiv:1801.01489 [stat.ME].
[9] L. Breiman, J. Friedman, R. Olshen, C. Strone, Classification and Regression Trees, Chapman & Hall/CRC, Belmont, New York, 1984.
[10] J. Blasius, P.H.C. Eilers, J. Gower, Better biplots, Comput. Stat. Data Anal. 53 (8) (2009) 3145–3158.
[11] J.C. Gower, Generalized biplots, Biometrika 79 (3) (1992) 475–493.
[12] J.C. Gower, 3-Dimensional biplots, Biometrika 77 (4) (1990) 773–785.
[13] J.C. Gower, S.A. Harding, Nonlinear biplots, Biometrika 75 (3) (1988) 445–455.
[14] N.L. Afanador, A. Smolinska, T.N. Tran, L. Blanchet, Unsupervised random forest: a tutorial with case studies, J. Chemometr. 30 (5) (2016) 232–241.
[15] M. Daszykowski, B. Walczak, D.L. Massart, Representative subset selection, Anal. Chim. Acta 468 (2002) 91–103.
[16] R.W. Kennard, L.A. Stone, Uniform subset selection - Kennard and Stone algorithm, Comput. Aided Des. Exp. Technometr. 11 (1969) 137–148.
[17] T. Shi, S. Horvath, Unsupervised learning with random forest predictors, J. Comput. Graph Stat. 15 (1) (2006) 118–138.
[18] S. Nembrini, On the behaviour of permutation-based variable importance measures in random forest clustering, J. Chemometr. 33 (8) (2019).
[19] K.K. Nicodemus, J.D. Malley, C. Strobl, A. Ziegler, The behaviour of random forest permutation-based variable importance measures under predictor correlation, BMC Bioinf. 11 (2010) 110.
[20] P.W. Krooshof, B. Ustun, G.J. Postma, L.M. Buydens, Visualization and recovery of the (bio)chemical interesting variables in data analysis with support vector machine classification, Anal. Chem. 82 (16) (2010) 7000–7007.
[21] A. Smolinska, L. Blanchet, L. Coulier, K.A. Ampt, T. Luider, R.Q. Hintzen, S.S. Wijmenga, L.M. Buydens, Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis, PloS One 7 (6) (2012), e38163.
[22] J. Gower, S. Harding, Nonlinear biplots, Biometrika 78 (1988) 445–455.
[23] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.J. van der Werff-van der Vat, R.H. Jellema, Fusion of mass spectrometry-based metabolomics data, Anal. Chem. 77 (20) (2005) 6729–6736.
[24] A. Smolinska, J.M. Posma, L. Blanchet, K.A. Ampt, A. Attali, T. Tuinstra, T. Luider, M. Doskocz, P.J. Michiels, F.C. Girard, L.M. Buydens, S.S. Wijmenga, Simultaneous analysis of plasma and CSF by NMR and hierarchical models fusion, Anal. Bioanal. Chem. 403 (4) (2012) 947–959.
[25] M. Forina, C. Armanino, M. Castino, M. Ubigli, Multivariate data analysis as a disriminating method of the origin of wines, Vitis -Geilweilerhof 25 (1986) 189–201.
[26] Wine Data Set, UCI, 1991. https://nam03.safelinks.protection.outlook.com/?url=https%3A%2F%2Farchive.ics.uci.edu%2Fml%2Fdatasets%2Fwine&data=02%7C01%7CA.Achuthan%40elsevier.com%7Ce01c257a12634b5f1d7008d821fc0e14%7C9274ee3f94254109a27f9fb15c10675d%7C0%7C0%7C637296713740644300&sdata=hAVZ78y1rLEATgs4RqfPtmFF%2Blaq05JmswHCBnRG6HI%3D&reserved=0.
[27] E.S. Wills, D.M. Jonkers, P.H. Savelkoul, A.A. Masclee, M.J. Pierik, J. Penders, Fecal microbial composition of ulcerative colitis and Crohn's disease patients in remission and subsequent exacerbation, PloS One 9 (3) (2014), e90981.
[28] D. McDonald, M.N. Price, J. Goodrich, E.P. Nawrocki, T.Z. DeSantis, A. Probst, G.L. Andersen, R. Knight, P. Hugenholtz, An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, ISME J. 6 (3) (2012) 610–618.
[29] D.I. Tedjo, A. Smolinska, P.H. Savelkoul, A.A. Masclee, F.J. van Schooten, M.J. Pierik, J. Penders, D.M. Jonkers, The fecal microbiota as a biomarker for

disease activity in Crohn's disease, Sci. Rep. 6 (2016), 35216.
[30] G.J. Postma, P.W. Krooshof, L.M. Buydens, Opening the kernel of kernel partial least squares and support vector machines, Anal. Chim. Acta 705 (1–2) (2011) 123–134.
[31] R. Vitale, D. Palací-Lópeza, H.H.M. Kerkenaard, G.J. Postma, L.M.C. Buydens, A. Ferrera, Kernel-Partial Least Squares regression coupled to pseudo-sample trajectories for the analysis of mixture designs of experiments, Chemometr. Intell. Lab. Syst. 175 (2018) 37–46.
[32] J. Engel, G.J. Postma, I. van Peufflik, L. Blanchet, L.M.C. Buydens, Pseudo-sample trajectories for variable interaction detection in dissimilarity partial least squares, Chemometr. Intell. Lab. Syst. 146 (2015) 89–101.
[33] R. Vitale, O.E. De Noord, A. Ferrer, A kernel-based approach for fault diagnosis in batch processes, J. Chemometr. 28 (8) (2014) S697–S707.
[34] R. Vitale, O.E. de Noord, A. Ferrer, Pseudo-sample based contribution plots: innovative tools for fault diagnosis in kernel-based batch process monitoring, Chemometr. Intell. Lab. Syst. 149 (2015) 40–52.
[35] S. Paramsothy, S. Nielsen, M.A. Kamm, N.P. Deshpande, J.J. Faith, J.C. Clemente, R. Paramsothy, A.J. Walsh, J. van den Bogaerde, D. Samuel, R.W.L. Leong, S. Connor, W. Ng, E. Lin, T.J. Borody, M.R. Wilkins, J.F. Colombel, H.M. Mitchell, N.O. Kaakoush, Specific bacteria and metabolites associated with response to fecal microbiota transplantation in patients with ulcerative colitis, Gastroenterology 156 (5) (2019) 1440–1454 e2.
[36] H. Nagao-Kitamoto, N. Kamada, Host-microbial cross-talk in inflammatory bowel disease, Immune Netw. 17 (1) (2017) 1–12.
[37] H.C. Mirsepasi-Lauridsen, K. Vrankx, J. Engberg, A. Friis-Moller, J. Brynskov, I. Nordgaard-Lassen, A.M. Petersen, K.A. Krogfelt, Disease-specific enteric microbiome dysbiosis in inflammatory bowel disease, Front. Med. 5 (2018) 304.
[38] S. Coufal, N. Galanova, L. Bajer, Z. Gajdarova, D. Schierova, Z. Jiraskova Zakostelska, K. Kostovcikova, Z. Jackova, Z. Stehlikova, P. Drastich, H. Tlaskalova-Hogenova, M. Kverka, Inflammatory bowel disease types differ in markers of inflammation, gut barrier and in specific anti-bacterial response, Cells 8 (7) (2019).



Lionel Blanchet received his M.Sc. from the engineering school Polytech'Lille, France, in 2005. He completed his PhD in 2008 on the application of chemometrics on spectroscopic data at the universities of Lille, France, and Barcelona, Spain. After multiple years of post-doctoral research at Radboud University (The Netherlands), Maastricht University (The Netherlands) and Dartmouth College (USA); he joined Philips in 2016 as a senior data scientist, currently focussing on the application of deep learning in the clinical workflow.

Raffaele Vitale graduated in Analytical Chemistry in 2012 (Università degli Studi di Roma "La Sapienza", Italy) and obtained his PhD in Statistics and Optimization in 2017 (Universitat Politècnica de València, Spain). He has authored more than 23 publications. He has been granted with several awards including Siemens Process Analytics Prize for Young Scientist in 2017, the III Jean-Pierre Huvenne Award for the Best PhD thesis in 2019 and the XVI European Network for Business and Industrial Statistics Young Statistician Award in 2020. His main research interests are statistics and multivariate data analysis/ chemometrics for complex data in applied sciences.

Robert van Vorstenbosch studied Forensic Science and Chemistry at the University of Amsterdam and VU University Amsterdam (The Netherlands). The topic of these studies ranged from the analysis of human hair for human provenancing, the detection of explosives, to studying Vacuum Ultra-Violet Spectroscopy, to finally correlating molecular information to sensory characteristics of foods and beverages. He has been working as PhD student in the field of volatile organic compounds and its relation to detect colorectal cancer at early stage. His main research of interest is multivariate data analysis for –omics related problems.

Georgios Stavropoulos studied chemistry at the University of Patras, department of chemistry, Greece, where he obtained his Bachelor degree. In August 2015, he moved to The Netherlands for his Master studies, where he studied Chemistry: Analytical Sciences at the University of Amsterdam, department of chemistry. Since 2017 he has been doing his PhD at the University of Maastricht, department of Pharmacology and Toxicology, where he is using advanced machine learning techniques to comprehensively understand and profile primary sclerosing cholangitis disease.

A/Prof John Penders is an expert in molecular epidemiology and microbial ecology. His research group integrates metagenomic methods within the context of prospective epidemiological studies using various longitudinal statistical and bioinformatics tools to elucidate the role of the microbiome in health and disease. His group (at Maastricht University, The Netherlands) is currently funded by The Netherlands Organization for Scientific Research, The Netherlands Organization for Health Research and Development and the Joint Programming Initiative on Healthy Diet for Healthy Living. He has authored more than 120 publications, including in leading journals like Nature Biotechnology, Lancet Infectious Diseases, Gastroenterology, Gut, Mucosal Immunology and Microbiome.

Prof Daisy Jonkers has studied Health Sciences at Maastricht University (1987-1992) and obtained her PhD-degree in 1997. Since March 2019, she is appointed as Professor of Intestinal health . Her main research topic is to understand intestinal health in the context of common intestinal disorders with a special focus on GI function, diet and the microbiome.She is PI of several cohort and (nutritional) intervention studies in IBD and IBS. She participates in the Carbohydrate Competence Center (wwwcccresearch.nl), is board member of the Experimental Section of Gastroentrology of the Dutch Society, PI of TKI project Well on Wheat and partner in H2020/EU DISCOVeRIE project.

Prof Frederik-Jan van Schooten Prof. Van Schooten studied biology at the Free University of Amsterdam, The Netherlands. In 1991 he obtained his PhD in Chemical Carcinogenesis at the University of Leiden, The Netherlands. Since 1999 he is Head of the Department, and since 2000 Leader of the Division IV within the Nutrition Toxicology and Environment Research Institute Maastricht (NUTRIM). His research interests are in unravelling nutrition-gene interactions with respect to the carcinogenic and atherosclerotic process as well the use of exhaled breath as new matrix for disease diagnosis and monitoring. He has authored more than 400 publications.

Agnieszka Smolinska studied Chemistry at Silesian University in Katowice, Poland. In 2012 she obtained her PhD from Radboud University in Nijmegen, The Netherlands in metabolomics filed and advanced machine learning/chemometrics. After her PhD, she has been working as postdoc at Dartmouth University in USA. Her current research group (at Maastricht University, The Netherlands) is focused on the multiple applications of volatile metabolites in exhaled air and other biofluids and finding their relation to the gut microbiome using machine learning/ multivariate statistics. She was granted with various awards (best PhD thesis, metabolomics young scientist) and has authored more than 45 publications.