Variable selection and importance in presence of high collinearity: an application to the prediction of lean body mass from multi-frequency bioelectrical impedance.

Camillo Cammarota*[1], Alessandro Pinto†[2]

[1]Department of Mathematics
[2]Research Unit on "Food Science and Human Nutrition",
Department of Experimental Medicine
"Sapienza" University of Rome
P.le A. Moro 5, 00185 Rome, Italy

May 14, 2020

Abstract. In prediction problems both response and covariates may have high correlation with a second group of influential regressors, that can be considered as background variables. An important challenge is to perform variable selection and importance assessment among the covariates in presence of these variables. A clinical example is the prediction of the lean body mass (response) from bioimpedance (covariates), where anthropometric measures play the role of background variables. We introduce a reduced dataset in which the variables are defined as the residuals with respect to the background, and perform variable selection and importance assessment both in linear and random forest models. Using a clinical dataset of multi-frequency bioimpedance, we show the effectiveness of this method to select the most relevant predictors of the lean body mass beyond anthropometry.

Keywords: variable selection, importance, linear model, random forests, bioimpedance, multi-frequency, anthropometric variables, lean body mass.

1. Introduction

In biomedical research a typical challenge is the prediction of a target variable of clinical interest, measured using invasive methods, from a set of covariates measured using non-invasive methods. Furthermore a variable selection among the covariates is also needed in order to select those that are most influential and to quantify their importance for applications. In this framework two types of problems may occur.

*cammar@mat.uniroma1.it
†alessandro.pinto@uniroma1.it

The first is that the covariates may have strong collinearity. The second is the role of a different group of variables: these variables are able to explain a large part of the variability of both the target and the covariates, so that they can be considered influential "background variables" [30]. In several cases the correlation between the target and the covariates disappears conditioning on these variables (spurious correlation).

The usual approach in the framework of linear models is to include the background variables in the regressors but often collinearity produce variance inflation and not reliable estimates. In the framework of linear models a variable importance can be associated to each regressor, obtained from an additive decomposition of the R^2, also if the variable is not significant in conjunction to the others [15, 16]. This approach can be subjected to severe limitations for the a priori assumed functional form (linearity) of the dependence.

A different approach is provided by the random forests, based on the tree-structured regression [22, 31]. Tree regression has a wide applicability in biomedical field [6, 32, 24, 9, 7], were it is useful for the interpretability of the results. Random forests are extensively used in prediction and classification tasks in order to reduce the variance of the tree regression [6, 34] and bias [37]. The main advantage of random forests with respect to other learning machine methods is the ability to identify relevant variables in high-dimensional data and to provide a quantitative measure of their importance [17, 12, 35, 19]. The variable selection task is more challenging if the predictors have large correlations and the capability of random forests to select the more influential predictors was extensively investigated using simulations [30, 3, 21]. A theoretical study on the impact of the correlation among predictors on the variable importance is in [4]. Theoretical and methodological aspects of variable selection and importance measures are reviewed in [5].

The data challenge motivating this work is the prediction of the lean body mass (LBM) obtained by an invasive method, dual-energy X-ray absorptiometry (DXA) [11, 26, 18]; the predictors are the electrical impedances of the body to an alternate current at different frequencies, measured by a safe and non-invasive procedure [11, 26, 20]. The background variables are anthropometric measures of the subject (gender, age, height, weight). It is obvious that, among background variables, at least height and weight are greatly influential both of the lean body mass and of the impedance, that depends linearly on the length of arms and legs.

Prediction tasks and selection of variables in clinical applications are subjected to severe limitations due to the interpretability of the results and their simple use in practice. First of all clinical databases that include a variable measured invasively are necessarily not large, hence it is not possible to perform a prediction conditional to the background variables. Second, a few influential variables are to be selected,

typically two, for easy of graphical representation. Third, if different set of variables are proposed, one has to select the one that provides the best prediction of the target beyond the background variables.

In previous clinical studies on body composition [33, 10, 8, 36, 27] these problems were investigated only in the framework of linear models and no systematic investigation was performed for variable selection and importance assessment. The major flaw of these studies is that the collinearity among all the covariates was not taken into account.

A possible approach to the problem of collinearity is to use the notion of partial correlation between two variables, defined as the correlation between the residuals of two linear models having as regressors the remaining covariates.

In our study we are interested in the residuals of the target and of a group of explanatory variables with respect to the background variables. This approach has been often adopted in econometrics literature since the FWL theorem [23]. We analyze models for the residuals obtained predicting on the background variables, in order to evaluate the importance of the explanatory variables and perform variable selection. We consider a standard linear model and a non parametric one, the random forests.

It is worth of mention that residuals with respect to anthropometry are used in clinical studies of body composition [8, 1]. We apply the above methodology to analyze a clinical database of 135 healthy subjects that underwent DXA examination, collecting LBM, anthropometric variables and 10 impedance measures, i.e. resistances and reactances at five frequencies.

In the next section we describe the methods of variable selection and importance for linear models and random forests. In the third section we apply the above methods to analyze the data. In the fourth we performa simulation study. In the last section we provide the conclusions.

2. Methodology

2.1. **The reduced dataset.** The complete dataset consists of the target variable Y and of two matrices X and B, where the columns $X_{.j}, j = 1, ..., p$ are a group of predictors, and the columns $B_{.k}, k = 1, ..., q$ are the background variables.

We consider the complete linear model

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{k=1}^{q} B_{ik}\alpha_k + \epsilon_i, \qquad i = 1, \ldots, n; j = 1, .., p; k = 1, \ldots, q \tag{1}$$

where ϵ is the noise term.

The main limitation of the use of linear model in presence of high collinearity in the standard least squared estimation is that the variance of the estimator of the h-th parameter is inflated by the factor $1/(1 - R_h^2)$ where R_h^2 is the multiple R-squared of the regression of the h-th covariate on the other covariates. This may cause that some of the variables are not significant according to the standard t-test, but the R-squared of the model is significant according to the Fisher test.

In order to perform the variable selection we consider the reduced dataset defined as follows. We denote $Y^{(B)}$ the residuals of the linear model of Y with respect to B, and $X^{(B)}$ the matrix whose columns are the residuals of the regression of the columns of X with respect to B. We call for brevity 'reduced dataset' the new dataset having as target the variable $Y^{(B)}$ and predictors the variables $X^{(B)}$.

2.2. **The reduced linear model.** We call 'reduced linear model' the linear regression of $Y^{(B)}$ on $X^{(B)}$. In this problem of prediction both the explanatory variables $X^{(B)}$ and the target $Y^{(B)}$ are residuals, i.e. quantities estimated and not observed. In the assumption of multivariate normality and independence for the complete observed dataset, also the residuals are multivariate normal, but the independence is no longer true. As a rule of thumb [28] the residuals can be considered approximately independent if the number of explanatory variables is much less than the number of samples. In the present application the explanatory variables are the background and their number is $q = 3$, and the number of samples is $n = 135$.

There is not a simple relationship between the R^2 of the complete model and the one of the reduced model. In our data the complete model has $R^2 = 0.90$, the reduced model has $R^2 = 0.50$. In the reduced model the response $Y^{(B)}$ and the covariates $X^{(B)}$ are both orthogonal to the columns of the background B; hence one could expect to obtain from analysis of the reduced model new informations on the most influential among the X variables, that are independent on the ones in B.

2.3. **The relative importance metrics.** The relative importance metrics for linear models are described in [16, 15] and they are implemented in the `R` [29] package `relaimpo` [14]. We use the metric `lmg` defined as following. For a regressor with index k among p regressors, the additional R_k^2 is computed as following: given a permutation π of $(1, \ldots, p)$ the $R_k^2(\pi)$ is the increment of R^2 for the addition of this regressor to the set of regressors preceding k in π. The R_k^2 is defined as the average over all the permutations π of this additional $R_k^2(\pi)$:

$$R_k^2 = \frac{1}{p!} \sum_{\pi} R_k^2(\pi) \tag{2}$$

The remarkable property of this metric is that it provides an additive decomposition of the model R^2 that is independent on the order of regressors:

$$R^2 = \sum_{k=1}^{p} R_k^2 \tag{3}$$

2.4. **Random forests.** The random forests algorithm is a non-parametric method based on the tree-structured regression [22, 31]. We apply this algorithm to the reduced dataset defined in sec. 2.1, where both the target and the explanatory variables are residuals, with the assumptions on normality and independence discussed in sec. 2.2.

The tree regression is implemented in several `R` packages; we have used the function `ctree` in the `party` package [32], that can be summarized as follows:

i) in a database where Y is the target variable and $X_1, ..., X_p$ are predictors, a test of the association of between Y and any single predictor is performed using as statistic the linear correlation. The global null hypothesis of no association between any of the predictors and the target is performed, with Bonferroni adjustment for multiple testing. Stop if this hypothesis cannot be rejected. Otherwise select the variable X_j that has the maximal association with Y, computed by 1-p-value exceeding 0.95.

ii) the range of X_j is split in two intervals to achieve the best piecewise constant fit of Y; more precisely the split value s in the range of X_j is chosen to get the following minimum

$$\min_s \sum_{i:X_{ij}\leq s} (Y_i - Y_1)^2 + \sum_{i:X_{ij}>s} (Y_i - Y_2)^2 \tag{4}$$

where Y_1, Y_2 are respectively the means of Y in the sets $\{i : X_{ij} \leq s\}$, $\{i : X_{ij} > s\}$.

iii) for each of the two sets of samples $\{i : X_{ij} \leq s\}$, $\{i : X_{ij} > s\}$ the previous steps are replicated until the process stops when no significant association of Y with any covariate is found. Different criteria for testing association, splitting and stopping can be chosen; details are in [32].

The trees constructed on a learning sample can be considered weak learners since they have low bias and high variance. A collection of trees, the forest, is constructed in order to obtain an unique predictor with reduced variance.

i) A bootstrap sample of the learning set is randomly selected.

ii) A tree is grown on this sample as before with the only difference that at each node m covariates X_j are randomly chosen out of the p available.

iii) The prediction in the remaining dataset called out-of-bag (OOB) sample is obtained as the average of all trees predictions. The number m and the number of bootstrap samples are the only parameters to be selected.

2.5. **Permutation importance.** The variable importance implemented in random forests framework is based on the idea that if X_j is a relevant predictor a permutation

of its values (or of the response) destroys the prediction accuracy. The importance is computed by the following steps:

i) A bootstrap sample consisting of 2/3 of the observations is selected and a tree is grown on it. The remaining observations, considered OOB observations, are used to test the prediction. The accuracy is computed as the mean squared error (MSE).

ii) For each variable X_j the importance is computed in terms of the difference of MSE between the prediction obtained using X_j and the permuted version X'_j (or of the response). More precisely, for a tree t the OOB- MSE is computed as

$$\text{OOB-MSE}_t = \frac{1}{|OOB_t|} \sum_{i \in OOB_t} (Y_i - \hat{Y}_i^{(t)})^2 \tag{5}$$

where OOB_t is the set of terminal nodes of the tree t and $\hat{Y}_i^{(t)}$ is the prediction according to t. The same quantity is computed for the permuted variable X'_j (or the response) and the difference with respect to the previous is computed.

ii) The operation is repeated for all bootstrap samples, typically 1000, and the average is computed. For details see [17, 34, 30]. The percentage increasing in MSE (%IncMSE) is also used, defined as MSE after permutation minus the one before permutation divided by the latter. This method produces an empirical null distribution of importance for each predictor; the p-value is extracted comparing with the original importance scores.

In this work we use a test of significance for the importance metric implemented in the package `rfPermute` [2]. The significance is obtained by permuting the response variable.

TABLE 1. Summary statistics of anthropometry, impedance data and lean body mass (LBM) of 135 subjects. Units: LBM (kg), height (m), weight (kg), age (years); R = logarithm of resistance (Ohm); X= reactance (Ohm)

Statistic	N	Mean	St. Dev.	Min	Max
LBM	135	55.15	8.18	36.59	74.95
height	135	1.62	0.06	1.45	1.80
weight	135	97.74	17.58	56.20	136.80
age	135	44.86	13.23	18	69
R5	135	6.31	0.14	5.94	6.68
R10	135	6.28	0.14	5.92	6.65
R50	135	6.18	0.14	5.83	6.54
R100	135	6.13	0.14	5.79	6.49
R250	135	6.05	0.14	5.73	6.42
X5	135	25.72	5.22	9.93	41.92
X10	135	35.79	6.89	18.84	59.69
X50	135	49.39	8.38	29.81	74.41
X100	135	44.08	7.21	26.14	62.02
X250	135	30.67	5.77	17.10	44.97

TABLE 2. Pearson correlations of the variables

	LBM	height	weight	age	R5	R10	R50	R100	R250	X5	X10	X50	X100	X250
LBM	1	0.42	0.87	-0.21	-0.59	-0.61	-0.62	-0.63	-0.63	-0.16	-0.24	-0.38	-0.45	-0.49
height	0.42	1	0.21	-0.21	0.17	0.17	0.16	0.16	0.17	0.10	0.10	0.12	0.13	0.15
weight	0.87	0.21	1	-0.16	-0.59	-0.59	-0.59	-0.60	-0.59	-0.19	-0.31	-0.48	-0.53	-0.53
age	-0.21	-0.21	-0.16	1	-0.13	-0.12	-0.07	-0.04	-0.005	-0.41	-0.43	-0.47	-0.43	-0.33
R5	-0.59	0.17	-0.59	-0.13	1	1.00	0.99	0.99	0.98	0.60	0.72	0.82	0.85	0.82
R10	-0.61	0.17	-0.59	-0.12	1.00	1	0.99	0.99	0.98	0.58	0.70	0.80	0.84	0.82
R50	-0.62	0.16	-0.59	-0.07	0.99	0.99	1	1.00	0.99	0.52	0.63	0.74	0.79	0.80
R100	-0.63	0.16	-0.60	-0.04	0.99	0.99	1.00	1	1.00	0.50	0.61	0.72	0.77	0.78
R250	-0.63	0.17	-0.59	-0.005	0.98	0.98	0.99	1.00	1	0.48	0.58	0.69	0.74	0.76
X5	-0.16	0.10	-0.19	-0.41	0.60	0.58	0.52	0.50	0.48	1	0.93	0.80	0.72	0.56
X10	-0.24	0.10	-0.31	-0.43	0.72	0.70	0.63	0.61	0.58	0.93	1	0.92	0.85	0.68
X50	-0.38	0.12	-0.48	-0.47	0.82	0.80	0.74	0.72	0.69	0.80	0.92	1	0.98	0.86
X100	-0.45	0.13	-0.53	-0.43	0.85	0.84	0.79	0.77	0.74	0.72	0.85	0.98	1	0.93
X250	-0.49	0.15	-0.53	-0.33	0.82	0.82	0.80	0.78	0.76	0.56	0.68	0.86	0.93	1

3. Application

3.1. **Measures.** We apply the above methods to perform variable selection and importance assessment for the bioimpedance data in the prediction of the lean body mass, using the anthropometric variables as background. The data are extracted from a database collected at the Food Science and Human Nutrition Research Unit of the Department of Experimental Medicine of Sapienza Rome University in the years 2017-2018. The dataset extracted for the present study is enclosed as supplementary material. This dataset contains a group of 135 overweight and obese women that underwent dual-energy X-ray absorptiometry (DXA) examination (Hologic 4500 RDR). This method [18] provides an accurate prediction of body composition that is commonly used as a reference to validate bioimpedance prediction equations [26]. Whole body bioimpedance measurements were performed according to the standardized protocol [26], using the multi-frequency device `Human im Touch` (Ds Medica, Milan, Italy). The database collected raw multi-frequency impedance data (resistance and reactance denoted respectively by R and X) measured at five frequencies (5, 10, 50, 100, 250 kHz). The anthropometric variables include for each subject height, weight, age.

3.2. **Description of the dataset.** The descriptive statistics of the variables included in the study are in Tab. 1. To assess the normality of the variables distribution we have used the `R` package `LambertW` [13] in which the Shapiro-Wilk, Shapiro-Francia and Anderson-Darling normality tests are used. The resistance data are generally non Gaussian (right skewed), and this can be corrected using a logarithmic transformation. This is coherent to what was observed by [25] i.e. that random effects related to impedances have a log-normal distribution. The reactance data are normal. The variables LBM, height, weight are normal; the age has a small deviation from normality not corrected.

Tab. 2 reports the Pearson correlations among the variables. Resistances show high collinearity having correlations greater than 0.98; the reactances are moderately correlated (greater than 0.53). The target variable LBM has a a strong correlation to the weight (0.85) as expected, and a moderate negative correlation to the resistances and reactances. The resistances have a negative correlation (-0.60) to the weight, and the reactances a negative correlation to the age.

3.3. **Linear models.** We fit to the complete dataset the standard linear model with ordinary least squares estimation in eq. (1) (complete linear model), using the t-test for the significance of the parameters and the Fisher test for the significance of R^2. The complete linear model has $R^2 = 0.90$ and the only significant variables are intercept, weight, height. This result can be explained by the high collinearity of the resistances that inflates the variance of their estimates. The model that uses

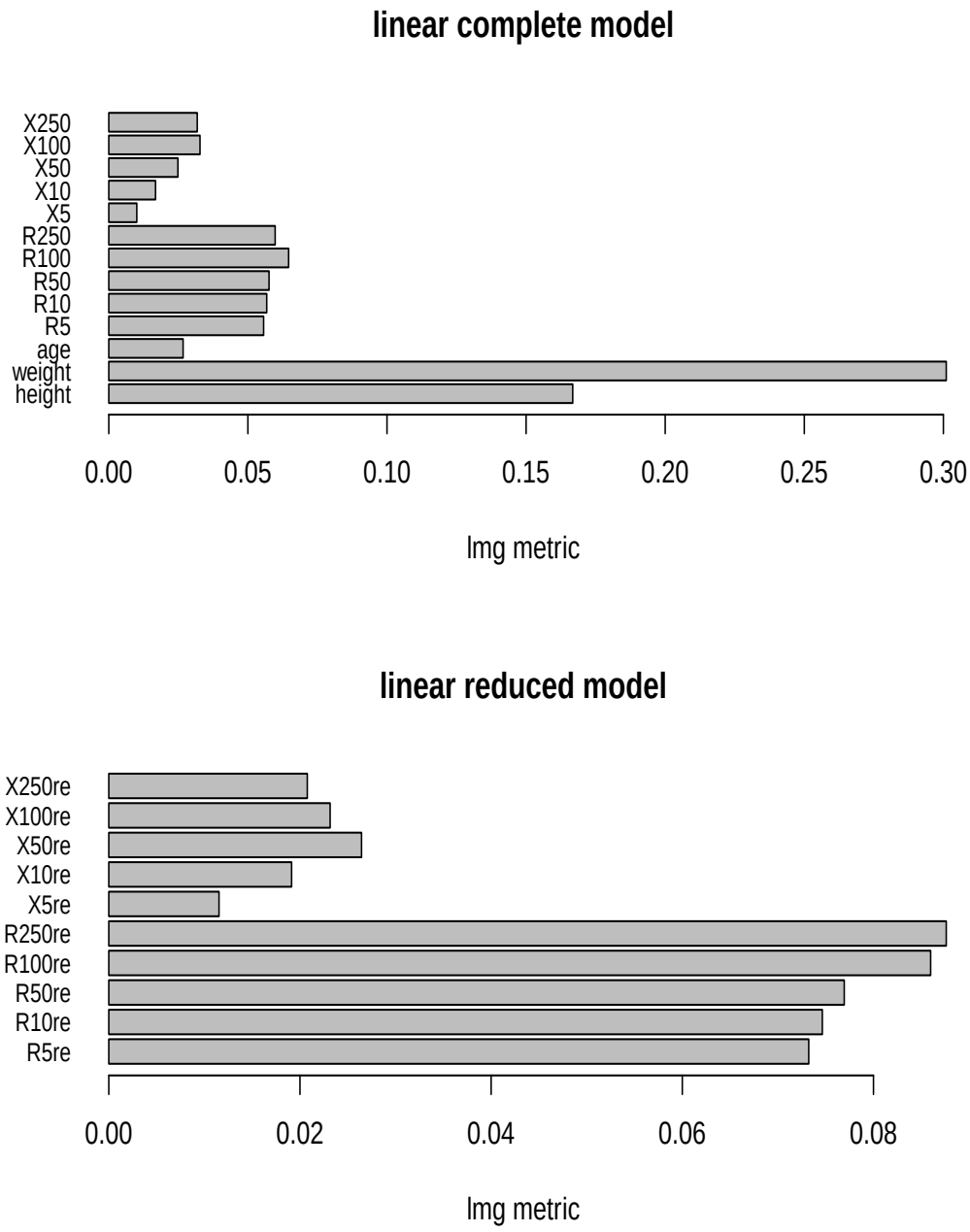

FIGURE 1. Variable importance of linear model prediction of LBM in complete dataset (upper) and reduced dataset (lower). The bar heights sum up to the R^2 model.

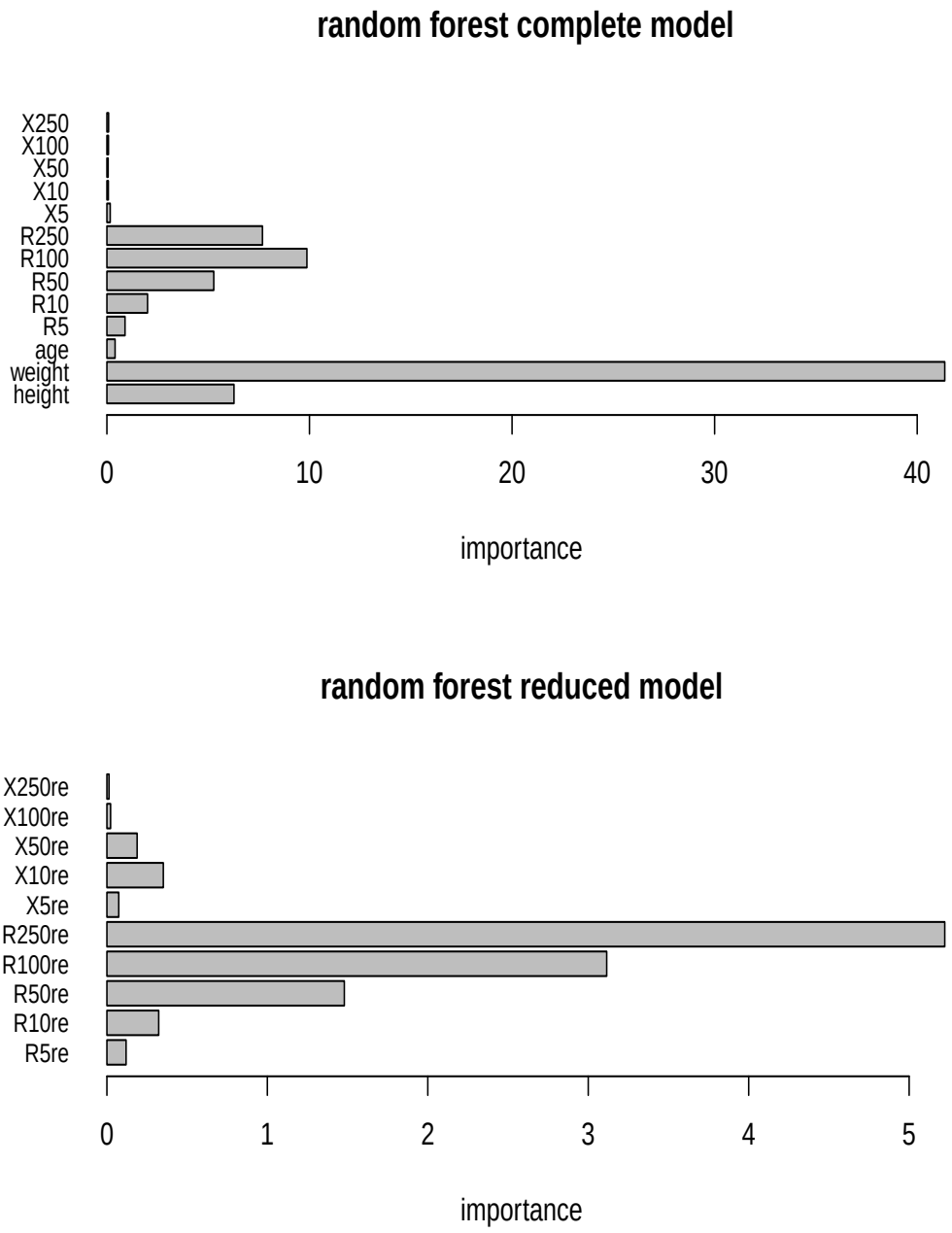

FIGURE 2. Variable importance of the random forest prediction of the LBM in complete dataset (upper) and reduced dataset (lower). Parameters setting: `ntree`=1000, `mtry`=4

only anthropometry as predictors has $R^2 = 0.81$. This suggests to investigate more deeply the role of the resistances and reactances in the prediction of the target beyond anthropometry.

We have analyzed the reduced linear model i.e. the linear model of the reduced dataset obtained computing the residuals with respect to anthropometry both of the target and of the other covariates, according to secs. 2.1 and 2.2. The normality of residuals is verified and the covariance matrix is computed (not shown here). This matrix reveals in turn collinearity. The reduced linear model is significant in the Fisher test for the R-squared with $R^2 = 0.50$, but none of the variables is significant in the standard t-test. This can explained as before from the inflated variance of the estimates.

The importance metric, that provides an additive decomposition of the R^2 with respect to the covariates, is summarized in fig. 1 both for complete and reduced models.

In the complete model having $R^2 = 0.90$ the anthropometric variables height and weight are the most important, and resistances are more important than reactances. In the reduced model having $R^2 = 0.50$ the importance of resistances over reactances results increased.

3.4. **Random forests.** We apply the random forest approach for the variable importance assessment both in the complete dataset and in the reduced dataset obtained according to the procedure described in sec. 2.1. In random forests approach the definition of importance is not related to a variance decomposition as in linear models, but to the permutation-based reduction of the MSE of prediction. From sec. 2.4 the parameter `ntree` (number of trees) was set to 1000 and the parameter `mtry` (number of variables randomly selected at each step) was changed from 1 to 6, without observing relevant differences in the importance allocations. In fig. 2 the upper panel shows that among the anthropometric variables the weight has larger importance, and among the other covariates the resistance at 100 KHz has an importance greater than others. The lower panel shows the results for the reduced model. The importances of the resistances is larger than reactances and the plot suggests that it is increasing with frequency, having its maximum at 250KHz.

3.5. **Permutation-test.** We have performed the test of significance of importance in the case of the reduced dataset, in order to select the variables (resistances and reactances) more influential beyond anthropometry. This test, usually adopted in the literature [12, 16, 17], is based on the permutation-increase of MSE, as described in section 2.5. In tab. 3 we report the results of the test. The first row shows the %IncMSE for each variable of the reduced model (average of 200 replicates) and the

second row the p-value obtained from the empirical distribution of 200 replicates. The only variables having significant importance are R250, R100, R50.

TABLE 3. Test of significance of the importance metric defined by % increase in mean squared error in the reduced dataset. Only the importances of three variables (R250, R100, R50) result significant.

	R250re	R100re	R50re	R10re	R5re	X50re	X10re	X100re	X5re	X250re
%IncMSE	5.38	5.07	2.91	1.55	1.07	0.85	0.60	0.49	0.33	0.28
%IncMSE.pval	0.00	0.00	0.01	0.34	0.74	0.66	0.69	0.88	0.69	0.68

4. Simulation study

This study is conducted to evaluate the ability of the proposed method to distinguish relevant from irrelevant variables in different correlation schemes, characterized by high collinearity. In the correlation table (tab. 2) of the observed dataset the resistances have very high correlations (the maximum is 0.99), that cannot be increased. Consequently we have investigated the performance of the method when this maximum is lowered preserving a correlation scheme similar to the observed one. We have used the following method. Given a pair of variables $x,\ y$ consider the new pair defined by $x' = x + w_x,\ y' = y + w_y$, where the two terms w_x, w_y are independent each other and from x, y, with $\mathbb{E}(w_x) = 0, \mathbb{V}ar(w_x) = \alpha^2 \mathbb{V}ar(x)$ and similar for y. Then the correlation of $x',\ y'$ is obtained from the correlation of $x,\ y$ lowered by the factor $1 + \alpha^2$.

In the simulations we have generated datasets of 135 observations in three different cases. In the first the observations are distributed according to a multivariate normal having mean and covariance obtained from the observed one. In the second we have added to each variable an amount of noise with standard deviation equal to 10% of the standard deviation of the variable (case $\alpha = 0.1$); in the third case have used $\alpha = 0.2$. In each case consisting of 100 simulations we have obtained the reduced dataset defined by the residuals with respect to the background and computed the importances from the linear and forest methods. The results are summarized in the figs. 3, 4, 5, where the box plots of the 100 simulations are shown for each of the predictors in the reduced dataset. These figures should be compared with the lower panels of figs. 1 and 2. The following features are preserved across the simulations: the resistances are more important than reactances both in linear and forest method; the forest method shows a differentiation among the resistances allocating greater importance to resistances at larger frequencies (R100, R250).

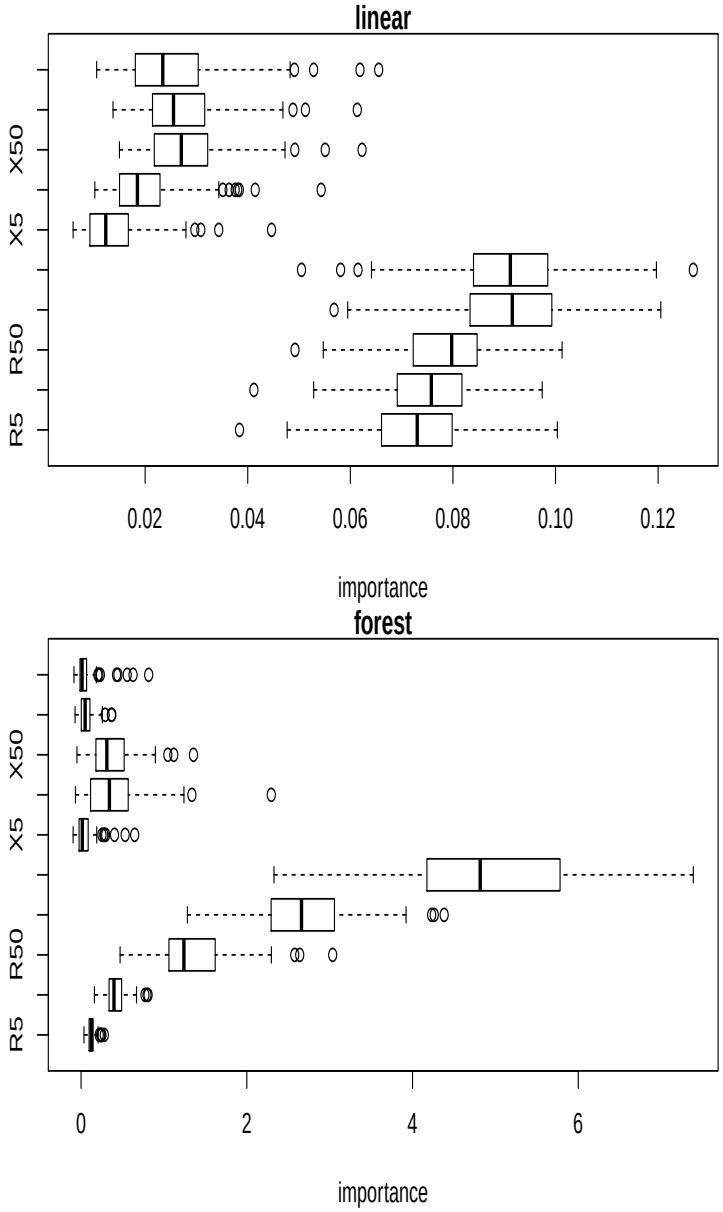

FIGURE 3. Simulation study 1 - Box plots of the importance of the predictors in the reduced dataset for 100 simulations of the lmg metric (upper) and permutation metric (lower).

5. Conclusion and discussion

We have considered a variable selection task in presence of high collinearity, for two groups of predictors, one of which plays the role of influential background variables and the other one are variables of clinical interest. We have considered a reduced dataset obtained from the residuals with respect to the linear fit on the

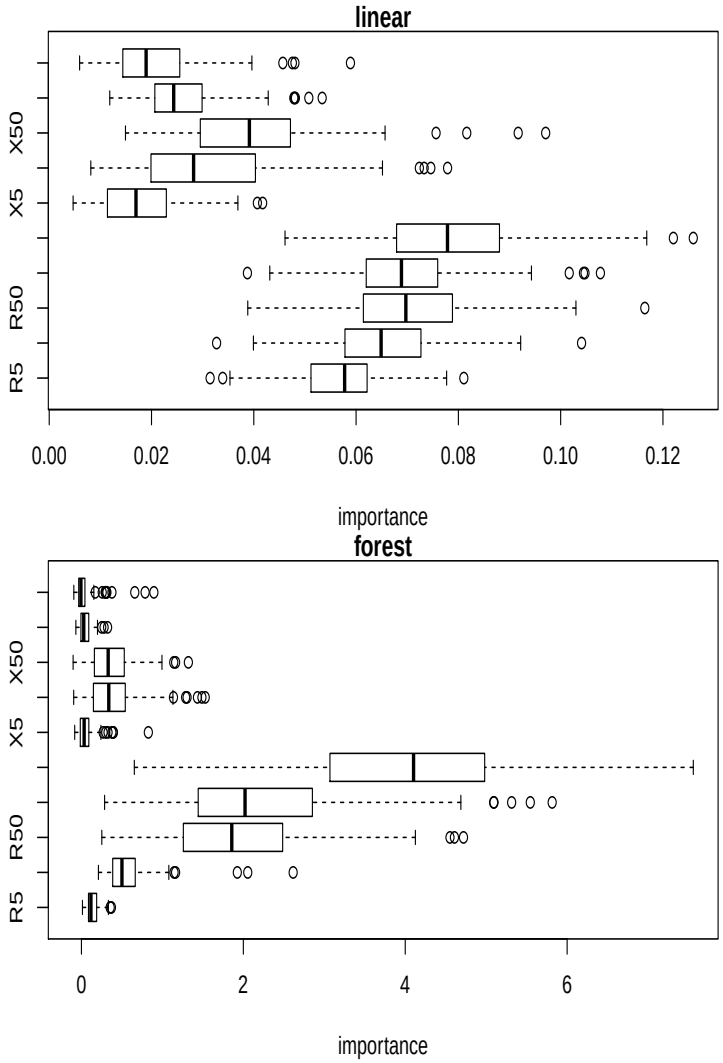

FIGURE 4. Simulation study 2 - Box plots of the importance of the predictors in the reduced dataset for 100 simulations of the lmg metric (upper) and permutation metric (lower) where each variable is added with noise with standard deviation 10% of the one of the variable.

background variables of all other variables. Two problems may occur. First, the collinearity may be present also in the reduced dataset and this prevents from using standard methods for variable selection. Second, when the background variables are able to explain a large part of the variability of the response, it is not obvious that the reduced dataset can reveal a residual dependence on the covariates, useful for applications.

We have applied two methods for variable selection: the relative importance metric in the framework of linear methods, and the permutation importance in the

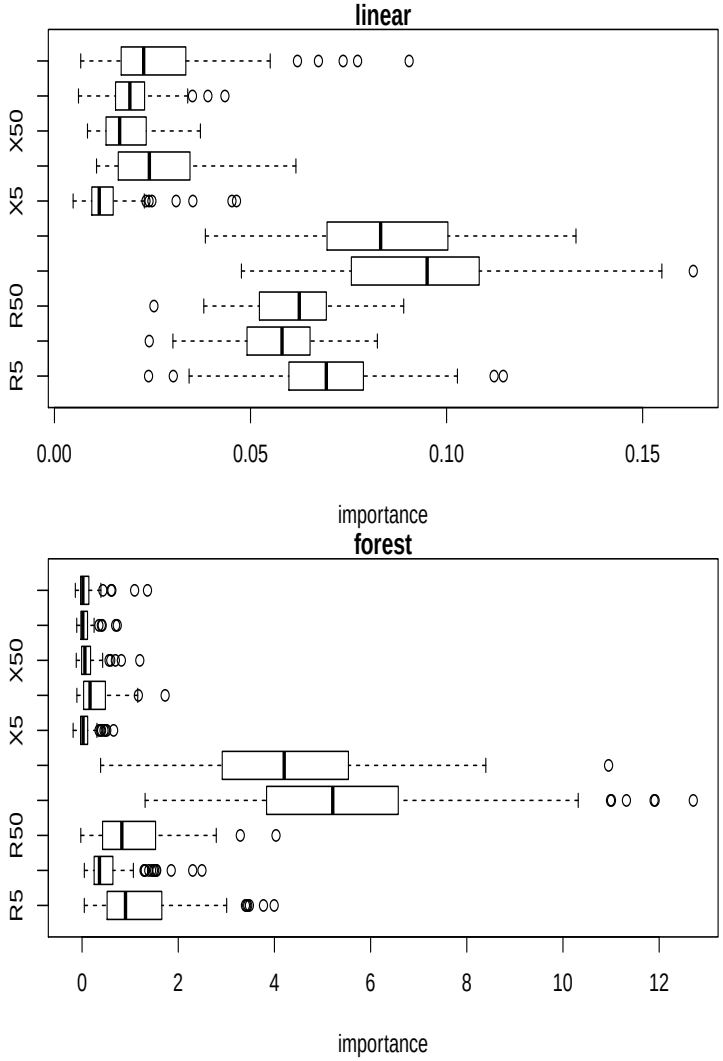

FIGURE 5. Simulation study 3 - Box plots of the importance of the predictors in the reduced dataset for 100 simulations of the lmg metric (upper) and permutation metric (lower) where each variable is added with noise with standard deviation 20% of the one of the variable.

framework of random forests. The application has been performed both in the complete and in the reduced dataset in order to compare the results.

The main objective of the paper was to select the most influential variables from bioimpedance beyond the anthropometry (background) in the prediction of lean body mass. The main results, shown in fig. 1 and fig. 2, are:

1) In the complete dataset the anthropometry has globally larger importance than bioimpedance; the most important variable is weight both in linear and in random

forest prediction. Actually, the prediction of the response in the linear model from the anthropometry has $R^2 = 0.81$ and from all the predictors has $R^2 = 0.90$.

2) In the reduced dataset ($R^2 = 0.50$), the resistances are more important than the reactances both in linear and in random forest prediction.

3) In the reduced dataset for both types of prediction the importance of the resistances is allocated increasingly with respect to the frequency, having its maximum at 250 KHz. The empirical test of significance selects only the resistances R50, R100 and R250 as having a significant importance.

4) Comparison between complete and reduced datasets reveals that the importance allocations may be different: in the complete dataset R100 is the most important and in the reduced dataset R250 is the most important. This inversion is present both in linear and random forest approaches.

The simulation study conducted in three different correlation schemes concerns the prediction in the reduced model and is aimed to give insights on points 2) and 3) above. The simulation confirms that the method is able to distinguish between two groups of predictors, allocating more importance on the resistances than on the reactances, and is able to select among the resistances the ones at high frequencies (R100, R250) having greater importance.

The theoretical and methodological aspects of the importance measures in random forests are still object of research, mainly focused to investigate the performance of the methods when the predictors are highly correlated [5]. In the particular case of additive regression models it is possible to describe the impact of the correlation on the permutation importance and to show the efficiency of the algorithm to select a small number of variables [4]. This and other methods consider special examples of covariance schemes and are not yet sufficiently general to include clinical applications as the present one. We are not aware of investigations on the role of background variables in the importance measure, with the exception of [30] where a conditional variable importance is defined. This method computes the permutation importance on subset of samples obtained from the splitting of the variables to be conditioned on during the grow of the trees. This obviously does not defines a reduced dataset that in our approach is used to apply a different method of variable selection and compare the results. The method here proposed can be justified by the fact that the number of predictors used for computing the residuals is much smaller than the number of samples (respectively 3 and 135) so the samples in the reduced dataset can be considered approximately independent. The simulation has confirmed the validity of this approach.

Our clinical oriented application suggests three open problems: 1) to give a theoretical justification of the use of residuals with respect to a group of background

variables to perform a variable selection in the remaining group of variables of clinical interest; 2) to provide a test to compare two permutation importances; 3) to select variables among different subgroups having different physiological roles, such as resistances and reactances.

The main contribution of this study to the prediction of the lean body mass is the evidence of increasing allocation of importance of resistances with respect to frequency observed both in linear and random forest approach. A possible explanation is that this increase of importance is due to the well known fact that the alternate current penetrates into intracellular water of lean mass increasingly with frequency. We conclude that R250, the resistance at 250 KHz, could be selected as the most influential predictor, beyond anthropometry. It is worth of mention that for prediction of body composition the traditional clinical practice of bioimpedance analysis uses measures obtained at a single frequency, typically the pair R50 and X50 [26].

6. Acknowledgements

We thank the Referees for valuable comments and suggestions.

References

[1] Newman AB, Kupelian V, Visser M, Simonsick E, Goodpaster B, Nevitt M, Kritchevsky SB, Tylavsky FA, Rubin SM, and Harris TB amd Health ABC Study Investigators. Sarcopenia: alternative definitions and associations with lower extremity function. *J Am Geriatr Soc.*, 51(11):1602–9, 2003.

[2] Eric Archer. *rfPermute: Estimate Permutation p-Values for Random Forest Importance Metrics*, 2018. R package version 2.1.6.

[3] Kellie J. Archer and Ryan V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249 – 2260, 2008.

[4] Gregorutti B., Michel B., and Saint-Pierre P. Correlation and variable importance in random forests. *Stat Comput*, 27:659–678, 2017.

[5] G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[6] Strobl C, Malley J, and Tutz G. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological methods*, 14(4):323–348, 2009.

[7] Guy Cafri, Luo Li, Elizabeth W. Paxton, and Juanjuan Fan. Predicting risk for adverse health events using random forest. *Journal of Applied Statistics*, 45(12):2279–2294, 2018.

[8] Paul Deurenberg, Anna Tagliabue, and Frans J. M. Schouten. Multi-frequency impedance for the prediction of extracellular water and total body water. *British Journal of Nutrition*, 73(3):349–358, 1995.

[9] Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre-Ghalila, and Mériem Jaïdane. Random forest-based approach for physiological functional variable selection for driver's stress level classification. *Statistical Methods & Applications*, 2018.

[10] Seoane F., Abtahi S., Abtahi F., Ellegard L., Johannsson G., Bosaeus I., and Ward L. C. Mean expected error in prediction of total body water: A true accuracy comparison between bioimpedance spectroscopy and single frequency regression equations. *BioMed Research International*, page 656323, 2015.

[11] Kyle Ursula G., Bosaeus Ingvar, De Lorenzo Antonio D., Paul Deurenberg, Marinos Elia, Jose Manuel Gomez, Berit Lilienthal Heitmann, Luisa Kent-Smith, Jean-Claude Melchior, Matthias Pirlich, Hermann Scharfetter, Annemie M.W.J. Schols, and Claude Pichard. Bioelectrical impedance analysis part i: review of principles and methods. *Clinical Nutrition*, 23(5):1226 – 1243, 2004.

[12] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225 – 2236, 2010.

[13] Georg M. Goerg. *LambertW: Probabilistic Models to Analyze and Gaussianize Heavy-Tailed, Skewed Data*, 2016. R package version 0.6.4.

[14] Ulrike Grömping. Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software*, 17(1):1–27, 2006.

[15] Ulrike Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

[16] Ulrike Grömping. Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, 63(4):308–319, 2009.

[17] A. Hapfelmeier and K. Ulm. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60:50 – 69, 2013.

[18] Ellis Kenneth J. Human body composition: In vivo methods. *Physiological Reviews*, 80(2):649–680, 2000.

[19] Silke Janitza, Ender Celik, and Anne-Laure Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 2016.

[20] Sami F. Khalil, Mas S. Mohktar, and Fatimah Ibrahim. The theory and fundamentals of bioimpedance analysis in clinical status monitoring and diagnosis of diseases. *Sensors*, 14(6):10895–10928, 2014.

[21] Nicodemus K.K., Malley J.D., Strobl C., and Ziegler A. The behaviour of random forest permutation based variable importance measures under predictor correlation. *BMC Bioinformatics*, (11:110), 2010.

[22] Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. *Classification and regression trees*. Chapman & Hall, Boca Raton, 1998.

[23] Michael C. Lovell. A simple proof of the fwl theorem. *The Journal of Economic Education*, 39(1):88–91, 2008.

[24] Tayefi Maryam, Esmaeili Habibollah, Karimian Maryam Saberi, Alireza Amirabadi Zadeh, Mahmoud Ebrahimi, Mohammad Safarian, Mohsen Nematy, Seyed Mohammad Reza Parizadeh, Gordon A. Ferns, and Majid Ghayour-Mobarhan. The application of a decision tree to establish the parameters associated with hypertension. *Computer Methods and Programs in Biomedicine*, 139(Supplement C):83 – 91, 2017.

[25] J M McGree, S B Duffull, J A Eccleston, and L C Ward. Optimal designs for studying bioimpedance. *Physiological Measurement*, 28(12):1465, 2007.

[26] Earthman Carrie P. Body composition tools for assessment of adult malnutrition at the bedside. *Journal of Parenteral and Enteral Nutrition*, 39(7):787–822, 2015.

[27] Georg P. Pichler, Omid Amouzadeh-Ghadikolai, Albrecht Leis, and Falko Skrabal. A critical analysis of whole body bioimpedance spectroscopy (bis) for the estimation of body compartments in health and disease. *Medical Engineering & Physics*, 35(5):616 – 625, 2013.

[28] Draper N. R. and Smith H. *Applied Regression Analysis*. Wiley, New York, 1998.

[29] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[30] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307), 2008.

[31] Hastie T, Tibshirani R, and Friedman J. *The elements of statistical learning*. Springer, 2001.

[32] Hothorn T, Hornik K, and Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

[33] H. van Baar, P.J.M. Hulshof, M. Tieland, and C.P.G.M. de Groot. Bioimpedance analysis for appendicular skeletal muscle mass assessment in (pre-) frail elderly people. *Clinical Nutrition ESPEN*, 10:e147 – e153, 2015.

[34] A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330 – 349, 2011.

[35] Qiang Wang, Thanh-Tung Nguyen, Joshua Z. Huang, and Thuy Thi Nguyen. An efficient random forests algorithm for high dimensional data classification. *Advances in Data Analysis and Classification*, 2018.

[36] Y. Yamada, Y. Watanabe, M. Ikenaga, K. Yokoyama, T. Yoshida, T. Morimoto, and M. Kimura. Comparison of single- or multifrequency bioelectrical impedance analysis and spectroscopy for assessment of appendicular skeletal muscle in the elderly. *J Appl Physiol*, 115(6):812–8, 2013.

[37] Guoyi Zhang and Yan Lu. Bias-corrected random forests in regression. *Journal of Applied Statistics*, 39(1):151–160, 2012.