How to choose the parameters and comparison with other models

The proposed method's R code is available in this repository, providing researchers with an accessible tool to implement the method. The R function allows for control over fundamental parameters: the significance level α, the maximum number of iterations, `mtry` the number of variables randomly selected at each split, and `ntree` the number of trees in random forest. A brief discussion on how to choose these parameters can be found in the next section.

This section is devoted to evaluating the proposed model's effectiveness, which we demonstrate through two simulated datasets and an industrial case study. We also include a comparison with existing models such as the original Boruta, Lasso, Elastic net, Knockoffs variable selection and VSURF. The first two models, Lasso and Elastic Net, are two commonly used regression models with regularization. Knockoffs variable selection [1, 2] is a variables selection model that uses a set of "knockoff" variables and it is designed to control for false discovery rate. Variable Selection Using Random Forest (VSURF) [3, 4] is also a variable selection method employing a random forest-based approach.

The first simulation aims at assessing the model's performance in detecting the most significant variables among all variables in the presence of multiple groups of correlated variables. Additionally, this simulation includes a brief discussion of the model's parameters. The second simulation evaluates the model's performance when the interactions among variables are present. The utilization of simulated data in our evaluation process serves two purposes. Firstly, it allows for greater control over the correlation among the variables, enabling a more systematic investigation of the model's performance under varying conditions. Secondly, using simulated data provides a comparison of various model selection approaches. All the analyses were performed in R language, version 4.2.3, using functions available at [5] and publicly available R packages.

Simulated Dataset. In the first simulation study, two groups of correlated variables are introduced in the dataset, and the simulations were run at increasing correlation levels from 0 to 0.9. The 20 input variables given in X are drawn from a joint normal distribution with mean 0 and variance 1. Two groups of correlated variables are introduced: $[X_1, X_2, X_3, X_4, X_5]$ and $[X_{10}, X_{11}, X_{12}, X_{13}, X_{14}]$, as shown in Figure 1. The response variable is generated as a linear combination of X' where $X' = [X_2, X_{11}, X_{19}, X_{20}]$ and the coefficients β_i for $i \in \{2, 11, 19, 20\}$ are such that $\beta_i / sd(\beta_i) = k$, where $sd(\cdot)$ is the standard deviation and $k = [4, 3, 3, 5]$. The model for Y is given as

$$Y = \beta_2 X_2 + \beta_{11} X_{11} + \beta_{19} X_{19} + \beta_{20} X_{20} + \epsilon \tag{1}$$

where ϵ represents the added noise with $\epsilon \sim N(0, 0.1)$.

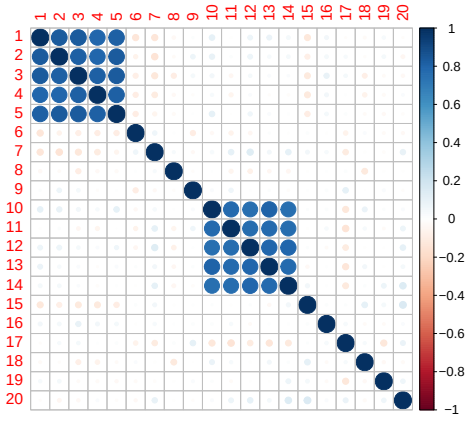
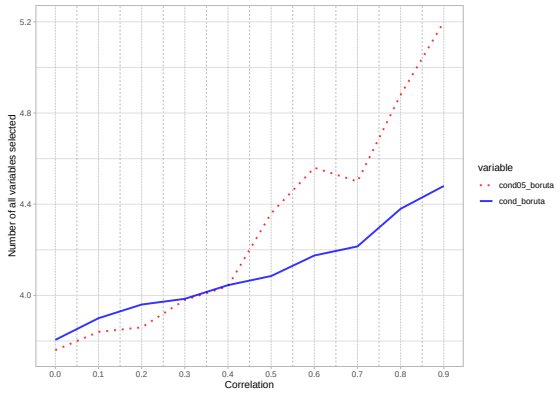
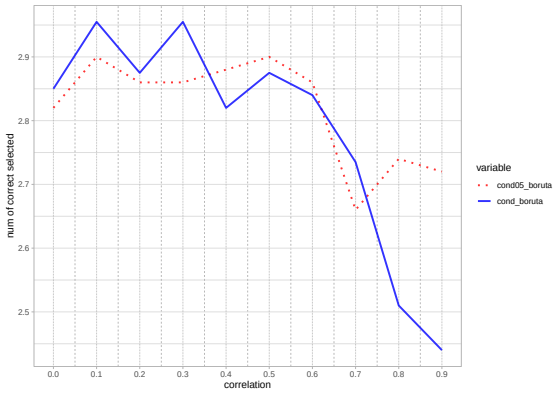


Figure 1: Correlation structure of the input variables.

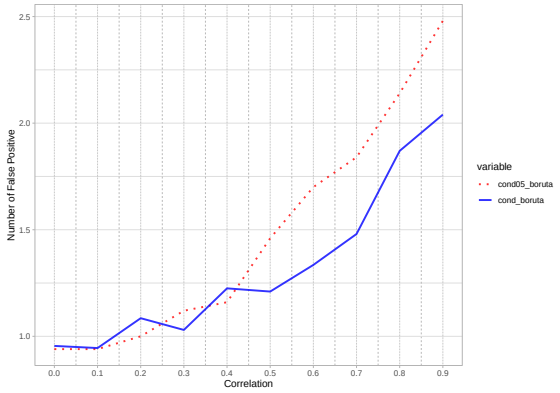
How the parameters. We performed an analysis to assess the model sensitivity to the choice of significance level α, the number of variables randomly selected at each split (`mtry`) and the number of trees in the random forest (`ntree`). In each case we performed 50 simulation runs. The first parameter being investigated is the significance level α, for



(a) Number of all selected variables versus the correlation level.



(b) Number of correctly selected variables.



(c) Number of wrongly selected variables.

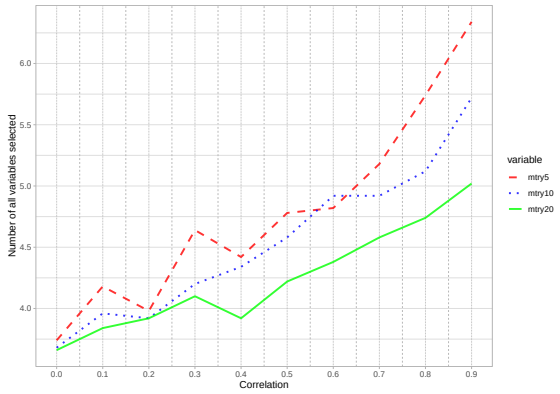
Figure 2: In the solid blue line, the default value $\alpha = 0.01$ and the dotted red line $\alpha = 0.05$. (*a*): Number of all selected variables vs correlation. (*b*): Mean number of correctly selected variables over 50 runs vs correlation. (*c*): Number of wrongly selected variables as the correlation increases.

which the default value is 0.01. Figure 2 depicts a comparison between the default value of 0.01 against 0.05, a level that is also commonly used in regression.

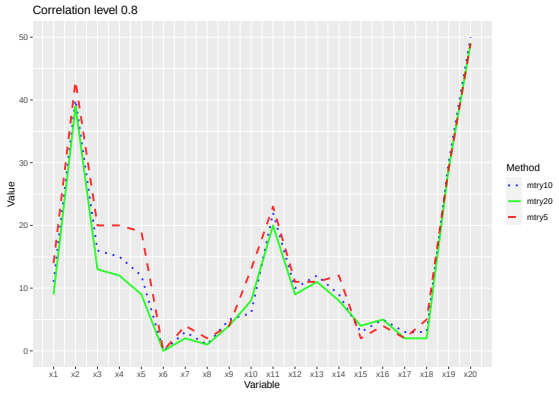
Figure 2 demonstrates that the total number of selected variables is expectantly higher for $\alpha = 0.05$ compared to $\alpha = 0.01$. This implies that on average a greater number of significant variables are being correctly selected. However, an increase in the number of selected variables also results in the selection of noise variables that are not relevant to the response variable, Y. Careful consideration of the positive and negative effects associated with different α-levels is necessary to make an informed decision. Determining the appropriate significance level depends on the specific case and objectives of the analysis.

The sensitivity of the model to the changes in `mtry` is depicted in Figure 3. As discussed in [6, 7], the CMDA method is moderately sensitive to the choice of `mtry`, albeit less so than the MDA method. This phenomenon is because correlated variables are favoured in the split selection process. Consequently, for low values of `mtry`, correlated variables are more likely to be chosen, resulting in a higher variable importance score than the uncorrelated and noise variables. As can be seen in Figure 3a, a higher number of variables are selected for low `mtry` values. This increase is partly due to the selection of noise variables that are highly correlated with the relevant variables. As shown in Figure 3b, in the case of correlation level 0.8 and low `mtry` values, the entire correlation group associated with $X2$ is selected, in addition to $X2$ itself. This effect is less pronounced for the second correlation group, as $X11$ has a lower regression coefficient. The sensitivity of the CMDA method to the choice of `mtry` suggests that careful consideration of this parameter is necessary to ensure an appropriate model.

We also assessed the impact of `ntree` parameter by evaluating the model's performance for `ntree = 500, 1000` and 2000 trees. We observed no substantial changes in the model's performance as the number of trees increased.



(a) Number of all selected variables versus the correlation level.



(b) Total number of selected variables for 50 runs.

Figure 3: In the solid green line `mtry=20`, the dotted blue line `mtry=10` and dashed red line `mtry=5`. (*a*): Number of all selected variables vs correlation. (*b*): Number of times each variable has been selected in 50 runs.

Given that the proposed method is computationally more demanding compared to the original method, we recommend using the default value of `ntree = 500`. This choice strikes a reasonable balance between model accuracy and computational efficiency.

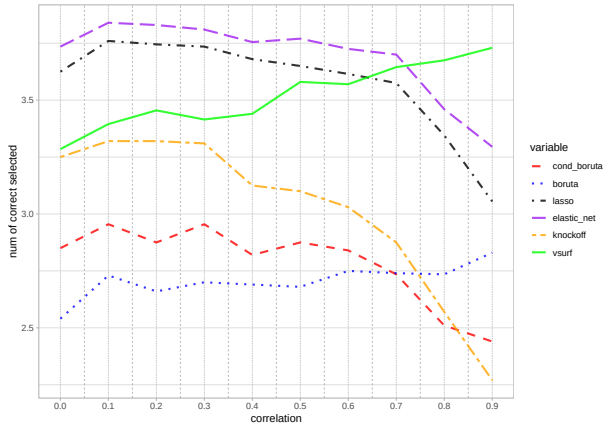
Comparison with other models. In the second simulation study, the data set is similar, except for the addition of an interaction term in the model. That is, the response variable is generated as a linear combination of X' where $X' = [X_2, X_7X_{11}, X_{11}, X_{19}, X_{20}]$ and the coefficients β_i is such that $\beta_i/sd(\beta_i) = k$, where $sd(\cdot)$ is the standard deviation and $k = [4, 5, 3, 3, 5]$. The model for Y is

$$Y = \beta_2 X_2 + \beta_{7,11} X_7 X_{11} + \beta_{11} X_{11} + \beta_{19} X_{19} + \beta_{20} X_{20} + \epsilon \tag{2}$$

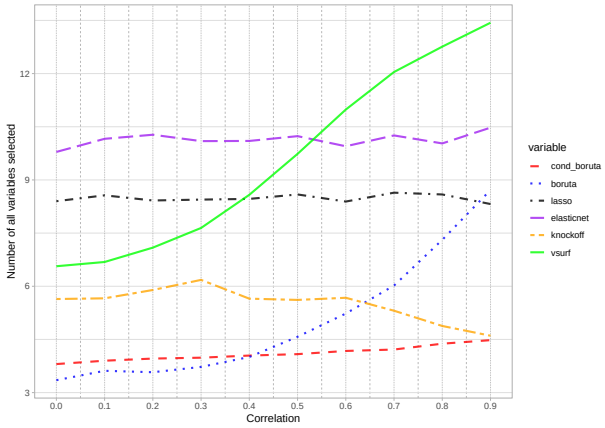
with ϵ represents the added noise with $N(0, 0.1)$. At each correlation level, the proposed model as well the original Boruta, Lasso, Elastic Net, Knockoffs and VSURF models were run 200 times each. The Lasso and Elastic Net model parameters were estimated through five-fold cross-validation. We show the model comparison results from the first simulation dateset for illustration purposes in Figure 4. The other simulation runs show similar results and hence omitted.

In both simulation studies, the Elastic Net and Lasso models gave similar performances. Both methods, on average, select one more correct variable than the other models, as can be seen in Figure 4a. However, Figure 4b shows that both models on average select a larger number of variables even for low correlation among variables. The number of selected variables is more than double the number of truly significant ones. Yet so, the number of the selected variables is nearly constant for increasing levels of correlation. This is further supported by Figure 4c, which shows that the ratio between variables correctly selected and the total number of selected variables related to the correlation level generally remains the same. The number of wrongly selected variables shown in Figure 4d, is stable for increasing correlation levels and this trend is also reflected in the ratio between wrongly selected and all selected variables as shown in Figure 4e.

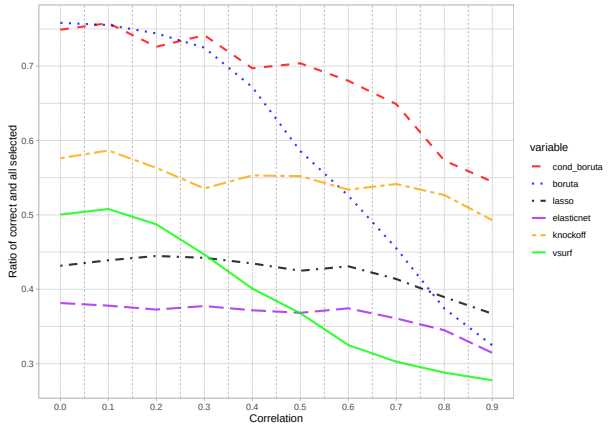
The Knockoffs variable selection method results are also presented in Figure 4. One of the key parameters of the model is the target false discovery rate (FDR), with a default value of 0.1. However, with this setting, only 0.5% of the models selected any variables, meaning that 99.5% of the models were empty and had selected 0 variables. Therefore, we had to increase the value of the FDR to 0.5 to ensure that at least one variable was selected in at least 90% of the models. We observed that the number of models with no terms increases proportionally with the correlation levels. In other words, only a few models contained no terms for low correlation levels, however for high correlation levels, an increasing number of models ended up with no terms. Therefore we conclude that, as the correlation between variables increases, the method may become less effective in identifying relevant variables and may result in an increasing number of cases with no terms being selected. Figure 4b shows that the Knockoffs model selects a constant number of variables. This is also reflected in the ratio between correctly selected and all selected



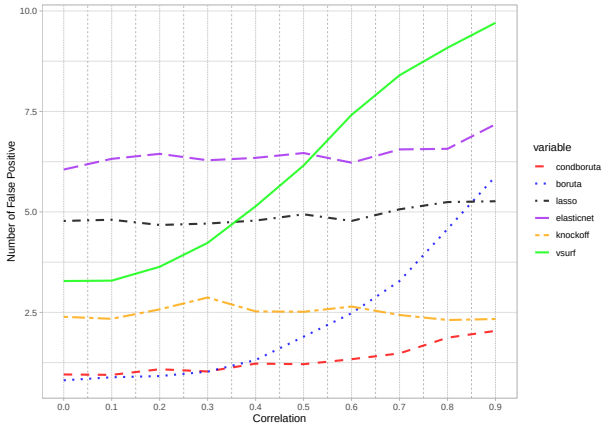
(a) Correctly selected variables.



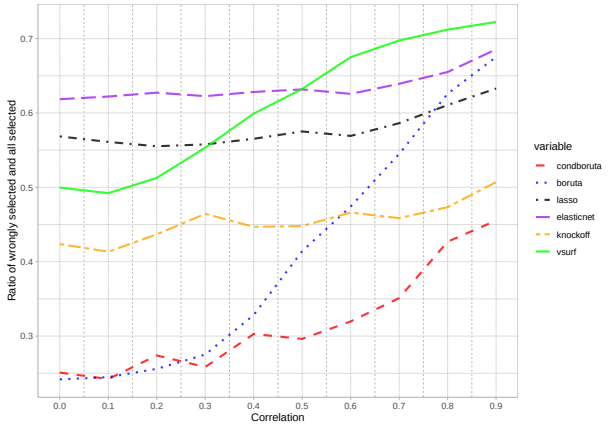
(b) All selected variables.



(c) Ratio between the correct selected and all selected variables.



(d) Falsely selected variables.



(e) Ratio between falsely and all selected variables.

Figure 4: (We present the results from only one of the simulated data sets for illustration.) (*a*): Number of correctly selected variables by six different methods vs the correlation among the variables. (*b*): Number of all selected variables vs the correlation among the variables. (*c*): The ratio between the correctly selected and all selected variables vs the correlation among the variables. (*d*): Number of falsely selected variables vs the correlation among the variables. (*e*): The ratio between falsely selected and all selected variables vs the correlation among the variables.

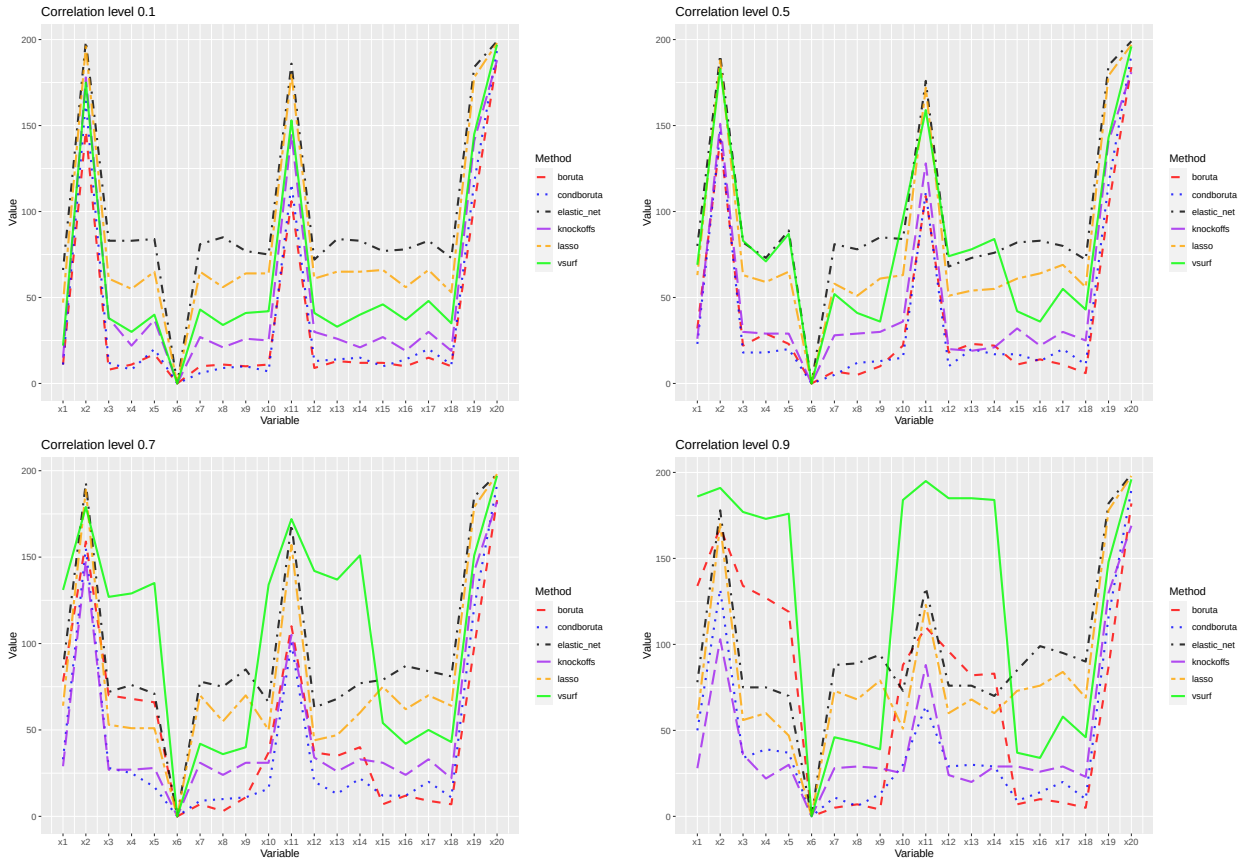


Figure 5: (We present the results from only one of the simulated data sets for illustration.) The frequency of the variables being selected in 200 runs for correlation levels: $cor = 0.1, 0.5, 0.7, 0.9$.

variables shown in Figure 4c, as well as the ratio between wrongly selected and all selected shown in Figure 4e. These ratios remain relatively constant but at a lower level than for the proposed model.

In Figure 4 we also present the results for the VSURF model. We notice that as the correlation increases, the number of selected variables rises drastically, negatively impacting the ratio in Figure 4c. As correlation increases, the number of incorrectly selected variables also increases, indicating a growing number of noise variables being selected. In Figure 4e, we can notice that more than half of the variables selected are noise variables.

We also observe a difference between the original Boruta algorithm and the proposed extension. On average, the extended Boruta algorithm correctly selects a greater number of significant variables. We can also see that as the correlation increases, the total number of variables selected by the original Boruta grows quickly. Instead, even in the case of high correlation, the proposed extension maintains a constant number of selected variables as shown in Figure 4b. This is reflected, also, in the ratio between correctly selected variables and all selected variables. The conditional Boruta selects a large number of significant variables while keeping the total number of variables selected at a low level. This is also supported by the ratio of correctly chosen variables to the total number of variables selected, shown in Figure 4c. Moreover, the proposed model exhibits a very low number of wrongly selected variables in Figure 4d, in fact the lowest among all the models studied in this work. The ratio between wrongly selected and all selected variables is also the lowest among all models. These findings suggest that the selected variables accurately reflect the variables used to construct Y, even in challenging scenarios with high correlation among variables for which other models tend to over select variables.

In Figure 5, we present the frequency of each variable being selected by different models. The input variables are displayed on the x-axis of each figure, and the y-axis represents the frequency with which each variable was selected by the algorithms across 200 iterations. We can see that all models, on average, pick the relevant variables correctly. However, even at low levels of correlation among variables, Lasso, Elastic Net and VSURF also select noise variables as significant. As the correlation increases, the original Boruta algorithm selects the entire group of correlated variables. In general, we also notice a difference in the frequency of selection of the two correlated variables $X2$ and $X11$. The variable $X11$ is selected less frequently than the variable $X2$, this is due to a lower coefficient used in (1). The same happens for the uncorrelated variables $X19$ and $X20$, for which the former is not selected in all models having a lower coefficient.

The Boruta algorithm, at each iteration, stores the variable importance value for the original variables and averages them at the end of the algorithm in a final table. The extension we propose in this paper gives us a more accurate selection of variables and a more representative final table. Further analysis could be carried out by exploiting the ranking of the variable's importance values in the final table, particularly the variables close to the Boruta threshold. This possibility can be pursued if Boruta's outcome is not entirely satisfactory. We can increase the number of variables or, on the contrary, be more restrictive in our decision by utilizing subject matter knowledge.

References

[1] R. F. Barber, E. J. Candès, Controlling the false discovery rate via knockoffs, The Annals of Statistics 43 (5) (2015) 2055 – 2085. doi:10.1214/15-AOS1337.
URL https://doi.org/10.1214/15-AOS1337

[2] E. J. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection, Vol. 1610, Department of Statistics, Stanford University Stanford, CA, USA, 2016.

[3] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern recognition letters 31 (14) (2010) 2225–2236.

[4] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Vsurf: an r package for variable selection using random forests, The R Journal 7 (2) (2015) 19–33.

[5] M. Rotari, Conditional-boruta, GitHub repository (2021).
URL https://github.com/MartaRotari/Conditional-Boruta

[6] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC bioinformatics 9 (1) (2008) 1–11.

[7] D. Debeer, C. Strobl, Conditional permutation importance revisited, BMC bioinformatics 21 (1) (2020) 1–30.