Estimating Variance of Simple Defined Variable Main and Low-Order Interaction Effects

Felix Kapulla

```
knitr::opts_chunk$set(fig.width=15, fig.height=8)
```

```
rm(list=ls())
library(Matrix)
library(tidyverse)
library(ggplot2)
library(ggpubr)
library(ranger)
library(MixMatrix)
library(mvtnorm)
library(stringr)
library(parallel)

cores <- detectCores()
clust <- makeCluster(6)

source('C:/Users/feix_/iCloudDrive/Studium Master/CQM - Thesis Internship/Thesis-VariableEffects/Baseli

parallel::clusterEvalQ(clust,
                       expr = {source('C:/Users/feix_/iCloudDrive/Studium Master/CQM - Thesis Internshi
```

Simulation

```
n <- c(40, 400, 4000) ; num.trees <- 2000 ; repeats <- 4; cor <- c(0, 0.8)
k <- c(0.2, 1); node_size <- c(1); pdp <- F; ale <- F
formulas <- c("2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4")
longest_latex_formula <- "2x_1+4x_2-3x_3+2.2x_4-x_3x_4"





#parallel::clusterExport(cl = clust, varlist = 'formulas')
scenarios <- data.frame(expand.grid(n, num.trees, formulas, repeats,
cor, k, node_size, pdp, ale))
colnames(scenarios) = c("N", "N_Trees", "Formula", "Repeats",
"Correlation", "k", "Node_Size", "pdp", "ale")
scenarios$k_idx <- (scenarios$k == unique(scenarios$k)[1])
scenarios[,"Formula"] <- as.character(scenarios[,"Formula"]) ### Formula became Factor
```

```
scenarios["Longest_Latex_formula"] <- longest_latex_formula
scenarios <- split(scenarios, seq(nrow(scenarios)))
#Run Simulation
system.time(result <- parLapply(cl = clust,
                                X = scenarios,
                                fun = sim_multi))
```

```
##    user  system elapsed
##    0.04    0.08  300.11
```

```
if (!pdp | !ale) {
 print_results(result)
}
```

```
## Setting 1: N = 40 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   0.692266 4.289887 -1.710529 0.7724662 -0.2121708
## Mean(s) of simulated LM Variable Effect(s):
##   2.001843 4.133343 -2.886753 2.219207 -0.9122996
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.7046328 1.728856 0.911333 0.6966893 0.1777904 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.236153 4.064989 1.797715 1.367491 0.6856083 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 2: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.472971 3.158968 -2.478718 2.401807 -0.8137937
## Mean(s) of simulated LM Variable Effect(s):
##   2.01567 3.985667 -2.976988 2.212592 -1.022138
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.7156143 1.192769 0.4281745 1.638789 1.387805 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.140364 1.699125 1.269949 1.357359 1.673076 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 3: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   2.057604 4.501444 -4.166826 1.96737 -2.662162
## Mean(s) of simulated LM Variable Effect(s):
##   2.002612 3.989011 -3.018654 2.197719 -1.01272
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
```

```
##   0.5765436 0.1863343 0.7286306 0.3641241 0.35162 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.8696853 1.017663 1.09981 0.4995938 3.042452 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 4: N = 40 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   0.5711579 3.731085 -0.3945913 1.31787 -0.2492237
## Mean(s) of simulated LM Variable Effect(s):
##   2.142268 3.734094 -2.858673 2.192911 -1.037049
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.9255343 2.630523 0.2693725 1.651451 0.1876233 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.231152 3.17737 0.3914101 1.575641 0.4857525 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 5: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.334911 3.72262 -2.534601 2.008005 0.33254
## Mean(s) of simulated LM Variable Effect(s):
##   2.003483 4.01181 -2.971746 2.189866 -0.8887123
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.6600891 1.012388 0.7364256 1.049385 1.078058 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.7738023 1.689333 0.8400468 0.9268083 0.4616386 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 6: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.422388 3.932398 -1.809773 2.619883 -0.5571645
## Mean(s) of simulated LM Variable Effect(s):
##   1.995831 3.98825 -2.992251 2.206607 -0.9726855
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   1.061917 1.48394 0.4972909 0.4231503 0.9749694 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.4750646 0.9628195 0.9924715 1.304933 2.856954 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 7: N = 40 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
```

```
##   1.348024 3.416792 -2.327757 0.7085222 -0.1562627
## Mean(s) of simulated LM Variable Effect(s):
##   1.913912 4.001763 -3.141249 2.151034 -1.020389
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.1600315 0.8182388 0.6852166 0.3293763 0.1541648 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.618551 0.8257999 0.6573426 0.4493547 0.2058337 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 8: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.429076 3.827246 -2.665673 2.023188 -0.9781332
## Mean(s) of simulated LM Variable Effect(s):
##   2.000324 4.000196 -3.017156 2.324219 -1.098443
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.253597 0.3803918 0.6282837 0.08053834 0.2558342 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3909373 0.5389392 0.5192057 0.3936789 0.224908 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 9: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.888578 4.010638 -2.958504 2.04854 -1.147339
## Mean(s) of simulated LM Variable Effect(s):
##   2.015524 4.002383 -3.005271 2.19903 -1.016528
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.08578729 0.2282935 0.1710634 0.1972844 0.2016033 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.06997343 0.2380533 0.3179004 0.1513601 0.2184951 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 10: N = 40 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.182486 2.514306 -0.2341016 1.317615 -0.0536546
## Mean(s) of simulated LM Variable Effect(s):
##   2.121273 4.053837 -3.001313 1.801601 -1.343464
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.7673413 0.6848835 0.1909346 0.2953787 0.0772001 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.6830223 0.6462718 0.2932953 0.6232892 0.1424712 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 11: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.733223 3.545522 -1.414955 1.439592 -0.3380721
## Mean(s) of simulated LM Variable Effect(s):
##   2.040582 4.038513 -3.010549 2.161018 -0.9257854
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.09463615 0.4659685 0.2711737 0.1834662 0.1389055 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3748434 0.4411671 0.2581761 0.3410316 0.138304 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 12: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4-x.3*x.4
## Mean(s) of simulated RF Variable Effect(s):
##   1.735404 3.816383 -2.243384 1.823509 -0.7293243
## Mean(s) of simulated LM Variable Effect(s):
##   2.032768 3.997636 -3.015051 2.179289 -0.9825022
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.2971016 0.1912254 0.2730666 0.2830091 0.07427923 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.07037594 0.2982607 0.1734664 0.1789616 0.2234698 .
## Number of Smaller Nulls:
##   0 0 0 0
```

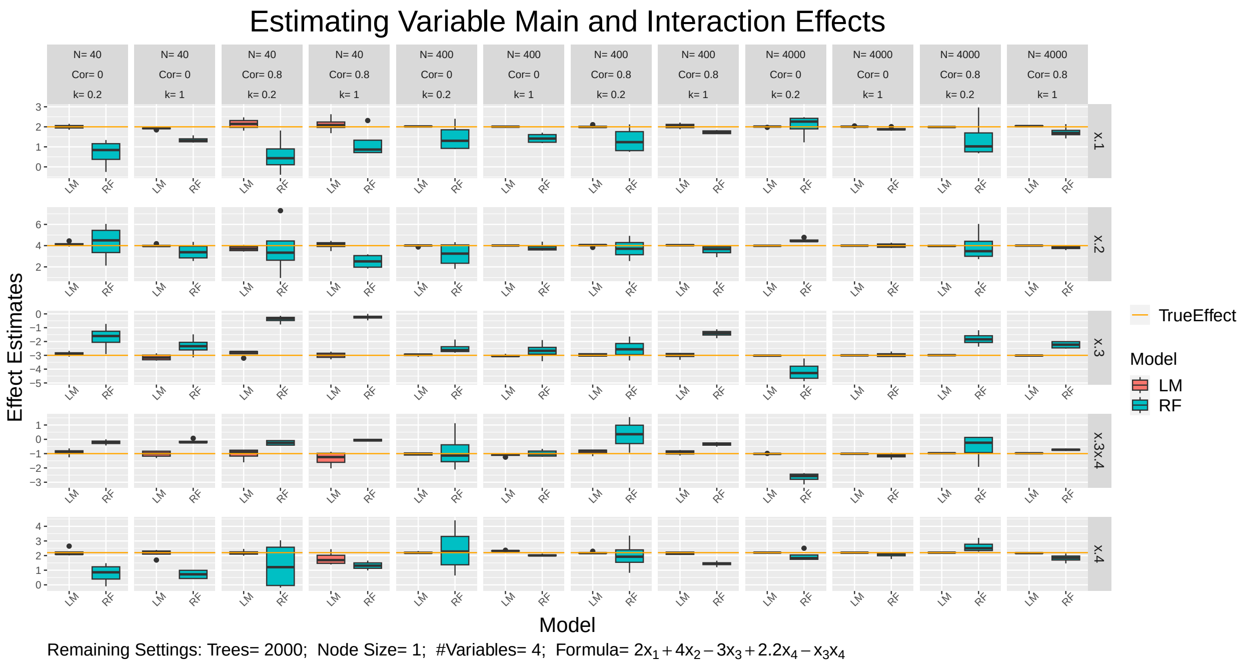
```
effect_plots <- plot_effects(result)
```

```
## ‘summarise()‘ has grouped output by ’N’, ’cor’, ’k’, ’num.trees’, ’node_size’,
## ’variable’. You can override using the ‘.groups‘ argument.
```

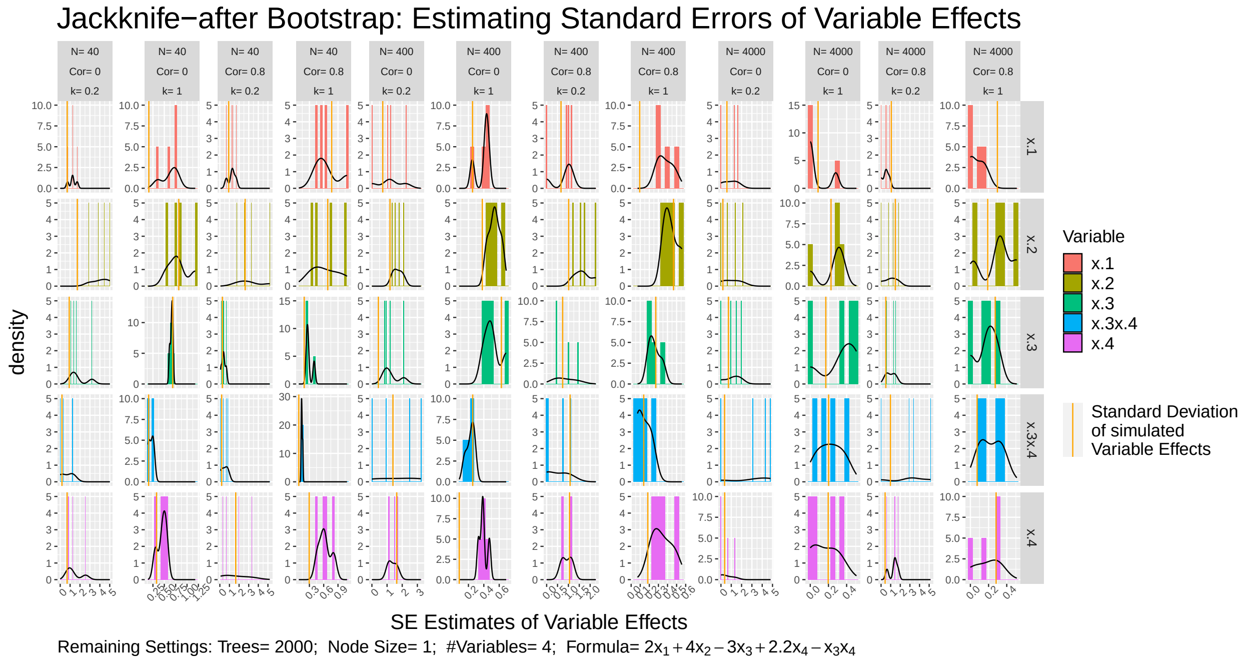
```
se_plot <- plot_se(result)
```

```
## ‘summarise()‘ has grouped output by ’N’, ’cor’, ’k’, ’num.trees’, ’node_size’,
## ’variable’. You can override using the ‘.groups‘ argument.
```

```
effect_plots
```


Estimating Variable Main and Interaction Effects
N= 40 Cor= 0 k= 0.2
N= 40 Cor= 0 k= 1
N= 40 Cor= 0.8 k= 0.2
N= 40 Cor= 0.8 k= 1
N= 400 Cor= 0 k= 0.2
N= 400 Cor= 0 k= 1
N= 400 Cor= 0.8 k= 0.2
N= 400 Cor= 0.8 k= 1
N= 4000 Cor= 0 k= 0.2
N= 4000 Cor= 0 k= 1
N= 4000 Cor= 0.8 k= 0.2
N= 4000 Cor= 0.8 k= 1
x.1
x.2
x.3
x.3x.4
x.4
LM
RF
Effect Estimates
Model
TrueEffect
Model
LM
RF
Remaining Settings: Trees= 2000; Node Size= 1; #Variables= 4; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - x_3x_4$

```
se_plot
```


Jackknife−after Bootstrap: Estimating Standard Errors of Variable Effects
N= 40 Cor= 0 k= 0.2
N= 40 Cor= 0 k= 1
N= 40 Cor= 0.8 k= 0.2
N= 40 Cor= 0.8 k= 1
N= 400 Cor= 0 k= 0.2
N= 400 Cor= 0 k= 1
N= 400 Cor= 0.8 k= 0.2
N= 400 Cor= 0.8 k= 1
N= 4000 Cor= 0 k= 0.2
N= 4000 Cor= 0 k= 1
N= 4000 Cor= 0.8 k= 0.2
N= 4000 Cor= 0.8 k= 1
x.1
x.2
x.3
x.3x.4
x.4
density
SE Estimates of Variable Effects
Variable
x.1
x.2
x.3
x.3x.4
x.4
Standard Deviation of simulated Variable Effects
Remaining Settings: Trees= 2000; Node Size= 1; #Variables= 4; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - x_3x_4$