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

Felix Kapulla

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

```
library(Matrix)
library(tidyverse)
library(ggplot2)
library(ggpubr)
library(ranger)
library(MixMatrix)
library(mvtnorm)
library(stringr)
library(parallel)

cores <- detectCores()
clust <- makeCluster(4)

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

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

Simulation

```
n <- c(40, 400, 4000) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
k <- c(0.2, 1); node_size <- c(1); pdp <- F; ale <- F
formulas <- c("x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2")

longest_latex_formula <- "x_1-2x_2^3+3e^{x_3}x_3-4(|x_4|>0.5)-2(x_3x_4)^2"



#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
##     1.67     1.89 13778.18
```

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

```
## Setting 1: N = 40 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.6641673 -0.701099 2.236787 -0.04655147 -0.0714578
## Mean(s) of simulated LM Variable Effect(s):
##   1.227987 -5.330202 8.886028 -0.08193873 -0.1964402
## True Variable Effect(s):
##   1 -0.08 3.0602 0 0
## Standard Error of simulated Variable Effects (RF):
##   1.922442 1.837782 2.530255 1.564437 1.168513 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.929173 1.877734 2.477782 1.625651 13.0651 .
## Number of Smaller Nulls:
##   3 2 0 4 0
##
## Setting 2: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.796454 -0.2753487 2.757118 0.0420498 0.02831961
## Mean(s) of simulated LM Variable Effect(s):
##   1.1242 -5.899072 9.394501 -0.05052571 -0.02465182
## True Variable Effect(s):
##   1 -0.08 3.0602 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.7543726 1.918664 1.275783 0.5769247 1.281978 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.7571048 0.7674196 1.299574 0.5885258 4.376032 .
## Number of Smaller Nulls:
##   25 22 4 21 0
##
## Setting 3: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.9959953 -0.1951789 3.147599 0.01555481 -0.07550225
## Mean(s) of simulated LM Variable Effect(s):
##   0.9971692 -5.990957 9.946521 0.03837723 0.08341096
## True Variable Effect(s):
##   1 -0.08 3.0602 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.849639 0.5404156 1.027319 0.4849181 1.679668 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.8739311 0.5692929 1.128943 0.5672807 3.612906 .
## Number of Smaller Nulls:
##   27 45 22 47 3
##
## Setting 4: N = 40 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.691709 -0.1151238 1.71005 0.1268125 -0.03627547
## Mean(s) of simulated LM Variable Effect(s):
##   1.147414 -4.738649 7.898808 0.02101721 -4.174663
## True Variable Effect(s):
##   1 -0.08 3.0602 0 0
## Standard Error of simulated Variable Effects (RF):
##   1.320489 1.04533 1.640295 1.362274 0.8118563 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.494659 1.174087 1.857536 1.289841 8.174283 .
## Number of Smaller Nulls:
##   0 0 0 0 0
##
## Setting 5: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.8753285 -0.1492494 2.700842 -0.003867327 -0.3358031
## Mean(s) of simulated LM Variable Effect(s):
##   1.013864 -5.883591 9.504606 0.06980629 -4.167608
## True Variable Effect(s):
##   1 -0.08 3.0602 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.8622485 0.5344919 1.200152 0.6703421 1.574143 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.923628 0.5817427 1.277777 0.7015999 3.901522 .
## Number of Smaller Nulls:
##   0 10 0 3 0
##
## Setting 6: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.9770652 -0.09680271 3.005015 0.01041914 -0.02575886
## Mean(s) of simulated LM Variable Effect(s):
##   1.005392 -6.016701 9.904372 -0.01730044 -4.615506
## True Variable Effect(s):
##   1 -0.08 3.0602 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.9890788 0.5320957 1.028377 0.6329077 2.008042 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.9127944 0.5784166 1.115158 0.6677385 3.453546 .
## Number of Smaller Nulls:
##   24 38 12 30 1
##
## Setting 7: N = 40 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.6140983 -1.863812 3.438514 -0.01623455 -0.02880704
## Mean(s) of simulated LM Variable Effect(s):
##   0.9454708 -5.035516 7.911423 -0.1307043 -0.04282408
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.7449266 1.089576 1.580844 0.9576219 0.3664355 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.8420074 1.226042 1.494358 0.9494844 1.087974 .
## Number of Smaller Nulls:
##   0 0 0 0 0
##
## Setting 8: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.8313656 -1.855977 3.982152 0.02628312 -0.02099948
## Mean(s) of simulated LM Variable Effect(s):
##   1.038198 -5.837441 9.789033 -0.007135302 0.1163802
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.2755989 0.4964508 0.6452502 0.3470272 0.3118002 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3089542 0.4545669 0.6106273 0.3424352 0.5267949 .
## Number of Smaller Nulls:
##   14 5 1 19 0
##
## Setting 9: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.9752444 -1.888314 4.501522 -0.02654018 0.02575786
## Mean(s) of simulated LM Variable Effect(s):
##   1.033162 -6.013907 9.911478 0.008793702 0.02360648
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.2473121 0.2649316 0.30097 0.2309091 0.2011303 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2361749 0.2611885 0.2598161 0.2502038 0.2700422 .
## Number of Smaller Nulls:
##   22 34 26 28 16
##
## Setting 10: N = 40 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.7938029 -0.2571195 2.385257 0.7682875 -0.1134458
## Mean(s) of simulated LM Variable Effect(s):
##   0.6527543 -4.632059 8.220022 0.2919347 -3.537226
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.7424401 0.5678043 0.9321887 0.7615806 0.2227306 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.7408824 0.6788823 1.064485 0.8294509 0.7095703 .
## Number of Smaller Nulls:
```

```
##   0 0 0 0 0
##
## Setting 11: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.7820516 -1.102668 3.796964 0.1311253 0.008818658
## Mean(s) of simulated LM Variable Effect(s):
##   0.9859096 -5.851488 9.660513 0.04618228 -4.181757
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.2707225 0.3569481 0.5845062 0.3321183 0.3252757 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3010676 0.3553125 0.5278203 0.3521334 0.4350818 .
## Number of Smaller Nulls:
##   0 3 1 1 0
##
## Setting 12: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2*(x.3*x.4)^2
## Mean(s) of simulated RF Variable Effect(s):
##   0.9160171 -1.627056 4.46406 -0.01016371 -0.1888156
## Mean(s) of simulated LM Variable Effect(s):
##   1.009256 -6.02657 9.827769 0.03408624 -4.760048
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.237418 0.2602983 0.2601699 0.2194287 0.2295502 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2423878 0.2699304 0.273734 0.2456212 0.3379412 .
## Number of Smaller Nulls:
##   8 26 18 18 8
```

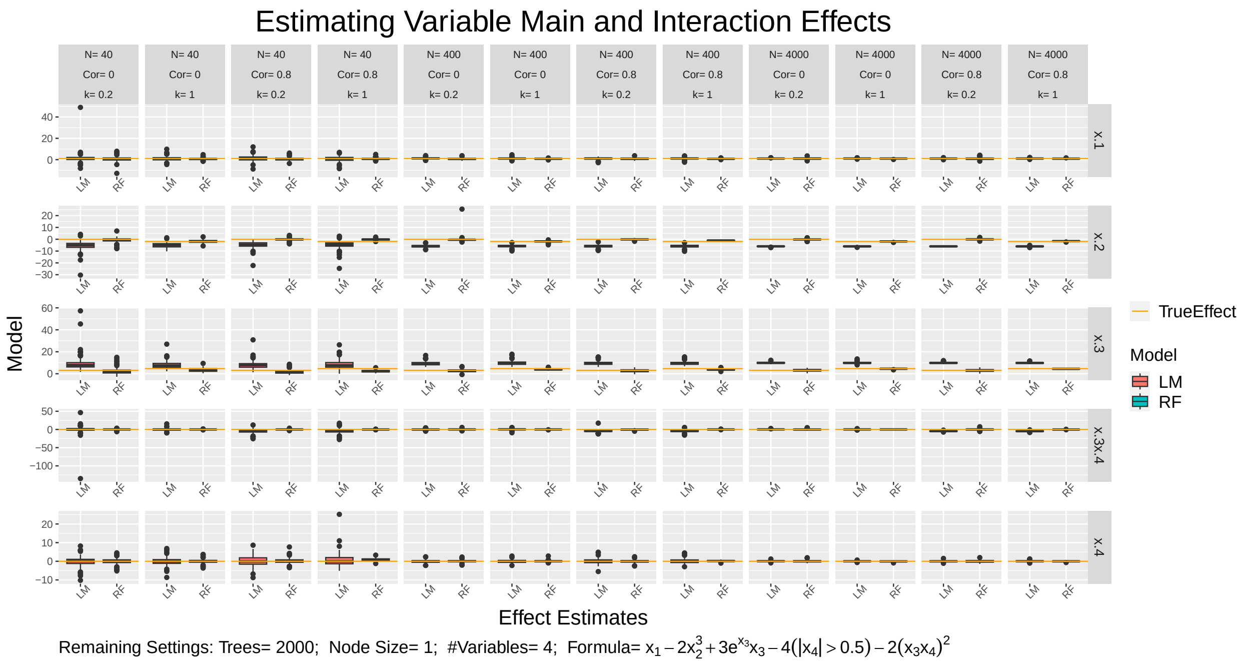
```
effect_plots <- plot_effects(result)
```

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

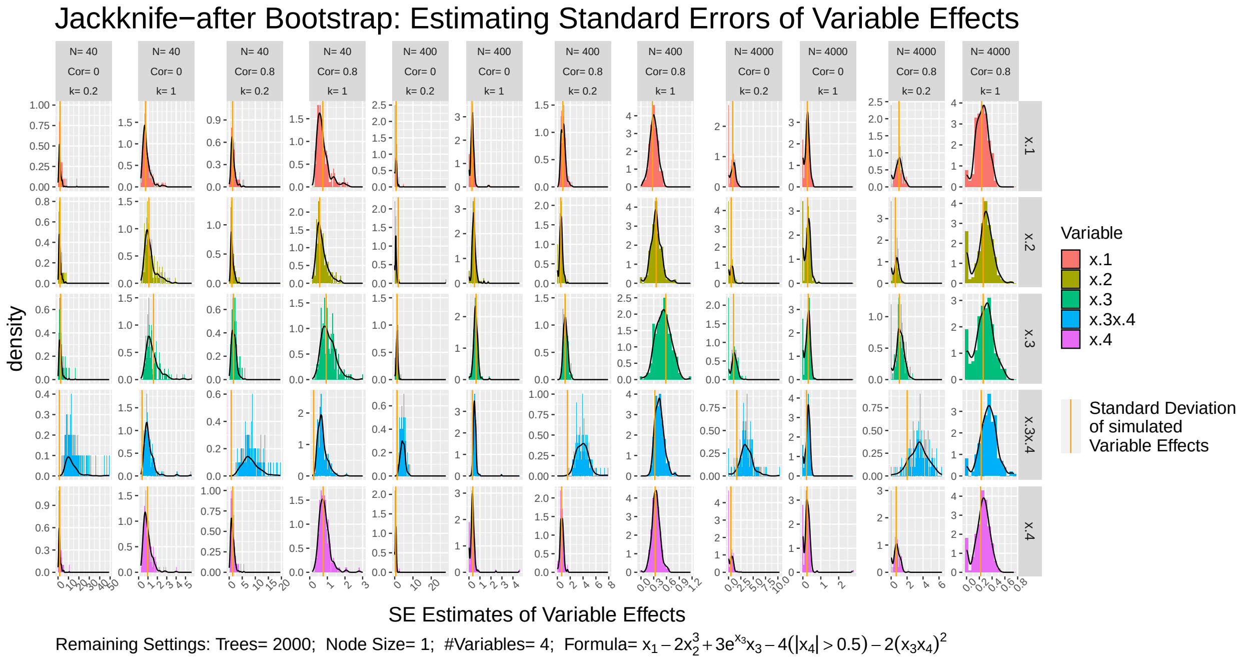
```
se_plot <- plot_se(result)
```

```
## ‘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
Model
TrueEffect
Model
LM
RF
Effect Estimates
Remaining Settings: Trees= 2000; Node Size= 1; #Variables= 4; Formula= $x_1 - 2x_2^3 + 3e^{x_3}x_3 - 4(|x_4| > 0.5) - 2(x_3x_4)^2$

```
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
Variable
x.1
x.2
x.3
x.3x.4
x.4
Standard Deviation of simulated Variable Effects
SE Estimates of Variable Effects
Remaining Settings: Trees= 2000; Node Size= 1; #Variables= 4; Formula= $x_1 - 2x_2^3 + 3e^{x_3}x_3 - 4(|x_4| > 0.5) - 2(x_3x_4)^2$