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/Baseli

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

Simulation

```
n <- c(400) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
k <- c(1); node_size <- c(1); pdp <- F; ale <- F
formulas <- c("2*x.1",
              "2*x.1+4*x.2",
              "2*x.1+4*x.2-3*x.3",
              "2*x.1+4*x.2-3*x.3+2.2*x.4",
              "2*x.1+4*x.2-3*x.3+2.2*x.4-1.5*x.5")

longest_latex_formula <- "2x_1+4x_2-3x_3+2.2x_4-1.5x_5"



#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")
```

```r
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
##    0.18    0.58 1425.58
```

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

```
## Setting 1: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1
## Mean(s) of simulated RF Variable Effect(s):
##   1.958066
## Mean(s) of simulated LM Variable Effect(s):
##   1.99808
## True Variable Effect(s):
##   2
## Standard Error of simulated Variable Effects (RF):
##   0.4853944 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5286185 .
## Number of Smaller Nulls:
##   0
##
## Setting 2: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2
## Mean(s) of simulated RF Variable Effect(s):
##   1.960933 3.964437
## Mean(s) of simulated LM Variable Effect(s):
##   2.002968 3.99832
## True Variable Effect(s):
##   2 4
## Standard Error of simulated Variable Effects (RF):
##   0.3305576 0.3321162 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3702053 0.3582485 .
## Number of Smaller Nulls:
##   0 0
##
## Setting 3: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3
## Mean(s) of simulated RF Variable Effect(s):
##   1.733368 3.740256 -2.726029
## Mean(s) of simulated LM Variable Effect(s):
##   1.996731 4.003024 -2.998691
## True Variable Effect(s):
##   2 4 -3
## Standard Error of simulated Variable Effects (RF):
##   0.312126 0.3381907 0.3413503 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.335785 0.3612116 0.3664631 .
## Number of Smaller Nulls:
##   8 11 4
##
## Setting 4: 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
## Mean(s) of simulated RF Variable Effect(s):
##   1.65273 3.756933 -2.736457 1.824704
## Mean(s) of simulated LM Variable Effect(s):
##   2.000989 3.998044 -2.993299 2.194985
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.292794 0.4055282 0.3810014 0.3211969 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3329405 0.4713182 0.4180504 0.3566684 .
## Number of Smaller Nulls:
##   8 0 2 6
##
## Setting 5: 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.506032 3.762602 -2.667698 1.746653 -0.9751166
## Mean(s) of simulated LM Variable Effect(s):
##   1.995371 4.006345 -2.997803 2.202154 -1.50225
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.2868668 0.4183913 0.3735371 0.3529272 0.2388462 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3141098 0.4613993 0.4108228 0.3477084 0.2430739 .
## Number of Smaller Nulls:
##   20 3 6 11 30
##
## Setting 6: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1
## Mean(s) of simulated RF Variable Effect(s):
##   2.010354
## Mean(s) of simulated LM Variable Effect(s):
##   2.000851
## True Variable Effect(s):
##   2
## Standard Error of simulated Variable Effects (RF):
##   0.4707787 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5427053 .
## Number of Smaller Nulls:
##   0
##
```

```
## Setting 7: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2
## Mean(s) of simulated RF Variable Effect(s):
##   2.052439 3.840571
## Mean(s) of simulated LM Variable Effect(s):
##   2.009404 3.997915
## True Variable Effect(s):
##   2 4
## Standard Error of simulated Variable Effects (RF):
##   0.329436 0.3416585 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3443271 0.3656753 .
## Number of Smaller Nulls:
##   1 0
##
## Setting 8: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3
## Mean(s) of simulated RF Variable Effect(s):
##   1.613955 3.103939 -1.816362
## Mean(s) of simulated LM Variable Effect(s):
##   2.001382 4.001561 -2.997004
## True Variable Effect(s):
##   2 4 -3
## Standard Error of simulated Variable Effects (RF):
##   0.279251 0.2950932 0.3057046 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3194457 0.3094639 0.3055222 .
## Number of Smaller Nulls:
##   2 10 5
##
## Setting 9: 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
## Mean(s) of simulated RF Variable Effect(s):
##   1.414728 3.368828 -1.434518 1.594788
## Mean(s) of simulated LM Variable Effect(s):
##   1.997398 3.995614 -2.979761 2.190104
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.3031074 0.3995199 0.273345 0.3049198 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3555216 0.4046906 0.2924424 0.3533295 .
## Number of Smaller Nulls:
##   0 0 2 0
##
## Setting 10: 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.258212 3.06722 -1.512062 1.408804 -0.6392068
## Mean(s) of simulated LM Variable Effect(s):
##   1.995062 4.008722 -3.013593 2.207166 -1.495582
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
```

```
##   0.2942092 0.3550284 0.2528728 0.3054132 0.2071744 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3016599 0.3866885 0.2808636 0.3105473 0.2078096 .
## Number of Smaller Nulls:
##   11 2 9 7 23
```

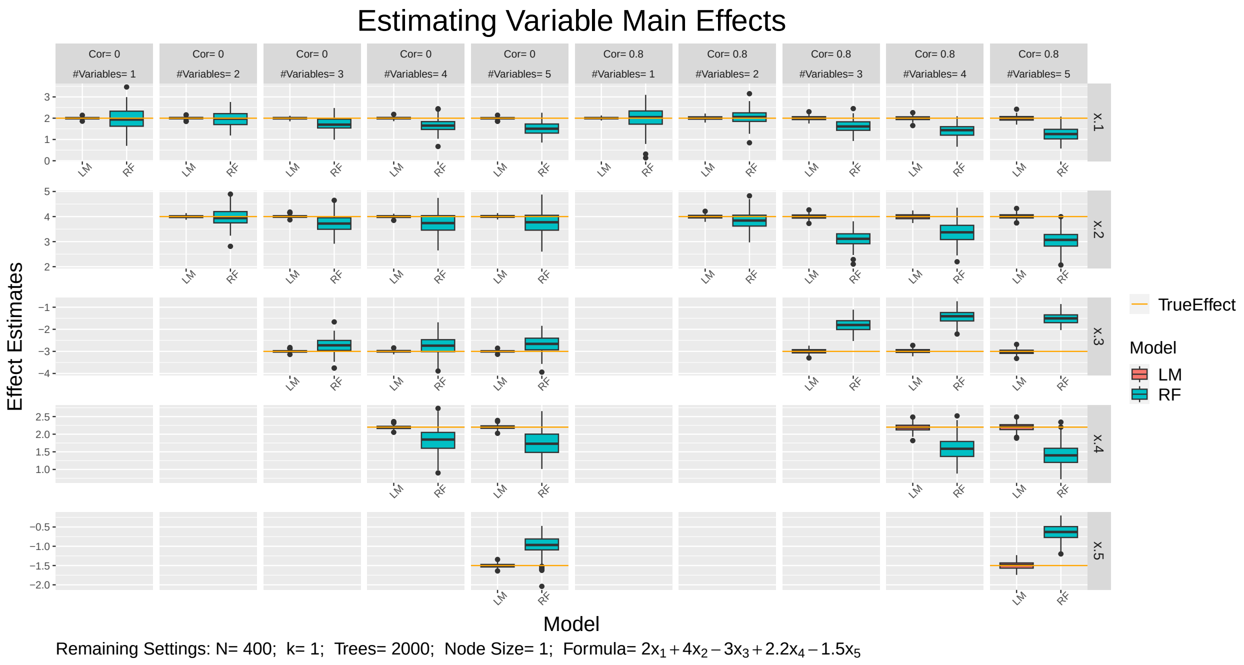
```
effect_plots <- plot_effects(result)
```

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

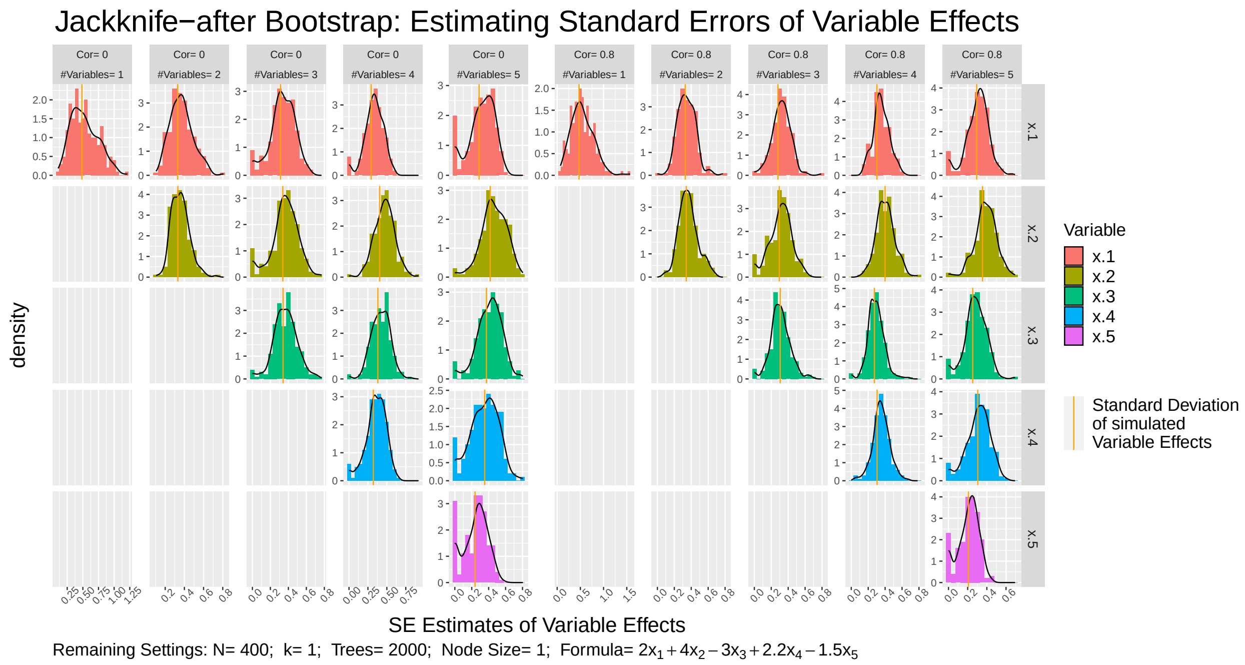
```
se_plot <- plot_se(result)
```

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

```
effect_plots
```



```
se_plot
```



Jackknife−after Bootstrap: Estimating Standard Errors of Variable Effects

Remaining Settings: N= 400; k= 1; Trees= 2000; Node Size= 1; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - 1.5x_5$

```r
n <- c(400) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
k <- c(1); node_size <- c(1); pdp <- F; ale <- F
formulas <- c("2*x.1",
              "2*x.1+4*x.2",
              "2*x.1+4*x.2-3*x.3",
              "2*x.1+4*x.2-3*x.3+2.2*x.4",
              "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
##    0.13    0.22 1149.23
```

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

```
## Setting 1: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1
## Mean(s) of simulated RF Variable Effect(s):
##   2.015129
## Mean(s) of simulated LM Variable Effect(s):
##   2.001201
## True Variable Effect(s):
##   2
## Standard Error of simulated Variable Effects (RF):
##   0.441608 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5496615 .
## Number of Smaller Nulls:
##   0
##
## Setting 2: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2
## Mean(s) of simulated RF Variable Effect(s):
##   1.937705 3.969382
## Mean(s) of simulated LM Variable Effect(s):
##   1.998388 3.999891
## True Variable Effect(s):
##   2 4
## Standard Error of simulated Variable Effects (RF):
##   0.3465069 0.3381912 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3765805 0.3812959 .
## Number of Smaller Nulls:
##   1 1
##
## Setting 3: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3
## Mean(s) of simulated RF Variable Effect(s):
##   1.783631 3.710747 -2.744414
## Mean(s) of simulated LM Variable Effect(s):
##   1.997837 4.007287 -3.003327
## True Variable Effect(s):
##   2 4 -3
## Standard Error of simulated Variable Effects (RF):
##   0.3321617 0.3108669 0.2956067 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3408379 0.3548829 0.350729 .
## Number of Smaller Nulls:
##   8 7 11
##
## Setting 4: 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
## Mean(s) of simulated RF Variable Effect(s):
##   1.596168 3.863291 -2.744103 1.862367
## Mean(s) of simulated LM Variable Effect(s):
##   1.995832 4.003421 -3.006802 2.196196
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2981809 0.4410149 0.3735673 0.3006591 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3430221 0.4758604 0.4166576 0.3632945 .
## Number of Smaller Nulls:
##   8 0 4 4
##
## Setting 5: 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):
##   1.573052 3.752885 -2.768455 1.845894 -0.8425802
## Mean(s) of simulated LM Variable Effect(s):
##   1.996731 4.000469 -2.999807 2.203286 -1.003665
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.2588735 0.3668176 0.4126275 0.3433345 0.2336144 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3422475 0.4676634 0.4296549 0.3602987 0.3770308 .
## Number of Smaller Nulls:
```

```
##   7 0 2 4 0
##
## Setting 6: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1
## Mean(s) of simulated RF Variable Effect(s):
##   1.972329
## Mean(s) of simulated LM Variable Effect(s):
##   2.000231
## True Variable Effect(s):
##   2
## Standard Error of simulated Variable Effects (RF):
##   0.5025023 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5320276 .
## Number of Smaller Nulls:
##   0
##
## Setting 7: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2
## Mean(s) of simulated RF Variable Effect(s):
##   2.05141 3.820282
## Mean(s) of simulated LM Variable Effect(s):
##   2.000915 3.995909
## True Variable Effect(s):
##   2 4
## Standard Error of simulated Variable Effects (RF):
##   0.2990789 0.3271407 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3501693 0.3630088 .
## Number of Smaller Nulls:
##   1 1
##
## Setting 8: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = 2*x.1+4*x.2-3*x.3
## Mean(s) of simulated RF Variable Effect(s):
##   1.585079 3.129259 -1.856186
## Mean(s) of simulated LM Variable Effect(s):
##   2.006671 4.004988 -3.014648
## True Variable Effect(s):
##   2 4 -3
## Standard Error of simulated Variable Effects (RF):
##   0.3036867 0.3036286 0.29742 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3204246 0.3296385 0.3041868 .
## Number of Smaller Nulls:
##   4 8 6
##
## Setting 9: 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
## Mean(s) of simulated RF Variable Effect(s):
##   1.43103 3.380769 -1.427979 1.597691
## Mean(s) of simulated LM Variable Effect(s):
##   2.015162 3.999762 -3.001355 2.19452
## True Variable Effect(s):
```

```
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.314795 0.3904439 0.2502286 0.3485577 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3572936 0.400769 0.2915544 0.3639874 .
## Number of Smaller Nulls:
##   0 0 2 1
##
## Setting 10: 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):
##   1.488124 3.411476 -1.505982 1.661854 -0.4095566
## Mean(s) of simulated LM Variable Effect(s):
##   2.003622 4.001579 -3.01719 2.206374 -1.013428
## True Variable Effect(s):
##   2 4 -3 2.2 -1
## Standard Error of simulated Variable Effects (RF):
##   0.3433223 0.4025713 0.3245275 0.3330436 0.1895165 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3530295 0.4284122 0.3017194 0.3526983 0.2832182 .
## Number of Smaller Nulls:
##   1 1 1 1 0
```

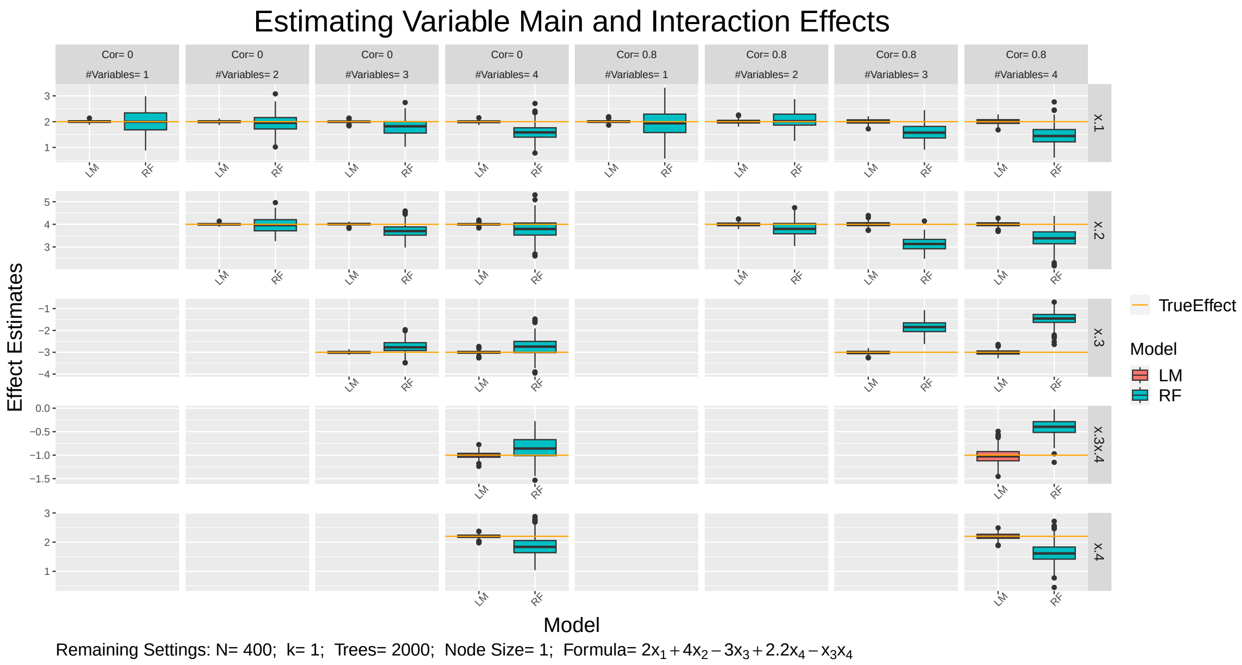
```
effect_plots <- plot_effects(result)
```

```
## ‘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.
```

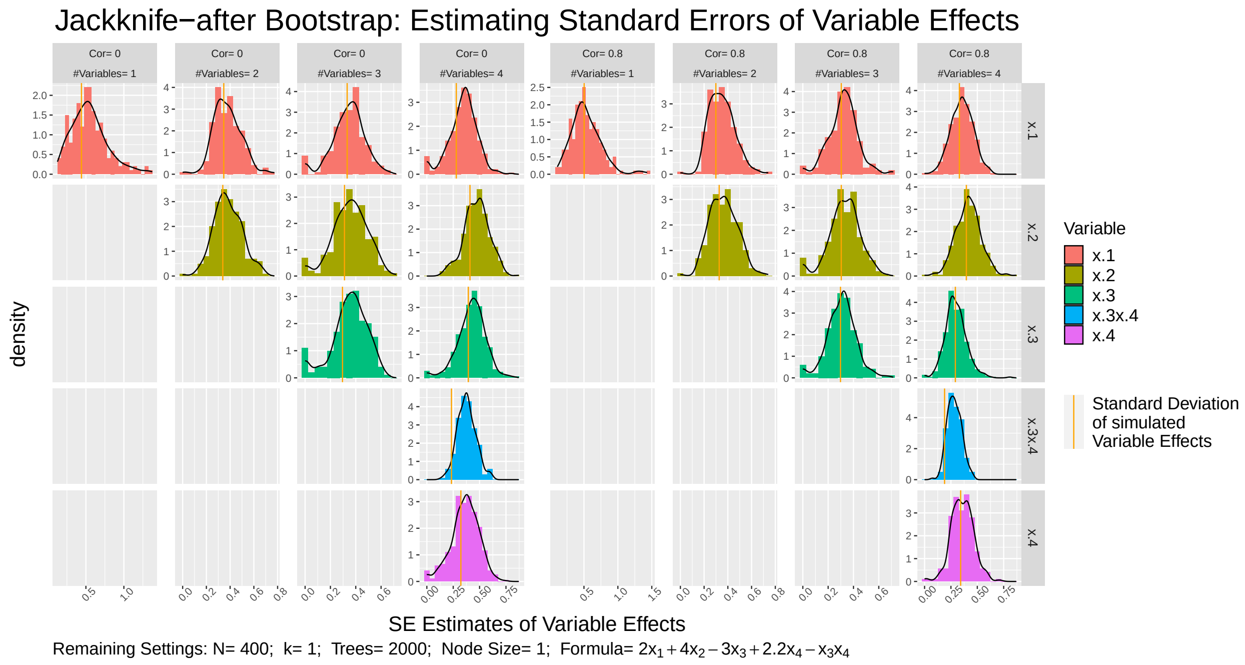
```
effect_plots
```



Estimating Variable Main and Interaction Effects

Remaining Settings: N= 400; k= 1; Trees= 2000; Node Size= 1; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - x_3x_4$

```
se_plot
```



Jackknife−after Bootstrap: Estimating Standard Errors of Variable Effects

Remaining Settings: N= 400; k= 1; Trees= 2000; Node Size= 1; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - x_3x_4$

```r
n <- c(400) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
k <- c(1); node_size <- c(1); pdp <- F; ale <- F
formulas <- c("x.1",
              "x.1-2*x.2^3",
              "x.1-2*x.2^3+3*exp(x.3)*x.3",
              "x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)",
              "x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)-2^x.5")

longest_latex_formula <- "x_1-2x_2^3+3e^{x_3}x_3-4(|x_4|>0.5)-2^{x_5}"




#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
##    0.11    0.21 1336.09
```

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

```
## Setting 1: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1
## Mean(s) of simulated RF Variable Effect(s):
##   0.9599868
## Mean(s) of simulated LM Variable Effect(s):
##   0.9998462
## True Variable Effect(s):
##   1
## Standard Error of simulated Variable Effects (RF):
##   0.4724047 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5349497 .
## Number of Smaller Nulls:
##   0
##
## Setting 2: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3
## Mean(s) of simulated RF Variable Effect(s):
```

```
##   0.953241 -1.861642
## Mean(s) of simulated LM Variable Effect(s):
##   0.9827743 -5.916559
## True Variable Effect(s):
##   1 -2
## Standard Error of simulated Variable Effects (RF):
##   0.4589376 0.3574527 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.4280032 0.3462527 .
## Number of Smaller Nulls:
##   0 1
##
## Setting 3: 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
## Mean(s) of simulated RF Variable Effect(s):
##   0.9911595 -1.751621 4.283762
## Mean(s) of simulated LM Variable Effect(s):
##   0.9769982 -5.890226 9.524896
## True Variable Effect(s):
##   1 -2 4.629242
## Standard Error of simulated Variable Effects (RF):
##   0.6462137 1.039248 0.591363 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5134707 0.4995584 0.5034343 .
## Number of Smaller Nulls:
##   14 26 6
##
## Setting 4: 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)
## Mean(s) of simulated RF Variable Effect(s):
##   0.8829454 -1.800354 4.070935 0.03011573
## Mean(s) of simulated LM Variable Effect(s):
##   1.021796 -5.879838 9.580489 0.1446039
## True Variable Effect(s):
##   1 -2 4.629242 0
## Standard Error of simulated Variable Effects (RF):
##   0.2831207 0.3847273 0.6231614 0.2850959 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3144195 0.4427991 0.5398417 0.3137784 .
## Number of Smaller Nulls:
##   7 6 1 12
##
## Setting 5: 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.5
## Mean(s) of simulated RF Variable Effect(s):
##   0.8647216 -1.694709 3.795957 -0.04135875 -0.6273796
## Mean(s) of simulated LM Variable Effect(s):
##   1.004455 -5.818259 9.718525 -0.02498201 -0.8078303
## True Variable Effect(s):
##   1 -2 4.629242 0 -0.75
## Standard Error of simulated Variable Effects (RF):
##   0.574849 0.4206488 0.6314149 0.2833824 0.3214726 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3315991 0.4149544 0.5259288 0.3250422 0.2990347 .
## Number of Smaller Nulls:
##   28 15 4 15 25
##
## Setting 6: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1
## Mean(s) of simulated RF Variable Effect(s):
##   0.9679437
## Mean(s) of simulated LM Variable Effect(s):
##   0.9950645
## True Variable Effect(s):
##   1
## Standard Error of simulated Variable Effects (RF):
##   0.4673022 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5301576 .
## Number of Smaller Nulls:
##   0
##
## Setting 7: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3
## Mean(s) of simulated RF Variable Effect(s):
##   0.7919032 -1.703348
## Mean(s) of simulated LM Variable Effect(s):
##   1.003787 -6.010384
## True Variable Effect(s):
##   1 -2
## Standard Error of simulated Variable Effects (RF):
##   0.3198184 0.390088 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3534157 0.3707056 .
## Number of Smaller Nulls:
##   0 0
##
## Setting 8: 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
## Mean(s) of simulated RF Variable Effect(s):
##   1.152408 -0.9484285 3.93081
## Mean(s) of simulated LM Variable Effect(s):
##   1.050688 -5.932469 9.825033
## True Variable Effect(s):
##   1 -2 4.629242
## Standard Error of simulated Variable Effects (RF):
##   0.3856811 0.3326398 0.5455683 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3948566 0.3564682 0.4276392 .
## Number of Smaller Nulls:
##   5 2 5
##
## Setting 9: 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)
## Mean(s) of simulated RF Variable Effect(s):
##   0.6992561 -1.176983 3.850825 0.148068
## Mean(s) of simulated LM Variable Effect(s):
##   1.012314 -5.885565 9.469685 0.0877609
## True Variable Effect(s):
##   1 -2 4.629242 0
## Standard Error of simulated Variable Effects (RF):
##   0.2633368 0.314028 0.5187603 0.3046569 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3191982 0.3579051 0.5055123 0.3452434 .
## Number of Smaller Nulls:
##   1 3 0 0
##
## Setting 10: 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.5
## Mean(s) of simulated RF Variable Effect(s):
##   0.6725765 -1.177684 3.574717 0.1359979 -0.4437693
## Mean(s) of simulated LM Variable Effect(s):
##   0.9070247 -5.755621 9.726925 0.05708183 -0.948389
## True Variable Effect(s):
##   1 -2 4.629242 0 -0.75
## Standard Error of simulated Variable Effects (RF):
##   0.2667945 0.3370431 0.5953978 0.3053197 0.2454209 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2809577 0.3271512 0.5061641 0.3186145 0.2346754 .
## Number of Smaller Nulls:
##   10 5 1 2 10
```

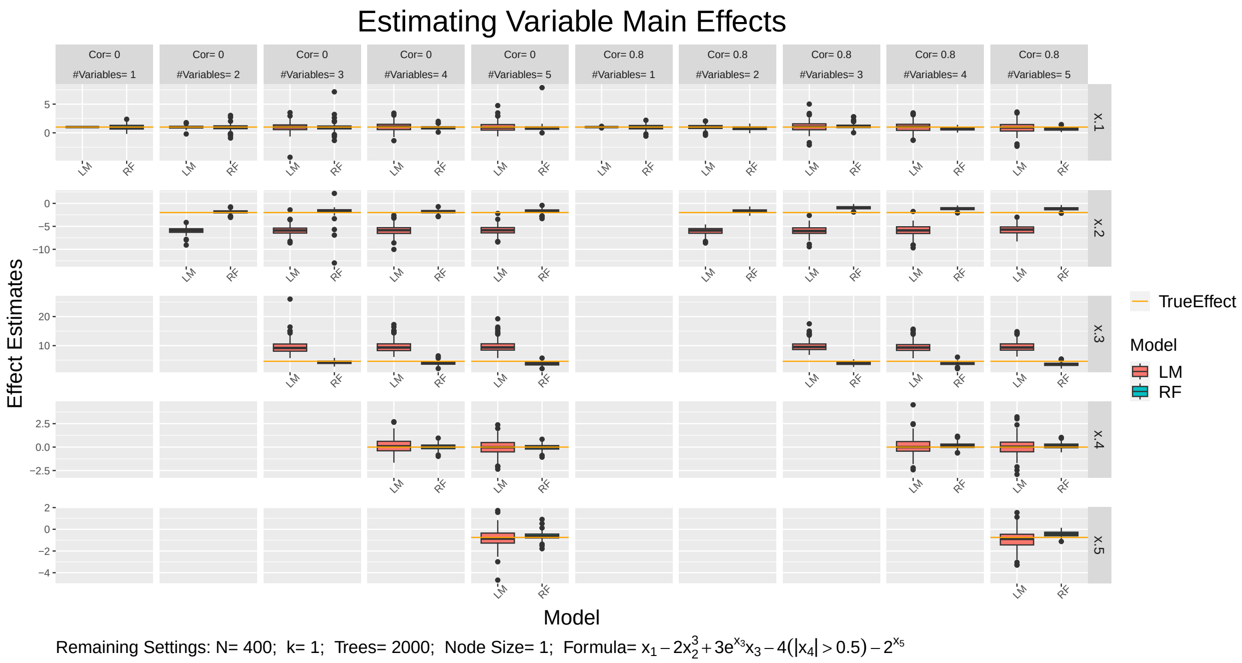
```
effect_plots <- plot_effects(result)
```

```
## ‘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.
```

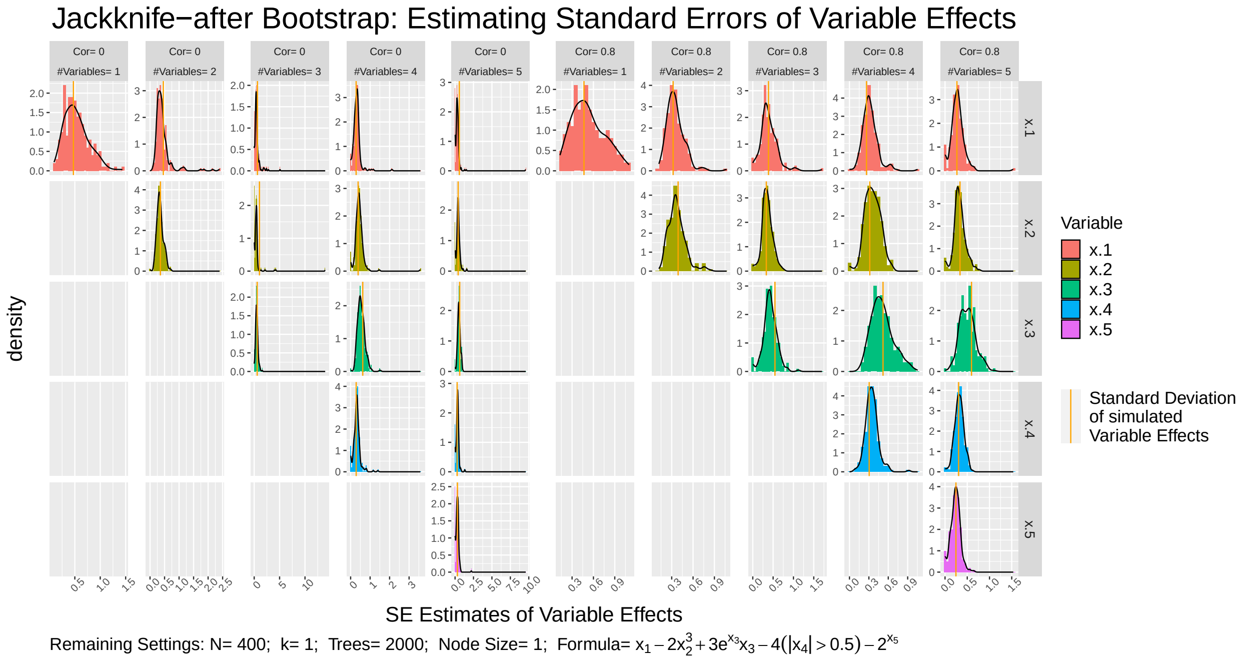
```
effect_plots
```



Estimating Variable Main Effects

Remaining Settings: N= 400; k= 1; Trees= 2000; Node Size= 1; Formula= $x_1 - 2x_2^3 + 3e^{x_3}x_3 - 4(|x_4| > 0.5) - 2^{x_5}$

```
se_plot
```



Jackknife−after Bootstrap: Estimating Standard Errors of Variable Effects

Remaining Settings: N= 400; k= 1; Trees= 2000; Node Size= 1; Formula= $x_1 - 2x_2^3 + 3e^{x_3}x_3 - 4(|x_4| > 0.5) - 2^{x_5}$

```r
n <- c(400) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
k <- c(1); node_size <- c(1); pdp <- F; ale <- F
formulas <- c("x.1",
              "x.1-2*x.2^3",
              "x.1-2*x.2^3+3*exp(x.3)*x.3",
              "x.1-2*x.2^3+3*exp(x.3)*x.3-4*(abs(x.4)>0.5)",
              "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
##    0.08    0.19 1202.14
```

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

```
## Setting 1: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1
## Mean(s) of simulated RF Variable Effect(s):
##   1.052878
## Mean(s) of simulated LM Variable Effect(s):
##   0.9982007
## True Variable Effect(s):
##   1
## Standard Error of simulated Variable Effects (RF):
##   0.4927623 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5397165 .
## Number of Smaller Nulls:
##   0
##
## Setting 2: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3
## Mean(s) of simulated RF Variable Effect(s):
```

```
##   1.029521 -1.892361
## Mean(s) of simulated LM Variable Effect(s):
##   0.9931885 -5.910137
## True Variable Effect(s):
##   1 -2
## Standard Error of simulated Variable Effects (RF):
##   0.3626514 0.3790081 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.4007091 0.3709278 .
## Number of Smaller Nulls:
##   1 1
##
## Setting 3: 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
## Mean(s) of simulated RF Variable Effect(s):
##   0.8779767 -1.724019 4.302348
## Mean(s) of simulated LM Variable Effect(s):
##   1.044238 -5.820051 9.761607
## True Variable Effect(s):
##   1 -2 4.629242
## Standard Error of simulated Variable Effects (RF):
##   0.6521187 0.5390552 0.603793 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5900749 0.4697802 0.4889034 .
## Number of Smaller Nulls:
##   22 11 7
##
## Setting 4: 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)
## Mean(s) of simulated RF Variable Effect(s):
##   0.8817755 -1.838348 3.98456 0.007259945
## Mean(s) of simulated LM Variable Effect(s):
##   0.9203077 -5.931264 9.883974 0.1223425
## True Variable Effect(s):
##   1 -2 4.629242 0
## Standard Error of simulated Variable Effects (RF):
##   0.6011641 0.4453489 0.6155018 0.4518265 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3797168 0.4684696 0.5111986 0.3593519 .
## Number of Smaller Nulls:
##   16 3 5 12
##
## Setting 5: 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):
##   0.8375151 -1.805701 4.065036 -0.0069358 0.01401671
## Mean(s) of simulated LM Variable Effect(s):
##   1.029929 -5.770524 9.649756 0.01367098 0.04777354
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.3095771 0.4386054 0.6724948 0.2797762 0.3322542 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3133396 0.4382195 0.5892393 0.3239208 0.5202119 .
## Number of Smaller Nulls:
##   10 6 3 12 0
##
## Setting 6: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1
## Mean(s) of simulated RF Variable Effect(s):
##   0.9883835
## Mean(s) of simulated LM Variable Effect(s):
##   0.9940773
## True Variable Effect(s):
##   1
## Standard Error of simulated Variable Effects (RF):
##   0.5044178 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.5399479 .
## Number of Smaller Nulls:
##   0
##
## Setting 7: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 1 ;
## Formula = x.1-2*x.2^3
## Mean(s) of simulated RF Variable Effect(s):
##   0.7981531 -1.750696
## Mean(s) of simulated LM Variable Effect(s):
##   1.032266 -5.873715
## True Variable Effect(s):
##   1 -2
## Standard Error of simulated Variable Effects (RF):
##   0.3374257 0.3961004 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3835735 0.3646038 .
## Number of Smaller Nulls:
##   0 0
##
## Setting 8: 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
## Mean(s) of simulated RF Variable Effect(s):
##   1.143704 -0.9768898 3.908077
## Mean(s) of simulated LM Variable Effect(s):
##   1.000353 -5.777315 9.413085
## True Variable Effect(s):
##   1 -2 4.629242
## Standard Error of simulated Variable Effects (RF):
##   0.3814564 0.3815756 0.5261335 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.368646 0.3422761 0.4155156 .
## Number of Smaller Nulls:
##   3 7 7
##
## Setting 9: 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)
## Mean(s) of simulated RF Variable Effect(s):
##   0.7595723 -1.202843 3.885992 0.1379597
## Mean(s) of simulated LM Variable Effect(s):
##   1.018391 -5.859212 9.680046 -0.05726719
## True Variable Effect(s):
##   1 -2 4.629242 0
## Standard Error of simulated Variable Effects (RF):
##   0.2857052 0.3434102 0.5471736 0.3108382 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3175196 0.3624527 0.5166523 0.3521749 .
## Number of Smaller Nulls:
##   1 1 0 0
##
## Setting 10: 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):
##   0.7725019 -1.135798 3.801906 0.1350401 0.03098075
## Mean(s) of simulated LM Variable Effect(s):
##   1.003857 -5.913045 9.478464 0.1370576 -4.72593
## True Variable Effect(s):
##   1 -2 4.629242 0 0
## Standard Error of simulated Variable Effects (RF):
##   0.3057994 0.3323579 0.5515299 0.3745343 0.3112674 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3060905 0.3633446 0.5300374 0.3523195 0.4449279 .
## Number of Smaller Nulls:
##   2 2 0 0 0
```

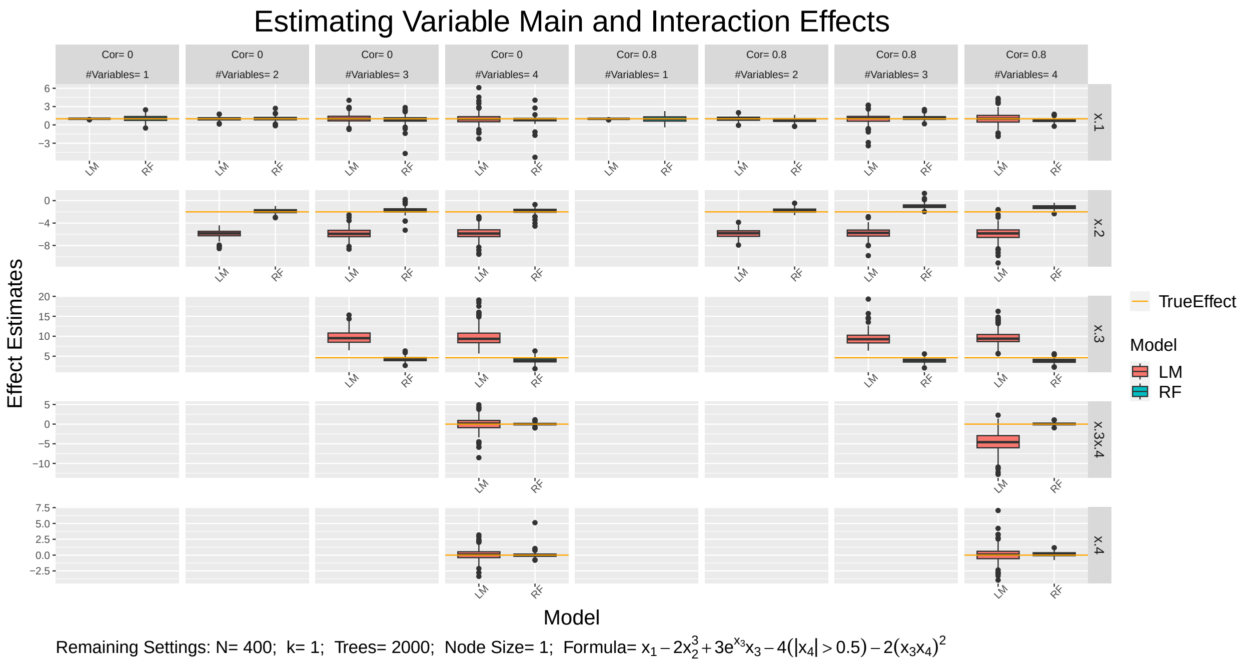
```
effect_plots <- plot_effects(result)
```

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

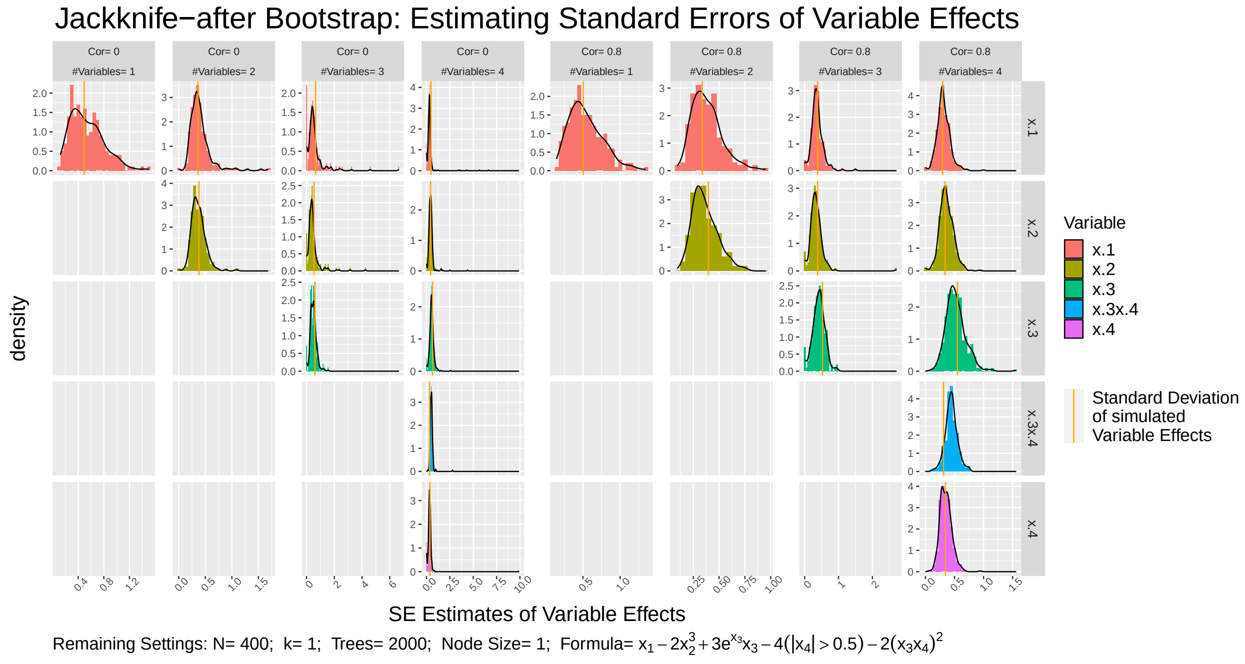
```
se_plot <- plot_se(result)
```

```
## ‘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
Cor= 0 #Variables= 1
Cor= 0 #Variables= 2
Cor= 0 #Variables= 3
Cor= 0 #Variables= 4
Cor= 0.8 #Variables= 1
Cor= 0.8 #Variables= 2
Cor= 0.8 #Variables= 3
Cor= 0.8 #Variables= 4
x.1
x.2
x.3
x.3x.4
x.4
LM
RF
Effect Estimates
Model
TrueEffect
Model
LM
RF
Remaining Settings: N= 400; k= 1; Trees= 2000; Node Size= 1; 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
Cor= 0 #Variables= 1
Cor= 0 #Variables= 2
Cor= 0 #Variables= 3
Cor= 0 #Variables= 4
Cor= 0.8 #Variables= 1
Cor= 0.8 #Variables= 2
Cor= 0.8 #Variables= 3
Cor= 0.8 #Variables= 4
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: N= 400; k= 1; Trees= 2000; Node Size= 1; Formula= $x_1 - 2x_2^3 + 3e^{x_3}x_3 - 4(|x_4| > 0.5) - 2(x_3x_4)^2$