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

Felix Kapulla

```
knitr::opts_chunk$set(fig.width=14, fig.height=8)
```

```
library(Matrix)
library(tidyverse)
library(ggplot2)
library(ggpubr)
library(ranger)
library(MixMatrix)
library(mvtnorm)
library(stringr)
library(parallel)
source('C:/Users/feix_/iCloudDrive/Studium Master/CQM - Thesis Internship/Thesis-VariableEffects/Baseli
# cores <- detectCores()
# clust <- makeCluster(cores-1)
# parallel::clusterEvalQ(clust,
#                        expr = {source('C:/Users/feix_/iCloudDrive/Studium Master/CQM - Thesis Interns
```

Simulation

```
n <- c(400, 4000) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
k <- c(0.2, 1); node_size <- 5
formulas <- c("2*x.1+4*x.2-0.5*x.3+2.2*x.4")
scenarios <- data.frame(expand.grid(n, num.trees, formulas, repeats,
                                    cor, k, node_size))
colnames(scenarios) = c("N", "N_Trees", "Formula", "Repeats",
                        "Correlation", "k", "Node_Size")
scenarios[,"Formula"] <- as.character(scenarios[,"Formula"]) ### Formula became Factor
scenarios <- split(scenarios, seq(nrow(scenarios)))
system.time(result <- lapply(X = scenarios, FUN = sim_multi))
```

```
##     user   system  elapsed
## 27287.22  1088.61  7400.47
```

```
#Run Simulation
# system.time(result <- parLapply(cl = clust,
#                                 X = scenarios,
#                                 fun = sim_multi))
#stopCluster(clust)
```

```
print_results(result)
```

```
## Setting 1: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.866874 4.368406 -0.30355 2.231225
## Mean(s) of simulated LM Variable Effect(s):
##   2.004825 4.001511 -0.5009547 2.199257
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.9434505 1.412226 0.2848474 1.016713 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.9664579 1.600096 0.2992765 1.069307 .
## Number of Smaller Nulls:
##   9 0 47 9
##
## Setting 2: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   2.06926 4.416613 -0.3941406 2.382714
## Mean(s) of simulated LM Variable Effect(s):
##   2.000248 3.999514 -0.5007153 2.200665
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.8535174 1.034873 0.3556811 0.8948765 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.8111324 1.09211 0.5008294 0.9661676 .
## Number of Smaller Nulls:
##   50 28 49 29
##
## Setting 3: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.784883 4.041052 -0.03981188 2.082522
## Mean(s) of simulated LM Variable Effect(s):
##   2.006854 4.000957 -0.5047086 2.197311
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.9544331 1.206937 0.3595162 0.9162381 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.9821396 1.306016 0.4031362 1.020449 .
## Number of Smaller Nulls:
##   1 0 16 0
##
## Setting 4: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.983109 3.977771 -0.2982558 2.199619
## Mean(s) of simulated LM Variable Effect(s):
##   2.000511 3.997801 -0.5015595 2.202556
```

```
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.8689132 0.9375509 0.3422117 0.8998986 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.9605602 1.017817 0.4609828 0.9685559 .
## Number of Smaller Nulls:
##   14 26 47 21
##
## Setting 5: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.803698 3.91215 -0.290887 1.980237
## Mean(s) of simulated LM Variable Effect(s):
##   2.003057 4.003963 -0.5004176 2.197447
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2792475 0.3287363 0.1328027 0.2821532 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3109121 0.371659 0.1488752 0.3280302 .
## Number of Smaller Nulls:
##   1 0 25 0
##
## Setting 6: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.952386 3.965611 -0.3942233 2.136422
## Mean(s) of simulated LM Variable Effect(s):
##   2.000791 3.998983 -0.5001806 2.198781
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2116108 0.2204729 0.144868 0.2199292 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2450483 0.2401344 0.1732931 0.2220997 .
## Number of Smaller Nulls:
##   26 23 42 30
##
## Setting 7: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.681723 3.738946 0.04964862 1.889897
## Mean(s) of simulated LM Variable Effect(s):
##   2.005528 3.998527 -0.5067161 2.198432
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2466799 0.2842898 0.1512438 0.2815353 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3050421 0.3387647 0.1599544 0.3073377 .
## Number of Smaller Nulls:
##   2 1 12 1
##
```

```
## Setting 8: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.779114 3.892531 -0.2251415 2.017318
## Mean(s) of simulated LM Variable Effect(s):
##   1.997328 3.998003 -0.4992711 2.202116
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2216304 0.217767 0.1347205 0.2252413 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2345328 0.248022 0.1478816 0.2219522 .
## Number of Smaller Nulls:
##   19 19 47 27
```

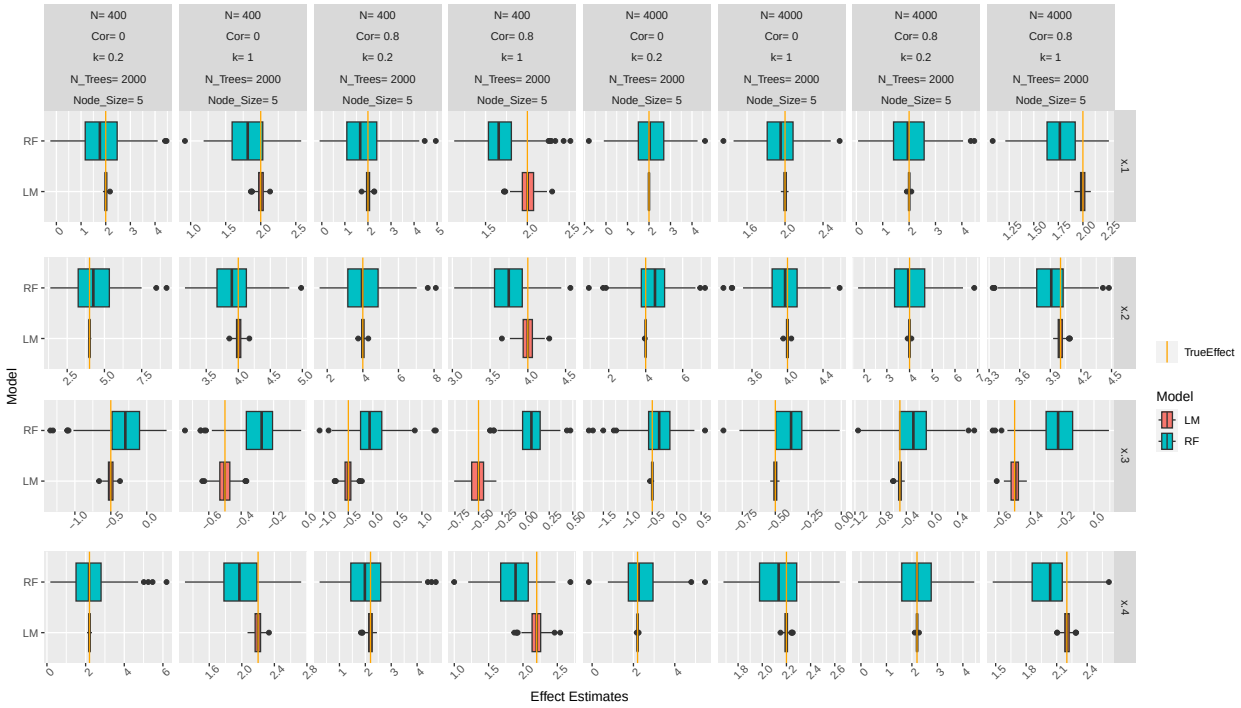
```
effect_plots <- plot_effects(result)
```

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

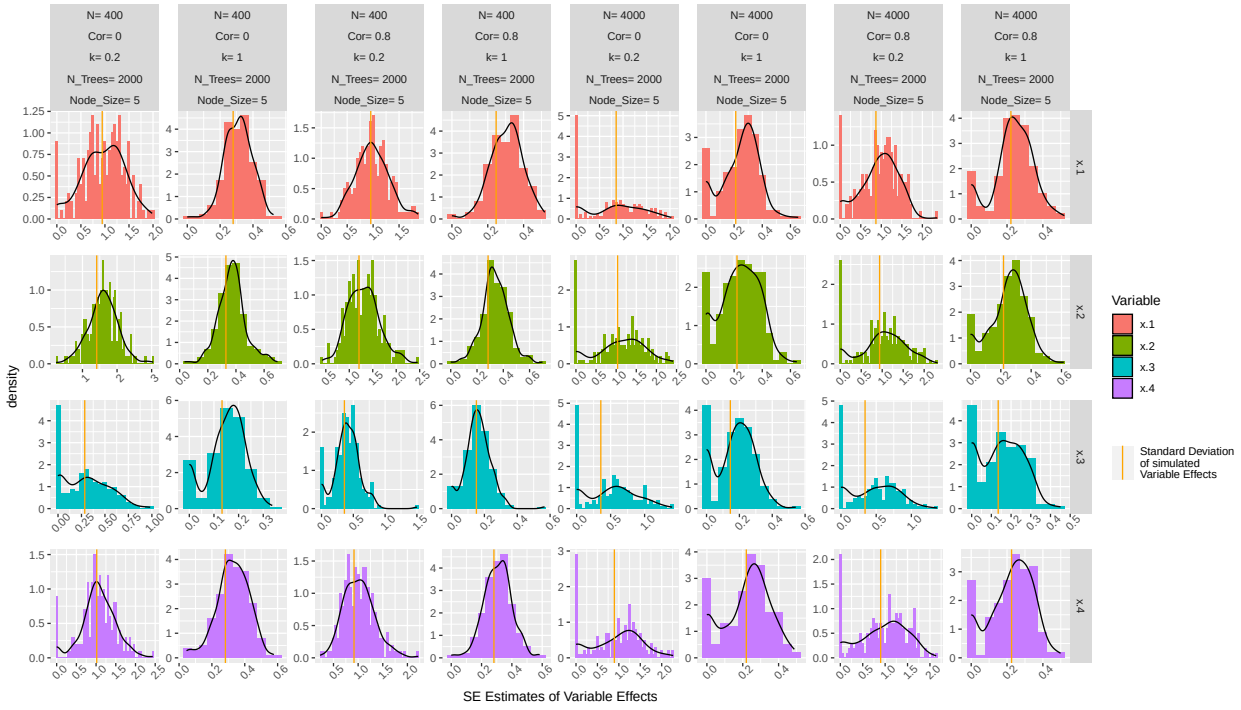
```
se_plot <- plot_se(result)
```

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

```
effect_plots
```

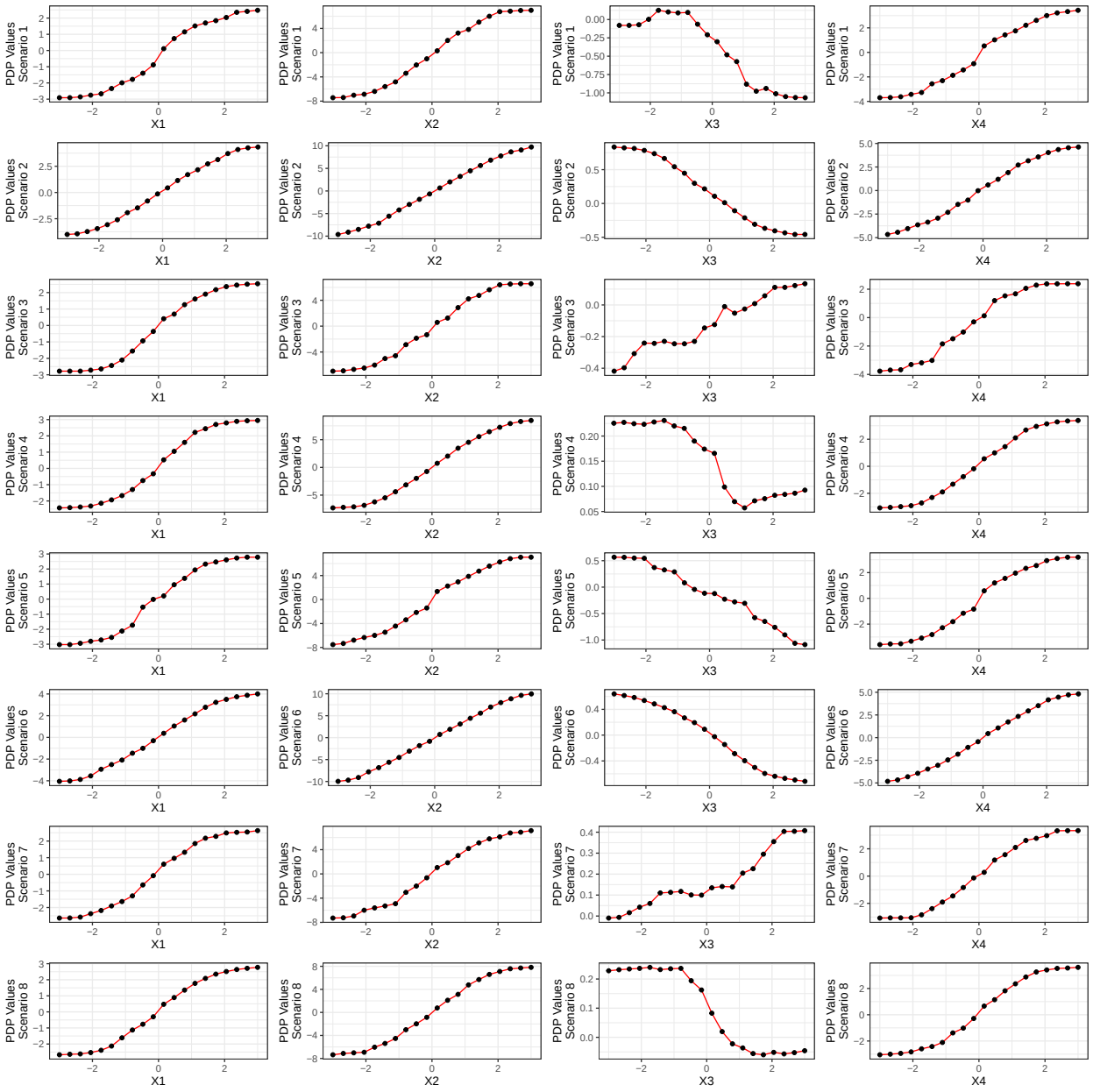


```
se_plot
```




```
plot_pdps(result)
```

```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with ‘aes()‘
```


PDP Values Scenario 1
PDP Values Scenario 2
PDP Values Scenario 3
PDP Values Scenario 4
PDP Values Scenario 5
PDP Values Scenario 6
PDP Values Scenario 7
PDP Values Scenario 8
X1
X2
X3
X4

```
n <- c(400) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0)
k <- c(1); node_size <- c(1, 5, 100)
formulas <- c("2*x.1+4*x.2-0.5*x.3+2.2*x.4")
scenarios <- data.frame(expand.grid(n, num.trees, formulas, repeats,
                                    cor, k, node_size))
colnames(scenarios) = c("N", "N_Trees", "Formula", "Repeats",
                        "Correlation", "k", "Node_Size")
scenarios[,"Formula"] <- as.character(scenarios[,"Formula"]) ### Formula became Factor
scenarios <- split(scenarios, seq(nrow(scenarios)))
system.time(result <- lapply(X = scenarios, FUN = sim_multi))
```

```
##    user  system elapsed
## 1013.22   36.88  320.28
```

```
#Run Simulation
# system.time(result <- parLapply(cl = clust,
#                                 X = scenarios,
#                                 fun = sim_multi))
#stopCluster(clust)
```

```
print_results(result)
```

```
## Setting 1: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 1 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.82291 3.894038 -0.2924567 2.02035
## Mean(s) of simulated LM Variable Effect(s):
##   2.007268 4.005154 -0.5037806 2.203316
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2856286 0.3624526 0.1352331 0.3004344 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3222181 0.3917451 0.163792 0.3307528 .
## Number of Smaller Nulls:
##   1 0 16 4
##
## Setting 2: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.760606 3.875843 -0.2953254 2.014947
## Mean(s) of simulated LM Variable Effect(s):
##   2.001734 4.004083 -0.4992996 2.200433
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2810826 0.3196873 0.1429669 0.2769429 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3239505 0.383674 0.1504479 0.3235231 .
## Number of Smaller Nulls:
##   0 0 25 0
##
```

```
## Setting 3: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 100 ;
##          Formula = 2*x.1+4*x.2-0.5*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.246611 3.528612 -0.08369733 1.508565
## Mean(s) of simulated LM Variable Effect(s):
##   2.00052 4.000157 -0.5017892 2.193928
## True Variable Effect(s):
##   2 4 -0.5 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.222309 0.3629937 0.04338138 0.2480373 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.244281 0.3910674 0.04646482 0.2694188 .
## Number of Smaller Nulls:
##   9 0 48 8
```

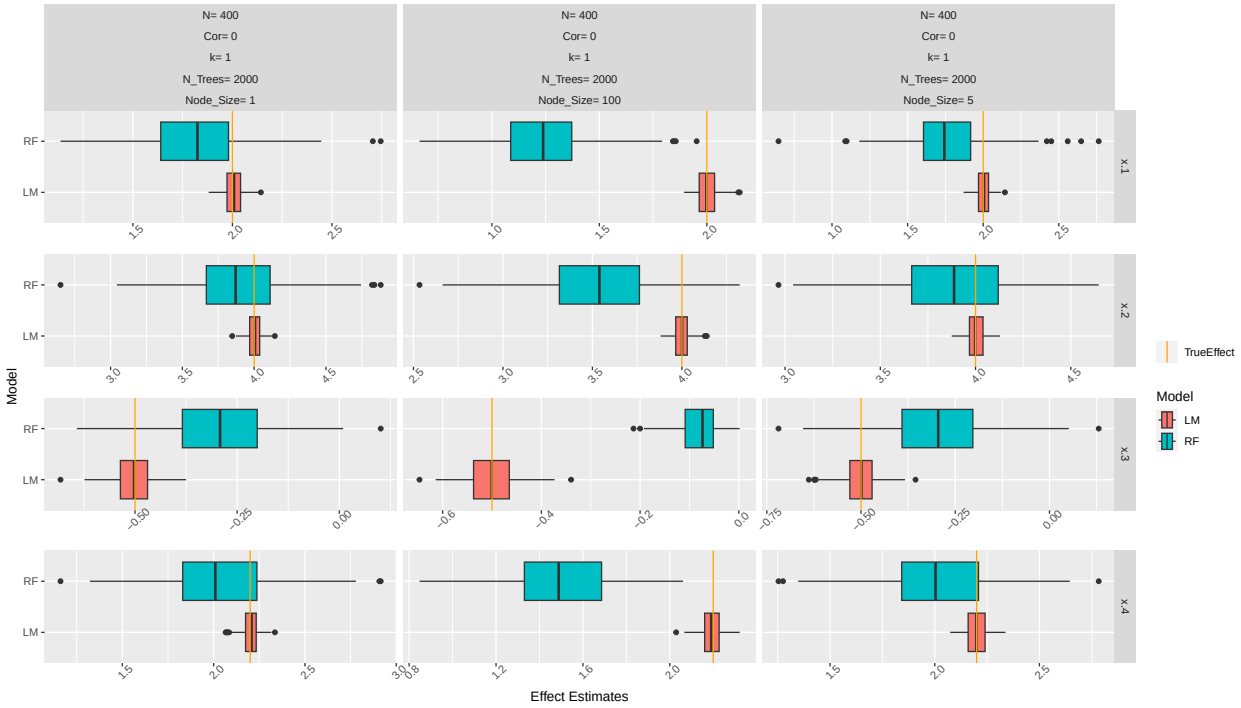
```
effect_plots <- plot_effects(result)
```

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

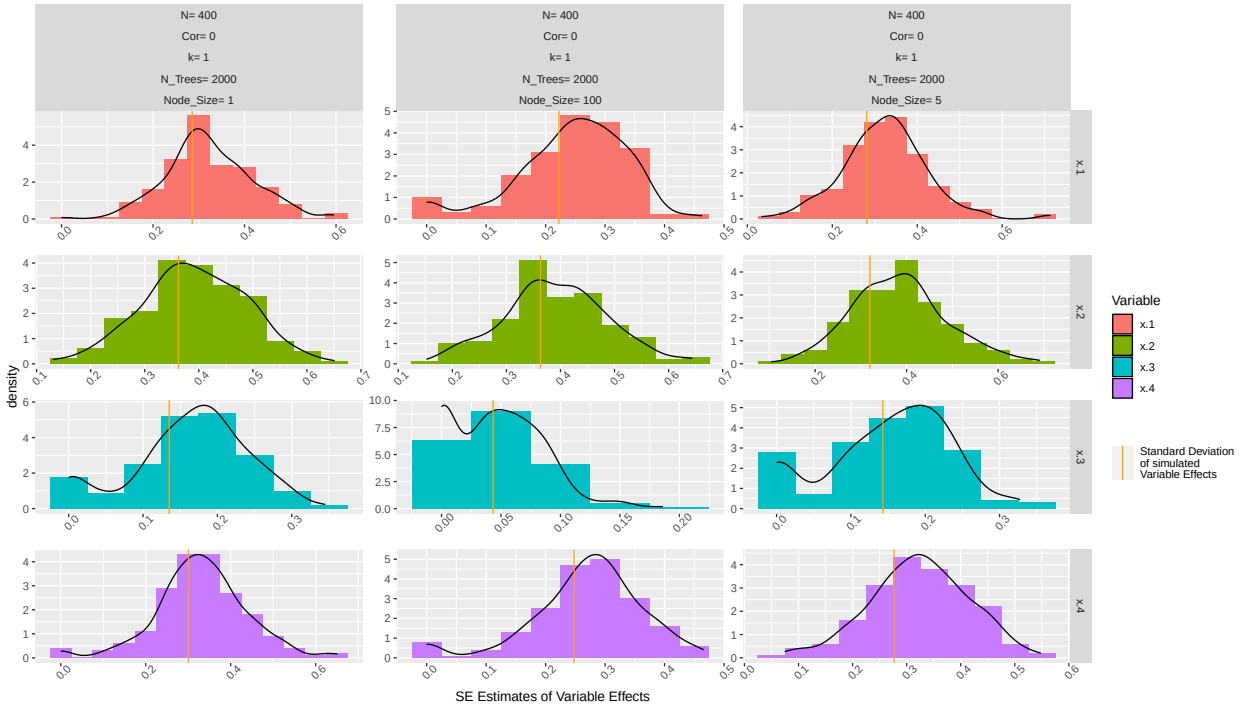
```
se_plot <- plot_se(result)
```

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

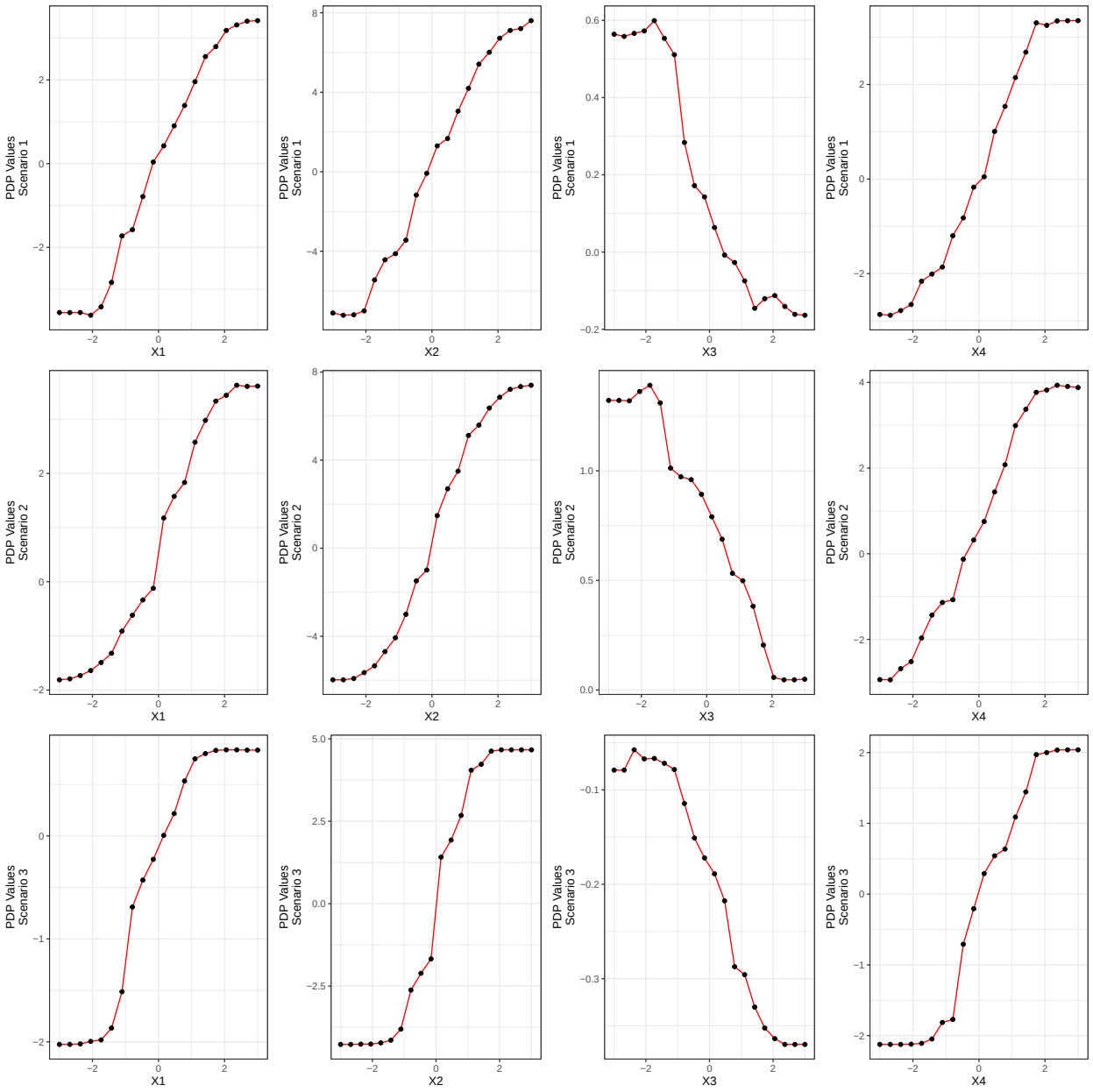
```
effect_plots
```



```
se_plot
```




```
plot_pdps(result)
```


PDP Values Scenario 1
X1
X2
X3
X4
PDP Values Scenario 2
PDP Values Scenario 3