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)

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, 4000) ; num.trees <- 2000 ; repeats <- 25; cor <- c(0, 0.8)
k <- c(0.2, 1); node_size <- 5

formulas <- c("2*x.1+4*x.2-3*x.3+2.2*x.4")
parallel::clusterExport(cl = clust, varlist = 'formulas')
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))
#Run Simulation
system.time(result <- parLapply(cl = clust,
                                X = scenarios,
                                fun = sim_multi))
```

```
##    user  system elapsed
##    0.03    0.14 9714.23
```

```
#n <- c(400) ; num.trees <- 2000 ; repeats <- 200; cor <- c(0, 0.8)
#k <- c(1); node_size <- c(1, 5, 100)
```

```
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.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.835421 4.693725 -2.501647 1.92564
## Mean(s) of simulated LM Variable Effect(s):
##   2.002947 3.995393 -2.984373 2.201051
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.6339431 1.578164 1.344074 0.79587 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.9310794 2.095856 1.410991 1.084546 .
## Number of Smaller Nulls:
##   1 0 0 1
##
## Setting 2: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   2.2325 4.46783 -3.442266 2.395639
## Mean(s) of simulated LM Variable Effect(s):
##   1.997088 3.995655 -3.001335 2.196187
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.6806877 0.8295604 1.165746 0.8939239 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.816083 1.160021 1.289947 0.9900736 .
## Number of Smaller Nulls:
##   7 4 4 5
##
## Setting 3: N = 400 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.733545 4.213461 -2.143972 1.833192
## Mean(s) of simulated LM Variable Effect(s):
##   1.977154 4.041815 -3.015121 2.202934
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   1.194506 1.669516 0.9871739 0.9822198 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.050537 1.454767 0.8621647 1.096022 .
## Number of Smaller Nulls:
##   0 0 0 0
##
```

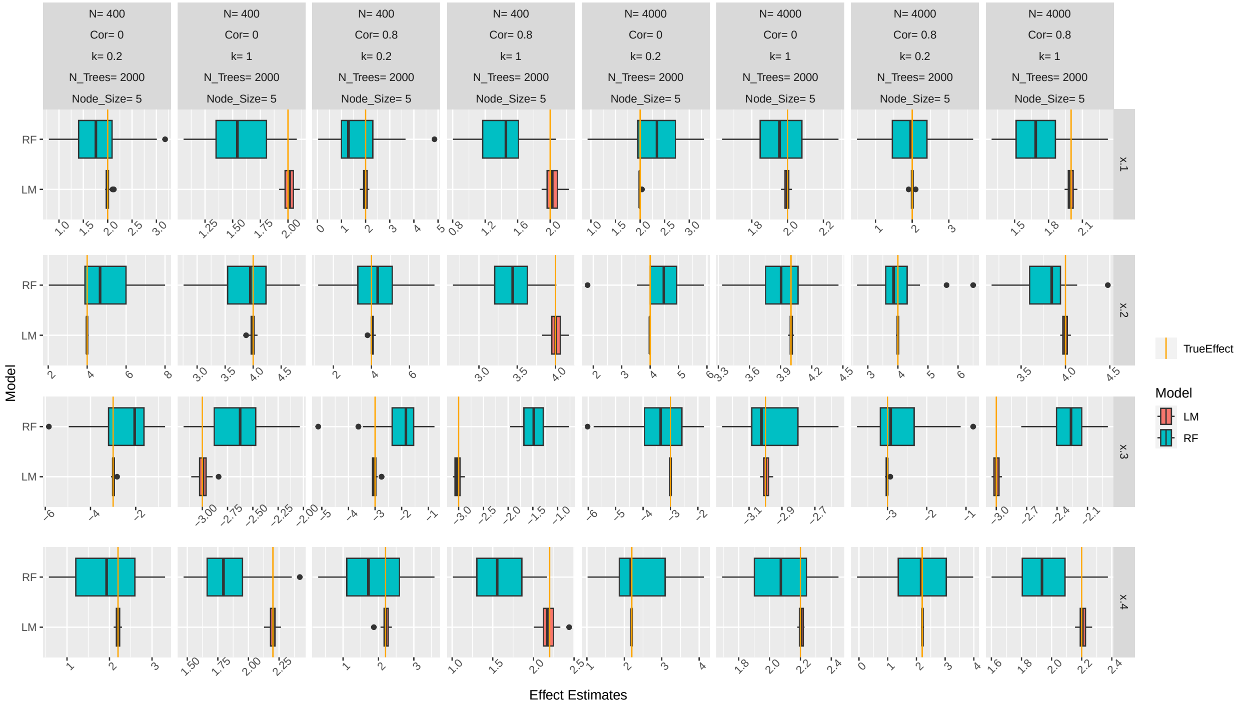
```
## Setting 4: N = 4000 ; k = 0.2 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.948415 3.965964 -2.682621 2.106681
## Mean(s) of simulated LM Variable Effect(s):
##   2.001287 4.002291 -3.009822 2.208757
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.7601325 0.8330654 0.8107465 1.189138 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.913656 1.033238 0.9626783 1.13456 .
## Number of Smaller Nulls:
##   4 2 5 4
##
## Setting 5: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.556956 3.904334 -2.653766 1.861803
## Mean(s) of simulated LM Variable Effect(s):
##   2.013372 3.983403 -2.993704 2.19536
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2912889 0.4796368 0.3333571 0.2574759 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3333026 0.5129965 0.4054587 0.3729918 .
## Number of Smaller Nulls:
##   1 0 0 1
##
## Setting 6: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.957841 3.898904 -2.94927 2.069039
## Mean(s) of simulated LM Variable Effect(s):
##   1.995241 4.001745 -2.996274 2.203811
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.1558078 0.2453017 0.2022808 0.2257155 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2779681 0.2873529 0.2770407 0.2872489 .
## Number of Smaller Nulls:
##   4 3 3 4
##
## Setting 7: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.404835 3.412498 -1.479015 1.575656
## Mean(s) of simulated LM Variable Effect(s):
##   2.027383 4.009386 -3.017576 2.17908
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
```

```
##   0.3138702 0.349494 0.3100941 0.3062274 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3569627 0.4257582 0.3125806 0.3450876 .
## Number of Smaller Nulls:
##   0 0 0 0
##
## Setting 8: N = 4000 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.711752 3.793111 -2.291378 1.944701
## Mean(s) of simulated LM Variable Effect(s):
##   1.996553 3.996669 -3.001624 2.205721
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2244863 0.2787307 0.1974503 0.2034594 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.265737 0.260076 0.2012942 0.1824444 .
## Number of Smaller Nulls:
##   5 5 8 6
```

```
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)
effect_plots
```

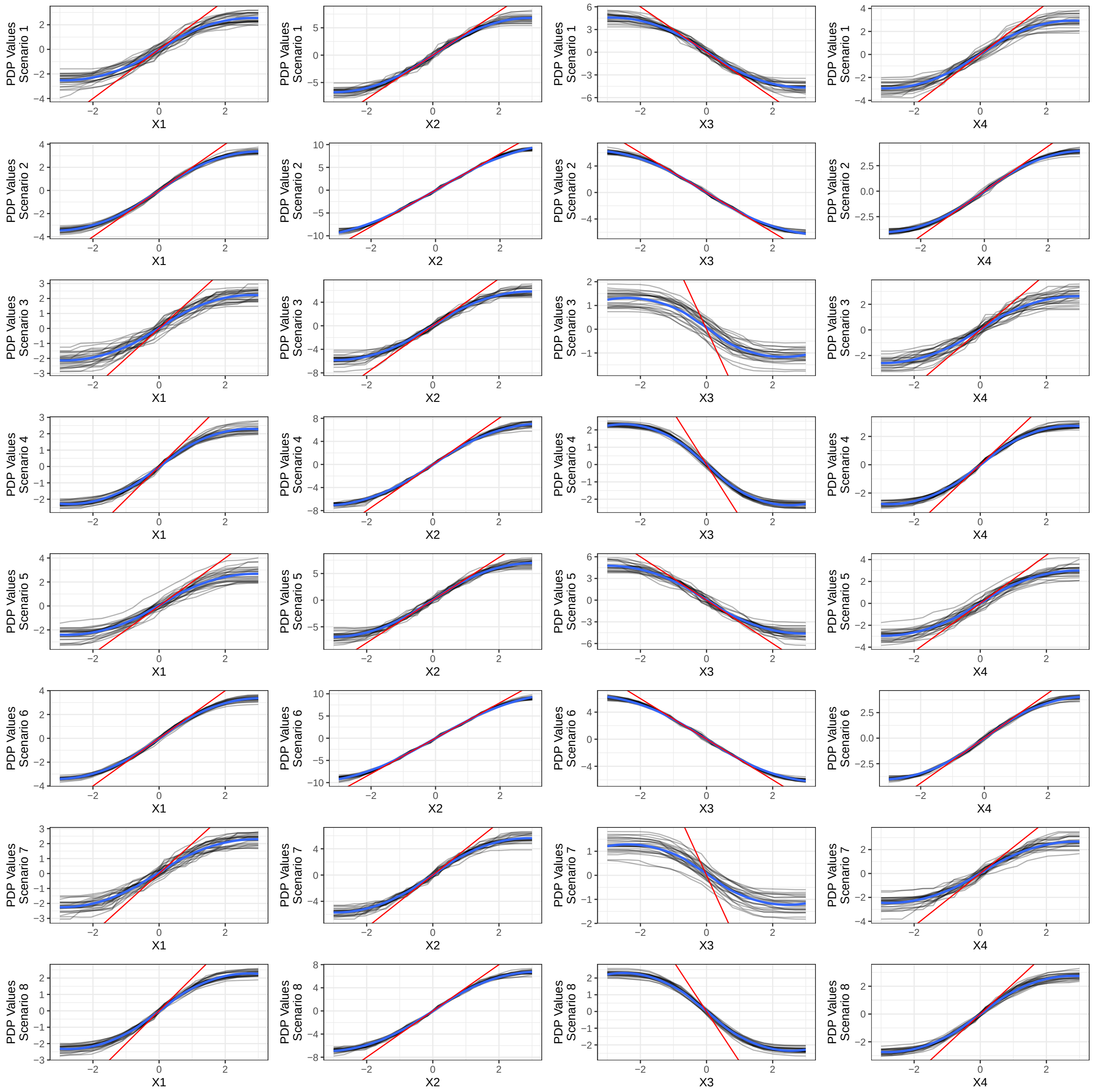


```
#se_plot
```

```
plot_pdps(result)
```

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

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



```
n <- c(400) ; num.trees <- 2000 ; repeats <- 25; cor <- c(0, 0.8)
k <- c(1); node_size <- c(1, 5, 100)

formulas <- c("2*x.1+4*x.2-3*x.3+2.2*x.4")
parallel::clusterExport(cl = clust, varlist = 'formulas')
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))
#Run Simulation
system.time(result <- parLapply(cl = clust,
                                X = scenarios,
                                fun = sim_multi))
```

```
##    user  system elapsed
##    0.00    0.05  522.08
```

```
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.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.559239 3.747535 -2.867026 1.919492
## Mean(s) of simulated LM Variable Effect(s):
##   1.99092 4.006588 -3.002631 2.210554
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2981234 0.3388409 0.3359494 0.3425085 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3094456 0.4647545 0.4232519 0.3657832 .
## Number of Smaller Nulls:
##   3 0 0 0
##
## Setting 2: 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 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.342273 3.354428 -1.353913 1.602446
## Mean(s) of simulated LM Variable Effect(s):
##   1.993904 4.011471 -2.993125 2.201093
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2464376 0.2741061 0.2437082 0.3385745 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.333981 0.4063587 0.2940455 0.3665163 .
## Number of Smaller Nulls:
```

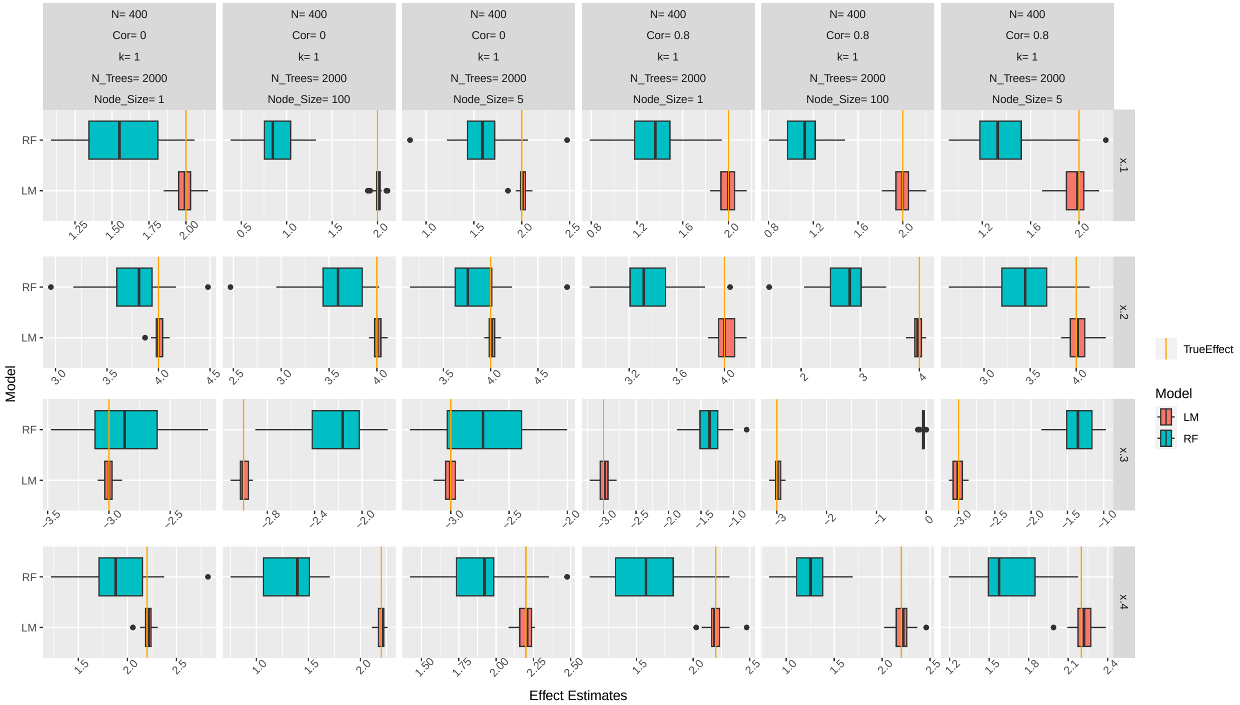
```
##   0 0 0 0
##
## Setting 3: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.588815 3.787651 -2.701698 1.924184
## Mean(s) of simulated LM Variable Effect(s):
##   2.009616 4.015745 -3.002201 2.196921
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.315269 0.36362 0.3798632 0.2671987 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3313573 0.479691 0.4558573 0.3360366 .
## Number of Smaller Nulls:
##   1 0 0 3
##
## Setting 4: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 5 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.390853 3.423811 -1.377179 1.641511
## Mean(s) of simulated LM Variable Effect(s):
##   1.972915 4.020617 -3.014216 2.222674
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.3518579 0.3559948 0.2549314 0.271613 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.324589 0.3867037 0.297948 0.3523888 .
## Number of Smaller Nulls:
##   1 0 0 2
##
## Setting 5: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0 ; Minimum Node Size = 100 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   0.8835802 3.55528 -2.218626 1.299753
## Mean(s) of simulated LM Variable Effect(s):
##   2.005709 4.009949 -2.999111 2.20119
## True Variable Effect(s):
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.2483698 0.3606988 0.2747809 0.2861535 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2930285 0.4735001 0.3283782 0.3072017 .
## Number of Smaller Nulls:
##   1 0 2 0
##
## Setting 6: N = 400 ; k = 1 N_Trees = 2000 ; Correlation = 0.8 ; Minimum Node Size = 100 ;
##          Formula = 2*x.1+4*x.2-3*x.3+2.2*x.4 ; N_Trees = 2000
## Mean(s) of simulated RF Variable Effect(s):
##   1.108685 2.737371 -0.06494871 1.253171
## Mean(s) of simulated LM Variable Effect(s):
##   1.988056 3.977815 -2.984589 2.207315
## True Variable Effect(s):
```

```
##   2 4 -3 2.2
## Standard Error of simulated Variable Effects (RF):
##   0.178626 0.4592597 0.03935917 0.2212652 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2439827 0.3662587 0.04401566 0.2769652 .
## Number of Smaller Nulls:
##   0 1 6 1
```

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

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

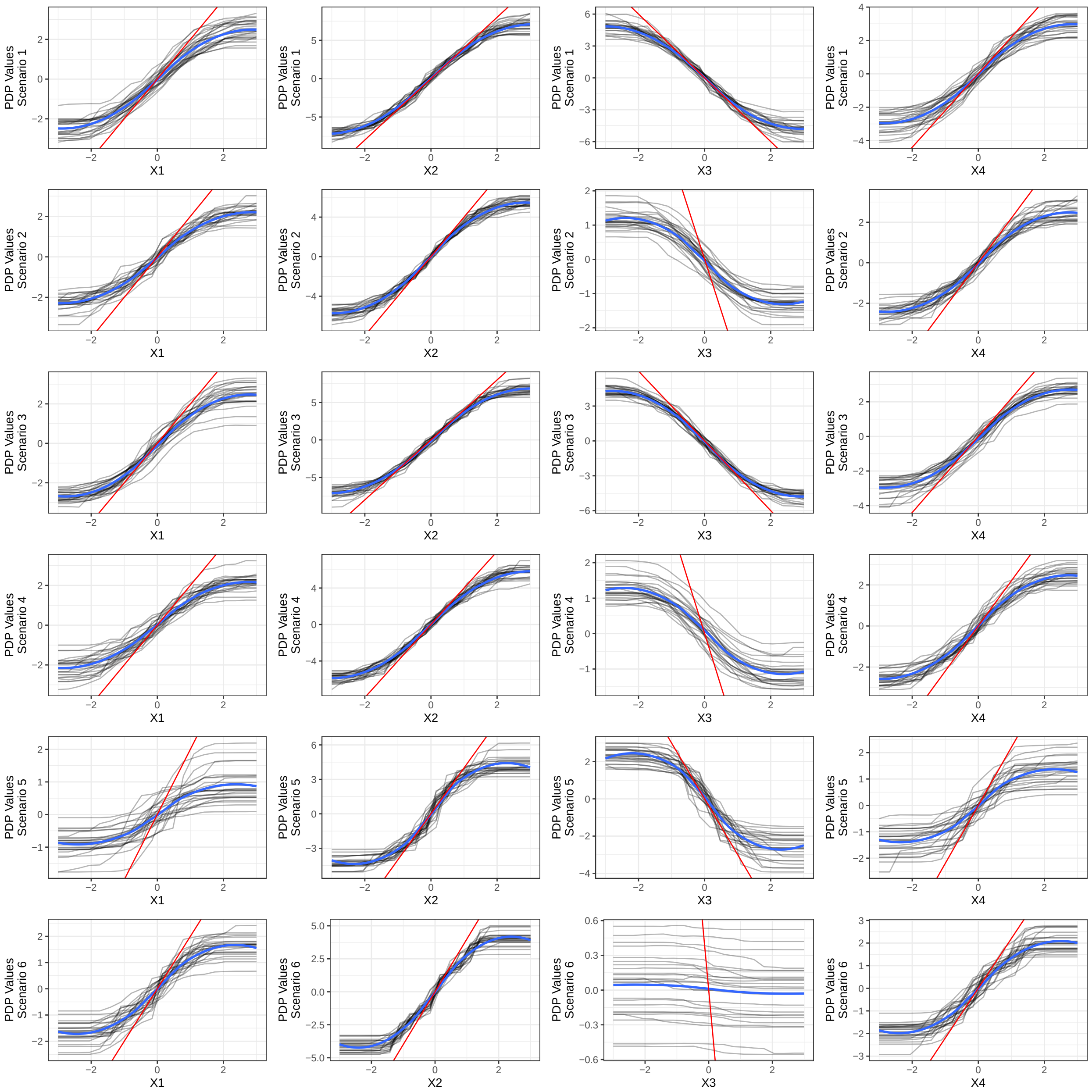
```
#se_plot <- plot_se(result)
effect_plots
```



```
#se_plot
```

```
plot_pdps(result)
```

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


PDP Values Scenario 1
2
0
−2
−2
0
2
X1
PDP Values Scenario 1
5
0
−5
−2
0
2
X2
PDP Values Scenario 1
6
3
0
−3
−6
−2
0
2
X3
PDP Values Scenario 1
4
2
0
−2
−4
−2
0
2
X4
PDP Values Scenario 2
2
0
−2
−2
0
2
X1
PDP Values Scenario 2
4
0
−4
−2
0
2
X2
PDP Values Scenario 2
2
1
0
−1
−2
−2
0
2
X3
PDP Values Scenario 2
2
0
−2
−2
0
2
X4
PDP Values Scenario 3
2
0
−2
−2
0
2
X1
PDP Values Scenario 3
5
0
−5
−2
0
2
X2
PDP Values Scenario 3
3
0
−3
−6
−2
0
2
X3
PDP Values Scenario 3
2
0
−2
−4
−2
0
2
X4
PDP Values Scenario 4
2
0
−2
−2
0
2
X1
PDP Values Scenario 4
4
0
−4
−2
0
2
X2
PDP Values Scenario 4
2
1
0
−1
−2
0
2
X3
PDP Values Scenario 4
2
0
−2
−2
0
2
X4
PDP Values Scenario 5
2
1
0
−1
−2
0
2
X1
PDP Values Scenario 5
6
3
0
−3
−2
0
2
X2
PDP Values Scenario 5
2
0
−2
−4
−2
0
2
X3
PDP Values Scenario 5
2
1
0
−1
−2
−2
0
2
X4
PDP Values Scenario 6
2
1
0
−1
−2
−2
0
2
X1
PDP Values Scenario 6
5.0
2.5
0.0
−2.5
−5.0
−2
0
2
X2
PDP Values Scenario 6
0.6
0.3
0.0
−0.3
−0.6
−2
0
2
X3
PDP Values Scenario 6
3
2
1
0
−1
−2
−3
−2
0
2
X4