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(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("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")
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
##     1.25     2.31 13486.73
```

```
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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   0.989074 3.066033 -2.153594 1.180682 -0.7227185
## Mean(s) of simulated LM Variable Effect(s):
##   2.005335 4.005195 -2.989008 2.208721 -1.510876
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   1.25806 2.359318 2.007152 1.36902 1.126388 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.327965 2.665076 1.982572 1.446042 1.174069 .
## Number of Smaller Nulls:
##   3 0 0 3 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.688745 4.337196 -2.98528 1.902715 -1.045087
## Mean(s) of simulated LM Variable Effect(s):
##   1.995359 4.001413 -3.002338 2.196875 -1.501501
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.8313032 1.677499 1.378223 0.9476404 0.5335667 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.8689221 1.952629 1.424156 1.00562 0.652478 .
## Number of Smaller Nulls:
##   36 4 13 28 41
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   2.203317 4.637699 -3.341859 2.415581 -1.446277
## Mean(s) of simulated LM Variable Effect(s):
##   1.9997 3.997598 -2.999612 2.201434 -1.499661
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.7303045 1.060518 0.9074625 0.819183 0.469704 .
```

```
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.041735 1.389075 1.305743 1.101308 0.9495438 .
## Number of Smaller Nulls:
##   59 47 53 55 60
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   0.9998642 2.328665 -0.5561285 1.284727 -0.2168764
## Mean(s) of simulated LM Variable Effect(s):
##   2.029693 3.996465 -3.034369 2.194971 -1.477556
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   1.167811 1.87856 0.6309279 1.481377 0.6690634 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   1.394882 1.973191 0.7679727 1.481587 0.8753965 .
## Number of Smaller Nulls:
##   0 0 1 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.513802 3.55965 -1.953174 1.687302 -0.796812
## Mean(s) of simulated LM Variable Effect(s):
##   2.008298 4.008031 -3.001425 2.190819 -1.500028
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.8839418 1.345093 0.7050937 0.9560777 0.5996694 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.9663524 1.442219 0.8757531 1.005854 0.5587766 .
## Number of Smaller Nulls:
##   5 1 5 7 25
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.836234 4.04844 -2.907274 2.171276 -1.196822
## Mean(s) of simulated LM Variable Effect(s):
##   1.999635 4.002419 -3.001451 2.197476 -1.497557
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.6952734 0.9207793 0.7122099 0.8001829 0.5448965 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.8770691 1.103529 0.9321232 0.9421134 0.7807929 .
## Number of Smaller Nulls:
##   48 41 36 46 43
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.103892 2.73916 -1.796903 1.179834 -0.7446299
```

```
## Mean(s) of simulated LM Variable Effect(s):
##   1.999031 4.031385 -3.017765 2.209433 -1.471938
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.5721092 0.6991086 0.6590858 0.5676593 0.4908361 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.6349754 0.8089658 0.7286235 0.6472656 0.5526473 .
## 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 = 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):
##   1.509893 3.669964 -2.696888 1.787641 -0.9954056
## Mean(s) of simulated LM Variable Effect(s):
##   1.996253 4.001392 -3.004535 2.199687 -1.50119
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.2939082 0.4180029 0.3860526 0.3200229 0.2374842 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3166911 0.43521 0.4193827 0.3606314 0.2622753 .
## Number of Smaller Nulls:
##   18 6 9 12 22
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.84725 3.903992 -2.886446 2.06235 -1.250031
## Mean(s) of simulated LM Variable Effect(s):
##   1.998591 3.997835 -3.002081 2.201138 -1.504111
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.2007446 0.2424615 0.2185068 0.2285574 0.180278 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2901323 0.3275556 0.2999634 0.2920165 0.2672256 .
## Number of Smaller Nulls:
##   44 44 49 57 56
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.017269 1.869628 -0.356509 1.096408 -0.05084701
## Mean(s) of simulated LM Variable Effect(s):
##   2.006207 3.97642 -3.003302 2.234925 -1.509224
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.409758 0.5228259 0.3051133 0.4296796 0.3222852 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.481472 0.5306717 0.336761 0.4895128 0.3489615 .
## 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 = 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):
##   1.284521 3.042519 -1.512648 1.400328 -0.6566918
## Mean(s) of simulated LM Variable Effect(s):
##   1.99927 4.005045 -3.003192 2.202088 -1.503436
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.2836394 0.3400942 0.2582548 0.2800804 0.20684 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.3107737 0.3773996 0.2742352 0.3088739 0.2303534 .
## Number of Smaller Nulls:
##   7 4 9 9 8
##
## 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.4-1.5*x.5
## Mean(s) of simulated RF Variable Effect(s):
##   1.551648 3.626522 -2.328929 1.789976 -0.9217441
## Mean(s) of simulated LM Variable Effect(s):
##   2.002031 4.002966 -3.001727 2.198887 -1.501813
## True Variable Effect(s):
##   2 4 -3 2.2 -1.5
## Standard Error of simulated Variable Effects (RF):
##   0.204398 0.2246956 0.2135522 0.2199342 0.1705223 .
## Mean of Standard Errors Estimates of Variable Effects (RF):
##   0.2334257 0.2757652 0.2404774 0.2482287 0.2187202 .
## Number of Smaller Nulls:
##   53 35 41 43 43
```

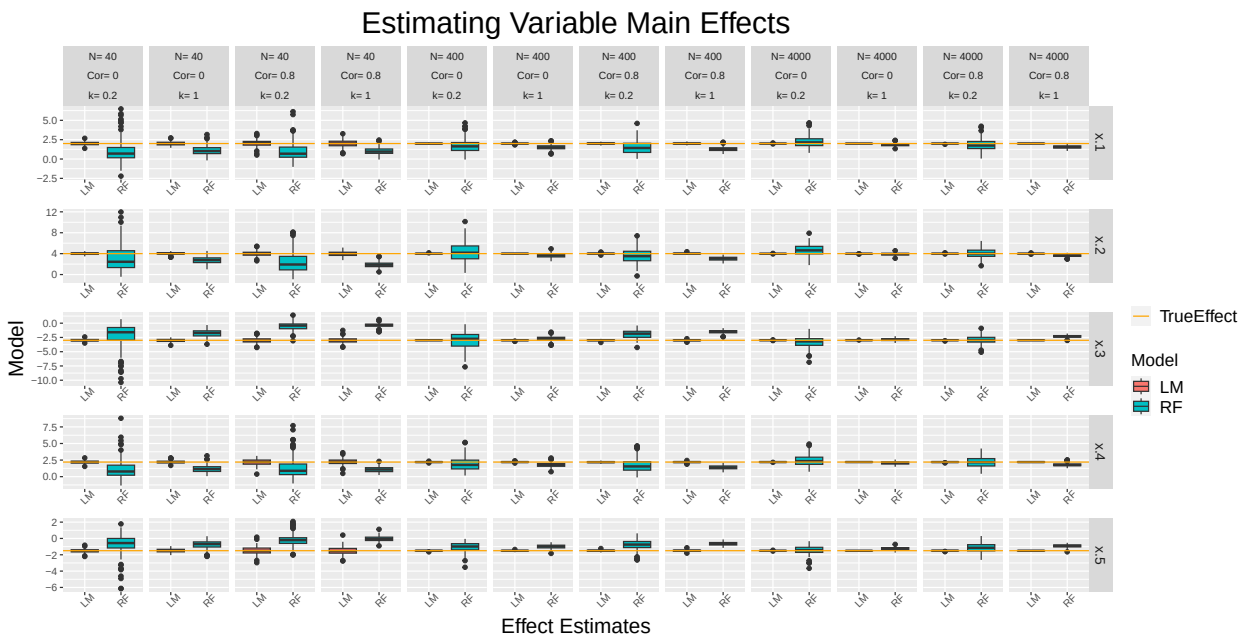
```
effect_plots <- plot_effects(result)
```

```
## `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)
```

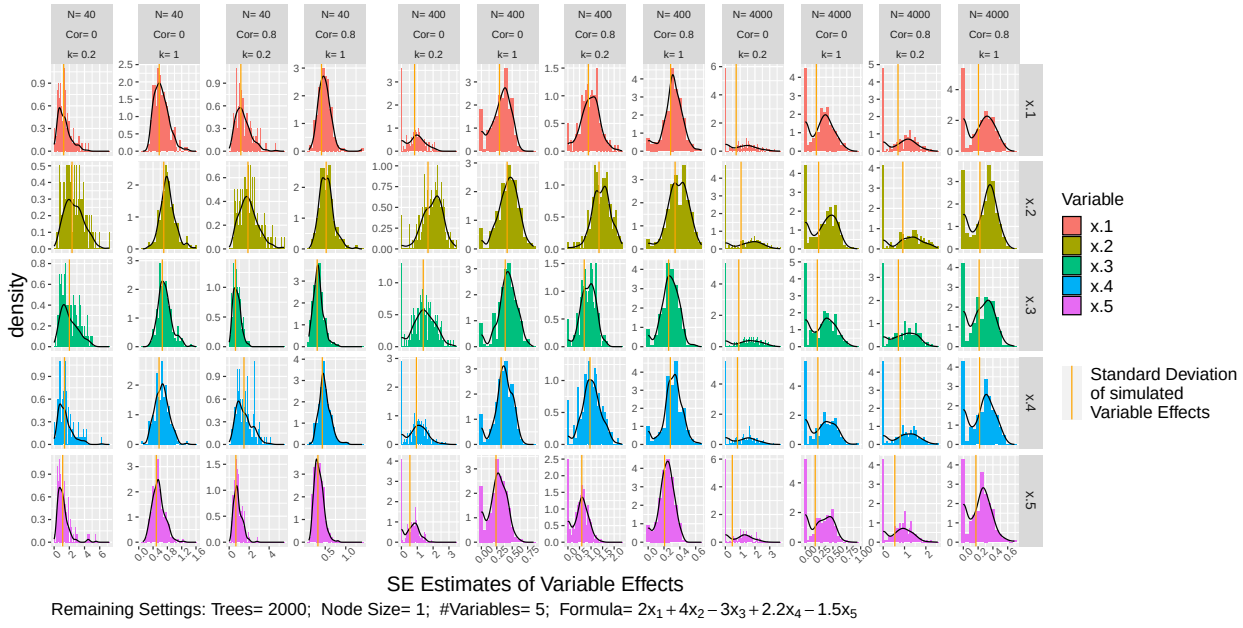
```
## `summarise()` has grouped output by 'N', 'cor', 'k', 'num.trees', 'node_size',
## 'variable'. You can override using the `.groups` argument.
```

```
effect_plots
```



Remaining Settings: Trees= 2000; Node Size= 1; #Variables= 5; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - 1.5x_5$

```
se_plot
```



Remaining Settings: Trees= 2000; Node Size= 1; #Variables= 5; Formula= $2x_1 + 4x_2 - 3x_3 + 2.2x_4 - 1.5x_5$