Estimate Scaling Parameter in SUCCOTASH

David Gerard

2016-04-28

Abstract

I compare scaled-variance SUCCOTASH to procedures that do not change the variance and to the ad-hoc doubling of the variance. Estimating the scaling parameter results in a slightly anti-conservative procedure. It seems to perform best at $\pi_0 = 0.5$ and not as well as the ad-hoc variance inflation procedures at $\pi_0 = 0.9$ or 1. MSE and AUC for this new procedure are very competitive.

Results

```r
library(knitr)
library(xtable)
library(dplyr)
library(reshape2)
library(ggplot2)
```

To view a description of these simulations and the results when the variance was not-inflated, please see http://dcgerard.github.io/flash_sims/analysis/flashr_v_succ.pdf.

In the plots below, "scale_succ1" refers to estimating the scaling parameter starting at a value of 1. "scale_succ2" is when the optimization starts at a value of 2. They give identical results.

Estimates of π_0 when estimating the scaling parameter are slightly anti-conservative. It seems that estimating the scaling parameter performs the best when $\pi_0 = 0.5$, but not as well as just doubling the variance when $\pi_0 = 0.9$ or 1.

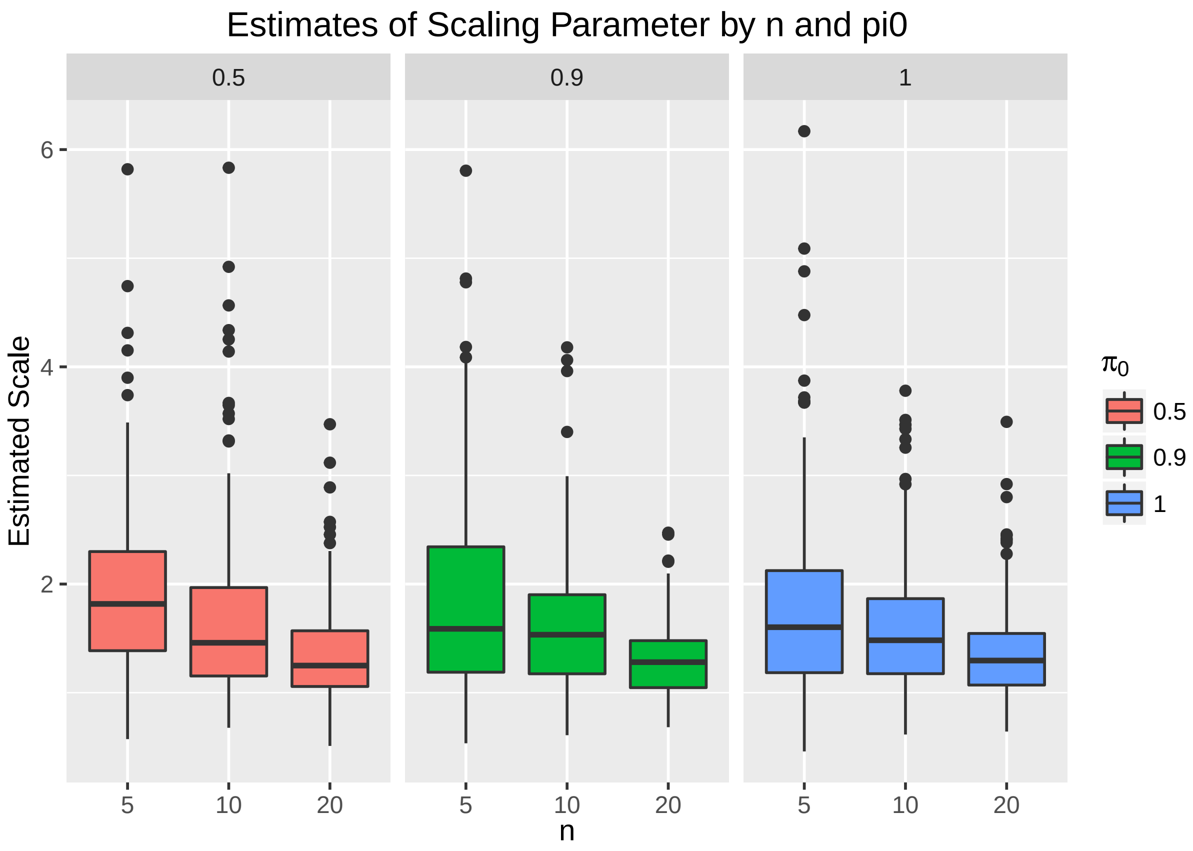
MSE and AUC when estimating the scaling parameter are as good or better than every other method.

Why does inflating the variance by 2 seem to work so well for this data set? On average, the variance was always inflated. For small $n$, this inflation was actually very nearly 2. Perhaps 2 isn't optimal, but better than no variance inflation.

```r
scale_est_mat <- tbl_df(read.csv("scale_est_ssuc.csv", header = TRUE))
kable(aggregate(scale_suc1 ~ nullpi + nsamp, FUN = mean, data = scale_est_mat),
      col.names = c("$\\pi_0$", "$n$", "Mean Estimated Scale"), digits = 1)
```

π_0	$n$	Mean Estimated Scale
0.5	5	1.9
0.9	5	1.8
1.0	5	1.8
0.5	10	1.7
0.9	10	1.6
1.0	10	1.6
0.5	20	1.3
0.9	20	1.3
1.0	20	1.4

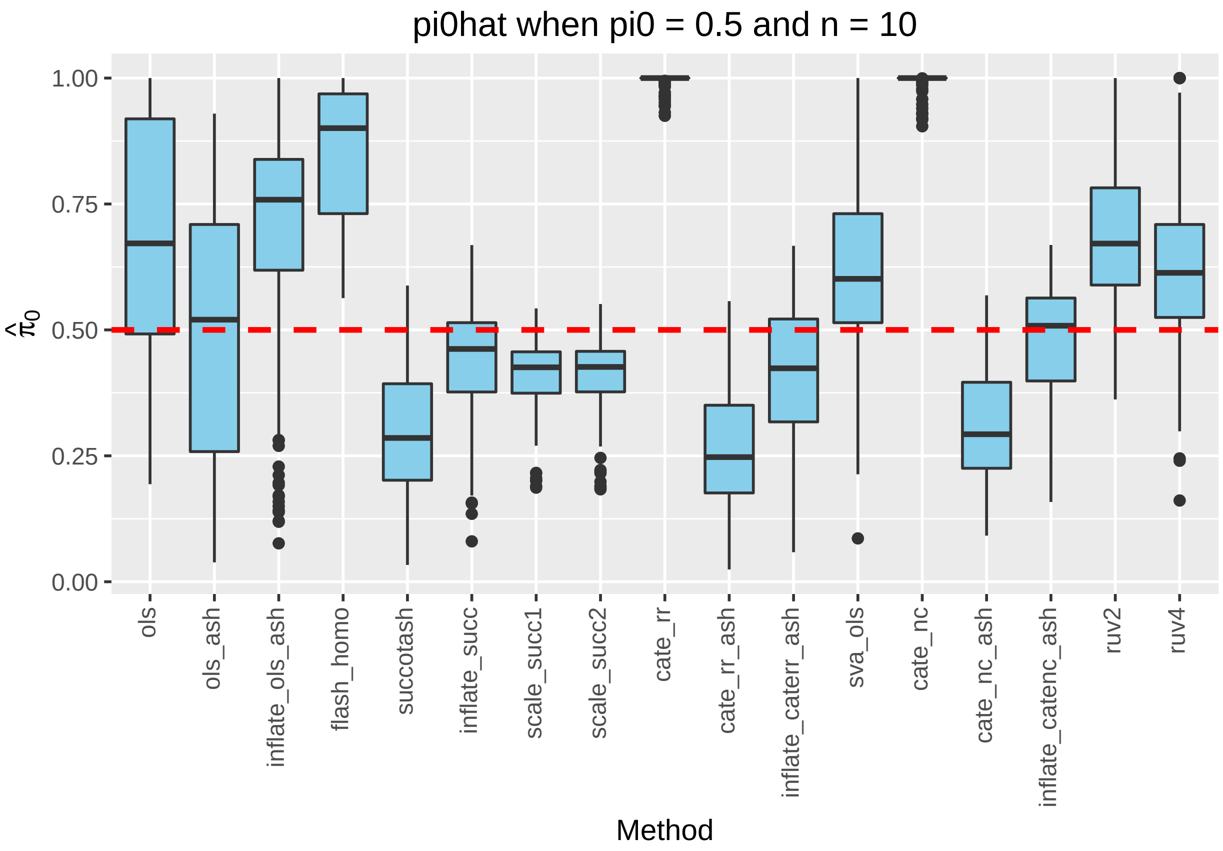
```
ggplot(data = scale_est_mat, mapping = aes(x = factor(nsamp), y = scale_suc1,
                                           fill = factor(nullpi))) +
    facet_grid(.~nullpi) +
    geom_boxplot() +
    xlab(expression(n)) + ylab("Estimated Scale") +
    scale_fill_discrete(name=expression(pi[0])) +
    ggtitle("Estimates of Scaling Parameter by n and pi0")
```

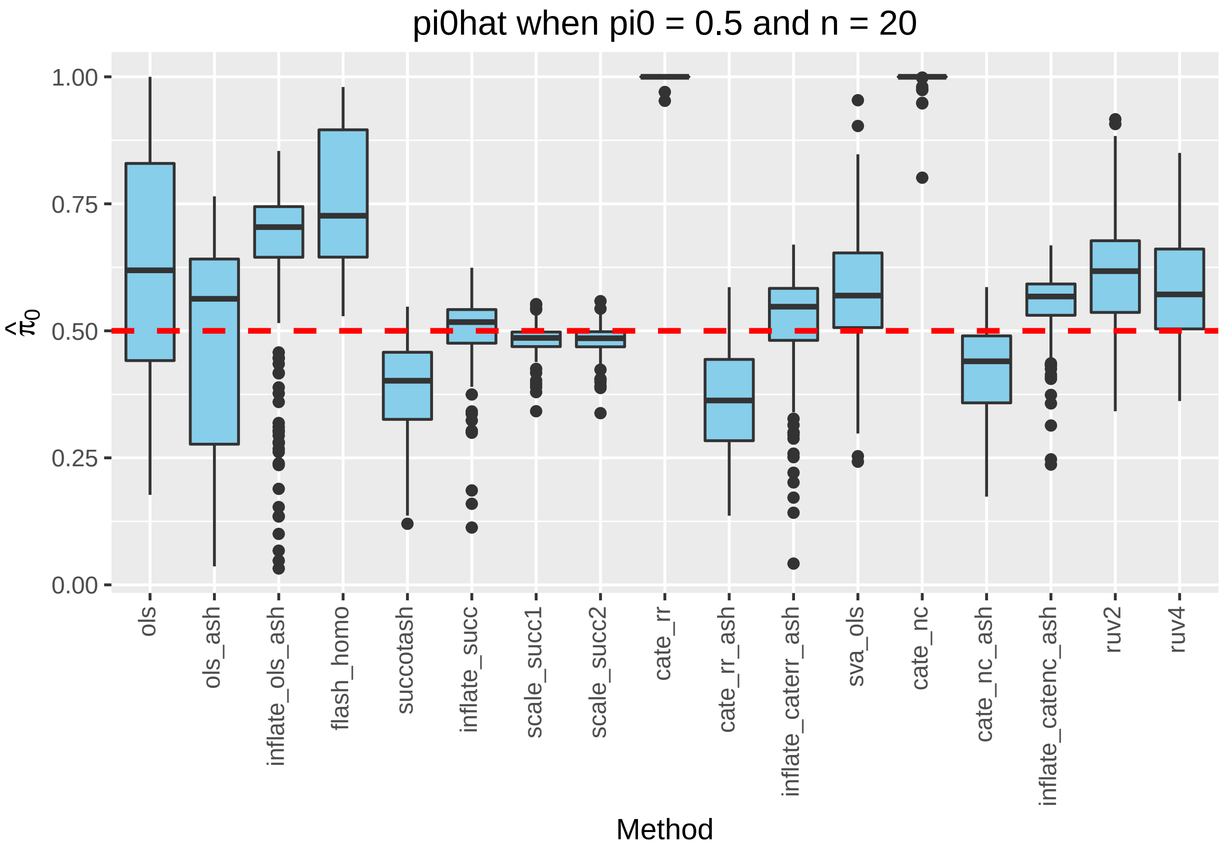


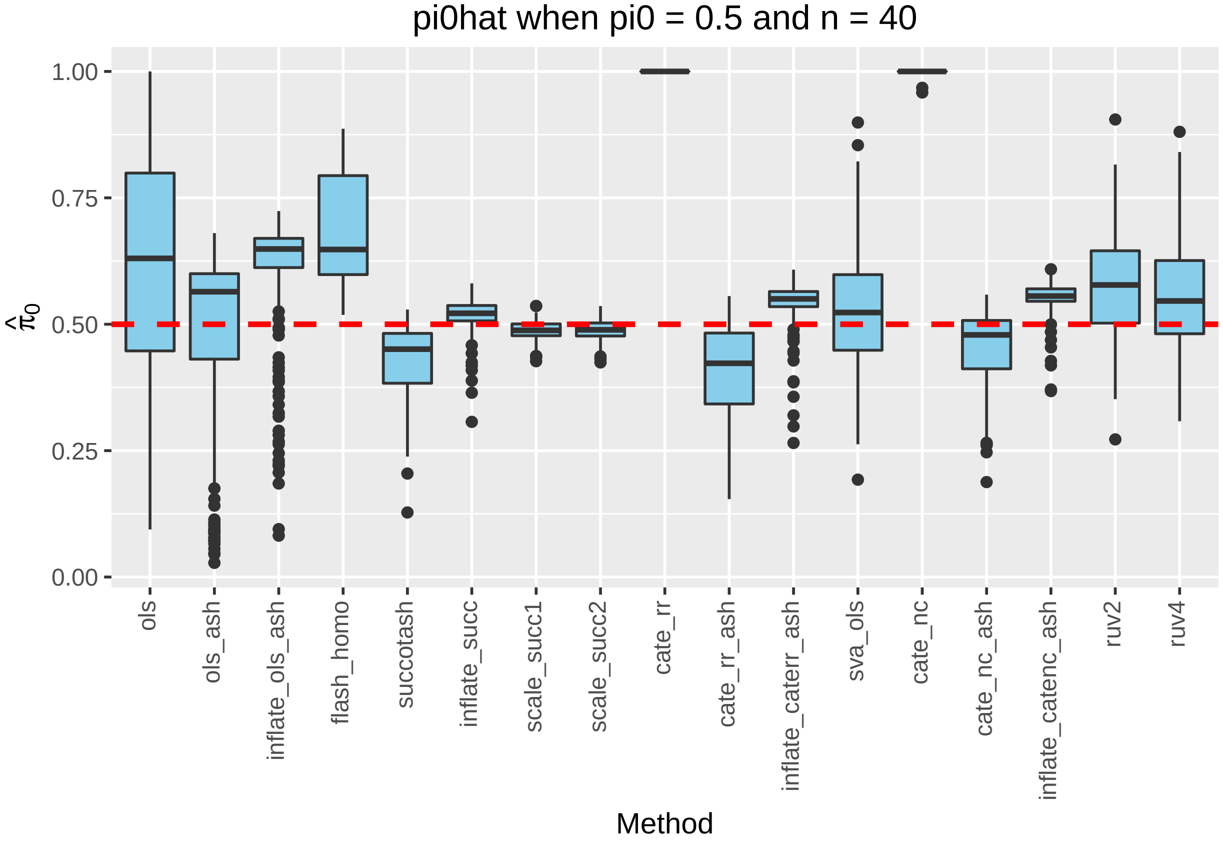
$\hat{\pi}_0$ Plots

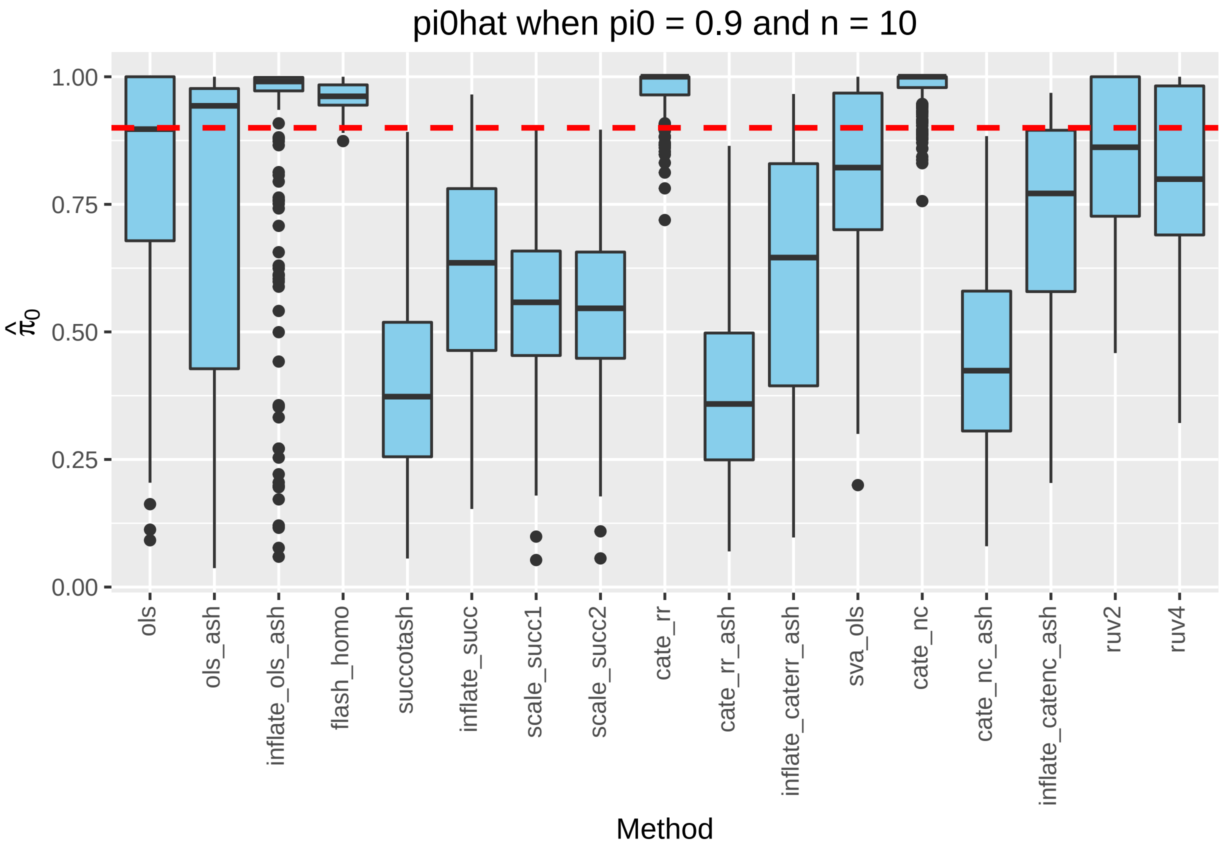
```
double_pi0 <- read.csv("../double_succ/pi0_mat.csv")
reg_pi0 <- read.csv("../flash_v_rest_using_package/pi0_mat.csv")
scale_pi0 <- read.csv("pi0_ssuc.csv")
reg_pi0$inflate_succ <- double_pi0$succotash
reg_pi0$inflate_caterr_ash <- double_pi0$cate_rr_ash
reg_pi0$inflate_catenc_ash <- double_pi0$cate_nc_ash
reg_pi0$inflate_ols_ash <- double_pi0$ols_ash
reg_pi0$scale_succ1 <- scale_pi0$scale_suc1
reg_pi0$scale_succ2 <- scale_pi0$scale_suc2
reg_pi0 <- tbl_df(reg_pi0)
reg_pi0 <- reg_pi0[, c(1:2, 17, 3:4, 14, 18:19, 5:6, 15, 7:9, 16, 10:13)]
nsamp_seq <- unique(reg_pi0$nsamp)
nullpi_seq <- unique(reg_pi0$nullpi)
```

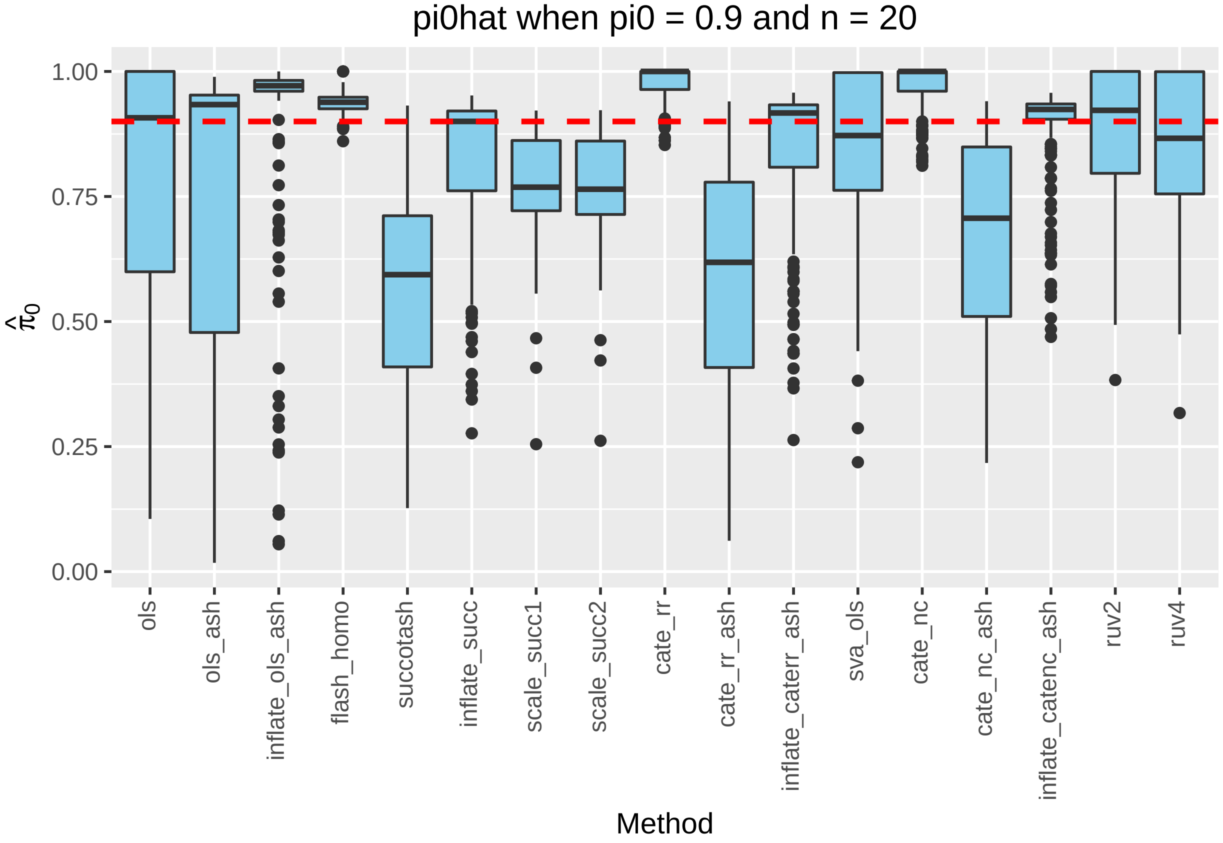
```
for (current_pi in nullpi_seq) {
    for (current_nsamp in nsamp_seq) {

        subdf <- select(
            filter(
                reg_pi0, nullpi == current_pi & nsamp == current_nsamp),
            -c(nsamp, nullpi)
        )

        melted_df <- melt(subdf, id.vars = NULL)

        p <- ggplot(data = melted_df, mapping = aes(x = variable, y = value)) +
            geom_boxplot(fill = I("skyblue")) +
            xlab(label = "Method") + ylab(label = expression(hat(pi)[0])) +
            geom_hline(yintercept = current_pi, color = I("red"), lty  = 2, lwd = 1) +
            ggtitle(paste("pi0hat when pi0 =", current_pi, "and n =", current_nsamp * 2)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```

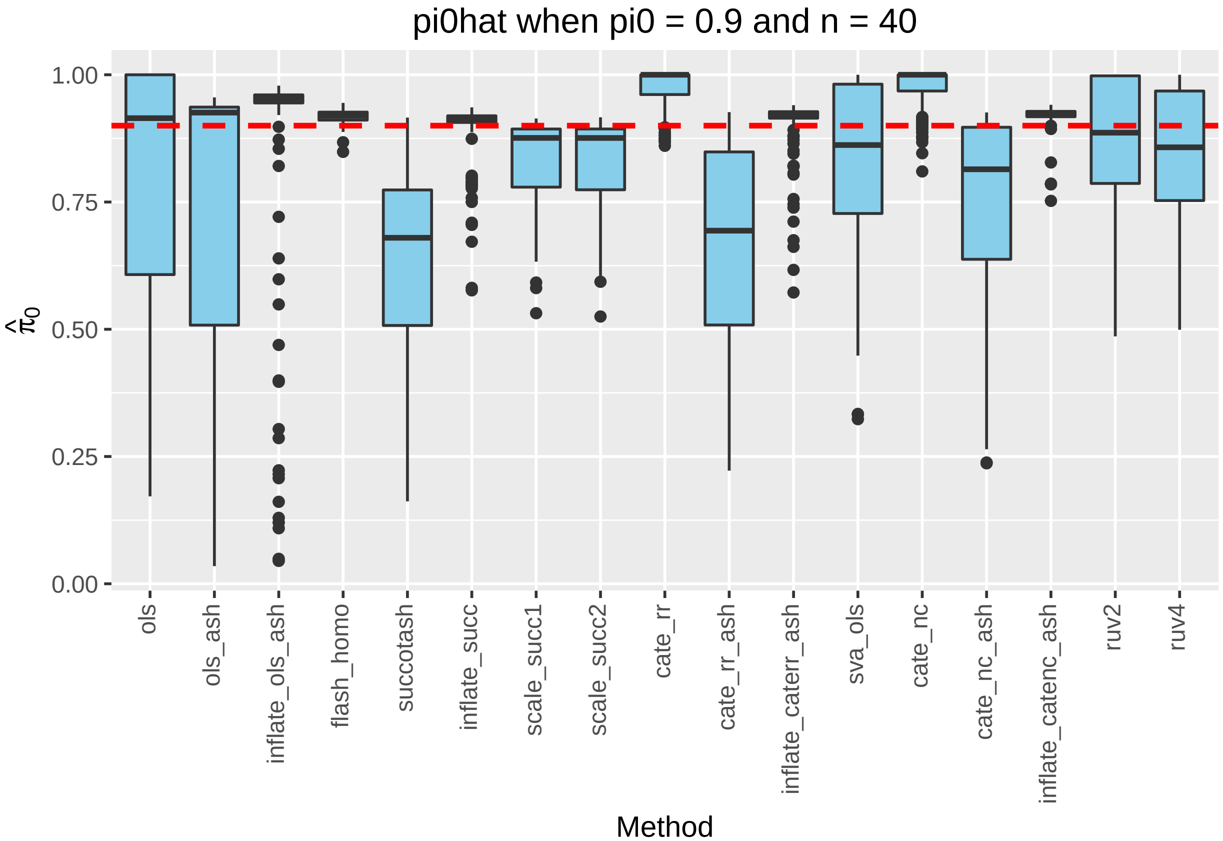


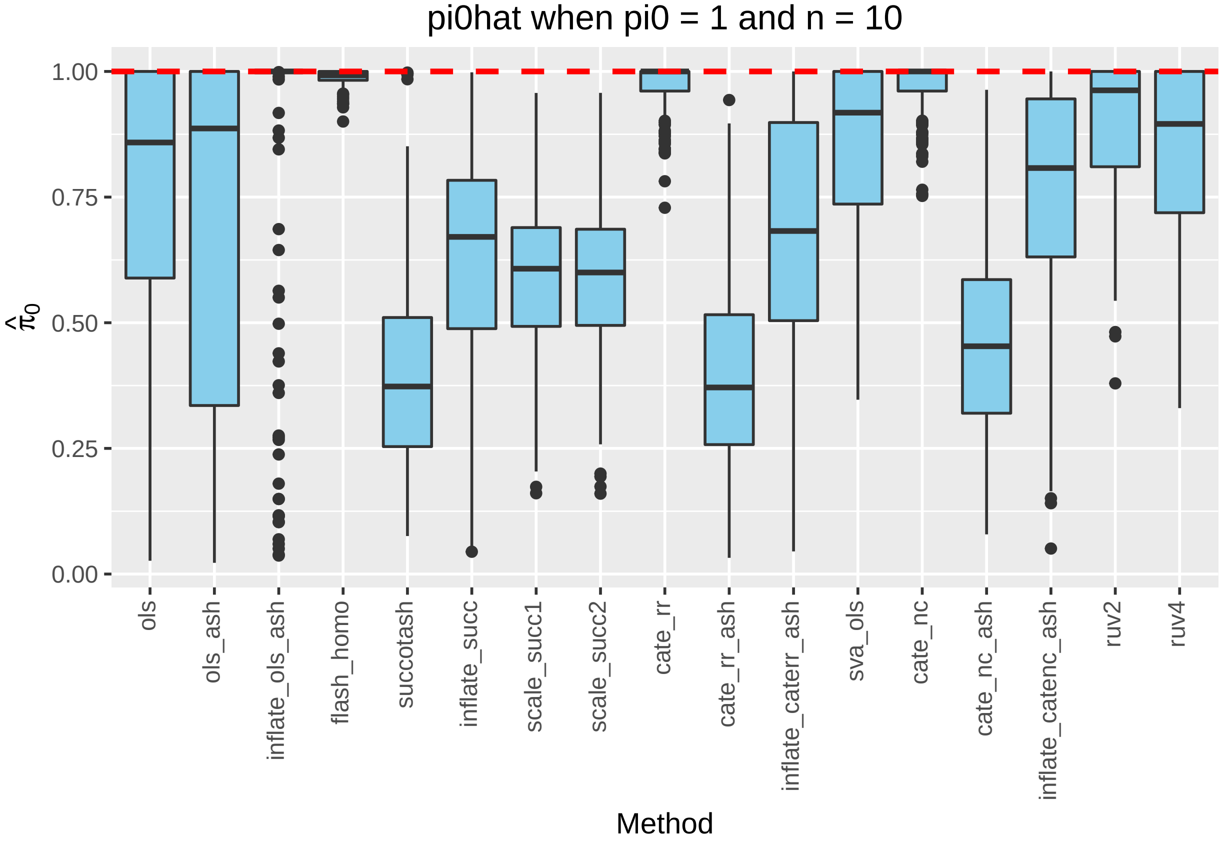

pi0hat when pi0 = 0.5 and n = 20
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

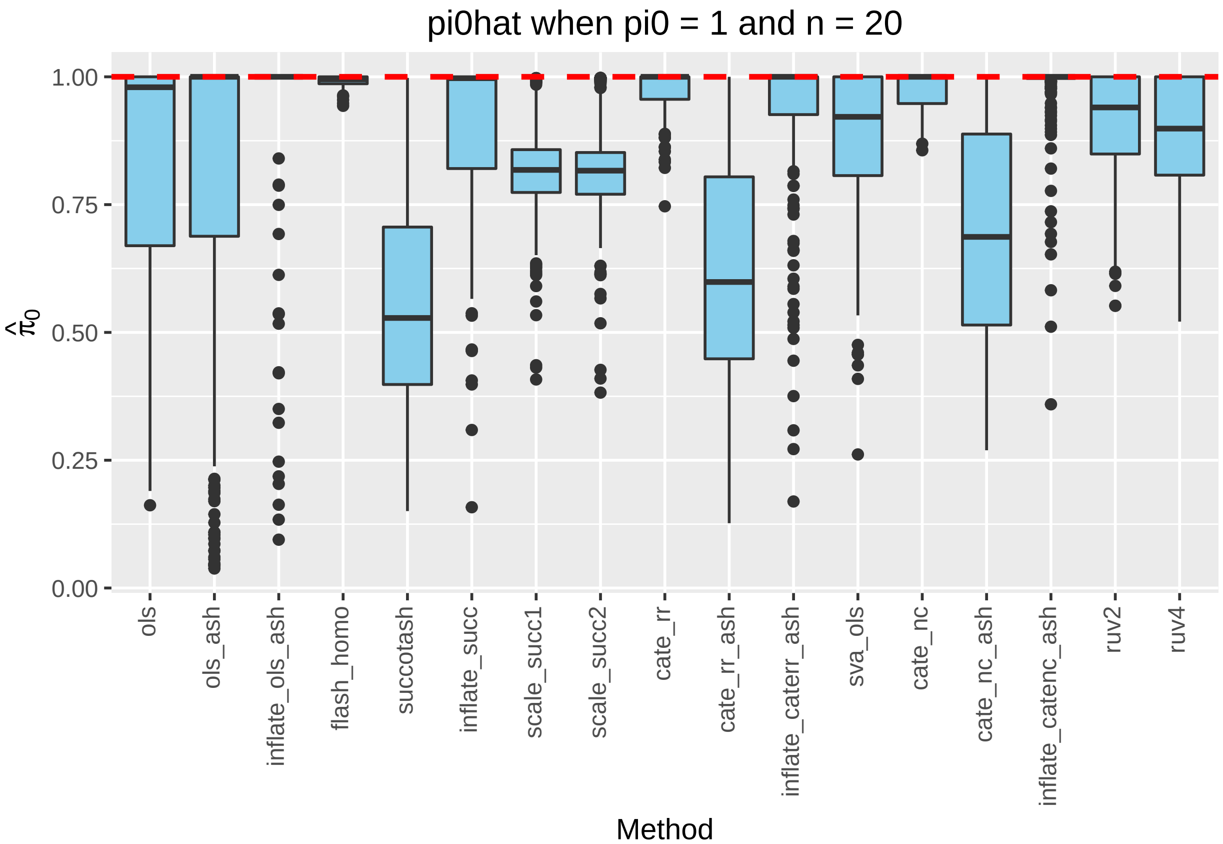

pi0hat when pi0 = 0.5 and n = 40
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

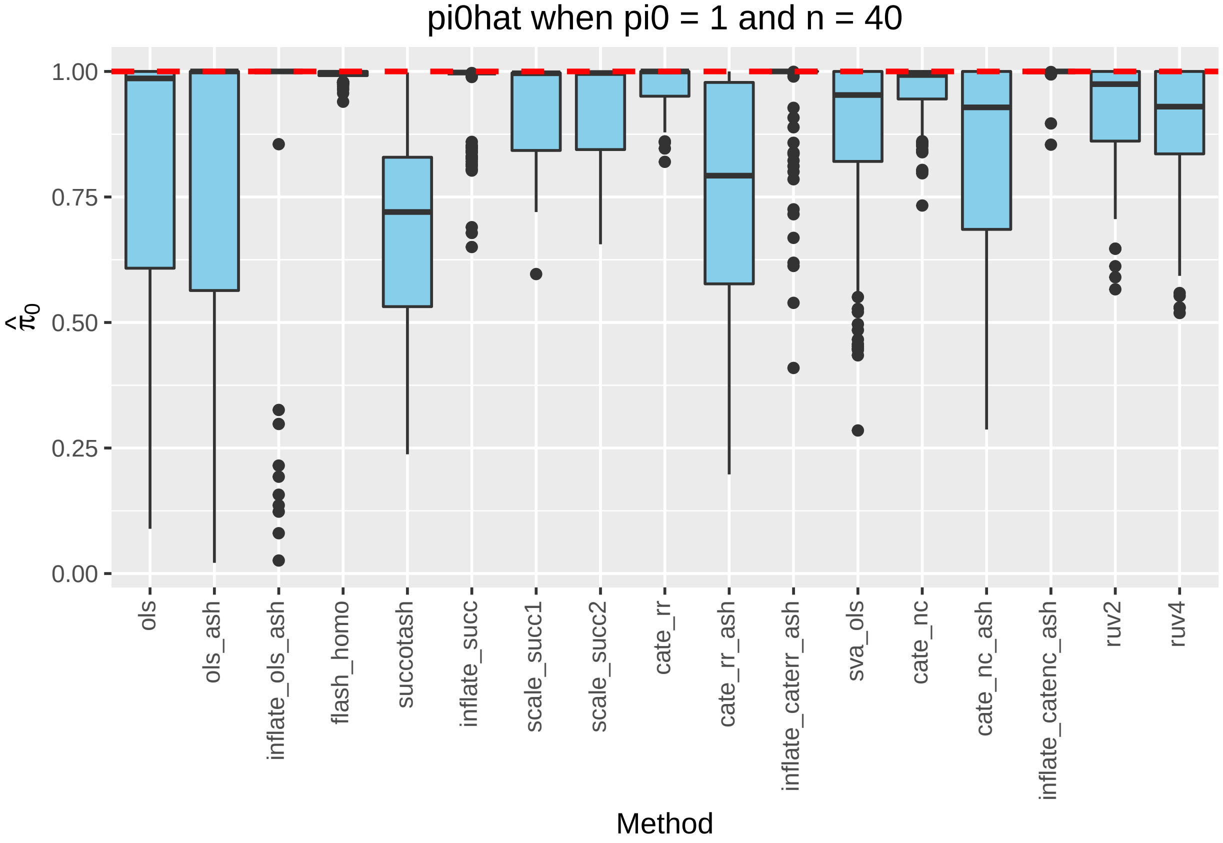

pi0hat when pi0 = 0.9 and n = 10
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


pi0hat when pi0 = 0.9 and n = 20
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


pi0hat when pi0 = 0.9 and n = 40
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

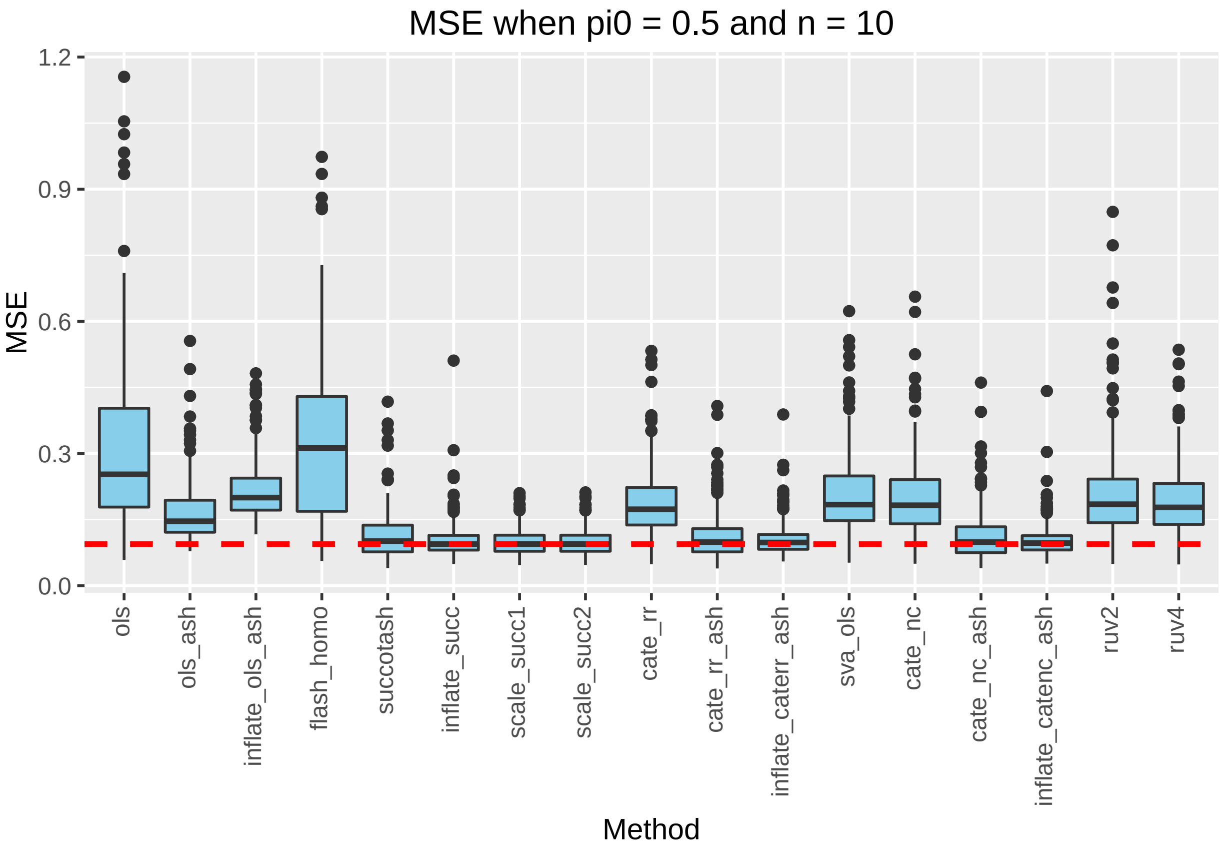

pi0hat when pi0 = 1 and n = 10
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


pi0hat when pi0 = 1 and n = 20
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

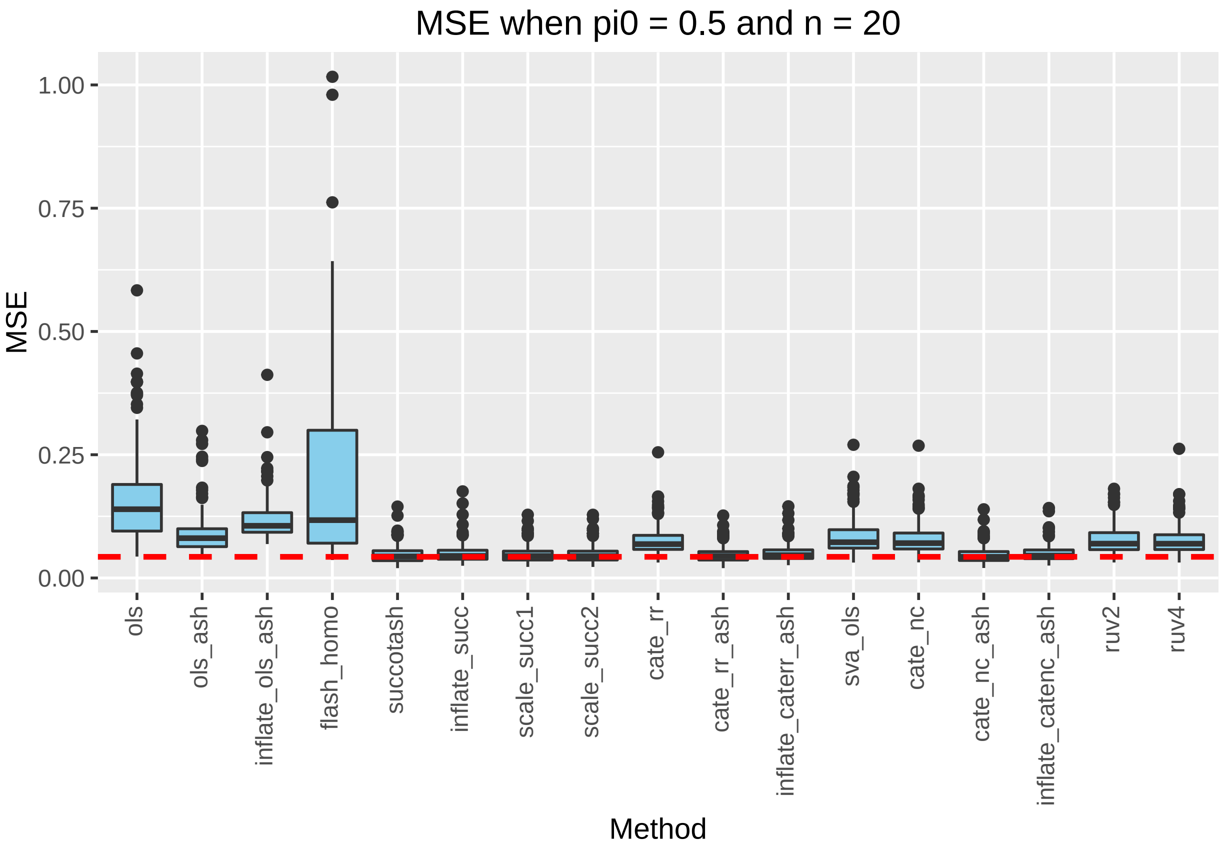

pi0hat when pi0 = 1 and n = 40
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

MSE Plots

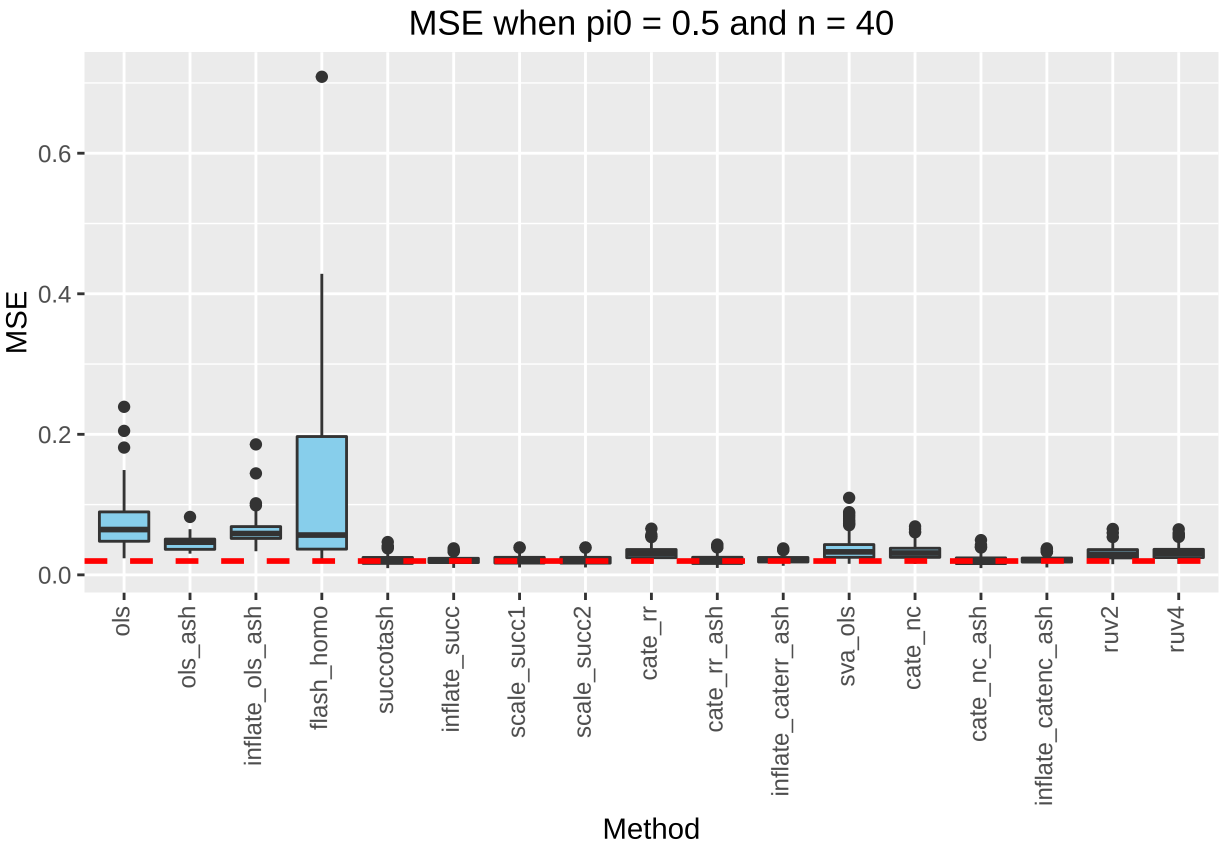
```
double_mse <- read.csv("../double_succ/mse_mat.csv")
reg_mse <- read.csv("../flash_v_rest_using_package/mse_mat.csv")
scale_mse <- read.csv("mse_ssuc.csv")
reg_mse$inflate_succ <- double_mse$succotash
reg_mse$inflate_caterr_ash <- double_mse$cate_rr_ash
reg_mse$inflate_catenc_ash <- double_mse$cate_nc_ash
reg_mse$inflate_ols_ash <- double_mse$ols_ash
reg_mse$scale_succ1 <- scale_mse$scale_suc1
reg_mse$scale_succ2 <- scale_mse$scale_suc2
reg_mse <- tbl_df(reg_mse)
reg_mse <- reg_mse[, c(1:2, 17, 3:4, 14, 18:19, 5:6, 15, 7:9, 16, 10:13)]
nsamp_seq <- unique(reg_mse$nsamp)
nullpi_seq <- unique(reg_mse$nullpi)
for (current_pi in nullpi_seq) {
    for (current_nsamp in nsamp_seq) {

        subdf <- select(
            filter(
                reg_mse, nullpi == current_pi & nsamp == current_nsamp),
            -c(nsamp, nullpi)
        )

        hval <- min(apply(subdf, 2, median))

        melted_df <- melt(subdf, id.vars = NULL)

        p <- ggplot(data = melted_df, mapping = aes(x = variable, y = value)) +
            geom_boxplot(fill = I("skyblue")) +
            xlab(label = "Method") + ylab(label = "MSE") +
            geom_hline(yintercept = hval, color = I("red"), lty  = 2, lwd = 1) +
            ggtitle(paste("MSE when pi0 =", current_pi, "and n =", current_nsamp * 2)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```



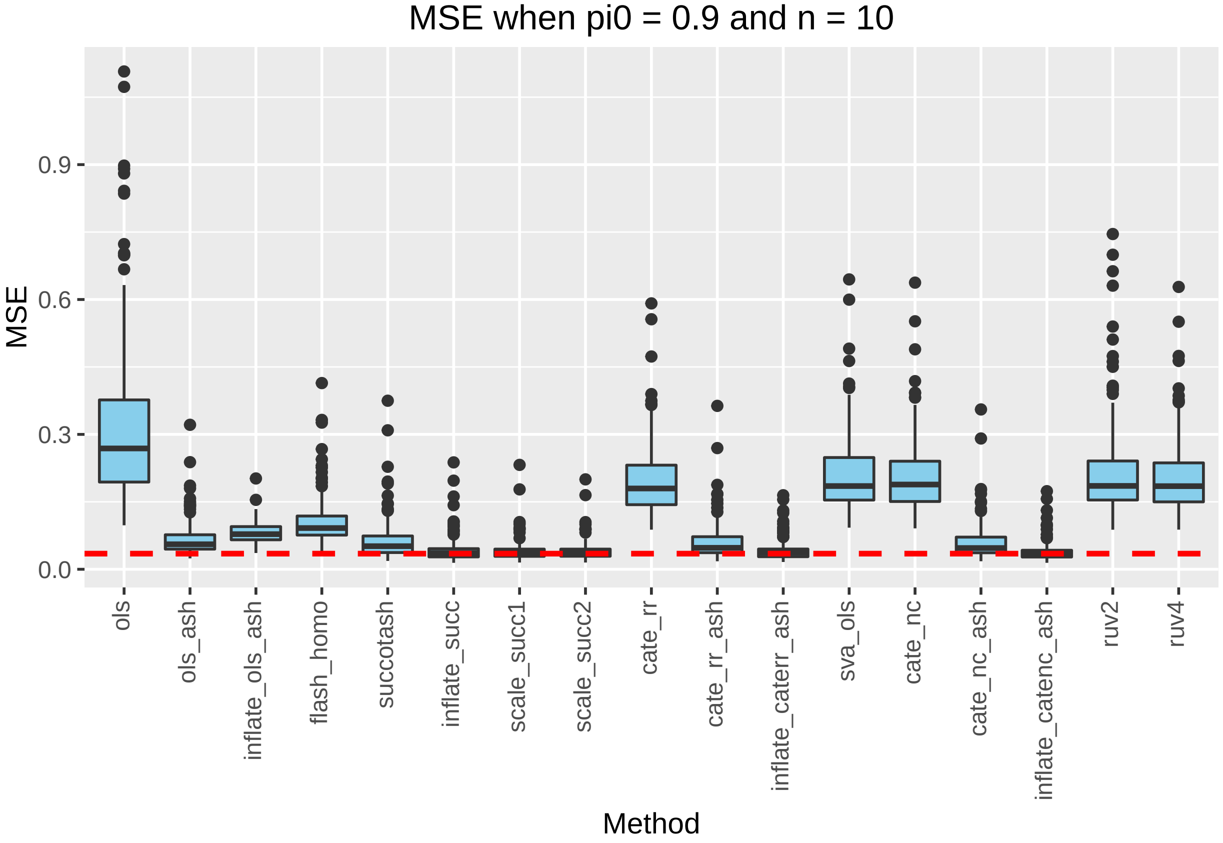
```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



```
## Warning: Removed 203 rows containing non-finite values (stat_boxplot).
```

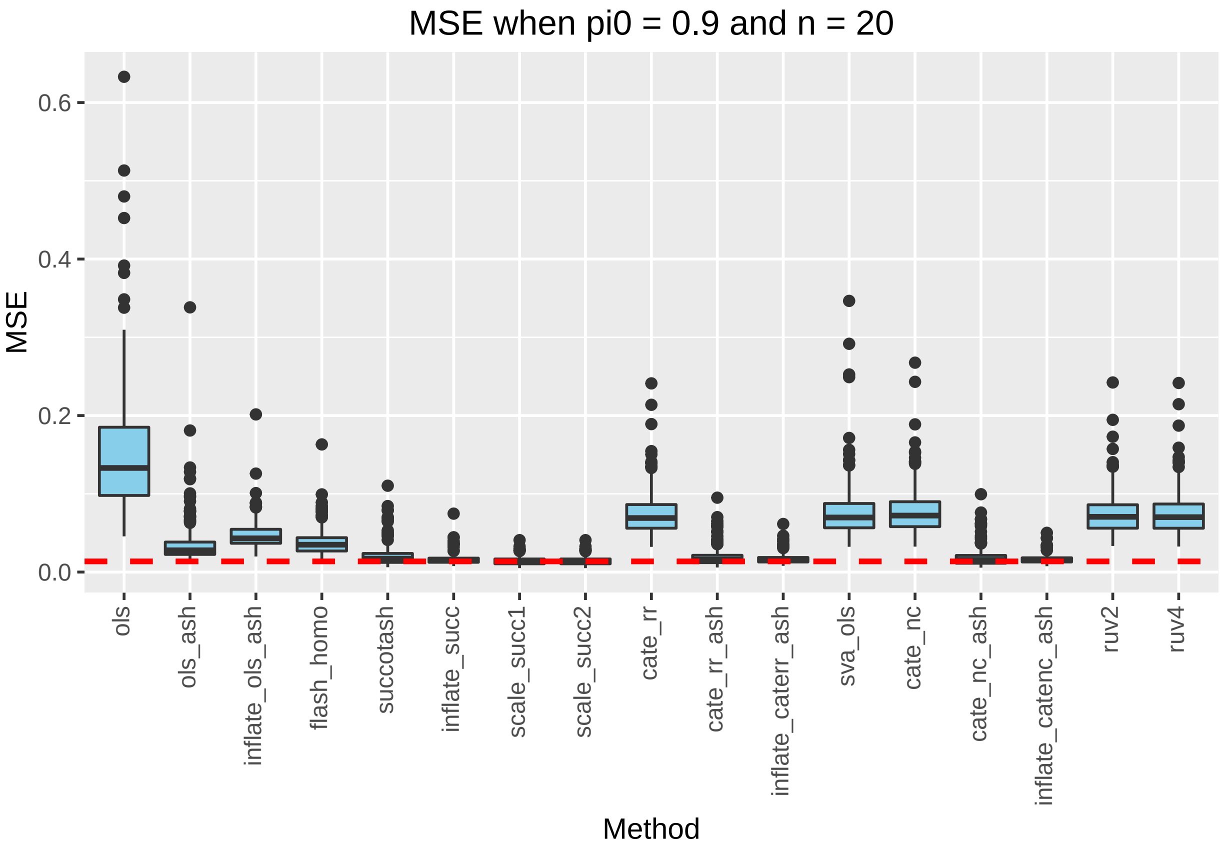


MSE when pi0 = 0.5 and n = 40

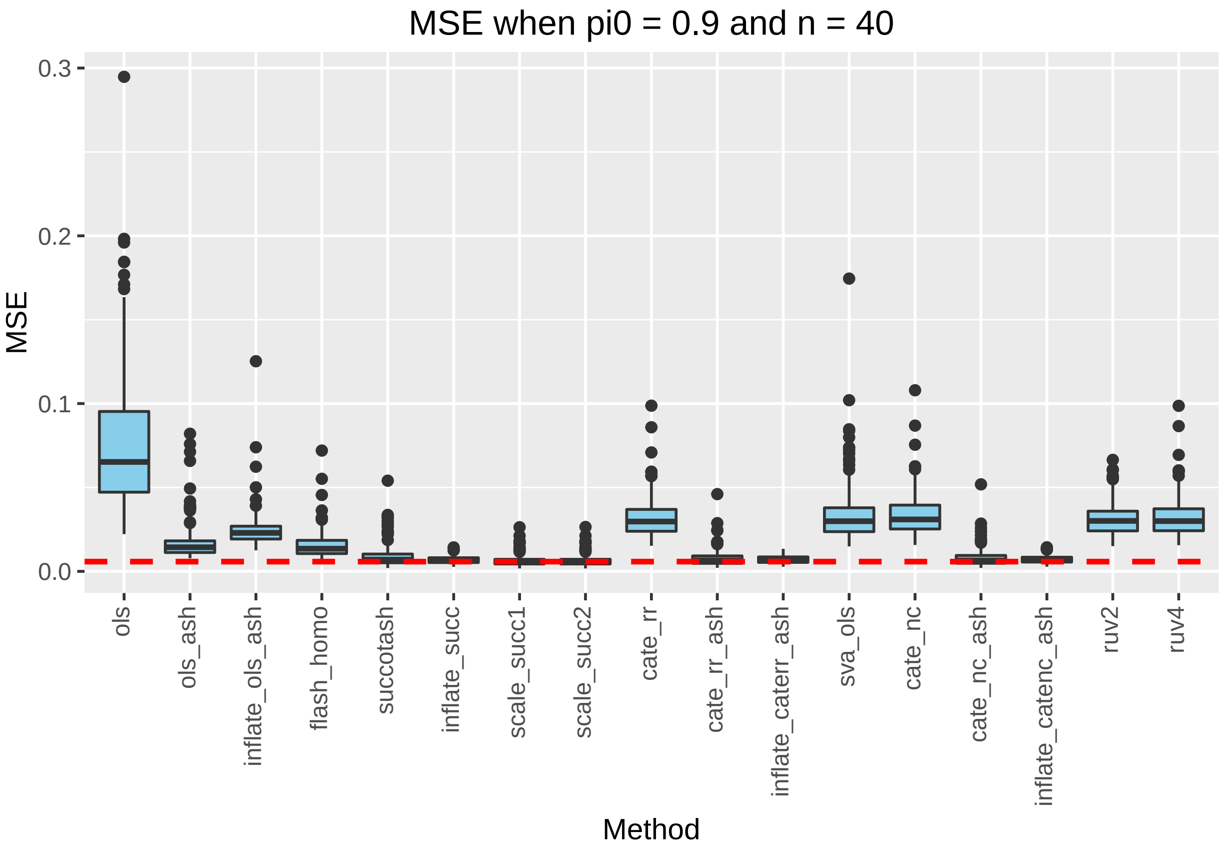


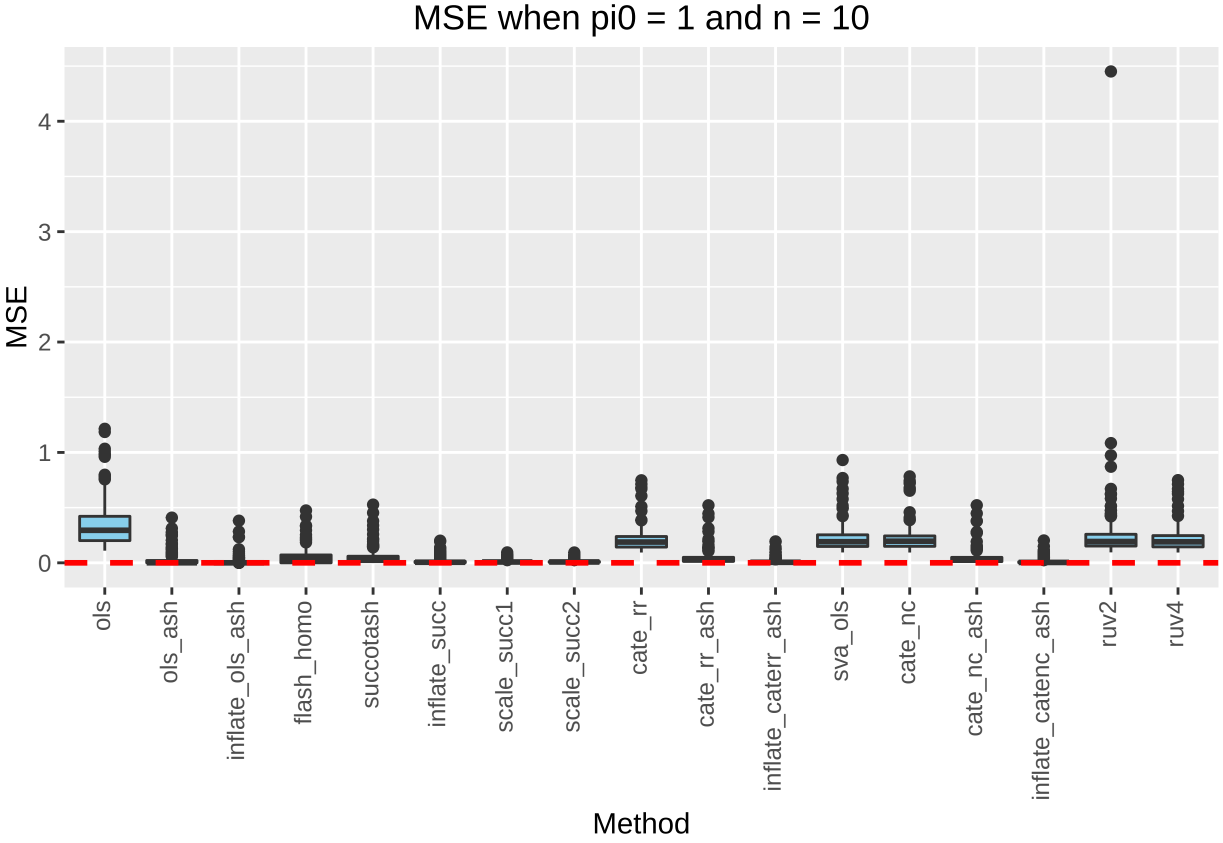
MSE when pi0 = 0.9 and n = 10

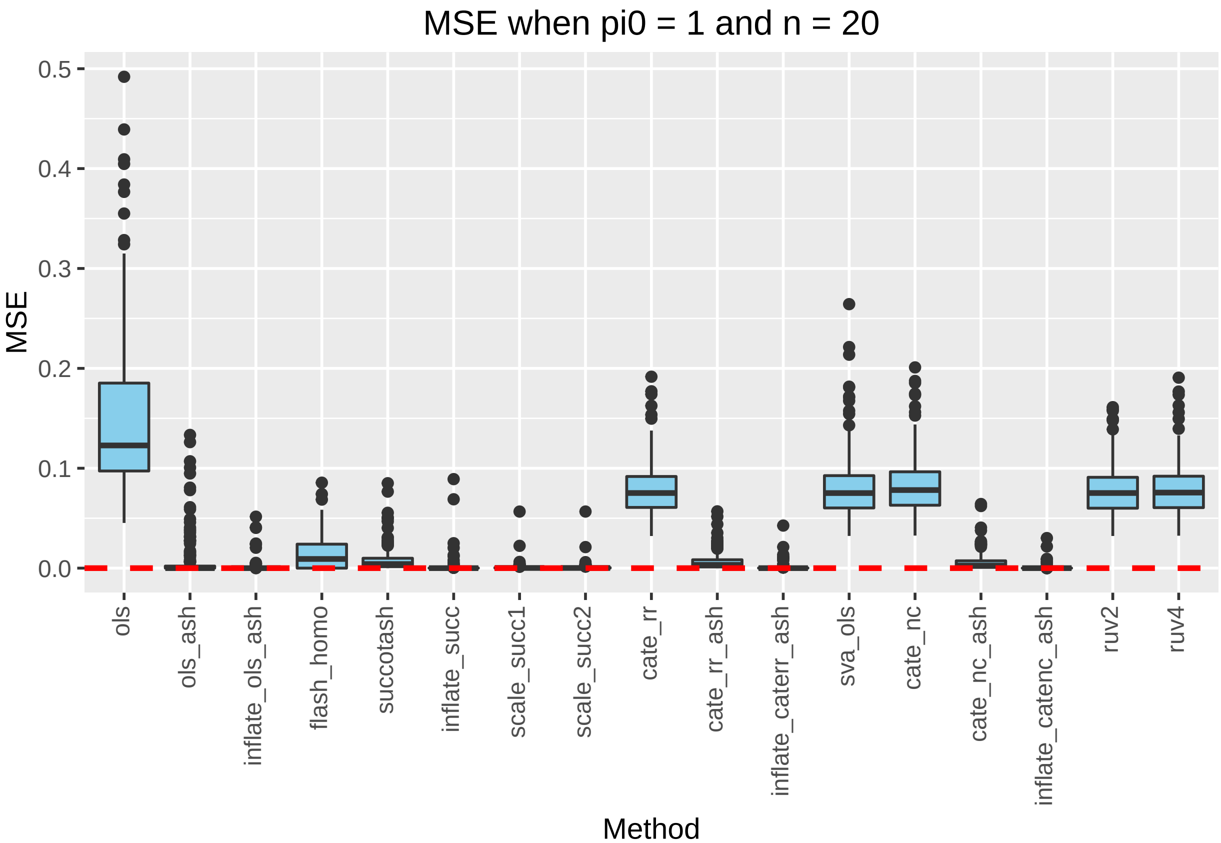
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

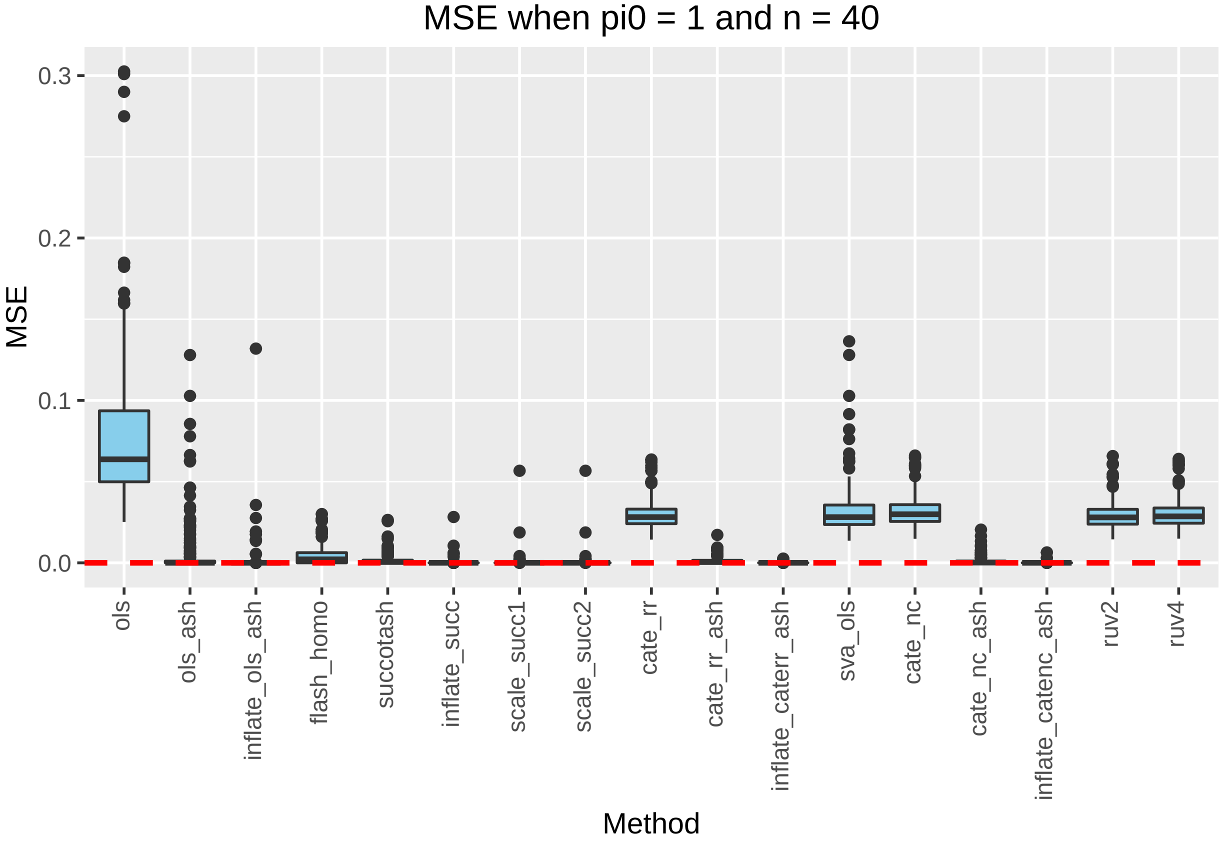


```
## Warning: Removed 89 rows containing non-finite values (stat_boxplot).
```


MSE when pi0 = 0.9 and n = 40
MSE
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

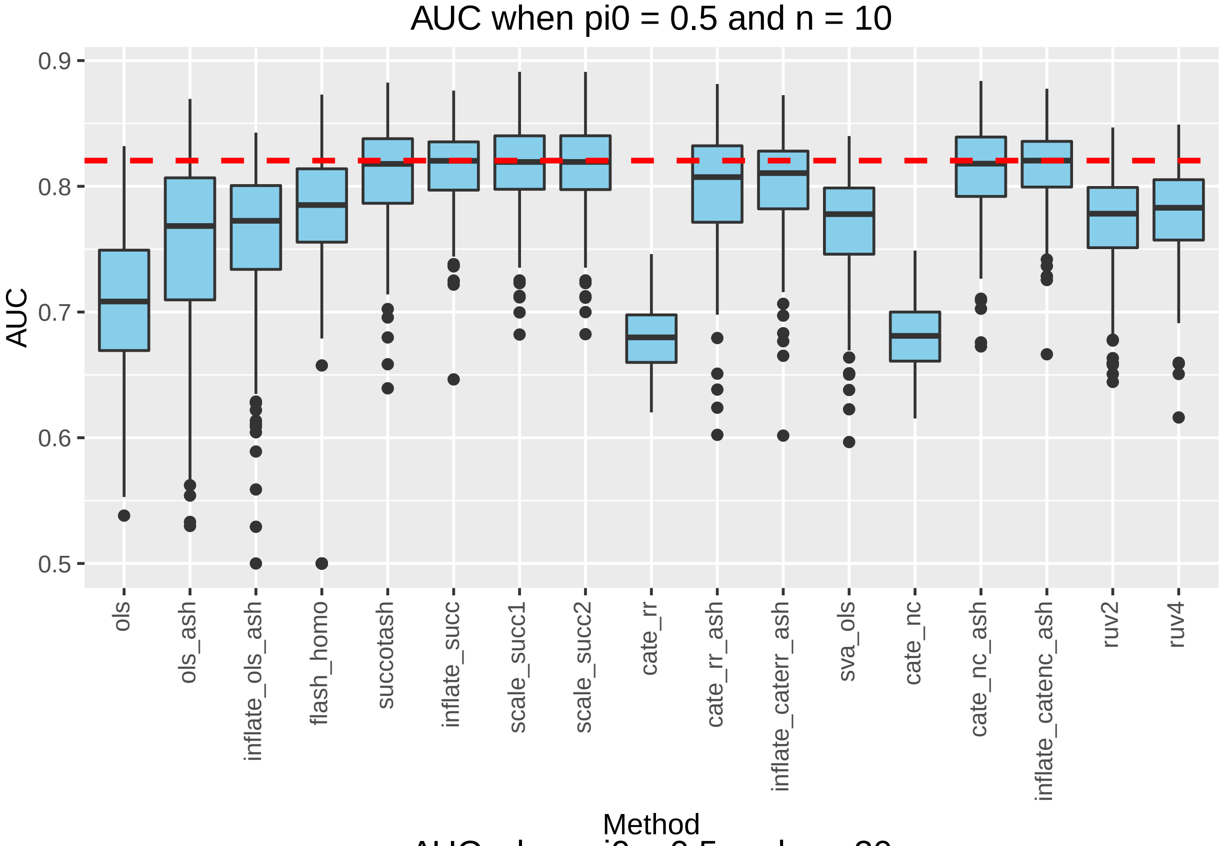

MSE when pi0 = 1 and n = 10
MSE
4
3
2
1
0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

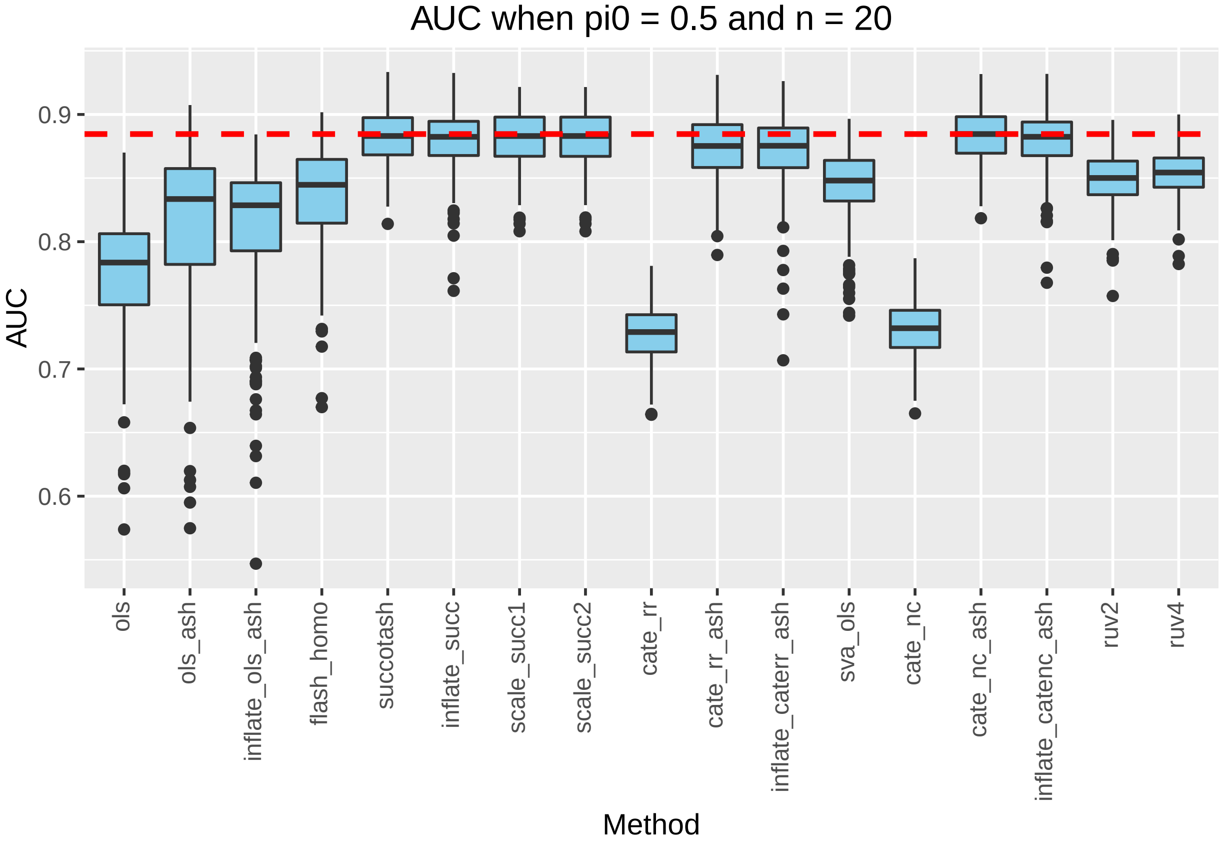

MSE when pi0 = 1 and n = 20
MSE
0.5
0.4
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

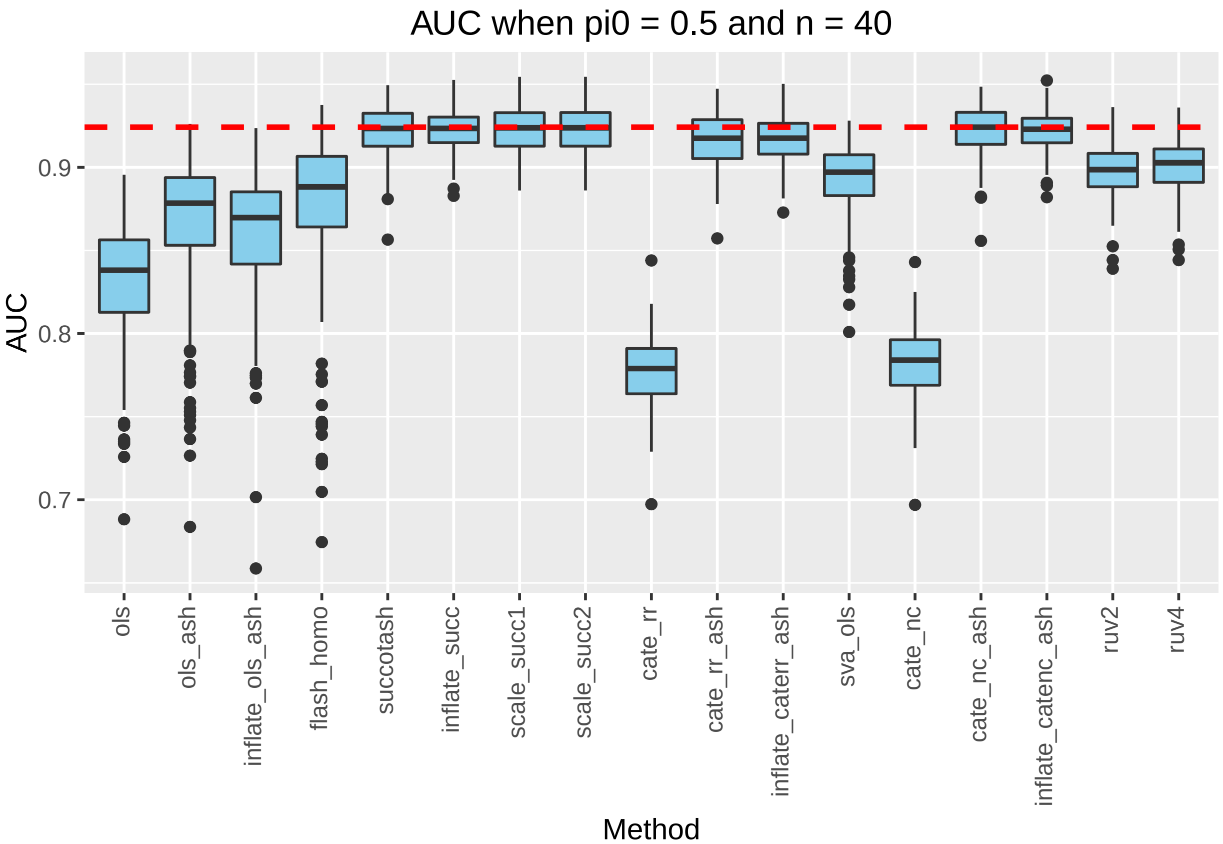

MSE when pi0 = 1 and n = 40
MSE
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

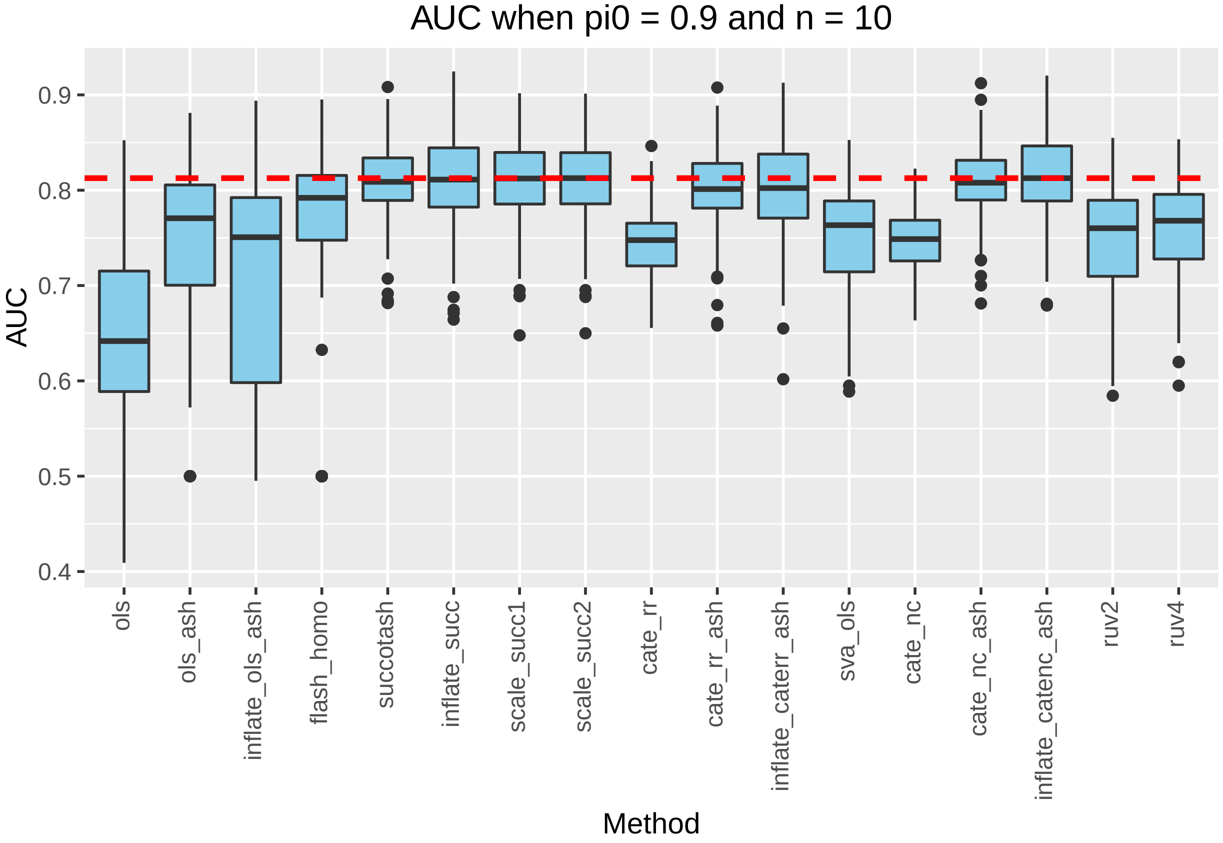
AUC Plots

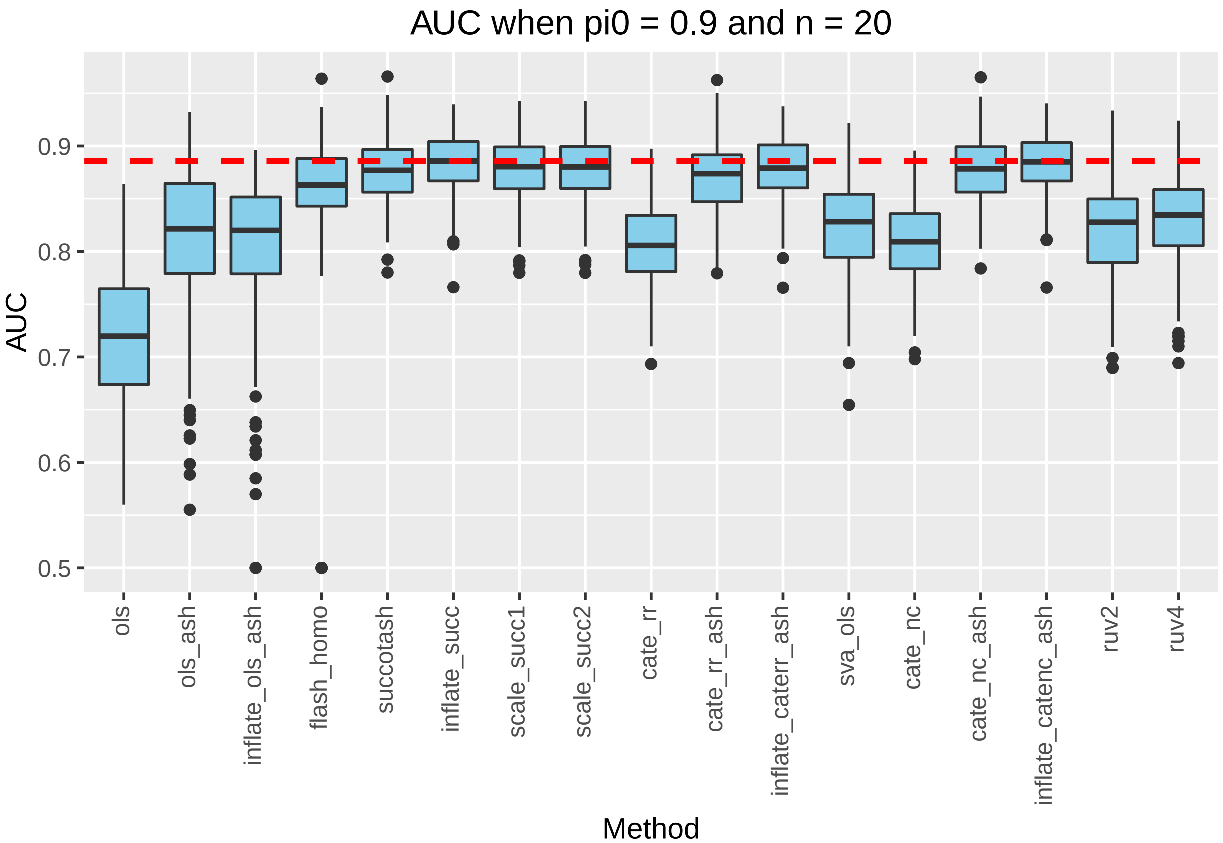
```r
double_auc <- read.csv("../double_succ/auc_mat.csv")
reg_auc <- read.csv("../flash_v_rest_using_package/auc_mat.csv")
scale_auc <- read.csv("auc_ssuc.csv")
reg_auc$inflate_succ <- double_auc$succotash
reg_auc$inflate_caterr_ash <- double_auc$cate_rr_ash
reg_auc$inflate_catenc_ash <- double_auc$cate_nc_ash
reg_auc$inflate_ols_ash <- double_auc$ols_ash
reg_auc$scale_succ1 <- scale_auc$scale_suc1
reg_auc$scale_succ2 <- scale_auc$scale_suc2
reg_auc <- tbl_df(reg_auc)
reg_auc <- reg_auc[, c(1:2, 17, 3:4, 14, 18:19, 5:6, 15, 7:9, 16, 10:13)]
nsamp_seq <- unique(reg_auc$nsamp)
nullpi_seq <- unique(reg_auc$nullpi)
for (current_pi in nullpi_seq) {
    for (current_nsamp in nsamp_seq) {

        subdf <- select(
            filter(
                reg_auc, nullpi == current_pi & nsamp == current_nsamp),
            -c(nsamp, nullpi)
        )

        hval <- max(apply(subdf, 2, median))

        melted_df <- melt(subdf, id.vars = NULL)

        p <- ggplot(data = melted_df, mapping = aes(x = variable, y = value)) +
            geom_boxplot(fill = I("skyblue")) +
            xlab(label = "Method") + ylab(label = "AUC") +
            geom_hline(yintercept = hval, color = I("red"), lty  = 2, lwd = 1) +
            ggtitle(paste("AUC when pi0 =", current_pi, "and n =", current_nsamp * 2)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```


AUC when pi0 = 0.5 and n = 10
AUC
0.9
0.8
0.7
0.6
0.5
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

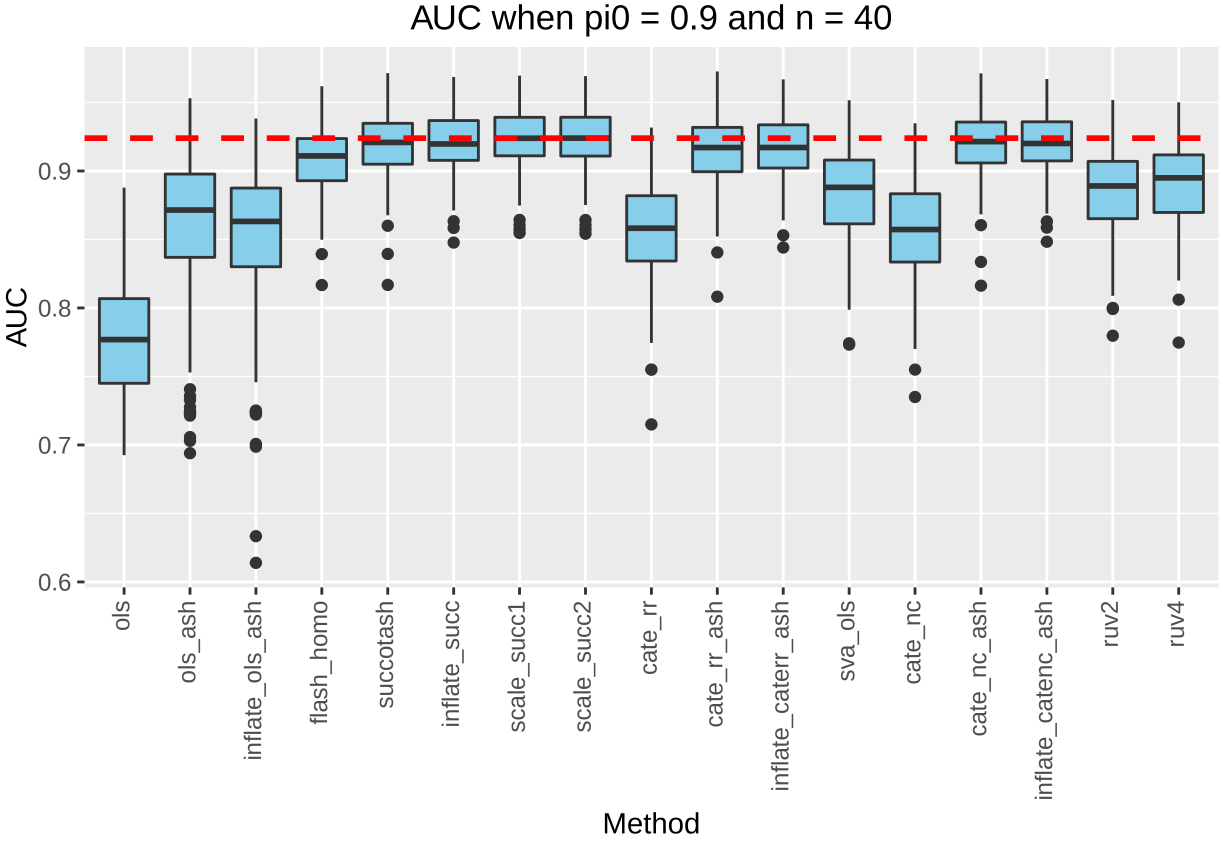

AUC when pi0 = 0.5 and n = 20
AUC
0.9
0.8
0.7
0.6
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


AUC when pi0 = 0.5 and n = 40
AUC
0.7
0.8
0.9
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


AUC when pi0 = 0.9 and n = 10
AUC
0.4
0.5
0.6
0.7
0.8
0.9
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method



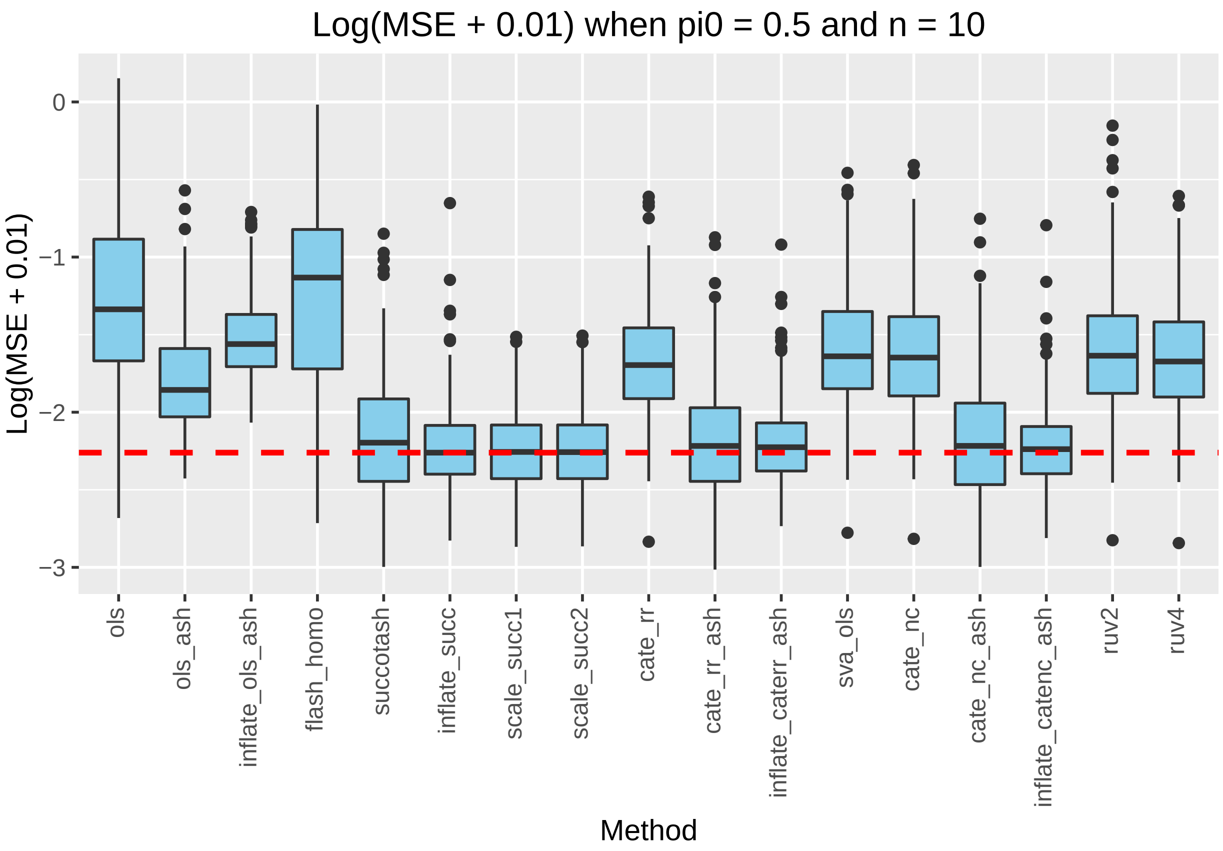
AUC when pi0 = 0.9 and n = 20



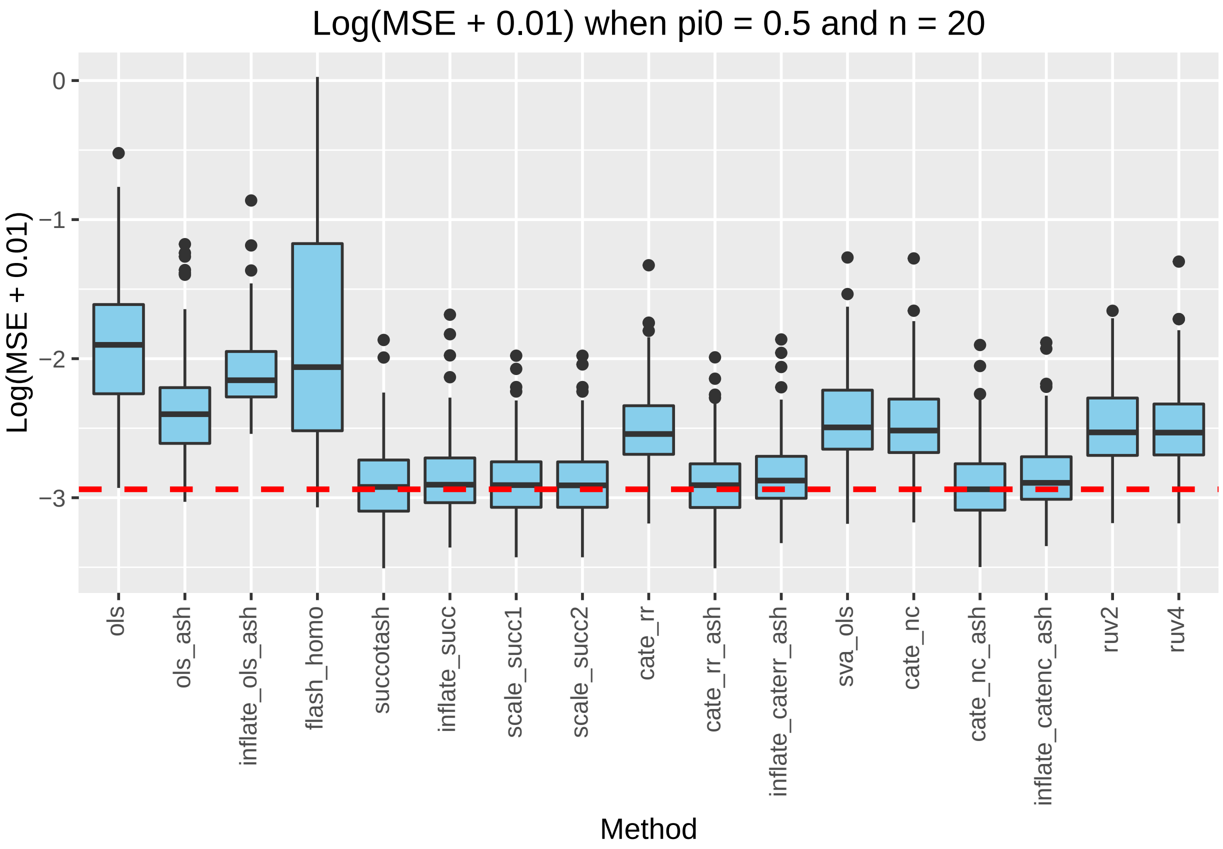
AUC when pi0 = 0.9 and n = 40

Log(MSE + 0.1) Plots

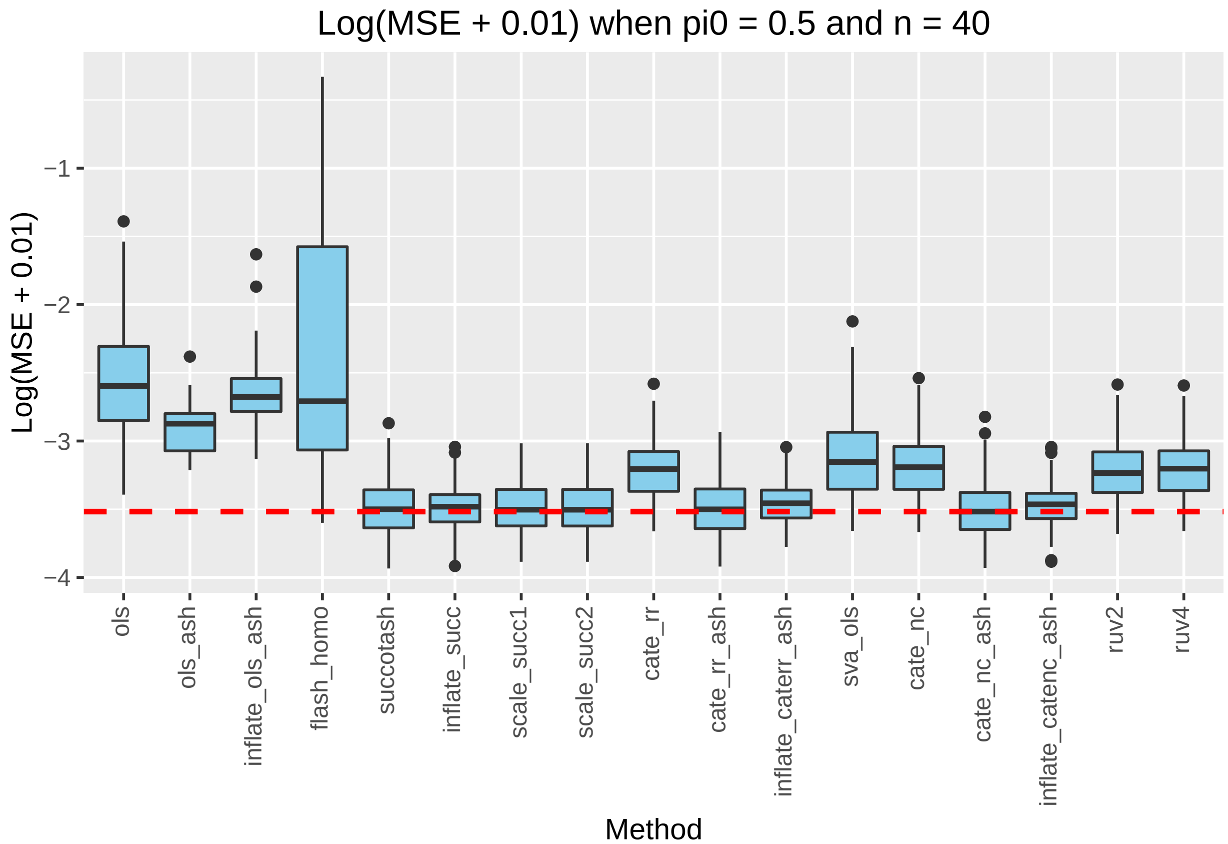
```
nsamp_seq <- unique(reg_mse$nsamp)
nullpi_seq <- unique(reg_mse$nullpi)
for (current_pi in nullpi_seq) {
    for (current_nsamp in nsamp_seq) {

        subdf <- select(
            filter(
                reg_mse, nullpi == current_pi & nsamp == current_nsamp),
            -c(nsamp, nullpi)
        )

        hval <- min(apply(log(subdf + 0.01), 2, median))

        melted_df <- melt(subdf, id.vars = NULL)

        p <- ggplot(data = melted_df, mapping = aes(x = variable, y = log(value + 0.01))) +
            geom_boxplot(fill = I("skyblue")) +
            xlab(label = "Method") + ylab(label = "Log(MSE + 0.01)") +
            geom_hline(yintercept = hval, color = I("red"), lty  = 2, lwd = 1) +
            ggtitle(paste("Log(MSE + 0.01) when pi0 =", current_pi, "and n =", current_nsamp * 2)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```

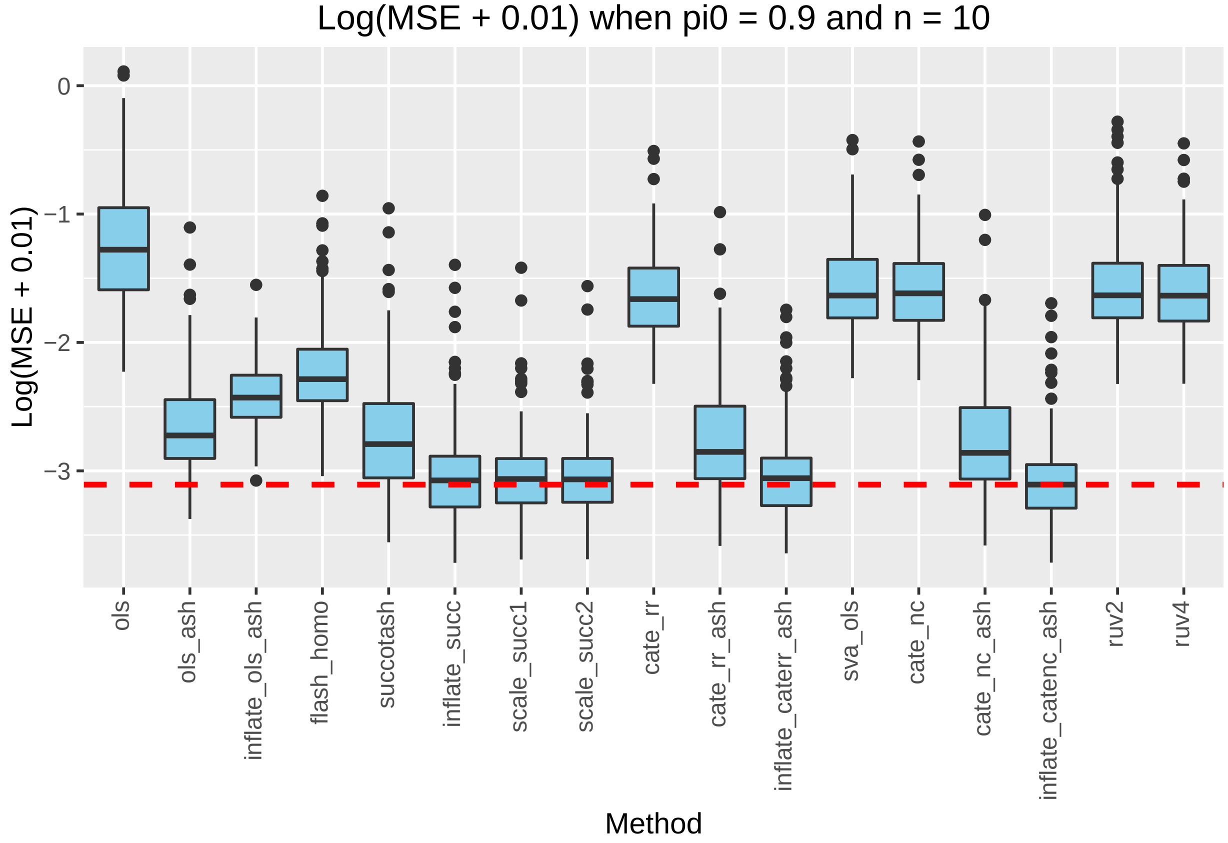


```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

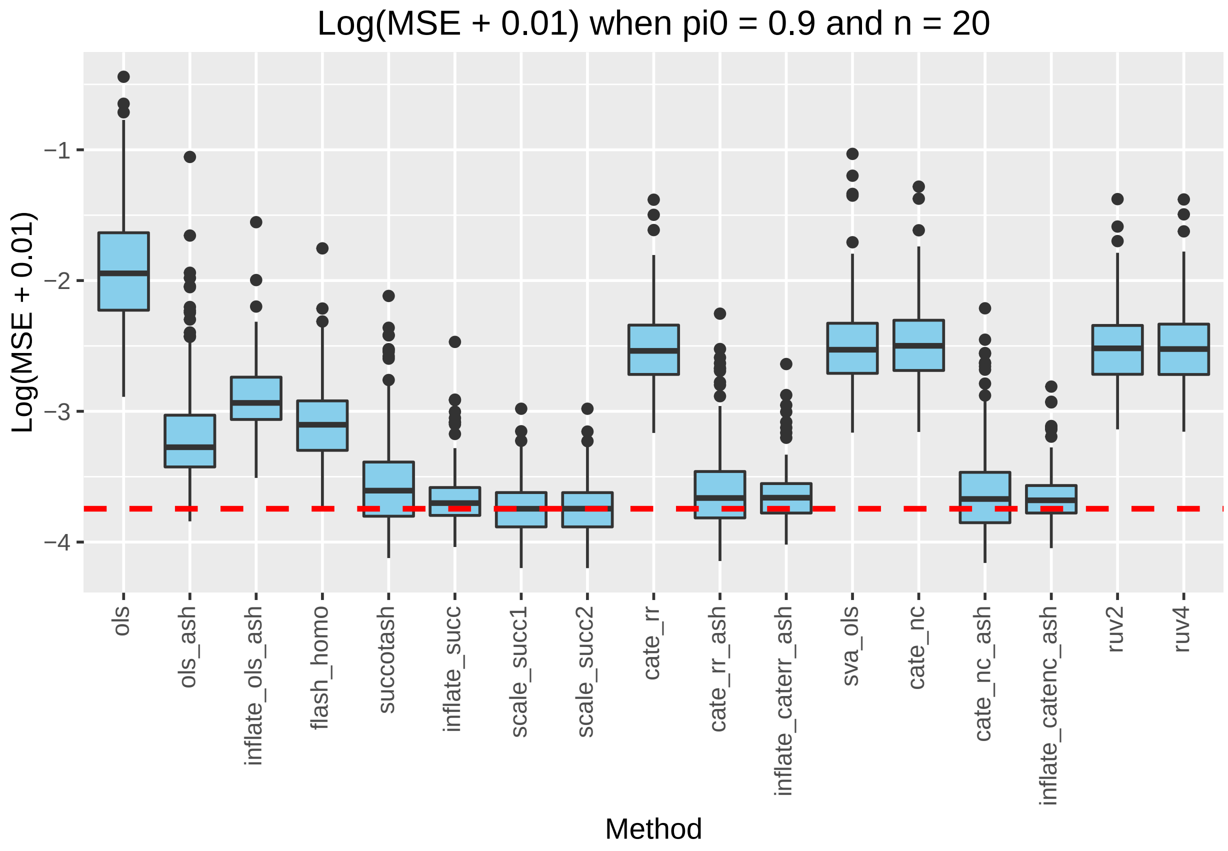


```
## Warning: Removed 203 rows containing non-finite values (stat_boxplot).
```

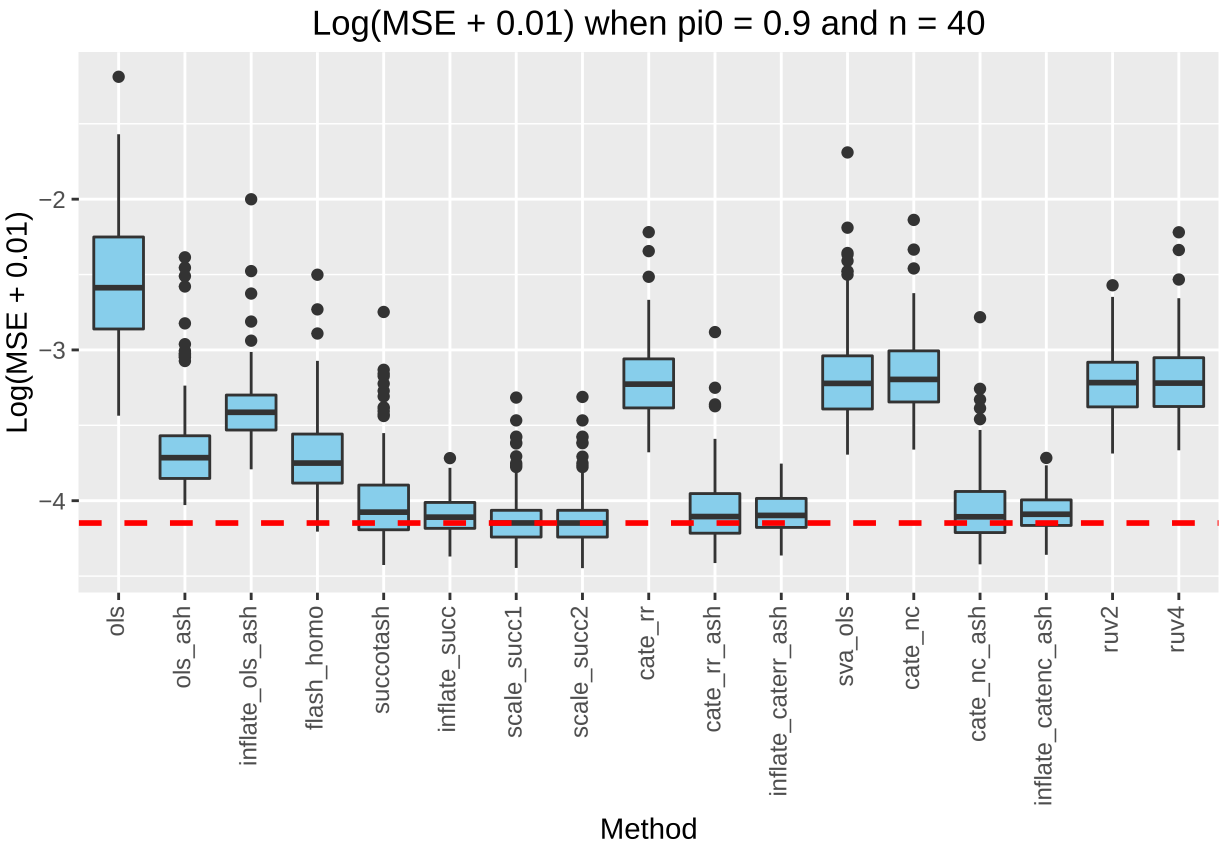




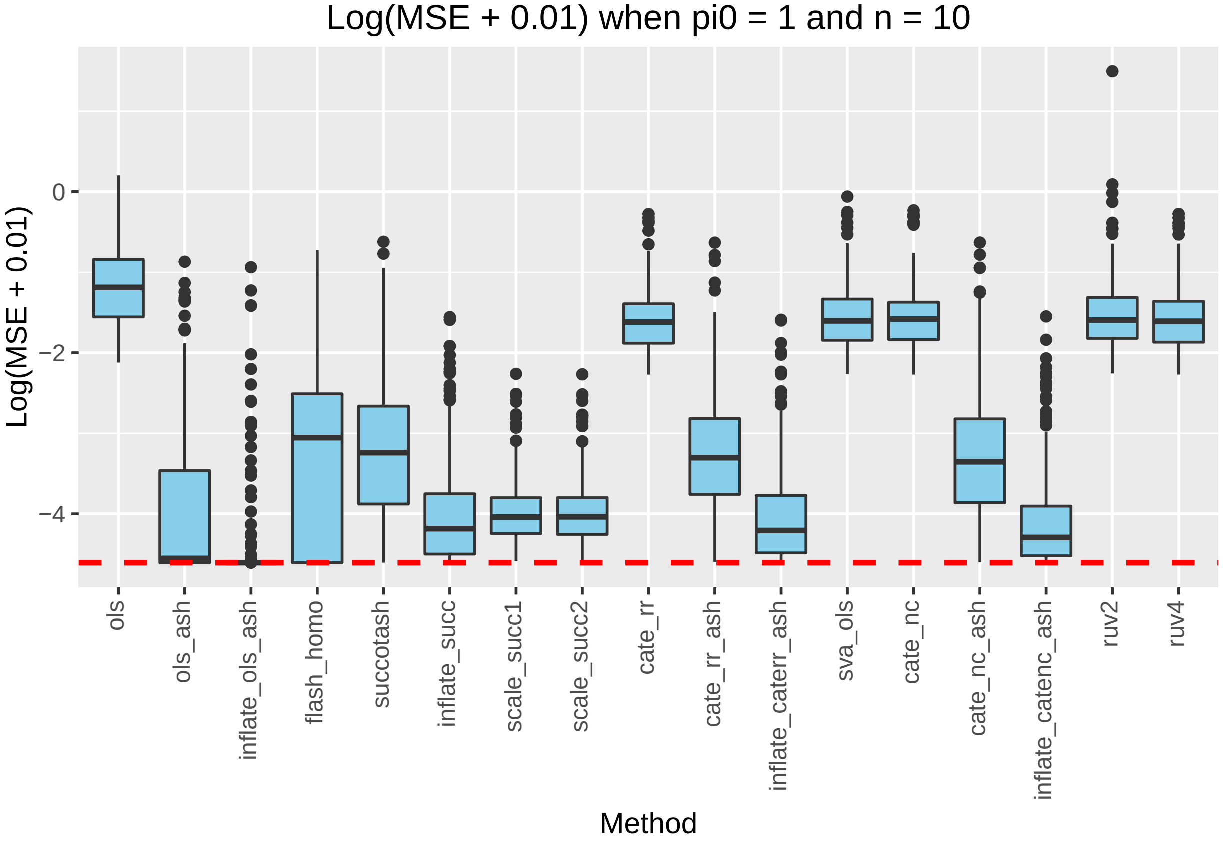
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



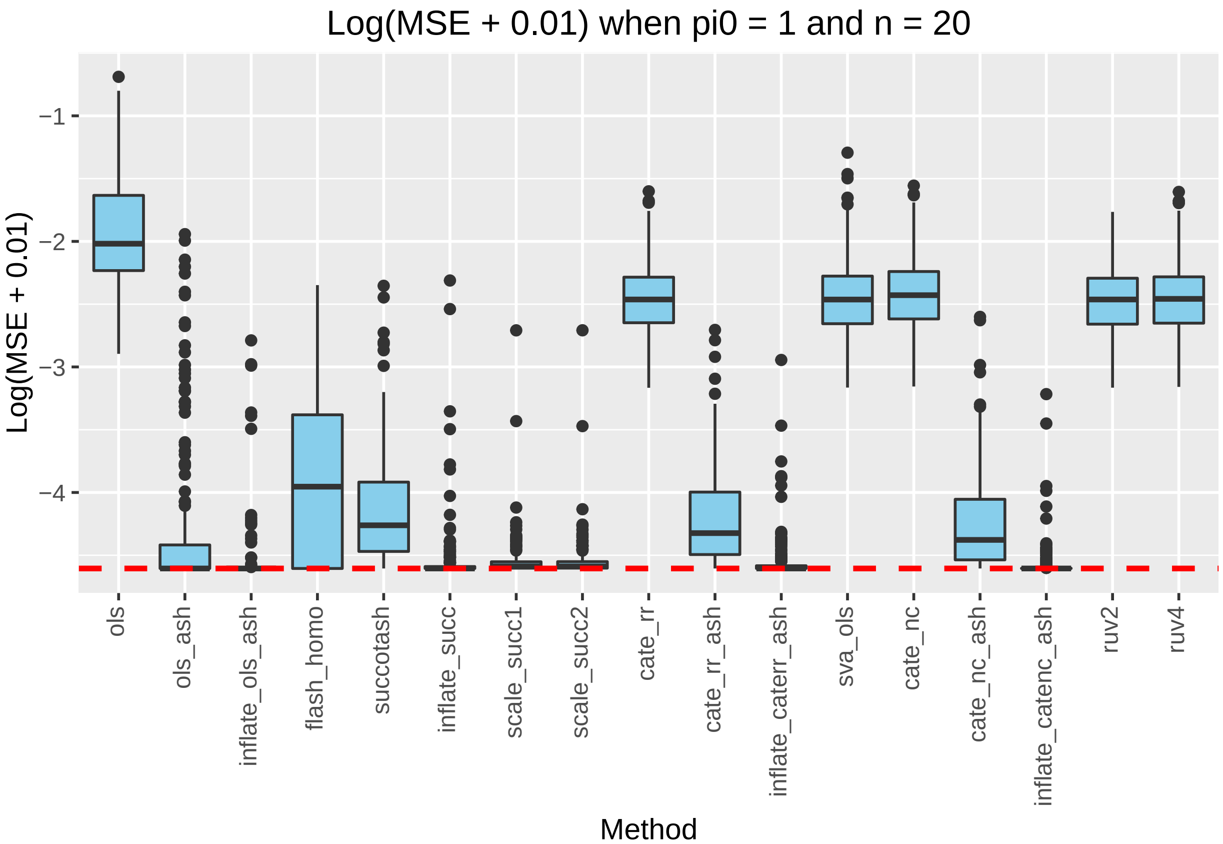
```
## Warning: Removed 89 rows containing non-finite values (stat_boxplot).
```

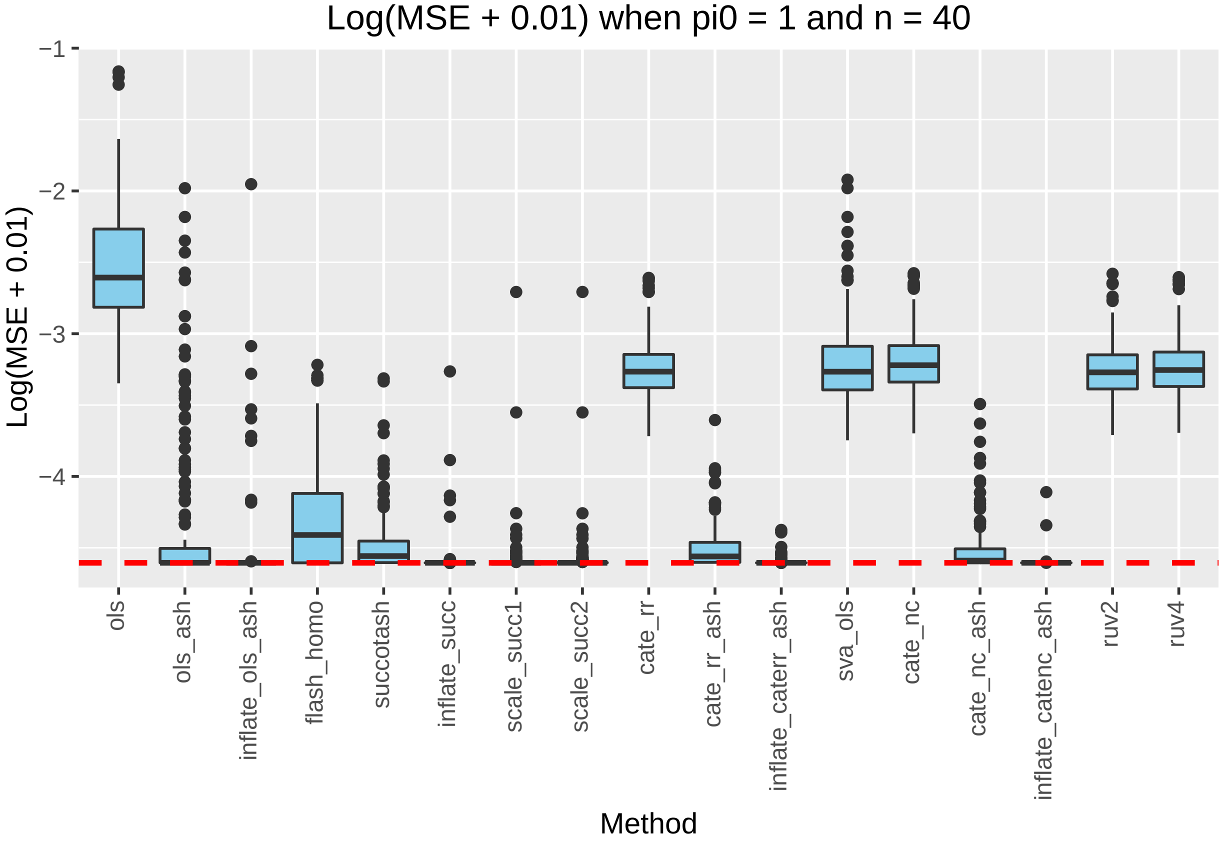


Log(MSE + 0.01) when pi0 = 0.9 and n = 40



Log(MSE + 0.01) when pi0 = 1 and n = 10


Log(MSE + 0.01) when pi0 = 1 and n = 20
Log(MSE + 0.01)
−1
−2
−3
−4
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


Log(MSE + 0.01) when pi0 = 1 and n = 40
Log(MSE + 0.01)
−1
−2
−3
−4
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ1
scale_succ2
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

```r
sessionInfo()
```

```
## R version 3.2.5 (2016-04-14)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_2.1.0  reshape2_1.4.1 dplyr_0.4.3    xtable_1.8-2
## [5] knitr_1.12.26
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.4        magrittr_1.5       munsell_0.4.3
##  [4] colorspace_1.2-6   R6_2.1.1           highr_0.5.1
##  [7] stringr_1.0.0      plyr_1.8.3         tools_3.2.5
## [10] parallel_3.2.5     grid_3.2.5         gtable_0.2.0
## [13] DBI_0.3.1          htmltools_0.3.5    yaml_2.1.13
## [16] lazyeval_0.1.10    assertthat_0.1     digest_0.6.9
## [19] formatR_1.3        codetools_0.2-14   evaluate_0.8.3
## [22] rmarkdown_0.9.5.12 labeling_0.3       stringi_1.0-1
## [25] compiler_3.2.5     scales_0.4.0
```