Inflate Variance in Confounder Adjustment Methods

David Gerard

2016-04-26

Abstract

Here, I inflate the variance for SUCCOTASH, CATE + ASH, and OLS + ASH, and compare them to uninflated procedures. Inflating the variance for SUCCOTASH and CATE + ASH results in improved MSE and improved estimates of π_0 with no detriment to AUC. Inflating the variance for OLS + ASH does not seem to do much.

Results

```
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.

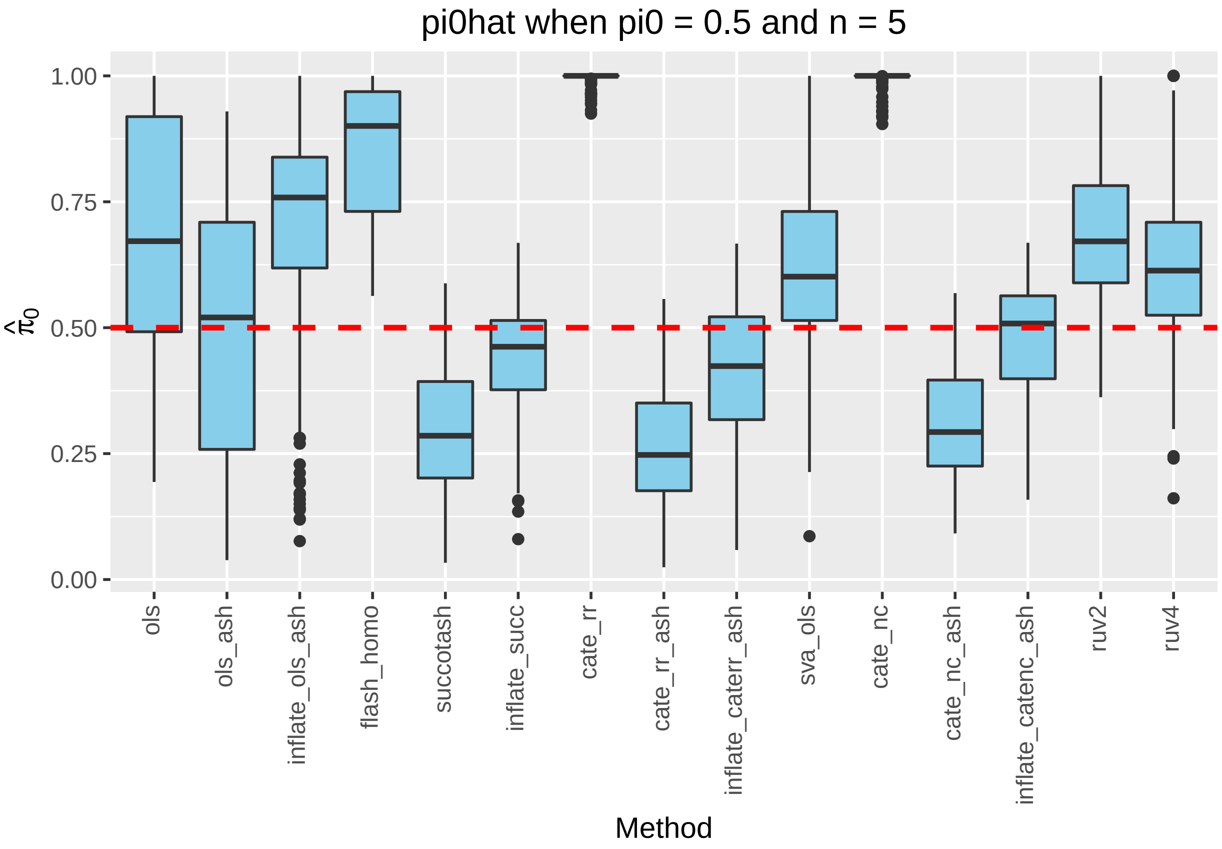
I just multiplied the variance estimates by 2 (multiplied the standard deviation estimates by $\sqrt{2}$) for the inflated variance versions.

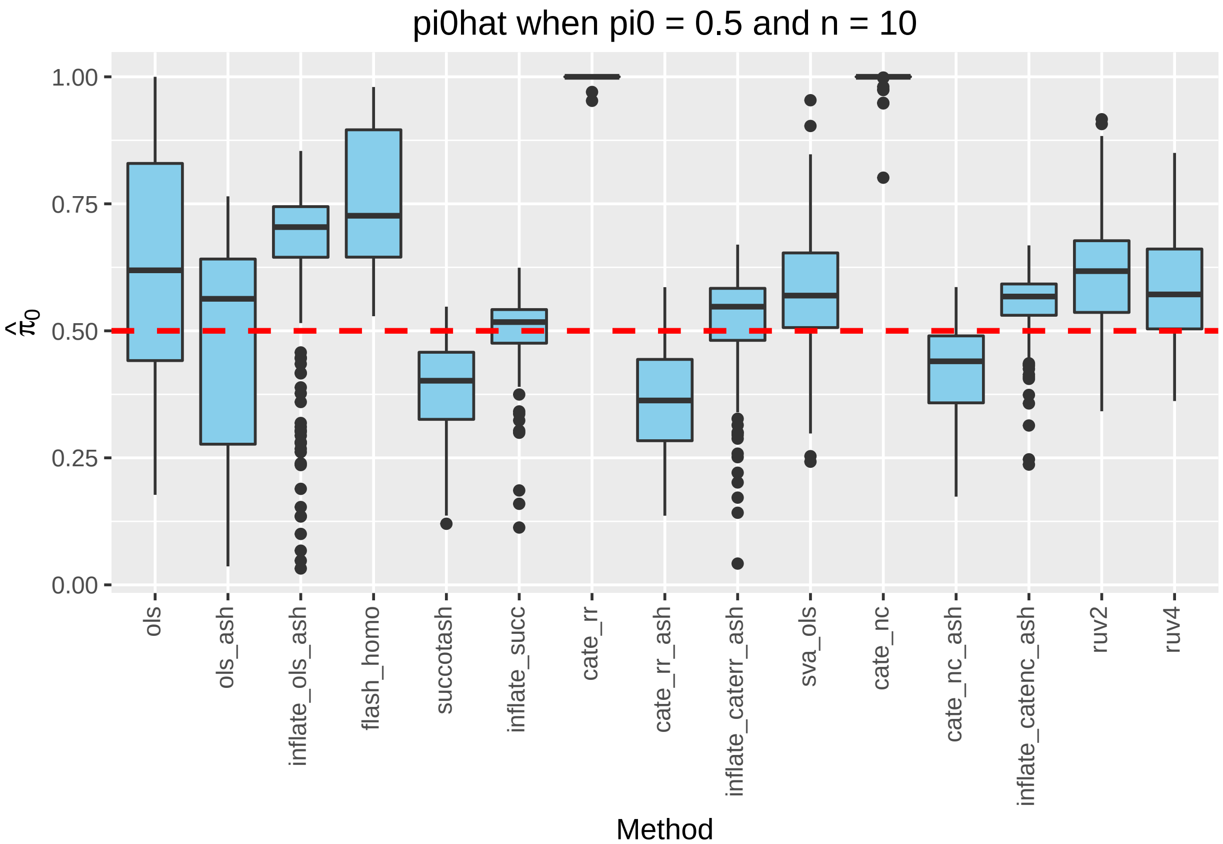
Inflating the variance in SUCCOTASH and CATE + ASH improve both the estimates of π_0 and the MSE. The AUC is almost the exact same for both the inflated and uninflated versions.

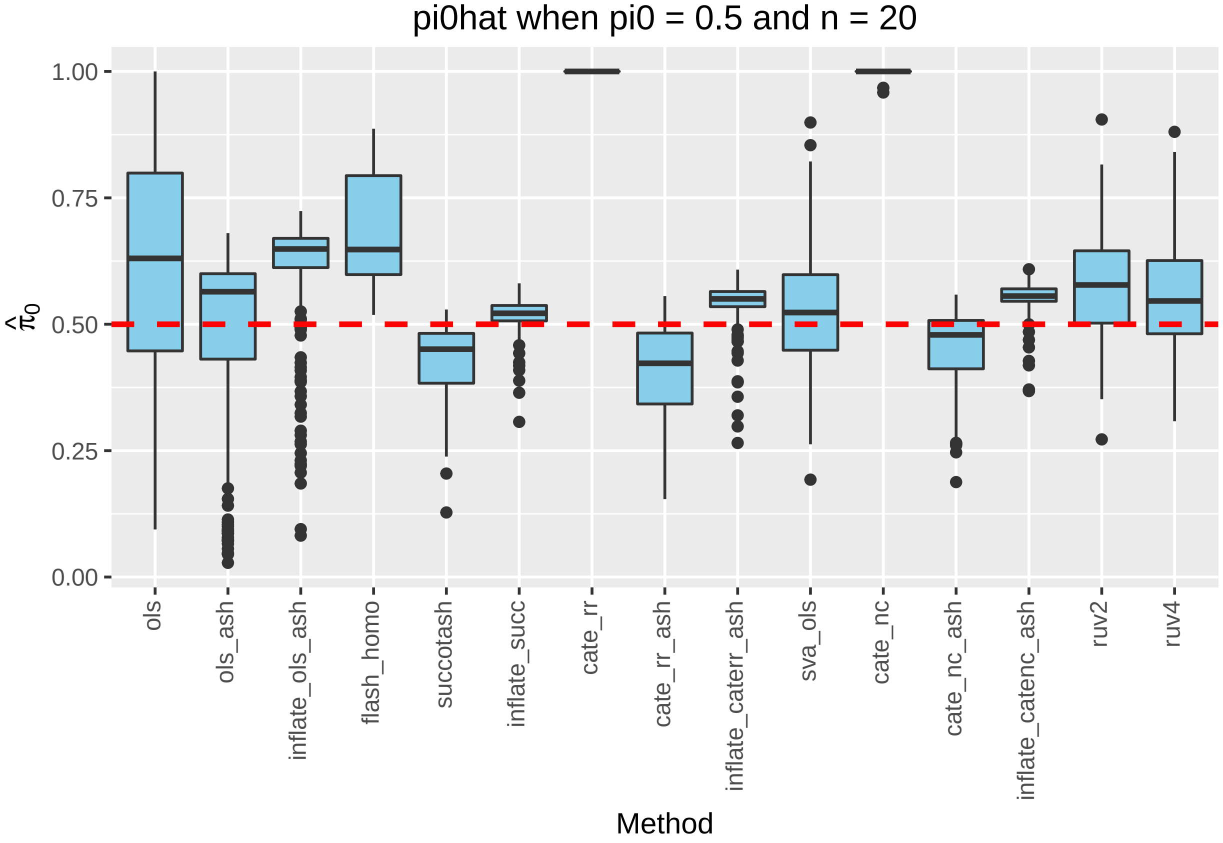
Interestingly, inflating the variance in OLS + ASH saw a much smaller improvement (if any). In particular, MSE got worse. The estimates of π_0 look a little better for each pair of π_0 and n, but the estimates have a very large left tail with about 10% of the values being deemed "outliers" in the boxplot for each combination.

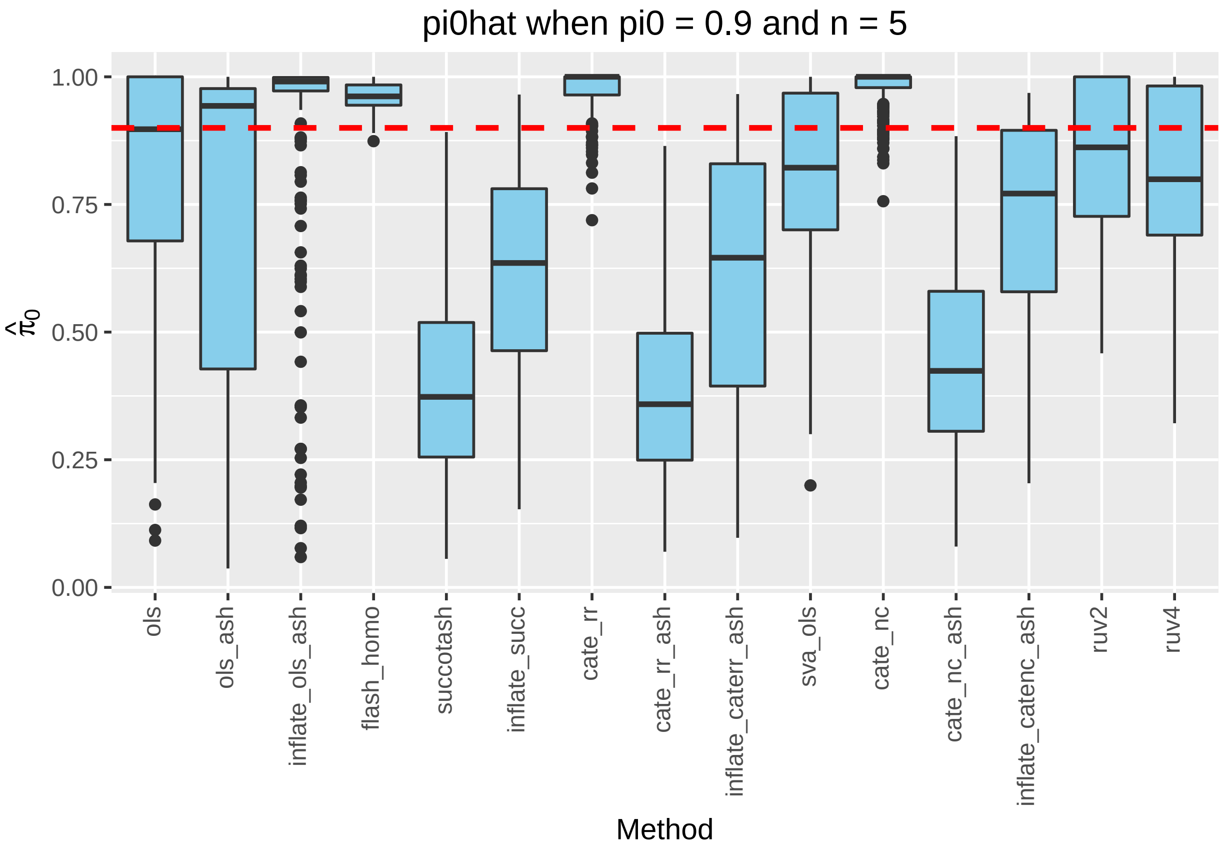
$\hat{\pi}_0$ Plots

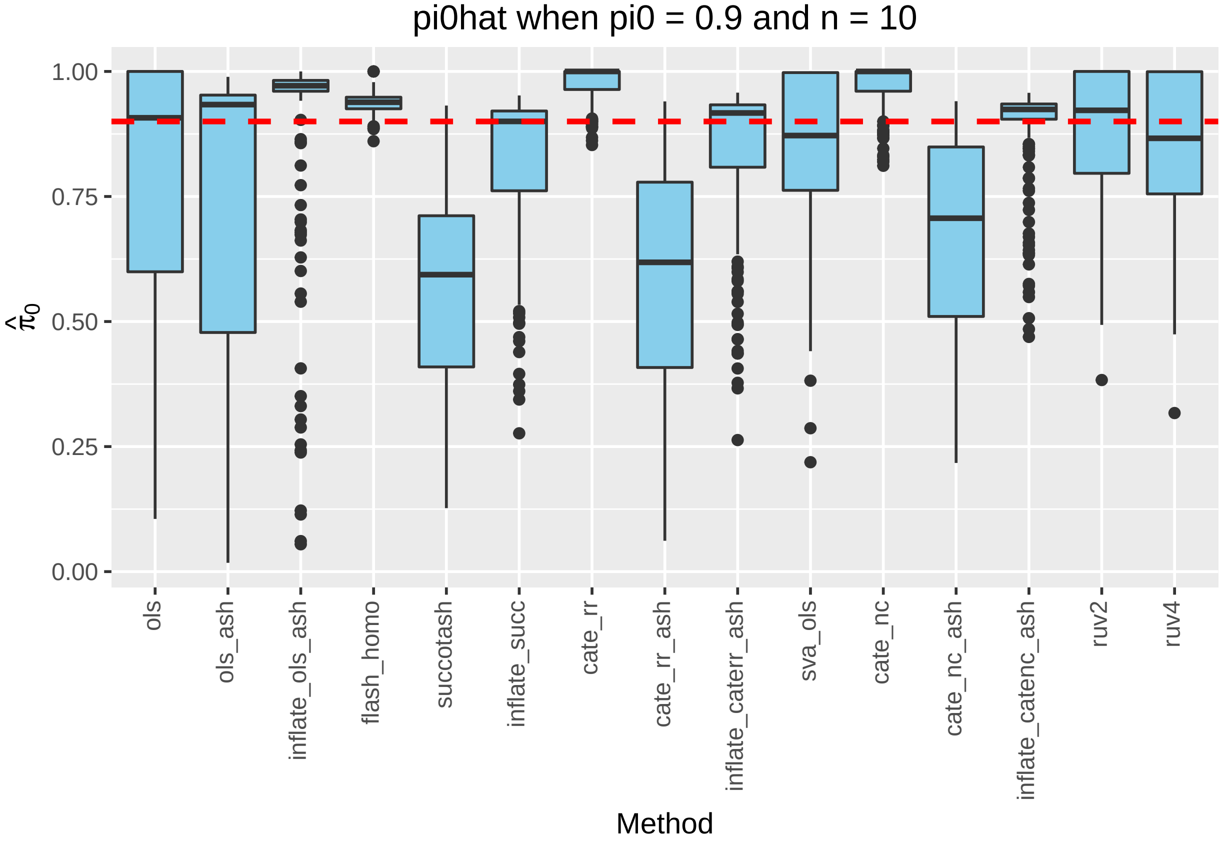
```
inflate_pi0 <- read.csv("pi0_mat.csv")
reg_pi0 <- read.csv("../flash_v_rest_using_package/pi0_mat.csv")
reg_pi0$inflate_succ <- inflate_pi0$succotash
reg_pi0$inflate_caterr_ash <- inflate_pi0$cate_rr_ash
reg_pi0$inflate_catenc_ash <- inflate_pi0$cate_nc_ash
reg_pi0$inflate_ols_ash <- inflate_pi0$ols_ash
reg_pi0 <- tbl_df(reg_pi0)
reg_pi0 <- reg_pi0[, c(1:2, 17, 3:4, 14, 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)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```

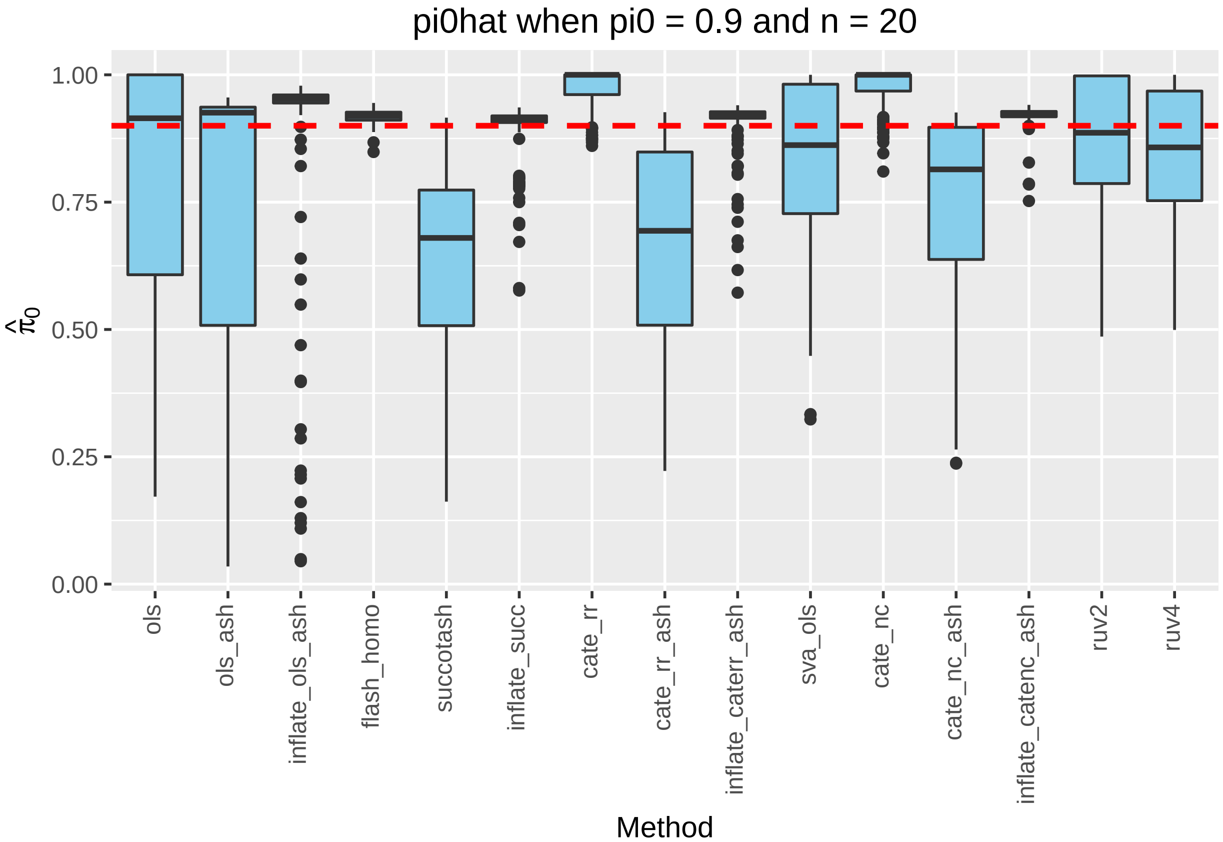


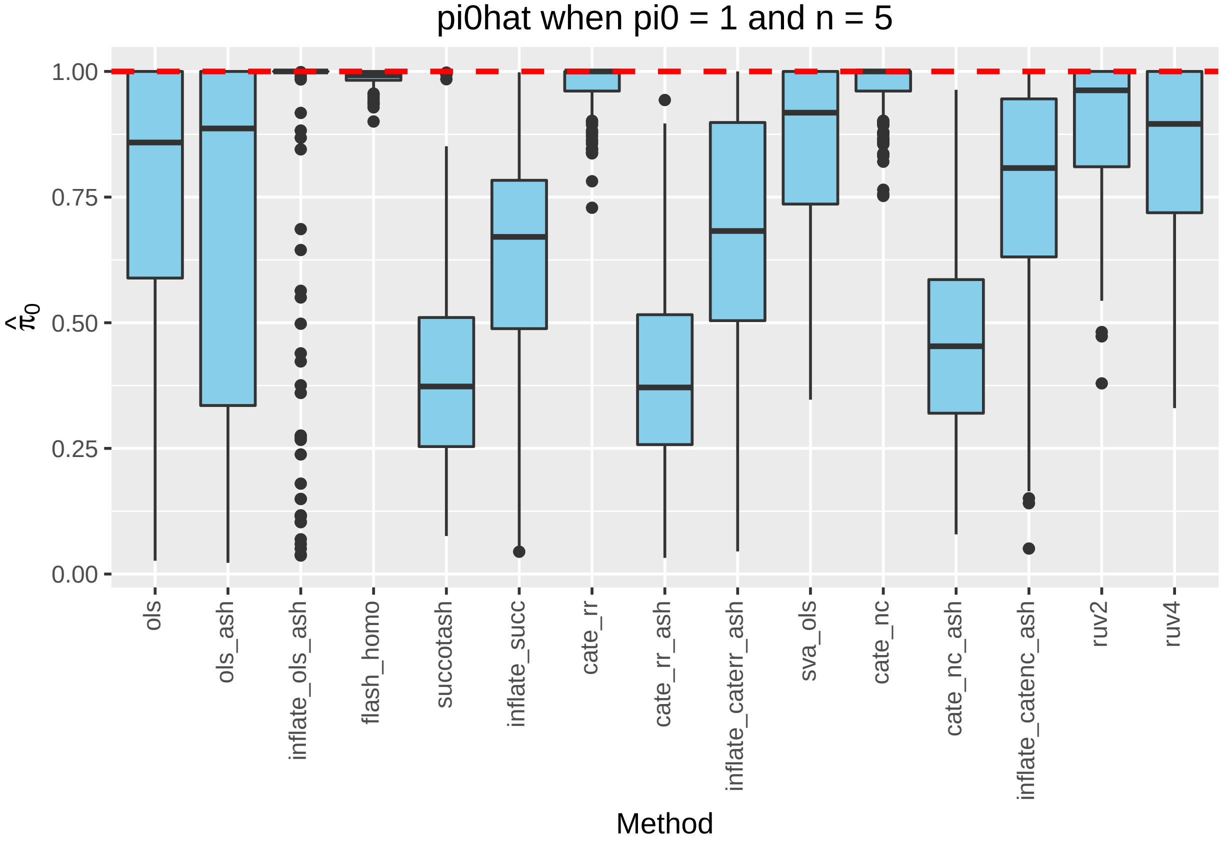

pi0hat when pi0 = 0.5 and n = 10

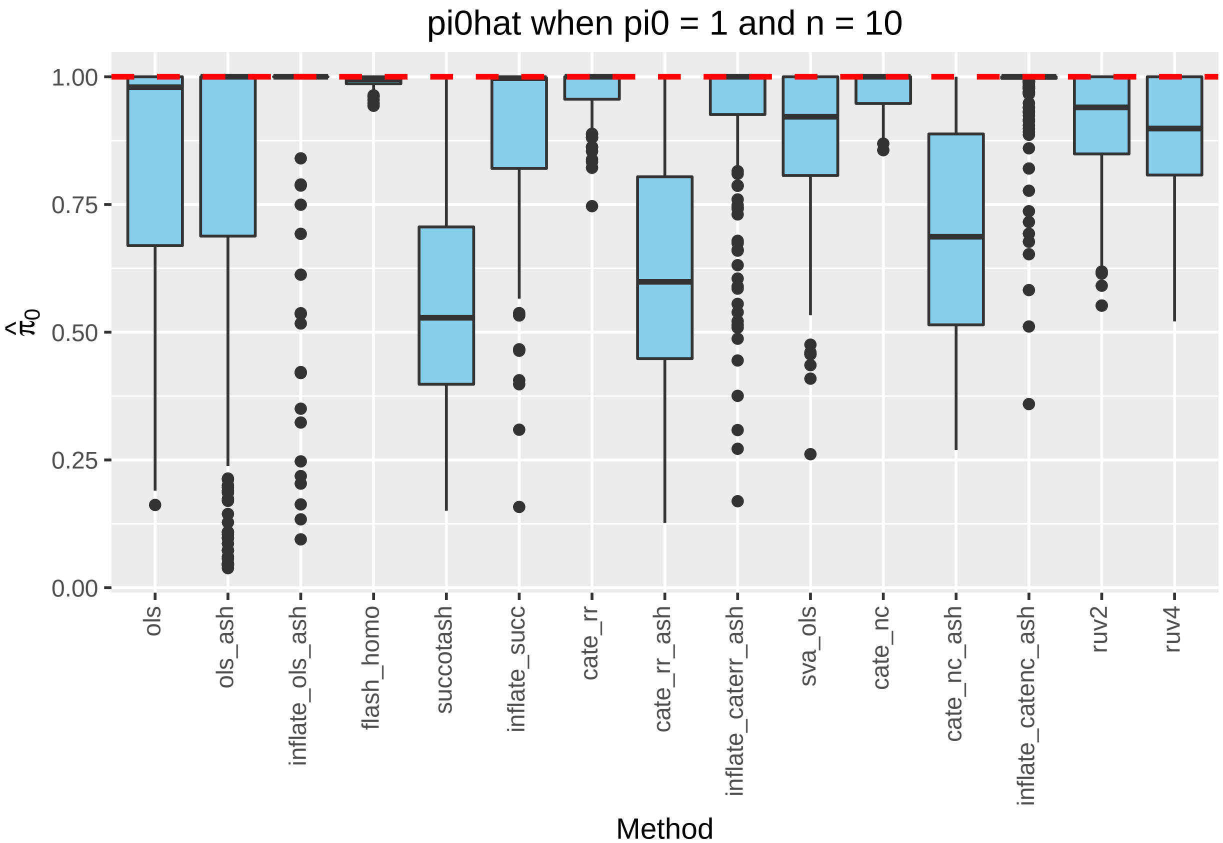

pi0hat when pi0 = 0.5 and n = 20

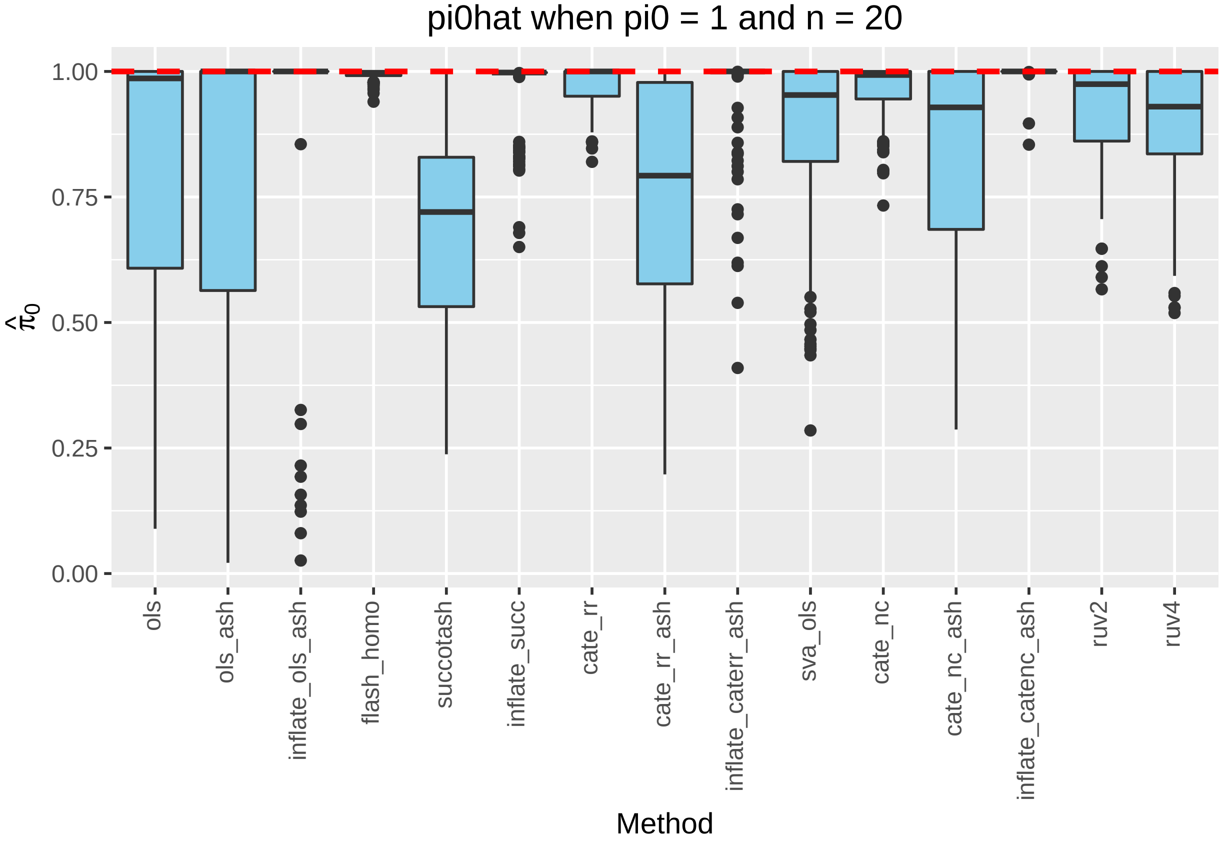

pi0hat when pi0 = 0.9 and n = 5
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


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
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
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

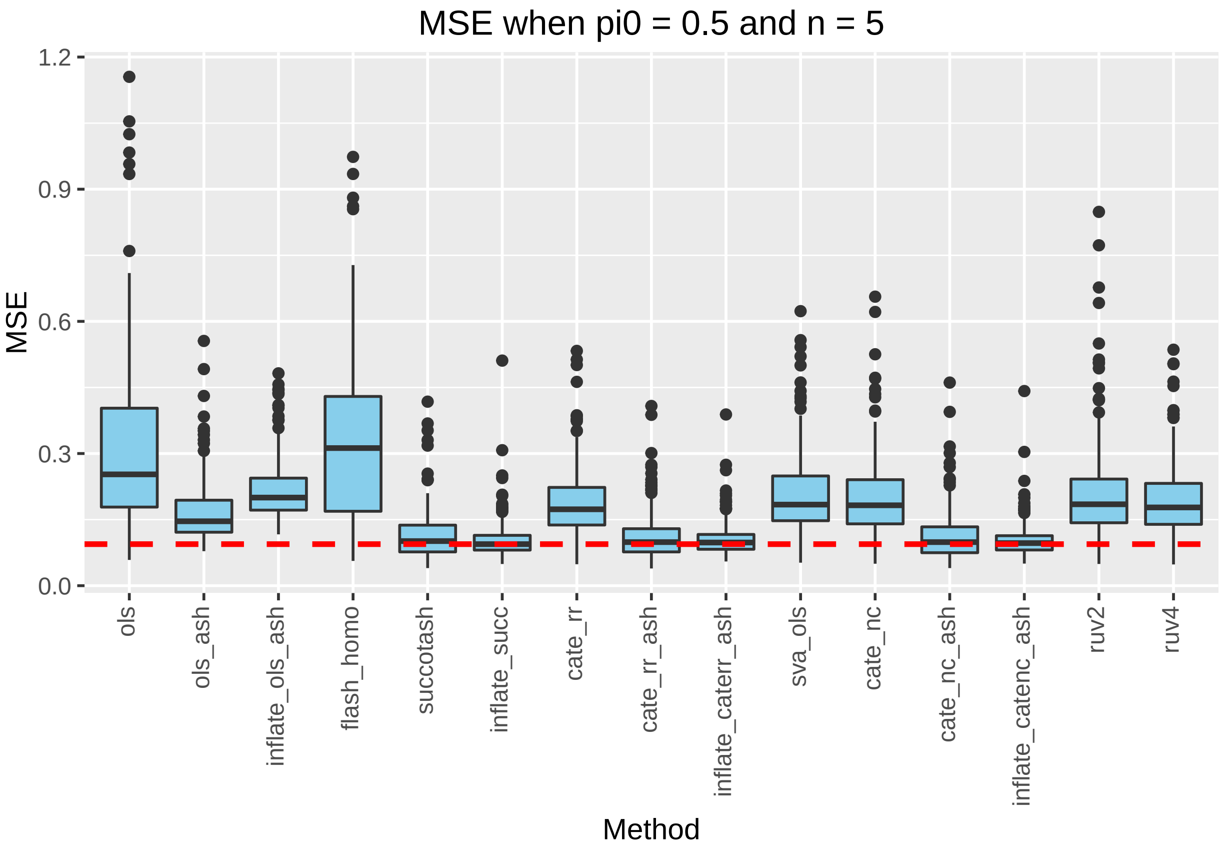

pi0hat when pi0 = 1 and n = 5
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


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
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

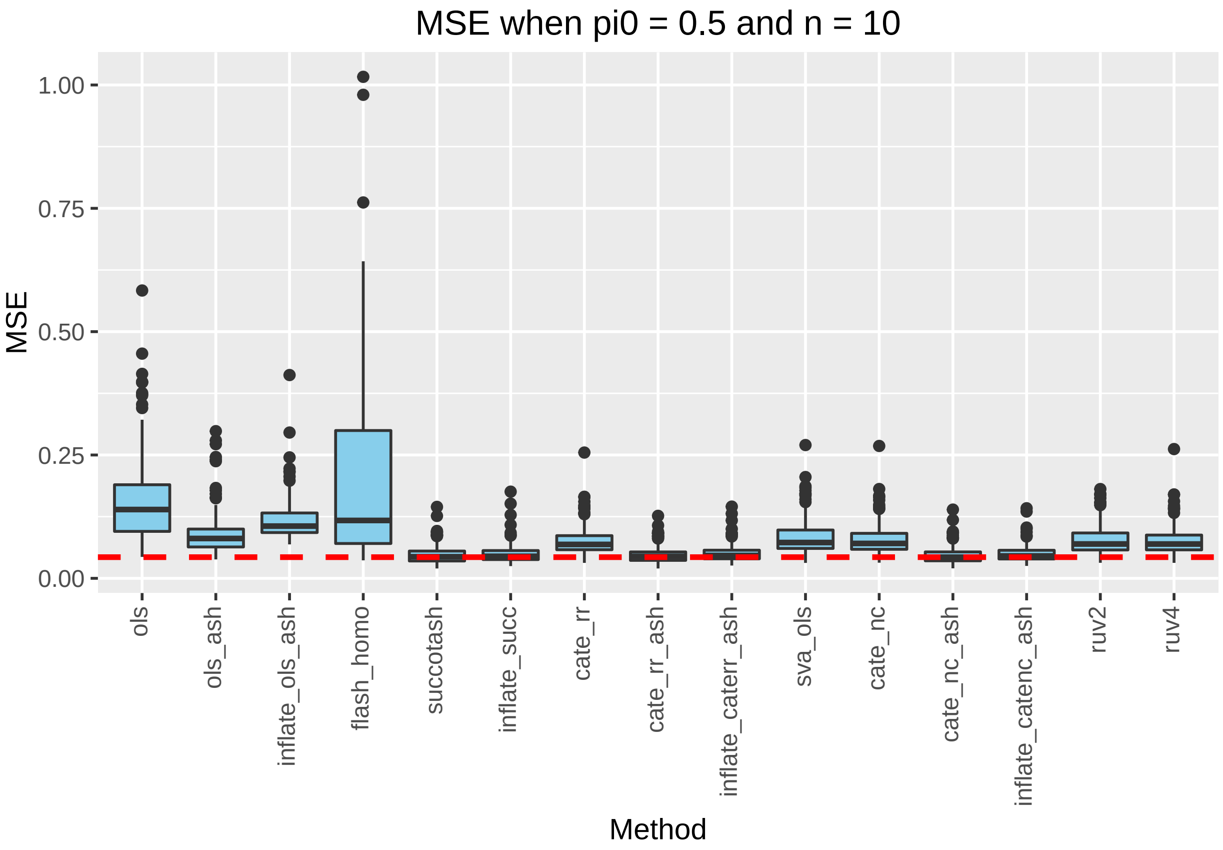

pi0hat when pi0 = 1 and n = 20
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

MSE Plots

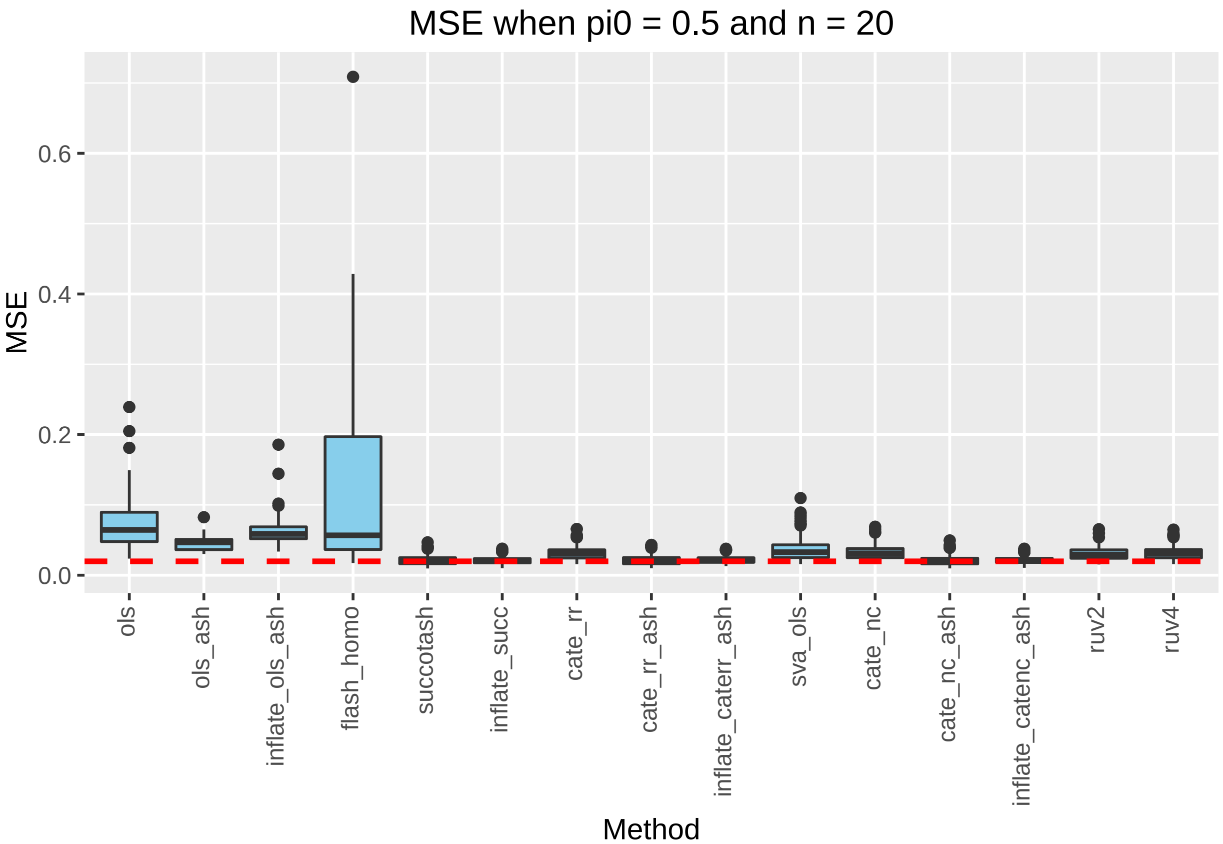
```
inflate_mse <- read.csv("mse_mat.csv")
reg_mse <- read.csv("../flash_v_rest_using_package/mse_mat.csv")
reg_mse$inflate_succ <- inflate_mse$succotash
reg_mse$inflate_caterr_ash <- inflate_mse$cate_rr_ash
reg_mse$inflate_catenc_ash <- inflate_mse$cate_nc_ash
reg_mse$inflate_ols_ash <- inflate_mse$ols_ash
reg_mse <- tbl_df(reg_mse)
reg_mse <- reg_mse[, c(1:2, 17, 3:4, 14, 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)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```

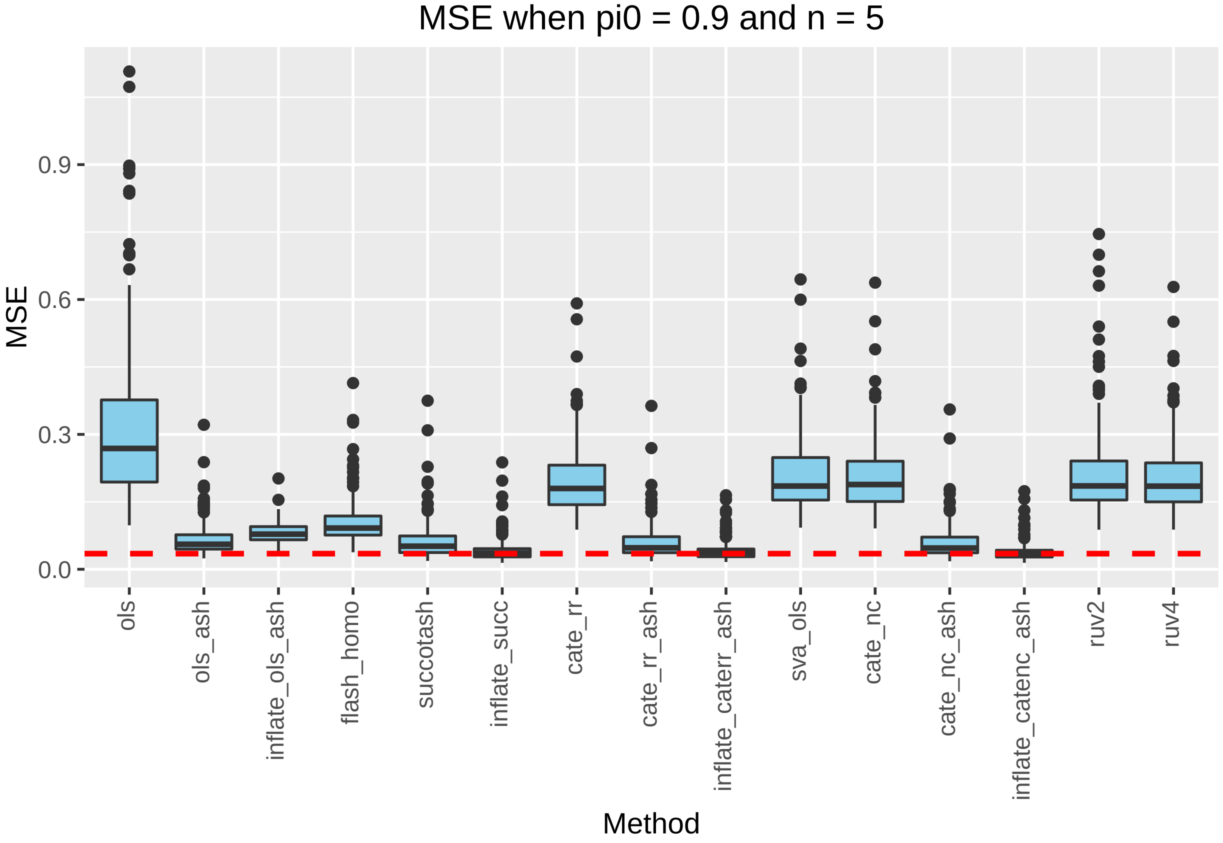


```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

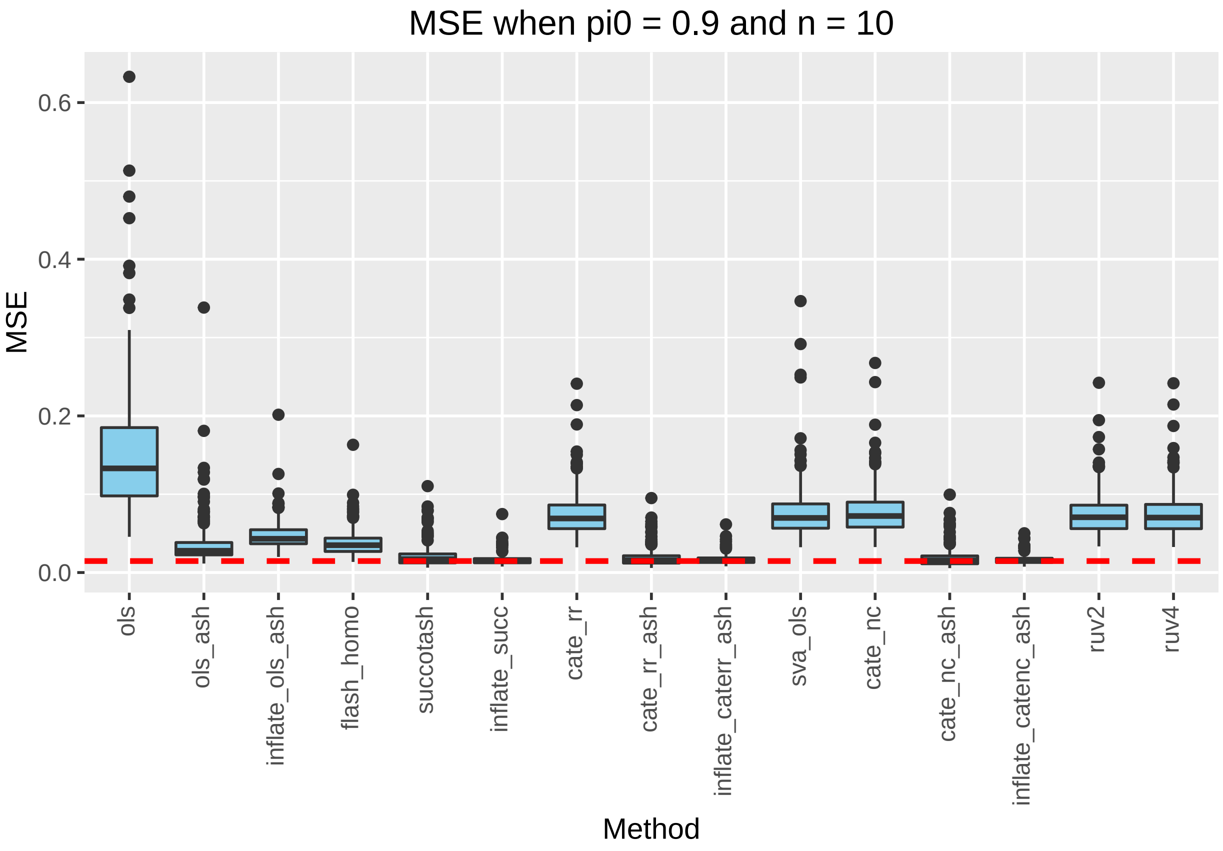


```
## Warning: Removed 203 rows containing non-finite values (stat_boxplot).
```

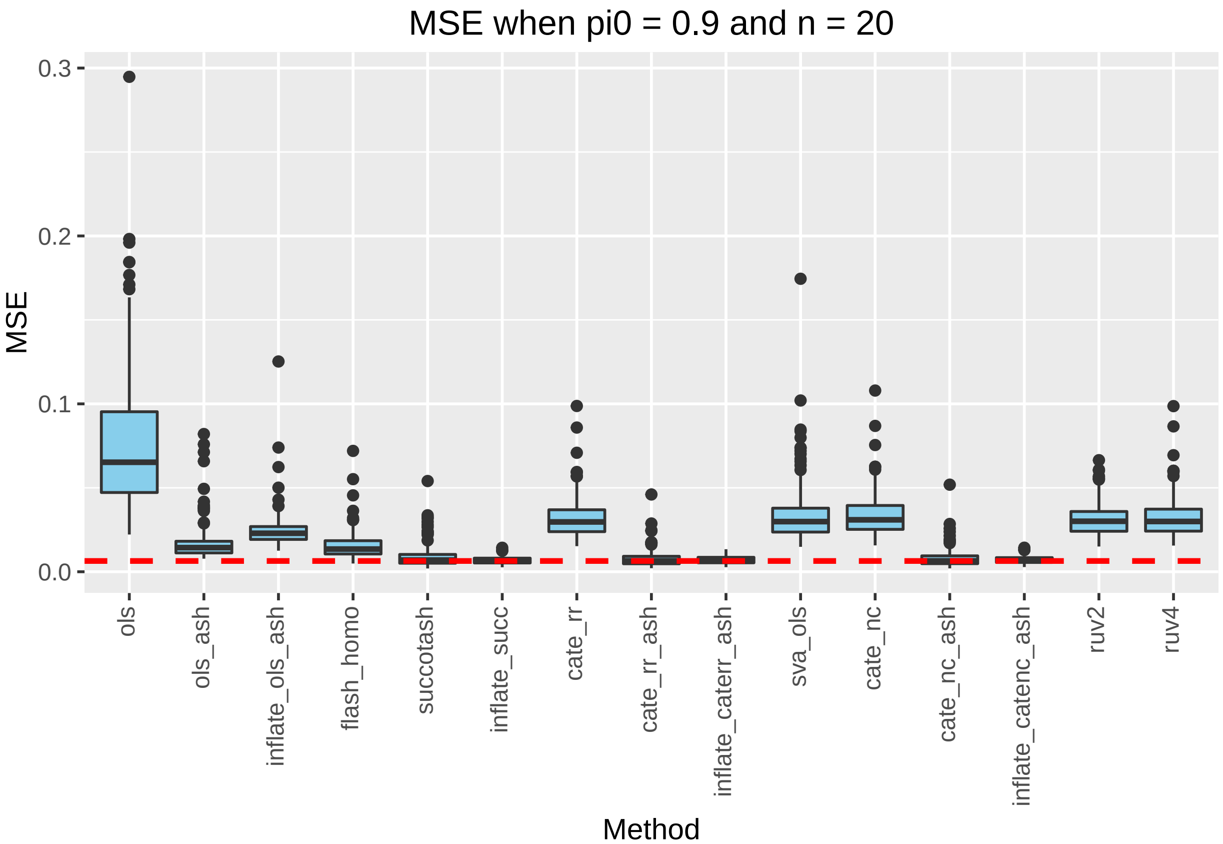


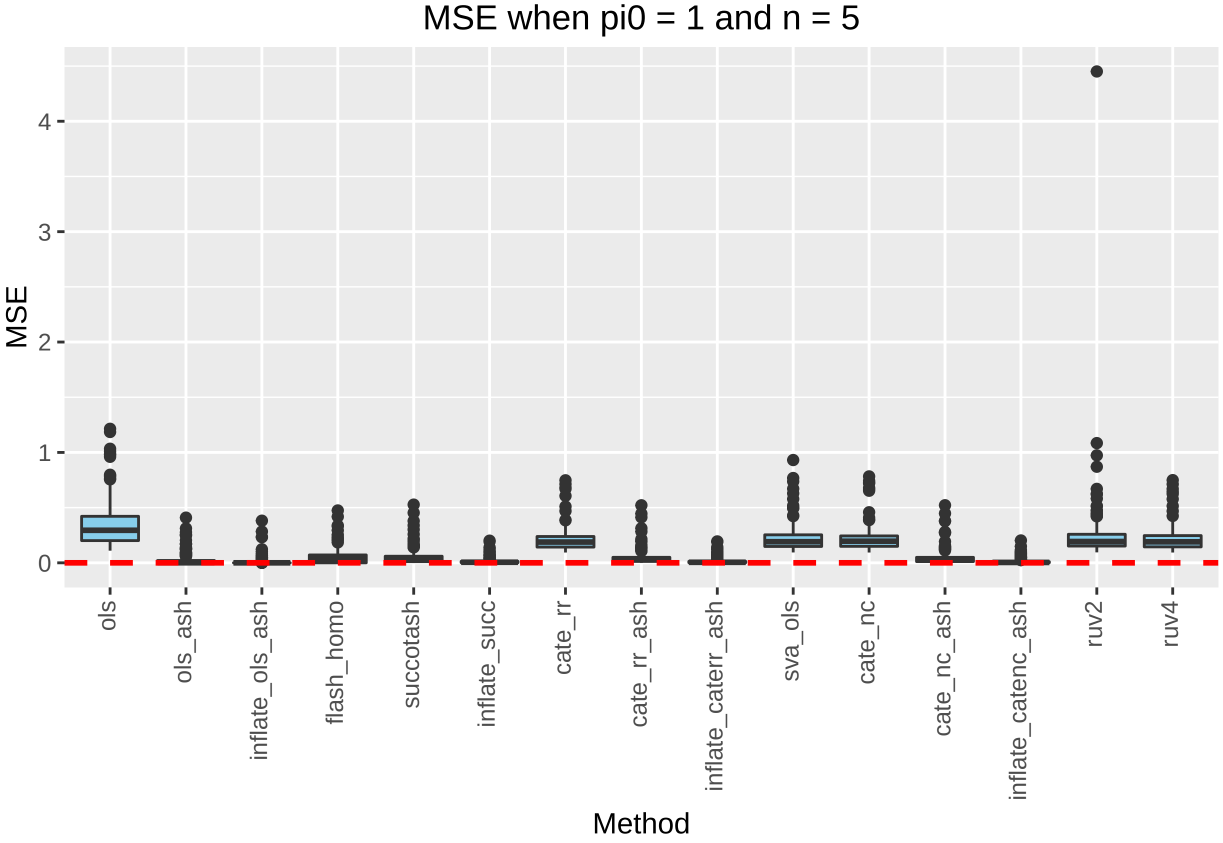


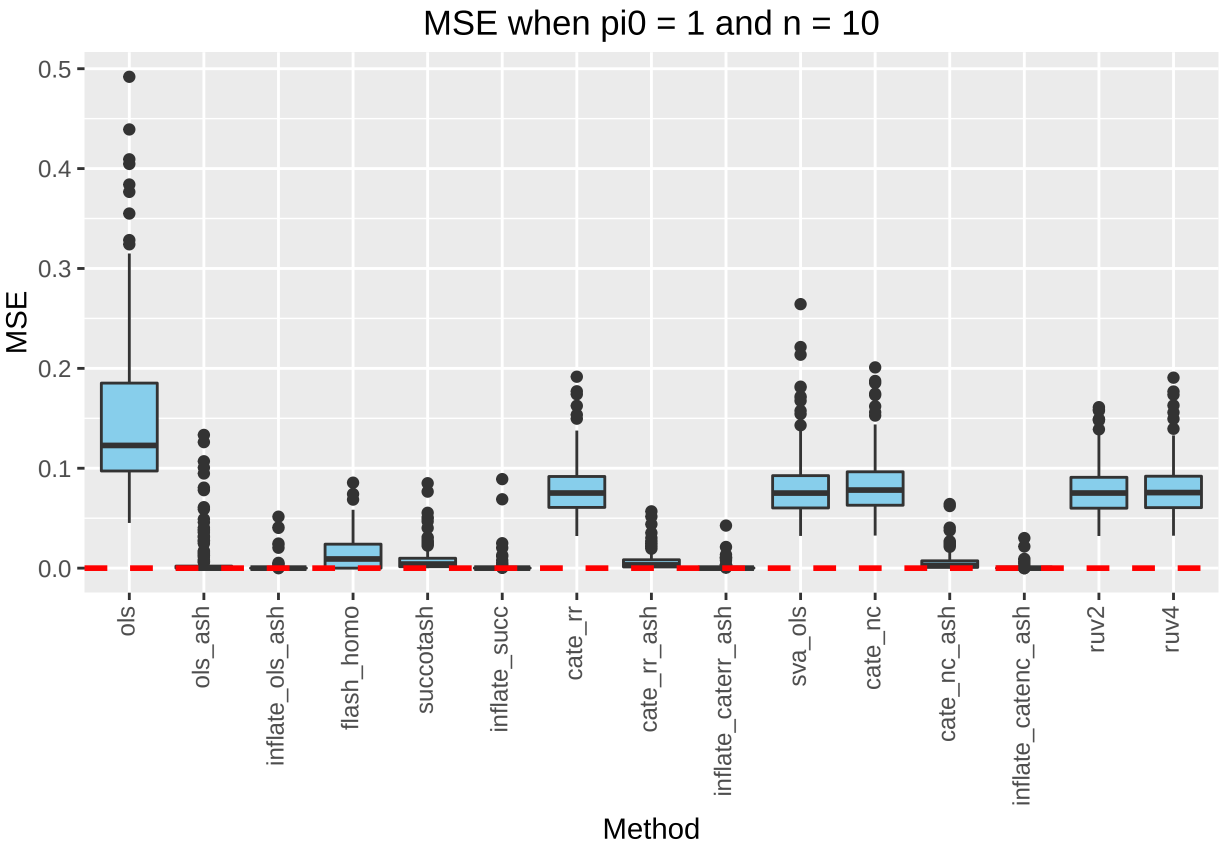
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

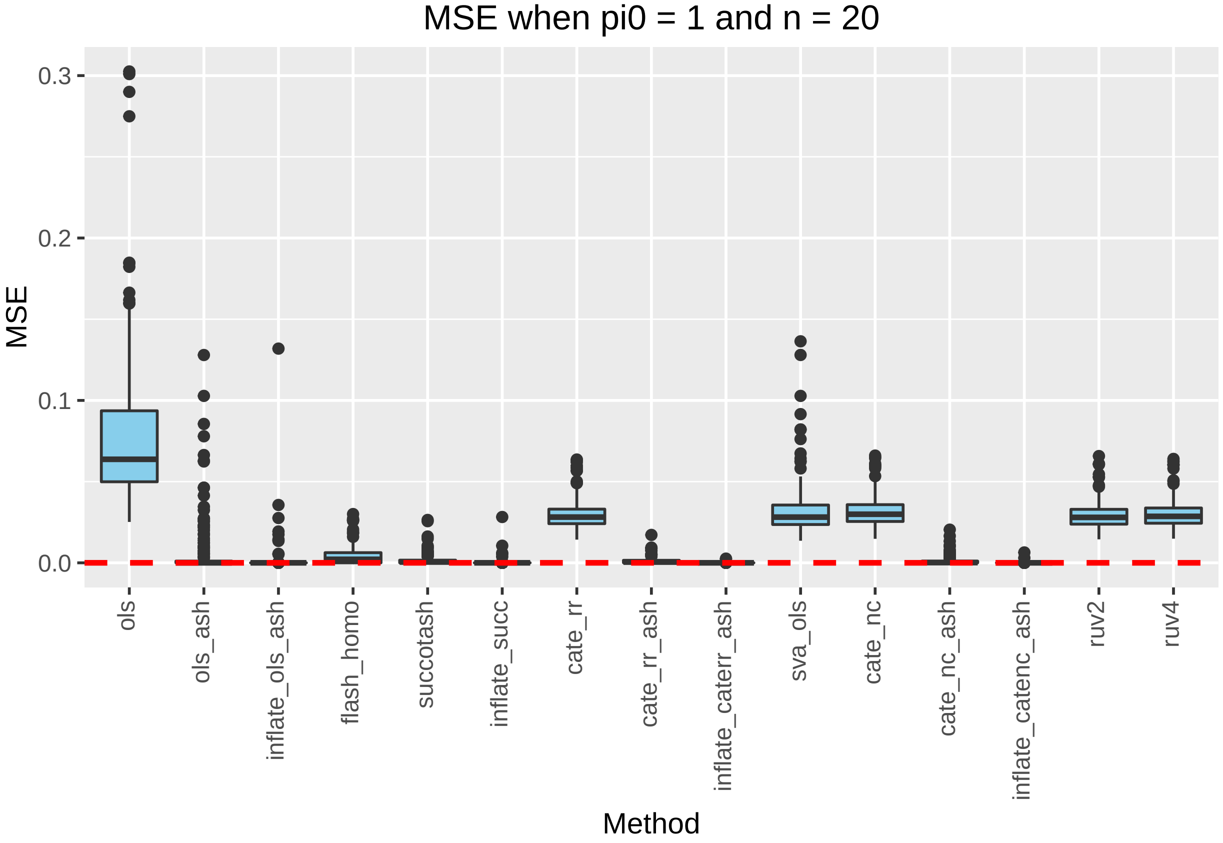


```
## Warning: Removed 89 rows containing non-finite values (stat_boxplot).
```


MSE when pi0 = 0.9 and n = 20

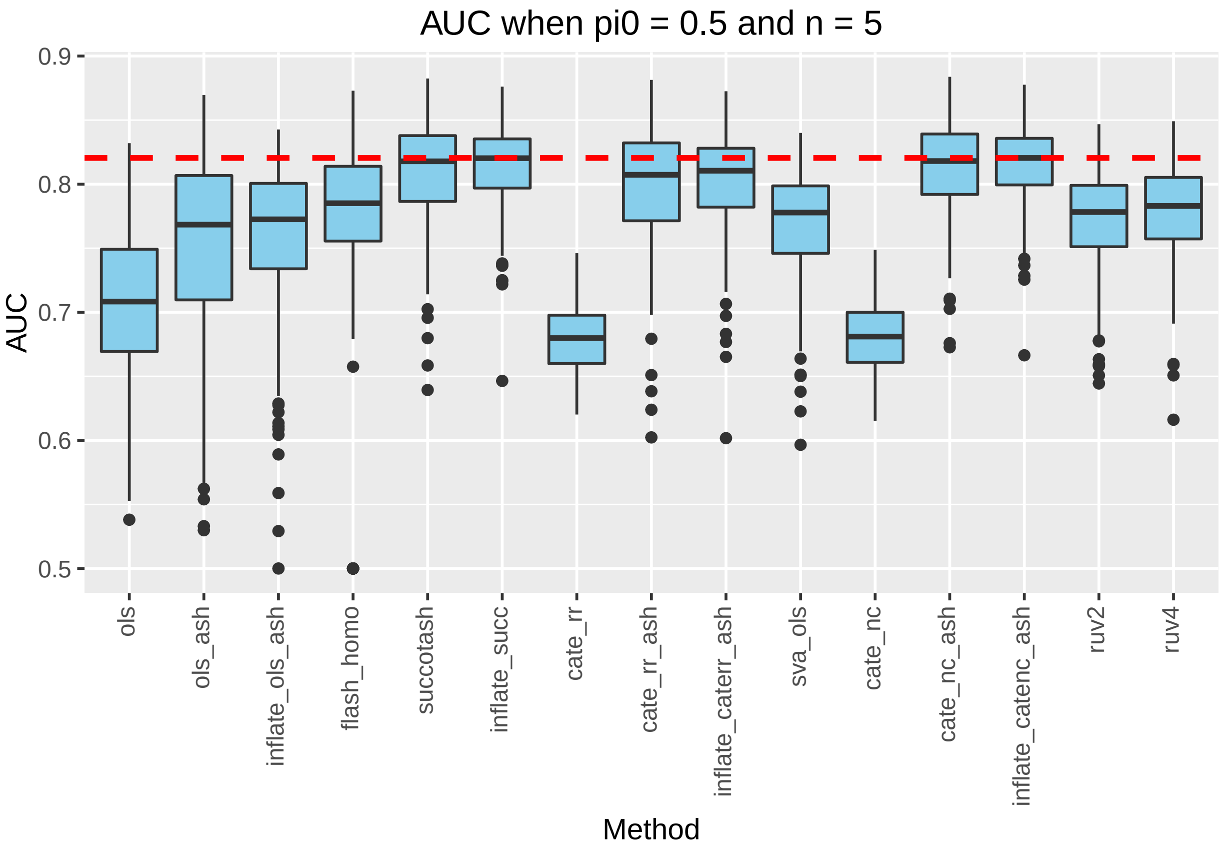

MSE when pi0 = 1 and n = 5

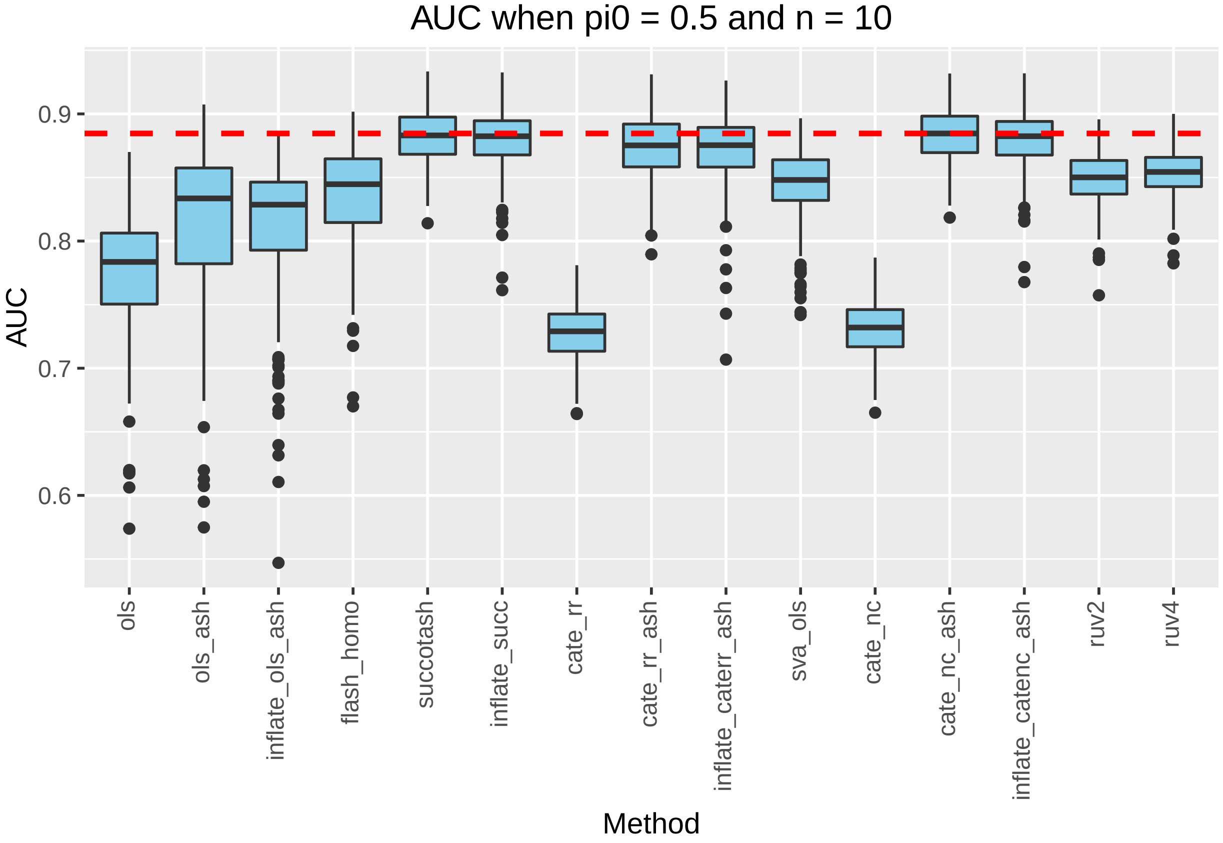

MSE when pi0 = 1 and n = 10
MSE
0.5
0.4
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

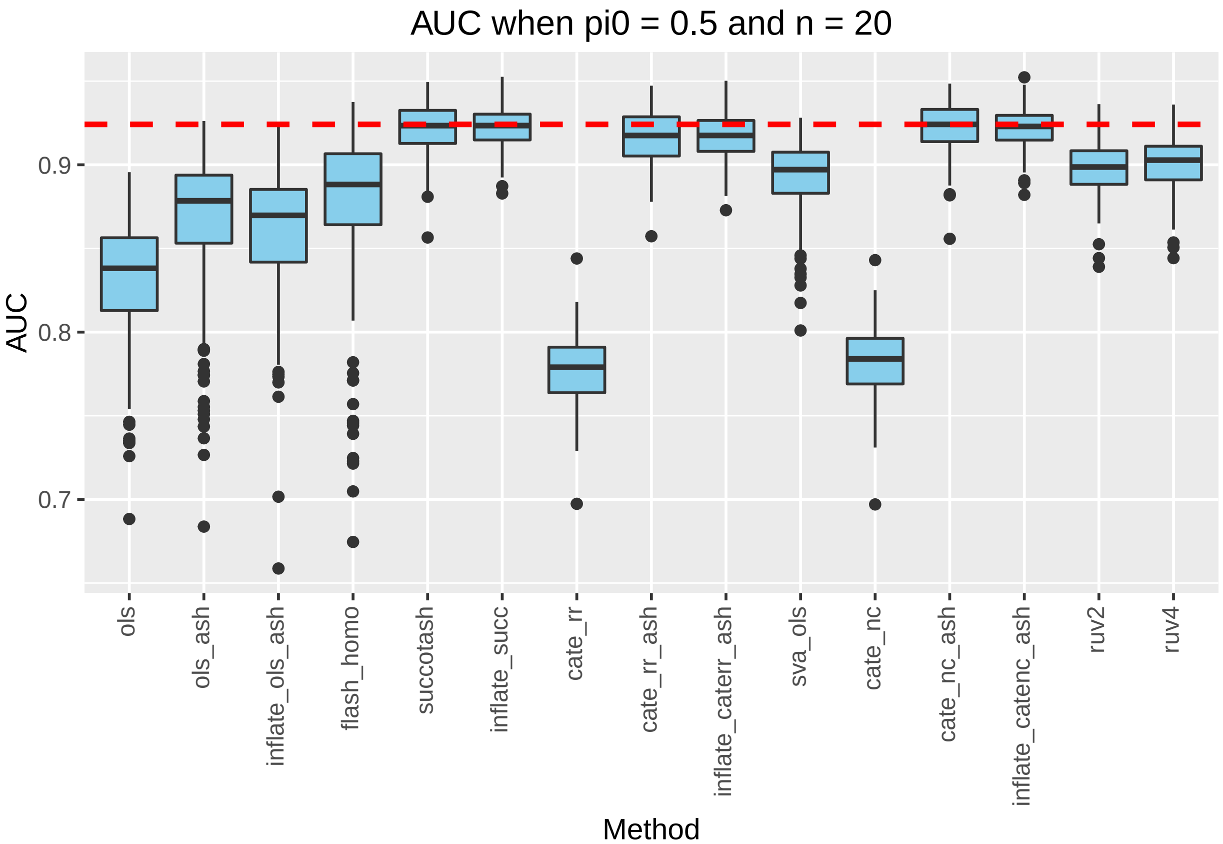

MSE when pi0 = 1 and n = 20
MSE
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

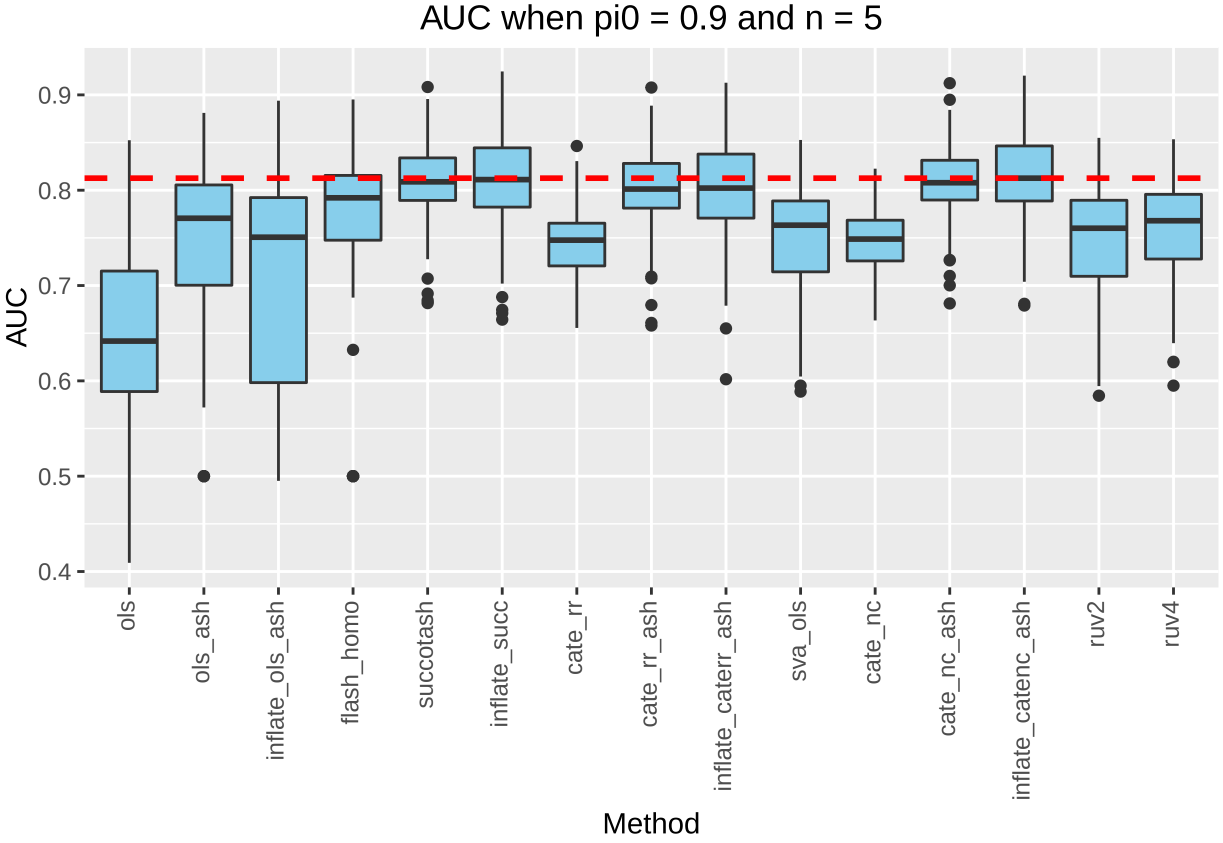
AUC Plots

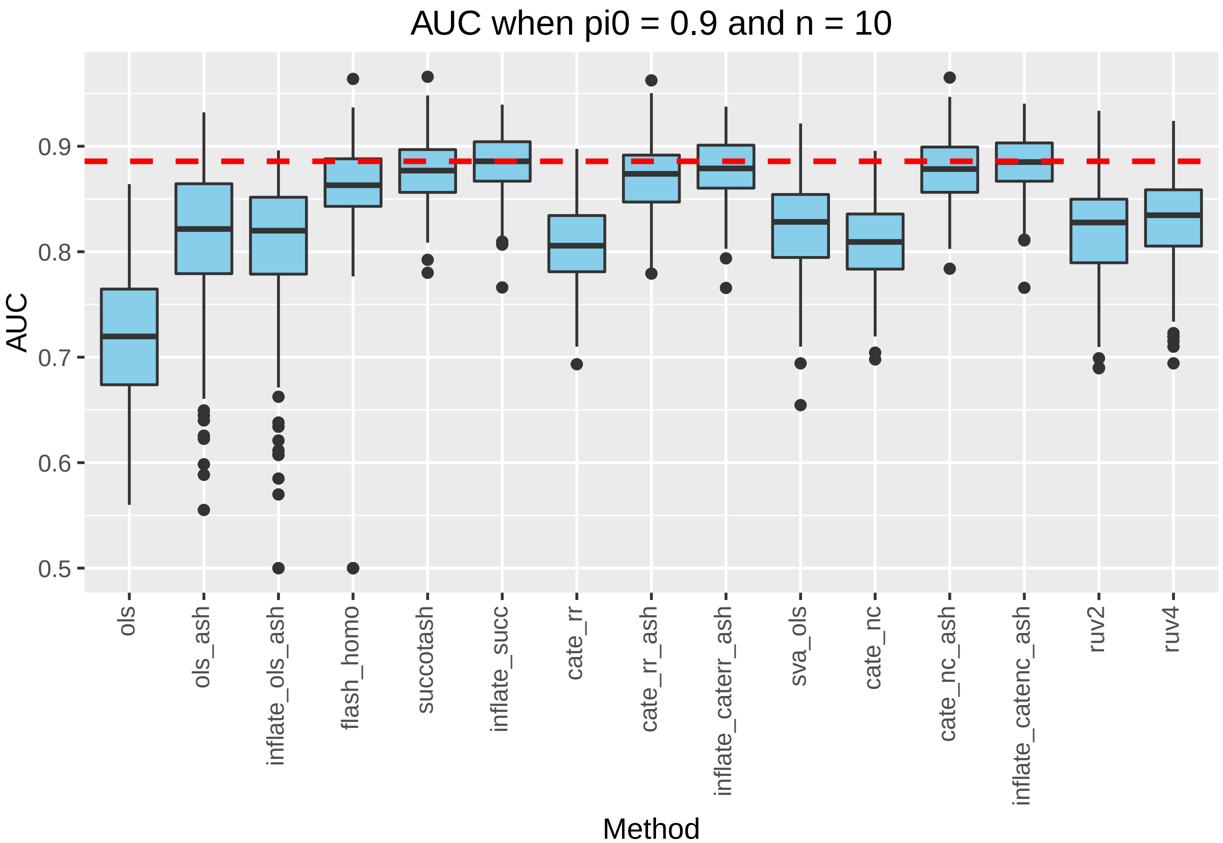
```r
inflate_auc <- read.csv("auc_mat.csv")
reg_auc <- read.csv("../flash_v_rest_using_package/auc_mat.csv")
reg_auc$inflate_succ <- inflate_auc$succotash
reg_auc$inflate_caterr_ash <- inflate_auc$cate_rr_ash
reg_auc$inflate_catenc_ash <- inflate_auc$cate_nc_ash
reg_auc$inflate_ols_ash <- inflate_auc$ols_ash
reg_auc <- tbl_df(reg_auc)
reg_auc <- reg_auc[, c(1:2, 17, 3:4, 14, 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)) +
            theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.3))
        print(p)
    }
}
```

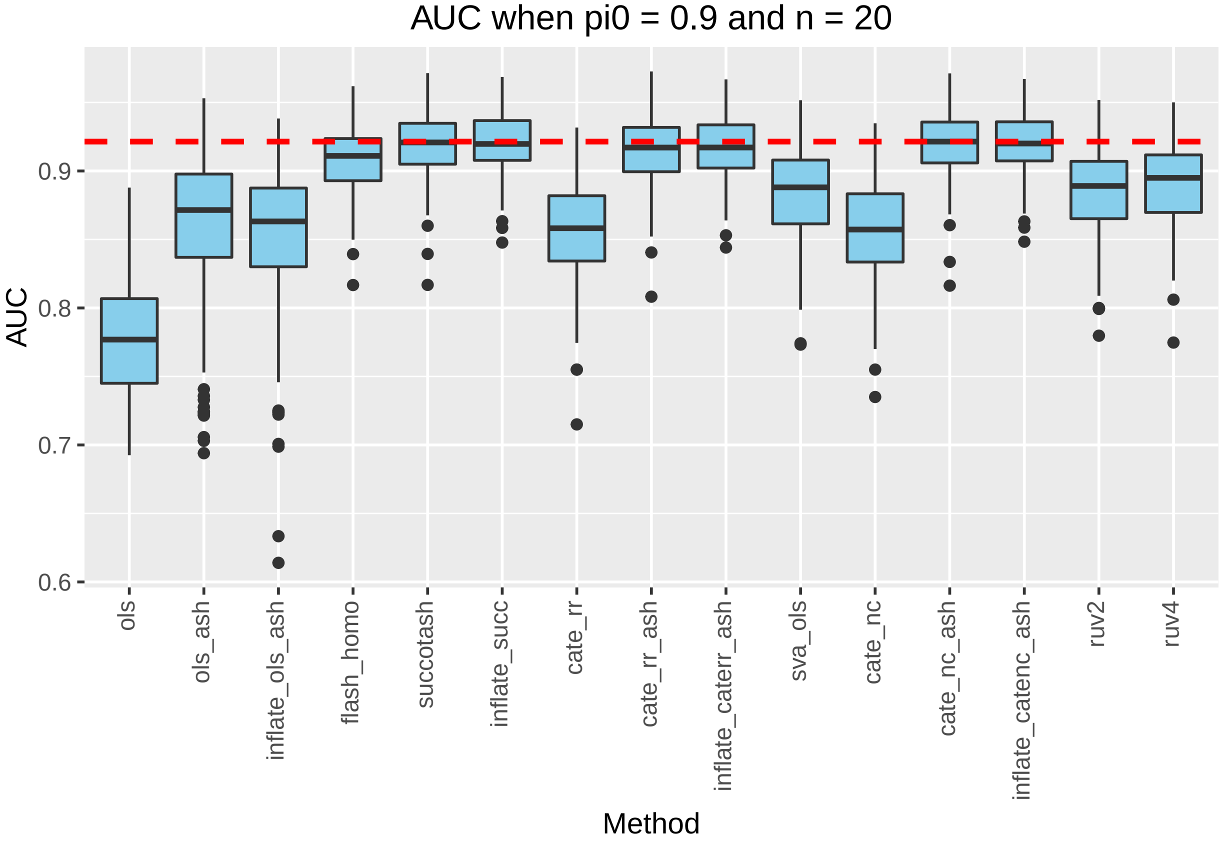

AUC when pi0 = 0.5 and n = 5


AUC when pi0 = 0.5 and n = 10


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

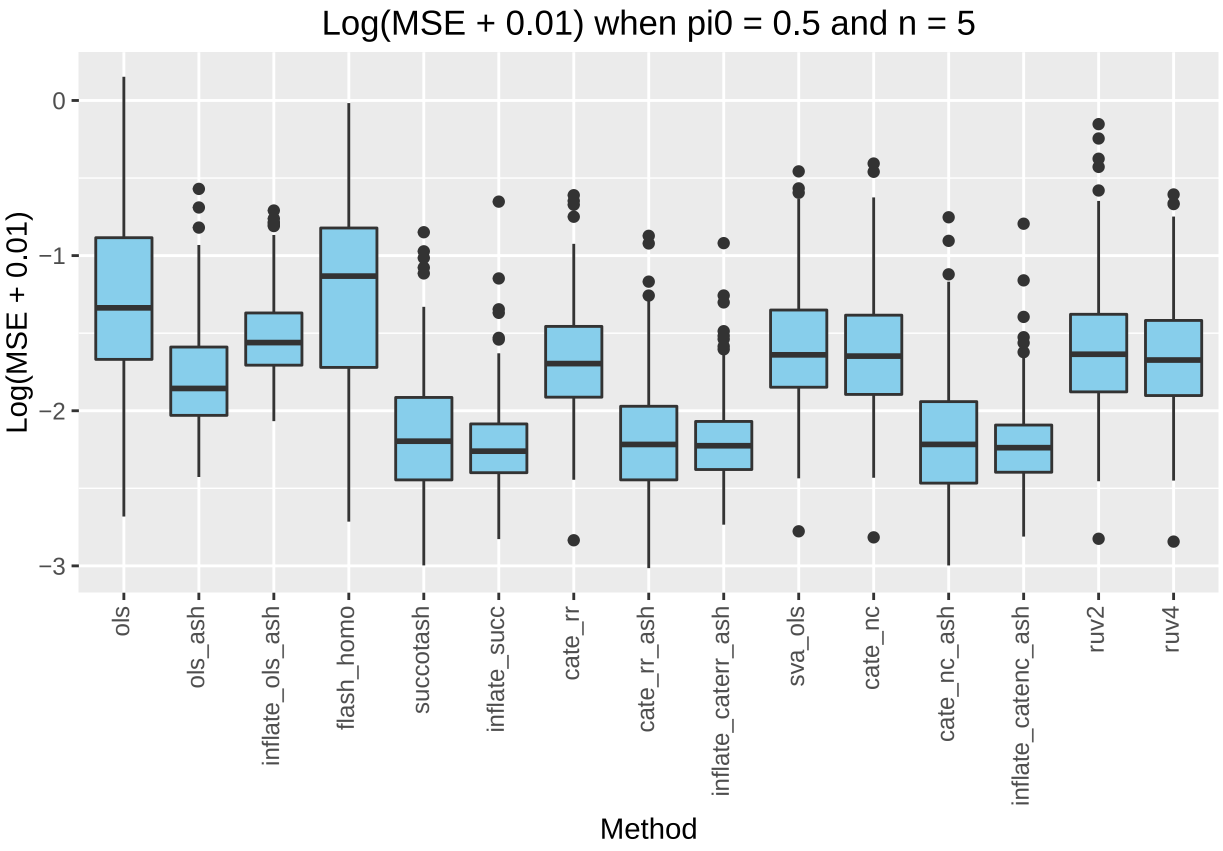

AUC when pi0 = 0.9 and n = 5
AUC
0.9
0.8
0.7
0.6
0.5
0.4
Method
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4


AUC when pi0 = 0.9 and n = 10

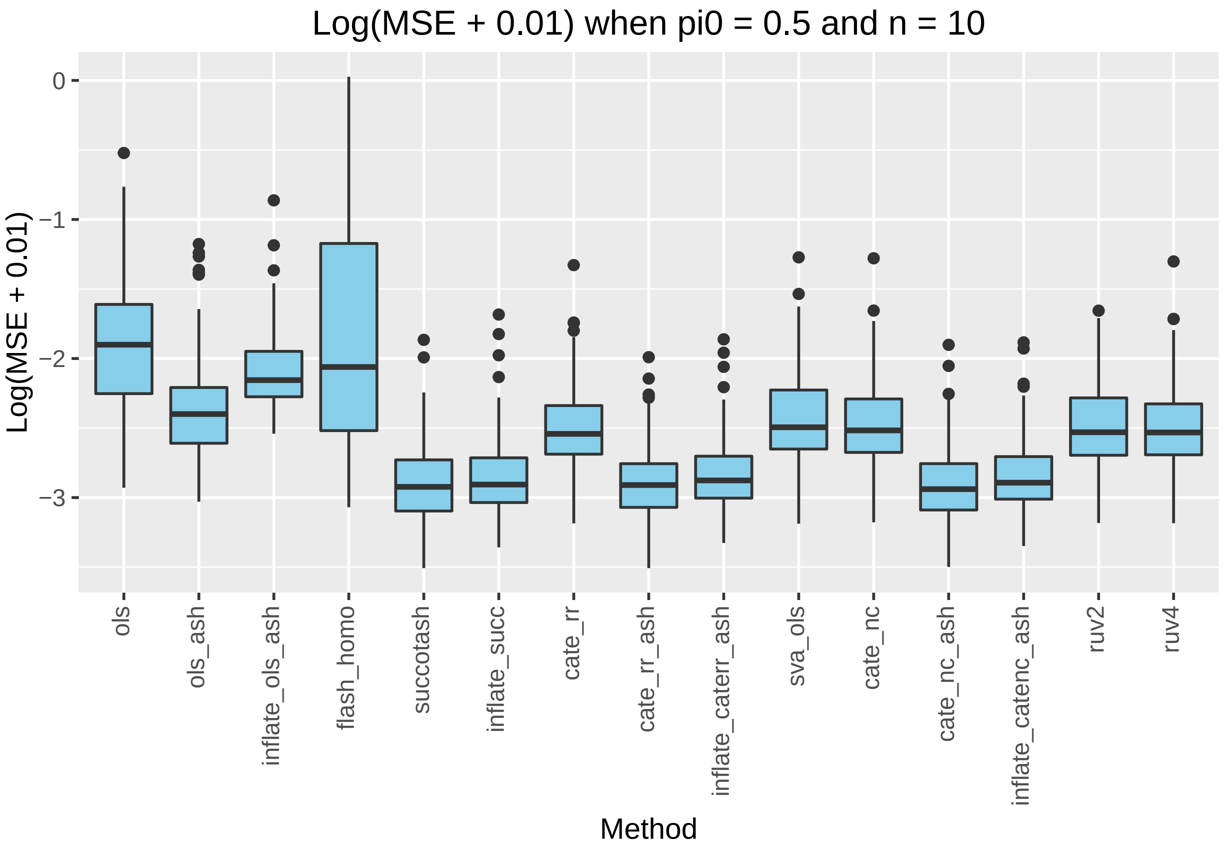

AUC when pi0 = 0.9 and n = 20

Log(MSE + 0.01) Plots

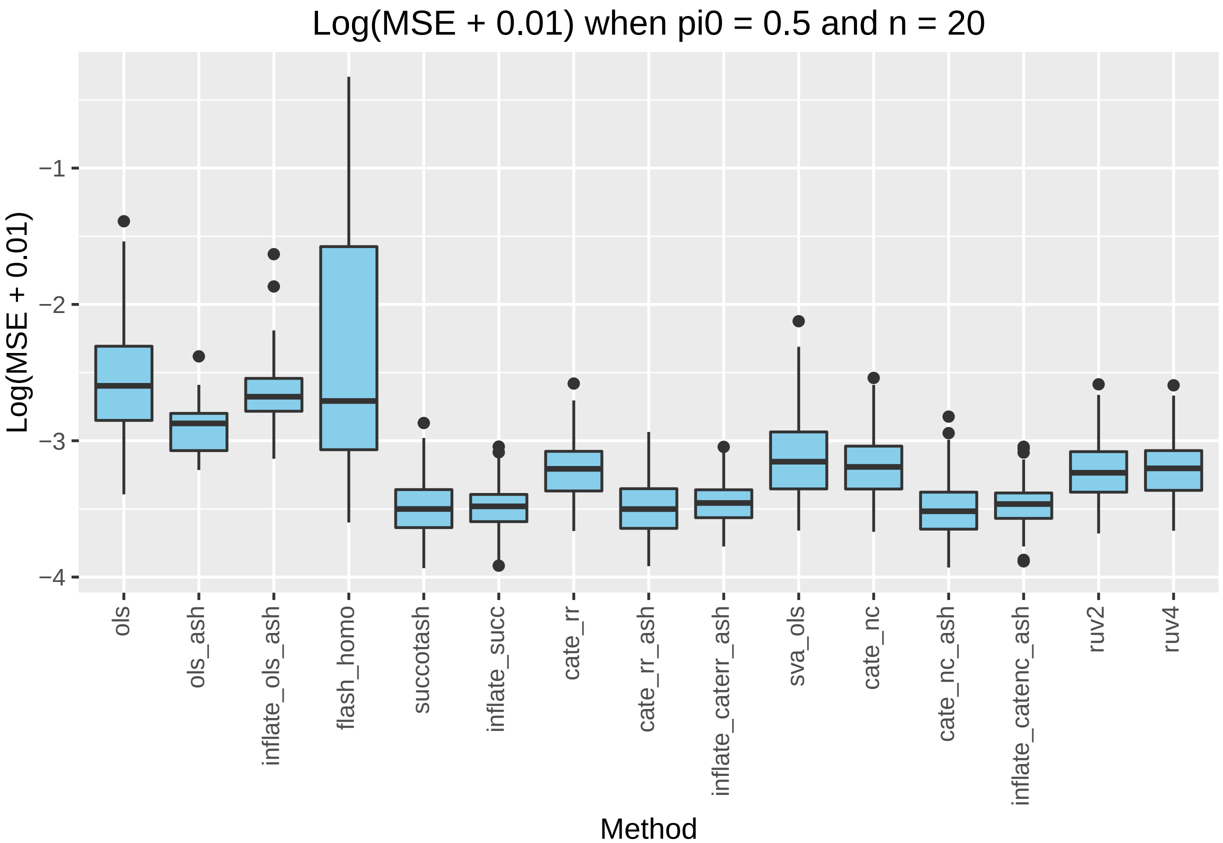
```
inflate_mse <- read.csv("mse_mat.csv")
reg_mse <- read.csv("../flash_v_rest_using_package/mse_mat.csv")
reg_mse$inflate_succ <- inflate_mse$succotash
reg_mse$inflate_caterr_ash <- inflate_mse$cate_rr_ash
reg_mse$inflate_catenc_ash <- inflate_mse$cate_nc_ash
reg_mse$inflate_ols_ash <- inflate_mse$ols_ash
reg_mse <- tbl_df(reg_mse)
reg_mse <- reg_mse[, c(1:2, 17, 3:4, 14, 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)
        )

        melted_df <- melt(subdf, id.vars = NULL)
        melted_df$value <- log(melted_df$value + 0.01)

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



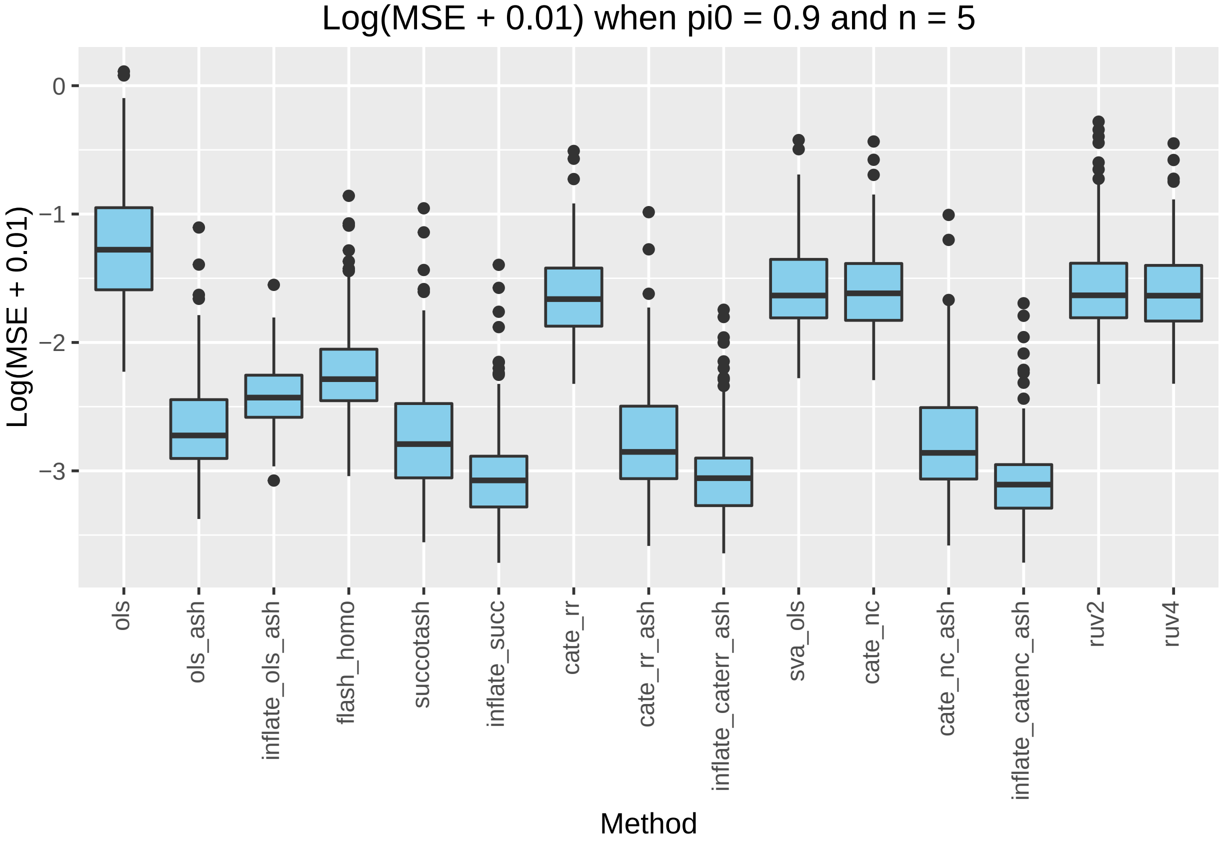
```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



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

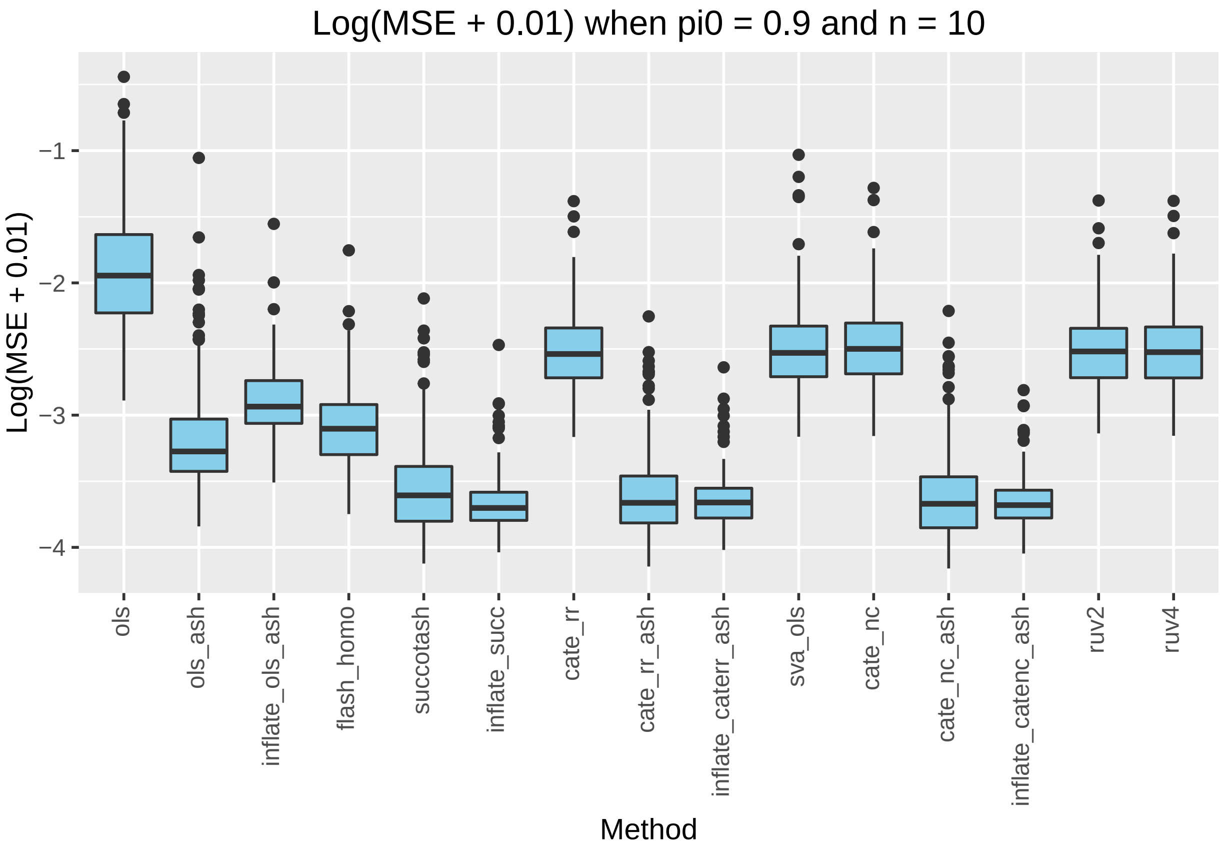


Log(MSE + 0.01) when pi0 = 0.5 and n = 20

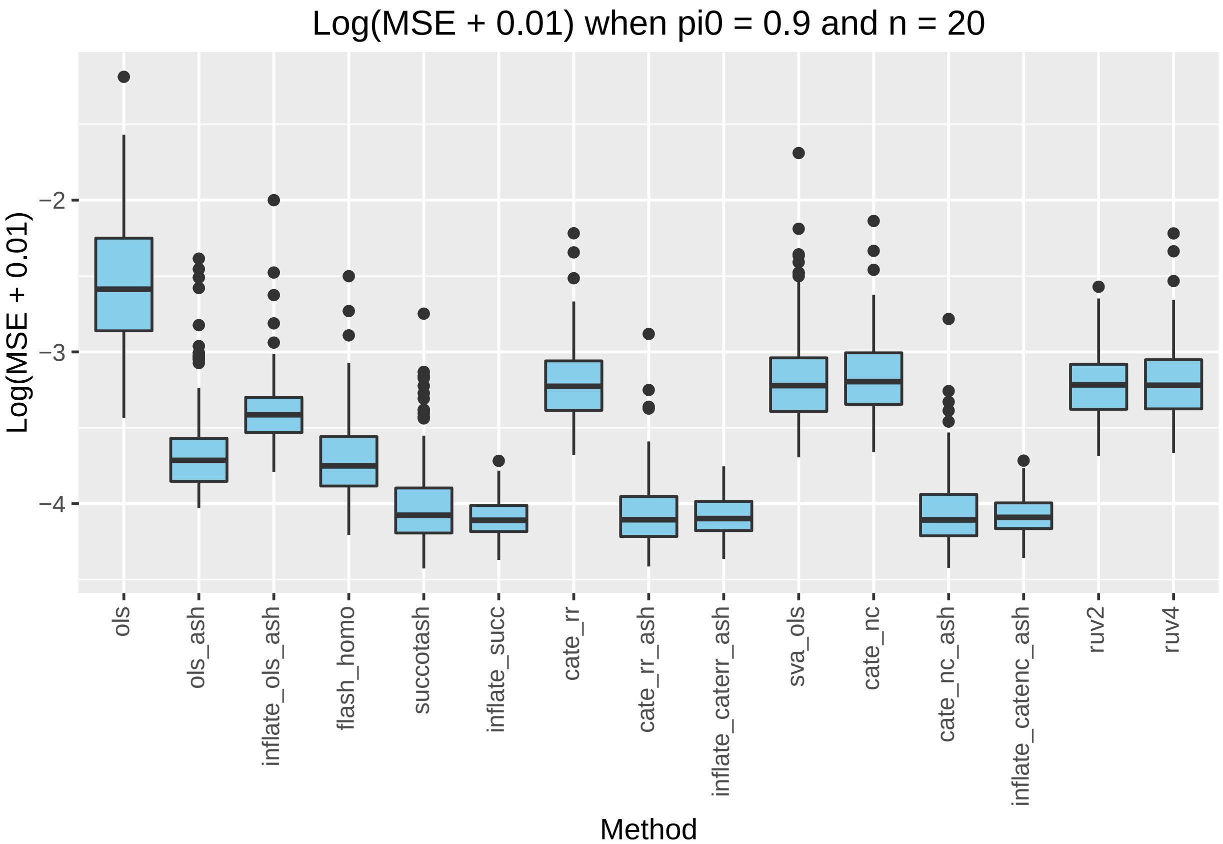


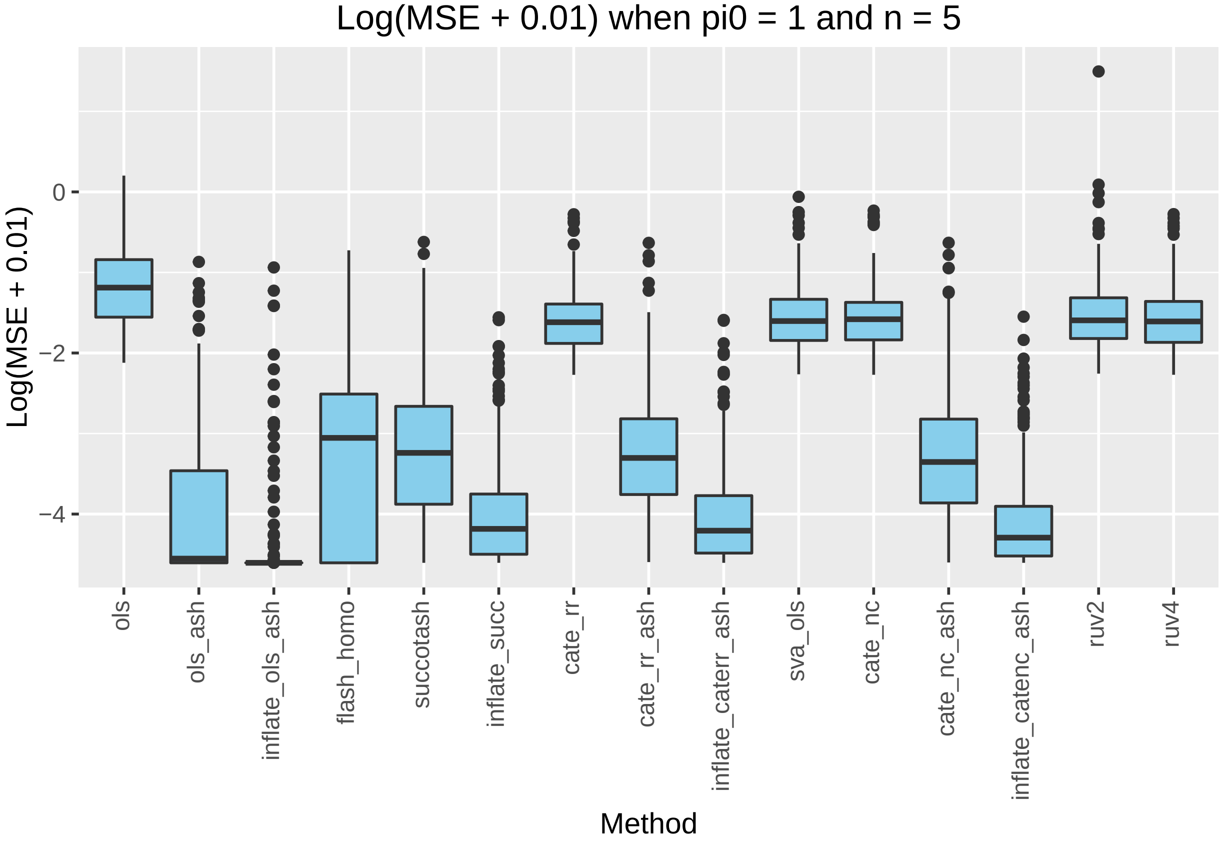
Log(MSE + 0.01) when pi0 = 0.9 and n = 5

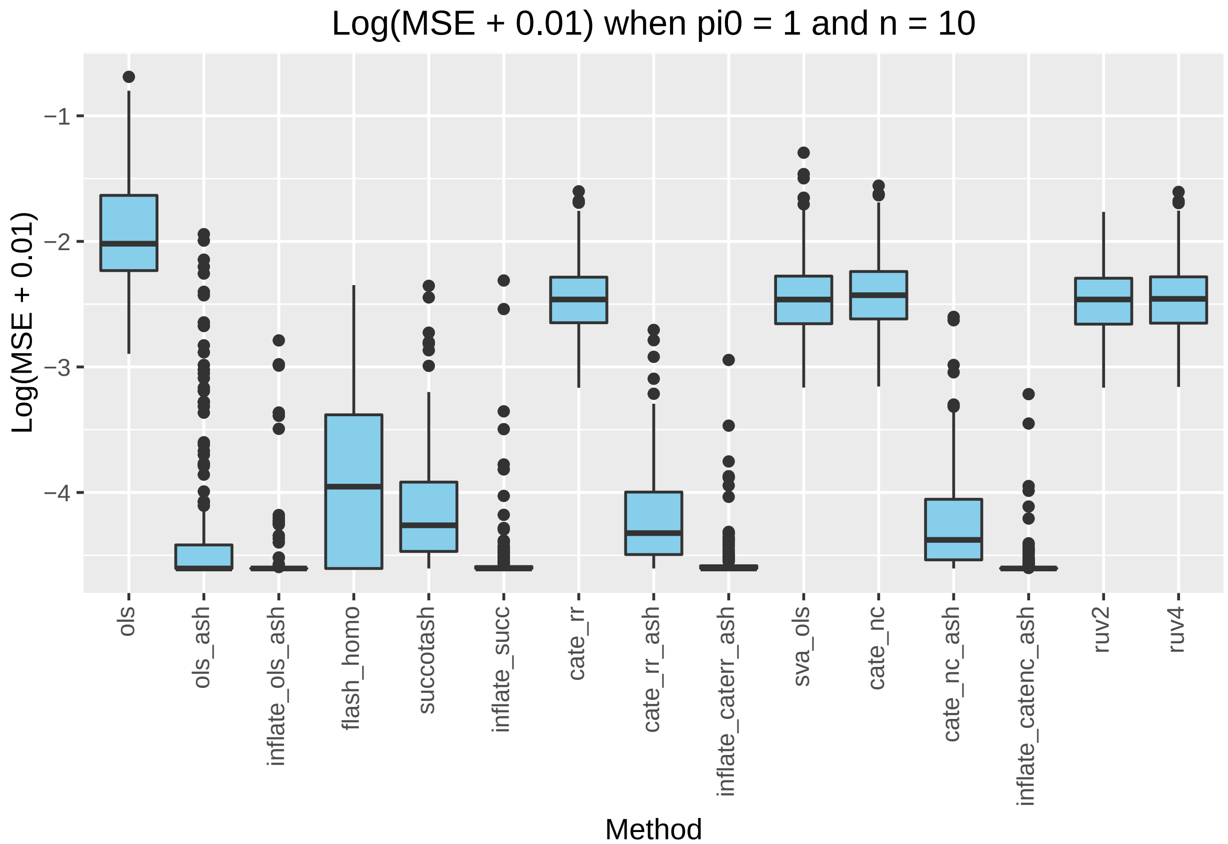
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

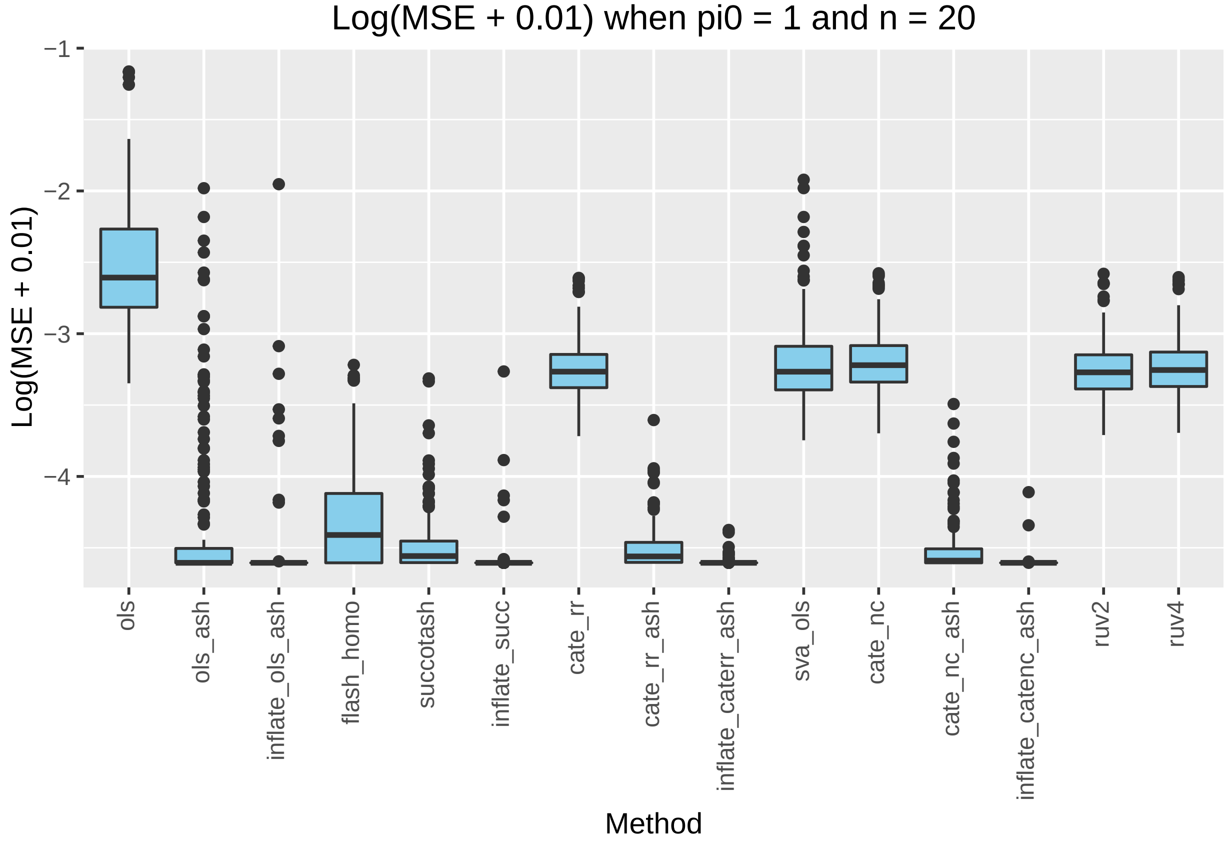


```
## Warning: Removed 89 rows containing non-finite values (stat_boxplot).
```


Log(MSE + 0.01) when pi0 = 0.9 and n = 20
Log(MSE + 0.01)
−2
−3
−4
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
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 = 5
Log(MSE + 0.01)
0
−2
−4
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
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 = 10


Log(MSE + 0.01) when pi0 = 1 and n = 20

```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] xtable_1.8-2   knitr_1.12.24  ggplot2_2.1.0  reshape2_1.4.1
## [5] dplyr_0.4.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.2       magrittr_1.5      munsell_0.4.3
##  [4] colorspace_1.2-6  R6_2.1.1          stringr_1.0.0
##  [7] plyr_1.8.3        tools_3.2.5       parallel_3.2.5
## [10] grid_3.2.5        gtable_0.2.0      DBI_0.3.1
## [13] htmltools_0.3.5   yaml_2.1.13       lazyeval_0.1.10
## [16] assertthat_0.1    digest_0.6.9      formatR_1.3
## [19] codetools_0.2-14  evaluate_0.8.3    rmarkdown_0.9.5.9
## [22] labeling_0.3      stringi_1.0-1     compiler_3.2.5
## [25] scales_0.4.0
```