Include Scaling Penalty in SUCCOTASH and Inflate Variance in ASH

David Gerard

2016-05-02

Abstract

This file contains simulations on two new approaches. One adds a small "penalty" to the variance inflation parameter of SUCCOTASH. The other applies SVA + OLS + ASH while estimating a variance inflation parameter. The SUCCOTASH method with a penalty on the variance inflation parameter worked only slightly better than not including the penalty. The new ASH method was much more anti-conservative than just doing OLS + ASH.

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.

After running simulations with this scaling of the variances, it was behaving slightly anti-conservative. My guess is that this is the usual biasedness of the MLE in variance estimation. One ad-hoc solution might be to add $q\log(\lambda)$ as a penalty to the optimization problem, where q is the number of covariates in X and λ is the variance inflation parameter. The intuition behind this is that if we use this "prior" under the normal means/variances problem then we get the UMVUE for the variance instead of the MLE. There might be more principled ways to do something like this.

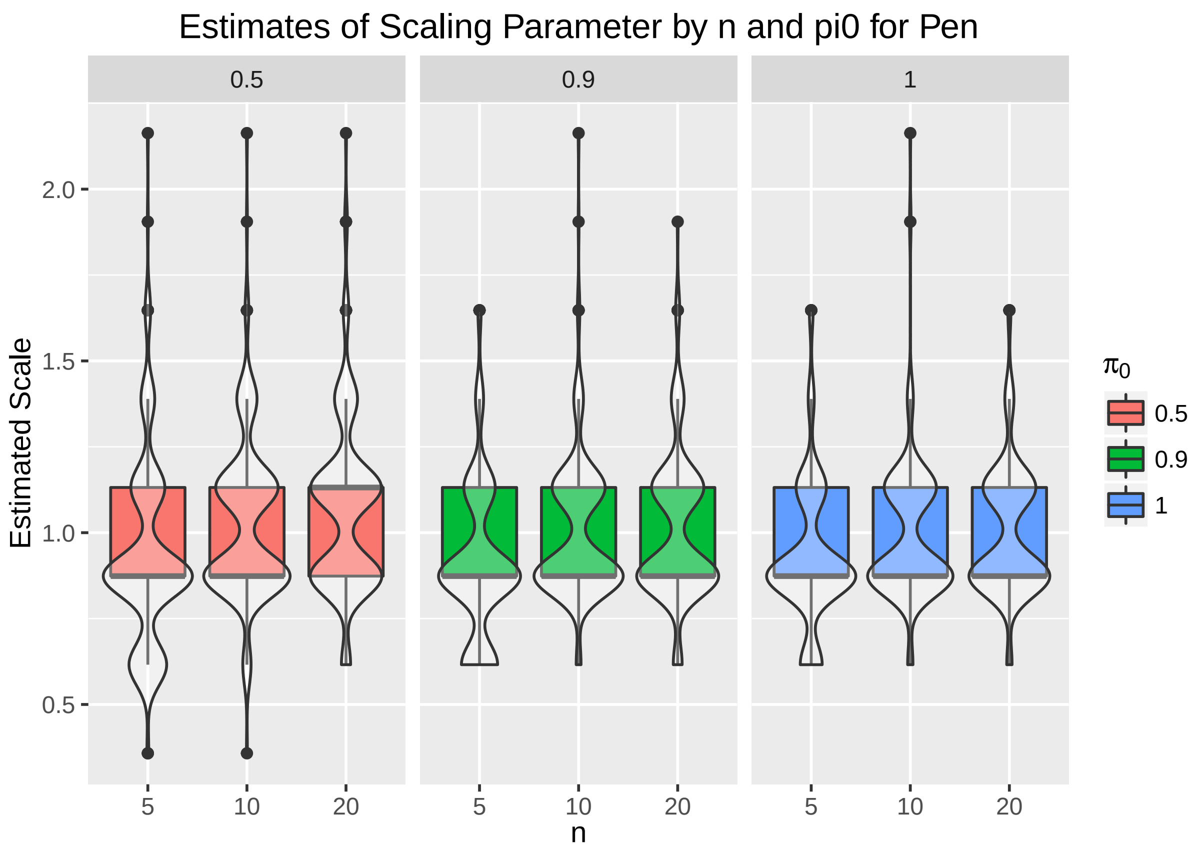
The results using this approach are labeled "succ_pen" in the plots below. There seems to be a slight improvement in estimating π_0, but it's very minor. Importantly, the estimates of π_0 are still anti-conservative. MSE and AUC are almost exactly the same as without this penalty.

As inflating the variance of CATE then using ASH worked very well, I tried to estimating this variance inflation when using SVA to estimate the surrogate variables. The pipeline was

1. Use SVA to get the surrogate variables.
2. Use OLS to get the $\hat{\beta}_i$'s and $\hat{s}_i$'s.
3. Use ASH with data $\hat{\beta}_i$ and standard devitiations $\lambda\hat{s}_i$, where λ varies on a grid of 20 from 0.1 to 4.
4. Choose the ASH output with the largest log-likelihood.

This method worked really poorly in terms of estimating π_0, but actually worked better in terms of MSE and AUC. It never does better than SUCCOTASH with the variance inflation parameter, though. One weird thing is that the distribution of the estimates of the variance inflation parameter looks the same under each scenario. Most of the time, the variance was actually SHRUNK.

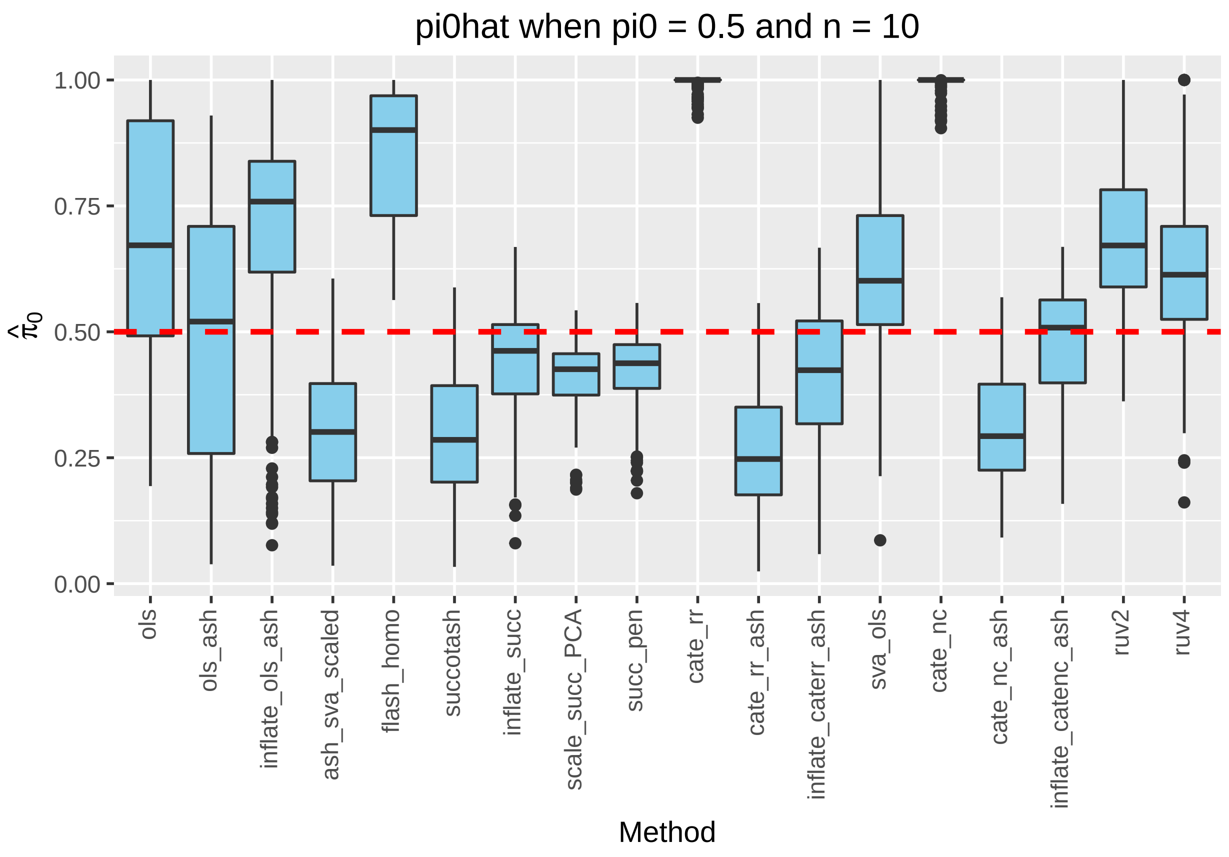
```r
scale_mat_ash <- read.csv("scale_est_svaash_mc.csv")
ggplot(data = scale_mat_ash, mapping = aes(x = factor(nsamp), y = post_inflate,
                                           fill = factor(nullpi))) +
    facet_grid(.~nullpi) +
    geom_boxplot() + geom_violin(fill = I("white"), alpha = 0.3) +
    xlab(expression(n)) + ylab("Estimated Scale") +
    scale_fill_discrete(name=expression(pi[0])) +
        ggtitle("Estimates of Scaling Parameter by n and pi0 for Pen")
```

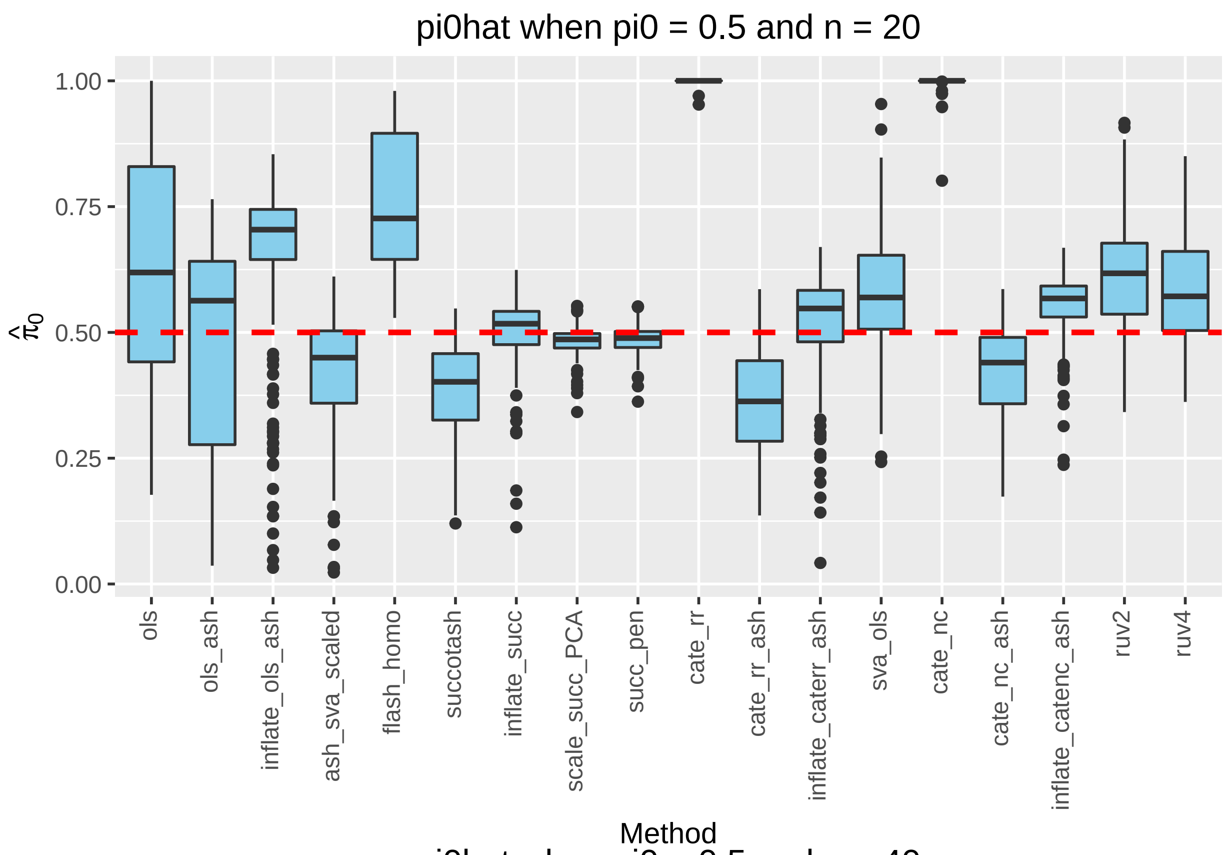


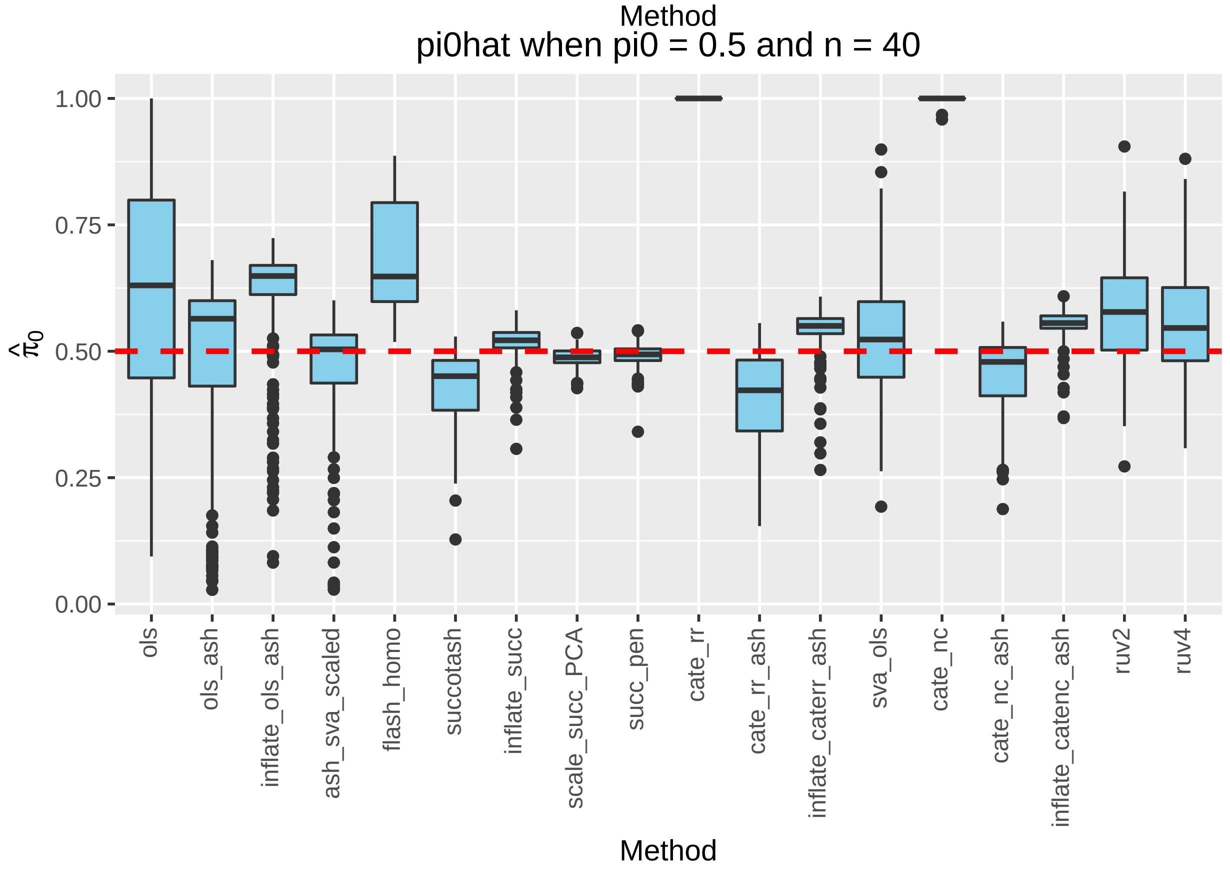
$\hat{\pi}_0$ Plots

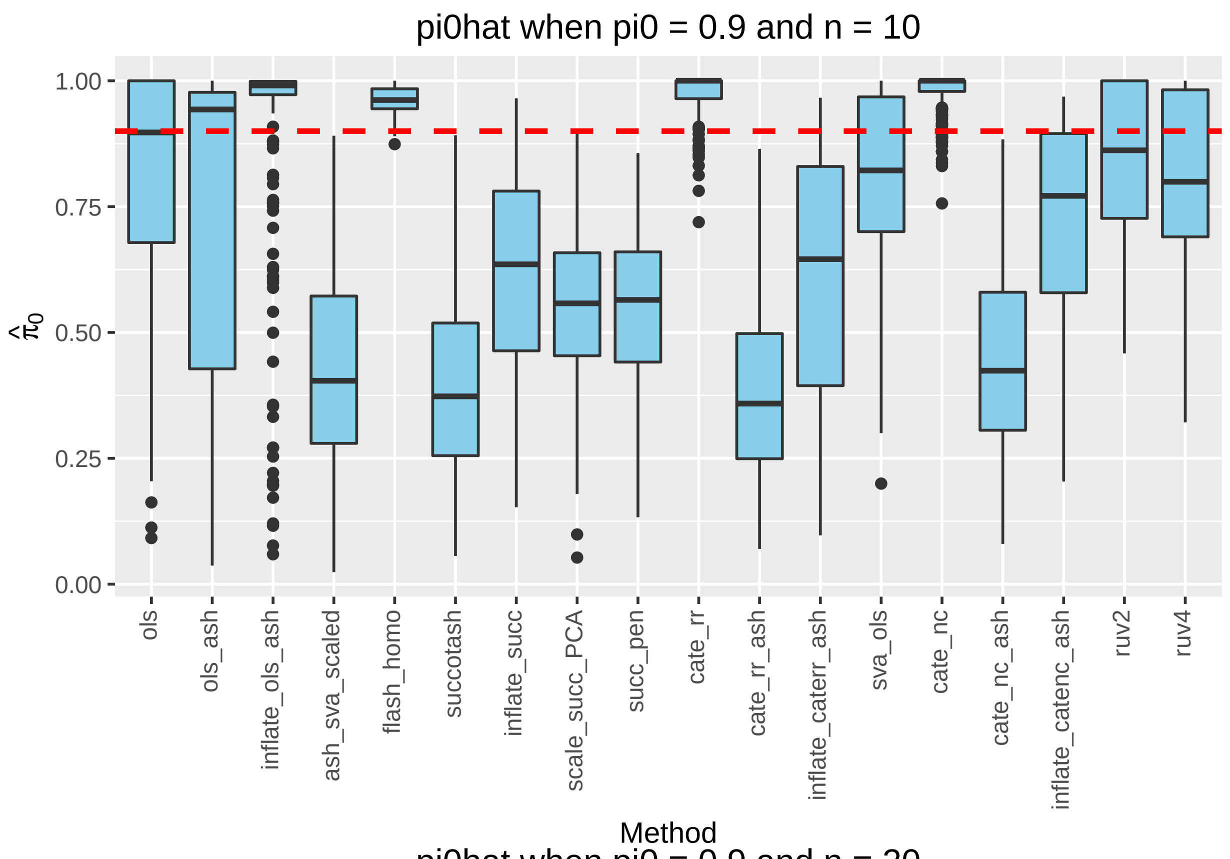
```r
double_pi0 <- read.csv("../double_succ/pi0_mat.csv")
reg_pi0 <- read.csv("../flash_v_rest_using_package/pi0_mat.csv")
scale_pi0 <- read.csv("../succ_scaled/pi0_ssuc.csv")
scale_pi0_pen <- read.csv("../succ_scaled_pen/pi0_ssuc_mc.csv")
ash_sva <- read.csv("pi0_svaash_mc.csv")
reg_pi0$inflate_succ <- double_pi0$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$scale_succ_PCA <- scale_pi0$scale_suc1
reg_pi0$succ_pen <- scale_pi0_pen$post_inflate
reg_pi0$ash_sva_scaled <- ash_sva$post_inflate
```

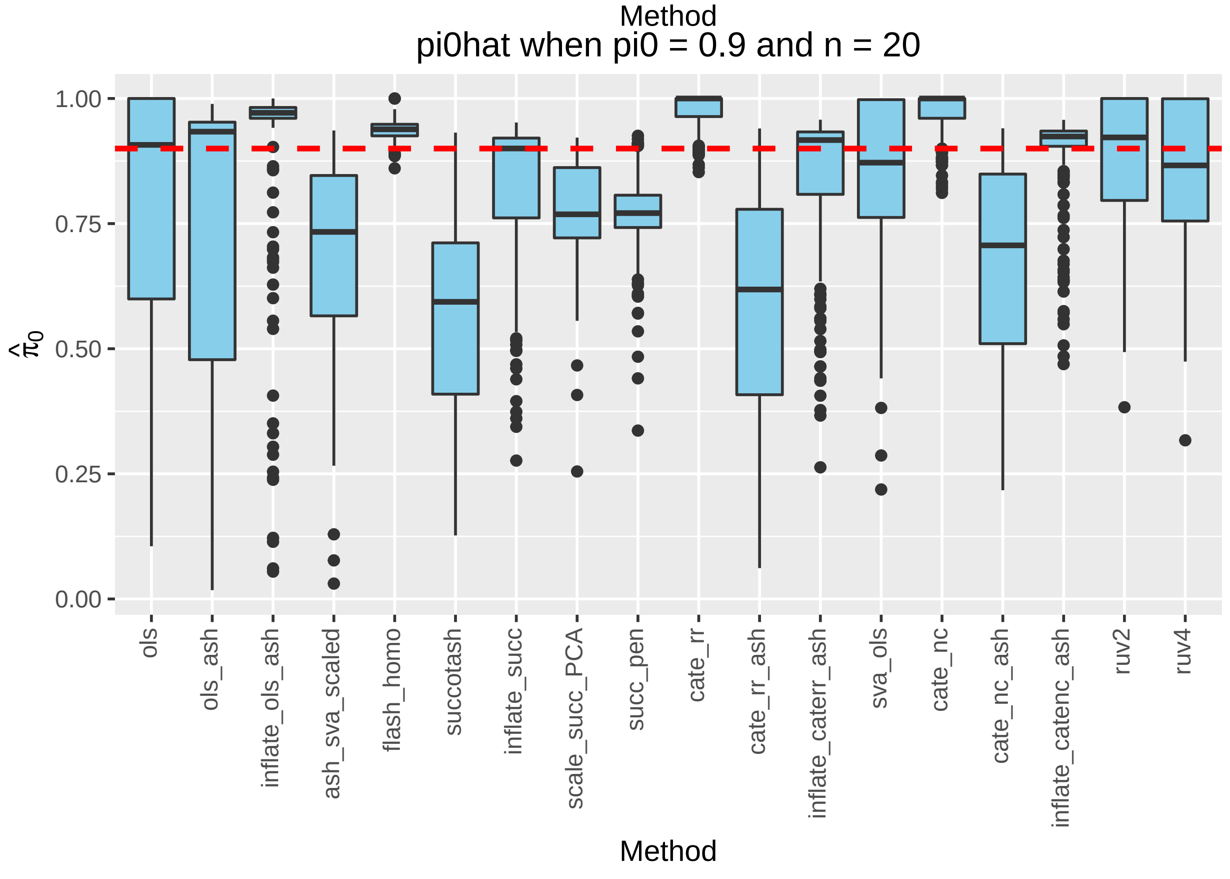
```r
reg_pi0 <- tbl_df(reg_pi0)
reg_pi0 <- reg_pi0[, c(1:2, 17, 20, 3:4, 14, 18:19, 5:6, 15, 7:9, 16, 10:13)]
nsamp_seq <- unique(reg_pi0$nsamp)
nullpi_seq <- unique(reg_pi0$nullpi)
for (current_pi in nullpi_seq) {
    for (current_nsamp in nsamp_seq) {


        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)
    }
}
```

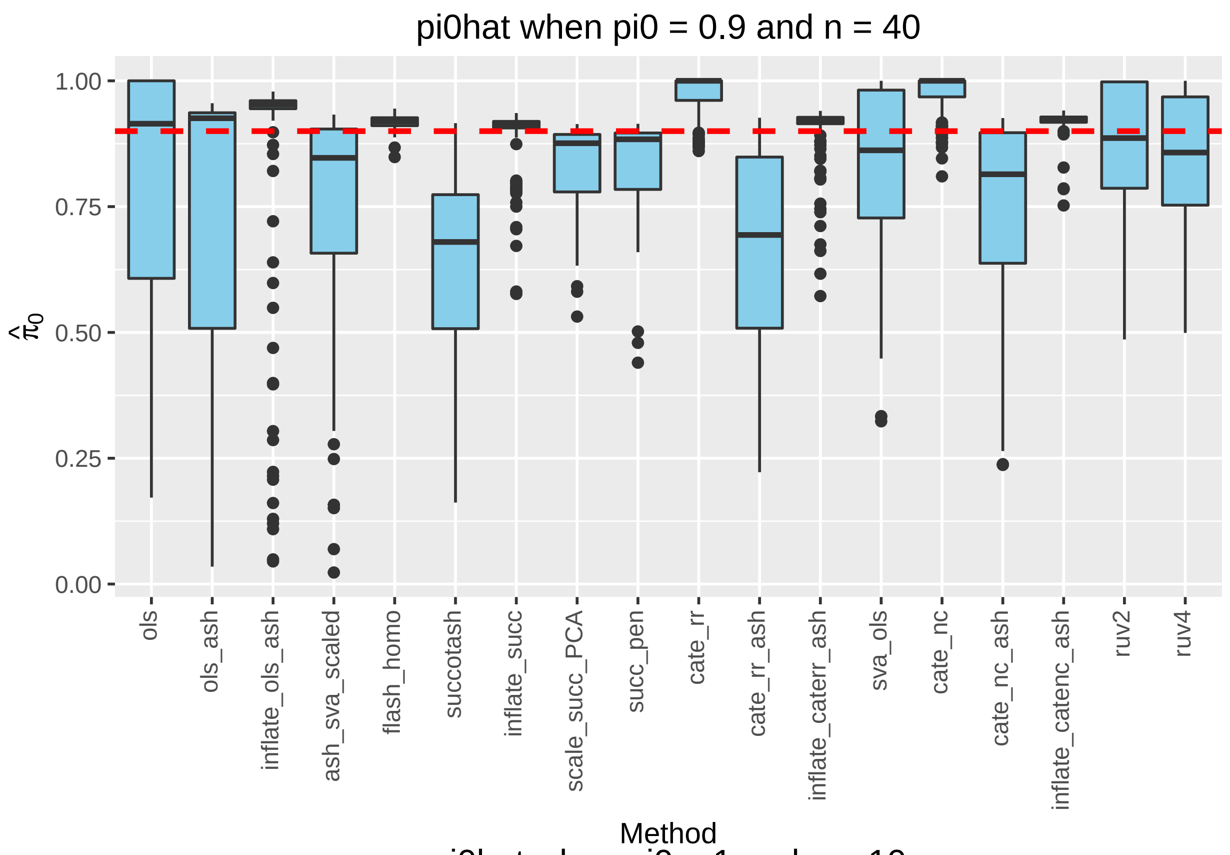


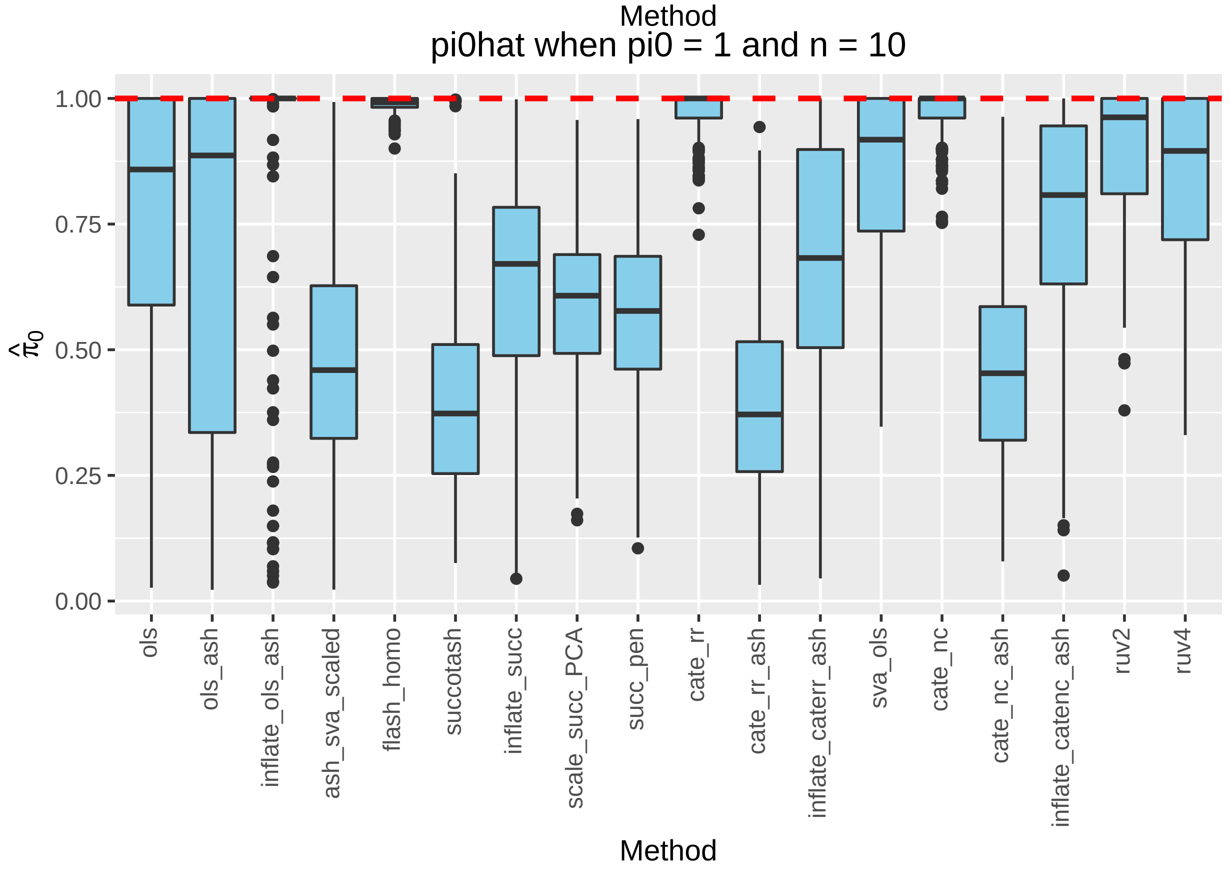

pi0hat when pi0 = 0.5 and n = 20
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

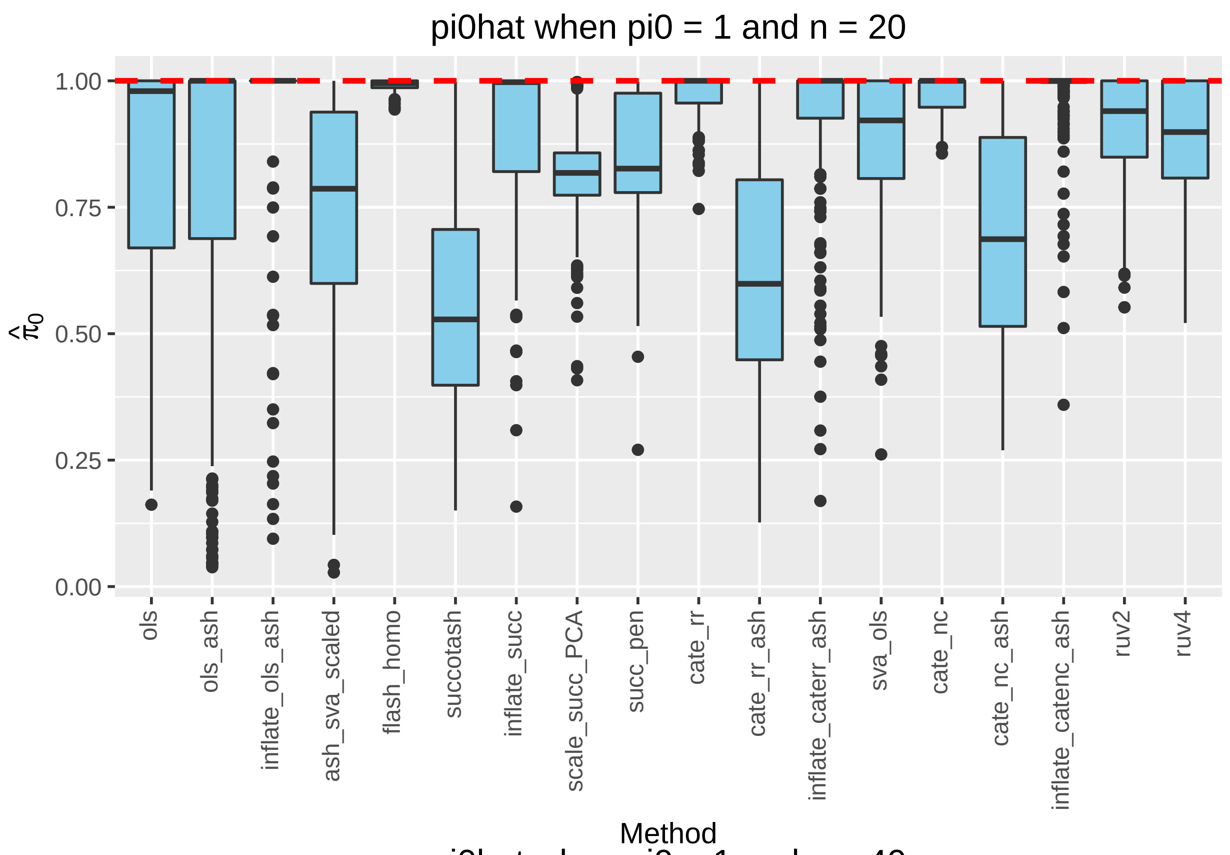

pi0hat when pi0 = 0.5 and n = 40
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

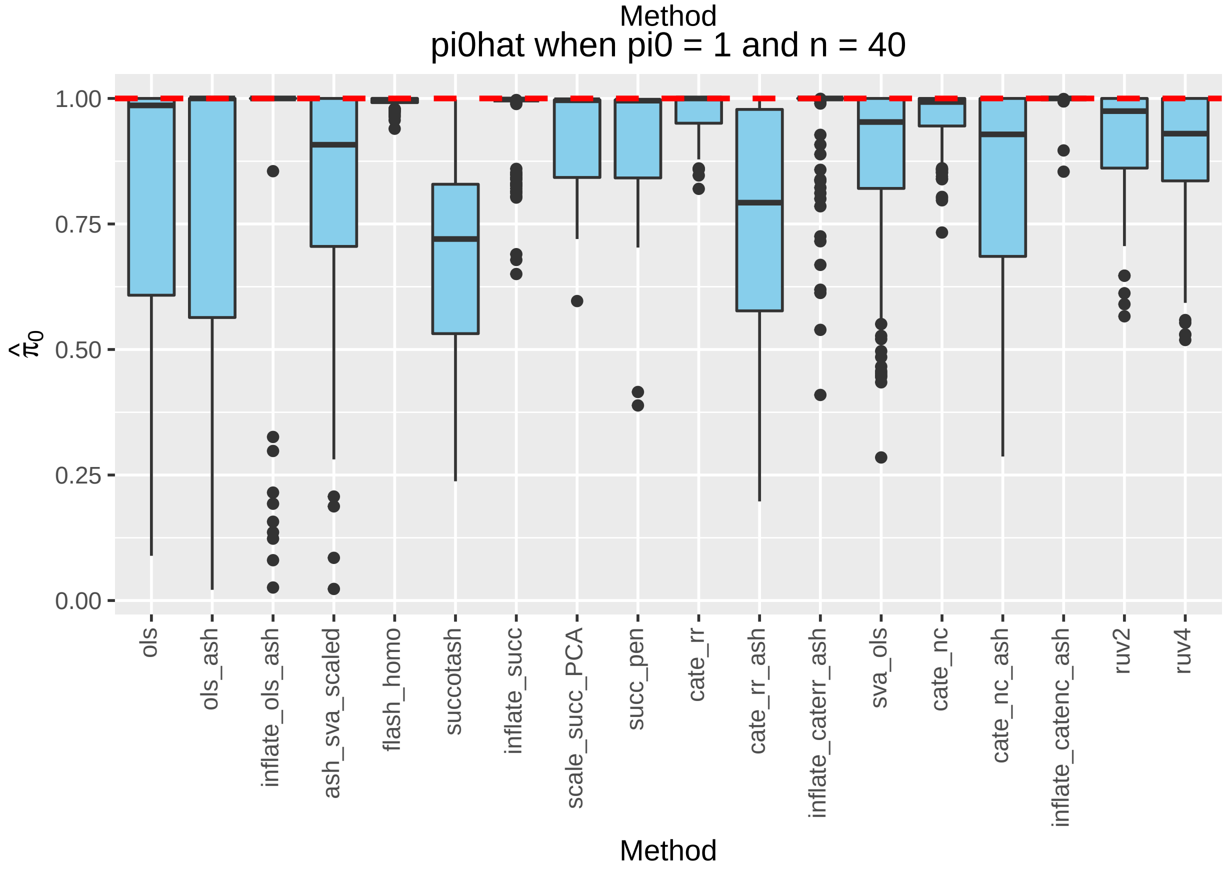

pi0hat when pi0 = 0.9 and n = 10
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


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
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
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
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

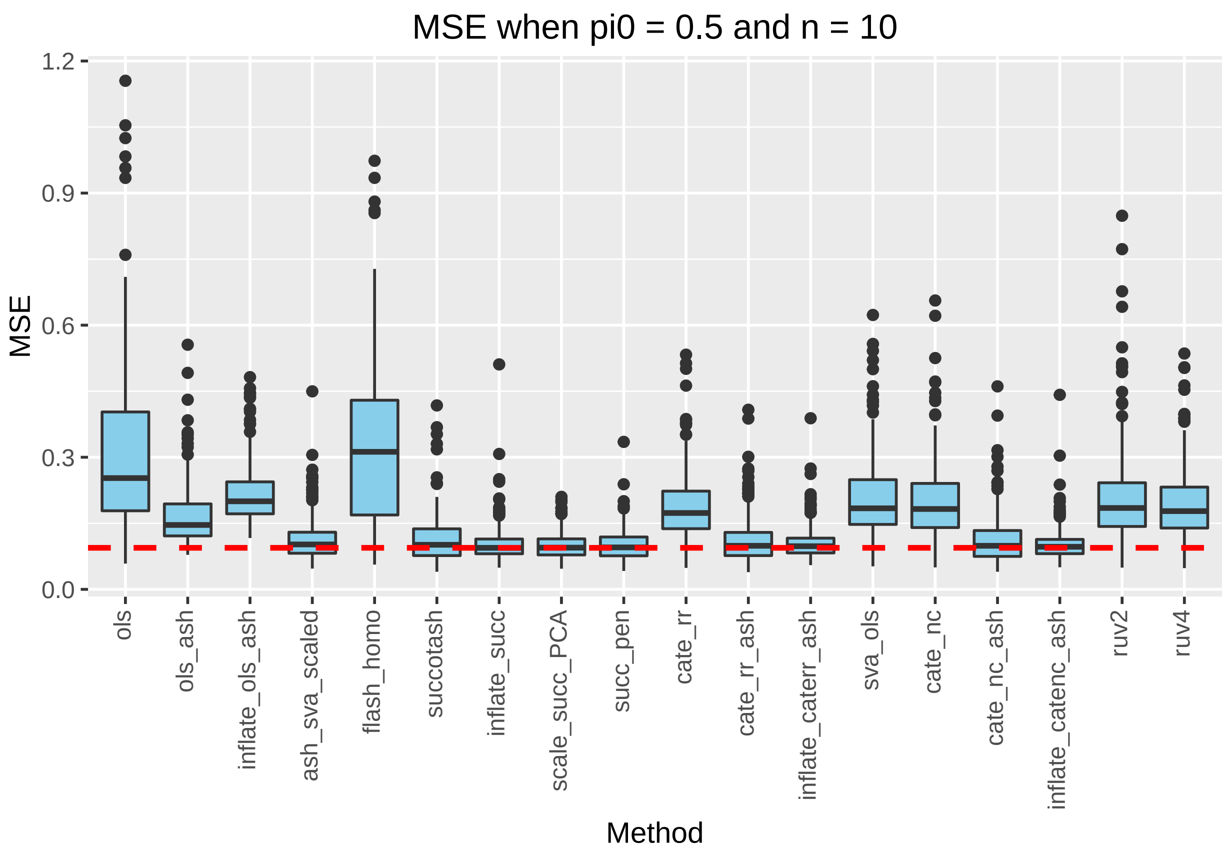

pi0hat when pi0 = 1 and n = 10
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


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
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

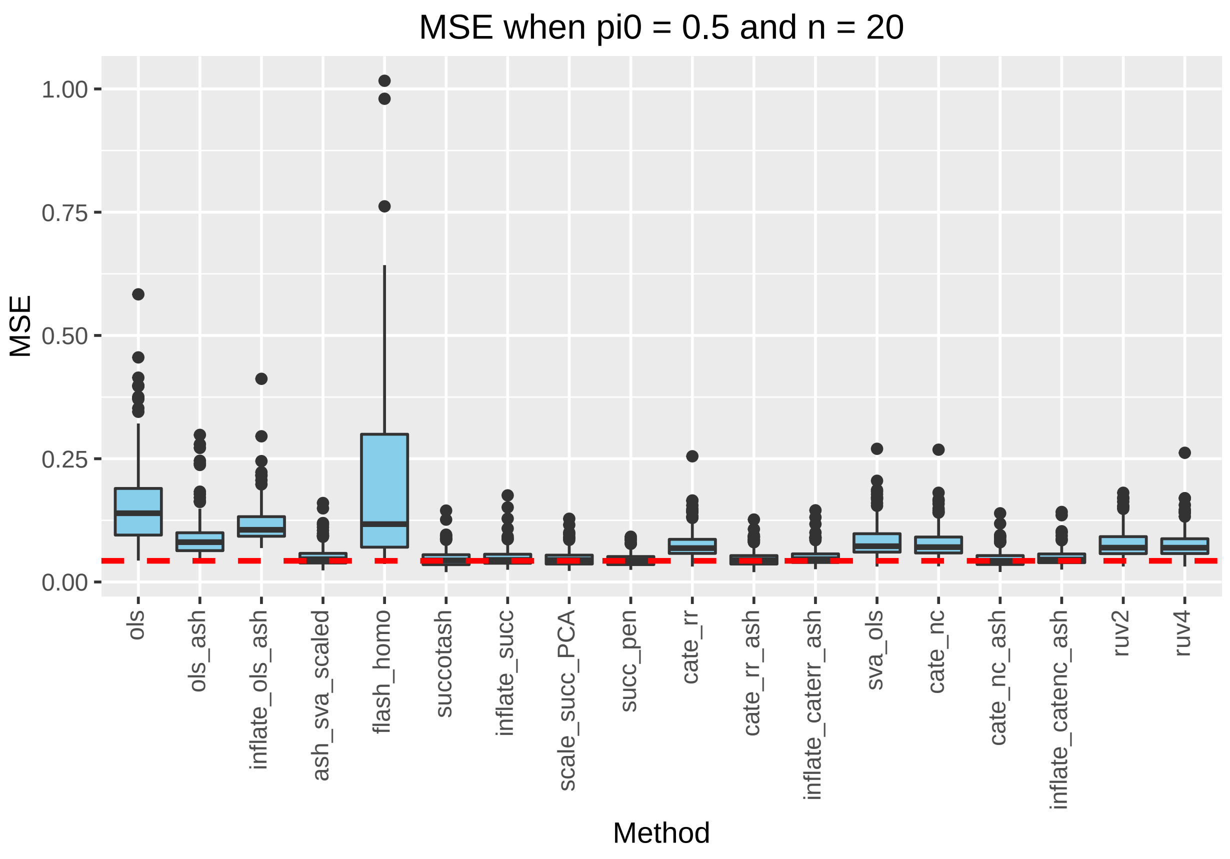

pi0hat when pi0 = 1 and n = 40
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

MSE Plots

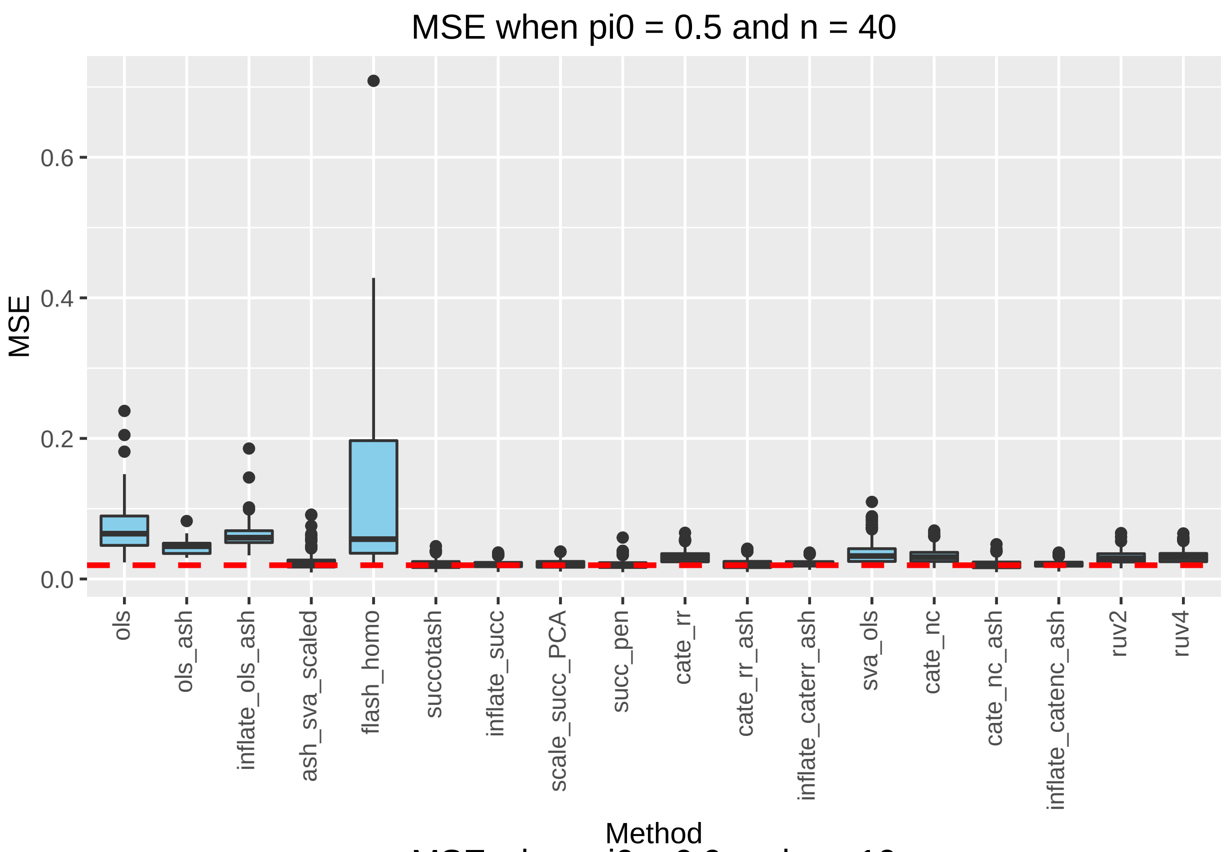
```r
double_mse <- read.csv("../double_succ/mse_mat.csv")
reg_mse <- read.csv("../flash_v_rest_using_package/mse_mat.csv")
scale_mse <- read.csv("../succ_scaled/mse_ssuc.csv")
scale_mse_pen <- read.csv("../succ_scaled_pen/mse_ssuc_mc.csv")
ash_sva <- read.csv("mse_svaash_mc.csv")
reg_mse$inflate_succ <- double_mse$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$scale_succ_PCA <- scale_mse$scale_suc1
reg_mse$succ_pen <- scale_mse_pen$post_inflate
reg_mse$ash_sva_scaled <- ash_sva$post_inflate
reg_mse <- tbl_df(reg_mse)
reg_mse <- reg_mse[, c(1:2, 17, 20, 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)
    }
}
```

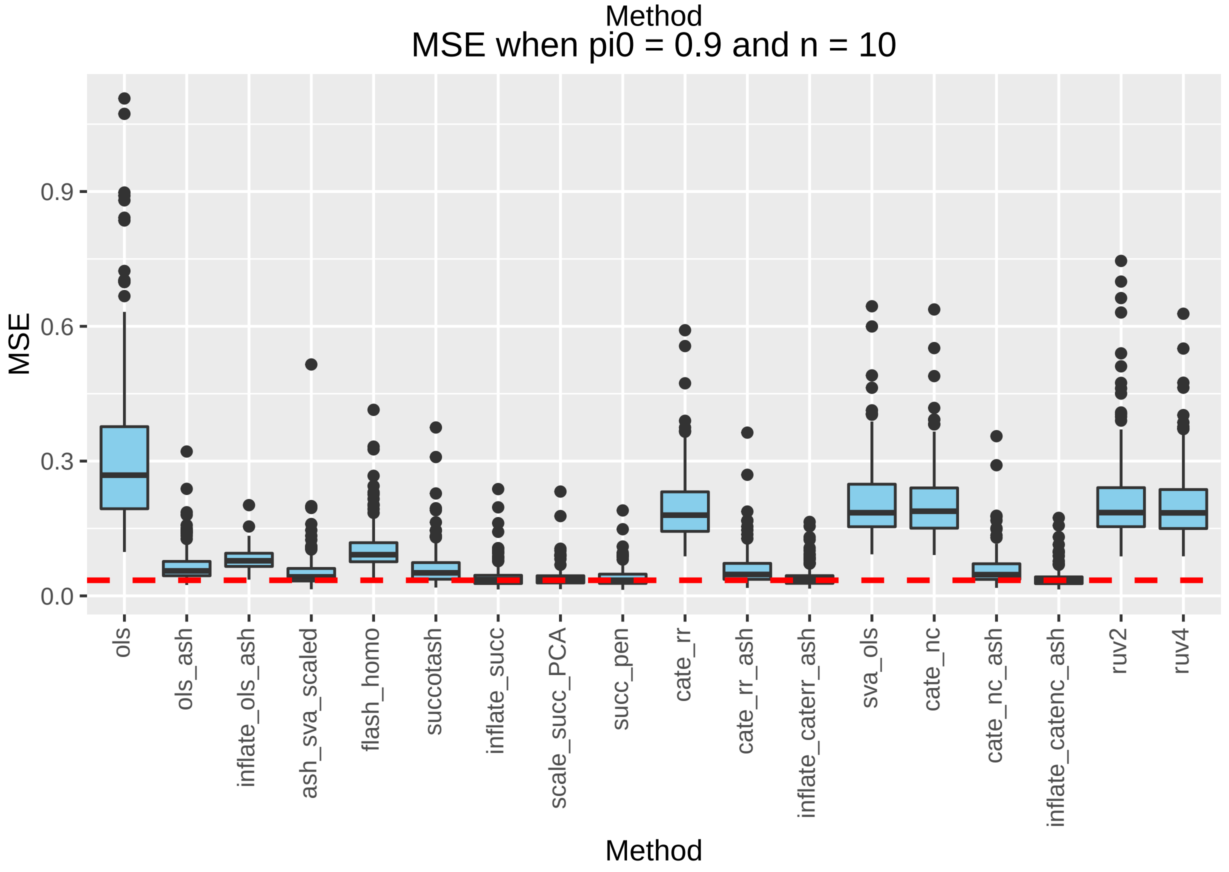


```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

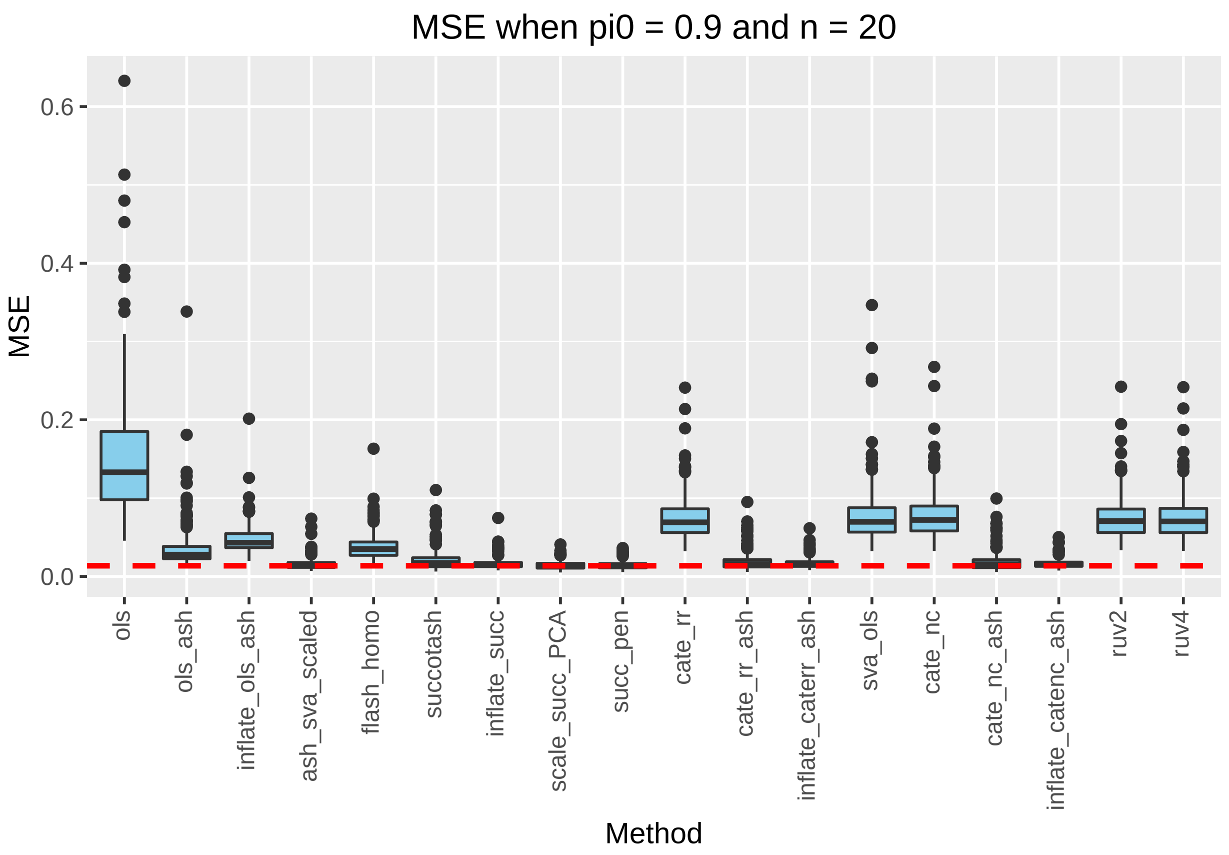


```
## Warning: Removed 203 rows containing non-finite values (stat_boxplot).
```

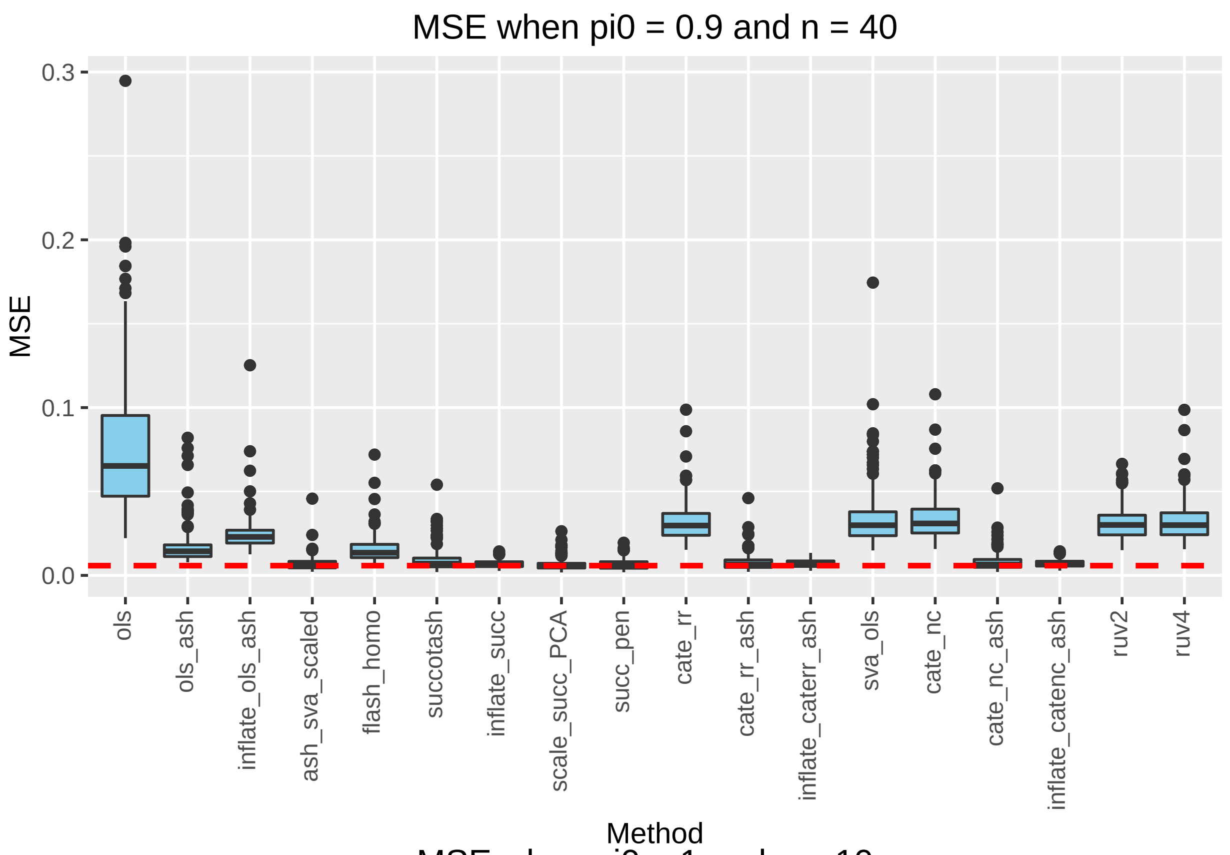


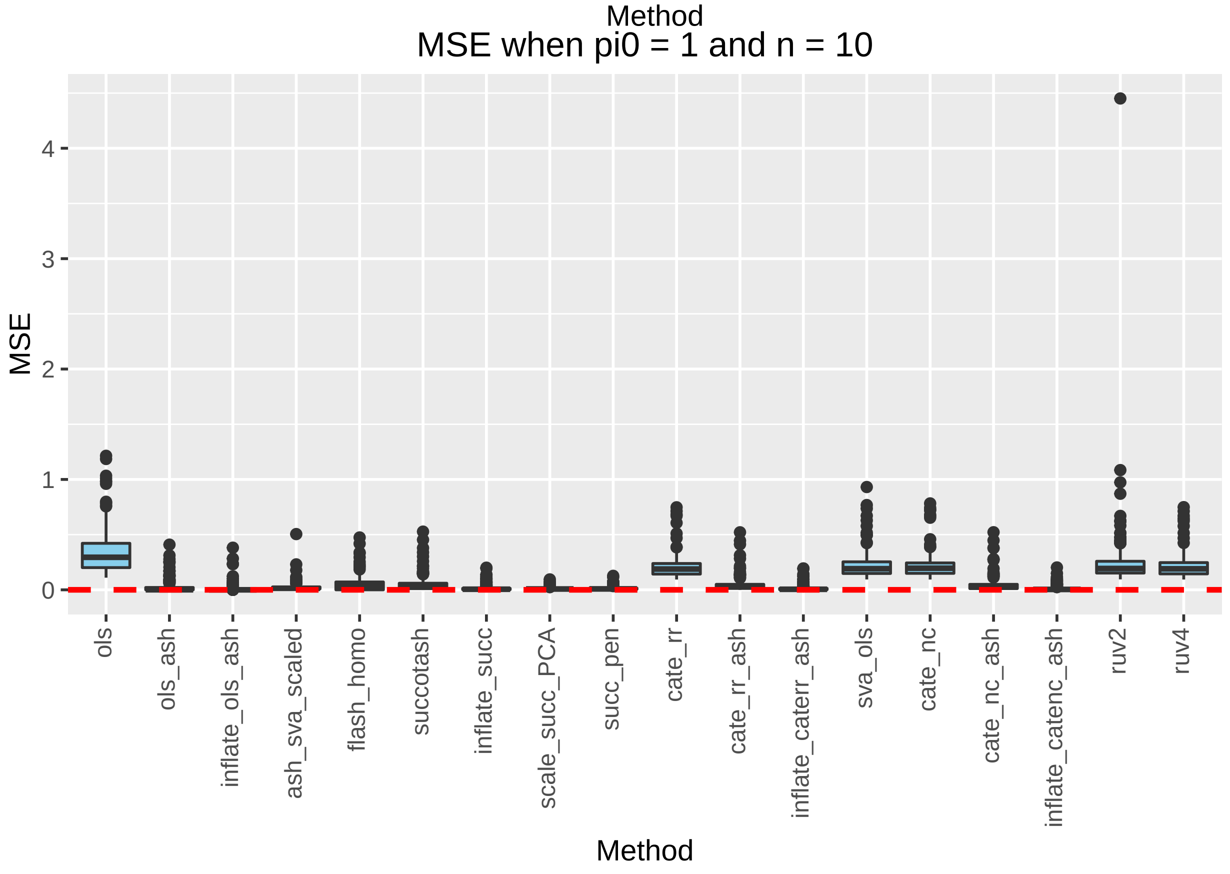


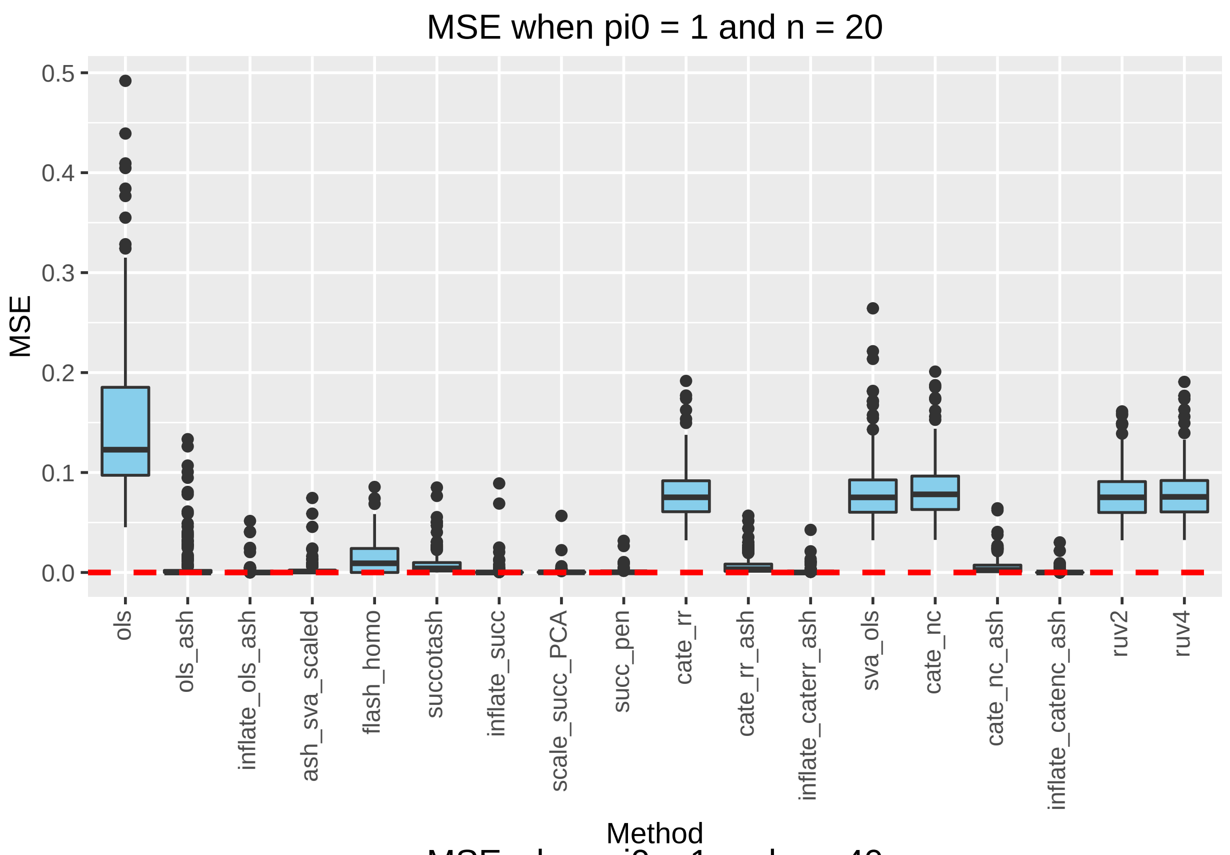
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

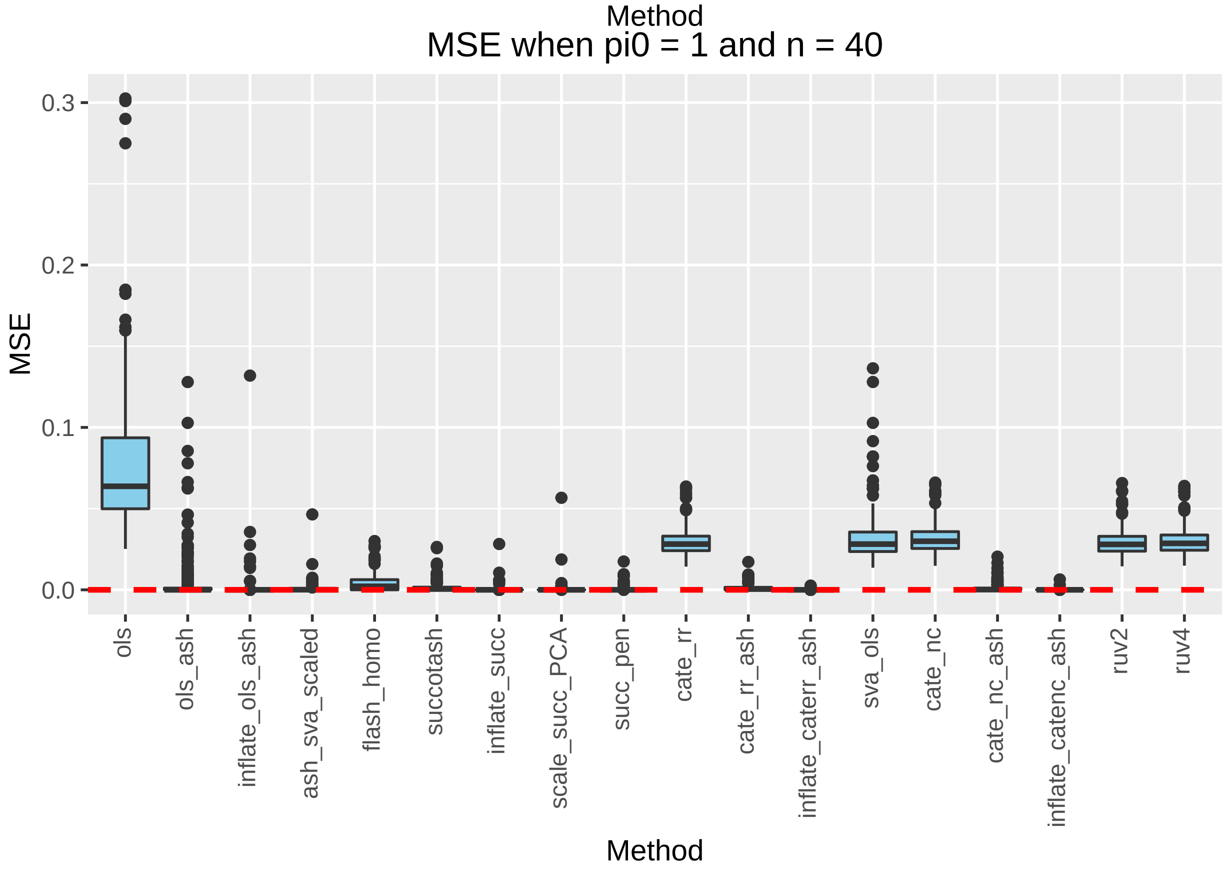


```
## Warning: Removed 89 rows containing non-finite values (stat_boxplot).
```


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

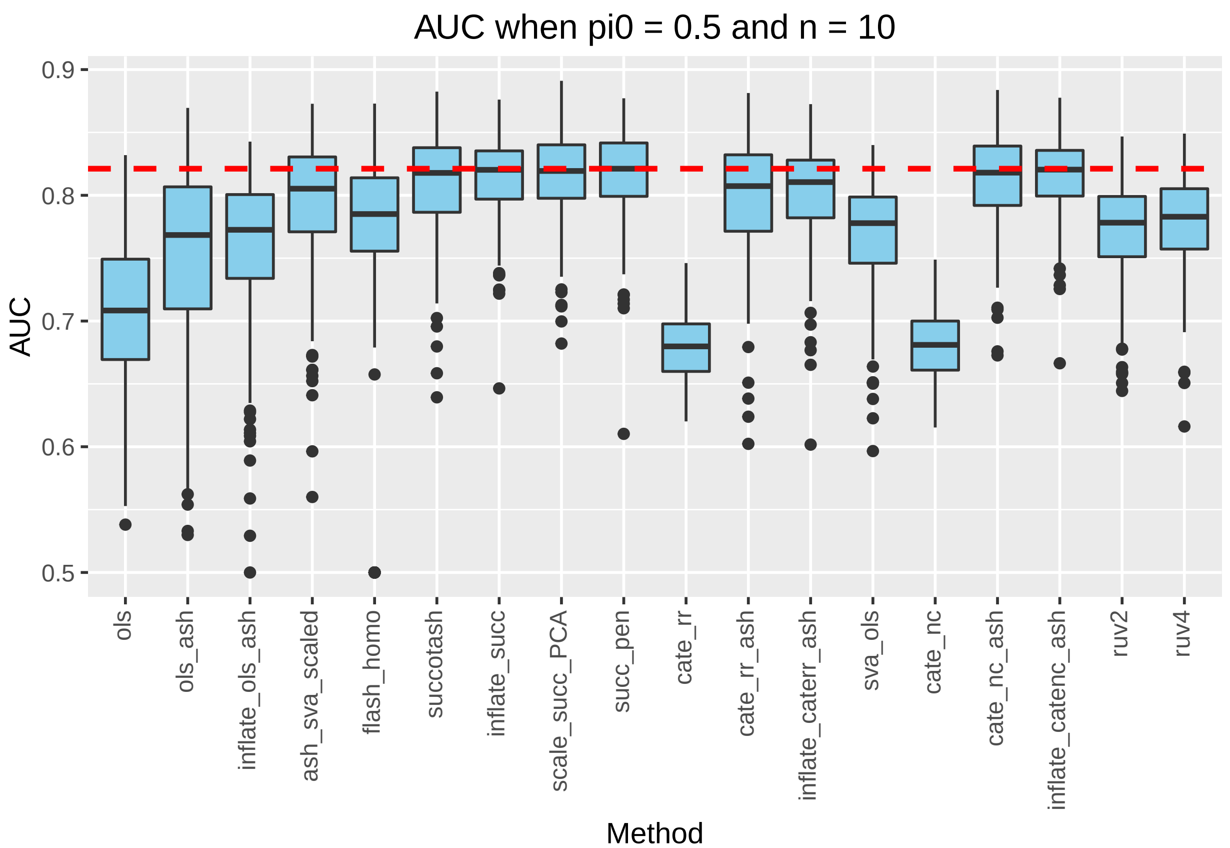

MSE when pi0 = 1 and n = 10
MSE
4
3
2
1
0
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

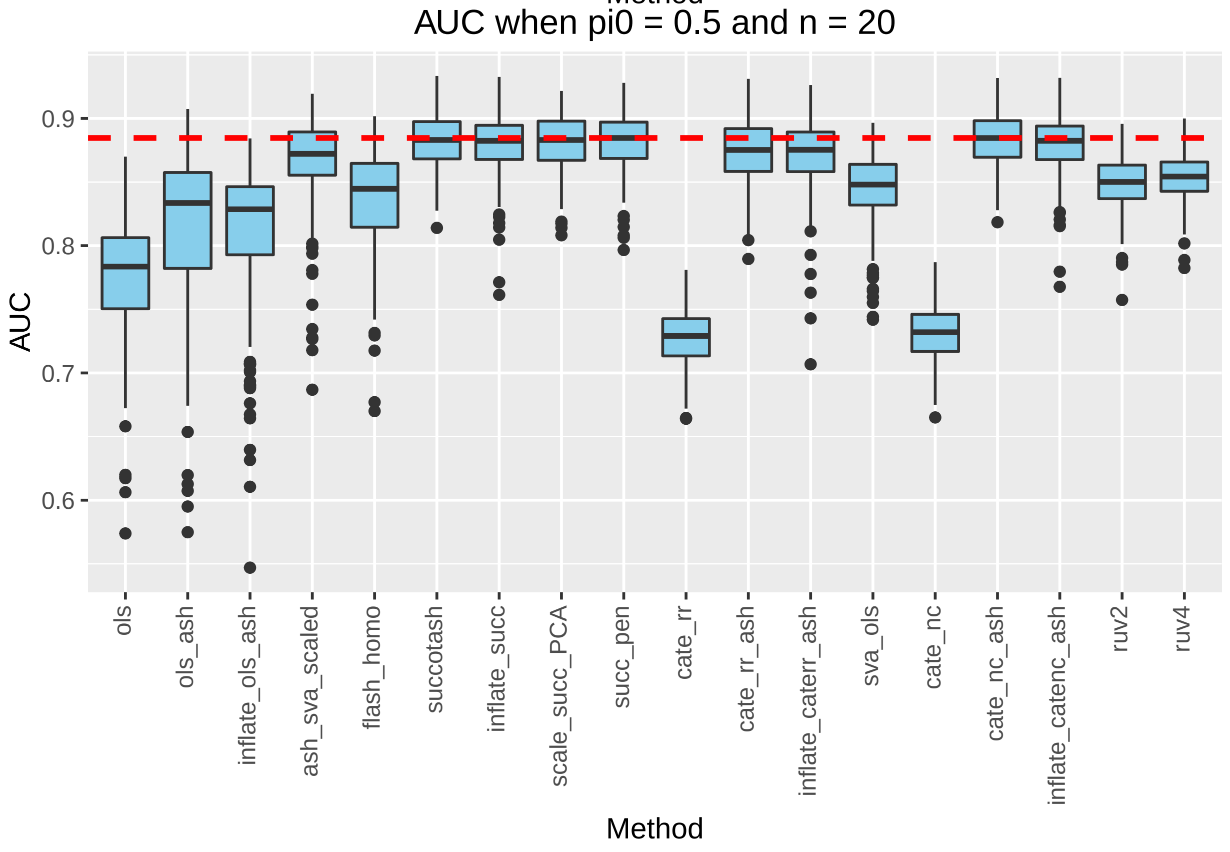

MSE when pi0 = 1 and n = 20
MSE
0.5
0.4
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

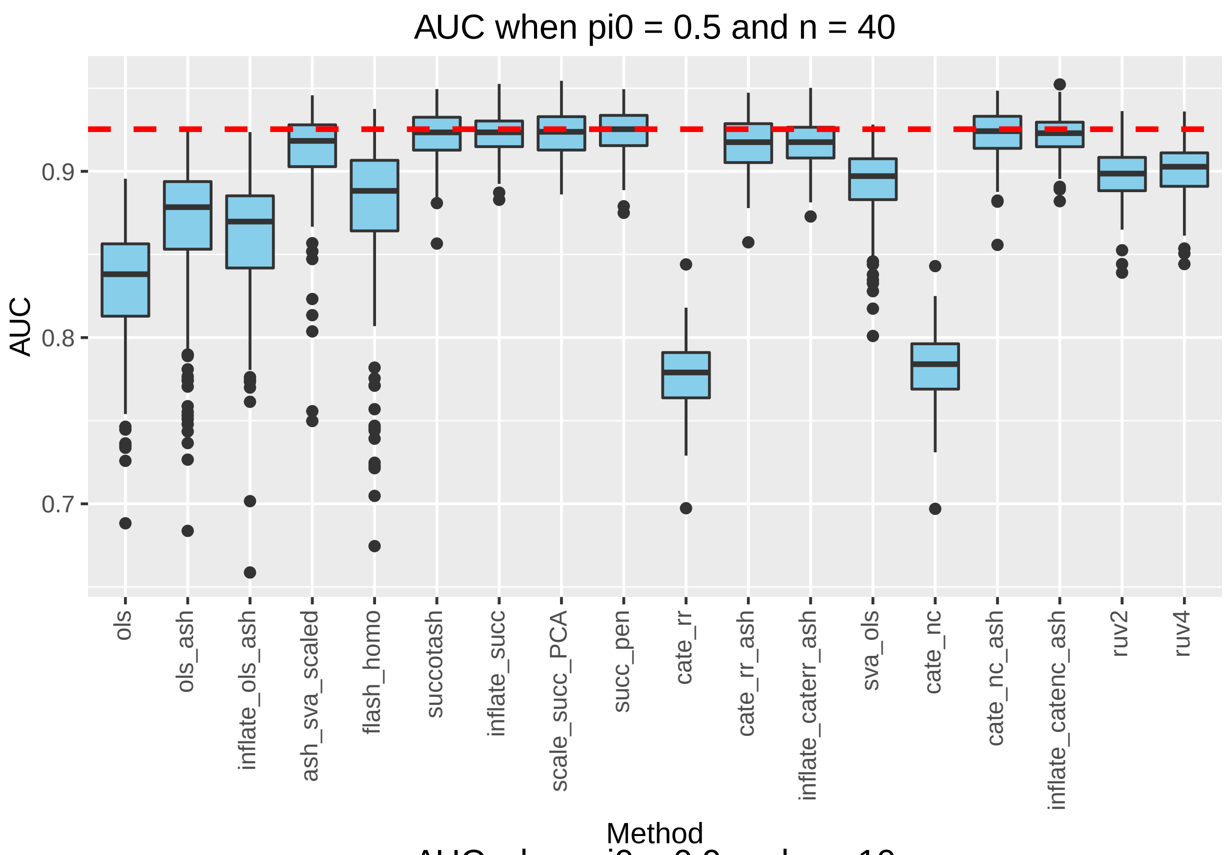

MSE when pi0 = 1 and n = 40
MSE
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

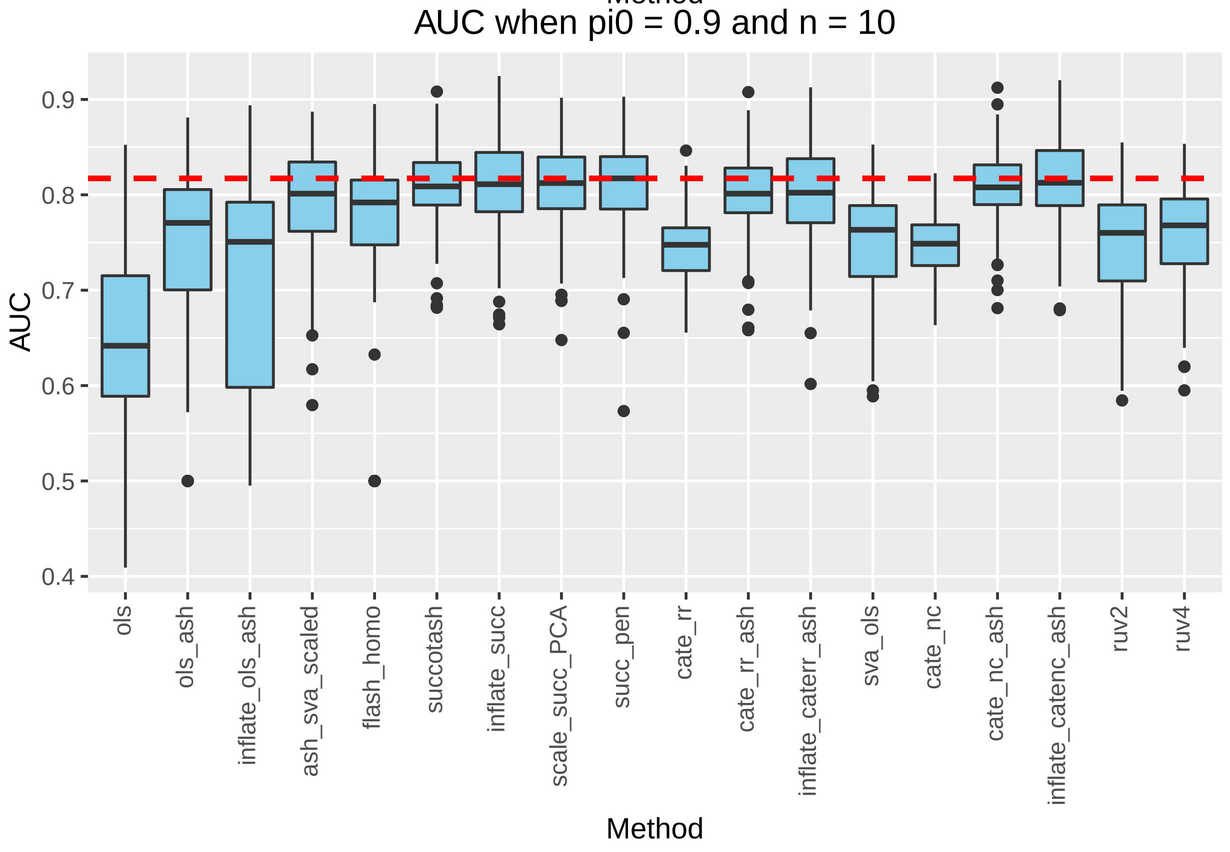
AUC Plots

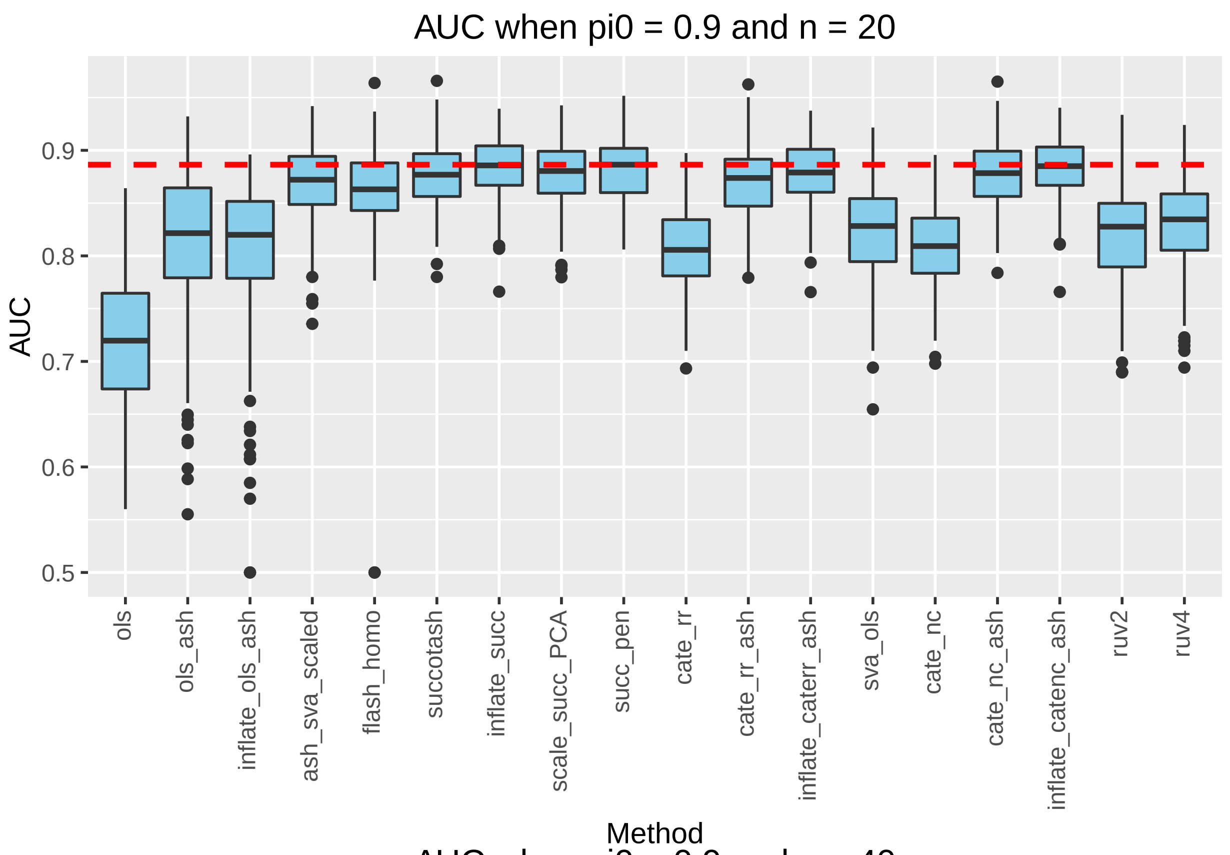
```r
double_auc <- read.csv("../double_succ/auc_mat.csv")
reg_auc <- read.csv("../flash_v_rest_using_package/auc_mat.csv")
scale_auc <- read.csv("../succ_scaled/auc_ssuc.csv")
scale_auc_pen <- read.csv("../succ_scaled_pen/auc_ssuc_mc.csv")
ash_sva <- read.csv("auc_svaash_mc.csv")
reg_auc$inflate_succ <- double_auc$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$scale_succ_PCA <- scale_auc$scale_suc1
reg_auc$succ_pen <- scale_auc_pen$post_inflate
reg_auc$ash_sva_scaled <- ash_sva$post_inflate
reg_auc <- tbl_df(reg_auc)
reg_auc <- reg_auc[, c(1:2, 17, 20, 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)
    }
}
```

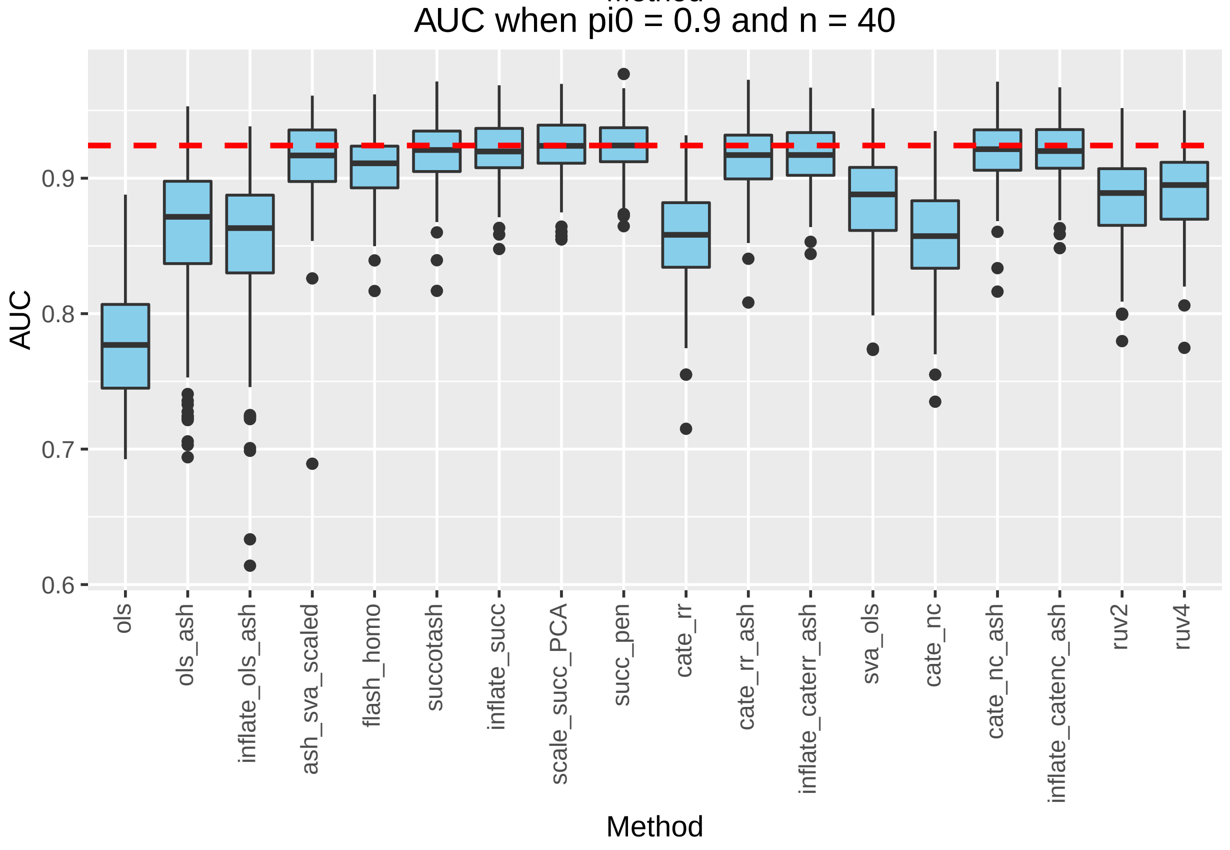

AUC when pi0 = 0.5 and n = 10
AUC
0.9
0.8
0.7
0.6
0.5
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


AUC when pi0 = 0.5 and n = 20
AUC
0.9
0.8
0.7
0.6
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
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
0.9
0.8
0.7
Method
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
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
0.9
0.8
0.7
0.6
0.5
0.4
Method
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
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 = 20
AUC
0.5
0.6
0.7
0.8
0.9
Method
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
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 = 40
AUC
0.6
0.7
0.8
0.9
Method
ols
ols_ash
inflate_ols_ash
ash_sva_scaled
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4

```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    colorspace_1.2-6
##  [5] R6_2.1.1         stringr_1.0.0    plyr_1.8.3       tools_3.2.5
##  [9] parallel_3.2.5   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  assertthat_0.1
## [17] digest_0.6.9     formatR_1.3      codetools_0.2-14 evaluate_0.8.3
## [21] rmarkdown_0.9.6  labeling_0.3     stringi_1.0-1    compiler_3.2.5
## [25] scales_0.4.0
```