Estimate Scaling Parameter in SUCCOTASH, then Inflate

David Gerard

2016-05-03

Abstract

I obtain the best results I've seen from any method by using a two-step procedure with SUCCOTASH:

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.

"nopen_then_inflate" below describes the following two-step procedure.

1. Estimate the variance scaling parameter λ using SUCCOTASH. Call this estimate $\hat{\lambda}$
2. Re-run SUCCOTASH, but with variance $\tilde{\lambda}\hat{\Sigma}$, where

$$\tilde{\lambda} = \frac{n}{n-k-q}\hat{\lambda},$$

   where n is the sample size, k is the number of unknown confounders (estimated with `num.sv`), and q is the number of covariates.

The idea here is that this is the same multiplicative correction you would apply to the MLE variances in a standard normal problem.

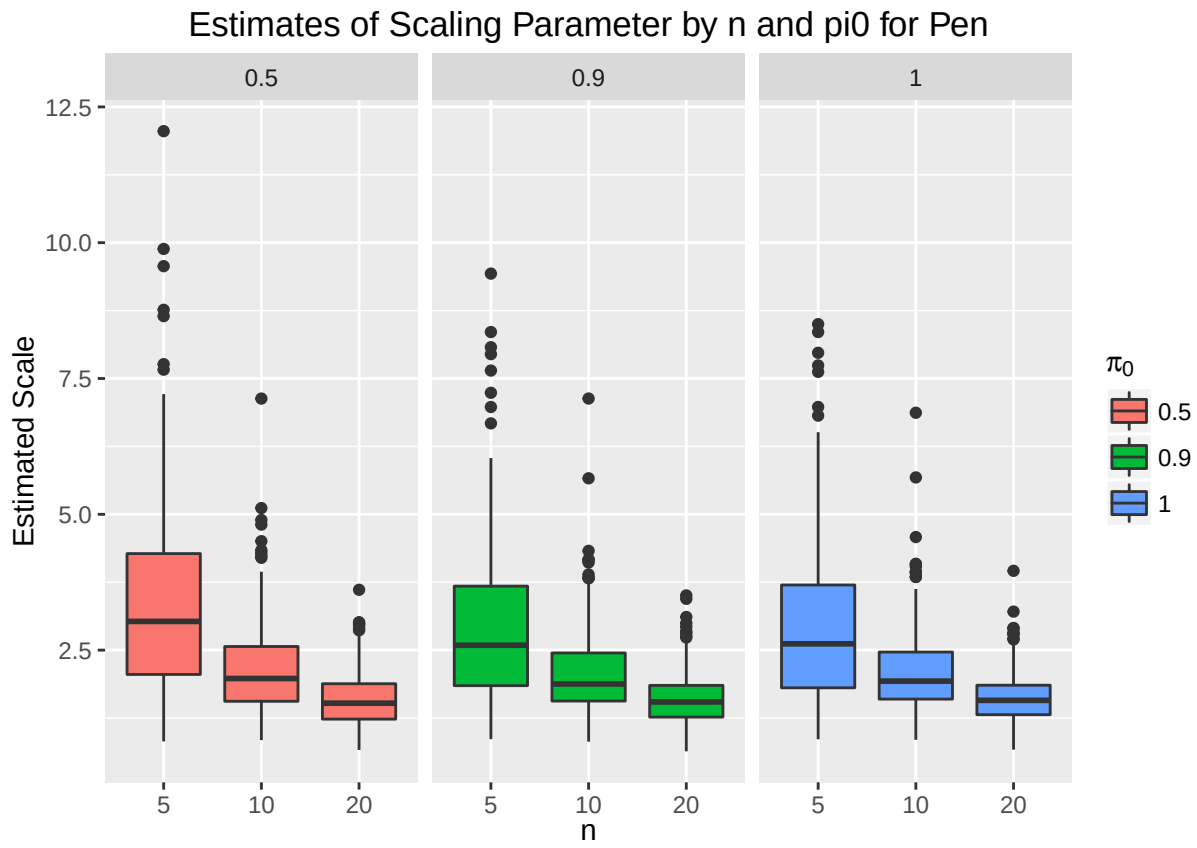
I see superior performance in estimating π_0 over every other method I have tried, including the ad-hoc inflation methods that I chanced upon by accident. It also has among the top AUC's and the MSE is competative and among the best performers.

This table has the mean inflation factor for the current method and for the first step above.

```
scale_est_mat <- tbl_df(read.csv("scale_est_ssuc_mc.csv", header = TRUE))
scale_est_mat_nopen <- tbl_df(read.csv("../succ_scaled/scale_est_ssuc.csv", header = TRUE))
temp_tab <- cbind(aggregate(succ_superpen ~ nullpi + nsamp, FUN = mean, data = scale_est_mat),
                  aggregate(scale_suc1 ~ nullpi + nsamp, FUN = mean, data = scale_est_mat_nopen)[,3])
kable(temp_tab, col.names = c("$\\pi_0$", "$n/2$", "nopen_then_inflate", "Raw Mean Scale Est"), digits =
```

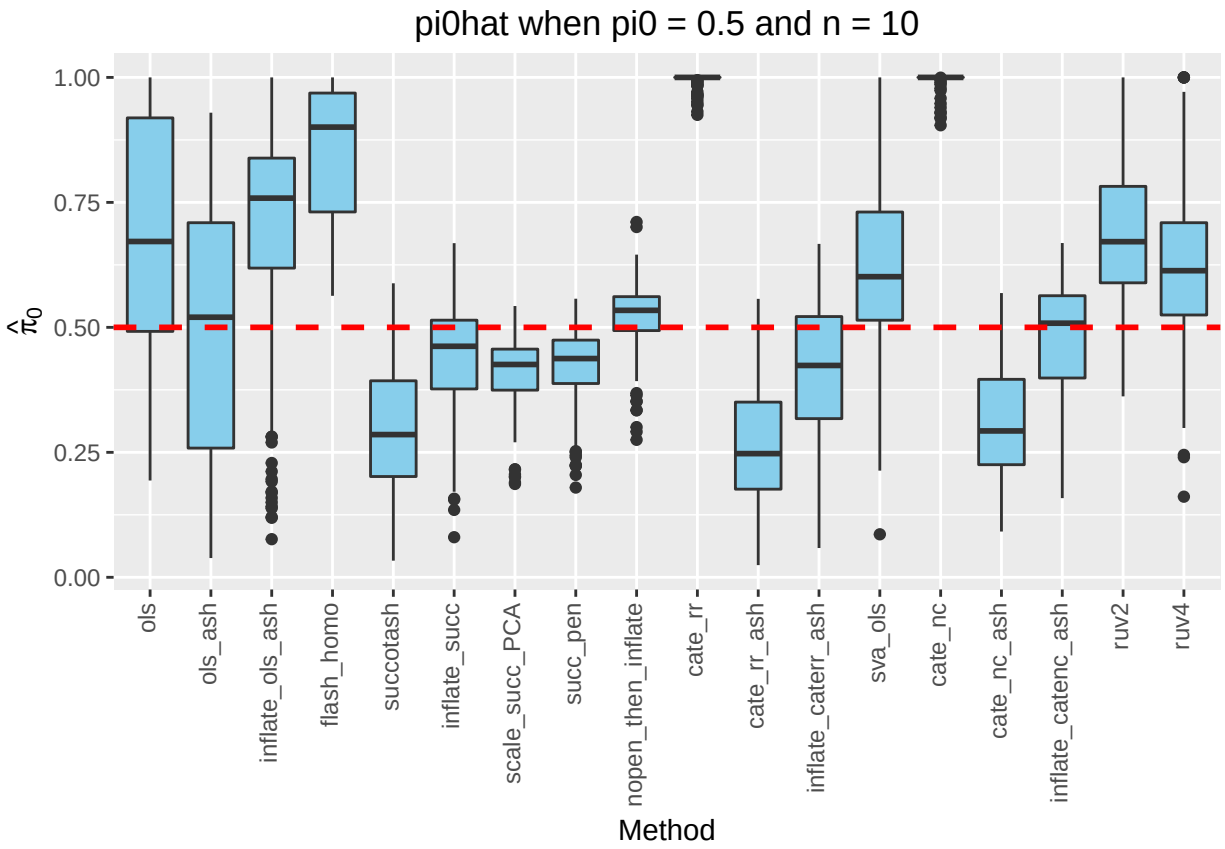
π_0	$n/2$	nopen_then_inflate	Raw Mean Scale Est
0.5	5	3.40	1.90
0.9	5	3.00	1.83
1.0	5	2.99	1.75
0.5	10	2.18	1.68
0.9	10	2.13	1.61
1.0	10	2.14	1.60
0.5	20	1.62	1.34
0.9	20	1.62	1.29
1.0	20	1.64	1.37

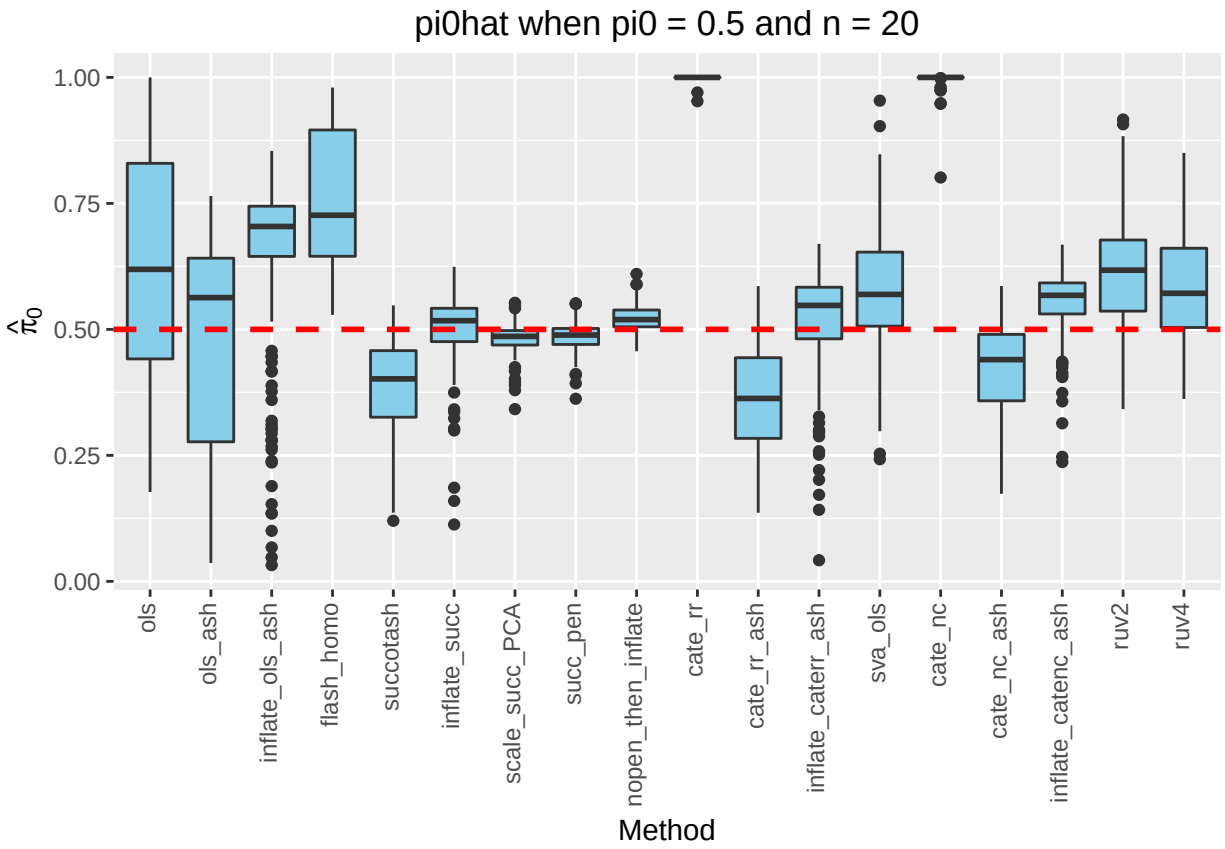
```
ggplot(data = scale_est_mat, mapping = aes(x = factor(nsamp), y = succ_superpen,
                                           fill = factor(nullpi))) +
    facet_grid(.~nullpi) +
    geom_boxplot() +
    xlab(expression(n)) + ylab("Estimated Scale") +
    scale_fill_discrete(name=expression(pi[0]))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

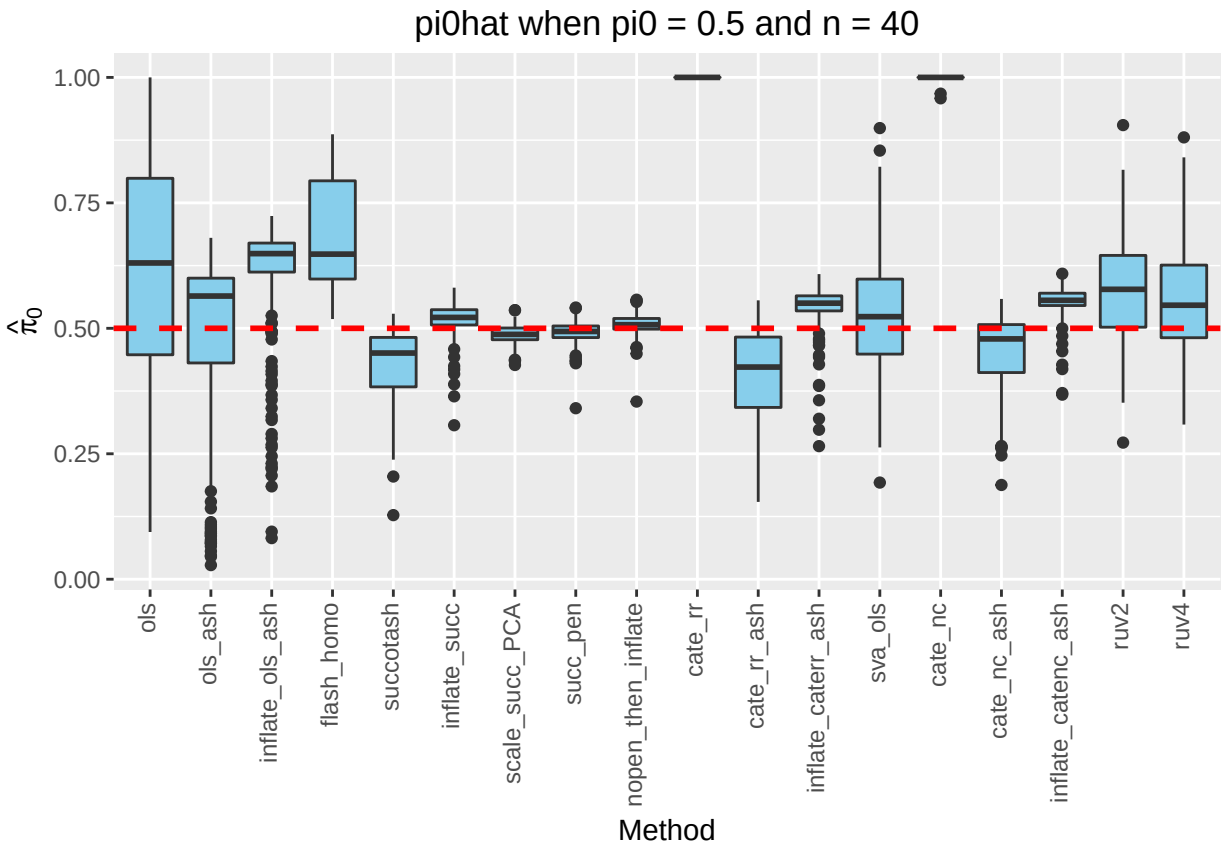


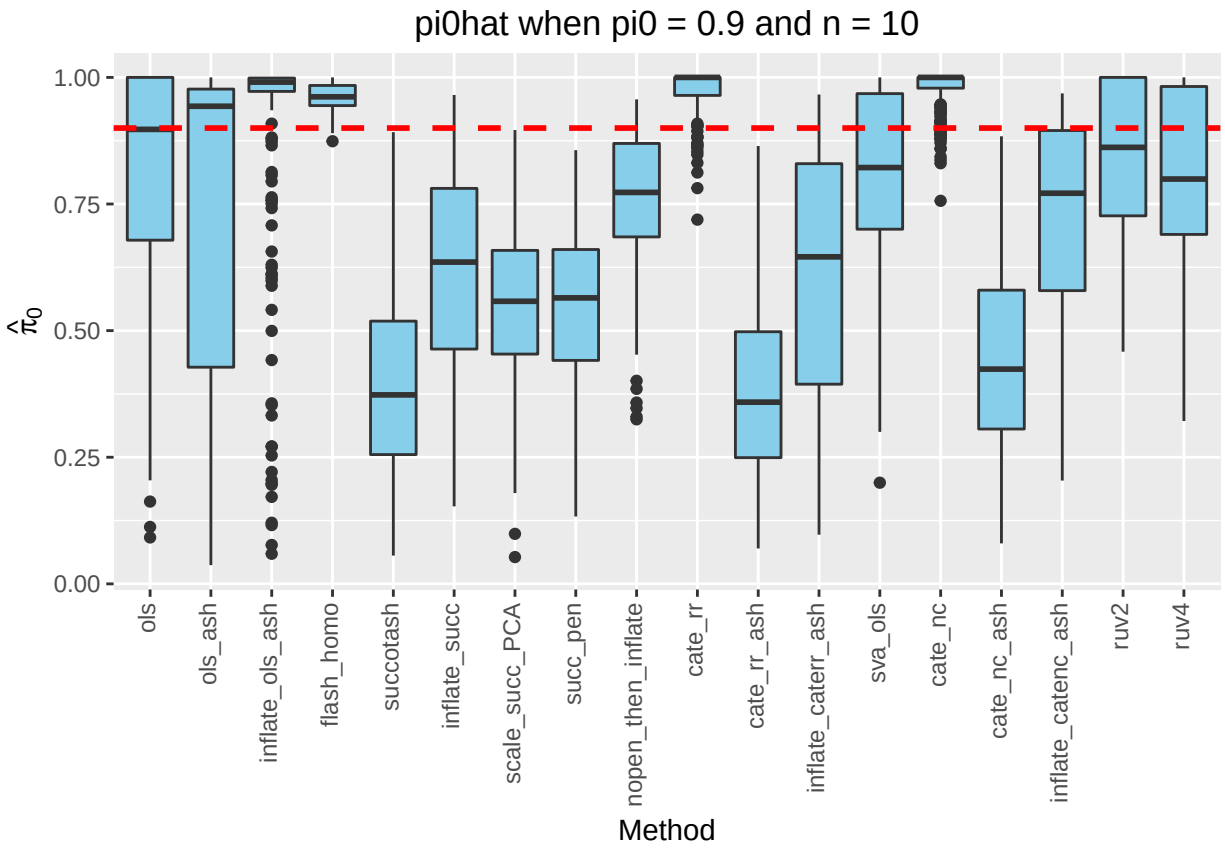
$\hat{\pi}_0$ Plots

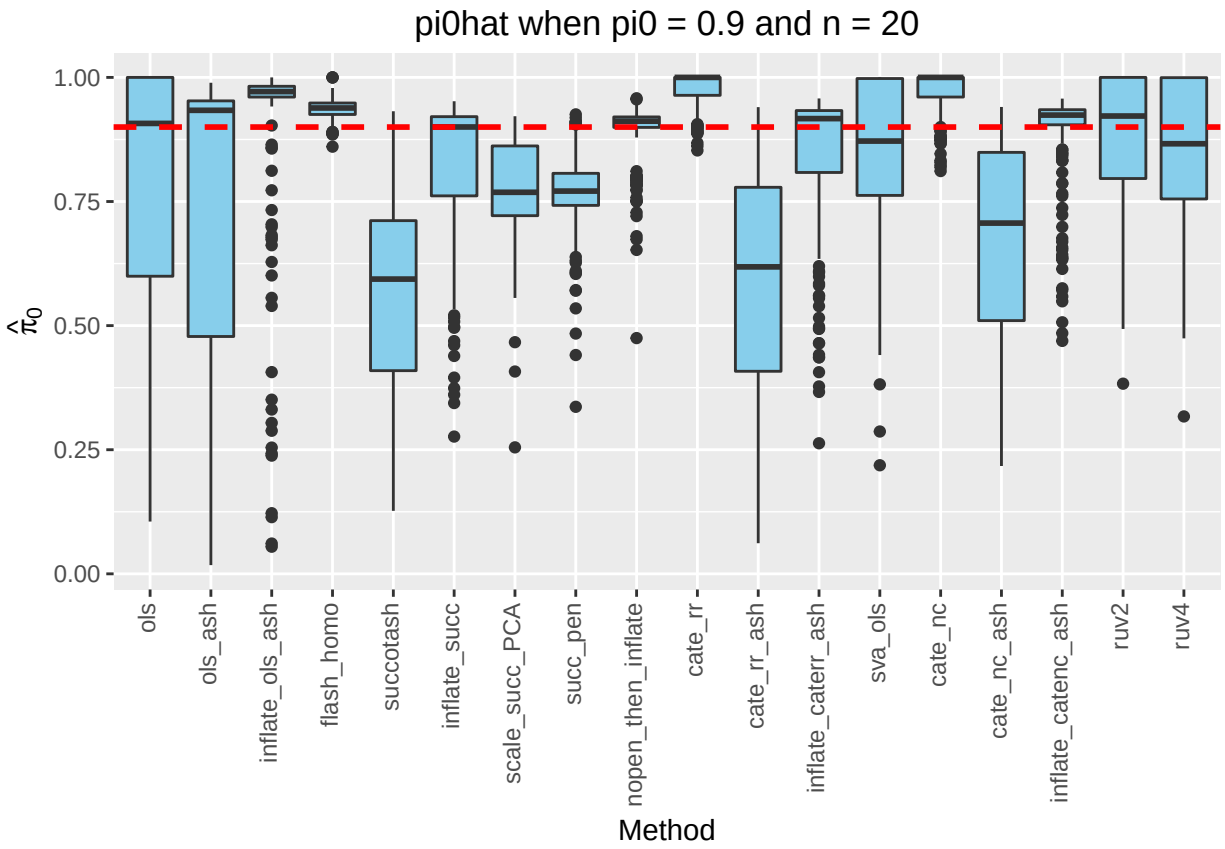
```
double_pi0         <- read.csv("../double_succ/pi0_mat.csv")
reg_pi0            <- read.csv("../flash_v_rest_using_package/pi0_mat.csv")
scale_pi0          <- read.csv("../succ_scaled/pi0_ssuc.csv")
scale_pi0_pen      <- read.csv("../succ_scaled_pen/pi0_ssuc_mc.csv")
nopen_then_inflate <- read.csv("pi0_ssuc_mc.csv")
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$nopen_then_inflate <- nopen_then_inflate$succ_superpen
reg_pi0 <- tbl_df(reg_pi0)
reg_pi0 <- reg_pi0[, c(1:2, 17, 3:4, 14, 18:20, 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)
    }
}
```

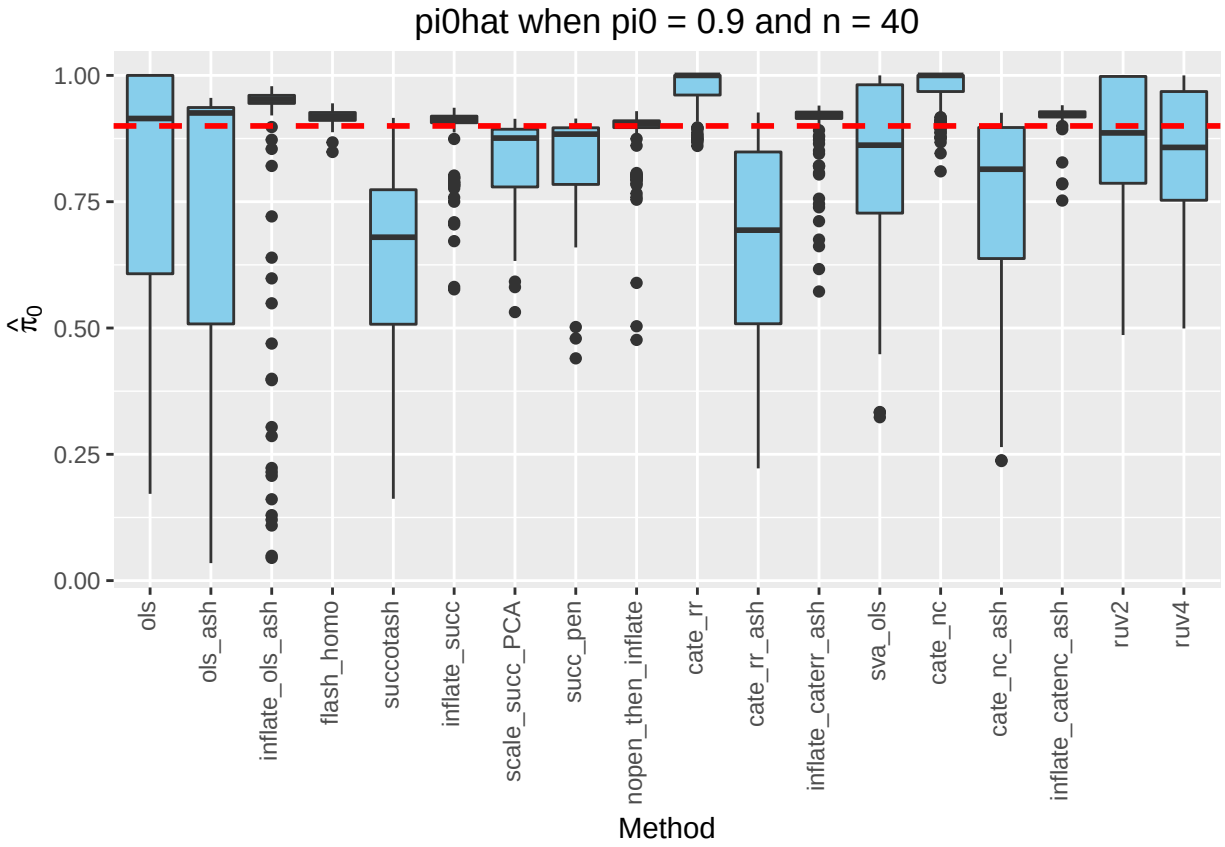

pi0hat when pi0 = 0.5 and n = 10
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

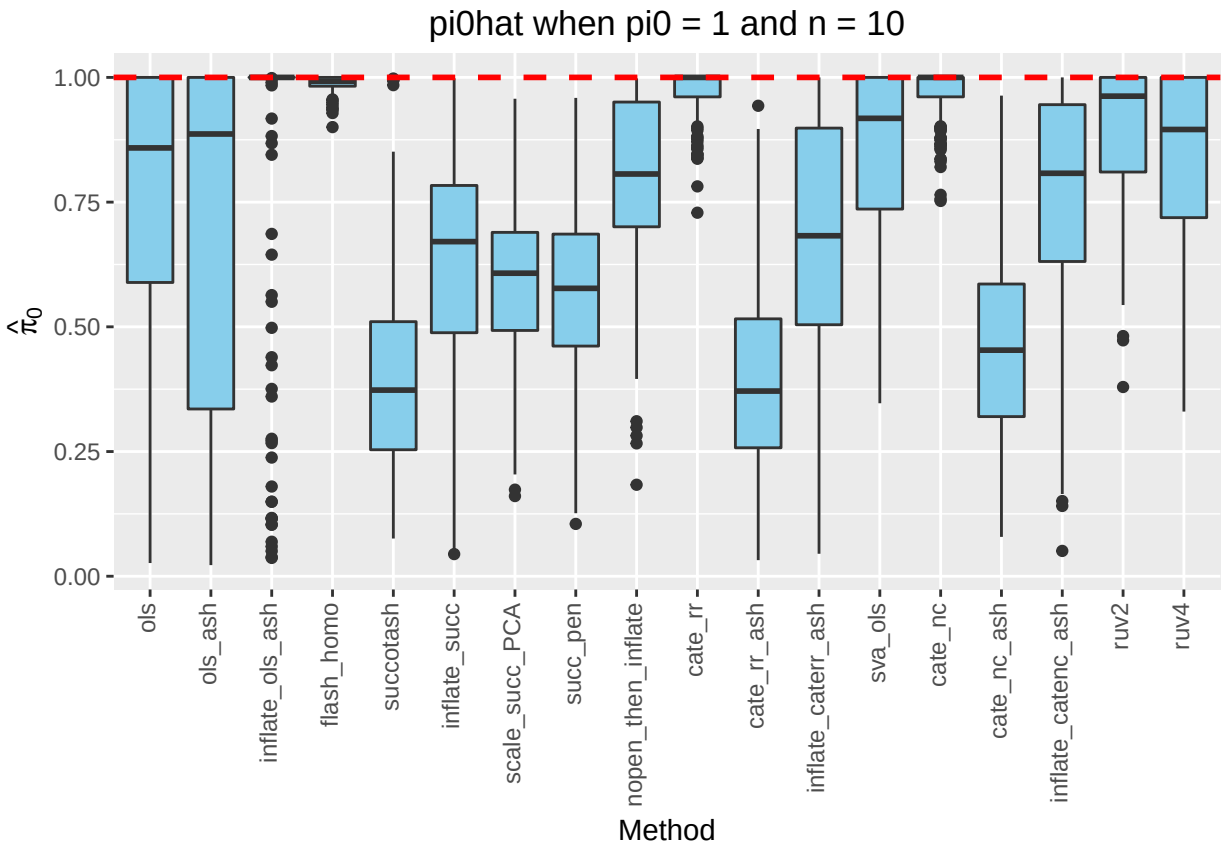

pi0hat when pi0 = 0.5 and n = 20
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

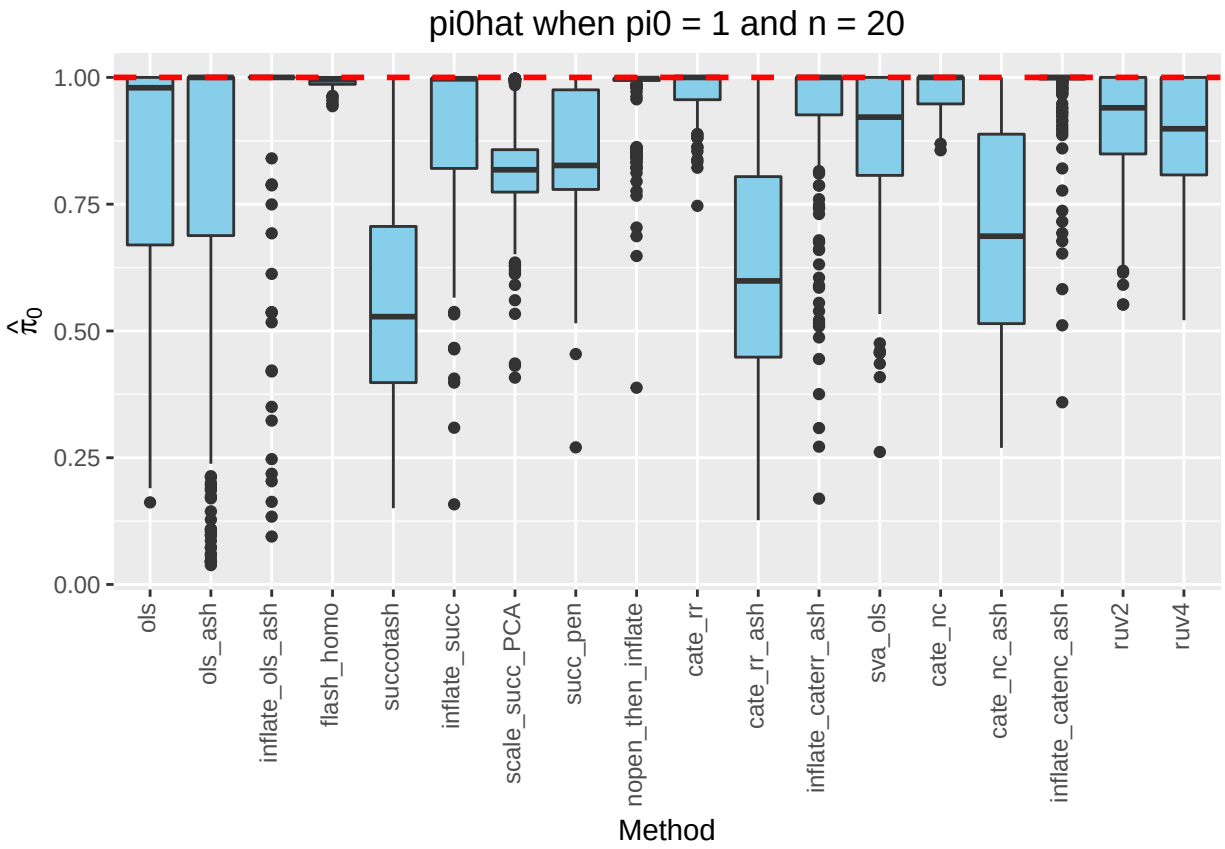

pi0hat when pi0 = 0.5 and n = 40
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

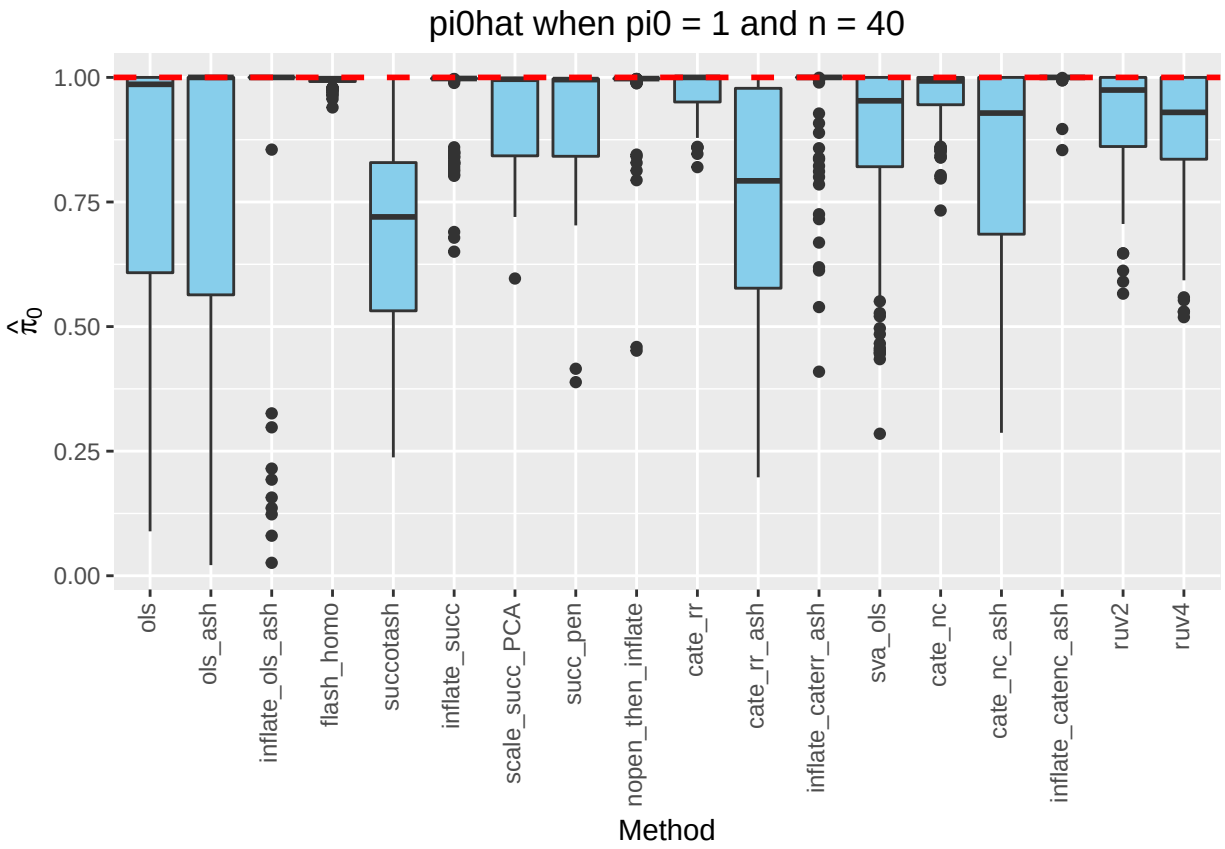

pi0hat when pi0 = 0.9 and n = 10
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


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

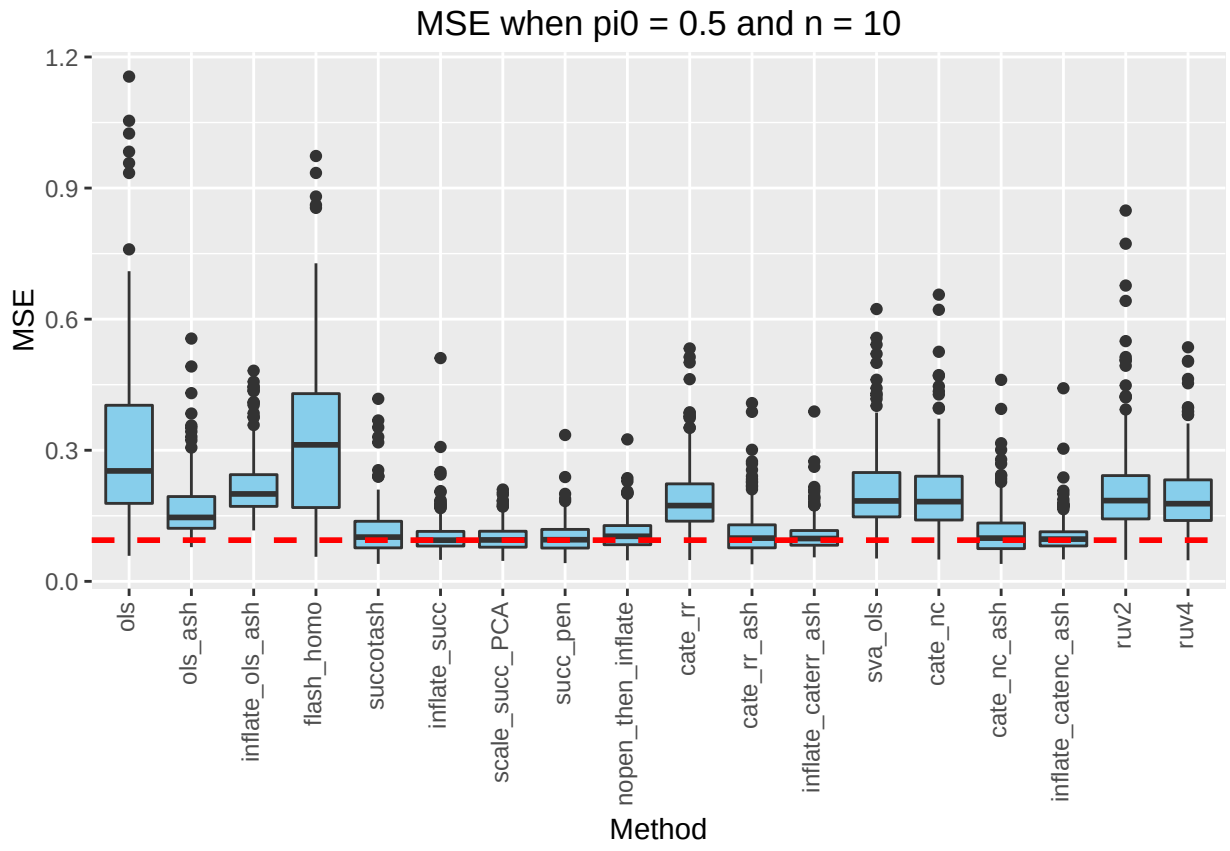

pi0hat when pi0 = 1 and n = 20
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method



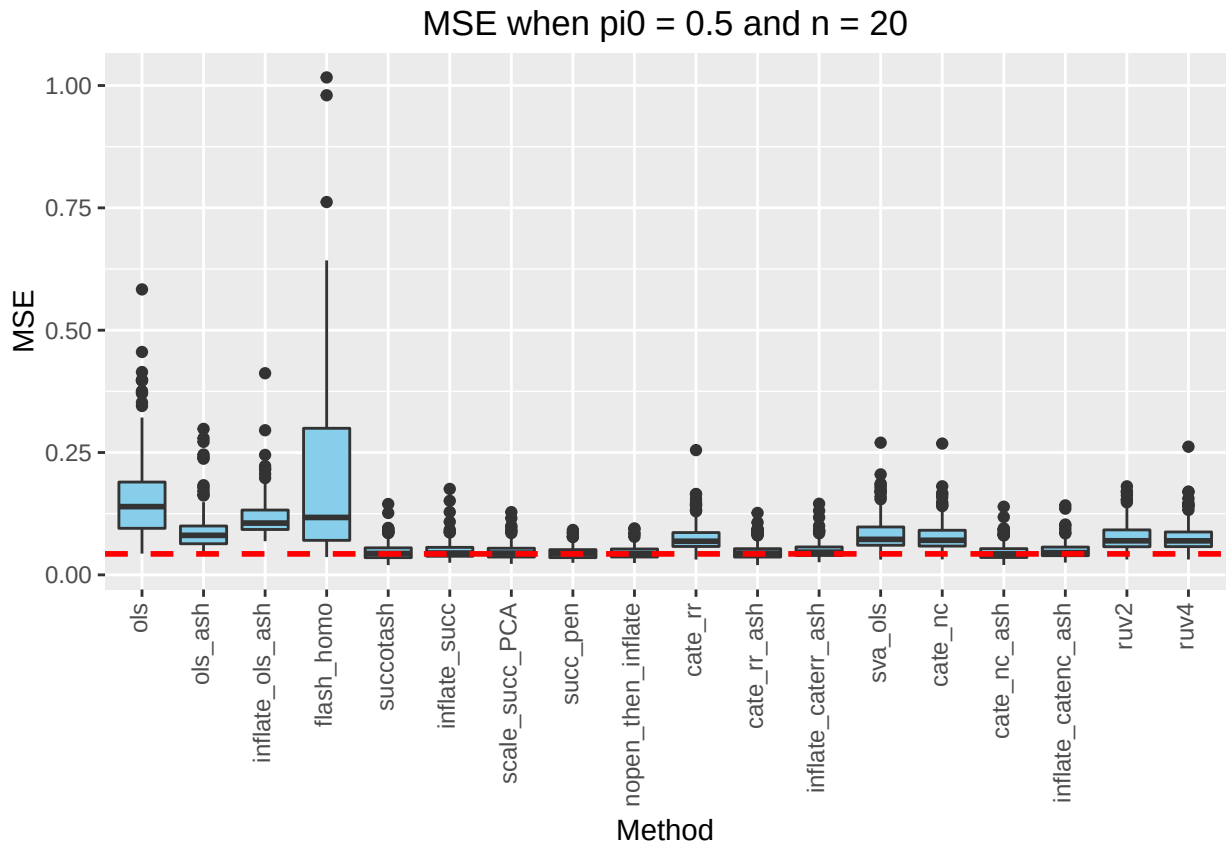
MSE Plots

```
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("../succ_scaled/mse_ssuc.csv")
scale_mse_pen      <- read.csv("../succ_scaled_pen/mse_ssuc_mc.csv")
nopen_then_inflate <- read.csv("mse_ssuc_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$nopen_then_inflate <- nopen_then_inflate$succ_superpen
reg_mse <- tbl_df(reg_mse)
reg_mse <- reg_mse[, c(1:2, 17, 3:4, 14, 18:20, 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(
```

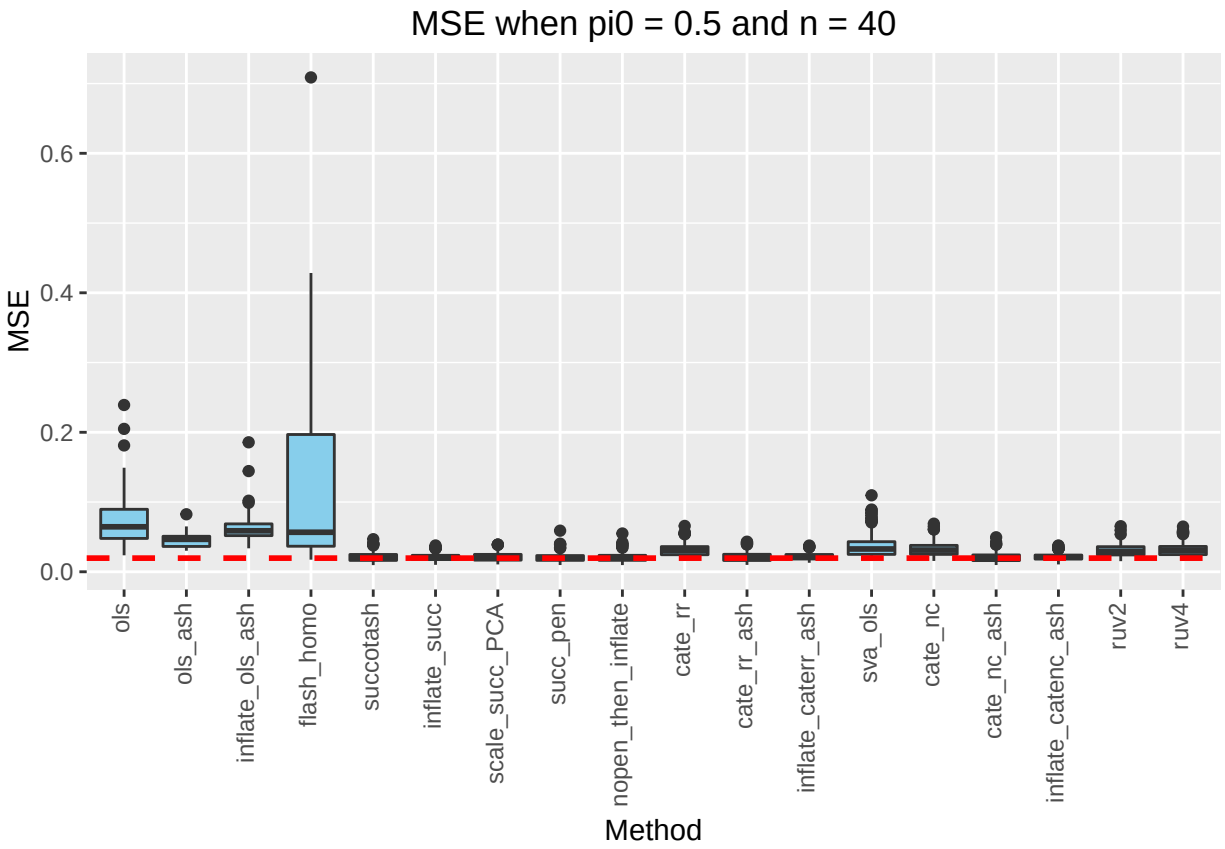
```
                reg_mse, nullpi == current_pi & nsamp == current_nsamp),
            -c(nsamp, nullpi)
        )

        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)
    }
}
```

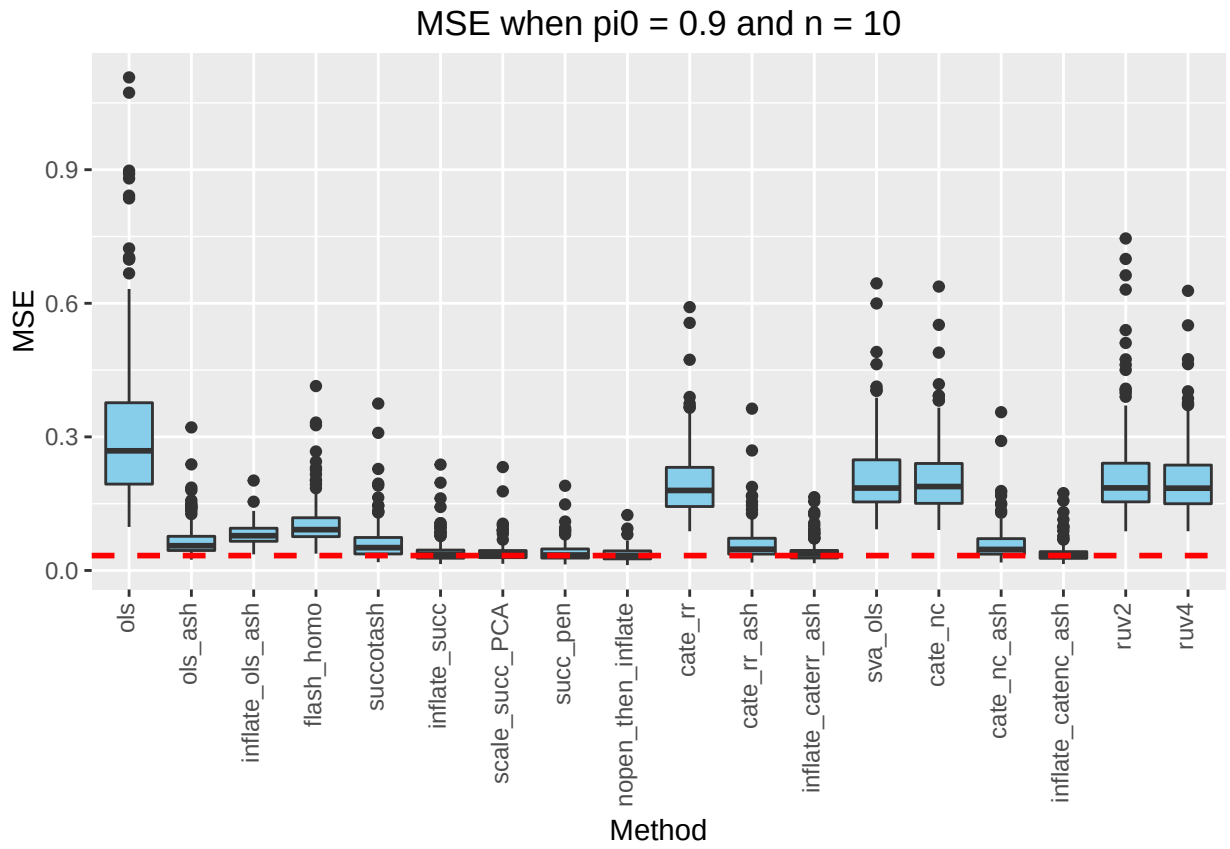



```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

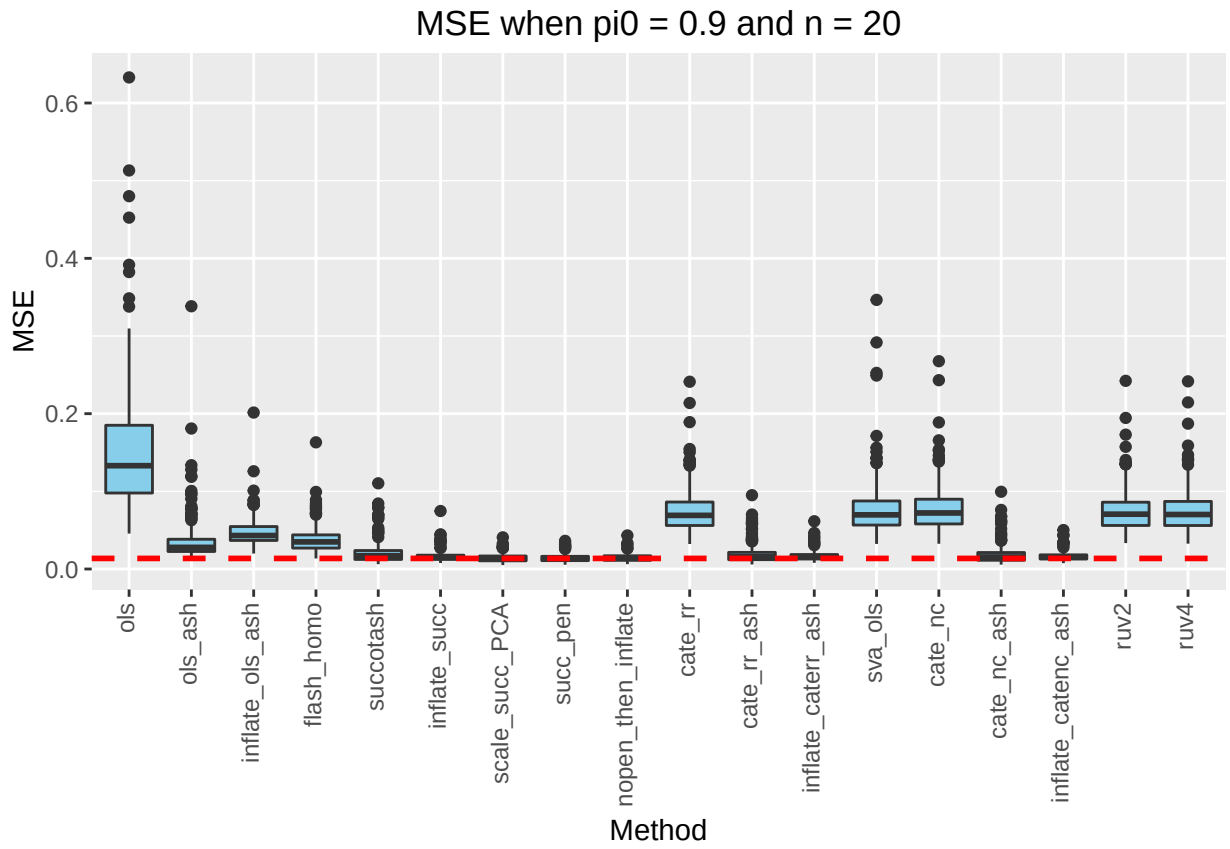



```
## Warning: Removed 203 rows containing non-finite values (stat_boxplot).
```

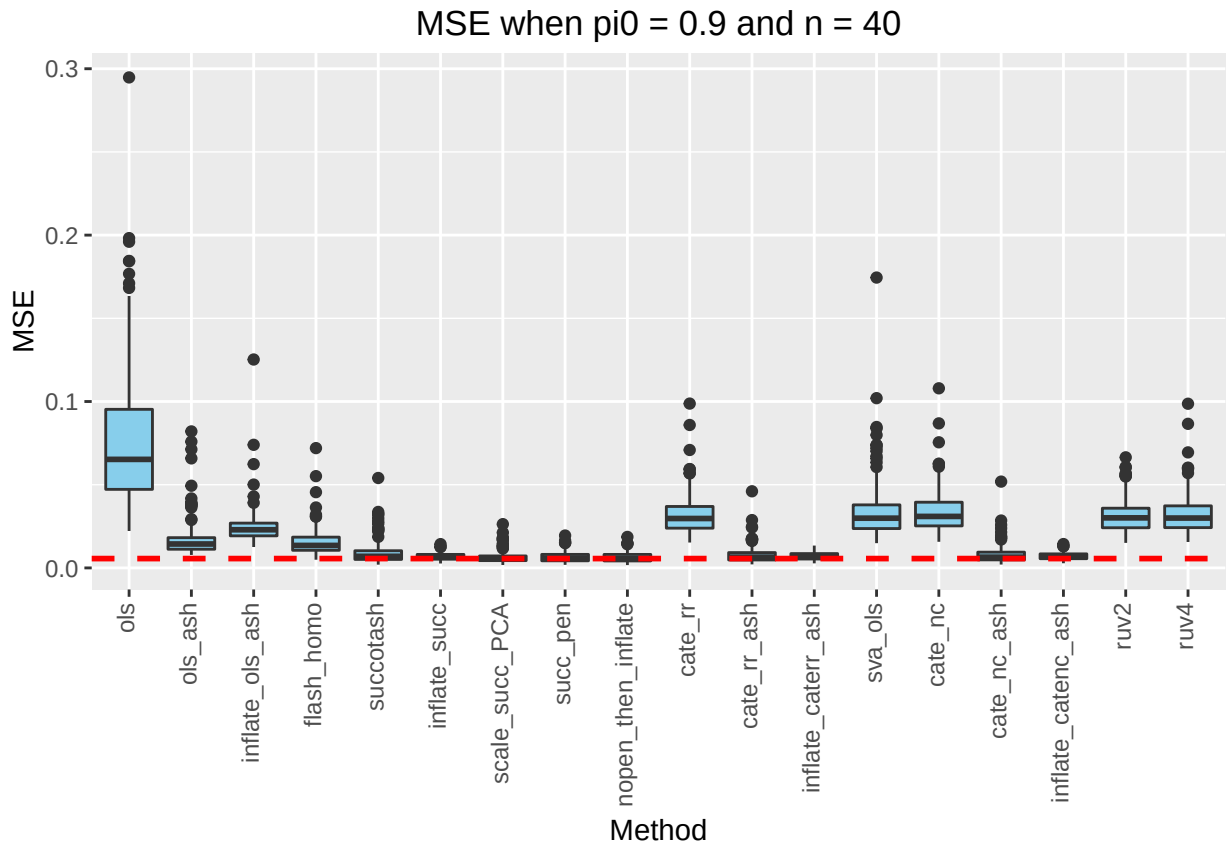

MSE when pi0 = 0.5 and n = 40
MSE
0.0
0.2
0.4
0.6
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

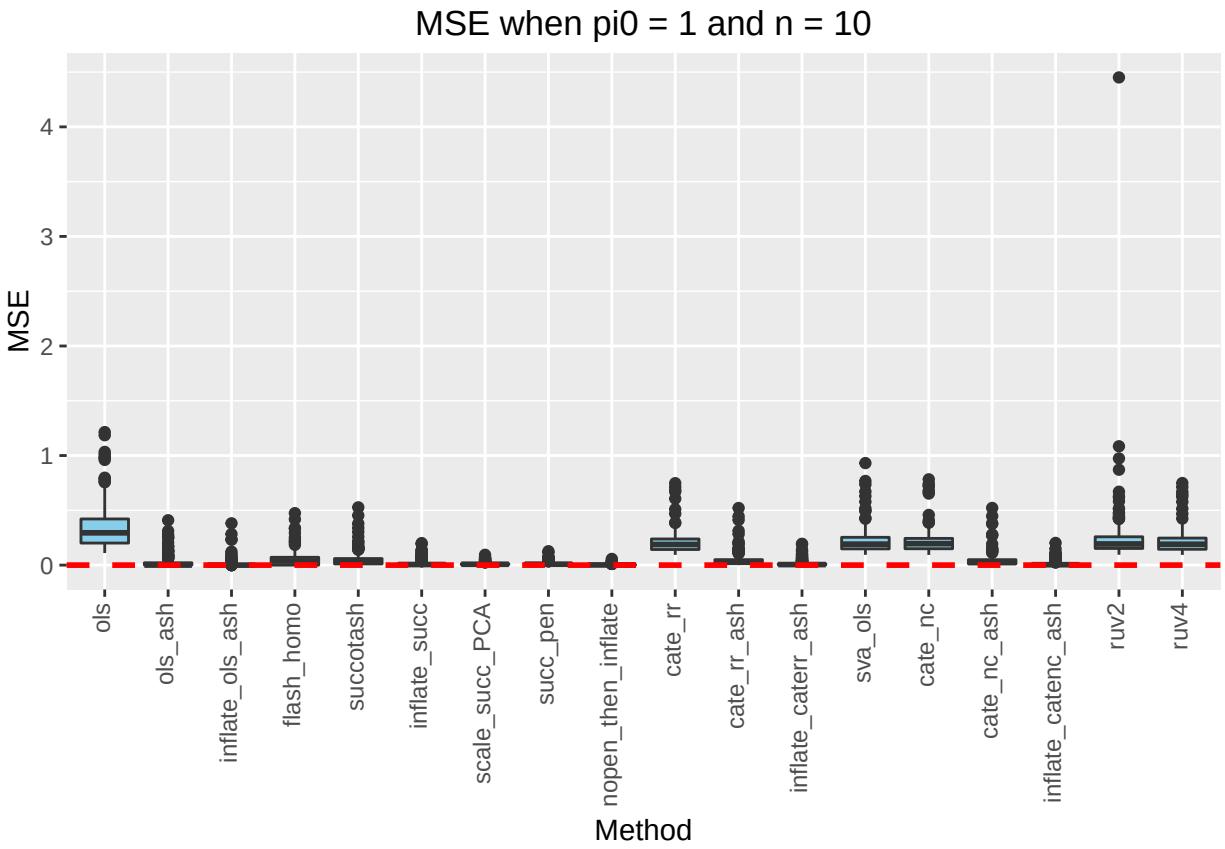


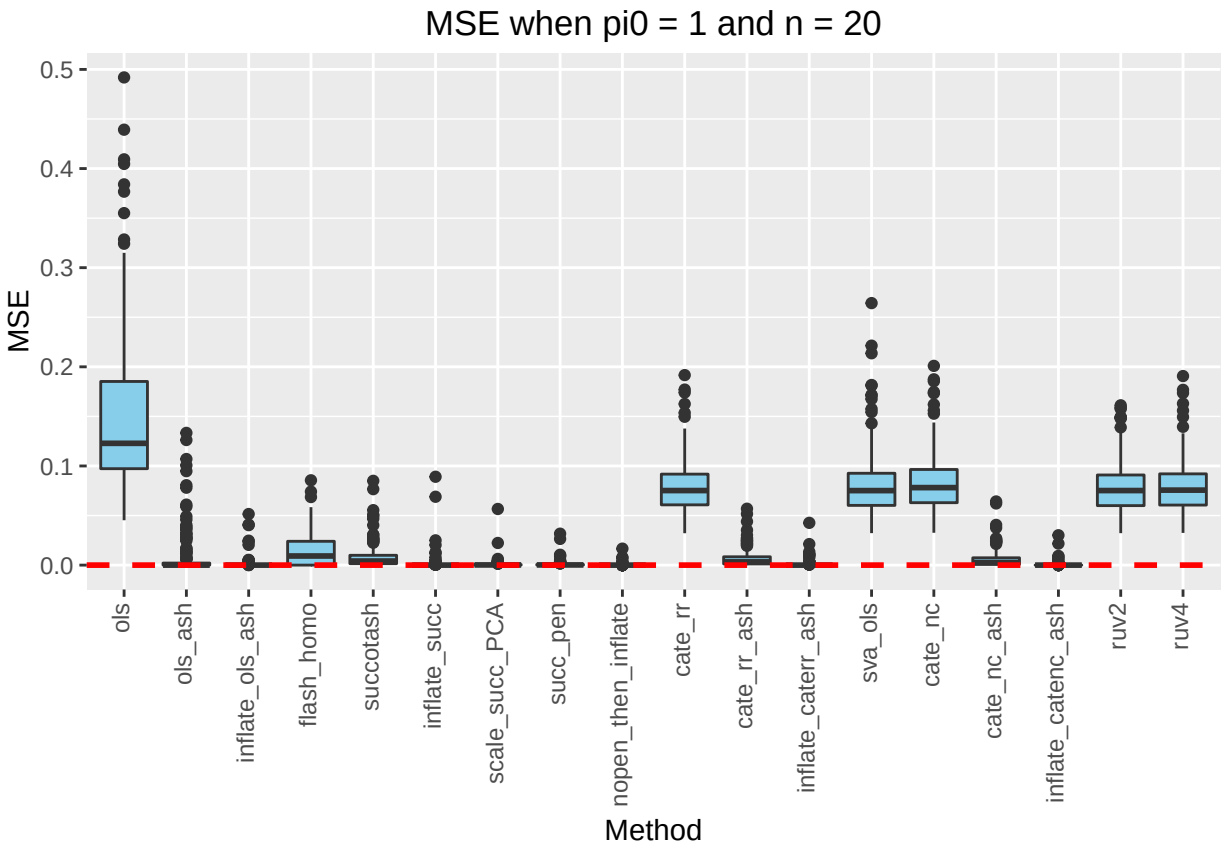

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

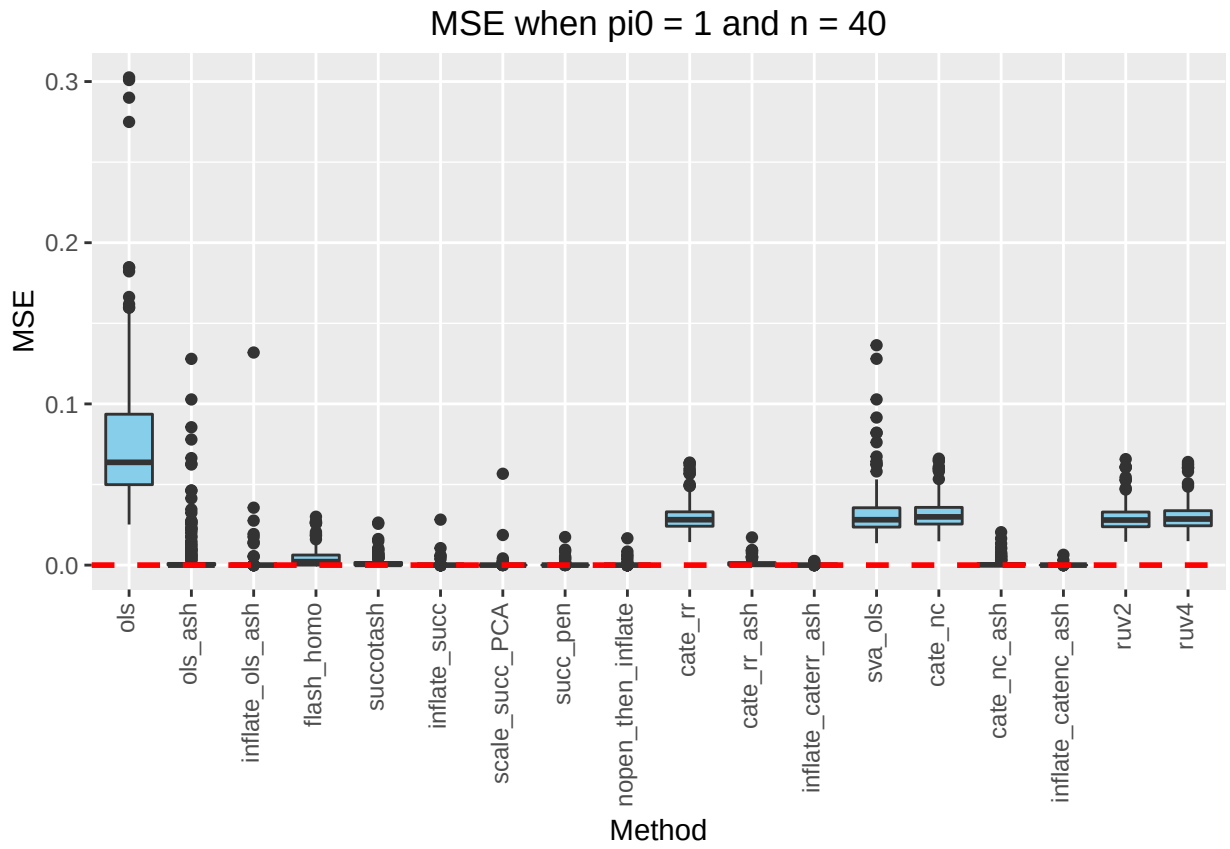



```
## Warning: Removed 89 rows containing non-finite values (stat_boxplot).
```


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_succ_PCA
succ_pen
nopen_then_inflate
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
0
1
2
3
4
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

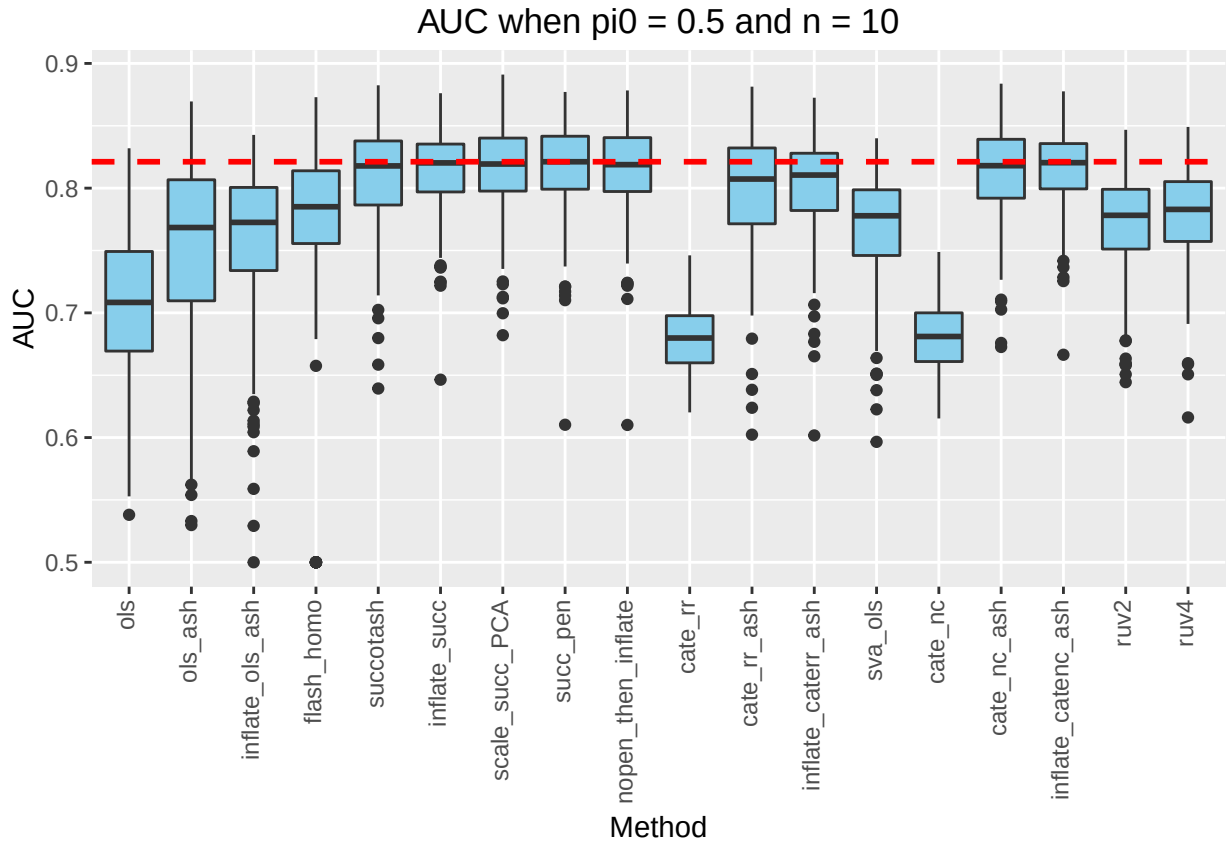

MSE when pi0 = 1 and n = 20
MSE
0.5
0.4
0.3
0.2
0.1
0.0
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method

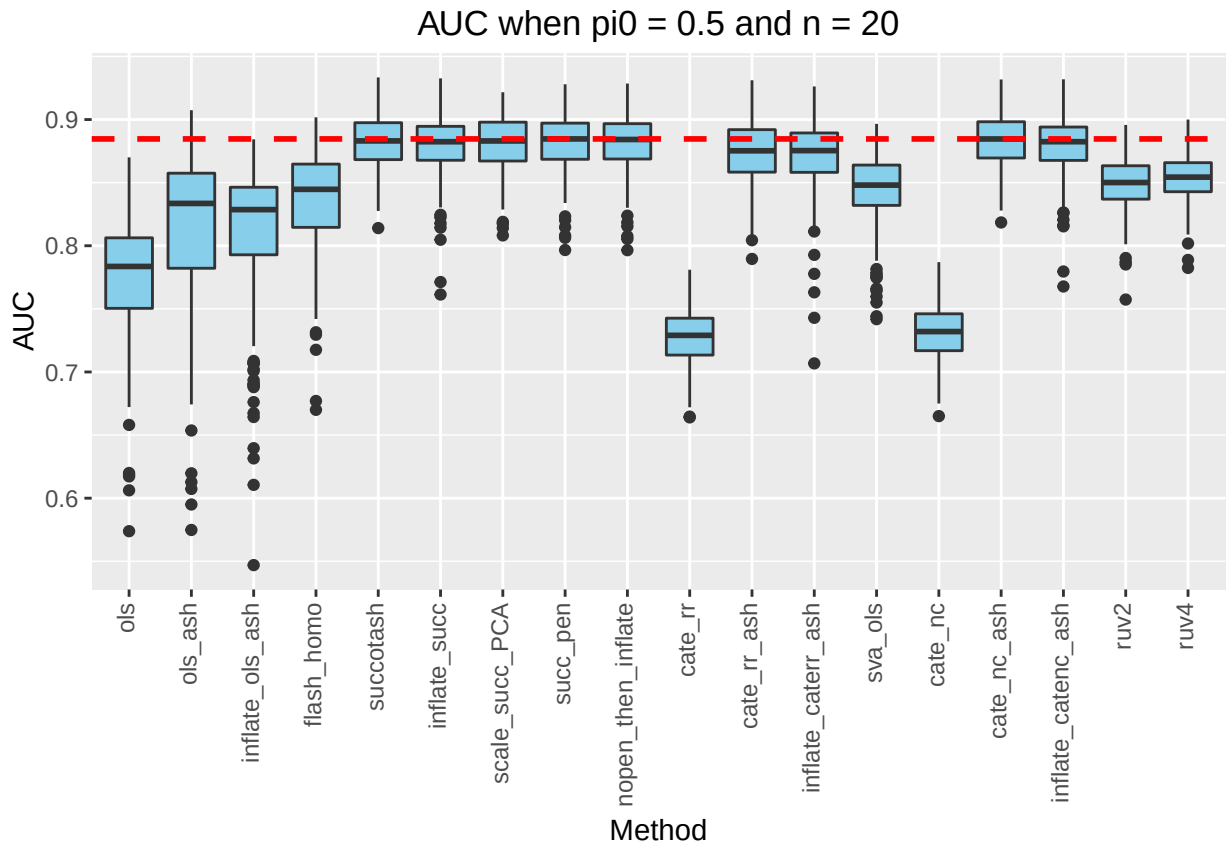


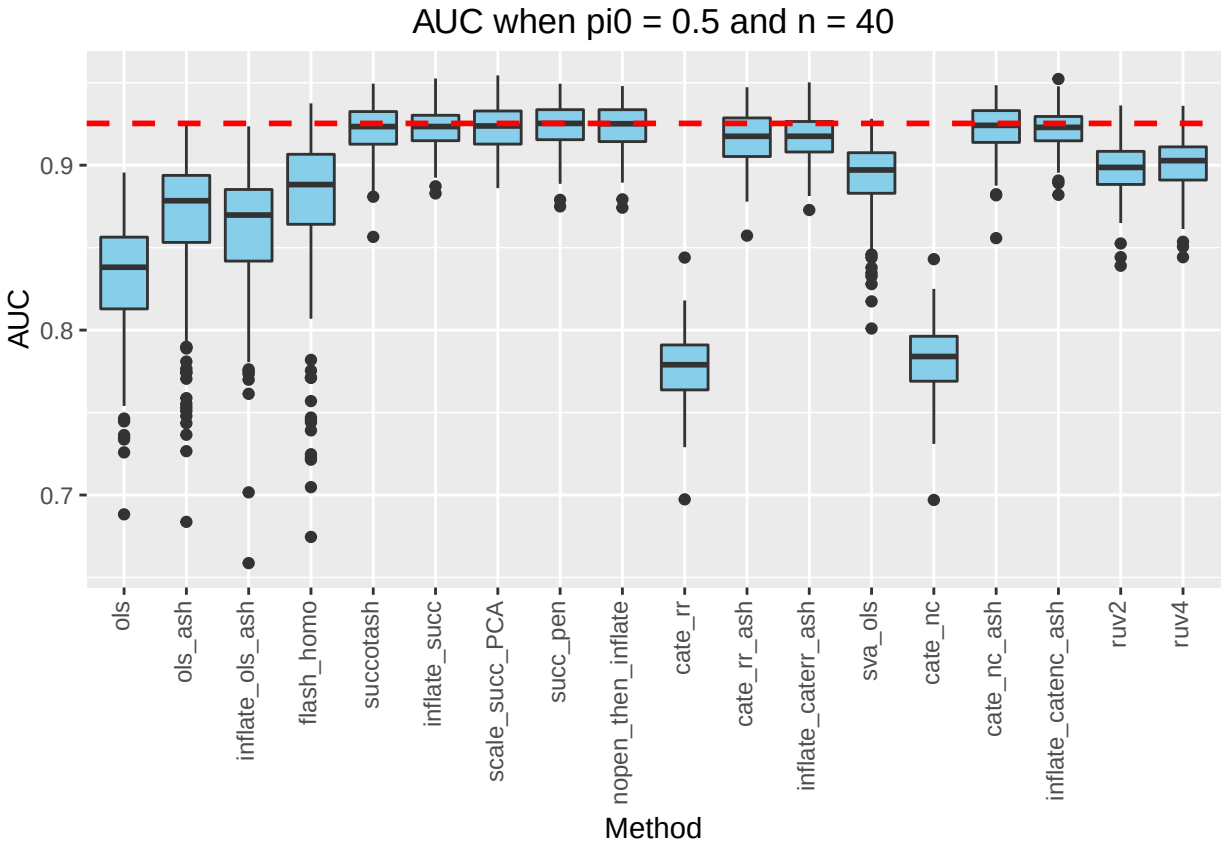
AUC Plots

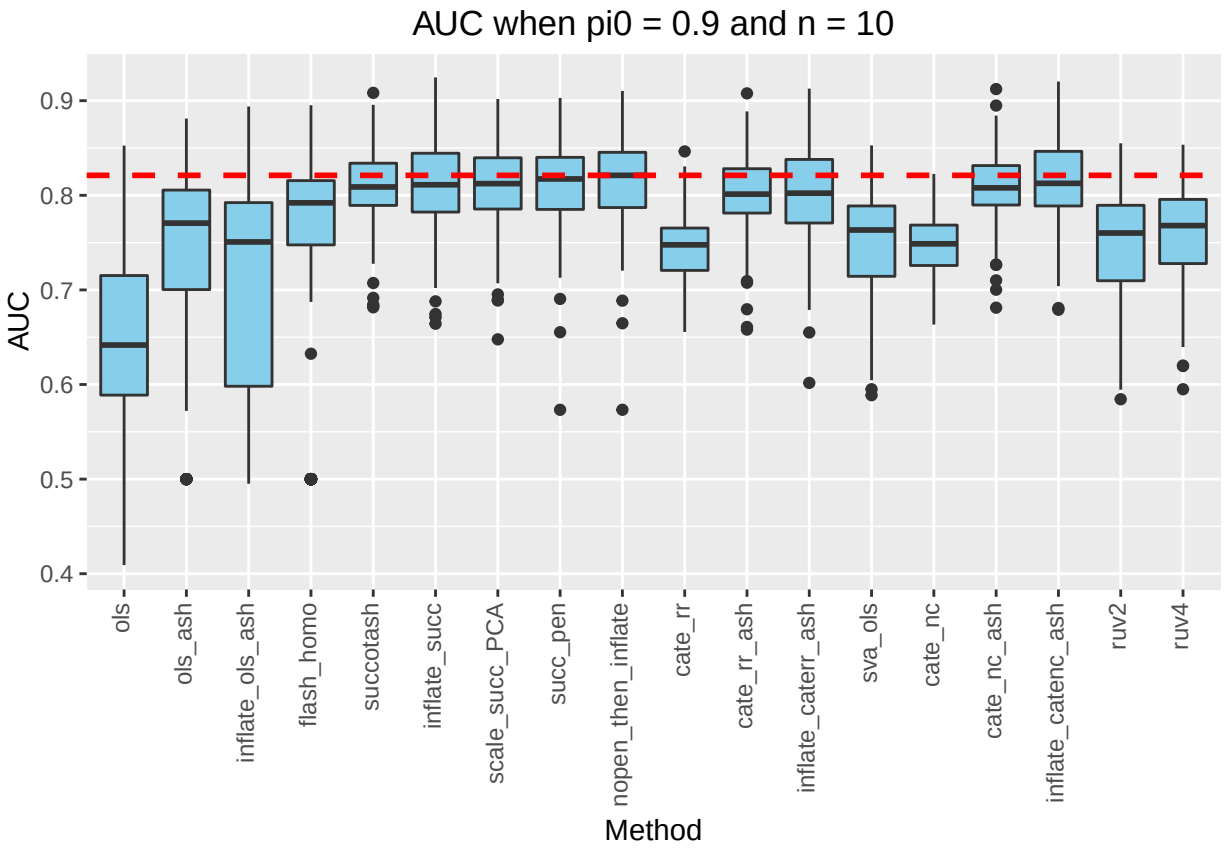
```
double_auc         <- read.csv("../double_succ/auc_mat.csv")
reg_auc            <- read.csv("../flash_v_rest_using_package/auc_mat.csv")
scale_auc          <- read.csv("../succ_scaled/auc_ssuc.csv")
scale_auc_pen      <- read.csv("../succ_scaled_pen/auc_ssuc_mc.csv")
nopen_then_inflate <- read.csv("auc_ssuc_mc.csv")
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$nopen_then_inflate <- nopen_then_inflate$succ_superpen
reg_auc <- tbl_df(reg_auc)
reg_auc <- reg_auc[, c(1:2, 17, 3:4, 14, 18:20, 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(
```

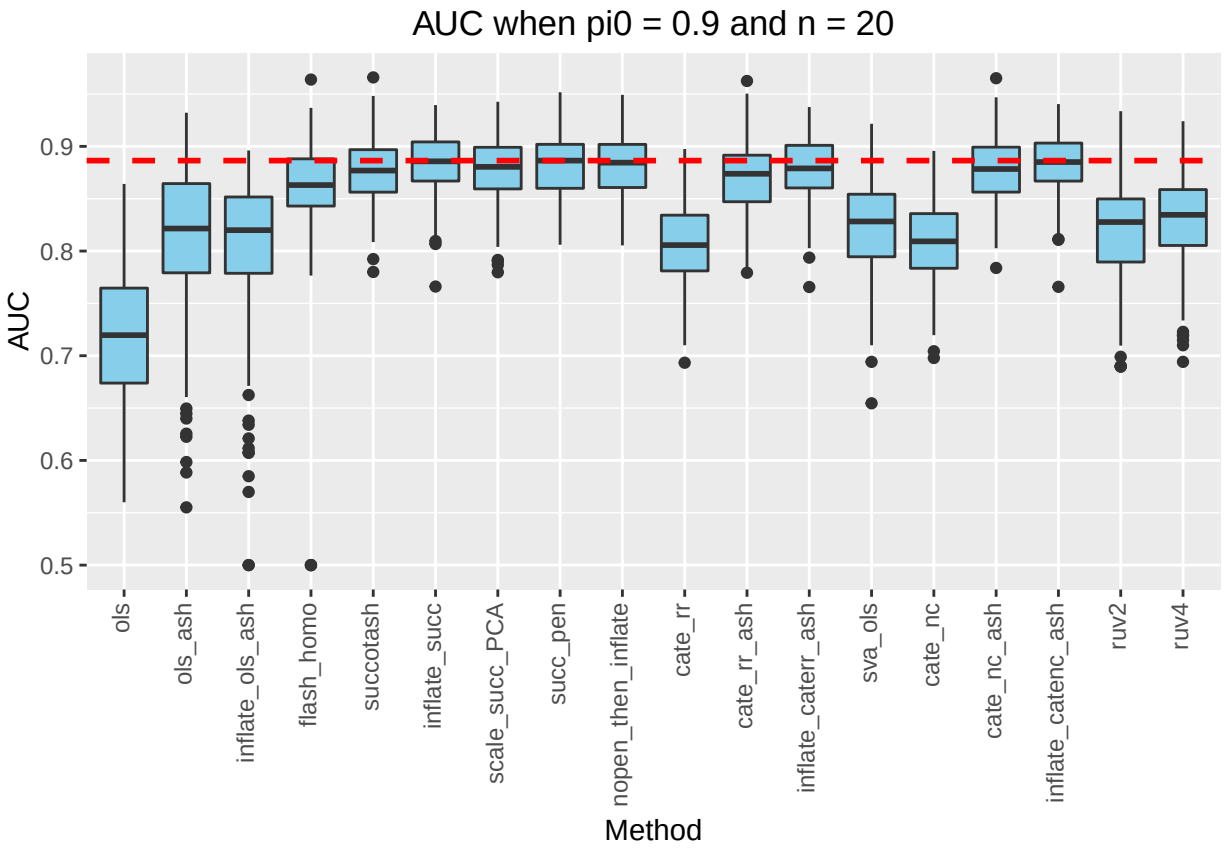
```
                reg_auc, nullpi == current_pi & nsamp == current_nsamp),
            -c(nsamp, nullpi)
        )

        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)
    }
}
```

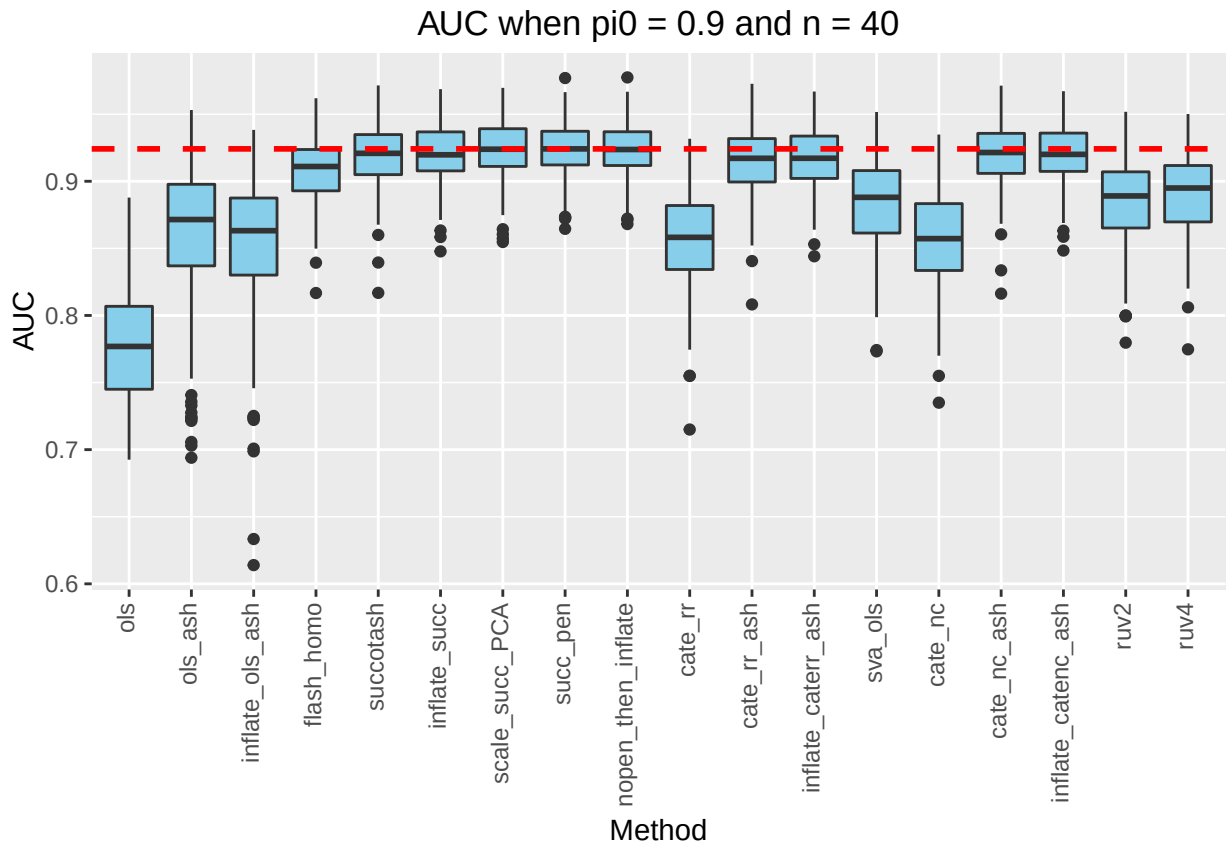



AUC when pi0 = 0.5 and n = 20
AUC
0.9
0.8
0.7
0.6
ols
ols_ash
inflate_ols_ash
flash_homo
succotash
inflate_succ
scale_succ_PCA
succ_pen
nopen_then_inflate
cate_rr
cate_rr_ash
inflate_caterr_ash
sva_ols
cate_nc
cate_nc_ash
inflate_catenc_ash
ruv2
ruv4
Method


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




```
sessionInfo()
```

```
## R version 3.2.4 Revised (2016-03-16 r70336)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 10586)
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## 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_1.12.23
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.4       digest_0.6.9      assertthat_0.1
##  [4] grid_3.2.4        plyr_1.8.3        R6_2.1.2
##  [7] gtable_0.2.0      DBI_0.3.1         formatR_1.3
## [10] magrittr_1.5      scales_0.4.0      evaluate_0.8.3
## [13] highr_0.5.1       stringi_1.0-1     rmarkdown_0.9.5.9
```

```
## [16] labeling_0.3      tools_3.2.4       stringr_1.0.0
## [19] munsell_0.4.3     yaml_2.1.13       parallel_3.2.4
## [22] colorspace_1.2-6  htmltools_0.3.5
```