

CATE + ASH and Different FA's + MOUTHWASH

David Gerard

March 10, 2016

Abstract

I look at the performance of different options of CATE and compare them against the performance of MOUTHWASH using different types of factor analysis. In terms of estimating π_0 , it looks like any homoscedastic version of factor analysis plus MOUTHWASH performs the best. Bayesian shrinkage of the heteroscedastic variances performs intermediately.

1 Methods

For CATE, I varied three parameters.

1. The factor analysis method: Either quasi-maximum likelihood (“ml”), PCA (“pc”), or an early stopping method I haven’t read about but is an option (“esa”).
2. Whether the p-values are calibrated using maximum absolute deviation (TRUE) or not (FALSE). This only matters for the qvalue methods and shouldn’t affect the ASH methods.
3. Whether we used the robust-regression version of CATE (“rr”) or the negative controls version of CATE (“nc”) using half of the null genes as the negative controls.

For each setting in CATE, I performed two methods. The first method consisted of a two-step procedure:

1. Estimate $\hat{\beta}_{[2,i]}$ and it’s corresponding standard error \hat{s}_i .
2. Run ASH on $\hat{\beta}_{[2,i]}$ and \hat{s}_i .

The second method was to use the p-values output by CATE.

I always ran CATE on $\log(COUNTS + 1)$.

The ASH methods provide an estimate of π_0 . I obtained an estimate of π_0 from the p-values by the `qvalue` package in R [Storey, 2002].

The number of hidden confounders was estimated using the methods of Buja and Eyuboglu [1992] implemented in the `num.sv()` function in the `sva` package in R. CATE doesn’t work sometimes when there is only one confounder, so I set the minimum number to 2 confounders.

For MOUTHWASH I used the mixture of normals version with the same regularization and grid-size choices as in the `ashr` package. I used the following factor analysis methods:

- PCA using the column-wise mean squared error to estimate the heteroscedastic variances.
- PCA using the overall mean squared error to estimate the homoscedastic variances.
- Homoscedastic FLASH.
- Heteroscedastic FLASH.
- PCA followed by variance shrinkage using the column-wise mean squared errors and the empirical Bayes shrinkage from `limma`.

More detail is in order for this last method. Suppose we assume the model

$$Y = \Theta + E \quad (1)$$

$$\text{rank}(\Theta) = k \text{ (known)} \quad (2)$$

$$E \sim N_{n \times p}(0, \Sigma \otimes I_n) \quad (3)$$

$$\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2). \quad (4)$$

We estimate Θ with $\hat{\Theta}$, the truncated SVD of Y of rank k . Then define the residual matrix to be $R = Y - \hat{\Theta}$. Let

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^n R_{ij}^2. \quad (5)$$

The estimates of the σ_j^2 's are obtained by fitting the Bayesian hierarchical model of [Smyth \[2004\]](#). That is, assume

$$s_j^2 | \sigma_j^2 \sim \frac{\sigma_j^2}{n-1} \chi_{n-1}^2 \quad (6)$$

$$\frac{1}{\sigma_j^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2, \quad (7)$$

then estimate d_0 and s_0 using a method of moments approach and obtaining $\hat{\sigma}_j^2$ by the posterior means.

2 Simulation Study

I ran through 100 repetitions of generating data from GTEX lung data under the following parameter conditions:

- $n \in \{10, 20, 40\}$,
- $p = 1000$,
- $\pi_0 \in \{0.5, 0.9\}$,
- $\sigma_{\log 2} \in \{1, 5\}$.

I extracted the most expressed p genes (excluding the top 5 expressed genes) from the GTEX lung data and n samples are chosen at random. Half of these samples are randomly given the “treatment” label 1, the other half given the “treatment” label 0. Of the p genes, $\pi_0 p$ were chosen to be non-null. Signal was added by the Poisson-thinning approach in Mengyin’s code with a mean log2-fold change of 0 and a standard deviation log2-fold change of $\sigma_{\log 2}$. That is

$$A_1, \dots, A_{p/2} \sim N(0, \sigma_{\log 2}^2) \quad (8)$$

$$B_i = 2^{A_i} \text{ for } i = 1, \dots, p/2. \quad (9)$$

If $A_i > 0$ then we replace $Y_{[1:(n/2), i]}$ with $\text{Binom}(Y_{[j, i]}, 1/B_i)$ for $j = 1, \dots, n/2$. If $A_i < 0$ then we replace $Y_{[(n/2+1):n, i]}$ with $\text{Binom}(Y_{[j, i]}, B_i)$ for $j = n/2 + 1, \dots, n$.

For each iteration, I calculated two things:

1. The AUC using either the lfdrs or p-values.
2. The estimates of π_0 .

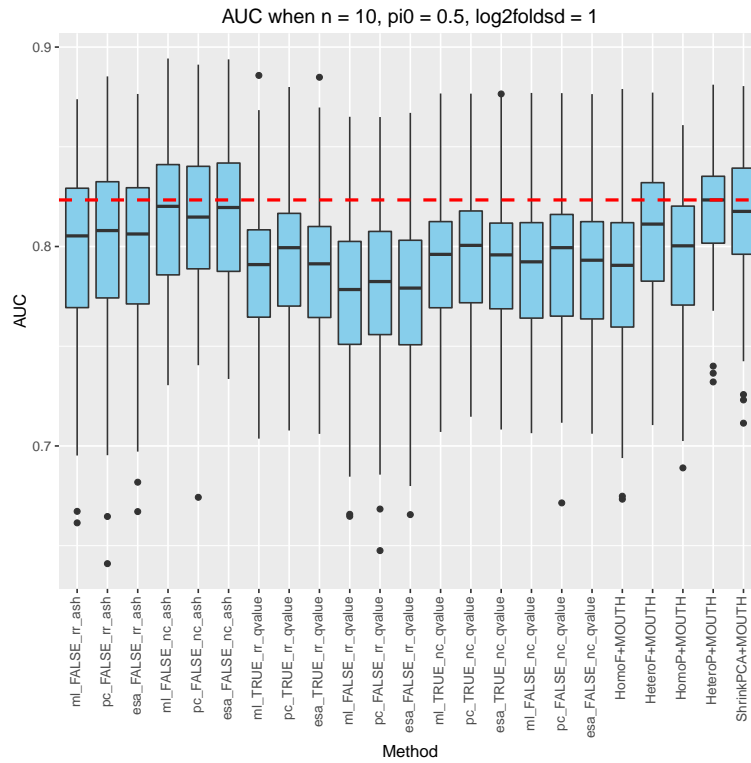
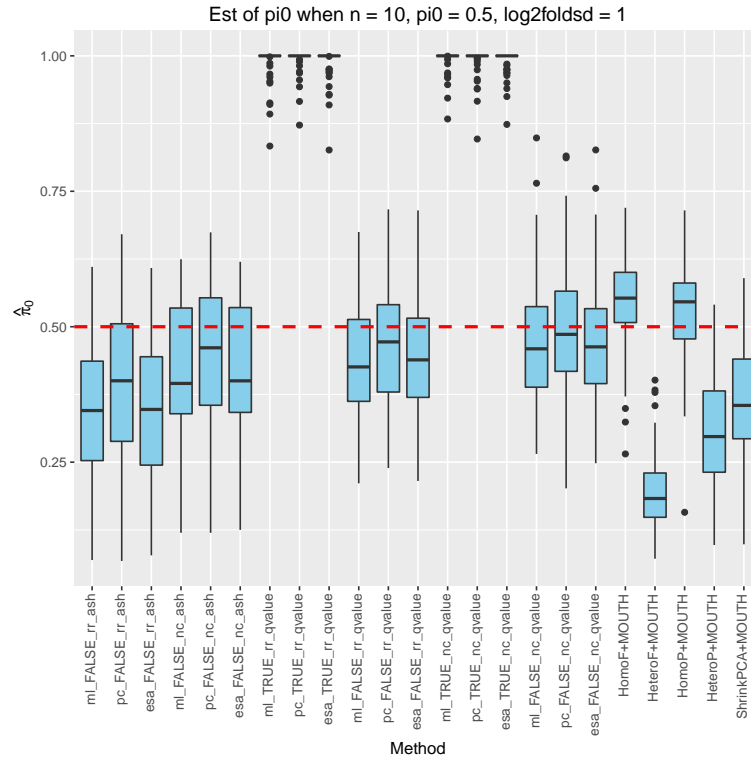
For the frequentist procedures, I used the vector of p-values as the predictions and I used the vector of lfdr's from the ASH-like procedures for prediction. These were used to create ROC curves and calculate AUCs.

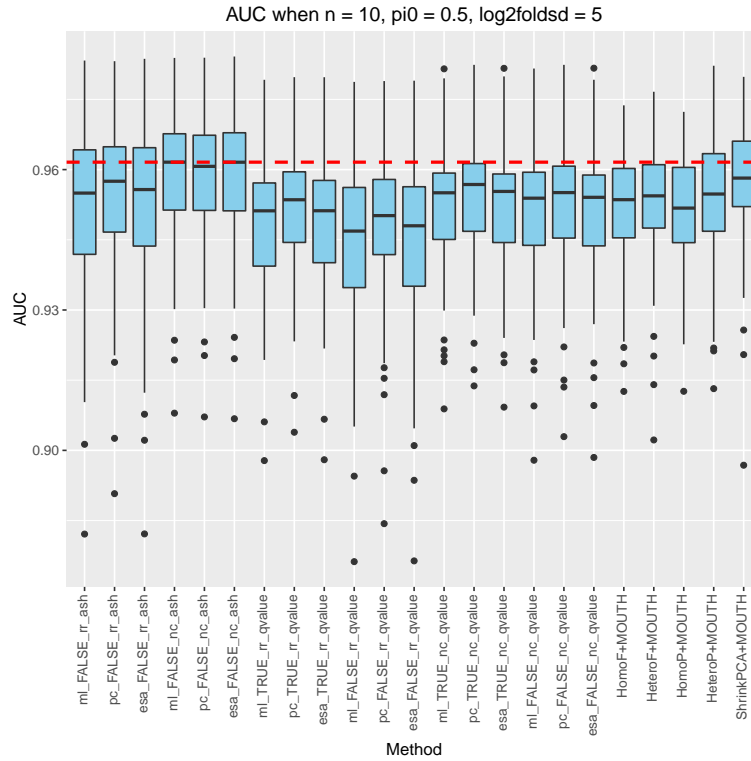
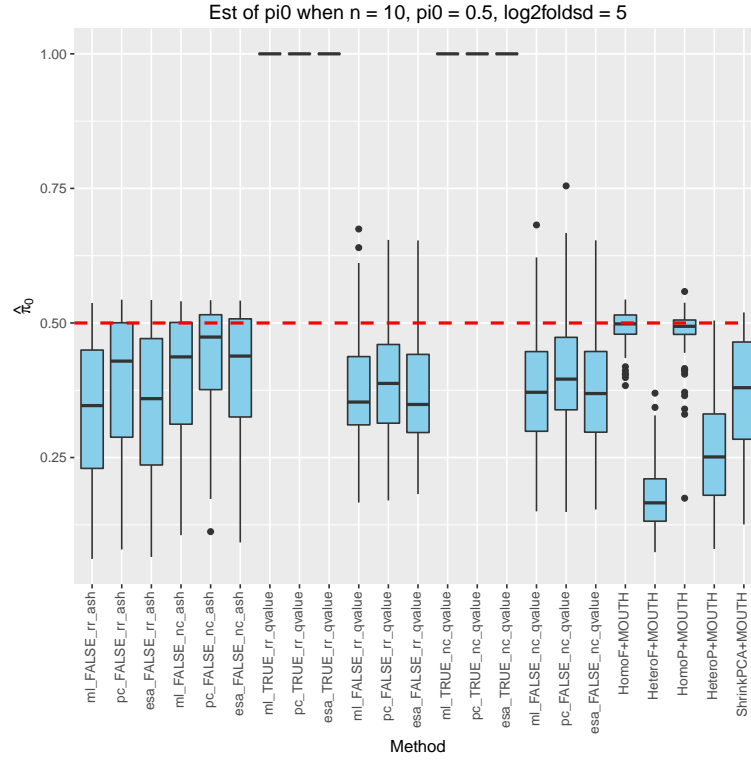
3 Results

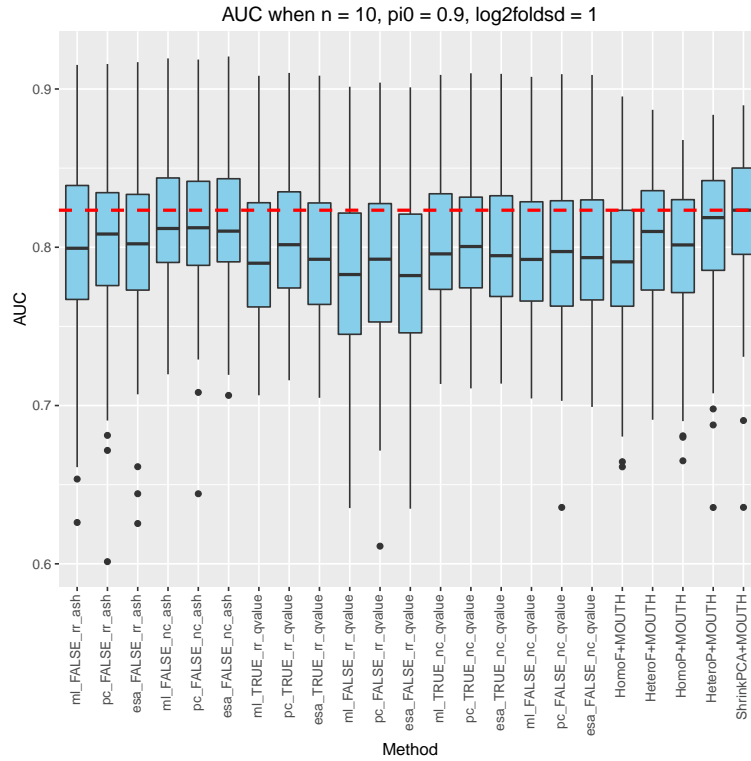
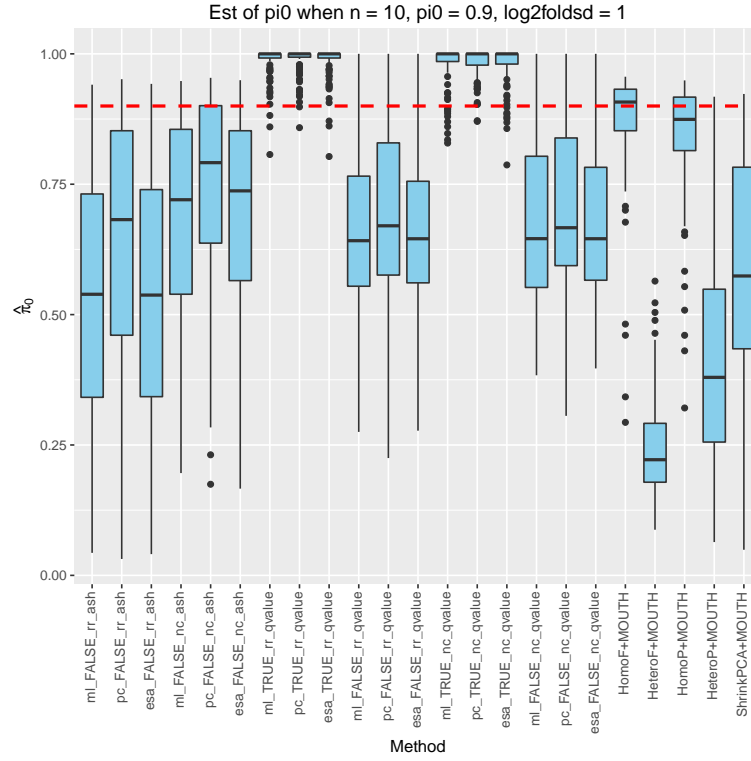
In terms of estimating π_0 , SUCCOTASH using the homoscedastic variance models works so much better than those using the heteroscedastic models. It appears that FLASH's superior calibration was just because of the variance modeling.

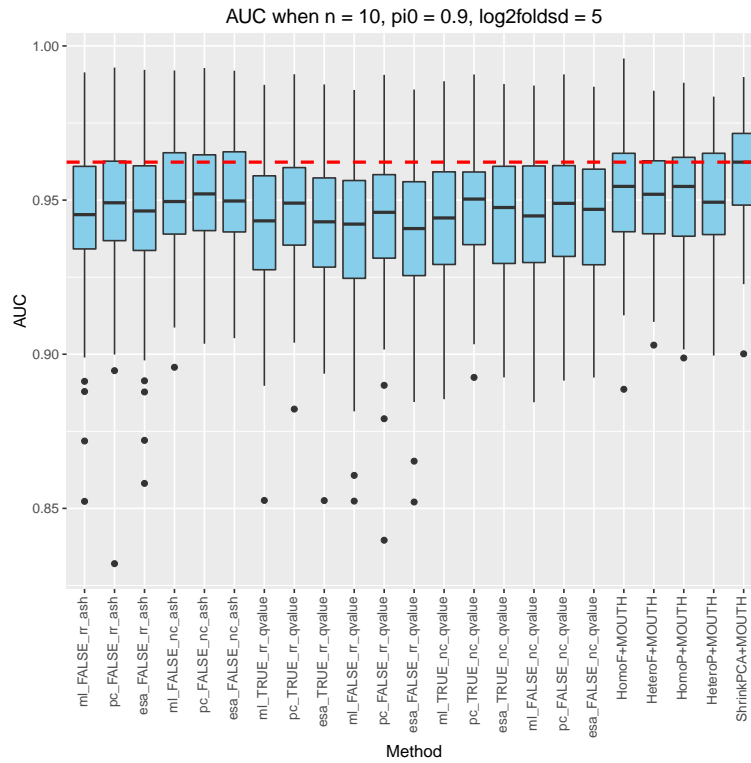
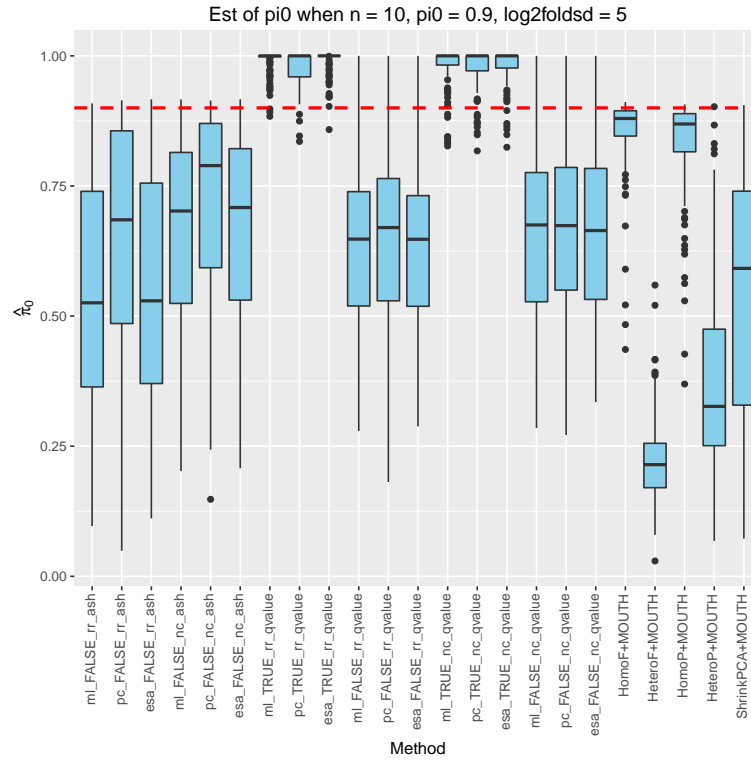
Shrinking the variances using the empirical Bayes approach of `limma` did better at estimating π_0 than just using the column wise mean squared errors, but worse than the homoscedastic assumption.

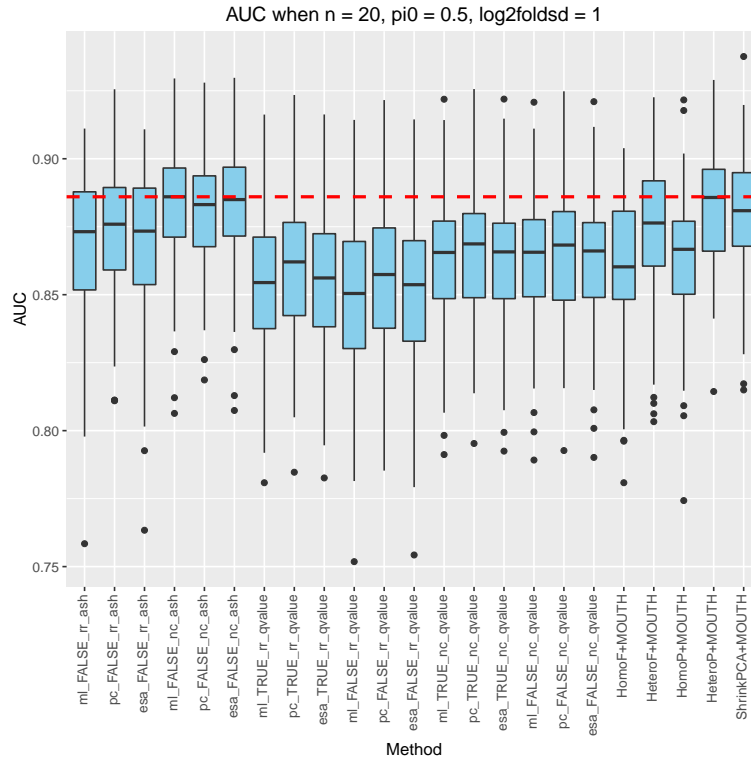
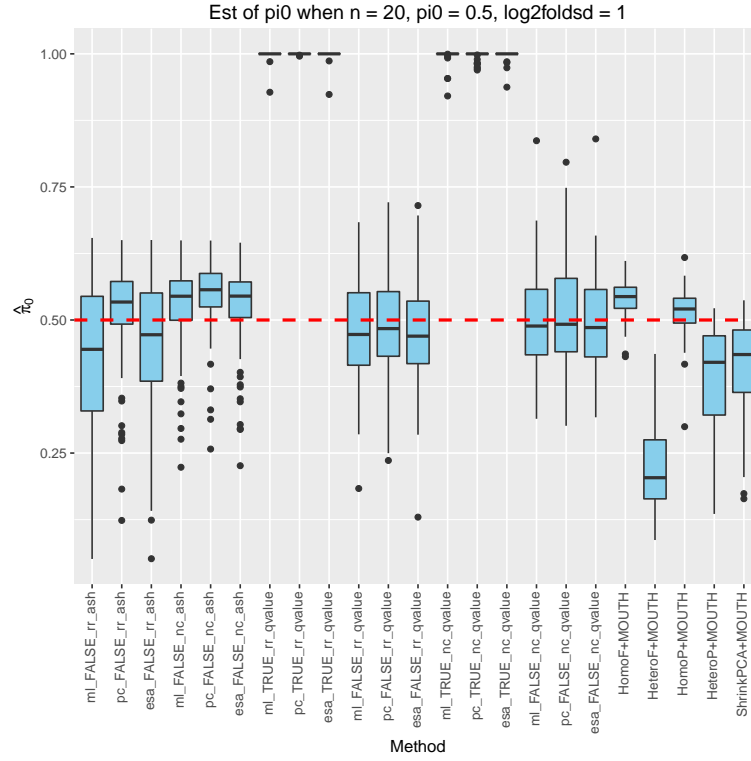
In terms of AUC, the heteroscedastic models work better than the homoscedastic models.

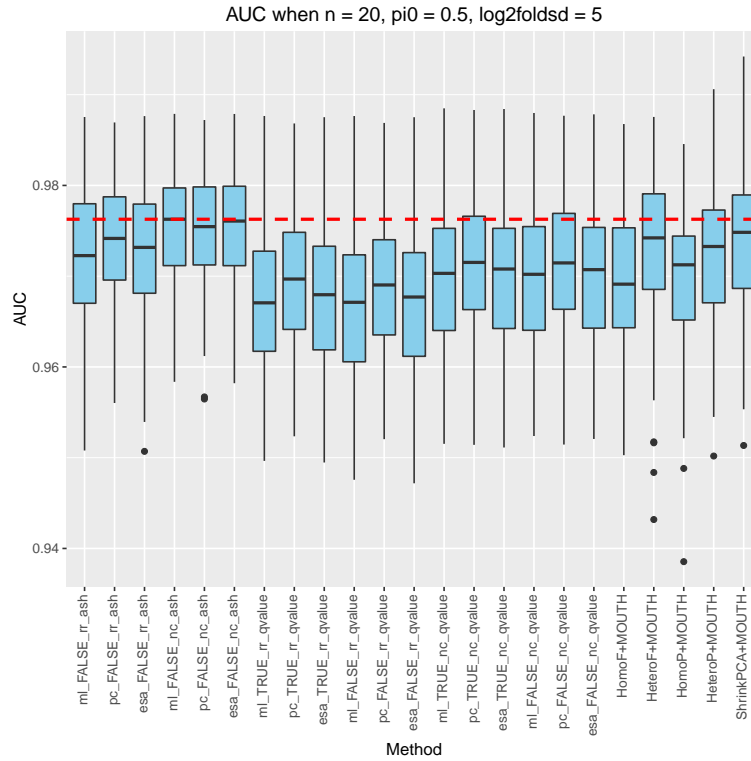
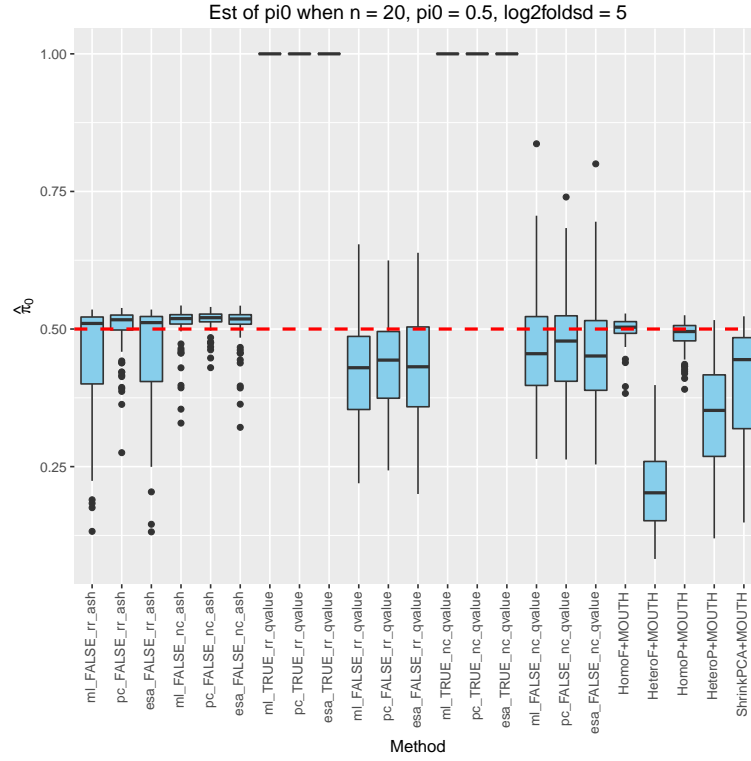


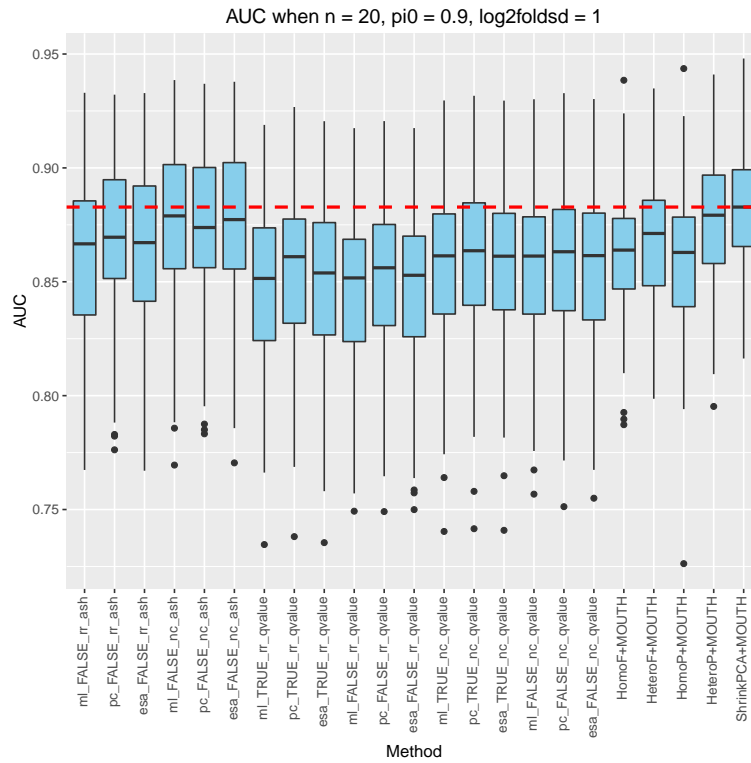
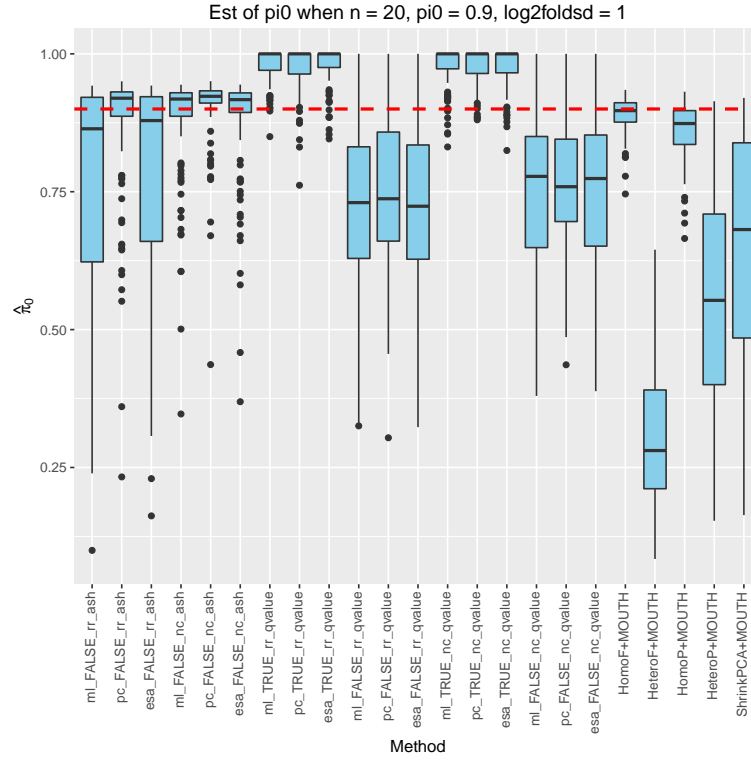


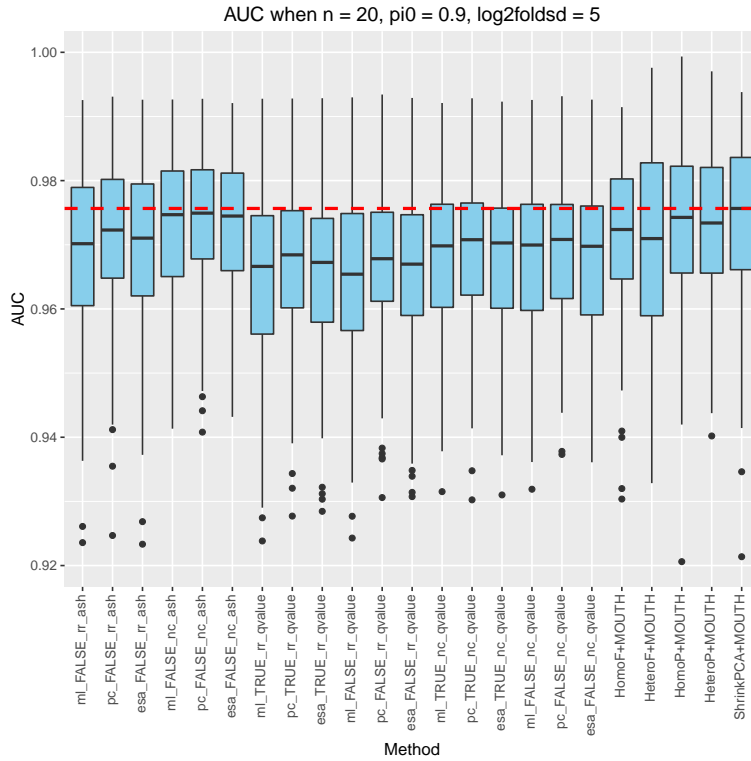
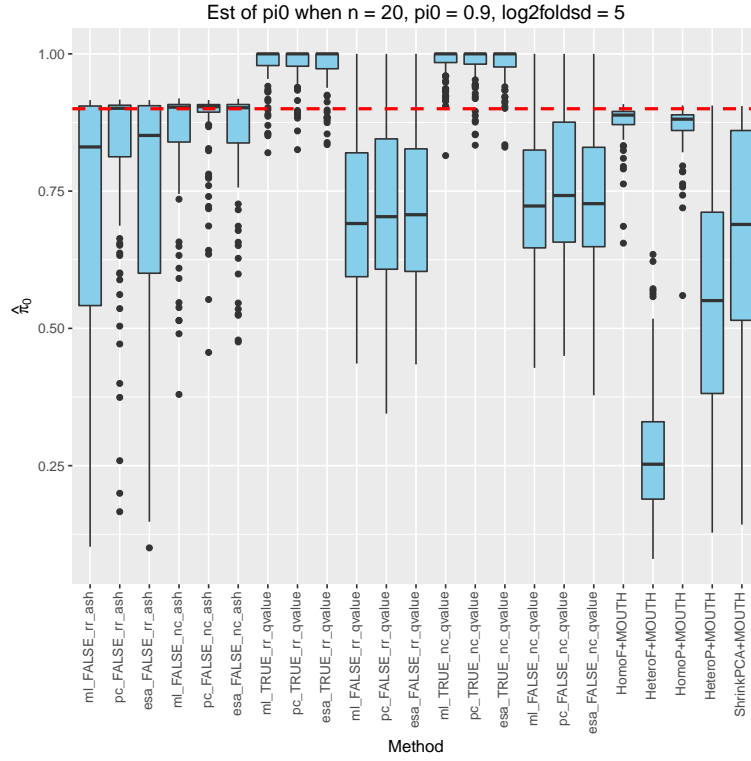


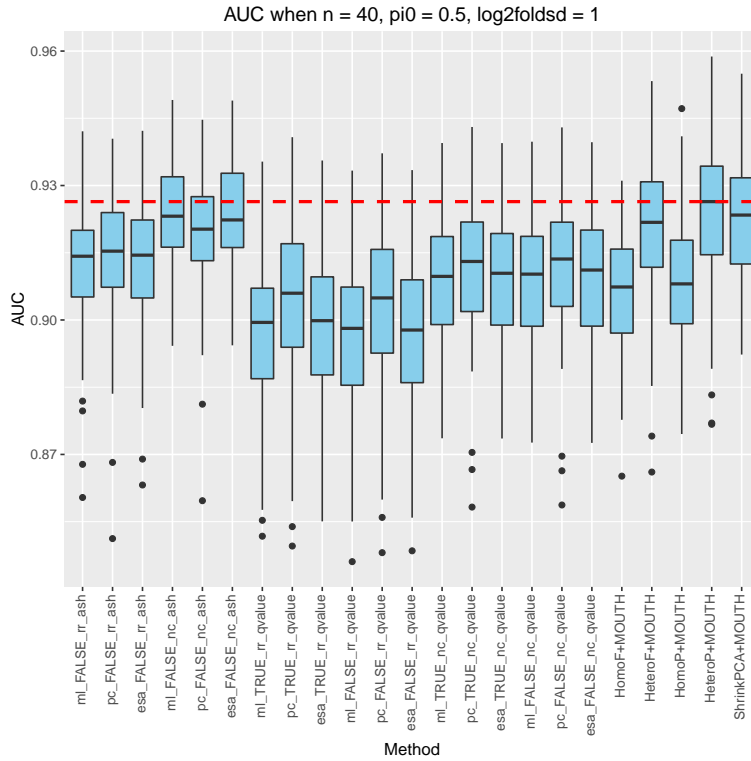
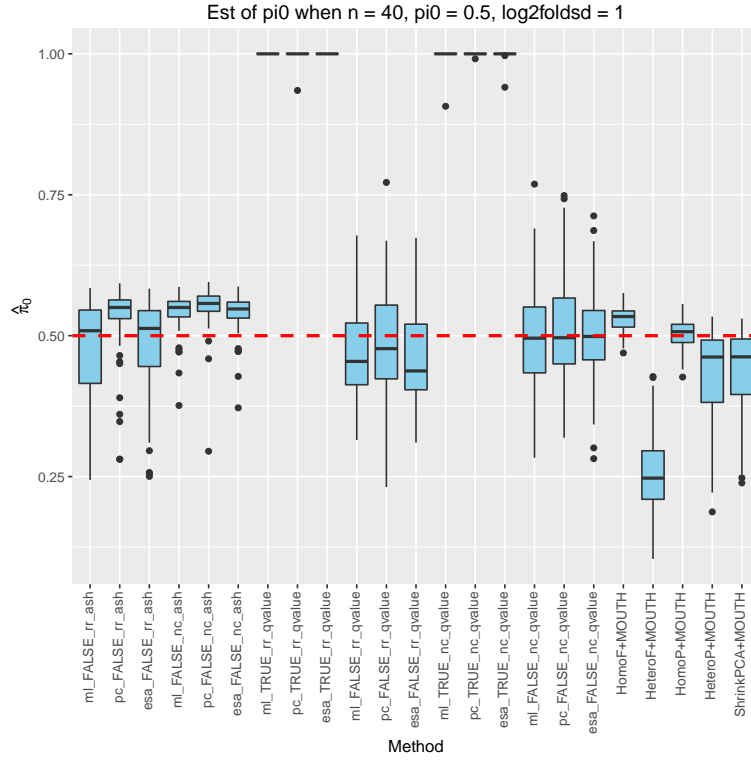


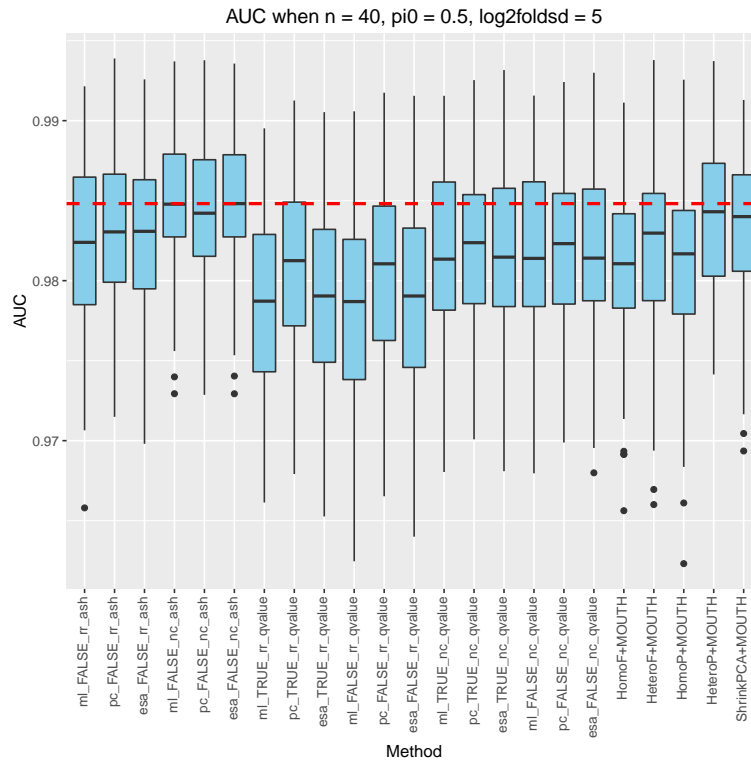
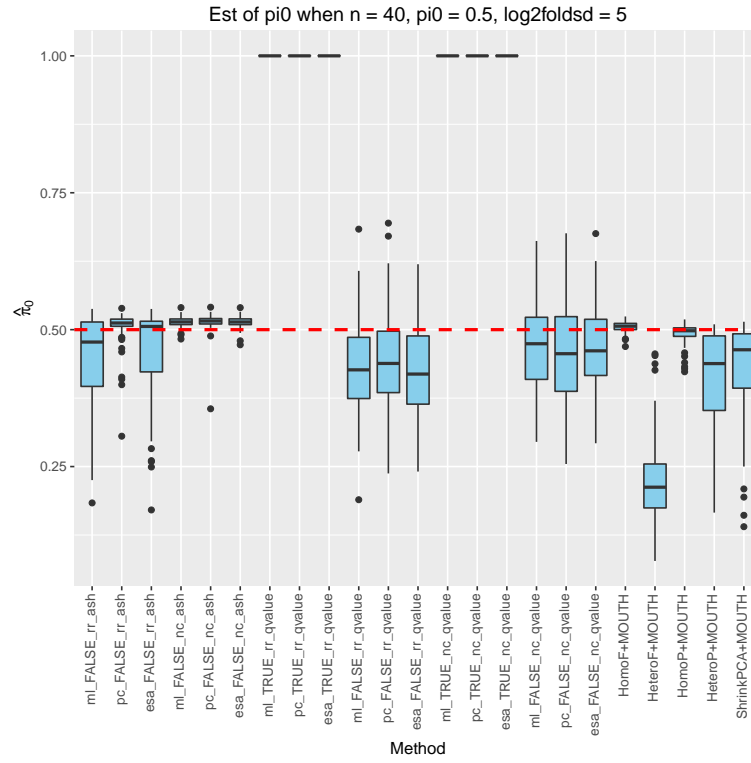


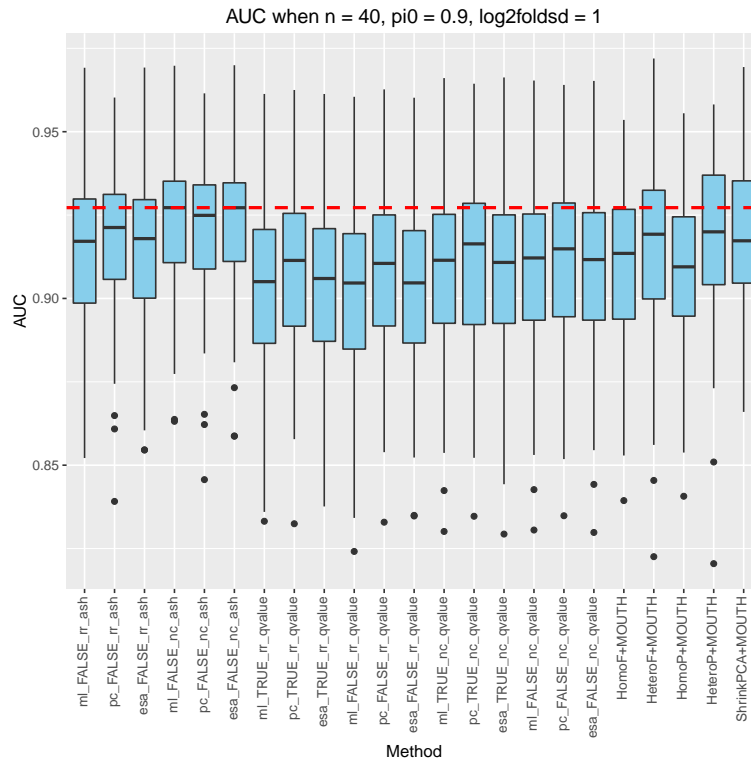
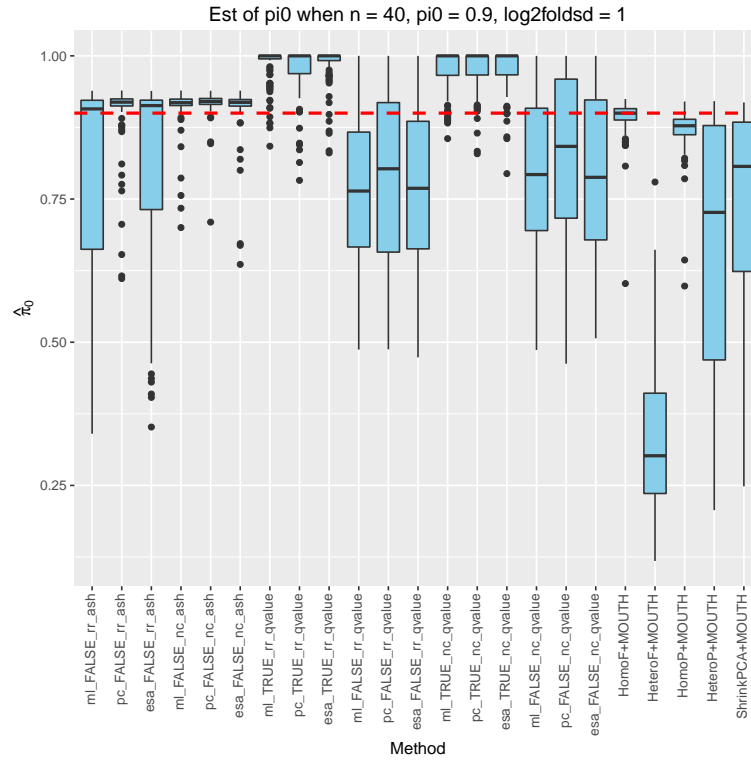


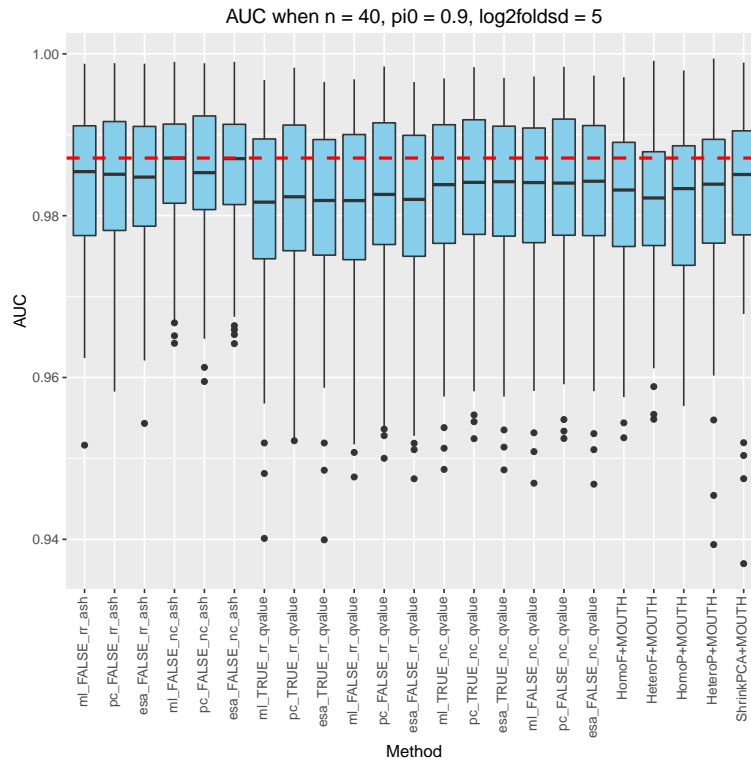
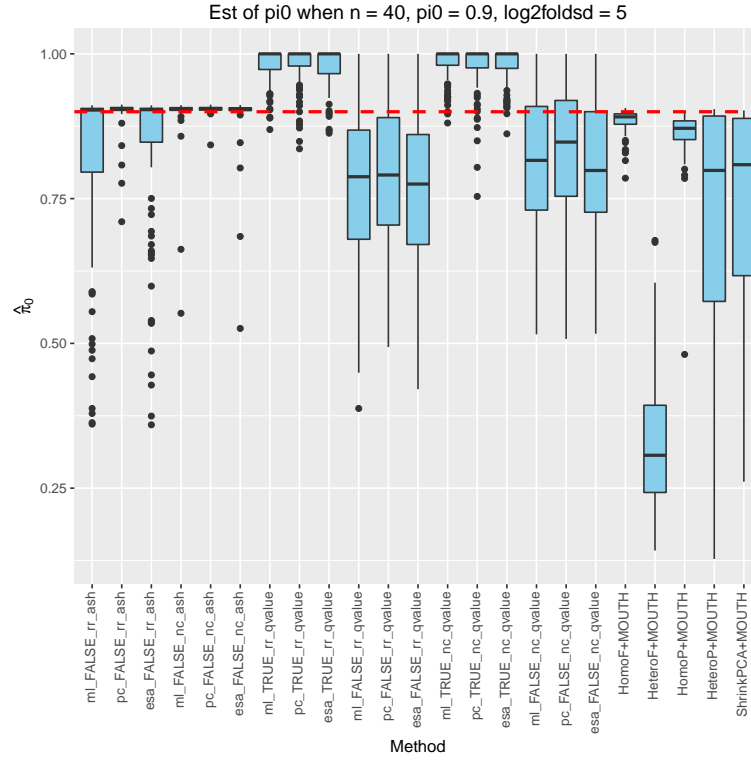












References

- Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540, 1992.
- Gordon K Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.
- John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.