MOUTHWASH with Uniform Mixtures

David Gerard

March 15, 2016

Abstract

I do initial simulations using mixtures of uniforms as the prior.

1 Methods

I first normalized the counts by $\log(COUNTS + 1)$. 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`.

I methods of MOUTHWASH I looked at were

- t-likelihood, uniform mixtures, using homoscedastic PCA.
- t-likelihood, uniform mixtures, using PCA with `limma`-shrunk variance estimates, adjusting with the estimated df.
- t-likelihood, uniform mixtures, using heteroscedastic PCA.
- normal likelihood, uniform mixtures, using heteroscedastic PCA.
- normal likelihood, normal mixtures, using homoscedastic PCA.
- normal likelihood, normal mixtures, using heteroscedastic PCA.

For the t-likelihoods, I always used $n-1$ degrees of freedom except when using the `limma`-shrunk variances.

2 Simulation Study

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

- $n \in \{10, 20, 40\}$,
- $p = 100$. A smaller number of genes was used because the uniform mixture implementation is much slower than the normal mixture implementation. So using more genes would have resulted in much longer running times.
- $\pi_0 \in \{0.5, 0.9\}$,

- $\sigma_{log2} = 1$. I was running into numerical issues at $\sigma_{log2} = 5$ for the uniform mixtures for some of the data sets. The error code wasn't too helpful and rerunning MOUTHWASH on the same data set would be fine. So it must be that for some starting values for some data sets I am running into numerical issues. This seems very hard to debug.

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 "control" 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 σ_{log2}. That is

$$A_1, \ldots, A_{p/2} \sim N(0, \sigma^2_{log2}) \tag{1}$$

$$B_i = 2^{A_i} \text{ for } i = 1, \ldots, p/2. \tag{2}$$

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

For each iteration, I calculated two things:

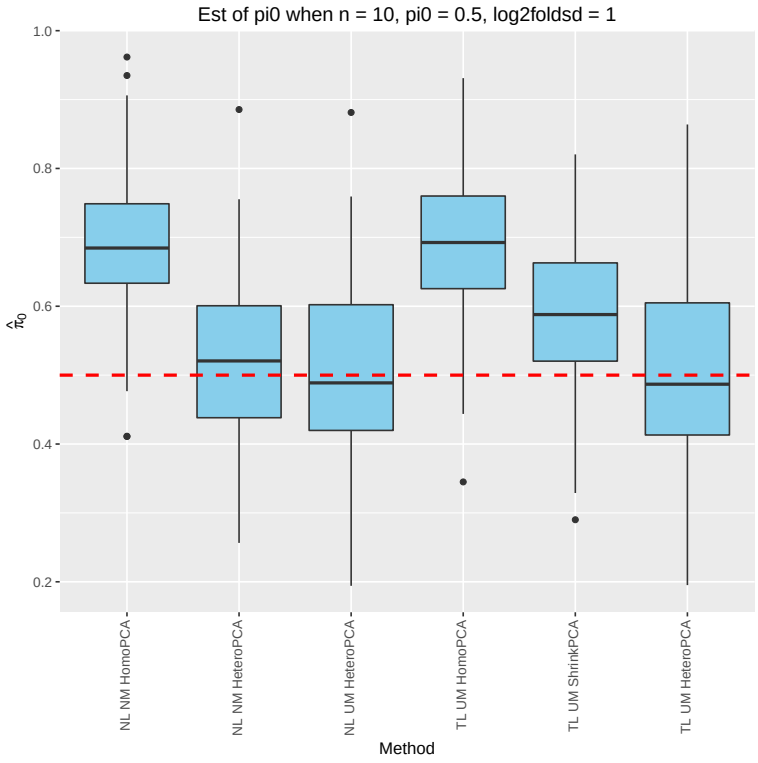
1. The AUC using either the lfdrs.
2. The estimates of π_0.

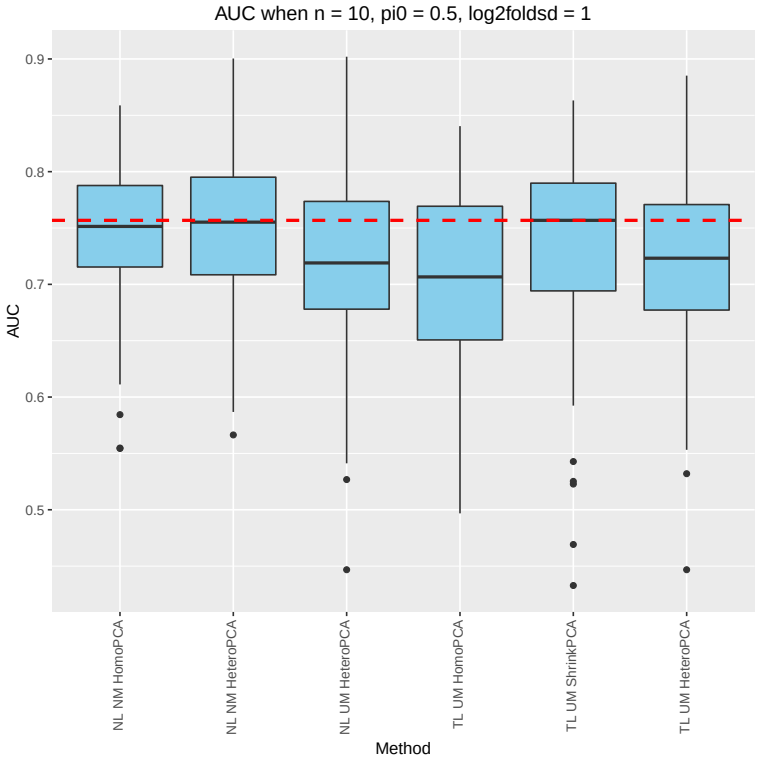
3 Results

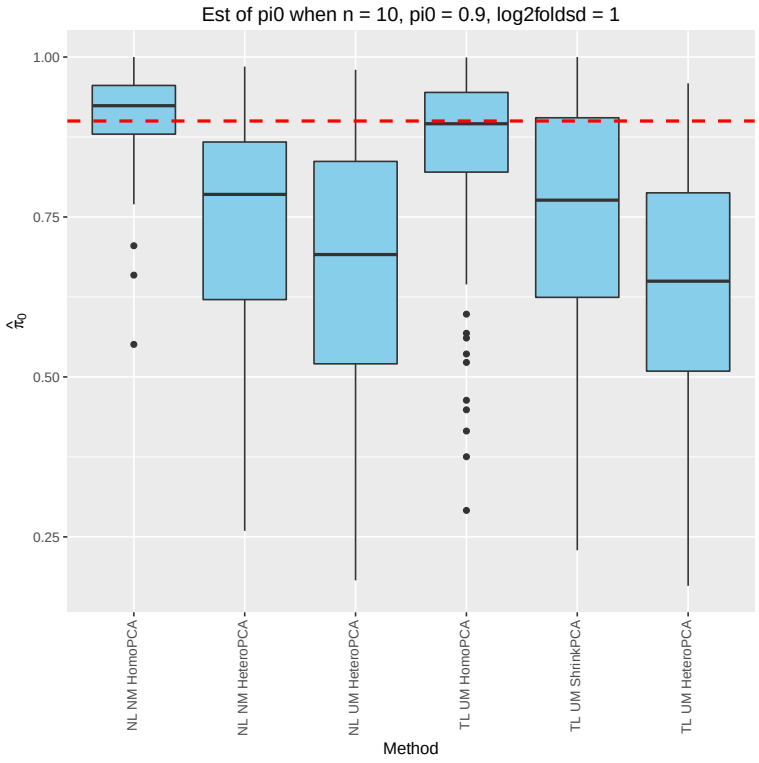
In the plots below, I use the following abbreviations:

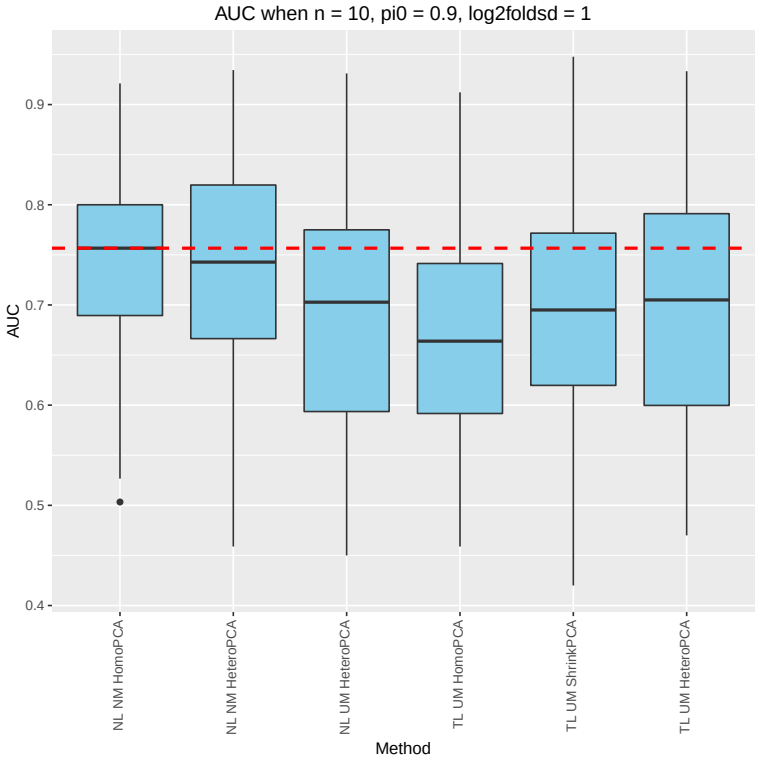
- NL = normal likelihood, TL = t-likelihood,
- NM = normal mixture, UM = uniform mixtures,
- HomoPCA = PCA + homoscedastic variance model.
- HeteroPCA = PCA + heteroscedastic variance model.
- ShrinkPCA = PCA + `limma`-shrunk variances.

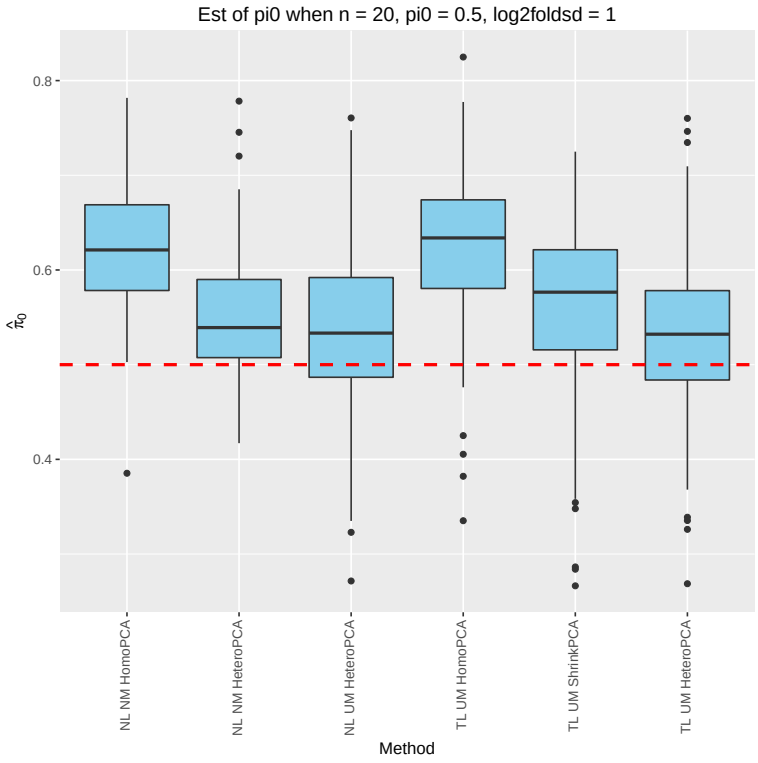
In general, the most important effect is the variance model. Again, I am seeing that homoscedasticity works really well. The t-likelihood doesn't noticeably improve estimating π_0. I am seeing that for large n, all of the methods start performing well at estimate π_0.

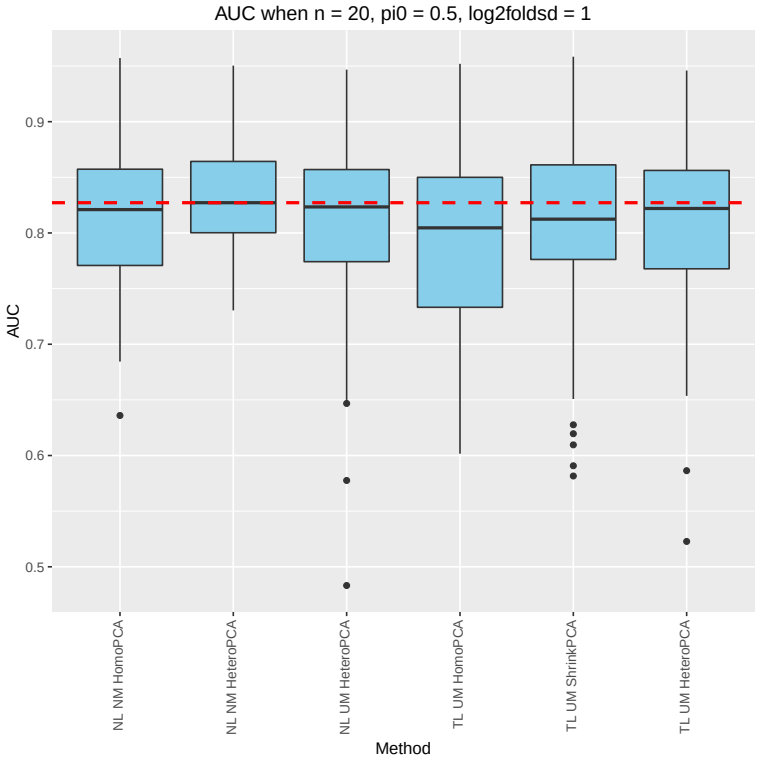

Est of pi0 when n = 10, pi0 = 0.5, log2foldsd = 1
1.0
0.8
0.6
0.4
0.2
$\hat{\pi}_0$
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

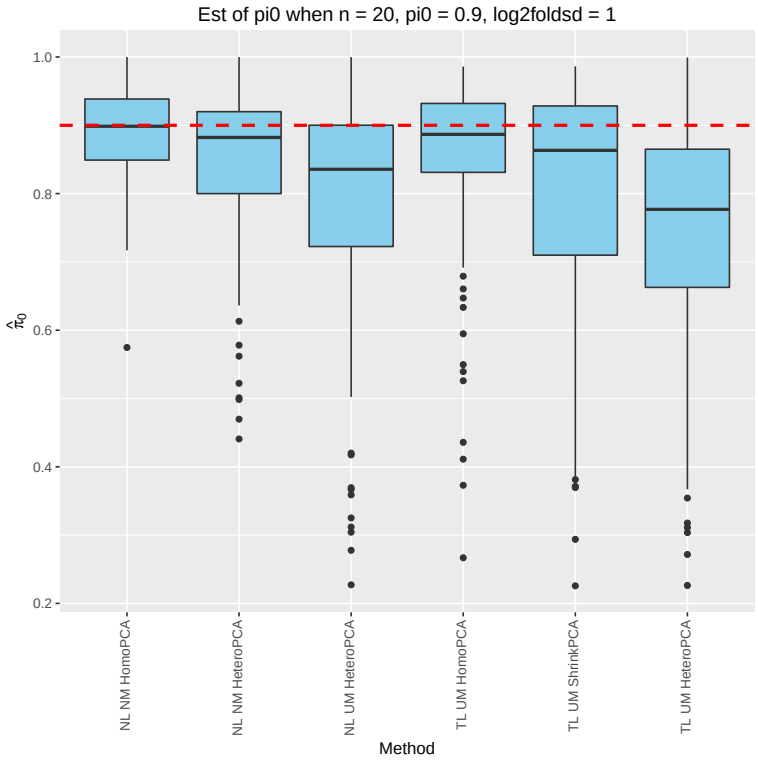

AUC when n = 10, pi0 = 0.5, log2foldsd = 1
0.9
0.8
0.7
0.6
0.5
AUC
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

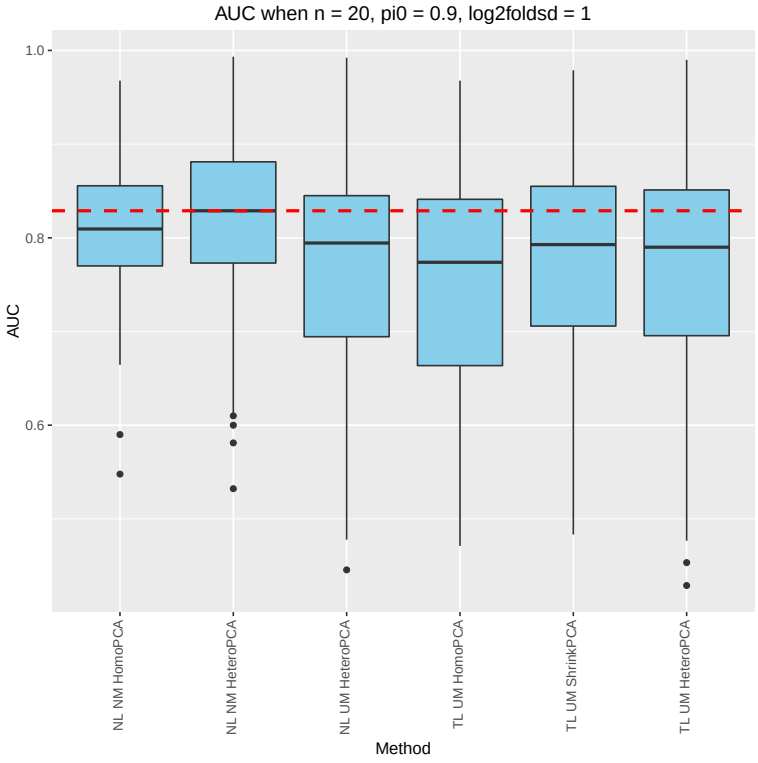

Est of pi0 when n = 10, pi0 = 0.9, log2foldsd = 1
1.00
0.75
0.50
0.25
$\hat{\pi}_0$
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

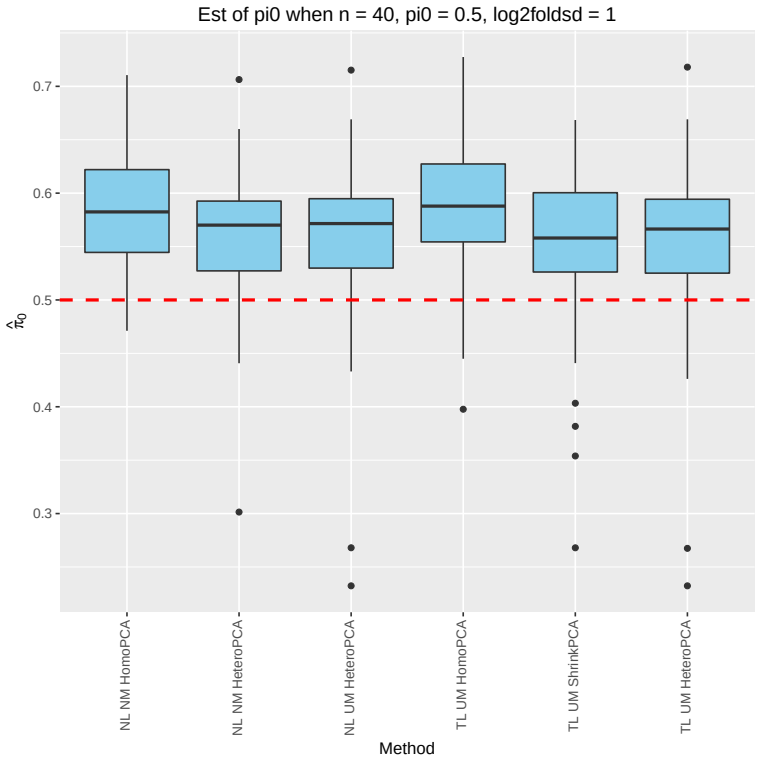

AUC when n = 10, pi0 = 0.9, log2foldsd = 1
0.9
0.8
0.7
0.6
0.5
0.4
AUC
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

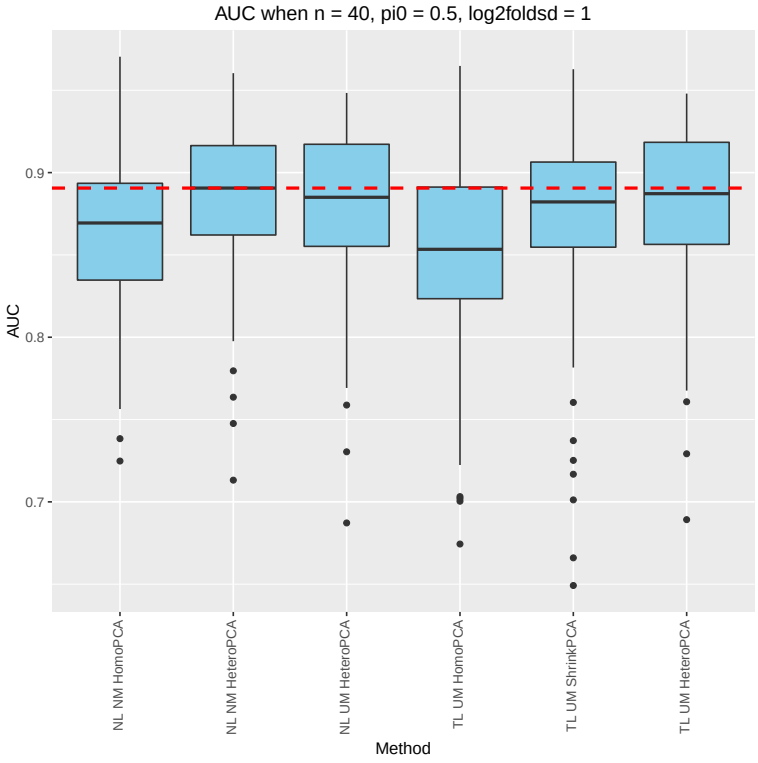

Est of pi0 when n = 20, pi0 = 0.5, log2foldsd = 1
0.8
0.6
0.4
$\hat{\pi}_0$
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

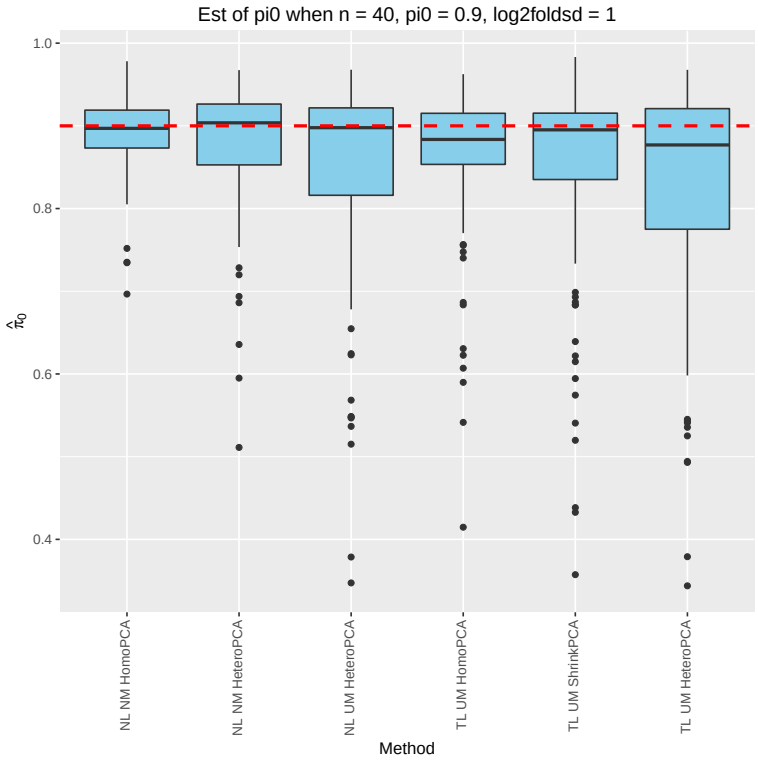

AUC when n = 20, pi0 = 0.5, log2foldsd = 1
0.9
0.8
0.7
0.6
0.5
AUC
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

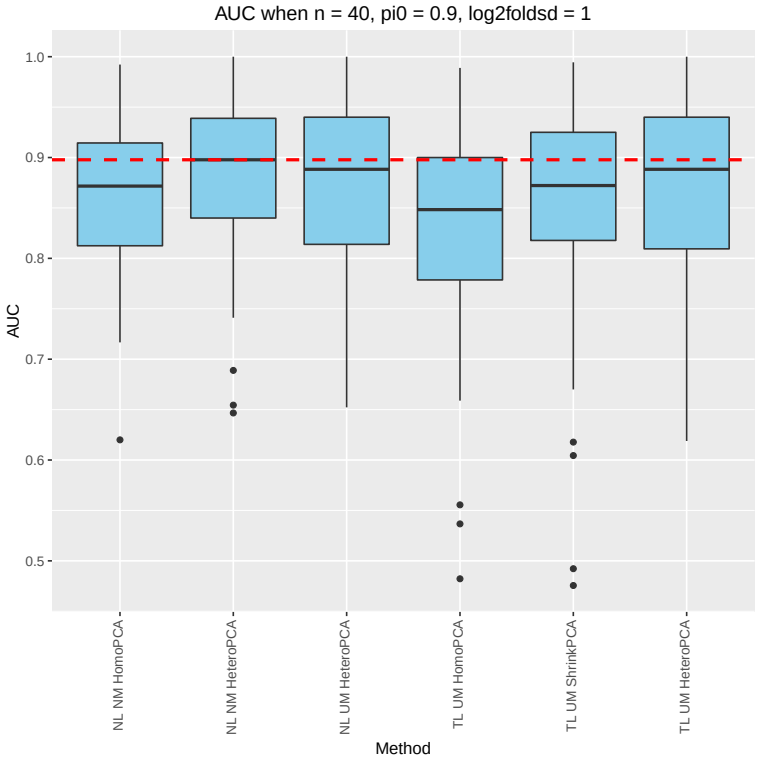

Est of pi0 when n = 20, pi0 = 0.9, log2foldsd = 1
1.0
0.8
0.6
0.4
0.2
$\hat{\pi}_0$
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method


AUC when n = 20, pi0 = 0.9, log2foldsd = 1
1.0
0.8
0.6
AUC
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method


Est of pi0 when n = 40, pi0 = 0.5, log2foldsd = 1
0.7
0.6
0.5
0.4
0.3
$\hat{\pi}_0$
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method


AUC when n = 40, pi0 = 0.5, log2foldsd = 1
0.9
0.8
0.7
AUC
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method


Est of pi0 when n = 40, pi0 = 0.9, log2foldsd = 1
1.0
0.8
0.6
0.4
$\hat{\pi}_0$
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method


AUC when n = 40, pi0 = 0.9, log2foldsd = 1
1.0
0.9
0.8
0.7
0.6
0.5
AUC
NL NM HomoPCA
NL NM HeteroPCA
NL UM HeteroPCA
TL UM HomoPCA
TL UM ShrinkPCA
TL UM HeteroPCA
Method

References

Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4):509–540, 1992.