

Different Regularization Values for MOUTHWASH, Compare to CATE

David Gerard

February 25, 2016

Abstract

I look at the performance of different regularizations of MOUTHWASH. I compare these against CATE's performance.

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 its 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 PCA to estimate the hidden confounders as this is the option that works best in CATE.
- I used the mixture of normals version with the same grid-size as in the `ashr` package.
- I varied λ in $\{10, 100, 150, 200, 250\}$.

2 Simulation Study

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

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

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

$$A_1, \dots, A_{p/2} \sim N(0, \sigma_{\log 2}^2) \quad (1)$$

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

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 lfders or p-values.
2. The estimates of π_0 .

3 Results

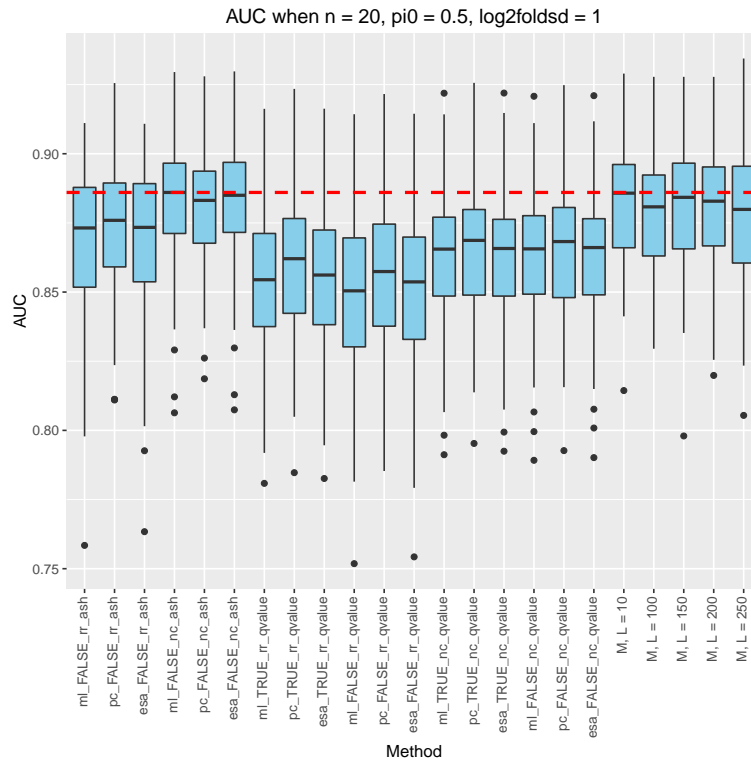
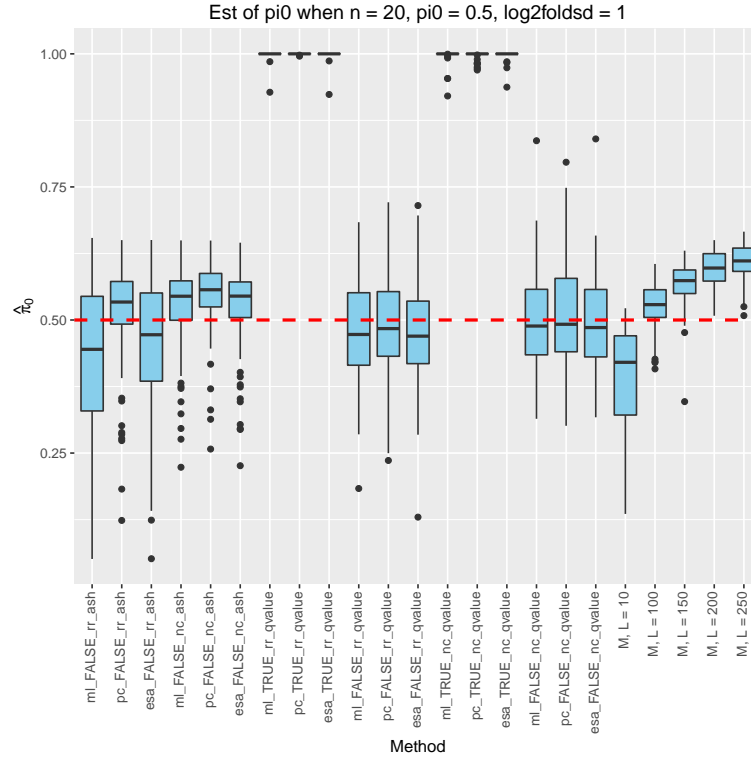
For the frequentist procedures, I used the vector of p-values as the predictions and I used the vector of lfders from the ASH-like procedures for prediction. These were used to create ROC curves and calculate AUCs. In general, ASH procedures performed better than just using the p-values and using negative controls worked better than the robust regression version. MOUTHWASH does OK in terms of AUC — sometimes the best (even compared to the methods that use negative controls), sometimes among the top.

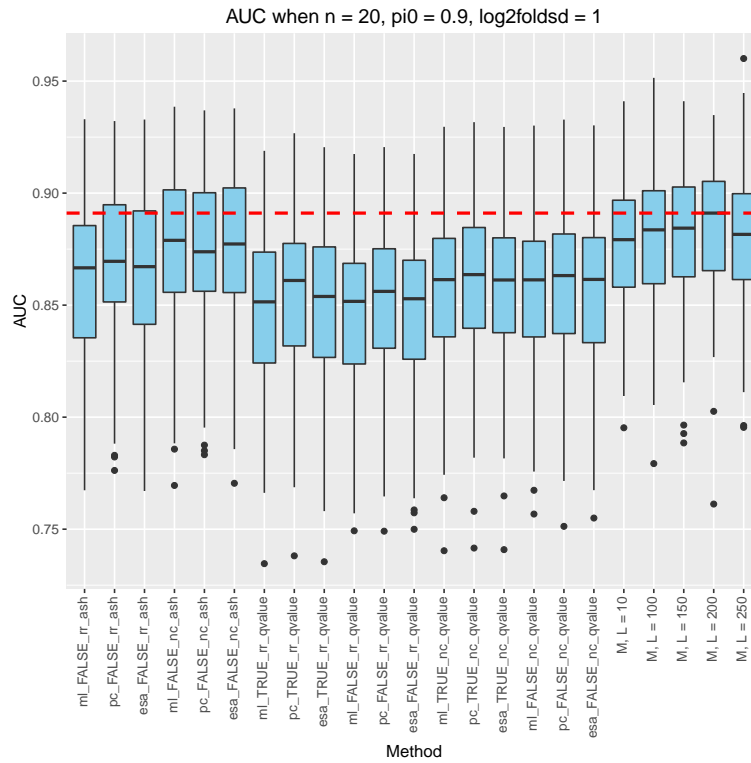
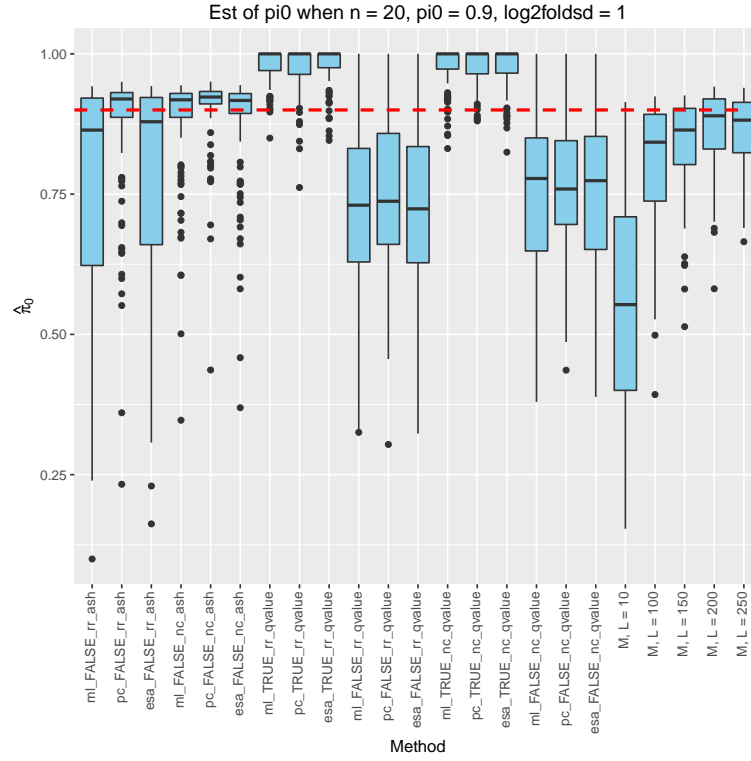
From the p-values, I used the `qvalue` package [Storey, 2002] to estimate π_0 . Estimates of π_0 are given from `ashr` for the ASH-like methods.

At $\pi_0 = 0.9$, every regularization I tried of MOUTHWASH still under estimated π_0 . But $\lambda = 100$ was enough to get conservative estimates of π_0 when $\pi_0 = 0.5$. AUC wasn’t much affected by the choice of λ .

Inserting the calibrated p-values of CATE into `qvalue` always resulted in very high estimates of π_0 .

In the figures below, “M” refers to MOUTHWASH and “L =” refers to the value of the regularization parameter. The rest of the methods are either CATE + ASH or CATE + `qvalue`.





References

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