SUCCOTASH vs Methods in Mengyin's Code when all Null, Sample Size of 40

David Gerard

February 9, 2016

Abstract

Here, I compare SUCCOTASH's lfdr to the lfdr's provided in all of the methods in Mengyin's code. I then compare SUCCOTASH's estimation performance to those of CATE, SVA, LEAPP, RUV, and OLS. These are all under the data generation process that Mengyin coded up using the GTEX data where all genes are null. This is the same simulation setup as in suc_v_rest_real_writeup.pdf except in a 20 versus 20 design.

1 Competitors

For each of procedure in Mengyin's code, I performed the following 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 methods available in Mengyin's code to get $\hat{\beta}_{[2,i]}$ and $\hat{s}_i$ were

- VOOM [Law et al., 2014].
- RUVseq [Risso et al., 2014] followed by VOOM [Law et al., 2014] with the estimated confounding factors. Half of the factors were used as control genes.
- SVAseq [Leek, 2014] followed by VOOM [Law et al., 2014] with the estimated confounding factors.
- RUVseq + quasi-binomial glm.
- SVAseq + quasi-binomial glm.
- MYRNA, which is just a quasi-binomial glm using the 75th percentile of the samples' counts as covariates [Langmead et al., 2010].
- MYRNA offset, which is just a quasi-binomial glm using the 75th percentile of the samples' counts as offsets [Langmead et al., 2010].
- EdgeR [Robinson et al., 2010].
- DESeq2glm [Love et al., 2014].

Note that "VOOM" means using VOOM [Law et al., 2014] to find weights for each observations, then fitting a linear model using LIMMA [Smyth, 2005].

I also compared the estimation performance of SUCCOTASH with

- LEAPP [Sun et al., 2012].
- The robust regression version of CATE [Wang et al., 2015].

- SVA [Leek and Storey, 2007] with the number of confounders estimated using the method of Bai et al. [2012].
- RUV4 [Gagnon-Bartsch et al., 2013] with 50% of the observations being control genes with the number of confounders estimated using the method of Bai et al. [2012].
- The ordinary least squares (OLS) estimator.

The factor analysis part of SUCCOTASH was done with the quasi-mle approach of Bai et al. [2012] with the number of hidden confounders using the methods of Buja and Eyuboglu [1992] implemented in the `num.sv()` function in the `sva` package in `R`.

2 Simulation Study

I ran through 100 repetitions of generating the data using Mengyin's code with

- $n = 40$,
- $p = 1000$.

That is, 1000 genes are chosen at random from the GTEX lung data and 40 samples are chosen at random. Twenty of these samples are randomly given the "treatment" label 1, the other twenty given the "treatment" label 0. Since the assignments are random, the true effect is 0 for all genes.

For SUCCOTASH, CATE, LEAPP, SVA, RUV, and OLS I merely applied a log transformation to the counts with a 1 offset before applying each method.

I calculated the sum of squared errors (SSE's) for SUCCOTASH, LEAPP, CATE, RUV, SVA, and OLS on the log(counts + 1) data. Since all effects are null, this is just the sum of squares for each method's estimates. I didn't look at the SSE of the other methods because they have different data normalization procedures.

I also compared the local false discovery rates (lfdr's) of SUCCOTASH, VOOM, RUVseq + VOOM, SVAseq + VOOM, RUVseq + quasi-binomial glm, SVAseq + quasi-binomial glm, MYRNA, MYRNA offset, EdgeR, and DESeq2. Specifically, I calculated the mean of the lfdr's at each iteration for all methods. The higher the mean lfdr's the better as all genes are null.

3 Results

SUCCOTASH performed the best in terms of MSE against the other confounder adjustment methods (Figure 1).

The results for lfdr are in Figure 2. SUCCOTASH seems to perform the worst among all methods tested. All of the other methods perform comparably to each other.

Note that for these data sets, some of the quasi-binomial glm methods were unable to provide lfdr's. For each of the 100 trials, the number of times each method failed to provide lfdr's is provided in Table **??**. The labels in Table **??** are the same as in Figure 2.

```
## Error in is.data.frame(x):  object 'lfdr_mat' not found
## Error in xtable(fail_mat, caption = paste0("Number of times each method failed to
provide lfdrs out of ", :  object 'fail_mat' not found
```

Figure 1: Mean squared erorrs (MSE) for SUCCOTASH (SUCC), LEAPP, CATE, RUV, SVA, and OLS

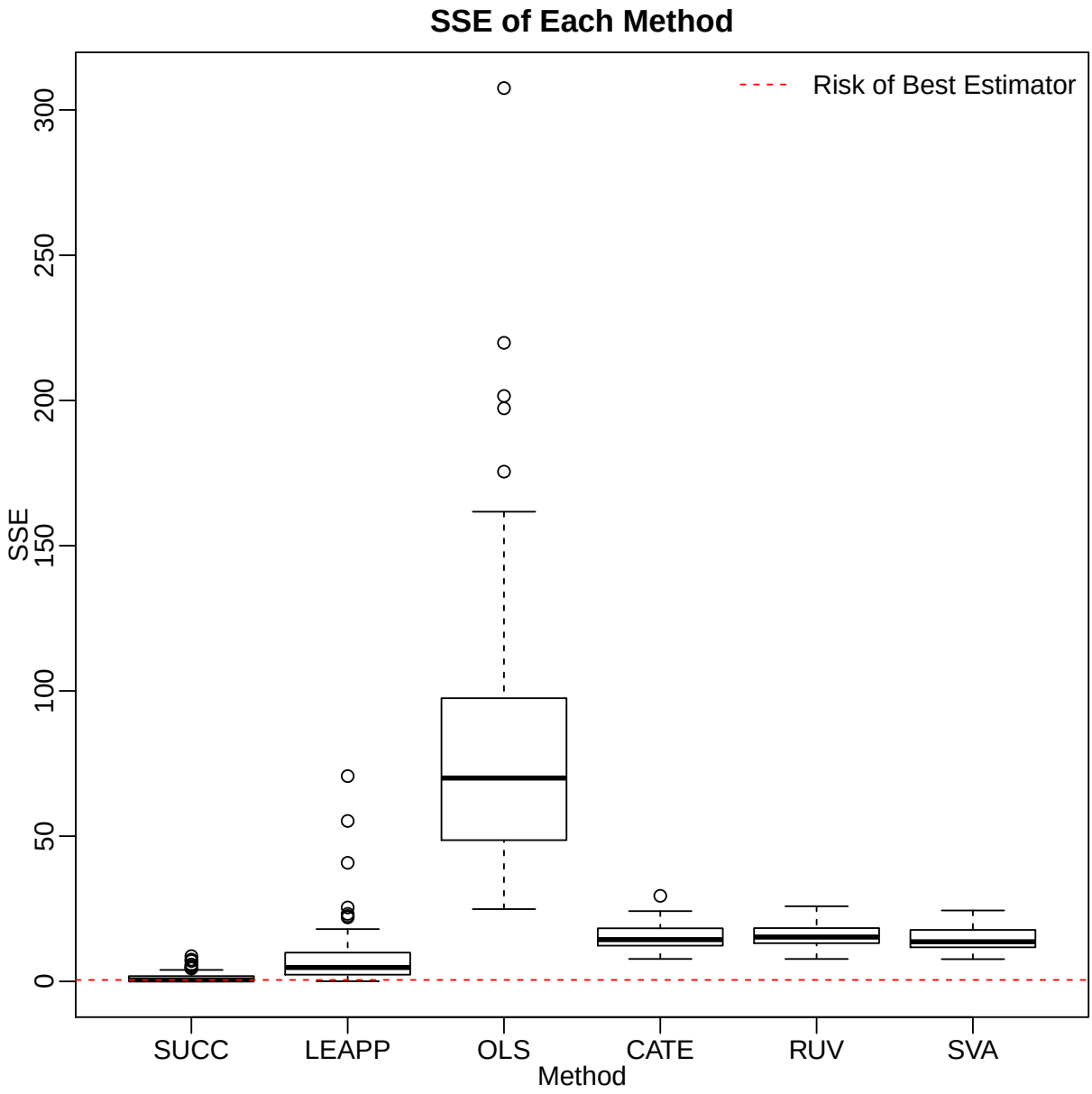
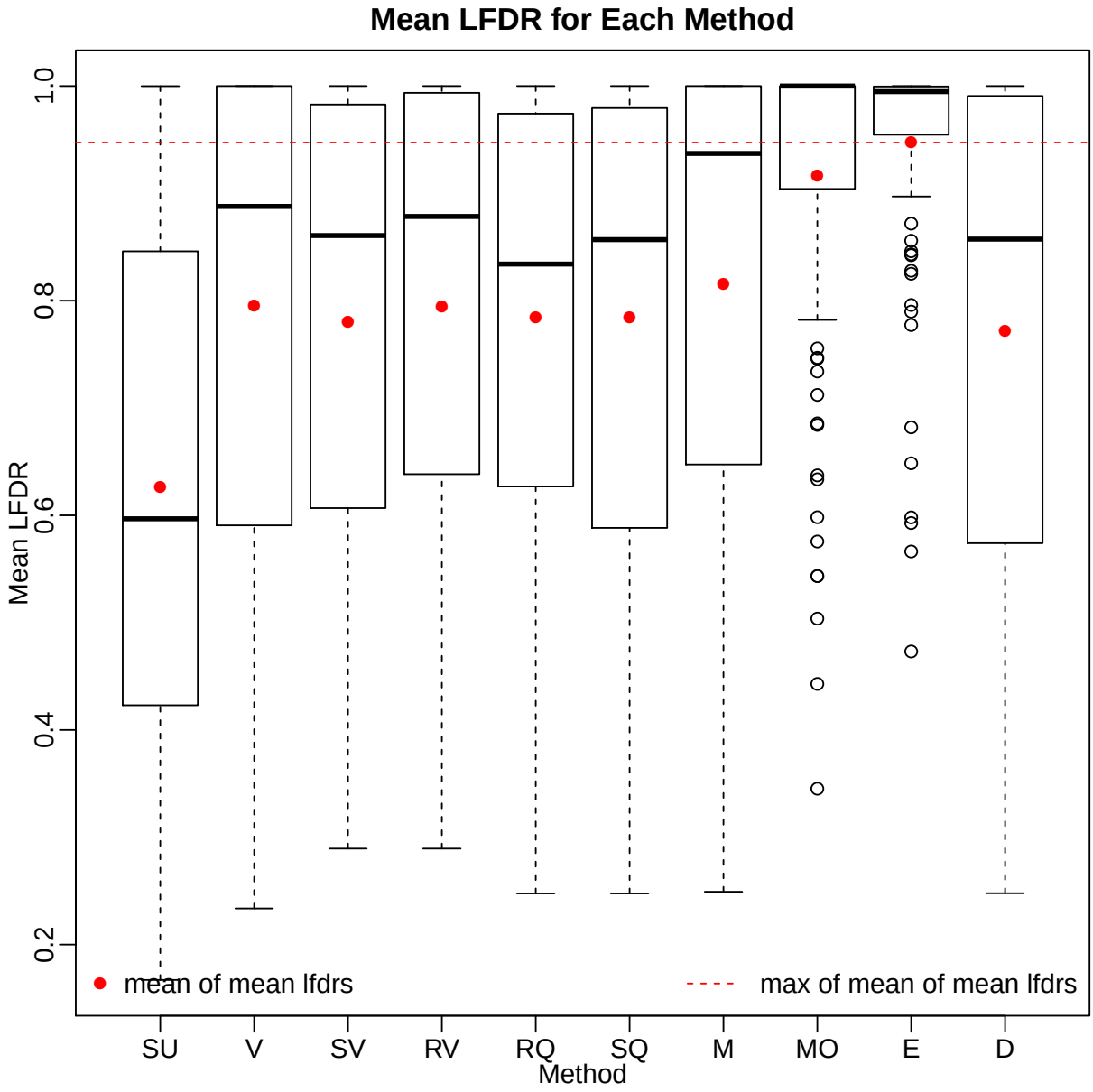


Figure 2: Mean lfdr for SUCCOTASH (SU), VOOM (V), SVA and VOOM (SV), RUV and VOOM (RV), RUV and quasi-binomial glm (RQ), SVA and quasi-binomial glm (SQ), Myrna (M), Myrna offset (MO), EDGER (E), DESEQ2 (D).



References

Jushan Bai, Kunpeng Li, et al. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465, 2012.

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

J Gagnon-Bartsch, L Jacob, and TP Speed. Removing unwanted variation from high dimensional data with negative controls. Technical report, Technical Report 820, Department of Statistics, University of California, Berkeley, 2013.

Ben Langmead, Kasper D Hansen, Jeffrey T Leek, et al. Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome Biol*, 11(8):R83, 2010.

Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.

Jeffrey T Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research*, page gku864, 2014.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–1735, 2007.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014.

Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

Yunting Sun, Nancy R Zhang, Art B Owen, et al. Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics*, 6(4):1664–1688, 2012.

Jingshu Wang, Qingyuan Zhao, Trevor Hastie, and Art B Owen. Confounder adjustment in multiple hypotheses testing. *arXiv preprint arXiv:1508.04178*, 2015.