SUCCOTASH vs Methods in Mengyin's Code when All Null

David Gerard

January 13, 2016

Abstract

Here, I compare SUCCOTASH to the various methods that Mengyin coded up when no genes are associated with the covariate of interest. I do this under an idealized situation where the data are generated according to the SUCCOTASH model.

1 Data Generation

$$\log_2(Y_{n\times p} + 1) = X_{n\times 2}\beta_{2\times p} + Z_{n\times k}\alpha_{k\times p} + E_{n\times p} \tag{1}$$

such that

- $X_{[,1]} = 1$, $X_{[1:(n/2),2]} = 0$, $X_{[(n/2+1):n,2]} = 1$,
- $\beta_{[1,]} = (10, \ldots, 10)$,
- $\beta_{[2,]} = (0, \ldots, 0)$,
- $Z_{ij} \overset{iid}{\sim} N(0, 1)$,
- $\alpha_{ij} \overset{iid}{\sim} N(0, 1)$,
- $E_{ij} \overset{iid}{\sim} N(0, 1)$,

The values of Y that were less than 0 were set to 0. All other values were rounded to their nearest integer values.

2 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 quasi-binomial GLM methods in Mengyin's code were having trouble converging with this data generating process, so I removed them from consideration. The methods I did use 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.
- EdgeR [Robinson et al., 2010].

I also compared the estimation performace 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 known.
- RUV4 [Gagnon-Bartsch et al., 2013] with 50% of the observations being control genes with the number of confounders known.

The factor analysis part of SUCCOTASH was done with the quasi-mle approach of Bai et al. [2012] with the number of hidden confounders known.

3 Simulation Study

I ran through 100 repetitions of generating the data as in Section 1 with $n = 100$, $p = 1000$, and $k = 25$. For each iteration, I compared the sum of squared errors (SSE) between SUCCOTASH, LEAPP, CATE, SVA, and RUV4. I did not compare SSE with VOOM, RUVseq + VOOM, SVAseq + VOOM, and EdgeR because I think there are different normalizations than $\log_2(Y_{ij}+1)$, so they shouldn't really be comparable. For example, VOOM uses log-counts per million:

$$\log_2\left(\frac{Y_{ij}+0.5}{\sum_{j=1}^{p} Y_{ij}+1} * 10^6\right). \tag{2}$$

Rather, I compared the local false discovery rates (lfdr's) of SUCCOTASH, VOOM, RUVseq + VOOM, SVAseq + VOOM, and EdgeR. Since lfdr is the posterior probability of 0, the higher the lfdr's the better since we are in the null case where all effects are 0. I compared the 25th, 50th, and 75th percentiles of the lfdr's for each estimator. I also looked at the mean lfdr of each estimator.

4 Results

SUCCOTASH performed better than SVA, CATE, RUV, and LEAPP in terms of SSE (Figure 1). I made an implementation mistake for SVA, so I didn't include it in Figure 1.

SUCCOTASH performs worse than all of the VOOM methods in terms of lfdr (Figure 2). But it performs much better than EDGR.

References

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

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.

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.

Figure 1: Sum of Squared Errors (SSE) for SUCCOTASH (SUCC), LEAPP, CATE, and RUV.

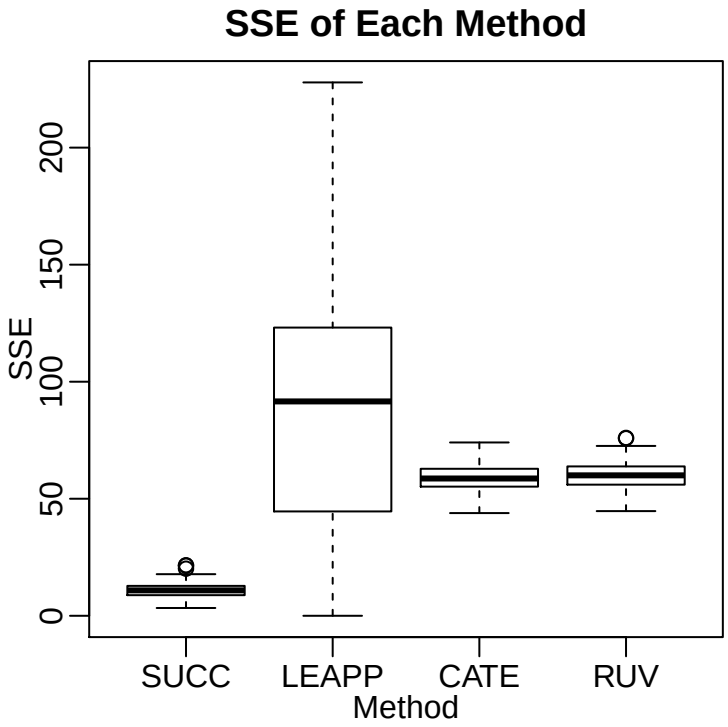
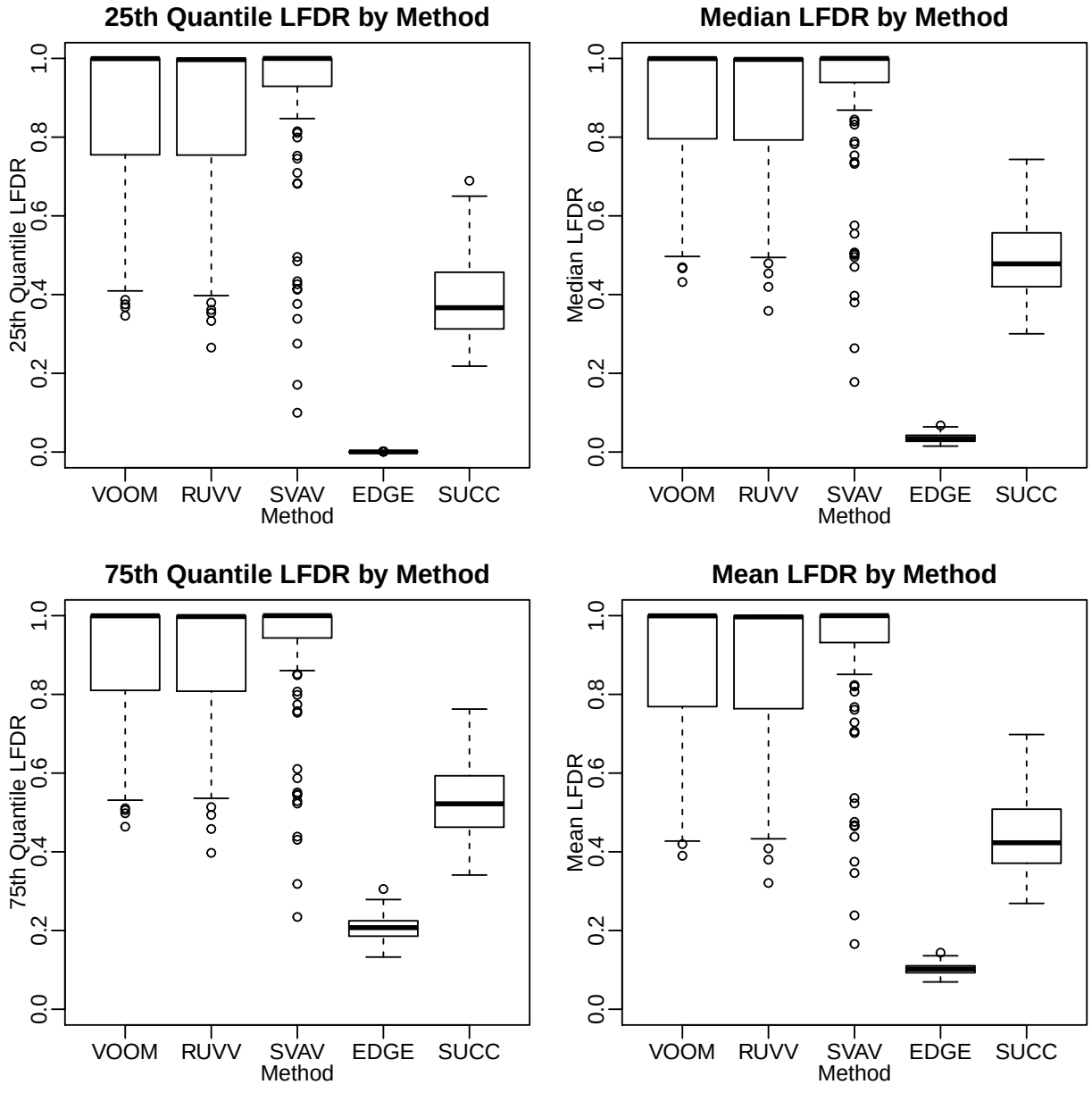


Figure 2: Local false discovery rate (lfdr) summaries for VOOM, RUV and VOOM (RUVV), SVA and VOOM (SVAV), EDGER, and SUCCOTASH (SUCC). Plotted are the 25th, 50th, and 75th quantiles of the lfdr's over the 100 iterations, and the mean lfdr.



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

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.

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.