

# SUCCOTASH vs Methods in Mengyin's Code when Nonnull

David Gerard

January 13, 2016

## Abstract

Here, I compare SUCCOTASH to the various methods that Mengyin coded up. I do this under an idealized situation where the data are generated according to the SUCCOTASH model. Here, I generate the data under a nonnull model.

## 1 Data Generation

$$W_{n \times p} = X_{n \times 2} \beta_{2 \times p} + Z_{n \times k} \alpha_{k \times p} + E_{n \times p} \quad (1)$$

such that

- $X_{[,1]} = 1$ ,  $X_{[1:(n/2),2]} = 0$ ,  $X_{[(n/2+1):n,2]} = 1$ ,
- $\beta_{[1,]} = (10, \dots, 10)$ ,
- $f(\beta_{[2,]}) = \prod_{i=1}^p \sum_{k=1}^K \pi_k N(\beta_{[2,i]} | 0, \tau_k^2)$ ,
- $Z_{ij} \stackrel{iid}{\sim} N(0, 1)$ ,
- $\alpha_{ij} \stackrel{iid}{\sim} N(0, 1)$ ,
- $E_{ij} \stackrel{iid}{\sim} N(0, 1)$ .

We then set

$$Y_{ij} = \frac{2^{W_{ij}} - 0.5}{\sum_{j=1}^p (2^{W_{ij}} - 0.5)} \times 10^7. \quad (2)$$

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 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 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

- $n = 100$ ,
- $p = 1000$ ,
- $k = 25$ ,
- $\pi = (0.5, 0.3, 0.2)$ ,
- $\tau = (0, 5, 10)$ .

For each iteration, I compared the mean squared errors (MSE) between SUCCOTASH, LEAPP, CATE, SVA, RUV4, VOOM, RUVseq + VOOM, SVASEq + VOOM, and EdgeR.

I also looked at the estimated  $\pi_0$ 's for SUCCOTASH, VOOM, RUVseq + VOOM, SVASEq + VOOM, and EdgeR. Note that for none of the ash-like methods do I assume that the grid is known.

Finally, I compared the true versus estimated local false discovery rates (lfdr's) of SUCCOTASH, VOOM, RUVseq + VOOM, SVASEq + VOOM, and EdgeR. Specifically, I calculated the sum of squared errors between the true and local false discovery rates for each method.

### 4 Results

SUCCOTASH performed the best in terms of MSE (Figure 1).

SUCCOTASH performs worse than RUV + VOOM and SVA + VOOM in terms of estimating  $\pi_0$  (Figure 3) and lfdr (Figure 2), but better than vanilla VOOM and EDGR. But RUV + VOOM and SVA + VOOM are both done under the unrealistic (I think?) scenario of having half of the genes being control genes (the version of SVA performed in Mengyin's code is fed control genes).

### References

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

Figure 1: Mean squared errors (MSE) for SUCCOTASH (SUCC), LEAPP, CATE, RUV, SVA, VOOM, RUV with VOOM (RUVV), SVA with VOOM (SVAV), and EDGER (EDGE).

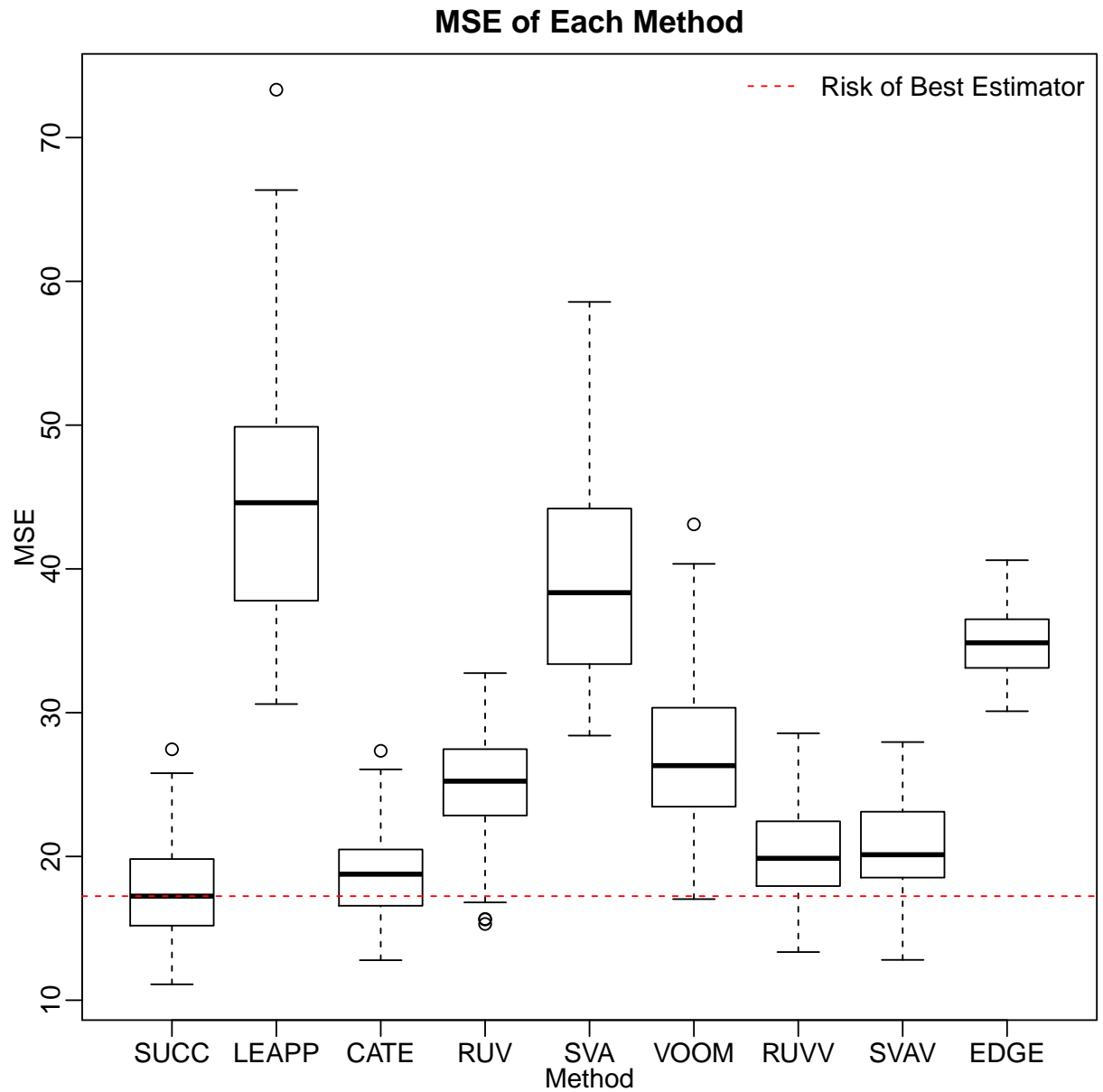


Figure 2: Sum of squared error s between true and estimated local false discovery rates (lfdr) for VOOM, RUV and VOOM (RUVV), SVA and VOOM (SVAV), EDGER (EDGE), and SUCCO-TASH (SUCC).

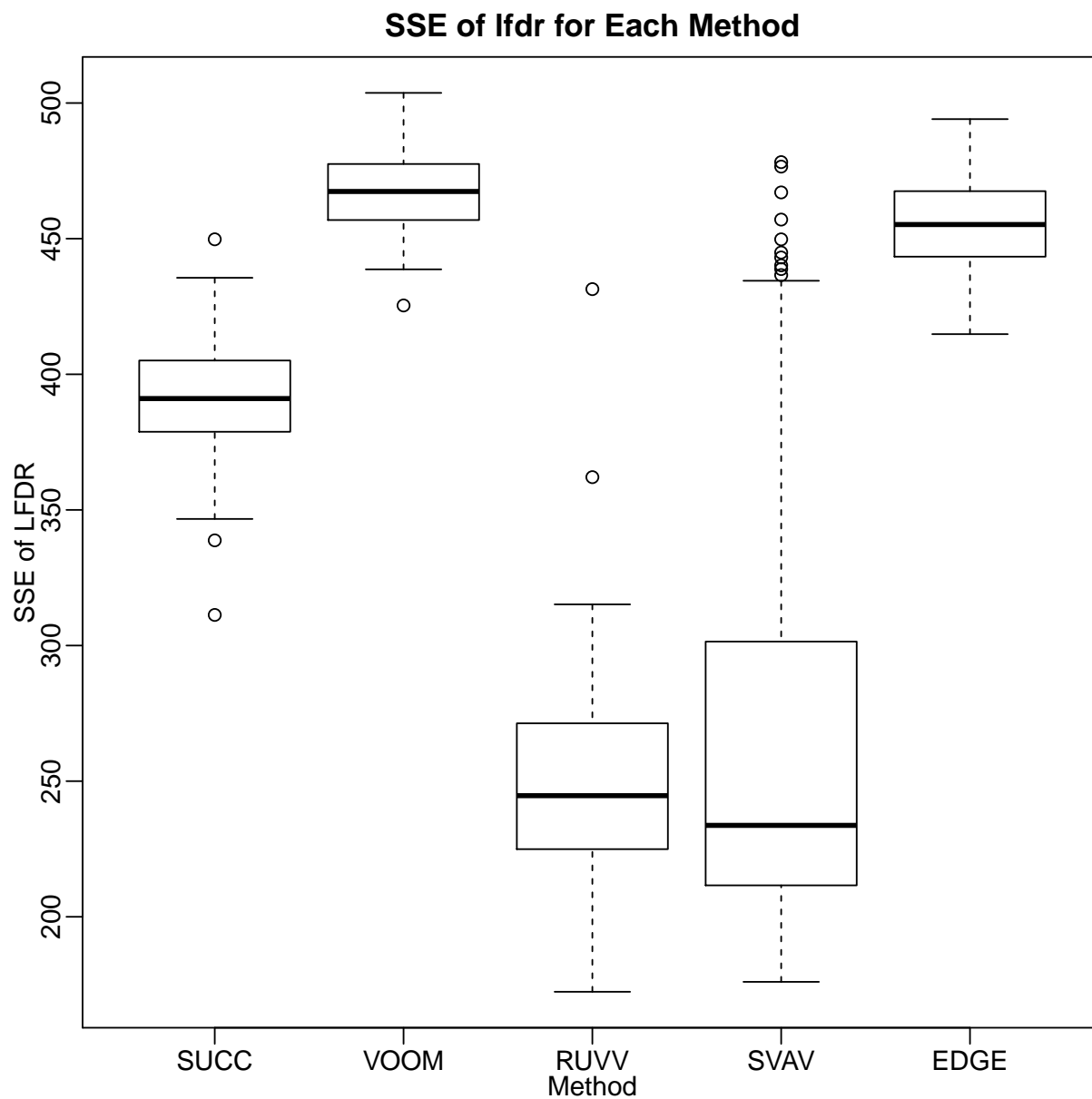
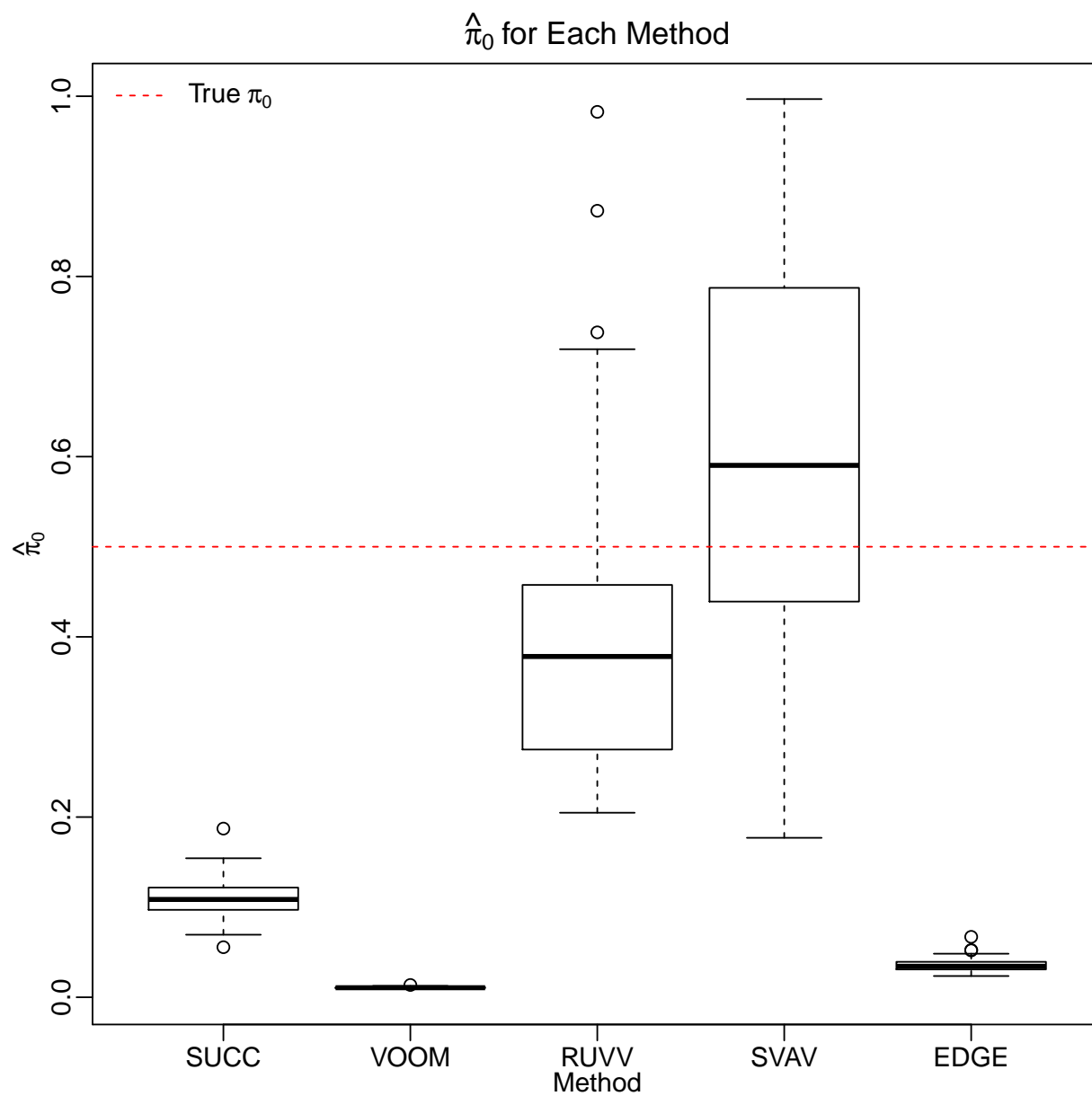


Figure 3: Estimates of  $\pi_0$  when  $\pi_0 = 0.5$  for VOOM, RUV and VOOM (RUVV), SVA and VOOM (SVAV), EDGER (EDGE), and SUCCOTASH (SUCC).



- 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.
- 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.