SUCCOTASH Sims when α and the covariance are known.

David Gerard

January 4, 2016

Abstract

This simulation study looks at the model assumed in the second step of SUCCOTASH. I compare SUCCOTASH with the second step of LEAPP. No other confounder adjustment procedure is applicable for comparison when assuming this model. SUCCOTASH outperforms LEAPP in terms of Sum of Squared Errors, but always overestimates π_0.

1 Model Description

$$Y_{p\times 1} = \beta_{p\times 1} + \alpha_{p\times k} Z_{k\times 1} + E_{p\times 1}, \tag{1}$$

such that

- α is known.
- $E \sim N_p(0, I_p)$, so we explore homoscedastic case.

2 Procedure

- $p = 100$,
- $k \in \{5, 10, 50\}$,
- $\beta_j \sim N(0, \tau_k^2)$ w.p. π_k,
- $\tau_k^2 = 0, 1, 4$ for $k = 0, 1, 2$,
- $\pi \in \{(1, 0, 0), (0.9, 0.1, 0), (0.9, 0, 0.1), (0.5, 0.5, 0), (0.5, 0, 0.5), (0.5, 0.25, 0.25)\}$,
- $Z_j \overset{i.i.d.}{\sim} N(0, 1)$,
- $\alpha_{ij} \overset{i.i.d.}{\sim} N(0, 1)$,
- 400 iterations for each π by k combination, sampling a new Z and α at each iteration.

- At each iteration, I calculated the Sum of Squared Errors (SSE) for the posterior means under SUCCOTASH, and the estimates of β given by the second step of LEAPP.
- I also calculated the SSE when using just Y to estimate β (called OLS in Figures and Tables below).
- I also calculated $\hat{\pi}_0$ given by SUCCOTASH and LEAPP at each iteration.
- LEAPP uses an L_1 penalty, so I called its $\hat{\pi}_0$ to just be the proportion of elements of β it sets to 0.

- The only comparable procedure using Model (1) is the second step of LEAPP [Sun et al., 2012].

- CATE [Wang et al., 2015] is not applicable because its model for its second step is

$$Y_{p\times 1} \sim N_p(\beta_{p\times 1} + \alpha_{p\times k}\gamma_{k\times 1}, \alpha\alpha^T + I_p), \tag{2}$$

  where γ describes the linear relationship between the observed and unobserved variables.
- RUV [Gagnon-Bartsch et al., 2013] is not applicable because we don't assume we have any control genes.
- SVA [Leek and Storey, 2008] is not applicable because it doesn't use this two-step procedure.

3 Results

SUCCOTASH always overestimates π_0 (Table 2). LEAPP overestimates π_0 when k is small and underestimates π_0 when k is large. In terms of Sum of Squared Errors (SSE), SUCCOTASH outperforms LEAPP in every scenario (Table 1), especially when k is large, where LEAPP performs extremely poorly, even worse than just using Y to estimate β.

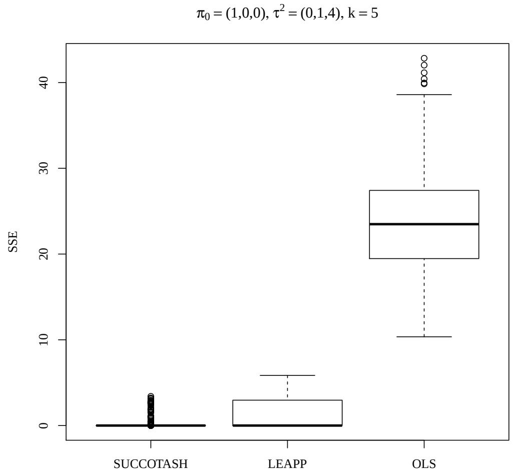
Table 1: Average Sum of Squared Errors for SUCCOTASH, LEAPP, and OLS at given k and π values.

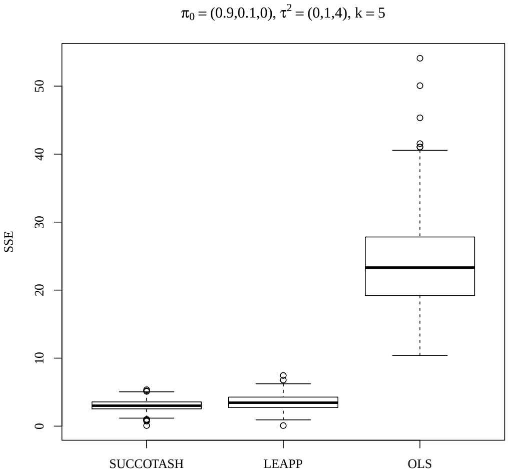
k	π_0	π_1	π_2	SUCCOTASH	LEAPP	OLS
5	1	0	0	0.2	0.9	23.7
5	.9	.1	0	3.0	3.5	23.9
5	.9	0	.1	4.4	4.9	24.0
5	.5	.5	0	6.3	7.2	23.4
5	.5	0	.5	8.3	10.7	23.7
5	.5	.25	.25	7.4	9.0	24.0
10	1	0	0	0.1	1.2	32.3
10	.9	.1	0	3.0	3.6	32.9
10	.9	0	.1	4.7	5.2	32.5
10	.5	.5	0	6.4	7.4	32.3
10	.5	0	.5	8.8	11.3	32.8
10	.5	.25	.25	7.7	9.3	32.2
50	1	0	0	0.1	98.9	71.0
50	.9	.1	0	3.1	99.4	70.8
50	.9	0	.1	5.9	98.4	70.6
50	.5	.5	0	7.3	98.9	71.1
50	.5	0	.5	14.6	98.8	71.0
50	.5	.25	.25	11.1	99.4	70.7

Table 2: Mean $\hat{\pi}_0$ for SUCCOTASH and LEAPP at given k and π values.

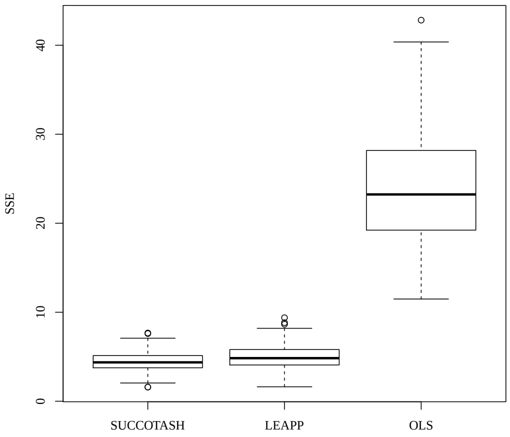
k	π_0	π_1	π_2	SUCCOTASH	LEAPP
5	1	0	0	1	1
5	.9	.1	0	.99	.99
5	.9	0	.1	.95	.98
5	.5	.5	0	.91	.98
5	.5	0	.5	.72	.91
5	.5	.25	.25	.81	.95
10	1	0	0	1	1
10	.9	.1	0	.99	.99
10	.9	0	.1	.96	.98
10	.5	.5	0	.93	.98
10	.5	0	.5	.74	.90
10	.5	.25	.25	.83	.94
50	1	0	0	1	.21
50	.9	.1	0	1	.21
50	.9	0	.1	.99	.21
50	.5	.5	0	1	.21
50	.5	0	.5	.94	.20
50	.5	.25	.25	.97	.20

4 SSE Boxplots

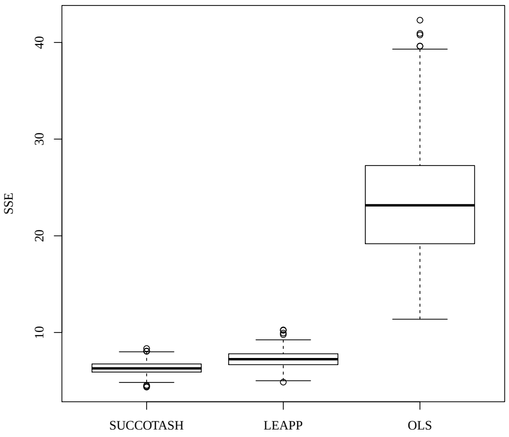




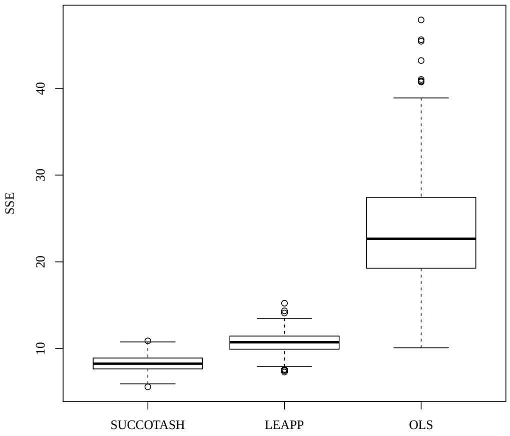
$\pi_0 = (0.9,0,0.1)$, $\tau^2 = (0,1,4)$, k = 5



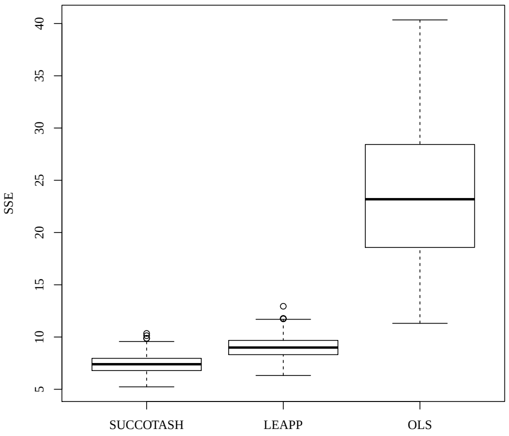
$\pi_0 = (0.5,0.5,0)$, $\tau^2 = (0,1,4)$, k = 5



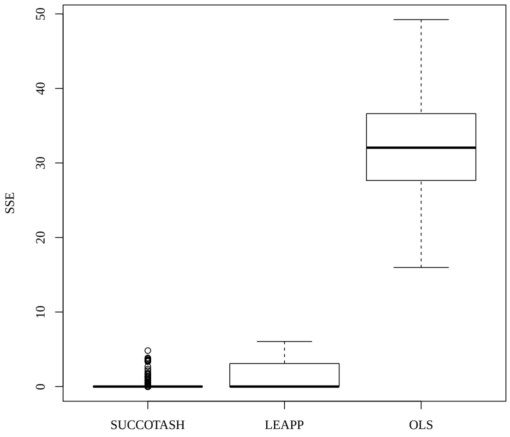
$\pi_0 = (0.5,0,0.5)$, $\tau^2 = (0,1,4)$, k = 5



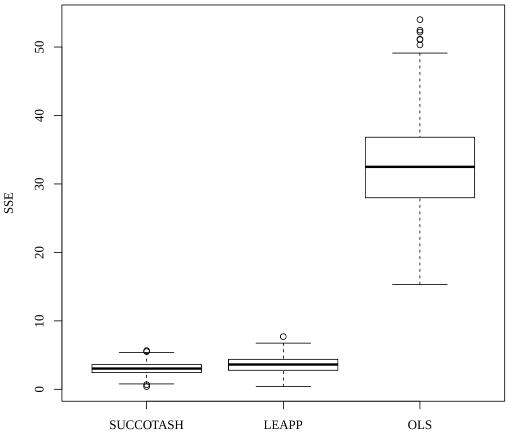
$\pi_0 = (0.5,0.25,0.25)$, $\tau^2 = (0,1,4)$, k = 5



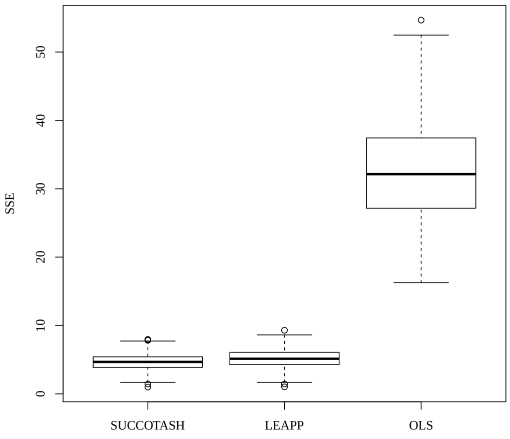
$\pi_0 = (1,0,0)$, $\tau^2 = (0,1,4)$, k = 10



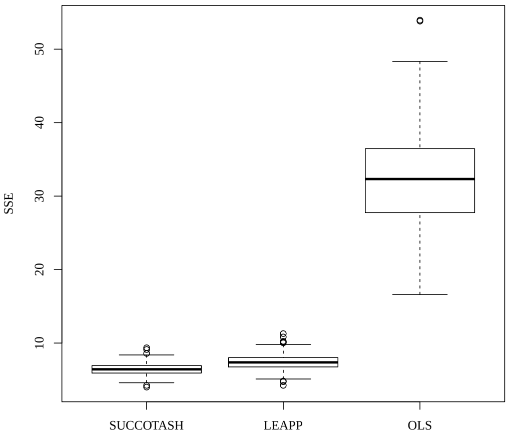
$\pi_0 = (0.9,0.1,0)$, $\tau^2 = (0,1,4)$, k = 10



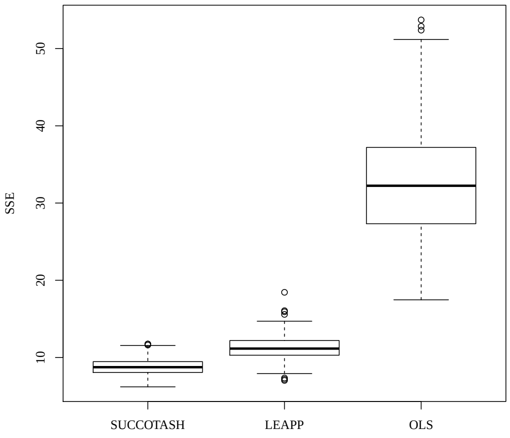
$\pi_0 = (0.9,0,0.1)$, $\tau^2 = (0,1,4)$, k = 10



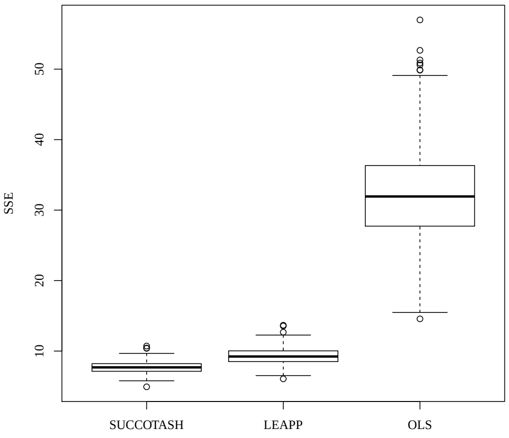
$\pi_0 = (0.5,0.5,0)$, $\tau^2 = (0,1,4)$, k = 10



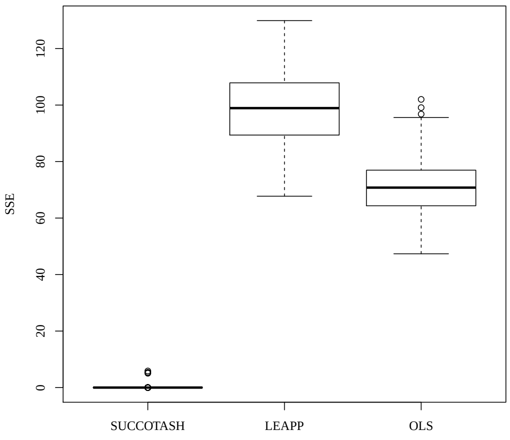
$\pi_0 = (0.5,0,0.5)$, $\tau^2 = (0,1,4)$, k = 10



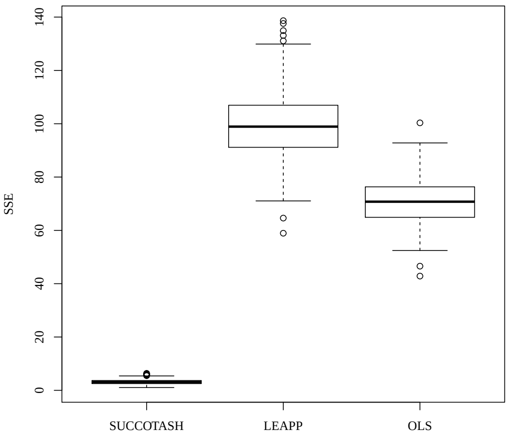
$\pi_0 = (0.5,0.25,0.25)$, $\tau^2 = (0,1,4)$, k = 10

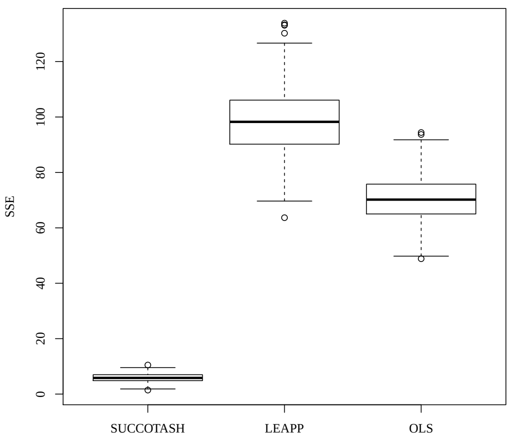


$\pi_0 = (1,0,0)$, $\tau^2 = (0,1,4)$, k = 50

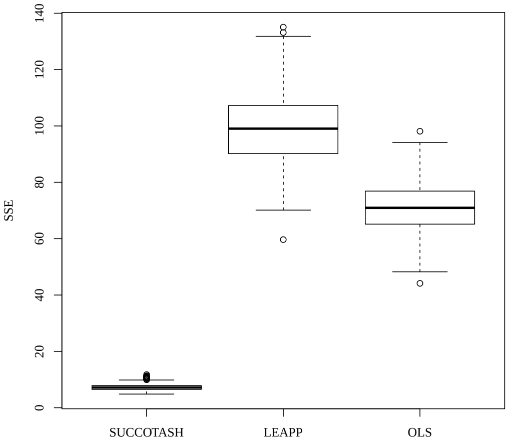


$\pi_0 = (0.9,0.1,0)$, $\tau^2 = (0,1,4)$, k = 50

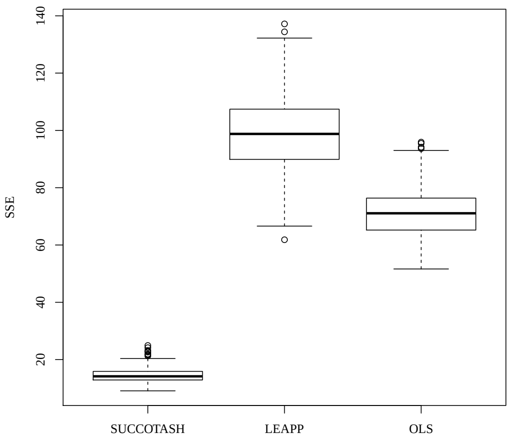




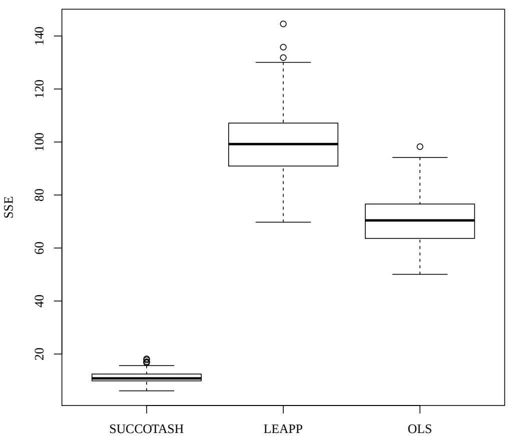
$\pi_0 = (0.9,0,0.1)$, $\tau^2 = (0,1,4)$, $k = 50$



$\pi_0 = (0.5,0.5,0)$, $\tau^2 = (0,1,4)$, $k = 50$

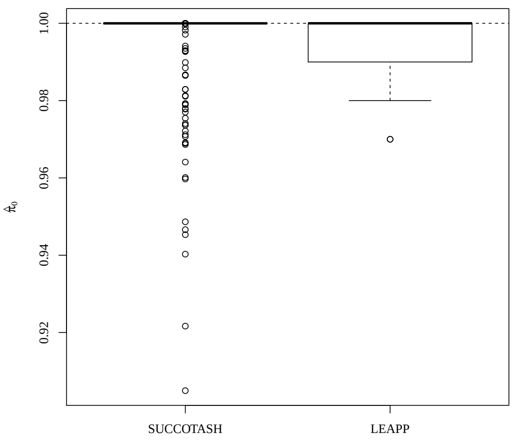


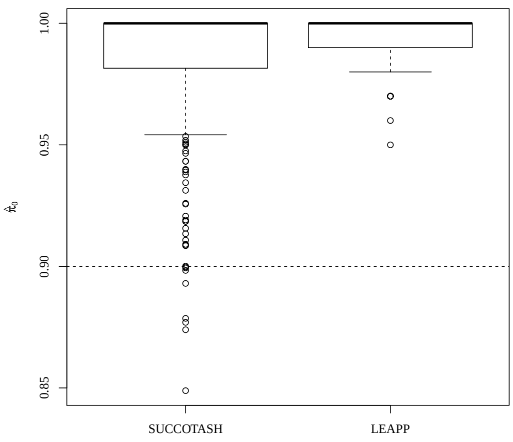
$\pi_0 = (0.5,0,0.5)$, $\tau^2 = (0,1,4)$, $k = 50$

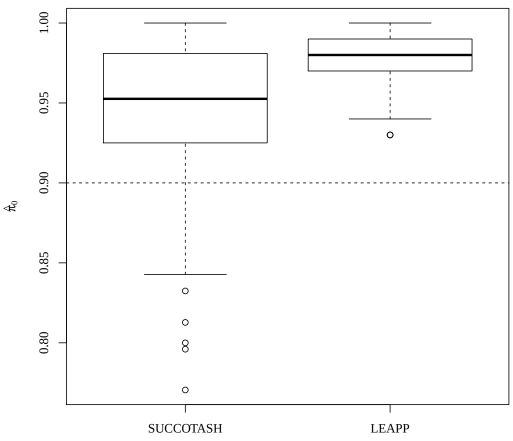


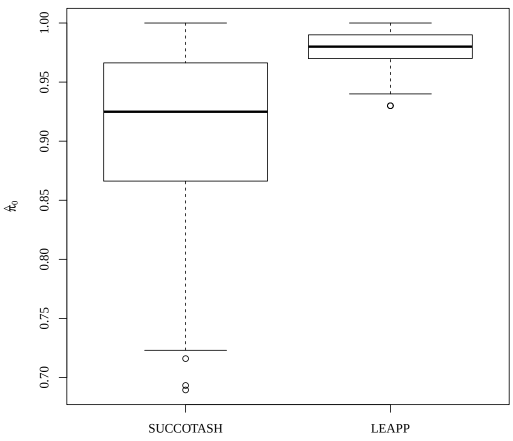
$\pi_0 = (0.5,0.25,0.25)$, $\tau^2 = (0,1,4)$, $k = 50$

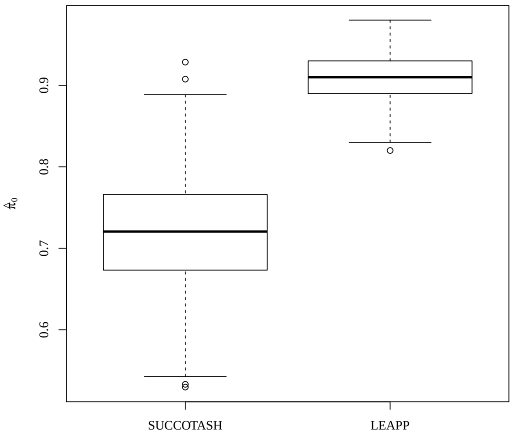
5 $\hat{\pi}_0$ Boxplots

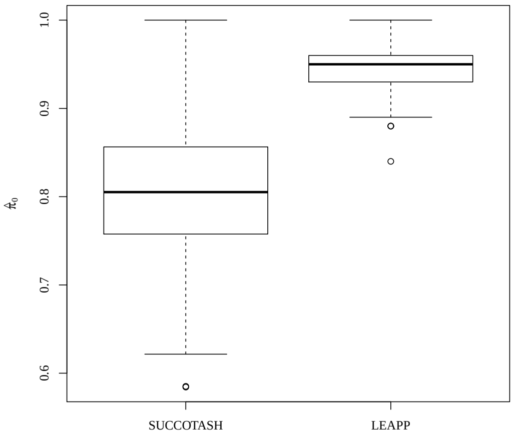

$\pi_0 = (1,0,0)$, $\tau^2 = (0,1,4)$, k = 5

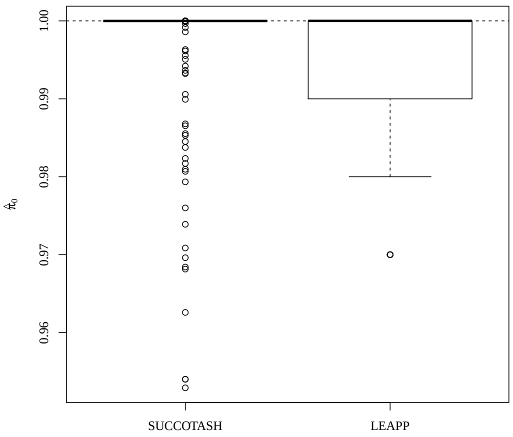

$\pi_0 = (0.9,0.1,0)$, $\tau^2 = (0,1,4)$, k = 5

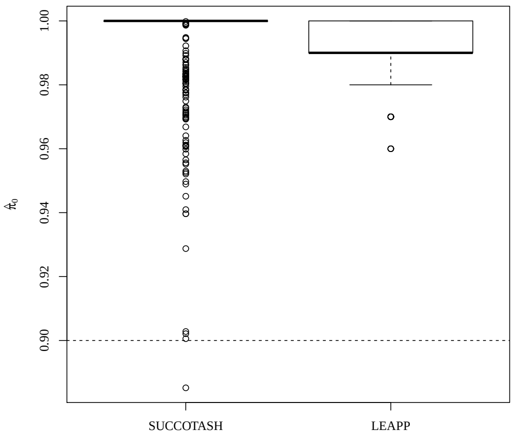

$\pi_0 = (0.9,0,0.1)$, $\tau^2 = (0,1,4)$, k = 5

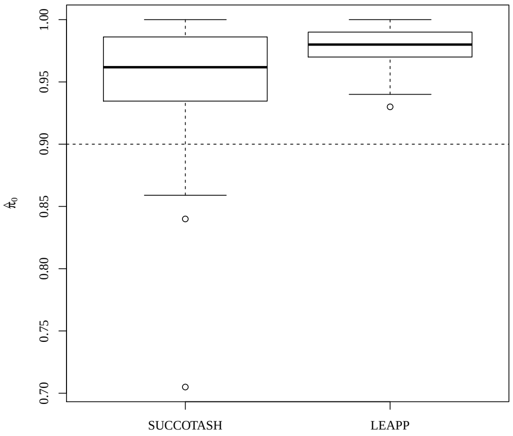

$\pi_0 = (0.5,0.5,0)$, $\tau^2 = (0,1,4)$, k = 5

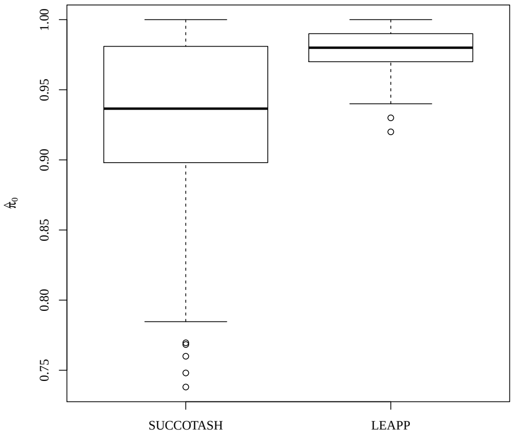

$\pi_0 = (0.5,0,0.5)$, $\tau^2 = (0,1,4)$, k = 5


$\pi_0 = (0.5,0.25,0.25)$, $\tau^2 = (0,1,4)$, k = 5

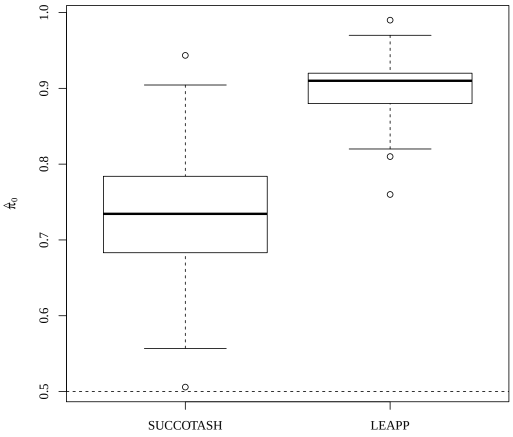

$\pi_0 = (1,0,0)$, $\tau^2 = (0,1,4)$, k = 10


$\pi_0 = (0.9,0.1,0)$, $\tau^2 = (0,1,4)$, k = 10

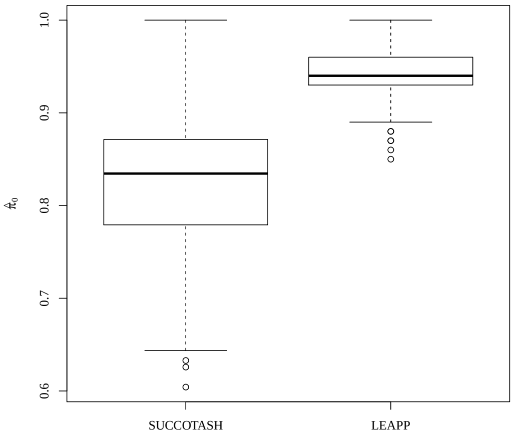

$\pi_0 = (0.9,0,0.1)$, $\tau^2 = (0,1,4)$, k = 10


$\pi_0 = (0.5,0.5,0)$, $\tau^2 = (0,1,4)$, k = 10

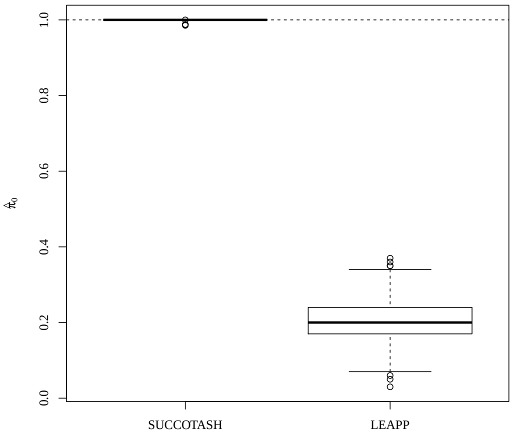
$\pi_0 = (0.5,0,0.5)$, $\tau^2 = (0,1,4)$, k = 10



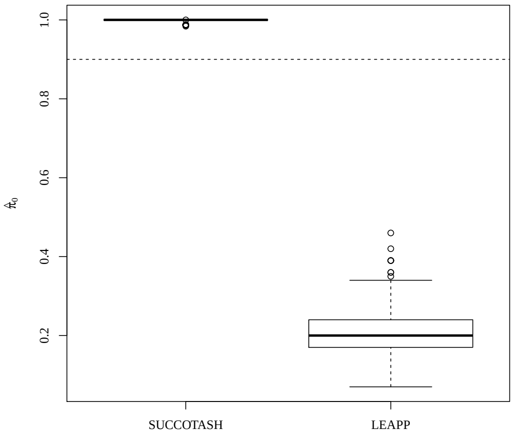
$\pi_0 = (0.5,0.25,0.25)$, $\tau^2 = (0,1,4)$, k = 10



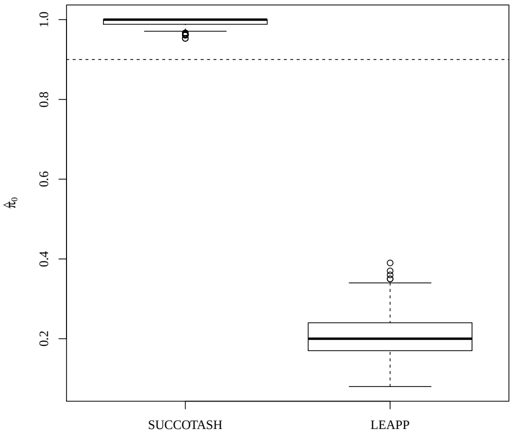
$\pi_0 = (1,0,0)$, $\tau^2 = (0,1,4)$, k = 50



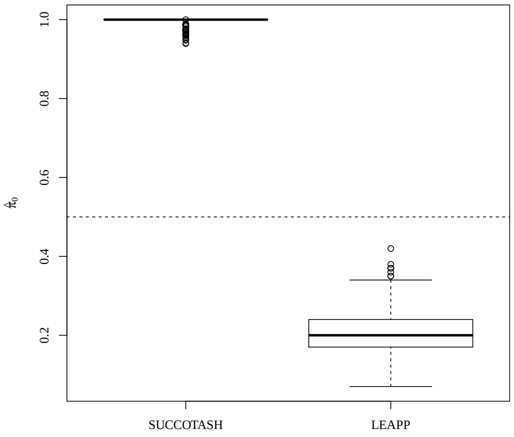
$\pi_0 = (0.9,0.1,0)$, $\tau^2 = (0,1,4)$, k = 50

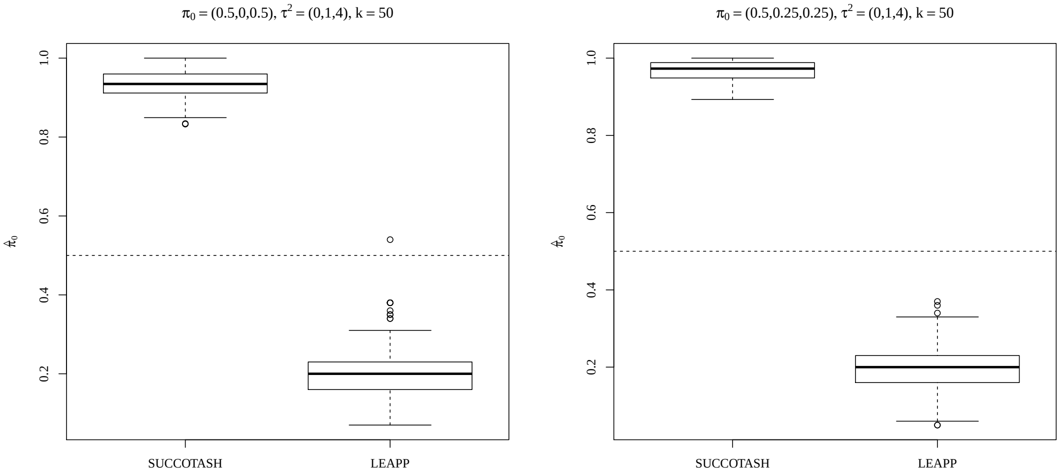


$\pi_0 = (0.9,0,0.1)$, $\tau^2 = (0,1,4)$, k = 50



$\pi_0 = (0.5,0.5,0)$, $\tau^2 = (0,1,4)$, k = 50





References

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.

Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.

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.