Cell-type specific mediation project update

Andrea Lane

2/24/2020

Outline

1. Project idea
2. EM initial value comparison
3. Bootstrap results using true values as initial values
4. Next steps

Project idea reminder

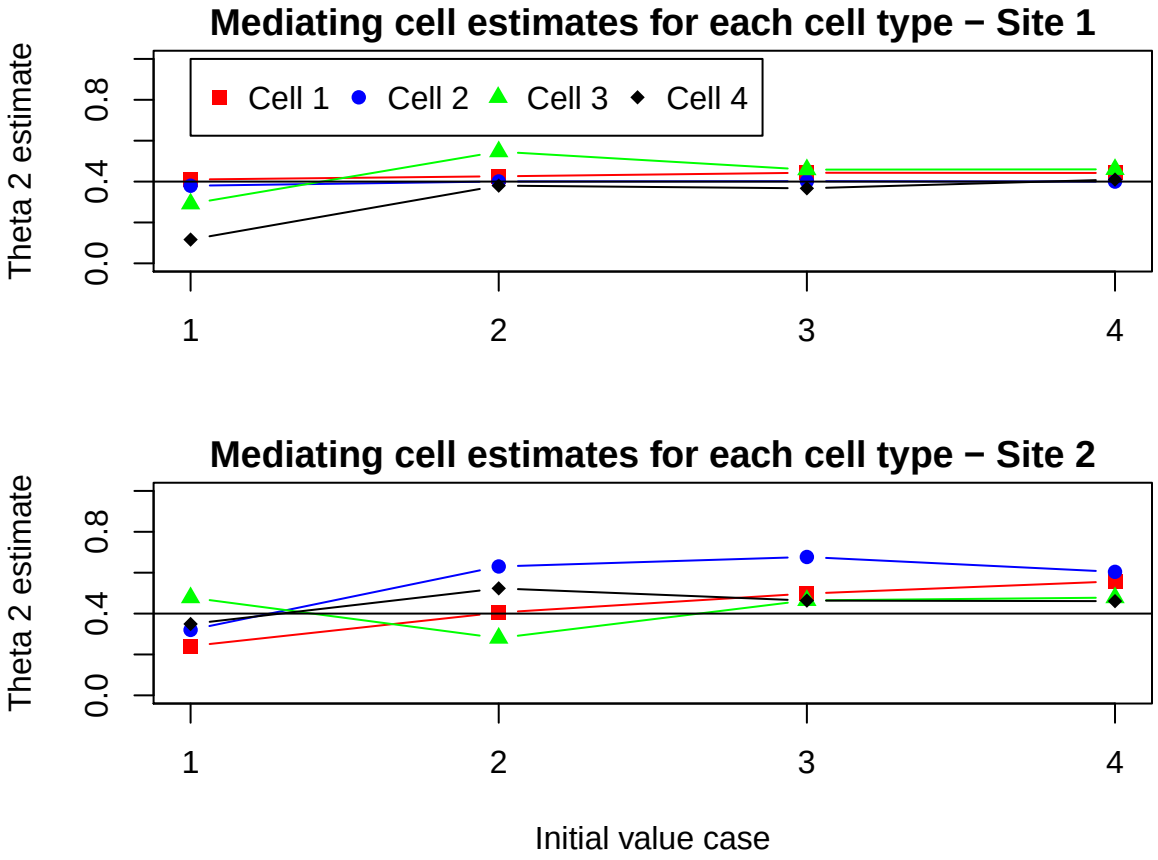
$$
\begin{aligned}
Y &\sim N(\theta_0 + \theta_1 E + \sum_k \theta_2^k m_k,\ \gamma^2) \\
M &\sim N(\sum_k m_k \pi_k,\ \sigma^2) \\
m_k &\sim N(\beta_0^k + \beta_1^k E,\ \tau_k^2)
\end{aligned}
$$

- Use TOAST to get β initial values and reduce number of sites to go through EM
- Run EM algorithm one site at a time to get estimates of Θ
- Calculate observed indirect effect $\hat{\beta}_1 \hat{\theta}_2$
- Use percentile bootstrap to assess significance of the indirect effect for each site and cell type

EM initial values

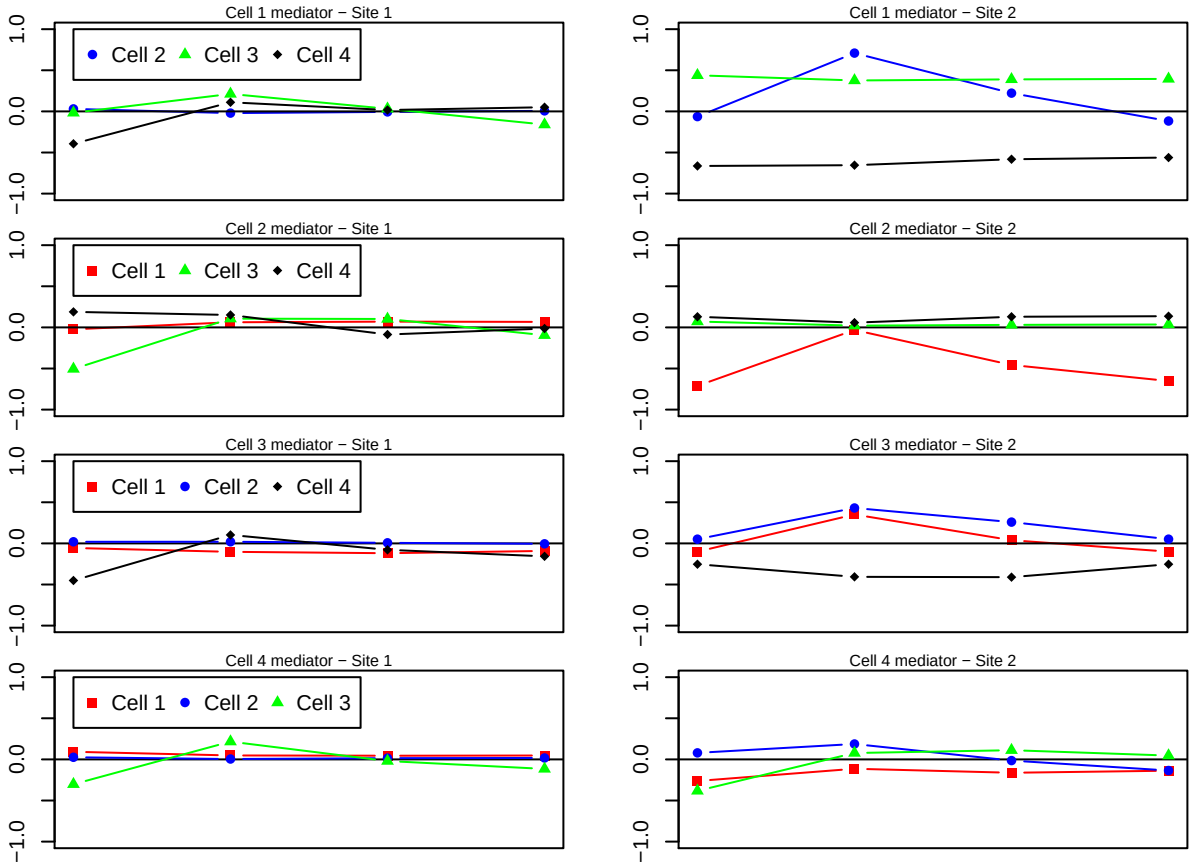
To evaluate the estimates obtained from EM depending on the initial values, I varied the mediating cell type and looked at 2 different mediating CpG sites. I used 4 different initial value cases (note that the true value for the mediating $\theta_2^k = 0.4$):

- Case 1: 0 for all θ_2^k
- Case 2: 0.2 for all θ_2^k
- Case 3: 0.1 for non-mediating cell types, 0.3 for mediating cell type
- Case 4: 0 for non-mediating cell types, 0.4 for mediating cell type (true values)



Non-mediating cell type estimates (true value 0 in all cases)

- Case 1: 0 for all θ_2^k
- Case 2: 0.2 for all θ_2^k
- Case 3: 0.1 for non-mediating cell types, 0.3 for mediating cell type
- Case 4: 0 for non-mediating cell types, 0.4 for mediating cell type (true values)



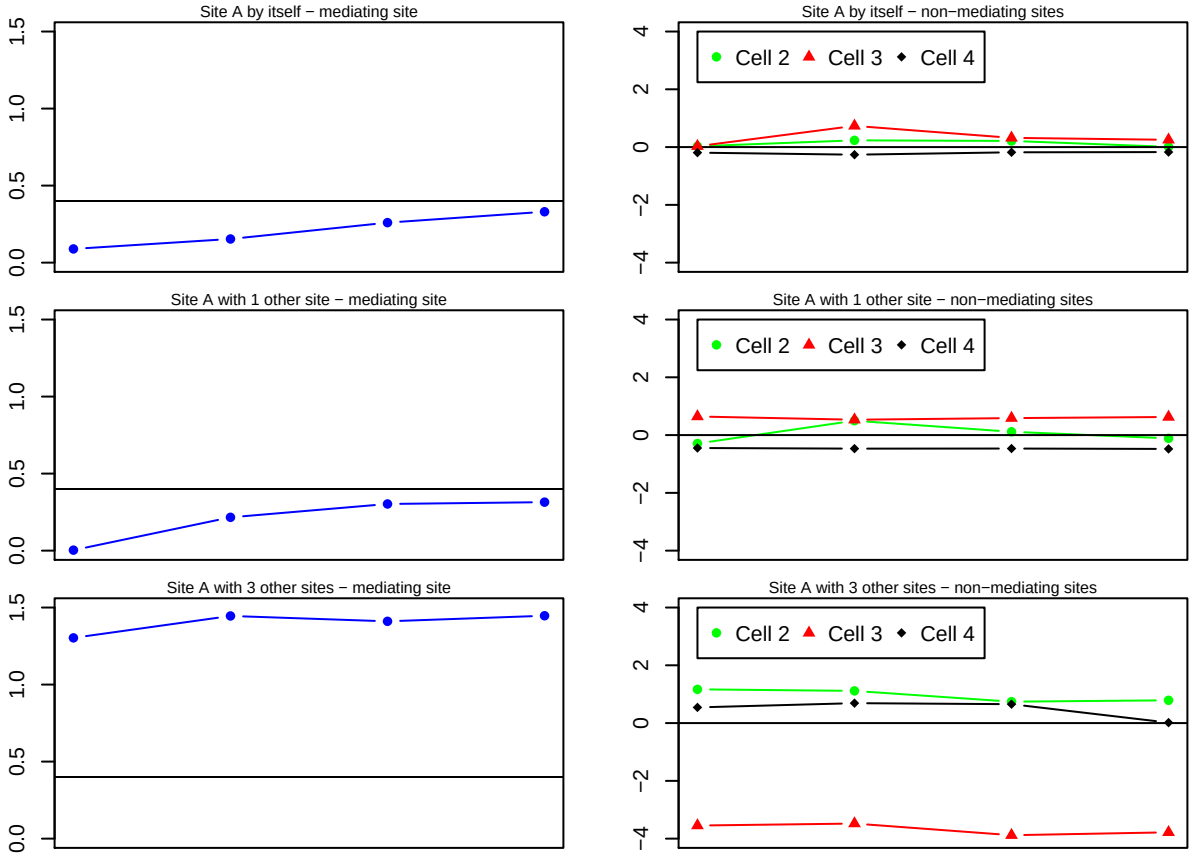
Comments:

- for the most part, it seems case 3 and case 4 results are pretty similar, which is good.
- I don't understand what causes differences between sites? For example in the top row of the non-mediator plots, the two sites are quite different. **See next page**

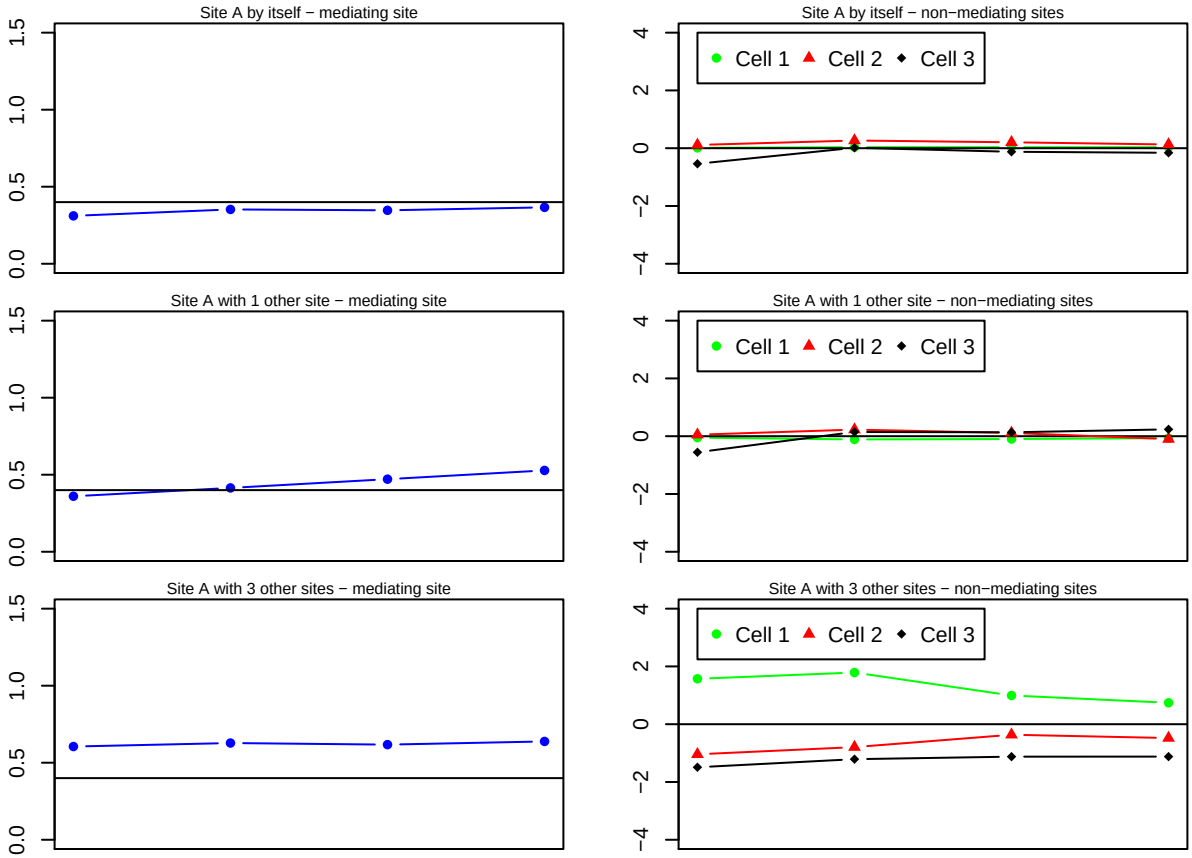
The difference in the results for the sites comes from the number of mediating sites. The outcome Y is generated by a linear combination of all of the mediating sites, even though EM is used for each site individually. Therefore, the relationship between the mediator and the outcome becomes weaker as the number of mediating sites increases.

The plots below show the estimates from EM for the mediating and non-mediating cell types when site A (site 2 from previous page) is the only mediating site, when it is one of two mediating sites, and one of four mediating sites.

Here, cell type 1 is the mediating cell type. Cell type 1 has the lowest proportion so this is sort of a "worst case scenario."



This set of plots show the estimates when the mediating cell type is cell type 4, which has the largest proportion. As expected, performance is better when the cell type has a larger proportion.



Results in table form, if that's preferable:

Table 1: Theta2 estimates

Site 1				Site 2			
Cell 1	Cell 2	Cell 3	Cell 4	Cell 1	Cell 2	Cell 3	Cell 4
Initial value case 1: 0 for all theta2							
0.41	0.03	-0.02	-0.39	**0.24**	-0.06	0.44	-0.66
-0.02	**0.38**	-0.51	0.19	-0.71	**0.32**	0.07	0.13
-0.06	0.02	**0.29**	-0.45	-0.10	0.05	**0.48**	-0.25
0.09	0.03	-0.30	**0.12**	-0.26	0.08	-0.38	**0.35**
case 2: 0.2 for all theta2							
0.43	-0.02	0.21	0.11	**0.40**	0.71	0.38	-0.65
0.06	**0.40**	0.11	0.15	-0.03	**0.63**	0.02	0.06
-0.10	0.02	**0.55**	0.10	0.35	0.43	**0.28**	-0.41
0.05	0.00	0.22	**0.38**	-0.11	0.19	0.08	**0.52**
case 3: 0.1 non-med, 0.3 med							
0.44	-0.01	0.03	0.02	**0.50**	0.22	0.39	-0.58
0.07	**0.40**	0.10	-0.09	-0.46	**0.68**	0.03	0.13
-0.12	0.01	**0.46**	-0.08	0.04	0.26	**0.46**	-0.41
0.04	0.02	-0.02	**0.37**	-0.16	-0.01	0.11	**0.46**
case 4: 0 non-med, 0.4 med (true values)							
0.44	0.01	-0.16	0.05	**0.56**	-0.12	0.40	-0.56
0.07	**0.40**	-0.10	-0.01	-0.65	**0.60**	0.04	0.14
-0.09	-0.01	**0.46**	-0.16	-0.10	0.05	**0.48**	-0.25
0.05	0.02	-0.12	**0.41**	-0.14	-0.14	0.05	**0.46**

Bootstrap results

To assess the EM/boostrap procedure, I ran it with only the mediating sites using the true values as initial values.

Settings/Notes:

- N=500
- To account for multiple testing, the (.025/Ncell∗Nsite,1-(.025/Ncell∗Nsite)) quantile is used for bootstrap samples.
- for EM, max.iter=500 and tol=0.001
- EM initial values:
  - beta: from TOAST
  - theta: true values
  - tau, sigma: sampled from inverse gamma distribution with mean 0.001 (based on HIRE, which states that variance initial values are sampled from inverse gamma distribution with small mean but does not specify the mean)
  - gamma: true value 0.0004
- Scripts used: "Lane_Mediation_sim12821_cluster.R" and "mediation_sim_functions_1120.R" (functions script contains EM julia code)

Tables 2 and 3 show results from 1 and 2 sites with mediating effect sizes 0.2 and 0.4. Table 4 summarizes the results from tables 2 and 3 by averaging the non-mediating sites (type 1 error) and the mediating sites (power).

Tables 5 and 6 show results for one CpG site when increasing the number of bootstrap samples from 200 to 1000.

Table 7 summarizes the results from my last report for comparison. Here the initial values were all 0 rather than the true values.

Table 2: 1 site - 200 bootstrap

Theta = 0.2				Theta = 0.4			
Cell 1	Cell 2	Cell 3	Cell 4	Cell 1	Cell 2	Cell 3	Cell 4
317	78	167	194	**874**	94	162	94
315	**678**	115	68	158	**998**	79	37
80	24	**942**	65	23	13	**998**	15
57	14	42	**984**	19	7	13	**999**

Table 3: 2 sites - 200 bootstrap

Theta = 0.2				Theta = 0.4			
Cell 1	Cell 2	Cell 3	Cell 4	Cell 1	Cell 2	Cell 3	Cell 4
358	112	209	213	**895**	71	163	65
317	2	2	0	**790**	3	3	0
320	**710**	165	77	352	**995**	159	79
4	**277**	1	2	7	**711**	2	2
160	52	**896**	51	40	16	**990**	79
83	37	**886**	14	45	31	**997**	12
51	12	21	**831**	19	8	52	**987**
155	47	36	**999**	101	33	37	**1000**

Table 4: Average across mediating and nonmediating sites/cell types - 200 bootstrap

	1 site	2 sites	1 site	2 sites
	Theta = 0.2		Theta = 0.4	
Power	0.730	0.659	0.967	0.921
Type 1 error	0.102	0.076	0.060	0.057

Table 5: 1 site - 1000 bootstrap

Theta = 0.2				Theta = 0.4			
Cell 1	Cell 2	Cell 3	Cell 4	Cell 1	Cell 2	Cell 3	Cell 4
307	65	133	166	**855**	73	132	82
273	**659**	104	47	113	**998**	51	28
61	15	**934**	50	20	6	**999**	12
35	14	36	**981**	9	4	11	**999**

Table 6: Average across mediating and nonmediating sites/cell types - 1000 bootstrap

	Theta = 0.2	Theta = 0.4
Power	0.720	0.963
Type 1 error	0.083	0.045

Table 7: Average across mediating and nonmediating sites/cell types - 0s as initial values (theta=0.4)

	1 site	2 sites	3 sites
Power	0.691	0.696	0.692
Type 1 error	0.308	0.160	0.155

Next steps

- Find a way to get initial values (i.e. try TOAST M ~ Y)
- Increase variance
- Use bootstrap with TCA estimates
- Add covariate(s)