Hanes Hall, B44
Chapel Hill, NC 27510
(919) 265-9114

February 10, 2016

Johns Hopkins University
Bloomberg School of Public Health
Office E3624
615 North Wolfe Street
Baltimore, MD 21205

Dear Prof Leek,

My names is Meilei Jiang. I am a third year Phd student in statistics at UNC Chapel Hill, currently working with Steve Marron. Recently I am doing research on adjusting for batch effects in high dimensional data. I have read your papers and I quite like your approach, Surrogate Variable Analysis(SVA), as a method for batch correction and for addressing the data heterogeneity in the context of multiple testing dependence. However, I am quite puzzled by one aspect of SVA and wonder whether you can explain the reason behind your choice of posterior probability as the basis of the methods. There appreas to be a simpler and better choice of posterior probabilityn, so I wonder if I am missing something.

In particular, I think it seems the performance of SVA can be improved by modifying one step: weighting the residual matrix R by the probability $\Pr(\gamma_i \neq \vec{0}|X, S, \hat{G})$ instead of weighting the data matrix X by the probability $\Pr(b_i = \vec{0}\ \&\ \gamma_i \neq \vec{0}|X, S, \hat{G})$. I am wondering whether you have considered this and rejected it for some reason?

Based on a simple simulation study, this new approach to SVA improves the performance of IRW-SVA in the case that there exist no genes (variables) which are strongly associated with unmeasured confounders but are not associated with measured factors. The two approaches seem to have similar performance when such a subset exists.

The simulation study is conducted using your package 'sva' in the R software. Details are in the attachment.

Thank you for your time and consideration.

I look forward to your reply.

Sincerely,

Meilei Jiang