Simulation study: Modified SVA versus SVA

Meilei Jiang
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

February 2, 2016

1 Modified SVA

In your papers [2, 3, 1], you proposed a factor model for the relationship between expression values, measured biological factors and unmeasured biological and non-biological factors:

$$X = BS + \Gamma G + U$$

In order to remove the batch effects, you proposed SVA on the data set to estimate G. An essential idea of SVA is to identify a subset of genes (tests) which are strongly associated with unmeasured confounders, but not with the group outcome. Especially, an empirical bayesian approach has been applied to estimate the probabilities

$$\begin{aligned}\pi_{iw} &= \Pr(b_i = \vec{0} \ \& \ \gamma_i \neq \vec{0} | X, S, \hat{G}) \\ &= \Pr(b_i = \vec{0} | \gamma_i \neq \vec{0}, X, S, \hat{G}) \Pr(\gamma_i \neq \vec{0} | X, S, \hat{G})\end{aligned}$$

Then use π_{iw} to weight the ith row of X and perform a singular value decomposition of the weighted X to reconstruct $\hat{G}$.

However, if we estimate $\hat{G}$ in this approach, IRW-SVA, I think it assumes there exists such subset of genes (tests) in the data set. When such subset of genes (tests) does not exist, IRW-SVA could fail. I think there is an easy way to overcome this problem:

we only need to estimate the probability $\pi_{i\gamma}$ to weight the ith row of residual matrix $R = X - \hat{B}S$. Then reconstruct $\hat{G}$ through the singular value decomposition of the weighted R.

In order to check this idea, I set up a simple simulation study to look at the performance of two approaches of SVA.

2 Simulation Settings

Data matrix is in the dimensions of 100 $\times$ 80, which contains 100 genes and 80 samples.

80 Samples come from two classes (measured factor) and two batches (unmeasured factor)

- Class 1: Sample 1 - 40; Class 2: Sample 41 - 80.
- Batch 1: Sample 1 - 20, 41 - 60; Batch 2: Sample 21 - 40, 61 - 80.

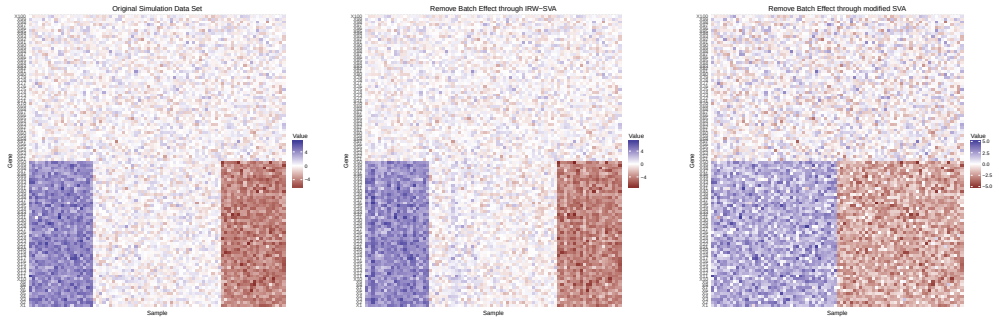
100 Gene has four types:

- Type A Gene: Gene associated with class label (measured factor) but not with batch label (unmeasured factor).
- Type B Gene: Gene associated with batch label (unmeasured factor) but not with class label (measured factor).
- Type C Gene: Gene associated with both class label (measured factor) and batch label (unmeasured factor).
- Type D Gene: Gene contains no signal.

3 Simulation Result

3.1 Case 1: The simulation data set contains Type B Gene and Type D Gene.

In the Case 1, IRW-SVA almost fails and modified SVA works pretty well. Moreover, samples from two batches are more distinguished under the direction of surrogate variable gained from modified SVA.

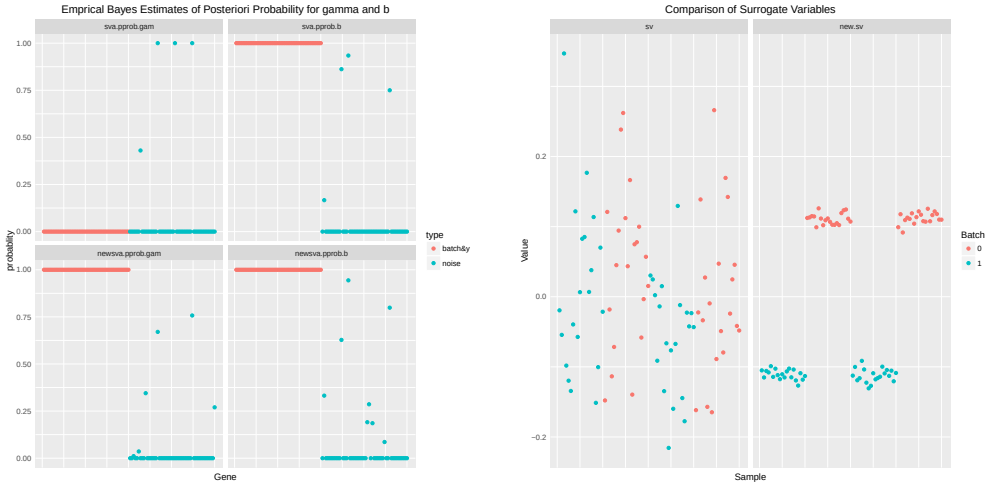


(a) Original Simulation Data Set in the Case 1

(b) Remove Batch Effect through IRW-SVA

(c) Remove Batch Effect through modified SVA

Figure 1: Remove Batch Effect Through Different Approaches of SVA in the Case 1



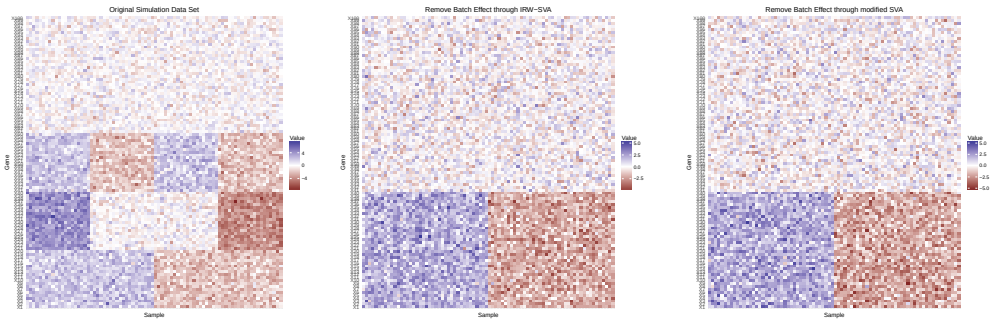
(a) Posterior probability of $\mathbb{P}(b \neq 0|X, S, \hat{G})$ and $\mathbb{P}(\gamma \neq 0|X, S, \hat{G})$

(b) Comparison of Surrogate Variables

Figure 2: Visualize The Analysis Through Sample Space And Gene Space in the Case 1

3.2 Case 2: The simulation data set contains all four types of genes.

In the Case 2, IRW-SVA and modified SVA both work pretty well and they produce similar surrogate variable.

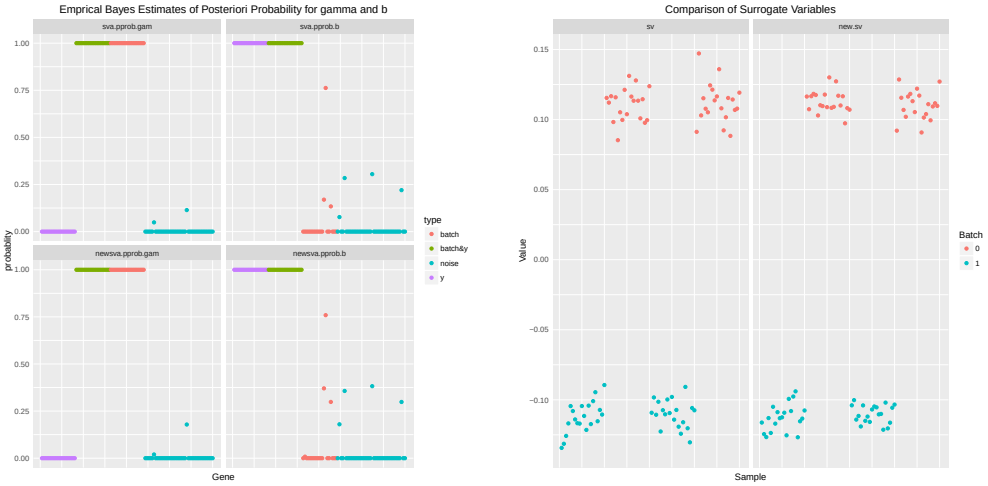


(a) Original Simulation Data Set in the Case 2

(b) Remove Batch Effect through IRW-SVA

(c) Remove Batch Effect through modified SVA

Figure 3: Remove Batch Effect Through Different Approaches of SVA in the Case 2



(a) Posterior probability of $\mathbb{P}(b \neq 0|X, S, \hat{G})$ and $\mathbb{P}(\gamma \neq 0|X, S, \hat{G})$

(b) Comparison of Surrogate Variables

Figure 4: Visualize The Analysis Through Sample Space And Gene Space in the Case 2

References

[1] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[2] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 2007.

[3] 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.