Simulation study: Modified SVA versus SVA

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

February 4, 2016

1 Modified SVA

Papers [1, 2, 3] 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, SVA is proposed to estimate G. An essential idea of SVA is to identify a subset of genes that are strongly associated with unmeasured confounders, but not with the group outcome. Especially, an empirical bayesian procedure has been applied to estimate the probabilities:

$$\pi_{i\gamma} = \Pr(\gamma_i \neq \vec{0}|X, S, \hat{G})$$
$$\pi_{ib} = \Pr(b_i \neq \vec{0}|\gamma_i \neq \vec{0}, X, S, \hat{G})$$

And then calculate the probability

$$\begin{aligned}\pi_{iw} &= \Pr(b_i = \vec{0} \ \& \ \gamma_i \neq \vec{0}|X, S, \hat{G}) \\ &= \Pr(b_i = \vec{0}|\gamma_i \neq \vec{0}, X, S, \hat{G})\Pr(\gamma_i \neq \vec{0}|X, S, \hat{G}) \\ &= (1 - \pi_{ib})\pi_{i\gamma}\end{aligned}$$

Next π_{iw} is used to weight the ith row of X and a singular value decomposition on the weighted X is performed to reconstruct $\hat{G}$.

However, if we estimate $\hat{G}$ using this approach, as done by IRW-SVA, I think it assumes there exists such a subset of genes in the data set. When such a subset does not exist, IRW-SVA can fail.

Therefore, I propose a simple way to overcome this problem: Use 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 investigate this idea, a simple simulation study is set up to compare the performance of two approaches to SVA.

2 Simulation Settings

Data matrix is in the dimensions of 100 $\times$ 80, i.e. 100 genes and 80 samples. The 80 Samples come from two classes (measured factor) and two batches (unmeasured factor)

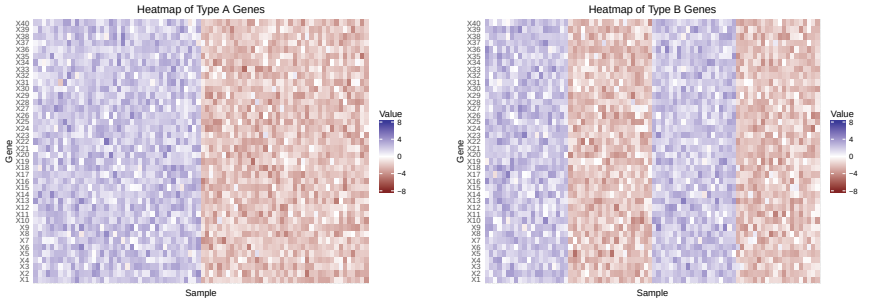
- Class 1: Sample 1 - 40; Class 2: Sample 41 - 80.
- Batch 1: Sample 1 - 20, 41 - 60; Batch 2: Sample 21 - 40, 61 - 80.

The 100 Gene have four types, whose patterns of heatmap are shown in Figure 1:

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

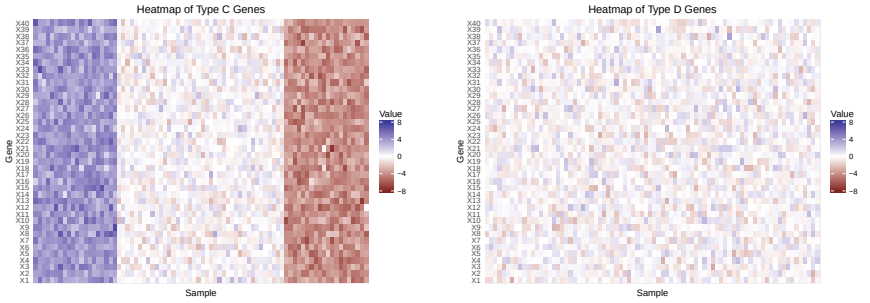
Type A and Type C genes are associated with the classed label. Type B and Type C genes are associated with the batch label. The key step of IRW-SVA is to identify the Type B genes. In the simulation study, two cases are typically investigated:

- Case 1: Simulation data set does not contain Type B genes.
- Case 2: Simulation data set contains Type B genes.



(a) Type A: genes associated with class label but not with batch label.

(b) Type B: genes associated with batch label but not with class label.



(c) Type C: genes associated with both class label and batch label.

(d) Type D: genes containing no signal.

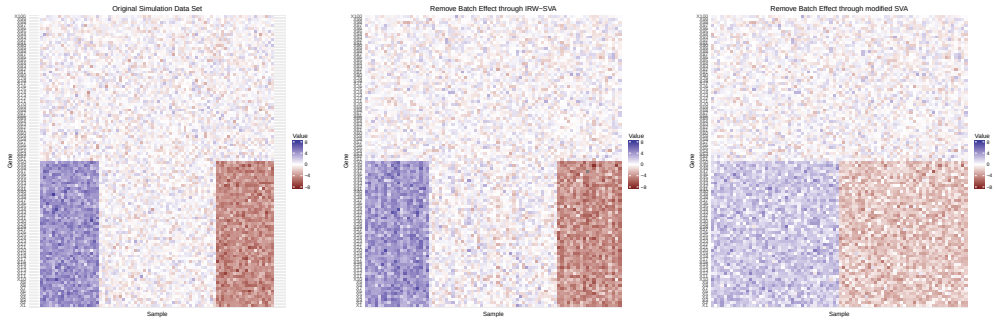
Figure 1: Heatmap of four types of genes

3 Simulation Result

3.1 Case 1: The simulation data set contains only Type C Genes and Type D Genes.

In the Case 1, Figure 2 shows that IRW-SVA fails to adjust batch effects, while modified SVA adjusts batch effects and recovers the pattern of measured factor pretty well. Figure 3(a) shows that IRW-SVA can not correctly identify the genes associated with batch label, while modified SVA are able to correctly identify the genes associated with batch label and class label. Moreover, Figure 3(b) shows samples from two batches are more separable under the direction of surrogate variable gained from modified SVA.

The simulation results in the Case 1 indicate that modified SVA has better performance than IRW-SVA when Type B genes does not exist in the data.



(a) Original simulation data set in the Case 1

(b) Remove batch effect through IRW-SVA

(c) Remove batch effect through modified SVA

Figure 2: Study batch effect through two approaches to SVA in the case 1

(a) pprob.b: posterior probability of genes affected by measured factor. pprob.gam: posterior probability of genes affected by unmeasured factor.

(b) Comparison of surrogate variables

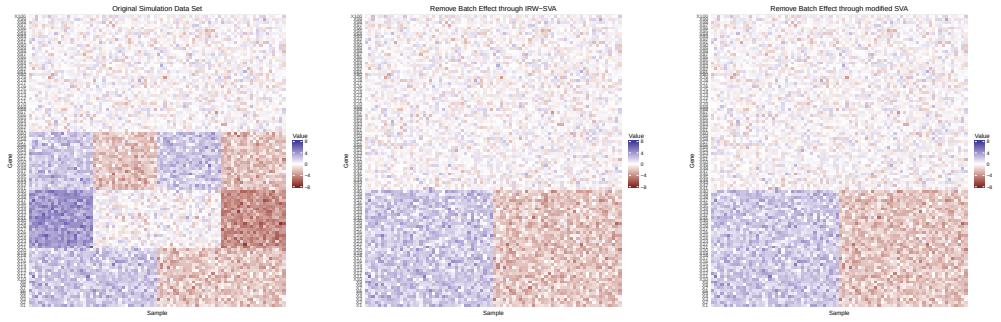
Figure 3: Visualize the analysis through sample space and gene space in the Case 1. In each subfigure, the panels on the left show the results from IRW-SVA and the panels on the right show the results form modified SVA.

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

Figure 4 shows that both IRW-SVA and modified SVA can adjust the batch effects in the Case 2. Figure 5(a) shows that both IRW-SVA and modified SVA are able to identify the genes associated with batch label and class label. Figure 5(b) shows that they produce similar surrogate variable which separate samples from two batches pretty well.

The simulation results in the Case 2 indicate that IRW-SVA and modified SVA

have similar performance when the data set contains Type B genes.

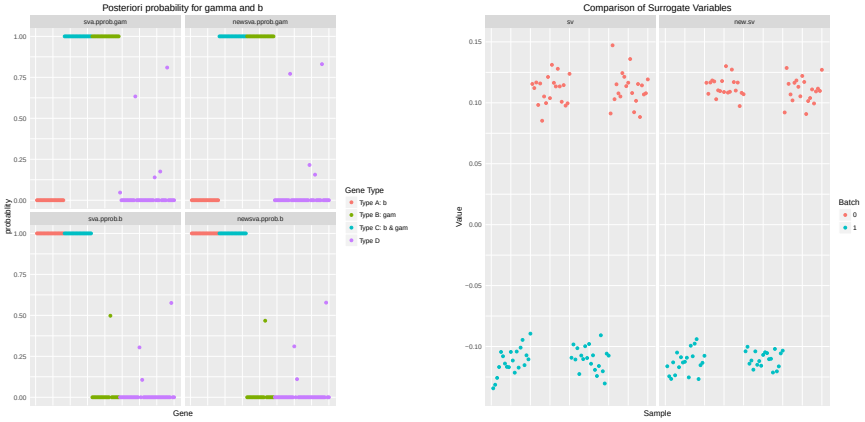


(a) Original simulation data set in the Case 2

(b) Remove batch effect through IRW-SVA

(c) Remove batch effect through modified SVA

Figure 4: Study batch effect through two approaches to SVA in the Case 2



(a) pprob.b: posterior probability of genes affected by measured factor. pprob.gam: posterior probability of genes affected by unmeasured factor.

(b) Comparison of surrogate variables

Figure 5: Visualize the analysis from sample space and gene space in the Case 2. In each subfigure, the panels on the left show the results from IRW-SVA and the panels on the right show the results form modified SVA.

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.