Simulation study: Modified SVA versus SVA

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

February 12, 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 posterior probabilities of each gene affected by unmeasured confounders and measured factors, namely:

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

And then calculate the probability that a gene is associated with unmeasured confounders, but not with the measured factors

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

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, as shown in Case 1 of the simulation study.

There seems to be 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 these two approaches to SVA.

2 Simulation Settings

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

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

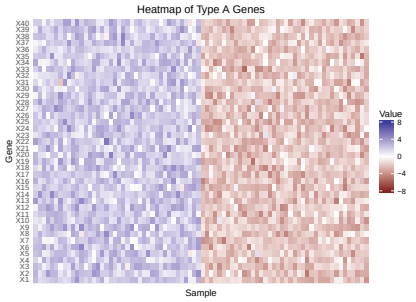
The 100 Genes have four types:

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

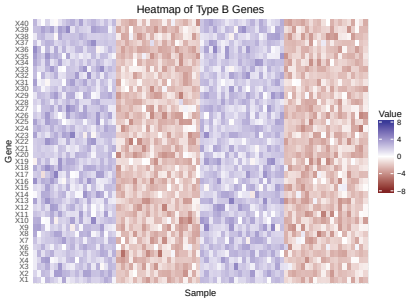
The heatmaps of the four types of genes are shown respectively in Figure 1. In Figure 1, rows are genes, columns are samples and entries are expression values.

The key step of IRW-SVA is to identify the Type B genes. In the simulation study, two cases are typically investigated:

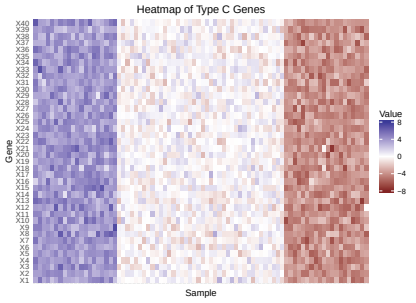
- Case 1: Simulation data set does not contain Type B genes.
- Case 2: Simulation data set contains Type B genes.



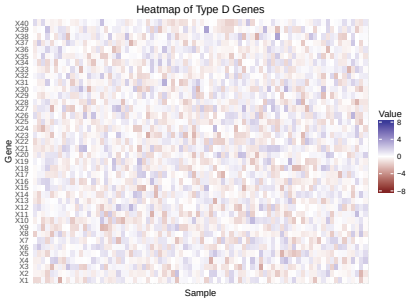
(a) Type A: genes with class label but not with batch label signal.



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



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



(d) Type D: genes with no signal.

Figure 1: Separate heatmaps for each of the four gene types

3 Simulation Result

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

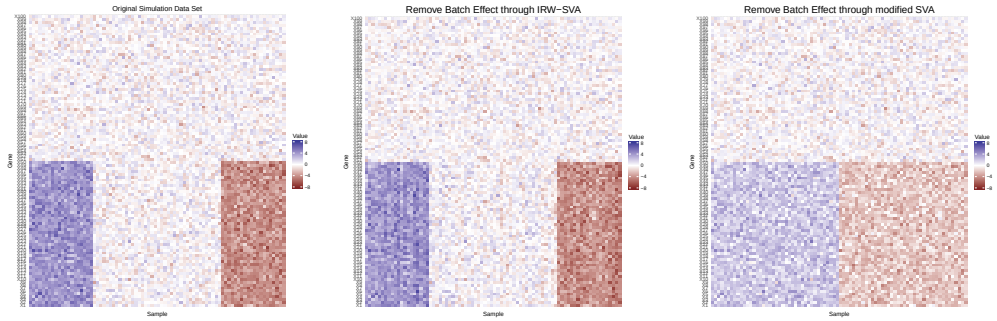
As shown in Figure 2(b), the input data matrix in Case 1 is the contamination of the matrices in Figure 1(c) and Figure 1(d). Figure 2(b) shows that the data matrix after removing batch effects through IRW-SVA has the same pattern as the original simulation data set, which indicate that IRW-SVA fails to adjust batch effects in this case. Figure 2(c) shows that after removing the batch effects through modified, the rows of Type C genes, which affected by both class and batch effects, have been recovered as the same signal in Figure 1(a). This indicates modified SVA adjusts batch effects and recovers the pattern of measured factor quite well in Case 1.

In Figure 6, the x-axis and y-axis represent gene index and probability respectively, and the point of each gene is colored by its type. Figure 6(b) shows the posterior probability of each gene affected by class effects and we expect that Type C genes have high probability while Type D genes have low probabilty. Figure 6(b) indicates that both IRW-SVA and modified SVA identify the genes affected by class

effects. Figure 6(a) shows the posterior probability of each gene affected by batch effects. The left panel of Figure 6(a) shows that IRW-SVA can not correctly identify the genes associated with batch label, while the right panel of Figure 6(a) indicates modified SVA is able to correctly identify the genes associated with batch label. This is not surprised since IRW-SVA relies on Type B Genes to recover the signal of batch effects while the data set in Case 1 does not contain Type B genes.

Moreover, Figure 4 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 do not exist in the data.

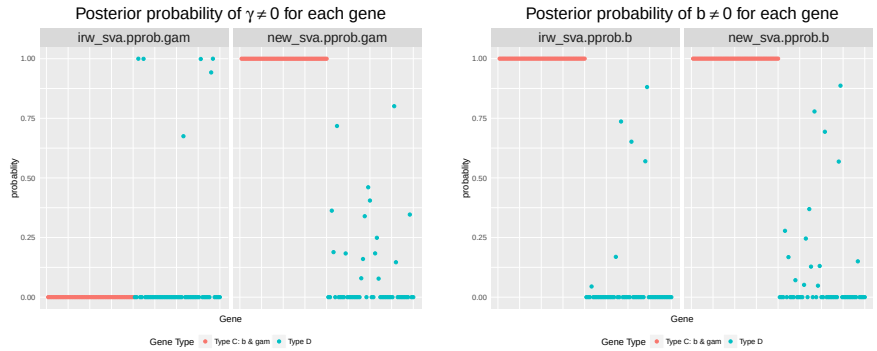


(a) Original simulation data set in the Case 1

(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) Posterior probability of genes affected by unmeasured factor.

(b) Posterior probability of genes affected by measured factor.

Figure 3: Visualize the posterior probabilities for genes 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.

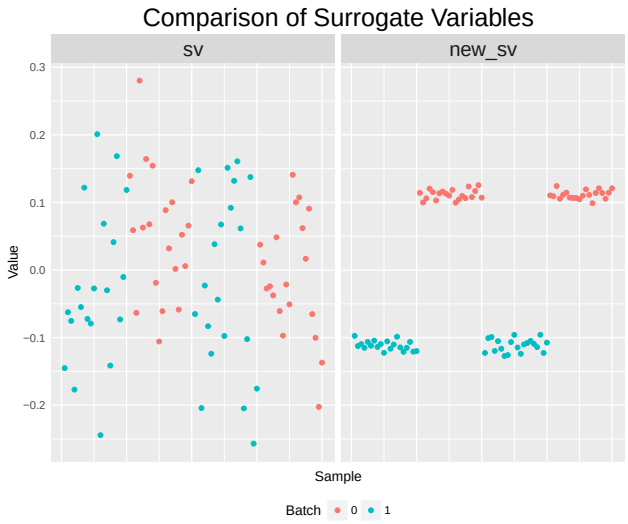
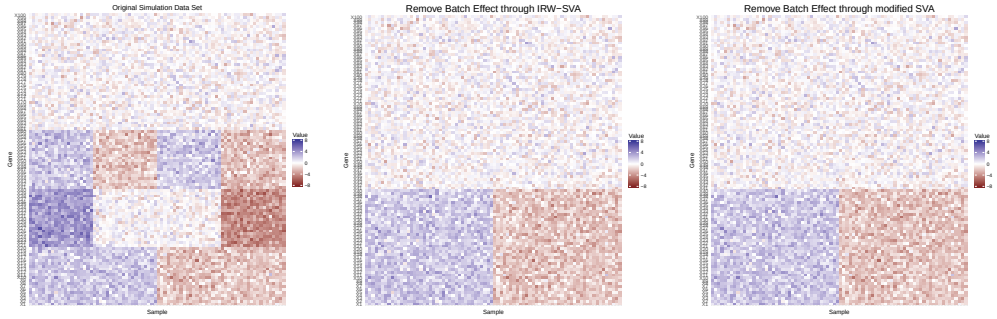


Figure 4: Comparison of surrogate variables

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

Figure 5 shows that both IRW-SVA and modified SVA can adjust the batch effects in the Case 2. Figure **????** shows that both IRW-SVA and modified SVA are able to identify the genes associated with batch label and class label. Figure **????** 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.

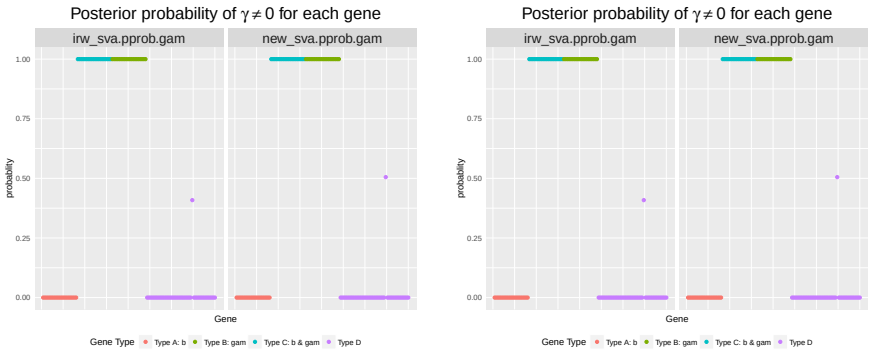


(a) Original simulation data set in the Case 2

(b) Remove batch effect through IRW-SVA

(c) Remove batch effect through modified SVA

Figure 5: Study batch effect through two approaches to SVA in the Case 2



(a) Posterior probability of genes affected by measured factor.

(b) Posterior probability of genes affected by unmeasured factor.

Figure 6: Visualize the posterior probabilities for genes 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.

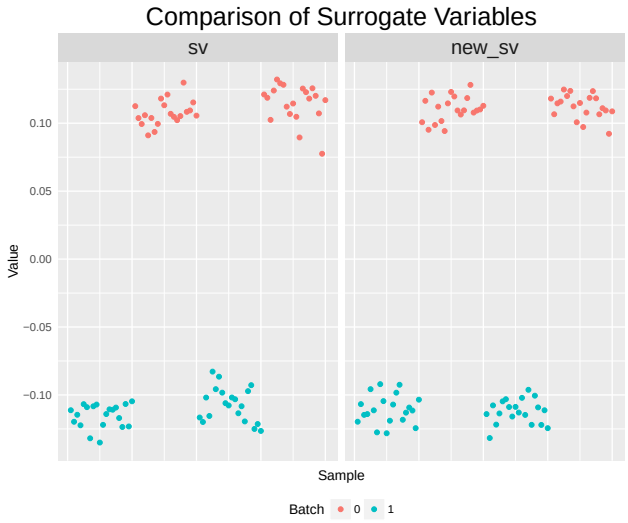


Figure 7: Comparison of surrogate variables

Figure 8: Visualize the scores of samples on the surrogate variable. 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.