Simulation study: Modified SVA versus IRW-SVA

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

February 18, 2016

1 Modified SVA

Papers [1, 2, 3] proposed a factor model for the relationship between expression values X, measured biological factors S, and unmeasured biological and non-biological factors G:

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

In the model, $X_{m\times n}$ is a data matrix where each entry x_{ij} is the value of gene i for sample j, $S_{p_1\times n}$ is a design model matrix of p_1 measured factors for the n samples, $B_{m\times p_1}$ are the coefficients for these variables on the m genes, $G_{p_2\times n}$ is an unknown matrix that parameterizes the effect of unmeasured confounders and $\Gamma_{m\times p_2}$ is the corresponding coefficient matrix for G.

In order to remove the batch effects, iteratively reweighted surrogate variable analysis (IRW-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:

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

Next the probability that a gene is associated with unmeasured confounders but not with the measured factors is calculated

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

In IRW-SVA, 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, 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 the 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 the dimensions of 100×80, i.e. 100 genes and 80 samples, as shown in Figure 1. The 80 Samples (columns) come from two classes (the measured factor) and two batches (the 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.

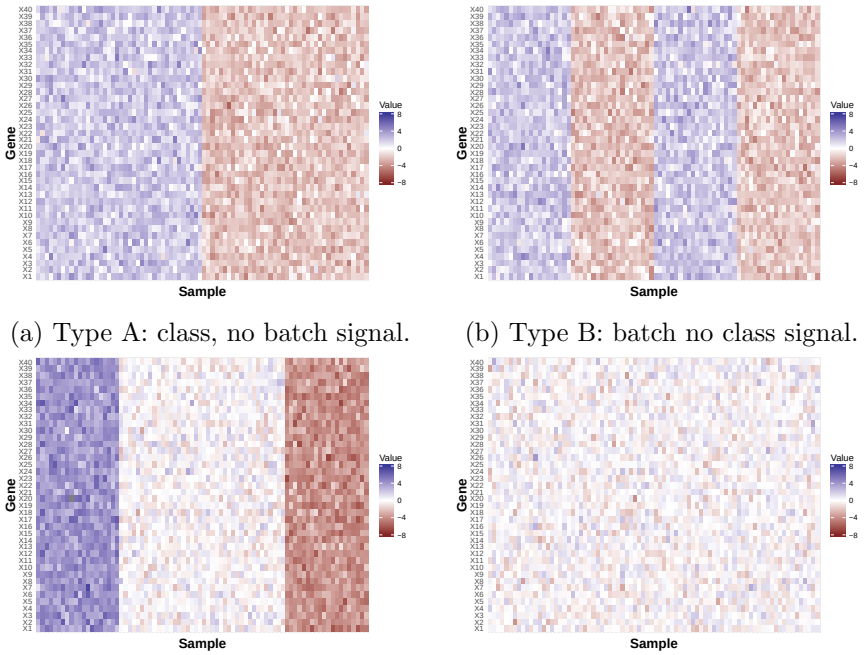
This results in the design model matrix $S = \begin{pmatrix} 1_{40} & 1_{40} \\ 1_{40} & 0_{40} \end{pmatrix}$ and the confounder matrix $G = \begin{pmatrix} 1_{20} & 0_{20} & 1_{20} & 0_{20} \end{pmatrix}$, where 1_k (0_k) is the $1 \times k$ row vector of all ones (zeros respectively).

The 100 Genes (rows) have four types, which are shown in separate panels of Figure 1:

- 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 key step of IRW-SVA is to identify the Type B genes and use them as the focus of estimating G. 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: class, no batch signal.

(b) Type B: batch no class signal.

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

(d) Type D: no signal.

Figure 1: Separate heatmaps for each of the four gene types determining on class and batch labels. The rows are genes, the columns are samples, and the entries are color coded expression values.

3 Simulation Result

3.1 Case 1: Type C and Type D Genes.

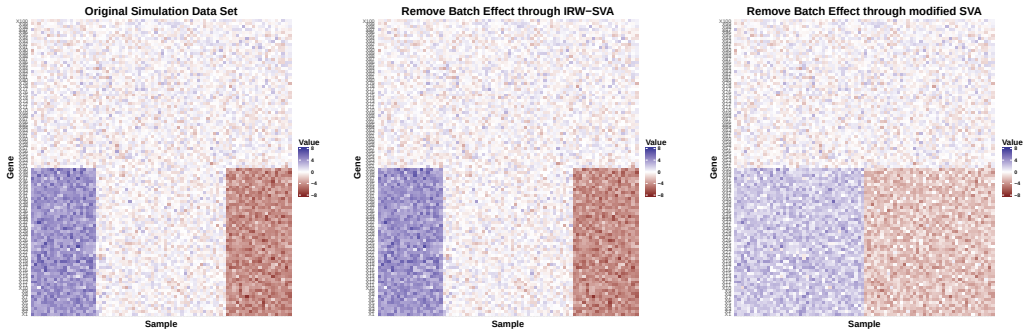
The input data matrix in Case 1 is shown in Figure 2(a), which is the concatenation of the matrices in Figure 1(c) and Figure 1(d). Figure 2(b) shows that the data matrix after batch adjusting using IRW-SVA still retains the same pattern as the original simulation data set, which indicates that IRW-SVA failed to adjust for batch effects in this case. Figure 2(c) shows that after batch adjustment using modified SVA, the rows of Type C genes in the data set, which are affected by both class and batch effects, have been recovered as the class signal in Figure 1(a). This indicates that modified SVA adjusts batch effects and recovers the pattern of the measured factors quite well in Case 1.

In order to understand the failure of IRW-SVA in Case 1, the posterior probabilities are investigated in the two approaches to SVA. In Figure 3, the x-axes and y-axes represent the gene index and the estimated probabilities respectively, and the genes are colored by their types. Figures 3(a) and 3(b) show the posterior probabilities,

i.e. π_{ib} and $\pi_{i\gamma}$, of each gene affected by class effects and batch effects respectively. In both subfigures we expect that Type C genes have high probabilities while Type D genes have low probabilities. In Figure 3(a) the Type C genes are all nearly 1 and the Type D genes are usually near 0, which indicates that both IRW-SVA and modified SVA correctly identify the genes affected by class effects. In the left panel of Figure 3(b) almost all the genes are near zero and only several Type D genes are nearly 1. This shows that IRW-SVA does not correctly identify the genes associated with batch label, while the right panel of Figure 3(b) indicates modified SVA is able to do that. This is not surprising since IRW-SVA relies on Type B Genes to recover batch label signal but the data set in Case 1 does not contain Type B genes.

Moreover, Figure 4 visualizes the values of the surrogate variables constructed by the two approaches to SVA respectively for each sample. In Figure 4, x-axes and y-axes represent samples and expression value respectively, and samples are colored by batch labels. The right panel of Figure 4 shows that the samples from two batches are quite separable using the surrogate variable calculated by the modified SVA, but the left panel of Figure 4 shows no useful separation by IRW-SVA.

All these results in Case 1 indicate that modified SVA has better performance than IRW-SVA when Type B genes do not exist in the data.

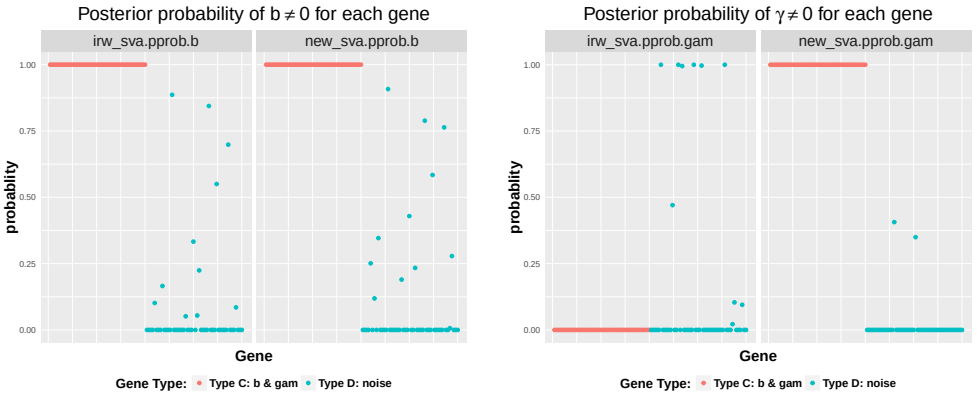


(a) Original simulation data set in the Case 1

(b) Batch adjustment by IRW-SVA

(c) Batch adjustment by modified SVA

Figure 2: Study batch effect through two approaches to SVA in Case 1. Show failure of IRW-SVA in this case.



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

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

Figure 3: The posterior probabilities for genes in 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 from modified SVA.

3.2 Case 2: All four types of genes.

In contrast to Case 1, the input data matrix in Case 2 contains all four types of genes, as shown in Figure 5(a), which is a typical situation we expect IRW-SVA works. Figure 5(b) and Figure 5(c) shows that the genes affected by class labels in the two adjusted matrices have the same pattern in Figure 1(a). There results demonstrate that both IRW-SVA and modified SVA effectively adjust the batch effects in Case 2.

To investigate why, Figure 6 shows the posterior probabilities in the two approaches. In Case 2, we expect that Type A and Type C genes have high posterior probabilities affected by class effects in Figure 6(a), and Type B and Type C genes have high posterior probabilities affected by batch effects in Figure 6(b). Then Figure 6(a) and Figure 6(b) shows that both IRW-SVA and modified SVA are able to identify the genes associated with batch label and class label as well.

The scores of samples on the surrogate variables from two approaches to SVA are shown in Figure 7.The two approaches to SVA produce similar surrogate variables which separate samples from two batches pretty well.

To sum up, the simulation results in the Case 2 indicate that IRW-SVA and modified SVA have similar performance and both are able to adjust the unmeasured factors when the data set contains Type B genes.

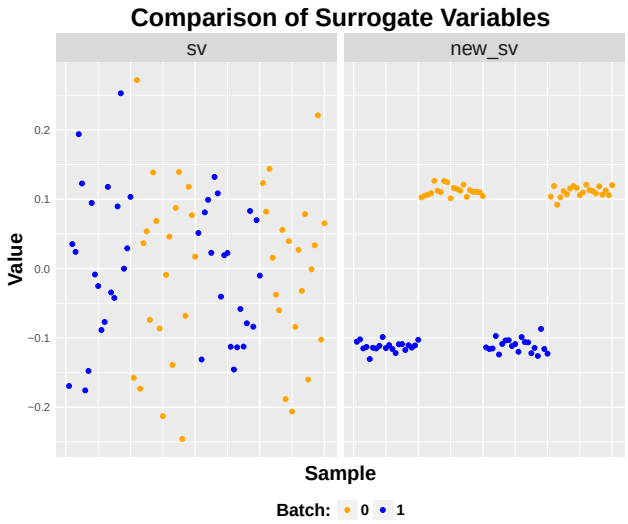
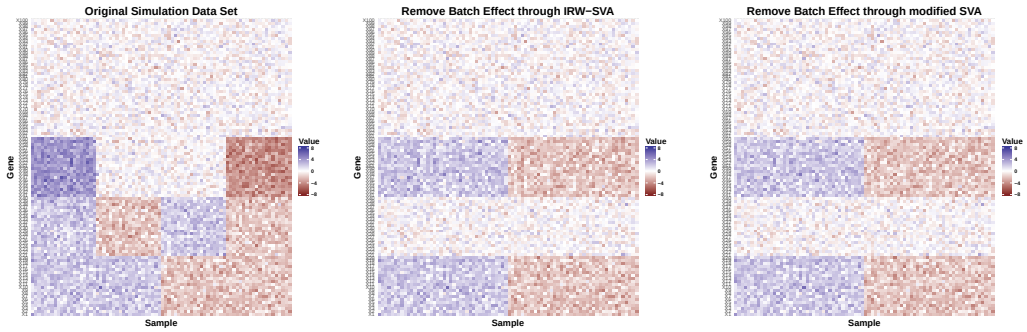


Figure 4: Comparison of surrogate variables. The left panel shows the surrogate variable constructed through IRW-SVA and the right panel shows the surrogate variable constructed through modified SVA. Modified SVA gives much better performance.

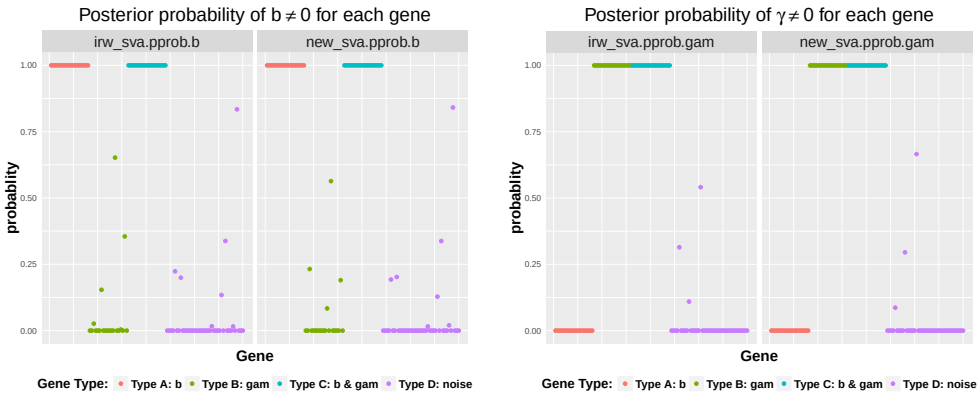


(a) Original simulation data set in the Case 2

(b) Batch adjustment by IRW-SVA

(c) Batch adjustment by modified SVA

Figure 5: Study batch effect through two approaches to SVA in Case 2. Show both approaches to SVA work well in this case.



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

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

Figure 6: 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.

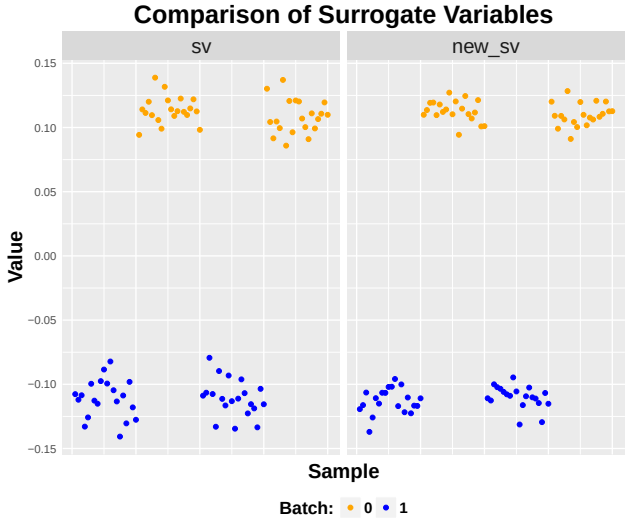


Figure 7: Comparison of surrogate variables. The left panel shows the surrogate variable constructed through IRW-SVA and the right panel shows the surrogate variable constructed through modified SVA, with both giving good performance.

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.