Summary and discussion of: "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis"

Statistics Journal Club, 36-825

Cong Lu and Wei Dai

1 Summary

1.1 Motivating problem

Gene expression study is well known to focus on finding association between expression levels of particular genes and some interesting variables, for example, a disease state. In such studies, besides the primary variable of interest, some other covariates are usually measured and included in the model of association tests. However, it is not possible to measure all the variables related to gene expression. In addition to the known and measured variables, there might be some other unknown and unmeasured factors, which would contribute to the heterogeneity of expression levels of particular genes. Their effect cannot be simply ignored as the unmodeled factors, either biological or technical, may induce extra variability in the gene expression, which would decrease the power to detect the association between gene expression and the primary variable. Or they might introduce spurious signals of the association as the variation on the gene expression could be confounded by a unmodeled factor and the primary variable.The model of Surrogate variable analysis(SVA) was designed to address this problem. In this paper, SVA was shown to able to identify and estimate the components of Expression Heterogeneity(EH), and parse the signal and noise more accurately and reproducibly.

1.2 Explanation of SVA procedure

Generally, there are four steps of a surrogate variable analysis.

First, from the gene expression matrix X we remove the signals due to the primary variable, to get a residual matrix R. Then the residual matrix R can be decomposed to a orthogonal basis of singular vectors, which are able to completely reproduce all signatures. Among all the singular vectors, we test each of them to see if it represents more variation of expression than by chance. The process is implemented by permutations. And the significant singular vectors will be picked out. This step would guarantee the signal is not from the primary variable.

Second, for each selected orthogonal singular vector, we identify a group of genes by significance analysis, that are significant associated with the vector. As an unmodeled factor may have an impact on only a subset of genes rather than the whole set of all genes, this step will lead to a more accurate estimation of a SV.

Third, for each identified group of genes, we built one Surrogate Variable(SV) from the original Expression matrix on those genes. By using the original expression matrix (full EH signature), we allow for the correlation between the SV and the primary variable.

Last, we will include the primary variable(PV) and all the significant SVs as covariates in the final model. For different genes, we allow for different coefficients for the SVs. And one gene could be affected by more than one SV.

1.3 Simulation studies of SVA

Results of simulation studies were shown to justify that SVA is able to account for unmeasured covariate and reduce the heterogeneity of gene expression levels.

1.3.1 Simulation setting

The simulation in the paper was designed to show the performance of SVA in the large-scale significance testing, i.e., for each gene, they tested whether it is significantly associated with the PV, disease states. In total, 1,000 studies were simulated, and in each of the studies there were 1,000 genes on 20 arrays. 10 arrays were from cases and the other 10 were from controls of a diseases. The first 300 genes were differentially expressed between cases and controls, and gene 200-500 were affected by an independent unobserved factor.

1.3.2 Results

The simulation results showed that SVA did a good job from the following four perspectives: (a) Estimation of SV, (b)Distribution of null p-values, (c)Estimation of FDR, (d)Robustness of confounding between the PV and SVs.

Accurate estimation of SV In the 1,000 studies, 99.5% of them identified one significant surrogate variable. And the average correlation between the estimated SV and the unobserved factor was 0.95 with standard deviation of 0.05. Each SV is a weighted average of expression over a subset of genes. On average, 30.5% of the 300 truly affected genes were identified as affected, while only 9.9% of the 700 truly unaffected genes were identified as affected. We can tell that on average, 160 genes were identified to be affected by the unobserved factor, and only a little more than half(56%) are the truly affected genes.

Correct p-value Distribution In a significance analysis, p-values corresponding to the null hypotheses should be uniformly distributed in [0, 1]. Therefore for each single null gene, the p-values over the 1,000 studies should be uniformly distributed. However, the p-values from the 1,000 genes in one study will be different from the p-values from one gene across 1,000 studies. Since in the former setting, there is dependence across the genes, which are affected by the unmodeled factor. i.e. for the 700 null genes in one simulation study, since 200 genes are affected by the same factor, the distribution of the 700 p-values will deviate from a uniform distribution. In general, if the unmodeled factor is correlated with the PV, the null p-values will be biased towards zero. While if the unmodeled factor is uncorrelated with the PV, the null p-values will be biased towards one.

Plots of 9 simulations studies were shown before and after applying SV, and so was a plot of p-value quantiles of 1,000 K-S tests. From these plots we can observe, SVA was able to correct the distribution of null gene p-values towards the uniform distribution. From Figure S4 in the paper we can see the distribution of p-values is corrected by SVA, where in the simulation the residuals were drawn from a published microarray study. This confirms that SVA is robust to the distribution of the gene specific error. The application on DE genes was shown in Figure 1C. From the simulation study, we can compute the power as we know which genes are the truly differentially expressed genes. From the plot, we can see SVA increases the power. However, in practice on the real data, it is hard to tell if SVA increases or decrease the power. It will depend on whether the null p-values are biased towards zero or one.

Stable and Accurate Gene Ranking SVA can make reproducible ranking of genes for differential expression.

Improved FDR estimation Large-scale dependence across genes has been shown to be a problem for estimating FDR since the dependence increases the variance of FDR. SVA reduces the variability in both the estimate of the proportion of null hypotheses and the q-values for each study. See details in Figure S6.

Robustness to confounding Another simulation was conducted to let the unmodeled factor to be correlated with the PV, with correlation coefficient of 0.5 and a standard deviation of 0.16 across 1,000 simulation studies. Now the unmodeled factor is both correlated with the PV and affect a subset of genes.

In this setting, SVA identified one SV in 94.5% of the simulation studies. And the correlation between the estimated SV and the unmodeled factor is 0.94 with a standard deviation 0.22. The author argued that the results showed SVA is robust to the strong dependence between the SV and the unmodeled factor. However, the author did not show for one simulation data set, among the identified genes affected by the SV, how many are truly affected and how many are not.

1.4 Case Studies on Real Data

In the paper, SVA was applied in three real data sets, for the studies of eQTL mapping and differential expression analysis.

1.4.1 Genetics of Gene Expression in Yeast: eQTL mapping

In the Yeast Data from Brem et al(2002, 2005), there are 112 segregants of a cross between two isogenic strains of yeast. And also the genotype data, which covers 99% of the genome, was also available.

Many of the gene expression traits are cis-linking (i.e. most of the associated loci are within a short physical distance from that gene, which we call open reading frame), while some of them are trans-linking (i.e. the associated loci are physically far from the open reading frame). For the trans-linking pairs of a gene and a genetic locus, there are several

"pivotal" loci, any one of which influences hundreds and thousands genes in their expression. Then the pivotal loci act as a major source of EH.

They fit two models:

$$\text{Model 1: Expression of a Gene} \sim \text{a Genetic Loci}$$
$$\text{Model 2: Expression of a Gene} \sim \text{a Genetic Loci + SVs}$$

From the results of the two models on all the combination pairs of a gene and a genetic locus , it showed the effects of the few pivotal loci could be captured and removed by SVA. See Figure 3A, we can observe by applying SVA, the majority of the trans-linkages to the pivotal loci were removed.

With the existence of EH caused by the pivotal loci, some cis-linkages, which are more interesting than trans-linkages, may not be detected as significant ones. They calculated the p-values from both Model 1 and Model 2, only on the loci within 3 centimorgans of the open reading frame for each gene. In Table 1, we see at the same FDR cutoff, Model 2 with SVA detected more cis-linkages, which suggests SVA has the potential to increase the power of eQTL mapping.

1.4.2 Human Expression: Differential Expression between two disease states

There are 15 tumor samples tumor samples in the data from Hedenfalk et al, 2001. 7 of them are with $BRCA1 mutations$ and 8 are with $BARCA2$ mutations. Previous biomedical research has shown that there are notable substructure within the BRCA2 samples. SAV was applied to capture this structure.

$$\text{Model 3: Expression of a Gene} \sim \text{Disease State (BRCA1 or BRCA2)}$$
$$\text{Model 4: Expression of a Gene} \sim \text{Disease State (BRCA1 or BRCA2) + SVs}$$

Fit the Model 3 and 4 on all the genes, Model 4 with SVA detected fewer differentially expressed genes at the same FDR cutoffs. Many of the declared extremely significant DE genes from Model 3 are highly associated with top SV in Model 4, and we believe the differential expression of those genes is mainly driven by EH. Figure S8 shows some genes with low ranks(large rank indices) in the unadjusted model, Model 3, would increase the ranking to high ranks (small indices) in the adjusted model, Model 4. This suggests that the small p-values detected from the unjust model would go to be very large in the adjusted model, i.e., they are not true signals.

1.4.3 Human Expression: Differential Expression on Age

In this study, they used data from Rodwell et al, 2004. There are 133 patients with gene expression measured in kidney tissue.

$$\text{Model 5: Expression of a Gene} \sim \text{Age + Tissue Type}$$
$$\text{Model 6: Expression of a Gene} \sim \text{Age}$$
$$\text{Model 7: Expression of a Gene} \sim \text{Age + SVs}$$

The top SV estimated has high correlation with 0.86 with Tissue Type. 84% of the detected genes associated with Tissue Type in Model 5 were also detected to be significantly associated with the first SV in Model 7.

At the same standard FDR cutoff, all age-associated genes detected in Model 4 were included in the list of the significant age-associated genes in Model 5. 96% of the age-associated genes detected in Model 7 with SVA were in the list from Model 5. Furthermore SVA-adjusted model (Model 7) detected more age-associated DE genes than unadjusted model(Model 6), among which 116 genes were significant both in Model 5 and 7, but not in Model 6. This suggests SVA is poetical to increase the power of detecting DE genes.

2 Discussion

In recent years, Surrogate Variable Analysis(SVA) has been very popular in the community of genetics on associate studies. People prefer using it for many reasons. For example, applying SVA may lead to increased number of eQTL detections and by the standard in the biology community more eQTL means improvement. Another truth of most associate studies is that usually very few related covariates are measured and SVA brings the possibility of reducing the heterogeneity. However, there are still arguments on that whether SVA could work in the way as it claimed.

Simulation Study I In Ryan's simulation (R code attached), he showed if SVA is able to accurately account for unmeasured covariates. Here are the details of the simulation: There assumes to be 50 samples (sample size n = 50) with 500 genes (p = 500) expressed, and 10 (k = 10) out of the 500 genes are affected by unmeasured covariates while the other 490 genes are not affected. More specifically, in those 10 genes, one gene is affected by 10 covariates, and there are 10 hidden covariates in total. None of the genes are affected by a primary variable Y. In another words, with respective to the primary variable Y, all are null genes. After applying SVA to this simulated data set, we let it generate 10 estimated SVs to account for the 10 hidden covariates, and for each gene we fit the model :

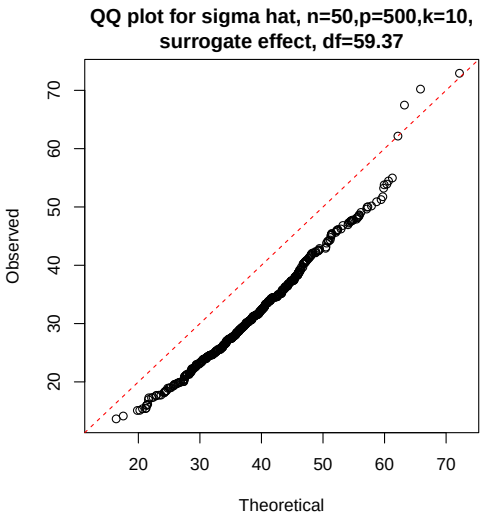
$$\text{Gene Expression} \sim \text{PV} + \text{SV1} + \text{SV2} + ... + \text{SV10}$$

We found that for the 490 gene not affected by the hidden covariates, none of the 10 SVs is significant for most cases as expected, while for the 10 genes that are affected by the hidden covariates, some of the SVs estimated are significant. After fitting the linear models, the sum of squared residuals (SSE) should follow a χ^2 distribution, i.e.

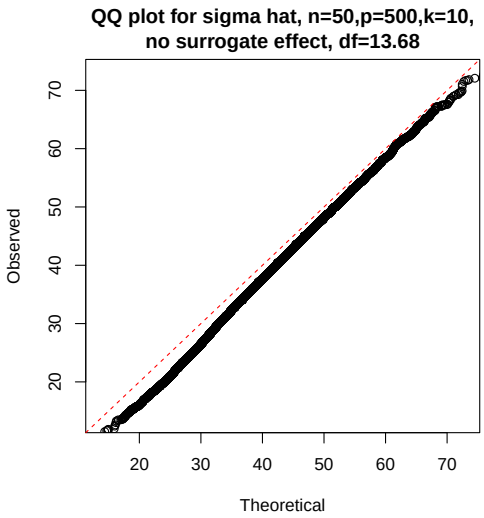
$$\frac{SSE}{\sigma^2} \sim \chi^2_{df=n-10-1},$$

We repeat the same simulation 50 times (R = 50). For each gene at one simulation, we got calculated one statistic $X^2 = SSE$, and in total 500 X^2's for the 10 genes affected by some hidden covariate and 24,500 (i.e. 490×50) X^2's for the other 490 genes. With the X^2's calculated, we can get the QQ-plot of σ^2 for two groups, 10 affected genes and 490 unaffected genes.

Figure 1 shows the QQ-plots respectively. From the one on the right, we observe all the points almost lie on the straight line $y = x$, which suggests SVA would have little impact on the unaffected genes. However, in the plot on the left for 10 affected genes, almost all the points are below the straight line $y = x$, which suggests by including 10 SVs, more variation would be removed than it should be in the linear models. From Figure 2 we

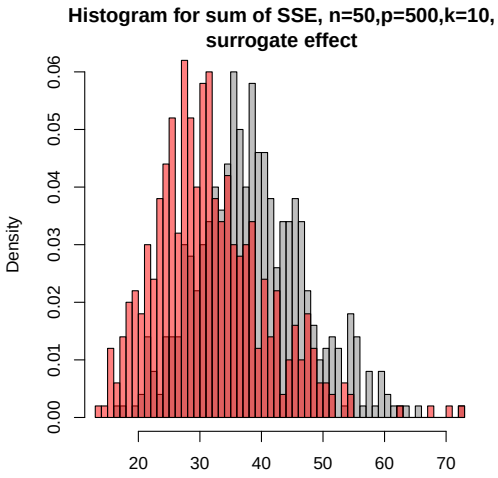


(a) 10 affected genes

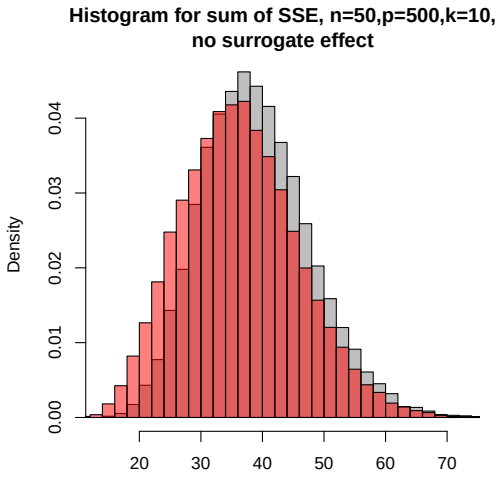


(b) 490 unaffected genes

Figure 1: QQ-plot

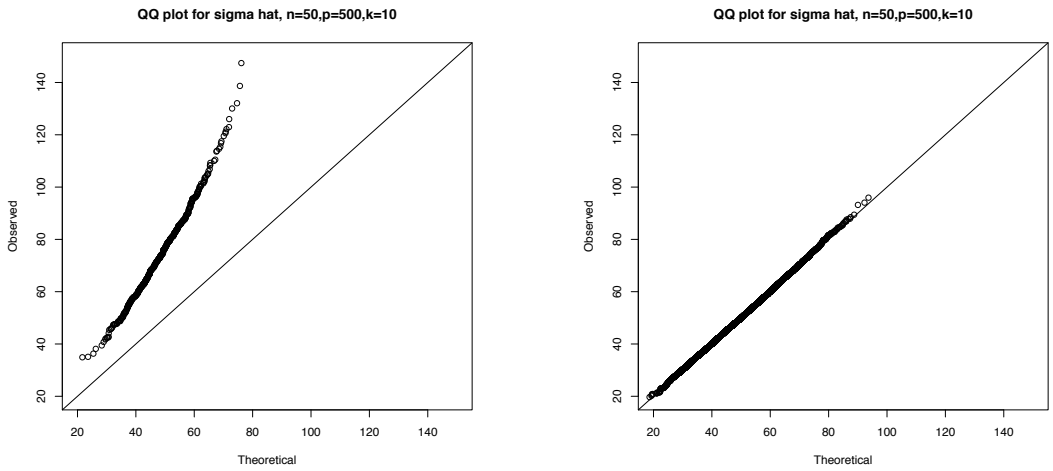


(a) 10 affected genes



(b) 490 unaffected genes

Figure 2: Histogram of $X^2 = SSE$ and χ^2_{39}



(a) 10 affected genes

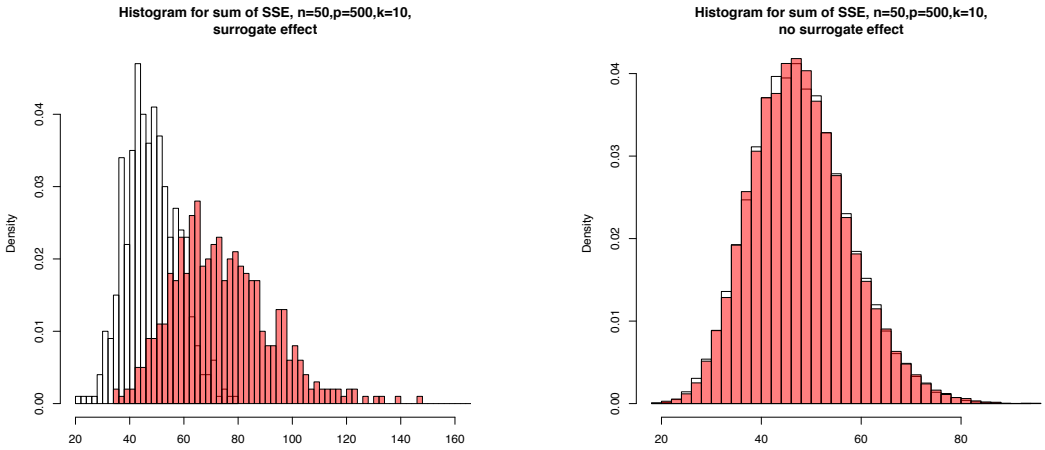
(b) 490 unaffected genes

Figure 3: QQ-plot

observe the same pattern as $X^2 = SSE$(in color red) is smaller than χ^2_{39}(in color grey) for the 10 affected genes, but the distributions of them are almost the same (slightly smaller than theoretical due to the number of SVs) for the 490 unaffected genes.

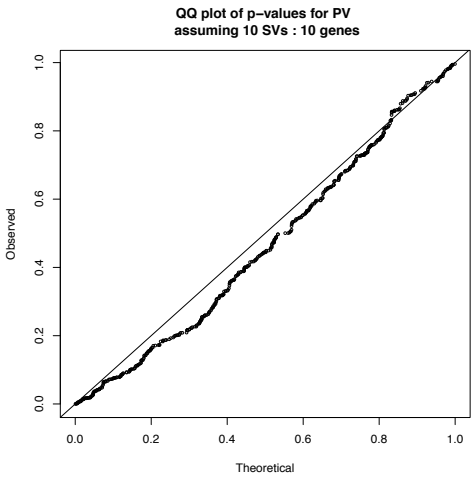
Simulation Study II In the last simulation study, we let SVA generate 10 estimated SVs as we simulated 10 hidden covariates. However, they may not be all significant. Therefore we conduct almost exactly the same simulation again but using the estimated number of significant SVs. As a result, only one significant SV is estimated. From the QQ-plot in Figure 3 for unaffected 490 genes, we find the observed SSE is perfectly following $\chi^2_{df=48}$. But we also observe deviations of SSE calculated from the linear model for 10 genes from χ^2 distribution. And also, as for the 490 genes, we observe no deviation at all compared to the histogram in Figure 2 since in Simulation Study I, we removed a little more variation when we used 10 SVs.

p-values of Study I and II Besides, the distribution of SSE in fitting the linear models, we take a look at the distribution of p-values for the primary variable(PV). Since all of the 500 genes are simulated to be null genes for the PV, we should expect to see a uniform distribution of p-values. In Figure 5, we can tell that using one estimated SV would generate closer p-value distributions than applying 10 SVs, although there are 10 hidden covariates.

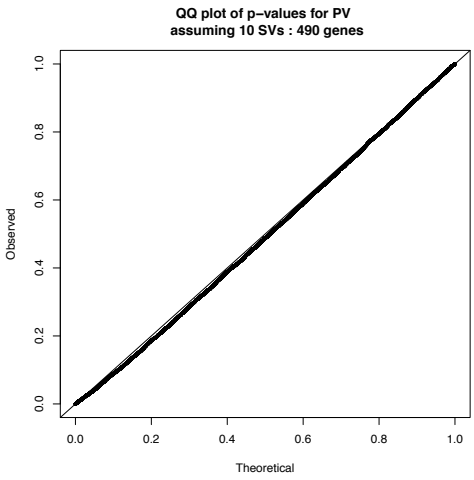


(a) 10 affected genes (b) 490 unaffected genes

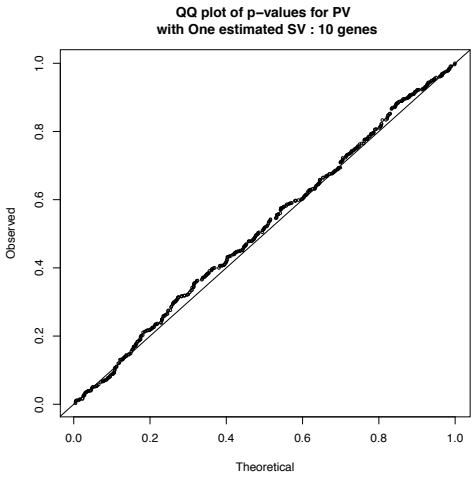
Figure 4: Histogram of $X^2 = SSE$ and χ^2_{48}



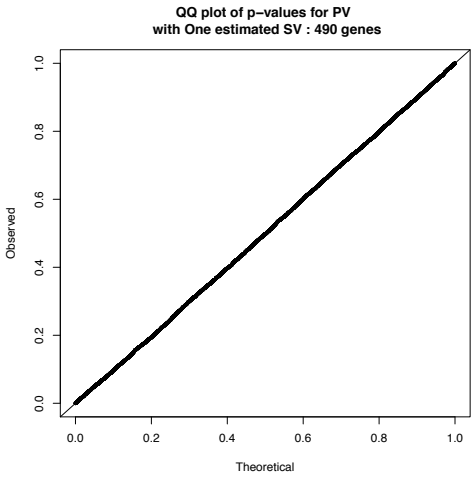
(a) 10 affected genes



(b) 490 unaffected genes



(c) 10 affected genes



(d) 490 unaffected genes

Figure 5: QQ-plot for p-values