Estimating batch effect in Microarray data with Principal Variance Component Analysis(PVCA) method

Pierre R. Bushel, Jianying Li

May 19, 2021

Contents

1 Introduction

Often times "batch effects" are present in microarray data due to any number of factors, including e.g. a poor experimental design or when the gene expression data is combined from different studies with limited standardization. To estimate the variability of experimental effects including batch, a novel hybrid approach known as principal variance component analysis (PVCA) has been developed. The approach leverages the strengths of two very popular data analysis methods: first, principal component analysis (PCA) is used to efficiently reduce data dimension with maintaining the majority of the variability in the data, and variance components analysis (VCA) fits a mixed linear model using factors of interest as random effects to estimate and partition the total variability. The PVCA approach can be used as a screening tool to determine which sources of variability

(biological, technical or other) are most prominent in a given microarray data set. Using the eigenvalues associated with their corresponding eigenvectors as weights, associated variations of all factors are standardized and the magnitude of each source of variability (including each batch effect) is presented as a proportion of total variance. Although PVCA is a generic approach for quantifying the corresponding proportion of variation of each effect, it can be a handy assessment for estimating batch effect before and after batch normalization.

The pvca package implements the method described in the book Batch Effects and Noise in Microarray Experiment, chapter 12 "Principal Variance Components Analysis: Estimating Batch Effects in Microarray Gene Expression Data": *Jianying Li, Pierre R Bushel, Tzu-Ming Chu, and Russell D Wolfinger 2010*

The pvca method was applied in the paper: *Michael J Boedigheimer, Russell D Wolfinger, Michael B Bass, Pierre R Bushel, Jeff W Chou, Matthew Cooper, J Christopher Corton, Jennifer Fostel, Susan Hester, Janice S Lee, Fenglong Liu, Jie Liu, Hui-Rong Qian, John Quackenbush, Syril Pettit and Karol L Thompson11 2008 "Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories" BMC Genomics 2008, June 12, 9:285*

2 Installation

Simply skip this section if one has been familiar with the usual Bioconductor installation process. Assume that a recent version of R has been correctly installed.

Install the packages from the Bioconductor repository, using the `biocLite` utility. Within R console, type

```
>   if (!requireNamespace("BiocManager", quietly=TRUE))
+       install.packages("BiocManager")
>   BiocManager::install("pvca")
```

Installation using the `biocLite` utility automatically handles the package dependencies. The `pvca` package depends on the packages like `lme4` etc., which can be installed when `pvca` package is stalled.

3 An example run on published Golub data

We use `Golub` dataset in the package `golubEsets` as an example to illustrate the PVCA batch effect estimation procedure. This dataset contains 7129 genes from Microarray

data on 72 samples from a leukemia study. It is a merged dataset and we are testing variability from each factor and their two-ways interactions. The pacakge performs PVCA on this merged data and produces an estimation of the proportion (as possible batch effect) each factor and interaction contribute. The figure shows the proportion in a bar chart. .

```
> library(golubEsets)
> library(pvca)
> data(Golub_Merge)
> pct_threshold <- 0.6
> batch.factors <- c("ALL.AML", "BM.PB", "Source")
> pvcaObj <- pvcaBatchAssess (Golub_Merge, batch.factors, pct_threshold)
>
```

We can plot the source of potential batch effect in proporting as shown in Figure 1.

```
> bp <- barplot(pvcaObj$dat,  xlab = "Effects",
+         ylab = "Weighted average proportion variance",
+         ylim= c(0,1.1),col = c("blue"), las=2,
+         main="PVCA estimation bar chart")
> axis(1, at = bp, labels = pvcaObj$label, xlab = "Effects", cex.axis = 0.5, las=2)
> values = pvcaObj$dat
> new_values = round(values , 3)
> text(bp,pvcaObj$dat,labels = new_values, pos=3, cex = 0.8)
>
```

4 Session

```
>   print(sessionInfo())
```

```
R version 4.1.0 RC (2021-05-16 r80304)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Mojave 10.14.6

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib

locale:
```

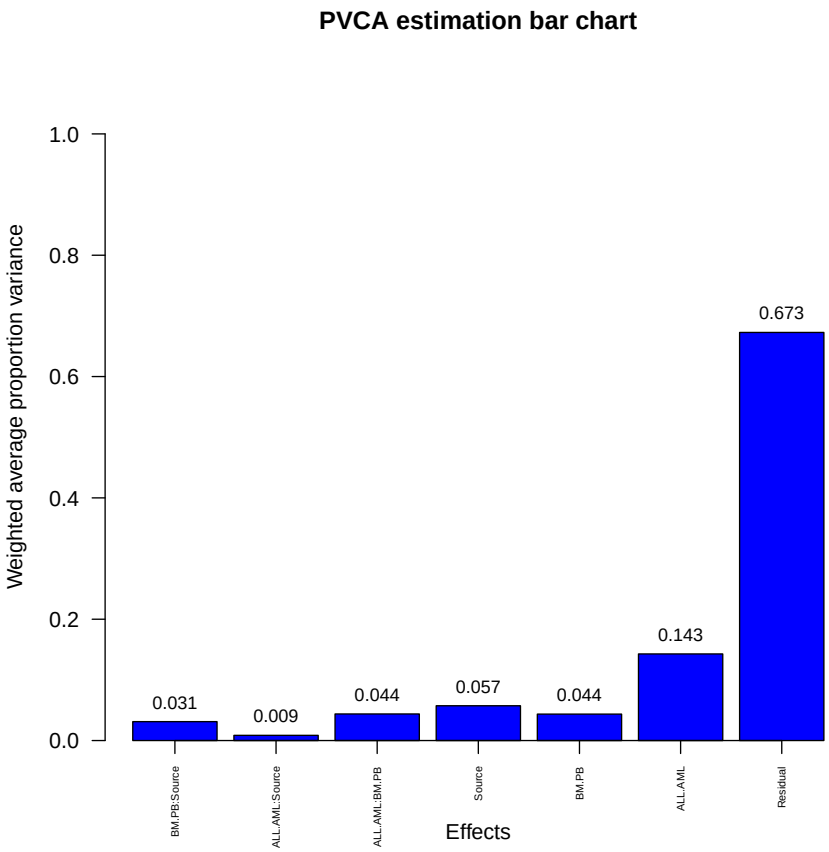

Figure 1: The bar chart shows the proportion of batch effect from possible source.

```
[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] pvca_1.33.0         golubEsets_1.33.0   Biobase_2.53.0
[4] BiocGenerics_0.39.0

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.6             nloptr_1.2.2.2         pillar_1.6.1
 [4] compiler_4.1.0         BiocManager_1.30.15    tools_4.1.0
 [7] zlibbioc_1.39.0        boot_1.3-28            lme4_1.1-27
[10] nlme_3.1-152           lifecycle_1.0.0        tibble_3.1.2
[13] preprocessCore_1.55.0 gtable_0.3.0           lattice_0.20-44
[16] pkgconfig_2.0.3        rlang_0.4.11           Matrix_1.3-3
[19] DBI_1.1.1              dplyr_1.0.6            generics_0.1.0
[22] vctrs_0.3.8            grid_4.1.0             tidyselect_1.1.1
[25] glue_1.4.2             R6_2.5.0               fansi_0.4.2
[28] minqa_1.2.4            limma_3.49.0           ggplot2_3.3.3
[31] purrr_0.3.4            magrittr_2.0.1         MASS_7.3-54
[34] splines_4.1.0          scales_1.1.1           ellipsis_0.3.2
[37] assertthat_0.2.1       colorspace_2.0-1       utf8_1.2.1
[40] affy_1.71.0            munsell_0.5.0          vsn_3.61.0
[43] crayon_1.4.1           affyio_1.63.0
```