We will now perform adjustment to the normalized data by removing the unwanted batch effects. We will be using RUV-2 and RUV-4 methods from the R package `ruv`. For the purpose of estimating the unwanted factors, we need to specify a set of control probes. The HumanMethylation 450k array have a small number of probes that are not positioned in CpG islands but rather these probes are position in the location known to harbor single-nucleotide polymorphism (SNP), a single-base mutation. We do not expect the intensity data from these probes to be affected by the factor of interest (AFB1 levels), so we will be using these probes as control probes.

1. Read in the datasets and the workspace from Lab 2 using the following command:

```
#datasets
load(file= paste('Datasets_','YOURNAME','.Rdata',sep=""))
# read workspace from previous lab
filaneme1 <- paste('Lab2_','YOURNAME','.Rdata',sep="")
load(filename1)
```

2. First, we will perform eigen-decomposition on the inter-sample covariance matrix, where the covariance matrix is calculated using only control probes. We will then plot the cumulative percentage of variations explained by the eigenvalues

```
anno.rs <- anno[grep('rs',anno$Name),]
control <- data$X %in% anno.rs$Name
eigen.control <- eigen(cov(y.norm[control,],use='pairwise'))
pct.var <- cumsum(eigen.control$values)/sum(eigen.control$values)
plot(pct.var)
```

3. From (a), we should finding about 15 eigenvalues explain about 65% of the variations. Hence, we will be performing RUV-2 algorithm, specifying $k = 15$ unwanted factors.

```
# perform RUV-2
require(ruv)
Xmat <- model.matrix(~as.factor(info$SoB)+info$AFB1)
ruv2.adj <- RUV2(Y=t(y.norm),X=Xmat[,-1],ctl=control,k=15)

# plot the first few unwanted factors and check whether there are
# clusters in terms of 'R' and 'C' batches
batch  <- strsplit(as.character(info$sample),split="_")
batch  <- grep('R',unlist(batch),value=T)
batch.R<- as.numeric(substr(batch,3,3))
batch.C<- as.numeric(substr(batch,6,6))
par(mfrow=c(2,2))
 plot(ruv2.adj$W[,1:2],col=batch.C,cex=0.5)
 plot(ruv2.adj$W[,3:4],col=batch.C,cex=0.5)
 plot(ruv2.adj$W[,1:2],col=batch.R,cex=0.5)
 plot(ruv2.adj$W[,3:4],col=batch.R,cex=0.5)
```

4. Use the `qvalue` command from `qvalue` package to estimate the proportion of probes not associated with (a) Season of Birth, (b) AFB1 levels. Compare the estimated proportions for the adjusted and unadjusted data. What can you conclude about the effect of unwanted variations? Plot the relationships between $\pi_0(\lambda)$ and λ for both datasets. After looking at these relationships, do you think the estimates of π_0 for both datasets reasonable?

5. If you still have time perform also the adjustment using RUV-4 method.

6. When you finished don't forget to save the commands you have been executing and the workspace containing all your works, but first you need to remove the three datasets `data,anno,info` so that the workspace size is not overwhelming. The following commands can be used:

```
# remove big datasets
rm(data);rm(anno);rm(info)
# save commands
filaneme1 <- paste('Lab3_','YOURNAME','.R',sep="")
savehistory(filename1)
# save workspace
filaneme2 <- paste('Lab3_','YOURNAME','.RData',sep="")
save.image(filename2)
```