covEB package

1 Introduction

In Bioinformatics, one commonly used tool in differential expression analysis has been an empirical Bayes approach to estimating variances. This method was shown to reduce the false positive rates, particularly at low sample sizes. Small sample sizes are by monetary necessity, common place in experiments. This method, combines information across genes, shrinking the gene -specific variances to a common variance across all genes. Because this approach is used in linear regressions, we hypothesised that a similar methodology could be used with correlations as these are linear regressions between two variables. Correlation matrices are important in inferring relationships and networks between regulatory or signalling elements. As the sample sizes for experiments are small, these correlations can be difficult to estimate and can exhibit high false positive rates. This package is designed to reduce these false positive rates and therefore direct the researcher to higher value relationships that are more likely to be validated experimentally. At a genome-wide scale estimation of correlation matrices can also be computationally demanding. This package provides an empirical Bayes approach to improve covariance estimates for gene expression, where we assume the covariance matrix has block diagonal form. These covariance matrices can be estimates from either microarray or RNA-seq data.

1.1 A Simple Example

We show a simple example of how to run the empirical Bayes estimation, these are trivial examples but serve to illustrate the syntax and parameters of the function. We use the package mvtnorm to simulate data from a multivariate normal distribution.

```
> library(covEB)
>          sigma <- matrix(c(4,2,2,3), ncol=2)
>         x <- rmvnorm(n=500, mean=c(1,2), sigma=sigma)
>          samplecov<-cov(x)
>          test<-covEB(samplecov,delta=0.05,shift=0.025,startlambda=0.4,n=500)
>
>
```

In this example there are three parameters excluding the covariance matrix, samplecov. These are delta, shift and startlambda. These three parameters together control the correlation ranges that define each of the blocks for which the empirical Bayes estimate is calculated. covEB is a novel iterative block diagonal algorithm for estimating the correlation matrix. Instead of one prior matrix for the full correlation matrix, we identify blocks of correlated elements for a given correlation range. Consequently we use a range of shrinkage values ($\theta's$) for each block and calculate the empirical Bayes estimate for each of these. We use a range to allow for noise in the data for example, if we assume the correlation for a block b is θ_b and we there is noise in our data of $delta$ then the range of correlation values would be $(\theta_b - delta/2, \theta_b + delta/2)$. Values outside the current range are set to zero to determine the diagonal block structure. From a starting threshold (startlambda) the two user selected parameters determine the size of the interval (delta) and the size of the shift in moving the starting threshold (shift). This shift value means that the range values used are overlapping rather than mutually exclusive. The block diagonal structure of the correlation matrix is determined using the current range and the empirical Bayesian estimation for each block performed separately. The final estimate is an average of each non zero estimation. Setting $startlambda > 0$ will aid

computation time as any correlations under this level are set to zero and may also be used to reduce noise in the data, by assuming any correlation below the value of startlambda is noise.

1.2 Example with biological data

Here we use a data set available from bioconductor to demonstrate how covEB can be used in the pipeline analysis of gene expression data. We load the data package curatedBladderData that contains gene expression from bladder cancer patients in the R object type expression set. We get the gene expression data matrix, this contains around 5,000 probes from microarrays with 40 samples and store this in the matrix Edata.

```
> library(curatedBladderData)
> data(package="curatedBladderData")
> data(GSE89_eset)
> Edata<-exprs(GSE89_eset)
```

We filter the data to include those that are 'expressed' as defined as being in the top 20th percentile according to variance across samples. This gives us just over a thousand genes, we then calculate the covariance matrix between the genes. This covariance matrix is our input into the covEB function.

```
> variances<-apply(Edata,1,var)
> edata<-Edata[which(variances>quantile(variances,0.8)),]
> covmat<-cov(t(edata))
> cormat<-cov2cor(covmat)
> #we are now able to use covmat as input into covEB:
> out<-covEB(covmat,0.2,0.1,startlambda=0.6,n=40)
Warning: Covariance matrix is not positive semi definite
```

We now provide an example of how the output may be used. We can visualise the correlations between genes using functions available in R and its associated packages. First we use simple thresholding to define significant correlations, we create adjacency matrices for graphs, setting correlations below 0.65 to zero.

```
> outmat<-out
> outmat[abs(out)<0.65]<-0
> outmat[abs(out)>=0.65]<-1
```

We find connected subgraphs in the adjacency matrix using the clusters function and then select one of the subgraphs (number 5) that has 12 genes in it. Finally, for visualisation purposes, we remove edges between nodes and themselves (i.e. the diagonal)

```
> clusth<-clusters(graph.adjacency(outmat))
> sel<-which(clusth$membership==5)
> subgraphEB<-outmat[sel,sel]
> subgraph<-cormat[sel,sel]
> subgraph[subgraph<0.65]<-0
> subgraph[subgraph>=0.65]<-1
> diag(subgraph)<-0
> diag(subgraphEB)<-0
```

We can now plot these graphs, there are 6 fewer edges after using covEB. On a larger network this would help the interpretability of the model further.

```
> plot(graph.adjacency(subgraph,mode="undirected"))
> plot(graph.adjacency(subgraphEB,mode="undirected"))
```

2 References

Champion, C. J. (2003). Empirical Bayesian estimation of normal variances and covariances. Journal of Multivariate Analysis, 87(1), 60-79. http://doi.org/10.1016/S0047-259X(02)00076-3

Benjamin Frederick Ganzfried, Markus Riester, Benjamin Haibe-Kains, Thomas Risch, Svitlana Tyekucheva, Ina Jazic, Xin Victoria Wang, Mahnaz Ahmadifar, Michael Birrer, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron. curatedOvarianData: Clinically Annotated Data for the Ovarian Cancer Transcriptome, Database 2013: bat013 doi:10.1093/database/bat013 published online April 2, 2013.