A correction for the LPE statistical test

Carl Murie <carl.murie@mcgill.ca>,
Robert Nadon <robert.nadon@mcgill.ca>

April 12, 2014

Contents

1 Introduction

LPEadj is a correction for the LPE statistical test [1] which consists of two additions to the LPE method. The LPE package documentation is still correct with the exception of the two additions listed below and should be consulted for more information on the LPE method.

The correction is in two parts. See [3] for more information on the correction.

1. The LPEadj method discontinues the LPE practice of setting all variances below the maximum variance in the ordered distribution of variances to the maximum variance. In certain cases this practice can set many variances to the maximum and lower the performance of this algorithm. If the assumption that there are only a few low variances to be adjusted is correct then it may be safe to use this procedure. This option is controlled by the doMax parameter (default is FALSE).

2. The LPEadj method replaces the $\pi/2$ adjustment of the variance with an empirically estimated adjustment based on the sample size of the data. The empirical adjustment values have been estimated for replicate sizes up to 10. This option is controlled by the doAdj parameter (default is TRUE).

The top level function LPEadj executes the first and second step by default. This is the recommended manner in which LPEadj should be run. One can use steps 1 and 2 independently if desired.

2 LPE variance adjustment

The LPE method pools variance estimates of genes with similar intensities in order to gain an improved error estimate and increased degrees of freedom. A calibration curve of variance versus mean intensity is generated for each group and the gene specific median intensity is used to obtain the gene's variance estimate from the calibration curve. It has been shown that the sampling variability of the median is proportionally higher (by $\pi/2$) than that of the mean [2]. Accordingly, a multiplicative adjustment of $\pi/2$ is applied to the variance estimate obtained from the calibration curve for the purpose of statistical testing.

The LPE z-statistic is as follows:

$$z = \frac{Med_1 - Med_2}{\sigma_{pool}} \tag{1}$$

where

$$\sigma^2_{pool} = \frac{\pi}{2}\left(\frac{\sigma^2_1(Med_1)}{n_1} + \frac{\sigma^2_2(Med_2)}{n_2}\right) \tag{2}$$

$\sigma^2_i(Med_i)$ are the variances derived from the calibration curve using the median of the gene intensities for a particular group. n is the number of replicates in the groups (assuming equal sample size). The associated probability of the z-statistic under the null hypothesis is calculated by reference to the standard normal distribution.

The Mood [2] proof shows that with normal data the ratio of the squared standard error of the median relative to that of the mean is asymptotically $\pi/2$. Figure 1 shows that the ratio converges to $\pi/2$ when the sample size is large, around 100, but is less than $\pi/2$ when the sample sizes are small, from three to ten. The ratio of variances at small sample sizes also oscillates lower to higher depending on whether the sample size is even or odd. This fluctuation is due to the difference in obtaining the median with even and odd sample sizes. The middle value of the ordered distribution is used as the median with odd sample sizes while the mean of the two middle values of the ordered distribution is used with even sample sizes. There is higher variability when taking the middle value of a distribution (with odd number of samples) than taking the average of the two middle values (with even number of samples).

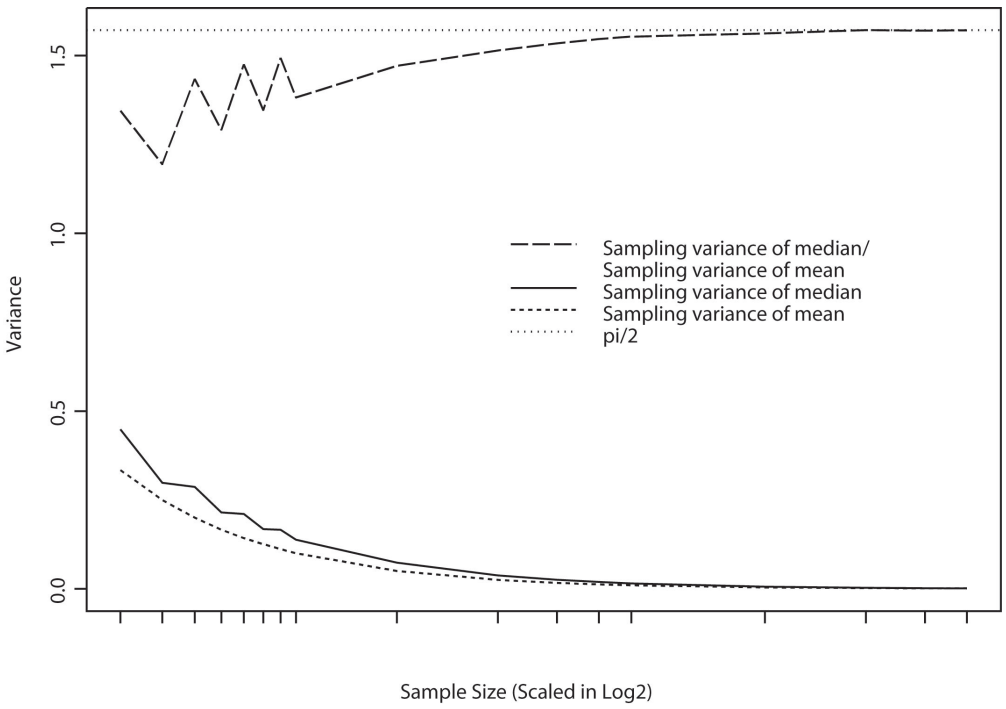

Figure 1: The ratio of the variance of sampling medians over sampling means across a range of sample sizes. The Sampling variance of mean line is the variance of taking the mean of a random sample from a standard normal distribution. This line corresponds to the Central Limit Theorem, which states that the sample variance of a distribution of means is σ^2/N. The Sampling variance of median line is the variance of taking the median from the same random samples. The Sampling variance of median/Sampling variance of mean line is the variance of the median divided by the variance of the mean at each sample size. The sampling was repeated 1000 times for each sample size (ranging from 3 to 1000).

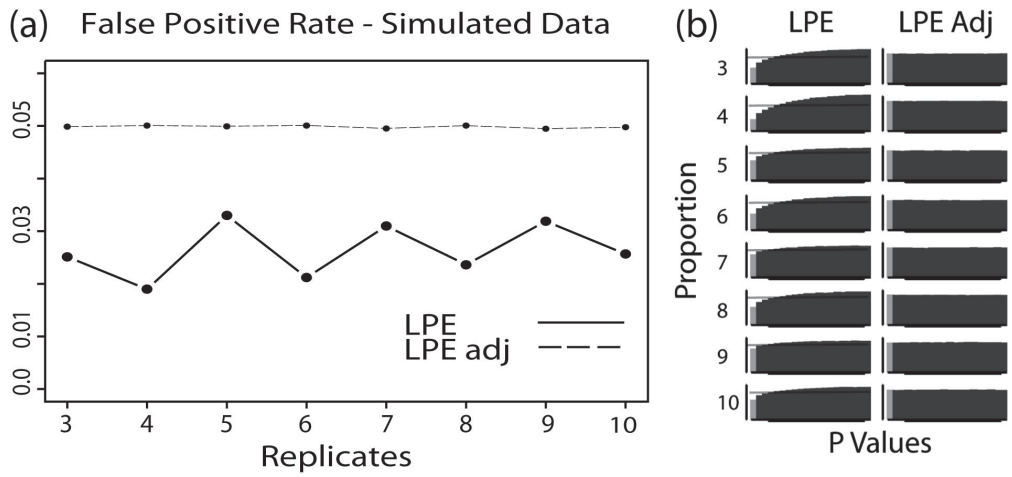

Figure 2: (a) False positive rates for the LPE and adjusted LPE methods using simulated data with no differentially expressed genes evaluated at p $\leq$ .05 threshold. The LPE showed variable and low false positive rates. In contrast, the adjusted LPE showed appropriate false positive rate for all sample sizes. (b) The adjusted LPE, but not the LPE, shows the theoretically expected uniform p-value distribution. Each data set had 10000 genes with each gene's replicate intensity drawn from a $N(\mu, 0.1)$ distribution. μ was drawn from a $N(7, 1)$ distribution.

3 Modification of LPE Method

The use of an empirically estimated variance ratio adjustment, c_i, based on sample size can correct the bias caused by the $\pi/2$ adjustment. The $\pi/2$ term in Equation 1 is replaced by the empirically generated ratio of the variance of sampling a median over the variance of sampling a mean. Equation 1 then becomes:

$$\sigma^2_{pool} = c_1 \frac{\sigma^2_1(Med_1)}{n} + c_2 \frac{\sigma^2_2(Med_2)}{n} \tag{3}$$

The parameters, c_1 and c_2, are the ratio of variances of sampling the median and mean based on the number of replicates for each group. c_1 and c_2 are the adjust1 and adjust2 variables in the calculateLpeAdj function.

Figure 2 shows that the LPE test has a lower than expected false positive rate (FPR) which fluctuates between even and odd sample sizes (average FPR with odd and even samples sizes is 0.030 and 0.022 respectively) in a similar manner as the ratio of variances in Figure 1. The LPE method also shows a non-uniform p-value distribution with fewer than expected small p-values. The $\pi/2$ adjustment increases the variance by an overly large proportion and causes the LPE test statistics to be smaller than they should be and skews the p-value distribution leftward. In contrast, the adjusted LPE test produced theoretically expected values.

Figure 3 summarizes the results of the LPE and adjusted LPE methods applied to the

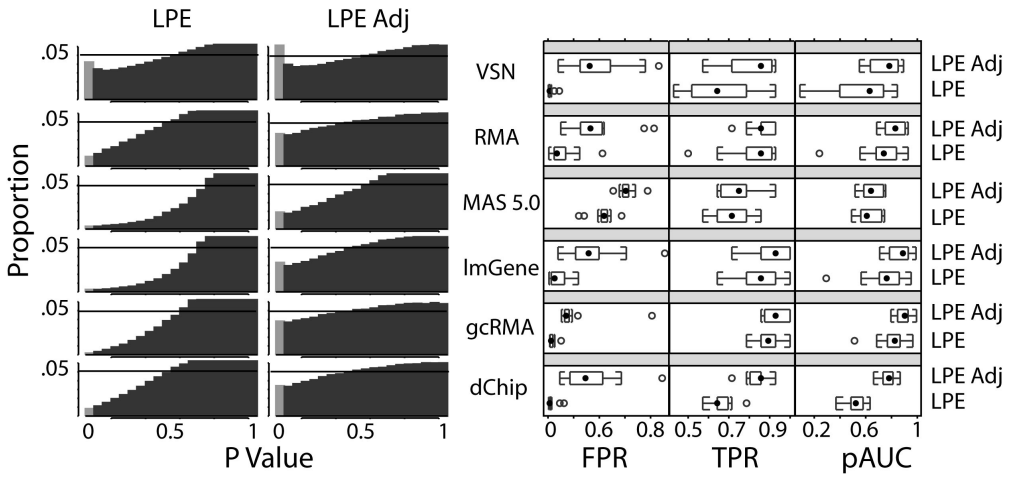

Figure 3: P-value histograms and boxplots of FPR, TPR, and pAUC from the LPE and adjusted LPE methods applied to the HGU95 latin square data set. The data were normalized using six different normalization methods (labeled by row).

HGU95 Affymetrix spike-in data set (www.affymetrx.com). The HGU95 data is based on a 14 x 14 Latin Square design of "spiked-in" transcripts (14 concentrations per microarray chip x 14 groups) with three replicates for each group. The concentrations for the "spiked-in" transcripts were doubled for each consecutive group (0 and 0.25 to 1024 pM inclusive). To assess the performance of the statistical tests we used the FPR, the true positive rate (TPR, which is the proportion of transcripts correctly identified as being differentially expressed), and the partial area under the curve (pAUC, which measures the area under a Receiver Operator Characteristic curve (ROC) below a false positive cutoff of 0.05). The pAUC has a value between 0 (worst performance) and 1 (perfect performance).

4 Example

The easiest way to apply the LPEadj statistical test is to use the LPEadj function. Setting doMax to false (the default value) stops the LPE method from setting the variances of low intensity genes to the maximum variance. Setting doAdj to true (the default value) makes LPE use a variance adjustment based on the number of replicates of each group rather than $\pi/2$. The code will print the values used for the variance adjustment. The variance adjustment values for replicate sizes up to 10 are precalculated within the lpeAdj function. For example if there are two groups with three replicates each the following line will be printed.

```
> # Loading the library and null dataset (two groups with three
> # replicates each).
> library(LPEadj)
> dat <- matrix(rnorm(6000), ncol=6)
```

```
>  # Applying LPE
>  lpe.result <- lpeAdj(dat, labels=c(0,0,0,1,1,1), doMax=FALSE, doAdj=TRUE)

variance adjustment values used: group 1:  1.345859  group 2 1.345859
```

If you want more control over low level variables you can call calculateLpeAdj directly. Note that in this case the results of this call are the same as the previous call to lpeAdj because the variance adjustment values are identical (the variance adjustment values defined in ADJ.VALUES is the same as the variance adjustment values defined in lpeAdj).

```
> # Loading the library and null dataset (two groups with three
>  # replicates each)
>  library(LPEadj)
>  dat <- matrix(rnorm(6000), ncol=6)
>  ADJ.VALUES <- c(1, 1, 1.34585905516761 ,1.19363228146169 ,1.436849413109
+                  ,1.289652132873 ,1.47658053092781 ,1.34382984852146
+                  ,1.49972130857404, 1.3835405678718)
>  # calculate base line error distributions
>  var1 <- adjBaseOlig.error(dat[,1:3], setMax1=FALSE, q=.05)
>  var2 <- adjBaseOlig.error(dat[,4:6], setMax1=FALSE, q=.05)
>  # The correct variance adjustments can be fetched using the replicate
>  # number for each group as in index for the ADJ.VALUES vector.
>  # eg: ADJ.VALUES[n] if there are n replicates in a group
>  results <- calculateLpeAdj(dat[,1:3],dat[,4:6],var1,var2,
+                  probe.set.name=c(1:1000), adjust1=ADJ.VALUES[3],
+                  adjust2=ADJ.VALUES[3])
```

References

[1] Jain et. al. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics*, 2003, Vol 19, No. 15, pp: 1945-1951.

[2] Mood et. al. Introduction to the theory of statistics Mcgraw-Hill, New York, 3rd Ed, 1974

[3] Murie et. al. A correction for estimating error when using the Local Pooled Error statistical test, *Bioinformatics*, in press.