How to use the PLPE Package

HyungJun Cho, and Jae K. Lee

October 29, 2019

Contents

1 Introduction

One of the critical demands in current proteomic research is the comparison of two or more complex samples in order to determine which proteins are differentially expressed. The *PLPE* package is designed to examine paired two groups with high-throughput data, such as mass spectrometry (MS) proteomic data and cDNA microarray data by using Paired t-test , Paired L-statistic and Paired L_w-statistic with their FDRs (Cho et al., 2007).

2 Methods

Suppose x_{ij} and y_{ij} are (a peptide) ion intensities for two conditions x and y, where replicates $i = 1, 2, \ldots, n$ and peptides (or proteins) $j = 1, 2, \ldots, m$. Note that prior to analysis the data may be log2-transformed in order to remedy the highly right-skewed distribution of protein intensity values

2.1 Paired t-test

Then for each protein j, the paired t-test statistic is:

$$t_j = \frac{d_j}{\sqrt{s_j^2/n}} \tag{1}$$

where $d_{ij} = (x_{ij} - y_{ij})$, $d_j = \sum_i d_{ij}/n$, and $s_j^2 = \sum_i (d_{ij} - d_j)/(n-1)$. The statistical significance of each protein can then be obtained from the observed t-statistics of which the p-value is often adjusted for multiple comparisons. Note that the sample variance s_j^2 is derived based only on the replicated observations of peptide j, which can be considerably variable and inaccurate with a small sample size. Due to this, the paired t-test is often underpowered and unreliable when data is lowly replicated.

2.2 Paired L-statistic

In order to more reliably identify differentially expressed peptides from pair-labeled LC-MS/MS data, the L-statistic is defined as:

$$L_j = \frac{\delta_j}{\sqrt{\tau_j^2}} \tag{2}$$

where δ_j is the median of paired differences. which reduces the effect of outliers. Borrowing the error information of adjacent-intensity proteins, the variance (τ_j^2) is estimated based on local pooled error estimates (Cho et al. 2007).

2.3 Paired L_w-statistic

The variance estimate for the above L-statistic is based solely on the pooled error variance of adjacent intensity proteins. While the LPE estimate has a shrinkage effect toward the mean of (local) error variances, this effect is not sensitive enough to capture the innate biological variability of individual proteins among different biological subjects. Thus, in order to optimize the error estimates between individual and LPEs, we introduce the Lw-statistic which uses a weighted variance estimate between the two variance estimates. That is, the Lw-statistic on the paired LC-MS/MS data with a weight, w, is defined as:

$$L_{wj} = \frac{\delta_j}{\sqrt{(1-w)\tau_j^2 + w s_j^2/n}} \tag{3}$$

where $0 \leq w \leq 1$ is the weight parameter between individual variance estimates or pooling variance estimates.

2.4 False discovery rates (FDRs)

Raw p-values corresponding to the above L- or L_w-statistics can be obtained for all observed proteins if an underlying data distribution is assumed to be well-behaved, e.g., a Gaussian distribution. We control the FDR to determine a threshold of L- or Lw-statistics based on a rank-invariant resampling technique. We estimate FDRs using the resampled null data sets, as decribed in Cho et al. (2007).

3 Example

For demonstration, we use LC-MS/MS data for platelet MPs. This data set consists of 62 peptides and three replicates of two different paired samples. For details, refer to the paper of Garcia et al. (2005).

To run *PLPE*, the data can be prepared as follows.

```
> library(PLPE)
> data(plateletSet)
> x <- exprs(plateletSet)
> x <- log2(x) #and any normalization
> cond <- c(1, 2, 1, 2, 1, 2)   #two different samples
> pair <- c(1, 1, 2, 2, 3, 3)   #pairing
> design <- cbind(cond, pair)
```

The above data was log-transformed with base 2, assuming to be normalized by an appropriate method. Two different samples and their pairing are indicated in the design matrix. Thus, data and design matrices (x and *design*) are the required inputs for the main function `lpe.paired`. Another useful argument is q, which is the percentage of interval partitions for pooling peptides with similar intensities. The value 0.1 indicates that each interval contains 10% of the data. The details can be found at the paper of Cho et al. (2007). The test statistics and false discovery rates (FDRs) can be computed by the following commands.

```
> out <- lpe.paired(x=x, design=design, q=0.1,data.type="ms") #Compute test statistic
> out.fdr <- lpe.paired.fdr(x, obj=out) #Compute FDRs
> out$test.out[1:10,]
```

```
          A           M     var.L      L.stat p.value.L       var.t
1  24.44489 -0.68674738 0.2043176 -1.51930205 0.1286865 0.118090527
2  24.26340  0.10855622 0.2043176  0.24016064 0.8102057 0.017343090
3  20.85696  0.02931575 0.6063909  0.03764651 0.9699695 0.170649087
4  21.06345  0.38155812 0.5555438  0.51191936 0.6087074 0.233067978
5  20.12792  1.31553951 0.8527625  1.42458985 0.1542758 0.517247714
6  21.52744  0.37835569 0.5075633  0.53107416 0.5953674 0.049571653
```

```
7  21.51976 -0.37160470 0.5076349 -0.52156143 0.6019757 0.279125178
8  19.68920  0.96394185 1.0237903  0.95267622 0.3407541 0.319240267
9  21.56502 -1.10398634 0.5074147 -1.54982326 0.1211839 0.013898401
10 21.72789  0.19906429 0.5090750  0.27899889 0.7802457 0.004570286
         t.stat   p.value.t    var.Lw    Lw.stat p.value.Lw
1  -3.543818738  0.07122412 0.1612041 -1.7104446 0.08718368
2            NA  2.00000000 0.1108303  0.3260809 0.74436313
3            NA  0.14514775 0.3885200  0.0470321 0.96248765
4            NA          NA 0.3943059  0.6076368 0.54342838
5   0.278728050          NA 0.6850051  1.5894870 0.11195048
6   0.000000000          NA 0.2785675  0.7168612 0.47345975
7   0.281718683  0.12854526 0.3933800 -0.5924818 0.55352798
8   1.203655035  0.00000000 0.6715153  1.1763130 0.23946984
9  -0.476697864  0.92687351 0.2606566 -2.1623669 0.03058991
10  0.003959108 -0.78566721 0.2568226  0.3928047 0.69446371
```

```
> out.fdr$FDR[1:10,]
```

```
              L     FDR.L          Lw   FDR.Lw
1  -1.51930205 0.2037037 -1.7104446 1.037037
2   0.24016064 1.0185185  0.3260809 1.037037
3   0.03764651 1.0015432  0.0470321 1.003036
4   0.51191936 1.0185185  0.6076368 1.037037
5   1.42458985 0.3395062  1.5894870 1.037037
6   0.53107416 1.0185185  0.7168612 1.037037
7  -0.52156143 1.0185185 -0.5924818 1.037037
8   0.95267622 1.0185185  1.1763130 1.037037
9  -1.54982326 0.2546296 -2.1623669 0.000000
10  0.27899889 1.0185185  0.3928047 1.037037
```

```
>
```

The output from `lpe.paired` contains MA tranformed data and several test statistics, including their variance estimates and naive p-values. The Lw statistics are computed by the weighted average of the variance estimates for the L- and t-tests (Cho et al. 2007). The default for the weight is 0.5, which can adjusted by a user. The output from `lpe.paired.fdr` contains FDRs for the L- and Lw-test, including their test statistics. Choosing a small FDR value, we can determine a corresponding cutoff value of the statistics.

Reference

Cho H, Smalley DM, Ross MM, Theodorescu D, Ley K and Lee JK (2007). Statistical Identification of Differentially Labelled Peptides from Liquid Chromatography Tandem Mass Spectrometry, Proteomics, 7:3681-3692.

Garcia BA, Smalley DM, Cho H, Shabanowitz J, Ley K and Hunt DF (2005). The Platelet Microparticle Proteome, Journal of Proteome Research, 4:1516-1521.