Polyfit Vignette

Conrad Burden

October 13, 2014

Polyfit [2] is an add-on to the negative-binomial based package `DESeq` [1] for two-class detection of differential expression. Its purpose is to ensure the p-value distribution is close to uniform over the interval [0, 1] for the subset of genes satisfying the null hypothesis of no differential expression. The first component of Polyfit is the function `pfNbinomTest` which replaces the function `nbinomTest` in `DESeq`. Its purpose is to smooth point singularities (or 'flagpoles'), particularly one at $p = 1$, in the p-value distribution caused by calculating calculating p-values from a discrete distribution. The output from this function should then be passed to the second component, the function `levelPValues`. Its purpose is to apply a variant of the Storey-Tibshirani procedure [3] to shift the p-values so that those corresponding to the null hypothesis have a unform distribution, and to calculate corresponding q-values (or 'adjusted p-values') for controlling errors via the false discovery rate.

To load and attach Polyfit, type

```
> library(Polyfit)
```

at the R prompt. edgeR and DESeq are dependencies and will be automatically loaded.

Removing the flagpoles

When calculating p-values, DESeq assumes as a null hypothesis that the total number of counts K_A and K_B summed over replicates in each of conditions A and B is distributed according to a negative binomial distribution with parameters estimated from the data. The distribution of K_A, conditonal on the observed value $k_A + k_B$ of the sum of counts in both conditions is thus a discrete distribution. DESeq calculates p-values as the sum of probabilities from this distribution less than or equal to the probability of the observed counts (k_A, k_B) (see Fig. 1). This method invariably causes a spikes in the histogram of generated p-values, the most notable one being at $p = 1$ from those observed counts which happen to hit the mode of the null hypothesis distribution.

Because it needs to fit a smooth polynomial to the p-value histogram, Polyfit redistributes the spikes by replacing the discrete distribution with a "squared off" continuous

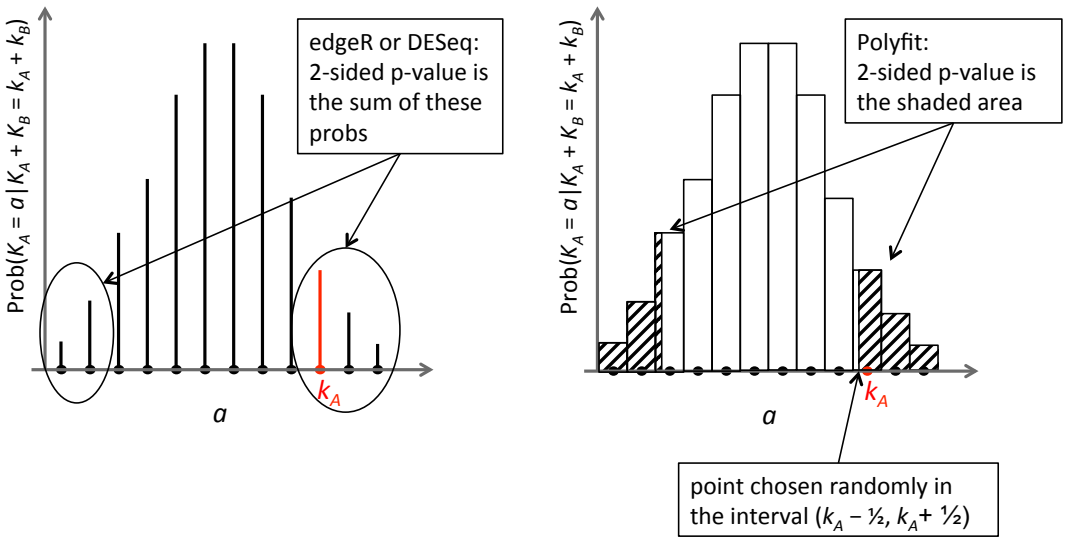

Figure 1: Calculation of p-values by the DESeq functions `nbinomTest` (left) and the replacement Polyfit function `pfNbinomTest` (right).

distribution, as shown in the right half of Fig. 1. If the data are generated according to the postulated null hypothesis, the p-values will then have a uniform distribution on $[0, 1]$.

In the following example we use simulated data generated by the DESeq function `makeExampleCountDataSet()`. To generate p-values with the flagploes removed, replace the DESeq function `nbinomTest` with its Polyfit replacement `pfNbinomTest`.

```
> cds <- makeExampleCountDataSet()
> cds <- estimateSizeFactors( cds )
> cds <- estimateDispersions( cds )
> nbT <- nbinomTest( cds, "A", "B" )
> pValuesDESeq <- nbT$pval # <-- Original DESeq code
> nbTPolyfit <- pfNbinomTest( cds, "A", "B" )
> pValuesPFDESeq <- nbTPolyfit$pval # <-- Polyfit replacement
```

Histograms of the resulting p-values are shown in Fig. 2.

Levelling the p-value histogram

Because the parameters of the negative binomial distribution for each gene must be estimated from the available count data, p-values reported by DESeq for those genes

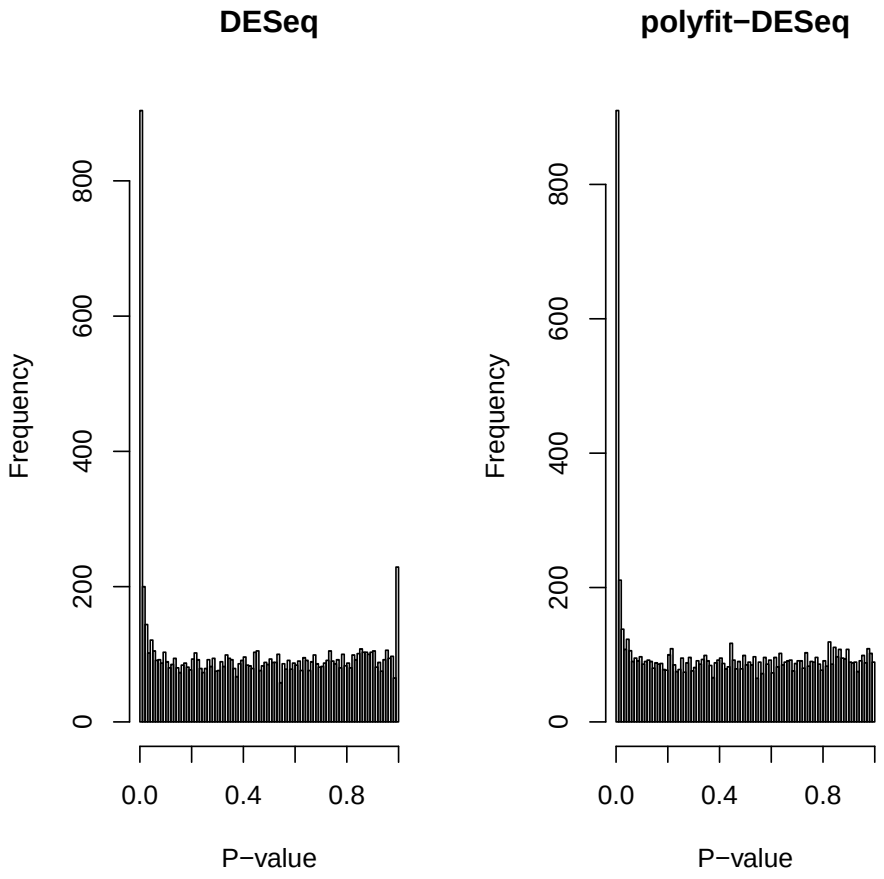

Figure 2: Histogram of p-values generated by `nbinomTest` and `pfNbinomTest`

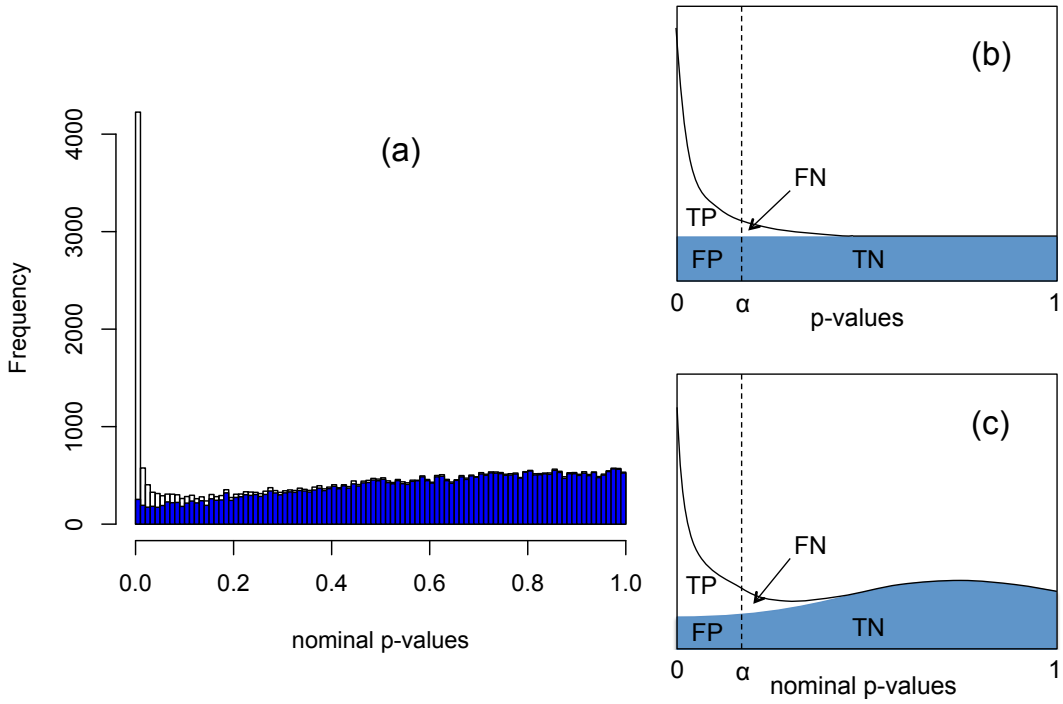

Figure 3: (a) Example p-value spectrum calculated by DESeq for synthetic data RNA-seq with 15% genes up- or down-regulated after removal of 'flagpoles' (taken from ref. [2]). The shaded histogram is the 85% of transcripts which are unregulated. (b) Schematic representation of the Storey-Tibshirani procedure for estimating false discovery rates, assuming correctly calculated p-values. (c) Schematic representation of the analogous Polyfit procedure. (TP = true positives, FP = false positives, FN = false negatives and TN = true negatives at a specified significance point α.)

which are not differentially expressed may not be uniformly distributed. A fairly extreme case is shown in Fig. 3(a) generated from DESeq using synthetic data after the flagpole has been removed as described above.

If p-values have been calculated exactly, the Storey-Tibshirani procedure calculates q-values, that is, estimates of the false discovery rate, essentially by fitting a uniform distribution to the right hand part of the p-value histogram (see Fig. 3(b)). Polyfit replaces the uniform distribution fit with a polynomial fit (see Fig. 3(c)), and estimates p-values and q-values for each gene by the formulae

$$\text{corrected p-value} = \frac{\text{FP}}{\text{FP} + \text{TN}}\,, \qquad \text{corrected q-value} = \frac{\text{FP}}{\text{FP} + \text{TP}}\,.$$

The function `levelPValues` generates the levelled p-values, estimated q-values calculated from the adapted Storey-Tibshirani procedure and, for comparison, also reports q-values calculated from the levelled p-values using the Benjamini-Hochberg procedure.

The following code calculates the corrected p-values and q-values from the nominal p-values generated in the DESeq example above:

```
> lP <- levelPValues(pValuesPFDESeq)
> outTable <- cbind(origPval= pValuesPFDESeq,
+                   levelledPval=lP$pValueCorr,
+                   levelledQval=lP$qValueCorr,
+                   BH_Qval=lP$qValueCorrBH)
> head(outTable)

       origPval levelledPval levelledQval   BH_Qval
[1,] 0.88318253   0.87550083    0.8767728 0.9816269
[2,] 0.31129123   0.30716855    0.7227408 0.8092118
[3,] 0.02410326   0.02430328    0.1847991 0.2069091
[4,] 0.08661783   0.08681295    0.4204964 0.4707195
[5,] 0.70459987   0.69269211    0.8548007 0.9565190
[6,] 0.59753546   0.58626342    0.8301943 0.9292607
```

If the option `plot=TRUE` is used a diagnostic plot in the format of Fig. 4 showing the p-value distribution before and after levelling is produced. The top left hand panel plots estimates of the fraction (π_0) of genes not DE obtained by fitting a quadratic to the nominal p-value historgam (without flagpoles) over the interval $[\lambda, 1]$. Beneath this is a density plot of obtained estimates $\hat{\pi}_0$. The optimal λ (the red dot) is obtained by choosing the $\hat{\pi}_0(\lambda)$ closest to the mode of the $\hat{\pi}_0$ density. The mode is also indicated by the dotted line in the top left panel. The original and corrected p-value histograms are shown on the right, together with optimally fitted quadratic (upper plot) and its image after correction (lower plot). The red part of the quadratic and its image below correspond to the interval over which the quadratic is fitted.

References

[1] Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106 (2010)

[2] Burden, C.J., Qureshi, S. and Wilson, S.R. Error estimates for the analysis of differential expression from RNA-seq count data. *PeerJ PrePrints 2:e400v1*, `http://dx.doi.org/10.7287/peerj.preprints.400v1` (2014)

[3] Storey, J. and Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Science*, 100(16):9440–9445 (2003)

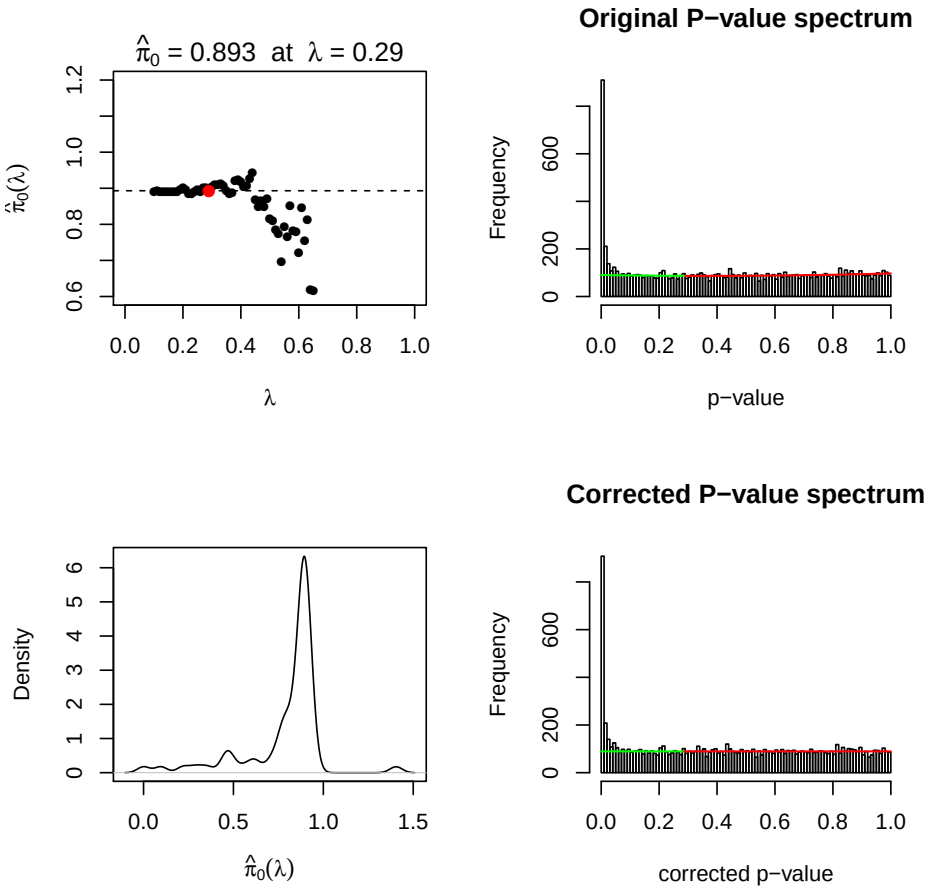

Figure 4: Diagnostic plot produced by `levelPValues`