Moment based gene set enrichment testing – the npGSEA package

Jessica L. Larson[1*] and Art Owen[2]

[1] Department of Bioinformatics and Computational Biology, Genentech, Inc.

[2] Department of Statistics, Stanford University

[*] larson.jessica (at) gene.com

April 24, 2017

Contents

1 Introduction

Gene set methods are critical to the analysis of gene expression data. The npGSEA package provides methods to run permutation-based gene set enrichment analyses without the typically computationally expensive permutation cost. These methods allow users to adjust for covariates and approximate corresponding permutation distributions. We are currently evaluating the applicability and accuracy of our method for RNA-seq expression data.

Our methods find the exact relevant moments of a weighted sum of (squared) test statistics under permutation, taking into account correlations among the test statistics. We find moment-based gene set enrichment p-values that closely approximate the permutation method p-values.

This vignette describes a typical analysis workflow and includes some information about the statistical theory behind npGSEA. For more technical details, please see Larson and Owen, 2015 .

2 Example workflow for GSEA

2.1 Preparing our gene sets and our dataset for analysis

For our example, we will use the ALL dataset. We begin by loading relevant libraries, subsetting the data, and running featureFilter on this data set. For details on these methods, please see the limma manual.

```
> library(ALL)
> library(hgu95av2.db)
> library(genefilter)
> library(limma)
> library(GSEABase)
> library(npGSEA)
> data(ALL)
> ALL <- ALL[, ALL$mol.biol %in% c('NEG','BCR/ABL') &
+      !is.na(ALL$sex)]
> ALL$mol.biol <- factor(ALL$mol.biol,
+      levels = c('NEG', 'BCR/ABL'))
> ALL <- featureFilter(ALL)
```

We adjust the feature names of the ALL dataset so that they match the names of our gene sets below. We convert them to entrez ids.

```
> featureNames(ALL) <- select(hgu95av2.db, featureNames(ALL),
+                             "ENTREZID", "PROBEID")$ENTREZID
```

We now make four arbitrary gene sets by randomly selecting from the genes in our universe.

```
> xData <- exprs(ALL)
> geneEids <- rownames(xData)
```

```
> set.seed(12345)
> set1 <- GeneSet(geneIds=sample(geneEids,15, replace=FALSE),
+              setName="set1",
+              shortDescription="This is set1")
> set2 <- GeneSet(geneIds=sample(geneEids,50, replace=FALSE),
+              setName="set2",
+              shortDescription="This is set2")
> set3 <- GeneSet(geneIds=sample(geneEids,100, replace=FALSE),
+              setName="set3",
+              shortDescription="This is set3")
> set4 <- GeneSet(geneIds=sample(geneEids,500, replace=FALSE),
+             setName="set4",
+             shortDescription="This is set4")
```

As a positive control, we also make three gene sets that include our top differentially expressed genes.

```
> model <- model.matrix(~mol.biol, ALL)
> fit  <-  eBayes(lmFit(ALL, model))
> tt <- topTable(fit, coef=2, n=200)
> ttUp <- tt[which(tt$logFC >0), ]
> ttDown <- tt[which(tt$logFC <0), ]
> set5 <- GeneSet(geneIds=rownames(ttUp)[1:20],
+         setName="set5",
+         shortDescription="This is a true set of the top 20 DE
+         genes with a positive fold change")
> set6 <- GeneSet(geneIds=rownames(ttDown)[1:20],
+         setName="set6",
+         shortDescription="This is a true set of the top 20 DE genes
+         with a negative fold change")
> set7 <- GeneSet(geneIds=c(rownames(ttUp)[1:10], rownames(ttDown)[1:10]),
+         setName="set7",
+         shortDescription="This is a true set of the top 10 DE genes
+         with a positive and a negative fold change")
```

We then collapse all of our gene sets into a `GeneSetCollection`. For more information on `GeneSets` and `GeneSetCollections`, see the `GSEABase` manual.

```
> gsc <- GeneSetCollection( c(set1, set2, set3, set4, set5, set6, set7) )
> gsc
```

```
GeneSetCollection
  names: set1, set2, ..., set7 (7 total)
  unique identifiers: 6527, 8600, ..., 2625 (686 total)
  types in collection:
    geneIdType: NullIdentifier (1 total)
    collectionType: NullCollection (1 total)
```

2.2 Running npGSEA

Now that we have both our gene sets and experiment, we are ready to run npGSEA and determine the level of enrichment in our experiment. We can use npGSEA with our eset or expression data (xData) directly. We call npGSEASummary to get a summary of the results. T_Gw is explained in more detail in Section 3.2.

```
> yFactor <- ALL$mol.biol
> res1 <- npGSEA(x = ALL, y = yFactor, set = set1)  ##with the eset
> res1
```

```
Normal Approximation for set1
T_Gw = 0.107
var(T_Gw) = 0.0239
pLeft = 0.755, pRight = 0.245, pTwoSided = 0.49
```

```
> res2_exprs <- npGSEA(xData, ALL$mol.biol, gsc[[2]])  ##with the expression data
> res2_exprs
```

```
Normal Approximation for set2
T_Gw = 0.159
var(T_Gw) = 0.0765
pLeft = 0.717, pRight = 0.283, pTwoSided = 0.566
```

npGSEA has several built in accessor functions to gather more information about the analysis of your set of interest in your experiment.

```
> res3 <- npGSEA(ALL, yFactor, set3)
> res3
```

```
Normal Approximation for set3
T_Gw = 0.356
var(T_Gw) = 0.444
pLeft = 0.703, pRight = 0.297, pTwoSided = 0.593
```

```
> geneSetName(res3)
```

```
[1] "set3"
```

```
> stat(res3)
```

```
[1] 0.3560641
```

```
> sigmaSq(res3)
```

```
[1] 0.4440907
```

```
> zStat(res3)
```

```
[1] 0.5343088
```

```
> pTwoSided(res3)
```

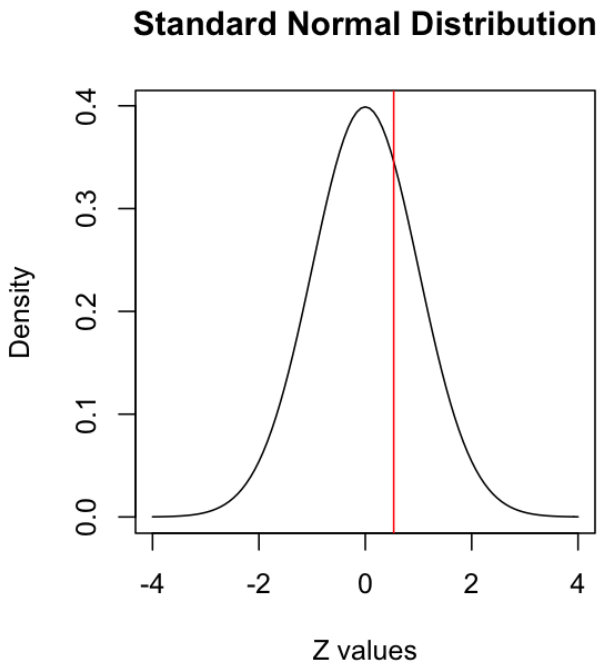

Figure 1: **Set3 normal approximation results** This plot displays the standard normal curve and our observed zStat for set3 in this analysis.

```
  [1] 0.5931279
> pLeft(res3)
  [1] 0.7034361
> pValues(res3)
  pLeft = 0.703, pRight = 0.297, pTwoSided = 0.593
> dim(xSet(res3))
  [1] 100 109
```

There is also a `npGSEA` specific plot function (`npGSEAPlot`) to visualize the results of your analysis. Highlighted in red on the plot is the corresponding `zStat` of our analysis.

```
> npGSEAPlot(res3)
```

2.3 Running npGSEA with the beta and chi-sq approximations

There are three types of approximation methods in `npGSEA`: "norm", "beta", and "chiSq". Each method is discussed in brief in Section 3. The "norm" approximation method is the default. Note that each of these methods has the same $\hat{\beta}_g$ (see methods section).

```
> res5_norm <- npGSEA(ALL, yFactor, set5, approx= "norm")
> res5_norm
```

```
Normal Approximation for set5
T_Gw = 5
var(T_Gw) = 0.394
pLeft = 1, pRight = 8.72e-16, pTwoSided = 1.74e-15
```

```
> betaHats(res5_norm)
```

```
             [,1]
1490   0.53814745
168544 0.10183475
1893   0.23559626
2013   0.11853248
2022   0.23358385
216    0.19554569
2273   0.35939613
23179  0.24618755
25     0.21597396
2549   0.21100011
2934   0.33968229
5445   0.23247513
55884  0.16183870
57556  0.37650630
687    0.46075522
6915   0.10897784
7277   0.34083849
92     0.15933797
9369   0.09865998
9788   0.26243503
```

```
> npGSEAPlot(res5_norm)
```

The beta approximation yields results quite similar to the normal approximation.

```
> res5_beta <- npGSEA(ALL, yFactor, set5, approx= "beta")
> res5_beta
```

```
Beta Approximation for set5
T_Gw = 5
var(T_Gw) = 0.394
pLeft = 1, pRight = 5.83e-29, pTwoSided = 1.17e-28
```

```
> betaHats(res5_beta)
```

```
             [,1]
1490   0.53814745
168544 0.10183475
```

```
1893   0.23559626
2013   0.11853248
2022   0.23358385
216    0.19554569
2273   0.35939613
23179  0.24618755
25     0.21597396
2549   0.21100011
2934   0.33968229
5445   0.23247513
55884  0.16183870
57556  0.37650630
687    0.46075522
6915   0.10897784
7277   0.34083849
92     0.15933797
9369   0.09865998
9788   0.26243503
```

```
> npGSEAPlot(res5_beta)
```

The chi-sq approximation method is only available for the two-sided test. Here we call `npGSEA` and then show how the `chiSqStat` is related to `C_Gw`. `C_Gw` is explained in more detail in Section 3.2.

```
> res5_chiSq <- npGSEA(ALL, yFactor, set5, approx= "chiSq")
> res5_chiSq
```

```
Chi-sq Approximation for set5
C_Gw = 1.52
df = 2.42, sigmaSq = 0.0172
pTwoSided = 0
```

```
> betaHats(res5_chiSq)
```

```
             [,1]
1490   0.53814745
168544 0.10183475
1893   0.23559626
2013   0.11853248
2022   0.23358385
216    0.19554569
2273   0.35939613
23179  0.24618755
25     0.21597396
2549   0.21100011
2934   0.33968229
5445   0.23247513
55884  0.16183870
```

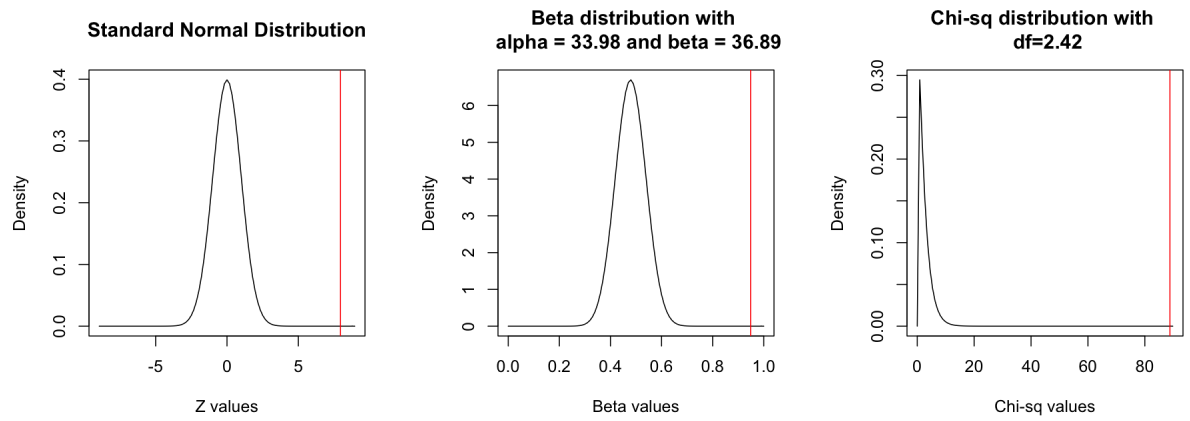

Figure 2: **Set5 normal, beta, and chi-sq approximation results** These plots displays the reference normal, beta, and chi-sq curves, and our observed zStat, betaStat, and chiSqStat for set5 in this analysis.

```
57556   0.37650630
687     0.46075522
6915    0.10897784
7277    0.34083849
92      0.15933797
9369    0.09865998
9788    0.26243503
```

```
> chiSqStat(res5_chiSq)
```

```
[1] 88.81123
```

```
> stat(res5_chiSq)
```

```
[1] 1.524982
```

```
> stat(res5_chiSq)/sigmaSq(res5_chiSq)
```

```
[1] 88.81123
```

```
> npGSEAPlot(res5_chiSq)
```

Note that, as we expected, `set5` is a significantly enriched in all three methods. In each of the three corresponding plots, the observed statistic is a very rare event.

2.4 Adding weights to the model

Sometimes we do not want to weigh each gene in our set equally. We want to assign a larger weight to genes that are of a particular interest, and a lower weight to genes that we know may behave poorly. In this example, we weight the genes in `set7` by their variance.

```
> res7_nowts <- npGSEA(x = ALL, y= yFactor, set = set7)
> res7_nowts
```

```
Normal Approximation for set7
T_Gw = 1.46
var(T_Gw) = 0.0576
pLeft = 1, pRight = 6.39e-10, pTwoSided = 1.28e-09
```

```
> wts <- apply(exprs(ALL)[match(geneIds(set7), featureNames(ALL)), ],
+                                                  1, var)
> wts <- 1/wts
> res7_wts <- npGSEA(x = ALL, y = yFactor, set = set7,  w = wts, approx= "norm")
> res7_wts
```

```
Normal Approximation for set7
T_Gw = 10.4
var(T_Gw) = 2
pLeft = 1, pRight = 8.87e-14, pTwoSided = 1.77e-13
```

By adding these weights, we get a slightly more significant result. We can add weights for the beta and chi-sq approximations, too. By default, `npGSEA` assigns a weight of 1 for all genes.

2.5 Adding covariates to model

Often we want to correct for confounders in our model. To do this with `npGSEA`, we provide a vector or matrix in the `covars` slot of our function. `npGSEA` then projects both the data (`x`) and the outcome of interest (`y`) against our covariate matrix/vector. The resulting residuals are used for further analysis.

In this example, we correct for the age and sex of the subjects in our experiment. For more details on model selection and its relation to inference, please see the `limma` manual.

```
> res3_age <- npGSEA(x = ALL, y = yFactor, set = set3, covars = ALL$age)
> res3_age
```

```
Normal Approximation for set3
T_Gw = 0.304
var(T_Gw) = 0.372
pLeft = 0.691, pRight = 0.309, pTwoSided = 0.618
```

```
> res3_agesex <- npGSEA(x = ALL, y = yFactor, set = set3, covars = cbind(ALL$age, ALL$sex
> res3_agesex
```

```
Normal Approximation for set3
T_Gw = 0.262
var(T_Gw) = 0.364
pLeft = 0.668, pRight = 0.332, pTwoSided = 0.664
```

By adjusting for these variables, we get a slight different result than above. Note that we can adjust for covariates in the beta and chi-sq approximation methods, too.

2.6 Running npGSEA with multiple gene sets

To explore multiple gene sets, we let `set` be a `GeneSetCollection`. This returns a list of `npGSEAResultNorm` objects, called a `npGSEAResultNormCollection`. We can access statistics for each `GeneSet` in our analysis through accessors of `npGSEAResultNormCollection`.

```
> resgsc_norm <- npGSEA(x = ALL, y = yFactor, set = gsc)
> unlist( pLeft(resgsc_norm) )
```

```
         set1          set2          set3          set4          set5          set6
7.547548e-01 7.168167e-01 7.034361e-01 6.021808e-01 1.000000e+00 4.247294e-11
         set7
1.000000e+00
```

```
> unlist( stat (resgsc_norm) )
```

```
      set1       set2       set3       set4       set5       set6       set7
 0.1065501  0.1585581  0.3560641  0.3661015  4.9973052 -3.7097502  1.4564761
```

```
> unlist( zStat (resgsc_norm) )
```

```
      set1       set2       set3       set4       set5       set6       set7
 0.6895291  0.5734107  0.5343088  0.2589959  7.9583579 -6.4915710  6.0700222
```

Note how quick our method is. We get results as accurate as permutation methods in a fraction of the time, even for multiple gene sets.

Using the `ReportingTools` package, we can publish these results to a HTML page for exploration. We first adjust for multiple testing.

```
> pvals <- p.adjust( unlist(pTwoSided(resgsc_norm)), method= "BH" )
> library(ReportingTools)
> npgseaReport <- HTMLReport (shortName = "npGSEA",
+         title = "npGSEA Results", reportDirectory = "./reports")
> publish(gsc, npgseaReport, annotation.db = "org.Hs.eg",
+         setStats = unlist(zStat (resgsc_norm)), setPValues = pvals)
> finish(npgseaReport)
```

3 Methods in brief

3.1 Disadvantages to a permutation approach

There are three main disadvantages to permutation-based analyses: cost, randomness, and granularity.

Testing many sets of genes becomes computationally expensive for two reasons. First, there are many test statistics to calculate in each permuted version of the data. Second, to allow for multiplicity adjustment, we require small nominal p-values to draw inference about our sets, which in turn requires a large number of permutations.

Permutations are also subject random inference. Because permutations are based on a random shuffling of the data, there is a chance that we will obtain a different p-value for our set of interest each time we run our permutation analysis.

Permutations also have a granularity problem. If we do M permutations, then the smallest possible p-value we can attain is $1/(M+1)$. When it is necessary to adjust for multiplicity, the permutation approach becomes very computationally expensive. Another aspect of the granularity problem is that permutations give us no basis to distinguish between two gene sets that both have the same p-value $1/(M+1)$. There may be many such gene sets, and they have meaningfully different effect sizes.

Because of each of these limitations of permutation testing, there is a need to move beyond permutation-based GSEA methods. The methods we present in `npGSEA` and discuss in brief below are not as computationally expensive, random, or granular than their permutation counterparts. More details on our method can be found in Larson and Owen (2015).

3.2 Test statistics

We present our notation using the language of gene expression experiments.

Let g and h denote individual genes and G be a set of genes. Our experiment has n subjects. The subjects may represent patients, cell cultures, or tissue samples. The expression level for gene g in subject i is X_{gi}, and Y_i is the target variable on subject i. Y_i is often a treatment, disease, or genotype. We center the variables so that $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X_{gi} = 0, \forall g$.

Our measure of association for gene g on our treatment of interest is

$$\hat{\beta}_g = \frac{1}{n}\sum_{i=1}^{n} X_{gi}Y_i.$$

We consider the linear statistic

$$T_{G,w} = \sum_{g\in G} w_g \hat{\beta}_g$$

and the quadratic statistic

$$C_{G,w} = \sum_{g\in G} w_g \hat{\beta}_g^2,$$

where w_g corresponds to the weight given to gene g in set G.

3.3 Moment based reference distributions

To avoid the issues discussed above, we approximate the distribution of the permuted test statistics $T_{G,w}$ by Gaussian or by rescaled beta distributions. For the quadratic statistic $C_{G,w}$ we use a distribution of the form $\sigma^2\chi^2_{(\nu)}$.

For the Gaussian treatment of $T_{G,w}$ we calculate $\sigma^2 = Var(T_{G,w})$ under permutation, and then report the p-value

$$p = Pr(N(0,\sigma^2) \le T_{G,w}).$$

The above is a left tail p-value. Two-sided and right tailed p-values are analogous.

When we want something sharper than the normal distribution, we can use a scaled Beta distribution, of the form $A + (B - A)Beta(\alpha, \beta)$. The $Beta(\alpha, \beta)$ distribution has a continuous density function on $0 < x < 1$ for $\alpha, \beta > 0$. We choose A, B, α and β by matching the upper and lower limits of $T_{G,w}$ under permutation, as well as its mean and variance. The observed left tailed p-value is

$$p = Pr\Big(Beta(\alpha, \beta) \leq \frac{T_{G,w} - A}{B - A}\Big).$$

For the quadratic test statistic $C_{G,w}$ we use a $\sigma^2\chi^2_{(\nu)}$ reference distribution reporting the p-value

$$\Pr(\sigma^2\chi^2_{(\nu)} \geq C_{G,w}),$$

after matching the first and second moments of $\sigma^2\chi^2_{(\nu)}$ to $E(C_{G,w})$ and $E(C^2_{G,w})$ under permutation, respectively.

Additional details on how σ^2, A, B, α, β, $E(C_{G,w})$, $E(C^2_{G,w})$, and ν are derived can be found in Larson and Owen (2015).

4 Session Info

- R version 3.4.0 (2017-04-21), `x86_64-apple-darwin15.6.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Running under: `OS X El Capitan 10.11.6`
- Matrix products: default
- BLAS: `/Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib`
- LAPACK: `/Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: ALL 1.17.0, AnnotationDbi 1.38.0, Biobase 2.36.0, BiocGenerics 0.22.0, GSEABase 1.38.0, IRanges 2.10.0, S4Vectors 0.14.0, XML 3.98-1.6, annotate 1.54.0, genefilter 1.58.0, graph 1.54.0, hgu95av2.db 3.2.3, limma 3.32.0, npGSEA 1.12.0, org.Hs.eg.db 3.4.1
- Loaded via a namespace (and not attached): BiocStyle 2.4.0, DBI 0.6-1, Matrix 1.2-9, RCurl 1.95-4.8, RSQLite 1.1-2, Rcpp 0.12.10, backports 1.0.5, bitops 1.0-6, compiler 3.4.0, digest 0.6.12, evaluate 0.10, grid 3.4.0, htmltools 0.3.5, knitr 1.15.1, lattice 0.20-35, magrittr 1.5, memoise 1.1.0, rmarkdown 1.4, rprojroot 1.2, splines 3.4.0, stringi 1.1.5, stringr 1.2.0, survival 2.41-3, tools 3.4.0, xtable 1.8-2, yaml 2.1.14

5 References

Larson and Owen. (2015). Moment based gene set tests. *BMC Bioinformatics*. **16**:132. http://www.biomedcentral.com/1471-2105/16/132