An Introduction to *POST*

Xueyuan Cao, Stanley Pounds

November 2, 2016

1 Introduction

POST, Projection onto Orthognal Space Test, is a general procedure to tes a set of genomic features that exhibit association with an endpoint variable. For each gene-set, POST represents the gene profiles as a set of eigenvectors and then uses statistical modeling to compute a set of (adjusted) z-statistics that measure the association of each eigenvector with the phenotype. The overall gene-set statistic is the sum of squared z-statistics weighted by the corresponding eigenvector. Finally, bootstrapping is used to compute a p-value.

In this document, we describe how to perform POST procedure using hypothetical example data sets provided with the package.

2 Requirements

The POST package depends on *Biobase*, *GSEABase*, *CompQuadForm* and *Matrix*. The understanding of *ExpressionSet* and *GeneSetCollection* is a prerequiste to perform the POST procedure.

The detailed requirements are illustrated below.

Load the POST package and the example data sets: sampExprSet and exmplGeneSet into R.

```
> library(POST)
> data(sampExprSet)
> data(sampGeneSet)
```

The *ExpressionSet* should contain at least two components: *exprs* (array data) and *phenoData* (endpoint data). *exprs* is a data frame with column names representing the array identifiers (IDs) and row names representing the probe (genomic feature) IDs. *phenoData* is an *AnnotatedDataFrame* with column names representing the endpoint variables and row names representing array. The array IDs of *phenoData* and *exprs* should be matched.

GeneSetCollection contains gene set definiton. This gene set collection can be from biological processes or ontologies. In this hypothetical example, we are interested in testing association of expression of 4 gene sets with a binary outcome and associatiion of expression of gene sets with a time-to-event endpoint.

3 POST Analysis

As mentioned in section 2, the *ExpressionSet* and *GeneSetCollection* are required by POST procedure. The code below performs a POST analysis at gene set level to detect assocaiton of gene set with binary outcome in logistic regression framework.

```
>  test<-POSTglm(exprSet=sampExprSet,
+                geneSet=sampGeneSet,
+                lamda=0.95,
+                seed=13,
+                nboots=100,
+                model='Group ~ ',
+                family=binomial(link = "logit"))
```

Gene set result:

```
> test
```

```
     GeneSet Nprobe Nproj Stat
[1,] "SetA"  "10"   "4"   "2.42341946595234"
[2,] "SetB"  "10"   "4"   "63.6341928599234"
[3,] "SetC"  "10"   "5"   "40.4788851952684"
[4,] "SetD"  "30"   "4"   "58.9759176199774"
     p.value
[1,] "0.738735024903039"
[2,] "0.00166225382835439"
[3,] "0.0195953770210454"
[4,] "0.142760313270624"
```

The code below performs POST analysis at gene set level to detect assocaiton of gene set with time to event endpoint in Cox proportional hazard model famework.

```
>  test2<-POSTcoxph(exprSet=sampExprSet,
+                geneSet=sampGeneSet,
+                lamda=0.95,
+                nboots=100,
+                model="Surv(time, censor) ~ ",
+                seed=13)

> test2
```

```
     GeneSet Nprobe Nproj Stat
[1,] "SetA"  "10"   "4"   "6.73347880247785"
[2,] "SetB"  "10"   "4"   "10.7974608126469"
[3,] "SetC"  "10"   "5"   "11.2005143113062"
[4,] "SetD"  "30"   "4"   "13.8964097387099"
```

```
     p.value
[1,] "0.354932076148768"
[2,] "0.248238953750029"
[3,] "0.367212521285667"
[4,] "0.57770769131475"
```