Reference Manual for Outlier and Subgroup Identification Statistics (OASIS)

Stan Pounds and Iwona Pawlikowska

February 6, 2014

1 Introduction

Modern high-throughput technologies allow researchers to idencorrelate with biologically or clinically important traits. It is also possible to discover new biological processes by identifying transcriptomic features that have outliers or multiple modes in their expression distributions. Outliers or subgroups in expression may flag features that affect disease biology or indicate a genomic abnormality such as translocation or deletion. The statistics and bioinformatics literature proposes several OASIS methods. Here, we adapted some of them, i.e. leave-one-out, least median squares, the dip test and maximum spacing test, and introduced a new method i.e. most informative spacing test for OASIS.

Leave-one-out (LOO) procedure leaves out one data value, compute the mean and standard deviation of the remaining data values, and then compare the left-out data value to those summary statistics. Rousseeuw [1] notes that LOO is an effective method for detection of single outliers but also shows that LOO is not an effective method for detection of multiple outliers. Thus, Rousseeuw proposes least median squares (LMS) as a robust method to detect multiple outliers.

LMS first identifies the narrowest interval that includes at least 50% of the data values and then uses the center and width of this interval that captures the 'bulk' of the data to determine whether other data values are outliers. Rousseeuw shows that LMS effectively identifies outliers even when up to 50% of the observations are outliers.

The dip test developed by Hartigan and Hartigan [2] is another potentially robust OASIS method that is not widely used in the bioinformatics and genomics literature. The dip test evaluates the null hypothesis that a set of data values is unimodal. The dip statistic is the largest difference between the empirical distribution function (EDF) and the unimodal distribution function that minimizes the maximum difference from the EDF. Thus, a significant dip statistic indicates compelling evidence that a particular set of data values has multiple modes.

Intuitively, the differences between consecutive ordered data values are very informative regarding the existence of outliers or multiple modes. Pyke [3] called these differences spacings and derived their theoretical properties under many statistical models. Pounds [4] successfully used Pyke's work to accurately estimate the fraction of clonable DNA. Therefore, we use Pyke's theory to develop two novel OASIS methods for analysis of transcriptomic expression data.

Here, we borrow ideas from the OASIS methods in the bioinformatics and statistics literatures and [5] to develop the most informative spacing test (MIST) [6]. For each individual expression variable, MIST computes the differences between consecutive order statistics (spacings) and multiplies each spacing by the geometric mean of the sizes of the two groups it defines. The spacing with the largest value of this statistic is considered to be the most informative spacing and its significance is determined by simulation.

In this document, we describe how to use OASIS package.

2 Installation

The latest version of OASIS package can be found in the link http://www.stjuderesearch.org/site/depts/biostats/software and R CRAN version will be available soon.

```
> # Install packages
> # Load the package
> library(OASIS)
```

3 M7example

Example data set *M7example* [7] contains mRNA-seq exon read counts for 14 patients with Acute Megakaryoblastic Leukemia (AMKL) treated at St Jude Children's Research Hospital. It is a data frame with 504 rows and 18 columns. The first four columns *gene*, *chrom*, *loc.start* and *loc.end* contain annotation information such as gene symbol, chromosomal location and also start and end location of exon. The remaining 14 columns are raw mRNA-seq exon count data for children with AMKL.

```
> # Load the example data set
> data(M7example)
> # show
> head(M7example)
```

```
    gene chrom loc.start loc.end SJAMLM7001.D SJAMLM7003.D SJAMLM7004.D
1 stoyru    16   4233394 4234121            2            7            1
2    SRL    16   4239370 4242965           61          254           40
3    SRL    16   4239375 4242965           61          254           40
4    SRL    16   4239377 4242965           61          254           40
5 TMEM8A    16    424005  424151          213          296          704
6 TMEM8A    16    424005  424174          213          296          706
  SJAMLM7005.D SJAMLM7006.D SJAMLM7007.D SJAMLM7008.D SJAMLM7009.D SJAMLM7010.D
1            1            3            2            0            2            0
2           39           13           17           42           19            0
3           39           13           17           42           19            0
4           39           13           17           42           19            0
5          256          237          196          184          429          289
6          256          237          198          184          429          289
  SJAMLM7011.D SJAMLM7012.D SJAMLM7013.D SJAMLM7014.D SJAMLM7015.D
1            1            0            0            2            1
2            8           18            5           18           99
3            8           18            5           18           99
4            8           18            5           18           99
5          281          298          229          238          742
6          281          300          230          238          743
```

4 Data preparation

The main function of the package, *row.oasis*, operates on a data frame with rows containing expression features and columns as subjects. Before we proceed with OASIS analysis, we need to normalize data unless each subject has expression in the interval (0,1). We propose positive quantile transformation (PQT) using *pq.transform*.

```
> # specify columns of data to transform
> data.columns=5:18
> # perform positive quantile transformation of specified columns of data
> pqt.data=pq.transform(M7example,data.columns,add.row.id=TRUE)
> head(pqt.data)
```

```
  row.id   gene chrom loc.start loc.end SJAMLM7001.D SJAMLM7003.D SJAMLM7004.D
1      1 stoyru    16   4233394 4234121   0.05042017    0.1680498   0.04347826
2      2    SRL    16   4239370 4242965   0.40756303    0.7344398   0.34096110
3      3    SRL    16   4239375 4242965   0.40756303    0.7344398   0.34096110
4      4    SRL    16   4239377 4242965   0.40756303    0.7344398   0.34096110
5      5 TMEM8A    16    424005  424151   0.71638655    0.7904564   0.87643021
6      6 TMEM8A    16    424005  424174   0.71638655    0.7904564   0.87871854
  SJAMLM7005.D SJAMLM7006.D SJAMLM7007.D SJAMLM7008.D SJAMLM7009.D SJAMLM7010.D
1   0.03534304    0.1284211   0.06004141    0.0000000   0.04192872    0.0000000
2   0.32016632    0.2505263   0.23809524    0.3948498   0.25786164    0.0000000
3   0.32016632    0.2505263   0.23809524    0.3948498   0.25786164    0.0000000
4   0.32016632    0.2505263   0.23809524    0.3948498   0.25786164    0.0000000
5   0.76091476    0.7747368   0.69772257    0.6824034   0.84486373    0.8402626
6   0.76091476    0.7747368   0.69979296    0.6824034   0.84486373    0.8402626
  SJAMLM7011.D SJAMLM7012.D SJAMLM7013.D SJAMLM7014.D SJAMLM7015.D
1   0.03404255    0.0000000    0.0000000   0.05208333   0.03151261
2   0.15744681    0.2542017    0.1120507   0.24791667   0.50630252
3   0.15744681    0.2542017    0.1120507   0.24791667   0.50630252
4   0.15744681    0.2542017    0.1120507   0.24791667   0.50630252
5   0.78936170    0.7394958    0.5961945   0.73333333   0.92647059
6   0.78936170    0.7415966    0.5983087   0.73333333   0.92857143
```

The input data *data.set* can be a data frame or name of a tab-delimited data file. If *data.columns* is NULL, then all columns of data set will be transformed. If data.columns is a character vector, then all columns of data set with a name found in *data.columns* will be transformed. If *data.columns* is a numeric vector, then the columns with those numeric indices will be transformed. By default *data.columns* are all column's names of input data set. The parameter *add.row.id* indicates whether to add a column with row identifiers and by default is set to TRUE. For each subject, PQT normalizes the raw expression values by determining their quantile against the positive raw expression values.

Now we can perform OASIS analysis on normalized data set.

```
> data.columns=data.columns+1  # since there is an additional column row.id
> # define s0
> s0=1/(2*(nrow(M7example)))
> # perform OASIS and calculate values of simulations statistics
> res=row.oasis(pqt.data,s0=s0,data.columns,sim.stats=NULL,nsim=1000)
```

Parameters *data.set* and *data.columns* are defined in the same way as in the function *pq.transform*. *s0* is a constant added to the scale estimate prior to computing the t-statistic in LOO and LMS. Default value for *s0* is 1/(2*(nrow(data.set))). Parameter *unitize*, by default set to FALSE, indicates whether to unitize rows of *data.set*. By default we assume that input data set is transformed using PQT or some other normalization is used, so that values of each row lie in (0,1). In the other case, we transform input data into values that lies along a line crossing y-axis in the point 1/(n+1), where n denotes sample size. The transformation is inspired by the fact that for n iid random variables $X_1, \ldots, X_n$ from the uniform(0,b), the scale b can be estimated as $(n+1)/nX_{(n)}$, where $X_{(n)}$ is the largest order statistic.

User can choose which OASIS method to perform using the parameter *method*:

- *mast* - maximum spacing,
- *mist* - most informative spacing,
- *dip* - dip test,
- *lms.mop* - least median squares with minimum outlier p-value,
- *lms.sst* - least median squares with sum of squared t-statistics,
- *loo.mop* - leave-one-out with minimum outlier p-value,
- *loo.sst* - leave-one-out with sum of squared t-statistics.

By default, *"all"*, function will compute all OASIS methods.
P-value for each test is based on simulation from the normal distribution and simulation from the uniform distribution for each test. For larger data sets simulations are computationally intensive and we recommend to perform them once and store the results. Using parameter *sim.stats=NULL* we can calculate a data frame with simulated OASIS statistics. Otherwise one can provide a data frame with previously calculated simulated statistics. The number of simulations is set to 10000 by default.

```
> names(res)

[1] "oasis.res" "sim.stats"

> names(res$oasis.res)

 [1] "row.id"          "gene"            "chrom"           "loc.start"
 [5] "loc.end"         "mast.stat"       "p.mast"          "mast.index"
 [9] "mast.hi"         "mist.stat"       "mist.index"      "mist.hi"
[13] "lms.bulk1"       "lms.bulk2"       "lms.bulk.index1" "lms.bulk.index2"
[17] "lms.bulk.size"   "lms.center"      "lms.scale"       "lms.maxt"
[21] "lms.mop"         "lms.sst"         "dip.stat"        "loo.maxt"
[25] "loo.maxt.index"  "loo.mop"         "loo.sst"

> names(res$sim.stats)

 [1] "mast.stat.usim.pvalue" "mist.stat.usim.pvalue" "lms.mop.usim.pvalue"
 [4] "lms.sst.usim.pvalue"   "dip.stat.usim.pvalue"  "loo.mop.usim.pvalue"
 [7] "loo.sst.usim.pvalue"   "mast.stat.zsim.pvalue" "mist.stat.zsim.pvalue"
[10] "lms.mop.zsim.pvalue"   "lms.sst.zsim.pvalue"   "dip.stat.zsim.pvalue"
[13] "loo.mop.zsim.pvalue"   "loo.sst.zsim.pvalue"
```

The result of function *row.oasis* depends on the parameter *sim.stats*. If *sim.stats=NULL* the result is given in the form of a list with names "oasis.res" and "sim.stats". "oasis.res" is a data frame with the following columns:

- *mast.stat* maximum spacing statistic (MAST)
- *p.mast* p-value for mast statistic
- *mast.index* index of maximum spacing
- *mast.hi* number of data points located to the right of the pair of points defining the MAST
- *mist.stat* most informative spacing test (MIST) statistic

- *mist.index* index of most informative spacing
- *mist.hi* number of data points located to the right of the pair of points defining the MIST
- *lms.bulk* narrowest interval covering half the data
- *lms.bulk.index* indices of sorted data values that define the endpoints of the narrowest interval covering half the data
- *lms.bulk.size* width of narrowests interval covering half the data
- *lms.center* least median square (LMS) estimate of center = mean of observations in the narrowest interval covering half the data
- *lms.scale* LMS scale estimate obtained by matching narrowest interval covering half the data to the first and third quartiles of the normal distribution
- *lms.maxt* maximum outlier t-statistic by LMS
- *lms.mop* minimum outlier p-value by LMS
- *lms.sst* sum of squared outlier t-statistics by LMS
- *dip.stat* dip statistic
- *loo.maxt* maximum outlier t-statistics by leave-one-out (LOO)
- *loo.maxt.index* index of maximum absolute outlier t-statistic
- *loo.mop* minimum outlier p-value by LOO
- *loo.sst* sum of squared outlier t-stats by LOO

```
> names(res$sim.stats)
```

```
 [1] "mast.stat.usim.pvalue" "mist.stat.usim.pvalue" "lms.mop.usim.pvalue"
 [4] "lms.sst.usim.pvalue"   "dip.stat.usim.pvalue"  "loo.mop.usim.pvalue"
 [7] "loo.sst.usim.pvalue"   "mast.stat.zsim.pvalue" "mist.stat.zsim.pvalue"
[10] "lms.mop.zsim.pvalue"   "lms.sst.zsim.pvalue"   "dip.stat.zsim.pvalue"
[13] "loo.mop.zsim.pvalue"   "loo.sst.zsim.pvalue"
```

"sim.stats" is a data frame with the same number of rows as the data set and 14 columns that represent p-values for each of the oasis statistic for uniform distribution ("usim") and p-values for each of the oasis statistic for normal distribution ("zsim").

4.1 MIST and MAST

These tests sort the data values and then compute the differences (spacings) between consecutive ordered data values. MAST: A very large spacing may indicate an extreme outlier or correspond to the difference between two very distinct subgroups. The maximum spacing is computed and its value is compared to the distribution of the maximum spacing of a set of independent uniform(0,1) observations to obtain a p-value. MIST: Each spacing is multiplied by a factor that is a function of its position relative the data values (number of data values to the left and number of data values to the right) to give spacing information. The factor is defined such that it has the largest value for spacing dividing two balanced subgroups of data and it has the smallest value if the spacing separates largest of smallest observation from the rest of the data. The MIST statistic is calculated as the maximum of values of spacing across all observations. The observed MIST statistic is compared to its distribution under the null model that all observations are independent identically distributed uniform(0,1) and normal(0,1).

4.2 LMS

Each observation is tested for possible outlier using least median squares (LMS). LMS identifies the narrowest interval that covers at least 50% of the data and assumes that interval defines the "bulk" of the observations. Next, t-tests are used to evaluate whether each individual observation is an outlier relative to a normal distribution with first and third quartiles corresponding to the endpoints of the interval. The sum of squared t-statistics (SST) and minimum outlier p-value (MOP) are computed to summarize the results. s_0 is a small positive number that the user may add to the LMS estimate of variance to avoid a large number ties for biologically not meaningful data (such as many zeros in RNA-seq data). Our implementation first determines the empirical null distribution for LMS-MOP and LMS-SST by computing the smallest p-value across samples and sum of squared t-statistics for each of a large number (default=10,000) of data sets with equal sample size generated from the normal(0,1) and uniform(0,1) distributions. Then, the observed MOP of each variable is compared to this empirical null distribution. The p-value is the proportion of empirical null MOPs that are greater than or equal to the observed MOP. The p-value for SST is calculated analougously. By default, the empirical null distributions are calculated (sim.stats=NULL).

4.3 DIP

This test evaluates whether the distribution of the data values appears multimodal. The dip test compares the observed empirical distribution function (EDF) of the data to the unimodal distribution function (UDF) that minimizes the maximum difference between the EDF and the UDF. Hartigan and Hartigan [2] call this minimax difference the dip statistic. Hartigan and Hartigan prove that the uniform distribution is asymptotically the "least favorable" unimodal distribution and thus recommend that the test be performed by comparing the observed value of the dip statistic to its empirical null distributions obtained by generating many data sets of equal size from the uniform(0,1) and normal(0,1) distributions. Then, the observed dip statistic of each variable is compared to this empirical null distribution. The p-value is the proportion of empirical null dip statistics that are greater than or equal to the observed dip statistic. User can insert empirical null distributions of the dip statistic (default sim.stats=NULL).

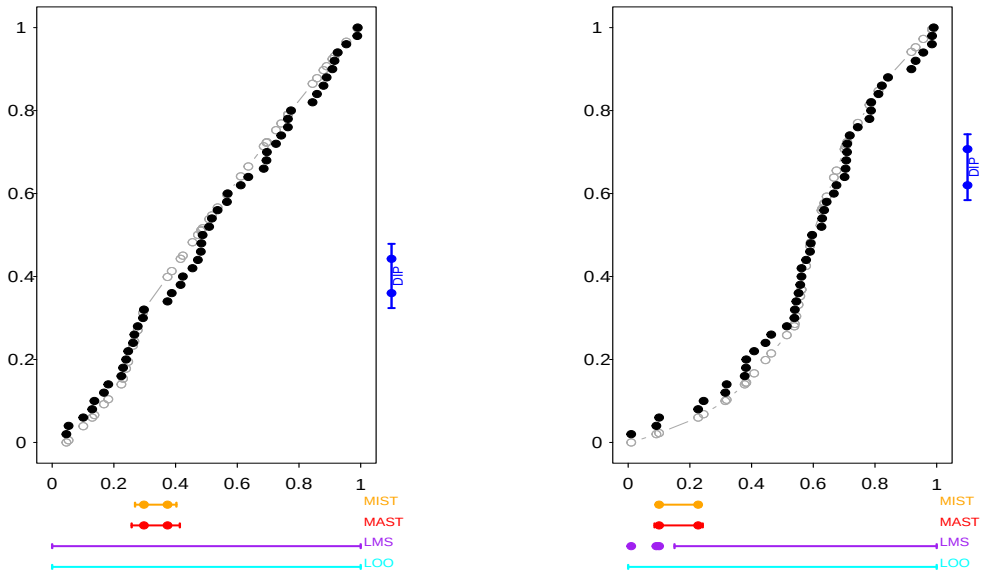
4.4 LOO

For each subject i, leave-one-out (LOO) computes the mean and standard deviation of the remaining observations (with subject i left out). Then, for each subject i, an outlier t-statistic and outlier p-value is calculated. Similarly to LMS, two statistics are calculated to summarize the collection of outlier p-values, minimum outlier p-value (MOP) and sum of squared t-statistics. There can be some cases that have large number of ties in the LOO procedure (such as many zeros in RNA-seq count data) so that leave-one-out variance estimate $s_i = 0$ for some subject i and $|t_i| = \infty$ and $p_i^{LOO} = 0$. Then, these features will be extremely significant but usually not of biological interest. To prevent these technical problems, the user may add a small positive constant s_0 to s_i. The statistical significance of LOO-MOP and LOO-SST is calculated in the similar way to LMS-MOP and LMS-SST.

```
> # or we can insert values of simulations statistics
> res.new=row.oasis(pqt.data,s0=s0,data.columns,sim.stats=res$sim.stats)
> names(res.new)
```

```
 [1] "row.id"          "gene"            "chrom"           "loc.start"
 [5] "loc.end"         "mast.stat"       "p.mast"          "mast.index"
 [9] "mast.hi"         "mist.stat"       "mist.index"      "mist.hi"
[13] "lms.bulk1"       "lms.bulk2"       "lms.bulk.index1" "lms.bulk.index2"
```

Figure 1: OASIS plot of 50 random variables from uniform(0,1) distribution (left) and 50 random variables from normal(0,1) distribution (right).




```
[17] "lms.bulk.size"   "lms.center"      "lms.scale"       "lms.maxt"
[21] "lms.mop"         "lms.sst"         "dip.stat"        "loo.maxt"
[25] "loo.maxt.index"  "loo.mop"         "loo.sst"
```

5 Generate plot for OASIS methods

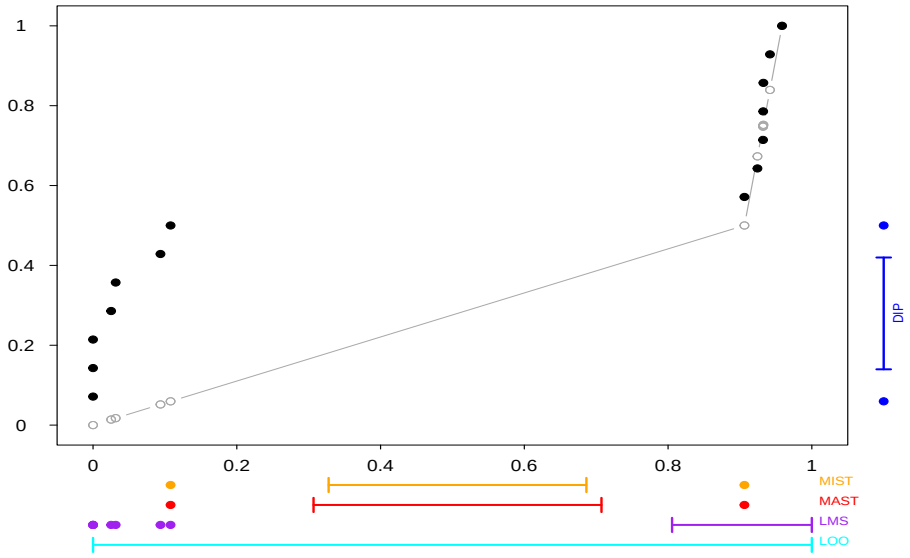
The package OASIS provides *one.oasis.plot* to visualize results of OASIS methods. This function plots data and confidence intervals for all OASIS methods for a vector of data values. The usage of *one.oasis.plot* is shown in the following examples:

```
> par(mfrow=c(1,2))
> u=runif(50)
> one.oasis.plot(u)
> z=rnorm(50)
> one.oasis.plot(z,unitize=TRUE)
```

User can specify significance level for confidence intervals in the parameter *alpha*, by default set to 0.01.

We can make a plot of an exon of GLIS2 gene that was found to have bimodal distribution [7] due to fusion with highly expressed gene CBFA2T3.

Figure 2: OASIS plot of an exon of GLIS2 from M7 example.



References

[1] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.

[2] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *The Annals of Statistics*, vol. 13, pp. 70–84, 1985.

[3] R. Pyke, "Spacings," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 27, no. 3, pp. 395–449, 1965.

[4] S. Pounds, "Estimating the fraction of clonable genomic DNA," *Bulletin of Mathematical Biology*, vol. 63, no. 5, pp. 995–1002, 2001.

[5] P. Tong, Y. Chen, X. Su, and K. R. Coombes, "SIBER: systematic identification of bimodally expressed genes using RNAseq data," *Bioinformatics*, vol. 29, no. 5, pp. 605–613, 2013.

[6] I. Pawlikowska, G. Wu, M. Edmonson, Z. Liu, T. Gruber, J. Zhang, and S. Pounds, "The most informative spacing test effectively discovers biologically relevant outliers or multiple modes in expression," *Bioinformatics*, 2014.

[7] T. A. Gruber, G. A. Larson, J. Zhang, C. S. Koss, S. Marada, H. Q. Ta, S.-C. Chen, X. Su, S. K. Ogden, J. Dang, G. Wu, V. Gupta, A. K. Andersson, S. Pounds, L. Shi, J. Easton, M. I. Barbato, H. L. Mulder, J. Manne, J. Wang, M. Rusch, S. Ranade, R. Ganti, M. Parker, J. Ma, I. Radtke, L. Ding, G. Cazzaniga, A. Biondi, S. M. Kornblau, F. Ravandi, H. Kantarjian, S. D. Nimer, K. Döhner, H. Döhner, T. L. Ley, P. Ballerini, S. Shurtleff, D. Tomizawa, S. Adachi, Y. Hayashi, A. Tawa, L.-Y. Shih, D.-C. Liang, J. E. Rubnitz, C.-H. Pui, E. R. Mardis, and J. Wilson, R. K. Downing, "An Inv (16)(p13. 3q24. 3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute Megakaryoblastic Leukemia," *Cancer Cell*, vol. 22, no. 5, pp. 683–697, 2012.