betaMix

Haim Bar

2020-12-02

The betaMix package is used to find edges in gene networks using co-expression data. It uses some plotting functions from the edgefinder package, and like edgefinder, it uses a mixture model to detect edges, but the mixture models are different. The betaMix method is built on the insightful results of (Frankl and Maehara 1990) who showed that two random vectors are approximately perpendicular with high probability, if the dimension of the space is sufficiently large. The pair-wise correlations between pairs of predictors are equal to the cosine of the angles between the pairs. From these angles we compute $z_j = \sin^2(\theta)$ and fit a mixture of two beta distributions. For pairs of random vectors (the null set) the distribution of z_j is Beta((N-1)/2, 1/2), where N is the sample size. The nonnull set is assumed to follow a Beta(a,b) distribution, and using the EM algorithm we estimate a,b, and the proportion, p0, of the null set of pairs. The betaMix function determines a threshold which will control the error rate given by the user. Any z_j below that threshold corresponds to a significantly correlated pair of predictors (an edge in the graphical model.) If the N samples can be assumed to be indepndent, set the parameter ind to TRUE. If it is set to FALSE (the default), the null set follows a Beta((nu-1)/2, 1/2) distribution and nu (the effective sample size) has to be estimated from the data.

The input to the program is a normalized expression matrix, with genes (nodes) in the columns, and samples in the rows.

With a large number of predictors, P, the estimation may be slow, so it is recommended to set the parameter subsamplesize to something smaller than choose(P,2). The minimum allowed by the program is 20,000. Using anything smaller will cause betaMix to fit the model to all choose(P,2) pairs.

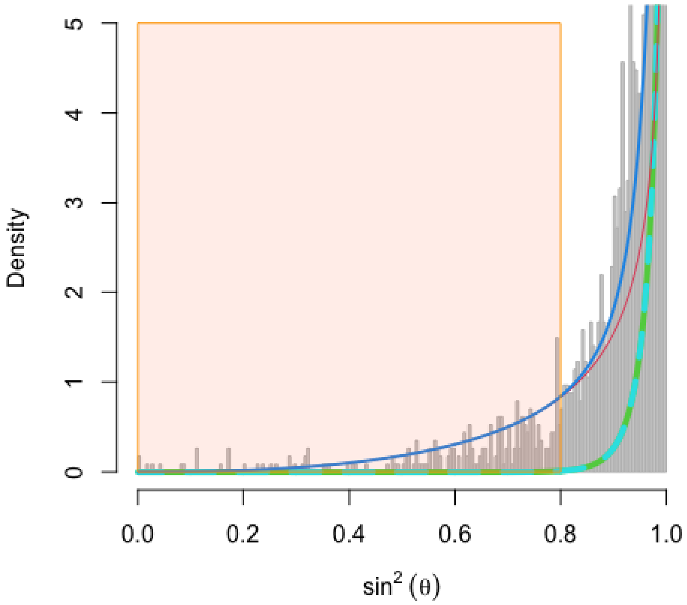
The betaMix package depends on the 'Matrix' package, which is also loaded by the edgefinder package, to allow for efficient storage and computation of large co-occurrence matrices.

Real data examples

We use metabolomic profile data from the www.metabolomicsworkbench.org website, Study ID ST000561. The data was introduced and analyzed in (Kazmi et al. 2017) and like their analysis we focus on 68 named metabolites. In the study, there are two groups of seeds - dry, and 6 hours imbibed. Each one has 50 samples, so P=68, N=50. We construct the network for each group separately. For example, for the dry-seed group we do the following:

```
library("betaMix")
data(DrySeeds)
res1 <- betaMix(DrySeeds, ind=T) # the samples can be assumed to be independent.
plotFittedBetaMix(DrySeeds,res1)
```

Note that we've used the ind=T option, since the 50 samples can be assumed to be independent. The fitted mixture model is shown below. The red line is the non-null component, and the green (solid) line is the null component under the independent sample assumption. If the model is fitted with ind=F, the effective sample size is estimated, and the null component is represented by the dashed line. The blue curve represents the mixture. The orange region shows the range of z_j's which are determined to be significant.

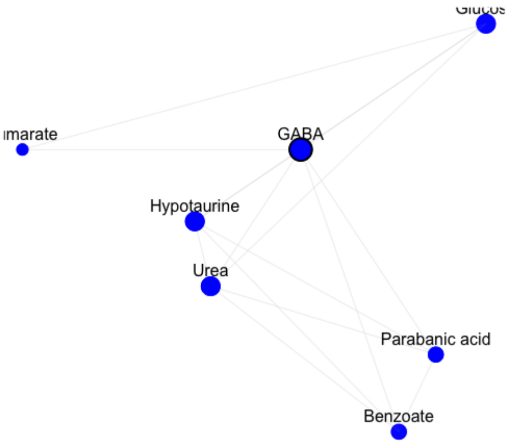


We can use some functions from the edgefinder package to summarize and visualize the results. For example, the following graphComponents code is used to create clusters of metabolites. The function summarizeClusters shows the number of nodes, edges, clusters, unclustered nodes, and summary statistics on the clusters. See the edgefinder documentation for more detail.

```
adjMat1 <- sin(res1$angleMat)^2 < res1$ppthr
diag(adjMat1) <- FALSE
graphComp1 <- graphComponents(adjMat1,minCtr = 2,type=1)
head(summarizeClusters(graphComp1))
plotCluster(adjMat1, 3, graphComp1, labels=TRUE, nodecol = "blue")
```

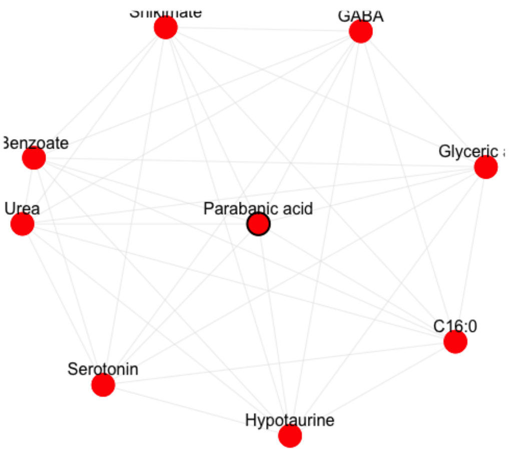
```
Num of nodes: 68
Num of edges: 236
Num of clusters: 3
Num of unclustered nodes: 27
     Cluster Nodes degreeMin degreeQ25 degreeMedian degreeQ75 degreeMax pctInClstMin pctInClstQ25
[1,]       1    17         5        10           12        14        16    0.4444444    0.9090909
[2,]       2    17         4         6           10        12        16    0.5000000    0.6666667
[3,]       3     7         3         4            5         5         6    0.6666667    0.9000000
```

A depiction of the third cluster is obtained by using the function plotCluster:



We can do the same for the SixHourImbibed dataset:

```
library("betaMix")
data(SixHourImbibed)
res2 <- betaMix(SixHourImbibed,ind=T)
plotFittedBetaMix(SixHourImbibed,res2)
adjMat2 <- sin(res2$angleMat)^2 < res2$ppthr
diag(adjMat2) <- FALSE
graphComp2 <- graphComponents(adjMat2,minCtr = 2,type=1)
summarizeClusters(graphComp2)
plotCluster(adjMat2, 3, graphComp2, labels=TRUE, nodecol = "red")
```

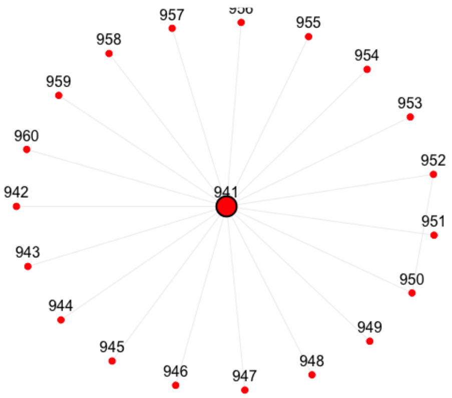


Simulated data

The following examples shows a simulated dataset (called SIM) with a hub structure, consisting of 1000 nodes and 50 hubs. The dataset is available in the edgefinder package. Note that since the simulated data was generated with variables in rows and samples in columns in the code below we pass the transpose of SIM to betaMix. Also note that in this case P=1000, which means that the number of putative edges is 499,500 so we use the subsamplesize = 30000 option. Also, because there are almost half a million tests, we set delta=1e-6 to have a proper control of the error rate. The ppr parameter is the Bayesian version of controlling the error rate by limiting the posterior probability rate.

```
data(SIM)
Simres <- betaMix(t(SIM), subsamplesize = 30000, ind=TRUE, delta=1e-6, ppr=0.01)
plotFittedBetaMix(SIM,Simres)
SimadjMat <- sin(Simres$angleMat)^2 < Simres$ppthr
diag(SimadjMat) <- FALSE
SimgraphComp <- graphComponents(SimadjMat,minCtr = 2,type=1)
summarizeClusters(SimgraphComp)
plotCluster(SimadjMat, 1, SimgraphComp, labels=TRUE, nodecol = "red")
```

We display the network of cluster 1, which shows that betaMix detects the correct cluster (hub) structure.



References

Frankl, Peter, and Hiroshi Maehara. 1990. “Some Geometric Applications of the Beta Distribution.” *Annals of the Institute of Statistical Mathematics* 42: 463–74.

Kazmi, Rashid, Leo Willems, Ronny Joosen, Noorullah Khan, Wilco Ligterink, and Henk Hilhorst. 2017. “Metabolomic Analysis of Tomato Seed Germination.” *Metabolomics* 13 (December). https://doi.org/10.1007/s11306-017-1284-x.