MBASIC: Matrix Based Analysis for State-space Inference and Clustering

Chandler Zuo and Sündüz Keleş
Departments of Statistics and of Biostatistics and Medical Informatics, University of Wisconsin-Madison

Contents

1 Introduction

This document provides an introduction to the power analysis of ChIP-seq data with the *MBASIC* package (**MBASIC** which stands for **M**atrix **B**ased **A**nalysis for **S**tate-space **I**nference and **C**lustering) [1]. *MBASIC* provides a Bayesian framework for clustering units based on their infered states over a set of experimental conditions.

MBASIC is especially useful for integrative analysis for ChIP-seq experiments. In this case, a set of prespecified loci is clustered based on their activities over celltypes and transcription factors. We build a pipeline in the *MBASIC* package and will focus on this pipeline in this vignette. We will introduce some advanced options in building the **MBASIC** model in Section 4.

2 MBASIC Pipeline for Sequencing Data

2.1 Overview

Applying the **MBASIC** framework to analyzing ChIP-seq data sets includes three steps:

1. *Matching each ChIP replicate data set with their input data set:* This step matches ChIP replicate files with their matching input files;
2. *Calculating mapped counts and the covariate on the target loci:* This step calculates the mapped counts from each ChIP and input replicate files on each of the target locus;
3. *Fitting MBASIC model:* This step fits the MBASIC model to identify the binding states for each locus and cluster the loci based on their binding states across different conditions.

MBASIC integrates Step 2-3 in a single function `MBASIC.pipeline`. For Step 1 *MBASIC* provides a function `ChIPInputMatch` that assists matching the ChIP files with input files based on the naming convention of the ENCODE datasets. We have found that in practice, more often than not, some violations to the ENCODE file name conventions always occur, and manual adjustments to the results of our automated matching are inevitable. Therefore, we do not integrate this function in `MBASIC.pipeline`.

2.2 Step 1: Match ChIP and Input Datasets

To illustrate Step 1 we first generate a set of synthetic data. *MBASIC* package provides a function `generateSyntheticData` to assist our demo. This function generates synthetic BED data for ChIP and input samples, as well as mappability and GC scores in a directory specified by the 'dir' argument. It also generates a target set of loci for our analysis. By default, the number of loci is 100, each with size 20 bp. All data are generated across 5 chromosomes, each with size 10K bp. ChIP data are from 2 celltypes, and for each celltype there are K=5 TFs. Under each condition randomly 1-3 replicates for the ChIP data are generated. All ChIP data from the same celltype are matched to the same set of 3 input replicates.

```
library(MBASIC)
```

```
target <- generateSyntheticData(dir = "syntheticData")
head(target)

## RangedData with 6 rows and 1 value column across 1 space
##      space       ranges |  chromosome
##   <factor>    <IRanges> | <character>
## 1        1 [7160, 7180] |        chr4
## 2        1 [8600, 8620] |        chr1
## 3        1 [ 560,  580] |        chr1
## 4        1 [4520, 4540] |        chr1
## 5        1 [9200, 9220] |        chr5
## 6        1 [ 760,  780] |        chr2

system("ls syntheticData/*/*", intern = TRUE)[1:5]

## [1] "syntheticData/chip/wgEncodeLabExpCell1Fac1CtrlAlnRep1.bed"
## [2] "syntheticData/chip/wgEncodeLabExpCell1Fac1CtrlAlnRep2.bed"
## [3] "syntheticData/chip/wgEncodeLabExpCell1Fac2CtrlAlnRep1.bed"
## [4] "syntheticData/chip/wgEncodeLabExpCell1Fac2CtrlAlnRep2.bed"
## [5] "syntheticData/chip/wgEncodeLabExpCell1Fac2CtrlAlnRep3.bed"
```

We have developed the `ChIPInputMatch` function to help match the ChIP and input data files. In the following example, we use this function to read all files with suffices ".bed" and ".bam" in directories specified by the argument "dir", and matches the files assuming ENCODE naming convention. It looks up files up to the number of levels of subdirectories specified by "depth". The output of this function contains multiple columns. The first column contains the file name for each ChIP replicate. The second column is the initial string for the matching input replicates, because for each ChIP replicate there are possibly multiple input replicates. The rest of the columns contains information for lab, experiment identifier, factor and control identifier. This information is parsed from the file names.

We acknowledge that the current function may not parse all data file names correctly. For practical data files, users are suggested to check its result, and manual corrections may be needed.

```
tbl <- ChIPInputMatch(dir = paste("syntheticData/", c("chip", "input"), sep = ""), celltypes = c("Cell1
    "Cell2"), suffices = c(".bam", ".bed"), depth = 5)
head(tbl)

##                                                     chipfile
## 1 syntheticData/chip/wgEncodeLabExpCell1Fac1CtrlAlnRep1.bed
## 2 syntheticData/chip/wgEncodeLabExpCell1Fac1CtrlAlnRep2.bed
## 3 syntheticData/chip/wgEncodeLabExpCell1Fac2CtrlAlnRep1.bed
## 4 syntheticData/chip/wgEncodeLabExpCell1Fac2CtrlAlnRep2.bed
## 5 syntheticData/chip/wgEncodeLabExpCell1Fac2CtrlAlnRep3.bed
## 6 syntheticData/chip/wgEncodeLabExpCell1Fac3CtrlAlnRep1.bed
##                                         inputfile lab experiment  cell factor control
## 1 syntheticData/input/wgEncodeLabExpCell1InputCtrl Lab        Exp Cell1   Fac1    Ctrl
## 2 syntheticData/input/wgEncodeLabExpCell1InputCtrl Lab        Exp Cell1   Fac1    Ctrl
## 3 syntheticData/input/wgEncodeLabExpCell1InputCtrl Lab        Exp Cell1   Fac2    Ctrl
## 4 syntheticData/input/wgEncodeLabExpCell1InputCtrl Lab        Exp Cell1   Fac2    Ctrl
## 5 syntheticData/input/wgEncodeLabExpCell1InputCtrl Lab        Exp Cell1   Fac2    Ctrl
## 6 syntheticData/input/wgEncodeLabExpCell1InputCtrl Lab        Exp Cell1   Fac3    Ctrl
##   chipformat inputformat
## 1        BED         BED
## 2        BED         BED
## 3        BED         BED
## 4        BED         BED
## 5        BED         BED
## 6        BED         BED
```

We also need to specify the experimental condition for each data set by a vector 'conds'.

```
conds <- paste(tbl$cell, tbl$factor, sep = ".")
```

2.3 Executing Step 2 and 3 Altogether

Now we are in a position to continue the next steps in the pipeline. There are two ways to execute these steps: (1) use function `MBASIC.pipeline` to execute the two steps together; or (2) execute each step separately. The following code calls the function `MBASIC.pipeline`:

```
## remove file 'data.Rda' since it will be generated try(file.remove('data.Rda'))
fit <- MBASIC.pipeline(chipfile = tbl$chipfile, inputfile = tbl$inputfile, input.suffix = ".bed",
    target = target, chipformat = tbl$chipformat, inputformat = tbl$inputformat, fragLen = 150,
    pairedEnd = FALSE, unique = TRUE, fac = conds, J = 3, S = 2, family = "negbin", datafile = "data.Rd
class(fit)
```

We list the meanings of the arguments of `MBASIC.pipeline` in Table 1. The above example fits a clustering model with 3 clusters. It returns an object of class *MBASICFit*. This class is described in more details in Section 3.

Alternatively, we can fit models with varying numbers of clusters simultaneously, and pick the one with the minimum BIC value:

```
allfits <- MBASIC.pipeline(chipfile = tbl$chipfile, inputfile = tbl$inputfile, input.suffix = ".bed",
    target = target, chipformat = tbl$chipformat, inputformat = tbl$inputformat, fragLen = 150,
    pairedEnd = FALSE, unique = TRUE, fac = conds, J = 3:10, S = 2, family = "negbin", datafile = "data
names(allfits)

## [1] "allFits" "BestFit" "Time"

class(allfits$BestFit)

## [1] "MBASICFit"
## attr(,"package")
## [1] "MBASIC"
```

Before we move on to describe the step-wise execution, we highlight the usage of the argument 'datafile'. In the above codes, when we compute 'fit', the file 'datafile' is not generated yet, so we process ChIP and input data and save the processed data in 'datafile'. When we compute 'allfits', `MBASIC.allfit` detects that 'datafile' already exists, so it automatically loads the preprocessed data and skip the step of processing the ChIP and input data. This can save substantially amount of time if the size of our data is large.

Table 1: Arguments for the `MBASIC.pipeline` function. For a comprehensive list of arguments and their details, users are recommended to read our manual.

Data Sources	
chipfile	A string vector for the ChIP files.
inputfile	A string vector for the matching input files. The length must be the same as "chipfile".
input.suffix	A string for the suffix of input files. If NULL, "inputfile" will be treated as the full names of the input files. Otherwise, all inputfiles with the initial "inputfile" and this suffix will be merged.
chipformat (inputformat)	A string specifying the type of all ChIP (input) files, or a vector of string specifying the types of each ChIP (input) file. Currently three file types are allowed: "BAM", "BED" or "TAGALIGN" ("TAGALIGN" and "BED" files are treated as same). Default: "BAM".
m.prefix (optional)	A string for the prefix of the mappability files. Default: NULL.
m.suffix (optional)	A string for the suffix of the mappability files. See our man files for more details. Default: NULL.
gc.prefix (optional)	A string for the prefix of the GC files. Default: NULL.
gc.suffix (optional)	A string for the suffix of the GC files. See our man files for more details. Default: NULL.
Genomic Information	
target	A RangedData object for the target intervals where the reads are mapped.
fragLen	Either a single value or a 2-column matrix of the fragment lengths for the chip and input files. Default: 150.
pairedEnd	Either a boolean value or a 2-column boolean matrix for whether each file is a paired-end data set. Currently this function only allows "BAM" files for paired-end data. Default: FALSE.
unique	A boolean value for whether only reads with distinct genomic coordinates or strands are mapped. Default: TRUE.
Model Parameters	
S	The number of states.
fac	A vector of length N for the experimental condition of each ChIP replicate.
J	A single number or a numeric vector of the numbers of clusters to be included in the model.
family	The distribution of family to be used. *MBASIC* currently support five distribution types: 'lognormal', 'negbin', 'binom', 'scaled-t', 'gamma-binom'. See our man files for more information.
Tuning Parameters	
maxitr	The maximum number of iterations in the E-M algorithm. Default: 100.
tol	Tolerance for the relative increment in the log-likelihood function to check the E-M algorithm's convergence. Default: 1e-10.
tol.par	Tolerance for the maximum relative change in parameters to check the algorithm's convergence. Default: 1e-5.
datafile	The location to save the count matrices, or load pre-computed count matrices.

2.4 Step 2: Generate the Data Matrices

We can execute Step 2 and 3 separately. In Step 2, we use the function `generateReadMatrices` to calculate the ChIP count at each locus for each ChIP replicate. We also calculate the count at each locus for each matching input.

```
## Step 2: Generate mapped count matrices
dat <- generateReadMatrices(chipfile = tbl$chipfile, inputfile = tbl$inputfile, input.suffix = ".bed",
    target = target, chipformat = tbl$chipformat, inputformat = tbl$inputformat, fragLen = 150,
    pairedEnd = FALSE, unique = TRUE)
```

```
conds <- paste(tbl$cell, tbl$factor, sep = ".")
```

We can directly use the matching input counts as the covariate data in our model. Advanced users of our package may want to normalize the input counts according to the mappability and GC scores and use the normalized counts as the covariate. In that case, we need call functions `averageMGC` and `bkng_mean`.

```
## Step 2': calculate the mappability and GC-content scores for each locus
target <- averageMGC(target = target, m.prefix = "syntheticData/mgc/", m.suffix = "_M.txt",
    gc.prefix = "syntheticData/mgc/", gc.suffix = "_GC.txt")
## Step 2': compute the normalized input counts
dat$input1 <- bkng_mean(inputdat = dat$input, target = target, family = "negbin")
```

Notice that such a normalization step is automatically executed if users specify the 'm.prefix', 'm.suffix', 'gc.prefix' and 'gc.suffix' arguments in `MASIC.pipeline`.

2.5 Step 3: Build the MBASIC Model

We can build an MBASIC model using function `MBASIC`:

```
## Step 3: Fit an MBASIC model
fit <- MBASIC(Y = t(dat$chip), Gamma = t(dat$input), S = 2, fac = conds, J = 3, maxitr = 10,
    family = "negbin")
```

We can also simultaneously fit models with different numbers of clusters:

```
## Step 3: Fit an MBASIC model
allfits <- MBASIC.full(Y = t(dat$chip), Gamma = t(dat$input), S = 2, fac = conds, J = 3:10,
    maxitr = 10, family = "negbin", ncores = 10)
```

3 Interpreting the Results

The outputs of both `MBASIC` and `MBASIC.pipeline` functions are of S-4 class *MBASICFit*.

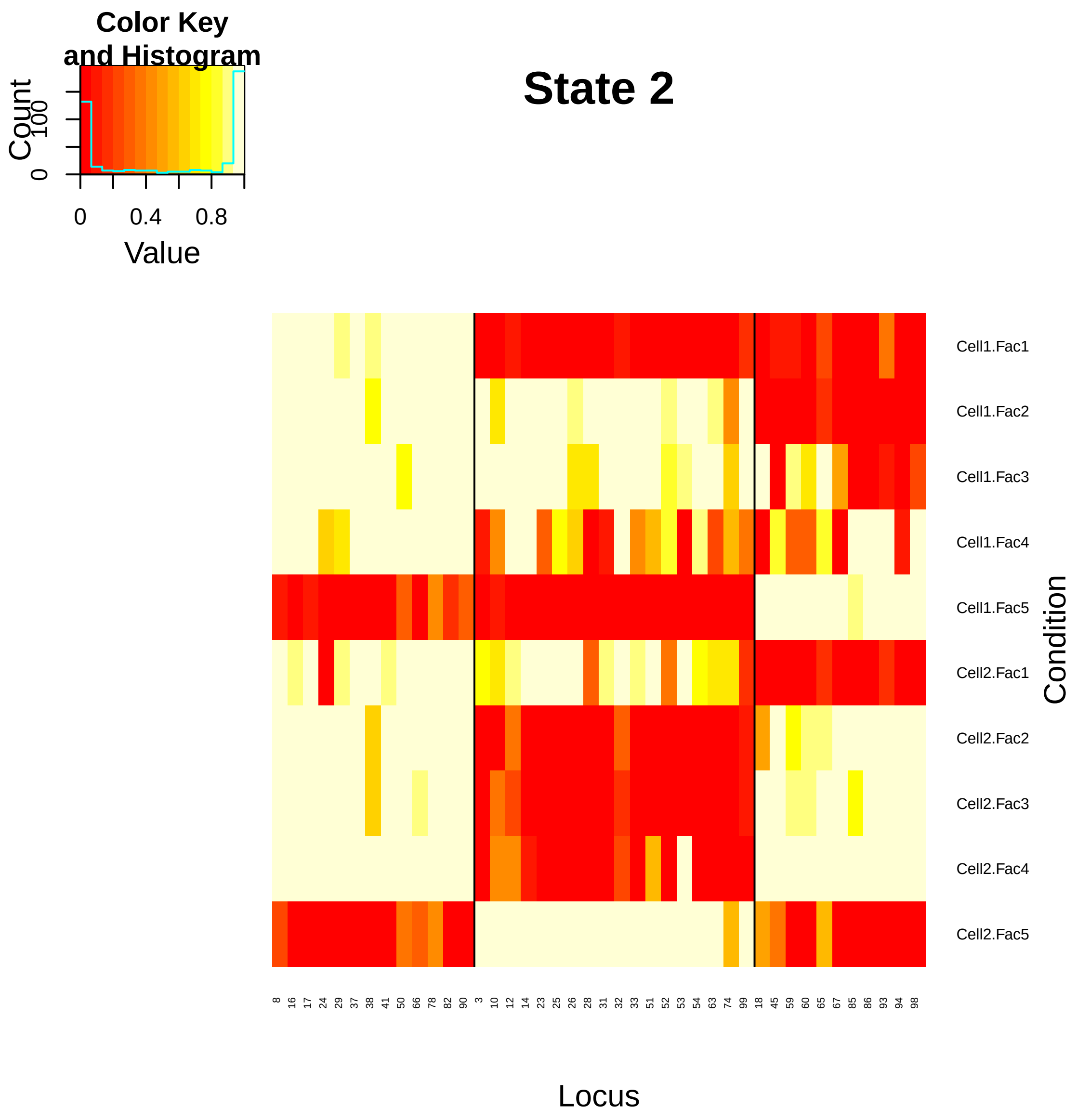
```
showClass("MBASICFit")
```

Slot 'Theta' records the probabilities for each locus to have each state under each condition. For a model with K experiment conditions, S states and I units, slot 'Theta' is a matrix of dimension KS by I, and the (K(s-1)+k, i)-th entry is the probability for the i-th locus to have state s under condition k.

```
dim(fit@Theta)

## [1]  20 100

rownames(fit@Theta)

##  [1] "Cell1.Fac1" "Cell1.Fac2" "Cell1.Fac3" "Cell1.Fac4" "Cell1.Fac5" "Cell2.Fac1"
##  [7] "Cell2.Fac2" "Cell2.Fac3" "Cell2.Fac4" "Cell2.Fac5" "Cell1.Fac1" "Cell1.Fac2"
## [13] "Cell1.Fac3" "Cell1.Fac4" "Cell1.Fac5" "Cell2.Fac1" "Cell2.Fac2" "Cell2.Fac3"
## [19] "Cell2.Fac4" "Cell2.Fac5"

head(fit@Theta[1, ])

## [1] 0.011700421 0.001076443 0.987588377 0.999539864 0.150268077 0.964401527
```

In our example, suppose state 1 corresponds to the un-enriched state, and state 2 the enriched state. We can use the function `plot` to draw a heatmap to visualize the enrichment states across all data.

```
plot(fit, slot = "Theta", xlab = "Locus", state = 2, cexRow = 0.6, cexCol = 0.4)
```



Slot 'clustProb' is a matrix for the posterior probablity of each locus to belong to each cluster. The first column contains the probability for each locus to be a singleton (i.e. not belong to any clusters). The (j+1)-th column contains the probabilities for each locus to be in cluster j.

```
dim(fit@clustProb)

## [1] 100   4

round(head(fit@clustProb), 3)

##       b.prob
## [1,]  0.583 0.412 0.000 0.005
## [2,]  1.000 0.000 0.000 0.000
## [3,]  0.038 0.000 0.962 0.000
## [4,]  0.999 0.000 0.000 0.001
## [5,]  0.878 0.004 0.000 0.118
## [6,]  0.761 0.000 0.239 0.000
```

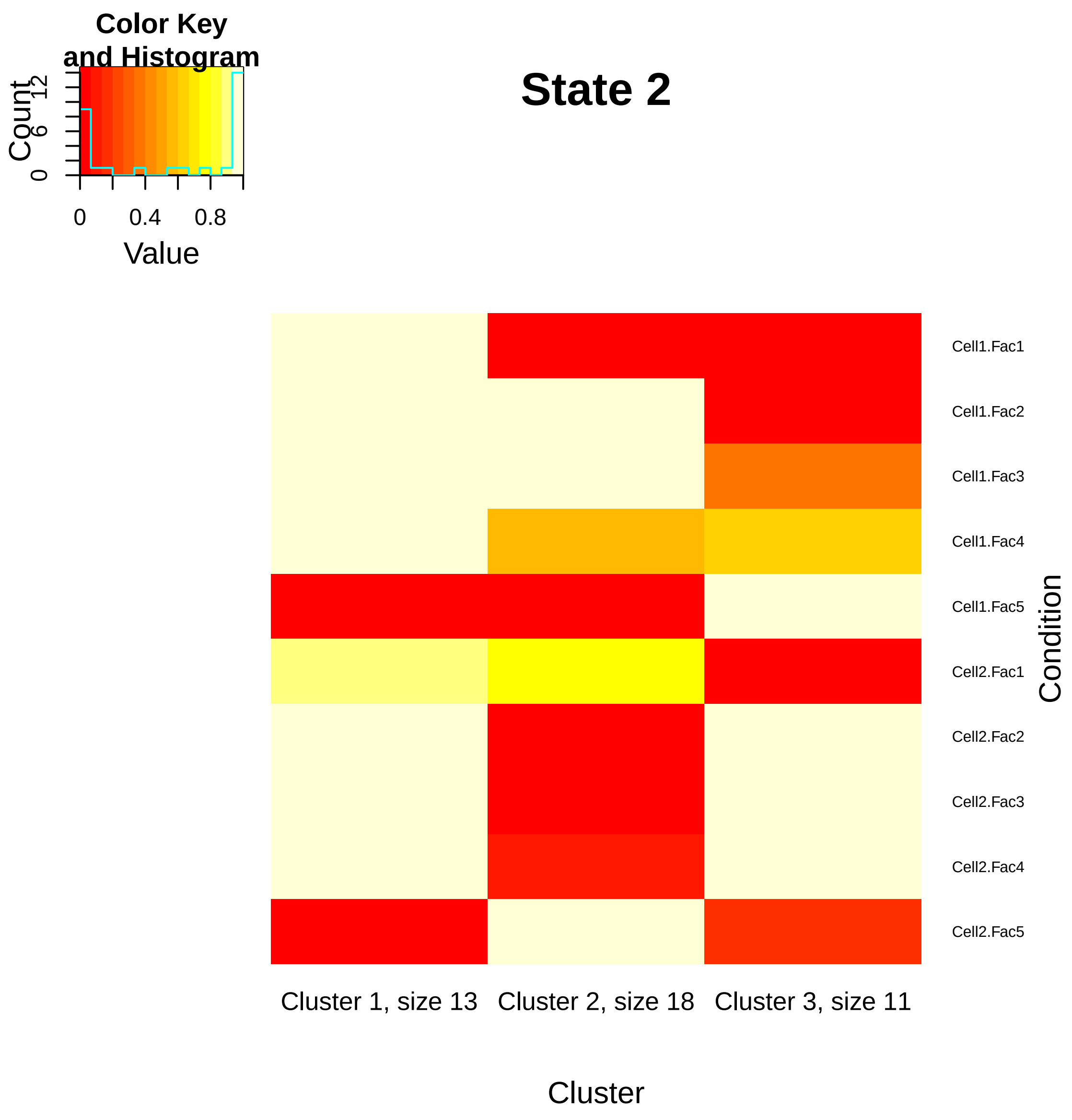
```
clusterLabels <- apply(fit@clustProb, 1, which.max) - 1
table(clusterLabels)

## clusterLabels
##  0  1  2  3
## 58 13 18 11
```

Slot 'W' is a matrix for the state-space profiles for all clusters. For a model with K conditions, S states and J clusters, this matrix has dimension KS by J, where the (k+K(s-1),j)-th entry contains the probability that a unit in cluster j has state s under condition k. For our example, the enriched probability for all clusters is in rows 11-20.

```
rownames(fit@W)

##  [1] "Cell1.Fac1" "Cell1.Fac2" "Cell1.Fac3" "Cell1.Fac4" "Cell1.Fac5" "Cell2.Fac1"
##  [7] "Cell2.Fac2" "Cell2.Fac3" "Cell2.Fac4" "Cell2.Fac5" "Cell1.Fac1" "Cell1.Fac2"
## [13] "Cell1.Fac3" "Cell1.Fac4" "Cell1.Fac5" "Cell2.Fac1" "Cell2.Fac2" "Cell2.Fac3"
## [19] "Cell2.Fac4" "Cell2.Fac5"

dim(fit@W)

## [1] 20  3

round(head(fit@W[seq(10) + 10, ]), 3)

##             [,1]  [,2]  [,3]
## Cell1.Fac1 0.999 0.013 0.012
## Cell1.Fac2 1.000 0.979 0.000
## Cell1.Fac3 0.951 0.986 0.349
## Cell1.Fac4 0.969 0.546 0.619
## Cell1.Fac5 0.061 0.000 0.999
## Cell2.Fac1 0.902 0.787 0.000
```

We can also use the `plot` function to visualize the probability for each cluster to have a particular state under all conditions.

```
plot(fit, slot = "W", state = 2, cexRow = 0.6, cexCol = 1, srtCol = 0, adjCol = c(0.5, 1))
```

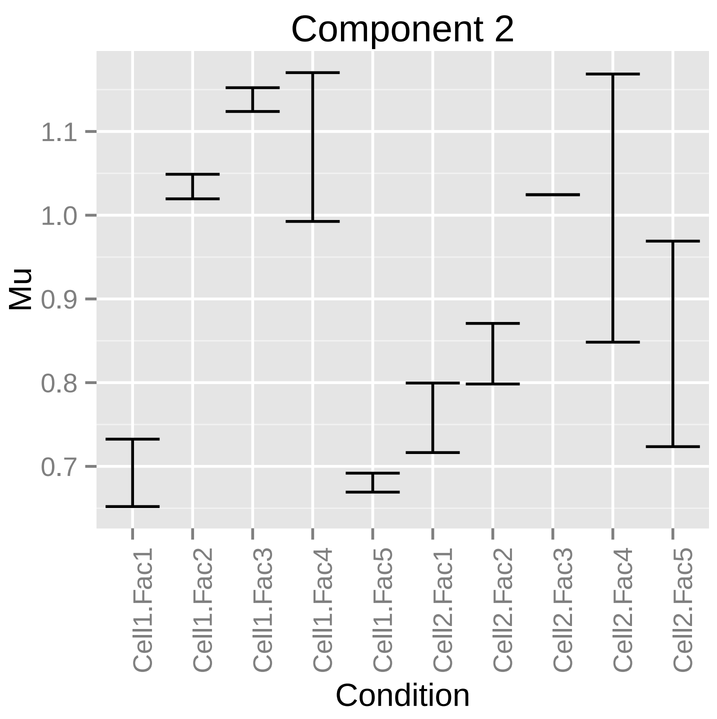


Slot 'Mu' and 'Sigma' are matrices for the distribution parameters for each replicate and each component (usually, one component is one state, see Section 4.2).

```
dim(fit@Mu)

## [1] 24  2

dim(fit@Sigma)

## [1] 24  2
```

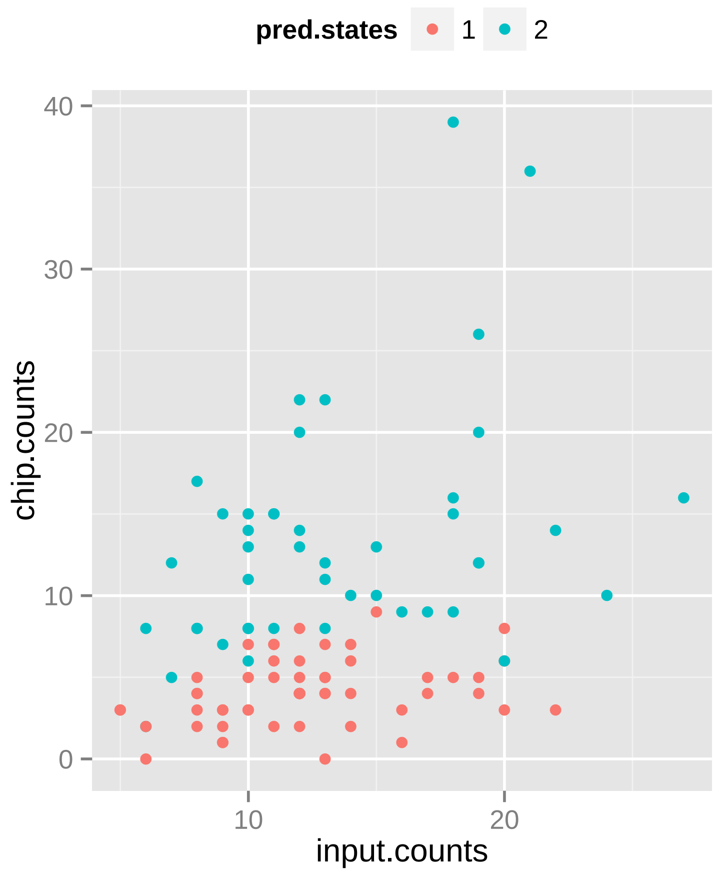
Function `plot` enables visualizing the range of the fitted parameter values across different replicates for the same condition. Notice that for these slots the function is implemented using *ggplot2*, so additional arguments can be passed by '+':

```
plot(fit, slot = "Mu", state = 2) + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Besides calling the plot function on our fitted object, we can also use ggplot to visualize how the loci within a single replicate are allocated to different states:

```
## which replicate to plot
repid <- 1
chip.counts <- dat$chip[, repid]
input.counts <- dat$input[, repid]
pred.states <- as.character(apply(fit@Theta[rownames(fit@Theta) == conds[repid], ], 2, which.max))
ggplot() + geom_point(aes(x = input.counts, y = chip.counts, color = pred.states)) + theme(legend.posit
```



4 Advanced Functions

4.1 Model Initialization

Function `MBASIC` accepts initial values for a number of parameters. This is especially useful when an initial run of `MBASIC` reaches the maximum number of iterations, but the algorithm is not converged. In this case, we can rerun the `MBASIC` using the previously fitted model parameters to initialize the model.

```
fit.update <- MBASIC(Y = t(dat$chip), Gamma = t(dat$input), S = 2, fac = conds, J = 3, maxitr = 10,
    family = "negbin", Mu.init = fit@Mu, Sigma.init = fit@Sigma, V.init = fit@V, ProbMat.init = fit@The
    W.init = fit@W, Z.init = fit@Z, b.init = fit@b, P.init = fit@P)
```

4.2 Multiple Components in a State

MBASIC model assumes that the distribution of each data is a mixture of several components, and the components are mapped to different states. In the simpliest cases, each component corresponds to a distinct state. In some cases, we might want to include multiple components in one state. For example, for ChIP-seq data, we may want to include two components for the enriched state to capture both the weakly enriched and strongly enriched loci. This can be done by specifying the 'statemap' value. In the following example, we assume there are three components for data set, the first component corresponds to the un-enriched state, the second and the third components both belong to the enriched states.

```
fit.mix <- MBASIC(Y = t(dat$chip), Gamma = t(dat$input), S = 2, fac = conds, J = 3, maxitr = 10,
    family = "negbin", statemap = c(1, 2, 2))
```

4.3 Simulation

The *MBASIC* package also provides functions to simulate and fit general MBASIC models.

Function `MBASIC.sim` simulates data with 'I' units and 'J' clusters. The 'S' argument specifies the number of different states, and 'zeta' is the proportion of singleton. 'fac' specifies the condition for each experiment. The 'xi' argument relates to the magnitude of the simulated data. For detailed description users are recommended to read our manual.

```
## Simulate data across I=1000 units with J=3 clusters There are S=3 states
dat.sim <- MBASIC.sim(xi = 2, family = "lognormal", I = 1000, fac = rep(1:10, each = 2), J = 3,
    S = 3, zeta = 0.1)
```

`MBASIC.sim` returns a list object. The 'Y' field contains the simulated data matrix at each unit (column) for each experiment (row). The 'Theta' field is the matrix for the states for each unit (column) and each experimental condition (column). The 'W' field is a matrix with dimensions KS $\times$ J, where the (S(k-1)+s,j)-th entry is the probability that units in the j-th cluster have state s under the k-th condition.

```
names(dat.sim)

##  [1] "Theta"      "Y"          "X"          "fac"        "W"          "Z"
##  [7] "V"          "delta"      "zeta"       "prior.mean" "prior.sd"   "stdev"
## [13] "Mu"         "bkng"       "snr"        "non.id"

dim(dat.sim$Y)

## [1]   20 1000

dim(dat.sim$W)

## [1] 30  3

dim(dat.sim$Theta)

## [1]   10 1000
```

We can apply `MBASIC` to this simulated data. If we pass the simulated data to the function through the 'para' argument, we can get the following slots on the estimation error:

- *ARI*: Adjusted Rand Index;
- *W.err*: The mean squared error in matrix W;
- *Theta.err*: The mean squared error in state estimation;
- *MisClassRate*: The mis-classification rate.

```
dat.sim.fit <- MBASIC(Y = dat.sim$Y, S = 3, fac = dat.sim$fac, J = 3, maxitr = 3, para = dat.sim,
    family = "lognormal")
```

```
dat.sim.fit@ARI

## [1] 0.9676749

dat.sim.fit@W.err

## [1] 0.1140384

dat.sim.fit@Theta.err

## [1] 0.222722

dat.sim.fit@MisClassRate

## [1] 0.03012732
```

4.4 Degenerate MBASIC Models

In a degenerate MBASIC model, the states for each unit under each condition are directly observed. `MBASIC.sim.state` and `MBASIC.state` functions allows users to simulate and fit such models. The usage of these functions are similar to functions `MBASIC.sim` and `MBASIC`.

`MBASIC.sim.state` simulates data from a degenerate MBASIC model. Different from `MBASIC.sim`, `MBASIC.sim.state` does not need arguments 'fac' and 'family', but it needs the 'K' argument, specifying the number of experimental conditions.

```
state.sim <- MBASIC.sim.state(I = 1000, K = 10, J = 4, S = 3, zeta = 0.1)
```

`MBASIC.state` fits a degenerate MBASIC model. Different to function `MBASIC`, it does not need arguments 'Y' and 'family'. Instead, it needs the argument 'Theta' to pass the observed states.

```
state.sim.fit <- MBASIC.state(Theta = state.sim$Theta, J = 4, zeta = 0.1)
```

5 Session Information

```
## R version 3.1.1 (2014-07-10)
## Platform: x86_64-redhat-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=zh_TW.UTF-8       LC_NUMERIC=C                 LC_TIME=en_US.UTF-8
##  [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] MBASIC_0.99.0        Rcpp_0.11.4          msm_1.5              mclust_4.4
##  [5] MASS_7.3-33          gtools_3.4.1         gplots_2.16.0        ggplot2_1.0.0
##  [9] GenomicRanges_1.14.4 XVector_0.2.0        IRanges_1.20.7       BiocGenerics_0.8.0
## [13] doMC_1.3.3           iterators_1.0.7      foreach_1.4.2        cluster_1.15.2
##
## loaded via a namespace (and not attached):
##  [1] BiocStyle_1.0.0    bitops_1.0-6       caTools_1.17.1     codetools_0.2-8
##  [5] colorspace_1.2-4   digest_0.6.8       evaluate_0.5.5     expm_0.99-1.1
##  [9] formatR_1.0        gdata_2.13.3       grid_3.1.1         gtable_0.1.2
## [13] highr_0.4          KernSmooth_2.23-12 knitr_1.9          labeling_0.3
## [17] lattice_0.20-29    Matrix_1.1-5       munsell_0.4.2      mvtnorm_1.0-2
## [21] plyr_1.8.1         proto_0.3-10       reshape2_1.4.1     scales_0.2.4
## [25] splines_3.1.1      stats4_3.1.1       stringr_0.6.2      survival_2.37-7
## [29] tools_3.1.1
```

References

[1] C. Zuo, Kyle Hewitt, Emery Bresnick, and S. Keleş. A hierarchical framework for state-space inference and clustering. Submitted, May 2015.