Bayesian BiClustering Manual

Jiajun Gu*
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA

Jun S. Liu
Department of Statistics
Harvard Univesity
Cambridge, MA, USA

1 Introduction

Bayesian BiClustering is an algoritm that searches for biclusters in gene expression data. It implements a Markov chain Monte Carlo procedure in C language.

2 Bayesian BiClustering Model

Consider a microarray dataset with N genes and P conditions (or samples), in which the expression value of the i^{th} gene and j^{th} condition is denoted as $y_{ij}, i = 1, 2, \cdots, N, j = 1, 2, \cdots, P$. We assume that

$$y_{ij} = \sum_{k=1}^{K}((\mu_k + \alpha_{ik} + \beta_{jk} + \epsilon_{ijk})\delta_{ik}\kappa_{jk}) + e_{ij}(1 - \sum_{k=1}^{K}\delta_{ik}\kappa_{jk}),$$

where K is the total number of clusters (unknown), μ_k is the main effect of cluster k, and α_{ik} and β_{jk} are the effects of gene i and condition j, respectively, in cluster k, ϵ_{ijk} is the noise term for cluster k, and e_{ij} models the data points that do not belong to any cluster. Here δ_{ik} and κ_{jk} are binary variables: $\delta_{ik} = 1$ indicates that row (gene) i belongs to cluster k, and $\delta_{ik} = 0$ otherwise; similarly, $\kappa_{jk} = 1$ indicates that condition (column) j is in cluster k, and $\kappa_{jk} = 0$ otherwise.

We assume *a priori* that

$$\begin{aligned}
\mu_k &\sim N(0, \sigma^2_{\mu k}) \\
\alpha_{ik}|\delta_{ik} = 1 &\sim N(0, \sigma^2_{\alpha k}) \\
\beta_{jk}|\kappa_{jk} = 1 &\sim N(0, \sigma^2_{\beta k}) \\
\epsilon_{ijk} &\sim N(0, \sigma^2_{\epsilon k}) \\
e_{ij} &\sim N(0, \sigma^2_{e}).
\end{aligned}$$

The hyperpriors for the $\sigma^2_{\mu k}, \sigma^2_{\alpha k}, \sigma^2_{\beta k}, \sigma^2_{\epsilon k}, \sigma^2_e$ are set to be inverse Gamma distributed.

3 Installation

The BBC software is free of charge for academic use. The program runs in LINUX/UNIX system. The program uses a few routines from the GNU Scientific Library (GSL), so users need install GSL first. For installation of the BBC program, please follow the two simple steps

[1]For questions about the software, please contact Jiajun Gu at jiajungu@fas.harvard.edu

- Download the source code (*.c and *.h files) and the Makefile to a folder.
- Compile and link by typing in "make". An executable file called "BBC" will be generated.

4 Bayesian BiClusterling program usage

After you install the BBC package and generate the executable file "BBC", you can start using the Bayesian BiClustering program. The BBC program command should be the following format:

BBC -i InputFileName -k NumberofClusters -o OutputFileName -n NormalizationMethod -r NormalizationAlphaValue

An example could be

BBC -i test.txt -k 2 -o out.txt -n iqrn -r 90

This command means the BBC algorithm is going to read in file "test.txt". It will use 90% inter-quartile range normalization (IQRN) on the dataset. Then the program will search for 2 biclusters and output the results in file "out.txt".

The exact meaning of each parameter is listed as below:

- InputFileName is the gene expression (Microarray) dataset. Please read Section 3 for the format of the input file.
- NumberofClusters means the number of clusters a user hopes to find in the datasets. Usually the exact number of clusters 'K' is not available, so we recommend searching for biclusters for a number of different 'K's and choose the best one according to Bayesian Information Criterion (BIC).
- OutputFileName is the file name for output file. Please read Section 4 for the format of the output file.
- NormalizationMethod refers to the normalization method to be used on the microarray data. We provide five different options:
  - None: No normalization will be done by the BBC program.
  - CSN: Column-wise standardization, where data in each column is re-centered and re-scaled, so that the sample mean of each column becomes 0, and the sample variance of each column becomes 1.
  - RSN: Row-wise standardization, where data in each row is re-centered and re-scaled, so that the sample mean of each row becomes 0, and the sample variance of each row becomes 1.
  - IQRN: Inter-quartile range normalization. One first sorts the data in each column, trims off $\alpha/2\%$ of the data from each tail, and computes the $\alpha\%$-trimmed mean and standard deviation. Then, all data in that column are standardized by subtracting the trimmed mean and being divided by the trimmed standard deviation. This normalization method can reduce the artificial normalization effect caused by outliers.

  - SQRN: The smallest range quartile normalization, one first finds for each column the shortest interval that contains a certain percentage, represented by α (for example, α could be 50, which means 50%) of the data. Then the data of that column is standardized by the sample mean and variance of the data inside the shortest quartile range. If distributions of the data in each column are symmetric and unimodal, then SQRN is equivalent to IQRN. But SQRN gives better results for skewed distributions.

- NormalizationAlphaValue (Optional): If a user chooses to run the BBC model with inter-quartile range normalization (IQRN) or the smallest quartile range (SQRN), then the alpha value for the normalization needs to be specified. According to the definition of the IQRN and SQRN, *alpha* value means the normalization is done based on the α% quartile of data in each column.

5 Input File Format

An input file should be a tab-delimited text file. The first row should contain sample/condition names. The first column should contain gene names. An example input file should look like:

Genes/Conditions	Exp 1	Exp 2	Exp 3	$\cdots$	Exp P
Gene 1	23.5	122	38.1	$\cdots$	57.3
Gene 2	88.09	277	-4.07	$\cdots$	6.33
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Gene N	11.09	-23.4	41.07	$\cdots$	121.3

6 Output File Format

An output file is a text file, which contains two parts. The first part summarizes the clustering results and the second part contains the genes and conditions for each of the K clusters. The summary information includes the following fields:

- Number of clusters: The number of clusters that a user specified
- Number of stable clusters: The number of non-empty clusters identified by the BBC algorithm. Sometimes if a user searchers for a large number of clusters, the algorithm will return some non-significant clusters. The non-significant clusters will be discarded automatically and only the significant clusters will be printed out in the output file.
- Likelihood: The likelihood of the model given the results
- Number of parameters: The total number of parameters in the results.
- BIC: Bayesian Information Criterion

For each cluster, both indicator and effect parameters are given. The first line gives the main effect of the bicluster. It follows by genes and the conditions information of the bicluster. Gene information includes row number, gene name and gene scores (the α_{ik} value of each gene). For example, if a line looks like

10 ABF 3.5

It means the 10th gene in the input file with gene name "ABF" belongs to this cluster, and the

gene score of this gene is 3.5. Condition information includes column number, condition name and condition scores (the β_{jk} value of each condition. For example, if a line looks like

10 Exp 10 0.3

It means the 10th condition in the input file with condition name "Exp 10" belongs to this cluster, and the condition score of this condition is 0.3.