MBBC: Model-Based Bayesian Clustering

Yongsung Joo[*], James G. Booth[†], George Casella[‡], and Younghwan Namgoong[§]

October 11, 2007

Abstract

The Bayesian product partition model in Booth *et al.* (2007) simultaneously searches for the optimal number of clusters, which is controlled by the tuning parameter in Crowely's prior, and clusters genes based on temporal changes of gene expressions. We developed MBBC v2.0 to make this method easily available for statisticians and scientists. MBBC v2.0 is built with three free computer language softwares, `OX`, `R` , and `C++`, taking own advantages of each language. Within MBBC, the search algorithm is implemented with `OX` and resulting graphs are drawn with R. User-friendly graphic interface is built with `DEV C++` to run `OX` and `R` programs internally. Thus, MBBC users aren't required to know how to use `OX`, `R`, or `DEV C++`. However, `OX` and `R` must be pre-installed to run MBBC properly.

Self-extractable zip file, MBBC20zip.exe, is available at MBBC webpage `www.phhp.ufl.edu/~yjoo/MBBC.html`. It contains MBBC.exe, source files, and all other related files. Free installation program for `OX` is available at `www.doornik.com/download_ oxcons.html` and overview is at `www.doornik.com/ox/`. Detailed installation guide for `OX` is provided in MBBC webpage and Help menu of MBBC v2.0, which is accessible without installing `OX`. Installation program for `R` is available at `www.r-project.org/`.

[*] Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32611

[†] Biological Statistics and Computational Biology, Cornell University, Ithaca NY 14853

[‡] Statistics, University of Florida, Gainesville, FL 32611, United States

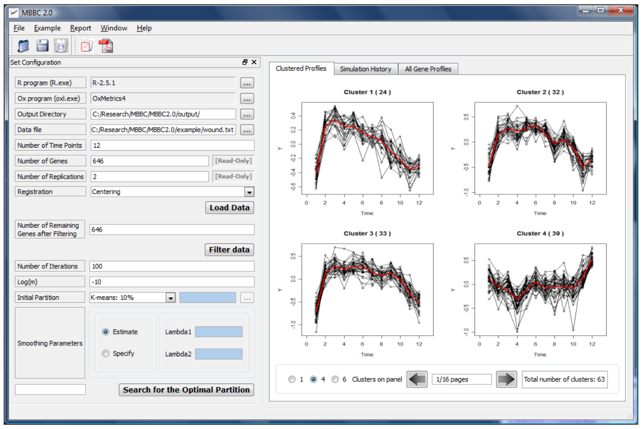
[§] Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, United States

1 Introduction

Since microarray analysis has become one of most important tools in genetics and genomics, clustering methods have been applied widely to select potential candidate genes for future research. In particular, it is increasingly common to measure gene responses over time, or as some function of experimental treatments. Booth *et al.* (2007) developed a novel Bayesian product partition model to cluster genes based on temporal changes of gene expressions. The split-merge algorithm in MBBC searches for a partition, ω, of genes that maximizes the posterior probability of the partition given the data, $\pi(\omega|y)$, which we call the Bayesian objective function. Unlike other clustering algorithms, such as k-means and mixture models (McLachlan and Basford 1988), MBBC does not require the user to predetermine the number of clusters in the optimal partition because partition ω is the parameter being searched for.

Searching for the optimal partition is a challenging problem. For a set of n genes or objects to be clustered, the number of all possible partitions is given by the Bell number, which grows super-exponentially. For example, for $n = 6$ the Bell number is 203, and for $n = 20$ it has 14 digits. Thus, an exhaustive search is infeasible in practical problems which typically involve hundreds, or even thousands, of gene profiles. MBBC uses MCMC optimization based on a split-merge algorithm for a stationary distribution proportional to the objective function. The Markov chain runs on the partition space, seeking partitions with large posterior probabilities. Such search algorithms can be quite effective (Jerrum and Sinclair 1996). Successful applications of the clustering algorithm in MBBC, along with detailed derivations and examples, are in Booth *et al.* (2007).

To make the algorithm available, MBBC has been developed with three freely available software packages: `OX`, `R`, and `DEV C++`. The `OX` is an matrix programming language like `R`. But, in our



R program (R.exe): R-2.5.1

Ox program (oxl.exe): OxMetrics4

Output Directory: C:/Research/MBBC/MBBC2.0/output/

Data file: C:/Research/MBBC/MBBC2.0/example/wound.txt

Number of Time Points: 12

Number of Genes: 646 [Read-Only]

Number of Replications: 2 [Read-Only]

Registration: Centering

Number of Remaining Genes after Filtering: 646

Number of Iterations: 100

Log(m): -10

Initial Partition: K-means: 10%

Smoothing Parameters: ● Estimate ○ Specify Lambda1 Lambda2

○ 1 ● 4 ○ 6 Clusters on panel 1/16 pages Total number of clusters: 63

Figure 1: Graphic user interface of MBBC v2.0, which shows input window and resulting clustered profiles.

experience, `OX` runs much faster than `R` and is comparable to `C++` in terms of speed. Considering the heavy computational burden of the clustering algorithm in MBBC, `OX` is used for all MCMC calculations. For easy implementation, `R` is used for graphics. User-friendly graphic interface, which is implemented with `DEV C++` (`www.bloodshed.net/devcpp.html`), runs `OX` and `R` programs internally; see Figure 1. In developing MBBC, own strength of `OX`, `R`, and `DEV C++` are used together for easy implementation of a fast and user-friendly program.

2 Approach

MBBC assumes that genes within each cluster follow a Bayesian linear mixed model with cluster specific and gene specific random effect terms. The mixed model design matrices are taken for each linear mixed model to be equivalent to the penalized quadratic spline regression with knots on all interior time points (Ruppert *et al.* 2003 and Booth *et al.* 2007). Genes in different clusters are assumed independent. Let $c(\omega)$ be the number of clusters in partition ω. Diffuse proper priors are used for all cluster-specific parameters in linear mixed models, $\theta = (\theta_1, \ldots, \theta_{c(\omega)})$. Crowley's prior (1997) is used for a partition parameter, ω. Then, the posterior distribution, $\pi(\theta, \omega|y)$, is constructed in the usual manner. Finally, the Bayesian objective function for the space of partitions, $\pi(\omega|y)$, is constructed by integrating out the cluster-specific parameter vector, θ. See Booth *et al.* (2005) for detailed calculations.

3 Methods

To find the optimal partition with respect to the Bayesian objective function, a number of options can be considered; Metropolis-Hastings, Gibbs sampling, biased random walk, and "heating" the chain by methods such as simulated tempering. All of these alternatives have been considered, and we have found that a well-tuned split-merge algorithm, the basis of MBBC, is simple and converges quite rapidly. Users prepare MBBC for a cluster analysis following through four conceptual steps; verification of `R` and `OX` installation, data loading, data filtering, and parameter specification.

Step1) If `R` and `OX` are installed properly, MBBC automatically detects the location of R.exe for `R` and oxl.exe for `OX`. If `OX` is not installed, users may refer to installation guide for `OX` in Help menu. If MBBC cannot find locations of those executive files in hard drive(s) even after proper installation, the user can specify locations manually.

Step2) Data matrix is the only input argument that MBBC user must specifies. MBBC assigns default values for other inputs in Step3 and 4. All genes in the data matrix must be measured at each time point with the same number of replicates. Then, user may choose to register the data; no registering, centering or standardization. If the user clicks "Load Data" button, the averages of each gene at each time point are calculated and then all average profiles are plotted on the right panel.

Step3) Even with an efficient search algorithm that we developed, it takes a long time to find the optimal partition when a large number of genes are considered to be clustered. Therefore, to speed up the algorithm still resulting in genetically meaningful clusters, we suggest reducing the number of genes to be clustered using one-way ANOVA models that have time as the predictor. In MBBC, the most time-varying genes are selected based on F-values. In MBBC, user can specify the number of most time-varying genes. The default number is 1000. By clicking "Filter Data", a reduced data set and a profile plot are generated.

Step4) Although the algorithm can run in fully automatic mode, MBBC allows the user to choose the number of MCMC iterations, an initial partition for the Markov chain, smoothing parameters in the linear mixed model, and the tuning parameter $\log(m)$ in Crowley's prior (1997), which controls the cluster sizes in the optimal model. (A smaller $\log(m)$ will favor larger clusters and vice versa.) As defaults, MBBC registers gene expressions with gene-specific centering, sets $\log(m) = -10$ and the number of iterations 10^6. The initial partition is generated using the k-means method with k set at one-tenth the number of genes. Smoothing parameters (λ_1 and λ_2 in Booth *et al.* 2007) are estimated based on the initial partition using the method of moments. By clicking "Search for the Optimal Partition", the user starts the search algorithm. The resulting clusters will be graphically illustrated in the right side panel; see Figure 1. Also, input parameter setup and results will be stored in pdf ("MBBCreport.pdf") and html("MBBCreport.html") files.

It is recommended to use MBBC with at least a 2.6 GHz CPU and 512MB memory. As a reference, for data consisting of 12 time point measurements on 646 genes (the Corneal Wound data set in Booth *et al.* 2006), MBBC takes 20 minutes to iterate the algorithm 10^6 times.

4 Discussion

When the convergence of search algorithm is examined, the choice of an initial partition is important. One wants the algorithm to fully explore the space of all partitions, and be able to reach correct convergence from any starting point. MBBC provides three noninformative choices of initial partitions (a uniformly chosen random partition, n-cluster or 1-cluster) and two informative choices (k-means algorithm or a user-defined partition). Pitman (1997) developed an algorithm to generate a uniformly chosen random partition, but it was numerically feasible for only a small number of objects (genes). We improved the applicability of this algorithm to generate uniformly random starting points with any number of objects. Alternatively, the user may specify an initial partition with n-cluster or 1-cluster, where each object starts as its own cluster or all clusters does as one cluster. It is also possible to provide a user-defined initial partition, presumably based on prior knowledge of possible clusters. Such a specification may allow the algorithm to converge faster, and this is a recommended method if feasible. Lastly, if the user has prior knowledge only on the number of clusters, the k-means algorithm may provide a reasonable initial partition. MBBC uses function kmeans($\cdot$) in R. For example, if "k-means: 10" is chosen, the kmeans algorithm searches for the best partition with 10 clusters and then MBBC uses it as an initial partition. Even though users are allowed to choose different initial partitions, we found that the stochastic optimization algorithm in MBBC is insensitive to these initial partitions once the same smoothing parameters are given. Simulation results are linked on the MBBC manual in Help menu.

The statistical model in MBBC has two smoothing parameters for cluster-specific and gene-specific random effects. The user may specify the smoothing parameters or allow MBBC to estimate them. If the initial partition has less than n clusters, MBBC estimates these using the method of moments in a frequentist manner based on the initial partition. The document with detailed explanation are linked on the MBBC manual.

5 Conclusion

MBBC makes a sophisticated Bayesian clustering method computationally practical and easy to use. It can handle a relatively large number of genes within a limited computational time. However, if the number of genes is excessively large, it may make simulation too slow or even cause out-of-memory errors. The feasible size of the data depends on the number of time points, the number of replicates, and the computational environment. It is always helpful for user to guess a proper number of iterations by experimenting runs with a small number of iterations; then, initiate a long simulation. MBBC draw simulation history plots of the objective function values at each iteration and the highest objective function values up to each iteration. If simulation is long enough, stabilized distribution must be observed in the first plot and convergence to the highest objective function value must be observed in the second plot.

References

Booth, J., Casella, G. and Hobert, J. (2007) Clustering Using Objective Functions and Stochastic Search. To-appear *Journal of Royal Statstical Society B.*, available at `www.bscb.cornell.edu/~booth/papers.html`.

Crowley, E.M. (1997). Product partition models for normal means. *Journal of the American Statistical Association*, 92, 192-198.

Jerrum, M. and Sinclair, A. (1996) The Markov Chain Monte Carlo method: An approach to approximate counting and integration, in *Approximation Algorithms for NP-hard Problems.* PWS Publishing, Boston.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture models: Inference and applications to clustering.* Marcel Dekker, Inc., New York.

Pitman, J. (1997). Some probabilistic aspects of set partitions. *American Mathematical Monthly*, 104, 201-209.

Ruppert, D., Wand, M.P. and Caroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press, New York, 2003.