BIOINFORMATICS

Vol. 00 no. 00 0000
Pages 1–2

Model-Based Bayesian Clustering (MBBC)

Yongsung Joo[1], James G. Booth[2], Younghwan Namkoong[3], and George Casella[4*]

[1]Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32611, United States
[2]Biological Statistics and Computational Biology, Cornell University, Ithaca NY 14853, United States
[3]Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611, United States
[4] Department of Statistics and Genetics Institute, University of Florida, Gainesville, FL 32611, United States

ABSTRACT

Motivation: The program MBBC 2.0 clusters time course microarray data using a Bayesian product partition model (Booth *et al.* 2007).

Results: The Bayesian product partition model in Booth *et al.* (2007) simultaneously searches for the optimal number of clusters, and assigns cluster memberships based on temporal changes of gene expressions. MBBC 2.0 to makes this method easily available for statisticians and scientists, and is built with three free computer language software packages: `Ox`, `R` , and `C++`, taking advantage of the strengths of each language. Within MBBC, the search algorithm is implemented with `Ox` and resulting graphs are drawn with `R`. A user-friendly graphical interface is built with `C++` to run the `Ox` and `R` programs internally. Thus, MBBC users are not required to know how to use `Ox`, `R`, or `C++`, but they must be pre-installed.

Availability: A self-extractable zip file, MBBC20zip.exe, is available at the MBBC webpage `www.stat.ufl.edu/~casella/mbbc/`, which contains MBBC.exe, source files, and all other related files. The current version works only in Windows operating system. A free installation program and overview for `Ox` is available at `www.doornik.com`. A detailed installation guide for `Ox` is provided by MBBC, and is accessible without installing `Ox`. `R` is available at `www.r-project.org/`.

Contact: `casella@stat.ufl.edu`

1 INTRODUCTION

Clustering methods have been applied widely in microarray studies to select potential candidate genes for future research, and it is increasingly common to measure gene responses over time. Booth *et al.* (2007) developed a 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 (Hartigan and Wong 1979) and mixture models (McLachlan and Basford 1988), MBBC does not require the specification of the number of clusters.

*to whom correspondence should be addressed

Searching for the optimal partition is a challenging problem, as the number of possible partitions of n objects 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 not feasible in practical problems which typically involve hundreds, or even thousands, of gene profiles. MBBC uses an MCMC optimization which searches a stationary distribution that is proportional to the objective function. The Markov chain runs on the partition space, seeking partitions with large posterior probabilities, $\pi(\omega|y)$. Such search algorithms can be quite effective (Jerrum and Sinclair 1996). Successful applications and derivation of the clustering algorithm in MBBC is in Booth *et al.* (2007).

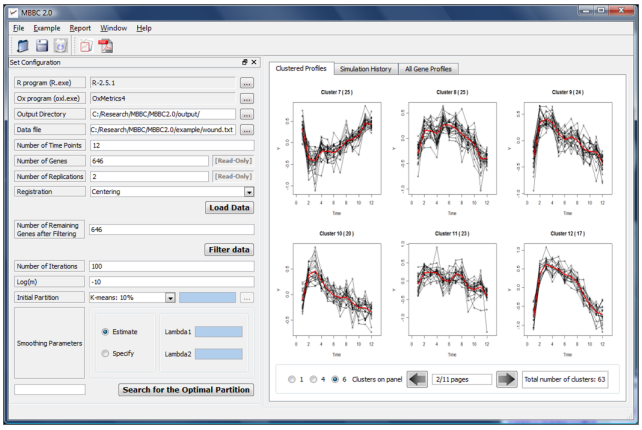

Fig. 1. Graphic user interface of MBBC 2.0, which shows input and resulting clustered profiles.

MBBC has been developed using the three software packages `Ox`, `R`, and `C++` (`http://trolltech.com`). `Ox` is a matrix programming language similar to `R`, but runs much faster and is comparable to `C++` in terms of speed. `Ox` is used for all MCMC calculations, and `R` is used for graphics. A user-friendly graphic interface, implemented with `C++`, runs all programs (Figure 1).

2 APPROACH

MBBC models the time-course of genes within each cluster nonparametrically, using a Bayesian linear mixed model with cluster and gene specific random effects. This is equivalent to penalized quadratic spline regression with knots at all interior time points (Ruppert *et al.* 2003; Booth *et al.* 2007). Genes in different clusters are assumed independent, but there can be correlation within clusters.

For a particular partition ω, $c(\omega)$ is the number of clusters in the partition and $\theta = (\theta_1, \ldots, \theta_{c(\omega)})$ are the cluster specific parameters. Diffuse proper priors are used for all cluster-specific parameters, and a prior given in Crowley (1997) is used for ω. The Bayesian objective function over the space of partitions, $\pi(\omega|y)$, is the posterior distribution $\pi(\theta, \omega|y)$ marginalized over the cluster-specific parameters. See Booth *et al.* (2007) for details.

3 METHODS

There are four steps to the MBBC cluster analysis; installation of R and Ox, data loading, data filtering, and parameter specification.

Step 1) MBBC automatically detects the location of R.exe for R and oxl.exe for Ox. If Ox is not installed, users may refer to the installation guide for Ox in the Help menu. If MBBC cannot find the locations of those executable files in the hard drive(s) even after proper installation, the user can specify the locations manually.

Step 2) The data matrix is the only required input to MBBC. All genes in the data matrix must be measured at each time point with the same number of replicates. The number of time points must be greater than 2, and the data may be centered or standardized. Clicking the "Load Data" button calculates the averages of each gene at each time point, and all average profiles are plotted on the right panel.

Step 3) Even an efficient search algorithm can take a long time to find the optimal partition when there are a large number of genes. To speed up the algorithm while still obtaining genetically meaningful clusters, we can reduce the number of genes to be clustered using polynomial regression models over time. With p time points the order of the regression model is min(p-2,3), where $p \geq 3$. The genes with the most variable time profiles are selected using the anova F-values. The user can specify the number of genes retained, with a default of 1000. By clicking "Filter Data", a reduced data set and a profile plot are generated.

Step 4) Although the algorithm can run in automatic mode, MBBC allows the choice of the number of iterations, an initial partition, smoothing parameters in the linear mixed model, and the tuning parameter log(m) in Crowley's prior (1997). (A smaller log(m) will favor larger clusters and vice versa.) This parameter is easy to specify as it is directly related to the *prior* mean number of the clusters, $m\sum_{i=1}^{n}(m+i-1)^{-1}$, which may be available from biological considerations. The user may specify this prior information with either log(m) or the prior mean. As defaults, MBBC registers gene expressions with gene-specific centering, sets $\log(m) = -10$ (which, for essentially all n, corresponds to a prior specification of one cluster), and the number of iterations to 10^4. 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) can either be specified or estimated based on the initial partition using the method of moments. By clicking "Search for the Optimal Partition", the search algorithm starts. The resulting clusters are graphically illustrated in the right side panel; see Figure 1. Also, the input parameter setup and results will be stored in the files ("MBBCreport.pdf" and "MBBCreport.html").

It is recommended to have at least a 2.6 GHz CPU and 512 MB 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 approximately 40 minutes to iterate the algorithm 10^6 times.

4 DISCUSSION

MBBC provides three noninformative choices of initial partitions (a uniformly chosen random partition, n-clusters or 1-cluster) and two informative choices (k-means algorithm or a user-defined partition). For example, if "k-means: 10" is chosen, the k-means algorithm searches for the best partition with 10 clusters and then MBBC uses it as an initial partition. Alternatively, the user may start each object in its own cluster or all objects in one cluster. Even though users are allowed to choose different initial partitions, we found that the stochastic optimization algorithm in MBBC is relatively insensitive to these choices (details given on the MBBC webpage).

The statistical model in MBBC has two smoothing parameters for cluster-specific and gene-specific random effects. The user may specify these smoothing parameters or allow MBBC to estimate them. This is also documented on the MBBC webpage.

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 *reasonable computation time*. In terms of computation time, 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 the user to estimate a proper number of iterations by experimenting with a small number of iterations, such as 10^4; then, initiate a long simulation. MBBC draws simulation history plots of the objective function values at each iteration and the highest objective function values up to each iteration. If a simulation is long enough, a stabilized distribution will be observed in the first plot and convergence to the highest objective function value will 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 the Royal Statstical Society Series B.*. Available at www.stat.ufl.edu/∼casella/Papers/clustering07.pdf.

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

Hartigan, J.A. and Wong, M.A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28, 100-108.

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.

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