Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach

Moysés Nascimento1*, Thelma Sáfadi2, Fabyano Fonseca e Silva1, Ana Carolina C. Nascimento1, Ana Carolina C. Nascimento1, Ana Carolina C. Nascimento1
1 Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brasil, 36570-000.
2 Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil, 37200-000.

ABSTRACT

Motivation: In a microarray time series analysis, due to the large number of genes evaluated, the first step towards understanding the complex time network is the clustering of genes that share similar expression patterns over time. Up until now, the proposed methods do not point simultaneously to the temporal autocorrelation of the gene expression and the model-based clustering. We present a Bayesian methodology that considers jointly the fit of autoregressive panel data models and hierarchical gene clustering.

Results: The proposed methodology was able to cluster genes that share similar expression over time, which was determined jointly by the estimates of autoregression parameters, by the average level of expression) and by the quality of the fitted model.

Availability and implementation: The R codes for implementation of the proposed clustering method and for simulation study, as well as the real and simulated data sets, are freely accessible on the Web http://www.det.ufv.br/~moyses/links.php.

Contact: moysesnascim@ufv.br, fabyanofonseca@ufv.br, safadi@dex.ufa.br, campanaaca@gmail.com
Supplementary Information: Supplementary Figures are available at Bioinformatics online.

1 INTRODUCTION

Microarray Time Series (MTS) analysis allows the researcher to characterize set of genes through their longitudinal pattern of expression. According to Schliep et al. (2003), the MTS data analysis methodologies can be divided into two classes. The first one assumes the observations on the expression at each time as independent variables, and so the usual methods such as hierarchical process (Eisen et al., 1998) and the optimization (Tavazoie et al., 1999) can be directly used to cluster genes with similar expression pattern. The second, performs clustering based jointly by the estimates of autoregression parameters, by the average level of expression) and by the quality of the fitted model. Considering these advantages of model-based clustering methods, and the difficulty of applying them to a large number of small series, here we propose a Bayesian method that considers simultaneously an autoregressive panel data model fit and a hierarchical clustering of the parameter estimates from this model.

Availability and implementation: The R codes for implementation of the proposed clustering method and for simulation study, as well as the real and simulated data sets, are freely accessible on the Web http://www.det.ufv.br/~moyses/links.php.

Contact: moysesnascim@ufv.br, fabyanofonseca@ufv.br, safadi@dex.ufa.br, campanaaca@gmail.com
Supplementary Information: Supplementary Figures are available at Bioinformatics online.
clusters contain series with similar gene expression results were always the same after the convergence. Under this framework whose expression series were modeled by an AR(2) panel data distributions, and due to this, the algorithm was run considering drawing random numbers from full conditional posterior is also guaranteed. Nevertheless, the Gibbs sampler involves optimal clustering of which the input data should be belonged to) convergence, the convergence to the globally optimal solution (the therefore this number can change in each iteration. The gene clustering was defined by Mojena’s criterion (Mojena, 1977), new clustering output. The number of clusters in each Ward analysis (Ward, 1963), then for each resulting cluster, the model 1 estimates in each Gibbs sampler iteration, and consequently in a was fitted again. This procedure results in a new set of parameter estimates for model 1 were obtained. These estimates were used as input variables in a Ward clustering which initially a single panel (only one cluster with all genes) from was simulated the gene expression time series for each gene into time points using the model described in (1). The function arima.sim of R software (R Development Core Team, 2011) was used to simulate the gene expression time series for each gene into each cluster as follow:

\[ Y_k = \mu_k + \text{arima.sim}(n=10, list(ar=c(\phi_{1k}, \phi_{2k}), sd=\text{sqrt}(\sigma^2_{0k}))) \]

where: \( Y_k \) is the time series (a vector with dimension 1 x 10) of gene \( i \) \((i=1,2,\ldots,I)\) of cluster \( k \) \((k=1,2,\ldots,5)\), \( \mu_k \) is the intercept (average of gene expression), the term \( ar \) indicates a autoregressive process (model 1) with parameters \( \phi_1 \) and \( \phi_2 \) (AR(2) process), and \( \sigma^2_{0k} \) is the residual variance.

The numbers of genes and longitudinal points were chosen taking into account the dimension of the real dataset presented in the next item, and the number of clusters (five) and the values of \( \mu_k, \phi_{1k}, \phi_{2k}, \) and \( \sigma^2_{0k} \) were determined in according to the results from the real data analysis that will be presented later. For example, the values assumed for \( \phi_{1k} \) in the simulation were the

\[
\begin{bmatrix}
X_1 & 0 & 0 & 0 \\
0 & X_2 & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & X_n
\end{bmatrix}
\]

being the dimensions of these two last matrices \( m(n-p) \times m(p+1) \) and \( (n-p) \times (p+1) \), respectively.

According to the Bayes theorem, the posterior distribution for the parameters of interest \( (\Phi, \sigma^2) \) is given by:

\[ P(\Phi, \sigma^2 | Y) \propto L(Y | \Phi, \sigma^2) \cdot P(\Phi | \sigma^2) \cdot P(\sigma^2) \]

where: \( L(Y | \Phi, \sigma^2) \) is the likelihood function presented in (2), and \( P(\Phi | \sigma^2) \) and \( P(\sigma^2) \) make a hierarchical Normal - Inverse Gamma prior distribution:

\[ \Phi | \sigma^2 \sim N(m(n-p) \times m(p+1), \sigma_0^2 I) \]

\[ \sigma^2 \sim IG(\alpha, \beta) \]

Under this approach, the Gibbs sampler algorithm was implemented in R software (R Development Core Team, 2011) using the following full conditional posterior distributions:

\[ \Phi | \sigma^2, Y \sim N(m(n-p) \times m(p+1), \sigma_0^2 \Sigma) \]

\[ \sigma^2 | \Phi, Y \sim GL\left( \frac{m(n+1)+2\alpha}{2}, D + \frac{1}{2} (\Phi - \Phi_0)^T \Sigma^{-1} (\Phi - \Phi_0) \right) \]

where:

\[ D = \beta + \left( Y^T Y + \mu^T I \right) - \left( X^T X + I \right)^T \left( X^T X + I \right)^{-1} \left( X^T Y + I \mu \right) \]

\[ \hat{\Phi}_0 = \left( X^T X + I \right)^{-1} \left( X^T Y + I \mu \right) \]

\[ \Sigma = \left( X^T X + I \right)^{-1} \]

and \( I \) is an identity matrix.

2.1 Iterative method for clustering genes with similar gene expression patterns

The gene clustering was performed using iterative process in which initially a single panel (only one cluster with all genes) from which the parameter estimates from model 1 were obtained. These estimates were used as input variables in a Ward clustering analysis (Ward, 1963), then for each resulting cluster, the model 1 was fitted again. This procedure results in a new set of parameter estimates in each Gibbs sampler iteration, and consequently in a new clustering output. The number of clusters in each Ward clustering was defined by Mojena’s criterion (Mojena, 1977), therefore this number can change in each iteration.

Thus, once guaranteed the Gibbs sampler chains convergence, the convergence to the globally optimal solution (the optimal clustering of which the input data should be belonged to) is also guaranteed. Nevertheless, the Gibbs sampler involves drawing random numbers from full conditional posterior distributions, and due to this, the algorithm was run considering three different sets of start values in order to verify if the clustering results were always the same after the convergence.

Figure 1 shows a scheme of the proposed method for genes whose expression series were modeled by an AR(2) panel data model, the simplest multi parametric model. Under this framework our main goal is that at the end of the algorithm, the resulting clusters contain series with similar gene expression patterns over time, according to the parameter estimates of the given model (Figure 1 at supplementary material).

Fig. 1. Flowchart summarizing the Bayesian model-based clustering using autoregressive panel data approach.

2.2 Implementation

We implemented our program using R statistical computing environment, with the computation conducted with an Intel Core 2 Duo E7500 2.0 GHz with 4 GB of RAM.

2.3 Application to simulated data

We evaluated the proposed clustering method performance using ten simulated datasets, each one with 117 genes divided into 5 clusters. The genes expression levels were generated over 10 time points using the model described in (1). The function arima.sim of R software (R Development Core Team, 2011) was used to simulate the gene expression time series for each gene into each cluster as follow:
average values of the estimates obtained from each resulting cluster $k$, i.e., $\phi_k = \sum_{i=1}^{I_k} \hat{\phi}_{ik} / I_k$, where $\hat{\phi}_{ik}$ is the estimate of $\phi_i$ for each gene belonging to cluster $k$, and $I_k$ is the number of genes in the cluster $k$. The values of $\mu_k, \phi_k, \sigma^2_k$ were obtained in same way as $\hat{\phi}_i$.

In order to compare the clustering method present here with a traditional clustering method, the ten simulated datasets also were analyzed using k-means algorithm, in which the input variables were the observed gene expression level in each time. Under this approach, given a set of time series gene expression observations: $Y_1 = [y_{11}, \ldots, y_{110}], \ldots, Y_{10} = [y_{11}, \ldots, y_{110}]$, the k-means algorithm aims to partition the 117 genes into $K$ ($K \leq 117$) sets $S_1, S_2, \ldots, S_K$ so as to minimize the within-cluster sum of squares: $\arg \min_S \sum_{k=1}^{K} \sum_{y_i \in S_k} \| y_i - \bar{S}_k \|^2$, where $\bar{S}_k$ is the mean of point in $S_k$. The choice of k-means is due to its general use in gene expression clustering (Oh et al., 2011 and Mar et al., 2011), but it is important emphasize that the success of this method depends of the stated number of clusters, which was the same number used in the simulation study (five), i.e., was considered the optimum condition for the application of this method.

The comparison between Bayesian model-based clustering using autoregressive panel data approach and k-means was evaluated by correct classification rate (CCR), which was computed as the ratio between the numbers of genes clustered into the true cluster (from simulation) and the total number of genes

2.4 Application to real data

We applied the proposed method to expressions of 117 genes that act on cell cycle of the *Saccharomyces cerevisiae* (Zhu et al., 2000). We used 10 points of the non-synchronized cells data, whose fold-change values came from the expression of mutant strains (treated) compared to wild strains (control), at each evaluated time (0, 15, 30, ..., 135 minutes). At first, the choice by non-synchronized data was a manner to show that the proposed methodology works well for not only highly selected datasets, but also for general class of microarray datasets. For example, if a cell culture is separated out by time, and the expression of genes is measured as a function of time, one can get an idea of which genes vary in expression over time using regression models (Ramoni et al., 2002 and Bar-Joseph et al., 2003) with time as a covariate. However, considering that non-synchronized cell populations contain cells at various points in the cell cycle, under this common regression approach, the time effect can be confounded with cell cycle. By the other hand, in the autoregressive regression model (1) the covariate is not time itself but instead is the expression levels measured at the previous time, thus avoiding the confounding between cell cycle and time. The used data set may be downloaded at: http://smd.stanford.edu.

3 RESULTS

3.1 Simulated data

Performance of the clustering methods based on the simulated data is presented in Figure 2, in which the proposed correct classification rate (CCR) in item 2.3 was used for evaluation over 10 replicated datasets.

Comparison of the approaches does show that k-means clustering performs poorly in relation to Bayesian model-based, once the CCR values were clearly lower. On average, these values were 47.5 (12.56) % and 67.42 (9.44) %, respectively for k-means and Bayesian model-based methods.

Although the simulated data set was generated under model (1) assumptions, the comparison between the proposed model-based clustering (using parameter estimates as input variables) and the traditional k-means (using observed gene expression level in each time as input variables) methods makes sense. It is because there are biological justifications to dependence between a gene expression at time $t$ and previous time, thus whether this dependence is disregarded in the clustering process (as occurred with k-means application) really the results will not be satisfactory.

We would like to make clear that the simulation study compared the clustering effectiveness of a new method, which take into account the temporal dependence of gene expression values using Ward algorithm, and a traditional method, k-means, that considers the independence between these values. Maybe a more honest comparison would be realized between k-means and Ward using this same Bayesian model-based approach, but it can be a topic for future research, once the use of k-means rather than Ward requires an iterative process (k-means optimization) into each MCMC iteration.

3.2 Real data

The algorithm had a running time of 24'46" considering Gibbs sampler chains with 10,000 iterations. The chains convergence was accessed by Geweke and Raftery-Lewis diagnostics using boa package (Smith, 2007) of R software (R Development Core Team, 2011). These results indicated that the iteration number was enough to ensure the convergence, and furthermore, it was found that the number of clusters ($k$) and individuals belonging to them stabilized from around 1,000 iterations. Also, it is relevant to
highlight that the three different sets of start values provided the same clustering output.

Results allowed genes to be partitioned into five distinct clusters, with 23, 32, 15, 24 and 23 genes in each one (Figures 3 and 4 at supplementary material). In general, it can be observed that the 5 clusters have very distinct expression patterns. Among the various differences, it can be noted that the genes that make up clusters 1 and 2 (Figure 3) have opposite average behavior during cell cycle, with observed fold-change values, respectively, for clusters 1 and 2, negative and positive until a given time, and then, after those times the signals are inverted (Figures 3B and 3D). Furthermore, genes belonging to cluster 5 (Figure 4) generally showed more expression in the control (wild strains) during the cell cycle of Saccharomyces.

![Fig. 3. Time serial expression of three gene clusters found by the proposed methodology. (A) cluster 1 expression profile, (B) cluster 1 average expression, (C) cluster 2 expression profile, (D) cluster 2 average expression, (E) cluster 3 expression profile, (F) cluster 3 average expression.](image)

![Fig. 4. Time serial expression of two gene clusters found by the proposed methodology. (A) cluster 4 expression profile, (B) cluster 4 average expression, (C) cluster 5 expression profile, (D) cluster 5 average expression.](image)

In summary, we can verify that the proposed methodology was able to cluster genes that share similar expression pattern over time. This similarity can be explained jointly by the magnitude of the dependence between expression at time t and the expressions at time t-1 and t-2 (estimates of autoregression parameters, $\phi_1$ and $\phi_2$), by the average level of expression ($\mu$) and by the quality of the fitted model ($\sigma^2$), once these values were considered as input variables in the clustering rather than the observed gene expression values over time.

4 CONCLUDING REMARKS

In both simulated and real datasets, the proposed Bayesian model-based clustering method performed very well, once allowed to cluster genes that share similar expression pattern over time. This method was based on the autoregressive panel data model, which is widely used in the econometrics field. Thus, one of the main contributions of the present study is to arouse interest of bioinformatician for models from this field in order to describe the time dependence of gene expressions. For example, other powerful models like ARIMA (autoregressive integrated moving average) and ARCH (autoregressive conditional heteroskedasticity) also can be considered. Furthermore, high-throughput sequencing data from time series RNA-seq assays also can be analyzed under this approach. However, because to the counting nature of these data (number of reads in a class), other distributions, such as Poisson and Negative Binomial, must be adopted in the likelihood function. With respect to inference methods, the maximum likelihood (Ideker et al., 2000) can be used rather than Bayesian method using generalizations of the EM algorithm

REFERENCES


