Model-based clustering of time-course RNA-seq data

Man-Kee Maggie Chu and Wenqing He
Department of Statistical and Actuarial Sciences,
The University of Western Ontario, London, ON, CANADA
Corresponding author: Man-Kee Maggie Chu, e-mail: mchu5@stats.uwo.ca

Abstract

The next generation sequencing technology (RNA-seq) provides absolute quantification of gene expression using counts of read. Transcriptome studies are switching to rely on RNA-seq rather than microarrays since RNA-seq has higher sensitivity and dynamic range, with lower technical variation and thus higher precision than microarrays. Limited work has been done on expression analysis of longitudinal RNA-seq data to account for the time-dependence nature of the count data with over-dispersion property. Functional clustering is an important method for examining gene expression patterns and thus discovering co-expressed genes to better understand the biological systems. We propose a model-based clustering method for identifying gene expression patterns using time-course RNA-seq data. A time-course genomic dataset is employed for illustration.

Keywords: Expectation-Maximization algorithm, longitudinal experiments, over-dispersed count data, mixture model.

1 Introduction

Microarrays and sequence-based methods are both often used in gene expression studies, with an increasing popularity of the use of RNA-seq over microarrays in transcriptome analyses. Statistical methods used for differential expression analysis with these two technologies are different because microarray intensities are continuously distributed, whereas RNA-seq gives discrete measurement of reads for each gene. Researchers have tested for differential expression in RNA-seq data using Poisson distributions (Bullard et al., 2010; Marioni et al., 2008) but it has been shown that the Poisson assumption of equivalent mean and variance ignores the extra variation arises from the differences in replicate samples (Nagalakshmi et al., 2008; Robinson and Smyth, 2007). Over-dispersed models are more suitable for accommodating the over-dispersion in RNA-seq data.

Functional clustering of genomic data can identify co-expressed genes with similar functions and help explain the complexities of biological systems. Exploring the patterns shown in genomic data from time-course experiments which provide us with important information on changes in expression levels over time. The development of clustering algorithms suitable for RNA-seq data becomes an important area of research since they allow for analysis of multiple treatment groups rather than simple two-group analyses. Clustering methods have been widely applied to time-course microarray data (Cooke et al., 2011; Grün et al., 2012; Ng et al., 2006;

Schliep et al., 2003; Yuan and He, 2008) but these clustering approaches are based on continuous distributions and thus not appropriate for the discrete-type RNA-seq data.

In order to effectively model the information provided by RNA-seq data, we consider an efficient data clustering method to identify patterns on gene expression data from time-course RNA-seq experiments. The goal is to use a model-based clustering approach to identify co-expressed genes and their expression patterns from gene expression levels measured by read counts over time. In the next section, we present the mixture model, the Expectation-Maximization (EM) algorithm and a hybrid-EM algorithm for our clustering approach. An application to real data is presented in section 3, and section 4 includes the conclusion to the proposed clustering approach.

2 Model and Estimation

The clustering of time-course RNA-seq data can be viewed as identifying developmental trajectories (or temporal pattern) within a RNA-seq measured gene expression dataset. The semi-parametric group-based trajectory modeling approach (Nagin, 1999) is an example of model-based clustering method for longitudinal data. The method models the data as a mixture of distinct groups/clusters defined by their trajectories, and differences that may explain individual- (or sample-) level variability can be expressed in terms of cluster differences. Since RNA-seq data suffers from the over-dispersion problem, a common approach is to model the count data using negative binomial distributions to accommodate over-dispersion. Here we develop an efficient model-based clustering method with mixtures of negative binomials to cluster time-course RNA-seq data using the semi-parametric group-based modelling approach proposed by Nagin (1999), and an EM algorithm for maximum likelihood estimation is incorporated to estimate the parameters in our clustering approach.

When working with time-course RNA-seq data, we denote $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ as the time-series read counts of gene 1 to gene n in the sample and $Pr(\boldsymbol{Y}_j)$ as the probability of observing a specific time-series sequence of read counts on gene j over time. With $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)^T$ denoting the observed random sample where $\boldsymbol{y}_j$ is the realization of the random variable $\boldsymbol{Y}_j$, we can represent the data as a standard g-component mixture model in the form

$$f(\boldsymbol{y}_j; \boldsymbol{\psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i),$$

where $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ is the component density for component i, and the corresponding likelihood is given by

$$L(\boldsymbol{\psi}) = \prod_{j=1}^{n} f(\boldsymbol{y}_j; \boldsymbol{\psi}).$$

The component density $f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ is the conditional density function of $\boldsymbol{Y}_j$ given component membership of the i^{th} component with component parameter $\boldsymbol{\theta}_i$, and $\boldsymbol{\psi} = (\pi_1, ..., \pi_g, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_g)$ is the set of model parameters from the different mixture

components. The goal is to obtain a set of estimates for the parameters such that the likelihood is maximized. The parameters $\boldsymbol{\psi}$ define the shapes of the expression pattern curves for the clusters and the probability of cluster memberships. The shape of each expression pattern curve (or trajectory) is described by a statistical model, and a separate set of parameters is estimated for each group to allow the shapes of curves to differ across groups.

Let $\boldsymbol{y}_j = (y_{j1}, ..., y_{jm})$ be the observed read counts at m time points for each gene j. Utilizing the flexibility provided by polynomial functions, we assume a quadratic relationship between time and read counts; and conditional on being in cluster i, each gene has independent observations over time. The cluster parameter $\boldsymbol{\theta}_i$ includes $\boldsymbol{\beta}^i$ and s_i, where $\boldsymbol{\beta}^i = (\beta_0^i, \beta_1^i, \beta_2^i)$ determines the shape of the trajectory and the parameter s_i describes the dispersion of the genes in cluster i, and these parameters are allowed to differ across clusters.

We use a negative binomial model for the read counts and conditional on being in group i, a gene j is assumed to have independent read counts over the m sampling time points, so we have

$$f_i(\boldsymbol{y}_j; \boldsymbol{\beta}^i, s_i) = \prod_{t=1}^{m} \left(\frac{\Gamma(y_{jt} + s_i)}{y_{jt}! \Gamma(s_i)} p_i^{s_i} (1 - p_i)^{y_{jt}} \right)$$

with mean

$$\lambda = exp(\boldsymbol{\beta}^i \mathbf{x}_{jt}) = exp(\beta_0^i + \beta_1^i time_{jt} + \beta_2^i time_{jt}^2)$$

and s_i being the dispersion parameter for the group i and probability $p_i = \frac{s_i}{s_i + \lambda}$. The mixture likelihood for the entire sample of n genes in g clusters is

$$L(\boldsymbol{\psi}) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y_j}; \boldsymbol{\beta}^i, s_i)$$

Defining the missing data vector $\boldsymbol{z} = (\boldsymbol{z}_1^T, ..., \boldsymbol{z}_n^T)^T$ with $\boldsymbol{z}_j = (z_{1j}, ..., z_{gj})^T$ reflecting the component membership of gene j, the complete-data likelihood for a sample of n genes can be written as

$$L_c(\boldsymbol{\psi}) = \prod_{j=1}^{n} \prod_{i=1}^{g} \pi_i^{z_{ij}} f_i(\boldsymbol{y_j}; \boldsymbol{\beta}^i, s_i)^{z_{ij}}$$

and the maximum likelihood estimates $\hat{\boldsymbol{\psi}} = (\hat{\pi_1}, ..., \hat{\pi_g}, \hat{\boldsymbol{\beta}^1}, ..., \hat{\boldsymbol{\beta}^g}, \hat{s}_1, ..., \hat{s}_g)$ based on the complete data can be obtained by maximizing the log-likelihood.

For our model-based clustering approach, the EM algorithm is implemented by treating the unknown component membership of the mixture population as missing data, so that the data is augmented with indicators of component membership. In the EM framework, starting from some current estimate for $\boldsymbol{\psi}$, say $\hat{\boldsymbol{\psi}}^{(k)}$, the E-step involves the calculation of the expectation of the complete-data log-likelihood, conditional on the observed data and the current estimate $\hat{\boldsymbol{\psi}}^{(k)}$. Since $\boldsymbol{y}$ and $\hat{\boldsymbol{\psi}}^{(k)}$ are treated as known, the complete-data log-likelihood is linear in the membership

variables so the conditional expectation depends only on the expectation of Z_{ij}. The **E-step** in the $(k+1)^{th}$ iteration involves the evaluation of

$$E(Z_{ij}|\boldsymbol{y}_j;\boldsymbol{\psi}^{(k)}) = \frac{\pi_i^{(k)} f_i(\boldsymbol{y}_j;\boldsymbol{\beta}^{i(k)},s_i)}{f(\boldsymbol{y}_j;\boldsymbol{\psi}_i^{(k)})} = \frac{\pi_i^{(k)} f_i(\boldsymbol{y}_j;\boldsymbol{\beta}^{i(k)},s_i)}{\sum_{i=1}^{g}\pi_i^{(k)} f(\boldsymbol{y}_j;\boldsymbol{\beta}^{i(k)},s_i)} = \hat{z}_{ij}^{(k)}.$$

This step is simply replacing the missing membership variables by the current values of their conditional expectations, i.e. the resulting estimate is the posterior probability that gene j belongs to cluster i. On the **M-step**, the value of $\boldsymbol{\psi}$ that maximizes the complete-data log-likelihood with each z_{ij} replaced by the corresponding posterior probability is evaluated, and the estimate of $\boldsymbol{\psi}$ is updated by

$$\hat{\boldsymbol{\psi}}^{(k+1)} = \arg\max_{\psi} E[\log L(\boldsymbol{\psi}|\boldsymbol{y};\hat{\boldsymbol{\psi}}^{(k)})].$$

The cluster proportions are given by

$$\hat{\pi}_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\hat{z}_{ij}^{(k)}.$$

There is no closed form solution to the evaluation of $\boldsymbol{\beta}$ and s in the M-step so the maximization requires numerical iteration, such as optimization procedures including Newton-type methods. Starting from the initial parameter value $\hat{\boldsymbol{\psi}}^{(0)}$, the E- and M-steps are repeated until convergence. After EM convergence is reached, a probabilistic clustering of the genes into g clusters are obtained through the posterior probabilities of component membership by assigning gene j to cluster k if $\hat{z}_{kj} = max(\hat{z}_{1j},...,\hat{z}_{gj})$.

There are some limitations to the application of EM algorithms. One major drawback is that the covariance matrix of the estimated parameters are not produced as an end-product of the algorithm. Another issue with the application of EM algorithm is that the speed of convergence can be very slow in some situations, for example, when the proportion of missing data is high. To speed up the estimation procedure, we propose a hybrid estimation algorithm for our model-based clustering approach. For a fixed number of components g, we use a combination of the EM algorithm and the quasi-Newton algorithm to obtain MLE's of parameters in our mixture model. Redner and Walker (1984) noted from their study that 95% of the change in log-likelihood from initial evaluation to the maximum value generally occurred within the first five EM iterations, thus we propose an estimation procedure which starts with running five (or ten) EM iterations to approach the near-neighbourhood of the ML estimates, and then switches to the quasi-Newton method for rapid convergence. These methods are referred to as EMQN5 and EMQN10 (with five and ten EM iterations before switching to quasi-Newton method respectively).

3 Application to real data

The goal of the *Drosophila* Transcriptome project (Graveley et al., 2011) is to examine developmental stages spanning the life cycle of *Drosophila melanogaster* (fruit

flies). The sample consisted of 542 genes and each of the 12 sampling time point corresponds to two hours. To account for the different gene lengths (to avoid gene-length bias), RPKM (reads per kilobase of exon model per million mapped reads) values were calculated and used as the gene expression measure for all genes. Upon obtaining the RPKM values for the genes, we applied the clusering algorithms using EM and hybrid-EM (EMQN5, EMQN10) to fit two- to six-component models to see the clustering effects and used BIC to select the optimal model.

The BIC values obtained from the different models showed that the four-component model can best represent the gene profiles over-time and the trajectories are shown in Figure 1. EM and EMQN10 both estimated very similar trajectories and cluster proportions, with three main groups of genes with decreasing expression patterns and one small group (7.6%) of genes having an almost bell-shaped expression pattern over time. The small cluster of genes showed an increasing expression level and then decreased back down to the initial starting level. On the other hand, EMQN5 estimated four groups at different cluster proportions with decreasing expression patterns. This might be due to the EMQN5 algorithm being trapped at a local maximum point during the ML estimation while EM and EMQN10 were able to find the global maximum point with more EM iterations performed than EMQN5.

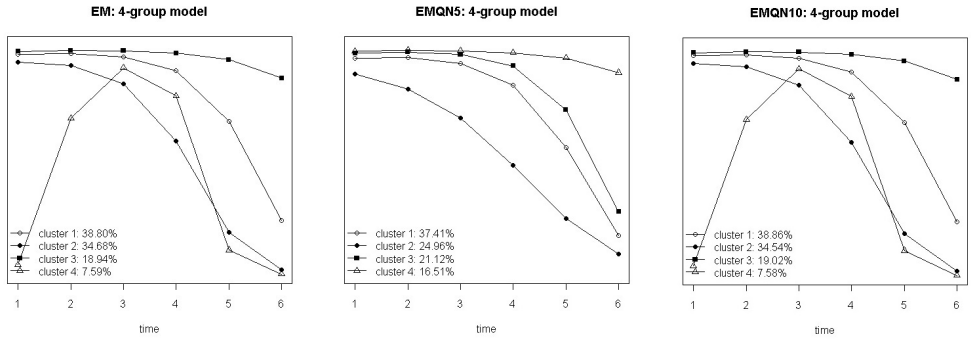

Figure 1: Real data analysis: four-component models.

4 Conclusion

Statistical models for analyzing RNA-seq data has recently become a popular research area in the literature of statistical genetics. To our knowledge, no model framework has been developed for cluster analysis of RNA-seq data focusing on the time-course experiment setting. We propose a clustering algorithm for discovering gene expression patterns in time-series RNA-seq data. The algorithm is based on negative binomial models in the time-course setting and can be applied to RNA-seq data, as well as other types of count data with over-dispersion. We propose an EM clustering method and two EM/quasi-Newton hybrid algorithms to improve on the speed of the EM clustering. We demonstrate that our proposed algorithms perform well on cluster analysis of time-course count data with over-dispersion. Applications to RNA-seq data illustrate that our model-based clustering approach

produces meaningful clustering results that can enhance researchers' understanding about gene expression patterns over time.

REFERENCES

Bullard, J., Purdom, E., Hansen, K. and Dudoit, S. (2010) "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, 11, 94.

Cooke, E., Savage, R., Kirk, P., Darkins, R. and Wild, D. (2011) "Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements," *BMC Bioinformatics* , 12, 399.

Grün, B., Scharl, T. and Leisch, F. (2012) "Modelling time course gene expression data with finite mixtures of linear additive models," *Bioinformatics*, 28, 222-228.

Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C. et al. (2011) "The developmental transcriptome of Drosophila melanogaster," *Nature*, 471, 473-479.

Marioni, J., Mason, C., Mane, S., Stephens, M. and Gilad, Y. (2008) "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, 18, 1509-1517.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, 320, 1344-1349.

Nagin, D. (1999) "Analyzing developmental trajectories: A semiparametric group-based approach," *Psychological Methods*, 4, 139-157.

Ng, S., McLachlan, G., Wang, K., Ben-Tovim Jones, L. and Ng, S. (2006) "A mixture model with random-effects components for clustering correlated gene-expression profiles," *Bioinformatics*, 22, 1745-1752.

Redner, R. and Walker, H. (1984) "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, 26, 195-239.

Robinson, M. and Smyth, G. (2007) "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, 23, 2881-2887.

Schliep, A., Schonhuth, A. and Steinhoff, C. (2003) "Using hidden Markov models to analyze gene expression time course data," *Bioinformatics*, 19, I283-289.

Yuan, A. and He, W. (2008) "Semiparametric clustering method for microarray data analysis," *Journal of Bioinformatics and Computational Biology*, 6, 261-282.