1 Introduction

Cluster analysis is the identification of natural groupings of observations that share certain characteristics. Classical, heuristic clustering methods, such as hierarchical agglomerative clustering and k-means clustering, are easy to understand and have been successfully implemented. However, despite considerable research in this area, there are few guidelines for solving basic practical questions that arise in cluster analysis, for instance, how many clusters to choose, which clustering method to use, and how to handle outliers and ties. The lack of knowledge on the statistical properties of the clusters also makes it difficult to measure the appropriateness of the resulting clustering (Fraley and Raftery, 2002).

Major advances in clustering methodology have been made through the introduction of statistical models, among which the most commonly proposed one is the finite mixture model (Johnson and Wichern, 12.5). In finite mixture models, each component probability distribution corresponds to a cluster (Fraley and Raftery, 2002). The problems of choosing the number of clusters and the method of clustering are reduced to the problem of choosing an appropriate statistical model.

Fraley and Raftery (2002) describes in great detail the procedures of performing model-based cluster analysis. The multivariate normal mixture model is used and different candidate models are acquired by varying constraints on its parameters. The EM algorithm is used to produce the classification of observations for each candidate model. The Bayesian Information Criteria (BIC) is used for model selection. Section 2 contains more details of this method.

Our goal is to cluster the genes based on their log fold changes over six time points, while accounting for the measurement error involved in estimating those log fold changes.

The log fold changes can be expressed as a linear combination of coefficient estimates from the negative binomial regression, and those estimates can be loosely considered as maximum likelihood estimates, which will allow us to use the inverse of (observed) information matrix as the measurement error.

We will use the model-based clustering methods developed by Fraley and Raftery (2002). The most important part of this project is to incorporate measurement error into their method.

In this paper, we extend the model-based clustering methods described in Fraley and Raftery (2002) so that measurement errors are accounted for in the cluster analysis. Specifically, we augmented the R package, `mclust`, developed by C. Fraley, A. Raftery and L. Scrucca, by replacing the EM algorithm part with our own version.

2 Model-Based Clustering

2.1 Multivariate Normal Mixture Model

Given data $\boldsymbol{y}$ consisting of an independent random sample $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}$, the likelihood of a mixture model is as follows:

$$L_{MIX}(\theta_1, ..., \theta_G; \tau_1, ..., \tau_G | \boldsymbol{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\boldsymbol{y}_i | \theta_k) \tag{1}$$

where f_k and θ_k are the density and parameters of the kth component in the mixture and τ_k is the probability that an observation belongs to the kth component, subject to $\sum_{k=1}^{G} \tau_k = 1$.

The most commonly used candidate for f_k is the multivariate normal distribution, whose PDF is denoted by ϕ_k and parametrized by mean μ_k and covariance matrices Σ_k as follows:

$$\phi_k\left(\boldsymbol{y}_i | \mu_k, \Sigma_k\right) = \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \mu_k)^T \Sigma_k^{-1} (\boldsymbol{y}_i - \mu_k)\right\}}{\sqrt{\det(2\pi\Sigma_k)}} \tag{2}$$

Data generated by this model will appear as ellipsoidal clusters, each of which corresponds to a component in (1). The center of each cluster is μ_k, while the shape, orientation and volume are characterized by Σ_k.

Banfield and Raftery (1993) used eigendecomposition to rewrite Σ_k in the following form:

$$\Sigma_k = \lambda_k D_k A_k D_k^T \tag{3}$$

where D_k is the orthogonal matrix of eigenvectors, A_k is the diagonal matrix of scaled eigenvalues and λ_k is the scale factor.

This parametrization of the covariance matrix allows us to model the geometric properties of each cluster. In fact, D_k governs the orientation of the kth cluster, A_k its shape, and λ_k its volume. Treating them as three independent sets of parameters, and allowing them to vary or constraining them to remain the same across clusters, we obtain parsimonious and easily interpreted models which are appropriate to describe various clustering situations. Celeux and Govaert (1995) describes in detail a total number of fourteen different models generated from the above parametrization.

2.2 The Expectation-Maximization Algorithm for Mixture Models

Let $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ be the complete data, where $\boldsymbol{y}_i$ is observed data and $\boldsymbol{z}_i = (z_{i1}, ..., z_{iG})$ is the unobserved portion, with

$$z_{ik} = \begin{cases} 1 \text{ if } x_i \text{ belongs to group } k \\ 0 \text{ otherwise} \end{cases}$$

The complete data log likelihood is:

$$l(\theta_k, \tau_k, z_{ik}|x) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik} \log[\tau_k f_k(y_i|\theta_k)] \tag{4}$$

The E step estimates the unobserved portion of the data, z_i by the following:

$$\hat{z}_{ik} = \frac{\hat{\tau}_k f_k(y_i|\hat{\theta}_k)}{\sum_{j=1}^{G} \hat{\tau}_j f_j(y_i|\hat{\theta}_j)} \tag{5}$$

while the M step maximizes the complete data log likelihood in terms of τ_k and θ_k, with z_{ik} fixed at the values obtained from the E step. To initialize the EM algorithm, we will first perform an initial clustering procedure using hierarchical clustering to acquire an initial classification, i.e. initial values for z_{ik}'s. And then iterate until convergence.

The ultimate goal of the EM algorithm is to obtain a classification rule of observations. Let z_{ik}^* denote the value of z_{ik} upon convergence, and the classification rule would be to assign observation i to group j such that $z_{ij}^* = \max_k(z_{ik}^*)$.

For multivariate normal mixture models, the M step involves estimating μ_k, τ_k and Σ_k for $k = 1, ..., G$. It turns out that the estimators for the first two parameters have nice, closed-form solutions:

$$\hat{\tau}_k = \frac{n_k}{n}; \quad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ik} y_i}{n_k}; \quad n_k = \sum_{i=1}^{n} \hat{z}_{ik} \tag{6}$$

Computation of the covariance estimate depends on its parametrization. Celeux and Govaert (1995) gives a detailed description of the estimation method.

2.3 Model Selection

For different values of G and different constraints on Σ_k, the problems of choosing an appropriate clustering method and choosing the number of clusters have been reduced to the problem of selecting an appropriate model. Fraley and Raftery (2002) suggests using the BIC as the information criteria, defined as follows:

Suppose that several models, $M_1, ..., M_K$ are considered, given data D, then the BIC for model M_k is given by

$$BIC_k = 2 \log p(D|\hat{\theta}_k, M_k) - \nu_k \log(n) \tag{7}$$

where ν_k is the number of independent parameters to be estimated in model M_k.

Fraley and Raftery (2002) points out that finite mixture models do not satisfy the regularity conditions required to prove (7), but results suggest its appropriateness and good performance in the model-based clustering context.

2.4 Cluster Analysis

Fraley and Raftery (2002) summarizes the model-based clustering method as the following four-step algorithm:

1. Determine a maximum number of clusters, M, and a set of mixture models to consider.
2. Perform hierarchical agglomeration to approximately maximize the classification likelihood for each model, and obtain the corresponding classifications up to M groups.
3. Apply the EM algorithm for each model and each number of clusters $2, ..., M$, starting with the classification from hierarchical agglomeration.
4. Compute BIC for the one-cluster case for each model and for the mixture model with the optimal parameters from EM for $2, ..., M$ clusters.

3 Incorporating Measurement Error into EM Algorithm

3.1 Negative Binomial Regression

3.2 Augmented Model

Here we state the model that includes measurement error for each observation:

For each observation $\boldsymbol{y}_i$, assume that there exists a latent variable $\boldsymbol{w}_i$ such that:

$$\begin{aligned}\boldsymbol{y}_i|\boldsymbol{w}_i &\sim N(\boldsymbol{w}_i, \Sigma_i)\\ \boldsymbol{z}_i|\theta_k &\sim N(\mu_k, \Sigma_k)\end{aligned}$$

Then we have:

$$\boldsymbol{y}_i|\theta_k \sim N(\mu_k, \Sigma_i + \Sigma_k)$$

For the data set under consideration, $\boldsymbol{y}_i$ stands for the log fold changes of expression level of gene i over six fixed time points, and is in fact a linear combination of estimated regression coefficients from a negative binomial (NB) regression model.

Let $\hat{\boldsymbol{\beta}}$ denote the estimated coefficients, and $\boldsymbol{a}_i\hat{\boldsymbol{\beta}}$ be the linear combination that represents the log fold change for gene i. Suppose $\hat{\boldsymbol{\beta}}$ is estimated via maximum likelihood, then by the asymptotic properties of MLE, the estimated variance (or measurement error) of $\boldsymbol{a}_i\hat{\boldsymbol{\beta}}$ can be approximated by:

$$\widehat{\mathrm{Var}}(\boldsymbol{a}_i\hat{\boldsymbol{\beta}}) = \boldsymbol{a}_i^T \left[\mathcal{I}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \boldsymbol{a}_i$$

where $\mathcal{I}\left(\hat{\boldsymbol{\beta}}\right)$ is the observed information.