Model-based Clustering with Measurement Errors

Wanli Zhang* **and Yanming Di****

Department of Statistics, Oregon State University, Corvallis OR, United States

**email*: zhangwa@stat.oregonstate.edu

***email*: diy@stat.oregonstate.edu

Model-based clustering has become a widely used clustering method. It imposes a finite mixture model on the data and provides guidance on practical questions such as the choice of the number of clusters. Implementation of this method has been realized by the MCLUST software. In most cases, observations to be clustered are assumed to have been accurately measured, free of errors. However, there are situations where this assumption may not be feasible, and since MCLUST does not explicitly consider measurement errors, there is an urgent need to develop new methods that account for said errors. In this paper, we first review the model and algorithm MCLUST adopts. Then, we propose a new model and use it to develop a new model-based clustering algorithm, called MCLUST-ME, that properly accounts for measurement errors. Next, through simulation studies, we evaluate the performance of the new algorithm and examine the differences between results produced by MCLUST and MCLUST-ME. Finally, we mention computational issues we encountered during implementation of our algorithm, as well as discuss possibility of larger scale implementations in the future.

1. Introduction

Cluster analysis identifies natural groupings of observations that share certain characteristics. Considerable research has been done in this area over the years, and yet there are few guidelines for answering practical questions such as how many clusters to choose and which clustering method to use (Fraley and Raftery, 2002). To address these problems, statistical models have been introduced into clustering methodology, among which the most commonly proposed one is the finite mixture model. Consequently, model-based clustering methods have turned the problems above into that of model selection. Fraley and Raftery (2002) reviewed their model-based approach to clustering, which had been implemented in the R package, MCLUST (Fraley and Raftery, 1999).

In most cases, observations to be clustered are assumed to have been precisely measured, whereas there are situations where this assumption is clearly not feasible. For example, when analyzing RNA-Seq data, researchers wish to group genes consistent in their expression profiles over time, measured by log fold changes at each time point. The log fold changes themselves, being functions of regression coefficient estimates, are inaccurately measured due to errors incurred in the estimation process. See Si, Liu, Li and Brutnell (2013) for such an example. A cluster analysis not accounting for these errors may fail to capture the most appropriate grouping of genes, rendering the result less meaningful or even misleading, prompting us to develop new clustering algorithms that take said errors into consideration.

In this paper, we introduce a new model-based clustering algorithm, which we name MCLUST-ME, that properly accounts for measurement errors in observations. A large part of this algorithm is inspired by MCLUST, which we will review in Section 2. For this study, we implemented MCLUST-ME for one particular case of covariance matrix structure, that is, where all component covariances are allowed to have distinct orientations, shapes and volumes (first introduced in Banfield and Raftery (1993), and called "VVV" in MCLUST).

One of the most used criteria to evaluate clustering performance is the Rand index (Rand, 1971), comparing two hard partitions of the same sample. Model-based clustering produces a fuzzy partition, which can be converted into a hard one by assigning each observation into the most probable group. Campello (2007) pointed out that this conversion results in loss of information, which causes the Rand index inappropriate for fuzzy clustering assessment. An extension of Rand index, called Fuzzy Rand index, was developed in the same paper, which we will use to assess simulation results in Section 5.

The organization of this article is as follows. Section 2 reviews the model and EM algorithm adopted by MCLUST. Section 3 introduces the new clustering scheme, MCLUST-

ME. Section 4 investigates decision boundaries of the two methods. Section 5 gives the settings and results of two simulations. Finally, conclusions and perspectives for future work are addressed in Section 6.

2. Review of model and EM algorithm of MCLUST

2.1 Model

Let $\boldsymbol{y}$ denote an i.i.d. d-dimensional sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ from a G-component mixture distribution with likelihood:

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

f_k, is often chosen to be multivariate normal density ϕ_k, with mean μ_k and covariance Σ_k and pdf:

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

2.2 EM algorithm

Let $\boldsymbol{x}_i = (\boldsymbol{y}_i, \boldsymbol{z}_i)$ be the complete data, where $\boldsymbol{y}_i$ is observable and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})$ is not, and $z_{ik} = \mathbf{1}(\boldsymbol{x}_i \text{ belongs to group } k)$. Suppose $\{\boldsymbol{z}_i\}$ are i.i.d. $\text{Mult}_G(1, \tau_1, ..., \tau_G)$, and that the density of $\boldsymbol{y}_i$ given $\boldsymbol{z}_i$ is $\prod_{k=1}^{G} \phi_k(\boldsymbol{y}_i | \mu_k, \Sigma_k)^{z_{ik}}$. The EM algorithm (Dempster, Laird and Rubin, 1977) proceeds as follows:

- *M step*: Given current estimate of z_{ik}, maximize the complete-data log-likelihood

$$l(\mu_k, \Sigma_k, \tau_k, z_{ik} | \boldsymbol{x}) = \sum_{i=1}^{N} \sum_{k=1}^{G} z_{ik} \log[\tau_k \phi_k(\boldsymbol{y}_i | \mu_k, \Sigma_k)] \tag{3}$$

  with respect to $(\tau_k, \mu_k, \Sigma_k)$.

- *E step*: Given MLEs $(\hat{\tau}_k, \hat{\mu}_k, \hat{\Sigma}_k)$ from *M step*, compute

$$\hat{z}_{ik} = \frac{\hat{\tau}_k \phi_k(\boldsymbol{y}_i|\hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^{G} \hat{\tau}_j \phi_j(\boldsymbol{y}_i|\hat{\mu}_j, \hat{\Sigma}_j)} \tag{4}$$

The two steps alternate until convergence. A fuzzy membership matrix is produced and if needed, converted into a hard partitioning by assigning each observation into the most probable group.

Celeux and Govaert (1995) have shown that $\hat{\mu}_k$ and $\hat{\tau}_k$ always have closed-form solutions, while the form of $\hat{\Sigma}_k$ depends on its parameterization. When the covariance structure is "VVV", a closed-form solution for $\hat{\Sigma}_k$ exists.

3. Model and EM algorithm of MCLUST-ME

3.1 Model

For each observed value $\boldsymbol{y}_i$, assume that there exists a latent variable $\boldsymbol{w}_i$, representing the "true" value of $\boldsymbol{y}_i$, such that:

$$\boldsymbol{y}_i|\boldsymbol{w}_i \sim N_d(\boldsymbol{w}_i, M_i)$$
$$\boldsymbol{w}_i|\mu_k, \Sigma_k \sim N_d(\mu_k, \Sigma_k)$$

where the measurement error covariance M_i is known; μ_k and Σ_k are the unknown mean and covariance of the kth component in the mixture.

The marginal distribution of $\boldsymbol{y}_i$ is:

$$\boldsymbol{y}_i|\mu_k, \Sigma_k \sim N_d(\mu_k, \Sigma_k + M_i) \tag{5}$$

with density function:

$$\phi_k^*\left(\boldsymbol{y}_i|\mu_k,\Sigma_k\right)=\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i-\mu_k)^T\left(\Sigma_k+M_i\right)^{-1}(\boldsymbol{y}_i-\mu_k)\right\}}{\sqrt{\det[2\pi(\Sigma_k+M_i)]}}=\phi_k(\boldsymbol{y}_i|\mu_k,\Sigma_k+M_i) \quad (6)$$

Substituting f_k in (1) by (6), given an i.i.d. sample $\boldsymbol{y}=(\boldsymbol{y}_1,...,\boldsymbol{y}_N)$, the (observed) likelihood for a modified G-component mixture model is:

$$L_{MIX}^*(\tau_k,\mu_k,\Sigma_k|\boldsymbol{y})=\prod_{i=1}^{N}\sum_{k=1}^{G}\tau_k\phi_k^*(\boldsymbol{y}_i|\mu_k,\Sigma_k)$$

3.2 EM algorithm

Based on the new model, we modify MCLUST's EM algorithm as follows:

- *M step*: Given current estimate of z_{ik}, maximize the complete-data log-likelihood

$$l^*(\mu_k,\Sigma_k,\tau_k,z_{ik}|\boldsymbol{y})=\sum_{i=1}^{N}\sum_{k=1}^{G}z_{ik}\log[\tau_k\phi_k^*(\boldsymbol{y}_i|\mu_k,\Sigma_k)] \quad (7)$$

  with respect to (τ_k,μ_k,Σ_k).

- *E step*: Given estimates $(\hat{\tau}_k,\hat{\mu}_k,\hat{\Sigma}_k)$ from last *M step*, compute

$$\hat{z}_{ik}=\frac{\hat{\tau}_k\phi_k^*(\boldsymbol{y}_i|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{j=1}^{G}\hat{\tau}_j\phi_k^*(\boldsymbol{y}_i|\hat{\mu}_j,\hat{\Sigma}_j)} \quad (8)$$

The two steps alternate until convergence. Similar to MCLUST, upon convergence, a fuzzy membership matrix is produced and if needed, can be converted into a hard partition.

It can be shown that a closed-form solution always exists for $\hat{\tau}_k$, while in general, not for $\hat{\mu}_k$ and $\hat{\Sigma}_k$. Nevertheless, when $M_i\equiv M$ for some constant covariance matrix M , MCLUST-ME essentially becomes MCLUST, and both $\hat{\mu}_k$ and $\hat{\Sigma}_k$ have closed-form

solutions if component covariances have "VVV" parameterization.

4. Decision boundaries for hard partitioning

In this section, we consider decision (or classification) boundaries for partitioning a sample into $G = 2$ clusters. Let $(\tilde{\tau}_k, \tilde{\mu}_k, \tilde{\Sigma}_k)$ and $(\tilde{\tau}_k^*, \tilde{\mu}_k^*, \tilde{\Sigma}_k^*)$ denote maximum likelihood estimates for $(\tau_k, \mu_k, \Sigma_k)$ upon convergence, produced by MCLUST and MCLUST-ME, respectively. Suppose the sample contains N observations $\{\boldsymbol{y}_i\}_{i=1}^N$.

The MCLUST decision boundary is given by:

$$B = \{\boldsymbol{t} : \tilde{\tau}_1 \phi_1(\boldsymbol{t}|\tilde{\mu}_1, \tilde{\Sigma}_1) - \tilde{\tau}_2 \phi_2(\boldsymbol{t}|\tilde{\mu}_2, \tilde{\Sigma}_2) = 0\}$$

where $\phi_k, k = 1, 2$ is defined in (2). Notice that a common boundary exists for all observations.

Formulation of MCLUST-ME decision boundary is somewhat different. The boundary for $\boldsymbol{y}_i$ is given by:

$$B^*(M_i) = \{\boldsymbol{t} : \tilde{\tau}_1^* \phi_1(\boldsymbol{t}|\tilde{\mu}_1^*, \tilde{\Sigma}_1^* + M_i) - \tilde{\tau}_2 \phi_2(\boldsymbol{t}|\tilde{\mu}_2^*, \tilde{\Sigma}_2^* + M_i) = 0\}$$

where $\phi_k, k = 1, 2$ is same as above, and M_i is measurement error of $\boldsymbol{y}_i$. Here, each observation $\boldsymbol{y}_i$ will have its own decision boundary $B^*(M_i)$. When two or more observations have measurement errors that coincide, they will share a common boundary. Therefore, if $M_i \equiv M$ for all $i = 1, ..., N$ for some constant matrix M, all observations share the same boundary, and MCLUST-ME essentially becomes MCLUST in this situation.

5. Simulation and results

5.1 Overview

Assume that the component covariances have structure "VVV". For both simulations, we fix the following parameters: number of clusters [$G = 2$], true centers [$\mu_1 = (0, 0)$, $\mu_2 = (8, 0)$], true covariances [$\Sigma_1 = 64I_2$, $\Sigma_2 = 16I_2$, where I_2=2-dim id matrix], dimension of data [$d = 2$], total sample size [$N = 200$], mixing proportion [$\tau = 0.5$].

5.2 Simulation 1: Large errors with varying error proportions

5.2.1 Setting We generate a random sample from a 2-component normal mixture model, with 2 unique values of measurement error, $M = 36I_2$ and 0. Let p denote the expected proportion of observations with measurement errors, with 5 possible values: 0.1, 0.3, 0.5, 0.7 and 0.9. For each choice of p, flip a coin with head probability p for N times to obtain indicators $\{h_i = \mathbf{1}(\text{head})\}_{i=1}^{N}$. Then randomly select 100 random seeds from $\{1, ..., 200\}$, and for each seed, generate each $\boldsymbol{y}_i, i = 1, ..., N$ according to either $N_d(\mu_1, \Sigma_1 + h_i M)$ or $N_d(\mu_2, \Sigma_2 + h_i M)$, based on the flip of a fair coin. For each sample, run MCLUST-ME and MCLUST. Initiate both algorithms by the true memberships of the observations.

5.2.2 Results Firstly, we plotted hard partitioning results from both methods, as well as the theoretical decision boundaries acquired in Section 4. Figure 1 shows groupings of the sample generated with $p = 0.5$ and with random seed 7. As discussed in Section 4, since there are 2 unique values of measurement errors ($36I_2$ and 0), MCLUST-ME has 2 distinct boundaries. On this particular sample, we make some interesting observations: (i) the two MCLUST-ME boundaries are relatively far apart; (ii) MCLUST boundary falls between the two MCLUST-ME boundaries; and (iii) clusters produced by MCLUST-ME cannot be separated by a single convex curve.

Then, we evaluate performances of MCLUST and MCLUST-ME separately, using both Rand index and its fuzzy version. Figure 2 shows that for both methods, no matter which criterion is used, clustering accuracy, as measured by Rand/Fuzzy Rand indices, tends to

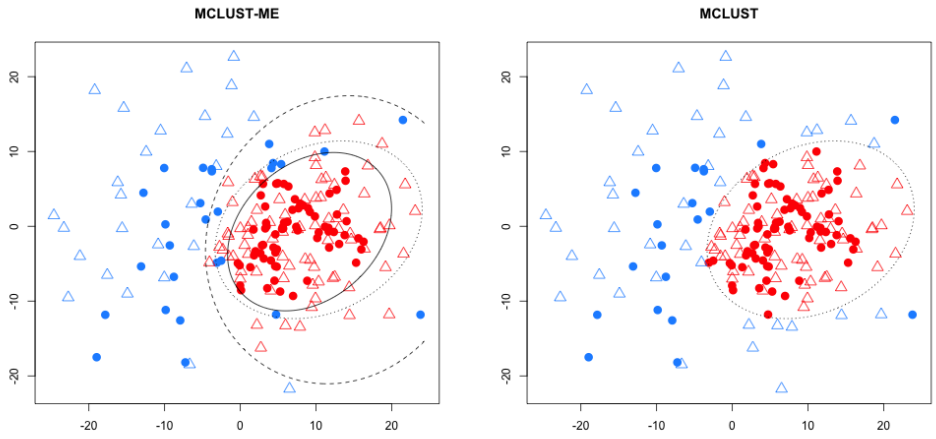

Figure 1: Sample generated with random seed=7 and $p = 0.5$. **Both plots**: solid dots represent observations with no measurement errors; empty triangles represent those generated with error M. Different colors represent different clusters. **Left**: clustering result produced by MCLUST-ME. Solid line represents classification boundary for error-free observations; dashed line represents boundary for those with errors M; dotted line represents boundary produced by MCLUST. **Right**: clustering result produced by MCLUST. Dotted line is the same as in the left plot.

decrease as error proportion p increases. This is intuitively reasonable, because observations with measurement errors are more easily misclassified due to their high variability, and a larger proportion of such observations means lower overall accuracy.

Finally, we compare the performances of MCLUST and MCLUST-ME by examining pairwise differences in Rand/Fuzzy Rand indices. Figure 3 shows that on average, MCLUST-ME has slight advantage in terms of accuracy, and it appears that this advantage is most significant when $p = 0.5$, and becomes smaller either as p increases or decreases. As p gets closer to 1 or 0, measurement errors will tend to become constant (all equal to $36I_2$ or 0) for all observations, meaning that MCLUST-ME will behave more and more like MCLUST, hence diminishing MCLUST-ME's advantage in accuracy.

5.3 Simulation 2: Small errors with varying error proportions

A second simulation was run, where all simulation parameters are the same as Simulation 1, except that M is set to be $9I_2$. We were able to make the following observations: Individual Rand/Fuzzy Rand indices show the same pattern as in Simulation 1, i.e. decreasing as p increases. Pairwise differences of Rand/Fuzzy Rand indices also demonstrate

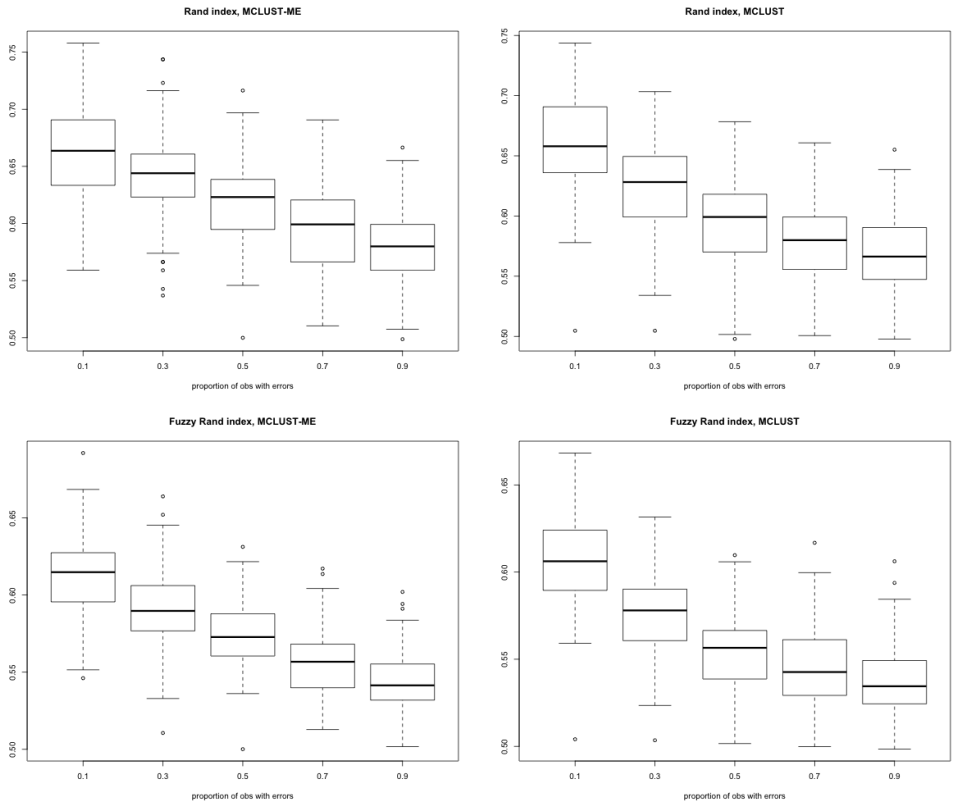

Figure 2: Rand/Fuzzy Rand indices for MCLUST-ME (first column) and for MCLUST (second column), for 5 different proportions of erroneous observations

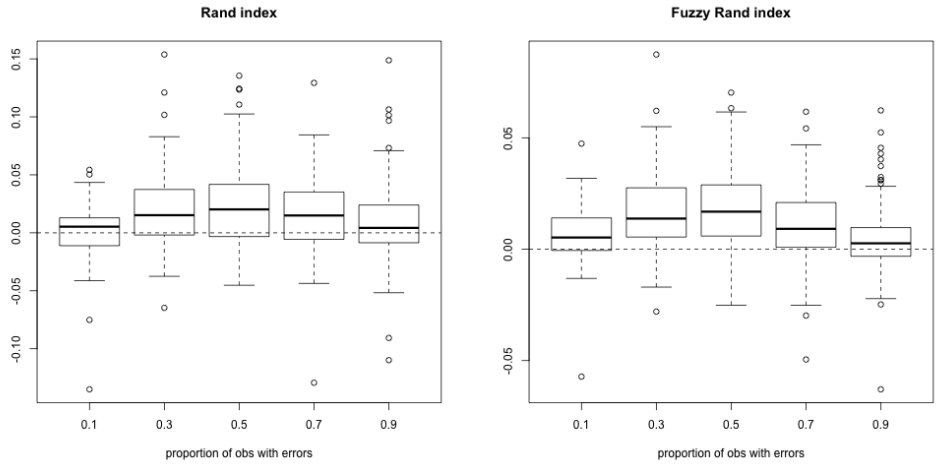

Figure 3: Pairwise difference in Rand/Fuzzy Rand indices (MCLUST-ME−MCLUST), for 5 different proportions of erroneous observations

the same pattern, except that the differences are more trivial than when $M = 36I_2$. This matches our intuition, because the smaller the measurement errors are, the more similar MCLUST and MCLUST-ME become to each other, and the more trivial their differences in clustering accuracy.

6. Discussion and future work

Using simulation, we have shown that whether or not accounting for measurement errors leads to different clustering results. Here we used Rand index to compare performances of MCLUST and MCLUST-ME. Although we have made some interesting observations, it's still early to conclude which one is superior to the other, as there are criteria other than clustering accuracy that define the performance of such methods. For example, the precision of parameter estimation may serve as such a criterion: if two clustering results have the same accuracy, the one with better parameter estimates will always be preferable. In future simulation studies, we shall consider using criteria such as this.

We assumed that the component covariance parameterization is always "VVV", and that the number of clusters had been determined prior to clustering. One of the next steps in our research is to incorporate model selection element into our method, just like MCLUST did, and investigate the potential effect of measurement errors on the choice of models.

In Section 5, we examined effect of changing proportion and magnitude of errors while fixing all other parameters. In the future, it would also be a good idea to investigate the effect of changing other parameters, such as distance between centers, mixing proportion, initial membership, etc. Furthermore, since the error covariances we used are all multiples of the identity matrix, it'll be tempting to try varying the spread at each dimension of covariance and see how clustering results will be affected.

References

[1] Banfiled, J. D., and Raftery, A. E. (1993). "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, **49**, 803-821

[2] Campello, R.J.G.B. (2007). "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment," *Patter Recognition Letters*, **28**, 833-841

[3] Celeux, G. and Govaert, G. (1995). "Gaussian parsimonious clustering models," *Patter Recognition*, **28**, 781-793

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). "Maximum likelihood for incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, **39**, 1-38

[5] Fraley, C. and Raftery, A. E. (2002). "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, **97**, 611-631

[6] Fraley, C. and Raftery, A. E. (1999). "MCLUST: Software for model-based cluster analysis," *Journal of Classification*, **16**, 297-306

[7] Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, **66**, 846-850

[8] Si, Y., Liu, P., Li, P. and Brutnell, T. P. (2014). "Model-based clustering for RNA-seq data," *Bioinformatics*, **30**, 197-205