PROD. TYPE: COM
PP: 1-19 (col.fig.: nil)

JSPI3256 MODC+(SC)

ED: Bhagya
PAGN: Veena – SCAN: nil

ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

journal of statistical planning and inference

www.elsevier.com/locate/jspi

A comparative study of the *K*-means algorithm and the normal mixture model for clustering: Univariate case

Dingxi Qiu, Ajit C. Tamhane*

Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208, USA

Abstract

This paper gives a comparative study of the *K*-means algorithm and the mixture model (MM) method for clustering normal data. The EM algorithm is used to compute the maximum likelihood estimators (MLEs) of the parameters of the MM model. These parameters include mixing proportions, which may be thought of as the prior probabilities of different clusters; the maximum posterior (Bayes) rule is used for clustering. Hence, asymptotically the MM method approaches the Bayes rule for known parameters, which is optimal in terms of minimizing the expected misclassification rate (EMCR).

The paper gives a thorough analytic comparison of the two methods for the univariate case under both homoscedasticity and heteroscedasticity. Simulation results are given to compare the two methods for a range of sample sizes. The comparison, which is limited to two clusters, shows that the MM method has substantially lower EMCR particularly when the mixing proportions are unbalanced. The two methods have asymptotically the same EMCR under homoscedasticity (resp., heteroscedasticity) when the mixing proportions of the two clusters are equal (resp., ~~not too~~ unequal), but for small samples the MM method sometimes performs slightly worse because of the errors in estimating unknown parameters.

MSC: 62H30; 62F10

Keywords: Bayes rule; Clustering; Data mining; EM algorithm; *K*-means algorithm; Misclassification rate; Mixture model; Prior and posterior probabilities

1. Introduction

The *K*-means algorithm (MacQueen, 1967) is one of the most popular methods for clustering multivariate quantitative data. This algorithm is non-parametric in nature as it does not assume any probability model for the data. Given a fixed number of clusters, it determines an assignment of the data vectors (observations) to the clusters so as to minimize the total of the squared distances between the observations assigned to the same cluster and summed over all clusters. See Everitt (1993) for a review of clustering methods.

The mixture model (MM) method provides a parametric approach to the clustering problem. The EM algorithm (Dempster et al., 1977) is a natural method for obtaining the maximum likelihood estimators (MLEs) of the unknown parameters of the MM. The parameters include the mixing proportions or the prior probabilities of the clusters since the

* Corresponding author. Tel.: +1 8474913577; fax: +1 8474918005.
E-mail address: ajit@iems.northwestern.edu (A.C. Tamhane).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.045

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the *K*-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

true cluster memberships of the observations are unobserved. Clustering is done by applying the maximum posterior (Bayes) rule.

The K-means algorithm makes "hard" (deterministic) assignments of the observations to the clusters, i.e., each observation is assigned to exactly one cluster. On the other hand, the MM method computes posterior probabilities (called *responsibilities*) of belonging to different clusters for individual observations. Hastie et al. (2002, p. 463) note that the MM method is a "soft" version of the K-means algorithm in that if the data from each cluster is assumed to be multivariate normal (MVN) with the mean vector depending on the cluster and a common covariance matrix $\sigma^2\boldsymbol{I}$, then as $\sigma^2 \to 0$, the MM method based on the EM algorithm converges to the K-means algorithm. Thus, as in the K-means algorithm, asymptotically the MM method assigns each observation to that cluster whose estimated mean is closest to the observation.

Although there is asymptotic convergence of the MM method and the K-means algorithm, it is under very restrictive conditions of homoscedasticity not only among the clusters, but also among the measured variables. More crucially, it assumes independence among the variables. These assumptions underlie the K-means algorithm, which ignores correlations and heteroscedasticity among the variables by using the simple Euclidean distance measure. Therefore, it is of interest to compare the performances of the two methods under the practical conditions of small samples, correlated responses and heteroscedasticity. In this paper we initiate this study by focusing on the univariate case for $K = 2$ under homoscedasticity and heteroscedasticity. The MVN case, which allows the study of how correlations between measured variables affect the performance of the competing algorithms, will be considered in a future paper. Surprisingly, even the univariate case has not been studied in this context to the best of our knowledge.

The outline of the paper is as follows. In Section 2 we formulate the problem and define the notation. In Section 3 we review the K-means algorithm and the MM method with the associated EM algorithm. The discussions in both sections are framed in the general setting of MVN data, but in the remainder of the paper we focus exclusively on univariate normal data. In Section 4 we give analytical results for comparing the two methods in the homoscedastic case. In Section 5 we extend these results to the heteroscedastic case. In Section 6 we present simulation results on misclassification rates of the two methods. Finally, Section 7 gives a discussion and conclusions.

2. Problem formulation and notation

Suppose that there are N subjects on each of whom m variables are measured resulting in observations $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})'$ $(1 \leqslant i \leqslant N)$. The goal of clustering is to group these N subjects into $K < N$ clusters, C_k $(1 \leqslant k \leqslant K)$, so that similar subjects are grouped into the same cluster and dissimilar subjects are grouped into different clusters. We will assume that K is the true known number of clusters and is fixed. (In practice, of course, K is not known. The problem of determining the optimal K will not be addressed here.) Let N_k denote the true number of subjects belonging to cluster C_k where $\sum_{k=1}^{K} N_k = N$. A clustering rule (denoted by R) is a many-to-one mapping, $R(\boldsymbol{x}_i) = C_k$ $(1 \leqslant i \leqslant N, 1 \leqslant k \leqslant K)$.

We will assume that the observations $\boldsymbol{x}_i$ are mutually independent and the observations from different clusters have MVN distributions with different mean vectors; the covariance matrices may be equal (homoscedastic) or unequal (heteroscedastic). Specifically, if subject i belongs to cluster C_k (denoted by $i \in C_k$) then

$$\boldsymbol{X}_i | i \in C_k \sim \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.1}$$

where $\boldsymbol{X}_i$ denotes the random variable (r.v.) corresponding to the observation $\boldsymbol{x}_i$, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix of the MVN distribution for cluster C_k.

3. Review of two clustering methods

3.1. K-means algorithm

The K-means algorithm uses the Euclidean squared distance measure:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 = \sum_{j=1}^{m} (x_{ij} - x_{i'j})^2,$$

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

which implies homoscedastic independent measurements. The algorithm aims to minimize the within-cluster scatter given by

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{R(\boldsymbol{x}_i)=C_k}\sum_{R(\boldsymbol{x}_{i'})=C_k}\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|^2=\sum_{k=1}^{K}\sum_{R(\boldsymbol{x}_i)=C_k}\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}_k\|^2,$$

where $\bar{\boldsymbol{x}}_k=(\bar{x}_{k1},\bar{x}_{k2},\ldots,\bar{x}_{km})'$ is the vector of sample means of the observations assigned to cluster C_k. Since $\bar{\boldsymbol{x}}_k$ minimizes $\sum_{i\in C_k}\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2$ with respect to $\boldsymbol{\mu}_k$, the K-means algorithm actually aims to minimize the overall objective function,

$$\sum_{k=1}^{K}\sum_{R(\boldsymbol{x}_i)=C_k}\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2, \tag{3.1}$$

with respect to the classification rule R and the true cluster means $\boldsymbol{\mu}_k$ $(1\leqslant k\leqslant K)$. Beginning with an initial assignment of observations to the clusters and the corresponding sample cluster means $\bar{\boldsymbol{x}}_k$, the K-means algorithm iterates through the following two steps until convergence.

Step 1: Reassign each observation to the cluster whose mean $\bar{\boldsymbol{x}}_k$ is closest to that observation. Thus,

$$R(\boldsymbol{x}_i)=C_k\iff k=\underset{1\leqslant\ell\leqslant K}{\operatorname{argmin}}\ \|\boldsymbol{x}_i-\bar{\boldsymbol{x}}_\ell\|^2.$$

Step 2: Calculate the new cluster means $\bar{\boldsymbol{x}}_k$.

Note that this algorithm does not guarantee the global minimum of the objective function (3.1). It also does not take into account any prior knowledge about the mixing proportions. Thus, the K-means algorithm is a purely computational algorithm with no probabilistic basis.

3.2. MM method

Consider the following MM: denote the MVN density function of $\boldsymbol{X}_i$ conditional on $i\in C_k$ by

$$f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)=(2\pi)^{-m/2}|\boldsymbol{\Sigma}_k|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\},$$

where $\boldsymbol{\theta}_k=(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ (note that we are not using $\boldsymbol{\theta}_k$ as a vector). Let $\eta_k=\Pr(i\in C_k)$ be the prior probability (mixing proportion) that a subject i belongs to cluster C_k with $\sum_{k=1}^{K}\eta_k=1$. Then the unconditional distribution of $\boldsymbol{X}_i$ is

$$f(\boldsymbol{x}_i;\boldsymbol{\theta},\boldsymbol{\eta})=\sum_{k=1}^{K}\eta_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k),$$

where $\boldsymbol{\theta}$ represents the collection of $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_K$ and $\boldsymbol{\eta}=(\eta_1,\eta_2,\ldots,\eta_K)'$.

The log-likelihood function for the MM is given by

$$\ln L(\boldsymbol{\theta},\boldsymbol{\eta})=\sum_{i=1}^{N}\ln\left[\sum_{k=1}^{K}\eta_k f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k)\right]. \tag{3.2}$$

The data are "incomplete" since the cluster memberships of the subjects are unobserved. Therefore, (3.2) is referred to as the *incomplete log-likelihood function*. The "complete" data vector is denoted by $(\boldsymbol{x}_i,\boldsymbol{z}_i)$ (with $(\boldsymbol{X}_i,\boldsymbol{Z}_i)$ as the corresponding random vector), where $\boldsymbol{z}_i=(z_{i1},z_{i2},\ldots,z_{iK})$ is a binary vector with $z_{ik}=1$ and $z_{i\ell}=0$ for $\ell\neq k$ if $i\in C_k$. If $\boldsymbol{z}_i$ is observed then the *complete log-likelihood function* is given by

$$\ln L(\boldsymbol{\theta})=\sum_{i=1}^{N}\sum_{k=1}^{K}z_{ik}\ln f_k(\boldsymbol{x}_i;\boldsymbol{\theta}_k). \tag{3.3}$$

The EM algorithm is used to maximize the incomplete log-likelihood function (3.2) through maximization of the complete log-likelihood function (3.3). Since the z_{ik} are unobserved, they are replaced by the current estimates of their

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

expected values, which are the *responsibilities*. Upon initialization, the EM algorithm iterates through the following two steps until convergence.

Step 1 (*Expectation*): Given the current estimates $\widehat{\boldsymbol{\theta}}_k = (\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$ $(1 \leqslant k \leqslant K)$ and $\widehat{\boldsymbol{\eta}}$, calculate the expected values of the sufficient statistics for the unobserved r.v.s Z_{ik}, which are the estimates of the posterior probabilities, $\Pr(i \in C_k | \boldsymbol{x}_i)$, by applying the Bayes rule:

$$\widehat{z}_{ik} = E[Z_{ik} | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}] = \Pr[Z_{ik} = 1 | \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}] = \frac{\widehat{\eta}_k f_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_k)}{\sum_{\ell=1}^{K} \widehat{\eta}_\ell f_\ell(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_\ell)}. \tag{3.4}$$

Step 2 (*Maximization*): Find the new estimates $\widehat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k$ $(1 \leqslant k \leqslant K)$ by maximizing the complete log-likelihood function (3.3). These estimates, $\widehat{\boldsymbol{\theta}}_k = (\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k)$, are given by (see McLachlan and Krishnan, 1997, p. 71)

$$\widehat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{N} \widehat{z}_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{N} \widehat{z}_{ik}} \quad (1 \leqslant k \leqslant K)$$

and

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{N} \widehat{z}_{ik} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)'}{\sum_{i=1}^{N} \widehat{z}_{ik}} \quad (1 \leqslant k \leqslant K).$$

If homoscedasticity is assumed then compute a pooled estimate of common $\boldsymbol{\Sigma}$:

$$\widehat{\boldsymbol{\Sigma}} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} \widehat{z}_{ik} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)'}{N}.$$

Finally compute the new estimate of η_k:

$$\widehat{\eta}_k = \frac{1}{N} \sum_{i=1}^{N} \widehat{z}_{ik} \quad (1 \leqslant k \leqslant K).$$

The final estimates of the posterior probabilities, $\widehat{z}_{ik}$, are used to assign the observations to clusters according to the maximum posterior rule:

$$R(\boldsymbol{x}_i) = C_k \Longleftrightarrow \widehat{z}_{ik} \geqslant \widehat{z}_{i\ell} \quad \forall \ell \neq k.$$

As $N_k \to \infty \forall k$, this rule approaches the Bayes rule for known parameters:

$$R(\boldsymbol{x}_i) = C_k \Longleftrightarrow \eta_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \geqslant \eta_\ell f_\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_\ell) \quad \forall \ell \neq k. \tag{3.5}$$

The *misclassification rate* (*MCR*), which is the proportion of misclassified observations, is generally used as the performance measure of any classification/clustering rule. Anderson (1958, Section 6.6) has shown that the Bayes rule minimizes the *expected misclassification rate* (*EMCR*) defined by

$$\text{EMCR} = \sum_{k=1}^{K} \eta_k \Pr\{R(\boldsymbol{x}_i) \neq C_k | i \in C_k\} = 1 - \sum_{k=1}^{K} \eta_k \Pr\{R(\boldsymbol{x}_i) = C_k | i \in C_k\}. \tag{3.6}$$

Therefore, the value of EMCR for the Bayes rule provides a lower bound on the EMCR for any other classification rule. We refer to this lower bound as the "gold standard." The maximum posterior rule (3.5) achieves this lower bound asymptotically (as $N_k \to \infty \forall k$) since the two rules then coincide.

Under the general MVN assumption, the Bayes rule classifies an observation $\boldsymbol{x}$ using the following rule:

$$R(\boldsymbol{x}) = C_k \Longleftrightarrow \frac{1}{2}[(\boldsymbol{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) - (\boldsymbol{x} - \boldsymbol{\mu}_\ell)' \boldsymbol{\Sigma}_\ell^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\ell)] \leqslant \ln \left(\frac{\eta_k |\boldsymbol{\Sigma}_\ell|^{1/2}}{\eta_\ell |\boldsymbol{\Sigma}_k|^{1/2}} \right) \quad \forall \ell \neq k. \tag{3.7}$$

This rule is quadratic in $\boldsymbol{x}$. Under homoscedasticity, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$, the rule becomes linear:

$$R(\boldsymbol{x}) = C_k \Longleftrightarrow (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)' \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \geqslant \frac{1}{2}[\boldsymbol{\mu}_k' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_\ell] - \ln \left(\frac{\eta_k}{\eta_\ell} \right) \quad \forall \ell \neq k. \tag{3.8}$$

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the *K*-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

An expression for the EMCR of this linear Bayes rule can be derived as follows. Denote $Y_{k\ell} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$, where $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. Then $Y_{k\ell} \sim \mathrm{N}(\xi_{k\ell}, \tau_{k\ell}^2)$, where

$$\xi_{k\ell} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \quad \text{and} \quad \tau_{k\ell}^2 = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell).$$

For notational convenience, denote

$$d_{k\ell} = \frac{1}{2}[\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell] - \ln\left(\frac{\eta_k}{\eta_\ell}\right).$$

Then from (3.6) we have

$$\begin{aligned}\mathrm{EMCR} &= 1 - \sum_{k=1}^{K} \eta_k \Pr(Y_{k\ell} > d_{k\ell} \forall \ell \neq k) \\ &= 1 - \sum_{k=1}^{K} \eta_k \Pr\left(Z_{k\ell} > \frac{d_{k\ell} - \xi_{k\ell}}{\tau_{k\ell}} \forall \ell \neq k\right),\end{aligned} \tag{3.9}$$

where the $Z_{k\ell}$ for $\ell \neq k$ are N(0, 1) r.v.s with

$$\mathrm{Corr}(Z_{k\ell}, Z_{k\ell'}) = \frac{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\ell'})}{\tau_{k\ell}\tau_{k\ell'}} \quad (\ell \neq \ell' \neq k).$$

This MVN probability can be evaluated given the values of all the parameters. As noted in the Introduction, the multivariate case will be studied in a future paper.

4. Univariate normal homoscedastic mixtures with two clusters

We now specialize to the univariate ($m = 1$) case with $K = 2$ clusters. Denote the cluster distributions by $\mathrm{N}(\mu_1, \sigma^2)$ and $\mathrm{N}(\mu_2, \sigma^2)$ and assume that $\mu_1 < \mu_2$. Let $\eta_1 = \eta$ and $\eta_2 = 1 - \eta$ be the mixing proportions. For this simple setting both the K-means algorithm and the MM method are defined by single thresholds, c and d, respectively, such that an observation x is classified to cluster C_2 if x exceeds the threshold and to cluster C_1 if x is less than the threshold. The MM method based on the EM algorithm approaches the Bayes rule asymptotically (as $N_k \to \infty \forall k$). In this section we will compare asymptotic EMCRs (which, for conveniences will be referred to simply as EMCRs) of the MM method and the K-means algorithm.

4.1. EMCR of the MM method

Asymptotically, the MM method clustering rule is equivalent to the Bayes rule (3.8):

$$R(x) = C_2 \iff x \geqslant d = \overline{\mu} + \frac{\sigma}{\delta}\ln\left(\frac{\eta}{1-\eta}\right), \tag{4.1}$$

where $\overline{\mu} = (\mu_1 + \mu_2)/2$ and $\delta = (\mu_2 - \mu_1)/\sigma > 0$. Note that d is an increasing and skew-symmetric function of η around $\eta = \frac{1}{2}$ where $d = \overline{\mu}$, i.e., if d and d' correspond to η and $\eta' = 1 - \eta$ then $d' = (\mu_1 + \mu_2) - d$. Fig. 1 shows d as a function of η for mixtures of N(1, 1) and N(3, 1) distributions. For comparison purposes, the threshold c of the K-means algorithm (studied analytically in the following subsection) is also plotted in the same figure. We see that the curves for c and d vary in opposite ways and cross at $\eta = \frac{1}{2}$ where $c = d = \overline{\mu}$.

The EMCR given by (3.9) simplifies to

$$\begin{aligned}\mathrm{EMCR} &= \eta \Pr_{\mu_1,\sigma}(X > d) + (1-\eta)\Pr_{\mu_2,\sigma}(X \leqslant d) \\ &= \eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d - \mu_2}{\sigma}\right).\end{aligned} \tag{4.2}$$

When $\eta = 0$, $d = -\infty$ and when $\eta = 1$, $d = +\infty$; in both cases, EMCR $= 0$. Additionally, when $\eta = \frac{1}{2}$, $d = \overline{\mu}$ and EMCR $= \Phi(-\delta/2)$. The following proposition gives a more detailed characterization of the EMCR.

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

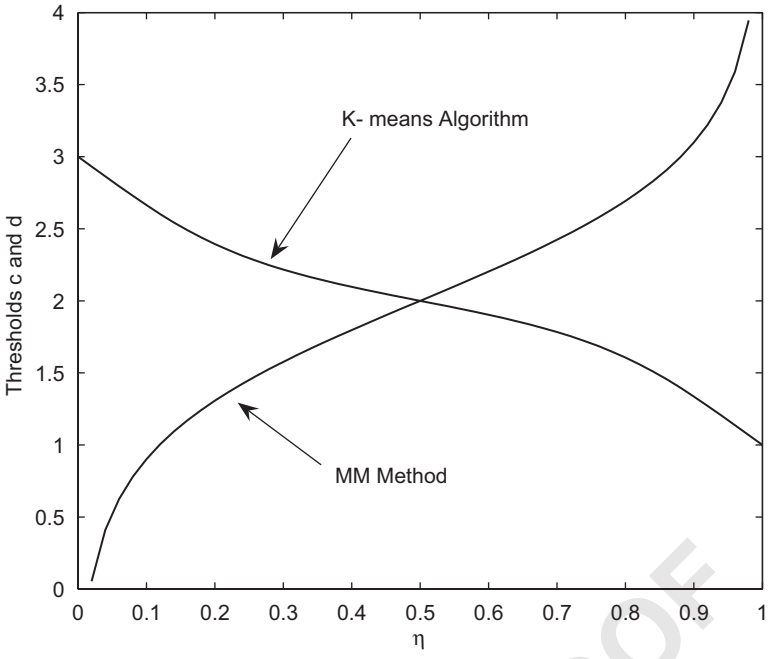

Fig. 1. Asymptotic thresholds c of the K-means algorithm and d of the MM method for mixtures of N(1, 1) and N(3, 1) distributions.

Proposition 1. *The* EMCR *of the MM method is symmetric in* η *around* $\frac{1}{2}$ *and is increasing for* $\eta < \frac{1}{2}$ *and decreasing for* $\eta > \frac{1}{2}$.

Proof. Let d and d' be the asymptotic threshold values of the MM method for the priors η and $\eta' = 1 - \eta$, respectively. As noted above, $d' = (\mu_1 + \mu_2) - d$. Let EMCR and EMCR$'$ be the corresponding EMCRs. Then

$$\begin{aligned}\text{EMCR}' &= \eta'\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) + (1-\eta')\Phi\left(\frac{d' - \mu_2}{\sigma}\right)\\ &= (1-\eta)\Phi\left(\frac{d-\mu_2}{\sigma}\right) + \eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right)\\ &= \text{EMCR}.\end{aligned}$$

To show that the EMCR is increasing for $\eta < \frac{1}{2}$, consider the Bayes rules for the priors η and $\eta' = \eta + \Delta\eta$ where $\Delta\eta > 0$ and $\eta, \eta' < \frac{1}{2}$. Denote by d and d' the threshold values for η and η', respectively, and the corresponding EMCRs by EMCR and EMCR$'$. Then from (4.2) we have

$$\begin{aligned}\text{EMCR}' - \text{EMCR} &= \left[(\eta+\Delta\eta)\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) + (1-\eta-\Delta\eta)\Phi\left(\frac{d'-\mu_2}{\sigma}\right)\right]\\ &\quad - \left[\eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d-\mu_2}{\sigma}\right)\right]\\ &= \left[\eta\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d'-\mu_2}{\sigma}\right)\right] - \left[\eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d-\mu_2}{\sigma}\right)\right]\\ &\quad + \Delta\eta\left[\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) - \Phi\left(\frac{d'-\mu_2}{\sigma}\right)\right]\\ &= T_1 + \Delta\eta T_2 \text{ (say)}.\end{aligned}$$

Now T_1 equals the difference between the EMCR of a non-optimal rule under η (since it uses the threshold d') and the EMCR of the optimal Bayes rule under η. Therefore, $T_1 \geqslant 0$. Next $T_2 > 0$ because $d' < \overline{\mu}$ when $\eta' < \frac{1}{2}$ and hence $\mu_1 - d' > d' - \mu_2$. Therefore, EMCR$'$ − EMCR > 0 as was to be shown. □

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

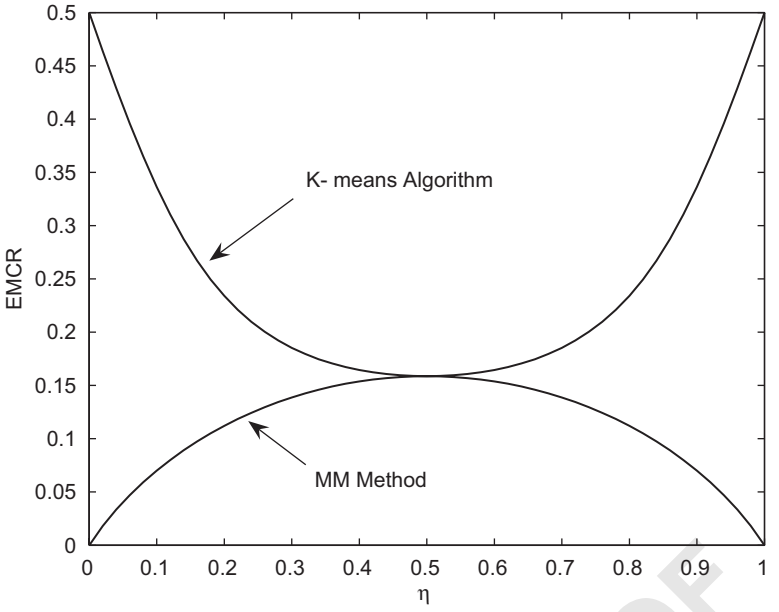

Fig. 2. EMCR of the K-means algorithm and the MM method for mixtures of N(1, 1) and N(3, 1) distributions.

Fig. 2 shows the EMCR of the MM method as a function of η for mixtures of N(1, 1) and N(3, 1) distributions. For comparison purposes the EMCR of the K-means algorithm (studied analytically in the following subsection) is also plotted in the same figure. We see that the two EMCR curves vary in opposite ways with equality at $\eta = \frac{1}{2}$; obviously, the EMCR of the Bayes rule is lower for all other η values.

4.2. EMCR of the K-means algorithm

For the K-means algorithm, the threshold c divides the data into two clusters such that the means of the two clusters are equidistant from the threshold. Asymptotically, the cluster means are weighted combinations of the conditional means of the data from each normal distribution, conditional on the data falling into the appropriate cluster. To evaluate these conditional means we use the following lemma.

Lemma 1. *Let* $X \sim \mathrm{N}(\mu, \sigma^2)$. *Then*

$$\alpha(c) = E_{\mu,\sigma}(X|X \leqslant c) = \mu - \frac{\sigma\phi((c-\mu)/\sigma)}{\Phi((c-\mu)/\sigma)} \quad \textit{and} \quad \beta(c) = E_{\mu,\sigma}(X|X > c) = \mu + \frac{\sigma\phi((\mu-c)/\sigma)}{\Phi((\mu-c)/\sigma)}, \tag{4.3}$$

where $\phi(\cdot)$ *and* $\Phi(\cdot)$ *are the standard normal p.d.f. and c.d.f., respectively.*

Proof. The p.d.f. of the $\mathrm{N}(\mu, \sigma^2)$ distribution equals $(1/\sigma)\phi((x-\mu)/\sigma)$. So

$$\begin{aligned}
\alpha(c) &= \frac{1}{\Phi((c-\mu)/\sigma)} \int_{-\infty}^{c} x \left(\frac{1}{\sigma}\right) \phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x \\
&= \frac{1}{\Phi((c-\mu)/\sigma)} \int_{-\infty}^{c} [\mu + (x-\mu)] \left(\frac{1}{\sigma}\right) \phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x \\
&= \mu + \frac{1}{\Phi((c-\mu)/\sigma)} \int_{-\infty}^{c} (x-\mu) \left(\frac{1}{\sigma}\right) \phi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x.
\end{aligned}$$

Make a change of variables $y = \phi((x-\mu)/\sigma)$. Then $(x-\mu)\phi((x-\mu)/\sigma)\,\mathrm{d}x = -\sigma^2\,\mathrm{d}y$. The expression for $\alpha(c)$ in (4.3) follows by making this substitution. The expression for $\beta(c)$ is derived in the same way. □

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

Let $\widetilde{\mu}_1(c)$ and $\widetilde{\mu}_2(c)$ denote the population means of the clusters formed of observations that are less than or greater than a specified threshold c, respectively. Then

$$\begin{aligned}\widetilde{\mu}_1(c) &= \frac{\eta \Pr_{\mu_1,\sigma}(X \leqslant c) E_{\mu_1,\sigma}(X|X \leqslant c) + (1-\eta) \Pr_{\mu_2,\sigma}(X \leqslant c) E_{\mu_2,\sigma}(X|X \leqslant c)}{\eta \Pr_{\mu_1,\sigma}(X \leqslant c) + (1-\eta) \Pr_{\mu_2,\sigma}(X \leqslant c)} \\ &= \alpha_1(c) p_\eta(c) + \alpha_2(c)[1 - p_\eta(c)], \end{aligned} \tag{4.4}$$

where

$$p_\eta(c) = \frac{\eta \Phi((c-\mu_1)/\sigma)}{\eta \Phi((c-\mu_1)/\sigma) + (1-\eta)\Phi((c-\mu_2)/\sigma)}, \tag{4.5}$$

and $\alpha_1(c)$ and $\alpha_2(c)$ are the values of $\alpha(c)$ from (4.3) when $\mu = \mu_1$ and μ_2, respectively. Similarly,

$$\widetilde{\mu}_2(c) = \beta_1(c) q_\eta(c) + \beta_2(c)[1 - q_\eta(c)], \tag{4.6}$$

where

$$q_\eta(c) = \frac{\eta \Phi((\mu_1 - c)/\sigma)}{\eta \Phi((\mu_1 - c)/\sigma) + (1-\eta)\Phi((\mu_2 - c)/\sigma)}, \tag{4.7}$$

and $\beta_1(c)$ and $\beta_2(c)$ are the values of $\beta(c)$ from (4.3) when $\mu = \mu_1$ and μ_2, respectively. Then c solves the equation

$$f_\eta(c) = \widetilde{\mu}_1(c) + \widetilde{\mu}_2(c) - 2c = 0. \tag{4.8}$$

Remark 1. Note that although the K-means algorithm does not explicitly take into account the prior η and the underlying probability model, the asymptotic threshold c used by it depends on these quantities through the above equation.

To prove the existence, uniqueness and monotonicity of the solution c to the above equation, we need the following two lemmas.

Lemma 2. *The function* $f(x) = \phi(x)/\Phi(x)$ *is decreasing in* $x\ \forall x \in (-\infty, \infty)$.

Proof. The derivative of $f(x)$ equals

$$f'(x) = \frac{-x\phi(x)\Phi(x) - \phi^2(x)}{\Phi^2(x)}.$$

Obviously, $f'(x) < 0$ for $x > 0$. For $x < 0$, put $x = -y$ where $y > 0$. Then the numerator of $f'(x)$ equals $\phi(y)[y\Phi(-y) - \phi(y)]$, which is < 0 since Mills' ratio $r(y) = \Phi(-y)/\phi(y) < 1/y$; see Johnson and Kotz (1970, p. 279). Hence $f'(x) < 0$ $\forall x \in (-\infty, \infty)$. □

Corollary 1. *For* $\mu_1 < \mu_2$, *we have* $\mu_1 - \alpha_1(c) < \mu_2 - \alpha_2(c)$.

Proof. The inequality is equivalent to

$$\frac{\phi((c-\mu_1)/\sigma)}{\Phi((c-\mu_1)/\sigma)} < \frac{\phi((c-\mu_2)/\sigma)}{\Phi((c-\mu_2)/\sigma)},$$

which follows by putting $x = (c-\mu)/\sigma$, and noting that $f(x)$ is decreasing in x and hence increasing in μ. □

Lemma 3. *The function* $g(x) = x + \phi(x)/\Phi(x)$ *is increasing in* $x\ \forall x \in (-\infty, \infty)$.

Proof. The derivative of $g(x)$ is

$$\begin{aligned} g'(x) &= 1 - \frac{x\phi(x)\Phi(x) + \phi^2(x)}{\Phi^2(x)} \\ &= \frac{\Phi^2(x) - x\phi(x)\Phi(x) - \phi^2(x)}{\Phi^2(x)}. \end{aligned}$$

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

So we only need to prove that the numerator of this derivative, $h(x) = \Phi^2(x) - x\phi(x)\Phi(x) - \phi^2(x)$, is positive. Taking the derivative of $h(x)$, we get

$$\begin{aligned} h'(x) &= 2\Phi(x)\phi(x) - \phi(x)\Phi(x) + x^2\phi(x)\Phi(x) - x\phi^2(x) + 2x\phi^2(x) \\ &= \Phi(x)\phi(x) + x^2\phi(x)\Phi(x) + x\phi^2(x). \end{aligned}$$

So $h'(x) > 0\ \forall x > 0$.

For $x < 0$, put $x = -y$ where $y > 0$. Then $h'(x) = \phi(y)[(1+y^2)\Phi(-y) - y\phi(y)]$. It follows that $h'(x) > 0 \forall x < 0$ since Mills' ratio $r(y) = \Phi(-y)/\phi(y) > y/(1+y^2)$; see Johnson and Kotz (1970, p. 279). To complete the proof we need to show that

$$\lim_{x\to-\infty} h(x) = \lim_{x\to-\infty} x^2\Phi^2(x)\left[\frac{1}{x^2} - \frac{\phi(x)}{x\Phi(x)} - \left(\frac{\phi(x)}{x\Phi(x)}\right)^2\right] \geqslant 0,$$

which together with the fact that $h'(x) > 0\ \forall x$ implies that $h(x) > 0\ \forall x$. Again putting $x = -y$, we see that the above inequality is equivalent to

$$\lim_{y\to\infty}\left[\frac{1}{y^2} + \frac{1}{yr(y)} - \left(\frac{1}{yr(y)}\right)^2\right] \geqslant 0.$$

But it is well known that $\lim_{y\to\infty} yr(y) = 1$. Hence the above limit equals 0. This completes the proof of $g'(x) > 0\ \forall x \in (-\infty, \infty)$. □

Corollary 2. *The function $\alpha(c) = E_{\mu,\sigma}(X|X \leqslant c)$ is increasing in μ for all c. Hence for $\mu_1 < \mu_2$, we have $\alpha_1(c) < \alpha_2(c)$ and $\beta_1(c) < \beta_2(c)$.*

Proof. Write

$$\alpha(c) = c - \sigma\left[\left(\frac{c-\mu}{\sigma}\right) + \frac{\phi((c-\mu)/\sigma)}{\Phi((c-\mu)/\sigma)}\right].$$

Now put $x = ((c-\mu)/\sigma)$. Then $\alpha(c)$ is decreasing in x and hence increasing in μ. The proof of $\beta_1(c) < \beta_2(c)$ is analogous. □

We are now ready to state and prove the following two propositions regarding the existence, uniqueness and monotonicity of c.

Proposition 2. *For $\mu_1 < \mu_2$ and $\eta \in [0,1]$, there exists a solution c to Eq.* (4.8).

Proof. We will show that $f_\eta(\mu_1) > 0$ and $f_\eta(\mu_2) < 0$, where $f_\eta(\cdot)$ is defined in (4.8). Then by the continuity of $f_\eta(\cdot)$ and the intermediate value theorem, the existence of the solution to $f_\eta(c) = 0$ for some $c \in [\mu_1, \mu_2]$ will be established.

Write

$$\begin{aligned} f_\eta(\mu_1) &= \widetilde{\mu}_1(\mu_1) + \widetilde{\mu}_2(\mu_1) - 2\mu_1 \\ &= [\alpha_1(\mu_1) - \mu_1]p_\eta(\mu_1) + [\alpha_2(\mu_1) - \mu_1][1 - p_\eta(\mu_1)] \\ &\quad + [\beta_1(\mu_1) - \mu_1]q_\eta(\mu_1) + [\beta_2(\mu_1) - \mu_1][1 - q_\eta(\mu_1)] \\ &= -\frac{\sigma\sqrt{2/\pi}(0.5\eta) + \sigma[-\delta + \phi(-\delta)/\Phi(-\delta)](1-\eta)\Phi(-\delta)}{0.5\eta + (1-\eta)\Phi(-\delta)} \\ &\quad + \frac{\sigma\sqrt{2/\pi}(0.5\eta) + \sigma[\delta + \phi(\delta)/\Phi(\delta)](1-\eta)\Phi(\delta)}{0.5\eta + (1-\eta)\Phi(\delta)} \\ &> -\frac{\sigma\sqrt{2/\pi}(0.5\eta) + \sigma[\delta + \phi(\delta)/\Phi(\delta)](1-\eta)\Phi(-\delta)}{0.5\eta + (1-\eta)\Phi(-\delta)} \\ &\quad + \frac{\sigma\sqrt{2/\pi}(0.5\eta) + \sigma[\delta + \phi(\delta)/\Phi(\delta)](1-\eta)\Phi(\delta)}{0.5\eta + (1-\eta)\Phi(\delta)}, \end{aligned} \tag{4.9}$$

where we have used the inequality

$$\delta + \frac{\phi(\delta)}{\Phi(\delta)} > -\delta + \frac{\phi(-\delta)}{\Phi(-\delta)},$$

which follows from Lemma 3. Now put

$$s = \sqrt{\frac{2}{\pi}}, \quad t = \delta + \frac{\phi(\delta)}{\Phi(\delta)}, \quad u = \Phi(-\delta) \quad \text{and} \quad v = \Phi(\delta).$$

Then simple algebra shows that the lower bound on $f_\eta(\mu_1)$ obtained in (4.9) is strictly > 0 iff $(t-s)(v-u) > 0$. This inequality holds because $t = g(\delta) > g(0) = s$ from Lemma 3 and $v > u$. Similarly we can show that $f_\eta(\mu_2) < 0$. This proves the existence of $c \in [\mu_1, \mu_2]$ such that $f_\eta(c) = 0$. □

Proposition 3. *The solution c to* (4.8) *is decreasing, skew-symmetric and one-to-one function of* η. *Hence c is unique with* $c = \mu_2$ *for* $\eta = 0$, $c = \mu_1$ *for* $\eta = 1$ *and* $c = \overline{\mu}$ *for* $\eta = \frac{1}{2}$.

Proof. Write $f_\eta(c)$ as

$$f_\eta(c) = \alpha_1(c)p_\eta(c) + \alpha_2(c)[1 - p_\eta(c)] + \beta_1(c)q_\eta(c) + \beta_2(c)[1 - q_\eta(c)] - 2c. \tag{4.10}$$

Note that $p_\eta(c), q_\eta(c)$ are increasing in η. Furthermore, from the corollary to Lemma 3 we have $\alpha_1(c) < \alpha_2(c)$ and $\beta_1(c) < \beta_2(c)$. Since, as η increases, more weight is put on smaller quantities, $\alpha_1(c)$ and $\beta_1(c)$, and less weight on larger quantities, $\alpha_2(c)$ and $\beta_2(c)$, it follows that $f_\eta(c)$ decreases in η. Therefore, if c is the root of Eq. (4.8), then $f_{\eta'}(c) < 0$ for $\eta' > \eta$. But in Proposition 2 we have shown that $f_{\eta'}(\mu_1) > 0$. By the intermediate value theorem, there exists $c' \in (\mu_1, c)$ such that $f_{\eta'}(c') = 0$, i.e., $c' < c$ is the root of Eq. (4.8) for $\eta' > \eta$. Hence the solution c to $f_\eta(c) = 0$ is decreasing in η.

To show that for any fixed $\eta \in [0, 1]$ the solution c is unique, first note that for $\eta = 0$ and $\eta = 1$ we have unique solutions $c = \mu_2$ and $c = \mu_1$, respectively. For example, for $\eta = 0$, we have $p_0(c) = q_0(c) = 0$, and the equation for c is

$$\begin{aligned} f_0(c) &= \widetilde{\mu}_1(c) + \widetilde{\mu}_2(c) - 2c \\ &= \alpha_2(c) + \beta_2(c) - 2c \\ &= 2\mu_2 - \sigma\phi\left(\frac{c-\mu_2}{\sigma}\right)\left[\frac{1}{\Phi((c-\mu_2)/\sigma)} - \frac{1}{\Phi((\mu_2-c)/\sigma)}\right] - 2c = 0. \end{aligned}$$

This last equation can be rewritten as $g(\delta) = g(-\delta)$ where the function $g(\cdot)$ is defined in Lemma 3 and $\delta = (\mu_2 - c)/\sigma$. But, as shown in that lemma, $g(\cdot)$ is a strictly increasing function and so the only solution to the above equation is $\delta = 0$, i.e., $c = \mu_2$. Similarly, $c = \mu_1$ is the unique solution for $\eta = 1$. Now suppose that for any other $\eta \in [0, 1]$ there are two distinct solutions, c_1 and c_2, such that $f_\eta(c_1) = f_\eta(c_2) = 0$. Then it must be the case that for some c there are two distinct η_1 and η_2 such that $f_{\eta_1}(c) = f_{\eta_2}(c) = 0$, which contradicts the just proven fact that $f_\eta(c)$ is a strictly decreasing function of η. Therefore, the solution c is unique for all $\eta \in [0, 1]$. For $\eta = \frac{1}{2}$, by symmetry we obtain $c = \overline{\mu}$ as the unique solution.

Finally, we will show the skew-symmetric property of c. Consider two priors η and $\eta' = 1 - \eta$, and let c and c' be the corresponding asymptotic threshold values of the K-means algorithm. Thus, c satisfies the equation $f_\eta(c) = 0$. We will show by direct substitution that $c' = (\mu_1 + \mu_2) - c = 2\overline{\mu} - c$ satisfies the equation $f_{\eta'}(c') = 0$. We can readily check the following relations:

$$\alpha_1(c) = 2\overline{\mu} - \beta_2(c'), \quad \alpha_2(c) = 2\overline{\mu} - \beta_1(c'), \quad \beta_1(c) = 2\overline{\mu} - \alpha_2(c'), \quad \beta_2(c) = 2\overline{\mu} - \alpha_1(c')$$

and

$$p_\eta(c) = 1 - q_{\eta'}(c'), \quad q_\eta(c) = 1 - p_{\eta'}(c').$$

Substituting these expressions in $f_\eta(c) = 0$ we get

$$\begin{aligned} 0 = f_\eta(c) &= [2\overline{\mu} - \beta_2(c')][1 - q_{\eta'}(c')] + [2\overline{\mu} - \beta_1(c')]q_{\eta'}(c') \\ &\quad + [2\overline{\mu} - \alpha_2(c')][1 - p_{\eta'}(c')] + [2\overline{\mu} - \alpha_1(c')]p_{\eta'}(c') - 2(\mu_1 + \mu_2) + 2c' \\ &= -\alpha_1(c')p_{\eta'}(c') - \alpha_2(c')[1 - p_{\eta'}(c')] - \beta_1(c')q_{\eta'}(c') - \beta_2(c')[1 - q_{\eta'}(c')] + 2c' \\ &= -f_{\eta'}(c'), \end{aligned}$$

which shows that $f_{\eta'}(c') = 0$. □

Remark 2. The behavior of c as a function of η is opposite to that of the asymptotic threshold d of the MM method. Fig. 1 shows c as a function of η for mixtures of N(1, 1) and N(3, 1) distributions.

The EMCR of the K-means algorithm is given by expression (4.2) with d replaced by c. Since $c = \mu_2$ for $\eta = 0$ and $c = \mu_1$ for $\eta = 1$, it follows that EMCR = 0.5 for $\eta = 0$ and $\eta = 1$. For $\eta = \frac{1}{2}$, the EMCR values of the Bayes rule and the K-means algorithm are equal to $\Phi(-\delta/2)$ since $c = d = \overline{\mu}$. For all other $\eta \in [0, 1]$, the EMCR of the MM method is smaller because of the optimality property of the associated Bayes rule referred to earlier. The following proposition shows that the EMCR of the K-means algorithm is a symmetric function of η as is the EMCR of the MM method.

Proposition 4. *The* EMCR *of the K-means algorithm is symmetric around* $\eta = \frac{1}{2}$ *and is decreasing in* η *for* $\eta < \frac{1}{2}$ *and increasing in* η *for* $\eta > \frac{1}{2}$.

Proof. The symmetry of the EMCR of the K-means algorithm follows from the skew-symmetry of c in the same manner as the symmetry of the EMCR of the MM method follows from the skew-symmetry of d. Now we will show that EMCR is decreasing in η for $\eta < \frac{1}{2}$. Let EMCR and EMCR$'$ correspond to η and $\eta' < \eta$, respectively, where $\eta < \frac{1}{2}$ and $\eta' = \eta - \Delta\eta$. Then

$$\begin{aligned} \text{EMCR}' &= \eta'\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1 - \eta')\Phi\left(\frac{c' - \mu_2}{\sigma}\right) \\ &= \eta\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1 - \eta)\Phi\left(\frac{c' - \mu_2}{\sigma}\right) + \Delta\eta\left[\Phi\left(\frac{c' - \mu_2}{\sigma}\right) - \Phi\left(\frac{\mu_1 - c'}{\sigma}\right)\right] \\ &> \eta\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1 - \eta)\Phi\left(\frac{c' - \mu_2}{\sigma}\right) \end{aligned}$$

since

$$\Phi\left(\frac{c' - \mu_2}{\sigma}\right) > \Phi\left(\frac{\mu_1 - c'}{\sigma}\right),$$

which follows from the fact that $c' > \overline{\mu}$ for $\eta' < \frac{1}{2}$. Hence, to prove that EMCR$'$ > EMCR, it suffices to show that

$$\begin{aligned} &\eta\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1 - \eta)\Phi\left(\frac{c' - \mu_2}{\sigma}\right) > \eta\Phi\left(\frac{\mu_1 - c}{\sigma}\right) + (1 - \eta)\Phi\left(\frac{c - \mu_2}{\sigma}\right) \\ &\iff (1 - \eta)\left[\Phi\left(\frac{c' - \mu_2}{\sigma}\right) - \Phi\left(\frac{c - \mu_2}{\sigma}\right)\right] > \eta\left[\Phi\left(\frac{\mu_1 - c}{\sigma}\right) - \Phi\left(\frac{\mu_1 - c'}{\sigma}\right)\right] \\ &\iff (1 - \eta)\int_c^{c'} \phi_{\mu_2,\sigma}(x)\,\mathrm{d}x > \eta\int_c^{c'} \phi_{\mu_1,\sigma}(x)\,\mathrm{d}x, \end{aligned}$$

where $\phi_{\mu,\sigma}(x)$ is the p.d.f. of the N(μ, σ^2) distribution. The last step follows because $\overline{\mu}$ is the point of intersection of $\phi_{\mu_1,\sigma}(x)$ and $\phi_{\mu_2,\sigma}(x)$, and since $\eta < \frac{1}{2}$ and c', $c > \overline{\mu}$, for $c \leqslant x \leqslant c'$ we have $(1 - \eta)\phi_{\mu_2,\sigma}(x) > \eta\phi_{\mu_1,\sigma}(x)$. □

Remark 3. The monotone behavior of the EMCR of the K-means algorithm as a function of η is opposite to that of the EMCR of the MM method. Fig. 2 shows the EMCR of the K-means algorithm as a function of η for mixtures of N(1, 1) and N(3, 1) distributions. It should be noted that for η close to 0 or 1, essentially we have a single cluster. The MM

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the *K*-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

can deal with this problem because it estimates η in a continuous manner. On the other hand, the K-means algorithm is forced to divide the data set into two clusters even if there are no observations from the cluster having the smaller value η or $1-\eta$. In practice, the user would generally perform a test of $K=1$ vs. $K=2$, which would improve the performance of the K-means algorithm. Therefore, the discrepancy in the EMCR functions of the two methods may not be as large in practice as shown in Fig. 2, especially for η-values in the extreme.

5. Univariate normal heteroscedastic mixtures with two clusters

Denote the two cluster distributions by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ and assume that $\mu_1 < \mu_2$ and $\sigma_1^2 < \sigma_2^2$ without loss of generality. In this section we carry out a comparison between the EMCRs of the MM method and the K-means algorithm paralleling that for the homoscedastic case.

5.1. EMCR of the MM method

In this case, asymptotically, the MM method clustering rule is equivalent to the Bayes rule (3.8):

$$R(x)=C_1 \iff \frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] \leqslant \ln\left(\frac{\eta_1\sigma_2}{\eta_2\sigma_1}\right), \tag{5.1}$$

where $\eta_1=\eta$ and $\eta_2=1-\eta$.

Consider the quadratic equation obtained by making the above inequality an equality. For convenience, we will refer to this quadratic equation by the same equation number. If there is no real root or a single root of this equation, then rule (5.1) is $R(x)=C_2$ for all x. The quadratic equation has two distinct real roots, say $d_1<d_2$, if its discriminant is >0, i.e., if

$$\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)^2-\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)\left[\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}-2\ln\left(\frac{\eta_1\sigma_2}{\eta_2\sigma_1}\right)\right]>0.$$

Denoting

$$k=\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}-\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)^{-1}\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)^2,$$

we see that the above condition is equivalent to

$$\eta>\eta^*=\frac{\sigma_1\exp(k/2)}{\sigma_1\exp(k/2)+\sigma_2}. \tag{5.2}$$

It is easy to check that $k<0$ and hence

$$\eta^*=\frac{\sigma_1\exp(k/2)}{\sigma_1\exp(k/2)+\sigma_2}<\frac{\sigma_1}{\sigma_1+\sigma_2}=\eta^{**}.$$

If $\eta>\eta^*$ then rule (5.1) is $R(x)=C_1$ if $d_1\leqslant x\leqslant d_2$; otherwise $R(x)=C_2$. The two real roots are the points of intersection of the prior-weighted p.d.f.s, $\eta\phi_{\mu_1,\sigma_1}(x)$ and $(1-\eta)\phi_{\mu_2,\sigma_2}(x)$. Fig. 3 depicts this graphically where the prior-weighted p.d.f. curves are shown by dotted lines for $\eta\in(\eta^*,\eta^{**})$. When $\eta=\eta^{**}$, we have

$$d_1=\frac{\mu_1\sigma_2-\mu_2\sigma_1}{\sigma_2-\sigma_1} \quad \text{and} \quad d_2=\frac{\mu_1\sigma_2+\mu_2\sigma_1}{\sigma_2+\sigma_1}. \tag{5.3}$$

These points of intersection are shown in the same figure with the prior-weighted p.d.f. curves for $\eta=\eta^{**}$ being shown by solid lines.

It is clear that as η decreases and $1-\eta$ increases, d_1 increases and d_2 decreases. In particular, if $\eta<\eta^{**}$ then

$$d_1>\frac{\mu_1\sigma_2-\mu_2\sigma_1}{\sigma_2-\sigma_1} \quad \text{and} \quad d_2<\frac{\mu_1\sigma_2+\mu_2\sigma_1}{\sigma_2+\sigma_1}. \tag{5.4}$$

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

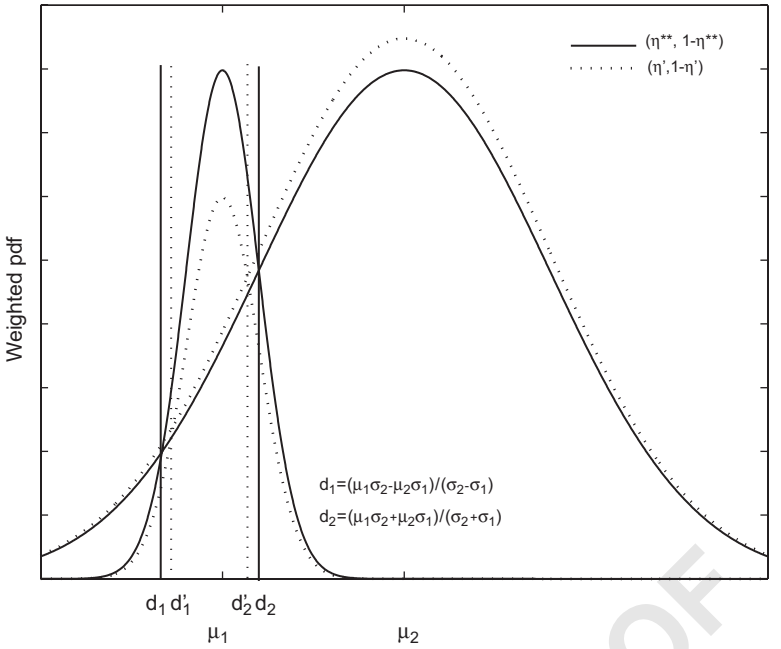

Fig. 3. Thresholds (d_1, d_2) and (d_1', d_2') of the MM method for mixtures of N(1, 1) and N(4, 4) distributions for two priors, $\eta = \eta^{**}$ and $\eta = \eta' < \eta^{**}$.

When $\eta = \eta^*$, d_1 and d_2 are equal. When η decreases further, real roots d_1 and d_2 do not exist since the two prior-weighted p.d.f. curves do not intersect or equivalently the quadratic curve in (5.1) lies completely in the upper half of the coordinate plane. As η increases for $\eta > \eta^*$, d_1 decreases and d_2 increases. When $\eta = 1$, we have $d_1 = -\infty$ and $d_2 = \infty$ (so that $R(x) = C_1 \forall x$).

Fig. 4 shows how d_1 and d_2 change with η for mixtures of N(1, 1) and N(4, 4) distributions. In this case η^* and η^{**} can be calculated to be $\eta^* = 0.1004$, $\eta^{**} = 0.3333$. The K-means algorithm uses a single threshold c (studied analytically in the following subsection) which is also plotted in the same figure for comparison purposes.

For $\eta \leqslant \eta^*$, since $R(x) = C_2 \forall x$, the EMCR of the MM method equals η. For $\eta > \eta^*$, this EMCR is given by

$$\begin{aligned}\text{EMCR} &= \eta \Pr_{\mu_1,\sigma_1}\{(X < d_1) \cup (X > d_2)\} + (1-\eta)\Pr_{\mu_2,\sigma_2}\{d_1 \leqslant X \leqslant d_2\} \\ &= \eta\left[\Phi\left(\frac{d_1-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1-d_2}{\sigma_1}\right)\right] + (1-\eta)\left[\Phi\left(\frac{d_2-\mu_2}{\sigma_2}\right) - \Phi\left(\frac{d_1-\mu_2}{\sigma_2}\right)\right].\end{aligned}$$

From the above we can conclude that EMCR $= 0$ for $\eta = 0$ and 1 (since $d_1 = -\infty$ and $d_2 = \infty$ in that case). The following proposition gives a more detailed characterization of the EMCR.

Proposition 5. *The EMCR of the MM method increases in η for $\eta < \eta^{**}$ and reaches a maximum at $\overline{\eta} > \eta^{**}$ where $\overline{\eta}$ solves the equation*

$$\Phi\left(\frac{d_1-\mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1-d_2}{\sigma_1}\right) - \Phi\left(\frac{d_2-\mu_2}{\sigma_2}\right) + \Phi\left(\frac{d_1-\mu_2}{\sigma_2}\right) = 0; \tag{5.5}$$

here d_1 and d_2 are the roots of the quadratic equation (5.1) *and hence depend on η.*

Proof. As shown before, for $\eta \leqslant \eta^*$, EMCR $= \eta$, which increases linearly in $\eta \leqslant \eta^*$. For $\eta \in (\eta^*, \eta^{**})$, the proof of monotonicity is similar to that of Proposition 1. Consider the Bayes rules for η and $\eta' = \eta + \Delta\eta < \eta^{**}$ where $\Delta\eta > 0$. Denote by d_1' and d_2' the threshold values for η'. Then EMCR$'$ − EMCR can be decomposed into two terms as in Proposition 1. The first term is the difference in the EMCR of a non-optimal rule under η that uses the thresholds d_1' and d_2', and the EMCR of the optimal rule under η that uses the thresholds d_1 and d_2; hence this term is positive.

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

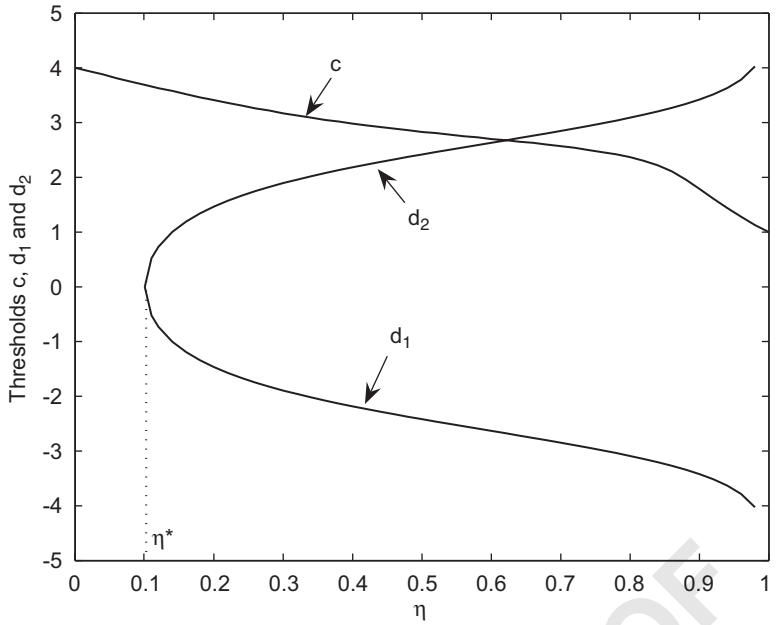

Fig. 4. Asymptotic thresholds c of the K-means algorithm and (d_1, d_2) of the MM method for mixtures of N(1, 1) and N(4, 4) distributions (For $\eta < \eta^*$, (d_1, d_2) do not exist).

The second term equals

$$\Delta\eta\left[\Phi\left(\frac{d_1'-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{\mu_1-d_2'}{\sigma_1}\right)-\Phi\left(\frac{d_2'-\mu_2}{\sigma_2}\right)+\Phi\left(\frac{d_1'-\mu_2}{\sigma_2}\right)\right].$$

This term is positive since $d_2' < (\mu_1\sigma_2+\mu_2\sigma_1)(\sigma_2+\sigma_1)$ if $\eta, \eta' < \eta^{**}$ as seen from (5.4). Hence

$$\Phi\left(\frac{\mu_1-d_2'}{\sigma_1}\right)-\Phi\left(\frac{d_2'-\mu_2}{\sigma_2}\right)>0.$$

Since, as noted before, d_1 decreases and d_2 increases with increasing η, the left-hand side of (5.5), regarded as a function of η and denoted by $g(\eta)$, is a decreasing function. It is easy to show using the (d_1, d_2) values from (5.3) for $\eta = \eta^{**}$ that $g(\eta^{**}) = 2\Phi((\mu_1-\mu_2)/(\sigma_2-\sigma_1)) > 0$ and $g(1) = -1$. Therefore, there exists $\overline{\eta} \in (\eta^{**}, 1)$ such that $g(\eta) > 0$ for $\eta < \overline{\eta}$, $g(\eta) < 0$ for $\eta > \overline{\eta}$ and $g(\overline{\eta}) = 0$. Now consider η and $\eta' = \eta - \Delta\eta$ such that $\eta, \eta' > \overline{\eta}$. The difference between the corresponding EMCR values is then

$$\begin{aligned}
\mathrm{EMCR}' - \mathrm{EMCR} = &\left\{(\eta-\Delta\eta)\left[\Phi\left(\frac{d_1'-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{\mu_1-d_2'}{\sigma_1}\right)\right]\right.\\
&\left.+(1-\eta+\Delta\eta)\left[\Phi\left(\frac{d_2'-\mu_2}{\sigma_2}\right)-\Phi\left(\frac{d_1'-\mu_2}{\sigma_2}\right)\right]\right\}\\
&-\left\{\eta\left[\Phi\left(\frac{d_1-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{\mu_1-d_2}{\sigma_1}\right)\right]+(1-\eta)\left[\Phi\left(\frac{d_2-\mu_2}{\sigma_2}\right)-\Phi\left(\frac{d_1-\mu_2}{\sigma_2}\right)\right]\right\}\\
&\geqslant -\Delta\eta\left[\Phi\left(\frac{d_1'-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{\mu_1-d_2'}{\sigma_1}\right)-\Phi\left(\frac{d_2'-\mu_2}{\sigma_2}\right)+\Phi\left(\frac{d_1'-\mu_2}{\sigma_2}\right)\right]\\
&\geqslant -\Delta\eta g(\eta')\geqslant 0,
\end{aligned}$$

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

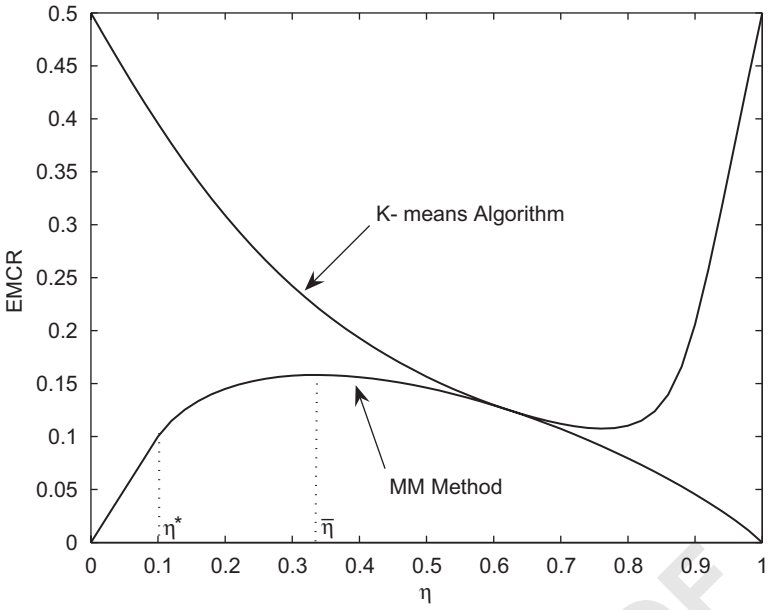

Fig. 5. EMCR of the *K*-means algorithm and the MM method for mixtures of N(1, 1) and N(4, 4) distributions.

since $g(\eta') \leqslant 0$. In the second to last step above, the inequality is obtained by dropping the term

$$\left\{\eta\left[\Phi\left(\frac{d_1'-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{\mu_1-d_2'}{\sigma_1}\right)\right]+(1-\eta)\left[\Phi\left(\frac{d_2'-\mu_2}{\sigma_2}\right)-\Phi\left(\frac{d_1'-\mu_2}{\sigma_2}\right)\right]\right\}$$
$$-\left\{\eta\left[\Phi\left(\frac{d_1-\mu_1}{\sigma_1}\right)+\Phi\left(\frac{\mu_1-d_2}{\sigma_1}\right)\right]+(1-\eta)\left[\Phi\left(\frac{d_2-\mu_2}{\sigma_2}\right)-\Phi\left(\frac{d_1-\mu_2}{\sigma_2}\right)\right]\right\},$$

which is positive because it is the difference between the EMCR of a non-optimal rule that uses d_1' and d_2' under the prior η and the EMCR of the optimal Bayes rule that uses d_1 and d_2 under the prior η. Therefore, we have shown that for $\eta > \overline{\eta}$, the EMCR decreases. Similarly, it can be shown that for $\eta < \overline{\eta}$, the EMCR increases. Therefore, the EMCR reaches a maximum when $\eta = \overline{\eta}$, which is the solution to Eq. (5.5). □

Fig. 5 shows a plot of the EMCR as a function of η for the MM method for mixtures of N(1, 1) and N(3, 4) distributions. The point of maximum EMCR obtained by solving Eq. (5.5) equals $\overline{\eta} = 0.3358$, which is slightly greater than η^{**}. The EMCR of the *K*-means algorithm (discussed in the following subsection) is also plotted in the same figure for comparison purposes.

5.2. EMCR of the K-means algorithm

The *K*-means algorithm does not distinguish between homoscedasticity and heteroscedasticity, and uses a single threshold c to assign observations to two clusters. Therefore, c is determined by the same Eq. (4.8), where now the quantities $\alpha_i(c)$, $\beta_i(c)$, $p_\eta(c)$ and $q_\eta(c)$ depend on both μ_i and σ_i $(i = 1, 2)$ in an obvious way.

Proposition 6. *For $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$, the solution c to* (4.8) *is a decreasing function of η. Furthermore, $c > \overline{\mu}$ for $\eta = \frac{1}{2}$.*

Proof. It is straightforward to see that the properties of the $f_\eta(\cdot)$ function shown in Propositions 2 and 3 for the homoscedastic case extend to its modification for the heteroscedastic case. In particular, $f_\eta(\mu_1) > 0$, $f_\eta(\mu_2) < 0$ and $f_\eta(x)$ is decreasing in η. From this it follows that c is a decreasing function of η; the proof is similar to that of Proposition 3.

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the *K*-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

To show the second part of the proposition, we will show that $f_{1/2}(\overline{\mu}) > 0$, so that if $f_{1/2}(c) = 0$ then $c > \overline{\mu}$. Denote $\Delta = (\mu_2 - \mu_1)/2$ and note that

$$\overline{\mu} - \alpha_1(\overline{\mu}) = \Delta + \frac{\sigma_1\phi(\Delta/\sigma_1)}{\Phi(\Delta/\sigma_1)}, \quad \overline{\mu} - \alpha_2(\overline{\mu}) = -\Delta + \frac{\sigma_2\phi(-\Delta/\sigma_2)}{\Phi(-\Delta/\sigma_2)},$$

$$\beta_1(\overline{\mu}) - \overline{\mu} = -\Delta + \frac{\sigma_1\phi(-\Delta/\sigma_1)}{\Phi(-\Delta/\sigma_1)}, \quad \beta_2(\overline{\mu}) - \overline{\mu} = \Delta + \frac{\sigma_2\phi(\Delta/\sigma_2)}{\Phi(\Delta/\sigma_2)}.$$

Substituting these values in the expression for $f_{1/2}(\overline{\mu})$ and recalling that $\eta = 1 - \eta$ get cancelled from the numerator and denominator, we get

$$\begin{aligned} f_{1/2}(\overline{\mu}) = & -\frac{[\overline{\mu} - \alpha_1(\overline{\mu})]\Phi(\Delta/\sigma_1) + [\overline{\mu} - \alpha_2(\overline{\mu})]\Phi(-\Delta/\sigma_2)}{\Phi(\Delta/\sigma_1) + \Phi(-\Delta/\sigma_2)} \\ & + \frac{[\beta_1(\overline{\mu}) - \overline{\mu}]\Phi(-\Delta/\sigma_1) + [\beta_2(\overline{\mu}) - \overline{\mu}]\Phi(\Delta/\sigma_2)}{\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2)} \\ = & -\frac{\Delta\Phi(\Delta/\sigma_1) + \sigma_1\phi(\Delta/\sigma_1) - \Delta\Phi(-\Delta/\sigma_2) + \sigma_2\phi(-\Delta/\sigma_2)}{\Phi(\Delta/\sigma_1) + \Phi(-\Delta/\sigma_2)} \\ & + \frac{-\Delta\Phi(-\Delta/\sigma_1) + \sigma_1\phi(-\Delta/\sigma_1) + \Delta\Phi(\Delta/\sigma_2) + \sigma_2\phi(\Delta/\sigma_2)}{\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2)}. \end{aligned}$$

Now, the numerators of the two terms are equal since $\phi(x) = \phi(-x)$ and $\Phi(\Delta/\sigma_1) - \Phi(-\Delta/\sigma_2) = -\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2)$. Hence $f_{1/2}(\overline{\mu}) > 0$ if $\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2) < \Phi(\Delta/\sigma_1) + \Phi(-\Delta/\sigma_2)$, which can be easily checked to be true. This completes the proof. □

Remark 4. Fig. 4 shows a plot of c as a function of η for mixtures of N(1, 1) and N(4, 4) distributions. From this figure we see that c and d_2 are equal for some η. This value of η can be found by solving Eqs. (4.8) and (5.1) simultaneously under the constraint that $c = d_2$. The common value is found to be 2.67667 at $\eta = 0.62095$. At this value, the EMCR values of the two methods are nearly (but not exactly) equal as seen from Fig. 5. The two EMCR values are 0.125291 for the MM method and 0.125377 for the K-means algorithm.

Proposition 7. *For $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$, the EMCR of the K-means algorithm is decreasing in η for $\eta < \eta^{**}$.*

Proof. The proof is similar to that of Proposition 4. Let EMCR and EMCR′ be the expected MCR values for the K-means algorithm corresponding to the priors $\eta < \eta^{**}$ and $\eta' = \eta - \Delta\eta < \eta$. Then using the fact that $c' > \overline{\mu} > m = (\mu_1\sigma_2 + \mu_2\sigma_1)/(\sigma_1 + \sigma_2)$ and hence

$$\Phi\left(\frac{c' - \mu_2}{\sigma_2}\right) > \Phi\left(\frac{\mu_1 - c'}{\sigma_1}\right),$$

it follows that

$$\text{EMCR}' > \eta\Phi\left(\frac{\mu_1 - c'}{\sigma_1}\right) + (1 - \eta)\Phi\left(\frac{c' - \mu_2}{\sigma_2}\right).$$

Then, as before, to prove that EMCR′ > EMCR it suffices to show that

$$(1 - \eta)\int_c^{c'} \phi_{\mu_2,\sigma_2}(x)\,\mathrm{d}x > \eta\int_c^{c'} \phi_{\mu_1,\sigma_1}(x)\,\mathrm{d}x.$$

This is true because for $\eta < \eta^{**}$, the point of intersection of $\eta\phi_{\mu_1,\sigma_1}(x)$ and $(1 - \eta)\phi_{\mu_2,\sigma_2}(x)$ is less than m as seen from (5.4). Since $\eta < \eta^{**} < \frac{1}{2}$ and $c', c > m$, for $c \leqslant x \leqslant c'$ we have $(1 - \eta)\phi_{\mu_2,\sigma_2}(x) > \eta\phi_{\mu_1,\sigma_1}(x)$. □

Remark 5. As Fig. 5 shows, the EMCR of the K-means algorithm continues to decrease past $\eta = 0.3333$ achieving a minimum at $\eta = 0.7628$ (determined numerically) and then increases rather steeply to 0.5 for $\eta = 1$. The EMCR of the K-means algorithm is plotted in Fig. 5 as noted earlier.

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

6. Simulation study

In this section we compare the performances of the K-means algorithm and the MM method via simulation. The study is restricted $K = 2$ clusters. The EMCR of the Bayes rule (the "gold standard") is used as a benchmark for comparison. (Note that because of the finite sample sizes used in the simulation study, the MCR of the MM method will be generally higher than that of the Bayes rule. The two converge asymptotically.) The *empirical* MCR of any rule is given by the observed proportion of misclassifications.

6.1. Univariate normal, homoscedastic mixture ~~with $K = 2$ clusters~~

A mixture of two normal distributions with $\mu_1 = 1$, $\mu_2 = 3$ and $\sigma_1 = \sigma_2 = 1$, was simulated. Since the misclassification rates are symmetric about $\eta = \frac{1}{2}$, we varied η only from 0.10 to 0.50. Also we varied the sample sizes from 50 to 50,000. Because the empirical MCR has a larger variance when the sample size is small, we replicated small samples until their overall total equaled 50,000, and computed the average misclassification rates. Thus, the simulation run for $N = 50,000$ was replicated once, while that for $N = 50$ was replicated 1000 times.

For the EM algorithm, we set the initial estimates of the cluster means equal to 0.5 and 4. The common initial estimate of the cluster variances was set equal to the overall sample variance. Initial estimate of η was set equal to 0.50. The simulation results are shown in Table 1.

The following conclusions emerge from these simulations:

1. The MCR of the Bayes rule increases as η increases from 0.10 to 0.50 as shown in Proposition 1.
2. The performance of the K-means algorithm is significantly worse than that of the MM method when η is away from 0.50, but gets closer as η gets closer to 0.50.
3. The sample size has a significant effect on the MCR of the MM method. Generally, the MCR decreases as the sample size increases because more accurate estimates are obtained using the EM algorithm with larger samples.

Table 1
Simulated misclassification rates of the MM method and the K-means algorithm for the univariate homoscedastic case ($K = 2$, $\mu_1 = 1$, $\mu_2 = 3$, $\sigma_1 = \sigma_2 = 1$)

η	EMCR of Bayes rule	N	Empirical MCR	
			MM method	K-means algorithm
0.10	0.0701	50	0.1299	0.3255
		500	0.0800	0.3415
		5000	0.0698	0.3552
		50 000	0.0710	0.3352
0.20	0.1121	50	0.1588	0.2415
		500	0.1202	0.2452
		5000	0.1133	0.2327
		50 000	0.1144	0.2351
0.30	0.1387	50	0.1851	0.1861
		500	0.1422	0.1872
		5000	0.1391	0.1865
		50 000	0.1381	0.1869
0.40	0.1538	50	0.1984	0.1686
		500	0.1601	0.1640
		5000	0.1561	0.1636
		50 000	0.1537	0.1644
0.50	0.1587	50	0.2077	0.1626
		500	0.1658	0.1622
		5000	0.1581	0.1577
		50 000	0.1589	0.1570

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

Table 2
Simulated misclassification rates of the MM method and the K-means algorithms for the univariate heteroscedastic case ($K = 2$, $\mu_1 = 1$, $\sigma_1 = 1$, $\mu_2 = 4$, $\sigma_2 = 2$)

η_1	EMCR of Bayes rule	N	Empirical MCR	
			MM method	K-means algorithm
0.10	0.1000	50	0.1789	0.3830
		500	0.1449	0.3881
		5000	0.1039	0.3891
		50 000	0.1003	0.3833
0.20	0.1450	50	0.2082	0.3075
		500	0.1682	0.3129
		5000	0.1491	0.3097
		50 000	0.1451	0.3075
0.30	0.1575	50	0.2288	0.2576
		500	0.1778	0.2568
		5000	0.1606	0.2587
		50 000	0.1563	0.2571
0.40	0.1561	50	0.2202	0.2105
		500	0.1701	0.2115
		5000	0.1584	0.2044
		50 000	0.1555	0.2131
0.50	0.1461	50	0.2057	0.1712
		500	0.1599	0.1734
		5000	0.1452	0.1705
		50 000	0.1462	0.1709
0.60	0.1295	50	0.1845	0.1402
		500	0.1356	0.1380
		5000	0.1318	0.1364
		50 000	0.1299	0.1362
0.70	0.1073	50	0.1513	0.1161
		500	0.1139	0.1096
		5000	0.1097	0.1096
		50 000	0.1069	0.1083
0.80	0.0795	50	0.1241	0.1180
		500	0.0840	0.0893
		5000	0.0799	0.0841
		50 000	0.0801	0.0836
0.90	0.0456	50	0.0917	0.2025
		500	0.0488	0.1619
		5000	0.0456	0.1472
		50 000	0.0450	0.1519

The MCR of the K-means algorithm is relatively unaffected by the sample size since it does not involve estimation of any parameters. When η is close to 0.5, the MCR of the MM method for small sample sizes can be sometimes higher than that of the K-means algorithm because of poor parameter estimates.

6.2. Univariate normal, heteroscedastic mixtures

Mixtures of two normal distributions with $\mu_1 = 1$, $\sigma_1 = 1$ and $\mu_2 = 4$, $\sigma_2 = 2$, were simulated. Because of unequal variances, the misclassification rates are not symmetric about $\eta = \frac{1}{2}$. Therefore, we varied η over its entire range from 0.10 to 0.90. We also varied the sample sizes from 50 to 50,000 as explained before.

Please cite this article as: Qiu, D., Tamhane, A.C., A comparative study of the K-means algorithm and the normal mixture model for clustering:.... J. Statist. Plann. Inference (2007), doi: 10.1016/j.jspi.2007.03.045

For the EM algorithm we set the initial estimates of the cluster means equal to 0.5 and 4.5, and initial estimates of the variances equal to 2.5 and 0.5, respectively. Initial η was set equal to 0.50. The results are shown in Table 2.

1. Many of the conclusions are qualitatively similar to those obtained in the univariate, homoscedastic case. For example, the performance of the K-means algorithm is relatively unaffected by the sample size, but that of the MM method is generally affected with higher empirical MCR values for small sample sizes that approach the EMCR of the Bayes rule as the sample size increases. The performance of the K-means algorithm gets progressively better as the mixing proportions become more balanced. The K-means algorithm beats the MM method in terms of the empirical MCR only when the sample size is small and η does not take extreme values (e.g., for $N = 50$, η is between 0.40 and 0.80 and for $N = 500$, $\eta = 0.7$). However, recall the caution expressed in Remark 3.
2. The EMCR of the Bayes rule, although not symmetric about $\eta = \frac{1}{2}$, shows a similar behavior, increasing with η up to $\overline{\eta} = 0.3358$ and then decreasing. The empirical MCR of the K-means algorithm, on the other hand, decreases until about $\eta = 0.8$ (more accurately until $\eta = 0.7628$) and then increases.

Finally, we note that all simulation results are in agreement with the analytical results derived in Sections 4 and 5.

7. Discussion and conclusions

In this paper we have analyzed the univariate case in thorough detail. The results show that the MM method is a preferred method in many cases for clustering since it yields smaller misclassification rates. Exceptions are those cases where the prior probabilities of the two clusters are not too different and the sample sizes are small. The EM algorithm is computationally more intensive and requires larger sample sizes to obtain accurate estimates of parameters. For multivariate data there are many more parameters that need to be estimated because of the $m(m+1)/2$ elements of the covariance matrix for each cluster. Therefore, the computational burden and the requirement of sample sizes increases with the dimension of the data vectors. However, the greater achieved accuracy is a worthwhile payoff especially in data mining applications where large sample sizes are prevalent and computing time is not a major restriction.

Analytical results for the multivariate case are much more difficult to obtain. However, for the homoscedastic case, the clustering problem can be shown to reduce to the univariate problem because the assignment rule is linear as seen in (3.8). We plan to study this problem in a future paper.

~~8. Uncited reference~~

~~McLachlan and Peel (2000).~~

Acknowledgment

We acknowledge helpful discussions with Professor Bruce Ankenman about the simulation results. We thank an anonymous referee for several helpful suggestions, especially Remark 3.

References

Anderson, T.W., 1958. An Introduction to Multivariate Statistical Analysis. Wiley, New York.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. Ser. B 39, 1–38.
Everitt, B.S., 1993. Cluster Analysis. third ed. Halsted Press, New York.
Hastie, T., Tibshirani, R., Friedman, J., 2002. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York.
Johnson, N.L., Kotz, S., 1970. Continuous Univariate Distributions-2. Wiley, New York.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations, Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley. pp. 281–297.
McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.
~~McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.~~