Mathematical Biosciences and Engineering

MBE, 17(1): 250–265.
DOI: 10.3934/mbe.2020014
Received: 23 May 2019
Accepted: 05 September 2019
Published: 30 September 2019

http://www.aimspress.com/journal/MBE

Research article

Fuzzy Gaussian Lasso clustering with application to cancer data

Miin-Shen Yang* and Wajid Ali

Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan

* **Correspondence**: Email: msyang@math.cycu.edu.tw.

Abstract: Recently, Yang et al. (2019) proposed a fuzzy model-based Gaussian (F-MB-Gauss) clustering that combines a model-based Gaussian with fuzzy membership functions for clustering. In this paper, we further consider the F-MB-Gauss clustering with the least absolute shrinkage and selection operator (Lasso) for feature (variable) selection, termed a fuzzy Gaussian Lasso (FG-Lasso) clustering algorithm. We demonstrate that the proposed FG-Lasso is a good clustering algorithm with better choice for feature subset selection. Experimental results and comparisons actually present these good aspects of the proposed FG-Lasso clustering algorithm. Cancer is a disease with growth of abnormal cells in a body. WHO reported that it is the first or second main leading cause of death. It spreads and affects the other parts of body if there is not properly diagnosed. In the paper, we apply the proposed FG-Lasso to cancer data with good feature selection and clustering results.

Keywords: fuzzy sets; model-based clustering; fuzzy model-based Gaussian; feature selection; Lasso; Fuzzy Gaussian Lasso (FG-Lasso) clustering

1. Introduction

Clustering is an unsupervised learning technique to divide data into similar groups/clusters. It has real applications in different areas such as biology, agriculture, economics, intelligent system, medical data and imaging [1–4]. It is a branch of multivariate analysis and briefly divided into two categories: non-parametric approaches and (probability) model-based clustering [5]. In non-parametric approaches, prototype-based clustering algorithms, such as k-mean [6], fuzzy c-means [7,8] and possibilistic c-means [9,10] are most used methods. In 1977, Dempster et al. [11] first proposed a probability mixture-model likelihood approach to clustering via the expectation and maximization (EM) algorithm. To

consider variable selection, Pan and Shen [12] combined EM [11] with the idea of least absolute shrinkage and selection operator (Lasso) [13].

In 1993, Banfield and Raftery [14] first proposed a so-called model-based Gaussian clustering to overcome the drawbacks of existing classification maximum likelihood approaches [15,16]. Banfield and Raftery [14] utilized eigenvalue decomposition for covariance matrix so that they can assign which feature to be common to all clusters, and which feature to be different between clusters for the model-based Gaussian clustering. It was widely applied in various areas, such as image segmentation [17], gene expression data [18], and background subtraction [19]. Recently, Yang et al. [20] proposed a fuzzy model-based Gaussian (F-MB-Gauss) clustering that combines the model-based Gaussian [14] with fuzzy membership functions [21,22] for clustering. However, F-MB-Gauss [20] treats data points with feature (variable) components under equal importance, and so it cannot distinguish these irrelevant feature components. In general, there exist some irrelevant features in a data set that may cause bad performance for clustering algorithms. In this paper, we further consider the F-MB-Gauss clustering with a Lasso penalty term. We then propose a fuzzy Gaussian Lasso (FG-Lasso) clustering algorithm. The proposed FG-Lasso algorithm becomes a clustering algorithm fitted for feature selection.

Medical data with gene expression in bioinformatics is an emerging systematic biological study. It is a discipline by combining biology, computer science, information engineering, mathematics and statistics to have better interpretation of data [23,24]. Bioinformatics is closely related to computational molecular biology, In a broad sense, computational biology covers all scientific operations related with biology that involve mathematics, computation, statistics, and algorithmic methods [25,26]. Genes/features selection is a significant task in bioinformatics due to having many irrelevant genes/features, and so discarding theses irrelevant genes/features may largely enhance clustering results and is suitable for further statistical/mathematical or any other treatment to get better results. Cancer is a disease in which WHO reported it is the first or second main leading cause of death. Cancer data are important medical data. Thus, to retain only relevant features is one of significant tasks and issues for researchers, especially in cancer data.

Since the proposed FG-Lasso algorithm is good for feature selection, we apply the FG-Lasso for cancer data, especially for feature selection. It is seen that the proposed FG-Lasso can perform both feature selection and regularization to increase accuracy and interpretability of clustering. It is also a good choice for high dimensional data set. The rest of the paper is organized as follows. In Section 2, we briefly review the F-MB-Gauss clustering and then propose the FG-Lasso algorithm. In Section 3, we present numerical results of the FG-Lasso clustering algorithm. In Section 4, we apply FG-Lasso for cancer data with feature selection. Conclusions are stated in Section 5.

2. Fuzzy Gaussian Lasso clustering algorithm

Model-based clustering is an essential technique to pertain data into similar and dissimilar groups/clusters by using mixtures of probability distributions. The model-based Gaussian clustering was initially proposed by Banfield and Raftery [14] to extend the classification maximum likelihood of Scott and Symons [15] and Symons [16]. Let a data set $X = \{x_1,..., x_n\}$ be a random sample from a *d*-variate Gaussian mixture with Gaussian distributions

$N\left(x;\mu_k,\Sigma_k\right)=(2\pi)^{-d/2}\,|\Sigma_k|^{-1/2}\exp(-(1/2)(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k))$. Let $P=\{P_1,...,P_c\}$ be a hard c-partition on X, where $\{P_1,...,P_C\}$ is equivalent to indicator functions $\{z_1,...,z_C\}$ with $z_k(x)=1$ as $x\in P_k$, and $z_k(x)=0$ otherwise. The objective function of the model-based Gaussian is given by $J(z,\theta)=\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}\ln f_k(x_i;\theta_k)$, where $f_k(x_i;\theta_k)=N\left(x_i;\mu_k,\Sigma_k\right)$, and $z_{ki}=z_k(x_i)$ is the membership function with $z_{ki}\in\{0,1\}$. The model-based Gaussian clustering algorithm is iterated by using the necessary conditions for maximizing the objective function $J(z,\theta)$ with $\hat{\mu}_k=\sum_{i=1}^{n}z_{ki}x_i\Big/\sum_{i=1}^{n}z_{ki}$ and $\hat{\Sigma}_k=\sum_{i=1}^{n}z_{ki}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^{\mathrm{T}}\Big/\sum_{i=1}^{n}z_{ki}$.

Zadeh [21] proposed fuzzy sets in 1965. Afterwards, Ruspini [27] extended the indicator functions $\{z_1,...,z_C\}$ to allow the membership $z_k(x)$ to be in the interval [0,1] with $\sum_{k=1}^{c}z_k(x)=1$ for all $x\in X$. These extended membership functions $\{z_1,...,z_C\}$ are called fuzzy c-partition. Recently, Yang et al. [20] proposed a fuzzy model-based Gaussian (F-MB-Gauss) clustering by combining the model-based Gaussian with fuzzy membership functions. The F-MB-Gauss objective function is as follows [20]:

$$J(z,\theta)=\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}^{m}\ln f(x;\theta)=\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}^{m}\ln N\left(x_i;\mu_k,\Sigma_k\right)$$

where z_{ki} is a fuzzy c-partition with the condition $\sum_{k=1}^{c}z_{ki}=1,\forall i$ and m is a fuzziness index with $m>1$ that determines the fuzziness level of clusters. However, the fuzziness index m may influence clustering results. To avoid m, the entropy term $-w\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}\ln z_{ki}$ of membership functions is added. Thus, the objective function becomes as

$$J(z,\mu,\Sigma)=\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}\ln N\left(x_i;\mu_k,\Sigma_k\right)-w\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ki}\ln z_{ki}$$

where $w\geq 0$ is a parameter whose value is determined by a suitable decreasing learning rate, such as 0.999^t, $e^{-t/100}$, $e^{-t/10}$, or e^{-t}. In Yang et al. [20], they considered the decreasing learning rate for w with $w^{(t)}=0.999^t$. We adopt it in this paper.

Although F-MB-Gauss [20] presents good clustering results for data sets, it always treats feature components of data points with equal importance. There exist some irrelevant features in most data sets that always affect performance of clustering algorithms with bad clustering results. However, the F-MB-Gauss cannot distinguish these irrelevant feature components. In this paper, we further study the F-MB-Gauss to have the algorithm be able to find out these irrelevant feature components. We use the idea of Lasso (least absolute shrinkage and selection operator) that was first proposed by Tibshirani [13] as variable selection in regression models. Note that Witten and Tibshirani [28] first proposed a feature selection framework using sparse clustering, where they use Lasso constraints of feature weights to shrink features toward 0 as feature selection. Witten and Tibshirani [28] defined the sparse clustering as the optimization of $\max_{w,\Theta} \sum_{j=1}^{d} w^T \Theta$ subject to $\|w\|_1 \le s,\ \|w\|^2 \le 1, w_j \ge 0, \forall j$, where $w = (w_1, w_2, \cdots, w_d) \in R^d$ are feature weights, and s is L_1 bound of w. They proposed the sparse k-means clustering by replacing the optimization of the k-means objective function. Castro and Pu [29] further proposed a simple approach to sparse k-means clustering based on the framework of Witten and Tibshirani [28]. Qiu et al. [30] extended the sparse k-means clustering to a sparse fuzzy c-means algorithm, and more recently, Chang et al. [31] proposed another sparse fuzzy c-means algorithm by extending the framework of Witten and Tibshirani [28] to $L_q (0 < q \le) - norm$ regularization for shrinking irrelevant feature weights to 0. However, all of these clustering algorithms for feature selection are based on Lasso constraints of feature weights. For the F-MB-Gauss clustering with Gaussian mixture distributions, it is no way in considering feature weights. However, we can use mean components μ_{kp} with $\lambda \sum_{k=1}^{c} \sum_{p=1}^{d} |\mu_{kp}|$. Thus, we consider the F-MB-Gauss with a Lasso penalty term, and then propose a fuzzy Gaussian Lasso (FG-Lasso) clustering algorithm. The FG-Lasso objective function is as follows:

$$J_{\text{FG-Lasso}}(z, \mu, \Sigma) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ki} \ln N(x_i; \mu_k, \Sigma_k) - w \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ki} \ln z_{ki} - \lambda \sum_{k=1}^{c} \sum_{p=1}^{d} |\mu_{kp}| \tag{1}$$

where $\lambda \ge 0$ is the regularization parameter that manages the amount of shrinkage and $\mu_k^T = (\mu_{k1}, \cdots, \mu_{kd})$, $N(x_i; \mu_k, \Sigma_k) = (2\pi)^{-d/2} |\Sigma_k|^{-1/2} \exp(-(1/2)(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k))$. The parameter λ can be used as a feature selection threshold. When the values of λ are increasing, more irrelevant features will be discarded. Of course, as $\lambda = 0$, the FG-Lasso becomes the F-MB-Gauss.

To obtain the necessary conditions for minimizing the FG-Lasso objective function $J_{\text{FG-Lasso}}(z, \mu, \Sigma)$, we use the Lagrangian as follows:

$$\tilde{J}_{\text{FG-Lasso}}(z, \mu, \Sigma) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ki} \ln N(x_i; \mu_k, \Sigma_k) - w \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ki} \ln z_{ki} - \lambda \sum_{k=1}^{c} \sum_{p=1}^{d} |\mu_{kp}| - \tau_1 \left(\sum_{k=1}^{c} z_{ki} - 1 \right)$$

Differentiating $\tilde{J}_{\text{FG-Lasso}}(z, \mu, \Sigma)$ with respect to the fuzzy membership function z_{ki} and setting it to be zero, we get the updating equation for z_{ki} as follows:

$$\hat{z}_{ki} = \frac{\left[N(x_i;\mu_{ki},\Sigma_k)\right]^{1/w}}{\sum_{s=1}^{c}\left[N(x_i;\mu_{ki},\Sigma_k)\right]^{1/w}} \tag{2}$$

To obtain the updating equation of μ_{kp} by differentiating $\tilde{J}_{\text{FG-Lasso}}(z,\mu,\Sigma)$ with respect to μ_{kp}, we only consider the case of Σ_k=Σ=$diag(\sigma_p^2), p=1,\cdots,d$. Thus, we can obtain that $\frac{\partial \tilde{J}_{\text{FG-Lasso}}(z,\mu,\Sigma)}{\partial \mu_{kp}} = \frac{\sum_{i=1}^{n} z_{ki}(x_i-\mu_{kp})}{\sigma_p^2} - \lambda sign(\mu_{kp})$. By using direct inspection of the FG-Lasso objective function $J_{\text{FG-Lasso}}(z,\mu,\Sigma)$, we can get the following solution for $\hat{\mu}_{kp}$:

$$\hat{\mu}_{kp} = \begin{cases} \tilde{\mu}_{kp} + \dfrac{\lambda\hat{\sigma}_p^2}{\sum_{i=1}^{n}\hat{z}_{ki}} & , if\ \tilde{\mu}_{kp} < -\dfrac{\lambda\hat{\sigma}_p^2}{\sum_{i=1}^{n}\hat{z}_{ki}} \\ 0 & , if\ \left|\tilde{\mu}_{kp}\right| \le \dfrac{\lambda\hat{\sigma}_p^2}{\sum_{i=1}^{n}\hat{z}_{ki}} \\ \tilde{\mu}_{kp} - \dfrac{\lambda\hat{\sigma}_p^2}{\sum_{i=1}^{n}\hat{z}_{ki}} & , if\ \tilde{\mu}_{kp} > \dfrac{\lambda\hat{\sigma}_p^2}{\sum_{i=1}^{n}\hat{z}_{ki}} \end{cases} \tag{3}$$

with

$$\tilde{\mu}_{kp} = \frac{\sum_{i=1}^{n}\hat{z}_{ki}x_{ip}}{\sum_{i=1}^{n}\hat{z}_{ki}} \tag{4}$$

where $\tilde{\mu}_{kp} = \sum_{i=1}^{n}\hat{z}_{ki}x_{ip} \Big/ \sum_{i=1}^{n}\hat{z}_{ki}$ is the maximum likelihood estimator (MLE) of the Gaussian means. As we increase the value of λ in Eq (3), it should have some $\hat{\mu}_{kp}$ = 0, otherwise it has the amount $\lambda\hat{\sigma}_p^2 \Big/ \sum_{i=1}^{n}\hat{z}_{ki}$ of shrinkage. Thus, during clustering processes, if $\left|\tilde{\mu}_{kp}\right| \le \lambda\hat{\sigma}_p^2 \Big/ \sum_{i=1}^{n}\hat{z}_{ki}$, then $\hat{\mu}_{kp}$ = 0. Otherwise $\hat{\mu}_{kp} = \tilde{\mu}_{kp} - \lambda\hat{\sigma}_p^2 \Big/ \sum_{i=1}^{n}\hat{z}_{ki}$. Note that $\tilde{\mu}_{kp} = \sum_{i=1}^{n}\hat{z}_{ki}x_{ip} \Big/ \sum_{i=1}^{n}\hat{z}_{ki}$ is the MLE of the normal mean μ_{kp}. However, the updating Eq (3) for $\hat{\mu}_{kp}$ represents the contribution of the *p*th feature to the cluster *k* through the regularization parameter λ. If $\hat{\mu}_{kp} = 0$ for all *k*, then the *pth* feature has no contribution to clustering that is non-informative and then discarded. This is why we consider the Lasso penalty $\lambda\sum_{k=1}^{c}\sum_{p=1}^{d}\left|\mu_{kp}\right|$ to the F-MB-Gauss objective function such that it becomes the FG-Lasso objective function (1). Thus, the FG-Lasso algorithm can have a behavior of feature selection.

To drive the updating Eq (3) of $\hat{\mu}_{kp}$, we use the FG-Lasso objective function $J_{\text{FG-Lasso}}(z,\mu,\Sigma)$. Taking the derivative of $J_{\text{FG-Lasso}}(z,\mu,\Sigma)$ with respect to μ_{kp}, we get $\frac{\partial J_{\text{FG-Lasso}}(z,\mu,\Sigma)}{\partial \mu_{kp}} = \frac{\sum_{i=1}^{n} z_{ki}(x_i-\mu_{kp})}{\sigma_p^2} - \lambda sign(\mu_{kp})$. Set it to be 0, and after simplification, we get $\sum_{i=1}^{n} \hat{z}_{ki}\sigma_p^{-2}(x_{ip}-\mu_{kp}) - \lambda sign(\mu_{kp}) = 0$, and have $\sum_{i=1}^{n} \hat{z}_{ki}x_{ip} = \lambda\sigma_p^2 sign(\mu_{kp}) + \sum_{i=1}^{n} \hat{z}_{ki}\mu_{kp}$. Dividing both side by $\sum_{j=1}^{n} \hat{z}_{kj}$, we obtain the Eq of $\hat{\mu}_{kp}$ with $\hat{\mu}_{kp} = \frac{\sum_{i=1}^{n} \hat{z}_{ki}x_{ip}}{\sum_{i=1}^{n} \hat{z}_{ki}} - \frac{\lambda\sigma_p^2 sign(\hat{\mu}_{kp})}{\sum_{i=1}^{n} \hat{z}_{ki}}$. Thus, we have $\hat{\mu}_{kp} = \tilde{\mu}_{kp} - \frac{\lambda\sigma_p^2 sign(\hat{\mu}_{kp})}{\sum_{i=1}^{n} \hat{z}_{ki}}$. As we know that $|\hat{\mu}_{kp}|$ is not differentiable at $\hat{\mu}_{kp}$ = 0, and so we need to cope up this problem by using subderivative or subgradient of a convex function. In case of the absolute value function $f(x) = \lambda|x|$, the subgradient or subderivative is defined by $\partial f(x) = \begin{cases} \{-\lambda\} & if\ x<0 \\ [-\lambda,\lambda] & if\ x=0 \\ \{+\lambda\} & if\ x>0 \end{cases}$. To see further explanation about the soft threshold operator λ for regression models, Hastie et al. [32] is a good reference. Thus, we can obtain the updating Eq (3) for $\hat{\mu}_{kp}$ in our case.

Similarly, for the common diagonal covariance matrix $\Sigma_k = \Sigma = diag(\sigma_p), p = 1,\cdots,d$, differentiating $\tilde{J}_{\text{FG-Lasso}}(z,\mu,\Sigma)$ with respect to $\sigma_p^2, p = 1,\cdots,d,$ we can get the following updating Eq:

$$\hat{\sigma}_p^2 = \frac{\sum_{k=1}^{c}\sum_{i=1}^{n} z_{ki}(x_{ip}-\tilde{\mu}_{kp})^2}{\sum_{k=1}^{c}\sum_{i=1}^{n} \hat{z}_{ki}} \tag{5}$$

Due to the expected singularity of matrix when the cluster number is large in the FG-Lasso algorithm, we use the following condition to overcome this problem:

$$\tilde{\sigma}_p^2 = (1-\gamma)\sigma_p^2 + \gamma\omega \tag{6}$$

where γ is a small positive number and ω is a diagonal matrix with a small positive number. Here we use γ = 0.0001, $\omega = d_{\min}^2$, $I_d = \min\{d_{ij}^2 = \|x_i - x_j\|^2 > 0, 1 \le i, j \le n\}$. For w, we use the same decreasing learning rate as Yang et al. [20] with

$$w^{(t)} = 0.999^t \tag{7}$$

Thus, the proposed FG-Lasso clustering algorithm can be summarized as follows.

FG-Lasso Algorithm

Step 1: Fix $\varepsilon > 0$. Give initials for μ and $\sigma_p^{2,(0)}$. Set λ = 1 and $t = 1$.

Step 2: Compute $\hat{z}_{kj}^{(0)}$ by using Eq (2).

Step 3: Compute $\tilde{\mu}_{kp}^{(t)}$ by using Eq (4).

Step 4: Compute $w^{(t)}$ by using Eq (7).

Step 5: Update $\hat{\sigma}_p^{2,(t)}$ with $\tilde{\mu}_{kp}^{(t)}$ and $\hat{z}_{kj}^{(t-1)}$ by Eqs (5) and (6).

Step 6: Update $\hat{z}_{kj}^{(t)}$ with $\tilde{\mu}_{kp}^{(t)}$, $w^{(t)}$ and $\hat{\sigma}_p^{2,(t)}$ by using Eq (2).

Step 7: Update $\tilde{\mu}_{kp}^{(t+1)}$ with $\hat{z}_{kj}^{(t)}$ by using Eq (4).

If $\max \left\| \tilde{\mu}_{kp}{}^{(t+1)} - \tilde{\mu}_{kp}{}^{(t)} \right\| < \varepsilon$ stop.

Else $t = t + 1$ and return to Step 3.

Step 8: Update $\hat{\sigma}_p^{2,(t+1)}$ with $\tilde{\mu}_{kp}^{(t+1)}$ and $\hat{z}_{kj}^{(t)}$ by using Eqs (5) and (6).

Step 9: Update $\hat{\mu}_{kp}^{(t)}$ with $\hat{z}_{kj}^{(t)}$, $\tilde{\mu}_{kp}^{(t+1)}$ and $\hat{\sigma}_{kp}^{2,(t+1)}$ by using Eq (3), that is,

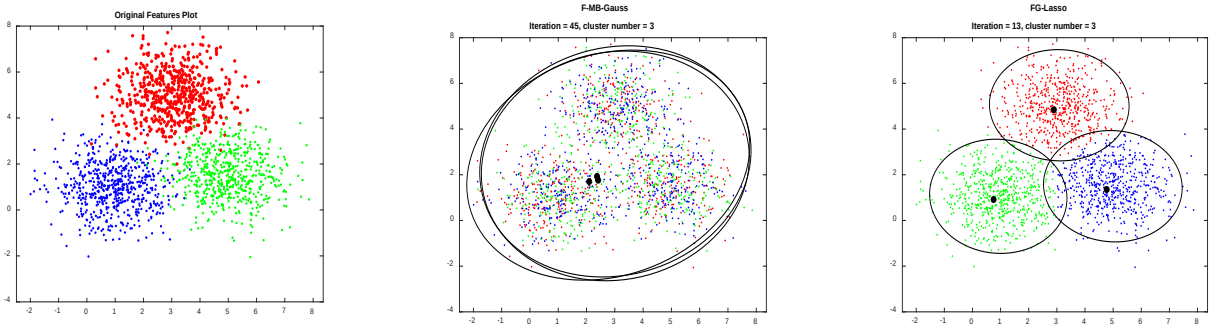
If $\left| \tilde{\mu}_{kp}^{(t+1)} \right| \le \dfrac{\lambda \hat{\sigma}_p^{2,(t+1)}}{\sum_{j=1}^{n} \hat{z}_{kj}^{(t)}}$, then let $\hat{\mu}_{kp}^{(t)} = 0$.

Else $\hat{\mu}_{kp}^{(t)} = \tilde{\mu}_{kp}^{(t+1)} - \dfrac{\lambda \hat{\sigma}_p^{2,(t+1)}}{\sum_{j=1}^{n} \hat{z}_{kj}^{(t)}}$.

Step 10: Increase λ and return to Step 3, or output results.

3. Experimental results using numerical and real data

In this section, we demonstrate the performance of the proposed FG-Lasso clustering algorithm. Several synthetic and real data sets are used to have more insights to the feature selection behaviors of the FG-Lasso algorithm. We also give the comparisons of the proposed FG-Lasso with F-MB-Gauss [20]. The accuracy rate (*AR*) is used as a criterion for evaluating the performance of a clustering algorithm. *AR* is the percentage of data points that are correctly identified by the clustering algorithm in which *AR* is defined as $AR = \sum_{i=1}^{k} r_i \big/ n$, where r_i is the number of points in $C_i^{'}$ that are also in C_i in which $C = \{C_1, C_2, \cdots, C_c\}$ is the set of *c* clusters for the given data set and $C^{'} = \{C_1^{'}, C_2^{'}, \cdots, C_c^{'}\}$ is the set of *c* clusters generated by the clustering algorithm.



(a) (b) (c)

Figure 1. (a) The original 3-cluster Gaussian data set; (b) F-MB-Gauss results after 45 iteration; (c) FG-Lasso results after 13 iterations.

Table 1. Comparisons of F-MB-Gauss and FG-Lasso for 3-cluser Gaussian data.

c	F-MB-Gauss		FG-Lasso	
	d	AR	d*	AR
3	4	0.3501	2	0.961

Table 2. Features selection by FG-Lasso for 3-cluser Gaussian data.

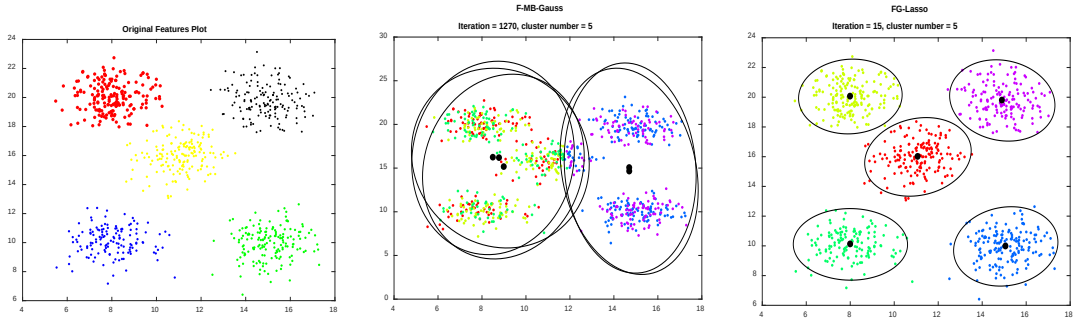
Features	λ	
	10	15
$feature_1$	-	-
$feature_2$	-	-
$feature_3$	$\hat{\mu}_{13} = \hat{\mu}_{23} = \hat{\mu}_{33} = 0$	×
$feature_4$	-	$\hat{\mu}_{14} = \hat{\mu}_{24} = \hat{\mu}_{34} = 0$

Example 1. In this example, a simulation data set is used to demonstrate the significance of the FG-Lasso algorithm, where the usefulness of λ for feature selection, especially to remove irrelevant features, is demonstrated. A data set, called 3-cluser Gaussian data, with having 1800 points are generated from a Gaussian mixture with $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$ where 600 points are from the normal distribution with $\mu_1 = (1\ 1)$ and $\Sigma_1 = (1\ \ 0; 0\ \ 1)$, 600 points are from the normal distribution with $\mu_2 = (3\ 5)$ and $\Sigma_2 = (1\ \ 0; 0\ \ 1)$ while the same size in the cluster three with $\mu_3 = (5\ 1.5)$ and $\Sigma_1 = (1\ \ 0; 0\ \ 1)$. We consider the two features, named as $feature_1$ and $feature_2$, as informative. We then add other two irrelevant features generated from the uniform distributions over the intervals $[-5,5]$ and $[-10,10]$, respectively, named as $feature_3$ and $feature_4$, that are considered as non-informative features. The original data set with two informative features $feature_1$ and $feature_2$ is

shown in Figure 1(a). Figure 1(b) represents the clustering results of the F-MB-Gauss algorithm with the 3 cluster centers after 45 iterations. The final clustering results of the FG-Lasso algorithm with the 3 clusters after 13 iterations are shown in Figure 1(c). Results of *ARs* from the FG-Lasso and F-MB-Gauss algorithms are shown in Table 1. It is seen that the proposed FG-Lasso is feasible for feature selection with the final feature number of d* = 2. However, the F-MB-Gauss algorithm cannot have feature selection and so it give the final feature number of d = 4. It is clearly that the irrelevant features of $feature_3$ and $feature_4$ actually distort the final clustering results for the F-MB-Gauss algorithm with d = 4 and average *AR* = 0.3501. However, the proposed FG-Lasso can discard these non-informative features of $feature_3$ and $feature_4$ with d* = 2 and a high average *AR* = 0.961. The details of discarded features as increasing the values of λ using FG-Lasso are shown in Table 2. When the value of λ is increasing as 10, we obtain $\hat{\mu}_{13} = \hat{\mu}_{23} = \hat{\mu}_{33} = 0$ and so the feature $feature_3$ is discarded. Similarly, when we increase the value of λ as 15, we obtain $\hat{\mu}_{14} = \hat{\mu}_{24} = \hat{\mu}_{34} = 0$ and, so the feature $feature_4$ is discarded. Thus, it successfully discards all irrelevant features of $feature_3$ and $feature_4$ when the value of λ is 15. From Table 2, we find that both features of $feature_1$ and $feature_2$ are informative, but features $feature_3$ and $feature_4$ are non-informative, and then discarded.

Example 2. We also use a simulation data set for the proposed FG-Lasso algorithm to demonstrate the significance of λ for feature selection, especially to remove irrelevant features. A data set, called 5-cluser Gaussian data, with 800 points are generated from a Gaussian mixture with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1/5$, $\mu_1 = (8\ \ 10)$, $\mu_2 = (8\ \ 20)$, $\mu_3 = (15\ \ 10)$, $\mu_4 = (15\ \ 20)$, and $\mu_5 = (11\ \ 16)$ with $\sum_1 = \sum_2 = \sum_3 = \sum_4 = \sum_5 = (1\ \ 0; 0\ \ 1)$. We consider the two features, named as $feature_1$ and $feature_2$, as informative. We then add one irrelevant feature generated from the uniform distributions over the intervals [0,65], named as $feature_3$, that is considered as a non-informative feature. The original data set with two informative features $feature_1$ and $feature_2$ is shown in Figure 2(a). Figure 2(b) represents the final clustering results of the F-MB-Gauss algorithm with 5 cluster centers after 1270 iterations. The final clustering results of the proposed FG-Lasso algorithm with 5 clusters after 15 iterations are shown in Figure 2(c). Results of average *ARs* from the FG-Lasso and F-MB-Gauss algorithms with 50 different initializations are shown in Table 3. It is seen that the proposed FG-Lasso is feasible for feature selection with the final feature number of d* = 2. It is clearly that irrelevant features actually distort the final clustering results for the F-MB-Gauss algorithm with average *AR* = 0.320 that cannot give a feature selection behavior with the final feature number of d = 3. However, the proposed FG-Lasso can discard the non-informative feature

$feature_3$ with a high average $AR = 0.810$. The details of discarded feature as increasing the values of λ using FG-Lasso are also shown in Table 4. When the values of λ is increasing as 80, we obtain $\hat{\mu}_{13} = \hat{\mu}_{43} = \hat{\mu}_{53} = 0$ and when we increase the value of λ is 105, we obtain $\hat{\mu}_{23} = \hat{\mu}_{33} = 0$. It is clearly the feature $feature_3$ is discarded. Thus, it successfully discards irrelevant feature $feature_3$ as the value of λ is 80 or 105. From Table 4, we find that both features $feature_1$ and $feature_2$ are informative, but feature $feature_3$ is non-informative, and then discarded.



(a) (b) (c)

Figure 2. (a) The original 5-cluster Gaussian data set; (b) F-MB-Gauss results after 1270 iteration; (c) FG-Lasso results after 15 iteratio.

Table 3. Comparisons of F-MB-Gauss and FG-Lasso for 5-cluser Gaussian data.

c	F-MB- Gauss		FG-Lasso	
	d	AR	d*	AR
5	3	0.320	2	0.810

Table 4. Features selection by FG-Lasso for 5-cluser Gaussian data.

Features	λ	
	80	105
$feature_1$	-	-
$feature_2$	-	-
$feature_3$	$\hat{\mu}_{13} = \hat{\mu}_{43} = \hat{\mu}_{53} = 0$	$\hat{\mu}_{23} = \hat{\mu}_{33} = 0$

Except the above two synthetic data sets, we also use a real data set, Pima indian, from UCI repository data [33].

Example 3 (Pima indian [33]). In this example, we consider the real data set of Pima indian [33]. This data set consists of 8 features, named as the number of times of Pregnant, Plasma glucose

concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, and Age (years), while class variable (Outcomes). The data set has two classes. By using F-MB-Gauss, we obtain the average AR = 0.45 when 30 different initializations are considered. As we increase the values of λ to λ = 50, the proposed FG-Lasso algorithm discards the features, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure, and 2-Hour serum insulin, with a higher average AR = 0.67. This shows the good aspect of the proposed FG-Lasso clustering algorithm for the Pima indian data set.

4. Application to cancer data

Cancer is uncontrolled growth of abnormal cells in a body found in a group of diseases [34]. According to the estimate from WHO (World Health Organization), it is the first or second main leading cause of death before 70 years in 91 out of 172 countries [35]. It spreads and affects the other parts of body if there is not properly diagnosed. This severe disease has many symptoms such as tumor, abnormal bleeding, long-term cough, more weight loss, etc. According to the Global Cancer Incidence, Mortality and Prevalence (GLOBOCAN), cancer has extended 36 types in which lung cancer is the most common diseases (11.6%) of the total cases in male and female [36]. Other alarming leading cancer diseases are breast cancer (11.6%), prostate cancer (7.1%), colorectal cancer (6.1%), stomach cancer (8.2%) and liver cancer (8.2%) [35,36]. Breast cancer is the most leading and commonly diagnosed cancer disease among females [36–38] and leading breast cancer issue in 154 out of 185 countries [38]. According to the WHO findings, there were 2.1 million newly women diagnosed breast cases in 2018. The highest statistics found countries are Australia/New-Zealand, United Kingdom, Sweden Finland, Denmark Belgium (Highest rate), the Netherlands and France. There are many risk factors of breast cancer like family history, physical activity, breast feeding, hormones intake, alcohol intake, greater weight and body fat [36,39]. There are three important techniques to diagnose breast cancer, namely as mammography, Fine Needle Aspirate (FNA) biopsy and surgical biopsy. To demonstrate the applicability of the proposed FG-Lasso clustering algorithm, we use the following three real cancer data sets, Breast cancer Wisconsin, Colon tissues, and Leukemia data sets. We implement the FG-Lasso and F-MB-Gauss algorithms to the three cancer data sets and compare their results.

Example 4 (Breast Cancer Wisconsin)**.** Breast cancer Wisconsin data was created in Street et al. [40]. This data consists of 569 FNA with 212 malignant (patients sample) and 357 benign (healthy samples) [36,40]. The 30 attributes are computed from a digitized image of a FNA of breast mass. They describe characteristics of the cell nuclei presented in the image. Real values are computed from each cell nucleus, namely as radius (from center to points on the perimeter), texture (gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 /area-1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension ("coastline approximation"-1), where each one in the 10 features has three components as mean, standard error and worst. That has 30 features. The FG-Lasso and F-MB-Gauss algorithms are implemented on Breast cancer Wisconsin where clustering results are shown in Table 5. From Table 5, it is seen that the F-MB-Gauss algorithm obtains a good accuracy rate with d = 30 and AR = 0.905 when 30 different initializations are considered. Using the same 30 initializations, we also implement

the proposed FG-Lasso algorithm on the data set. There are 27 from 30 features to be selected with d* = 27 and a little better accuracy rate of AR = 0.925 when the λ value is 12, where the features of area mean, area standard error and area worst are considered as unimportant features and then discarded.

Table 5. Comparisons of F-MB-Gauss and FG-Lasso for Breast cancer Wisconsin data.

c	F-MB-Gauss		FG-Lasso	
	d	AR	d*	AR
2	30	0.905	27	0.925

Next, we implement the FG-Lasso and F-MB-Gauss algorithms to the colon tissue data set [41].

Example 5 (Colon Tissues). The Colon tissue data set consists of colon cancer which contains 62 samples from the microarray experiments of colon tissues samples with 2000 genes and two classes (40 tumor tissues and 22 normal tissues) [41]. Colorectal cancer is the kind of cancer that starts tissue or tumor growth on the inner lining of the colon. It is the third incidence of death and second in term of mortality in the world [41]. Most suffered in colon cancer countries are Hungry, Slovenia, Netherlands, Norway, Australia New Zealand, North America, Japan, Republic of Korea, and Singapore (in Females). Among these countries, Hungry is ranked first in males and Norway is ranked first in females, while highest colon incidence rates are found in Republic of Korea among males and Macedonia among females [36,42]. We first standardize the data set, and then apply the FG-Lasso and F-MB-Gauss algorithms to the Colon tissue data set where the clustering results are shown in Table 6. From Table 6, it is seen that the F-MB-Gauss algorithm obtains the average accuracy rate AR = 0.601 with 30 different initializations. When the proposed FG-Lasso algorithm is implemented on the data set with the same 30 different initializations, there are 1384 from 2000 features to be selected with the final feature number of d* = 1384 when the λ value is increasing to be 90 in which it obtains a better average accuracy rate AR = 0.653, as shown in Table 6. This demonstrates that the proposed FG-Lasso algorithm is significant for feature selection on the Colon tissue data set.

Table 6. Comparisons of F-MB-Gauss and FG-Lasso for Colon Tissue data.

c	F-MB-Gauss		FG-Lasso	
	d	AR	d*	AR
2	2000	0.601	1384	0.653

Finally, we implement the FG-Lasso and F-MB-Gauss algorithms to the Leukemia cancer data set. Leukemia is the type of cancer as the uncontrolled growth of hematopoietic stem cell in the bone marrow occurs. Leukemia is a well-known data that the type of cancer occurs as the uncontrolled growth of hematopoietic stem cell in the bone marrow. It is the most common in white male people and increase according to ages [43,44]. Broadly there are four subtypes of leukemia, acute lymphoblastic, acute myelogenous, chronic lymphocytic, and chronic myelogenous. Acute lymphoblastic leukemia (ALL) is most commonly found in children while the other three types occur

in adults. ALL causes fever, lethargy, bleeding, musculoskeletal pain or dysfunction. On the other hand, fever, fatigue, weight loss, bleeding or bruising are the most commonly symptoms of acute myelogenous leukemia (AML) [44]. According to the global cancer statistics [36], leukemia has 2.4% rate of new cases, and 3.2% of deaths rate worldwide in 2018.

Example 6 (Leukemia Data)**.** Leukemia data had originally considered by Golub et al. [45]. The data consists of 38 patients considered as observations each from leukemia patients with their biological sample array while 7129 genes are considered as features. Among these samples, 27 are acute lymphoblastic leukemia (ALL) and 11 are acute myelogenous leukemia (AML). Golub et al. [45] distinguished two types of patients due to their isolation clinical treatments. We only sort 2000 genes according to their variances and we also standardize data so each attribute has mean 0 and variance 1. We implement the FG-Lasso and F-MB-Gauss algorithms to the Leukemia data set where the clustering results are shown in Table 7. From Table 7, it is seen that the F-MB-Gauss algorithm obtains the low average accuracy rate AR = 0.393 with 30 different initializations. This shows that these irrelevant features in the Leukemia data set actually affects clustering results. When the proposed FG-Lasso algorithm is applied to the data set, there are 654 from 2000 features to be selected with the final feature number of d* = 654 when the λ value is increasing to be 350 in which it promotes the accuracy rate to AR = 0.615. That is, the proposed FG-Lasso algorithm is quite significant for feature selection on the Leukemia data set by selecting 654 from 2000 features.

Table 7. Comparisons of F-MB-Gauss and FG-Lasso for Leukemia data.

c	F-MB-Gauss		FG-Lasso	
	d	AR	d*	AR
2	2000	0.393	654	0.615

5. Conclusions

The F-MB-Gauss clustering proposed by Yang et al. [20] always treats feature components in data points with equal importance, and so it does not have a feature selection behavior. However, there generally exist irrelevant features in data that may badly affect the performance of clustering algorithms. In this paper, we extended the F-MB-Gauss clustering to the fuzzy Gaussian Lasso (FG-Lasso) using a Lasso penalty term of Gaussian means components. The FG-Lasso algorithm is then proposed for clustering data sets with feature selection. The proposed FG-Lasso has good behaviors with better choice for feature selection. Several experimental results and comparisons have actually demonstrated the feature selection aspect of the proposed FG-Lasso algorithm. According to the estimate from WHO, cancer is the first or second main leading cause of death. This severe disease has many symptoms such as tumor, abnormal bleeding, long-term cough, and more weight loss. Cancer data are important medical data where they are high dimensional and exist many irrelevant features. In this paper, we also apply the proposed FG-Lasso algorithm to the three cancer data, Breast cancer Wisconsin, Colon tissues, and Leukemia. According to clustering results, it is seen that the proposed FG-Lasso can select these important features with a higher accuracy as increasing of the threshold λ. However, our question is what value of the threshold λ should be to have an optimal number of features in the FG-Lasso clustering algorithm. That is, to find a good estimate for the threshold parameter λ should be important and will be our further research topic. On the other

hand, to consider a whole covariance matrix, not only a diagonal matrix, is also another problem in FG-Lasso, and it would be also our future work.

Acknowledgements

The authors would like to thank the anonymous referees for their helpful comments in improving the presentation of this paper.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, New York, 2009.
2. J. C. Bezdek, *Pattern Recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.
3. D. Jiang, C. Tang and A. Zhang, Cluster analysis for gene expression data: A survey, *IEEE Trans. Knowl. Data Eng.*, **16** (2004), 1370–1386.
4. J. M. T. Wu, C. W. Lin, P. Fournier-Viger, et al., The density-based clustering method for privacy-preserving data mining, *Math. Biosci. Eng.*, **16** (2019), 1718–1728.
5. M. S. Yang, C. Y. Lai and C. Y. Lin, A robust EM clustering algorithm for Gaussian mixture models, *Pattern Recognit.*, **45** (2012), 3950–3961.
6. A. K. Jain, Data clustering: 50 years beyond k-means, *Pattern Recognition Lett*., **31** (2010), 651–666.
7. A. Baraldi and P. Blonda, A survey of fuzzy clustering algorithms for pattern recognition-part I and part II, *IEEE Trans. Syst. Man Cybern. B*, **29** (1999), 778–785.
8. M. S. Yang and Y. Nataliani, Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters, *Pattern Recognit.*, **71** (2017), 45–59.
9. R. Krishnapuram and J. M. Keller, A possibilistic approach to clustering, *IEEE Trans. Fuzzy Syst.*, **1** (1993), 98–110.
10. M. S. Yang, S. J. Chang-Chien and Y. Nataliani, A fully-unsupervised possibilistic c-means clustering method, *IEEE Access*, **6** (2018), 78308–78320.
11. A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Stat. Soc. Ser. B*, **39** (1977), 1–38.
12. W. Pan and X. Shen, Penalized model-based clustering with application to variable selection, *J. Mach. Learn. Res.*, **8** (2007), 1145–1164.
13. R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B*, **58** (1996), 267–288.
14. J. D. Banfield and A. E. Raftery, Model-based Gaussian and non-Gaussian Clustering, *Biometrics*, **49** (1993), 803–821.
15. A. J. Scott and M. J. Symons, Clustering methods based on likelihood ratio criteria, *Biometrics*, **27** (1971), 387–397.

16. M. J. Symons, Clustering criteria and multivariate normal mixtures, *Biometrics*, **37** (1981), 35–43.
17. R. Wehrens, L. M. C. Buydens, C. Fraley, et al., Model-based clustering for image segmentation and large datasets via sampling, *J. Classif.*, **21** (2004), 231–253.
18. W. C. Young, A. E. Raftery and K. Y. Yeung, Model-based clustering with data correction for removing artifacts in geneexpression data, *Ann. Appl. Stat.*, **11** (2017), 1998–2026.
19. T. Akilan, Q. M. J. Wu and Y. Yang, Fusion-based foreground enhancement for background subtraction using multivariate multi-model Gaussian distribution, *Inf. Sci*., **430–431** (2018), 414–431.
20. M. S. Yang, S. J. Chang-Chien and Y. Nataliani, Unsupervised fuzzy model-based Gaussian clustering, *Inf. Sci.*, **481** (2019), 1–23.
21. L. A. Zadeh, Fuzzy sets, *Inf. Control*, **8** (1965), 338–353.
22. M. S. Yang and Y. Nataliani, A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy, *IEEE Trans. Fuzzy Syst.*, **26** (2018), 817–835.
23. K. Voevodski, M. F. Balcan, H. Röglin, et al., Active clustering of biological sequences, *J. Mach. Learn. Res.*, **13** (2012), 203–225.
24. D. Gaweł and K. Fujarewicz, On the sensitivity of feature ranked lists for large-scale biological data, *Math. Biosci. Eng.*, **10** (2013), 667–690.
25. J. Xiong, *Essential Bioinformatics*, Cambridge University Press, New York, 2006.
26. R. Jiang, X. Zhang, M. Q. Zhang, *Basics of Bioinformatics*, Springer-Verlag Berlin An, 2013.
27. E. H. Ruspini, A new approach to clustering, *Inf. Control*, **15** (1969), 22–32.
28. D. M. Witten and R. Tibshirani, A framework for feature selection in clustering, *J. Am. Stat. Assoc.*, **105** (2010), 713–726.
29. E. A. Castro and X. Pu, A simple approach to sparse clustering, *Comput. Stat. Data Anal.*, **105** (2017), 217–228.
30. X. Qiu, Y. Qiu, G. Feng, et al., A sparse fuzzy c-means algorithm base on sparse clustering framework, *Neurocomputing*, **157** (2015), 290–295.
31. X. Chang, Q. Wang, Y. Liu, et al., Sparse regularization in fuzzy c-means for high-dimensional data clustering, *IEEE Trans. Cybern.*, **47** (2017), 2616–2627.
32. T. Hastie, R. Tibshirani and M. Wainwright, Statistical Learning with Sparsity: The lasso and Generalization, Chapman and Hall/CRC press, New York, (2015).
33. C. L. Blake and C. J. Merz, UCI repository of machine learning database, a huge collection of artificial and real-world data sets, (1988).
34. N. K. Phan, Biological therapy: A new age of cancer treatment, *Biomed. Res. Ther.*, **1** (2014), 32–34.
35. *Global Health Observatory (GHO) data*, World Health Organization, Geneva, 2018. Available from: https://www.who.int/gho/en/.
36. F. Bray, J. Ferlay, I. Soerjomataram, et al., A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA A Cancer J. Clin*., **68** (2018), 394–424.
37. D. N. K. Boulos and R. R. Ghali, Awareness of breast cancer among female students at Ain Shams University, *Egypt, Glob. J. Health Sci.*, **6** (2014), 154–161.
38. K. McPherson, C. M. Steel and J. M. Dixon, Breast cancer—epidemiology, risk factors, and genetics, *BMJ*, **321** (2000), 624–628.

39. R. R. Janghel, A. Shukla, R. Tiwari, et al., *Intelligent decision support system for breast cancer*, International Conference in Swarm Intelligence, Beijing, China, 2010, 351–358. Available from: https://link_springer.gg363.site/chapter/10.1007/978-3-642-13498-2_46#citeas.
40. W. N. Street, W. H. Wolberg and O. L. Mangasarian, *Nuclear feature extraction for breast tumor diagnosis*, Biomedical image processing and biomedical visualization, **1905** (1993), 861–870. Available from: https://doi.org/10.1117/12.148698.
41. A. R. Marley and H. Nan, Epidemiology of colorectal cancer, *Int. J. Mol. Epidemiol. Genet.*, **7** (2016), 105–114.
42. M. Arnold, M. S. Sierra, M. Laversanne, et al., Global patterns and trends in colorectal cancer incidence and mortality, *Gut*, **66** (2017), 683–691.
43. *Cancer Stat Facts: Leukemia*, National Cancer Institute, Surveillance Epidemiology and End Results Program, 2006–2010. Available from: http://seer.cancer.gov/statfacts/html/leuks.html.
44. A. S. Davis, A. J. Viera and M. D. Mead, Leukemia: An overview for primary care, Am. Fam. Physician, **89** (2014), 731–738.
45. T. R. Golub, D. K. Slonim, P. Tamayo, et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286** (1999), 531–537.

AIMS Press