Mach Learn (2015) 98:7–30
DOI 10.1007/s10994-013-5338-7

A framework to uncover multiple alternative clusterings

Xuan Hong Dang · James Bailey

Received: 25 May 2012 / Accepted: 1 March 2013 / Published online: 22 March 2013

Abstract Clustering is often referred to as unsupervised learning which aims at uncovering hidden structures from data. Unfortunately, though widely being used as one of the principal tools to understand the data, most conventional clustering techniques are limited in achieving this goal since they only attempt to find a single clustering solution from the data. For many real-world applications, especially those being described in high dimensional data, it is common to see that the data can be grouped into different yet meaningful ways. This gives rise to the recently emerging research area of mining alternative clusterings. In this paper, we propose a framework named MACL that is capable of discovering multiple alternative clusterings from a given dataset. MACL seeks alternative clusterings in sequence and a novel solution is found by conditioning on all previously known clusterings. The framework takes a mathematically appealing approach by combining the maximum likelihood framework and mutual information. Consequently, its resultant clustering quality is achieved by the likelihood maximization over the data whereas the dissimilarity is ensured by the minimization over the information sharing amongst alternatives. We test the proposed algorithm on both synthetic and real-world datasets and the experimental results demonstrate its potential in discovering multiple alternative clusterings from data.

Keywords Unsupervised learning · Alternative clustering · Expectation maximization · Mutual information

Editors: Emmanuel Müller, Ira Assent, Stephan Günnemann, Thomas Seidl, and Jennifer Dy.

Majority of this work was done while the first author was with The University of Melbourne.

X.H. Dang (✉)
Department of Computer Science, Aarhus University, 8200 Aarhus N, Denmark
e-mail: dang@cs.au.dk

J. Bailey
Department of Computing and Information Systems, The University of Melbourne, Melbourne, VIC 3010, Australia
e-mail: baileyj@unimelb.edu.au

1 Introduction

Cluster analysis has long been identified as one of the core tasks in data mining. Many clustering techniques have been developed so far including k-means (Lloyd 1982; Arthur and Vassilvitskii 2007), hierarchical agglomerative clustering (Day and Edelsbrunner 1984), mixture densities (Dempster et al. 1977), spectral partitioning (Ng et al. 2001), and density-based clustering (Ester et al. 1996; Ankerst et al. 1999). Although it is common to produce only a single clustering from a given empirical data (that these algorithms have extensively focused on), it is observed in many cases that the data can be clustered along many different yet reasonable ways. For example, while most conventional work on text clustering widely attempts to classify documents according to the topics, it is conceivable that grouping them to writing styles is also valid and meaningful. Likewise, extensive research in bio-information has largely been focused on categorizing data proteins according to their structures, it is possible to see that grouping them by their functions is also useful. In both these applications and many other ones, one may see that the natural structure behind high dimensional data is not unique and there exist many different ways to interpret the data. Therefore, to further understand the data and to achieve the ultimate goal of data exploration of unsupervised clustering, there is a strong demand to devise novel techniques that are able to generate multiple different yet high qualitative clusterings from the data.

In addressing this problem, several algorithms have been developed in the literature and based on whether or not prior information is required during the clustering process, it is possible to classify them into two different approaches: unsupervised (Jain et al. 2008; Dang and Bailey 2010a; Niu et al. 2010) and semi-supervised (Gondek and Hofmann 2003; Bae and Bailey 2006; Cui et al. 2007; Davidson and Qi 2008) strategies. In the former approach, two alternative clusterings are sought at the same time whereas in the latter one, a novel alternative clustering is found by conditioning on a given solution. Although being demonstrated to work well in some applications, it is unclear how to extend these algorithms to find multiple alternative clusterings since their objective functions are only suitable to find up to two alternative clusterings from the data. It is also worth mentioning that seeking alternative clustering can be considered related to ensemble clustering (Strehl and Ghosh 2002; Topchy et al. 2005; Fern and Lin 2008). However, there is a significant difference in the clustering objective of these two areas. While alternative clustering aims at finding different clustering solutions from the data, the final objective of ensemble clustering remains searching for a *single* clustering, where each cluster can be picked up from a clustering solution, that is most consistent throughout the entire data (Cui et al. 2007).

We develop in this paper a framework to uncover multiple alternative clusterings from an input data. The proposed algorithm, namely MACL (Multiple Alternative Clusterings), takes an iterative procedure to search for alternative clusterings and at each iteration, a novel clustering is uncovered by conditioning on all previously found clusterings. Though this work can be considered as an extension from our previous one (Dang and Bailey 2010a), a clear distinction in this work is that we address a more general problem by searching for multiple possible alternative clusterings, not limited to two alternative clusterings as tackled in Dang and Bailey (2010a) (and in most of the work aforementioned above). Moreover, while (Dang and Bailey 2010a) addresses the problem in an unsupervised manner, the work in this paper seeks alternative clusterings in sequence by conditioning on all previously found clusterings. In other words, it is only able to ensure the alternative clustering's novelty if all previous clustering solutions were taken into account. For this reason, compared to those developed in Dang and Bailey (2010a), Jain et al. (2008), MACL is considered to be more closely related to the semi-supervised learning techniques and we thus provide experimental comparisons against most of these algorithms in Sect. 5 of the paper.

In summary, in this work we make the following contributions:

- We propose a framework for handling the problem of discovering multiple alternative clusterings over data. Specifically, we develop an efficient EM-based algorithm that well optimizes a dual-objective function of both clustering quality and dissimilarity.
- Unlike most of the algorithms that exploit the orthogonality between clustering solutions, we exploit the mutual information, which is firmly rooted from information theory, to minimize the uncorrelation amongst alternative clusterings. Such a measure directly manipulates over data distributions and further enables practical computation when being combined with the maximum likelihood framework.
- We conduct experiments over both synthetic and real world benchmark datasets, compare our proposed approach against most well-known algorithms in the literature. The experimental results demonstrate the effectiveness of our approach in uncovering multiple alternative clusterings.

The remaining of the paper is organized as follows. We review related work to our study in Sect. 2 and provide the preliminaries along with the formal definition of our problem in seeking multiple alternative clusterings from data in Sect. 3. We describe our framework to address this problem in Sect. 4 by first constructing the clustering objective function, then developing an algorithm relied on the expectation-maximization technique to optimize it. The convergence property of the algorithm is also proved in this section. In Sect. 5, we present the experimental results of our proposed solution on a number of synthetic and real-life datasets and in Sect. 6, we conclude the paper.

2 Related work

The problem of discovering alternative clusterings is relatively young and recently it has drawn much attention from both data mining and machine learning communities. As mentioned in the previous section, one can generally divide most of algorithms developed in this area into two approaches: unsupervised (Jain et al. 2008; Dang and Bailey 2010a; Niu et al. 2010) and semi-supervised (Gondek and Hofmann 2003; Bae and Bailey 2006; Cui et al. 2007; Davidson and Qi 2008) strategies. The algorithms developed in Jain et al. (2008), Niu et al. (2010) and Dang and Bailey (2010a) are unsupervised learning techniques which attempt to seek two alternative clusterings at the same time and without requiring any prior provided clustering. In these techniques, the objective function of a partitioning method is adapted by incorporating a measure of the uncorrelation between two disparate clusterings. Such a quantity in Jain et al. (2008) is the dot product between pairwise mean vectors of two clustering solutions whereas in Dang and Bailey (2010a), it is the information sharing between two solutions. For example, when minimizing the cluster means' inner products along with the objective function of k-means technique, one can ensure that two solutions (represented by cluster-means) not only approach orthogonality but also have good quality (in terms of the k-means' objective). The work in Niu et al. (2010) takes a different approach by combining the uncorrelated subspace learning into the process of spectral clustering. In quantifying for the independence between two subspaces, it uses the Hilbert-Schmidt Independence Criterion (HSIC) (Arthur et al. 2005) and such a combination results in a nice objective function represented in matrix trace forms. Consequently, an iterative approach can be employed to learn two matrices of projections of which solutions based on spectral partitioning (Weiss 1999; Ng et al. 2001) are generally supported.

On the other hand, the approaches developed in Gondek and Hofmann (2003), Bae and Bailey (2006), Cui et al. (2007), Davidson and Qi (2008) are semi-supervised as they require an existing clustering solution to be provided as prior information, and search for another clustering that is uncorrelated (i.e., different) from that given one. While the CIB technique developed in Gondek and Hofmann (2003) is an extension of the information bottleneck method (Tishby et al. 1999; Slonim et al. 2006) in which the mutual information between data features and the new clustering is maximized conditioning on the given clustering, COALA proposed in Bae and Bailey (2006) generates a set of cannot-link constraints (Wagstaff and Cardie 2000; Wagstaff et al. 2001) based on the provided clustering and it builds up an alternative clustering by conforming these constraints in the agglomerative clustering process. Two algorithms developed in Cui et al. (2007) take a different approach by exploring the property of orthogonality. They first characterize the existing clustering by either a set of centroids or data features, and then form a subspace orthogonal to these representative vectors. An alternative clustering is then simply found by partitioning the data projected on this new orthogonal subspace. It is noticed that though (Cui et al. 2007) can discover for more than one alternative clustering, they only condition on a single previous clustering and thus may not ensure the novelty of the alternative clustering. The ADFT algorithm developed in Davidson and Qi (2008) adopts an approach that employs a distance metric (Xing et al. 2002) for clustering's representatives rather than the clusters' means. Compared to the work (Cui et al. 2007), this approach is more advantageous as it can further handle the case in which the data dimension can be smaller than the number of clusters (e.g., spatial datasets). The work developed in Günnemann et al. (2012) proposes an interesting approach which combines a model-based clustering paradigm and a subspace projection to discover alternative clusterings and further allows the overlapping between alternative clusterings. The study (de Bie 2011) proposes a quantity named self-information defined over a cluster or set of clusters. It then seeks clusters/clusterings iteratively in which a subsequent one is maximally interesting given the previously found patterns. The technique developed in Dang and Bailey (2010b) takes a different approach stemmed from information theory which aims to maximize the mutual information between data observations and the cluster labels of the alternative clustering while at the same time to minimize such information between the alternative and the given clustering. A resemble clustering objective is also adopted in Nguyen and Epps (2010) yet is optimized by using an iterative approach, in contrast to the hierarchical technique adopted in Dang and Bailey (2010b). We provide experimental comparison to most of these reviewed algorithms in Sect. 5.

3 Preliminaries and problem definition

In information theory, the entropy quantity plays a central role as a measure of uncertainty or information. Let X be a continuous random variable and associated with X is the probability density function $p(x)$, then the entropy of X is mathematically defined as:

$$H(X) = -\int p(x) \log p(x) dx \tag{1}$$

This definition for a single variable can be extended for a pair of random variables and in such case, we have a joint entropy between two continuous random variables defined as:

$$H(X, Y) = -\iint p(x, y) \log p(x, y) dx dy \tag{2}$$

of which $p(x, y)$ is the joint density function of two variables X and Y. When a variable is known and the other is not, the remaining information (uncertainty) is measured by the conditional entropy:

$$H(X|Y) = -\iint p(x, y) \log p(x|y) dx dy \tag{3}$$

A closely related concept with the entropy is the mutual information which is defined as the relative entropy (also called Kullback-Leibler distance) between the joint distribution $p(x, y)$ and the product of two marginal distributions $p(x)p(y)$:

$$I(X; Y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \tag{4}$$

Since $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, it can be said that the mutual information quantifies for the amount of information that one random variable contains about another variable. When this measure is large, two random variables are closely correlated and conversely when it is small, the two variables are highly uncorrelated. And it is not difficult to prove that X and Y are independent if and only if the mutual information between them is equal to zero. These definitions and relationships are straightforwardly extended for multiple random variables (Cover and Thomas 1991).

We apply the concepts above into our clustering problem by treating each clustering solution as a random variable. With this setting, it is possible to formulate the problem of uncovering multiple alternative clusterings as follows:

Problem definition We are given a set of N data points $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ with each data instance x_n $(1 \leq n \leq N)$ is a vector in the D-dimensional space. The task is to seek for a set of non-redundant alternative clusterings $\mathbb{C} = \{C^{(1)}, C^{(2)}, \ldots\}$ from $\mathcal{X}$ such that the clustering quality of each $C^{(s)}$ is high (e.g., fulfilled by an objective function), while at the same time, each of them is pairwise uncorrelated to one another, i.e. $I(C^{(r)}; C^{(s)})$ is minimized and as close to zero as possible for all $C^{(r)}, C^{(s)} \in \mathbb{C}$ and $C^{(r)} \neq C^{(s)}$.

4 Multiple alternative clusterings

4.1 Clustering objective function

In many practical machine learning and pattern recognition problems, the maximum likelihood is widely used as a statistical technique to estimate the parameters of a density mixture model. Under the framework of maximum likelihood, one aims to maximize the following log-likelihood function:

$$L(\Theta|\mathcal{X}) = \log P(\mathcal{X}|\Theta) = \sum_{n=1}^{N} \log p(x_n|\Theta) \tag{5}$$

where the set of data instances x_i is assumed to be independently drawn from the distribution $p(x|\Theta)$ parameterized by Θ. The function $L(\Theta|\mathcal{X})$ can also be thought of as the likelihood of the parameters Θ given the data observation $\mathcal{X}$. The goal of maximum likelihood is thus to find the Θ that maximizes $L(\Theta|\mathcal{X})$.

The cluster analysis problem turns out to be a special case of estimating parameters for a density mixture model. From this view, one may model a clustering solution as a mixture density model of K probability distributions and associate each individual distribution (referred as component distribution) with a cluster. For most cases, all component distributions have the same functional form and often the Gaussian probability density function is used. The cluster analysis is therefore equivalent to the process of maximizing the parameters of the density mixture model which has the form below:

$$\widehat{\Theta} = \arg\max_{\Theta} L(\Theta|\mathcal{X}) = \arg\max_{\Theta} \sum_{n=1}^{N} \log \sum_{i=1}^{K} \alpha_i \, p(x_n|\theta_i) \tag{6}$$

in which $\alpha_1, \alpha_2, \ldots, \alpha_K$ are the prior or mixing probabilities, and

$$p(x|\theta_i) = \mathcal{G}(x - \mu_i, \Sigma_i) = \frac{\exp\{-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)\}}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \tag{7}$$

is the ith Gaussian function in the mixture model, which is completely identified by the parameters $\theta_i = (\mu_i, \Sigma_i)$. Together with α_i's, these parameters need to be found (as being explicitly presented in the next section).

Our clustering objective will be formulated under this framework of maximum likelihood for a mixture model. Specifically, our clustering framework is iterative and we seek for a novel alternative clustering (represented by a mixture model) at each time. Information of all previously uncovered clusterings will be used as the background knowledge to derive a novel alternative clustering and this process is repeated until the new clustering has high sharing information with any of the reference clusterings.[1] This signifies that most of the high quality groupings from the data have been uncovered and all of them are independent from each other. We therefore shall regularize the likelihood function by the mutual information sharing between the novel clustering and each of the reference clusterings. This ensures that the resultant clustering is pairwise uncorrelated from any of the previously known clusterings. The selection of mutual information accounting for the clustering dissimilarity is advantageous in two folds. First, unlike most orthogonal projection/transformation techniques (reviewed in the introduction section) which explicitly simplify clusterings by some forms of representatives (e.g., clusters' means (Cui et al. 2007), or distance metric (Davidson and Qi 2008)), mutual information naturally manipulates directly on the data distribution and thus it does not loose important details in the data. Second, as also being defined as the function over the probability density distributions, mutual information is completely comparable with the likelihood term by measuring the clusterings' dissimilarity in the same unit of clustering quality. We therefore naturally formulate our alternative clustering objective function as follows:

$$\begin{aligned} \widehat{\Theta} &= \arg\max_{\Theta} \widetilde{L}(\mathcal{X}|\Theta) \\ &= \arg\max_{\Theta} \left\{ \sum_{n=1}^{N} \log \sum_{i=1}^{K} \alpha_i \, p(x_n|\theta_i) - \gamma \sum_{s} I\left(C; C^{(s)}\right) \right\} \end{aligned} \tag{8}$$

where $C^{(s)}$'s are known (reference) clusterings and the novel C is parameterized by Θ. This dual objective function ensures that the clustering quality of the alternative clustering C is

[1] In our experiment, we consider the sharing or mutual information of 0.5 as a high value.

maximized (by the first term) while at the same time its similarity with respect to all previously found solutions is minimized (assured by the second term). The tradeoff between these dual objectives is controlled by the regularization parameter γ. It is noted here that whether our combination of mutual information and the maximum likelihood is reasonable—i.e. do they measure comparable quantities? As shortly seen in the next section, our formulation based on mutual information will be consistent with the objective of maximum likelihood since both are defined based on the probability density functions over the observed data. In other words, they quantify the same unit and thus make our combination between clustering quality and clustering dissimilarity feasible to be optimized.

4.2 The EM based algorithm

Our objective of optimizing a set of parameters Θ characterized for the novel alternative clustering C can be achieved by using the Expectation Maximization (EM) technique. Generally, EM interprets $\mathcal{X}$ as the incomplete data and it views the cluster label C as an additional but unknown variable. The complete-data likelihood is therefore maximized and the EM involves two E- and M-steps. In the E-step, it computes a lower bound approximation to the likelihood function and maximizes it with respect to the distribution of the unobserved data. This leads to the finding of the distribution of C given the observed data $\mathcal{X}$ and the current parameter estimates. In the M-step, the algorithm determines a new set of parameters that maximizes this lower bound provided the distribution of the cluster label computed in the E-step. This procedure is iterated until the algorithm converges (i.e., when the variation of the log-likelihood is small enough).

We employ this technique to solve our clustering objective function proposed in Eq. (8) and it is noticed that minimizing the mutual information between C and any of the reference clusterings $C^{(s)}$'s is equivalent to maximizing its conditional entropy with respect to each of these solutions (cf. Sect. 3). Hence, the second term in our objective function can be replaced by:

$$\sum_s H\big(C|C^{(s)}\big) = -\sum_s \sum_{i,j} \alpha_j\, p(c_i|c_j) \log \frac{p(c_i, c_j)}{\alpha_j} \tag{9}$$

where c_i's denote for the set of clusters in our novel alternative clustering C and c_j's denote for clusters in each of the reference clusterings $C^{(s)}$'s. In estimating the joint probability $p(c_i, c_j)$, it is possibly assumed that c_i and c_j are *conditionally* independent given observed data x_n's (i.e., $p(c_i, c_j, x_n) = p(x_n)p(c_i|x_n)p(c_j|x_n)$ as widely used in graphical learning models (Bishop 2006)). Therefore,

$$\begin{aligned} p(c_i, c_j) &= \sum_{n=1}^{N} p(c_i|x_n)p(c_j|x_n)p(x_n) \\ &= \sum_{n=1}^{N} \frac{p(c_i)p(x_n|c_i)}{p(x_n)} \frac{p(c_j)p(x_n|c_j)}{p(x_n)} p(x_n) \\ &= p(c_i)p(c_j) \sum_{n=1}^{N} \frac{p(x_n|c_i)p(x_n|c_j)}{p(x_n)} \end{aligned} \tag{10}$$

in which we have used the Bayes' theorem. Additionally, since $p(x|c_i)$, $p(x|c_j)$ and $p(x_n)$ are all non-negative, it is always true that $\sum_{n=1}^{N} \frac{p(x_n|c_i)p(x_n|c_j)}{p(x_n)} \geq \frac{\sum_{n=1}^{N} p(x_n|c_i)p(x_n|c_j)}{\sum_{n=1}^{N} p(x_n)}$.[2] We thus approximate $p(c_i, c_j)$ by its lower bound, and by replacing the integrals for the summations (due to continuous values of x_n's), we have:

$$\begin{aligned} p(c_i, c_j) &\geq \frac{p(c_i)p(c_j)\int p(x|c_i)p(x|c_j)dx}{\int p(x)dx} \\ &= p(c_i)p(c_j)\mathcal{G}(\mu_i - \mu_j, \Sigma_i + \Sigma_j) \end{aligned} \tag{11}$$

Our strategy of optimizing the lower bound of the objective is in line with the philosophy of the standard EM technique, which also aims at optimizing the log-likelihood lower bound (as shortly shown in Theorem 1 below). The corresponding regularized log-likelihood function therefore can be written as:

$$\begin{aligned} Q\big(\Theta|\Theta^{(t)}\big) = &\sum_{n=1}^{N}\sum_{i=1}^{K} p(c_i|x_n;\Theta)\log\frac{\alpha_i\mathcal{G}(x_n - \mu_i, \Sigma_i)}{p(c_i|x_n;\Theta)} \\ &- \gamma\sum_{s}\sum_{i,j}\alpha_j p(c_i|c_j;\Theta)\log\alpha_i\mathcal{G}(\mu_j - \mu_i, \Sigma_j + \Sigma_i) \end{aligned} \tag{12}$$

The expectation step in the EM technique can thus be separated into two terms. The first one is the conditional probability of c_i with respect to each observed data x_n:

$$p\big(c_i|x_n;\Theta^{(t)}\big) = \frac{\alpha_i\mathcal{G}(x_n - \mu_i, \Sigma_i)}{\sum_m \alpha_m\mathcal{G}(x_n - \mu_m, \Sigma_m)} \tag{13}$$

The second one is the conditional probability of c_i with respect to each known cluster c_j of each reference clustering $C^{(s)}$:

$$p\big(c_i|c_j;\Theta^{(t)}\big) = \frac{\alpha_i\alpha_j\mathcal{G}(\mu_j - \mu_i, \Sigma_j + \Sigma_i)}{\sum_m \alpha_m\alpha_j\mathcal{G}(\mu_j - \mu_m, \Sigma_j + \Sigma_m)} \tag{14}$$

Notice that $\sum_i p(c_i|x_n;\Theta^{(t)}) = 1$, and $\sum_i p(c_i|c_j;\Theta^{(t)}) = 1$ within each of solution $C^{(s)}$.

In the M-step, we maximize the lower bound with respect to the parameters of the mixture model. This procedure involves more computation. First, we need to differentiate the lower bound with respect to the prior probabilities subject to the constraints $\alpha_i > 0$ and $\sum_i \alpha_i = 1$. This requirement can be handled by replacing α_i as a function of unconstrained variables as follows:

$$\alpha_i = \frac{exp(\beta_i)}{\sum_{i'} exp(\beta_{i'})} \tag{15}$$

which enforces both constraints automatically (Bishop 1995). Notice that:

$$\frac{\partial\alpha_i}{\partial\beta_{i'}} = \begin{cases} \alpha_i - \alpha_i^2 & \text{if } i' = i \\ -\alpha_{i'}\alpha_i & \text{otherwise} \end{cases} \tag{16}$$

[2]Notice that $\frac{a}{b} + \frac{c}{d} \geq \frac{a+c}{b+d}$ for all non-negative a, b, c, d.

For each data instance x_n in the first term and each cluster c_j of each solution $C^{(s)}$ in the second term in Eq. (12), we have from the chain rule that:

$$\begin{aligned}\frac{\partial Q(\Theta|\Theta^{(t)})_{x_n}}{\partial \beta_i} &= \sum_{i'} \frac{\partial Q(\Theta|\Theta^{(t)})_{x_n}}{\partial \alpha_{i'}} \frac{\partial \alpha_{i'}}{\partial \beta_i} \\ &= \sum_{i'} \frac{p(c_{i'}|x_n)}{\alpha_{i'}} (\alpha_{i'}\delta_{ii'} - \alpha_{i'}\alpha_i) \\ \frac{\partial Q(\Theta|\Theta^{(t)})_{c_j}}{\partial \beta_i} &= \sum_{i'} \frac{\partial Q(\Theta|\Theta^{(t)})_{c_j}}{\partial \alpha_{i'}} \frac{\partial \alpha_{i'}}{\partial \beta_i} \\ &= \sum_{i'} \frac{\alpha_j p(c_{i'}|c_j)}{\alpha_{i'}} (\alpha_{i'}\delta_{ii'} - \alpha_{i'}\alpha_i)\end{aligned}$$

in which $\delta_{ii'} = 1$ if and only if $i' = i$. With this in mind, the expansion of the first above derivative directly leads to $p(c_i|x_n) - \alpha_i \sum_{i'} p(c_{i'}|x_n)$ whereas of the second one leads to $\alpha_j p(c_i|c_j) - \alpha_i \alpha_j \sum_{i'} p(c_{i'}|c_j)$. Summing over all data instances x_n's and clusters c_j's of all reference solutions $C^{(s)}$, it straightforwardly follows that:

$$\alpha_i = \frac{\sum_n p(c_i|x_n) - \gamma \sum_s \sum_j \alpha_j p(c_i|c_j)}{\sum_n \sum_{i'} p(c_{i'}|x_n) - \gamma \sum_s \sum_{j,i'} \alpha_j p(c_{i'}|c_j)} \tag{17}$$

The expression for the new update of a mean vector can be found by taking the derivative of $Q(\Theta|\Theta^{(t)})$ with respect to μ_i. It notes that the term $\sum_{n=1}^{N} \sum_{i=1}^{K} p(c_i|x_n;\Theta) \times \log p(c_i|x_n;\Theta)$ in deploying the logarithm in the first sum of Eq. (12) can be omitted due the availability of $p(c_i|x_n;\Theta)$ computed in the E-step. Second, we only concern the terms related to μ_i which exists in two logarithms $\log \mathcal{G}(x_n - \mu_i, \Sigma_i)$ and $\log \mathcal{G}(\mu_j - \mu_i, \Sigma_j + \Sigma_i)$. Thus, the derivative of $Q(\Theta|\Theta^{(t)})$ w.r.t. μ_i can be simplified as follows:[3]

$$\begin{aligned}&\frac{\partial}{\partial \mu_i}\Bigg[\sum_{n=1}^{N}\sum_{i}^{K} p(c_i|x_n) \log \mathcal{G}(x_n - \mu_i, \Sigma_i) \\ &\quad - \gamma \sum_s \sum_{i,j} \alpha_j p(c_i|c_j) \log \mathcal{G}(\mu_j - \mu_i, \Sigma_j + \Sigma_i)\Bigg] \\ &= \sum_{n=1}^{N} p(c_i|x_n)\left(-\frac{x_n - \mu_i}{\Sigma_i}\right) - \gamma \sum_{s,j} \alpha_j p(c_i|c_j)\left(-\frac{\mu_j - \mu_i}{\Sigma_j + \Sigma_i}\right)\end{aligned}$$

Setting this derivative equal to 0, we obtain:

$$\mu_i = \frac{\sum_n p(c_i|x_n)\Sigma_i^{-1} x_n - \gamma \sum_s \sum_j \alpha_j p(c_i|c_j)(\Sigma_i + \Sigma_j)^{-1}\mu_j}{\sum_n p(c_i|x_n)\Sigma_i^{-1} - \gamma \sum_s \sum_j \alpha_j p(c_i|c_j)(\Sigma_i + \Sigma_j)^{-1}} \tag{18}$$

In calculating the new update of the covariance matrix Σ_i, we need to take the derivative of Eq. (8) with respect to Σ_i. Nevertheless, it is observed that the derivative of $Q(\Theta|\Theta^{(t)})$

[3]For clarity, we omit the term Θ in $p(.|.)$.

with respect to Σ_i cannot be solved directly due to the existence of the inverse matrix $(\Sigma_i + \Sigma_j)^{-1}$ appearing in the Gaussian kernel. One solution is to use the Cauchy-Schwartz inequality to find a new bound for the function. Particularly, since the Gaussian kernel is always nonnegative, we can write (based on the Cauchy-Schwartz inequality):

$$\begin{aligned}
&\log\big(\mathcal{G}(\mu_j - \mu_i, \Sigma_i + \Sigma_j)\big) \\
&= \frac{1}{2}\log\left(\int \mathcal{G}(x - \mu_i, \Sigma_i)\mathcal{G}(x - \mu_j, \Sigma_j)dx\right)^2 \\
&\leq \frac{1}{2}\log\int \big(\mathcal{G}(x - \mu_i, \Sigma_i)\big)^2 dx \int \big(\mathcal{G}(x - \mu_j, \Sigma_j)\big)^2 dx \\
&= \frac{1}{2}\log \mathcal{G}(0, 2\Sigma_i)\mathcal{G}(0, 2\Sigma_j)
\end{aligned} \tag{19}$$

It follows that the lower bound for the covariance matrix is given by:

$$\begin{aligned}
Q\big(\Theta|\Theta^{(t)}\big)_{\Sigma_i} = &\sum_{n=1}^{N}\sum_i p(c_i|x_n)\log\mathcal{N}(x_n - \mu_i, \Sigma_i) \\
&- \frac{\gamma}{2}\sum_i\sum_{i,j}\alpha_j p(c_i|c_j)\log\mathcal{G}(0, 2\Sigma_i)\mathcal{G}(0, 2\Sigma_j)
\end{aligned} \tag{20}$$

Taking the derivative of this equation respect to Σ_i and let it equal to 0, the new estimate for the covariance matrix is followed:

$$\begin{aligned}
&\frac{\partial}{\partial\Sigma_i}\left(\sum_{n=1}^{N}\sum_i p(c_i|x_n)\log\mathcal{G}(x_n - \mu_i, \Sigma_i)\right. \\
&\quad\left. - \frac{\gamma}{2}\sum_s\sum_{i,j}\alpha_j p(c_i|c_j)\log\mathcal{G}(0, 2\Sigma_i)\mathcal{G}(0, 2\Sigma_j)\right) \\
&= \sum_{n=1}^{N} p(c_i|x_n)\left(\frac{-1}{2\Sigma_i} + \frac{1}{2\Sigma_i}(x_n - \mu_i)(x_n - \mu_i)^T\frac{1}{\Sigma_i}\right) \\
&\quad - \frac{\gamma}{2}\sum_s\sum_j\alpha_j p(c_i|c_j)\left(-\frac{1}{2\Sigma_i}\right) = 0
\end{aligned}$$

or

$$\Sigma_i = \frac{\sum_{n=1}^{N} p(c_i|x_n)(x_n - \mu_i)(x_n - \mu_i)^T}{\sum_{n=1}^{N} p(c_i|x_n) - \frac{\gamma}{2}\sum_s\sum_j\alpha_j p(c_i|c_j)} \tag{21}$$

Theorem 1 *Let Θ_{i+1} and Θ_i be the parameter estimates of two successive iterations, the proposed algorithm always ensures that $\widetilde{L}(\mathcal{X}|\Theta^{(i+1)}) \geq \widetilde{L}(\mathcal{X}|\Theta^{(i)})$ with its E- and M-steps and thus is converged at certain point.*

Proof We prove the convergence of the proposed algorithm under the framework of the classical EM technique (McLachlan 1997).

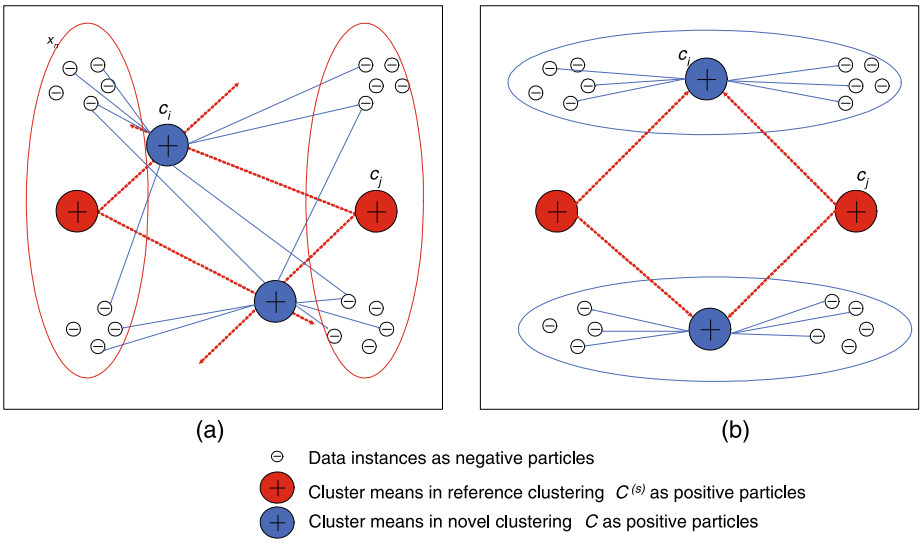

Fig. 1 An illustration based on interactions amongst positive and negative particles. Data points are represented as negative particles whereas cluster means are represented as positive ones

In the M-step, given the fixed observation data $\mathcal{X}$ and reference clusterings $C^{(s)}$, the updates computed in Eqs. (17), (18) and (21) minimize the lower bound of the regularized likelihood function with respect to the set of parameters. Thus, we always have $Q(\Theta^{(i+1)}|\Theta^{(i)}) \geq Q(\Theta^{(i)}|\Theta^{(i-1)})$. Furthermore, $Q(\Theta|\Theta^{(t)})$ is derived by applying the Jensen inequality solely on the first term (likelihood) of the objective function. This implies that $\widetilde{L}(\mathcal{X}|\Theta^{(i+1)})$ is equal to or greater than $Q(\Theta^{(i+1)}|\Theta^{(i)})$.

On the other hand, given $\Theta^{(t)}$ is fixed, updating the new distributions of $p(c_i|x_n;\Theta^{(t)})$ and $p(c_i|c_j;\Theta^{(t)})$ (Eqs. (13) and (14) respectively) in the E-step makes $Q(\Theta^{(i)}|\Theta^{(i-1)}) = \widetilde{L}(X|\Theta^{(i)})$. Thus, in summary we have $\widetilde{L}(\mathcal{X}|\Theta^{(i+1)}) \geq Q(\Theta^{(i+1)}|\Theta^{(i)}) \geq Q(\Theta^{(i)}|\Theta^{(i-1)}) = \widetilde{L}(\mathcal{X}|\Theta^{(i)})$ and thus $\widetilde{L}(\mathcal{X}|\Theta^{(i+1)}) \geq \widetilde{L}(\mathcal{X}|\Theta^{(i)})$, which confirms the regularized likelihood function is monotonically increased after each iteration.

It is therefore if $\widetilde{L}(\mathcal{X}|\Theta)$ has a local maximum, we are bound to reach that maximum at some point. □

Interpretation In an attempt to interpret our computations, we borrow terms from physics to explain the intuition behind the E- and M-steps above. As illustrated in Fig. 1, let us assume that data instances x_n's are *negative* particles (black small points in the figure) and there is a given reference clustering $C^{(s)}$ with its two cluster means c_j's (i.e., $K = 2$) as *positive* particles (red big points in the figure). The red dotted ellipses in Fig. 1(a) represents this reference clustering $C^{(s)}$. Likewise, we may consider cluster means c_i's in our seeking alternative C as also *positive* particles (blue big points in the figure). It is noticed that while the particles x_n's and c_j's are fixed in the space, c_i's are free to move according to the forces imposed on them from both x_n's and c_j's. However, since the polarities of x_n's and c_i's are oppositive, c_i's are pulled close to x_n's with the corresponding magnitude/intensity of the force computed in Eq. (13) of the E-step. In contrast, the polarities between c_i's and c_j's are the same (i.e., positive), c_i's will be pushed far apart from c_j's with the pushing magnitude given in Eq. (14). Consequently, the new position of c_i's are identified by both kinds of forces, yet with different directions (i.e., opposite signs) as shown in Eq. (18) of the M-step.

The movements of c_i's are stabilized once all these imposed forces on c_i's are balanced (as visualized in Fig. 1(b)) which is equivalent to the convergence status of the algorithm.

It is worth noting that, compared to a conventional EM technique, MACL has an extra step in computing the conditional probability of c_i w.r.t. c_j and incorporates this quantity into the calculations of three parameters in the M-step. For each epoch, the E-step computes $N \times K$ conditional probabilities of each cluster w.r.t. each data instance and $J \times K$ w.r.t. each known cluster (where J denotes the total number of all previous clusters). Similarly, within each epoch, the M-step involves the computation over $(N + J)$ entries of $p(c_i|x_n)$'s and $p(c_i|c_j)$'s which also amounts to $(N + J) \times K$ for K clusters. However, computing Σ_i, μ_i in the M-step and Gaussian distributions in the E-step can be cubically proportional to the number of data dimensions (due to involving the matrix-vector multiplication and determinant computation). The overall computation of both steps is thus $O((N + J)KD^3)$. Certainly, this evaluation is only within each epoch of the algorithm. The overall computation of MACL, analogous to k-means or a classical EM technique, is also dependent on the number of E and M iterations (until convergence) which can be varied across different initial parameters, structures of data distributions as well as the accuracy degree of a Gaussian mixture model assumption (Nathan Srebro and Roweis 2005; Dasgupta and Schulman 2000).

5 Experiments

5.1 Experimental setup

We provide experimental results on both synthetic and real-world benchmark datasets. The proposed MACL algorithm is compared against six alternative clustering algorithms: the CIB method (Gondek and Hofmann 2003), COALA (Bae and Bailey 2006), two methods from (Cui et al. 2007) denoted by Algo1 and Algo2, the ADFT algorithm (Davidson and Qi 2008), and the mSC technique recently developed in Niu et al. (2010).

We set the maximum number of iteration in MACL to 100 and consider it converged when the difference in two consecutive likelihoods is smaller than $1e \star 10^{-3}$. MACL's outputs are post-processed by assigning each data instance to the cluster to which it has the highest probability. For γ parameter, we varied its value from 0.1 to 0.2 and found that the range between 0.12 and 0.18 often leads to good outcomes. We therefore report our results when γ is set at 0.15 for most datasets examined, except Syn2 dataset (described below) at 0.13. An alternative clustering is considered novel if its mutual information with any previous clustering is no more than 0.5. We run the algorithm 5 times and choose the best result which has the highest likelihood value.

For ADFT, we implement the gradient descent method integrated with the iterative projection technique (in learning the full family of the Mahalanobis distance matrix) (Xing et al. 2002). We also use the EM technique as the core clustering technique for the approaches developed in Cui et al. (2007), Davidson and Qi (2008). Similar to MACL, we run each algorithm 5 times and choose its best clustering solution. Also, for each run, we initialize the prior probabilities of all clusters equally, same covariance matrices (equal to the data covariance) yet randomly selected cluster means within the data space.

For the CIB method, we implement the iterative version (Gondek and Hofmann 2003) and its output clustering is also post-processed by assigning data points to clusters with highest probability. Following the suggestion by the authors (Niu et al. 2010), for the mSC technique, we initialize the subspace views by grouping dependent features (measured by the HSIC) into the same view. The kernel function is Gaussian and we use the spectral clustering technique (Ng et al. 2001) for mSC's core clustering algorithm.

5.2 Clustering evaluation

We evaluate the clustering results based on both clustering dissimilarity and clustering quality measures. For measuring dissimilarity between two clusterings, we report the values of two different measures. The first also most popular one is the normalized mutual information (Law et al. 2004; Topchy et al. 2004; Gondek et al. 2005) defined by: $\mathrm{NMI}(C^{(r)}; C^{(s)}) = I(C^{(r)}; C^{(s)})/(H^{(r)}H^{(s)})$, where $I(C^{(r)}; C^{(s)}) = \sum_i \sum_j \frac{n_{ij}}{n} \log(\frac{n.n_{ij}}{n_i.n_j})$ and n_{ij} denotes the number of shared data instances between two clusters $c_i \in C^{(r)}$, $c_j \in C^{(s)}$. In addition to NMI which might favor techniques with information theory approaches, we use another measure, the Jaccard index (JI), which is defined as: $\mathrm{J}(C^{(r)}; C^{(s)}) = \frac{n_{11}}{n_{11}+n_{01}+n_{10}}$, where n_{11} is the number of pairs of data instances in the same cluster for both $C^{(r)}$ and $C^{(s)}$, n_{01} and n_{10} are the number of samples' pairs belonging to the same cluster in one solution, but not in the other.

For measuring clustering quality we divide into two cases: if ground truth class labels are known, the agreement between clustering results and the correct labels is calculated by the F-measure: $\mathrm{F} = 2P \times R/(P+R)$, in which P and R are respectively the precision and recall. If the true labels are not known, we use the Dunn Index, similar to (Bae and Bailey 2006; Davidson and Qi 2008): $\mathrm{DI}(C) = \frac{\min_{i \neq j}\{\delta(c_i, c_j)\}}{\max_{1 \leq k \leq K}\{\Delta(c_k)\}}$, where C is a clustering, $\delta : C \times C \rightarrow \mathbb{R}_0^+$ is the cluster-to-cluster distance and $\triangle : C \rightarrow \mathbb{R}_0^+$ is the cluster diameter measure. In addition to Dunn Index, we also use the vector quantization error VQE (Davidson and Qi 2008) to evaluate clustering quality. These measures are widely used in alternative clustering (Bae and Bailey 2006; Davidson and Qi 2008; Jain et al. 2008; Niu et al. 2010) and it is worth to remind that for the NMI and JI measures, *a smaller value is desirable*, indicating higher dissimilarity between clusterings, while for the F-measure and Dunn Index, *a larger value is desirable* and for VQE, *a smaller value is expected*, indicating a better clustering quality.

5.3 Results on synthetic datasets

We generate two synthetic datasets in order to evaluate the performance of our proposed algorithm against other alternative clustering techniques. For the first dataset Syn1, we extend the popular one from Bae and Bailey (2006), Cui et al. (2007), Davidson and Qi (2008) into three dimensions to include more clustering solutions. As such, Syn1 consists of 8 Gaussian sub-classes, each having 200 data points located at each corner of a cube. The goal of using this dataset, when setting the number of clusters to 2, is to test whether our algorithm is able to uncover three alternative clusterings that are pairwise orthogonal. For the second synthetic dataset Syn2, we use a more complicated scenario of which 6 Gaussians are generated and located in a ring shape. Though it is not always true in practice, we assume that the number of clusters within each alternative clustering is equal to 2 and thus there are three different yet equally important clustering structures embedded in this dataset. It is noticed that, unlike the Syn1 where alternative clusterings can be found by projecting the data onto different subspaces (dimensions), clustering structures in Syn2 are not orthogonal and simply projecting data on any subspace does not reveal solutions. Moreover, we assume that no feature selection/extraction is applied to Syn2 and it is directly provided to MACL and other algorithms.

For these two synthetic datasets, we first run MACL without any reference clustering (i.e., EM algorithm). Once the first clustering is found, it is incorporated into the objective function as a reference clustering and we iterate MACL to find another alternative clustering. This process is repeated until a newly generated clustering is found having the NMI value

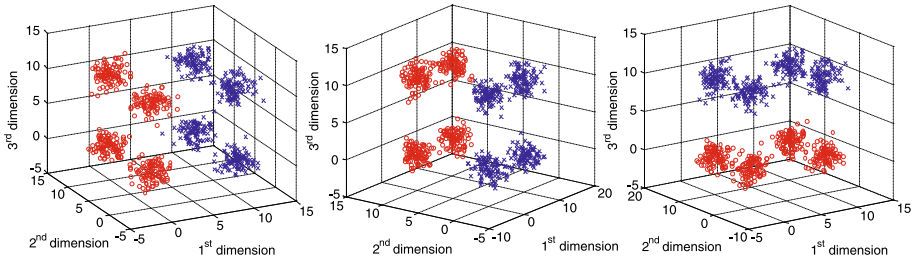

Fig. 2 Three alternative clusterings returned by MACL on Syn1 dataset

Table 1 Clustering performance of all algorithms on two synthetic datasets Syn1 and Syn2

Methods	NMI	JI	F	NMI	JI	F
	Syn1			Syn2		
COALA	0.00	0.33	1.00	0.12	0.40	0.84
CIB	0.15	0.39	0.89	0.14	0.42	0.83
ADFT	0.10	0.36	0.94	0.10	0.39	0.78
Algo1	0.12	0.37	0.92	0.15	0.41	0.76
Algo2	0.14	0.39	0.90	0.15	0.42	0.74
mSC	0.05	0.34	0.97	0.16	0.43	0.83
MACL	0.10	0.35	0.95	0.09	0.38	0.98

higher than 0.5 (thus, not considered dissimilar) with respect to any reference clustering. In Fig. 2, we show all clustering solutions returned by MACL on Syn1 dataset. It is observed that three orthogonal yet equally important clusterings have been successfully uncovered by our algorithm. Its average results computed from these clusterings are reported in Table 1 (under Syn1). The F-measure is used for clustering quality evaluation as the ground truth labels are known and it is averaged on the second and third alternative clusterings. We also apply a similar computation to mSC and the algorithms developed in Cui et al. (2007) and similarly, their results are averaged and reported in Table 1). However, as the remaining techniques are only able to uncover a single alternative clustering, we thus run them with a random reference clustering from Fig. 2. Their F-measure values reported in Table 1 are computed based on their single alternative clustering. One may observe that these algorithms also perform well with this Syn1 dataset. The performance of mSC technique is better than MACL as Gaussian sub-classes presented in the eigenvector space are quite separated and mSC's results are only slightly affected by the k-means's initialization applied on the eigenvector space. However amongst all algorithm, COALA achieves the highest results since its core technique is based on agglomerative hierarchical clustering and thus is not sensitive to initial parameters.

We show the clustering results returned by all algorithms on Syn2 dataset in Fig. 3 and their corresponding average clustering measures are reported in Table 1 (under Syn2). Figure 3(a) shows three alternative clusterings returned by MACL. For COALA, CIB and ADFT which are capable of producing only a single alternative clustering, we demonstrate two solutions corresponding the first and second clusterings (first two graphs in Fig. 3(a))

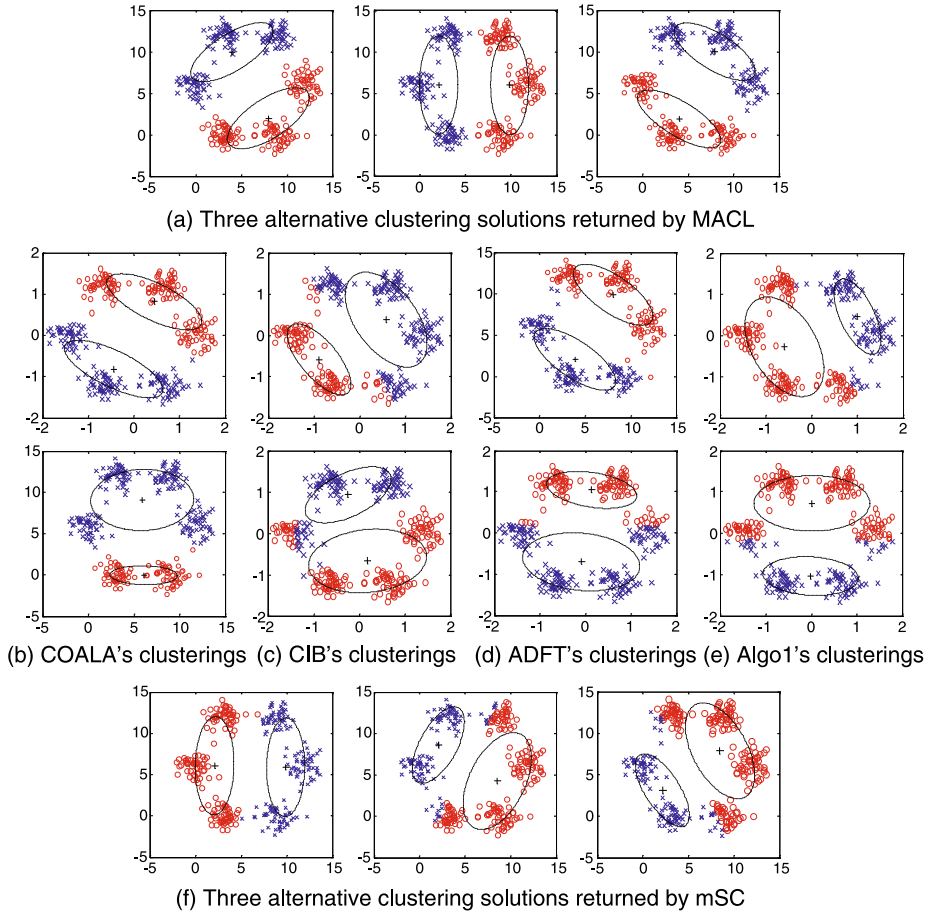

Fig. 3 Alternative clusterings returned by all algorithms on Syn2 dataset (see text for explanation)

provided as a reference clustering. We omit the case where the third clustering is given as a pre-defined clustering since the results are similar (yet opposite) to the case where the first clustering is provided. Their alternative clusterings are shown in Fig. 3(b–d). The alternative clusterings of Algo1 is shown in Fig. 3(e) and those of mSC is shown in Fig. 3(f).

It is observed that while MACL can easily discover three uncorrelated clusterings by minimizing the pairwise information, all other algorithms perform limitedly with this dataset. COALA seems to work well if the first alternative clustering is given but its clustering result is poor if the second clustering is provided (Fig. 3(b)). Likewise, both alternative clusterings returned by CIB are quite different from the true ones as seen in Fig. 3(c). Unlike MACL where we directly minimize the mutual information between the alternative and all existing clusterings, it is noted that CIB only conditions on the reference clustering in its process of maximizing the information between the new clustering and the set of data features. This might explain for its inferior performance.

Observed from Fig. 3(d) that the resultant clustering of ADFT is close to the true one if the first clustering is provided. However, we also see that its alternative clustering on the second case is not as expected. This is probably explained by the property of the stretcher

matrix where its diagonal elements are actually the stretching factors along each dimension. Thus, varying any of the elements (corresponding to dimensions) does not make the alternative clusterings easier to be discovered. It is also observed in Fig. 3(e) that, the resultant clusterings returned by Algo1 are far from the true structures for both cases. This can be justified by the step of data projection orthogonal to the set of provided clusters' means that has made the data being distorted and more overlapping in the orthogonal space. Moreover, though reported to be able to uncover multiple alternative clusterings, Algo1 solely conditions on the previous clustering to find a new clustering. For this low dimensional dataset, we have found that it is unable to find the third alternative clustering as the third alternative is highly overlapping with the first provided one (thus not shown in the figure). Also, we do not show the results returned by Algo2 since they are quite similar to that of Algo1. However, it is worth to note that the second transformation performed by Algo2 is undefined (since the PCA solution reduces the number of dimensions to obtain a new subspace). For the results of mSC reported in Fig. 3(f), notice that mSC seeks 3 alternative clusterings concurrently and as observed, except the first solution, the other two ones are less successful (in term of clustering quality) though they are quite orthogonal to each other. These algorithms might get more advantageous in high dimensional data, especially when clusterings exist in subspaces (Parsons et al. 2004), but less advantageous in cases of low dimensions. Finally, it is worth noting that for both Syn1 and Syn2 datasets, our algorithm is only able to uncover up to 3 alternative clusterings. When γ is set to 15% and slightly higher values (to maintain clustering quality), keeping searching for the fourth one results in a clustering having high overlapping (i.e., large value of NMI) with one of the previously found solutions. Therefore, in both datasets, our algorithm terminates with the number of alternative clusterings at 3, which is intuitive to our observation from Figs. 2 and 3.

5.4 Results on pen digit dataset

We use the hand written pen digit dataset from Davidson and Qi (2008), which consists of 1602 data samples and each single sample corresponds to a hand written digit from 0 to 9. A digit is written in a pen-based pressure sensitive tablet and 8 x, y positions of the pen are recorded to form 16 attributes of the digit (the stylus pressure level values are ignored). Certainly, the most prominent partitioning over this dataset is the one based on the ten digits. Nonetheless, for the purpose of generating multiple alternative clusterings and for explanation, analogous to the one adopted in Davidson and Qi (2008), our objective of using this dataset is to observe how our algorithm can interpret the ways that the digits have been written. It has been found that, by setting the cluster number to 2, our MACL algorithm is able to uncover up to three alternative clusterings from this dataset. Also, we use the Dunn Index and Vector Quantization Error to evaluate its clustering quality and compare against other algorithms. Moreover, since many trials of MACL, ADFT, Algo1 and Algo2 often return one similar clustering, we thus view it as a dominant clustering, denoted by $C^{(1)}$, and provide it to other algorithms, except mSC, as the first reference clustering.

We report the clustering results on this dataset of all algorithms in Table 2 and in Fig. 4, we demonstrate three alternative clusterings uncovered by our MACL algorithm. Each picture in the figure corresponds to a cluster's centroid. It is observed that three resultant clusterings provide three different interpretations regarding how the digits have been written. Notice that though it might not be really convincing when cluster's means are used for visualizing the writing styles since the most frequently occurring digits appeared in each cluster may not be much in common. However, the visualization can somewhat show the difference between data clusters as well as the contrast amongst clustering solutions. Also, it is noticed

Table 2 Clustering performance of all algorithms on Pen Digit datasets. Other than MACL, most algorithms find $C^{(3)}$ close to $C^{(1)}$ as indicated by the high values of NMI(C1, C3)

Method	COALA	CIB	ADFT	Algo1	Alog2	mSC	MACL
DI(C1)	1.7	1.7	1.7	1.7	1.7	1.67	1.7
DI(C2)	1.66	1.67	1.6	1.58	1.57	1.60	1.62
DI(C3)	1.69	1.6	1.71	1.67	1.66	1.55	1.8
VQE(C1)	1924	1924	1924	1924	1924	1931	1924
VQE(C2)	1932	1919	1920	1918	1925	1923	1919
VQE(C3)	1923	1921	1926	1919	1927	1919	1915
JI(C1, C2)	0.62	0.4	0.42	0.38	0.36	0.36	0.37
JI(C1, C3)	0.91	0.78	0.91	0.83	0.86	0.44	0.41
JI(C2, C3)	0.36	0.42	0.45	0.37	0.39	0.49	0.44
NMI(C1, C2)	0.4	0.02	0.01	0.01	0.02	0.02	0.01
NMI(C1, C3)	0.84	0.7	0.83	0.74	0.74	0.34	0.03
NMI(C2, C3)	0.01	0.12	0.01	0.01	0.02	0.06	0.2

that x and y positions of all digits are recorded at fixed time intervals (i.e., sampling rates). Therefore, different person may write differently for the same stroke of the same digit or even strokes of the same digit might be written in various sequences. This might cause some pen-digits with the same identity possibly being grouped into different clusters.

As seen from the first clustering $C^{(1)}$, the writing style of the digits seems to follow clockwise trend with a slightly constant speed for digits grouped in the first cluster but increasing writing speed for those grouped in the second cluster. For the second clustering $C^{(2)}$, it is possible to observe from the first cluster that the digit writing style is in counter-clockwise, as opposed to the first clustering, with a smooth speed for most of the strokes. For clustering $C^{(3)}$, we further observe that two clusters' centroids demonstrate two different novel writing styles. While the digit writing manner in the first cluster starts with a stroke from left to right, then with strokes going down to create a very far distance of two ends, the writing style in the second cluster begins with a stroke from right to left, going down then up again to create an almost closed-end circle. These two ways of grouping digit writing are not only themselves contrasted to each other but are also clearly distinguished from those discovered from the first two previous found clustering partitions $C^{(1)}$ and $C^{(2)}$.

For comparison against other techniques over this dataset, we observe the clustering performance reported in Table 2. It is noticed that COALA and CIB are unable to uncover multiple alternative clusterings. The results reported related to $C^{(3)}$ in Table 2 therefore were computed by providing $C^{(2)}$ as the reference clustering for these algorithms. It is seen that, for these algorithms, $C^{(3)}$ was found indeed very close to that of clustering $C^{(1)}$. It is also observed that the similar results are with Algo1 and Algo2. As observed from Table 2, the NMI and Jaccard Index between $C^{(1)}$ and $C^{(3)}$ of these algorithms are very large, which demonstrate the highly overlapped clustering structure between $C^{(3)}$ and $C^{(1)}$. The mSC seeks three uncorrelated subspaces along with clusterings simultaneously and its performance is better than these algorithms in searching for clustering $C^{(3)}$. However, its Dunn index over three resultant clusterings is still smaller than that of MACL while both the NMI and Jaccard index is higher. By conditioning on both $C^{(1)}$ and $C^{(2)}$ in searching for a new alternative clustering, our MACL algorithm has successfully discovered $C^{(3)}$. This alternative solution is not only highly independent from both $C^{(1)}$ and $C^{(2)}$ as indicated by the low values of NMI and Jaccard Index, its clustering quality is also high confirmed by the

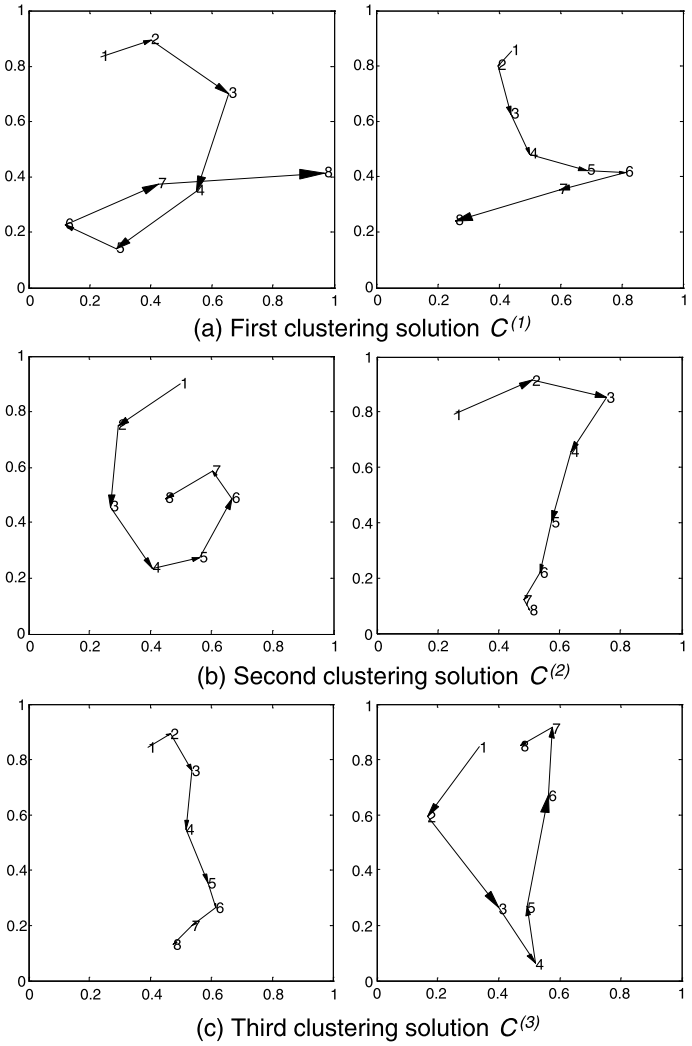

Fig. 4 Three alternative clusterings returned by MACL on Pen Digit dataset

large value of Dunn Index as well as the small one of VQE. It is worth to mention that the first two clustering solutions are also found and reported by the ADFT (Davidson and Qi 2008). However, our MACL technique can further uncover the third alternative which has meaningful interpretation. Similar to the two synthetic datasets, the fourth alternative clustering returned by MACL is often highly overlapped with the first clustering solution. The algorithm is thus terminated with the number of alternative clusterings at 3.

5.5 Results on CMUFace dataset

The CMUFace data obtained from the UCI KDD repository (Asuncion and Newman 2007) is an interesting dataset, since its data samples can be partitioned in several different ways (e.g. by individual, by pose, etc.). The dataset consists of images of 20 people taken at vari-

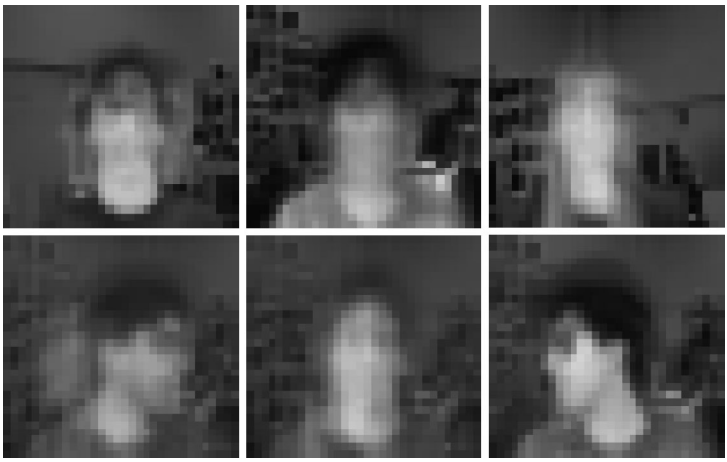

Fig. 5 MACL's clustering results on the CMUFace dataset. *First row's images* correspond to the cluster means in the reference clustering. *Second row's images* correspond to the cluster means found in MACL's alternative solution

ous features such as facial expressions (neutral, happy, sad, angry), head positions (left, right or straight), and eye states (open or sunglasses). Each person has 32 images captured in every combination of these features. Though it is possible to select all images for experiments, we have found that clustering result might be affected by the chosen number of clusters. For example if $K = 3$, an algorithm may derive different alternative clusterings. However if setting $K = 20$, any algorithm can only be able to derive a single clustering solution that is based on different people. This implies that meaningful clusterings are very much dependent on how K is chosen. For this reason, we therefore randomly select 3 people along with all their images to create the dataset in order to alleviate the effect of K selection on the alternative clustering results. In addition, since it is known which image comes from which person, this forms an existing partition over the set of images. We thus run MACL and other algorithms (except mSC) with this reference clustering. As the dimension of this dataset is 960 which is substantially high compared to the number of data instances, we use the PCA technique as a preprocessing step to reduce the number of dimensions, in which we retain the number of first principal components that cover 90% of the original data variance.

Given the reference clustering based on person, MACL is able to find another different clustering from this dataset. For visualization purpose, we show the mean vectors of the reference clustering in the first row and those of the alternative clustering returned by MACL in the second row of Fig. 5. Graphically, it is possible to observe that the uncovered alternative clustering returned by MACL provides another different, yet equally important clustering on this set of image data. While pictures in the first row show that they represent for different individuals, pictures in the second row clearly reveal that images have been partitioned according to different poses. This obviously provides another meaningful interpretation about the data. Despite being able to find multiple alternative clusterings, we found that the third one returned by MACL was highly overlapped with either of two solutions above and thus stopped running the algorithm. In order to compare against other techniques, we report in Table 3 the clustering measures returned by MACL as well as by other algorithms. As observed from this table, COALA and CIB perform slightly better than Algo1 and Algo2, which attempt to find alternative clusterings in an orthogonal transformation space. However, their clustering results are still worse than those of MACL. The clustering dissimilarity

Table 3 Clustering performance of all algorithms on the CMUFace dataset. Values in the last two rows are reported for the alternative clusterings with a reference clustering provided by a conventional EM technique

Methods	NMI	JI	F(pose)	F(person)
COALA	0.27	0.32	0.71	0.87
CIB	0.28	0.34	0.69	0.86
ADFT	0.29	0.33	0.69	0.89
Algo1	0.31	0.34	0.68	0.87
Algo2	0.33	0.36	0.67	0.84
mSC	0.32	0.36	0.59	0.87
MACL	0.23	0.31	0.74	0.91
MACL(0.86/0.64)	0.27	0.34	0.72	0.81
MACL(0.82/N/A)	0.3	0.35	0.7	N/A

returned by MACL is slightly better than that of COALA when it is measured in term of Jaccard Index, but clearly better in term of normalized mutual information. Its clustering accuracy is also better than all of algorithms examined. We also test another strategy by which the clustering labels based on poses are provided as the reference clustering. The clustering accuracy for the person based partitioning of all algorithms is summarized in the fourth column of Table 3.

Since we know both ground truth clusterings (one is based on persons and the other is based on poses) of this dataset, we test another scenario on the influence of the provided knowledge. More specifically, we want to see how well MACL can uncover two inherent clusterings when it is not provided by a proper ground truth clustering but the one that is close to it found by a conventional EM technique. Across multiple runs, it was found that a grouping close (in terms of F-measure) to the person-based clustering is often returned by the classical EM. We thus use it as the reference clustering for MACL and see how close it can uncover a clustering based on poses. In the last two rows of Table 3, we report the results corresponding to two cases: one uses the reference having F-measure of 0.86 and the other of 0.82 (first number in the bracket). One can observe that in two cases, the alternative clustering returned by MACL is still close to the ground truth pose-based clustering as indicated by the high F-measure values (under the F(pose) column) while also independent from the person-based clustering as revealed by the small values of NMI and JI. These values are slightly less successful compared to the case in which the proper person-based grouping is provided (e.g., F-measure of 0.72 and 0.7 compared to 0.74).

We also test the circumstance when a clustering close to the pose-based clustering is provided as prior knowledge. Amongst multiple runs, we found that there was only one clustering whose F-measure w.r.t. the ground truth pose-based clustering is above 0.5 and reaches 0.64[4] (precision of 0.84 and recall of 0.51). Using this grouping as the reference knowledge, MACL has found an alternative clustering that has F-measure of 0.81 w.r.t. the person-based clustering, which is somewhat much less successful compared to the ideal case (conditioning on the ground truth pose-based clustering). These experiments also reveal an interesting result that while it is hard to uncover a clustering based on poses by running a conventional EM multiple times (F-measure only achieves 0.64), we still can find it with

[4]The second number in the bracket of the last two rows of Table 3. Values for the last row are not available as only one clustering having F-measure above 0.5. was found.

Table 4 Clustering performance of all algorithms on Vehicle and Vowel datasets

Methods	NMI	JI	DI	VQE	NMI	JI	DI	VQE
	Vehicle				Vowel			
Algo1	0.27	0.31	1.15	6.21	0.47	0.39	1.66	4296
Algo2	0.27	0.32	1.09	6.12	0.49	0.41	1.62	4271
ADFT	0.29	0.33	1.3	6.03	0.48	0.46	1.63	4241
COALA	0.31	0.34	1.2	6.43	0.44	0.4	1.62	4331
CIB	0.32	0.39	1.16	6.61	0.47	0.38	1.59	4352
mSC	0.28	0.31	1.63	7.46	0.34	0.52	1.62	4283
MACL	0.25	0.27	1.34	5.82	0.37	0.31	1.7	4203

much higher clustering quality (F-measure of 0.72) by conditioning on the first prominent person-based clustering.

5.6 Results on real world datasets

We further compare seven algorithms on two real-world datasets selected from the UCI repository: the Vehicle Silhouette and the Vowel. Though it is not always practical, we make an assumption that the existing clusterings are the ones defined by the class labelled attributes of these datasets and limit the number of alternative clusterings to 3 (including the ground truth clustering). Also, as we do not have ground truth for alternative clusterings, the Dunn Index and VQE (averaged on the two novel alternative clusterings) are used for clustering quality comparison amongst the seven clustering techniques. For COALA, CIB and ADFT, the third alternative clustering is found by conditioning on the second alternative clustering. Moreover, since mSC does not require pre-identified clusterings, we select two out of its three alternative clusterings that are most uncorrelated (measured in NMI) from the pre-defined class labels for comparison. We report the results of all techniques on these datasets in Table 4.

It can be seen that MACL also performs well on these datasets. Its clustering results, both in terms of quality and dissimilarity, are better than those of COALA, CIB and ADFT. This is obvious since MACL conditions on all previously known clusterings to find a novel clustering while these algorithms are only able to condition on a single clustering. It is also seen that the VQE values of COALA and CIB are sightly higher than those of ADFT. This might be explained by the core clustering techniques that these algorithms have been used. While CIB optimizes an objective function purely based on mutual information and COALA is a hierarchical clustering technique, ADFT is an EM-based technique and thus implicitly minimizes the VQE measure. We also find that our MACL's clustering performance is better than that of Algo1, Algo2 and mSC, which all attempt to search alternative clusterings indirectly via transformed spaces. The performance of mSC is better than MACL on the Vowel dataset if measuring in the NMI. However, it is observed that its resultant clusterings are quite imbalanced as revealed by the large values of Jaccard index. MACL's clustering quality, measured in term of Dunn Index and VQE, is slightly better than that of Algo1 in the Vowel dataset, but there is a large difference between two algorithms in the Vehicle dataset. Our clustering algorithm also achieves better clustering dissimilarity compared to Algo1 and 2 in both datasets.

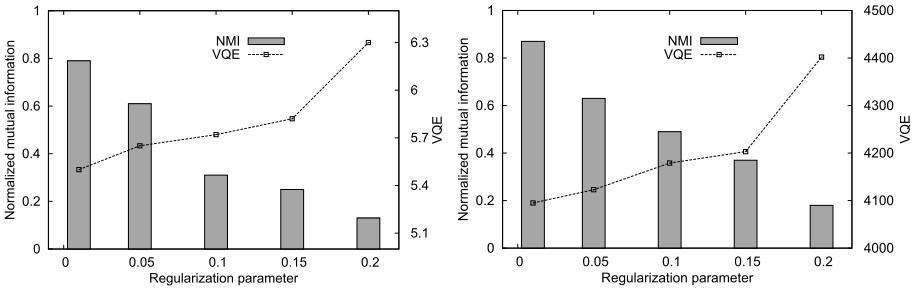

Fig. 6 Impact on MACL when varying the regularization parameter γ on the clustering performance. For an ideal result, both NMI and VQE should be small

5.7 Impact of regularization parameter

As mentioned in Sect. 3, the parameter γ is used to regularize the trade-off between the degree of the dissimilarity of a novel alternative clustering with respect to all previously found clusterings and its clustering quality. We next report the behavior of MACL when this parameter is varied.

In order to be consistent with the expectation maximization framework used in MACL, we do not use the available class labels, instead, the conventional EM technique is run to obtain the first clustering from a dataset. It is then supplied to MACL as a reference clustering and we evaluate how the alternative clustering is different from the first one when γ is changed. In Fig. 6, the relationship between the normalized mutual information, the VQE measure, and the regularization parameter γ is shown for two real world datasets: Vehicle and Vowel. The results are reported when γ is varied between 1 % and 20 % of each dataset's size. As we expected, when the regularization parameter is small, MACL usually converges to an alternative clustering that is highly overlapped with the provided clustering. This is indicated by the high value of the normalized mutual information between two solutions. As we increase the value of γ, the normalized mutual information is decreased, implying that the resultant alternative clustering is also more dissimilar from the provided one. However, its clustering quality, in term of VQE, is somewhat compromised and increased. This inverse relationship between clustering quality and dissimilarity is intuitive and visualizing it can suggest ways to choose an appropriate value of γ. As observed from two graphs in Fig. 6, both requirements of high qualitative and dissimilar clusterings can be achieved when the value of γ is set around 15 %, since the value of VQE in this range is relatively small, whereas that value of the NMI is also not high. It is noted that though there is no proper value of γ working for all datasets, this experiment suggests a general way to find it by tracking down both values of clustering quality (e.g., VQE) and clustering dissimilarity (e.g., NMI) and choose one that best compromises between these two objectives. This strategy should also be applied when more reference clusterings are involved in the objective function. However, as done in the previous sections with the range between 10 % and 20 %, we still found that setting γ to 15 % remains good in searching for the second alternative clustering.

6 Conclusion and discussion

In this paper, we have proposed a novel framework called MACL to discover multiple alternative clusterings that are both of high quality and distinctively different from each other. We address this important problem by combining two mathematically well founded areas of maximum likelihood framework learning and mutual information. Consequently, a dual-objective function is devised and we develop an expectation maximization technique to optimize it. The clustering quality of alternatives is thus achieved by the maximization over the data likelihood whereas the dissimilarity amongst them is ensured by the minimization over their mutual information. Interestingly, the computations in both E- and M-steps of the proposed technique are all intuitive and they resemble the world of force interaction amongst physical particles. We evaluated the performance of the proposed framework on both synthetic and real-world datasets and compared against most well-known algorithms in the literature. The experimental results demonstrated the appealing performance of MACL in searching for multiple alternative clusterings and thus confirmed the potential approach of combining maximum likelihood framework and mutual information.

Nevertheless, we observe that MACL also suffers from several drawbacks. First, being based on the assumption of Gaussian mixture models, MACL's solutions thus converge to convex shaped clusters. For datasets where clustering structures do not strictly follow this assumption (e.g. when clustering boundary boundaries are non-linear), its performance may be compromised. Second, in the circumstance when there is no background information regarding the number of clusters within each alternative clustering, MACL assumes the number to the same across alternatives, which might not be practical in some real world applications. Finally, although MACL is able to seek multiple alternative clusterings, it still may not ensure every possible alternative clustering is uncovered. In our work, we have opted to use a comparison of the similarity (via the NMI measure) between the novel clustering and all previous ones as a criterion to terminate the search process of the algorithm. However, a significant difference in the likelihood could also be a good factor to stop searching for a novel clustering if the number of clusters is the same across all alternative clusterings. In the general case, nonetheless, the likelihood quantity can be biased if the number of clusters is not the same for different clustering solutions. Therefore, seeking to optimise both the number of clusters within each alternative and the total number of alternatives truly embedded in the data is particularly challenging. We believe that these issues are worth further exploration as part of future work.

References

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: ordering points to identify the clustering structure. In *SIGMOD conference* (pp. 49–60).

Arthur, D., & Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035).

Arthur, G., Olivier, B., Alexander, S., & Bernhard, S. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*.

Asuncion, A., & Newman, D. (2007) UCI machine learning repository.

Bae, E., & Bailey, J. (2006). Coala: a novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *IEEE international conference on data mining* (pp. 53–62).

Bishop, C. (1995). *Neural networks for pattern recognition*. USA: Oxford University Press.

Bishop, C. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Cover, T. M., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Cui, Y., Fern, X., & Dy, J. (2007). Non-redundant multi-view clustering via orthogonalization. In *IEEE international conference on data mining*.

Dang, X. H., & Bailey, J. (2010a). Generation of alternative clusterings using the cami approach. In *SIAM international conference on data mining* (pp. 118–129).
Dang, X. H., & Bailey, J. (2010b). A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 573–582).
Dasgupta, S., & Schulman, L. J. (2000). A two-round variant of em for Gaussian mixtures. In *UAI conference* (pp. 152–159).
Davidson, I., & Qi, Z. (2008). Finding alternative clusterings using constraints. In *IEEE international conference on data mining* (pp. 773–778).
Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*, 7–24.
De Bie, T. (2011). Subjectively interesting alternative clusters. In *MultiClust workshop (ECML PKDD)*.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 226–231).
Fern, X., & Lin, W. (2008). Cluster ensemble selection. *Statistical Analysis and Data Mining*, *1*(3), 128–141.
Gondek, D., & Hofmann, T. (2003). Conditional information bottleneck clustering. In *3rd IEEE international conference on data mining, workshop on clustering large data sets* (pp. 36–42).
Gondek, D., Vaithyanathan, S., & Garg, A. (2005). Clustering with model-level constraints. In *SIAM international conference on data mining*.
Günnemann, S., Färber, I., & Seidl, T. (2012). Multi-view clustering using mixture models in subspace projections. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 132–140).
Jain, P., Meka, R., & Dhillon, I. (2008). Simultaneous unsupervised learning of disparate clusterings. In *SIAM international conference on data mining* (pp. 858–869).
Law, M., Topchy, A., & Jain, A. (2004). Multiobjective data clustering. In *IEEE computer society conference on computer vision and pattern recognition* (pp. 424–430).
Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*, 129–137.
McLachlan, T. K. G. J. (1997) *Wiley series in probability and statistics: The EM algorithm and extensions*.
Nathan Srebro, G. S., & Roweis, S. (2005). When is clustering hard? In *PASCAL workshop on statistics and optimization of clustering*.
Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In *In advances in neural information processing systems* (Vol. 14, pp. 849–856). Cambridge: MIT Press.
Nguyen, X. V., & Epps, J. (2010). Mincentropy: a novel information theoretic approach for the generation of alternative clusterings. In *IEEE international conference on data mining conference* (pp. 521–530).
Niu, D., Dy, J. G., & Jordan, M. I. (2010). Multiple non-redundant spectral clustering views. In *ICML conference* (pp. 831–838).
Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: a review. *SIGKDD Explorations*, *6*(1), 90–105.
Slonim, N., Friedman, N., & Tishby, N. (2006). Multivariate information bottleneck. *Neural Computation*, *18*, 1739–1789.
Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.
Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method (pp. 368–377).
Topchy, A., Jain, A., & Punch, W. (2004). A mixture model for clustering ensembles. In *SIAM international conference on data mining*.
Topchy, A., Jain, A., & Punch, W. (2005). Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12).
Wagstaff, K., & Cardie, C. (2000). Clustering with instance-level constraints. In *Proceedings of the seventeenth international conference on machine learning* (pp. 1103–1110).
Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the seventeenth international conference on machine learning*.
Weiss, Y. (1999). Segmentation using eigenvectors: a unifying view. In *IEEE international conference on computer vision* (pp. 975–982).
Xing, E., Ng, A., Jordan, M., & Russell, S. (2002). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 505–512).