# Multiple Clustering Views via Constrained Projections *

Xuan Hong Dang, Ira Assent† James Bailey‡

**Abstract**

Clustering, the grouping of data based on mutual similarity, is often used as one of principal tools to analyze and understand data. Unfortunately, most conventional techniques aim at finding only a single clustering over the data. For many practical 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 discovering alternative clusterings. In this preliminary work, we propose a novel framework to generate multiple clustering views. The framework relies on a constrained data projection approach by which we ensure that a novel alternative clustering being found is not only qualitatively strong but also distinctively different from a reference clustering solution. We demonstrate the potential of the proposed framework using both synthetic and real world datasets and discuss some future research directions with the approach.

## 1 Introduction.

Cluster analysis has been widely considered as one of the most principal and effective tools in understanding the data. Given a set of data observations, its objective is to categorize those observations that are similar (under some notion of similarity) into the same cluster whilst separating dissimilar ones into different clusters. Toward this goal, many algorithms have been developed by which some clustering objective function is proposed along with an optimization mechanism such as k-means, mixture models, hierarchical agglomerative clustering, graph partitioning, and density-based clustering. A general observation with these algorithms is that they only attempt to produce a single partition over the data. For many practical applications, especially those being characterized in high dimensional data, this objective seems to be not sufficient since it is common that the data observations can be grouped along different yet equally meaningful ways. For example, when analyzing a document dataset, one may find that it is possible to categorize the documents according to either topics or writing styles; or when clustering a gene dataset, it is found that grouping genes based on their functions or structures is equally important and meaningful [4]. In both these applications and many other ones, we may see that the natural structure behind the observed data is often not unique and there exist many alternative ways to interpret the data. Consequently, there is a strong demand to devise novel techniques that are capable of generating multiple clustering views regarding the data.

In the literature, several algorithms have been developed to seek alternative clusterings and it is possible to categorize them into two general approaches: seeking alternative clusterings simultaneously [13, 8, 22] and seeking alternative clusterings in sequence [3, 6, 10, 9]. In the former approach, all alternative clusterings are sought at the same time whereas in the latter one, each novel clustering is found by conditioning on all previous clusterings. From a modeling view point, the latter approach has a major advantage that it limits the number of cluster parameters needed to be optimized concurrently.

In this paper, we develop a novel framework to find multiple clustering views from a provided dataset. Given the data and a reference clustering $C^{(1)}$ as inputs, our objective is to seek a novel alternative clustering that is not only qualitatively strong but also distinctively different from $C^{(1)}$. The proposed algorithm achieves this dual objective by adopting a graph-based mapping approach that preserves the local neighborhood proximity property of the data and further conforms the constraint of clustering independence with respect to $C^{(1)}$. Though our research can be categorized into the second approach of sequentially seeking alternative clusterings, it goes beyond the work in the literature by further ensuring that the geometrical proximity of the data is retained in the lower mapping subspace and thus naturally reveals clustering structures which are often masked in the high dimensional data space. We formulate our clustering objective in the framework of a constrained eigendecomposition problem and thus it has a clear advantage that a closed form solution for the learning subspace always exists and is guaranteed to be globally optimal. This property contrasts the proposed

*Part of this work has been supported by the Danish Council for Independent Research - Technology and Production Sciences (FTP), grant 10-081972.

†Dept. of Computer Science, Aarhus University, Denmark. Email: {dang,ira}@cs.au.dk.

‡Dept. of Computing and Information Systems, The University of Melbourne, Australia. Email: baileyj@unimelb.edu.au.

framework to most existing algorithms, which solely focus on optimizing the single condition of decorrelation and may only achieve a local optimal subspace. We demonstrate the potential of the proposed framework using both synthetic and real world data and empirically compare it against several major algorithms in the literature. Finally, some future studies for our framework are discussed and we shortly suggest some potential approaches to deal with them.

## 2 Related Work.

Though being considered related to subspace clusterings [1, 15, 16, 28, 20], the problem of discovering multiple alternative clusterings is relatively young and recently it has has been drawing much attention from both data mining and machine learning communities. In [19], the authors give an excellent tutorial regarding different approaches to the problem. In this section, we adopt the taxonomy that can generally divide the majority of work in the literature into two approaches: those seeking alternative clusterings concurrently and those seeking alternative clusterings in sequence, and briefly review the algorithms that closely related to our research in this paper.

In the first approach, two algorithms named Dec-kmeans and ConvEM are developed in [13] to find two disparate clusterings at the same time. In Dec-kmeans, the concept of representative vectors is introduced for each clustering solution. Subsequently, the objective function of the k-means method is modified by adding terms to account for the orthogonality between mean vectors of one clustering, with respect to the representative vectors of the other. In the ConvEM algorithm, a similar approach is applied by assuming that the data can be modeled as a sum of mixtures and this work associates each mixture with a clustering solution. This leads to the problem of learning a convolution of mixture distributions by which the expectation maximization method can be employed to find the distributions' parameters. Another algorithm called CAMI based on mixture models is developed in [8]. However, instead of trying to orthogonalize two sets of cluster means, CAMI takes into account a more general concept of mutual information to quantify for the decorrelation between two clustering models. The algorithm thus attempts to minimize this quantity while at the same time maximizing the likelihood of each respective model.

In the second approach, an algorithm named COALA is proposed in [3]. Given a known clustering, COALA generates a set of pairwise cannot-link constraints and it attempts to find a disparate data partition by using these constraints within an agglomerative clustering process. The NACI algorithm developed in [9] takes a different approach purely stemming from information theory. Its clustering objective is thus to maximize the mutual information between data instances and cluster labels of the alternative clustering while minimizing such information between that alternate and the provided clustering. However, instead of using the traditional Shannon entropy [5], this work is developed based on the use of Renyi's entropy, with the corresponding quadratic mutual information [14, 24]. Such an approach allows the MI to be practically approximated when combined with the non-parametric Parzen window technique [23]. Recently, this dual-optimized clustering objective is also exploited in work [21] with an iterative approach, in contrast to the hierarchical technique adopted in [9]. Another line in the second approach are the algorithms developed in [6, 10, 25] which address the alternative clustering problem via the use of subspace projection. Work in [6] develops two techniques to find an alternative clustering using orthogonal projections. Intuitively, one can characterize a data partition by a set of representatives (e.g., cluster means). It is then possible to expect that a dissimilar partition might be found by clustering the data in a space orthogonal to the space spanned by such representatives. In the first algorithm in [6], clustering means learnt from a given partition are used as representatives, whilst in the second algorithm, principal components extracted from such centroid vectors are used. A similar approach is developed in [10] in which the transformation is applied on the distance matrix learnt from the provided clustering. Compared to the two methods developed in [6], this work has a benefit that it can further handle the problem of which the data dimension can be smaller than the number of clusters (e.g., spatial datasets). Another approach based on data transformation is developed in [25]. This work attempts to transform the data such that data instances belonging to the same cluster in the reference clustering are mapped far apart in the newly transformed space. Our research in this paper is also to adopt the data projection approach. However, we go beyond these related algorithms by ensuring that the novel projection subspace is not only decorrelated from the provided clustering solution, the local similar property of the data is also preserved in the projection subspace to strongly support uncovering a novel clustering structure behind the data.

## 3 Problem Definition.

We define our problem of generating multiple clustering views from the data as follows.

Given a set $\mathcal{X}$ of $d$-dimensional data instances/observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ and a clustering solution $C^{(1)}$ (i.e., a data partition over $\mathcal{X}$ found

by any clustering algorithm) over $\mathcal{X}$, we seek an algorithm that can learn $C^{(2)}$, a novel clustering solution from $\mathcal{X}$, whose clusters $C_i^{(2)}$ satisfy $\bigcup_i C_i^{(2)} = \mathcal{X}$ and $C_i^{(2)} \cap C_j^{(2)} = \emptyset$ for $\forall i \neq j;\ i,j \leq K$, where $K$ is the number of expected clusters in $C_i^{(2)}$. The quality of the alternative clustering should be high and also be distinctively different from $C^{(1)}$. We will here mostly focus on the case where only a single reference clustering is provided but the framework can be straightforwardly extended to the general case of generating multiple alternative clusterings.

## 4 The Proposed Framework.

**4.1 Subspace Learning Objective with Constraint.** In generating an alternative clustering, our study in this work makes use of a subspace learning approach. Given $C^{(1)}$ as a reference clustering, we aim to map the data from the original space into a new subspace in which it is uncorrelated from $C^{(1)}$ whereas the mapping also well captures and retains certain properties of the data. The dimension of the subspace is usually smaller than the original one and thus the clustering structure behind the data can be more easily uncovered. Clearly, for our clustering problem, we would like that if two instances $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the original space, they should be mapped also close to each other in the lower dimensional space. This idea is also behind several methods including Local Linear Embedding [26], Laplacian Eigenmap [18] and Locality Preserving Projection [12]. However, it is worth mentioning that these algorithms are naturally developed to learn a single manifold embedded in a high dimensional space and to find ways to represent it in an optimal subspace. Our research in this work, despite sharing a similar mapping objective, is clearly different as we aim to seek a set of clusters from the data and more importantly, further require the mapping data to be uncorrelated from one or more reference clusterings.

Following this subspace learning approach, we formulate our problem using graph theory. Let $G = \{V, E\}$ be an undirected graph, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and $E = \{e_{ij}\}$ is a set of edges, each connecting two vertices $(v_i, v_j)$. A vertex $v_i$ corresponds to a data instance $\mathbf{x}_i$ in the dataset $\mathcal{X}$ and the edge $e_{ij}$ between $v_i$ and $v_j$ exists if the respective points $\mathbf{x}_i, \mathbf{x}_j$ are close to each other. Under this setting, the closeness between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be defined using the $\ell$-nearest neighbor concept. Specifically, we define $\mathbf{x}_i$ to be close to $\mathbf{x}_j$ if it is among the $\ell$-nearest neighbors of $\mathbf{x}_j$ or vice versa. Moreover, for the edge connecting two respective vertices $v_i$, $v_j$ of such $\mathbf{x}_i$ and $\mathbf{x}_j$, we associate $K_{ij}$ (computed by the Gaussian function [9]) as a measure of their closeness degree. For two vertices that are not connected, the respective $K_{ij}$ is set to zero. Consequently, we denote $K$ as the $n \times n$ matrix having $K_{ij}$ as its elements and it is easy to observe that $K$ is symmetric and typically sparse, since each vertex is only connected to a limited number of its nearest neighbors.

Given the weight matrix $K$ derived from the graph $G$ and a reference clustering $C^{(1)}$, our objective is to learn a novel set $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, where each $\mathbf{y}_i \in R^q$ is the mapping of $\mathbf{x}_i \in R^d$, *via a projection matrix* $F$, that optimally retains the local neighborhood proximity of the original data yet taking the decorrelated requirement over the reference solution $C^{(1)}$ into account. Essentially, let us denote $X$ and $Y = F^T X$ two matrices respectively having $\mathbf{x}_i$'s and $\mathbf{y}_i$'s as their column vectors and let $\mathbf{f}$ be a column in $F$, then we can consider $\mathbf{f}$ as a transformation vector that *linearly* combines $X$'s dimensions into a 1-dimensional vector $\mathbf{y}^T = \{y_1, \ldots, y_n\} = \mathbf{f}^T X$. That means $\mathbf{y}$ is one feature in the mapping space $Y$. We are now able to define our subspace learning task with the following optimization function:

$$\arg\min_{\mathbf{f}} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{f}^T \mathbf{x}_i - \mathbf{f}^T \mathbf{x}_j)^2 K_{ij} \quad \text{s.t. } S^T X^T \mathbf{f} = 0 \tag{4.1}$$

in which $S$ is a feature subspace that best captures the reference solution $C^{(1)}$ and adding the constant $1/2$ does not affect our optimization objective. The constraint $S^T X^T \mathbf{f} = 0$ (or equivalently $S^T \mathbf{y} = 0$) is crucial since it ensures that the mapping dimension is independent from $S$. Also notice that here we assume $X$ is projected onto $\mathbf{f}$ to form an $R^1$ optimal subspace, yielding 1-dimensional vector $\mathbf{y}^T = \{y_1, \ldots, y_n\}$. The generalization to $q$ optimal dimensions will be straightforward once we derive the solution for $\mathbf{f}$ and subsequently $\mathbf{y}$.

Observing from Eq.(4.1) that if $K_{ij}$ is large (implying $\mathbf{x}_i$ and $\mathbf{x}_j$ are geometrically close in the original space), the objective function will be large if the two respective points $y_i = \mathbf{f}^T \mathbf{x}_i$ and $y_j = \mathbf{f}^T \mathbf{x}_j$ are mapped far apart. Therefore, finding an optimal vector $\mathbf{f}$ that minimizes Eq.(4.1) is equivalent to optimally retaining the local proximity of the data, subject to the constraint $S^T X^T \mathbf{f} = 0$ to ensure the decorrelation from $C^{(1)}$.

**4.2 Fisher Linear Discriminant for $S$ Selection.** We now discuss how to select $S$, a subspace that best captures the provided clustering solution $C^{(1)}$. That means the clusters as components of $C^{(1)}$ can be well discriminated when data is represented in $S$. In achieving this goal, the Fisher's linear discriminant (FDA) can be a natural choice. Briefly, FDA is a

supervised learning technique that seeks a direction $\mathbf{w}$ as a linear combination of the features over $X$ so that the within cluster variances are minimized while at the same time the variances between cluster means and the total sample mean are maximized. Mathematically, its objective function is represented by:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{4.2}$$

of which $S_B$ and $S_W$ respectively being called the between-cluster scatter matrix and within-cluster scatter matrix. They are calculated by:

$$S_B = \sum_k n_k (\mathbf{m}^{(k)} - \mathbf{m})(\mathbf{m}^{(k)} - \mathbf{m})^T \text{ and}$$
$$S_W = \sum_k \sum_i^{n_k} (\mathbf{x}_i^{(k)} - \mathbf{m}^{(k)})(\mathbf{x}_i^{(k)} - \mathbf{m}^{(k)})^T$$

with $\mathbf{m}^{(k)}$ and $\mathbf{m}$ are respectively the cluster mean of the $k$-th cluster and the total sample mean, and $n_k$ is the number of instances grouped in $k$-th cluster. Solving this problem results in that $\mathbf{w}$ is the eigenvector corresponding to the largest eigenvalue of the matrix $S_W^{-1} S_B$. Usually in practice, when the number of clusters is only 2, one may need only a single dimension $\mathbf{w}$ to capture the solution. For a more general case where the number of clusters is $q+1$, a set of $q$ eigenvectors $\mathbf{w}$'s corresponding to the $q$ largest eigenvalues of $S_W^{-1} S_B$ are selected. Therefore in our approach, we choose such $q$ eigenvectors as the optimal subspace $S$ encoding for $C^{(1)}$. Each row in $S$ corresponds to a projected data instance and the number of columns in $S$ equals to the number of retained eigenvectors. Recall that the projected data instances in $S$ strongly support the reference clustering $C^{(1)}$ (i.e., highly correlated to the cluster labels in $C^{(1)}$). Consequently, by taking the orthogonal condition of $S^T\mathbf{y}$, the newly mapped data, the $y_i$'s, should be decorrelated from the reference solution $C^{(1)}$.

**4.3 Solving the Constrained Objective Function.** For solving the objective function with the constraint in Eq.(4.1), we can use the Lagrange method. First, let $D$ be the diagonal matrix with $D_{ii} = \sum_j K_{ij}$ and let $L = D - K$, then by expanding the sum in Eq.(4.1), we can obtain the following function:

$$\begin{aligned} &\sum_{i,j} \mathbf{f}^T \mathbf{x}_i K_{ij} \mathbf{x}_i^T \mathbf{f} - \sum_{i,j} \mathbf{f}^T \mathbf{x}_i K_{ij} \mathbf{x}_j^T \mathbf{f} \\ &= \sum_i \mathbf{f}^T \mathbf{x}_i D_{ii} \mathbf{x}_i^T \mathbf{f} - \mathbf{f}^T X K X^T \mathbf{f} \\ &= \mathbf{f}^T X D X^T \mathbf{f} - \mathbf{f}^T X K X \mathbf{f} \\ &= \mathbf{f}^T X L X^T \mathbf{f} \end{aligned} \tag{4.3}$$

We need to put another constraint $\mathbf{f}^T X D X^T \mathbf{f} = 1$ to remove $\mathbf{f}$'s scaling factor. Then using the Lagrange method with two Lagrange multipliers $\alpha$ and $\beta$, we solve the following function:

$$\mathcal{L}(\alpha, \beta, \mathbf{f}) = \mathbf{f}^T X L X^T \mathbf{f} - \alpha(\mathbf{f}^T X D X^T \mathbf{f} - 1) - \beta S^T X^T \mathbf{f} \tag{4.4}$$

For simplicity, let us denote:

$$\begin{cases} \widetilde{L} = XLX^T \\ \widetilde{D} = XDX^T \\ \widetilde{S} = XS \end{cases} \text{ and}$$

It is easy to verify that $\widetilde{D}$ is symmetric and positive semi-definite. Moreover, there exists $\widetilde{D}^{-1/2}$ and its transpose being identical. We therefore change the variable $\mathbf{f} = \widetilde{D}^{-1/2}\mathbf{z}$. It follows that:

$$\mathbf{f}^T \widetilde{L} \mathbf{f} = \mathbf{z}^T \widetilde{D}^{-1/2} \widetilde{L} \widetilde{D}^{-1/2} \mathbf{z} = \mathbf{z}^T Q \mathbf{z}$$

and the two constraints respectively are:

$$\mathbf{f}^T \widetilde{D} \mathbf{f} = \mathbf{z}^T \mathbf{z} = 1$$
$$\widetilde{S}^T \mathbf{f} = \widetilde{S}^T \widetilde{D}^{-1/2} \mathbf{z} = 0$$

Hence, our Lagrange function can be re-written as follows:

$$\mathcal{L}(\alpha, \beta, \mathbf{z}) = \frac{1}{2}\mathbf{z}^T Q \mathbf{z} - \frac{1}{2}\alpha(\mathbf{z}^T \mathbf{z} - 1) - \beta U^T \mathbf{z} \tag{4.5}$$

of which we have used $U^T$ to denote $\widetilde{S}^T \widetilde{D}^{-1/2}$ and adding the constant $1/2$ does not affect our optimization objective. Taking the derivative of $\mathcal{L}(\alpha, \beta, \mathbf{z})$ with respect to $\mathbf{z}$ and setting it equal to zero give us:

$$\frac{\delta \mathcal{L}}{\delta \mathbf{z}} = Q\mathbf{z} - \alpha \mathbf{z} - \beta U = 0 \tag{4.6}$$

Left multiplying $U^T$ to both sides results in $\beta = (U^T U)^{-1} U^T Q \mathbf{z}$ and substituting it into Eq.(4.6), we derive:

$$\begin{aligned}\alpha\mathbf{z} &= Q\mathbf{z} - U(U^TU)^{-1}U^TQ\mathbf{z}\\ &= \left(I - U(U^TU)^{-1}U^T\right)Q\mathbf{z}\\ &= PQ\mathbf{z}\end{aligned}$$

which is an eigenvalue problem with $P = I - U(U^TU)^{-1}U^T$. It is worth mentioning that $PQ$ might not be symmetric albeit each of its individual matrices being symmetric. However, it is observed that $P^T = P$ and $P^2 = P$, so $P$ is a projection matrix. Consequently, it is true that $\alpha(PQ) = \alpha(PQP)$ or equivalently the eigenvalues of both matrices $PQ$ and $PQP$ are the same. So instead of directly solving $PQ\mathbf{z} = \alpha\mathbf{z}$, we solve $PQP\mathbf{v} = \alpha\mathbf{v}$, with $\mathbf{v} = P^{-1}\mathbf{z}$.

Notice that the eigenvalues $\alpha_i$'s of $PQP$ are always no less than zero and the smallest eigenvalue is indeed $\alpha_0 = 0$, corresponding to the eigenvector $\mathbf{v}_0 = P^{-1}\widetilde{D}^{1/2}\mathbf{1}$, where $\mathbf{1}$ is the unit vector. We thus remove such trivial eigenvalues/vectors from the solution. Consequently, the first nontrivial eigenvector $\mathbf{v}$ will correspond to the smallest non-zero eigenvalue $\alpha$. This leads to our first optimum transformation vector:

$$\mathbf{f} = \widetilde{D}^{-1/2}P\mathbf{v}$$

and subsequently the optimal mapping feature $\mathbf{y}^T = \mathbf{f}^TX$. Generally, in the case where we want to use $q$ transformation vectors to transform $X$ into an $q$-dimensional subspace $Y$, i.e. $Y = F^TX$, we can select the set of $q$ vectors $\mathbf{f} = \widetilde{D}^{-1/2}P\mathbf{v}$ corresponding to the $q$ smallest positive eigenvalues of $PQP$. Similar to the FDA approach, we select $q$ equal to the number of clusters desired for the alternative clustering $C^{(2)}$ minus 1. Given such novel mapping data, we apply the k-means to obtain the alternative clustering $C^{(2)}$.

It is worth mentioning that our algorithm can be extended to find multiple alternative clusterings based on the observation that it aims to find a subspace supporting each clustering solution. Therefore, it is straightforward to include all subspaces of previously found solutions as columns in the $S$ matrix when searching for a novel alternative clustering. Certainly, the number of alternative clusterings can be given by the user or we can iterate the process until the total sum of square distances (computed in k-means) for a novel clustering is significantly larger than those of previously found clusterings.

## 5 Experiments.

In this section, we provide some initial experimental results regarding our method, which we name ACCP (Alternative Clustering with Constrained Projection) on synthetic and real-world datasets. Since our algorithm explores the subspace transformation approach, we mainly compare it against other techniques which also adopt this direction. In particular, its clustering performance is compared against two methods from [6] which we denote by Algo1, Algo2 respectively and the ADFT algorithm from [10]. For ADFT, we implement the gradient descent method integrated with the iterative projection technique (in learning the full family of the Mahalanobis distance matrix) [29, 30]. For all algorithms, including ours, we use k-means as the clustering algorithm applied in the transformed subspace.

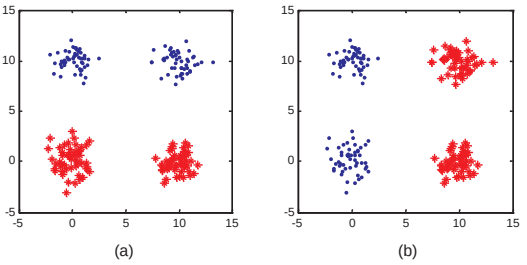


Figure 1: Alternative Clustering returned by our algorithm on the synthetic dataset (best visualization in color)

We evaluate the clustering results based on clustering dissimilarity and clustering quality measures. For measuring the dissimilarity/decorrelation between two clusterings, we use the normalized mutual information (NMI) that has been widely used in [11, 17, 27] and the Jaccard index (JI) which is used in [3, 10]. For measuring clustering quality, we use the Dunn Index (DI) [3, 10], which is defined by $DI(C) = \frac{\min_{i\neq j}\{\ \delta(c_i,c_j)\}}{\max_{1\leq\ell\leq k}\{\triangle(c_\ell)\}}$ where $C$ is a clustering, $\delta:\ C\times C \to \mathbb{R}_0^+$ is the cluster-to-cluster distance and $\triangle: C\to\mathbb{R}_0^+$ is the cluster diameter measure. Note that for the NMI and JI measures, a smaller value is desirable, indicating higher dissimilarity between clusterings, whereas for the DI measure, a larger value is desirable, indicating a better clustering quality.

**5.1 Experiments on Synthetic Data.** For testing the performance of our algorithm on a synthetic dataset, we take a popular one from [10, 3] that is often used for alternative clusterings. This dataset consists of 4 Gaussian sub-classes, each containing 200 points in a 2-dimensional data space. The goal of using this synthetic dataset, when setting $k = 2$, is to test whether our algorithm can discover an alternative clustering that is

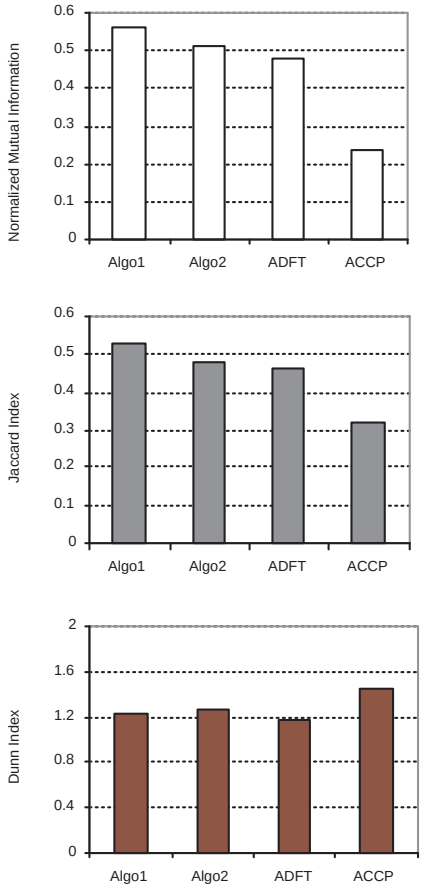


Figure 2: Clustering performance returned by four algorithms on the Cloud data

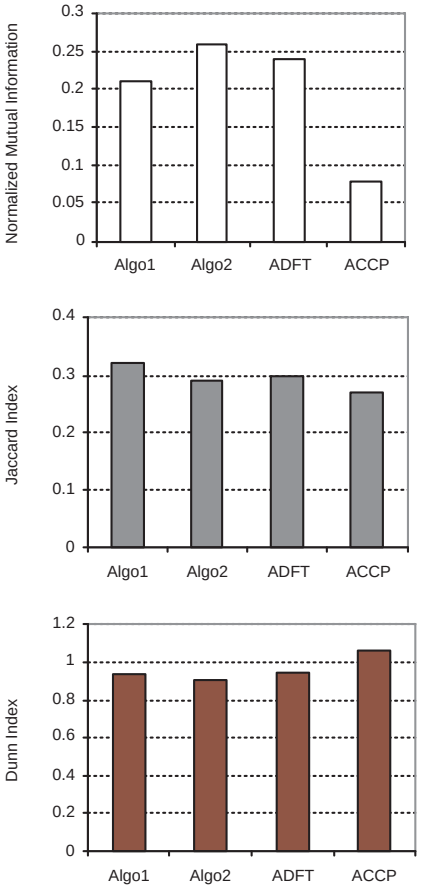


Figure 3: Clustering performance returned by four algorithms on the Housing data

orthogonal to the existing one. In Figure 1, we show the clustering results returned by our algorithm. Given the reference clustering $C^{(1)}$ which groups two Gaussians on the top and two Gaussians in the bottom as each cluster (shown in Figure 1(a)), ACCP successfully uncovers the alternative clustering $C^{(2)}$ which categorizes two Gaussians on the left and two ones on the right as clusters (shown in Figure 1(b)).

**5.2 Experiments on the UCI data.** We use two real-world datasets from the UCI KDD repository [2] to compare the performance of our algorithm against the other techniques. The first dataset is the Cloud data which consists of data collected from a cloud-seeding experiment in Tasmania in the time between mid-1964 and January 1971. For a reference clustering, we run the k-means algorithm (with $K = 4$ as the number of clusters) and its resultant clustering is used as the reference solution for all algorithms. The second dataset is the Housing data which consists information about the housing values collected in Boston's suburbs. Similar to the Cloud data, we apply k-means algorithm on this dataset and use it as the reference solution for all algorithms. We show the clustering performance of all techniques in Figures 2 and 3 respectively for the Cloud and Housing data.

As observed from all figures, our algorithm performs better than two techniques developed in [6] and the ADFT one in [10]. Its clustering dissimilarity measuring by the normalized mutual information and the Jaccard Index is lower than that of Algo1, Algo2 and ADFT whereas the clustering quality quantified by the Dunn Index is higher. This advantage can be justified by the approach of ACCP. Though all algorithms adopt a linear projection technique to derive a novel subspace that is decorrelated from the reference clustering, the

ACCP further takes into account the local neighborhood proximity of the data by retaining that property in the novel low dimensional subspace. The novel clustering structure in this learnt subspace therefore is more prominent compared to that of the other techniques which only attempt to minimize the correlation between two subspaces. Moreover, though the derived subspace of all techniques is ensured to be independent from the provided clustering, it is still possible that the alternative learnt from that subspace might not strongly decorrelated since there is a need to optimize for the clustering objective function as well. This generally explains for the difference in the NMI and JI measures of all methods.

## 6 Conclusions and Future Studies.

In this paper, we have addressed an important problem of uncovering multiple clustering views from a given dataset. We propose the ACCP framework to learn a novel lower dimensional subspace that is not only decorrelated from the provided reference clustering but the local geometrical proximity of the data is also being retained. By the second objective, our work goes beyond the others, which also adopt a subspace learning approach for seeking alternative clusterings, yet only focus on the dissimilarity property of the novel clustering. More importantly, we have demonstrated our dual-objective can be formulated via a constrained subspace learning problem of which a global optimum solution can be achieved.

Several initial empirical results of the proposed framework over the synthetic and real world datasets are given. Certainly, these results remain preliminary and more experimental work should be done in order to provide more insights into its performance. Additionally, though our framework is claimed to be capable of finding multiple clustering views, its performance on some suitable datasets has not been verified. Determining an ideal number of alternative clusterings still needs to be formally addressed. We leave these problems as an important and immediate task of the future work.

In addition to the above issues, our research in this paper opens up some potential and interesting directions for future studies. Specifically, despite advancing the subspace based learning approach for the alternative clustering problem, our research has exploited a hard constraint solution to ensure alternatives' dissimilarity. Such an approach might be too strict in some practical applications and thus some novel approaches based on soft constraints could be interesting. This implies that a trade-off factor between two criteria of subspaces' independence and intrinsic data structure retaining can be introduced, or we can constrain the subspace's decorrelation criteria to no less than a given threshold. In either of the two cases, a data-adaptive learning technique is required and the compromise between two factors is worth to study both theoretically and empirically. Furthermore, our research has not yet focused on the interpretation regarding the resultant alternative clusterings. This is an important problem, especially from the user perspective. Which features are best to describe a clustering solution and which ones are least correlated with respect to that solution are all informative to the user in understanding the data. Compared to a nonlinear subspace learning approach [7], our research direction in this work is beneficial due to its linear subspace learning approach. That means the linear combination amongst the original features is explicitly represented in the transformation matrix $F$. However, seeking for an optimal and concise set of features that best characterizes for a clustering solution is still a challenging issue given the fact that the number of original data features is usually huge. We believe that these open issues are worth to be further explored and studied.

## References


[1] I. Assent, E. Müller, S. Günnemann, R. Krieger, and T. Seidl. Less is more: Non-redundant subspace clustering. In *1st MultiClust International Workshop in conjunction with 16th ACM SIGKDD.*, 2010.

[2] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[3] E. Bae and J. Bailey. COALA: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *The IEEE International Conference on Data Mining.*, pages 53–62, 2006.

[4] G. Chechik and N. Tishby. Extracting relevant structures with side information. In *International Conference on Neural Information Processing Systems (NIPS)*, 2002.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, August 1991.

[6] Y. Cui, X. Fern, and J. Dy. Non-redundant multi-view clustering via orthogonalization. In *The IEEE International Conference on Data Mining*, pages 133–142, 2007.

[7] X.H. Dang and J. Bailey. Generating multiple alternative clusterings via globally optimal subspaces. Under Review.

[8] X.H. Dang and J. Bailey. Generation of alternative clusterings using the cami approach. In *SIAM International Conference on Data Mining (SDM)*, pages 118–129, 2010.

[9] X.H. Dang and J. Bailey. A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In *ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 573–582, 2010.

[10] I. Davidson and Z. Qi. Finding alternative clusterings using constraints. In *The IEEE International Conference on Data Mining*, pages 773–778, 2008.

[11] X. Fern and W. Lin. Cluster ensemble selection. *Stat. Anal. Data Ming*, 1(3):128–141, 2008.

[12] X. He and P. Niyogi. Locality preserving projections. In *International Conference on Neural Information Processing Systems (NIPS)*, 2003.

[13] P. Jain, R. Meka, and I. Dhillon. Simultaneous unsupervised learning of disparate clusterings. In *SIAM International Conference on Data Mining (SDM)*, pages 858–869, 2008.

[14] J. Kapur. *Measures of Information and their Application*. John Wiley, 1994.

[15] H.-P. Kriegel, E. Schubert, and A. Zimek. Evaluation of multiple clustering solutions. In *2nd MultiClust International Workshop in conjunction with ECML/PKDD*, 2011.

[16] H.-P. Kriegel and A. Zimek. Subspace clustering, ensemble clustering, alternative clustering, multiview clustering: What can we learn from each other. In *1st MultiClust International Workshop in conjunction with 16th ACM SIGKDD*, 2010.

[17] M. Law, A. Topchy, and A. Jain. Multiobjective data clustering. In *CVPR Conference*, pages 424–430, 2004.

[18] B. Mikhail and N. Partha. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 585–591, 2001.

[19] E. Müller, S. Günnemann, I. Färber, and T. Seidl. Discovering multiple clustering solutions: Grouping objects in different views of the data (tutorial). In *SIAM International Conference on Data Mining (SDM)*, 2011.

[20] Emmanuel Müller, Ira Assent, Stephan Günnemann, Patrick Gerwert, Matthias Hannen, Timm Jansen, and Thomas Seidl. A framework for evaluation and exploration of clustering algorithms in subspaces of high dimensional databases. In *BTW*, pages 347–366, 2011.

[21] X. V. Nguyen and J. Epps. minCEntropy: a novel information theoretic approach for the generation of alternative clusterings. In *The IEEE International Conference on Data Mining*, pages 521–530, 2010.

[22] D. Niu, J. G. Dy, and M. I. Jordan. Multiple non-redundant spectral clustering views. In *ICML*, pages 831–838, 2010.

[23] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[24] J. Principe, D. Xu, and J. Fisher. *Information Theoretic Learning*. John Wiley & Sons, 2000.

[25] Z. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. In *ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 717–726, 2009.

[26] T. R. Sam and K. S. Lawrence. Nonlinear dimensionality reduction by locally linear embedding. *Science Journal*, 290(5500):2323–2326, 2000.

[27] A. Topchy, A. Jain, and W. Punch. A mixture model for clustering ensembles. In *SDM Conference*, 2004.

[28] J. Vreeken and A. Zimek. When pattern met subspace cluster - a relationship story. In *2nd MultiClust International Workshop in conjunction with ECML/PKDD*, 2011.

[29] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *International Conference on Neural Information Processing Systems (NIPS)*, pages 505–512, 2002.

[30] L. Yang and R. Jin. Distance metric learning: A comprehensive survey, 2006.