Identifying Information-Rich Subspace Trends in High-Dimensional Data

Snehal Pokharkar[*] Chandan K. Reddy[†]

Abstract

Identifying information-rich subsets in high-dimensional spaces and representing them as order revealing patterns (or trends) is an important and challenging research problem in many science and engineering applications. The information quotient of large-scale high-dimensional datasets is significantly reduced by the curse of dimensionality which makes the traditional clustering and association analysis methods unsuitable. Most interesting patterns cannot be revealed using global methods which consider the entire data and feature spaces during their analysis. Identifying some interesting patterns in large scale high-dimensional data is usually accomplished using popular techniques such as dimensionality reduction, feature selection and subspace clustering. Though these methods are successfully able to identify the groupings in the feature subsets and localized neighborhood data subspaces, none of these methods extract the latent patterns that are present in local information-rich subsets of the data. In this paper, we seek an information-revealing representation of the data subsets and features that may contain local patterns. We formalize the problem of identifying '*subspace trends*' in high-dimensional datasets focusing on information-rich subsets and develop a new algorithm to extract such subspace trends. We demonstrate our results on both synthetic and real-world datasets and show the superiority of the proposed methodology over traditional clustering and dimensionality reduction techniques.

Keywords: Clustering, subspaces, dimensionality reduction, trend analysis, regression, feature selection

1 Introduction

With the advancements in data collection and storage technologies, there has been an exponential increase in the availability and usage of large, high-dimensional datasets. Many practical (biomedical, financial, web transactions and others) applications produce datasets that contain thousands of records and several hundreds of features. In such a high-dimensional space, it is a tedious task to identify the continuous structural patterns indicating the correlation in the data in a reduced subspace of data within only the relevant set of features. High-dimensional datasets, although high in global information quotient, do not reveal many locally relevant correlations with respect to features and subsets of data points. In many of these datasets, it is certainly important to be able to identify those subsets of data (and features) which form locally relevant patterns and also be able to identify which features contribute to this phenomenon. This would enable researchers to focus their attention on these local subsets and make it easy to identify the important and most informative aspects of the data. One of the many objectives of data exploration is to find correlation in the data, uncovering hidden patterns and trends in the data distribution, thus providing additional insights about the data [22, 21]. To achieve these goals, one of the effective solution is to order the data and give a continuous representation. Such an ordered dataset will reveal a non-discrete structural pattern indicating the correlation in the data, which can be further analyzed to extract inferences that were previously unknown.

The main goal of this paper to identify such hidden local patterns or 'trends' over a subset of datapoints within the reduced feature subset. This work is an extension of our previous work on identifying trends to local subspaces [17]. As we consider a subspace of data points and features, the patterns found are hence referred to as '*subspace trends*'. Identifying such subspace trends is a three-fold problem:

1. We need to identify the dominant features that lead to creation of such patterns.

2. We must limit the data points to a locally relevant subsets.

3. We need to formulate 'trends' in these subspaces (of data and features) and represent them using a reduced dimension space, if necessary.

Our method finds trends containing strongly correlated data points present in locally relevant data spaces and not in the entire global feature space which might or might not yield optimal trends. Some datasets are also predominantly informative only in subsets of locally relevant data points. If we find 'global trends' in such datasets, the ordering (or correlation) could be very weak and the trends obtained might be less informative. Hence, it is essential to focus on finding local correlations and ordering in such high-dimensional datasets.

[*] Department of Computer Science at Wayne State University.

[†] Corresponding Author. E-mail Address: reddy@cs.wayne.edu. Department of Computer Science at Wayne State University.

The proposed solution is initiated with an unsupervised subspace clustering method providing a similarity preserving base to build a structure for further analysis. Subspace clustering identifies the feature subset and the data points that will form the most fundamental building blocks of the 'subspace trends'. Subspace clusters that share a common feature space are then combined to form what are termed as *'Information-Rich Subsets'* (IRS) of the data. Using a graph-based framework, each IRS is represented by local proximity structures of the data points. This representation is used to generate the hypotheses of trends and the most prominent trends are then selected. These hypotheses are then weighted based on an objective function and the optimal trends that have complete coverage of the Information-Rich Subset are chosen [17]. Such trends represent the ordering and continuity information of the data points and have the potential to explain the linear or non-linear correlations in the subspaces. This continuous representation will provide a more powerful model of these subspaces compared to the traditional subspace cluster representation.

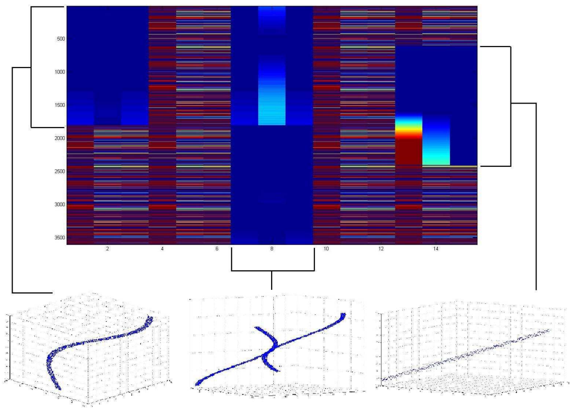
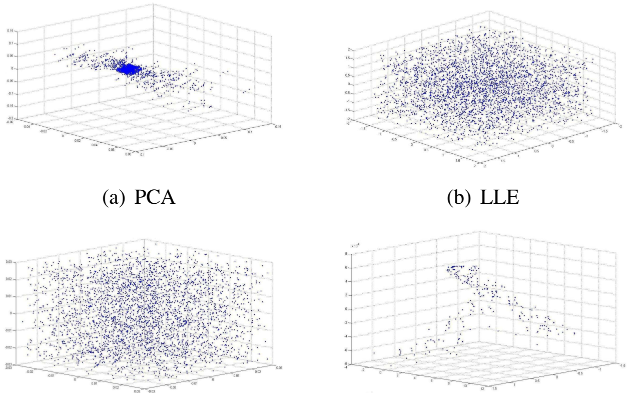


Figure 1: Correlation matrix representation of a simple synthetic dataset that contains local correlations.

1.1 Motivating Example: Consider a simple synthetic dataset (shown in Fig. 1) with 15 features $\{F_1,...,F_{15}\}$ and 3602 data points $\{DP_1,...,DP_{3602}\}$. Features $\{F_1,F_2,F_3\}$ spread over data points $\{DP_1,...,DP_{1801}\}$ have embedded in them a non-linear trend (sine wave). Features $\{F_7,F_8,F_9\}$ and data points $\{DP_1,...,DP_{3602}\}$ contain latent intersecting sine waves. Features $\{F_{13},F_{14},F_{15}\}$ and data points $\{DP_{600},...,DP_{2401}\}$ contain a latent linear pattern in them. The rest of the data points and features are random noise. The results of various dimensionality reduction methods on this dataset are shown in Fig. 2. Classical methods such as Principal Component Analysis (PCA) [19], Multidimensional Scaling (MDS) [8], Locally Linear Embedding (LLE) [18], Laplacian Eigenmap (LE) [4] and Isometric Mapping (Isomap) [12] are unable to extract the locally embedded 'subspace trends'. These dimensionality reduction methods used for linear and non-linear embedding of the data can only interpret the hidden geometry of the entire dataset. They are unable to provide any information about the local subspace trends. Clearly, it is important to be able to extract such local subspace trends in large high-dimensional datasets. Our method seeks to extract such local 'subspace trends' that are spread over a subset of features and datapoints. These subspace trends can be further analyzed with respect to their inherent properties such as length, continuity, overlap, coverage and ordering, amongst others. Various trends identified in a given dataset can be studied and compared for different variabilities in the features and their effect on data distributions.



(a) PCA

(b) LLE

(c) Laplacian Eigenmap

(d) Isomap

Figure 2: Results of dimensionality reduction methods on the synthetic dataset.

Analyzing individual trends can yield more information about the structural arrangement of the data points with some continuity information. We are able to represent data subsets in order-preserving patterns by identifying optimal 'subspace trends' which might yield some useful insights for domain experts for any further analysis. The rest of this paper is organized as follows: Section 2 explains the shortcomings of different methods proposed in the literature. Section 3 describes the problem formulation and explains the key concepts needed to comprehend our algorithm. The proposed algorithm for finding subspace trends along with some implementation details is given in Section 4. Section 5 shows the experimental results of the proposed algorithm on various synthetic and real-world datasets. Section 6 concludes our discussion and gives the future research directions.

2 Related Work

We discuss some important methods studied widely in literature which are relevant to our problem. Specifically, we explain some of the widely used techniques like dimensionality reduction, feature selection and subspace clustering and discuss their shortcomings.

2.1 Dimensionality Reduction: Dimensionality reduction has become a well studied topic in data mining and statistics to find meaningful interpretations of the data sets from the high-dimensional representation. Dimensionality reduction is one of the areas which attempts to extract the meaningful dimensions from the large pool of features or to develop a meaningful low-dimensional representation that will preserve the information quotient of the data in the entire feature space. Classical methods used for linear dimensionality reduction that are widely used in many practical applications are PCA and MDS. PCA is an eigenvector based dimensionality reduction method which preserves the variance from the high-dimensional setting of the data and represents it in a new low-dimensional coordinate system. MDS is a distance preserving technique that preserves the pairwise distance between the data points while representing the embedding in a low dimensional space. Both PCA and MDS, though widely in many applications, can only produce a linear mapping into a low-dimensional space. But there are many datasets where the underlying variability of the features creates a highly non-linear structure. For these datasets, methods such as ISOMAP, LLE, Laplacian Eigenmap and other manifold learning algorithms focus on preserving the inherent structural geometry of the dataset if the data lies on a subspace manifold. All these methods are geometry preserving dimensionality reduction methods which are able to identify the hidden structure of the entire dataset and preserve it in a low-dimensional space. They are unable to extract the locality preserving structures that are present in the subspaces of the datasets. Although these methods succeed in identifying the structure, they are essentially dimensionality reduction methods and do not identify the trends that may be present in the subspaces. They are able to only provide a guideline to generate a basis to do preliminary investigation about any positive correlations and cannot give any information about some of the subspace correlations hidden in these datasets.

2.2 Feature Selection: In the high-dimensional feature space, feature selection methods will allow us to extract those relevant (useful) features and separate them out from the redundant, repetitive and noisy features [5, 15]. Feature Selection algorithms seek to identify such a subset of features primarily to improve the interpretability and reduce the unnecessary complexities that might arise during the data mining process. All the same, these algorithms can only extract the relevant features, but are not capable of selecting a subset of data points or provide any information about some of the local latent structures, geometry or ordering of the data points within those feature sets. A more comprehensive study about the different feature selection algorithms proposed in the literature is given in [14].

2.3 Subspace Clustering Clusters of data are identified in the entire data/feature space and this is not suitable for high-dimensional spaces that contain many irrelevant features [16]. In order to avoid sub-optimal cluster formation, subspace clustering algorithms find locally relevant clusters in subspaces in a low-dimensional feature space. Based on the search strategy of the method, these algorithms are classified as top-down approaches or bottom-up approaches. Top-down approaches [2, 1, 10] perform clustering in the original space and iteratively evaluate the clusters based on their subspaces to identify the most similar data points in reduced feature space. Bottom up approaches [3, 9, 6] find locally dense data subspaces and iteratively combine them to form clusters based on improving the quality of the clusters. Although successful in grouping data points, subspace clustering algorithms do not provide any continuous representation of latent patterns in these subspaces. In our algorithm, we use subspace clusters of the dataset to identify the information-rich data subsets.

The two main drawbacks of subspace clustering algorithms that motivated the need for the proposed methodology are that the subspace algorithms:

- Simultaneously optimize the data and features to obtain localized clusterings and in this process, they tend to provide local dense clusters and do not preserve patterns in the data.

- Give a discrete set of clusterings which are hard to interpret. Especially, when one is looking for certain correlation patterns it is important to group some of these subspace clusters to represent trends in the data.

In essence, we extend the notion of subspace clusters to ‘subspace trends’ in this paper. We solve these problems by merging similar subspace clusters and identify trends in these subspaces.

3 Problem Description

In this section, we provide the mathematical formulation of the problem of identifying Information-Rich Subspaces (feature subsets and data subspaces) followed by generating ‘subspace trends’ in them. We will now formally define the notion of *Information-Rich Subspaces - IRS* and ‘subspace

Table 1: Notations used in this paper.

Notation	Description
n	Number of datapoints
m	Number of features
F	Feature Set F=$\{F_1$,...,$F_m\}$
DP	Data Points DP=$\{DP_1$,...,$DP_n\}$
X	Input dataset X = $\{F$,$DP\}$
f_i	Subset of Features for i^{th} subspace cluster f_i = (F_i,...,F_*,...,F_l)$\subset F$
dp_i	Subset of Data Points i^{th} subspace cluster dp_i = (DP_k,...,DP_*,...,DP_r)$\subset DP$
SC_i	Subspace Cluster SC_i = $\{f_i$,$dp_i\}$
IRS_I	Information-Rich Subset IRS_I = $\{SC_i \cup SC_j \cup ... \cup SC_m\}$ for $f_i \approx f_j \approx ... \approx f_m$
k	Number of cluster centroids
CC_i	Cluster centroids
tn_i	Terminal Nodes of the Pattern Graph
cn_i	Composite Nodes of the Pattern Graph
CL_{ij}	Edges connecting two Centroids CC_i and CC_j
PG	Pattern Graph : MST connecting CC_is with CL_{ij}s
SP_I	Subpaths of PG = $(tn_i, ..., cn_*, ..., tn_j)$
AM	Adjacency matrix of the MST
l_I	Length of SP_I
c_I	Curvature of SP_I
w	Input weight parameter for Γ_I
Γ_I	Subpath Selection Factor (w*l_I+(1-w)*c_I)
Υ	Subspace Trends

trends' and then provide the problem statement. Table 1 gives the notations used in this paper.

3.1 Information-Rich Subspaces *(IRS_I)* is a subset of selected set of features (F) and data points (DP), extracted from the original dataset such that a strong local structure is present. It is comprised of very dense regions in the selected feature set. This forms the building blocks for the identification of 'subspace trends' which are embedded in that subspace.

The *Subspace Clusters (SC_i)* are the groups of those data points that have strong similarity in a selected feature subspace [16]. A high-dimensional dataset X = $\{F$,$DP\}$ can contain several subspace clusters denoted by SC_i = $\{f_i$,$dp_i\}$ where $f_i \subset F$ and $dp_i \subset DP$. These subspace clusters have locally dense and highly correlated data points and yet, only by themselves, do not provide any continuity information which is essential for extracting correlation information. Subspace clusters are often small in size (have fewer data points in comparison to the original dataset) and at best give the similarity information in a local neighborhood. The simple subspace grouping of these clusters has to be enhanced and made more coherent before they can be further analyzed for any 'subspace trends'. This forms the basis on which we will identify the Information-Rich Subsets in the dataset.

IRS are the union of those similar subspace clusters that share some common feature subspaces. IRS_I = $\{SC_i \cup SC_j \cup ... \cup SC_m\}$ for $f_i \approx f_j \approx ... \approx f_m$. IRS are a collection of similar subspace clusters sharing some common features and contain significant number of data points compared to any independent subspace cluster and offer more flexibility for the detection of subspace trends. Each IRS is composed of a selected and highly correlated set of features and only those data points that have a high similarity value with each other spread over those features that are correlated in that subspace are selected to belong to the IRS. This grouping makes them more informative and are hence referred to as 'Information-Rich Subsets' of data.

3.2 Subspace Trends (Υ) are the inherent, latent sequences of patterns present in the subsets of datasets. They are the ordered representation of the locally informative subset of data points. These patterns are represented in a continuous form in a feature space that is filled with data points to a pre-designated density in subspaces. Each latent pattern has properties of length and curvature based on which they are selected and identified as 'subspace trends'. We will now introduce some of the terminologies used in this paper.

DEFINITION 1. ***Pattern Graph** (PG) is a Minimum Spanning Tree(MST) representation of the Cluster Centroids (CC_i) joined by the edges (CL_{ij}).*

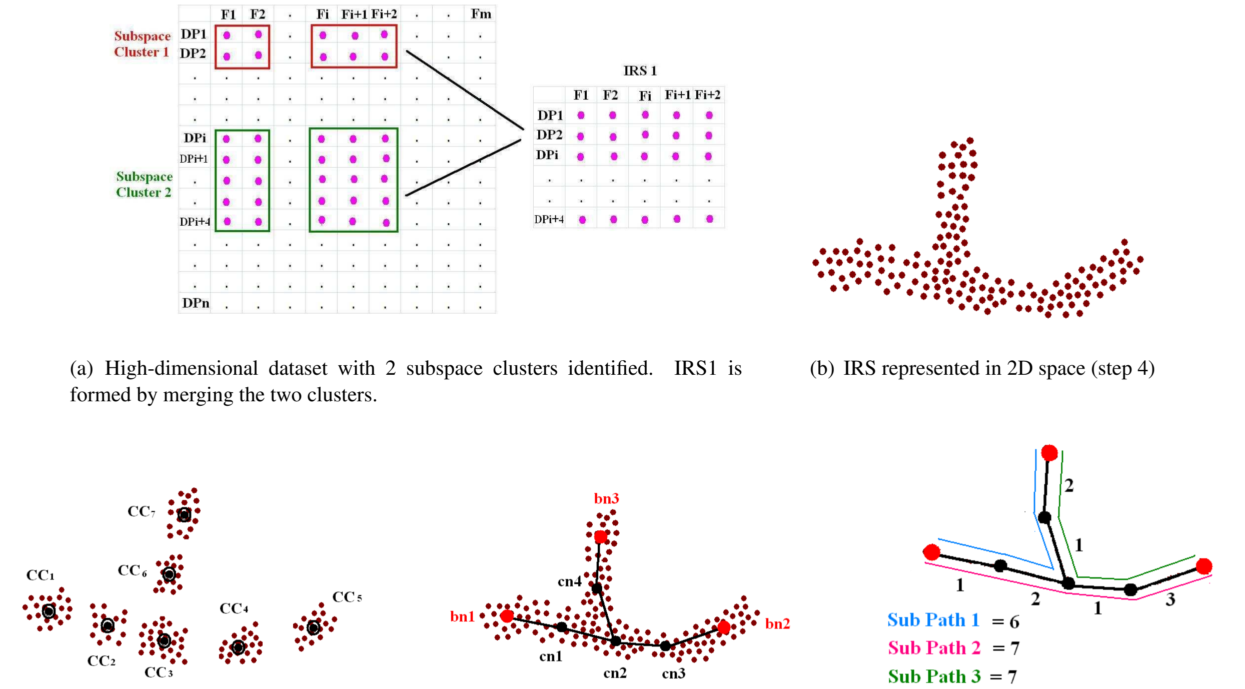
The MST offers a neighborhood preserving and mapping based structural representation of the Cluster Centroids. This forms the pattern graph from which the hypothesis of 'subspace trends' are identified and explored.

DEFINITION 2. ***Terminal Nodes** (tn_i) are those CC_i which have just one edge in PG ($|CL_i| = 1$).*

DEFINITION 3. ***Composite Nodes** (cn_i) are those CC_i which are not the Terminal nodes. They have 2 or more edges. $|CL_i| \geq 2$.*

DEFINITION 4. ***SubPaths** (SP_{ij}) are those paths in PG that contain several Composite Nodes (cn_i) which are bounded by exactly two Terminal Nodes (tn_i and tn_j).*

The Terminal nodes are the end nodes of various possible subspace trends. The entire 'subspace trend' is bound between these tn_i's. The subpaths are represented as (tn_i, cn_1, cn_2, ..., cn_m, tn_j). The total number of SubPaths depends on the number of Terminal Nodes in PG. There are $\binom{n}{2}$ subpaths for n Terminal Nodes.



(a) High-dimensional dataset with 2 subspace clusters identified. IRS1 is formed by merging the two clusters.

(b) IRS represented in 2D space (step 4)

(c) Various Clusters and their Centroids (step 5)

(d) Minimum Spanning Tree on IRS (step 6)

(e) The subpaths and their lengths as identified from the MST (steps 7,8).

Figure 3: Demonstration of the different steps of our algorithm.

Length (l_I) of SP_I - is the Euclidian Distance measured between the Terminal Nodes, as we move along the subpath over each one of its Composite Node in the order they are traversed.
Curvature (c_I) of SP_I - is the measure of the non-linearity of the subpath. It is calculated as the summation of the cos values of each Composite Nodes present on the subpath and normalized by averaging it over the length of the subpath. For a given subpath SP_I =(tn_1,cn_1,cn_2,cn_3,tn_2) and l_I, curvature c_I is calculated as follows:

$$(cos(cn_1)+cos(cn_2)+cos(cn_3))/l_I \tag{3.1}$$

We illustrate these concepts with an example shown in Fig. 3 which on a high-dimensional dataset $X = \{F,DP\}$. Fig. 3(a) represents two distinct subspace clusters over the same feature subset. They are merged to form IRS1 which now is the Information-Rich dense Subset of X. Fig. 3(b) is the 2-dimensional representation of IRS1. Fig. 3(c) shows the results of the clustering algorithm where the IRS data points are clustered and represented by their cluster centroids (CC_1,...,CC_7). Fig. 3(d) shows the Pattern Graph which joins all the Cluster Centroids. There are 3 Terminal Nodes (tn_1,tn_2,tn_3) and 4 Composite Nodes (cn_1,cn_2,cn_3,cn_4). Based on the Terminal nodes and Composite Nodes, there are 3 possible subpaths, shown in Fig. 3(e), along with the lengths of each subpaths is measured based on the edges of the Nodes belonging to that subpath.

Problem Statement: Given a Pattern Graph PG =(CC_I, CL_{IJ}), we would like to identify all the end to end SubPaths SP_I and find the minimal set of SP_Is that cover all the CC_is in PG and maximize the objective function Γ_I given below:

$$\Gamma_I = w * l_I + (1-w) * c_I \tag{3.2}$$

where the parameter Path Selection Factor (Γ_I) is the weighted sum of $\{l_I$,$c_I\}$ for each subpath and is used to optimize the selection criteria for the subpaths. Such a subset of optimal most informative subpaths form the ‘subspace trends’ that represent the correlations in the subspaces.

4 Algorithm for Finding Subspace Trends

We will now propose the following algorithm to generate subspace trend hypotheses in an IRS:

1. *IRS Identification* - using a subspace clustering algorithm with a data merging subroutine which assimilates the data points and relevant features to form Information-Rich Subspaces (IRS).

2. *Pattern Graph Generation* - using a local similarity preserving algorithm for clustering of IRS followed by

fitting a MST to connect the local cluster centroids in each IRS.

3. *Identification of Independent Paths* - implementing *Find_Trends* algorithm to obtain different 'subspace trends' hypotheses.

The high-level pseudocode is shown in Algorithm 1. Now, we will elaborate on each of the steps in the proposed algorithm mentioned above.

Algorithm 1 *Finding_Subspace_Trends*

```
**Input:** Dataset X
**Output:** Subspace Trends $\Upsilon$
**Algorithm:**
IRS $\leftarrow IRS\_Identification(X)$
$\sigma \leftarrow local\_clusters(IRS)$
$AM \leftarrow fit\_MST(\sigma)$
$[SP_I, l_I, c_I] \leftarrow SubPath\_Values(AM)$
$SubP \leftarrow \{SP_I, \Gamma_I\}$ (using Eq. (3.2))
$\Upsilon \leftarrow Find\_Trends(SubP)$
```

(1) IRS Identification - As the first step, we use a subspace clustering algorithm to identify the subspace clusters present in the entire data and feature space. The output of this module is the subspace cluster ID (SC_i), the data points (dp_i) and the feature subset (f_i) of each subspace cluster. Potentially, any subspace algorithm that is suitable for a given dataset can be used to obtain these subspace clusters. The algorithm returns the biclusters written in an output file. In our implementation, we used the BiVisu[1] biclustering algorithm, which is an efficient method for finding subspace clusters using parallel coordinate visualization [7].

Each subspace cluster has a distinct subset of features associated with it. Based on the collective number of data points belonging to each such subset of features, the clusters are identified and merged to form an IRS. There are as many IRS's as many distinct feature subsets of the subspace clusters. For example, let us consider the 4 distinct subspace clusters $SC_1 = \{f_1, dp_1\}$, $SC_2 = \{f_2, dp_2\}$, $SC_3 = \{f_3, dp_3\}$ and $SC_4 = \{f_4, dp_4\}$. If $f_1 \approx f_2$ and $f_3 \approx f_4$, then as the distinct and unique feature sets are 2 (f_1 and f_3), there will be only 2 IRS's. The routine UNIQUE(f) identifies this unique set of features. Based on this number of unique feature sets, as described in Algorithm 2, the IRS's are identified and used for further analysis.

Algorithm 2 *IRS_Identification*

```
**Input:** Subspace Clusters $SC_i, SC_j ...$
**Output:** Information-Rich Subspaces $IRS_I$
**Pseudocode:**
UniqueFS $\leftarrow$ UNIQUE(f)
$IRS = \phi$
**for** I$\leftarrow$ 1 : size(UniqueFS) **do**
    F = $UniqueFS(I)$
    **for** i$\leftarrow$ 1 : size($SC$) **do**
        **if** $f_i \approx$ F **then**
            $IRS_I \leftarrow IRS_I \cup dp_i$
        **end if**
    **end for**
**end for**
```

(2) Pattern Graph Generation - The building blocks of the IRS must be identified in order to formulate a continuity based Pattern Graph. To this end, the IRS is clustered into a suitable number (k) of clusters using any appropriate clustering algorithm (such as k-means or hierarchical) which groups each IRS into desired local clusterings needed for the next step of our algorithm. Each cluster is represented by its centroid CC_i. The clusters provide the similarity preserving basis on which the subspace trends hypotheses can be built while the centroids are the single point summarizations. We choose k-means for its high run-time efficiency, simplicity and its ability to provide good local (spherical) clusterings. Once the building blocks (clusters represented by their centroids (CC_i)) have been identified, a continuity based structure representation of the internal neighborhood preserving mapping of the datapoints is generated by fitting a MST (using Kruskal's algorithm) on the cluster centroids which will give a graph with edges CL_i (joining the cluster centroids CC_i) in the form of an Adjacency matrix (AM). Based on this adjacency matrix, the algorithm *SubPath_Values* identifies all the possible subpaths (end to end paths - SP_I) along with their lengths (l_I) and curvatures (c_I).

(3) Identifying the Optimal Trends - This is one of the main components of our approach that will examine every possible subpath and identifies a minimal subset which contains the entire IRS within themselves while optimizing the properties of length and curvature represented in their structure. *Find_Trends* is a greedy procedure which takes the subpaths matrix (SubP = $\{SP_I, \Gamma_I\}$) and the Path Selection Factor (Γ_I) (see Eq. (3.2)). Each subpath is examined in the descending order of their Γ_I values. The first subpath is automatically selected and is also the best representative of the 'subspace trends' of that IRS. This is the most optimal path with respect to length and curvature. Any new subpath that needs to be considered as a trend must include at least one new cluster centroid that is not present in the already selected subpaths. After all the cluster centroids are considered in the set of subpaths, the algorithm is terminated with

[1] www.eie.polyu.edu.hk/~nflaw/Biclustering/

selected subpaths as the 'subspace trends.'

Algorithm 3 *Find_Trends*

```
Input: Subpath matrix SubP
Output: Set of Optimal Subpaths $\tau$
Pseudocode:
sort (SubP, $\Gamma_I$)
Ind = $\sigma$ // Set of Indicator variables for $CC_{is}$
for I=1:size(SubP) AND Ind $\neq \phi$ do
    $SPC_I \leftarrow Get\_CC(SP_I)$
    for j=1:size($SPC_I$) do
        for k=1:size(Ind) do
            if $SPC_I$(j) =Ind(k) then
                Ind(k)$\leftarrow \phi$
                $\tau(i) \leftarrow SP_I$
            end if
        end for
    end for
end for
return ($\tau$)
```

The primary objective of our approach is to obtain the most optimal set of subpaths (τ_i) that covers all the data points and identifies the minimal set of subpaths that within themselves will cover all the cluster centroids while optimizing their lengths and curvature. The algorithm *Find_Trends* generates this minimal set τ using a maximization solution where the aim is to maximize the length and curvature in the selected subpaths using a weight parameter (w). Algorithm 3 gives the implementation of *Find_Trends* subroutine. All Subpaths (SP_I) are ranked in ascending order according to their Γ value. The function *Get_CC* obtains the corresponding set of Nodes (Cluster Centroids) for each SP_I. The paths are ranked in the descending order of Γ_I, are checked for the CC_is that they include. Those CC_is are eliminated from the *Ind* set which is a set of Indicator variables for CC_is. The paths are sequentially checked for any new CC_i that they may add and the process terminates once all CC_is are eliminated, that is, $(CC_i \in \tau) \cap \sigma = \phi$.

Each SP_I includes a sequence of CC_is that represent the clusters of the data. Hence each SP_I is associated with a set of data points included in those clusters. Thus, each SP_I is the representation of a *trend* in that data set in an order-preserving and continuous form. The 'subspace trends' have the latent patterns which can be seen when they are represented using direct plots or in the case of high-dimensional data, using suitable dimensionality reduction and visualization methods. Most importantly, these subspace trends have much stronger correlations which can be mathematically modeled.

5 Experimental Results

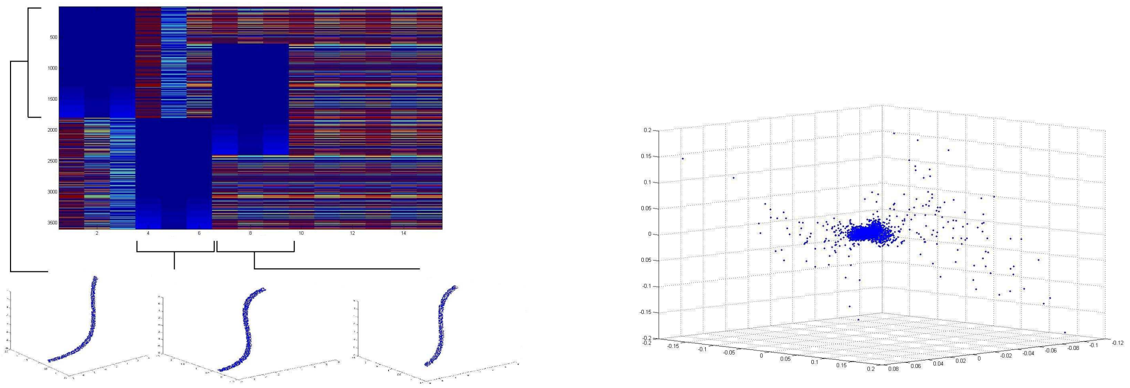
All programs were written in MATLAB Version 6.5 and run on pentium Dual Core 2.8 GHz machines. Experiments were performed using both synthetic and real-world datasets.

5.1 Synthetic Data sets Our algorithm was tested successfully on various synthetic data sets that inherently contain *subspace trends*. Several data sets were created with various embedded patterns hidden in the original global data space and the algorithm was successfully able to identify the IRS and the 'subspace trends' in them. We will demonstrate the results for two such datasets.

Synthetic dataset 1 : The dataset consists of 3602 data points $\{DP_1,...,DP_{3602}\}$ spread over 15 features $\{F_1,...,F_{15}\}$. There are three latent non-linear patterns embedded in 3 different IRS. The first IRS spans over data points $\{DP_1,...,DP_{1081}\}$ with features $\{F_1,F_2,F_3\}$, the second IRS spans over data points $\{DP_{1082},...,DP_{3602}\}$ with features $\{F_4,F_5,F_6\}$ and finally, the third IRS spans over data points $\{DP_{600},...,DP_{2401}\}$ with features $\{F_7,F_8,F_9\}$. All these IRS have in them non-linear correlation with sinewave form. The remaining features $\{F_{10},...,F_{15}\}$ and data are randomly generated noise points. The plot in Fig. 4(a) shows the strong correlation between the IRS feature sets $\{F_1,F_2,F_3\}$, $\{F_4,F_5,F_6\}$ and $\{F_7,F_8,F_9\}$ over their respective regions. There were 10 distinct subspace clusters ($SC_1,...,SC_{10}$) identified and were reduced to 3 IRS's $\{IRS_1,IRS_2,IRS_3\}$. The algorithm was able to identify all the IRS's suitably and represent their sine correlation in the final representation (see Fig 4(a)). As shown in Fig 4(b), dimensionality reduction methods like PCA cannot provide any information about the embedded subspace trends.

Synthetic dataset 2 : This dataset consists of 3602 data points $\{DP_1,...,DP_{3602}\}$ spanning 164 features $\{F_1,...,F_{164}\}$. There is one IRS present in the subset $\{DP_1,...,DP_{1081}\}$ across features $\{F_1,F_2,F_3\}$ as shown in Fig 5(a). The underlying structure of this IRS is that of two intersecting sine waves and we created this dataset to test the ability of the algorithm to separate the two underlying subspace trends as two distinct sine waves from the same IRS. The plot in Fig. 5(a) shows the strong correlation between $\{F_1,F_2,F_3\}$ over the IRS $\{DP_1,...,DP_{1081}\}$. Our algorithm was able to identify the IRS and separate the embedded subspace trends, as shown in Fig 5(b,c). Fig 5(d,e) shows the results of the traditional dimensionality reduction methods such as PCA and Laplacian Eigenmap which do not provide any hints about the subspace correlation.

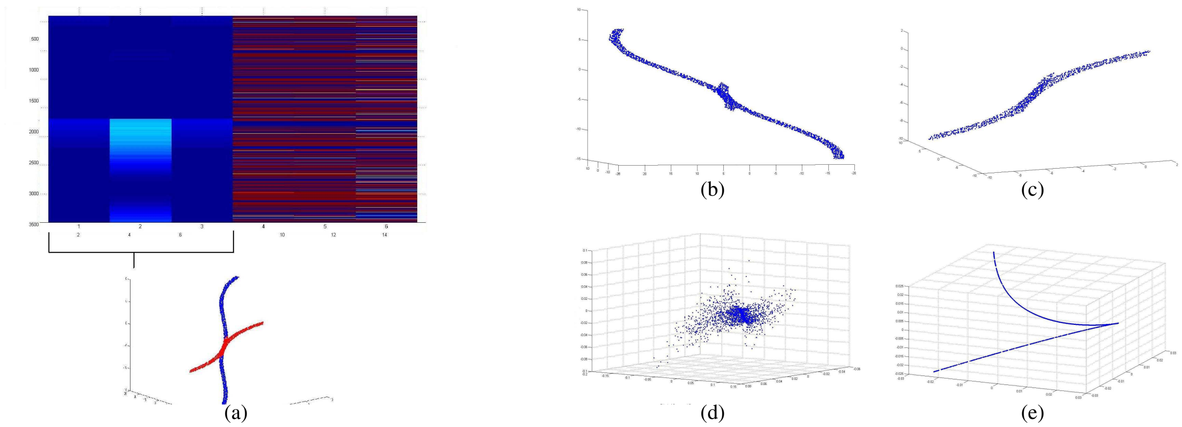
5.2 Real-world Data sets We also evaluated the performance of the proposed algorithm on three real-world datasets. We obtained promising results that demonstrate the



(a)

(b)

Figure 4: Results on Synthetic dataset 1. Results of (a) the proposed algorithm (b) PCA.



(a)

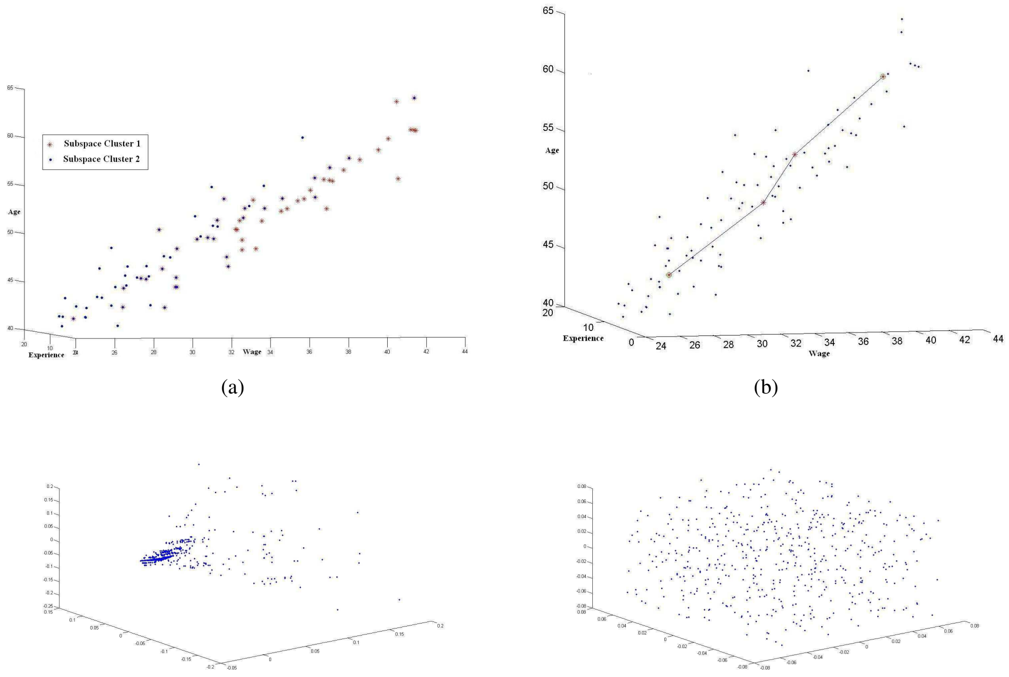
(b)

(c)

(d)

(e)

Figure 5: (a) Correlation matrix and subspace trends in synthetic dataset 2. Results on Synthetic dataset 2 - (b,c) Embedded subspace trends identified by the proposed algorithm. Results of (d) PCA and (e) Laplacian Eigenmap.



(a)

(b)

(c)

(d)

Figure 6: Results on the Wages dataset (a) IRS with merged subspace clusters (b) Subspace trends obtained using the proposed algorithm. Result of (c) PCA (d) Laplacian Eigenmap.

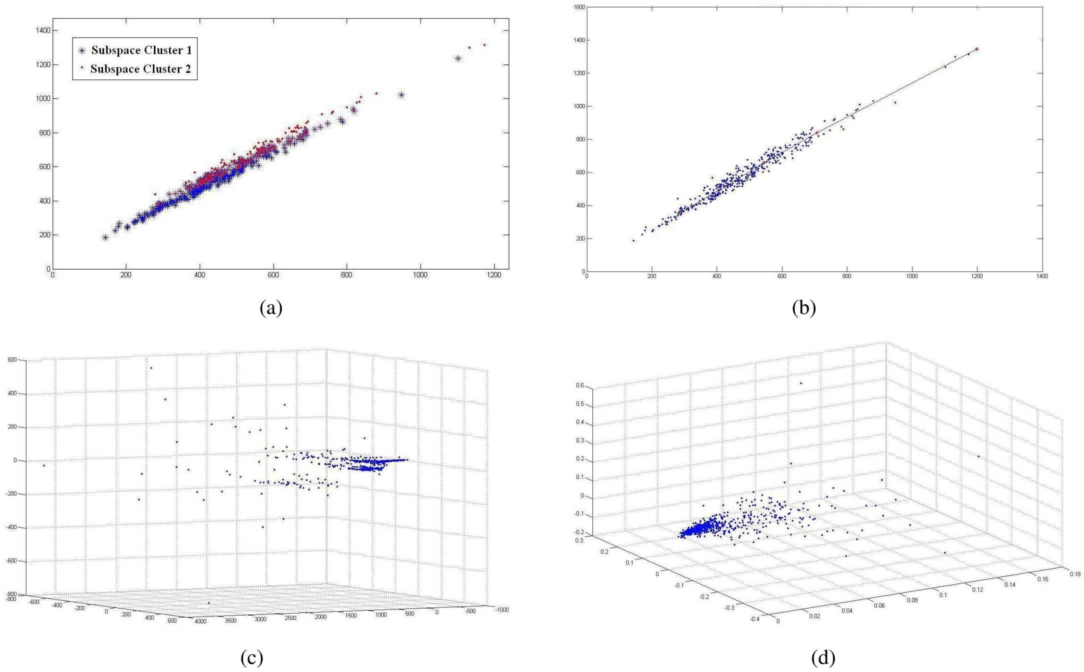
ability to offer some useful insights about information-rich subspace trends.

(1) *Wages dataset* :
The wages dataset contains the statistics of the determinants of Wages from the 1985 Current Population Survey. It contains 534 observations on 11 features sampled from the original Current Population Survey of 1985 and can be downloaded from StatLib Data archive[2]. Out of these 11 features, 4 are numerical [EDUCATION: Number of years of education, EXPERIENCE: Number of years of work experience, WAGE: Wage (dollars per hour) and AGE: Age (years)]. The other 7 are categorical which are converted to the corresponding numerical values. Our algorithm gave 14 subspace clusters $\{SC_1,...,SC_{14}\}$ spanning various feature subsets. We were able to obtain one prominent IRS spanning the features $\{F_4,F_6,F_7\}$ (Experience, Wage and Age) and 2 subspace clusters $|SC_1| = 54$ and $|SC_2| = 63$. The IRS and the 'subspace trend' identified in that is shown in Figs 6(a,b). One can see the inherent linear structure to the subset of the data. In Fig 6(c,d), we can see that the dimensionality reduction methods like PCA and Laplacian Eigenmap do not provide any information about these subspace trends.

(2) *Breast Cancer Dataset* :
We also tested our algorithm on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset[3] from the UCI Machine Learning Repository. In this dataset, there were 32 features computed from a digitized image describing the characteristics of the cell nuclei present in the image with 569 data points. Excluding patient ID and diagnosis (class label) columns, we used only 30 features for our analysis. Five distinct subspace clusters were identified out of which we were able to identify one IRS with strong correlation. This IRS is composed of 2 different subspace clusters $|SC_1|$ = 290 and $|SC_2|$ = 203 over features [F_4, F_{24}] This is shown in Fig. 7(a). As we can see from the figure, merging the two subspace clusters enhances the IRS and makes it more suitable for identifying informative trends. Individually, both these subspace clusters would provide very weak trends and hence it is important to identify a collective Information-Rich Subset of data when looking for subspace trends. We were able to identify a prominent 'subspace trend' over the IRS identified as shown in Fig. 7(b). The Isomap and PCA results shown in Figs. 7(c,d) clearly show that the traditional dimensionality reduction methods cannot identify these inherent subspace trends.

[2] http://lib.stat.cmu.edu/datasets/CPS_85_Wages

[3] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)



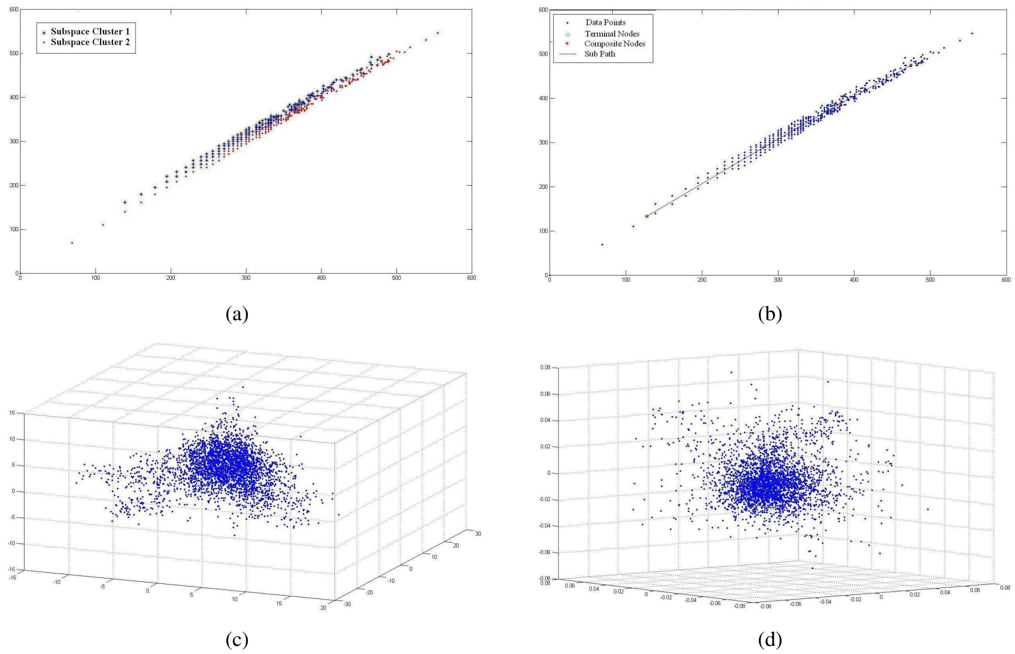
(a)

(b)

(c)

(d)

Figure 7: Results on UCI Breast Cancer dataset. (a) IRS with merged subspace clusters (b) Subspace trends obtained by the proposed algorithm. Results of (c) Isomap (d) PCA.



(a)

(b)

(c)

(d)

Figure 8: Results on Yeast Cell Gene expression data. (a) IRS with merged subspace clusters (b) Subspace trends obtained by the proposed algorithm. Result of (c) Isomap (d) PCA.

(3) *Yeast Expression Dataset* :
The Yeast Expression dataset[4] describes the systematic determination of genetic network architecture's dataset of yeast cells and contains 8224 samples (genes) and 17 features (expression levels). Out of all the subspace clusters found, we are able to identify a distinct IRS with strong correlation. This IRS is composed of 2 different subspace clusters $|SC_1| = 642$ and $|SC_2| = 1164$ over features $\{F_{11},F_{12}\}$ which are experimental conditions, as shown in Fig. 8(a). The IRS spans over 1675 data points and features $\{F_{11},F_{12}\}$. We were able to identify a 'subspace trend' in the IRS and as in Fig. 8(b). Figs. 8(c,d) shows the results of Isomap and PCA on this dataset and one can see that these embedded trends are not identified using the traditional methods.

5.3 Discussion In the previous section, we have shown the results in terms of visual plots that clearly highlight the subspace trends identified by the proposed algorithm. For some applications, it is important to go beyond these measures and quantify the results by modeling them using continuous functions such as principal curves [13]. Since the concept of subspace trends is novel and is not previously available in the literature, we compared our results quantitatively with that of the subspace clustering results. We know that clusters are groupings of data points based on a similarity criteria and can be represented by their centroids. The centroids are the single point summarizations of these subspace clusters. In order to compare the measure of summarization, we calculated the error measures for the clusters as Sum of Errors (SE) and Sum Square Error (SSE) [19]. For a data point x_i in a subspace cluster *SC* containing *nc* data points and with centroid *m*, SE and SSE are calculated as follows:

$$SE = \sum_{i=1}^{nc} |m - x_i| \tag{5.3}$$

$$SSE = \sqrt{\sum_{i=1}^{nc} (m - x_i)^2} \tag{5.4}$$

Though subspace cluster centroids are considered to be representative points, they provide very limited information in terms of the correlations. Since, we are able to extract these data points that follow a trend, we can take advantage of continuous modeling using principal curves to model such data. Hence, we will compare the effectiveness of clusterings with that of using principal curves [20]. Principal curves are the one-dimensional representation of the data that defines a line which passes through the most dense regions of the dataset. Each data point has a corresponding projection point on the principal curve and this projection distance can be used as the error measure for summarization [11]. The subspace trends that we identify in a reduced feature and data space can use principal curves to give a one dimensional representation and thus the principal curves will provide a mathematical formalization of trends which can model the data in terms of the subspace (linear or non-linear) correlations of the attributes. This gives rise to an effective measure to evaluate the best fit curve using the distance between the data points and the principal curve. For each data point $x_{i,j}$, let p($x_{i,j}$) be the projection onto the principal curve. The L_1 and L_2 error measure is defined as follows:

$$L_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} |x_{i,j} - p(x_{i,j})| \tag{5.5}$$

$$L_2 = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{i,j} - p(x_{i,j}))^2} \tag{5.6}$$

All these error measures for the three real-world datasets are calculated and shown in Table 2. One can see the effectiveness of using these principal curves for the continuous representation of the subspaces rather than using the clustering itself. For examples, the correlation between the attributes in the wages dataset is non-linear and principal curve will provide a mathematical framework for modeling these correlations.

Table 2: Comparison results for trends and clusters on various real-world datasets in terms of L_1 and L_2 evaluation measures.

	L1 Measure		L2 Measure	
Dataset	SE	Trend	SSE	Trend
Wages	657	4.4	403.7	11.7
Breast Cancer	55000	43.1	39300	7210.4
Yeast Gene	167000	19.3	119000	787.1

6 Conclusion and Future Research

In spite of the vast literature in high-dimensional data analysis, not many efforts were made in identifying information-rich subsets and creating an order-preserving representation of such subsets. In this paper, we developed a new algorithm that takes advantage of different subspace clusters and identifies an information-revealing representation for subsets of data and features that may contain local patterns. Our approach is transparent to the underlying subspace clustering algorithms used and will enhance the understanding of these subspaces by providing an integrated framework that defines the notion of trends in these subspaces. Analyzing such 'subspace trends' can provide very useful insights for domain

[4] http://cheng.ececs.uc.edu/biclustering/yeast.matrix

experts to further analyze such high-dimensional datasets for understanding the local linear and non-linear correlations occurring in these subspaces. One of the main future research directions of this work is to extend the algorithm to optimize for the feature subsets and the trends simultaneously. One can also evaluate the effectiveness of our algorithm using different subspace clustering algorithms.

Acknowledgments

We would like to thank Tin Kam Ho, Bell Laboratories for some useful discussions. We would also like to thank the anonymous reviewers for providing some valuable comments for our paper.

References

[1] C. Agarwal and P. Yu. Finding generalized projected clusters in high dimensional spaces. In *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 70–81, 2000.

[2] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park. Fast algorithms for projected clustering. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 61–72, 1999.

[3] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 94–105. ACM Press, 1998.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[5] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. In *Artificial Intelligence*, volume 97, pages 245–271, 1997.

[6] C. Cheng, A. W. Fu, and Y. Zhang. ENCLUS: Entropy-based subspace clustering for mining numerical data. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , pages 84–93, 1999.

[7] K. O. Cheng, N. F. Law, W. C. Siu, and A. W. Liew. Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. In *BMC Bioinformatics*, 2008.

[8] Trevor F. Cox and Michael A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, 1994.

[9] S. Goil, H. Nagesh, and A. Choudhary. MAFIA: efficient and scalable subspace clustering for very large data sets. Technical report, 2001.

[10] Kyoung gu Woo, Jeong hoon Lee, Myoung ho Kim, and Yoon joon Lee. FINDIT: a fast and intelligent subspace clustering algorithm using dimension voting. *Information & Software Technology*, 46(4):255–271, 2004.

[11] T. Hastie and W. Stuetzle. Principal curves. *Journal of American Statistical Association*, 84:502–516, 1989.

[12] J.B.Tenenbaum, V. de Silva, and J.C.Langford. A global geometric framework for nonlinear dimensionality reduction. In *Science magazine*, volume 290, pages 2319–2323, 2005.

[13] B. Kegl, A. Krzyzak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(3):281–297, 2000.

[14] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[15] L.Yu and H.Liu. Feature selection for high dimensional data: a fast correlation-based filter selection. In *International Conference on Machine Learning*, pages 856–863, 2003.

[16] L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 6(1):90–105, 2004.

[17] C. K. Reddy, S. Pokharkar, and T. K. Ho. Generating hypotheses of trends in high-dimensional data skeletons. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology* , pages 139–146, 2008.

[18] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. In *Science magazine*, volume 290, pages 2323–2326, 2005.

[19] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Addison Wesley, 2006.

[20] J. J. Verbeek, N. Vlassis, and B. Krse. A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, 23(8):1009–1017, 2002.

[21] X. Zhang, F. Pan, and W. Wang. CARE: Finding local linear correlations in high dimensional data. In *Proceedings of the 24th International Conference on Data Engineering*, pages 130–139, 2008.

[22] X. Zhang, F. Pan, and W. Wang. REDUS: Finding reducible subspaces in high dimensional data. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 961–970, 2008.