REDUS: Finding Reducible Subspaces in High Dimensional Data

Xiang Zhang, Feng Pan, and Wei Wang
Department of Computer Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
{xiang, panfeng, weiwang}@cs.unc.edu

ABSTRACT

Finding latent patterns in high dimensional data is an important research problem with numerous applications. The most well known approaches for high dimensional data analysis are feature selection and dimensionality reduction. Being widely used in many applications, these methods aim to capture global patterns and are typically performed in the full feature space. In many emerging applications, however, scientists are interested in the local latent patterns held by feature subspaces, which may be invisible via any global transformation.

In this paper, we investigate the problem of finding strong linear and nonlinear correlations hidden in feature subspaces of high dimensional data. We formalize this problem as identifying reducible subspaces in the full dimensional space. Intuitively, a reducible subspace is a feature subspace whose intrinsic dimensionality is smaller than the number of features. We present an effective algorithm, REDUS, for finding the reducible subspaces. Two key components of our algorithm are finding the overall reducible subspace, and uncovering the individual reducible subspaces from the overall reducible subspace. A broad experimental evaluation demonstrates the effectiveness of our algorithm.

Categories and Subject Descriptors: H.2.8 [Database Applications]: Data Mining

General Terms: Algorithm, Performance

Keywords: Reducible Subspace, High Dimensional Data

1. INTRODUCTION

Many real life applications deal with high dimensional data. In bio-medical domain, for example, advanced microarray techniques [2, 11, 17] allow to monitor the expression levels of hundreds to thousands of genes simultaneously. By mapping each gene to a feature, gene expression data can be treated as data points distributed in a very high dimensional feature space. To make sense of such high dimensional data, extensive research has been done in finding the latent structures hidden in the large number of features. Two well known approaches in analyzing high dimensional data are feature selection and dimensionality reduction.

The goal of feature selection methods [6, 22, 31, 33] is to find a single representative subset of features that are most relevant for the data mining task at hand, such as classification. The selected features generally have low correlation with each other but have strong correlation with the target feature.

Dimensionality reduction [4, 8, 18, 27, 29] is widely used as a key component of many approaches in analyzing high dimensional data. The insight behind dimensionality reduction methods is that a high dimensional dataset may exhibit interesting patterns on a lower dimensional subspace due to correlations among the features. Though very successful in finding the low dimensional structures embedded in a high dimensional space, these methods are usually performed in the full feature space. They aim to model the global latent structure of the data and do not separate the impact of any original features nor identify latent patterns hidden in some feature subspaces. Please refer to Section 2 for a more detailed discussion of the related work.

1.1 Motivating Example

In many emerging applications, the datasets usually consist of thousands to hundreds of thousands of features. In such high dimensional dataset, some feature subsets may be strongly correlated, while others may not have any correlation at all. In these applications, it is more desirable to find the correlations that are hidden in feature subspaces. For example, in gene expression data analysis, a group of genes having strong correlation is of high interests to biologists since it helps to infer unknown functions of genes [11] and gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network [17]. We refer to such correlation among a subset of features as a *local correlation* in comparison with the global correlation found by the full dimensional feature reduction methods. Since such local correlations only exist in some subspaces of the full dimensional space, they are invisible to the full dimensional reduction methods. In [32], an algorithm is proposed to find local linear correlations in high dimensional data. However, in real applications, the feature subspace can be either linearly or nonlinearly correlated. The problem of finding linear and nonlinear correlations in feature subspaces remains open.

For example, Figure 1 shows a data sets consisting of 12 features, $\{f_1, f_2, \cdots, f_{12}\}$, and 1000 data points. Embedded

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-59593-991-3/08/10 ...$5.00.

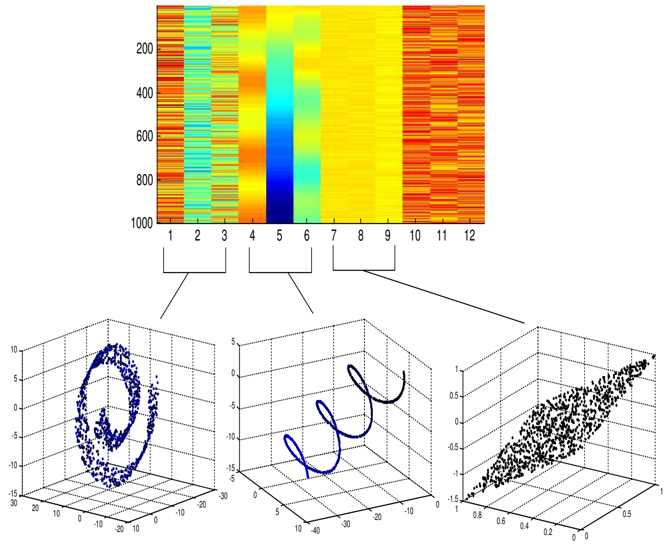
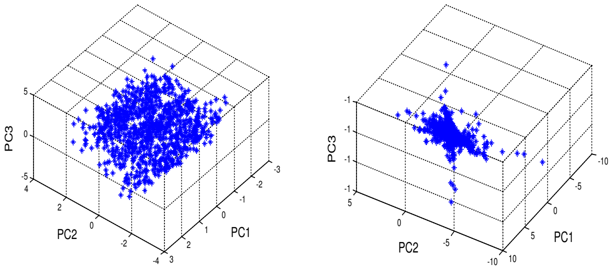


Figure 1: An example dataset



(a) Result of PCA

(b) Result of ISOMAP

Figure 2: Applying dimensionality reduction methods to the full dimensional space of the example dataset

in the full dimensional space, features subspaces $\{f_1, f_2, f_3\}$ and $\{f_4, f_5, f_6\}$ are nonlinearly correlated, $\{f_7, f_8, f_9\}$ are linearly correlated. Features $\{f_{10}, f_{11}, f_{12}\}$ contain random noises.

Performing dimensionality reduction methods to the full dimensional space cannot uncover these local correlations hidden in the full feature spaces. For example, Figure 2(a) shows the result of applying Principal Component Analysis (PCA)[18] to the full dimensional space of the example dataset shown in Figure 1. In this figure, we plot the point distribution on the first 3 principal components found by PCA. Clearly, we cannot find any pattern that is similar to the patterns embedded in the dataset. Similarly, Figure 2(b) shows the results of applying ISOMAP [29] to reduce the dimensionality of the dataset down to 3. There is also no desired pattern found in this low dimensional structure.

How can we identify these local correlations hidden in the full dimensional space?

This question is two-fold. First, we need to identify the strongly correlated feature subspaces, i.e., a subset of features that are strongly correlated and actually have low dimensional structures. Then, after these locally correlated feature subsets are found, we can apply the existing dimensionality reduction methods to identify the low dimensional structures embedded in them.

Many methods have been proposed to address the second aspect of the question, i.e., given a correlated feature space, finding the low dimensional embedding in it. The first aspect of the question, however, is largely untouched. In this paper, we investigate the first aspect of the question, i.e., identifying the strongly correlated feature subspaces.

1.2 Challenges and Contributions

(1) In this paper, we investigate the problem of finding correlations hidden in the feature subspaces of high dimensional data. The correlations can be either linear or nonlinear. To our best knowledge, our work is the first attempt to find local linear and nonlinear correlations hidden in feature subspaces.

Many methods for modelling correlations can be found in the literature, such as mutual information [10], Pearson correlation [26], and rank correlation [19]. However, the commonly used measurements are for correlations between two features. In high dimensional data, the correlation may involve a large number of features, i.e., a high order correlation. Note that a strong high order correlation does not necessarily imply that there are strong correlations between feature pairs. For example, the 3 features, $\{f_1, f_2, f_3\}$ in Figure 1, are strongly correlated, since the 3-dimensional Swiss roll structure is actually on a 2-dimensional manifold. Figures 3(a) to 3(c) show the projections of the Swiss roll onto the spaces of two features. As we can see from the figures, there are no clear pairwise correlations between any two features.

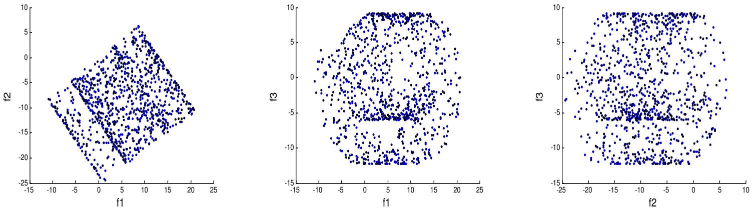
We adopt the concept of intrinsic dimensionality [13] to model the high dimensional correlation. We formalize this problem as finding *reducible subspaces* in the full dimensional space. Informally, a feature subspace is reducible if its intrinsic dimensionality is smaller than the number of features. Various intrinsic dimensionality estimators have been developed [9, 14, 21]. Our problem formalization does not depend on any particular method for estimating the intrinsic dimensionality. We show two necessary properties that any estimator should satisfy in order to be a generalization of the well-defined concepts in linear case.

(2) We develop an effective algorithm REDUS[1] to find the reducible subspaces in the dataset. REDUS consists of the following two steps.

It first finds the union of all reducible subspaces, i.e., the *overall reducible subspace*. The second step is to uncover the individual reducible subspaces in the overall reducible subspace. The key component of this step is to examine if a feature is strongly correlated with a feature subspace. We develop a method utilizing point distributions to distinguish the features that are strongly correlated with a feature subspace and those that are not. Our method achieves similar accuracy to that of directly using intrinsic dimensionality estimators, but with much less computational cost.

Extensive experiments on synthetic and real life datasets demonstrate the effectiveness of REDUS.

[1] REDUS stands for REDUcible Subspaces.



(a) f_1 and f_2 (b) f_1 and f_3 (c) f_2 and f_2

Figure 3: Pairwise correlations of the Swiss roll in the example dataset

2. RELATED WORK

Feature Selection Feature selection methods [6, 22, 31, 33] try to find a subset of features that are most relevant for certain data mining task, such as classification. In order to find the relevant feature subset, these methods search through various subsets of features and evaluate these subsets according to some criteria. Feature selection methods can be further divided into two groups according to their evaluation criteria: wrapper and filter. Wrapper models evaluate feature subsets by their predictive accuracy using statistical re-sampling or cross-validation. In filter techniques, the feature subsets are evaluated by their information content, using statistical dependence or information-theoretic measures.

Feature selection finds one feature subset for the entire dataset. The selected feature subset usually contains the features that have low correlation with each other but have strong correlation with the target feature. The correlation measurements are usually defined for feature pairs, such as mutual information and Pearson correlation. Our work, on the other hand, is to find the subsets of features, in which the features are strongly correlated. Moreover, the correlations are not limited to feature pairs.

Dimensionality reduction Dimensionality reduction methods can be categorized into linear methods, such as Multi-Dimensional Scaling (MDS) [8] and Principal Component Analysis (PCA)[18], and non-linear methods, such as Local Linear Embedding (LLE) [27], ISOMAP [29], and Laplacian eigenmaps [4]. For high dimensional datasets, if there exist low dimensional subspaces or manifolds embedded in the full dimensional spaces, these methods are successful in identifying these low dimensional embeddings.

Dimensionality reduction methods are usually applied on the full dimensional space to capture the independent components among all the features. They are not designed to address the problem of identifying correlation in feature subspaces. It is reasonable to apply them to the feature spaces that are indeed correlated. However, in very high dimensional datasets, different feature subspaces may have different correlations, and some feature subspace may not have any correlation at all. In this case, dimensionality reduction methods should be applied after such strongly correlated feature subspaces have been identified.

Correlation Clustering The goal of correlation clustering methods is to find clusters hidden in projected feature spaces [1, 7, 30]. They can be viewed as combinations of clustering methods and dimensionality reduction methods. Both ORCLUS [1] and 4C [7] can be treated as PCA-lized clustering methods. To find the low dimensional clusters, ORCLUS and 4C apply PCA on subsets of data points in the full dimensional space and merge the point subsets having similar orientations. Therefore, they do not touch the problem of finding reducible subspaces. Instead, they implicitly assume the full dimensional space is reducible for certain subsets of data points. CURLER [30] finds clusters having non-linear correlations in subspaces. It first applies EM clustering algorithm to find a large number of micro-clusters, and then merges clusters with large overlaps. Since in its first step, micro-clusters are formed by applying EM to the full dimensional space, CURLER also does not address the problem of finding the reducible subspaces.

Intrinsic Dimensionality Due to correlations among features, a high dimensional dataset may lie in a subspace with dimensionality smaller than the number of features [9, 14, 21]. The intrinsic dimensionality can be treated as the minimum number of free variables required to define the data without any significant information loss [13]. For example, as shown in Figure 1, in the 3-dimensional space of $\{f_1, f_2, f_3\}$, the data points lie on a Swiss roll, which is actually a 2-dimensional manifold. Therefore, its intrinsic dimensionality is 2.

The concept of intrinsic dimensionality has many applications in the database and data mining communities, such as clustering [3, 15], outlier detection [24], nearest neighbor queries [23], and spatial query selectivity estimation [5, 12]. Different definitions of intrinsic dimensionality can be found in the literature. For example, in linear cases, matrix rank [16] and PCA [18] can be used to estimate intrinsic dimensionality. For nonlinear cases, estimators such as *box counting dimension*, *information dimension*, and *correlation dimension* have been developed. These intrinsic dimensionality estimators are sometimes collectively referred to as *fractal dimension*. Please see [25, 28] for good coverage of the topics of intrinsic dimensionality estimation and its applications.

Local Linear Correlation In [32], the CARE algorithm has been proposed for finding local linear correlations. Adopting a similar criterion used in PCA, the strongly correlated feature subsets are formalized as feature subspaces having small residual variances. However, this work only focuses on linear correlations. The problem of finding non-linear local correlations remains uninvestigated.

3. PROBLEM FORMALIZATION

In this section, we utilize intrinsic dimensionality to formalize the problem of finding strongly correlated feature subspaces.

Suppose that the dataset Ω consists of N data points and M features. Let $\Omega_P = \{p_1, p_2, \cdots, p_N\}$ denote the point set, and $\Omega_F = \{f_1, f_2, \cdots, f_M\}$ denote the feature set in Ω respectively. We use $ID(V)$ to represent the intrinsic dimensionality of the feature subspace $V \in \Omega_F$.

Intrinsic dimensionality provides a natural way to examine whether a feature is correlated with some feature subspace: if a feature $f_a \in \Omega_F$ is strongly correlated with a feature subspace $V \subseteq \Omega_F$, then adding f_a to V should not cause much change of the intrinsic dimensionality of V. The following definition formalizes this intuition.

Definition 3.1. (Strong Correlation)
A feature subspace $V \subseteq \Omega_F$ and a feature $f_a \in \Omega_F$ have strong correlation, if

$$\Delta ID(V, f_a) = ID(V \cup \{f_a\}) - ID(V) \leq \epsilon.$$

In this definition, ϵ is a user specified threshold. Smaller ϵ value implies stronger correlation, and larger ϵ value implies weaker correlation. If V and f_a have strong correlation, we also say that they are strongly correlated.

Definition 3.2. (Redundancy)
Let $V = \{f_{v_1}, f_{v_2}, \cdots, f_{v_m}\} \subseteq \Omega_F$. $f_{v_i} \in V$ is a redundant feature of V, if f_{v_i} has strong correlation with the feature subspace consisting of the remaining features of V, i.e.,

$$\Delta ID(\{f_{v_1}, \cdots, f_{v_{i-1}}, f_{v_{i+1}}, \cdots, f_{v_m}\}, f_{v_i}) \leq \epsilon.$$

We say V is a *redundant* feature subspace if it has at least one redundant feature. Otherwise, V is a *non-redundant* feature subspace.

Note that in Definitions 3.1 and 3.2, $ID(V)$ does not depend on a particular intrinsic dimensionality estimator. Any existing estimator can be applied when calculating $ID(V)$. Moreover, we do not require that the intrinsic dimensionality estimator reflects the exact dimensionality of the dataset. However, in general, a good intrinsic dimensionality estimator should satisfy two basic properties.

First, if a feature is redundant in some feature subspace, then it is also redundant in the supersets of the feature subspace. We formalize this intuition as the following property.

Property 3.3. *For $V \in \Omega_F$, if $\Delta ID(V, f_a) \leq \epsilon$, then $\forall U$ $(V \subseteq U \subseteq \Omega_F)$, $\Delta ID(U, f_a) \leq \epsilon$.*

This is a reasonable requirement, since if f_a is strongly correlated with $V \subseteq U$, then adding f_a to U will not greatly alter its intrinsic dimensionality.

From this property, it is easy to see that, if feature subspace U is non-redundant, then all of its subsets are non-redundant, which is clearly a desirable property for the feature subspaces.

Corollary 3.4. *If $U \subseteq \Omega_F$ is non-redundant, then for $\forall V \subseteq U$, V is also non-redundant.*

The following property extend the concept of basis [20] in a linear space to nonlinear space using intrinsic dimensionality. In linear space, suppose that V and U contain the same number of vectors, and the vectors in V and U are all linearly independent. If the vectors of U are in the subspace spanned by the vectors of V, then the vectors in V and the vectors in U span the same subspace. (A span of a set of vectors consists of all linear combinations of the vectors.) Similarly, in Property 3.5, for two non-redundant feature subspaces, V and U, we require that if the features in U are strongly correlated with V, then U and V are strongly correlated with the same subset of features.

Property 3.5. *Let $V = \{f_{v_1}, f_{v_2}, \cdots, f_{v_m}\} \subseteq \Omega_F$ and $U = \{f_{u_1}, f_{u_2}, \cdots, f_{u_m}\} \subseteq \Omega_F$ be two non-redundant feature subspaces. If $\forall f_{u_i} \in U$, $\Delta ID(V, f_{u_i}) \leq \epsilon$, then for $\forall f_a \in \Omega_F$, $\Delta ID(U, f_a) \leq \epsilon$ iff $\Delta ID(V, f_a) \leq \epsilon$.*

Intuitively, if a feature subspace Y $(Y \subseteq \Omega_F)$ is redundant, then Y should be reducible to some subspace, say V $(V \subset Y)$. Concerning the possible choices of V, we are most interested in the smallest one that Y can be reduced to, since it represents the intrinsic dimensionality of Y. We now give the formal definitions of reducible subspace and its core space.

Definition 3.6. (Reducible Subspace and Core Space)
$Y \subseteq \Omega_F$ is a reducible subspace if there exists a non-redundant subspace V $(V \subset Y)$, such that
(1) $\forall f_a \in Y$, $\Delta ID(V, f_a) \leq \epsilon$, and
(2) $\forall U \subset Y$ $(|U| \leq |V|)$, U is non-redundant.
We say V is the core space of Y, and Y is reducible to V.

Criterion (1) in Definition 3.6 says that all features in Y are strongly correlated with the core space V. The meaning of criterion (2) is that the core space is the smallest non-redundant subspace of Y with which all other features of Y are strongly correlated.

Among all reducible subspaces, we are most interested in the maximum ones. A maximum reducible subspace is a reducible subspace that includes all features that are strongly correlated with its core space.

Definition 3.7. (Maximum Reducible Subspace)
$Y \subseteq \Omega_F$ is a maximum reducible subspace if
(1) Y is a reducible subspace, and
(2)$\forall f_b \in \Omega_F$, if $f_b \notin Y$, then $\Delta ID(V, f_b) > \epsilon$, where V is the core space of Y.

Let $\{Y_1, Y_2, \cdots, Y_S\}$ be the set of all maximum reducible subspaces in the dataset. The union of the maximum reducible subspaces $OR = \bigcup_{i=1}^{S} Y_i$ is referred to as the *overall reducible subspace.*

Note that Definition 3.7 works for the general case where a feature can be in different maximum reducible subspaces. In this paper, we focus on the special case where maximum reducible subspaces are non-overlapping, i.e., each feature can be in *at most one* maximum reducible feature subspace.

To find the maximum reducible subspaces in the dataset, REDUS adopts a two-step approach. The first step is to find the overall reducible subspace. Then, from the overall reducible subspace, it identifies the individual maximum reducible subspaces. In the next section, we present the algorithm for finding the overall reducible subspace. In Section 5, we discuss the method for finding individual maximum reducible subspaces.

4. OVERALL REDUCIBLE SUBSPACE

In this section, we present the algorithm for finding the overall reducible subspace. We first give a short introduction to the intrinsic dimensionality estimator. Then we present the algorithm for finding the overall reducible subspace and the proof of its correctness.

4.1 Intrinsic Dimensionality Estimator

To find the overall reducible subspace in the dataset, we adopt correlation dimension [25, 28], which can measure both linear and nonlinear intrinsic dimensionality, as our intrinsic dimensionality estimator since it is computationally more efficient than other estimators while its quality of estimation is similar to others. In practice, we observe that correlation dimension satisfies Properties 3.3 and 3.5, although we do not provide the proof here. In what follows, we give a brief introduction of correlation dimension.

Let Y be a feature subspace of the dataset, i.e., $Y \subseteq \Omega_F$. Suppose that the number of points N in the dataset approaches infinity. Let $dis(p_i, p_j, Y)$ represent the distance between two data points p_i and p_j in feature subspace Y. Let $B_Y(p_i, r)$ be the subset of points contained in a ball of radius r centered at point p_i in subspace Y, i.e.,

$$B_Y(p_i, r) = \{p_j | p_j \in \Omega_P, dis(p_i, p_j, Y) \le r\}.$$

The average fraction of pairs of data points within distance r is

$$C_Y(r) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{p_i \in \Omega_P} |B_Y(p_i, r)|.$$

The *correlation dimension* of Y is then defined as

$$ID(Y) = \lim_{r,r'\to 0} \frac{\log[C_Y(r)/C_Y(r')]}{\log[r/r']}.$$

In practice, N is a finite number. C_Y is estimated using $\frac{1}{N^2} \sum_{p_i \in Y_P} |B(p_i, r)|$. The correlation dimension is the growth rate of the function $C_Y(r)$ in log-log scale, since $\frac{\log[C_Y(r)/C_Y(r')]}{\log[r/r']} = \frac{\log[C_Y(r)] - \log[C_Y(r')]}{\log r - \log r'}$. The correlation dimension is estimated using the slope of the line that best fits the function in least squares sense.

The intuition behind the correlation dimension is following. For points that are arranged on a line, one expects to find twice as many points when doubling the radius. For the points scattered on 2-dimensional plane, when doubling the radius, we expect the number of points to increase quadratically. Generalizing this idea to m-dimensional space, we have $C_Y(r)/C_Y(r') = (r/r')^m$. Therefore, the intrinsic dimensionality of feature subspace Y can be simply treated as the growth rate of the function $C_Y(r)$ in log-log scale.

4.2 Finding Overall Reducible Subspace

The following theorem sets the foundation for the efficient algorithm to find the overall reducible subspace.

THEOREM 4.1. *Suppose that $Y \subseteq \Omega_F$ is a maximum reducible subspace and $V \subset Y$ is its core space. We have $\forall U \subset Y$ $(|U| = |V|)$, U is also a core space of Y.*

PROOF. We need to show that U satisfies the criteria in Definition 3.7. Let $V = \{f_{v_1}, f_{v_2}, \cdots, f_{v_m}\}$ and $U = \{f_{u_1}, f_{u_2}, \cdots, f_{u_m}\}$.

Since $U \subset Y$, from the definition of reducible subspace, U is non-redundant, and for every $f_{u_i} \in U$, $\Delta ID(V, f_{u_i}) \le \epsilon$. For every $f_a \in Y$, we have $\Delta ID(V, f_a) \le \epsilon$. Thus from Property 3.5, we have $\Delta ID(U, f_a) \le \epsilon$. Similarly, for every $f_b \notin Y$, $\Delta ID(V, f_b) > \epsilon$. Thus $\Delta ID(U, f_a) > \epsilon$.

Therefore, U is also a core space of Y. □

Theorem 4.1 tells us that any subset $U \subset Y$ of size $|V|$ is also a core space of Y.

Suppose that $\{Y_1, Y_2, \cdots, Y_S\}$ is the set of all maximum reducible subspaces in the dataset and the overall reducible subspace is $OR = \bigcup_{i=1}^{S} Y_i$. To find OR, we can apply the following method. For every $f_a \in \Omega_F$, let $RF_{f_a} = \{f_b | f_b \in \Omega_F, b \ne a\}$ be the remaining features in the dataset. We calculate $\Delta ID(RF_{f_a}, f_a)$. The overall reducible subspace $OR = \{f_a | \Delta ID(RF_{f_a}, f_a) \le \epsilon\}$.We now prove the correctness of this method.

Algorithm 1: REDUS

```
   Input: Dataset $\Omega$, input parameters $\epsilon$, $n$, and $\tau$,
   Output: $Y$: the set of all maximum reducible
           subspaces
1  $OR = \emptyset$;
2  for each $f_a \in \Omega_F$ do
3      $RF_{f_a} = \{f_b | f_b \in \Omega_F, b \ne a\}$;
4      if $\Delta ID(RF_{f_a}, f_a) \le \epsilon$ then
5          $OR = OR \cup \{f_a\}$;
6      end
7  end
8  sample $n$ points $P = \{p_{s_1}, p_{s_2}, \cdots, p_{s_n}\}$ from $\Omega$.
9  for $d = 1$ to $|OR|$ do
10     for each candidate core space $C \subset OR$ $(|C| = d)$ do
11         $T = \{f_a | f_a$ is strongly correlated with $C$,
           $f_a \in OR, f_a \notin C\}$;
12         if $T \ne \emptyset$ then
13             $Y \leftarrow T$;
14             update $OR$ by removing from $OR$ the
               features in $T$;
15         end
16     end
17 end
18 return Y.
```

COROLLARY 4.2. $OR = \{f_a | \Delta ID(RF_{f_a}, f_a) \le \epsilon\}$.

PROOF. Let f_y be an arbitrary feature in the overall reducible subspace. From Theorem 4.1, we have $\forall f_y \in Y_i \subseteq OR$, $\exists V_i \subset Y_i$ $(f_y \notin V_i)$, such that V_i is the core space of Y_i. Thus $\Delta ID(V_i, f_y) \le \epsilon$. Since $f_y \notin V_i$, we have $V_i \subseteq RF_{f_y}$. From Property 3.3, we have $\Delta ID(RF_{f_y}, f_y) \le \epsilon$.

Similarly, if $f_y \notin OR$, then $\Delta ID(RF_{f_y}, f_y) > \epsilon$.

Therefore, we have $OR = \{f_y | \Delta ID(RF_{f_y}, f_y) \le \epsilon\}$. □

The algorithm for finding the overall reducible subspace is shown in Algorithm 1 from Line 1 to Line 7. Note that the procedure of finding overall reducible subspace is linear to the number of features in the dataset.

5. MAXIMUM REDUCIBLE SUBSPACE

In this section, we present the second component of REDUS, i.e., identifying the maximum reducible subspaces from the overall reducible subspace found in the previous section.

5.1 Intrinsic Dimensionality Based Method

From Definition 3.7 and Theorem 4.1, we have the following property concerning the reducible subspaces.

COROLLARY 5.1. *Let $Y_i \subseteq OR$ be a maximum reducible subspace, and $V_i \subset Y_i$ be any core space of Y_i. We have*

$$Y_i = \{f_a | \Delta ID(V_i, f_a) \le \epsilon, f_a \in OR\}.$$

Therefore, to find the individual maximum reducible subspaces $Y_i \subseteq OR$ $(1 \le i \le S)$, we can use any core space $V_i \subset Y_i$ to find the other features in Y_i. More specifically, a *candidate* core space of size d is a feature subset $C \subset OR$ $(|C| = d)$. From size $d = 1$ to $|OR|$, for each candidate core space, let $T = \{f_a | \Delta ID(C, f_a) \le \epsilon, f_a \in OR, f_a \notin C\}$. If $T \ne \emptyset$, then T is a maximum reducible subspace with core space of size d. The overall reducible subspace OR is then updated by removing the features in T. Note that the size of $|OR|$ decreases whenever some maximum reducible subspace is identified. We now prove the correctness of this method.

COROLLARY 5.2. *Any candidate core space is non-redundant.*

PROOF. It is easy to see any candidate core space of size 1 is non-redundant. Now, assume that all candidate core spaces of size $d-1$ are non-redundant, we show all candidate core spaces of size d are non-redundant. We prove this by contradiction.

Let $V = \{f_{v_1}, f_{v_2}, \cdots, f_{v_d}\}$ be an arbitrary candidate core space of size d. Without loss of generality, assume that f_d is the redundant feature in V. Let $V' = \{f_1, f_2 \cdots, f_{v_{d-1}}\}$. We have $\Delta ID(V', f_{v_d}) \leq \epsilon$. Since $|V'| = d-1$, V' is non-redundant according to the assumption. Moreover, we have $T = \{f_a | \Delta ID(V', f_a) \leq \epsilon, f_a \in OR, f_a \notin V'\} \neq \emptyset$, since $f_{v_d} \in T$. Therefore, $f_{v_d} \in T$ would have been removed from OR before the size of the candidate core spaces reaches d. This contradicts the assumption of f_{v_d} being in the candidate core space V. Therefore, we have that any candidate core space is non-redundant. □

COROLLARY 5.3. *Let C be a candidate core space. If $\exists f_a \in OR$ such that $\Delta ID(C, f_a) \leq \epsilon$, then C is a true core space of some maximum reducible subspace in OR.*

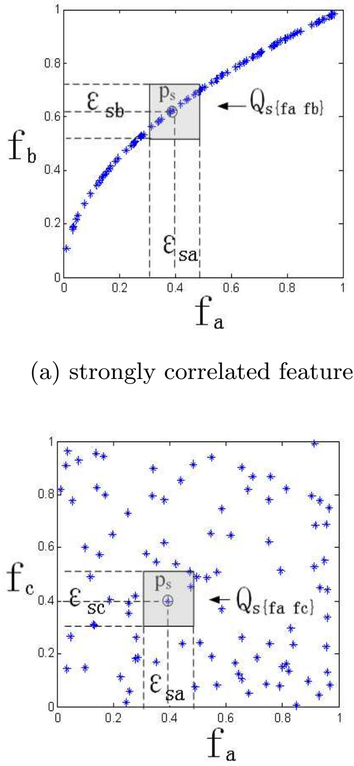
PROOF. Let $Y = \{f_y | \Delta ID(C, f_y) \leq \epsilon, f_y \in OR\}$. Following the process of finding OR, we know that Y includes all and only the features in Ω_F that are strongly correlated with C. Thus $\exists C \subset Y$, such that C satisfies Criterion (1) in Definition 3.6, and Criterion (2) in Definition 3.7. Moreover, according to Corollary 5.2, C is non-redundant. Hence C also satisfies Criterion (2) of Definition 3.6. Thus Y is a maximum reducible subspace with core space C. □

In this method, for each candidate core space, we need to calculate $\Delta ID(C)$ and $\Delta ID(C \cup \{f_a\})$ for every $f_a \in OR$ in order to get the value of $\Delta ID(C, f_a)$. However, the intrinsic dimensionality calculation is computationally expensive. Since the intrinsic dimensionality estimation is inherently approximate, we propose in the following section a method utilizing the point distribution in feature subspaces to distinguish whether a feature is strongly correlated with a core space.

5.2 Point Distribution Based Method

After finding the overall reducible subspace OR, we can apply the following heuristic to examine if a feature is strongly correlated with a feature subspace. The intuition behind our heuristic is similar to the one behind the correlation dimension.

Assume that the number of data points N in the dataset approaches infinity, and the features in the dataset are normalized so that the points are distributed from 0 to 1 in each dimension. Let $p_s \in \Omega_P$ be an arbitrary point in the dataset, and $0 < l < 1$ be a natural number. Let ξ_{sy} represent the interval of length l on feature f_y centered at p_s. The expected number of points within the interval ξ_{sy} is lN. For d features $C = \{f_{c_1}, f_{c_2}, \cdots, f_{c_d}\}$, let Q_{sC} be the d-dimensional hypercube formed by the intervals ξ_{sc_i} ($f_{c_i} \in C$). If the d features in C are totally uncorrelated, then the expected number of points in Q_{sC} is $l^d N$. Let f_m be another feature in the dataset, and $C' = \{f_{c_1}, f_{c_2}, \cdots, f_{c_d}, f_m\}$. If f_m is determined by $\{f_{c_1}, f_{c_2}, \cdots, f_{c_d}\}$, i.e., f_m is strongly correlated with C, then C' has intrinsic dimensionality d. The expected number of points in the d-dimensional hypercube, $Q_{sC'}$, which is embedded in the $(d+1)$-dimensional space of C', is still $l^d N$. If, on the other hand, f_m is uncorrelated with any feature subspace of $\{f_{c_1}, f_{c_2}, \cdots, f_{c_d}\}$, then C' has dimensionality $d+1$, and the expected number of points in the $(d+1)$-dimensional hypercube $Q_{sC'}$ is $l^{(d+1)}N$. The difference between the number of points in the cubes of these two cases is $l^d(1-l)N$.



(a) strongly correlated features

(b) uncorrelated features

Figure 4: Point distributions in correlated feature subspace and uncorrelated feature subspace

Figure 4(a) and 4(b) show two examples on 2-dimensional spaces. In both examples, $d = 1$ and $C = \{f_a\}$. In Figure 4(a), feature f_b is strongly correlated with f_a. Feature f_c is uncorrelated with f_a, as shown in Figure 4(b). The randomly sampled point p_s is at the center of the cubes $Q_{s\{f_a,f_b\}}$ and $Q_{s\{f_a,f_c\}}$. The point density in cube $Q_{s\{f_a,f_b\}}$ is clearly much higher than the point density in cube $Q_{s\{f_a,f_c\}}$ due to the strong correlation between f_a and f_b.

Therefore, for each candidate core space, we can check if a feature is correlated with it in the following way. We randomly sample n points $P = \{p_{s_1}, p_{s_2}, \cdots, p_{s_n}\}$ from the dataset. Suppose that $C = \{f_{c_1}, f_{c_2}, \cdots, f_{c_d}\}$ is the current candidate core space. For feature $f_a \in OR$ ($f_a \notin C$), let $C' = \{f_{c_1}, f_{c_2}, \cdots, f_{c_d}, f_a\}$. Let $\delta_{s_iC'}$ represent the number of points in the cube $Q_{s_iC'}$. $P' = \{p_{s_i} | \delta_{s_iC'} \geq l^{(d+1)}N\}$ is the subset of the sampled points such that the cube centered at them have more points than expected if f_a is uncorrelated with C. We say f_a is strongly correlated with C if $\frac{|P'|}{|P|} \geq \tau$, where τ is a threshold close to 1.

Concerning the choice of l, we can apply the following reasoning. If we let $l = (\frac{1}{N})^{\frac{1}{d+1}}$, then the expected number of points in the cube $Q_{s_i C'}$ is 1, if f_a is uncorrelated with C. If f_a is correlated with C, then the expected number of points in the cube $Q_{s_i C'}$ is greater than 1. In this way, we can set l according to the size of the candidate core space.

The second step of REDUS is shown in Algorithm 1 from Line 8 to Line 18. Note that in the worst case, the algorithm needs to enumerate all possible feature subspaces. However, in practice, the algorithm is very efficient since once an individual reducible subspace is found, all its features are removed. Only the remaining features need to be further examined.

6. EXPERIMENTS

To evaluate REDUS, we apply it on both synthetic datasets and real datasets. REDUS is implemented using Matlab 7.0.4. The experiments are performed on a 2.4 GHz PC with 1G memory running WindowsXP system.

6.1 Parameter Setting

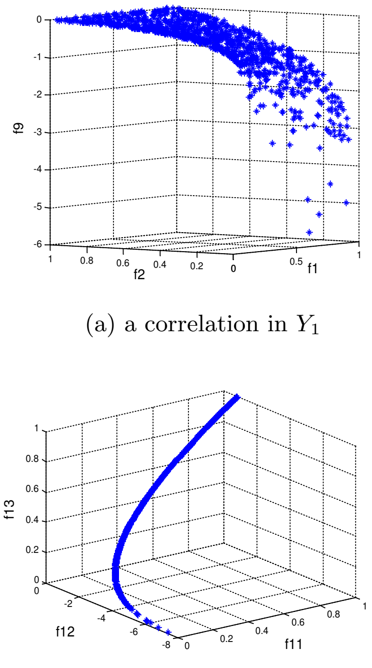
As shown in Algorithm 1, REDUS generally requires three input parameters: ϵ, n, and τ. In the first step of finding the overall reducible subspace, ϵ is the threshold to filter out the irrelevant features. Since features strongly correlated with some core space can only change intrinsic dimensionality a small amount, the value of ϵ should be close to 0. According to our experience, a good starting point is 0.1. After finding the reducible subspaces, the user can apply the standard dimensionality reduction methods to see if the are really correlated, and the adjust ϵ value accordingly to find stronger or weaker correlations in the subspaces. In all our experiments, we set ϵ between 0.002 to 0.25. In the second step, n is the point sampling size and τ is the threshold to determine if a feature is strongly correlated with a candidate core space. In our experiments, n is set to be 10% of the total number of data points in the dataset, and τ is set to be 90%.

6.2 Synthetic Datasets

6.2.1 *Effectiveness Evaluation*

To evaluate the effectiveness of the REDUS, we generate two synthetic datasets.

Synthetic dataset 1: The first synthetic dataset is as shown in Figure 1. There are 12 features, $\{f_1, f_2, \cdots, f_{12}\}$, and 1000 data points in the dataset. 3 reducible subspaces: a 2-dimensional Swiss roll, a 1-dimensional helix-shaped line, and a 2-dimensional plane, are embedded in different 3-dimensional spaces respectively. The overall reducible subspace is $\{f_1, f_2, \cdots, f_9\}$. Let c_i $(1 \leq i \leq 4)$ represent constants and r_j $(1 \leq j \leq 3)$ represent random vectors. The generating function of the Swiss roll is: $t = \frac{3}{2}\pi(1 + 2r_1)$, $s = 21r_2$, $f_1 = t\cos(t)$, $f_2 = s$, $f_3 = t\sin(t)$. The roll is then rotated 45° counter clockwise on feature space $\{f_2, f_3\}$. The helix-shaped line is generated by: $f_4 = c_1 r_3$, $f_5 = c_2 \sin(r_3)$, $f_6 = c_2 \cos(r_3)$. The 2-dimensional plane is generated by $f_9 = c_3 f_7 + c_4 f_8$. The remaining 3 features $\{f_{10}, f_{11}, f_{12}\}$ are random vectors consisting of noise data points.



(a) a correlation in Y_1

(b) a correlation in Y_2

Figure 5: Examples of embedded correlations in synthetic dataset 2

ϵ	Precision	Recall
0.06	83%	100%
0.05	91%	100%
0.04	96%	100%
0.03	100%	100%
0.02	100%	100%
0.01	100%	100%
0	100%	90%

Table 1: Accuracy of finding the overall reducible subspace when varying ϵ

In the first step, with $\epsilon = 0.25$, REDUS successfully uncovers the overall reducible space. The parameter setting for the second step is $\tau = 90\%$, and point sampling size 10%. We run REDUS 10 times. In all 10 runs, REDUS successfully identifies the individual maximum reducible subspaces from the overall reducible subspace.

Synthetic dataset 2: We generate another larger synthetic dataset as follows. There are 50 features, $\{f_1, f_2, \cdots, f_{50}\}$ and 1000 data points in the dataset. There are 3 reducible subspaces: $Y_1 = \{f_1, f_2, \cdots, f_{10}\}$ reducible to a 2-dimensional space, $Y_2 = \{f_{11}, f_{12}, \cdots, f_{20}\}$ reducible to a 1-dimensional space, and $Y_3 = \{f_{21}, f_{22}, \cdots, f_{30}\}$ reducible to a 2-dimensional space. The remaining features contain random noises. Figures 5(a) and 5(b) show two examples of the embedded correlations in 3-dimensional subspaces. Figure 5(a) plots the point distribution on feature subspace $\{f_1, f_2, f_9\}$ of Y_1, and Figure 5(b) plots the point distribution on feature subspace $\{f_{11}, f_{12}, f_{13}\}$ of Y_2.

τ	**maximum reducible subspaces identified**
0.94	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{11, 13, 15, 19\}$ $\{12, 14, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30\}$
0.92	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 22, 24, 25, 26, 27, 28, 29, 30\}$
0.90	$\{1, 2, 3, 4, 5, 6, 7, 9\}$ $\{8, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 22, 24, 25, 27, 28, 29, 30\}$
0.88	$\{1, 2, 3, 4, 5, 6, 7, 8\}$ $\{9, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 23, 25, 26, 27, 30\}$ $\{24, 29\}$

(a) varying τ

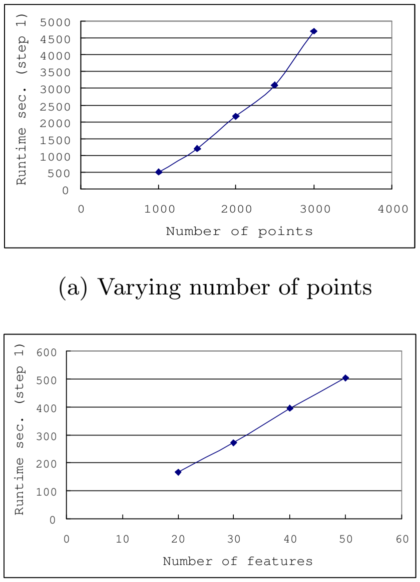
n/N	**maximum reducible subspaces identified**
20%	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 22, 24, 25, 26, 27, 28, 29, 30\}$
15%	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 22, 24, 25, 26, 27, 28, 29, 30\}$
10%	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 22, 24, 25, 26, 27, 28, 29, 30\}$
5%	$\{1, 2, 3, 4, 5, 6, 7, 9\}$ $\{8, 10\}$ $\{11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$ $\{21, 22, 24, 25, 27, 28, 29\}$ $\{23, 26, 30\}$

(b) varying n

Table 2: Accuracy of identifying the maximum reducible subspaces from the overall reducible subspace when varying τ and n

We apply REDUS on this synthetic dataset using various parameter settings. Table 1 shows the accuracy of finding the overall reducible subspace when ϵ taking different values. The recall is defined as $TP/(TP+FN)$, and the precision is defined as $TP/(TP+FP)$, where TP represents the number of true positive, FP represents the number of false positive, and FN represents the number of false negative. As we can see, REDUS is very accurate and robust to ϵ.

Tables 2(a) and 2(b) show the identified maximum reducible subspaces when varying τ and n. Table 2(a) shows results under different settings of τ. The point sampling size n in this table is the default value, i.e., 10% of the total number of data points. Although changing τ may cause some mis-classified features, Table 2(a) still shows that REDUS achieves reasonably high accuracy under different settings of τ. The reason for different decompositions of maximum reducible subspaces is that features in different maximum reducible subspaces may still have moderate correlations. If these correlated features are identified, they will be removed from the overall reducible subspace, since REDUS focuses on the case where the maximum reducible subspaces are non-overlapping. If we allow each feature to be in different maximum reducible subspaces, the findings under different τ should be more similar to each other. Finding overlapping reducible subspaces is computationally more demanding, and is an interesting problem worth further exploration.



(a) Varying number of points

(b) Varying number of features

Figure 6: Efficiency evaluation of finding the overall reducible subspace

Table 2(b) shows the identified maximum reducible subspaces when varying the number of the sampled points n. τ is set to be 0.92 in this table. As shown in the table, REDUS is not sensitive to the size of the sampled data points.

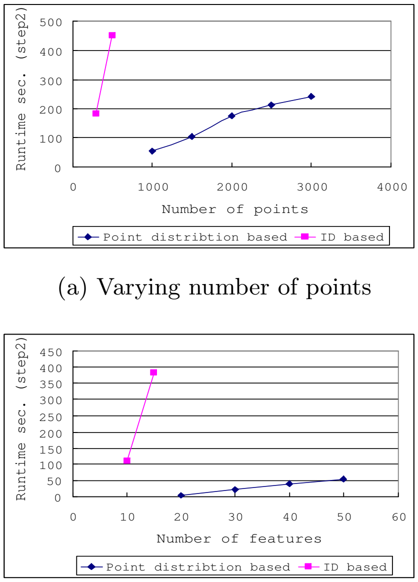
6.2.2 *Efficiency Evaluation*

To evaluate the efficiency and scalability of REDUS, we apply it to synthetic dataset 2. The default dataset for efficiency evaluation contains 1000 points and 50 features if not specified otherwise. The default values for the parameters are the same as before.

Figure 6(a) shows the runtime of finding the overall reducible subspace when varying the number of data points. The runtime scales roughly quadratically. This is because when computing the correlation dimensions, we need to calculate all pairwise distances between the data points, which is clearly quadratic to the number of points.

Figure 6(b) shows that the runtime of finding the overall reducible subspace is linear to the number of features. This is because REDUS only scans every feature once to examine if it is strongly correlated with the subspace of the remaining features. This linear scalability is desirable for the datasets containing a large number of features.

Figures 7(a) and 7(b) show the runtime comparisons between using the correlation dimension as intrinsic dimensionality estimator and the point distribution heuristic to identify the individual maximum reducible subspaces from the overall reducible subspaces. Since the calculation of intrinsic dimensionality is relatively expensive, the program



(a) Varying number of points

(b) Varying number of features

Figure 7: Efficiency evaluation of identifying maximum reducible subspaces from the overall reducible subspace

often cannot finish in a reasonable amount of time. Using the point distribution heuristics, on the other hand, is much more efficient and scales linearly to the number of points and features in the dataset.

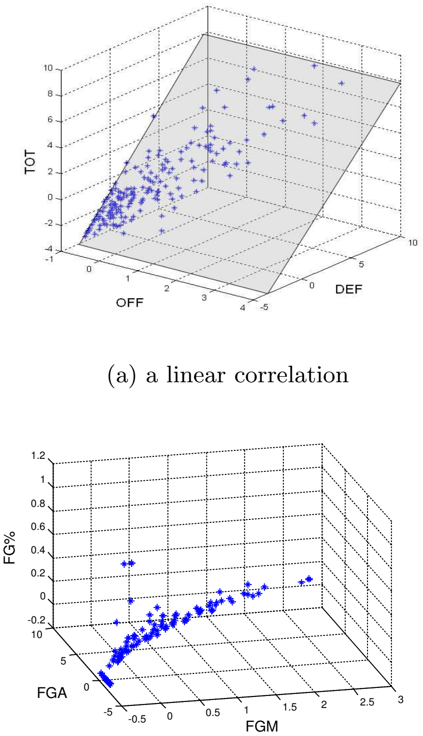
6.3 Real Life Datasets

6.3.1 NBA dataset

We apply REDUS on the NBA statistics dataset. The dataset can be downloaded from $http : //sports.espn.go.com/nba/teams/stats?team = Bos\&year = 2007\&season = 2$. It contains the statistics of 28 features for 200 players of season 2006-2007. Since the features have different value scales, we normalized each feature such that points are distributed between 0 and 1. We report two interesting correlations found in the dataset in Figures 8(a) and 8(b). Note that the features shown in the figures are mean-centered.

The correlation shown in Figure 8(a) says that the feature subspace of three features: defence rebounds (DEF), the offense rebounds (OFF), and the total number of rebounds (TOT), is strongly correlated and reducible to a 2-dimensional space. This is an obvious correlation that one would expect. As shown in the figure, the points are linearly distributed on a 2-dimensional plane in the 3-dimensional subspace.

Figure 8(b) shows a nonlinear correlation identified by REDUS. The feature subspace of three features: field goal made (FGM), field goal attempted (FGA), and the percentage of field goal (FG%) is strongly correlated and reducible to 2-dimensional space. Clearly, the data points on this 3-dimensional space are distributed on a 2-dimensional manifold.



(a) a linear correlation

(b) a nonlinear correlation

Figure 8: Correlations identified in the NBA dataset

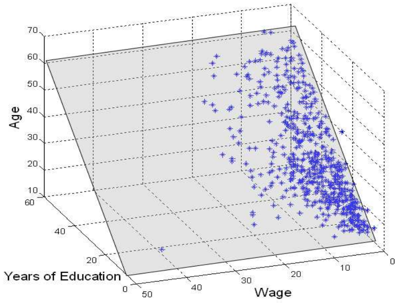


Figure 9: A linear correlation in the wage dataset

6.3.2 Wage dataset

The wage dataset from the 1985 Current Population Survey consists of 11 features in 534 data points. The dataset is available at $http : //lib.stat.cmu.edu/datasets/CPS\_85\_Wages$. The numerical features are age, years of education, years of work experience, and wage. We apply REDUS on this dataset to find the strongly correlated feature subsets.

REDUS identifies one correlation between the features: age, years of education, and wage. Figure 9 shows the point distribution of the data points in this 3-dimensional feature subspace. From this figure, we can see that wage is clearly a linear function of age and years of education. Therefore, this 3-dimensional space is reducible to the 2-dimensional plane embedded in it.

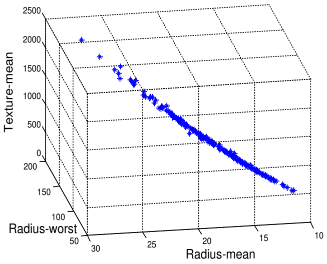


Figure 10: A correlation in the breast cancer dataset

6.3.3 Breast cancer dataset

We apply REDUS to the breast cancer dataset which is available at the UCI Machine Learning Archieve. There are 569 data points and 30 features in this dataset. The features include the statistics of radius, texture, perimeter, area, smoothness, compactness concavity, concave points, symmetry, and fractal dimension. These features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image.

Figure 10 shows one of the nonlinear correlations identified by REDUS. The three features are the mean of radius, largest radius, and the mean of texture. From the figure, we can see these three features are strongly correlated and reducible to 1-dimensional space.

7. CONCLUSION

In this paper, we investigate the problem of finding strongly correlated feature subspaces in high dimensional datasets. The correlation can be linear or nonlinear. Such correlations hidden in feature subspace may be invisible to the global feature transformation methods, such as PCA and ISOMAP. Utilizing the concepts of intrinsic dimensionality, we formalize this problem as the discovery of maximum reducible subspaces in the dataset. An effective algorithm, REDUS, is presented to find the maximum reducible subspaces. The experimental results show that REDUS can effectively and efficiently find these interesting local correlations.

Our work reported in this paper focuses on the case where the maximum reducible subspaces are non-overlapping. For future work, one interesting direction is to extend current work to the general case where a feature can be in multiple maximum reducible subspaces. Another interesting direction is finding the feature subspaces that are strongly correlated on a subset of data points. This is a more general problem and has wider applications.

8. ACKNOWLEDGMENTS

This research was partially supported by EPA grant STAR-RD832720, NSF grants IIS-0448392, IIS-0534580, and a Microsoft New Faculty Fellowship.

9. REFERENCES

[1] C. Aggarwal and P. Yu. Finding generalized projected clusters in high dimensional spaces. *SIGMOD*, 2000.

[2] A. Alizadeh and et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–11, 2000.

[3] D. Barbara and P. Chen. Using the fractal dimension to cluster datasets. *KDD*, 2000.

[4] M. Belkin and P. Niyogi. Şlaplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.

[5] A. Belussi and C. Faloutsos. Self-spacial join selectivity estimation using fractal concepts. *ACM Transactions on Information Systems*, 16(2):161–201, 1998.

[6] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[7] C. Bohm, K. Kailing, P. Kroger, and A. Zimek. Computing clusters of correlation connected objects. *SIGMOD*, 2004.

[8] I. Borg and P. Groenen. *Modern multidimensional scaling.* New York: Springer, 1997.

[9] F. Camastra and A. Vinciarelli. Estimating intrinsic dimension of data with a fractal-based approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002.

[10] T. M. Cover and J. A. Thomas. *The Elements of Information Theory.* Wiley & Sons, New York, 1991.

[11] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–68, 1998.

[12] C. Faloutsos and I. Kamel. Beyond uniformity and independence: analysis of r-trees using the concept of fractal dimension. *PODS*, 1994.

[13] K. Fukunaga. Intrinsic dimensionality extraction. *Classification, Pattern recongnition and Reduction of Dimensionality, Volume 2 of Handbook of Statistics*, pages 347–360, P. R. Krishnaiah and L. N. Kanal eds., Amsterdam, North Holland, 1982.

[14] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 20(2):165–171, 1976.

[15] A. Gionis, A. Hinneburg, S. Papadimitriou, and P. Tsaparas. Dimension induced clustering. *KDD*, 2005.

[16] G. Golub and A. Loan. *Matrix computations.* Johns Hopkins University Press, Baltimore, Maryland, 1996.

[17] V. Iyer and et. al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

[18] I. Jolliffe. *Principal component analysis.* New York: Springer, 1986.

[19] M. Kendall and J. D. Gibbons. *Rank Correlation Methods.* New York: Oxford University Press, 1990.

[20] D. C. Lay. *Linear Algebra and Its Applications.* Addison Wesley, 2005.

[21] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 2005.

[22] H. Liu and H. Motoda. *Feature selection for knowledge discovery and data mining.* Boston: Kluwer Academic Publishers, 1998.

[23] B.-U. Pagel, F. Korn, and C. Faloutsos. Deflating the dimensionality curse using multiple fractal dimensions. *ICDE*, 2000.

[24] S. Papadimitriou, H. Kitawaga, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. *ICDE*, 2003.

[25] S. N. Rasband. *Chaotic Dynamics of Nonlinear Systems.* Wiley-Interscience, 1990.

[26] H. T. Reynolds. *The analysis of cross-classifications.* The Free Press, New York, 1977.

[27] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

[28] M. Schroeder. *Fractals, Chaos, Power Lawers: Minutes from an Infinite Paradise.* W. H. Freeman, New York, 1991.

[29] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500):2319–2323, 2000.

[30] A. K. H. Tung, X. Xin, and B. C. Ooi. Curler: Finding and visualizing nonlinear correlation. *SIGMOD*, 2005.

[31] L. Yu and H. Liu. Feature selection for high-dimensional data: a fast correlation-based filter solution. *ICML*, 2003.

[32] X. Zhang, F. Pan, and W. Wang. Care: Finding local linear correlations in high dimensional data. *ICDE*, 2008.

[33] Z. Zhao and H. Liu. Searching for interacting features. *IJCAI*, 2007.