Two-stage Framework for Visualization of Clustered High Dimensional Data

Jaegul Choo*
College of Computing
Georgia Institute of Technology
266 Ferst Drive, Atlanta, GA 30332, USA

Shawn Bohn†
National Visualization and Analytics Center
Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA 99354, USA

Haesun Park*
College of Computing
Georgia Institute of Technology
266 Ferst Drive, Atlanta, GA 30332, USA

ABSTRACT

In this paper, we discuss dimension reduction methods for 2D visualization of high dimensional clustered data. We propose a two-stage framework for visualizing such data based on dimension reduction methods. In the first stage, we obtain the reduced dimensional data by applying a supervised dimension reduction method such as linear discriminant analysis which preserves the original cluster structure in terms of its criteria. The resulting optimal reduced dimension depends on the optimization criteria and is often larger than 2. In the second stage, the dimension is further reduced to 2 for visualization purposes by another dimension reduction method such as principal component analysis. The role of the second-stage is to minimize the loss of information due to reducing the dimension all the way to 2. Using this framework, we propose several two-stage methods, and present their theoretical characteristics as well as experimental comparisons on both artificial and real-world text data sets.

Keywords: dimension reduction, linear discriminant analysis, principal component analysis, orthogonal centroid method, 2D projection, clustered data, regularization, generalized singular value decomposition

Index Terms: H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces—Theory and methods

1 INTRODUCTION

Within the visual analytics community, various types of information content are represented using high dimensional signatures. To make these signatures useful they often need to be transformed into a lower dimension (i.e., 2D or 3D) for a variety of visual representations such as scatter plots. Many researchers in this community have used a wide assortment of dimension reduction techniques, e.g., self-organizing map (SOM) [12], principal component analysis (PCA) [11], multidimensional scaling (MDS) [2], etc. However, it is not always clear why a certain technique has been chosen over another, especially to the end user. Typically, its goal can be viewed in terms of two aspects: efficiency and accuracy. Efficiency as defined here is the time to compute the reduction, but accuracy may not be as simple to quantify. Many would amiably agree to quantify accuracy as a measure of the relationship preservation in the high dimensional space to the reduced dimensional space. Note that most techniques either directly or indirectly work on this principle.

There are other properties that are important to those interpreting the semantics of the reduced space. Specifically, we note that while local neighbor preservation is important it depends upon the analysis task. No single reduction technique will provide the complete view as various properties of the space are obscured or lost. We have mentioned that typically the primary objective is relationship preservation. However, there are at least two others: outlier and macro structure visualization. Outliers are conceptually easy (i.e., a variance beyond some threshold), but more difficult to quantify, as we do not necessarily know which set of outliers are important to accentuate to the user. Certain techniques (e.g., PCA) tend to show outliers more readily, however tend to compress the reduced space at the expense of showcasing the outliers. Other techniques (e.g., SOM) maximize space usage well, but do so at the expense of masking or even hiding those outliers. Likewise, macro structures of the high dimensional space may be masked or massively distorted during the reduction. Macro structures are those larger order groupings (e.g., clusters) that exist in the original dimensional space. We recognize they are important in dimension reduction research and to those in the visual analytics community. However, few of them focus on data representation especially for visualization of the clustered data [20, 13, 3].

We propose theoretical measures for these properties and efficient algorithms which will aid not only the researchers but ultimately the users/analysts to better understand which balance of properties are important and for which analytic tasks.

2 MOTIVATION

The focus of this paper is the fundamental characteristics of dimension reduction techniques for visualizing high dimensional data in the form of a 2D scatter plot when the data has cluster structure. The role of dimension reduction here is to give a 2-dimensional representation of data while preserving cluster structure as much as possible. To this end, supervised dimension reduction methods that incorporate cluster information such as linear discriminant analysis (LDA) [4] or orthogonal centroid method (OCM) [10] can be naturally considered.

However, one of the issues is that with many dimension reduction methods designed to preserve the cluster structure in the data, the theoretically optimal reduced dimension, which is the smallest dimension that is acceptable with respect to the optimization criteria of the dimension reduction method, is usually larger than 2. For example, in LDA, the minimum reduced dimension that preserves the cluster structure quality measure defined as a trace maximization problem is one less than the number of clusters in the data in general [8, 7].

In this case, one may simply choose the two dimensions that contribute most to such a measure. However, with only two dimensions, such a measure may become significantly smaller than the original quantity after dimension reduction. This results in loss of information that hinders visualization in properly reflecting the true cluster relationship of the data. A similar situation may occur when using PCA for visualizing the data not having a cluster structure. Even though PCA finds the principal axes that maximally capture the variance of the data, when the resulting 2-dimensional representation of the data maintains only a small fraction of the total variance, the relationships of the data in 2 dimension are likely to be highly inconsistent with those in the original dimension.

Such loss of information is inevitable in that the dimension has to be reduced to 2. Our main motivation is to deal with such loss more carefully by separating the loss-introducing stage from the original dimension reduction methods. Based on this idea, we propose the two-stage framework of dimension reduction for visualization.

*e-mail: {joyfull, hpark}@cc.gatech.edu
†e-mail:shawn.bohn@pnl.gov

In this framework, a supervised dimension reduction method is applied in the first stage so that the original dimension is reduced to the minimum dimension achievable while preserving the quality of cluster measure as defined in a dimension reduction method. The reduced dimension achieved in the first stage is often larger than 2. Thus in the second stage, we find another dimension reducing transformation that minimizes the loss introduced in further reducing the dimension all the way to 2. This two-stage framework provides us with a means to flexibly apply different types of dimension reduction techniques in each stage and to systematically analyze their effects, which provides understanding the effects of the overall dimension reduction methods for visualization of clustered data. The issues then are the design of the most appropriate dimension reduction methods, the modeling of optimization criteria, and the corresponding solution methods.

In this paper, we present both theoretical and empirical answers to these issues. Specifically, we propose several two-stage methods utilizing linear dimension reduction methods such as LDA, orthogonal centroid method (OCM), and principal component analysis (PCA), and we present their theoretical justifications by modeling the optimization criteria for which each method provides the optimal solution. Also, we illustrate and compare the effectiveness of the proposed methods by showing empirical visualization on synthetic and real-world data sets.Although nonlinear dimension reduction methods such as MDS or other manifold learning methods such as isometric feature mapping [16] and locally linear embedding [14] may also be utilized for the effective 2D visualization of high dimensional data, our focus in this paper is on linear methods. The linear methods are computationally more efficient in general, and unlike most of the manifold learning methods, they also provide dimension reducing transformations that can be applied to map and visualize unseen data points in the same space where the existing data are visualized.

Our approach to successively apply two dimension reduction methods should be discerned from the previous works [18, 19, 21] in that they usually aim for improving computational efficiency, scalability, or applicability of a certain dimension reduction method, e.g., LDA.

The rest of this paper is organized as follows. In Section 3, LDA, OCM, and PCA are described based on a unified framework of the scatter matrices and their trace optimization problems. In Section 4, we formulate two-stage dimension reduction methods, and in Section 5, several two-stage methods for visualization are proposed and compared along with their criteria. Experimental comparisons are given using artificial and real-world data sets in Section 6, and conclusion and future work are addressed in Section 7.

3 Dimension Reduction as Trace Optimization Problem

In this section, we introduce the notions of scatter matrices used in defining cluster quality and optimization criteria for dimension reduction.

Suppose a dimension reducing linear transformation $G^T \in \mathbb{R}^{l\times m}$ maps an m-dimensional data vector x to a vector z in an l-dimensional space ($m > l$):

$$G^T : x \in \mathbb{R}^{m\times 1} \rightarrow z = G^T x \in \mathbb{R}^{l\times 1}. \tag{1}$$

Suppose also that a data matrix $A = [a_1\, a_2 \cdots a_n] \in \mathbb{R}^{m\times n}$ is given where the columns a_j, $j = 1, \ldots, n$, of A represent n data items in an m-dimensional space, and they are partitioned into k clusters. Without loss of generality, for simplicity of notations, we further assume that A is partitioned as

$$A = [A_1 \quad A_2 \quad \cdots \quad A_k], \text{ where } A_i \in \mathbb{R}^{m\times n_i} \text{ and } \sum_{i=1}^{k} n_i = n.$$

Let $\mathcal{N}_i$ denote the set of column indices that belong to cluster i, and n_i the size of $\mathcal{N}_i$. The i-th cluster centroid $c^{(i)}$ and the global centroid c are defined, respectively, as

$$c^{(i)} = \frac{1}{n_i}\sum_{j\in\mathcal{N}_i} a_j \text{ and } c = \frac{1}{n}\sum_{j=1}^{n} a_j.$$

The scatter matrix within the i-th cluster $S_w^{(i)}$, the within-cluster scatter matrix S_w, the between-cluster scatter matrix S_b, and the total (or mixture) scatter matrix S_t are defined [9, 15], respectively, as

$$\begin{aligned}
S_w^{(i)} &= \sum_{j\in\mathcal{N}_i} (a_j - c^{(i)})(a_j - c^{(i)})^T, \\
S_w &= \sum_{i=1}^{k} S_w^{(i)} = \sum_{i=1}^{k}\sum_{j\in\mathcal{N}_i} (a_j - c^{(i)})(a_j - c^{(i)})^T, && (2)\\
S_b &= \sum_{i=1}^{k}\sum_{j\in\mathcal{N}_i} (c^{(i)} - c)(c^{(i)} - c)^T = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T \\
&= \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j (c^{(i)} - c^{(j)})(c^{(i)} - c^{(j)})^T, \text{ and} && (3)\\
S_t &= \sum_{j=1}^{n} (a_j - c)(a_j - c)^T. && (4)
\end{aligned}$$

Note that the total scatter matrix S_t is related to S_w and S_b as [9]

$$S_t = S_w + S_b. \tag{5}$$

When G^T in Eq. (1) is applied to the matrix A, the scatter matrices S_w, S_b, and S_t in the original dimensional space are reduced to the $l \times l$ matrices

$$G^T S_w G,\ G^T S_b G, \text{ and } G^T S_t G,$$

respectively. By computing the trace of the scatter matrices as

$$\begin{aligned}
\text{trace}(S_w) &= \sum_{i=1}^{k}\sum_{j\in\mathcal{N}_i} (a_j - c^{(i)})^T (a_j - c^{(i)}) \\
&= \sum_{i=1}^{k}\sum_{j\in\mathcal{N}_i} \|a_j - c^{(i)}\|_2^2, && (6)\\
\text{trace}(S_b) &= \sum_{i=1}^{k}\sum_{j\in\mathcal{N}_i} (c^{(i)} - c)^T (c^{(i)} - c) \\
&= \sum_{i=1}^{k} n_i \|c^{(i)} - c\|_2^2 && (7)\\
&= \frac{1}{n}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j \|c^{(i)} - c^{(j)}\|_2^2, \text{ and} && (8)\\
\text{trace}(S_t) &= \sum_{j=1}^{n} (a_j - c)^T (a_j - c) = \sum_{j=1}^{n} \|a_j - c\|_2^2, && (9)
\end{aligned}$$

we obtain values that can be used to measure the cluster quality. Note that from Eqs. (7) and (8), trace(S_b) can be viewed as the squared sum of the pairwise distances between cluster centroids as well as that of the distances between each centroid and the global centroid.

The cluster structure quality can be defined by analyzing how well each cluster can be discriminated from each other. High quality clusters usually have small trace(S_w) and large trace(S_b), relating to the small variance within each cluster and the large distances between clusters. Subsequently, dimension reduction methods may be intended to maximize trace$(G^T S_b G)$ and minimize

$\text{trace}(G^T S_w G)$ in the reduced dimensional space. This simultaneous optimization can be approximated to a single criterion as

$$J_{b/w}(G) = \max \text{trace}((G^T S_w G)^{-1}(G^T S_b G)), \tag{10}$$

which is the criterion of LDA. In addition, one may focus on maximizing the distances between clusters, which can be represented as the criterion of OCM, i.e.,

$$J_b(G) = \max_{G^T G = I} \text{trace}(G^T S_b G). \tag{11}$$

On the other hand, regardless of cluster dependent terms, S_w and S_b, the trace of the total scatter matrix S_t can be maximized as

$$J_t(G) = \max_{G^T G = I} \text{trace}(G^T S_t G), \tag{12}$$

which turns out to be the criterion of PCA. In Eqs. (11) and (12), without the constraint, $G^T G = I$, $J_b(G)$ and $J_t(G)$ can become arbitrarily large.

In what follows, LDA, OCM, and PCA are discussed based on such maximization criteria, and their properties relevant to visualization are identified.

3.1 Linear Discriminant Analysis (LDA)

Conceptually, in LDA, we are looking for a dimension reducing transformation that keeps the between-cluster relationship as remote as possible by maximizing $\text{trace}(G^T S_b G)$ while keeping the within cluster relationship as compact as possible by minimizing $\text{trace}(G^T S_w G)$. As shown in Eq. (10), the criterion of LDA can be written as

$$J_{b/w}(G) = \max \text{trace}((G^T S_w G)^{-1}(G^T S_b G)). \tag{13}$$

It can be shown that for any $G \in \mathbb{R}^{m \times l}$ where $m > l$,

$$\text{trace}((G^T S_w G)^{-1}(G^T S_b G)) \leq \text{trace}(S_w^{-1} S_b), \tag{14}$$

meaning that the cluster structure quality measured by $\text{trace}(S_w^{-1} S_b)$ cannot be increased after dimension reduction [4]. By setting the derivative of Eq. (13) with respect to G to zero, which gives the first order optimality condition, it can be shown that the solution of LDA, where we denote it as G_{LDA}, has the columns which are the leading generalized eigenvectors u of the generalized eigenvalue problem,

$$S_b u = \lambda S_w u. \tag{15}$$

Since the rank of S_b is at most $k-1$, LDA achieves the upper bound of $\text{trace}((G^T S_w G)^{-1}(G^T S_b G))$ in Eq. (14) for any l such that $l \geq k-1$, i.e.,

$$\begin{aligned}\text{trace}((G_{LDA}^T S_w G_{LDA})^{-1}(G_{LDA}^T S_b G_{LDA})) \\ = \text{trace}(S_w^{-1} S_b) \text{ for } l \geq k-1,\end{aligned} \tag{16}$$

which indicates $\text{trace}(S_w^{-1} S_b)$ is preserved between the original space and the reduced dimensional space obtained by G_{LDA}.

3.2 Orthogonal Centroid Method (OCM)

Orthogonal centroid method (OCM) [10] focuses only on maximizing $\text{trace}(G^T S_b G)$ under the constraint of $G^T G = I$. The criterion of OCM is shown as

$$J_b(G) = \max_{G^T G = I} \text{trace}(G^T S_b G). \tag{17}$$

It is known that for any $G \in \mathbb{R}^{m \times l}$ where $m > l$ such that $G^T G = I$,

$$\text{trace}(G^T S_b G) \leq \text{trace}(S_b), \tag{18}$$

which means the cluster structure quality measured by $\text{trace}(S_b)$ cannot be increased after dimension reduction. The solution of Eq. (17) can be obtained by setting the columns of G as the leading eigenvectors of S_b. Since S_b has at most $k-1$ nonzero eigenvalues, the upper bound of $\text{trace}(G^T S_b G)$ in Eq. (18) can be achieved for any l such that $l \geq k-1$, i.e.,

$$\text{trace}(G^T S_b G) = \text{trace}(S_b) \text{ for } l \geq k-1. \tag{19}$$

Eq. (19) indicates $\text{trace}(S_b)$ is preserved between the original and the reduced dimensional spaces.

An advantage of OCM is that it achieves an upper bound of $\text{trace}(G^T S_b G)$ more efficiently by using QR decomposition, avoiding the eigendecomposition. The algorithm of OCM is as follows. First the centroid matrix C is formed so that each column of C is composed of each cluster's centroid vector, i.e., $C = \begin{bmatrix} c_1 & c_2 & \cdots & c_k \end{bmatrix}$. Then the reduced QR decomposition [5] of C is computed for $C = Q_k R$ where $Q_k \in \mathbb{R}^{m \times k}$ with $Q_k^T Q_k = I$ and $R \in \mathbb{R}^{k \times k}$ is upper triangular. The solution of OCM, G_{OCM}, is found as

$$G_{OCM} = Q_k.$$

Note that the columns of G_{OCM} are composed of the orthogonal bases for the subspace spanned by the centroids, and $l = k$ in this case. Finally, OCM achieves

$$\text{trace}(G_{OCM}^T S_b G_{OCM}) = \text{trace}(S_b), \text{ where } l = k.$$

By using the equivalence between Eqs. (3) and (3), one can prove that each pairwise distance between cluster centroids is also preserved in the reduced dimensional space obtained by OCM.

Another important property of OCM is that by projecting data into the subspace spanned by the centroids, the order of similarities between any particular point and centroids are preserved in terms of Euclidean norm and cosine similarity measure [10, 7]. In other words, for any vector $q \in \mathbb{R}^{m \times 1}$ and cluster centroids $c^{(i)}$ and $c^{(j)}$, we have

$$\begin{aligned}&\|q - c^{(i)}\|_2 < \|q - c^{(j)}\|_2 \Rightarrow \\ &\|G_{OCM}^T q - G_{OCM}^T c^{(i)}\|_2 < \|G_{OCM}^T q - G_{OCM}^T c^{(j)}\|_2, \text{ and} \\ &\frac{q^T c^{(i)}}{\|q\|_2 \|c^{(i)}\|_2} < \frac{q^T c^{(j)}}{\|q\|_2 \|c^{(j)}\|_2} \Rightarrow \\ &\frac{\left(G_{OCM}^T q\right)^T G_{OCM}^T c^{(i)}}{\|G_{OCM}^T q\|_2 \|G_{OCM}^T c^{(i)}\|_2} < \frac{\left(G_{OCM}^T q\right)^T G_{OCM}^T c^{(j)}}{\|G_{OCM}^T q\|_2 \|G_{OCM}^T c^{(j)}\|_2}.\end{aligned}$$

3.3 Principal Component Analysis (PCA)

PCA is a well-known dimension reduction method that captures the maximal variance in the data. The criterion of PCA can be written as

$$J_t(G) = \max_{G^T G = I} \text{trace}(G^T S_t G). \tag{20}$$

For any $G \in \mathbb{R}^{m \times l}$ where $m > l$ such that $G^T G = I$, we have

$$\text{trace}(G^T S_t G) \leq \text{trace}(S_t), \tag{21}$$

which means $\text{trace}(S_t)$ cannot be increased after dimension reduction. The solution of Eq. (20), where we denote it as G_{PCA}, can be obtained by setting the columns of G as the leading eigenvectors of S_t. Since the rank of S_t is at most $\min(m, n)$, PCA achieves the upper bound of $\text{trace}(G^T S_t G)$ in Eq. (21) for any l such that $l \geq \min(m, n)$, i.e.,

$$\text{trace}(G_{PCA}^T S_t G_{PCA}) = \text{trace}(S_t) \text{ for } l \geq \min(m, n).$$

Table 1: Comparison of dimension reduction methods. It is assumed S_b and S_t are full rank.

	LDA	OCM	PCA
Optimization Criterion ($x \in \mathbb{R}^{m\times 1} \xrightarrow{G^T} y \in \mathbb{R}^{l\times 1}$)	$J_{b/w}(G) =$ $\max \text{trace}((G^T S_w G)^{-1}(G^T S_b G))$	$J_b(G) = \max_{G^T G=I} \text{trace}(G^T S_b G)$	$J_t(G) = \max_{G^T G=I} \text{trace}(G^T S_t G)$
Algorithm	generalized eigendecomposition	QR decomposition	symmetric eigendecomposition
Smallest dimension achieving the criterion upper bound	$k-1$	k	$\min(m,n)$

In many applications of PCA, however, l is usually chosen as a fixed value less than the ranke of S_t for the purpose of dimension reduction or noise reduction. This noisy subspace corresponds to the smallest eigenvectors of S_t, and they are removed by PCA for better representation of the data.

Although S_t is related to S_b and S_w as in Eq. (5), S_t as it is does not contain any information on cluster labels. That is, unlike LDA and OCM, PCA ignores the cluster structure represented by S_b and/or S_w, which is why PCA is considered as an unsupervised dimension reduction method.

Usually, PCA assumes that the global centroid is zero by subtracting the empirical mean of the data from each data vector. The centered data can be represented as $A - ce^T$, where e is n-dimensional vector whose components are all 1's.

PCA has a unique property that, given a fixed l, it produces the best reduced dimensional representation that minimizes the difference between the centered matrix $A - ce^T$ and its projection to the reduced dimensional space $GG^T(A - ce^T)$ where G has orthonormal columns, i.e.,

$$G_{PCA} = \arg \min_{G, G^T G = I_l} \|GG^T(A - ce^T) - (A - ce^T)\|,$$

where the matrix norm $\|\cdot\|$ is either a Frobenius norm or a Euclidean norm.

The three discussed methods are summarized in Table 1.

4 Formulation of Two-stage Framework for Visualization

Suppose we want to find a dimension reducing linear transformation $V^T \in \mathbb{R}^{2\times m}$ that maps an m-dimensional data vector x to a vector z in a 2-dimensional space ($m \gg 2$):

$$V^T : x \in \mathbb{R}^{m\times 1} \to z = V^T x \in \mathbb{R}^{2\times 1}. \tag{22}$$

Further assume that it is composed of two stages of dimension reductions as follows. In the first stage, a dimension reducing linear transformation $G^T \in \mathbb{R}^{l\times m}$ maps an m-dimensional data vector x to a vector y in the l-dimensional space ($l \ll m$):

$$G^T : x \in \mathbb{R}^{m\times 1} \to y = G^T x \in \mathbb{R}^{l\times 1}, \tag{23}$$

where l is fixed as its minimum optimal dimension by the first-stage criterion. When $l \leq 2$, we have no further dimension reduction to do after the first step. However, an optimal l in many methods and for many data sets is larger than 2, and so we assume that $l > 2$.

In the second stage, another dimension reducing linear transformation $H^T \in \mathbb{R}^{2\times l}$ maps an l-dimensional data vector y to a vector z in the 2-dimensional space($l > 2$):

$$H^T : y \in \mathbb{R}^{l\times 1} \to z = H^T y \in \mathbb{R}^{2\times 1}. \tag{24}$$

Such consecutive dimension reductions performed by G^T followed by H^T can be combined, resulting in a single dimension reducing transformation V^T as

$$V^T = H^T G^T. \tag{25}$$

In the next section, discussion will be focused on various ways for choosing the first stage dimension reducing transformation G and the second stage dimension transformation H with a purpose to construct combined dimension reducing transformation $V^T = H^T G^T$ for 2-dimensional visualization according to various optimization criteria.

5 Two-stage Methods for 2D Visualization

All the proposed two-stage methods start from one of the supervised dimension reduction methods such as LDA or OCM that are designed for clustered data. In the first stage (by $G^T \in \mathbb{R}^{l\times m}$ in Eq. (23)), the dimension is reduced by LDA or OCM to the smallest dimension that satisfies Eq. (16) or (19), respectively. Therefore in the first stage, the cluster structure quality measured either by $\text{trace}(S_w^{-1} S_b)$ or $\text{trace}(S_b)$ is preserved. Then we perform the second-stage dimension reduction (by $H^T \in \mathbb{R}^{2\times l}$ in Eq. (24)) that minimizes the loss of information either by applying the same criterion used in the first stage or by using J_t in Eq. (20), i.e., that of PCA. As seen in Section 3.3, Eq. (20) gives the best approximation of the first-stage results that minimize the difference in terms of Frobenius/Euclidean norm.

In what follows, we describe each of the two-stage methods in detail, and derive their equivalent single-stage methods (by $V^T \in \mathbb{R}^{2\times m}$ in Eq. (22)) in case they exist.

5.1 Rank-2 LDA

In this method, LDA is applied in the first stage, and $\text{trace}(S_w^{-1} S_b)$ is preserved in the l-dimensional space where $l = k-1$. In the second stage, the same criterion $J_{b/w}(H)$ is used to reduce the l-dimensional first-stage results to 2-dimensional data.

The criterion of the second-stage dimension reducing matrix H can be formulated as

$$H_{b/w} = \max_{H\in\mathbb{R}^{l\times 2}} \text{trace}((H^T(G_{LDA}^T S_w G_{LDA})H)^{-1} (H^T(G_{LDA}^T S_b G_{LDA})H)). \tag{26}$$

Assuming the columns of G_{LDA}, which are generalized eigenvectors of Eq. (15), are in decreasing order of their corresponding generalized eigenvalues, i.e., $G_{LDA} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k-1} \end{bmatrix}$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{k-1}$, the solution of Eq. (26) is

$$H_{b/w} = \begin{bmatrix} e_1 & e_2 \end{bmatrix},$$

where e_1 and e_2 are the first and the second standard unit vectors, i.e., $e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \in \mathbb{R}^{l\times 1}$ and $e_2 = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}^T \in \mathbb{R}^{l\times 1}$. This solution is equivalent to choosing two dimensions with the most leading generalized eigenvalues from the first stage result, and the resulting two-stage method can be represented as a single-stage dimension reduction method by $V \in \mathbb{R}^{m\times 2}$ which directly maximize $J_{b/w}$, i.e.,

$$\begin{aligned} V_{b/w} &= \arg \max_{V\in\mathbb{R}^{m\times 2}} J_{b/w}(V) \\ &= \arg \max_{V\in\mathbb{R}^{m\times 2}} \text{trace}((V^T S_w V)^{-1}(V^T S_b V)). \end{aligned} \tag{27}$$

Table 2: Summary of the optimization criteria of the two-stage dimension reduction methods.

	Rank-2 LDA	LDA + PCA	OCM+PCA	Rank-2 PCA on S_b
Stage 1: Preservation ($x \in \mathbb{R}^{m\times 1} \xrightarrow{G^T} y \in \mathbb{R}^{l\times 1}$)	$\text{trace}((G^T S_w G)^{-1}(G^T S_b G)) = \text{trace}(S_w^{-1} S_b)$		$\text{trace}(G^T S_b G) = \text{trace}(S_b)$	
Stage 2: Maximization ($y \in \mathbb{R}^{l\times 1} \xrightarrow{H^T} z \in \mathbb{R}^{2\times 1}$)	$\text{trace}((H^T(G^T S_w G)H)^{-1}(H^T(G^T S_b G)H))$	$\underset{H^T H=I}{\text{trace}}(H^T(G^T S_t G)H)$	$\underset{H^T H=I}{\text{trace}}(H^T(G^T S_t G)H)$	$\underset{H^T H=I}{\text{trace}}(H^T(G^T S_b G)H)$

The solution of Eq. (27) becomes

$$V_{b/w} = G_{LDA} H_{b/w} = \left[\begin{array}{cc} u_1 & u_2 \end{array} \right],$$

where $u_1$and u_2 are the leading generalized eigenvectors of Eq. (15). This solution is also known as reduced-rank linear discriminant analysis [6].

5.2 LDA followed by PCA

In this method, LDA is applied in the first stage, and $\text{trace}(S_w^{-1} S_b)$ is preserved in the l-dimensional space where $l = k-1$. In the second stage, PCA is applied in order to obtain the best approximation of the l-dimensional first-stage results in terms of Frobenius/Euclidean norm.

The second-stage dimension reducing matrix H is obtained by solving

$$H_t = \arg \max_{H\in\mathbb{R}^{l\times 2}, H^T H=I} \text{trace}(H^T (G_{LDA}^T S_t G_{LDA}) H), \qquad (28)$$

where the solution is the two leading eigenvectors of the total scatter matrix of the first-stage result, $G_{LDA}^T S_t G_{LDA}$.

From Eq. (5), we have

$$G_{LDA}^T S_t G_{LDA} = G_{LDA}^T (S_b + S_w) G_{LDA}. \qquad (29)$$

Since LDA conceptually maximizes $\text{trace}(G^T S_b G)$ and minimizes $\text{trace}(G^T S_w G)$, the result is expected to satisfy

$$\text{trace}(G_{LDA}^T S_b G_{LDA}) \gg \text{trace}(G_{LDA}^T S_w G_{LDA})),$$

which means that $G_{LDA}^T S_t G_{LDA}$ is dominated by $G_{LDA}^T S_b G_{LDA}$, i.e.,

$$G_{LDA}^T (S_b + S_w) G_{LDA} \simeq G_{LDA}^T S_b G_{LDA}.$$

In this case, the principal axes that PCA gives in the second stage better reflect those of the between-cluster matrix of the first-stage result, $G_{LDA}^T S_b G_{LDA}$, and they may in turn discriminate the clusters clearly in the 2-dimensional space. In this sense, LDA followed by PCA achieves a clear cluster structure as well as a good approximation of the first-stage result.

5.3 OCM followed by PCA

In this method, OCM is applied in the first stage, and $\text{trace}(S_b)$ is preserved in the l-dimensional space where $l = k$. In the second stage, PCA is applied in order to obtain the best approximation of the l-dimensional first-stage results in terms of Frobenius/Euclidean norm.

As in Section 5.2, the second-stage dimension reducing matrix H is obtained by solving

$$H_t = \arg \max_{H\in\mathbb{R}^{l\times 2}, H^T H=I} \text{trace}(H^T (G_{OCM}^T S_t G_{OCM}) H), \qquad (30)$$

where the solution is the two leading eigenvectors of the total scatter matrix of the first-stage result, $G_{OCM}^T S_t G_{OCM}$.

From Eq. (5), we have

$$G_{OCM}^T S_t G_{OCM} = G_{OCM}^T (S_b + S_w) G_{OCM}. \qquad (31)$$

Unlike LDA, OCM does not minimize $\text{trace}(G^T S_w G)$ as shown in Eq. (17). Therefore the following may not be the case:

$$\text{trace}(G_{OCM}^T S_b G_{OCM}) \gg \text{trace}(G_{OCM}^T S_w G_{OCM}),$$

which means that $G_{OCM}^T S_b G_{OCM}$ does not necessarily dominate $G_{OCM}^T S_t G_{OCM}$. Then the two principal axes of $G_{OCM}^T S_t G_{OCM}$ obtained by PCA in the second stage tend to fail to reflect those of $G_{OCM}^T S_b G_{OCM}$, which may rather scatter the data points within each cluster, eventually preventing the visualization results from showing a clear cluster structure.

5.4 Rank-2 PCA on S_b

In this method, OCM is applied in the first stage, and $\text{trace}(S_b)$ is preserved in the l-dimensional space where $l = k$. In the second stage, the same criterion $J_b(H)$ is used to reduce the l-dimensional first-stage results to 2-dimensional data.

The second-stage dimension reducing matrix H is obtained by solving

$$H_b = \arg \max_{H\in\mathbb{R}^{l\times 2}, H^T H=I} \text{trace}(H^T (G_{OCM}^T S_b G_{OCM}) H), \qquad (32)$$

where the solution is the two leading eigenvectors of the between-scatter matrix of the first-stage result, $G_{OCM}^T S_b G_{OCM}$. The columns of G_{OCM} form the subspace spanned by centroids, and this subspace includes the range space of S_b. Accordingly, one can easily show that the eigenvector $u_i^Y \in \mathbb{R}^{l\times 1}$ of $G_{OCM}^T S_b G_{OCM}$ is related to eigenvectors $u_i \in \mathbb{R}^{m\times 1}$ of S_b as

$$u_i^Y = G_{OCM}^T u_i$$

with their corresponding eigenvalues matched as well, i.e., $\lambda_i^Y = \lambda_i$. Hence, the solution of Eq. (32) can be written as

$$H_b = \left[\begin{array}{cc} u_1^Y & u_2^Y \end{array} \right] = G_{OCM}^T \left[\begin{array}{cc} u_1 & u_2 \end{array} \right]. \qquad (33)$$

Using Eq. (33) and the relationship shown in Eq. (25), the single-stage dimension reducing transformation V_b can be built as

$$\begin{aligned} V_b^T &= H_b^T G_{OCM}^T = \left[\begin{array}{c} u_1^T \\ u_2^T \end{array} \right] G_{OCM} G_{OCM}^T \\ &= \left[\begin{array}{c} u_1^T \\ u_2^T \end{array} \right] \qquad (34) \\ &= \arg \max_{V\in\mathbb{R}^{m\times 2}} J_b(V) \\ &= \arg \max_{V\in\mathbb{R}^{m\times 2}} \text{trace}(V^T S_b V). \qquad (35) \end{aligned}$$

Eq. (34) holds since the eigenvectors of S_b, u_1 and u_2, are in the range space of G_{OCM}. The criterion of Eq. (35) has been used in one of the successful visual analytic systems, IN-SPIRE, for 2D representation of document data [17].

The discussed two-stage methods are summarized in Table 2.

6 Experiments

In this section, we present visualization results using the proposed methods for several data sets, especially focusing on undersampled text data visualization where the data item is represented in m-dimensional space and the number of the data items n is less than m ($m > n$).

Table 3: Description of data sets.

	GAUSSIAN	MEDLINE	NEWSGROUPS	REUTERS
Original dimension, m	1100	22095	16702	3907
Number of data items, n	1000	500	770	800
Number of clusters, k	10	5	11	10

6.1 Regularization on LDA for undersampled data

In undersampled cases, the LDA criterion shown in Eq. (13) cannot be applied directly because S_w is singular. In order to overcome this singularity problem, Howland et al. proposed a universal algorithmic framework of LDA using the generalized singular value decomposition (LDA/GSVD) [8, 7]. Specifically, for the case when $m \gg n \gg k$, which is usual for most undersampled problems, LDA/GSVD gives the solution for G such that $G^T S_w G = 0$ while maintaining the maximum value of trace$(G^T S_b G)$. This solution makes sense since LDA criterion is formulated to minimize trace$(G^T S_w G)$. However, it means that all of the data points belonging to a specific cluster are represented as a single point in the reduced dimensional space, which lessens the generalization ability for classification as well as for visualizing the individual data relationship within each cluster.

On the contrary, the fact that LDA makes $G^T S_w G = 0$ can be viewed as an advantage for visualization purposes since LDA has the capability to fully minimize trace$(G^T S_w G)$. By means of regularization on S_w one can control trace$(G^T S_w G)$, which determines the scatter of the data points within each cluster. In regularized LDA which was originally proposed to avoid the singularity of S_w in classification context, S_w is replaced by a nonsingular matrix $S_w + \gamma I$ where I is an identity matrix, and γ is a control parameter. In general, as γ is increased, the within-cluster distance, trace$(G^T S_w G)$, also becomes larger with data points being more scattered around their corresponding centroids. As γ is decreased, the within-cluster distance becomes smaller, and the data points gather closer around their centroids. Such manipulation of γ can be exploited in a visualization context because one can choose an appropriate value of γ so that the second-stage method such as PCA, which maximizes trace$(G^T S_t G) =$ trace$(G^T S_b G + G^T S_w G)$, does not focus too much on trace$(G^T S_w G)$. The results that follow are based on such choices of γ.

6.2 Data Sets

The data sets tested are composed of one artificially-generated Gaussian-mixture dataset (GAUSSIAN) and three real-world text data sets (MEDLINE, NEWSGROUPS, and REUTERS) that are clustered based on their topics. All the text documents are encoded as term-document matrices where each dimension corresponds to a particular word, and the value of a certain dimension represents the frequency of the corresponding word shown in the document. Each data set is set to have an equal number of data per cluster, and have a mean of zero which is attained by subtracting the global mean. (See Section 6.3.)

The descriptions of data sets, which are also summarized in Table 3, are as follows.

The GAUSSIAN data set is a randomly generated Gaussian mixture with 10 clusters. Each cluster is made up of 100 data vectors, which add up to 1000 in total, and the dimension is set to 1100, which is slightly more than the number of the data items. In its visualization shown in Fig. (1), the clusters are labeled using letters as

- 'a', 'b', ..., and 'j'.

The MEDLINE data set is a document corpus related to medical science from the National Institutes of Health[1]. The original dimension is 22095, and the number of clusters is 5, where each cluster has 100 documents. The cluster labels that correspond to the document topics are shown as

- heart attack ('h'), colon cancer ('c'), diabetes ('d'), oral cancer ('o'), and tooth decay ('t'),

where the letters in parentheses are used in the visualization shown in Fig. (2).

The NEWSGROUPS data set [1] is a collection of newsgroup documents, and originally composed of 20 topics. However, we have chosen 11 topics for visualization, and each cluster is set to have 70 documents. The original dimension is 16702, and the cluster labels are shown as

- comp.sys.ibm.pc.hardware ('p'), comp.sys.mac.hardware ('a'), misc.forsale ('f'), rec.sport.baseball ('b), sci_crypt ('y'), sci.electronics ('e'), sci.med ('d'), soc.religion.christian ('c'), talk.politics.guns ('g'), talk.politics.misc ('p'), and talk.religion.misc ('r'),

where the letters in parentheses are used in the visualization shown in Fig. (3).

The REUTERS data set [1] is the document collection that appeared in the Reuters newswire in 1987, and originally composed of hundreds of topics. Among them, 10 topics related to economic subjects are chosen for visualization, and each cluster has 80 documents. The original dimension is 3907, and the cluster labels are shown as

- earn ('e'), acquisitions ('a'), money-fx ('m'), grain ('g'), crude ('r'), trade ('t'), interest ('i'), ship ('s'), wheat ('w'), and corn ('c'),

where the letters in parentheses are used in the visualization shown in Fig. (4).

6.3 Effects of Data Centering

Fig. 5 is the example of applying OCM+PCA to the MEDLINE data set with and without data centering. Once the MEDLINE data set is encoded as a term-document matrix, every component has a non-negative value, which results in the global centroid that is significantly far from the origin. Then performing PCA without data centering might give the first principal axis as the one reflecting the global centroid rather than that discriminating clusters. If we consider projecting the data onto each of the horizontal and the vertical axes in Fig. 5, the former, which corresponds to the first principal axis, does not help in showing the cluster structure clearly, and only the vertical axis, which corresponds to the second principal axis from PCA, discriminates clusters. We have found that such undesirable behavior is common in many cases without data centering, which is why we assume that data is centered throughout this paper. Accordingly, all the results shown in Figs 1-4 are obtained after data centering.

6.4 Comparison of Visualization Results

The results of four two-stage methods for the tested data sets are shown in Figs.1-4[2].

In all cases, LDA-based methods show cluster structures more clearly than OCM-based methods. This proves the effectiveness of LDA that considers both within- and between-cluster measures while OCM only takes into account the latter. Due to this difference, OCM generally produces a widely-scattered data representation within each cluster. As a result, in the NEWSGROUPS dataset, such a wide within-cluster variance significantly deteriorates the

[1] http://www.cc.gatech.edu/~hpark/data.html

[2] Those figures can be arbitrarily magnified without losing the resolution in the electronic version of this paper.

cluster structure visualization even if OCM still attempts to maximize the between-cluster distance.

In the MEDLINE and the REUTERS data sets, all of the four methods produce relatively similar results. However, we have controlled the within-cluster variance in LDA-based methods using the regularization term γI. In addition, the fact that rank-2 LDA and LDA+PCA produce almost identical results indicates that $G_{LDA}^T S_t G_{LDA}$ is dominated by $G_{LDA}^T S_b G_{LDA}$ after LDA is applied in the first stage as we expected.

Rank-2 LDA represents each cluster most compactly by minimizing the within-cluster radii both in the first and the second stage. However, it may reduce the between-cluster distances as well because $J_{b/w}$ maximizes the conceptual ratio of two scatter measures. As can be seen in the two LDA-based methods applied to the NEWGROUPS data set, while rank-2 LDA minimizes the within-cluster radii, it also places the centroids closer to each other as compared to those in LDA+PCA. Due to this effect, which one is preferable between rank-2 LDA and LDA+PCA depends on the data set to be visualized.

Overall, OCM+PCA and Rank-2 PCA on S_b show similar results. It means $G^T S_b G \simeq G^T S_t G$ in that the difference between two methods lies in whether PCA is applied to $G^T S_b G$ or $G^T S_t G$ in the second stage. Since performing PCA on $G^T S_b G$ is computationally more efficient than PCA on $G^T S_t G$, Rank-2 PCA on S_b can be a good alternative to OCM+PCA in case efficient computation is important.

Finally, these visualization results reveal the interesting cluster relationships underlying in the data. In Fig. (2), the clusters for colon cancer ('c') and oral cancer ('o') are shown close to each other. In Fig. (3), the clusters of soc.religion.christian ('c') and talk.religion.misc ('r'), those of comp.sys.ibm.pc.hardware ('p') and comp.sys.mac.hardware ('a'), and those of sci.crypt ('y') and sci.med ('d') are closely located respectively in LDA-based methods. In addition, the two clusters, misc.forsale ('f') and rec.sport.baseball ('b'), are shown to be the most distinctive, which makes sense because those topics are quite irrelevant to the others. In Fig. (4), the clusters of grain ('g'), wheat ('w'), and corn ('c') as well as those of money-fx ('m') and interest ('i') are visualized very close.

7 Conclusion and Future Work

According to our results, LDA-based methods are shown to be superior to OCM-based methods since both within- and between-cluster relationships are taken into account in LDA. Especially, combined with PCA in the second stage, LDA+PCA achieves a clear discrimination between clusters as well as the best approximation of the results of LDA when the distance between data is measured in terms of Frobenius/Euclidean norm.

However, many classes except for few of them that are clearly unrelated tend to be overlapped especially when dealing with large numbers of data points and clusters. This is inherently due to the nature of the second-stage dimension reduction in which only the two axes are chosen so that the classes which contribute most to the second stage criteria can be well-discriminated. Such behavior can exaggerate the distances between particular clusters, and more elaboration towards new criteria that fits in visualization is required. In the MEDLINE and the REUTERS datasets, visualization results seem to have a tail-shape along specific directions. We often found this phenomenon to occur in many other data sets. It is still unclear as to what causes this and how it affects the visualization, e.g. characteristics of information loss in the second stage. Finally, in order to determine how much loss of information is introduced by each method, more rigorous analysis based on various quantitative measures such as pairwise between-cluster distance and within-cluster radii should be conducted.

Acknowledgements

The work of these authors was supported in part by the National Science Foundation grants CCF-0732318 and CCF-0808863. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] A. Asuncion and D. Newman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007.

[2] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

[3] I. S. Dhillon, D. S. Modha, and W. S. Spangler. Class visualization of high-dimensional data with applications. *Computational Statistics & Data Analysis*, 41(1):59 – 90, 2002.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition, second edition*. Academic Press, Boston, 1990.

[5] G. H. Golub and C. F. van Loan. *Matrix Computations, third edition*. Johns Hopkins University Press, Baltimore, 1996.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[7] P. Howland, M. Jeon, and H. Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.

[8] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(8):995–1006, Aug. 2004.

[9] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

[10] M. Jeon, H. Park, and J. B. Rosen. Dimensional reduction based on centroids and least squares for efficient processing of text data. In *Proceedings of the First SIAM International Workshop on Text Mining*. Chiago, IL, 2001.

[11] I. Jolliffe. *Principal component analysis*. Springer, 2002.

[12] T. Kohonen. *Self-organizing maps*. Springer, 2001.

[13] Y. Koren and L. Carmel. Visualization of labeled data using linear transformations. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 23–30, Oct. 2003.

[14] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[15] D. L. Swets and J. J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[16] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[17] J. A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1224–1233, 1999.

[18] J. Ye and Q. Li. A two-stage linear discriminant analysis via qr-decomposition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):929–941, June 2005.

[19] H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, 34:2067–2070, 2001.

[20] X. Zhang, C. Myers, and S. Kung. Cross-weighted fisher discriminant analysis for visualization of dna microarray data. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04). IEEE International Conference on*, volume 5, pages V–589–92 vol.5, May 2004.

[21] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:336, 1998.

Figure 1: Comparison of the two-stage methods in the GAUSS data set.

(a) Rank-2 LDA (b) LDA+PCA (c) OCM+PCA (d) Rank-2 PCA on S_b

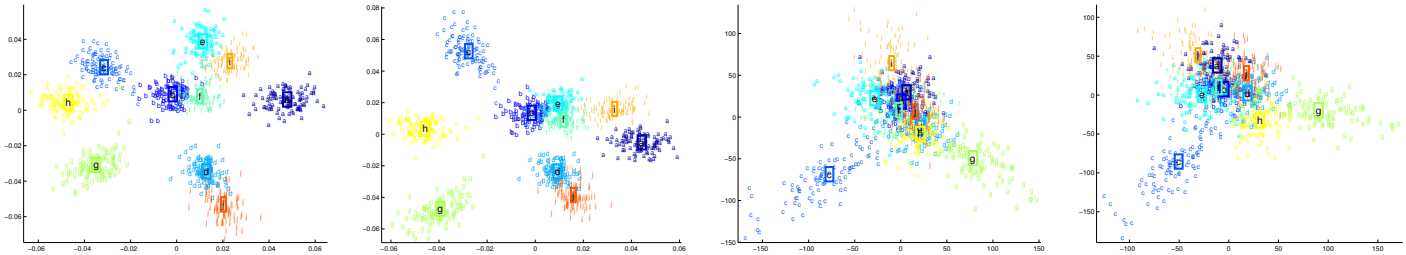

Figure 2: Comparison of the two-stage methods in the MEDLINE data set.

(a) Rank-2 LDA (b) LDA+PCA (c) OCM+PCA (d) Rank-2 PCA on S_b

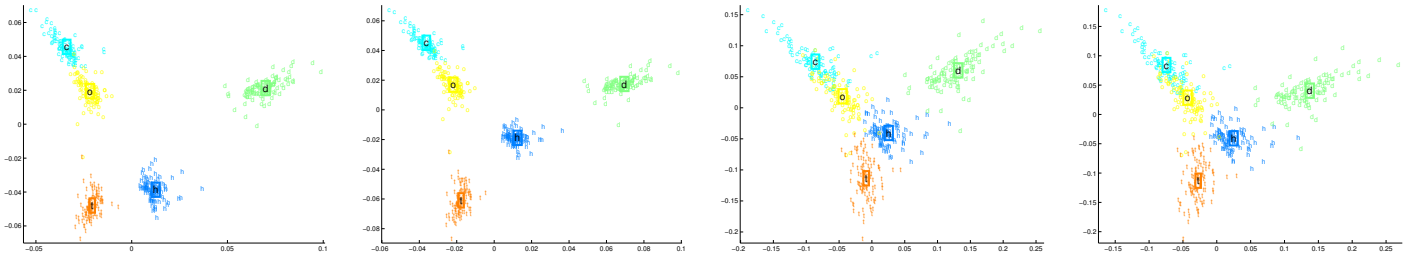

Figure 3: Comparison of the two-stage methods in the NEWSGROUPS data set.

(a) Rank-2 LDA (b) LDA+PCA (c) OCM+PCA (d) Rank-2 PCA on S_b

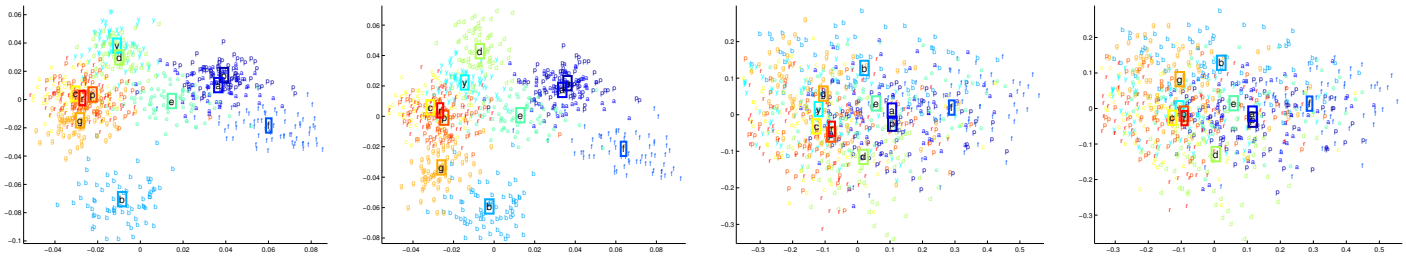

Figure 4: Comparison of the two-stage methods in the REUTERS data set.

(a) Rank-2 LDA (b) LDA+PCA (c) OCM+PCA (d) Rank-2 PCA on S_b

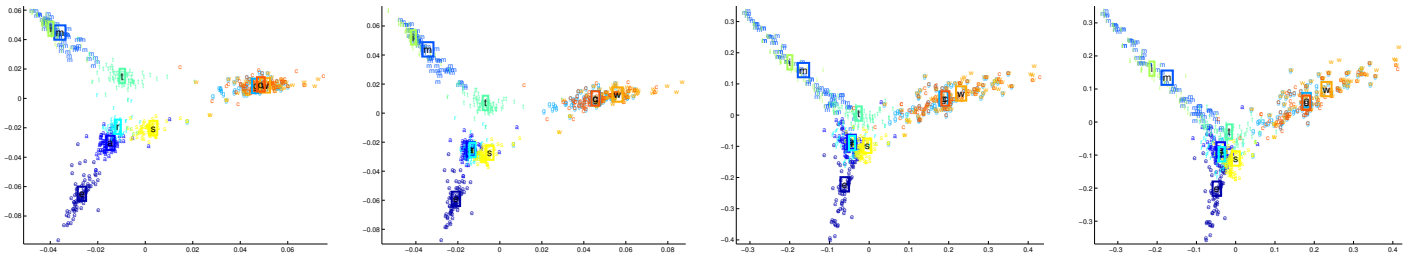

Figure 5: Example of effects of data centering in the MEDLINE data set.

(a)OCM+PCA with data centering (b)OCM+PCA without data centering