Heterogeneous Data Fusion via Space Alignment Using Nonmetric Multidimensional Scaling*

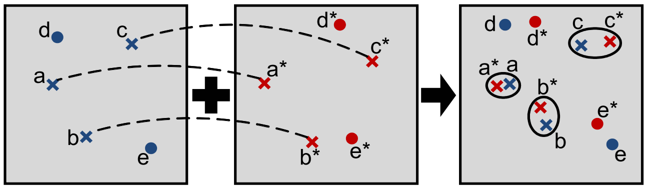
Jaegul Choo[†], Shawn Bohn[‡], Grant C. Nakamura[‡], Amanda M. White[‡], and Haesun Park[†]

Abstract

Heterogeneous data sets are typically represented in different feature spaces, making it difficult to analyze relationships spanning different data sets even when they are semantically related. Data fusion via space alignment can remedy this task by integrating multiple data sets lying in different spaces into one common space. Given a set of reference correspondence data that share the same semantic meaning across different spaces, space alignment attempts to place the corresponding reference data as close together as possible, and accordingly, the entire data are aligned in a common space. Space alignment involves optimizing two potentially conflicting criteria: minimum deformation of the original relationships and maximum alignment between the different spaces. To solve this problem, we provide a novel graph embedding framework for space alignment, which converts each data set into a graph and assigns zero distance between reference correspondence pairs resulting in a single graph. We propose a graph embedding method for fusion based on nonmetric multidimensional scaling (MDS). Its criteria using the rank order rather than the distance allows nonmetric MDS to effectively handle both deformation and alignment. Experiments using parallel data sets demonstrate that our approach works well in comparison to existing methods such as constrained Laplacian eigenmaps, Procrustes analysis, and tensor decomposition. We also present standard cross-domain information retrieval tests as well as interesting visualization examples using space alignment.

1 Introduction

Our data are often composed of multiple data types from various sources. To analyze them, raw data types are typically encoded into high-dimensional feature vectors, but such encoding schemes are entirely dependent on the data type. For example, text documents are represented using a bag-of-words model, and audio/speech data can be represented as mel-frequency cepstral coefficients. Even within the same types of data, e.g., text documents, different languages form different feature spaces because of their own vocabulary sets.



(a) Data A (b) Data B (c) Aligned data

Figure 1: Overview of the space alignment process. Dashed lines between two data sets show reference correspondence pairs to place as close as possible in the aligned space.

Although these different types of data may contain common semantic meanings, they cannot be directly compared with each other due to the fact that they lie in different feature spaces. To tackle this problem, we aim for fusion of heterogeneous data sets into a single space, which maintains the semantic coherence regardless of their data types and, in turn, makes it possible to seamlessly perform analysis across different data sets. For instance, suppose one searches for a term, ‘chicken.’ The integrated space would be able to retrieve its relevant data items of various types without any additional conversion processes.

One approach to achieve such fusion of different data sets or spaces would be to incorporate domain knowledge. For instance, when integrating multi-lingual data, we can match them in a feature level by comparing the terms between different languages [2] or even use off-the-shelf translation system. However, we deal with the problem by learning only from the data level since such domain knowledge or translation tools may not be available. In our approach, we assume that some pairs of data items that share the same semantic meaning between different spaces are available as *reference correspondence pairs*. Figure 1 depicts this idea. The dashed

*The work of these authors was supported in part by the National Science Foundation grants CCF-0728812 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.

[†]School of Computational Science and Engineering, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332, USA, {jaegul.choo, hpark}@cc.gatech.edu

[‡]Pacific Northwest National Laboratory, Richland, WA 99352, USA, {shawn.bohn, grant.nakamura, Amanda.White}@pnl.gov

lines indicate reference correspondence pairs, and the alignment process places each pair as close together as possible in the aligned space. Correspondingly, non-reference data points are aligned as well, and as a result, the proximity between them can also be revealed, as shown in the upper-left and the bottom-right parts in Figure 1(c).

In short, the goal of data fusion via space alignment is, given multiple data sets and their reference correspondence pairs, to provide the new representations of the data in a common space, which enables distance-based comparison across different data sets. Recently, this problem has been actively studied [23, 30, 32, 13, 34], and most of studies combine the problem with manifold learning approaches such as Laplacian eigenmaps [3] and Diffusion maps [7], which find a reduced dimensional space that redefines the distances along manifold surfaces.

In general, however, there is no guarantee that our data sets have such manifold structures with lower dimensions. In this respect, our alignment approach does not intend to reduce the data dimension nor discover the intrinsic manifold structure. Instead, we regard the distances computed in the original space as ideal relationships between the data, which we do not want to compromise during the alignment process.

However, as the reference points move towards their corresponding points in the aligned space, the original pairwise distances within each space go through some *deformation*. To be specific, the point configuration in Figure 1(a) is different from that of the same points in Figure 1(c). In general, as the reference correspondence pairs get tied closer for a better alignment between them, one can expect more deformation of the original space. Based on this idea, this paper focuses on how to attain the best alignment between different spaces while allowing a minimum deformation within each space. To tackle this problem, we introduce a novel graph embedding framework and propose one such method based on nonmetric multidimensional scaling (MDS) [21], which can effectively handle both deformation and alignment. Unlike metric MDS, which tries to preserve given pairwise dissimilarities[1] in a given dimensional space, nonmetric MDS tries to preserve only the rank ordering of the dissimilarities. For quantitative evaluation, we introduce neighborhood-based measures for deformation and alignment and present experimental results that compare our method with other existing ones using several real-world data sets. In addition, we also present standard cross-domain information retrieval tests as well as interesting visualization case studies using space alignment.

The rest of this paper is organized as follows. Section 2 describes a graph embedding framework for space alignment and propose a nonmetric MDS-based method under this framework. Section 3 discusses other existing approaches from our own perspective of deformation and alignment. Section 4 presents the quantitative and qualitative experimental results. Finally, Section 5 draws conclusions.

2 Space Alignment using Graph Embedding

In this section, we describe how to cast the space alignment problem into a graph embedding one. Typically, graph embedding takes input as a graph in which the vertices are data items and the edge weights are their pairwise dissimilarities or distances. As output, graph embedding produces the coordinates of the data in a given dimensional space, which best preserve the relationships described in the graph.

Suppose we have two data matrices $A = \{a_i\} \in \mathbb{R}^{m_A \times n_A}$ and $B = \{b_j\} \in \mathbb{R}^{m_B \times n_B}$ whose column vectors a_i's and b_j's represent the data items (Figure 2(a)). Without loss of generality, the reference correspondence pairs are assumed to be the first r items in each data set, i.e., (a_1, b_1), (a_2, b_2), …, and (a_r, b_r). The output of space alignment is the new vector representations of A and B, denoted as $A^f = \{a_i^f\} \in \mathbb{R}^{m^f \times n_A}$ and $B^f = \{b_j^f\} \in \mathbb{R}^{m^f \times n_B}$, respectively, in a single m^f-dimensional space.[2] In the following, the graph embedding approach for space alignment is described, which is also illustrated in Figure 2.

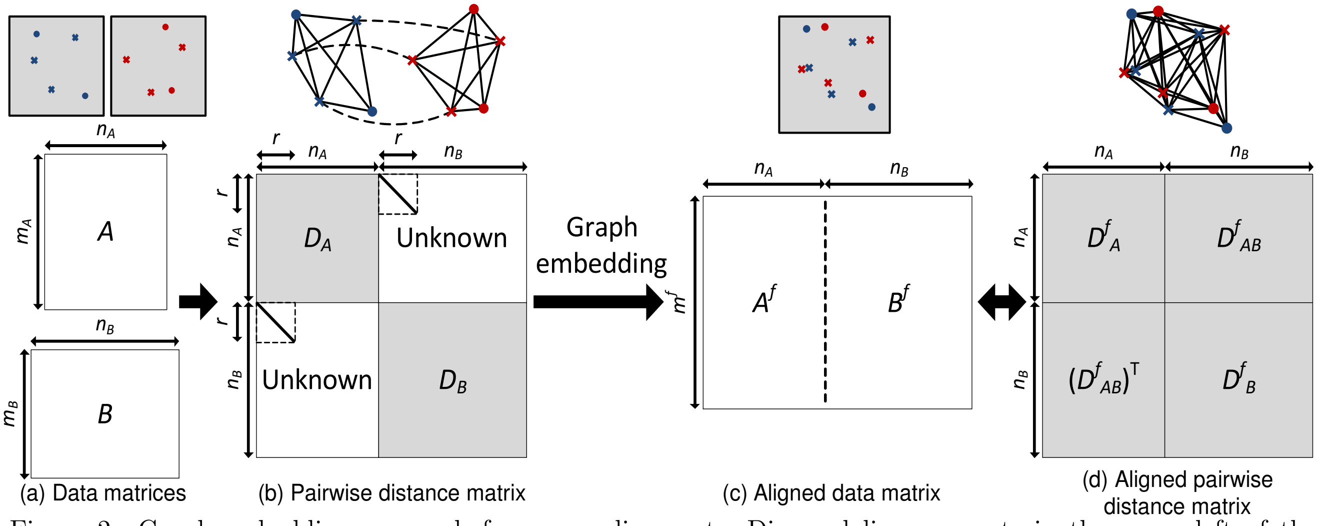
1. *Forming distance graphs from feature vectors*. Each of the data matrices A and B is turned into a dissimilarity graph, which is specified as pairwise distance matrices D_A and D_B, respectively. This way, we no longer care about the original feature spaces but only about the relationship of the data.

2. *Normalizing graphs*. This step takes care of the cases in which the distance values in one graph are significantly larger than those in the other graph, and each of the two graphs is normalized to have the same scale. Many approaches may exist, but we have used an isotropic scaling based on the mean edge weights in order not to change the relative relationships between the data. To be specific, every distance value in D_A (and D_B) is divided by the mean distance in D_A (and D_B).

3. *Assigning edges between two graphs*. Zero-distance edges are assigned between the reference correspon-

[1] Throughout this paper, we consider dissimilarities as (Euclidean) distances between data vectors.

[2] A superscript 'f' refers to an aligned (or fused) space.



(a) Data matrices (b) Pairwise distance matrix (c) Aligned data matrix (d) Aligned pairwise distance matrix

Figure 2: Graph embedding approach for space alignment. Diagonal line segments in the upper-left of the off-diagonal blocks in (b) represents zero-distance edges assigned between reference correspondence pairs.

dence pairs between the two graphs, which, in turn, forms a single connected graph involving both data sets (Figure 2(b)).

4. *Running graph embedding algorithm*. Graph embedding produces the optimal coordinates, A^f and B^f, in the aligned space, which preserve the relationships given in this graph. Zero-distance edges act as a driving force to merge two graphs in graph embedding.

In terms of the pairwise distance relationships of the aligned data, we can think of three kinds: D^f_A, D^f_B, and D^f_{AB}, which are the relationships within each data set, A^f and B^f, and those between different data sets (Figure 2(d)). Let us recall two potentially conflicting criteria, *minimum deformation* and *maximum alignment*. In the graph embedding framework, the deformation of the original data can be measured by quantifying how much D^f_A (and/or D^f_B) differs from D_A (and/or D_B). On the other hand, the alignment performance concerns how two different data sets are related to each other, i.e., D^f_{AB}. If any desirable characteristic about D^f_{AB} is available, the performance may be measured based on it. In Section 4, we further describe how we design the experiments and the data to handle such evaluation.

The graph embedding perspective of the space alignment problem opens up a wide possibility to apply a variety of methods in this problem. In fact, numerous dimension reduction methods that utilize pairwise distance relationships fall into this category. Such methods include MDS [8, 29], force-directed graph layout [20], kernel principal component analysis, and various manifold learning methods such as Laplacian eigenmaps [3], isometric feature mapping [28], maximum variance unfolding [33], and Sammon's mapping [26]. Not all methods are, however, directly applicable to space alignment since they usually assume a complete graph as input, which is not the case in space alignment, as the input graph shown in Figure 2(b) contains 'unknown' parts.

Furthermore, the primary purpose in this paper is neither dimension reduction nor discovery of intrinsic manifold, but finding a common space that reveals the relationships between different data sets by filling out these 'unknown' parts in Figure 2(b). For example, suppose we duplicate a certain data matrix, i.e., A is the same as B in Figure 2(a). Then we form a pairwise distance matrix just like Figure 2(b) by hiding the off-diagonal blocks except for some of zero-distance edges as reference correspondence pairs. If we run a particular dimension reduction method on this graph by giving the number of original dimensions as a target dimension, we expect that the original data vectors are recovered possibly up to scaling, translation, and rotation, which may not be the case for some of the above-mentioned manifold learning methods that focus on preserving locality.

In other words, unlike the previous work [23, 30, 32, 13, 34], our perspective is that the cause of deformation of the original data has to be solely the discrepancies between the configurations of two graphs due to the additional zero-distance edges between the reference points, but not the reduced dimensionality nor the emphasis on the local geometry. In this respect, which methods work well and what kinds of modification they require in data fusion should be carefully identified in both theoretical and experimental ways.

2.1 Nonmetric Multidimensional Scaling

Let us briefly describe metric MDS [8]. Metric MDS, which tries to preserve the given dissimilarity values, is formulated as

$$\min_{x_1,\,\ldots,\,x_n} \sum_{1\le i<j\le n} \left(d_{ij} - \delta_{ij}\right)^2 , \tag{2.1}$$

where d_{ij} is the dissimilarity between the reduced dimensional vectors x_i and x_j, and δ_{ij} is a given dissimilarity. Usually, the Euclidean distance, i.e., $\|x_i - x_j\|_2$, is used as d_{ij}. Incorporating a nonnegative weight w_{ij} corresponding to d_{ij}, Eq. (2.1) can be extended as

$$\min_{x_1,\,\ldots,\,x_n} \sum_{1\le i<j\le n} w_{ij}\left(d_{ij} - \delta_{ij}\right)^2 . \tag{2.2}$$

Except for a special case called classical MDS [29], which has an analytical solution via eigendecomposition, Eq. (2.2) is a non-convex problem and usually solved iteratively to reach its local minimum. One of the widely-used algorithms is stress majorization [9, 10], which simplifies Eq. (2.2) into a convex function by which Eq. (2.2) is upper-bounded and minimizes such a convex function at each iteration.

We now formulate the MDS criteria of Eq. (2.2) for space alignment. The main modification is to set the weights, w_{ij}'s in Eq. (2.2), of the unknown entries in Figure 2(b) to zero so that MDS can ignore these entries. This way, we have a metric MDS-based criterion for space alignment as

$$\begin{aligned} \min_{A^f,\,B^f} & \sum_{1\le i<j\le n_A} w_A(i,\,j)\left(D_A^f(i,\,j) - D_A(i,\,j)\right)^2 \\ & + \sum_{1\le i<j\le n_B} w_B(i,\,j)\left(D_B^f(i,\,j) - D_B(i,\,j)\right)^2 \\ & + \sum_{1\le i\le r} w_R(i)\left(D_{AB}^f(i,\,i) - 0\right)^2 , \end{aligned} \tag{2.3}$$

where $w_A(i,\,j)$ and $w_B(i,\,j)$ are the weights for the penalties on the changes of the distances between the data within A and B, respectively, and $w_R(i)$ is the weight for the penalty on the distance itself between the i-th reference correspondence pair, i.e., $(a_i^f,\,b_i^f)$, for $1\le i\le r$.

Nonmetric MDS [21], which tries to keep the rank orders of given distances in the reduced dimensional space, has a similar criterion to metric MDS. From Eq. (2.2), it replaces δ_{ij} with its approximated distance $\hat{\delta}_{ij}$, which is also referred to as a *disparity*, and such an approximation is subject to monotonicity constraints between δ_{ij}'s and $\hat{\delta}_{ij}$'s. That is, given w_{ij}'s and δ_{ij}'s, the nonmetric MDS criterion is expressed as

$$\min_{\hat{\delta}_{ij}\text{'s},\,x_1,\,\ldots,\,x_n} \sum_{1\le i<j\le n} w_{ij}\left(d_{ij} - \hat{\delta}_{ij}\right)^2 , \tag{2.4}$$

subject to $\hat{\delta}_{ij} \le \hat{\delta}_{kl}$ for all i, j, k, and l such that $\delta_{ij} < \delta_{kl}$. Notice $\hat{\delta}_{ij}$'s are also variables to be determined with the same orders as those of δ_{ij}'s. To prevent a trivial solution of all $\hat{\delta}_{ij}$'s being zeros, $\hat{\delta}_{ij}$'s are normalized to have their sum-of-squares equal to one.

This problem is basically solved by alternating an update of x_i's with an update of $\hat{\delta}_{ij}$'s. In other words, with fixed $\hat{\delta}_{ij}$'s, x_i's are obtained in the same way as in metric MDS by, say, stress majorization. To update $\hat{\delta}_{ij}$'s given fixed x_{ij}'s and their derived distances d_{ij}'s, nonmetric MDS uses a technique called isotonic regression [22, 4], which works as follows. If the order of d_{ij}'s is the same as that of δ_{ij}'s, i.e., $d_{ij} \le d_{kl}$ for all i, j, k, and l such that $\delta_{ij} < \delta_{kl}$, $\hat{\delta}_{ij}$'s would be set exactly as d_{ij}'s so that the objective function value of Eq. (2.4) can be zero, and the optimization is done at this point. When the order is not the same, the isotonic regression uses the "up-and-down blocks" algorithm [4]. Suppose $d_{ij} > d_{kl}$ and $\delta_{ij} < \delta_{kl}$ for some i, j, k, and l. Ignoring weights, the optimal value for $\hat{\delta}_{ij}$ and $\hat{\delta}_{kl}$ that minimizes $\left(d_{ij} - \hat{\delta}_{ij}\right)^2 + \left(d_{kl} - \hat{\delta}_{kl}\right)^2$ subject to $\hat{\delta}_{ij} \le \hat{\delta}_{kl}$ is $\hat{\delta}_{ij} = \hat{\delta}_{kl} = \left(d_{ij} + d_{kl}\right)/2$. Similarly, if the order of three d_{ij}'s are reversed with respect to that of δ_{ij}'s, the corresponding $\hat{\delta}_{ij}$'s would be set as the mean value of three d_{ij}'s. Involving the weights w_{ij}'s, the solution is generalized as their weighted average instead of their mean.

Now we formulate the nonmetric MDS-based criterion for space alignment. Note that in space alignment, we have ties for δ_{ij}'s since we assign the same zero distances between all the reference correspondence pairs. The tied δ_{ij}'s do not usually happen when δ_{ij}'s are computed as Euclidean distances of the original data. A natural approach would be to keep the corresponding distances as equal as possible in a target space, but in space alignment, it does not make much sense to force the same distances between all the reference correspondence pairs. In this respect, by having zero distances between the reference correspondence pairs, we only force their distances to be smaller than any nonzero distances given in D_A and D_B while not caring about which reference point pair has to be closer than another reference pair. Finally, the criterion of nonmetric MDS for space alignment can be represented as

$$\begin{aligned} \min_{A^f,\,B^f,\,\hat{D}_A,\,\hat{D}_B,\,\hat{D}_{AB}} & \\ & \sum_{1\le i<j\le n_A} w_A(i,\,j)\left(D_A^f(i,\,j) - \hat{D}_A(i,\,j)\right)^2 \\ & + \sum_{1\le i<j\le n_B} w_B(i,\,j)\left(D_B^f(i,\,j) - \hat{D}_B(i,\,j)\right)^2 \\ & + \sum_{1\le i\le r} w_R(i)\left(D_{AB}^f(i,\,i) - \hat{D}_{AB}(i,\,i)\right)^2 , \end{aligned} \tag{2.5}$$

subject to the constraints

$$\hat{D}_{AB}(i,\, i) \le \hat{D}_A(j,\, k),\, \forall(i,\, j,\, k) : 1 \le i \le r$$

$$\hat{D}_{AB}(i,\, i) \le \hat{D}_B(j,\, k),\, \forall(i,\, j,\, k) : 1 \le i \le r$$

$$\hat{D}_A(i,\, j) \le \hat{D}_A(k,\, l),\, \forall(i,\, j,\, k,\, l) : D_A(i,\, j) < D_A(k,\, l)$$

$$\hat{D}_B(i,\, j) \le \hat{D}_B(k,\, l),\, \forall(i,\, j,\, k,\, l) : D_B(i,\, j) < D_B(k,\, l)$$

Next, we discuss why nonmetric MDS is suitable for space alignment. As mentioned earlier, assigning zero-distance edges between reference correspondence pairs may cause nontrivial deformation of the original relationships within each space. Nonmetric MDS arbitrarily allows such deformation with no degradation in terms of the objective function value, Eq. (2.4), as long as the rank order remains the same. By relaxing the notion of deformation in this way, nonmetric MDS is able to focus more on maximizing alignment.

In addition, the isotonic regression described earlier provides an interesting interpretation in space alignment. Let us assume we have three semantic objects encoded in two different vector spaces. For example, we have three objects, i.e., apples, oranges, and bananas, and each of them is represented as an image and a text document. After encoding three images into one space and three documents into another space, suppose all the pairwise distances are computed. However, a disagreement between two spaces might arise. That is, in the image space, the distance from apples to oranges is shown farther than that from apples to bananas, but the opposite is shown in the document space. In this situation, suppose we perform space alignment by placing the image and the document of the same object as closely as possible, as shown in Figure 1. Then in the aligned space, between the distance from apples to oranges and that from apples to bananas, it is not clear which one to put farther than the other. Given such a disagreement, the safest option would be to make two distances approximately equal, which is the same as the way the isotonic regression behaves in order to deal with disagreements, as described above.

3 Other approaches

Recently, there have been an increasing number of results about how to align different data sets [23, 13, 32, 31], and a majority of them have been trying to impose this alignment problem in the manifold learning framework, which reveals the intrinsic low dimensional manifold structure by focusing the local neighborhood relationship. Another different stream of space alignment approaches is to seek for linear transformations that map each space into a single space [30, 5, 15, 27]. The advantage of linear methods is that out-of-sample extension is straightforward. That is, the unseen data that are not involved in an alignment process can be directly mapped into the aligned space. Furthermore, linear methods generally give an idea about the feature-level relationships between different spaces from an analysis of the linear transformation coefficients [31, 30].

In this section, we discuss three methods, one for manifold-based methods and two for linear transformation methods, in more detail from the novel perspective of the trade-off relationship between maximum alignment and minimum deformation.

3.1 Constrained Laplacian Eigenmaps

Ham et al. [13] have proposed a constrained version of Laplacian eigenmaps, which adds its original criterion with a penalty on the distances between the reference correspondence pairs. Later, this Laplacian-eigenmaps-based formulation has been further generalized to the case of more than two data sets as well as that with no known reference correspondence pairs [31, 32].

Interestingly enough, this work can be interpreted from our MDS-based graph-embedding perspective. Using the notations used in this paper, the constrained Laplacian eigenmaps criterion for space alignment is expressed as

$$\begin{aligned}(3.6)\qquad \min_{A^f,\, B^f} & \sum_{1\le i<j\le n_A} w_A(i,\, j)\left(D_A^f(i,\, j) - 0\right)^2 \\ & + \sum_{1\le i<j\le n_B} w_B(i,\, j)\left(D_B^f(i,\, j) - 0\right)^2 \\ & + \sum_{1\le i\le r} w_R(i)\left(D_{AB}^f(i,\, i) - 0\right)^2,\end{aligned}$$

where $w_A(i,\, j) = \exp\left(-\|a_i - a_j\|^2/2\sigma^2\right)$ if a_i and a_j have a neighborhood relationship or zero otherwise, and $w_B(i,\, j) = \exp\left(-\|b_i - b_j\|^2/2\sigma^2\right)$ if b_i and b_j have a neighborhood relationship or zero otherwise, and $w_R(i) = \mu$ for $\forall 1 \le i \le r$. Just as in the original Laplacian eigenmaps, A^f and B^f are subject to have a certain scaling so that it can avoid a trivial solution of all a_i's and b_i's being zero. The difference of Eq. (3.6) from the original Laplacian eigenmaps is the additional third term, which penalizes the distances between the reference correspondence pairs. Similar to the Laplacian eigenmaps formulation, Eq. (3.6) is a convex problem, which can be solved analytically as a generalized eigenvalue problem.

A comparison of Eq. (3.6) with Eq. (2.3) indicates that Eq. (3.6) tries to fit all the distances to zero instead of their original distances, and the information about the original distances is carried only in the weights $w_A(i,\, j)$'s and $w_B(i,\, j)$'s. In this way, the points whose original distances are closer than the others get bigger penalties on their distances in a target space. However, we claim that it might be too much of information loss when trying to preserve the distances in the resulting data, as we will show its experimental results in Section 4.

3.2 Orthogonal Procrustes Analysis

Orthogonal Procrustes analysis or, in short, Procrustes analysis is a well-established area of study [12, 19], and it has been widely used in image registration [18], shape analysis [11], and so on. It gives an optimal linear transformation represented as an orthogonal matrix, which best maps one matrix to the other. To be specific, given two matrices $X, Y \in \mathbb{R}^{m \times n}$, it finds an orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ that solves

$$\min_{Q} \|X - QY\|_F . \tag{3.7}$$

Being an orthogonal matrix, Q allows only high-dimensional rotations to fit Y to X. By incorporating additional translation and isotropic scaling factors, Eq. (3.7) can be extended as

$$\min_{Q,\, \mu_X,\, \mu_Y,\, k} \left\| \left(X - \mu_X \mathbf{1}_n^T\right) - kQ\left(Y - \mu_Y \mathbf{1}_n^T\right) \right\|_F , \tag{3.8}$$

where μ_X and μ_Y are m-dimensional column vectors, $\mathbf{1}_n$ is an n-dimensional column vector whose elements are all 1's, and k is a scalar.

Eq. (3.8) has an analytical solution [14, 30, 12] as follows. First, μ_X and μ_Y can be set to row-wise mean vectors of X and Y, respectively. k and Q are obtained from a singular value decomposition (SVD) of $\left(Y - \mu_Y \mathbf{1}_n^T\right)\left(X - \mu_X \mathbf{1}_n^T\right)^T$. That is, when $\left(X - \mu_X \mathbf{1}_n^T\right)\left(Y - \mu_Y \mathbf{1}_n^T\right)^T = U \Sigma V^T$, $Q = UV^T$ and $k = \text{trace}\left(\Sigma\right) / \text{trace}\left(YY^T\right)$.

Previously, Wang et al. [30] have applied Procrustes analysis in the space alignment problem, but they first apply manifold learning methods such as Laplacian eigenmaps. This approach sounds reasonable when we need to first generate the vector representations of data when only their relationships are available. In practice, however, such relationships are usually computed from the vector representations of data, and in this case, manifold learning prior to space alignment may tend to act as an artifact that changes the original relationships.

In this respect, assuming the original vector representations of data are available, we directly apply Procrustes analysis on these original data. In other words, we replace X and Y in Eq. (3.8) with the reference correspondence data within each data set, i.e., $A_r = \{a_1, \ldots, a_r\} \in \mathbb{R}^{m_A \times r}$ and $B_r = \{b_1, \ldots, b_r\} \in \mathbb{R}^{m_B \times r}$, respectively.

One potential problem in doing so is that Procrustes analysis requires the dimensions of A_r and B_r, i.e., m_A and m_B, to be the same. One way to handle this would be to apply principal component analysis, which preserves the squared sum of distances (or a variance) as much as possible, in order to obtain the same dimensional data sets. After solving Eq. (3.8) for these data, the transformation composed of Q, μ_X, μ_Y, and k is then applied to the entire data in A and B, resulting in

$$A^f = A - \mu_X \mathbf{1}_{n_A}^T \text{ and } B = kQ\left(B - \mu_X \mathbf{1}_{n_B}^T\right).$$

In terms of deformation, Procrustes analysis does not change the configuration of the original data within each space at all by allowing only rigid transformations and isotropic scaling, as shown in Eq. (3.8). Consequently, the original data relationships within A and B remain the same up to a scaling. However, the alignment performance may rather be limited when the configurations of the two data sets are significantly different, as we will show in Section 4.4.

3.3 PARAFAC2

Another linear method that can be used for space alignment is PARAFAC2 [17], a multi-way generalization of SVD for tensors. PARAFAC2 has been applied in the context of cross-language information retrieval [5], but it can be easily applied in general space alignment problems for any data types.

Given two matrices $X \in \mathbb{R}^{m_X \times n}$, $Y \in \mathbb{R}^{m_Y \times n}$, and a rank k where $k \leq \min(\text{rank}(X), \text{rank}(X))$, PARAFAC2 decomposes two matrices as

$$X \simeq U_X H \Sigma_X V^T \text{ and } Y \simeq U_Y H \Sigma_Y V^T, \tag{3.9}$$

where $U_X \in \mathbb{R}^{m_X \times k}$, $U_Y \in \mathbb{R}^{m_Y \times k}$, and $V \in \mathbb{R}^{n \times k}$ have orthonormal columns, $H \in \mathbb{R}^{k \times k}$ is a nonsingular matrix that leads to a certain invariance condition for uniqueness of the solution, i.e., $\left(U_X H\right)^T\left(U_X H\right) = \left(U_Y H\right)^T\left(U_Y H\right)$, and $\Sigma_X \in \mathbb{R}^{k \times k}$ and $\Sigma_Y \in \mathbb{R}^{k \times k}$ are diagonal matrices. As shown in Eq. (3.9), PARAFAC2 decomposes X and Y similarly to SVD with an additional constraint that their right singular vectors have to be the same. Eq. (3.9) can be rewritten as

$$V^T \simeq \Sigma_X^{-1} H^{-1} U_X^T X \simeq \Sigma_Y^{-1} H^{-1} U_Y^T Y,$$

which gives linear transformations that produce a common representation, V^T, of both X and Y.

For space alignment, we can replace X and Y with $A_r = \{a_1, \ldots, a_r\} \in \mathbb{R}^{m_A \times r}$ and $B_r = \{b_1, \ldots, b_r\} \in \mathbb{R}^{m_B \times r}$, respectively, in Eq. (3.9) to obtain their PARAFAC2 decomposition. The resulting transformation can then be applied to the entire data as

$$A^f = \Sigma_X^{-1} H^{-1} U_X^T A,\; B^f = \Sigma_Y^{-1} H^{-1} U_Y^T B.$$

The difference of the PARAFAC2 approach from Procrustes analysis is that its linear transformation is no longer orthogonal, which allows deformation of each space to some extent. At the same time, at least for reference correspondence pairs, PARAFAC2 achieves a perfect alignment by exactly co-locating each pair of points in the aligned space as V^T. As we will show in Section 4.4, however, the perfect alignment between the reference correspondence pairs does not always lead to a good alignment for the rest of the data, which is analogous to the notion of overfitting in the context of classification.

4 Experiments

In this section, we present the experimental study on various methods such as metric and nonmetric MDS, constrained Laplacian eigenmaps, Procrustes analysis, and PARAFAC2, for space alignment and information retrieval using real-world data sets. We excluded several traditional methods such as canonical correlation analysis, in our experimental results since they showed significantly poor performances.

4.1 Data sets To facilitate the quantitative evaluation of the alignment process, we use parallel data sets with different data types or languages. We test two cases of parallel data sets: English-Spanish document data sets and document-speech data sets. For English-Spanish document data sets, we first selected 1,000 Associated Press (AP) newswire articles published in 1989 from TIPSTER Volume 1 corpus [16]. Then, we ran a machine translator, SYSTRAN,[3] to convert them from English to Spanish, and each of the English and Spanish document sets is encoded using a bag-of-words model. In this way, we obtain two term-document matrices in separate spaces while maintaining one-to-one correspondences between data item pairs. Then, each matrix is preprocessed by a text analysis tool, IN-SPIRE [25],[4] which involves tf-idf (term frequency–inverse document frequency) weighting, normalization, and feature selection, resulting in 200-dimensional vector representations of 1,000 data items in each language.

Similarly, for document-speech data sets, we used the same 1,000 English documents as in the previous case. To obtain their parallel speech data, we converted each document to a sequence of phonemes through a text-to-speech tool, Festival.[5] Because a phoneme is a much smaller unit than a word, we extracted an n-gram-like phoneme features from the sequence, e.g., a trigram in our case. By treating this phoneme-level n-gram just like a word, we then followed the same procedure of the bag-of-words encoding and the same preprocessing as in the document data.

Let us point out an important difference between the two above-described cases. In order to see how much discrepancy exists between the parallel data sets in terms of their configurations, we compared their Kendall tau distances. That is, we first computed all the pairwise distances among 1,000 data points within each set and converted them to rank orders. The Kendal tau distance counts the number of pairwise disagreements between these rank orders. As a result, English-Spanish document sets were shown to have less disagreements than document-speech data sets, say, 15% vs. 20% out of the total pairs, probably because of the different features used, i.e., words vs. phonemes. It implies that the alignment of the latter case is more challenging than that of the former due to significantly different configurations of the data sets to be aligned, and our analysis in Section 4.4 will be described partly based on this observation.

4.2 Evaluation Measures The previous studies have evaluated their methods by using existing measures such as precision/recall or in a qualitative way, but none of them have analyzed deformation and alignment performances at the same time. Given the aligned vector representations, A^f and B^f, from a pair of parallel data sets, A and B, respectively, we define the nearest-neighbor-based measures for deformation and alignment:

Deformation To assess the amount of deformation, we focus on how well the original neighborhood relationships underlying in each data set are preserved after alignment. Let us define $N_k(x,\, X)$ the k-th nearest neighbor of x in a set X. Similarly, we define $N_{\le k}(x,\, X)$ the set of the k nearest neighbors of x in a set X. Given K and K^f, our deformation measure is defined as

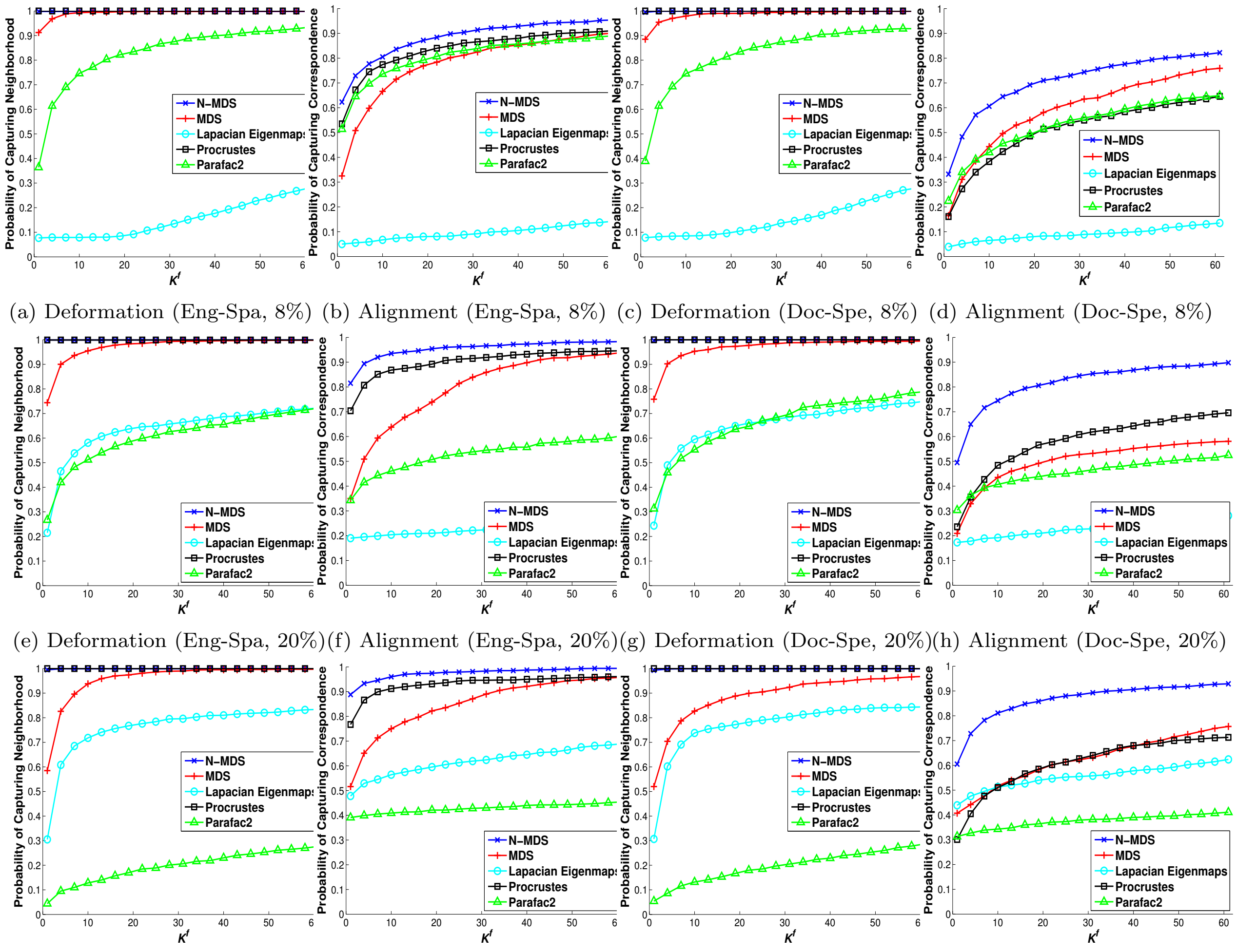
$$\text{(4.10)}\quad \frac{1}{K\,(n_A+n_B)} \left(\sum_{i=1}^{n_A}\sum_{k=1}^{K} \mathbf{1}\left(N_k(a_i,A) \in N_{\le K^f}\left(a_i^f, A^f\right)\right) + \sum_{j=1}^{n_B}\sum_{k=1}^{K} \mathbf{1}\left(N_k(b_j,B) \in N_{\le K^f}\left(b_j^f, B^f\right)\right)\right),$$

where $\mathbf{1}(\cdot)$ is 1 if $\cdot$ is true, and 0 otherwise. It basically computes the probability of capturing the original K nearest neighbors of a point within its K^f nearest neighbors after alignment. A higher value indicates a

[3] http://www.systransoft.com

[4] http://in-spire.pnl.gov

[5] http://www.cstr.ed.ac.uk/projects/festival



(a) Deformation (Eng-Spa, 8%) (b) Alignment (Eng-Spa, 8%) (c) Deformation (Doc-Spe, 8%) (d) Alignment (Doc-Spe, 8%)

(e) Deformation (Eng-Spa, 20%) (f) Alignment (Eng-Spa, 20%) (g) Deformation (Doc-Spe, 20%) (h) Alignment (Doc-Spe, 20%)

(i) Deformation (Eng-Spa, 40%) (j) Alignment (Eng-Spa, 40%) (k) Deformation (Doc-Spe, 40%) (l) Alignment (Doc-Spe, 40%)

Figure 3: Alignment and deformation results of English-Spanish document (Eng-Spa) and document-speech (Doc-Spe) data sets. The numbers in a percent sign represent the ratio of the number of used reference correspondence pairs to the total 1,000 pairs. K is set to 1 in Eq. (4.10).

lesser deformation.

Alignment To measure the alignment, we check how close the corresponding data pairs are in the aligned space. Let us assume A and B are parallel, i.e., a_i and b_i are the correspondence pair for all i's. Given K^f, our alignment measure is defined as

$$(4.11)\frac{1}{n_A + n_B}\left(\sum_{i=1}^{n_A}\mathbf{1}\left(b_i \in N_{\leq K^f}\left(a_i^f, A^f \cup B^f\right)\right) + \sum_{j=1}^{n_B}\mathbf{1}\left(a_i \in N_{\leq K^f}\left(b_i^f, A^f \cup B^f\right)\right)\right),$$

which indicates the probability of capturing the parallel data item of a point within its K^f nearest neighbors in the aligned space.

4.3 Experimental setup For metric and nonmetric MDS, we use a built-in function, 'mdscale', in MATLAB. In Eqs. (2.3) and (2.5), we set $w_A(i, j)$'s and $w_B(i, j)$'s to 1 and $w_R(i)$'s to 1,000 so that the weight between each reference correspondence pair can be comparable to its other edge weights.[6] To be specific, each reference correspondence data item has 999 edges going to the other data items within its data set while having only one edges connecting to its counterpart in the other data set.

Since metric and nonmetric MDS are non-convex problems, the results may depend on initial configurations. We use the solution of Procrustes analysis as the initial configuration. Even so, it would not be an un-

[6] The results were not sensitive to a wide range of values of $w_R(i)$'s, e.g., $[10^2, 10^8]$.

fair comparison between Procrustes analysis and MDS in that MDS significantly changes the initial solution, as will be seen from their performance results.

For constrained Laplacian eigenmaps, we modified the code found in the MATLAB toolbox for dimensionality reduction.[7] We set σ^2 to 0.5 and μ to 50, which is about $1,000$ times larger than the average value of $w_A(i, j)$'s and $w_B(i, j)$'s in Eq. (3.6).

For PARAFAC2, we obtained the code from the author of [5].

In all cases, we set the target dimension of an aligned space to the same as the original dimension of the data, i.e., 200, in order to ignore the information loss due to a smaller dimension in the aligned space than that in the original space.

4.4 Quantitative Comparisons Figure 3 shows the results of deformation and alignment measures in Eqs. (4.10) and (4.11) for two cases of parallel data sets with 1,000 pairs described in Section 4.1 when varying the number of reference correspondence pairs, e.g., 80, 200, and 400 out of the total 1,000 pairs. For deformation measures, we fixed K to 1,[8] and the horizontal axis in each graph represents various values of K^f in Eqs. (4.10) and (4.11).

Table 1 shows the results of alignment measures computed only using the reference correspondence pairs, which shows how well the algorithm aligns the given correspondences. On the other hand, the alignment results in Figure 3 show the generalization ability of aligning unknown correspondences.

In the following, we discuss the behavior of each method based on these results.

Constrained Laplacian eigenmaps Constrained Laplacian eigenmaps shows relatively poor performances both in deformation and alignment in most cases (Figures 3(a)-(l)). Even if our deformation measure only looks at locality information, which is supposedly emphasized in constrained Laplacian eigenmaps, its deformation performance is significantly low particularly when a small amount of reference correspondence pairs, e.g., 8%, are used (Figures 3(a)(c)).

The results of alignment are also far from being satisfactory. In addition, the alignment performance curve of constrained Laplacian eigenmaps has the smallest increase rate (Figures 3(b)(d)(f)(h)(j)(l)) while it achieves relatively good performances on the reference correspondence pairs already at $K^f = 1$ (Table 1). It indicates that constrained Laplacian eigenmaps lacks the generalization ability to align and reveal unknown correspondences.

Table 1: Alignment results only for reference correspondence pairs in English-Spanish document (Eng-Spa) and Document-Speech (Doc-Spe) data sets. The numbers in a percent sign represent the ratio of the number of used reference correspondence pairs to the total 1,000 pairs. K^f is set to 1 in Eq. (4.11).

	Eng-Spa			Doc-Spe		
	8%	20%	40%	8%	20%	40%
N-MDS	1	.97	.96	.98	.91	.88
MDS	1	1	1	1	1	1
Lapl. eig.	.63	.97	1	.51	.88	1
Procrustes	.98	.93	.88	.60	.52	.45
Parafac2	1	1	.96	1	1	.78

We have actually tried various parameter settings in order to improve its performance, but it was very hard to find such settings. Although not presented in this paper, one of the few cases it shows reasonable performances is when the total number of data is small, e.g., less than 100. As discussed in Section 3.1, we suspect that it is due to the fact that the original distance information is only reflected in the weights $w_A(i, j)$'s and $w_B(i, j)$'s in Eq. (3.6).

Procrustes analysis In terms of deformation, Procrustes analysis always maintains no deformation at all, i.e., keeping its original nearest neighbor as it is (Figures 3(a)(c)(e)(g)(i)(k)) due to the rigid transformations and isotropic scaling, as discussed in Section 3.2.

However, its alignment results are not the best in general. In English-Spanish document sets, it works relatively well (Figures 3(b)(f)(j)), but its performance is significantly degraded in document-speech data sets (Figures 3(d)(h)(l)). In addition, as shown in Table 1, as the number of reference correspondence pairs increases, the alignment performance of Procrustes analysis on them becomes worse. These observations reveal the limitation of Procrustes analysis in handling the substantial discrepancy underlying in the data sets to be aligned, as discussed in Section 4.1.

PARAFAC2 Unlike the previous linear method, Procrustes analysis, a non-orthogonal transformation of PARAFAC2 is shown to deform the spaces significantly (Figures 3(a)(c)(e)(g)(i)(k)) in return for the nearly perfect alignment of given correspondence pairs (Table 1). Even with such a good alignment of given correspondences, however, it does not properly align the entire

[7] http://homepage.tudelft.nl/19j49/MatlabToolboxforDimensionalityReduction.html

[8] Throughout the range of K from 1 to 10, the results were identical when it comes to comparisons between different methods.

data (Figures 3(b)(d)(f)(h)(j)(l)).

In addition, it is interesting that both performances in deformation and alignment of PARAFAC2 turn out to get worse as the number of given correspondence pairs increases. It can be explained as a kind of *overfitting* in the context of space alignment. To achieve the nearly perfect alignment of a larger number of given correspondence pairs, PARAFAC2 has to deform the spaces more significantly (Figures 3(i)(k)), and its generalization ability to align the unknown correspondences decreases, as shown in the graphs with the minimal increasing rates in 3(j)(l).

Metric and nonmetric MDS In all cases, nonmetric MDS shows the best performance in both deformation and alignment. The deformation amount of nonmetric MDS is very little or none, competing with zero deformation of Procrustes analysis (Figures 3(a)(c)(e)(g)(i)(k)). At the same time, its alignment is by far better than any other methods, particularly in document-speech data sets (Figures 3(d)(h)(l)). These results clearly demonstrate the suitability of nonmetric MDS in aligning the data sets that are hard to align due to their significant discrepancies, as discussed in Section 2.1.

On the contrary, although started with the same initial configurations as nonmetric MDS, the overall performance of metric MDS is not impressive. Its deformation results are much better than constrained Laplacian eigenmaps and PARAFAC2 but definitely inferior to nonmetric MDS and Procrustes analysis (Figures 3(a)(c)(e)(g)(i)(k)). Its alignment performances are worse than those of Procrustes analysis in most cases except for Figure 3(d), and they are not comparable with those of nonmetric MDS. It indicates that metric MDS suffers from trying to exactly fit the original distances within each space as well as the zero-distances between the reference correspondence pairs.

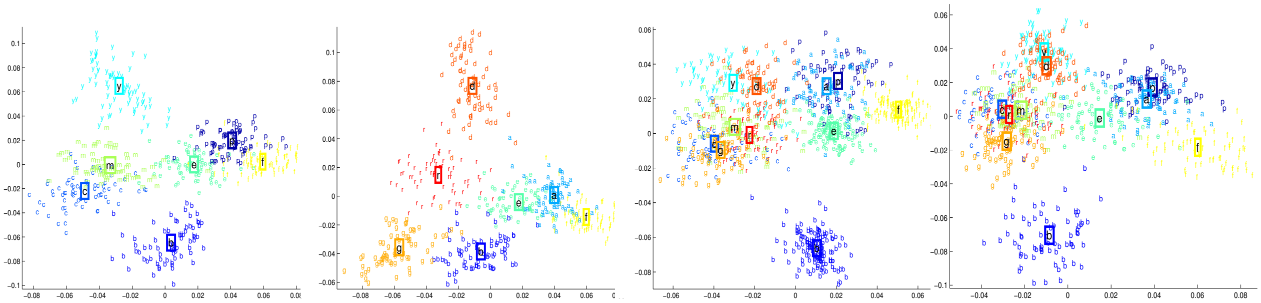
4.5 Cross-domain Information Retrieval In order to verify that our approach works well in practical scenarios, we have also conducted cross-domain information retrieval tests using one of the standard measures, mean average precision (MAP).

The data we used in this work have topic labels [24], and we have selected two of the most frequently shown classes, ‘israel’ and ’bush,’ to which about 2% and 1.4% of the 1,000 document data belong, respectively. For each class, by using each data item in this class from one data set as a query, we have retrieved the data from the other data set based on their Euclidean distances to the query item in the aligned space. Then, the MAP values for the compared methods in both English-Spanish document sets and document-speech data sets are computed. In this process, we have excluded the counterpart of the query item in order to compare the MAP values in the cross-domain setting with those in the single-domain setting in which the information retrieval is done within a single data set. Besides, how close each data item is to its counterpart has already been discussed in Section 4.4 using our own alignment measure.

Table 2: Mean average precision (MAP) results in the percentage for single-domain and cross-domain information retrieval using class label information. For English-Spanish document (Eng-Spa) and Document-Speech (Doc-Spe) data sets, each data item with a particular class label, e.g., ‘israel’ or ’bush,’ in one data set is given as a query, and the data items of the other data set are retrieved from the aligned space. The numbers in a percent sign represent the ratio of the number of used reference correspondence pairs to the total 1,000 pairs.

Class‘israel’						
Single-	English		Spanish		Speech	
domain	46.4		45.6		45.6	
Cross-	Eng-Spa			Doc-Spe		
domain	8%	20%	40%	8%	20%	40%
N-MDS	45.7	45.5	45.3	44.2	44.5	45.1
MDS	43.7	40.1	40.1	42.2	37.9	36.0
Lapl. eig.	1.9	2.1	4.6	1.8	2.2	5.9
Procrustes	38.6	37.5	38.6	27.2	27.5	27.0
Parafac2	36.8	17.7	8.7	32.5	15.7	5.8
Class‘bush’						
Single-	English		Spanish		Speech	
domain	19.8		17.4		17.0	
Cross-	Eng-Spa			Doc-Spe		
domain	8%	20%	40%	8%	20%	40%
N-MDS	17.0	17.8	19.1	14.0	16.7	18.6
MDS	13.8	15.2	17.4	8.2	13.2	17.2
Lapl. eig.	1.4	1.3	3.9	1.6	1.4	2.4
Procrustes	15.3	15.3	16.6	11.9	13.6	12.7
Parafac2	10.9	6.8	2.7	14.2	5.3	2.8

The MAP results for single-domain and cross-domain information retrieval are shown in Table 2. One can regard the single-domain MAP results as baseline or ideal values. In general, the cross-domain MAP results are less than the baseline, and a higher cross-domain MAP value indicates a better alignment of spaces. Just like the previous analysis in Section 4.4, nonmetric MDS consistently shows the best MAP results among the tested methods in Table 2. The behaviors of the other methods also conform to the previous analysis.



(a) Data set 1 (b) Data set 2 (c) Aligned data set (d) Integrated data set

Figure 4: Visualization of newsgroups data sets. Each of (a) and (b) contains seven clusters of data while sharing three of them, 'b', 'e', and 'f.' Aligned data sets shown in (c) successfully reveal the original data relationships shown in (d).

4.6 Visual Knowledge Discovery via Space Alignment In this section, we further apply the space alignment technique in knowledge discovery scenarios via visualization. To clearly show its benefit, we use a well-clustered newsgroups data set [1], a collection of newsgroup documents of 20 topic clusters. However, we have chosen 11 topics for visualization, and each cluster is set to have 70 documents. These 11 cluster data are then split into two data sets with three shared clusters, 'b', 'e', and 'f,' as follows:

- Data set 1 only: comp.sys.ibm.pc.hardware ('p'), sci_crypt ('y'), soc.religion.christian ('c'), talk.politics.misc ('m'),
- Data set 2 only: comp.sys.mac.hardware ('a'), sci.med ('d'), talk.religion.misc ('r'), talk.politics.guns ('g'),
- Shared: rec.sport.baseball ('b'), sci.electronics ('e'), misc.forsale ('f'),

where the letters in parentheses are used in the visualization shown in Fig. (4). The two data sets are encoded as term-document matrices, respectively, using their own vocabulary, which are then followed by dimension reduction to 400 via principal component analysis. Note that even though the shared clusters between the two sets contain the same documents, independent preprocessing leads to significant topology differences between the two sets, which makes their alignment nontrivial. Using the data items in three common clusters as reference correspondence pairs, the nonmetric MDS-based alignment has been performed on these two sets.

Figure (4) shows their visualization results due to space alignment. To visualize high-dimensional data, we applied the rank-2 linear discriminant analysis [6] using the given cluster labels. Figures (4)(a) and (b) show the individual visualizations of the two sets while Figure (4)(c) represents the visualization after their alignment, which reveals various interesting relationships across different data sets. The clusters from different data sets, e.g., soc.religion.christian ('c')-talk.religion.misc ('r'), comp.sys.ibm.pc.hardware ('p')-comp.sys.mac.hardware ('a'), and sci.crypt ('y')-sci.med ('d'), are shown to be closely related with each other.

In addition, Figure (4)(d) shows the visualization of the entire data covering all the 11 clusters. One can see that Figure (4)(c) is similar to Figure (4)(d), indicating that the space alignment well recovers the true underlying relationships.

5 Conclusions

In this paper, we have focused on data fusion problems towards a seamless analysis between different data sets. This work is, as far as we know, the first and unique work that presents a classification-like perspective on the space alignment problem, e.g., training/test accuracy and overfitting, as well as qualitative studies in interactive visualization scenarios. Our contributions include the following:

- A graph-embedding framework for fusion and a nonmetric MDS-based method.
- Discussions on existing methods from the perspective of deformation and alignment.
- Quantitative evaluations between various methods using neighborhood-based measures and cross-domain information retrieval tests.
- Visualization case studies for knowledge discovery from multiple data sets.

We plan to develop a visual analytics tool for heterogeneous data in which the presented methodology would play a foundational role. Along this direction, various interesting research topics would arise such as novel interfaces as to how to supervise the alignment process, e.g., how the users set the reference points interactively.

References

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

[2] L. Ballesteros and W. B. Croft. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications*, DEXA '96, pages 791–801, London, UK, 1996. Springer-Verlag.

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

[4] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Verlag, 2005.

[5] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali. Cross-language information retrieval using parafac2. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 143–152, New York, NY, USA, 2007. ACM.

[6] J. Choo, S. Bohn, and H. Park. Two-stage framework for visualization of clustered high dimensional data. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009.*, pages 67 –74, 2009.

[7] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006.

[8] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, London, 2000.

[9] J. De Leeuw. Applications of convex analysis to multidimensional scaling. *Recent developments in statistics*, pages 133–146, 1977.

[10] E. R. Gansner, Y. Koren, and S. North. Graph drawing by stress majorization. In *Graph Drawing*, volume 3383 of *Lecture Notes in Computer Science*, pages 239–250. Springer Berlin / Heidelberg, 2005.

[11] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991.

[12] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*. Oxford University Press, USA, 2004.

[13] J. Ham, D. Lee, and L. Saul. Semisupervised alignment of manifolds. In *Proceddings of the International Workshop on Artifcial Intelligence and Statistics*, pages 120–127, 2005.

[14] R. J. Hanson and M. J. Norris. Analysis of measurements based on the singular value decomposition. *SIAM Journal on Scientific and Statistical Computing*, 2(3):363–373, 1981.

[15] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[16] D. Harman. Overview of the first trec conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 36–47, New York, NY, USA, 1993. ACM.

[17] R. Harshman. PARAFAC2: Mathematical and Technical Notes. *UCLA working papers in phonetics*, page 30, 1972.

[18] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Physics in medicine and biology*, 46:R1–R45, 2001.

[19] J. R. Hurley and R. B. Cattell. The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7(2):258–262, 1962.

[20] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, April 1989.

[21] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[22] J. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964.

[23] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1784–1797, 2006.

[24] D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 298–306, New York, NY, USA, 1996. ACM.

[25] K. A. Pennock and N. E. Miller. System and method for interpreting document contents, U.S. Patent No. 6,772,170, August 2004.

[26] J. Sammon, John W. A nonlinear mapping for data structure analysis. *Computers, IEEE Transactions on*, C-18(5):401 – 409, may. 1969.

[27] X. Shi, Q. Liu, W. Fan, P. Yu, and R. Zhu. Transfer learning on heterogenous feature spaces via spectral transformation. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 1049 –1054, dec. 2010.

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

[29] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[30] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. pages 1120–1127, 2008.

[31] C. Wang and S. Mahadevan. A General Framework for Manifold Alignment. In *AAAI Fall Symposium on Manifold Learning and its Applications*, 2009.

[32] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *IJCAI'09: Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1273–1278, 2009.

[33] K. Weinberger and L. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70:77–90, 2006.

[34] L. Xiong, F. Wang, and C. Zhang. Semi-definite manifold alignment. *Machine Learning: ECML 2007*, 4701:773–781, 2007.