Aligning Mixed Manifolds

Thomas Boucher[†], **CJ Carey** and **Sridhar Mahadevan**
School of Computer Science
University of Massachusetts
Amherst, MA 01003
[†]boucher@cs.umass.edu

M. Darby Dyar
Department of Astronomy
Mount Holyoke College
South Hadley, MA 01075

Abstract

Current manifold alignment methods can effectively align data sets that are drawn from a non-intersecting set of manifolds. However, as data sets become increasingly high-dimensional and complex, this assumption may not hold. This paper proposes a novel manifold alignment algorithm, low rank alignment (LRA), that uses a low rank representation (instead of a nearest neighbor graph construction) to embed and align data sets drawn from mixtures of manifolds. LRA does not require the tuning of a sensitive nearest neighbor hyperparameter or prior knowledge of the number of manifolds, both of which are common drawbacks with existing techniques. We demonstrate the effectiveness of our algorithm in two real-world applications: a transfer learning task in spectroscopy and a canonical information retrieval task.

Introduction

As machine learning practitioners tackle increasingly large and complex data sets, better data representations are necessary to improve task performance while reducing computational burden. Manifold learning techniques, which aim to uncover the underlying geometry of a data set, have proven to be especially useful in combating the *curse of dimensionality* (Ma and Fu 2011).

One task that has especially benefited from manifold-based approaches is data set alignment, a semi-supervised task in which correspondences are learned between multiple data sets based on intra-set geometry and a provided partial set of pairwise correspondences between the data sets. Manifold alignment is a class of techniques that solves the alignment problem when these data sets are assumed to share a common underlying structure, by embedding each input set into a shared latent manifold space (Ham, Lee, and Saul 2005).

Manifold alignment was introduced as a semi-supervised, nonlinear extension of canonical correlation analysis (CCA) (Hotelling 1936) that aimed to preserve both local geometry and inter-set correspondences (Wang and Mahadevan 2009). One drawback of existing manifold alignment approaches is their assumption that all data sets are drawn from one or more non-overlapping manifolds. This assumption stems from the use of distance-weighted local neighborhoods for embedding construction, a technique that fails when the data are drawn from an intersecting mixture of manifolds. As data sets begin to use representations that employ thousands or millions of features (Ma et al. 2009), the assumption of non-mixing manifolds becomes increasingly tenuous.

This mixture of manifolds problem is demonstrated in Figure 1: when input manifolds are poorly separated, local neighborhood information is insufficient for recovering the true structure at manifold junctions. Thus, the use of traditional nearest neighbor graph construction algorithms induce incorrect connections at these intersection points, distorting the manifold representation.

This shortcoming of conventional manifold methods has given rise to a number of unsupervised clustering algorithms which attempt to segment input data by identifying the individual manifold components of a mixed manifold structure (Souvenir and Pless 2005; Wang et al. 2010). One such algorithm is low rank embedding (LRE) (Liu, Hao, and Su 2011), which notably avoids the construction of a nearest neighbor graph.

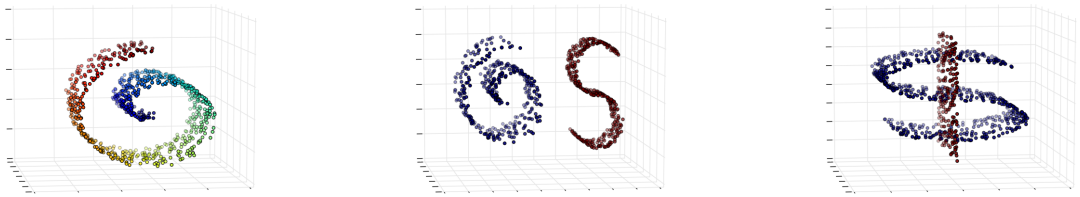
In this paper, we present a novel manifold learning algorithm, low rank alignment (LRA), which builds on the ideas of manifold alignment and LRE to align data sets drawn from mixtures of manifolds. LRA does not suffer from the sensitive nearest neighbor hyperparameter present in traditional manifold alignment, nor does it require prior knowledge of the number of manifolds, a common input to many manifold clustering techniques.

Manifold Alignment

The general manifold alignment framework for two data sets (Ma and Fu 2011; Wang and Mahadevan 2009) is the following. We are given the data sets X and Y of shapes $N_X \times D_X$ and $N_Y \times D_Y$, where each row is a sample (or instance) and each column is a feature, and a correspondence matrix $C^{(X,Y)}$ of shape $N_X \times N_Y$, where

$$C_{i,j}^{(X,Y)} = \begin{cases} 1 & : X_i \text{ is in correspondence with } Y_j \\ 0 & : \text{otherwise} \end{cases} . \quad (1)$$

Manifold alignment calculates the embedded matrices $F^{(X)}$ and $F^{(Y)}$ of shapes $N_X \times d$ and $N_Y \times d$ for $d \leq$



(a) Single "Swiss roll" manifold.

(b) Non-overlapping manifolds.

(c) Mixture of manifolds.

Figure 1: Types of manifold data.

$min(D_X, D_Y)$ that are the embedded representation of X and Y in a shared, low-dimensional space. These embeddings aim to preserve both the intrinsic geometry within each data set and the sample correspondences among the data sets. More specifically, the embeddings minimize the loss function $\mathcal{V}$,

$$\begin{aligned}\mathcal{V}\left(F^{(X)}, F^{(Y)}\right) = &\frac{\mu}{2}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}||F_i^{(X)} - F_j^{(Y)}||_2^2 C_{i,j}^{(X,Y)} \\ &+ \frac{1-\mu}{2}\sum_{i,j=1}^{N_X}||F_i^{(X)} - F_j^{(X)}||_2^2 W_{i,j}^{(X)} \\ &+ \frac{1-\mu}{2}\sum_{i,j=1}^{N_Y}||F_i^{(Y)} - F_j^{(Y)}||_2^2 W_{i,j}^{(Y)},\end{aligned} \tag{2}$$

where N is the total number of samples $N_X + N_Y$, $\mu \in [0,1]$ is the correspondence tuning parameter, and $W^{(X)}, W^{(Y)}$ are the calculated similarity matrices of shapes $N_X \times N_X$ and $N_Y \times N_Y$, such that

$$W_{i,j}^{(X)} = \begin{cases} k(X_i, X_j) & : X_j \text{ is a neighbor of } X_i \\ 0 & : \text{otherwise} \end{cases} \tag{3}$$

for a given kernel function $k(\cdot,\cdot)$. $W_{i,j}^{(Y)}$ is defined in the same fashion. Typically, k is set to be the nearest neighbor set member function or the heat kernel $k(X_i, X_j) = \exp\left(-|X_i - X_j|^2\right)$.

In the loss function of equation (2), the first term corresponds to the alignment error between corresponding samples in different data sets. The second and third terms correspond to the local reconstruction error for the data sets X and Y respectively. This equation can be simplified using block matrices by introducing a joint weight matrix W and a joint embedding matrix F, where

$$W = \begin{bmatrix} (1-\mu)W^{(X)} & \mu C^{(X,Y)} \\ \mu C^{(Y,X)} & (1-\mu)W^{(Y)} \end{bmatrix} \tag{4}$$

and

$$F = \begin{bmatrix} F^{(X)} \\ F^{(Y)} \end{bmatrix}. \tag{5}$$

In Wang and Mahadevan (2009) it is shown that the loss function $\mathcal{V}$ can be reduced to a matrix trace formulation,

$$\underset{F:F^\top DF=I}{\arg\min}\ \mathcal{V}(F) = \underset{F:F^\top DF=I}{\arg\min}\ tr(F^\top LF). \tag{6}$$

where $tr(\cdot)$ is the matrix trace and L is the combinatorial graph Laplacian $L = D - W$, where D is the diagonal matrix of row sums $D(i,i) = \sum_j W(i,j)$. See Chung (1996) for a comprehensive introduction to the graph Laplacian and its variants.

The constraint $F^\top DF = I$ ensures that the problem is well posed and removes arbitrary scaling factors in the embedding. Wang and Mahadevan (2009) showed that the d columns of the embedding matrix F in equation (6) are equal to the d *smallest* eigenvectors, the eigenvectors associated with the smallest non-zero eigenvalues, of the Laplacian matrix in the generalized eigenvalue problem $LF = \lambda DF$.

Low Rank Embedding

Low rank embedding (LRE) is a variation on locally linear embedding (LLE) (Roweis and Saul 2000) that uses low rank matrix approximations instead of LLE's nearest neighbor approach to calculate a reconstruction coefficients matrix (Liu, Hao, and Su 2011). LRE is a two part algorithm. Given a data set X, LRE begins by calculating the reconstruction coefficients matrix R by minimizing the loss function,

$$\min_R \frac{1}{2}||X - XR||_F^2 + \lambda||R||_*, \tag{7}$$

where $\lambda > 0$, $||X||_F = \sqrt{\sum_i \sum_j |x_{i.j}|^2}$ is the Frobenius norm, and $||X||_* = \sum_i \sigma_i(X)$ is the spectral norm, for singular values σ_i. In Candès and Tao (2010) it was shown that the spectral norm is a convex relaxation of the rank minimization problem, and so the solution XR is a low rank representation of the original data matrix X. To solve equation (7), the alternating direction method of multipliers (ADMM) (Boyd et al. 2011) is used.

To apply ADMM a new variable Z is introduced and equation (7) becomes

$$\min_{Z,R} \frac{1}{2}||X - XR||_F^2 + \lambda||Z||_*, \text{ s.t. } R = Z. \tag{8}$$

To solve the constrained optimization problem of equation (8), the augmented Lagrangian function $\hat{\mathcal{L}}$ is introduced,

$$\begin{aligned}\hat{\mathcal{L}}(Z,R,G) = &\frac{1}{2}||X - XR||_F^2 + \lambda||Z||_* \\ &+ \langle G, R - Z\rangle + \frac{\beta}{2}||R - Z||_F^2,\end{aligned} \tag{9}$$

where G is the Lagrange multiplier and β is the penalty parameter that controls the convergence of the ADMM algorithm.

The second step of LRE preserves the point-wise linear reconstruction by holding R fixed while minimizing the reconstruction loss in the embedded space,

$$\min_{F^{(X)}} \frac{1}{2}||F^{(X)} - F^{(X)}R||_F^2 \; s.t. \; (F^{(X)})^\top F^{(X)} = I, \quad (10)$$

where $F^{(X)}$ is the embedding of X and I is the identity matrix. The constraint $(F^{(X)})^\top F^{(X)} = I$ ensures that it is a well-posed problem. In Saul and Roweis (2003) it was shown that equation (10) can be minimized by calculating the d *smallest* non-zero eigenvectors of the Gram matrix $(I - R)^\top (I - R)$.

Low Rank Alignment

Low rank alignment (LRA) is a novel algorithm for the manifold alignment task that uses a variant of LRE to embed the data sets to a joint manifold space, unlike previous alignment methods that have been based on Laplacian eigenmaps (Belkin and Niyogi 2001; Wang and Mahadevan 2008; 2009) and Isomap (Tenenbaum, De Silva, and Langford 2000; Wang and Mahadevan 2013). These methods rely on nearest neighbor graph construction algorithms, and are thus prone to creating spurious inter-manifold connections when mixtures of manifolds are present. These so-called *short-circuit connections* are most commonly found at junction points between manifolds. In contrast, LRA is able to avoid this problem, successfully aligning data sets drawn from a mixture of manifolds. Figure 2 shows an example of this phenomena using a noisy dollar sign data set.

LRA differs from other manifold alignment algorithms in several key aspects. While some previous algorithms embed data using exclusively the eigenvectors of the graph Laplacian to preserve both inter-set correspondences and intra-set local geometry, LRA uses the eigenvectors of the sum of the Laplacian and the Gram matrix of low rank representations to preserve the inter-set correspondences and the intra-set *local linearity*. Moreover, previous manifold alignment algorithms require a reliable measure of similarity between nearest neighbor samples, whereas LRA relies on the linear weights used in sample reconstruction. Lastly, because LRA uses the global property of rank to calculate its reconstruction matrix, it can better discern the global structure of mixing manifolds (Liu, Hao, and Su 2011).

We now describe the low rank alignment algorithm for two data sets. It begins with the same setup as manifold alignment: two data sets X and Y are given, along with the correspondence matrix $C^{(X,Y)}$ describing inter-set correspondences (see equation 1). The goal of LRA is to calculate a set of embeddings $F^{(X)}$ and $F^{(Y)}$ to a joint, low-dimensional manifold subspace that best preserves both inter-set correspondences and intra-set geometries.

In the first step of LRA, the reconstruction weight matrices $R^{(X)} \in \mathbb{R}^{N_X \times N_X}$ and $R^{(Y)} \in \mathbb{R}^{N_Y \times N_Y}$ are calculated individually according to equation (7). In this step, the low rank constraint defines a barycentric coordinate for each sample that preserves locally linear relationships between samples. In Favaro, Vidal, and Ravichandran (2011) it is shown that the low rank representation problem in equation (7) can be solved in closed form. This avoids the iterative ADMM calculation found in the original LRE algorithm.

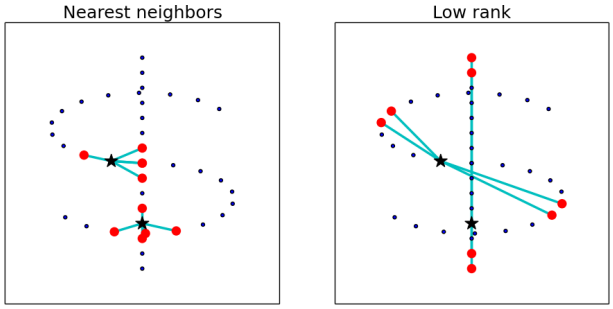


Figure 2: Manifold construction on synthetic data. For each star, the large points are the four neighbors used to define the manifold. The lines match the stars with their neighbors. Notice that traditional nearest neighbor construction (on the left) has *short-circuits* incorrectly connecting the two manifolds, whereas the low rank construction (on the right) selects points that correctly differentiate the mixed manifolds.

We begin by decomposing X using singular value decomposition (SVD), $X = USV^\top$. Next, the columns of V and S are partitioned into $V = [V_1 V_2]$ and $S = [S_1 S_2]$ according to the sets

$$I_1 = \{i : s_i > 1 \;\; \forall s_i \in S\} \text{ and } I_2 = \{i : s_i \leq 1 \;\; \forall s_i \in S\}.$$

Then an optimal closed-form solution to equation (7) is

$$R^{(X)} = V_1 (I - S_1^{-2}) V_1^\top. \quad (11)$$

$R^{(X)}, R^{(Y)}$ are calculated independently and so may be computed in parallel to reduce compute time. We next define the block matrices $R, C \in \mathbb{R}^{N \times N}$ as

$$R = \begin{bmatrix} R^{(X)} & 0 \\ 0 & R^{(Y)} \end{bmatrix} \text{ and } C = \begin{bmatrix} 0 & C^{(X,Y)} \\ C^{(Y,X)} & 0 \end{bmatrix} \quad (12)$$

and $F \in \mathbb{R}^{N \times d}$ as

$$F = \begin{bmatrix} F^{(X)} \\ F^{(Y)} \end{bmatrix}. \quad (13)$$

The second step of LRA is to calculate the embedding F of X, Y by minimizing the loss function,

$$\mathcal{Z}(F) = (1 - \mu) \|F - RF\|_F^2 + \mu \sum_{i,j=1}^{N} ||F_i - F_j||^2 C_{i,j}, \quad (14)$$

where $\mu \in [0, 1]$ is the hyperparameter that controls the importance of inter-set correspondences. The first term of the sum in equation (14) accounts for the local geometry within each data set, and the second term accounts for the correspondences between sets. We can reduce this loss function

to a sum of matrix traces:

$$\begin{aligned}\mathcal{Z}(F) &= (1-\mu)tr((F-RF)^\top(F-RF)) \\ &+ \mu\sum_{k=1}^{d}\sum_{i,j=1}^{N}||F_{i,k}-F_{j,k}||_2^2 W_{i,j} \\ &= (1-\mu)tr\left(((I-R)F)^\top(I-R)F\right) \\ &+ 2\mu\sum_{k=1}^{d}F_{\cdot,k}^\top L F_{\cdot,k} \\ &= (1-\mu)tr(F^\top(I-R)^\top(I-R)F) \\ &+ 2\mu\, tr(F^\top LF). \end{aligned} \tag{15}$$

Similarly to LLE and LRE, we introduce the constraint $F^\top F = I$ to ensure that the minimization of the loss function $\mathcal{Z}$ is a well-posed problem. Thus, we have

$$\underset{F:F^\top F=I}{\arg\min}\,\mathcal{Z} = \underset{F:F^\top F=I}{\arg\min}\,(1-\mu)tr(F^\top MF) + 2\mu\, tr(F^\top LF), \tag{16}$$

where $M = (I-R)^\top(I-R)$. To construct a loss function from equation (16), we take the right hand side and introduce the Lagrange multiplier Λ,

$$\begin{aligned}\mathcal{L}(F,\Lambda) &= (1-\mu)tr(F^\top MF) + 2\mu\, tr(F^\top LF) \\ &+ \langle\Lambda, F^\top F - I\rangle. \end{aligned} \tag{17}$$

To minimize equation (17), we find the roots of its partial derivatives,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial F} &= 2(1-\mu)MF + 4\mu LF - 2\Lambda F = 0 \\ \frac{\partial\mathcal{L}}{\partial\Lambda} &= F^\top F - I = 0. \end{aligned} \tag{18}$$

From this system of equations, we are left with the matrix eigenvalue problem

$$\left((1-\mu)M + 2\mu L\right)F = \Lambda F \quad \text{and} \quad F^\top F = I. \tag{19}$$

Therefore, to solve equation (16), we calculate the *d smallest* non-zero eigenvectors of the matrix

$$(1-\mu)M + 2\mu L. \tag{20}$$

This eigenproblem can be solved efficiently because the matrix $M + L$ is guaranteed to be symmetric, positive semidefinite (PSD), and sparse. These properties arise from the construction

$$\begin{aligned}M + L = &\begin{bmatrix} \left(I - R^{(X)}\right)^2 & 0 \\ 0 & \left(I - R^{(Y)}\right)^2 \end{bmatrix} \\ &+ \begin{bmatrix} D^X & -C^{(X,Y)} \\ \left(-C^{(X,Y)}\right)^\top & D^Y \end{bmatrix}, \end{aligned} \tag{21}$$

where by construction $D = \begin{bmatrix} D^X & 0 \\ 0 & D^Y \end{bmatrix}$ is a PSD diagonal matrix and $C^{(X,Y)}$ is a sparse matrix.

The time complexity of LRA is dominated by two operations: the full singular value decomposition necessary for step 1 and the sparse eigenvector decomposition in step 2. The SVD runs in cubic time proportional to the number of samples. This cost may be somewhat mitigated with a parallel approach, as the reconstruction matrix is calculated separately for each data set. The eigenvector decomposition also runs in cubic time proportional to the number of samples.

Algorithm 1: Low Rank Alignment

Input: data matrices X, Y, embedding dimension d, correspondence matrix $C^{(X,Y)}$ and weight μ.
Output: embeddings matrix F.

Step 0: Column normalize X & Y *(optional but recommended if X and Y differ largely in scale).*
Step 1: Compute the reconstruction coefficient matrices $R^{(X)}, R^{(Y)}$:
$USV^\top = \text{SVD}(X)$
$R^{(X)} = V_1(I - S_1^{-2})V_1^\top$
$\hat{U}\hat{S}\hat{V}^\top = \text{SVD}(Y)$
$R^{(Y)} = \hat{V}_1(I - \hat{S}_1^{-2})\hat{V}_1^\top$
Step 2: Set F equal to the d *smallest* eigenvectors of the matrix in equation (20).

Experimental Results

To evaluate the effectiveness of LRA, experiments were performed on two very different real-world data sets. For comparison, we implemented three state of the art alignment techniques: (instance-level/non-linear) manifold alignment (Wang and Mahadevan 2009), affine matching alignment (Lafon, Keller, and Coifman 2006), and Procrustes alignment (Wang and Mahadevan 2008). All of the methods evaluated align data sets by embedding the sets into a shared low-dimensional space. Affine matching and Procrustes alignment can only align two data sets at a time, and while LRA and manifold alignment do not suffer this limitation, we chose to limit our experimentation to alignment problems involving pairs of data sets.

All experiments were implemented in Python by the authors, with help from the machine learning library Scikit-learn (Pedregosa et al. 2011). An implementation of LRA is available for download on the author's website. [1]

Calibration Transfer

The first data set was a suite of laser-induced breakdown spectra (LIBS) acquired from 100 different geological samples under Mars-like atmospheric conditions at Mount Holyoke College. LIBS instruments are spectrometers composed of a high energy laser that pulses a sample to create plasma, which is observed by a charge-coupled device (CCD) that records the energy emitted. This data set was created in support of the Mars Science Laboratory mission

[1] https://github.com/all-umass/low_rank_alignment

for the ChemCam instrument, a LIBS spectrometer on the rover *Curiosity*.

The task of this experiment is calibration transfer (CT). CT is a transfer learning problem well-studied in chemometrics (Feudalea et al. 2002; Zheng et al. 2014; Ridder, Ver Steeg, and Price 2014), but largely ignored by the machine learning community. The general setup of the problem is the following. The spectra of a set of samples (e.g., rock powders) are recorded on two different instruments or on the same instrument under two varying conditions. The goal is to find a mapping or an alignment between the two sets of spectra. Frequently in all types of spectroscopic studies there is a need to ensure that possible differences in environmental or experimental conditions are mitigated or negated. CT provides an excellent solution to the task of reconciling data in inter- and intra-lab comparisons on Earth and in extraterrestrial applications.

In this experiment, the samples were recorded on the same instrument under a low and high laser power setting, 3% and 5% power respectively. The spectra were first preprocessed according to Wiens et al. (2013). The resultant spectra are 5485-dimensional vectors where each feature corresponds to the response of a particular wavelength channel between 225-925 nm.

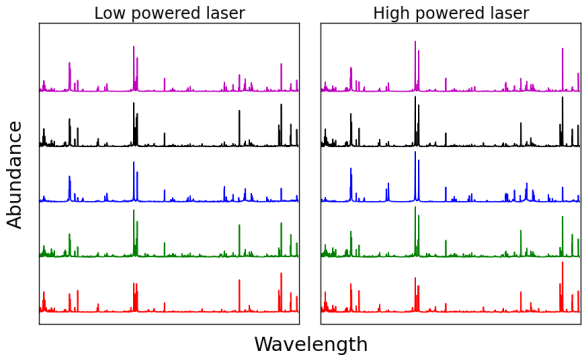


Figure 3: Five mineral spectra selected at random from the LIBS data set. The left hand side shows the spectra recorded with a low power laser, and the right hand side shows the corresponding spectra recorded at a high power.

The task of this experiment was to align the set of low power spectra with the set of high power spectra. A low power spectrum was considered correctly aligned if the corresponding high power spectrum was within its 3-nearest neighbors in the embedded space.

For all models evaluated, the correspondence weight was set to $\mu = 0.8$, based upon the ratio of train/test data. All competing models required an additional nearest neighbor hyperparameter. This hyperparameter was optimized using grid search and cross validation. For affine matching and Procrustes alignment the number of neighbors used was $k = 10$, and for traditional manifold alignment $k = 4$. For all of these competing methods, a binary weight was used in the graph construction because it proved more accurate than the heat kernel for this experiment.

The 5-fold cross validation results are shown in Figure 4. In each iteration, correspondences are provided for 80 spectra while the other 20 spectra are used for evaluation. The experiment was repeated 30 times with a random partitioning of folds. LRA outperformed all other models tested achieving an accuracy of 46.6% at $d = 8$, while the next best performing model, affine matching, had an accuracy of 42% at $d = 7$.

The error bars in this experiment (Figure 4) are larger than those in the following language experiments (Figure 5) due to the relatively small size of the 100 spectra data set.

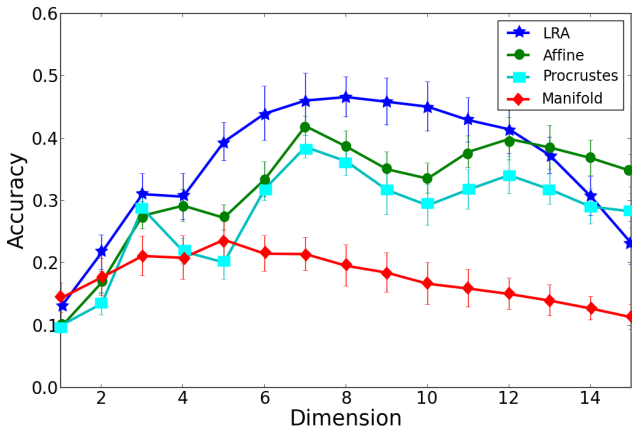


Figure 4: Cross validation results of 100 sample spectra alignment with 1-σ error bars.

In this last test, it was assumed that all of the spectra were recorded at both power settings, but in reality CT is often used when only a portion of the sample set is recorded under both conditions. For example, a researcher may have a large database recorded at high power that he or she uses to fit a regression model for predicting the chemical compositions (% weight) of the spectra. As commonly occurs, the researcher also has a smaller *calibration set* recorded at both high and low powers. Unfortunately, an unforeseen instrument malfunction occurrs allowing the spectrometer to only use low power. To predict subsequent low power spectra using the high power database, an alignment must be calculated.

To simulate this situation, we calculated an alignment using 30 samples at both powers (the calibration set), 50 samples at only high power (the large database), and 20 samples at only low power. The 20 low power samples represent the *out-of-sample* spectra recorded after instrument malfunction. To note, this results in a non-square correspondence matrix $C^{(X,Y)}$.

Next, a multivariate linear regression model was trained to predict 10 major elements of the minerals (e.g., SiO_2, Al_2O_3, CaO) using the embedded high power database and the embedded calibration samples. To evaluate the regression model, the compositions of the 20 embedded low power spectra were predicted and compared to ground-truth composition values.

Setting $d = 8$, the experiment was repeated 30 times with randomized sets. The regression model trained on LRA achieved on average a 1.8%, 4.8%, and 8.1% improvement in RMSEP over affine matching, Procrustes alignment, and

traditional manifold alignment, respectively. This shows that the high accuracy of LRA in alignment translates to improved performance in the final predictive model.

European Parliament Proceedings

In this second set of experiments, we used the transcribed proceedings of the European Parliament (Koehn 2005) for a standard cross-language document retrieval task. The task is simply stated: given a document in one language, find its matching document in the second language. The parliament corpus was collected between April 1996 and November 2011 and transcribed into 21 European languages. In the corpus, each utterance of a speaker was transcribed into paragraphs of typically 2-5 sentences. This data set is commonly used when comparing manifold alignment algorithms (Wang and Mahadevan 2008; 2013).

In the first experiment, we align the German corpus with the English corpus, and in the second experiment we align the Italian corpus with the English corpus. We chose these languages because each had approximately 1.9 million sentence pairs.

To represent the utterances, a bag-of-words model was used, where the 2500 most frequently occurring words were considered, after filtering for stop-words. Unlike methods like Gale and Church (1993) and Resnik and Smith (2003) that used domain knowledge in their model preprocessing, we used a simple statistical model to compare the different alignment methods more directly. To pare down the data set for efficient experimentation, only sentences with more than 45 words were used, resulting in a subset of approximately 2500 sentence pairs for both English/German and English/Italian experiments. For accurate method comparison, we used 5-fold cross validation. In each fold, 80% of the sentence correspondences were provided and the remaining 20% of the sentences were used for evaluation. To evaluate a sentence alignment, we define a correct translation as a sentence embedding where the true correspondence pair appears within the 10-nearest neighbors in the embedded space.

All methods used the same default correspondence weight $\mu = 0.5$. Grid search and cross validation were used to tune the number of nearest neighbors for all competing models. For affine matching and Procrustes alignment $k = 125$, and for manifold alignment $k = 5$.

The results of the text alignment test are shown in Figure 5. LRA outperformed all other evaluated models in the English-German experiment, with an accuracy of 90% at embedding dimension $d = 100$ and 88.4% at $d = 140$. In contrast, the best accuracy achieved by any competing method was 63.1% using affine matching. In the English-Italian test, the accuracy of every method was greater, but LRA continued to outperform all other methods tested.

Traditional manifold alignment was clearly the worst-performing model. Affine matching and Procrustes alignment are both two-step algorithms in the sense that they rely on a second transformation after the embedding step. In contrast, manifold alignment and LRA are one-step algorithms that incorporate these constraints into their embeddings and so could be seen to place more importance on their ability to recover the shared manifold of the data sets. This skewed performance between methods suggests that the corpora are drawn from mixtures of manifolds.

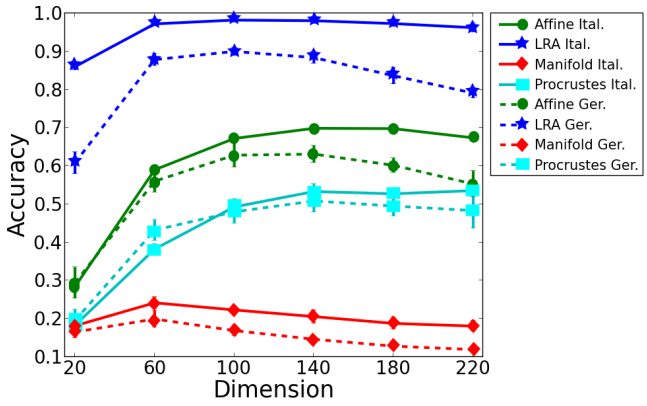


Figure 5: Cross validation results of EU parallel corpus with 2410 Italian-English sentences pairs and 2110 German-English sentences pairs.

Conclusion

This paper presents a novel algorithm for manifold alignment that can align data sets drawn from a mixture of manifolds. Unlike previous manifold alignment algorithms that rely on nearest neighbor graph construction, LRA instead uses a low rank matrix constraint to calculate its reconstruction weight matrix, which has been demonstrated to be less prone to short-circuit connections. We have demonstrated the effectiveness of the algorithm at two diverse real-world tasks: calibration transfer for spectroscopic data and cross-language information retrieval. In both tasks, LRA outperformed all other evaluated methods.

As manifold learning and alignment techniques are applied to more complex tasks, the mixture of manifolds problem has become increasingly apparent. The proposed Low Rank Alignment is the first such alignment method to gracefully handle arbitrary manifold mixtures, a benefit which is reflected in the task performance comparisons.

In future work, we plan to adapt existing manifold alignment methods for mixtures of manifolds and to create novel mixture-friendly regression methods.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under grants CHE-1307179 and CHE-1306133.

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems* 585–591.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3(1):1–122.

Candès, E., and Tao, T. 2010. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on* 56(5):2053–2080.

Chung, F. 1996. *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society.

Favaro, P.; Vidal, R.; and Ravichandran, A. 2011. A closed form solution to robust subspace estimation and clustering. *IEEE Conference on Computer Vision and Pattern Recognition* 1801–1807.

Feudalea, R.; Woodya, N.; Tana, H.; Mylesa, A.; Brown, S.; and Ferreb, J. 2002. Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems* 64:181–192.

Gale, W., and Church, K. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19.

Ham, J.; Lee, D.; and Saul, L. 2005. Semisupervised alignment of manifolds. *10 th International Workshop on Artificial Intelligence and Statistics* 120-127.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28:321–377.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *In MT Summit*.

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

Liu, R.; Hao, R.; and Su, Z. 2011. Mixture of manifolds clustering via low rank embedding. *Journal of Information & Computational Science* 8:725–737.

Ma, Y., and Fu, Y. 2011. *Manifold Learning Theory and Applications*. CRC Press.

Ma, J.; Saul, L.; Savage, S.; and Voelker, G. 2009. Identifying suspicious URLs: An application of large-scale online learning. *Proceedings of the International Conference on Machine Learning* 681–688.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Resnik, P., and Smith, N. 2003. The web as a parallel corpus. *Computational Linguistics* 29:349–380.

Ridder, T.; Ver Steeg, B.; and Price, G. 2014. Robust calibration transfer in noninvasive ethanol measurements, part I: Mathematical basis for spectral distortions in fourier transform near-infrared spectroscopy. *Applied Spectroscopy* 68:852–864.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.

Saul, L., and Roweis, S. 2003. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research* 4:119–155.

Souvenir, R., and Pless, R. 2005. Manifold clustering. In *Tenth IEEE International Conference on Computer Vision*, volume 1, 648–653. IEEE.

Tenenbaum, J.; De Silva, V.; and Langford, J. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323.

Wang, C., and Mahadevan, S. 2008. Manifold alignment using procrustes analysis. *Proceedings of the International Conference on Machine Learning* 1120–1127.

Wang, C., and Mahadevan, S. 2009. A general framework for manifold alignment. *AAAI Fall Symposium on Manifold Learning and its Applications*.

Wang, C., and Mahadevan, S. 2013. Manifold alignment preserving global geometry. *The 23rd International Joint conference on Artificial Intelligence*.

Wang, Y.; Jiang, Y.; Wu, Y.; and Zhou, Z. 2010. Multi-manifold clustering. In *PRICAI 2010: Trends in Artificial Intelligence*. Springer. 280–291.

Wiens, R.; Maurice, S.; Lasue, J.; Forni, O.; Anderson, R.; and et al. 2013. Pre-flight calibration and initial data processing for the chemcam laser-induced breakdown spectroscopy instrument on the mars science laboratory rover. *Spectrochimica Acta Part B* 82:1–27.

Zheng, K.; Zhang, X.; Iqbal, J.; Fan, W.; Wu, T.; Du, Y.; and Liang, Y. 2014. Calibration transfer of near-infrared spectra for extraction of informative components from spectra with canonical correlation analysis. *Journal of Chemometrics* 28:773–784.