An Interpretable Joint Nonnegative Matrix Factorization-Based Point Cloud Distance Measure

Hannah Friedman
Department of Mathematics
Harvey Mudd College
Claremont, CA, USA
hfriedman@hmc.edu

Amani R. Maina-Kilaas
Department of Computer Science
Harvey Mudd College
Claremont, CA, USA
amainakilaas@hmc.edu

Julianna Schalkwyk
Department of Mathematics
Harvey Mudd College
Claremont, CA, USA
jschalkwyk@hmc.edu

Hina Ahmed
Department of Mathematics
Scripps College
Claremont, CA, USA
hahmed@hmc.edu

Jamie Haddock
Department of Mathematics
Harvey Mudd College
Claremont, CA, USA
jhaddock@hmc.edu

2023 57th Annual Conference on Information Sciences and Systems (CISS) | 978-1-6654-5181-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CISS56502.2023.10089765

***Abstract*—In this paper, we propose a new method for determining shared features of and measuring the distance between data sets or point clouds. Our approach uses the joint factorization of two data matrices X_1, X_2 into non-negative matrices $X_1 = AS_1, X_2 = AS_2$ to derive a similarity measure that determines how well the shared basis A approximates X_1, X_2. We also propose a point cloud distance measure built upon this method and the learned factoriI zation. Our method reveals structural differences in both image and text data. Potential applications include classification, detecting plagiarism or other manipulation, data denoising, and transfer learning.**

***Index Terms*—nonnegative matrix factorization, topic modeling, point cloud distance, data set distance**

I. Introduction

Identifying similar or dissimilar features in the underlying structures of point clouds is a useful technique across a wide array of fields. There has been significant effort in developing measures of dataset similarity, but much less in developing measures that can also provide information about *how the datasets are similar or dissimilar.* This paper proposes a dataset similarity measure that also naturally provides information about their primary similarities or dissimilarities.

Point cloud comparison methods and point cloud distances have applications in document clustering [1], [2] and in computer vision such as object classification [3], object detection [4], and semantic segmentation [5]. These applications often turn to classical metrics for dataset or point cloud comparison, such as the Chamfer distance [6], which is defined as the sum of the averages of the minimum distances between points in data matrices X_1 and X_2,

$$d_{\text{cham}}(X_1, X_2) = \frac{1}{|X_1|} \sum_{\mathbf{x} \in X_1} \min_{\mathbf{y} \in X_2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \frac{1}{|X_2|} \sum_{\mathbf{y} \in X_2} \min_{\mathbf{x} \in X_1} \|\mathbf{x} - \mathbf{y}\|_2^2, \tag{1}$$

This work is supported by NSF DMS award #2211318.

where, in a slight abuse of notation, we let $\mathbf{x} \in X$ mean that $\mathbf{x}$ is a column of the data matrix X. Other popular metrics include the Hausdorff distance [7], [8], the earth mover's distance [9], and recent metrics based on optimal transport [10] However, these measures can be computationally expensive or ineffective [11]. Recent efforts for point cloud comparison include density-aware approaches [12], [13]. Additionally, data set or point cloud distances have applications in measuring generalizability of machine learning models [14]–[16] and transfer learning [17].

Nonnegative matrix factorization (NMF) is a useful tool for interpretable dimension reduction. Many types of data, including documents and images, can be represented by nonnegative matrices, making NMF a widely applicable method for data analysis [18]. NMF has been previously used in the creation of data set similarity metrics: Shahnaz et al. [2] clusters semantic features or topics in document data and uses NMF to preserve nonnegativity. Liu et al. [19] introduces a multi-view clustering approach based on NMF.

Joint non-negative matrix factorization (jNMF) allows for joint factorization of two data sets with a common basis [20], [21]. Additionally, jNMF for topic modeling has been used to identify similarities across data sets. Kim et al. [20] proposes a jNMF method for identifying both common and distinct topics among document data sets, although they do not use these topics to measure overall similarity of the datasets.

In this paper, we propose a new method for evaluating similarity between a pair of distinct point clouds or data sets. Our method analyzes the outputs of jNMF and measures the contributions of the basis vectors to each set. In our method, we first run jNMF on the two data matrices and then, motivated by statistical distribution comparison tests, we compare the empirical distribution functions that represent the jNMF coefficient factor matrices. We use this learned information to propose a point cloud distance measure *and* to provide information about how the data are structurally similar

978-1-6654-5181-9/23/$31.00 ©2023 IEEE

and dissimilar via the learned jNMF basis vectors and their measured association with each data set.

In Section II, we give a brief overview of NMF and jNMF. In Section III, we propose a method for determining shared features in data sets, a distance measure based on this method, and we list some desirable properties of a distance measure. In Section IV, we present examples of this method applied to real world data and experimentally verify our desired properties.

II. Overview of NMF and jNMF

Given a nonnegative $m \times n$ matrix X, the goal of nonnegative matrix factorization (NMF) is to find nonnegative matrices A and S such that

$$X \approx AS$$

where A is $m \times k$ and S is $k \times n$ [18]. One typically chooses k so that AS is a low rank approximation of X; there are many heuristics for choosing the model rank k, which are beyond the scope of this paper. NMF produces A and S by attempting to minimize the non-convex objective function,

$$\|X - AS\|_F^2 = \sum_{ij} (X_{ij} - (AS)_{ij})^2.$$

NMF models can be trained with many methods. One of the most popular methods is multiplicative updates, which is a variant of gradient descent that ensures entrywise nonnegativity in the factors [22]. We typically interpret columns of A as a set of "basis" vectors and the ith column of S as the coefficients of the conic combination of those basis vectors that approximates the ith data point. We do not focus on how basis vectors are combined to create individual elements, but instead analyze the rows of S to measure the contribution of each basis vector to the entire data set.

Joint NMF (jNMF) finds low-rank, non-negative approximations for two matrices, X and Y, that share a common factor matrix [20]. When applied to supervised NMF, jNMF typically factors X (the data) and Y (e.g., class labels) as $X \approx A_1 S$ and $Y \approx A_2 S$, so that S is shared between the factorizations. To control the emphasis put on the labels, a weighting factor λ can be introduced into the objective function,

$$\|X - A_1 S\|_F^2 + \lambda \|Y - A_2 S\|_F^2,$$

but we focus on the cases where the approximation terms are weighted equally, $\lambda = 1$. When $\lambda = 1$, the factorization can easily be learned by performing NMF on the matrix obtained by stacking X on top of Y, resulting in the factorization

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} S.$$

Lee et al. [23] and Haddock et al. [24] apply jNMF to semi-supervised tasks like document classification. Like NMF, this model is nonconvex and methods typically employ per-iteration efficient methods that optimize convex alternative problems for each factor matrix.

We apply jNMF via NMF on the matrix obtained by stacking data matrices X_1 and X_2 side-by-side, denoted by $[X_1\ X_2]$, which we factorize by $[X_1\ X_2] = A[S_1\ S_2]$. This model may be represented as minimizing objective

$$\|X_1 - AS_1\|_F^2 + \|X_2 - AS_2\|_F^2 \tag{2}$$

with respect to the factor matrices A, S_1, S_2. For the method to be well-defined, the data points in the two sets X_1, X_2 must have the same dimension. Geometrically, NMF can be interpreted as learning basis vectors such that the cone of these vectors best approximates a given data set [25]. Thus, jNMF attempts to learn basis vectors (columns of A) so that the cone of these vectors contains good approximations of all data points in X_1 and X_2; see Figure 1 for a visualization of a joint NMF learned for two datasets (X_1 in red and X_2 in blue).

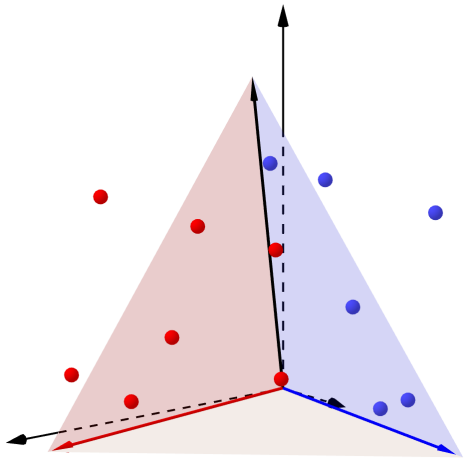


Fig. 1: Visualization of a joint NMF learned for two datasets (X_1 in red and X_2 in blue). Note that the data points in X_1 are well approximated by the basis elements visualized in black and red (their conic span is given in red), while the data points in X_2 are well approximated by the basis elements visualized in black and blue (their conic span is given in blue).

III. Proposed Method and Distance Measure

Our proposed distance measure identifies how well two data sets, $X_1 \in \mathbb{R}_{\geq 0}^{m \times n_1}$ and $X_2 \in \mathbb{R}_{\geq 0}^{m \times n_2}$, can be approximated by the common basis learned through jNMF. The existence of such a set of basis vectors implies a similar underlying structure between the data sets. However, one may obtain a basis set in which some elements primarily contribute to the first data set and some contribute to the second, but very few are shared; see Figure 1 for a visualization of such a scenario. This scenario suggests some structural differences in the data. We analyze the contributions of the basis vectors to the different data sets to identify features in one data set that are not expressed well by a basis for the other data set.

A. Proposed Similarity Method

Given a rank-k jNMF approximation $[X_1\ X_2] \approx A[S_1\ S_2]$, our method produces a length-k vector $\bar{\mathbf{p}}$ where each element represents the ratio of the corresponding basis vector's contribution to each of the data sets. We compute $\bar{\mathbf{p}}_i$ from the

ith rows of S_1 and S_2, because the magnitudes of the entries in row i of S_1 and S_2 indicate how much $A_{:,i}$, the ith basis vector, contributes to each data matrix. The entries of $\bar{\mathbf{p}}$ are between -1 and 1; $\bar{\mathbf{p}}_i$ is positive if $A_{:,i}$ appears more often and in higher intensity in decompositions of the points in X_1 and negative if $A_{:,i}$ appears more often and in higher intensity in decompositions of the points in X_2. Pseudocode for our method is provided in Algorithm 1.

Algorithm 1 jNMF similarity

Require: data matrices $X_1 \in \mathbb{R}^{m\times n_1}_{\geq 0}$ and $X_2 \in \mathbb{R}^{m\times n_2}_{\geq 0}$, number of samples K for averaging, model rank k

1: Scale each column in X_1, X_2 to be mean one.[1]

2: Learn rank-k jNMF approximation via (2),

$$[X_1 \ X_2] \approx A[S_1 \ S_2].$$

3: For $i = 1, \cdots, k$, define

$$s_i = \max\left(\{s_{ij}^{(1)}\}_{j=1}^{n_1} \cup \{s_{ij}^{(2)}\}_{j=1}^{n_2}\right)$$

where $s_{i1}^{(1)}, s_{i2}^{(1)}, \cdots, s_{in_1}^{(1)}$ and $s_{i1}^{(2)}, s_{i2}^{(2)}, \cdots, s_{in_2}^{(2)}$ are the entries of the ith rows of S_1 and S_2, respectively.

4: **for** $j = 1, ..., K$ **do**

5: **for** $i = 1, \cdots, k$ **do**

6: Choose $T_i \sim \text{unif}([0, s_i])$ for $i = 1, 2, \cdots, m$.

7: Compute $\mathbf{p}_i^{(j)} := F_i^{(2)}(T_i) - F_i^{(1)}(T_i)$, where

$$F_i^{(1)}(T_i) := \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{1}[s_{ij}^{(1)} < T_i]$$

and

$$F_i^{(2)}(T_i) := \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{1}[s_{ij}^{(2)} < T_i]$$

are the empirical distribution functions (EDFs) of $\{s_{ij}^{(1)}\}_{j=1}^{n_1}$ and $\{s_{ij}^{(2)}\}_{j=1}^{n_2}$ evaluated at T_i, respectively.

8: **return** $\bar{\mathbf{p}} = \frac{1}{K}\sum_{j=1}^{K} \mathbf{p}^{(j)}$

We note that Step 7 in the Algorithm 1 compares the fraction of sample $\{s_{ij}^{(1)}\}_{j=1}^{n_1}$ below a randomly sampled threshold to the fraction of sample $\{s_{ij}^{(2)}\}_{j=1}^{n_2}$ below the same threshold. That is, we measure the difference between these samples by calculating the difference between their EDFs $F_i^{(1)}(T)$ and $F_i^{(2)}(T)$ [26]; this is akin to the fundamental idea of the Kolmogorov-Smirnov test [27], [28] and Cramer-von Mises criterion [29], [30]. See Figure 2 for an example visualization of the samples $\{s_{ij}^{(1)}\}_{j=1}^{n_1}$ and $\{s_{ij}^{(2)}\}_{j=1}^{n_2}$, and their empirical distribution functions $F_i^{(1)}(T)$ and $F_i^{(2)}(T)$; these histograms and EDFs might correspond to the third (blue) basis vector in Figure 1, as the entries indicate that this basis vector is more heavily used by the blue data set. Note that $\bar{\mathbf{p}}_i$ is simply the difference between these EDFs averaged over random samples taken uniformly from the data interval.

[1]To address the case where $\mathbf{v} \approx \mathbf{0}$, we add a threshold so that $\mathbf{v} \in X_i$ is only normalized if $||\mathbf{v}|| \geq 0.05 * \text{avg}_{\mathbf{u}\in X_i}||\mathbf{u}||$.

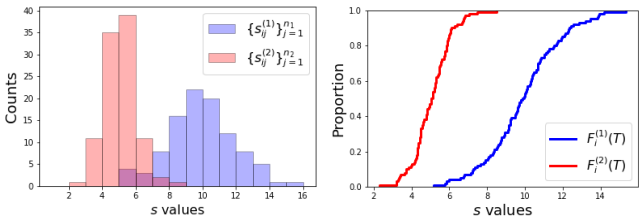


Fig. 2: Example histogram of the values $s_{i1}^{(1)}, \cdots, s_{in_1}^{(1)}$ and $s_{i1}^{(2)}, \cdots, s_{in_2}^{(2)}$ encountered in Step 6 and the corresponding empirical distribution functions $F_i^{(1)}(T)$ and $F_i^{(2)}(T)$.

We additionally note that while in Step 6 of Algorithm 1 we choose to uniformly sample from the interval $[0, s_i]$, one could instead evenly subdivide this interval and iterate through these break points or instead iterate through the ordered values $\{s_{ij}^{(1)}\}_{j=1}^{n_1}$ and $\{s_{ij}^{(2)}\}_{j=1}^{n_2}$.

Example. For illustrative purposes, suppose a rank 3 jNMF approximation of datasets X_1 and X_2 produces the vector $\bar{\mathbf{p}} = [-0.500, 0.001, 0.998]$. The first basis vector contributes to both data sets, but appears with higher frequency in X_2, the second basis vector contributes equally to both data sets, and the third basis vector appears almost exclusively in X_1. In the toy visualization of Figure 1, the entries of $\bar{\mathbf{p}} = [-0.500, 0.001, 0.998]$ might correspond to the blue, black, and red basis vectors, respectively.

B. Distance Measure

Although the basis vectors and the $\bar{\mathbf{p}}$ vector are interpretable, a single scalar value that measures similarity or distance between two data sets is often useful. We define the distance measure between the two data matrices X_1, X_2 as

$$d(X_1, X_2) := \|\bar{\mathbf{p}}\|_1,$$

where $\bar{\mathbf{p}}$ is computed via the Algorithm 1 in Subsection III-A.

We list here a few desirable properties of a distance measure $d(X_1, X_2)$, which are satisfied by the Chamfer distance, and of the vector $\bar{\mathbf{p}}$. These properties will be experimentally verified for our proposed measure in Subsection IV-A and could likely be proven for globally optimal solutions to jNMF. Let X_1 be a data matrix with n columns.

- **(P1)** Symmetry: $d(X_1, X_2) = d(X_2, X_1)$ with $\bar{\mathbf{p}}_1 = -\bar{\mathbf{p}}_2$ where $\bar{\mathbf{p}}_1$ corresponds to the comparison $d(X_1, X_2)$ and $\bar{\mathbf{p}}_2$ corresponds to the comparison $d(X_2, X_1)$.
- **(P2)** Self-similarity: $d(X_1, X_1) = 0$ and $\bar{\mathbf{p}} = \mathbf{0}$.
- **(P3)** Permutation invariance: If P_π is an $n \times n$ permutation matrix, $d(X_1, X_1P_\pi) = 0$ and $\bar{\mathbf{p}} = \mathbf{0}$.
- **(P4)** Scaling invariance: $d(X_1, \lambda X_1) = 0$, $\bar{\mathbf{p}} = \mathbf{0}$ for $\lambda > 0$.
- **(P5)** Large subsets: If the columns of $\tilde{X}_1$ are a large subset of those of X_1, then $d(X_1, \tilde{X}_1) \approx 0$, and $d(X_1, \tilde{X}_1)$ decreases monotonically as the number of columns of $\tilde{X}_1$ approaches n.
- **(P6)** Additive noise: If $\epsilon > 0$ is small and N is a noise matrix, $d(X_1, X_1 + \epsilon N) \approx 0$ and $d(X_1, X_1 + \epsilon N)$ grows monotonically with ϵ.

IV. Empirical Results

In this section, we illustrate the proposed method and distance measure on a toy image dataset called the Swimmer dataset [31], which is composed of 11×20-pixel images such as that of Figure 6a, and the 20 Newsgroups dataset [32]. Let X_1 be the matrix with columns that are the vectorized images from the Swimmer dataset and let N be a noise matrix of the same size as X_1, with entries sampled i.i.d. from unif([0,1]). All Swimmer jNMF experiments are run with rank $k = 10$.

A. Distance measure properties

In this section, we verify some of the desired properties from Subsection III-B experimentally. We note that our distance measure $d(X_1, X_2)$ appears to exhibit the self-similarity property (**P2**), the permutation invariance property (**P3**), and the scaling invariance property (**P4**). Our measure, like the Chamfer measure, produces $d(X_1, X_2) = 0$ and $\bar{\mathbf{p}} = \mathbf{0}$ when applied to X_1 and X_2, an identical, permuted, or scaled copy of X_1. Note that we scale the data prior to applying either distance measure as indicated in Step 1 of Algorithm 1. See Table I. We note that when comparing to the Chamfer distance, we do not seek to produce lower errors, simply to exhibit that our proposed distance measure exhibits similar qualitative behavior as the Chamfer distance, which is well-used.

TABLE I: Average value of our distance measure, $d(X_1, X_2)$, and the Chamfer distance, $d_{\text{cham}}(X_1, X_2)$, over fifty trials, where P_π represents the permutation matrix corresponding to a randomly sampled permutation π, and $\lambda > 0$ represents a randomly sampled value in $\{0.1, 1, 10, 100\}$, $\tilde{X}_1$ is X_1 with 10% of its columns randomly removed, and N is a matrix with entries i.i.d. from unif([0, 1]).

X_2	X_1	$X_1 P_\pi$	λX_1	$\tilde{X}_1$	$X_1 + N$	N
$d(X_1, X_2)$	0.000	0.000	0.000	0.052	1.509	2.297
$d_{\text{cham}}(X_1, X_2)$	0.000	0.000	0.000	0.000	0.741	1.560

It appears that the distance measure exhibits the large subsets property (**P5**) since $d(X_1, \tilde{X}_1)$ remains small when $\tilde{X}_1$ is formed as a large column subset of X_1. While $d(X_1, \tilde{X}_1)$ is small, our method is still able to distinguish between X_1 and $\tilde{X}_1$, whereas $d_{\text{cham}}(X_1, \tilde{X}_1) = 0$ for all the $\tilde{X}_1$ we examine. To verify this, we form $\tilde{X}_1$ as a random sample of $q\%$ of columns in X_1, where $q \in [88, 98]$ and plot $d(X_1, X_2)$ and $d_{\text{cham}}(X_1, \tilde{X}_1)$ for each value of q; see Figure 3. The distance measure also appears to exhibit the additive noise property (**P6**). In Figure 4, we see that $d(X_1, X_1 + \epsilon N)$, like $d_{\text{cham}}(X_1, X_1 + \epsilon N)$, grows monotonically with $\epsilon \in [0, 1]$. All experimental values are averaged over fifty trials.

Our method not only satisfies convenient distance properties like the Chamfer method, but also provides additional insight into the relationships between the datasets via the basis vectors produced by jNMF. Figure 5 shows the basis vectors and their associated $\bar{\mathbf{p}}$ scores produced by our method applied to X_1 and $X_1 + N$. The images with the dark background are good approximations for the basis vectors of the Swimmer data set. The primarily white basis vector is used almost exclusively in the noisy data, so this vector is the primary difference between our two data sets. The method was able to isolate the basis vectors for the original dataset from the noise.

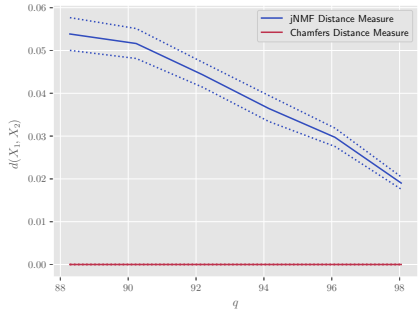


Fig. 3: The jNMF distance measure (blue) and Chamfer distance (red) $d(X_1, \tilde{X}_1)$ where X_1 is the Swimmer data matrix and $\tilde{X}_1$ is formed as a random sample of $q\%$ of columns in X_1, for $q \in [88, 98]$, decrease monotonically as the q grows. Values are averaged over fifty trials.

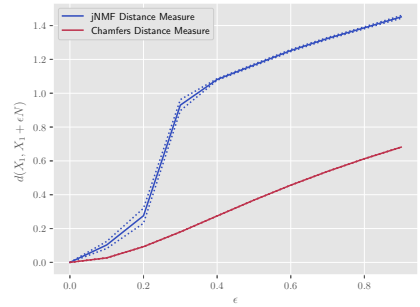


Fig. 4: The jNMF distance measure (blue) and Chamfer distance (red) $d(X_1, X_1 + \epsilon N)$, where X_1 is the Swimmer data matrix, N is a matrix with entries sampled i.i.d. from unif($[0, 1]$), and $\epsilon \in [0, 1]$, grow monotonically with ϵ. Values are averaged over fifty trials.

B. Swimmer experiments

Before we can use our measure to compare data sets, we must consider what it means for $d(X_1, X_2)$ to be small. Since $0 \leq |\bar{\mathbf{p}}_i| \leq 1$, the maximum value of the proposed distance measure is the rank of the jNMF approximation, k. However, $d(X_1, X_2)$ is frequently significantly below this value. As a benchmark for considering the significance of these distance values, we measure the distance between our structured Swimmer image matrix X_1 and noise matrix N to be $d(X_1, N) = 2.297$. While this value is larger than other distance values observed in our previous experiments, is it far below the upper bound of $k = 10$. Despite this relatively low distance measure, $d(X_1, N)$ can serve as a baseline for interpreting $d(X_1, X_2)$ for other matrices X_2.

We measure the distance between X_1 and a matrix X_2 formed by swapping the zeros and ones in X_1; see Figure 6a. Note that the data points in X_1 can be constructed by starting

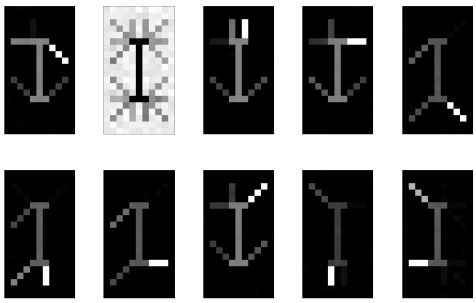
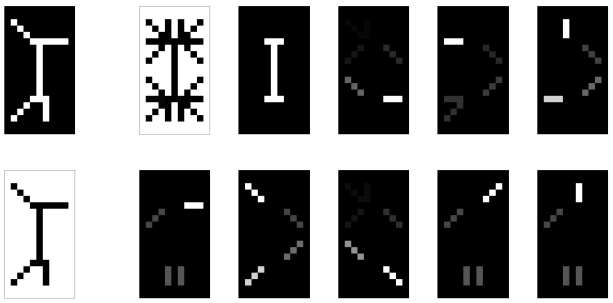


Fig. 5: Basis vectors produced by our method applied to Swimmer data matrix X_1 and $X_1 + N$ where the entries in N are sampled i.i.d. from unif($[0, 1]$). The associated $\bar{\mathbf{p}}$ values for each basis image are, reading left to right and top to bottom, $[0.063, -0.901, 0.076, 0.065, 0.069, 0.058, 0.058, 0.069, 0.079, 0.079]$ and $d(X_1, X_1+N) = 1.517$ (note that this value is computed on a single trial while the corresponding entry of Table I is averaged over 50 trials). All basis vectors contribute roughly equally to both data sets, with the exception of the basis vector in the second position of the first row, which contributes almost exclusively to the noisy data set $X_1 + N$.



(a) (b)

Fig. 6: (a) Sample data points from the Swimmer data set (top) and the modified swimmer data set, where all zeros and ones are switched (bottom). (b) Basis vectors learned by jNMF with rank $k = 10$ on the Swimmer data set (X_1) with an inverted copy of the Swimmer data set such that all the zeros and ones are switched (X_2). Ordering the basis vectors left to right and top to bottom, $\bar{\mathbf{p}} = [-0.999, 1.000, 0.010, -0.017, 0.003, -0.004, 0.015, 0.004, -0.001, -0.000]$ and $d(X_1, X_2) = 2.054$. The leftmost basis vector in the top row contributes almost exclusively to X_2 and the basis vector in the second position of the top row contributes almost exclusively X_1. The other basis vectors contribute roughly evenly to both, adding limbs to X_1 and removing them from X_2.

with the body and adding in limbs, while those in X_2 can be constructed by starting with a body with all possible limbs and covering the limbs that are not being used in a particular data point. Figure 6b shows that this data set can be represented well with eight common basis vectors and two additional basis vectors, each of which is strongly associated with one data set. The common basis vectors are used differently by the two data sets; in X_1, they are used to add limbs to the body, while in X_2, they are used to cover up limbs. The method both identifies similar structures between the two datasets and extracts the features necessary distinguish them.

C. 20 Newsgroups experiments

As a final illustration of the promise of our proposed method and distance measure, we measure distances between the term frequency-inverse document frequency (tf-idf) representations of the various newsgroups (categories) in the 20 Newsgroups dataset [32]. The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents. The data set consists of six groups partitioned roughly according to subjects, with a total of 20 subgroups, and is an experimental benchmark for document classification and clustering; see e.g., [23].

In Figures 7 and 8, we present heatmaps with colors corresponding to average jNMF distances and average Chamfer distances, respectively, between samples of 100 documents of each of the twenty newsgroups, averaged over 50 trials. We remove headers, footers, and quotes from the 20 Newsgroups dataset, and then apply the tf-idf transformation to the entire set. In each trial, we sample 100 documents (represented as tf-idf vectors with length equal to the size of the entire data corpus) uniformly from each newsgroup and calculate pairwise distances between each sample. The rows and columns of the resulting distance matrix are then re-ordered and line-segregated according to cluster labels assigned by k-means with $k = 6$ applied to the columns of the distance matrix.

Applying clustering to the jNMF and Chamfer distance matrices reveals existing block structure. While neither distance clustering respects the newsgroups divisions, the identified clusters represent highly related topics. The clustering applied to the Chamfer distance matrix correctly identifies "comp" newsgroup, while the jNMF clustering adds the "for sale" group to this cluster. Both distance clusterings group the "hockey" and "baseball" groups. Each clustering has an "atheism"/"politics" cluster, but the jNMF clustering separates these into two clusters and includes the "sci.med" group, while Chamfer groups "sci.med" with "sci.space" but places "sci.crypt" into the "atheism"/"politics" cluster.

Qualitatively, the clusters identified by the jNMF distance and the Chamfer distance are coherent. However, the jNMF distance produces clusters with significantly lower relative intra-cluster to inter-cluster distance ratio than that of the Chamfer distance.

Conclusion

In this work, we proposed a promising distance measure for datasets based on shared features learned by jNMF. Our proposed distance measure indicates similarity of two datasets and the proposed method learns which basis components are shared between the datasets and which are not. As one would hope, our proposed distance measure exhibits permutation and

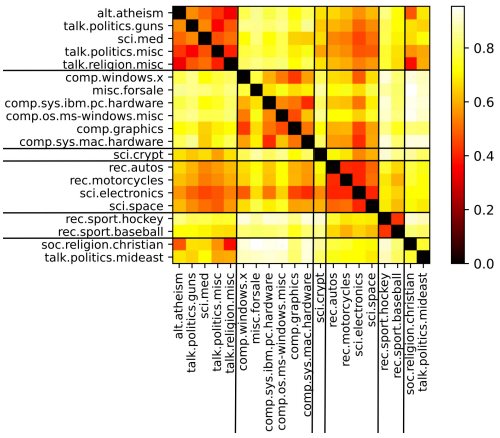


Fig. 7: Average jNMF based distance between samples of 100 documents of the twenty newsgroups (averaged over 50 trials).

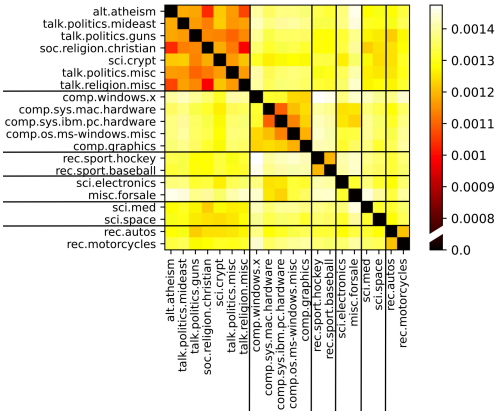


Fig. 8: Average Chamfer distance between samples of 100 documents of the twenty newsgroups (averaged over 50 trials).

scaling invariance, symmetry, and monotonicity over subset relationships and additive noise in the data.

Future work includes applying the measure in tasks like anomaly or plagiarism detection, investigating hyperparameter choice, and exploring distance measures derived from different matrix factorizations or low-dimensional approximations.

Acknowledgments

We thank April Zhao for helpful discussions early on. We are also deeply appreciative to the reviewers whose helpful comments have improved the manuscript.

References

[1] J. P. McCrae and P. Buitelaar, "Linking datasets using semantic textual similarity," *Cybernetics Inform. Tech.*, vol. 18, no. 1, pp. 109–123, 2018.

[2] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Inform. Process. Manag.*, vol. 42, no. 2, pp. 373–386, 2006.

[3] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, 2017, pp. 652–660.

[4] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3D object detection in point clouds," in *Proc. IEEE/CVF Int. Conf. Comp. Vis.*, 2019, pp. 9277–9286.

[5] Q.-H. Pham, T. Nguyen, B.-S. Hua, G. Roig, and S.-K. Yeung, "JSIS3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields," in *Proc. IEEE/CVF Conf. Comp. Vis. Patt. Recog.*, 2019, pp. 8827–8836.

[6] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, and D. Lin, "Balanced Chamfer distance as a comprehensive metric for point cloud completion," *Adv. Neur. In.*, vol. 34, pp. 29 088–29 100, 2021.

[7] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE T. Pattern Anal.*, vol. 15, no. 9, pp. 850–863, 1993.

[8] A. Javaheri, C. Brites, F. Pereira, and J. Ascenso, "A generalized Hausdorff distance based quality metric for point cloud geometry," in *Int. Conf. Qual. Multimedia Exp.* IEEE, 2020, pp. 1–6.

[9] C. L. Mallows, "A note on asymptotic joint normality," *Ann. Math. Stat.*, pp. 508–515, 1972.

[10] D. Alvarez-Melis and N. Fusi, "Geometric dataset distances via optimal transport," *Lect. Notes Comput. Sc.*, vol. 33, pp. 21 428–21 439, 2020.

[11] T. Nguyen, Q.-H. Pham, T. Le, T. Pham, N. Ho, and B.-S. Hua, "Point-set distances for learning representations of 3D point clouds," in *Proc. IEEE/CVF Int. Conf. Comp. Vis.*, 2021, pp. 10 478–10 487.

[12] D. Urbach, Y. Ben-Shabat, and M. Lindenbaum, "DPDist: Comparing point clouds using deep point cloud distance," in *Eur. Conf. Comp. Vis.* Springer, 2020, pp. 545–560.

[13] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, and D. Lin, "Density-aware Chamfer distance as a comprehensive metric for point cloud completion," *arXiv:2111.12702*, 2021.

[14] M. P. Khambete, W. Su, J. Garcia, and M. A. Badgeley, "Quantification of BERT diagnosis generalizability across medical specialties using semantic dataset distance," *Proc. AMIA Joint Summ. Trans. Sci.*, 2021.

[15] S. Ben-David, J. Blitzer, K. Crammer, and F. C. Pereira, "Analysis of representations for domain adaptation," in *Lect. Notes Comp. Sc.*, 2006.

[16] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv:0902.3430*, 2009.

[17] X. Liu, Y. Bai, Y. Lu, A. Soltoggio, and S. Kolouri, "Wasserstein task embedding for measuring task similarities," *arXiv:2208.11726*, 2022.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, pp. 788– 791, 1999.

[19] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SIAM Int. Conf. Data Mining*. SIAM, 2013, pp. 252–260.

[20] H. Kim, J. Choo, J. Kim, C. K. Reddy, and H. Park, "Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2015, pp. 567–576.

[21] X. Ma, D. Dong, and Q. Wang, "Community detection in multi-layer networks using joint nonnegative matrix factorization," *IEEE T. Knowl. Data En.*, vol. 31, no. 2, pp. 273–286, 2018.

[22] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Lect. Notes Comput. Sc.*, vol. 13, 2000.

[23] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Proc. Let.*, vol. 17, no. 1, pp. 4–7, 2009.

[24] M. Ahn, R. Grotheer, J. Haddock, L. Kassab, A. Kryshchenko, K. Leonard, S. Li, A. Madushani, T. Merkh, D. Needell, E. Sizikova, and C. Wang, "Semi-supervised nonnegative matrix factorization models for topic modeling in learning tasks," in *Proc. Asilomar Conf. Sig. Sys. Comp.*, 2020.

[25] N. Gillis, *Nonnegative matrix factorization*. SIAM, 2020.

[26] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer, 2005, vol. 488.

[27] A. Kolmogorov, "Sulla determinazione empirica di una lgge di distribuzione," *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[28] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Ann. Math. Stat.*, vol. 19, no. 2, pp. 279–281, 1948.

[29] H. Cramér, *On the composition of elementary errors: Statistical applications*. Almqvist and Wiksell, 1928.

[30] R. Von Mises, "Statistik und wahrheit," *Julius Springer*, vol. 20, 1928.

[31] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" *Lect. Notes Comput. Sc.*, vol. 16, 2003.

[32] J. Rennie. (2008) 20 Newsgroups. [Online]. Available: http://qwone.com/ jason/20Newsgroups/