Determining whether two datasets cluster similarly without determining the clusters

Frederik Van Eeghem, Lieven De Lathauwer, *Fellow, IEEE*

***Abstract*—When several datasets of observations in the same space are available, such as customer data in various countries, a key question is whether these datasets cluster similarly and if they can be merged for further analysis. For this reason, assessing whether two related datasets share the same cluster centers is an important problem. This paper tackles this problem using a novel method based on tensor similarity, which allows us to compare the cluster centers of two datasets without explicitly computing these clusters.**

***Index Terms*—Tensors, similarity, clustering, comparison, comparing**

I. Introduction

In data analysis, several applications call for a comparison of the cluster patterns of two datasets. Consider customer profiling based on consumer data from various countries. If the same profiles hold in different countries, the different datasets can be merged before further analysis. If not, each country must be analyzed separately. Other examples can be found in sets of scientific experiments conducted at different points in time and by different researchers (see [2]), or in survey results to see whether the results from a one-hour survey can be replaced by the results of a ten-minute survey.

In this paper, a measure is constructed that compares the cluster centers of different datasets in the same space. A naive way to tackle this problem is to cluster the datasets separately using any algorithm ([23]) and then to compare the cluster centers. However, this approach is sensitive to the cluster algorithm and how it is initialized.

Metrics to compare various clusterings of the same dataset have been treated many times in literature [15], [22], [21]. These approaches can be roughly divided into three groups: counting pairs of elements, checking set overlaps and information-theoretical mutual information [22]. However, these metrics assume the same underlying dataset, which is not the case in the applications considered in this paper.

Methods that compare cluster centers across datasets are scarce in the literature. In [14], a Bayesian approach is considered that determines whether both datasets can emerge from a statistical model with the same model parameters. The downside of this approach is that statistical modeling is needed. Another approach translates the comparison of datasets to a binary classification problem [16]. First, the datasets are merged and equipped with a source indicator. If a binary classifier is unable to reliably predict this source indicator, the datasets are considered different. Note that this approach requires a classifier to be trained and tuned.

The approach presented in this paper is based on tensor analysis but requires only simple linear algebra. Unlike matrix decompositions, tensor decompositions are unique up to trivial indeterminacies under mild conditions [13], [17]. Recent results in tensor similarity show that it is possible to compare the underlying factors of two tensor decompositions without explicitly computing them [20], [7], [5], [19]. This property can be used to assess whether two separate datasets in the same space share the same cluster centers, yielding a conceptually simple yet powerful method.

Notations: Following notation are used throughout this paper: Scalars are denoted by lowercase letters, vectors by bold lowercase letters, matrices by bold uppercase letters, and tensors by uppercase calligraphic letters (e.g., $\mathcal{A}$). Subsets of vectors, matrices and tensors are denoted by indices (e.g. a_{ij} for the entry of the matrix $\mathbf{A}$). Following symbols are used: outer product ($\circ$), Kronecker product ($\otimes$), Khatri–Rao product ($\odot$), expected value ($E[\cdot]$), Kronecker delta ($\delta(\cdot)$), vector of all ones ($\mathbf{1}$), ith canonical basis vector of a space ($\mathbf{e}_i$).

II. Problem statement

Consider two datasets, $\mathbf{V} \in \mathbb{C}^{I\times M}$ and $\mathbf{W} \in \mathbb{C}^{I\times N}$, representing a set of observations in the same I-dimensional feature space. The jth column of $\mathbf{V}$ is denoted by $\mathbf{v}(j)$ and represents the jth data point of the dataset (similarly for $\mathbf{w}(j)$ of $\mathbf{W}$). Assume both datasets display R clusters with cluster centers $\mathbf{\Pi} = [\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_R]$ and $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_R]$ respectively. The probability that a randomly chosen data point $\mathbf{v}(j)$ is part of the rth cluster of dataset $\mathbf{V}$ is denoted by ξ_r (and ζ_r for dataset $\mathbf{W}$). The points of both sets can now be written as

$$\mathbf{v}(j) = \boldsymbol{\pi}(j) + \mathbf{p}(j),$$
$$\mathbf{w}(j) = \boldsymbol{\gamma}(j) + \mathbf{q}(j).$$

In this equation $\boldsymbol{\pi}(j), \boldsymbol{\gamma}(j) \in \mathbb{C}^I$ represent functions that map the index of a data point to its corresponding cluster center $\boldsymbol{\pi}_r$ or $\boldsymbol{\gamma}_r$. The functions $\mathbf{p}(j), \mathbf{q}(j) \in \mathbb{C}^I$ represent the deviation of the jth data point from its corresponding cluster center. Assume following conditions hold:

- Deviations from cluster centers are zero-mean for each cluster, i.e. $E[\mathbf{p}(j)] = \mathbf{0}$ and $E[\mathbf{q}(j)] = \mathbf{0}$.

This research is funded by (1) Research Council KU Leuven: C1 project c16/15/059; (2) FWO: projects: G.0830.14N, G.0881.14N, EOS Project no 30468160 (SeLMA); (3) This research received funding from the Flemish Government under the Onderzoeksprogramma Artificile Intelligentie (AI) Vlaanderen program.

Frederik Van Eeghem and Lieven De Lathauwer are with both the Group of Science, Engineering and Technology, KU Leuven Kulak, E. Sabbelaan 53, B-8500 Kortrijk, Belgium and with the Department of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. (e-mail: frederik.vaneeghem@kuleuven.be, lieven.delathauwer@kuleuven.be)

- The cluster distributions are independent across clusters and dimensions, i.e., $p_k(j)$ is independent from $p_l(j)$ for $k \neq l$. This implies that the covariance matrix of $\mathbf{p}(j)$ (and $\mathbf{q}(j)$) is diagonal. For cluster r of dataset $\mathbf{V}$, the diagonal of the covariance matrix is given by $\boldsymbol{\sigma}^{(r,v)}$.

The goal is to assess whether the cluster centers are the same for both datasets without explicitly determining the clusters.

III. Tensors

A. Rank and Canonical Polyadic Decomposition

A rank-1 tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times \cdots \times I_N}$ of order N is defined as the outer product of N nonzero vectors $\mathbf{u}^{(n)} \in \mathbb{C}^{I_n}$, i.e., $\mathcal{A} = \mathbf{u}^{(1)} \circ \cdots \circ \mathbf{u}^{(N)}$. The rank of a tensor $\mathcal{T}$ is then defined as the minimum number of rank-1 terms that yield $\mathcal{T}$ in a linear combination. Such a (minimal) linear combination of rank-1 terms is called a canonical polyadic decomposition (CPD). Mathematically, the CPD of a rank-R tensor $\mathcal{T} \in \mathbb{C}^{I_1 \times \cdots \times I_N}$ can be written as

$$\mathcal{T} = \sum_{r=1}^{R} \lambda_r \mathbf{u}_r^{(1)} \circ \cdots \circ \mathbf{u}_r^{(N)} = \left[\!\left[\boldsymbol{\lambda}; \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)} \right]\!\right]. \quad (1)$$

In which the factor vectors $\mathbf{u}_r^{(n)} \in \mathbb{C}^{I_n}$ are concatenated in factor matrices $\mathbf{U}^{(n)} = \left[\mathbf{u}_1^{(n)}, \ldots, \mathbf{u}_R^{(n)}\right] \in \mathbb{C}^{I_n \times R}$ and the weighting factors in $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_R] \in \mathbb{C}^R$ to allow a more concise notation [13].

One of the key properties of tensors are that their decompositions are unique under mild conditions:

Theorem III.1. *Consider the polyadic decomposition of $\mathcal{X} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$ as defined in* (1)*. If*

$$\begin{cases} \mathbf{U}^{(3)} \in \mathbb{C}^{I_3 \times R} \text{ has full column rank}, \\ \mathcal{C}_2\left(\mathbf{U}^{(1)}\right) \odot \mathcal{C}_2\left(\mathbf{U}^{(2)}\right) \text{ has full column rank}, \end{cases}$$

then the rank of $\mathcal{X}$ is R and the CPD of $\mathcal{X}$ is essentially unique [11], [4], [8], [9].

In this theorem, $\mathcal{C}_2(\mathbf{A})$ denotes the 2nd compound matrix of an $I \times R$ matrix $\mathbf{A}$, which is the $\binom{I}{2} \times \binom{R}{2}$ matrix containing the determinants of all 2×2 submatrices of $\mathbf{A}$ [8]. The determinants are arranged such that the submatrix index sets are in lexicographic order.

B. Tensor similarity

Recent research on tensor similarity shows that it is possible to compare the factors of the decompositions of two (different) tensors without explicitly having to compute these decompositions [5], [7], [20], [19]. The Theorem III.2 below shows one specific result of [20] that considers similarity in the first two modes of two rank-R CPDs. The theorem indicates that, under specific (mild) conditions, the factors in the first two modes are essentially equal if and only if the unfolded tensors in the first two modes span the same space:

Theorem III.2. *Let two tensors $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}$ admit a rank-R CPD as follows:*

$$\mathcal{T}^{(1)} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \in \mathbb{C}^{I \times J \times K},$$
$$\mathcal{T}^{(2)} = [\![\mathbf{D}, \mathbf{E}, \mathbf{F}]\!] \in \mathbb{C}^{I \times J \times K}.$$

If

$$(i) \quad \mathbf{C} \text{ and } \mathbf{F} \text{ have full column rank},$$
$$(ii) \quad \mathcal{C}_2(\mathbf{A}) \odot \mathcal{C}_2(\mathbf{B}) \text{ has full column rank},$$

then $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$ are essentially equal in the first two modes if and only if span $\left\{\mathbf{T}^{(1)}_{IJ \times K}\right\}$ = span $\left\{\mathbf{T}^{(2)}_{IJ \times K}\right\}$.

The matrices $\mathbf{T}^{(1)}_{IJ \times K} \in \mathbb{C}^{IJ \times K}$ and $\mathbf{T}^{(2)}_{IJ \times K} \in \mathbb{C}^{IJ \times K}$ are matricized versions of the tensors $\mathcal{T}^{(1)}$ and $\mathcal{T}^{(2)}$ obtained by stacking the columns of each frontal matrix slice [13].

An important note for this theorem lies in the exact definition of essential equality. Since the CPD has trivial scaling and permutation ambiguities, essential equality means that the factors in the first two modes are the same up to column scaling and permutation [20].

IV. Comparing cluster centers

A. Step-by-step algorithm

The method presented in this paper consists of three main steps: fixing scaling ambiguities, tensorization and comparison of the tensors.

First, the datasets are extended to fix any scaling ambiguities. Because of the scaling and permutation ambiguities of tensor decompositions, tensor similarity methods can only compare factor matrices up to scaling and permutation. This scaling ambiguity can be fixed by adding normalization data to $\mathbf{V}$ and $\mathbf{W}$ and constructing $\widehat{\mathbf{V}} = \left[\mathbf{V}^{\mathrm{T}}, \mathbf{1}\right]^{\mathrm{T}}$ and $\widehat{\mathbf{W}} = \left[\mathbf{W}^{\mathrm{T}}, \mathbf{1}\right]^{\mathrm{T}}$. The additional row of ones fixes the scaling ambiguity, which is easy to see as follows: vectors $[2, 2]^{\mathrm{T}}$ and $[3, 3]^{\mathrm{T}}$ are the same up to scaling, whereas vectors $[2, 2, 1]^{\mathrm{T}}$ and $[3, 3, 1]^{\mathrm{T}}$ are not.

In a second step, the extended datasets are tensorized using one of the techniques presented in Sections IV-B and IV-C depending on the assumptions. This yields two tensors $\mathcal{V} \in \mathbb{C}^{I \times I \times X}$ and $\mathcal{W} \in \mathbb{C}^{I \times I \times Y}$ that contain the respective cluster centers in the factor matrices in the first two modes. The exact dimensions X and Y of the tensors may vary depending on the tensorization method that is used.

Finally, the factor matrices of both tensors $\mathcal{V}$ and $\mathcal{W}$ in the first two modes can be compared using tensor similarity. The high-level steps for one specific tensor similarity method from [20] are briefly repeated here for later reference:

1) Verify whether the conditions of Theorem III.2 hold, in which the challenge is that the factor matrices are not available since no explicit decomposition is computed. See [20] and [19] for detailed instructions on how to verify these conditions without directly accessing the factor matrices.
2) Determine the number of clusters R. This value can stem from knowledge of the application or can be estimated. If R is smaller than I^2 and X (or Y), a good estimate for R is the number of dominant singular values of the unfolded tensor $\mathbf{V}_{I^2 \times X}$ (or $\mathbf{W}_{I^2 \times Y}$). The underlying rationale for this estimate is that under the given conditions an unfolded tensor that admits a rank-R CPD is a rank-R

matrix since the matrix dimensions are not limiting the rank.

3) Find bases for the R-dimensional dominant column spaces of the unfolded tensors $\mathbf{V}_{I^2\times X}$ and $\mathbf{W}_{I^2\times Y}$. This can for instance be done by taking the left singular vectors corresponding to the R largest singular values of that matrix.
4) The final step consists of computing the R principal angles between the computed bases of the unfolded tensors. These principal angles are the minimum angles between both subspaces. Further details on principal angles can be found in [3], [12].

By going through these steps, the similarity of the R cluster centers of $\mathbf{V}$ and $\mathbf{W}$ has been reduced to a set of R principal angles. In brief, all principal angles will be zero if the cluster centers of both datasets are the same since the scaling ambiguity has been fixed [20].

Note that the approach presented in this paper can easily be adapted for the case where $\mathbf{V}$ and $\mathbf{W}$ only share part of their cluster centers. Without going in full detail, the interested reader is referred to [20] for a theorem that expands on this partial similarity.

B. Tensorization via Kronecker products

To tensorize the given datasets, compute the Khatri–Rao products $\mathbf{V}\odot\mathbf{V}\in\mathbb{C}^{I^2\times M}$ and $\mathbf{W}\odot\mathbf{W}\in\mathbb{C}^{I^2\times N}$:

$$\mathbf{V}\odot\mathbf{V}=\left[\mathbf{v}(1)\otimes\mathbf{v}(1),\ldots,\mathbf{v}(M)\otimes\mathbf{v}(M)\right],$$
$$\mathbf{W}\odot\mathbf{W}=\left[\mathbf{w}(1)\otimes\mathbf{w}(1),\ldots,\mathbf{w}(N)\otimes\mathbf{w}(N)\right].$$

Now consider the jth column of $\mathbf{V}\odot\mathbf{V}$ and assume without loss of generality that the corresponding jth data point is in cluster r. The $(k+I(l-1))$th entry of this vector is then given by

$$\begin{aligned}&(\mathbf{v}(j)\otimes\mathbf{v}(j))_{k+I(l-1)}\\&\quad=(\pi_l(j)+p_l(j))\,(\pi_k(j)+p_k(j))\\&\quad=\pi_{lr}\pi_{kr}+p_l(j)p_k(j)+\pi_{kr}p_l(j)+\pi_{lr}p_k(j)\end{aligned}\quad(3)$$

Analogous expressions can be found for entries of $\mathbf{W}\odot\mathbf{W}$.

Generally, all terms in equation (3) are nonzero. However, one step in the tensor similarity procedure computes the dominant subspace of the unfolded tensors $\mathbf{V}\odot\mathbf{V}$ and $\mathbf{W}\odot\mathbf{W}$, which has an averaging effect on the entries of both matrices. During this operation, the terms $\pi_{kr}p_l(j)$ and $\pi_{lr}p_k(j)$ in (3) will average to zero because of the assumed zero-mean deviations from cluster centers. This leaves following expressions for the expected value of entries of $\mathbf{V}\odot\mathbf{V}$:

$$E\left[(\mathbf{v}(j)\otimes\mathbf{v}(j))_{k+I(l-1)}\right]=\pi_{lr}\pi_{kr}+\delta(k-l)\left(\sigma_k^{(r,v)}\right)^2\quad(4)$$

For $j\neq k$, the term $p_l(j)p_k(j)$ averages to zero over the data points of the rth cluster because of the independence of both variables and their zero mean. For $j=k$ however, this term becomes $p_k^2(j)$ and will average to $\left(\sigma_k^{(r,v)}\right)^2$, the variance of the rth cluster in the kth dimension.

If the cluster variances remain relatively small in comparison to the norms of the cluster centers, the first term of (4) is dominant. This in turn implies that the R-dimensional dominant subspace of span $\{\mathbf{V}\odot\mathbf{V}\}$ can be approximated by $\mathbf{\Pi}\odot\mathbf{\Pi}$ (and similarly for span $\{\mathbf{W}\odot\mathbf{W}\}$). Mathematically, this can be written as

$$\mathbf{V}\odot\mathbf{V}\approx(\mathbf{\Pi}\odot\mathbf{\Pi})\,\mathbf{K}^{\mathrm{T}},\quad(5)$$
$$\mathbf{W}\odot\mathbf{W}\approx(\mathbf{\Gamma}\odot\mathbf{\Gamma})\,\mathbf{L}^{\mathrm{T}},\quad(6)$$

in which $\mathbf{K}\in\mathbb{C}^{M\times R}$ and $\mathbf{L}\in\mathbb{C}^{N\times R}$ can be interpreted as matrices that select the correct cluster center for each data point and capture the spread around that cluster center. Alternatively, the rows of $\mathbf{V}\odot\mathbf{V}$ and $\mathbf{W}\odot\mathbf{W}$ that are effected by the cluster variances can be omitted. This does not change anything in the computations, but affects the conditions given in Theorem III.2 as the corresponding submatrix must be considered there as well. Details of this "fiber sampling" approach and the impact on uniqueness conditions can be found in [18].

The expressions (5) and (6) are matricized CPDs, which implies that $\mathbf{V}\odot\mathbf{V}$ and $\mathbf{W}\odot\mathbf{W}$ can be interpreted as matricized tensors $\mathcal{V}\in\mathbb{C}^{I\times I\times M}$ and $\mathcal{W}\in\mathbb{C}^{I\times I\times N}$ that approximately admit a CPD. Because the information in the first two modes consist purely of the cluster centers, it suffices to compare just these modes.

Large scale case: If the clusters contain many data points, meaning $M,N\gg I$, the computation of $\mathbf{V}\odot\mathbf{V}$ and $\mathbf{W}\odot\mathbf{W}$ may become memory-intensive. Here, several approaches that can be used to circumvent this problem are briefly touched. First, only a subset of the available data points can be considered, though this will reduce the accuracy. Second, subspace updating can be used to update the dominant subspace of e.g. $\mathbf{V}\odot\mathbf{V}$ every time a new column is added to avoid constructing the full matrix [6]. A third way to handle large-scale data is to compute the matrices $(\mathbf{V}\odot\mathbf{V})(\mathbf{V}\odot\mathbf{V})^{\mathrm{T}}\in\mathbb{C}^{I^2\times I^2}$ and $(\mathbf{W}\odot\mathbf{W})(\mathbf{W}\odot\mathbf{W})^{\mathrm{T}}\in\mathbb{C}^{I^2\times I^2}$ instead of $\mathbf{V}\odot\mathbf{V}$ and $\mathbf{W}\odot\mathbf{W}$, since their column spaces are the same. This is advantageous because the former matrices are smaller and can be constructed without storing all columns of e.g. $\mathbf{V}\odot\mathbf{V}$. The downside of this approach is that the condition number of the problem is squared.

C. Tensorization via third-order moments

The previous section focused on the usage of second-order information for the tensorization. It is also possible to consider third-order or even higher-order information. Assume an additional assumption is imposed: the clusters are all spherical Gaussians with variances $\sigma^{(1,v)},\ldots,\sigma^{(R,v)}$. In this case, it was shown that the third-order moment of the vector $\mathbf{v}(t)$ (or similarly $\mathbf{w}(t)$) reduces to a simple expression due to the Gaussianity of the clusters and terms averaging to zero [17], [10], [1]. More specifically, an entry of the third-order moment tensor of $\mathbf{v}(t)$ can be written as:

$$m_{ijk}^{(3)}=\sum_{r=1}^{R}\xi_r\pi_{ir}\pi_{jr}\pi_{kr}+\ldots$$
$$\sum_{r=1}^{R}\xi_r\left(\sigma^{(r,v)}\right)^2\left(\pi_{ir}\delta(j-k)+\pi_{jr}\delta(i-k)+\pi_{kr}\delta(i-j)\right).$$

The first term in this equation is a decomposition in rank-1 terms, in which each cluster center contributes one term. The second term contains the influence of the cluster distributions for a select number of entries where two indices coincide. This second term can be precomputed by calculating (see [10]):

$$\begin{aligned}\mathbf{d} &= E\left[\mathbf{v}(t)\left(\mathbf{u}^{\mathrm{T}}(\mathbf{v}(t)-E\left[\mathbf{v}(t)\right])\right)^2\right] \\ &= \sum_{r=1}^{R}\xi_r\left(\sigma^{(r,v)}\right)^2\boldsymbol{\pi}_r, \end{aligned} \tag{7}$$

in which $\mathbf{u}$ represents any unit norm eigenvector corresponding to the smallest eigenvalue of the covariance matrix $E\left[(\mathbf{v}(t)-E\left[\mathbf{v}(t)\right])(\mathbf{v}(t)-E\left[\mathbf{v}(t)\right])^{\mathrm{T}}\right]$. Once $\mathbf{d}$ is computed using expression (7), tensor $\mathcal{D}\in\mathbb{C}^{I\times I\times I}$ can be constructed:

$$\mathcal{D}=\sum_{i=1}^{I}\left(\mathbf{d}\circ\mathbf{e}_i\circ\mathbf{e}_i+\mathbf{e}_i\circ\mathbf{d}\circ\mathbf{e}_i+\mathbf{e}_i\circ\mathbf{e}_i\circ\mathbf{d}\right).$$

By subtracting this tensor from the third-order moment tensor, it then immediately follows that:

$$\mathcal{M}^{(3)}-\mathcal{D}=\sum_{r=1}^{R}\xi_r\boldsymbol{\pi}^{(r)}\circ\boldsymbol{\pi}^{(r)}\circ\boldsymbol{\pi}^{(r)}.$$

This resulting tensor admitting a CPD can be constructed for both datasets $\mathbf{V}$ and $\mathbf{W}$, and subsequently used to compare the cluster centers in both datasets since each factor matrix contains the cluster centers.

V. Numerical experiments

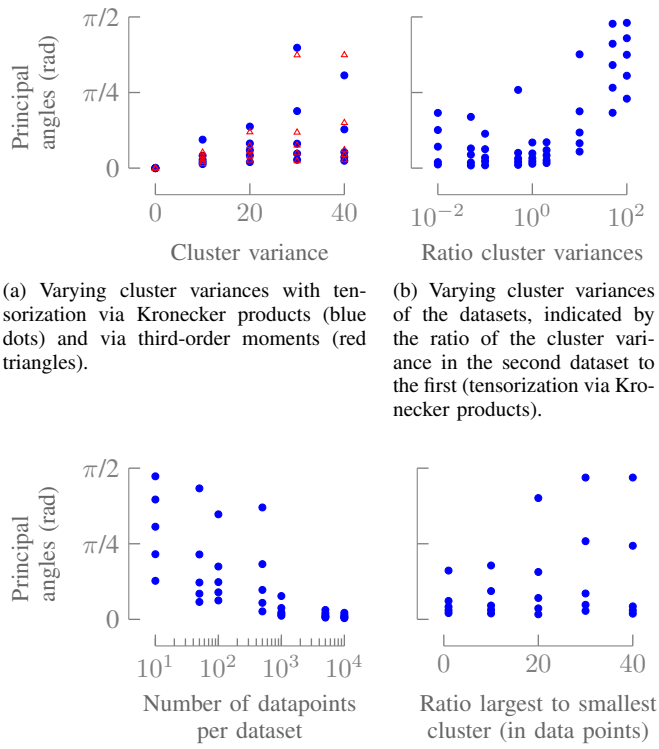
The presented method is illustrated by comparing cluster centers in the context of Gaussian mixture estimation. Consider a mixture of five spherical multivariate Gaussians $\mathcal{N}(\boldsymbol{\pi}_r,\sigma^2\mathbf{I})$. The cluster centers $\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_5\in\mathbb{C}^8$ are chosen randomly from a multivariate normal distribution with zero mean and standard deviation 5. By default, following parameter values are used unless specified otherwise: the cluster variances σ^2 are the same for all clusters and equal to 10. Two datasets are generated from this Gaussian mixture by randomly sampling 1000 data points for each. The probability of a data point belonging to cluster r is the same for all clusters, i.e. $\xi_1=\ldots=\xi_5=0.2$. The cluster centers of these two resulting data sets $\mathbf{V}\in\mathbb{C}^{8\times1000}$ and $\mathbf{W}\in\mathbb{C}^{8\times1000}$, which are the same by construction, are compared following the method presented in this paper. Figure 1 shows the median principal angles over 1000 repetitions of four numerical experiments:

(a) Varying the cluster variances in which all clusters have the same variance. The principal angles are bigger for larger cluster variances since the underlying cluster centers are implicitly estimated less accurately. This experiment shows that both tensorization strategies yield similar results, which is why only one of the methods is depicted in the other experiments.

(b) Varying cluster variances in the second dataset with respect to the first (in which the cluster variances are fixed to 10). The figure shows that even though the principal angles are largely determined by the highest variances, the results slightly worsen for smaller ratios as well. Overall, the principal angles are the smallest when all clusters have approximately the same variance.

(c) Varying the number of datapoints per dataset, which clearly illustrates that more points lead to more accurate comparisons because the underlying estimation of the cluster centers is more accurate.

(d) Varying number of points in the clusters so that the probability of a data point belonging to a specific cluster is no longer uniform. This is expressed as the ratio of the largest to the smallest cluster, where the number of points in each cluster increases linearly and the total number of datapoints per dataset remains fixed at 1000 points. For larger ratios, the performance decreases since it becomes harder to (implicitly) accurately estimate the centers of the smallest clusters.



(a) Varying cluster variances with tensorization via Kronecker products (blue dots) and via third-order moments (red triangles).

(b) Varying cluster variances of the datasets, indicated by the ratio of the cluster variance in the second dataset to the first (tensorization via Kronecker products).

(c) Varying the number of data points per dataset (tensorization via Kronecker products).

(d) Varying cluster sizes within datasets, indicated by the ratio of the largest cluster to the smallest (tensorization via third-order moments).

Figure 1: Effects of various parameters on the principal angles when the cluster centers of both datasets are the same.

VI. Conclusion

This papers introduces a novel method to compare cluster centers of two datasets. The method builds on tensor similarity, which allows us to compare the cluster centers without actually determining them. Two different tensorization options have been presented in the paper, including possible strategies to reduce the memory load of large-scale data. Numerical experiments in a Gaussian mixture estimation context show the effectiveness of the method under various parameter settings.

REFERENCES

[1] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.

[2] K. A. Baggerly, J. S. Morris, and K. R. Coombes, "Reproducibility of seldi-tof protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, vol. 20, no. 5, pp. 777–785, 2004.

[3] A. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of computation*, vol. 27, no. 123, pp. 579–594, 1973.

[4] L. De Lathauwer, "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 28, no. 3, pp. 642–666, 2006.

[5] L. De Lathauwer, "Characterizing higher-order tensors by means of subspaces," 2011, tech. report 11–32, ESAT–STADIUS, KU Leuven, Belgium.

[6] R. DeGroat, E. Dowling, and D. Linebarger, "Subspace tracking," in *The Digital Signal Processing Handbook*, V. Madisetti, Ed. CRC Press, 1998.

[7] I. Domanov and L. De Lathauwer, "From computation to comparison of tensor decompositions," *arXiv:1912.04694*.

[8] I. Domanov and L. De Lathauwer, "On the uniqueness of the canonical polyadic decomposition of third-order tensors — Part I: Basic results and uniqueness of one factor matrix," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 855–875, 2013.

[9] ——, "On the uniqueness of the canonical polyadic decomposition of third-order tensors — Part II: Overall uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 876–903, 2013.

[10] D. Hsu and S. Kakade, "Learning mixtures of spherical gaussians: moment methods and spectral decompositions," in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, 2013, pp. 11–20.

[11] T. Jiang and N. D. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear model with constant modulus constraints," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2625–2636, Sep. 2004.

[12] A. V. Knyazev and M. E. Argentati, "Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates," *SIAM Journal on Scientific Computing*, vol. 23, no. 6, pp. 2008–2040, 2002.

[13] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[14] P. Marshall, N. Rajguru, and A. Slosar, "Bayesian evidence as a tool for comparing datasets," *Physical Review D*, vol. 73, no. 6, p. 067302, 2006.

[15] M. Meilă, "Comparing clusterings: an information based distance," *Journal of multivariate analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[16] R. Pearson. Using many models to compare datasets. [Online]. Available: https://cran.r-project.org/web/packages/datarobot/vignettes/ComparingSubsets.html

[17] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[18] M. Sørensen and L. De Lathauwer, "Fiber sampling approach to canonical polyadic decomposition and application to tensor completion," *SIAM J. Matrix Anal. Appl.*, vol. 40, no. 3, pp. 888–917, 2019.

[19] F. Van Eeghem and L. De Lathauwer, "Tensor similarity in chemometrics," in *Chemistry, Molecular Sciences and Chemical Engineering*, R. Tauler, Ed. Elsevier, 2020, accepted for publication.

[20] F. Van Eeghem, O. Debals, and L. De Lathauwer, "Tensor similarity in two modes," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1273–1285, 2018.

[21] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2837–2854, 2010.

[22] S. Wagner and D. Wagner, *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[23] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.