Project or Factorize? A case study of Multiview CCA and PARAFAC2 tensor factorization

Jia Chen
Dept. of Electrical and Computer Engineering
University of California Riverside
jiac@ucr.edu

Evangelos E. Papalexakis
Dept. of Computer Science and Engineering
Uuniversity of Californiav Riverside
epapalex@cs.ucr.edu

***Abstract*—Consider learning the shared representations from multiple unlabeled views. Previous work either projects different views to the same space while enforcing the agreement among the projected views such as multiview canonical correlation analysis (MCCA), or factorizes different views while ensuring the common latent components across the views such as PARAFAC2, a tensor decomposition method. In this paper we first investigate a fundamental question: "Do these two approaches learn different representations?" Preliminary numerical results suggest that in practice they do, which, in turn, begs the question of how we can leverage this observation in order to compute a superior representation. In this paper, we present a simple proof-of-concept scheme which augments MCCA with PARAFAC2 representations and vice-versa, and we demonstrate on multiple real datasets that such scheme can improve upon the baseline representation, paving the way for future research on optimally combining the strengths of projection and factorization methods for multiview representation learning.**

***Index Terms*—Canonical correlation analysis, PARAFAC2, tensor methods, projection models, factorization models, multiview machine learning.**

I. Introduction

Learning data representations by analyzing multiple views jointly is arguably rendering more promising performance and having better generalization ability than single-view learning [26]. Interesting applications of multiview learning (MVL) have been found in information retrieval, clustering, and classification/recognition [30].

Typical approaches are based on either *projection models* such as Canonical Correlation Analysis (CCA) [13] or *factorization models* including tensor decomposition. Both have demonstrated promising results for downstream tasks [24].

Focusing on representation learning from multiview data, projection models, on the one hand, transform different views to new spaces and fuse them to learn a single representation, which include CCA variants [6], [11], [13], multi-kernel learning (MKL) based methods [8], [25], deep autoencoder variants [19], [29]. Multiview CCA (MCCA) searches for shared lower-dimensional representations of multiple views by minimizing the distance between the shared representations and the projected data from each view. MKL algorithms combine data-driven kernels for different views together to improve learning performance. The objective of autoencoder variants is to learn a lower-dimensional representation of multiple views by reconstructing one view from the other views. On the other hand, factorization models such as tensor methods [4], [10], [16], [20], [24] and matrix factorization-based methods [15], [17], [18], [31], [32], exploit latent subspace shared by multiple views by assuming that different views are generated from the same subspaces. Matrix factorization-based approaches decompose multiple views layer by layer to obtain complementary information. Tensor-based methods model multi-modal interactions among multiple views as a tensor structure and factorize it to extract meaningful and hidden information.

Despite the abundance of both projection-based and factorization-based MVL models, identifying the model that fits a data better and how to leverage the advantages of both types models are rarely studied and are challenging topics. More specifically, for example, the fundamental question of whether those two classes of methods ultimately compute the similar or different representations has received surprisingly little attention. In this paper, we make a first attempt at answering that question, providing numerical evidence that in practice those two classes of MVL learn different representations. Towards that end, we use the UCI dataset (the details about this dataset can be found in Sec. III) to elaborate such difference between two representative multiview learning models: MCCA (a projection model) and PARAFAC2 (a factorization model). The five views of the UCI data with dimensions between 47 and 240 are fed into MCCA and PARAFAC2 to extract two latent representations, respectively. In Fig. 1, the singular values of the concatenated matrix of the two latent representations are depicted, where d is the dimension of the latent representation from MCCA and PARAFAC2. Given that the covariance of the latent representation from MCCA is an identity matrix, the rank of the concatenated matrix should be at least d. We make the following fascinating observations:

1) For different choices of d, the rank of each concatenated

Research was supported by NSF under CAREER grant no. IIS 2046086 and CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) grant no. 2112650, NIFA-AFRI Sustainable Agricultural Systems (SAS) Grant no 2020-69012-31914 and the University Transportation Center for Railway Safety (UTCRS) at the University of Texas Rio Grande Valley through the USDOT UTC Program under Grant No. 69A3552348340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

matrix is higher than d. Thus, PARARAFC2 and MCCA indeed learn different subspaces.

2) As d increases, there appears to be a point after which we observe a few repeated singular values which may indicate some mutual redundancy in the learnt representations. However, even after the d-th position, the singular values are non-negligible, pointing to useful non-noisy signal captured in the stacked representation.

As a result, our preliminary results indicate that in practice MCCA and PARAFAC2 learn different representations.

Given our observation, for a certain multiview dataset, whether a projection or factorization model is preferred for the downstream task is unknown, and forcing the latent representations from both models to be close or identical may make both models lose their unique advantages. Thus, in the remainder of this paper, we will investigate a simple augmentation-based proof-of-concept scheme that allows us to augment one model using the representations of the other, and study the behavior of this approach, towards understanding how to best blend the best of both representations into a superior representation.

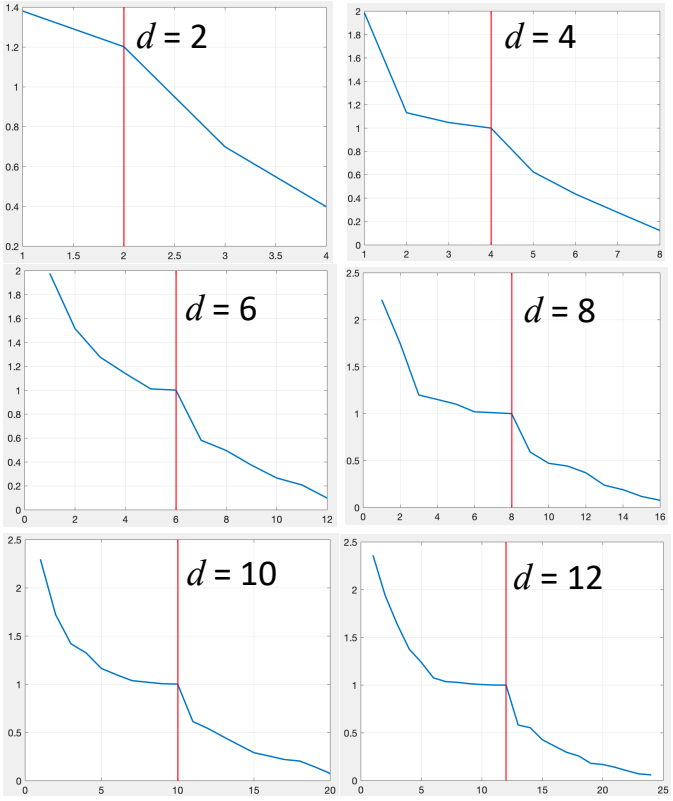

Fig. 1. MCCA and PARAFAC2 learn different subspaces.

II. Proof-of-concept Proposed Method

First, we discuss the two baseline models used, MCCA and PARAFAC2, and subsequently describe our simple proof-of-concept proposed method.

A. Projection model: MCCA

Multiview Canonical Correlation Analysis (MCC) seeks to find a shared latent representation of two or more datasets by enforcing that the projected representations are maximally correlated [12], [13].

Given N views of a dataset $\{\mathbf{X}_n \in \mathbb{R}^{d_n \times M}\}_{n=1}^N$ where M is the number of data samples per view and d_n is the dimention of the n-th view, MCCA computes a latent representation $\mathbf{V} \in \mathbb{R}^{d \times M}$ by solving

$$\min_{\{\mathbf{U}_n\}_{n=1}^N, \mathbf{V}} \quad \sum_{n=1}^{N} \left\| \mathbf{U}_n^\top \bar{\mathbf{X}}_n - \mathbf{V} \right\|_F^2 \tag{1}$$

subject to $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$, where $\bar{\mathbf{X}}_n$ is the centered version of $\mathbf{X}$, and matrices $\{\mathbf{U}_n \in \mathbb{R}^{d_n \times d}\}_{n=1}^N$ are projectors.

B. Factorization model: PARAFAC2

PARAFAC2 [14], [23] is known in the tensor literature for computing a factorization of so-called "irregular" tensors, i.e., datasets which form a multiset $\{\mathbf{X}_n\}_{n=1}^N$ but where one of the dimensions is not consistent across view (thus the "irregular" characterization). Given the irregular tensor or multiset $\{\mathbf{X}_n\}_{n=1}^N$, PARAFAC2 solves the following optimization problem:

$$\begin{aligned} \min_{\{\mathbf{U}_n, \mathbf{S}_n, \mathbf{Q}_n\}_{n=1}^N, \mathbf{H}, \mathbf{G}} \quad & \sum_{n=1}^{N} \|\mathbf{X}_n - \mathbf{U}_n \mathbf{S}_n \mathbf{G}\|_\mathbf{F}^2 \\ \text{s. to} \quad & \mathbf{U}_n = \mathbf{Q}_n \mathbf{H},\ \mathbf{Q}_n^\top \mathbf{Q}_n = \mathbf{I},\ \forall n \end{aligned} \tag{2}$$

where $\mathbf{S}_n \in \mathbb{R}^{R \times R}$ is diagonal, $\mathbf{U}_n \in \mathbb{R}^{d_n \times R}$ is factorized into an orthonormal matrix $\mathbf{Q}_n \in \mathbb{R}^{d_n \times d_n}$ and a matrix $\mathbf{H} \in \mathbb{R}^{d_n \times R}$, and the latent representation $\mathbf{G} \in \mathbb{R}^{R \times M}$.

C. Proposed Augmented Models

In order to leverage the best of both representations, we propose a very simple augmentation-based model which is meant to measure the effect of combining MCCA and PARAFAC2 representations in a manner that does not force them to be equal, which would run the risk of eliminating what is unique with respect to those representations:

Given a multivew dataset $\{\mathbf{X}_n\}_{n=1}^N$, we propose

- **MCCA with an auxiliary view**: we compute PARAFAC2 on $\{\mathbf{X}_n\}_{n=1}^N$ and we use latent representation $\mathbf{G}$ as a new $N+1$ view of that dataset. Given the new augmented dataset, we compute its MCCA representation.
- **PARAFAC2 with an auxiliary view**: we compute MCCA on $\{\mathbf{X}_n\}_{n=1}^N$ and we use latent representation $\mathbf{V}$ as a new $N+1$ view of that dataset. Given the new augmented dataset, we compute its PARAFAC2 representation.

For both of the above methods, the latent dimensionality d and R for each step of MCCA and PARAFAC2, respectively, are user-defined parameters.

III. Numerical Tests

The effectiveness of our proposed methods is tested on the K-means clustering task using the following four real-world datasets.

- **Cora Dataset** consisting of $2,708$ nodes, $5,429$ edges, and 7 labels is a citation network for scientific publications [7]. As in [5], we generate five views from the give graph embeddings of the binary Cora graph including Node2Vec embeddings with 32 and 64 dimensions [9], DeepWalk embeddings with 32 and 64 dimensions [22], and Line (Large-scale information network embedding) embedding with 32 dimensions [27].
- **UCI Dataset** includes mfeat-fac, mfeat-fou, mfeat-kar, mfeat-pix, mfeat-zer features of handwritten digit images with dimensions 216, 76, 64, 240, and 47, respectively [3]. We use the seven clusters representing digits 1, 2, 3, 4, 7, 8, and 9 with randomly chosen 100 images' five features/views per cluster/digit.
- **US Highway Rail Road Crossing Accident Dataset**, sourced from the Federal Railroad Administration of the US Department of Transportation, records $239,487$ railroad accidents spanning from 1975 to 2021, which is available at Kaggle [2]. We randomly select 500 accidents that occurred with death and 500 accidents without death involved. We generate two views where one contains environmental features, such as temperature and train speed and other one contains train features, such as the number locomotive units and the number of train cars in the train that is involved in the accident. There are two labels (with and without death) used. Even though there are numerous variables recorded during an accident report, many of which can be predictive of an accident [28], we opted for simplicity at first by including a small subset of numerical variables, in order to be compatible with the standard formulations of MCCA and PARAFAC2, however, we reserve a more complete investigation of accident features in the context of multiview learning for future work.
- **Hyperspectral Imaging Dataset**: We use the `Salinas-A` dataset from the widely used Hyperspectral Remote Sensing Scenes data repository [1]. The original format of the dataset is an 86×83 image over 204 frequency bands, collected over the Salinas Valley, California. Each pixel has a unique label belonging into one of seven classes, depending on the type of vegetation. For the purposes of creating a simple multiview dataset, we flatten the image into a set of pixels, which are our data points, and we divide the 204 frequency bands into two views, the first one containing the first 102 bands and the second the subsequent 102. We randomly select 300 pixels per class and we generate our dataset (we selected 300 due to limitations of the least represented class). Note that in the current formulation, we ignore any spatial information since the pixels are flattened, which understandably does not take full advantage of the information available. This is done for simplicity and compatibility with the vanilla MCCA and PARAFAC2 models. However, as previous work on such data has shown [21], there is important structure that can be captured among the pixels, and we reserve introducing this into our investigation as future work.

In Figures 2, 3, 4, 5, 6, 7, 8, and 9 we report the average K-means (with the true K) clustering accuracy as well as the standard derivation of 10 Monte Carlo tests of MCCA, PARAFAC2, MCCA with an auxiliary view from PARAFAC2 with the best R among a few candidates, PARAFAC2 with an auxiliary view from MCCA with the best d among a few candidates on the four datasets w.r.t. d or R. Clearly, adding the shared representation of multiple views extracted from PARAFAC2 to MCCA as an auxiliary view increases the clustering accuracy of MCCA compared against the prformance of MCCA on the raw views. Similary statement holds true for MCCA to PARAFAC2.

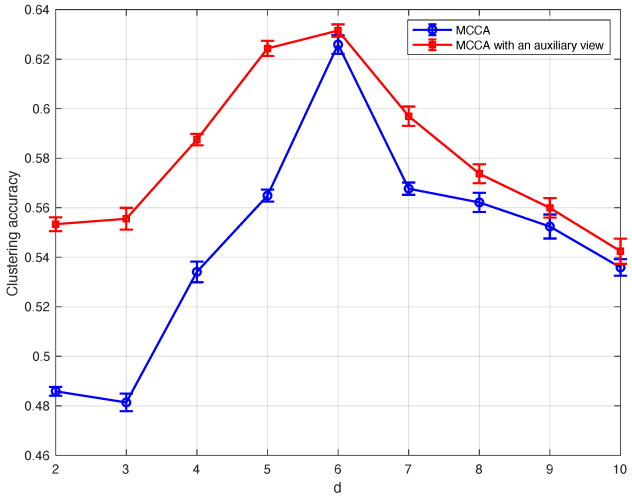

Fig. 2. Clustering performance on Cora dataset.

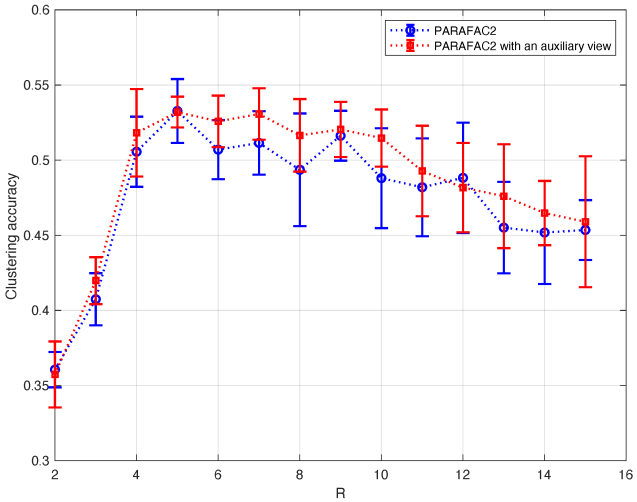

Fig. 3. Clustering performance on Cora dataset.

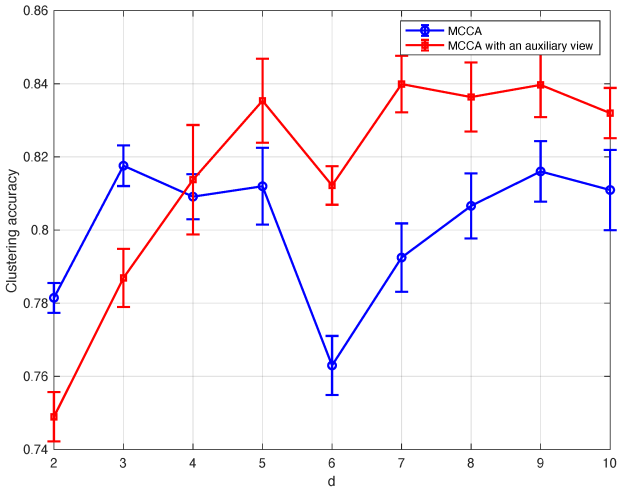

Fig. 4. Clustering performance on UCI dataset.

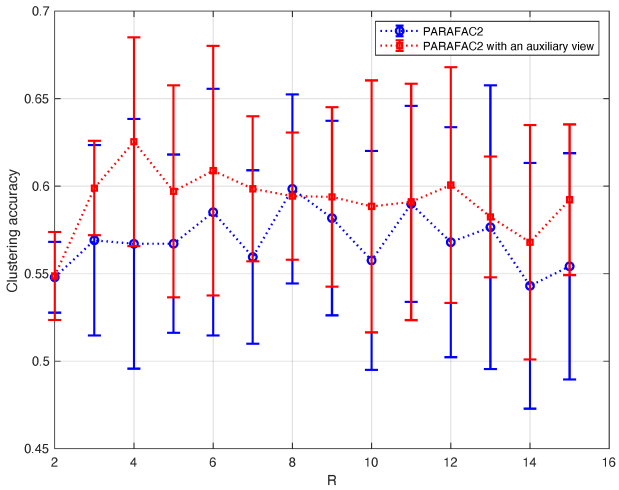

Fig. 5. Clustering performance on UCI dataset.

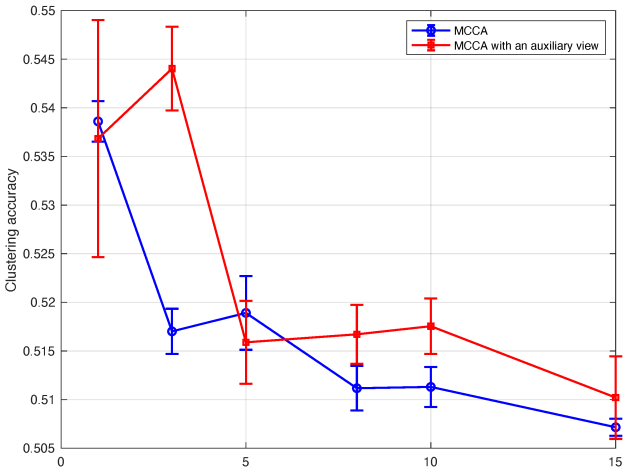

Fig. 6. Clustering performance on Rail Road Crossing Accident dataset.

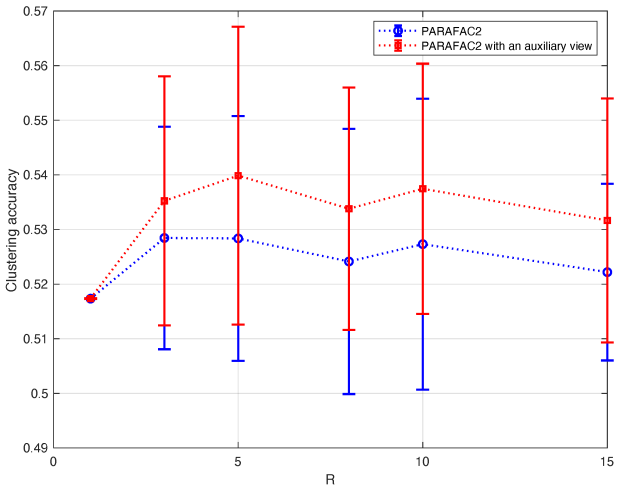

Fig. 7. Clustering performance on Rail Road Crossing Accident dataset.

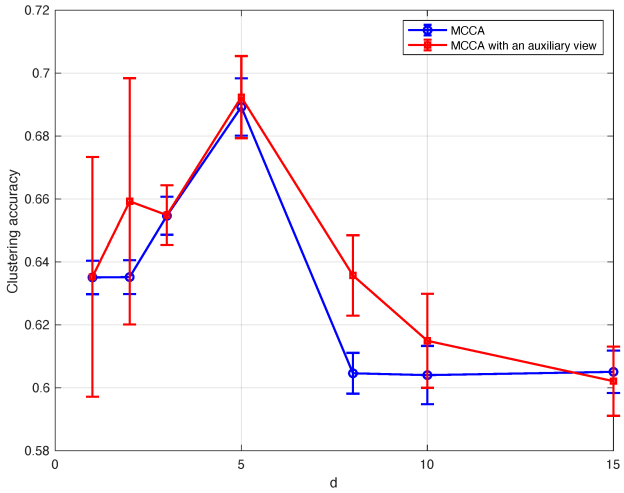

Fig. 8. Clustering performance on hyperspectral imaging dataset.

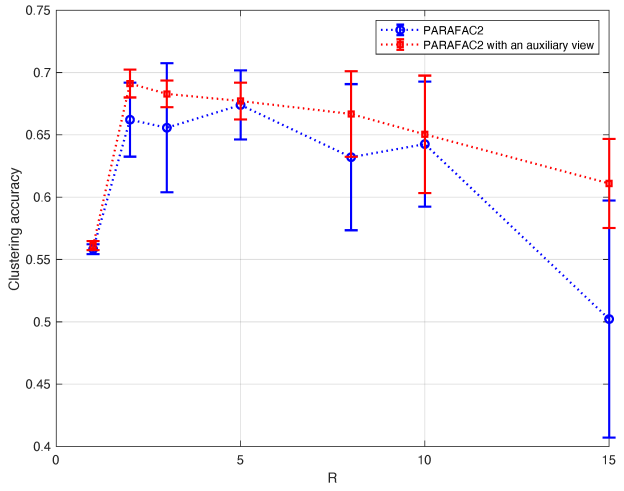

Fig. 9. Clustering performance on hyperspectral imaging dataset.

IV. Conclusions

In this work we study the following fundamental question: Do projection-based MVL methods, such as MCCA, learn different latent representations than their factorization-based counterparts, such as PARAFAC2? Our preliminary results indicate that in practice the two types of approaches learn different representations. Subsequently, we pose the following question: How can we best combine the projection-based and factorization-based representations in a way that achieves superior performance? We propose a simple augmentation-based proof-of-concept scheme which demonstrates that in a number of real multiview datasets even such simple means of combining representations can yield performance improvements. Even though our results are encouraging, there are a lot of important and interesting research questions that lie ahead. We, thus, view this paper as small first step towards investigating the best of both worlds, projection and factorization, in learning multiview data representations.

References

[1] Hyperspectral remote sensing scenes. https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

[2] Us highway rail road crossing accident dataset. https://www.kaggle.com/datasets/yogidsba/us-highway-railgrade-crossing-accident.

[3] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

[4] Bokai Cao, Chun-Ta Lu, Xiaokai Wei, S Yu Philip, and Alex D Leow. Semi-supervised tensor factorization for brain network analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 17–32. Springer, 2016.

[5] Jia Chen, Dalia Orozco, Lizeth Figueroa, and Evangelos Papalexakis. Unsupervised multiview embedding of node embeddings. In *Asilomar Conference on Signals, Systems, and Computers*, 2022.

[6] Jia Chen and Ioannis D Schizas. Distributed information-based clustering of heterogeneous sensor data. *Signal Processing*, 126:35–51, 2016.

[7] Lise Getoor. Link-based classification. In *Advanced methods for knowledge discovery from complex data*, pages 189–207. Springer, 2005.

[8] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[9] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[10] Ekta Gujral and Evangelos E Papalexakis. Smacd: semi-supervised multi-aspect community detection. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 702–710. SIAM, 2018.

[11] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[12] Paul Horst. *Generalized canonical correlations and their application to experimental data*. Number 14. Journal of clinical psychology, 1961.

[13] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[14] Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. Parafac2-part i. a direct fitting algorithm for the parafac2 model. volume 13, pages 275–294, 1999.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[17] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260. SIAM, 2013.

[18] Jing Liu, Yu Jiang, Zechao Li, Zhi-Hua Zhou, and Hanqing Lu. Partially shared latent factor learning with multiview data. *IEEE transactions on neural networks and learning systems*, 26(6):1233–1246, 2014.

[19] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[20] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 809–816, 2011.

[21] Ravdeep S Pasricha, Pravallika Devineni, Evangelos E Papalexakis, and Ramakrishnan Kannan. Tensorized feature spaces for feature explosion. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6298–6304. IEEE, 2021.

[22] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[23] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. Spartan: Scalable parafac2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2017.

[24] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. volume 65, pages 3551–3582. IEEE, 2017.

[25] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7(Jul):1531–1565, 2006.

[26] Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23(7-8):2031–2038, 2013.

[27] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *WWW*. ACM, 2015.

[28] Ethan Villalobos, Constantine Tarawneh, Jia Chen, Evangelos E Papalexakis, and Ping Xu. Kernel ridge regression in predicting railway crossing accidents. In *ASME/IEEE Joint Rail Conference*, volume 87776, page V001T05A013. American Society of Mechanical Engineers, 2024.

[29] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092, 2015.

[30] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[31] Wei Zhao, Cai Xu, Ziyu Guan, and Ying Liu. Multiview concept learning via deep matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[32] Linlin Zong, Xianchao Zhang, Long Zhao, Hong Yu, and Qianli Zhao. Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks*, 88:74–89, 2017.