Towards Aligned Canonical Correlation Analysis: Preliminary Formulation and Proof-of-Concept Results

Biqian Cheng
UC Riverside
Riverside, CA, USA
bchen158@ucr.edu

Evangelos E. Papalexakis
UC Riverside
Riverside, CA, USA
epapalex@cs.ucr.edu

Jia Chen
UC Riverside
Riverside, CA, USA
jiac@ucr.edu

ABSTRACT

Canonical Correlation Analysis (CCA) has been a tried and tested analytical model which seeks to jointly embed two or more views of the data in a maximally correlated latent space. In addition to a powerful data mining tool, CCA has many implications for cutting-edge self-supervised representation learning approaches, as one can cast a number of recent approaches as variants of that model. In this work we explore what happens when a fundamental assumption of that model breaks: what if the alignment across different views of the data is unknown? Typically, we would first attempt to align the different views, and subsequently apply CCA or any other model in order to embed the data views. Can we do better if we align and embed at the same time? In this work, we seek to jointly solve the alignment and embedding by formulating and solving both problems under the same umbrella of Aligned Canonical Correlation Analysis (ACCA). We present a preliminary formulation and alternating optimization algorithm and proof-of-concept results.

KEYWORDS

Canonical Correlation Analysis, Alignment, Matching, Data Integration

1 INTRODUCTION

Canonical Correlation Analysis (CCA) [Harold 1936; Kettenring 1971] is a classical model which, given two different views of the same set of entities, e.g., two different bipartite graphs of (user, product) and (user, video) interactions or different feature representations for those entities in general, seeks to project those entities (users) in a low-dimensional space where the different projected views are maximally correlated. Essentially, CCA can jointly embed heterogeneous datasets in a common low-dimensional space, as it can be extended to more than two views [Chen et al. 2019, 2022]. Even though CCA has been in and out of the spotlight for

many decades and has been around for quite a long time, it is still extremely relevant, not only as a standalone data mining tool, but also due to the fact that cutting-edge self-supervised representation learning techniques, such as Barlow Twins [Balestriero et al. 2023; Bielak et al. 2022; Zbontar et al. 2021; Zhang et al. 2021], can be essentially seen as variations of CCA, where the goal is to embed two views of the data (the original view and the augmentation) in a latent space where correlation is maximized.

Traditionally, in CCA-style analysis, we assume that entities across views have one-to-one correspondence across the two or more views of the data, and there is a wealth of algorithms that study different formulations for solving the problem of projecting those views in that desired maximally correlated space, both linearly and non-linearly [Andrew et al. 2013]. What if, however, this one-to-one correspondence is unknown? In this case, we are faced with two problems: (1) entity alignment and (2) CCA embedding. Motivated by recent results [Wu et al. 2022] in the related problem of misaligned joint tensor factorization, it turns out that formulating and solving the alignment and embedding problems jointly yields better results than solving each problem separately in multiple steps, as it appears that the two sub-problems work synergistically to produce better quality alignment and embeddings. In this work, we explore this type of formulation for CCA and we propose a new formulation, the Aligned Canonical Correlation Analysis (ACCA), where we seek to jointly compute the alignment and the embedding space.

The closest formulation to our proposed model is found in [Sahbi 2018] where the author is considering linear transformation of the two views in CCA, however, is not seeking to recover the precise alignment matrix as our formulation does. In our on-going work we will consider scenarios where we can fairly compare the two formulations and understand pros and cons for either one.

The list of contributions in this preliminary work are:

- **Novel Formulation**: We propose the Aligned Canonical Correlation Analysis (ACCA) model, which seeks to jointly identify the best entity alingment and latent embedding for the dataset views.
- **Proof of Concept**: We derive an Alternating Optimization algorithm and show preliminary results for solving the problem, demonstrating the feasibility of our effort.

2 BACKGROUND

Canonical correlation analysis (CCA) is a powerful tool to learn the shared latent components of two datasets by projecting them to the same space and enforcing the similarity of the projected data. Given two centered and aligned datasets $\mathbf{X} \in \mathbb{R}^{D_x \times N}$ and

$\mathbf{Y} \in \mathbb{R}^{D_y \times N}$ where N is the number of samples, D_x and D_y represent the dimensions of $\mathbf{X}$ and $\mathbf{Y}$, respectively, one popular CCA formulation is seeking for the two projection matrices $\mathbf{U} \in \mathbb{R}^{d \times D_x}$ and $\mathbf{V} \in \mathbb{R}^{d \times D_y}$ with $d \ll \min(D_x, D_y)$, and shared representation/embedding $\mathbf{S} \in \mathbb{R}^{d \times N}$ by solving the following problem

$$\min_{\mathbf{U},\mathbf{V},\mathbf{S}} ||\mathbf{UX} - \mathbf{S}||_F^2 + ||\mathbf{VY} - \mathbf{S}||_F^2 \quad (1)$$

under the constraint that $\mathbf{SS}^\top = \mathbf{I}$ which avoids the trivial solution, i.e., $\mathbf{U} = \mathbf{0}$, $\mathbf{V} = \mathbf{0}$, and $\mathbf{S} = \mathbf{0}$, and ensures the d latent components assembled in the rows of $\mathbf{S}$ are uncorrelated to each other. Here, the symbols $\top$ and $\|\cdot\|_F$ respectively stand for matrix transpose and Frobenius norm operators, and $\mathbf{I}$ is identity matrix with the suitable size. The minimization problem in Eq. (1) admits global optimal solution: the rows of $\mathbf{S}$ are the d eigenvectors corresponding to the top-d eigenvalues of $\mathbf{X}^\top(\mathbf{XX}^\top)^{-1}\mathbf{X} + \mathbf{Y}^\top(\mathbf{YY}^\top)^{-1}\mathbf{Y}$ with $(\cdot)^{-1}$ denoting the matrix inverse operator, $\mathbf{U} = \mathbf{SX}^\top(\mathbf{XX}^\top)^{-1}$, and $\mathbf{V} = \mathbf{SY}^\top(\mathbf{YY}^\top)^{-1}$, e.g., [Harold 1936].

3 PROPOSED METHOD

The traditional CCA formulations require the entities/samples from both $\mathbf{X}$ and $\mathbf{Y}$ to be aligned, i.e., the i-th columns of $\mathbf{X}$ and $\mathbf{Y}$ correspond to the two views/observations of the same latent data sample which is the groundtruth of the i-th column of $\mathbf{S}$. However, if such entity alignment is imperfect, CCA is not able to learn the meaningful latent representations shared by two datasets. Toward this end, we propose a novel model, namely *aligned canonical correlation analysis* (ACCA), to jointly learn the latent representations of two views and recover the entity alignment between the two views.

3.1 Proposed Formulation for ACCA

Consider two centered datasets $\mathbf{X} \in \mathbb{R}^{D_x \times N}$ and $\mathbf{Y} \in \mathbb{R}^{D_y \times N}$ are two views in one dataset, and the alignment between the columns of the two views, denoted as $\bar{\mathbf{P}} \in \mathbf{R}^{N \times N}$, is unknown. Our goal is to learn the latent component representation $\mathbf{S}$ and predict the alignment matrix $\bar{\mathbf{P}}$ iteratively. Let's denote the estimation of $\bar{\mathbf{P}}$ as $\mathbf{P} \in \mathbf{R}^{N \times N}$. Theoretically, $\mathbf{P}$ should be a (binary) permutation matrix, and the sum of row or column is one, which shows that $\mathbf{P}$ is an orthogonal matrix. Mathematically, we will minimize $||\mathbf{UX} - \mathbf{S}||_F^2 + ||\mathbf{VYP} - \mathbf{S}||_F^2$ under the constraints of $\mathbf{P}$ as well as the constraints from CCA, i.e., $\mathbf{SS}^\top = \mathbf{I}$. To address the computational limitation in such optimization problem, we define a list of constraints to describe different aspects of a permutation matrix instead of enforcing it to be one, for a tractable optimization solution. By relaxing the constraints on $\mathbf{P}$, the optimization formulation of our proposed ACCA is shown as:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{S},\mathbf{P}} ||\mathbf{UX} - \mathbf{S}||_F^2 + ||\mathbf{VYP} - \mathbf{S}||_F^2 + \gamma_1||\mathbf{PP}^\top - \mathbf{I}||_F^2 + \gamma_2||\mathbf{P}^\top\mathbf{P} - \mathbf{I}||_F^2$$

$$\quad (2)$$

$$\text{S. T.} \quad \mathbf{SS}^\top = \mathbf{I}, \text{(uncorrelatedness)} \quad (3)$$

$$0 \le p_{i,j} \le 1, \forall i, j, \text{(nonnegativity)} \quad (4)$$

$$\sum_{j=1}^{N} p_{i,j} = 1, \forall i, \text{(row-wise sum)} \quad (5)$$

$$H(\mathbf{p}_i) \le \lambda, \forall i(\text{ entropy}) \quad (6)$$

Algorithm 1: Aligned Canonical Correlation Analysis

1: **Input:** centered datasets $\mathbf{X}$ and $\mathbf{Y}$; dimension of the latent representation d; hyperparameters γ_1, γ_2, and λ; and initialization of $\mathbf{P}$.
2: **Repeat**
   Update $\mathbf{S}$: the rows of $\mathbf{S}$ are the d eigenvectors corresponding to the top-d eigenvalues of
   $\mathbf{X}^\top(\mathbf{XX}^\top)^{-1}\mathbf{X} + (\mathbf{YP})^\top(\mathbf{YPP}^\top\mathbf{Y}^\top)^{-1}\mathbf{YP}$.
   Update $\mathbf{U}$: $\mathbf{U} = \mathbf{SX}^\top(\mathbf{XX}^\top)^{-1}$.
   Update $\mathbf{V}$: $\mathbf{V} = \mathbf{S}(\mathbf{YP})^\top(\mathbf{YPP}^\top\mathbf{Y}^\top)^{-1}$.
   Update $\mathbf{P}$ using *scipy.optimize.minimize* solver.
3: **Until** the objective Eq. (2) is below a threshold or the number of iterations is beyond another threshold.
4: **Output:** $\mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{P}$.

---

where $p_{i,j}$ is the (i, j)-th entry of $\mathbf{P}$, $\mathbf{p}_i$ is the i-th row of $\mathbf{P}$, $H(\mathbf{p}_i)$ is the entropy of $\mathbf{p}_i$ by viewing the N entries of $\mathbf{p}_i$ as a discrete probability distribution, and the hyperparamters γ_1, γ_2, and λ are nonnegative. Enforcing the low entropy of $\mathbf{p}_i$ guarantees that the distribution is closed to a deterministic distribution, and the second and third terms in Eq. (2) will promote the orthogonality of $\mathbf{P}$

3.2 Alternating Optimization for ACCA

To solve the ACCA formulation, as the solution of CCA is dependent on the estimation of permutation matrix, and vice versa, we adopt alternating optimization method, shown in Algorithm 1.

4 EXPERIMENTAL EVALUATION

To validate the effectiveness of our proposed model ACCA, we will generate synthetic data with groundtruth $\mathbf{P}$ and investigate the performance of estimated $\mathbf{P}$ in terms of the matching accuracy between the entities in $\mathbf{X}$ and $\mathbf{Y}$. In all numerical tests, we set the hyperparameters γ_1 and γ_2 to be 0.0001. The initial $\mathbf{P}$ is obtained by solving the optimal matching directly using $\mathbf{X}$ and $\mathbf{Y}$ without considering the canonical correlation between the two datasets, i.e., solving the following minimization problem

$$\min_{\mathbf{P}} ||\mathbf{X} - \mathbf{YP}||_F^2 + \gamma_1||\mathbf{PP}^\top - \mathbf{I}||_F^2 + \gamma_2||\mathbf{P}^\top\mathbf{P} - \mathbf{I}||_F^2 \quad (7)$$

under the constraints specified in Eqs.(4), (5), and (6). We use the *scipy.optimize.minimize* solver to find the optimal $\mathbf{P}$.

4.1 Synthetic Data Generation

We first generate the groundtruth latent representation of the two datasets, namely $\mathbf{Z} \in \mathbb{R}^{\bar{d} \times N}$, where the columns of $\mathbf{Z}$ are N i.i.d. samples drawn from multivariate normal distribution with zero mean and identity covariance of size $\bar{d} \times \bar{d}$. Next, two aligned datasets $\mathbf{X}$ and $\bar{\mathbf{Y}} \in \mathbb{R}^{D_y \times N}$ are generated from their shared latent representation $\mathbf{Z}$ through two independent random projections: $\mathbf{X} = \mathbf{WZ}$ and $\mathbf{Y} = \mathbf{QZ}$ where $\mathbf{W} \in \mathbb{R}^{D_x \times \bar{d}}$ and $\mathbf{Q} \in \mathbb{R}^{D_y \times \bar{d}}$. For each experiment, the groundtruth $\bar{\mathbf{P}}$ is a random permutation matrix with only one entry in each row and column to be 1 and the rest to be 0s. Next, we have two *unaligned* datasets: $\mathbf{X}$ and $\mathbf{Y} = \bar{\mathbf{Y}}\bar{\mathbf{P}}$. The involved parameters are set as follows: $N = 20$, $\bar{d} = 2$, $d = 7$, $D_x = 15$, and $D_y = 10$.
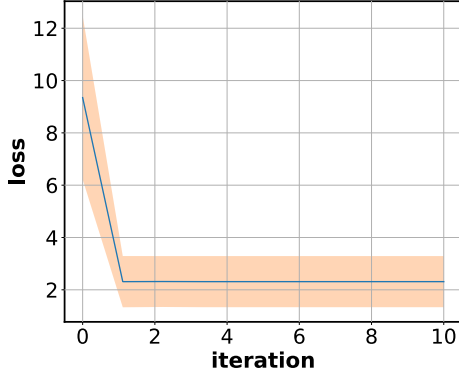
4.2 Experimental Results

Figure 1: Loss as a function of iterations

After setting the entropy upper bound hyperparameter λ to be 0.1, we run 10 times of Monte Carlo experiments and report the loss of Eq. (2) for each iteration in Figure 1. The curve in Figure 1 represents the average loss per iteration and the width of the shade stands for the standard derivation of the loss. Clearly, our proposed ACCA converges to a stable point using the generated synthetic data.
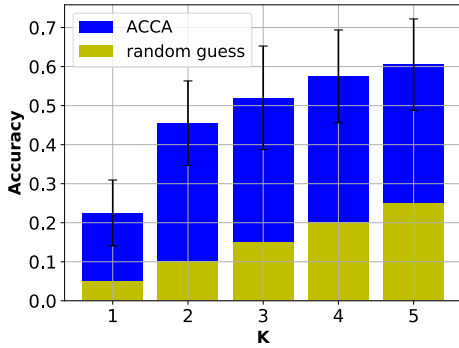

Figure 2: Top-k Accuracy of ACCA and Random guess

In Figure 2, we report the top-k matching accuracy with mean and standard deviation, defined as the percentage of rows in the estimated permutation $\mathbf{P}$ whose top k entries' index set includes the nonzero entry index of the true permutation $\bar{\mathbf{P}}$, with $k = 1, 2, 3, 4$, and 5, in comparison with such accuracy from random guess which is k/N. According to our experimental records as shown in figure 2, it's obvious that our ACCA framework has significantly better performance in predicting the potential alignment between two datasets, than that obtained from the random guess.

Next, we visualize the alignment performance with respect to different values of the hyperparameter λ in Figure 3 where we plot the real permutation matrix $\bar{\mathbf{P}}$ and the estimated $\mathbf{P}$ as gray-scale images with darker grid blocks representing higher values of the

corresponding entries of $\bar{\mathbf{P}}$ or $\mathbf{P}$. As uniform distribution leads to the highest entropy, λ can not exceed $log(N)$ ($=N \times 1/N \times log(1/N)$). With λ increasing, more nonzero entries are showing up in $\mathbf{P}$ as expected. With proper setup of entropy bound hyperparameter, the performance of ACCA will be further improved, with the comparison of prediction accuracies related to different entropy cases in Figure 3.

5 CONCLUSION & FUTURE WORK

In this preliminary work we investigated the joint CCA-style embedding of multiview data and the simultaneous alignment of the embedded entities, by breaking the traditional assumption in CCA that predicates a known one-to-one matching across views. We proposed an initial formulation for Aligned Canonical Correlation Analysis (ACCA) and derived an alternating optimization algorithm that produces proof-of-concept results for the viability of this formulation. However, there is still a lot of work to be done, and we hope that our preliminary results can serve as a stepping stone to further research in this direction.

In our on-going and future work we will investigate variations of the formulation and improvements of the optimization scheme, especially as it pertains to solving for the alignment matrix, which, even though has been radically simplified compared to solving for a permutation matrix, is still a major challenge both in terms of scalability as well as in terms of finding a precise alignment matrix. Furthermore, we would like to study the alignment matrix as a graph and introduce graph-based constraints which may further improve optimization. Finally, we will investigate connections between our proposed Aligned Canonical Correlation Analysis model and self-supervised representation learning models.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs. *Knowledge-Based Systems*, 256:109631, 2022.

Jia Chen, Gang Wang, and Georgios B Giannakis. Graph multiview canonical correlation analysis. *IEEE Transactions on Signal Processing*, 67(11):2826–2838, 2019.

Jia Chen, Lizeth Figueroa, Dalia Orozco, and Evangelos E Papalexakis. Unsupervised multiview embedding of node embeddings. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2022.

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

Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3): 433–451, 1971.

(a) True **P**   (b) Entropy = 0.1; top-3 acc.: 0.519   (c) Entropy = 0.5; top-3 acc.: 0.59   (d) Entropy = 1; top-3 acc.: 0.575   (e) Entropy = 2; top-3 acc.: 0.31
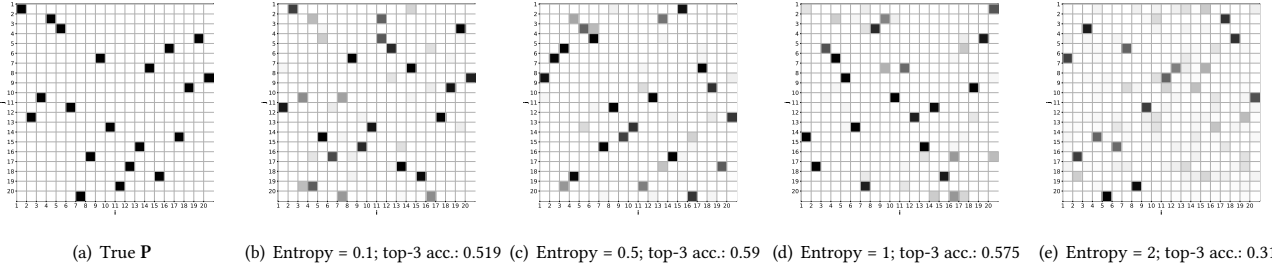
Figure 3: Estimated alignment matrix for different Entropy bounds.

Hichem Sahbi. Learning cca representations for misaligned data. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

Yunshu Wu, Uday Singh Saini, Jia Chen, and Evangelos E Papalexakis. Tenalign: Joint tensor alignment and coupled factorization. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 568–577. IEEE, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S Yu. From canonical correlation analysis to self-supervised graph neural networks. *Advances in Neural Information Processing Systems*, 34:76–89, 2021.