Vol.2019-AVM-107 No.22
2019/12/6

エントロピー正則化Wasserstein距離に基づくマルチビューWasserstein判別法

笠井 裕之[1]

Abstract：Multi-view data analysis has recently garnered increasing attention because multi-view data frequently appear in real-world applications, which are collected or taken from many sources or captured using various sensors. A simple and popular promising approach is to learn a latent subspace shared by multi-view data. Nevertheless, because one sample lies in heterogeneous types of structures, many existing multi-view data analyses show that discrepancies in within-class data across multiple views have a larger value than discrepancies within the same view from different views. To evaluate this discrepancy, this paper presents a proposal of a multi-view Wasserstein discriminant analysis, designated as MvWDA, which exploits a recently developed optimal transport theory.

Multi-view Wasserstein discriminant analysis with entropic regularized Wasserstein distance

1. Introduction

Many real-world applications such as image classification, item recommendation, web page link analysis, and bioinformatics analysis usually exhibit heterogeneous features from *multiple views*. For example, each web page includes two views of text and images and multiple labels such as sports and entertainment. Moreover, each image has multiple features such as frequency features: wavelet coefficients and color histograms. One category of successful techniques to handle multi-view data is *multi-view learning* which includes techniques to learn a *shared subspace* across multi-view data [1], [2], [3], [4], [5]. Many algorithms in this category originate from single-view linear discriminant analyses such as Fisher linear discriminant analysis (LDA or FDA) [6]. They include, for example, canonical correlation analysis (CCA) [7], [8], partial least squares (PLS) [9], bilinear model (BLM) [10], generalized multi-view analysis (GMA) [11], multi-view discriminant analysis (MvDA) [12], and standard linear multi-view discriminant analysis (S-LMvDa) [13]. Also, MvHE has been proposed to address cases in which the multi-view data are sampled from nonlinear manifolds or where they are adversely affected by heavy outliers [14]. However, because one sample lies in different and heterogeneous types of structures, many existing multi-view data analysis show that discrepancies in within-class data across multiple views are greater than discrepancies within the same view from different classes. To this end, building on recently proposed Wasserstein discriminant analysis (WDA) [15], the study described herein presents a proposal of a multi-view Wasserstein discriminant analysis, designated as MvWDA. The main contribution is exploitation of the recently developed *optimal transport* theory [16], [17] to evaluate discrepancies across multi-view data. It is noteworthy that the fundamental characteristics of *optimal transport matrix* amplify small discrepancies, which is necessary to evaluate the discrepancy within the same class and views. Numerical evaluations using several real-world multi-view datasets demonstrate the effectiveness of the proposed MvWDA.

2. Linear subspace discriminant analysis and multi-view extensions

This section presents explanation of linear subspace discriminant analysis and its multi-view extensions. For this purpose, some notations are summarized before the details. We denote scalars with lower-case letters $(a, b, \ldots)$, vectors as bold lower-case letters $(\boldsymbol{a}, \boldsymbol{b}, \ldots)$, and matrices as bold-face capitals $(\mathbf{A}, \mathbf{B}, \ldots)$. $\mathbf{1}_d$ is used for the d-dimensional vector of ones. Here, n, d, C, and V re-

[1] 早稲田大学基幹理工学部情報通信学科

spectively represent the number of sample data, data dimension, classes, and views. The sample data matrix is denoted as $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$; also, $\mathbf{X}$ is assumed to be centralized. The dimensions and the number of sample data of the c ($\in [C]$)-th class in the v ($\in [V]$)-th view, $\mathbf{X}^{v,c} \in \mathbb{R}^{d_v \times n_{v,c}}$, are denoted respectively as d_v and $n_{v,c}$. In addition, $\mathbf{W}_v \in \mathbb{R}^{d_v \times p}$ represents the *projection matrix* onto the p-dimensional subspace of the d_v dimensional data of the v-th view. We also define $\langle \mathbf{X}, \mathbf{Z} \rangle := \mathrm{Tr}(\mathbf{X}^T \mathbf{Z})$.

2.1 Linear subspace discriminant analysis (LDA)

A representative algorithm is (Fisher) linear discriminant analysis (LDA) [6], which maximizes the discrepancy between different classes while minimizing that within the same classes, where evaluations are performed on a *projected space* by $\mathbf{W}$. Therefore, LDA maximizes the ratio of the *between-class scatter (cross-covariance) matrix* $\mathbf{P}$ and the *within-class scatter matrix* $\mathbf{Q}$. Concretely, denoting the average of the sample data of the c-th class, $\mathbf{X}^c$, as $\boldsymbol{\mu}^c$, and the average of the sample data of the entire the classes, $\mathbf{X}$, as $\boldsymbol{\mu}$, the maximization problem is defined as

$$\max_{\mathbf{W} \in \mathbb{R}^{d \times p}} \frac{\mathrm{Tr}(\mathbf{W}^T \mathbf{P} \mathbf{W})}{\mathrm{Tr}(\mathbf{W}^T \mathbf{Q} \mathbf{W})},$$

where $\mathbf{P} = \sum_{c=1}^{C} n_c (\boldsymbol{\mu}^c - \boldsymbol{\mu})(\boldsymbol{\mu}^c - \boldsymbol{\mu})^T$ and $\mathbf{Q} = \sum_{c=1}^{C} \sum_{i=1}^{n_c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T$.

2.2 Multi-view extension of LDA

Canonical correlation analysis (CCA) [7], [8]: CCA maximizes the correlation between $\mathbf{X}_1 \in \mathbb{R}^{d_1 \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{d_2 \times n}$ as $\max_{\mathbf{W}_1, \mathbf{W}_2} \frac{\mathbf{W}_1^T \boldsymbol{\Sigma}_{12} \mathbf{W}_2}{\sqrt{\mathbf{W}_1^T \boldsymbol{\Sigma}_{11} \mathbf{W}_1} \cdot \sqrt{\mathbf{W}_2^T \boldsymbol{\Sigma}_{22} \mathbf{W}_2}}$ for the case of $V = 2$, where $\boldsymbol{\Sigma}_{st} = \frac{1}{n} \mathbf{X}^s (\mathbf{X}^t)^T$. CCA requires pairwise evaluations when $V > 2$ and requires that the number of samples of the two views be the same. To alleviate that difficulty, multi-view CCA (MCCA) algorithms have been proposed [18], [19].

Along another line of research, GMA and MvLDA have been proposed for general-purpose discriminant analysis ($V \geq 3$). It is noteworthy that, in the following approaches, the *concatenated* projection matrix to be calculated is denoted as $\mathbf{W} = [\mathbf{W}_1; \ldots; \mathbf{W}_V] \in \mathbb{R}^{d \times p}$, where $d = \sum_{v=1}^{V} d_v$.

Generalized multi-view analysis (GMA) [11]: GMA is a unified framework that includes various dimension-reduction algorithms. It considers maximization of discriminant information within a single view, but does not consider that between multiple views. The difficulty is maximization with respect to $\mathbf{W} \in \mathbb{R}^{d \times p}$, under $\sum_{v=1}^{V} \mathbf{W}_v^T \mathbf{Q}_v \mathbf{W}_v = \mathbf{I}$ as $\mathrm{Tr}(\sum_s^V \sum_{t>s}^V \lambda_{st} \mathbf{W}_s^T \mathbf{X}^s (\mathbf{X}^t)^T \mathbf{W}^t + \sum_{s=1}^{V} \mu_s (\mathbf{W}^s)^T \mathbf{P}_s \mathbf{W}_s)$.

Multi-view LDA (MvLDA) [12], [13]: MvLDA is regarded as a straightforward extension of LDA in **Section 2.1**. This section presents brief overviews of *three* representative methods. Before that, it is noteworthy that the formulated optimization problem of the three methods is uniformly defined as $\max_{\mathbf{W} \in \mathrm{St}(p,d)} \frac{\mathrm{Tr}(\mathbf{S}_B)}{\mathrm{Tr}(\mathbf{S}_W)}$, where $\mathbf{W}$ is calculated using the *generalized eigenvalue problem*. Also, $\mathrm{St}(p, d)$ is the Stiefel manifold, which is the Riemannian submanifold of orthonormal matrices $\mathcal{M} = \{\mathbf{X} \in \mathbb{R}^{d \times p} : \mathbf{X}\mathbf{X}^T = \mathbf{I}_p\}$. The first method, MvDA, maximizes discrepancies between classes between and within multiple views [12]. The between-class scatter matrix $\mathbf{S}_B$ and the within-class scatter matrix $\mathbf{S}_W$ are defined respectively as

$$\begin{aligned} \mathbf{S}_B &= \sum_{c=1}^{C} n_c (\boldsymbol{m}^c - \boldsymbol{m})(\boldsymbol{m}^c - \boldsymbol{m})^T, \\ \mathbf{S}_W &= \sum_{v=1}^{V} \sum_{c=1}^{C} \sum_{i=1}^{n_{v,c}} (\mathbf{W}_v^T \boldsymbol{x}_i^{v,c} - \boldsymbol{m}^c)(\mathbf{W}_v^T \boldsymbol{x}_i^{v,c} - \boldsymbol{m}^c)^T, \end{aligned} \quad (1)$$

where $\boldsymbol{m}^c$ is the averaged vector of the c-th class in the entire views, i.e., $\boldsymbol{m}^c = \frac{1}{n_c} \sum_{v=1}^{V} \sum_{i=1}^{n_{v,c}} \mathbf{W}_v^T \boldsymbol{x}_i^{v,c}$, and $\boldsymbol{m}$ represents the average of all sample data. It is noteworthy that both are considered on the *projected space*.

The second method, standard linear multi-view discriminant analysis (S-LMvDA), is intended to seek an optimal $\mathbf{W}$ to maximize the distance between two classes [13]. Here, $\mathbf{S}_B$ is defined as

$$\mathbf{S}_B = \sum_{s=1}^{V} \sum_{t=1}^{V} \sum_{k=1}^{C} \sum_{l=1}^{C} \left(\boldsymbol{m}^{s,k} - \boldsymbol{m}^{t,l}\right) \left(\boldsymbol{m}^{s,k} - \boldsymbol{m}^{t,l}\right)^T \quad (2)$$

where $\boldsymbol{m}^{v,c}$ is the averaged vector of the c-th class in the v-th view, which is calculated as $\boldsymbol{m}^{v,c} = \frac{1}{n_{v,c}} \sum_{i=1}^{n_{v,c}} \mathbf{W}_v^T \boldsymbol{x}_i^{v,c}$.

The last method, linear multi-view modular discriminant analysis (L-MvMDA), maximizes the distance between classes in different views [13]. Consequently, $\mathbf{S}_B$ is defined as

$$\mathbf{S}_B = \sum_{s=1}^{V} \sum_{t=1}^{V} \sum_{k=1}^{C} \sum_{l=1}^{C} (\boldsymbol{m}^{s,k} - \boldsymbol{m}^{s,l})(\boldsymbol{m}^{t,k} - \boldsymbol{m}^{t,l})^T. \quad (3)$$

3. Proposed multi-view Wasserstein discriminant analysis

After introducing the Wasserstein discriminant analysis briefly, details of the proposed MvWDA are given.

3.1 Wasserstein discriminant analysis (WDA)

Wasserstein discriminant analysis (WDA) applies discriminant analysis exploiting the Wasserstein distance with an entropic regularizer [15]. Given $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$ which form $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ and $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m] \in \mathbb{R}^{d \times m}$, respectively, when two empirical distributions $\nu = \frac{1}{n} \sum_i \delta_{\boldsymbol{x}_i}$ and $\xi = \frac{1}{m} \sum_j \delta_{\boldsymbol{z}_j}$ are defined, the Wasser-

stein distance with an entropic regularizer is defined as $W_\lambda(\xi,\mu) := W_\lambda(\mathbf{X},\mathbf{Z}) = \langle \mathbf{T}_\lambda, \mathbf{M}_{\mathbf{X},\mathbf{Z}} \rangle$. Defining $\mathbf{M}_{\mathbf{X},\mathbf{Z}} = [\|\boldsymbol{x}_i - \boldsymbol{z}_j\|_2^2]_{ij} \in \mathbb{R}^{n\times m}$, $\mathbf{T}_\lambda$ is obtainable as a solution of a entropy-smoothed optimal transport problem as

$$\mathbf{T}_\lambda = \arg\min_{\mathbf{T}\,\in\,\mathcal{U}_{nm}} \ \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{X},\mathbf{Z}} \rangle - \Omega(\mathbf{T}),$$

where $\Omega(\mathbf{T})$ is defined as a discrete joint probability distribution $\Omega(\mathbf{T}) := -\sum_{ij} t_{ij}\log(t_{ij})$. This minimization problem is solvable efficiently using Sinkhorn's fixed-point iterations [20]. Here, let $\mathcal{U}_{nm}$ be the polytope of $n \times m$ nonnegative matrices such that their row and column marginals are respectively equal to $\mathbf{1}_n/n$ and $\mathbf{1}_m/m$. Then, we have $\mathcal{U}_{nm} := \{\mathbf{T} \in \mathbb{R}_+^{n\times m} : \mathbf{T}\mathbf{1}_m = \mathbf{1}_n/n, \mathbf{T}^T\mathbf{1}_n = \mathbf{1}_m/m\}$. Finally, the problem formulation of WDA is defined as

$$\max_{\mathbf{W}\in \mathrm{St}(p,d)} J(\mathbf{W}, \mathbf{T}(\mathbf{W})) = \frac{\langle \mathbf{W}^T\mathbf{W}, \mathbf{P}\rangle}{\langle \mathbf{W}^T\mathbf{W}, \mathbf{Q}\rangle},$$

where

$$\mathbf{P} = \sum_{k,l>k}\sum_{i,j} [\mathbf{T}_k^l]_{ij}(\boldsymbol{x}_i^k - \boldsymbol{x}_j^l)(\boldsymbol{x}_i^k - \boldsymbol{x}_j^l)^T,$$
$$\mathbf{Q} = \sum_{k}\sum_{i,j} [\mathbf{T}_k^k]_{ij}(\boldsymbol{x}_i^k - \boldsymbol{x}_j^k)(\boldsymbol{x}_i^k - \boldsymbol{x}_j^k)^T,$$

and $\mathbf{T}_k^l = \arg\min_{\mathbf{T}\,\in\,\mathcal{U}_{n_k n_l}} \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{W}\mathbf{X}^k, \mathbf{W}\mathbf{X}^l}\rangle - \Omega(\mathbf{T})$.

It is noteworthy that WDA considers *global* and *local* interplays between classes. It must be emphasized that $\mathbf{T}$ itself amplifies the small errors, which is more necessary for within-class evaluation.

3.2 Proposed multi-view WDA (MvWDA)

This section presents a multi-view extension of Wasserstein discriminant analysis, designated herein as MvWDA. The main motivation is consideration of the optimal transport of the discrepancies between and within multi-view datasets. For this particular purpose, a transport matrix $\mathbf{T}_{s,k}^{t,l}$ from the k-th class of the s-th view to the l-th class of the t-th view is newly introduced. Furthermore, and more importantly, *concatenated* scatter between-matrices and within-matrices across multiple views are newly constructed. Then, designating two such matrices as $\mathbf{G} \in \mathbb{R}^{d\times d}$ and $\mathbf{H} \in \mathbb{R}^{d\times d}$ and following the single-view WDA, we formally define a minimization problem of MvWDA as

$$\max_{\mathbf{W}\in \mathrm{St}(p,d)} \quad J(\mathbf{W}, \mathbf{T}(\mathbf{W})) = \frac{\langle \mathbf{W}^T\mathbf{W}, \mathbf{G}\rangle}{\langle \mathbf{W}^T\mathbf{W}, \mathbf{H}\rangle}, \tag{4}$$
$$\text{s.t.} \quad \mathbf{T}_{s,k}^{t,l} = \underset{\mathbf{T}\,\in\,\mathcal{U}_{n_{s,k}n_{t,l}}}{\arg\min} \ \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{W}_s\mathbf{X}^{s,k}, \mathbf{W}_t\mathbf{X}^{t,l}}\rangle - \Omega(\mathbf{T}),$$

where $\mathbf{G}$ and $\mathbf{H}$ are defined respectively in (5) and (6).

Algorithm 1 MvWDA algorithm

Require: Hyper-parameter λ, # of maximum iterations $T_{\max}$.

```
Initialize W_0.
for t = 1, 2, ..., T_max do
  for s = 1, 2, ..., V do
    for t = s, s+1 ..., V do
      for k = 1, 2, ..., C do
        for l = k, k+1 ..., C do
          Calculate Ψ in (7), Φ in (8) and ϒ in (9).
        end for
      end for
      Update G_st and H_st as in Table 1.
    end for
  end for
  Construct G in (5) and H in (6) .
  Update W_t by Riemannian steepest descent.
end for
```

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \cdots & \cdots & \cdots & \mathbf{G}_{1V} \\ \vdots & \cdots & \mathbf{G}_{st} & \cdots & \vdots \\ \mathbf{G}_{V1} & \cdots & \cdots & \cdots & \mathbf{G}_{VV} \end{bmatrix} \quad (\in \mathbb{R}^{d\times d}), \tag{5}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{11} & \cdots & \cdots & \cdots & \mathbf{H}_{1V} \\ \vdots & \cdots & \mathbf{H}_{st} & \cdots & \vdots \\ \mathbf{H}_{V1} & \cdots & \cdots & \cdots & \mathbf{H}_{VV} \end{bmatrix} \quad (\in \mathbb{R}^{d\times d}). \tag{6}$$

The core contribution is a proposal of new formulations of the elements of $\mathbf{G}_{st} \in \mathbb{R}^{d_s\times d_t}$ and $\mathbf{H}_{st} \in \mathbb{R}^{d_s\times d_t}$ between the s-th and t-th views by exploiting optimal transport matrices. The three proposed new matrices are detailed below.

The first two methods are derived from L-MvMDA in **Section 2.2**. The first approach specifically considers maximization of the distance between difference class centers across different views. Concretely, we address that the first term in (3) represents the difference between the averaged vector of the k-th class in the s-th view, $\boldsymbol{\mu}^{s,k}$, and that of the l-th class in the s-th view, $\boldsymbol{\mu}^{s,l}$. This approach newly considers $\mathbf{T}_{s,k}^{s,l}$ of such averaged vectors. Analogously, the second term is obtained by $\mathbf{T}_{t,k}^{t,l}$. Thus, a new scatter matrix $\boldsymbol{\Psi}_{klij}^{st} \in \mathbb{R}^{d_s\times d_t}$ is created as

$$\boldsymbol{\Psi}_{klij}^{st} = \sqrt{\frac{\sum_{i,j}[\mathbf{T}_{s,k}^{s,l}]_{ij}}{n_{s,k}\cdot n_{s,l}} \frac{\sum_{i,j}[\mathbf{T}_{t,k}^{t,l}]_{ij}}{n_{t,k}\cdot n_{t,l}}} \cdot(\boldsymbol{\mu}^{s,k} - \boldsymbol{\mu}^{s,l})(\boldsymbol{\mu}^{t,k} - \boldsymbol{\mu}^{t,l})^T, \tag{7}$$

where $\boldsymbol{\mu}^{v,c} = \frac{1}{n_{v,c}}\sum_i \boldsymbol{x}_i^{v,c}$. As the second approach, addressing the error between the samples, another new scatter matrix $\boldsymbol{\Phi}_{klij}^{st}$ is obtained as

$$\boldsymbol{\Phi}_{klij}^{st} = \sqrt{\left[\mathbf{T}_{s,k}^{s,l}\right]_{ij}\left[\mathbf{T}_{t,k}^{t,l}\right]_{ij}} \cdot(\boldsymbol{x}_i^{s,k} - \boldsymbol{x}_j^{s,l})(\boldsymbol{x}_i^{t,k} - \boldsymbol{x}_j^{t,l})^T. \tag{8}$$

Table 1 Between-/within-class matrices in MvWDA.

Proposal	$\mathbf{G}_{st}$	$\mathbf{H}_{st}$
MvWDA-A	$\sum_{k,l>k}\sum_{i,j}\mathbf{\Phi}^{st}_{klij}$	$\sum_{k}\sum_{i,j}\mathbf{\Phi}^{st}_{kkij}$
MvWDA-B	$\sum_{k,l>k}\sum_{i,j}\mathbf{\Psi}^{st}_{klij}$	$\sum_{k}\sum_{i,j}\mathbf{\Upsilon}^{st}_{kkij}$
MvWDA-C	$\sum_{k,l>k}\sum_{i,j}\mathbf{\Phi}^{st}_{klij}$	$\sum_{k}\sum_{i,j}\mathbf{\Upsilon}^{st}_{kkij}$
MvWDA-D	$\sum_{k,l>k}\sum_{i,j}\mathbf{\Psi}^{st}_{klij}$	$\sum_{k}\sum_{i,j}\mathbf{\Phi}^{st}_{kkij}$

The final approach is to extend the within-class matrix of $\mathbf{S}_w$ in (1) , which engenders $\mathbf{\Upsilon}^{st}_{kkij}$ defined as

$$\mathbf{\Upsilon}^{st}_{kkij} = \begin{cases} \left[\mathbf{T}^{t,k}_{s,k}\right]_{ij}(\boldsymbol{x}^{s,k}_i - \boldsymbol{x}^{t,k}_j)(\boldsymbol{x}^{s,k}_i - \boldsymbol{x}^{t,k}_j)^T & (s=t), \\ -\dfrac{n_{s,k}\cdot n_{t,k}}{n_{s,k}+n_{t,k}}\boldsymbol{\mu}^{s,k}(\boldsymbol{\mu}^{t,k})^T & (s\neq t). \end{cases} \tag{9}$$

It is noteworthy that an extension of S-LMvDA defined in (2) is not trivial because of the different dimensions of data samples to calculate $\mathbf{T}^{t,l}_{s,k}$. Also, because [13] reports that L-MvMDA achieves experimentally better results than S-LMvDA, this paper addresses the two approaches described above derived from L-MvMDA. Hence, from these three approaches in (7), (8) and (9), $\mathbf{G}_{st}, \mathbf{H}_{st} \in \mathbb{R}^{d_s\times d_t}$ are calculated as presented in **Table 1**.

As pointed out in Theorem 1 of [20], $\mathbf{M}_{\mathbf{X},\mathbf{Z}}$ must be a valid metric. In addition, considering the structural distance, the *square root cosine distance* proposed in [21] is used instead of ℓ_2 distance, which is defined as

$$\mathbf{M}_{\mathbf{X},\mathbf{Z}} = \left[\sqrt{2-2\cos(\boldsymbol{x}_i,\boldsymbol{z}_j)}\right]_{ij} = \left[\left\|\frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_2} - \frac{\boldsymbol{z}_j}{\|\boldsymbol{z}_j\|_2}\right\|_2\right]_{ij}.$$

Finally, for the optimization perspective, because $\mathbf{T}$ depends on $\mathbf{W}$, $\mathbf{W}$ cannot be obtained from the generalized eigenvalue problem. Therefore, we use an alternative optimization algorithm between a manifold optimization on the Stiefel manifold $\mathrm{St}(p,d)$ [22] and $\mathbf{W}$. The overall algorithm of MvWDA is presented in **Algorithm 1**.

4. Conclusion

This paper presented a novel multi-view Wasserstein discriminant analysis, designated as MvWDA. The main contribution is exploitation of a recently developed optimal transport theory to evaluate the discrepancy across multi-view data. The presentation will show some numerical evaluations using several real-world datasets which demonstrate the effectiveness of the proposed MvWDA.

References

[1] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[2] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[3] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, 2014.

[4] J. Zhao, X. Xie, X. X., and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[5] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2019.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, 2009.

[7] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[8] B. Thompson, "Canonical correlation analysis," *Encyclopedia of Statistics in Behavioral Science*, vol. 1, no. 1, pp. 192–196, 2005.

[9] R. Rosipal and N. Kramer, "Overview and recent advances in partial least squares," in *Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection (SLSFS'05)*, 2005.

[10] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[11] A. Sharma, A. Kumar, H. Daume, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[12] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188 – 194, 2016.

[13] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2542–2555, 2016.

[14] J. Xu, S. Yu, X. You, M. Leng, X.-Y. Jing, and C. Chen, "Multiview hybrid embedding: A divide-and-conquer approach," in *IEEE transactions on cybernetics*, 2019.

[15] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy, "Wasserstein discriminant analysis," vol. 107, no. 12, 2018, pp. 1923–1945.

[16] C. Villani, *Optimal transport: Old and new*. Springer, 2008.

[17] G. Peyre and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[18] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Transactions on Image Processing*, vol. 11, no. 3, pp. 293–305, 2002.

[19] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *ACM SiKDD Conference on Knowledge Discovery and Data Mining (KDD201)*, 2010.

[20] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

[21] R. Xu, Y. Yang, N. Otani, and Y. Wu, "Unsupervised cross-lingual transfer of word embedding spaces," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[22] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.