Generalized Canonical Correlation Analysis: A Subspace Intersection Approach

Mikael Sørensen, Charilaos I. Kanatsoulis, *Student Member, IEEE,* and Nicholas D. Sidiropoulos, *Fellow, IEEE*

***Abstract*—Generalized Canonical Correlation Analysis (GCCA) is an important tool that finds numerous applications in data mining, machine learning, and artificial intelligence. It aims at finding 'common' random variables that are strongly correlated across multiple feature representations (views) of the same set of entities. CCA and to a lesser extent GCCA have been studied from the statistical and algorithmic points of view, but not as much from the standpoint of linear algebra. This paper offers a fresh algebraic perspective on GCCA based on a (bi-)linear generative model that naturally captures its essence. It is shown that from a linear algebra point of view, GCCA is tantamount to subspace intersection; and conditions under which the common subspace of the different views is identifiable are provided. A novel GCCA algorithm is proposed based on subspace intersection, which scales up to handle large GCCA tasks. Synthetic as well as real data experiments are provided to showcase the effectiveness of the proposed approach.**

***Index Terms*—Canonical Correlation Analysis, Generalized Canonical Correlation Analysis, Subspace Intersection, Multi-view Learning, Identifiability, Algebraic Algorithm, Common Subspace Analysis.**

I. Introduction

CANONICAL Correlation Analysis (CCA) is a classical statistical tool for two-set / two-view factor analysis [1], [2]. It aims at extracting a common latent structure of a set of entities observed in two different feature domains, which are usually referred as the 'views' of the entities. For example, an English document and its French translation is an entity represented in two different language-views. CCA can be naturally extended to the multi-view case, where more than two views are available for processing. Then it is referred as generalized CCA (GCCA) or multi-view CCA (MCCA) [3]. CCA/GCCA can also be considered as an extension of principal component analysis (PCA) to the case where multiple views of the data are available. On one hand PCA seeks for a feature representation that maximizes the variance explained, thus keeping the strong / *principal* feature components. On the other hand, CCA/GCCA extracts the *common* components between the views and ideally ignores even strong components that are not present in all the views.

(G)CCA is a powerful set of tools with diverse applications in machine learning [4]–[7], data mining [8]–[11], signal processing [12]–[16], biomedical engineering [17]–[21], health care data analytics [22], and genetics [23], [24], among others.

First submission July 29, 2020; revised February 11, 2021. M. Sørensen and N. D. Sidiropoulos are with the Department of ECE, University of Virginia, Charlottesville, VA 22904, USA email: ms8tz@virginia.edu; nikos@virginia.edu; C.I. Kanatsoulis is with the Department of ECE, University of Minnesota, Minneapolis, MN 55455, USA email: kanat003@umn.edu.

In the two view case, CCA can be optimally solved via generalized eigenvalue decomposition [2]. Furthermore, several algorithms exist that solve the CCA problem when big and high dimensional datasets are involved, and eigenvalue solutions are computationally prohibitive, e.g., [25], [26]. The multi-view scenario, on the other hand, is more complicated. There exist a number of different GCCA formulations, e.g., SUMCOR, MAXVAR, SUQUAR, etc; see [3], [27], and the majority of them are not solvable in polynomial time. SUMCOR and MAXVAR are the most popular formulations and various algorithms have been developed for them, e.g., [9]–[11], [14], [22].

Although CCA and GCCA are well-known and broadly-used tools with a long history, there still exist intriguing questions and open challenges related to (G)CCA theory and practice. First, our understanding of CCA/GCCA from an algebraic perspective is limited. The majority of the literature focuses on the statistical interpretation of CCA, e.g., [1], [3], [28], where each view is considered as a set of random vector realizations, and/or on algorithmic aspects. Interpreting (G)CCA from an algebraic viewpoint is important, since in practice the matrix views involved in (G)CCA do not necessarily follow a statistical model. Second, identifiability of CCA/GCCA, i.e., conditions under which the common latent components can be recovered, has only been partially studied. An identifiability condition for CCA was derived in [16], but only for the two view case. Also, identifiability of CCA was established in a statistical sense in [29], albeit under stringent statistical assumptions. Finally, there is limited analysis regarding the effect of multiple views compared to just using two views. Despite the rapid developments in data acquisition and cross-platform data availability, which enable leveraging multiple views of a given set of entities, researchers often work with just two views due to the more complicated nature of GCCA.

A. Organization and contributions of the paper

In this work we give answers to the above research questions. First, we show that from an algebraic point of view, GCCA amounts to *subspace intersection*, i.e., it computes the intersection of the subspaces of the given matrix views. Next, we provide both deterministic and generic conditions under which the common subspace between the views is identifiable. Our conditions show that having access to more views which share a common subspace benefits the identifiability of that subspace. Finally, we propose a simple and effective subspace intersection algorithm for GCCA which works for any number of views greater than or equal to two. The algorithm is algebraic

and it exploits knowledge of the desired rank (useful signal rank, i.e., the dimension of the dominant information-bearing 'signal subspace') of the matrix views. We also develop a large-scale approximation algorithm which works for big and high-dimensional data, both dense and sparse. Extensive simulations with synthetically generated and real datasets showcase the effectiveness of our proposed framework. The contributions of the paper can be summarized as follows:

- A subspace intersection interpretation of CCA and GCCA.
- Deterministic and generic identifiability conditions for GCCA.
- Subspace intersection based algorithms for CCA and GCCA.

The rest of the introduction will present the notation used throughout the paper. In Section II we review CCA and GCCA and present a generative model for GCCA. As our first contribution, in Section III we present a subspace intersection interpretation of CCA and GCCA. As our second contribution, in Section IV we present new identifiability conditions for GCCA. Based on the obtained identifiability conditions for GCCA, in Section V we discuss the benefits of processing more than two views. As our third contribution, in Section VI we present an algebraic framework for GCCA that is scalable to high-dimensional data. In Section VII we report numerical experiments, based on both synthetic and real data, that corroborate the benefits of processing more than two views. Section VIII summarizes our findings and concludes the paper.

B. Notation

The notation used throughout the paper is summarized in Table I.

TABLE I: Overview of notation.

a	$\triangleq$	scalar
$\boldsymbol{a}$	$\triangleq$	vector
$\boldsymbol{A}$	$\triangleq$	matrix
A	$\triangleq$	subspace
$\boldsymbol{a}_r$	$\triangleq$	r-th column of matrix $\boldsymbol{A}$
$\boldsymbol{A}^T$	$\triangleq$	transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^H$	$\triangleq$	conjugate-transpose of matrix $\boldsymbol{A}$
$\|\boldsymbol{A}\|_F$	$\triangleq$	Frobenius norm of matrix $\boldsymbol{A}$
trace$(\boldsymbol{A})$	$\triangleq$	trace of matrix $\boldsymbol{A}$
rank$(\boldsymbol{A})$	$\triangleq$	rank of matrix $\boldsymbol{A}$
range $(\boldsymbol{A})$	$\triangleq$	range of matrix $\boldsymbol{A}$
ker$(\boldsymbol{A})$	$\triangleq$	kernel of matrix $\boldsymbol{A}$
$\otimes$	$\triangleq$	Kronecker product of two matrices
$\oplus$	$\triangleq$	direct sum of two subspaces
dim(A)	$\triangleq$	dimension of subspace A
$\binom{m}{n}$	$\triangleq$	binomial coefficient, i.e., $\binom{m}{n} = \frac{m!}{n!(m-n)!}$

II. Generalized Canonical Correlation Analysis

In Sections II-A and II-B we first review CCA and GCCA. Next, in Section II-C we present a generative model for GCCA that enables us to study GCCA using tools from linear algebra.

A. Review of CCA

In CCA we consider a pair of zero-mean random vectors $\boldsymbol{x}_1 \in \mathbb{C}^{J_1}$ and $\boldsymbol{x}_2 \in \mathbb{C}^{J_2}$. The goal of the simplest version of CCA is to find linear combinations $\boldsymbol{\phi}_1^H \boldsymbol{x}_1$ and $\boldsymbol{\phi}_2^H \boldsymbol{x}_2$ that are maximally correlated, i.e., we seek two nonzero vectors $\boldsymbol{\phi}_1 \in \mathbb{C}^{J_1}$ and $\boldsymbol{\phi}_2 \in \mathbb{C}^{J_2}$ that maximize the absolute value of the cosine angle, also known as the *canonical correlation*:

$$\rho(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \frac{\boldsymbol{\phi}_1^H \mathbb{E}[\boldsymbol{x}_1 \boldsymbol{x}_2^H] \boldsymbol{\phi}_2}{\sqrt{\boldsymbol{\phi}_1^H \mathbb{E}[\boldsymbol{x}_1 \boldsymbol{x}_1^H] \boldsymbol{\phi}_1} \sqrt{\boldsymbol{\phi}_2^H \mathbb{E}[\boldsymbol{x}_2 \boldsymbol{x}_2^H] \boldsymbol{\phi}_2}}, \quad (1)$$

where $\mathbb{E}[\ ]$ denotes expectation and $-1 \leq \rho(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \leq 1$. We say that $\boldsymbol{\phi}_1^H \boldsymbol{x}_1$ and $\boldsymbol{\phi}_2^H \boldsymbol{x}_2$ are *coherent* when $\rho(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \pm 1$. In practice, only realizations of the random vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are observed. Let the rows of the matrices $\boldsymbol{X}_1 \in \mathbb{C}^{I \times J_1}$ and $\boldsymbol{X}_2 \in \mathbb{C}^{I \times J_2}$ correspond to realizations of the random vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. The empirical version of the correlation measure (1) is given by

$$\hat{\rho}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \frac{\boldsymbol{\phi}_1^H \boldsymbol{X}_1^H \boldsymbol{X}_2 \boldsymbol{\phi}_2}{\sqrt{\boldsymbol{\phi}_1^H \boldsymbol{X}_1^H \boldsymbol{X}_1 \boldsymbol{\phi}_1} \sqrt{\boldsymbol{\phi}_2^H \boldsymbol{X}_2^H \boldsymbol{X}_2 \boldsymbol{\phi}_2}}. \quad (2)$$

Observe that, from the Cauchy-Schwartz inequality, $\hat{\rho}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \pm 1$ means that $\boldsymbol{X}_1 \boldsymbol{\phi}_1 \propto \boldsymbol{X}_2 \boldsymbol{\phi}_2$ and consequently range$(\boldsymbol{X}_1 \boldsymbol{\phi}_1)$ = range$(\boldsymbol{X}_2 \boldsymbol{\phi}_2)$. A pair of components $(\boldsymbol{X}_1 \boldsymbol{\phi}_1, \boldsymbol{X}_2 \boldsymbol{\phi}_2)$ is said to be *coherent* if $\boldsymbol{X}_1 \boldsymbol{\phi}_1 \propto \boldsymbol{X}_2 \boldsymbol{\phi}_2$. Assume that we are interested in the R components $(\boldsymbol{X}_1 \boldsymbol{\phi}_{11}, \boldsymbol{X}_2 \boldsymbol{\phi}_{21}), \ldots, (\boldsymbol{X}_1 \boldsymbol{\phi}_{1R}, \boldsymbol{X}_2 \boldsymbol{\phi}_{2R})$ with the largest canonical correlation values, where $\boldsymbol{\phi}_{1r} \in \mathbb{C}^{J_1}$ and $\boldsymbol{\phi}_{2r} \in \mathbb{C}^{J_2}$. The extension of (2) to the case of multiple components yields the CCA formulation [2], [5], [30]:

$$\max_{\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2} \ \text{trace}\left(\boldsymbol{\Phi}_1^H \boldsymbol{X}_1^H \boldsymbol{X}_2 \boldsymbol{\Phi}_2\right) \quad (3a)$$

$$\text{s.t. } \boldsymbol{\Phi}_n^H \boldsymbol{X}_n^H \boldsymbol{X}_n \boldsymbol{\Phi}_n = \boldsymbol{I}_R, \ n \in \{1, 2\}, \quad (3b)$$

where $\boldsymbol{\Phi}_n = [\boldsymbol{\phi}_{n1}, \ldots, \boldsymbol{\phi}_{nR}] \in \mathbb{C}^{J_n \times R}$ has full column rank and $\boldsymbol{I}_R$ is the $R \times R$ identity matrix. Hence, CCA aims to extract the R principal canonical correlation components from the two matrix "views" $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The trace maximization formulation (3) of CCA is equivalent to the following minimization problem

$$\min_{\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2} \|\boldsymbol{X}_1 \boldsymbol{\Phi}_1 - \boldsymbol{X}_2 \boldsymbol{\Phi}_2\|_F \Leftrightarrow \min_{\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2} \left\| [\mathbf{X}_1, -\mathbf{X}_2] \begin{bmatrix} \boldsymbol{\Phi}_1 \\ \boldsymbol{\Phi}_2 \end{bmatrix} \right\|_F \quad (4a)$$

$$\text{s.t. } \boldsymbol{\Phi}_n^H \boldsymbol{X}_n^H \boldsymbol{X}_n \boldsymbol{\Phi}_n = \boldsymbol{I}_R, \ n \in \{1, 2\}. \quad (4b)$$

From (4a) it can be verified that the number of coherent canonical correlation components is equal to the dimension of ker$([\boldsymbol{X}_1, -\boldsymbol{X}_2])$. Thus, in the case where R components have maximal correlation, we have range$(\boldsymbol{X}_1 \boldsymbol{\Phi}_1)$ = range$(\boldsymbol{X}_2 \boldsymbol{\Phi}_2)$. Therefore in the ideal case where the two views share a common subspace of dimension R, the optimal CCA solution gives:

$$\boldsymbol{X}_1 \boldsymbol{\Phi}_1 = \boldsymbol{X}_2 \boldsymbol{\Phi}_2 \quad (5a)$$

$$\text{s.t. } \boldsymbol{\Phi}_n^H \boldsymbol{X}_n^H \boldsymbol{X}_n \boldsymbol{\Phi}_n = \boldsymbol{I}_R, \ n \in \{1, 2\}. \quad (5b)$$

Note that the quadratic constraints (3b), (4b) and (5b) simply say that the columns of $\boldsymbol{X}_1 \boldsymbol{\Phi}_1$ and $\boldsymbol{X}_2 \boldsymbol{\Phi}_2$ must form column-wise orthonormal bases for the obtained subspaces. However, nonorthogonal bases can be used and this constraint is strictly speaking not necessary for CCA.

B. Review of GCCA

Several extensions of CCA to the case of multiple views $N \geq 2$ have been proposed; see [3], [27] for details. When $N \geq 2$ matrix "views" $\boldsymbol{X}_1 \in \mathbb{C}^{I \times J_1}, \ldots, \boldsymbol{X}_N \in \mathbb{C}^{I \times J_N}$ are considered, then the problem of finding canonical correlation components is referred to as GCCA. SUMCOR [3], [27] is a popular formulation for GCCA, which is an extension of the trace maximization formulation (3) to the multiview case:

$$\max_{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_N} \sum_{1 \leq n_1 < n_2 \leq N} \text{trace}\left(\boldsymbol{\Phi}_{n_1}^H \mathbf{X}_{n_1}^H \mathbf{X}_{n_2} \boldsymbol{\Phi}_{n_2}\right), \quad (6a)$$

$$\text{s.t.} \quad \boldsymbol{\Phi}_n^H \mathbf{X}_n^H \mathbf{X}_n \boldsymbol{\Phi}_n = \mathbf{I}_R, \ n \in \{1, \ldots, N\}, \quad (6b)$$

where $\boldsymbol{\Phi}_n = [\boldsymbol{\phi}_{n1}, \ldots, \boldsymbol{\phi}_{nR}] \in \mathbb{C}^{J_n \times R}$. $n \in \{1, \ldots, N\}$. Note that an N-tuple of components $(\boldsymbol{X}_1 \boldsymbol{\phi}_{1r}, \ldots, \boldsymbol{X}_N \boldsymbol{\phi}_{Nr})$, with $\boldsymbol{\phi}_{nr} \in \mathbb{C}^{J_n}$, $n \in \{1, \ldots, N\}$, is now said to be *coherent* if $\boldsymbol{X}_{n_1} \boldsymbol{\phi}_{n_1} \propto \boldsymbol{X}_{n_2} \boldsymbol{\phi}_{n_2}$, $\forall n_1, n_2 \in \{1, \ldots, N\}$. Similar to (4), the maximizer of (6) corresponds to the minimizer of

$$\min_{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_N} \sum_{1 \leq n_1 < n_2 \leq N} \left\| [\mathbf{X}_{n_1}, -\mathbf{X}_{n_2}] \begin{bmatrix} \boldsymbol{\Phi}_{n_1} \\ \boldsymbol{\Phi}_{n_2} \end{bmatrix} \right\|_F, \quad (7a)$$

$$\text{s.t.} \quad \boldsymbol{\Phi}_n^H \mathbf{X}_n^H \mathbf{X}_n \boldsymbol{\Phi}_n = \mathbf{I}_R, \ n \in \{1, \ldots, N\}. \quad (7b)$$

Assume that there exists R coherent canonical correlation components $(\boldsymbol{X}_1 \boldsymbol{\phi}_{1r}, \ldots, \boldsymbol{X}_N \boldsymbol{\phi}_{Nr})$, $r \in \{1, \ldots, R\}$ that can be extracted from the N matrix "views" $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, so that $\text{range}(\boldsymbol{X}_1 \boldsymbol{\Phi}_1) = \cdots = \text{range}(\boldsymbol{X}_N \boldsymbol{\Phi}_N)$. Then, similar to (5), the solution to (7) will in the ideal case satisfy:

$$\boldsymbol{X}_{n_1} \boldsymbol{\Phi}_{n_1} = \boldsymbol{X}_{n_2} \boldsymbol{\Phi}_{n_2}, \ n_1 \neq n_2 \quad (8a)$$

$$\text{s.t.} \ \boldsymbol{\Phi}_n^H \boldsymbol{X}_n^H \boldsymbol{X}_n \boldsymbol{\Phi}_n = \boldsymbol{I}_R, \ n \in \{1, \ldots, N\}. \quad (8b)$$

Similar to CCA, the quadratic constraints (8b) simply say that the columns of $\boldsymbol{X}_n \boldsymbol{\Phi}_n$ must form a columnwise orthonormal basis, which is strictly speaking not necessary for GCCA.

It is important to note that other extensions of CCA to the multiview GCCA case have been proposed. We mention the MAXVAR formulation [3], [27], which will be reviewed in Section VI-B. In the next section we will propose a generative model for GCCA.

C. A generative model for GCCA

a) Definition of generative model for GCCA: Assume that relation (8a) is satisfied and that $\cap_{n=1}^N \text{range}(\boldsymbol{X}_n) = R$, then there only exist R linearly independent and maximally correlated components. Let the columns of $\boldsymbol{A} \in \mathbb{C}^{I \times R}$ form a basis for the subspace spanned by the R coherent canonical correlation components, i.e., $\text{range}(\boldsymbol{A}) = \text{range}(\boldsymbol{X}_n \boldsymbol{\Phi}_n)$, $\forall n \in \{1, \ldots, N\}$. Then there *always exist* matrices $\mathbf{B}_n \in \mathbb{C}^{J_n \times R}$ $\boldsymbol{C}_n \in \mathbb{C}^{I \times L_n}$ and $\mathbf{D}_n \in \mathbb{C}^{J_n \times L_n}$ such that

$$\begin{aligned} \mathbf{X}_n &= \mathbf{A}\mathbf{B}_n^T + \mathbf{C}_n \mathbf{D}_n^T \\ &= [\boldsymbol{A}, \boldsymbol{C}_n] \boldsymbol{S}_n^T \in \mathbb{C}^{I \times J_n}, \quad n \in \{1, \ldots, N\}, \end{aligned} \quad (9)$$

where $\boldsymbol{S}_n = [\mathbf{B}_n, \mathbf{D}_n] \in \mathbb{C}^{J_n \times (R + L_n)}$ and $\text{rank}(\mathbf{X}_n) = R + L_n$. Note that since $\cap_{n=1}^N \text{range}(\boldsymbol{X}_n) = R$, we may assume, without loss of generality (w.l.o.g.), that $\cap_{n=1}^N \text{range}(\boldsymbol{C}_n) = \{\mathbf{0}\}$. Note also that when $\dim(\cap_{n=1}^N \text{range}(\boldsymbol{C}_n)) \geq 0$ is permitted, then (9) can more generally be interpreted as a coupled low-rank factorization. Thus, the difference between the discussed generative GCCA model and a coupled low-rank factorization model is that the former model requires that $\dim(\cap_{n=1}^N \text{range}(\boldsymbol{C}_n)) = 0$. To summarize, when $\dim(\cap_{n=1}^N \text{range}(\boldsymbol{C}_n)) = 0$, then (9) is referred to as a generative GCCA model for $\mathbf{X}_1, \ldots, \mathbf{X}_N$ and when $\dim(\cap_{n=1}^N \text{range}(\boldsymbol{C}_n)) \geq 0$, then (9) is referred to as a coupled low-rank factorization of $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Note that the generative GCCA model (9) does not prevent that $R = 0$.

Our first observation is that since $\boldsymbol{A}$ in (9) is a shared factor matrix, we can w.l.o.g. assume that the matrices $\{[\boldsymbol{A}, \boldsymbol{C}_n]\}$ in (9) have full column rank (see Appendix A for a detailed proof) and that $\text{range}(\boldsymbol{X}_n) = \text{range}([\boldsymbol{A}, \boldsymbol{C}_n])$, $\forall n \in \{1, \ldots, N\}$. The latter implies that w.l.o.g. we can also assume that the matrices $\{\boldsymbol{S}_n\}$ in (9) have full column rank. Hence, w.l.o.g. we can always assume that the matrices $\mathbf{X}_1, \ldots, \mathbf{X}_N$ admit the factorization in (9), where R denotes the dimension of the common subspace $\cap_{n=1}^N \text{range}(\boldsymbol{X}_n)$, which is equal to the number of coherent canonical components, and $R + L_n$ denotes the dimension of the individual subspace $\text{range}(\boldsymbol{X}_n)$, $n \in \{1, \ldots, N\}$. We also note in passing that w.l.o.g. it can be assumed that $J_n = R + L_n$, i.e., if $J_n > R + L_n$, then $\boldsymbol{X}_n$ of size $(I \times J_n)$ can be replaced by a compressed version of size $(I \times (R + L_n))$.

To summarize, GCCA with R coherent canonical correlation components can w.l.o.g. be interpreted as the problem of finding the common subspace $\text{range}(\mathbf{A})$ of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ which can be factored as in (9), where the factor matrices

$$[\mathbf{A}, \mathbf{C}_n] \text{ and } \mathbf{S}_n = [\mathbf{B}_n, \mathbf{D}_n] \text{ have full column rank.} \quad (10)$$

We briefly mention that in the non-ideal case where there do not exist R coherent canonical correlation components, then $\mathbf{A}$ in (9) can be chosen to be the columnwise orthonormal matrix that minimizes $\| \sum_{n=1}^N \mathbf{P}_{\boldsymbol{X}_n}^{\perp} \mathbf{A} \|_F^2$, where $\mathbf{P}_{\boldsymbol{X}_n}^{\perp}$ denotes the orthogonal projector onto the orthogonal complement of $\text{range}(\boldsymbol{X}_n)$. This is related to the MAXVAR formulation of GCCA, as will be explained in Section VI.

It is important to note that the factorization (9) together with the full column rank properties of $\{[\boldsymbol{A}, \boldsymbol{C}_n]\}$ and $\{\boldsymbol{S}_n\}$ allow us to assume, w.l.o.g. that [1]

$$\text{range}(\boldsymbol{X}_n) = \text{range}(\boldsymbol{A}) \oplus \text{range}(\boldsymbol{C}_n), \quad n \in \{1, \ldots, N\}, \quad (11)$$

which implies that

$$R \leq I - \max_{1 \leq n \leq N} L_n. \quad (12)$$

Note that equation (11) is a key point of our approach and follows naturally from the fact that matrix $[\boldsymbol{A}, \boldsymbol{C}_n]$ has full column rank and that the subspaces $\text{range}(\boldsymbol{A})$ and $\text{range}(\boldsymbol{C}_n)$ are complementary (see Appendix A for details). We also make use of the convention

$$L_1 \leq L_2 \leq \cdots \leq L_N. \quad (13)$$

[1]Recall that if U and V are subspaces of the vector space W, then $W = U \oplus V$ if and only if $U \cap V = \{\mathbf{0}\}$ and $W = U + V$. Equivalently, $W = U \oplus V$ if and only if for any $\boldsymbol{w} \in W$ there exists a unique vector $\boldsymbol{u} \in U$ and a unique vector $\boldsymbol{v} \in V$ such that $\boldsymbol{w} = \boldsymbol{u} + \boldsymbol{v}$.

b) Definition of identifiability of range($\mathbf{A}$): Consider a set of N views $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. Assume that there exist R coherent canonical correlation components. Let the columns of $\mathbf{A} \in \mathbb{C}^{I \times R}$ form a basis for the span of the R coherent canonical correlation components. Note that range($\mathbf{A}$) can only be identified from $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ if $\dim(\cap_{n=1}^N \text{range}(\mathbf{X}_n)) = R$. Thus,

$$\text{range}(\mathbf{A}) \text{ is identifiable } \Leftrightarrow \dim(\cap_{n=1}^N \text{range}(\mathbf{X}_n)) = R. \tag{14}$$

As an example, assume that we want to extract two coherent canonical correlation components from the two views $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Let the model parameters of the coupled low-rank factorization (9) be $I = 3$, $R = 2$, $N = 2$ and $L_1 = L_2 = 1$. Since $\dim(\text{range}(\mathbf{X}_1) \cap \text{range}(\mathbf{X}_2)) = 3 > 2$, the two coherent canonical correlation components of interest cannot be identified from the two views $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

Observe that (14) expresses the identifiability of range($\mathbf{A}$), with $\dim(\text{range}(\mathbf{A})) = R$, in terms of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. We can also express it in terms of the factor matrices $\mathbf{A}$ and $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$ in the model (9). More precisely, let the columns of range($\mathbf{A}$) in (9) form a basis for the span of the R coherent canonical correlation components. Then identifiability of range($\mathbf{A}$) means that the views $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ admit the decompositions (9) with property $\cap_{n=1}^N \text{range}(\boldsymbol{C}_n) = \{\mathbf{0}\}$. This formulation of identifiability of range($\mathbf{A}$) will be important in the next sections and for that reason we state it again below

$$\text{range}(\mathbf{A}) \text{ is identifiable } \Leftrightarrow \dim(\cap_{n=1}^N \text{range}(\mathbf{C}_n)) = 0. \tag{15}$$

III. A SUBSPACE INTERSECTION APPROACH FOR CCA AND GCCA

In this section we provide a range subspace intersection approach for finding range($\boldsymbol{A}$) via the observed matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ with decompositions of the form (9), where $\dim(\bigcap_{n=1}^N \text{range}\,(\boldsymbol{C}_n)) \geq 1$ is permitted. The full column rank property of the matrices $\{[\boldsymbol{A}, \boldsymbol{C}_n]\}$ and $\{\boldsymbol{S}_n\}$ in (9) imply that

$$\begin{aligned} &\text{range}(\boldsymbol{X}_{n_1}) \cap \text{range}(\boldsymbol{X}_{n_2}) \\ &= \text{range}([\boldsymbol{A}, \boldsymbol{C}_{n_1}]) \cap \text{range}([\boldsymbol{A}, \boldsymbol{C}_{n_2}]) \\ &= \text{range}(\boldsymbol{A}) + (\text{range}(\boldsymbol{C}_{n_1}) \cap \text{range}(\boldsymbol{C}_{n_2}))\,, \\ &\qquad\qquad 1 \leq n_1 < n_2 \leq N, \end{aligned}$$

where the last equality follows from (11). More generally, we have that

$$\begin{aligned} Y &:= \bigcap_{n=1}^N \text{range}(\boldsymbol{X}_n) \\ &= \text{range}(\boldsymbol{A}) \oplus \left(\bigcap_{n=1}^N \text{range}(\boldsymbol{C}_n) \right) \\ &= \text{range}(\boldsymbol{A}) \oplus C, \end{aligned} \tag{16}$$

where $C := \bigcap_{n=1}^N \text{range}\,(\boldsymbol{C}_n)$. Thus, relation (11) implies that

$$\text{range}(\mathbf{A}) \subseteq Y \text{ and that } \dim(Y) \geq R. \tag{17}$$

Observe that $\dim(Y) = R$ implies that $Y = \text{range}(\boldsymbol{A})$. As a result, the study of the identifiability of range($\boldsymbol{A}$) reduces to the study of $\dim(Y)$. To put it differently,

$$\dim(Y) = R \Leftrightarrow Y = \text{range}(\mathbf{A}). \tag{18}$$

Note that (18) means that $\dim(C) = 0$, which in turn means that (9) corresponds to a generative GCCA model. From (15) we also know that this means that range($\mathbf{A}$) is identifiable via the N views $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. Thus, if the dimension of the subspace spanned by a basis for Y is R-dimensional, then range($\boldsymbol{A}$) can be uniquely determined from Y via GCCA.

What remains to be answered is how to determine the dimension of Y. Consider a nonzero vector $\boldsymbol{z} \in \mathbb{C}^I$. Let the columns of $\boldsymbol{U}_n \in \mathbb{C}^{I \times (R+L_n)}$ form a basis for range($\boldsymbol{X}_n$). We know that $\boldsymbol{z} \in Y$ if and only if there exist nonzero vectors $\boldsymbol{q}_1 \in \mathbb{C}^{R+L_1}, \ldots, \boldsymbol{q}_N \in \mathbb{C}^{R+L_N}$ such that

$$\boldsymbol{z} = \boldsymbol{U}_1 \boldsymbol{q}_1 = \cdots = \boldsymbol{U}_N \boldsymbol{q}_N. \tag{19}$$

Define $\boldsymbol{q} = [\boldsymbol{q}_1^T, \ldots, \boldsymbol{q}_N^T]^T \in \mathbb{C}^{(NR + \sum_{n=1}^N L_n)}$. Then a vector $\boldsymbol{q}$ with property (19) can be obtained by solving the system of homogenous linear equations

$$\left[\mathbf{0}_{I \times \alpha_{n_1}}, \boldsymbol{U}_{n_1}, \mathbf{0}_{I \times \beta_{n_1,n_2}}, -\boldsymbol{U}_{n_2}, \mathbf{0}_{I \times \omega_{n_2}}\right] \boldsymbol{q} = \mathbf{0}_I, \tag{20}$$

for $1 \leq n_1 < n_2 \leq N$, where
$\alpha_{n_1} = (n_1 - 1)R + \sum_{i=1}^{n_1 - 1} L_i$,
$\beta_{n_1,n_2} = (n_2 - n_1 - 1)R + \sum_{i=n_1+1}^{n_2-1} L_i$,
$\omega_{n_2} = (N - n_2)R + \sum_{i=n_2+1}^{N} L_i$.

We can now conclude that if the subspace

$$Z^{(N)} := \bigcap_{1 \leq n_1 < n_2 \leq N} \ker([\mathbf{0}, \boldsymbol{U}_{n_1}, \mathbf{0}, -\boldsymbol{U}_{n_2}, \mathbf{0}]) \tag{21}$$

$$= \bigcap_{1 \leq n_1 < n_2 \leq N} \ker([\mathbf{0}, \boldsymbol{A}, \boldsymbol{C}_{n_1}, \mathbf{0}, -\boldsymbol{A}, -\boldsymbol{C}_{n_2}, \mathbf{0}]) \tag{22}$$

is R-dimensional, i.e., there exist only R linearly independent vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_R$ in the range of $Z^{(N)}$, then Y is also R-dimensional and $Y = \text{range}(\boldsymbol{A})$. Note that the dimensions of the zero matrices in (22) are as in (20), but the subscripts have been omitted due to space limitations.

Therefore the dimension of Y can be expressed in terms of the factor matrices $\boldsymbol{A}, \{\boldsymbol{C}_n\}_{n=1}^N$. For example, when $N = 3$, the dimension of Y is equal to the dimension of the kernel of the following matrix (the same reasoning holds true when $N > 3$):

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & -\boldsymbol{A} & -\boldsymbol{C}_2 & \mathbf{0} & \mathbf{0} \\ \boldsymbol{A} & \boldsymbol{C}_1 & \mathbf{0} & \mathbf{0} & -\boldsymbol{A} & -\boldsymbol{C}_3 \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A} & \boldsymbol{C}_2 & -\boldsymbol{A} & -\boldsymbol{C}_3 \end{bmatrix}. \tag{23}$$

IV. IDENTIFIABILITY CONDITIONS FOR GCCA

The goal of GCCA is to find the subspace range($\boldsymbol{A}$), observing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$. We consider the exact case where $\boldsymbol{X}_n$ admits the factorization (9). As far as identifiability is concerned, [16] studied the two-view CCA ($N = 2$) and proved that if the matrices $[\boldsymbol{A}, \boldsymbol{C}_1, \boldsymbol{C}_2]$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ have full column rank, then range $(\boldsymbol{A})$ can be obtained via CCA, as reviewed in Section IV-A. In Section IV-B we move a step forward and provide an identifiability condition for the general

case of GCCA ($N \geq 2$). More precisely, using the proposed range subspace intersection approach for GCCA, we present an identifiability condition that does not require any of the matrices in the set $\{[\boldsymbol{A}, \boldsymbol{C}_{n_1}, \boldsymbol{C}_{n_2}]\}$ to have full column rank.

A. Review of CCA identifiability conditions

Without loss of generality, we assume that $\mathbf{X}_1$ and $\mathbf{X}_2$ in (5a)–(5b) admit factorizations

$$\begin{cases} \mathbf{X}_1 = \mathbf{A}\mathbf{B}_1^T + \mathbf{C}_1\mathbf{D}_1^T \in \mathbb{C}^{I \times J_1}, \\ \mathbf{X}_2 = \mathbf{A}\mathbf{B}_2^T + \mathbf{C}_2\mathbf{D}_2^T \in \mathbb{C}^{I \times J_2}, \end{cases} \tag{24}$$

where $\mathbf{A} \in \mathbb{C}^{I \times R}$, $\mathbf{B}_n \in \mathbb{C}^{J_n \times R}$, $\mathbf{C}_n \in \mathbb{C}^{I \times L_n}$ and $\mathbf{D}_n \in \mathbb{C}^{J_n \times L_n}$. Note that (24) is a special case of (9) with $N = 2$. The question is now when does the CCA solution yield range$(\mathbf{A})$. Theorem IV.1 below answers this question.

Theorem IV.1. *[16] Consider the two-view factorization of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ given by* (24)*. If*

$$\begin{cases} [\mathbf{B}_1, \mathbf{D}_1] \in \mathbb{C}^{J_1 \times (R+L_1)} \text{ has full column rank}, \\ [\mathbf{B}_2, \mathbf{D}_2] \in \mathbb{C}^{J_2 \times (R+L_2)} \text{ has full column rank}, \\ [\mathbf{A}, \mathbf{C}_1, \mathbf{C}_2] \in \mathbb{C}^{I \times (R+L_1+L_2)} \text{ has full column rank}, \end{cases} \tag{25}$$

then the common subspace range$(\mathbf{A})$ is identifiable via (24) *and the CCA solution* (5a)–(5b) *has the property range$(\mathbf{A})$ = range$(\boldsymbol{X}_1\boldsymbol{\Phi}_1)$ = range$(\boldsymbol{X}_2\boldsymbol{\Phi}_2)$.*

Note that condition (25) does not require that $\boldsymbol{\Phi}_n^H \boldsymbol{X}_n^H \boldsymbol{X}_n \boldsymbol{\Phi}_n = \boldsymbol{I}_R$. It is important to note that if condition (25) is not satisfied, then in general we have range$(\mathbf{A}) \neq$ range$(\boldsymbol{X}_1\boldsymbol{\Phi}_1) =$ range$(\boldsymbol{X}_2\boldsymbol{\Phi}_2)$, even if relations (5a)–(5b) are satisfied, and consequently range$(\mathbf{A})$ cannot be obtained via CCA. (This happens when $\dim(\bigcap_{n=1}^{2} \text{range}(\boldsymbol{C}_n)) \geq 1$.) A nice property of condition (25) is that it is easy to check and it is generically[2] satisfied if

$$J_1 \geq R + L_1, \;\; J_2 \geq R + L_2 \;\; \text{and} \;\; I \geq R + L_1 + L_2. \tag{26}$$

However, a drawback of condition (25) is that it is limited to the two-view case ($N = 2$), i.e., in the multi-view case it does not exploit all $N \geq 2$ observation matrices $\mathbf{X}_1, \ldots, \mathbf{X}_N$. For this reason, we consider GCCA, so that all $N \geq 2$ observation matrices $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are taken into account.

B. GCCA identifiability conditions

We start our identifiability analysis for GCCA by providing a short proof that explains that if range$(\mathbf{A})$ can be obtained via a two-view CCA based method, then range$(\mathbf{A})$ can also be obtained by a multi-view GCCA based method.

Proposition IV.2. *Consider the multi-view factorization of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ given by* (9)*. If condition* (25) *is satisfied for some pair $(\boldsymbol{X}_{n_1}, \boldsymbol{X}_{n_2})$, where $1 \leq n_1 < n_2 \leq N$, then dim(Y) = R, Y = range$(\boldsymbol{A})$ and the GCCA solution* (8a)–(8b) *has the property range$(\mathbf{A})$ = range$(\boldsymbol{X}_n\boldsymbol{\Phi}_n)$, $\forall n \in \{1, \ldots, N\}$.*

[2]We say that factor matrices $\mathbf{A}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{C}_1, \mathbf{C}_2, \mathbf{D}_1, \mathbf{D}_2$ in (24) are generic when their entries can be assumed to have been drawn from an absolutely continuous joint probability distribution.

Proof. Condition (25) implies that range$(\mathbf{A})$ = range$(\mathbf{X}_1) \cap$ range$(\mathbf{X}_2)$ and that $\dim(Y) \leq R$. Consequently, from (17), we conclude that if range$(\mathbf{A})$ = range$(\mathbf{X}_1) \cap$ range$(\mathbf{X}_2)$, then $Y = \cap_{n=1}^{N}$range$(\mathbf{X}_n)$ = range$(\mathbf{X}_1) \cap$ range$(\mathbf{X}_2)$ = range$(\mathbf{A})$. □

Hence, in terms of identifiability, a multi-view GCCA method cannot do worse than a two-view CCA method. In this section we explain that by taking all views $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ into account, the identifiability condition for a multi-view GCCA method is in fact more relaxed than the identifiability condition for a two-view CCA method, i.e., even if condition (25) is not satisfied, range$(\boldsymbol{A})$ can still be obtained via a GCCA method.

It is important to note that even if there exist matrices $\{\boldsymbol{\Phi}_n\}$ such that relations (8a)–(8b) are satisfied, it does not necessarily mean that range$(\boldsymbol{A})$ = range$(\boldsymbol{X}_n\boldsymbol{\Phi}_n)$. (This happens when $\dim(\bigcap_{n=1}^{N} \text{range}(\boldsymbol{C}_n)) \geq 1$.) We will now develop conditions that ensure that the following implication is satisfied

$$\boldsymbol{x} \in \cap_{n=1}^{N}\text{range}(\boldsymbol{X}_n) \Rightarrow \boldsymbol{x} \in \text{range}(\boldsymbol{A}), \tag{27}$$

so that the common subspace range$(\boldsymbol{A})$ can be obtained via GCCA, observing $\boldsymbol{X}_n$, $n \in \{1, \ldots, N\}$.

Since range$(\boldsymbol{A}) \subseteq Y$, the minimal dimension of the subspace Y given by (16) is R. This also means that if the subspace $Z^{(N)}$ given by (22) is R-dimensional, then $C = \{\mathbf{0}\}$.[3] This fact leads to the common subspace identifiability condition presented in Theorem IV.3 below.

Theorem IV.3. *Consider the multi-view factorization of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ given by* (9)*. If*

$$\begin{cases} Z^{(N)} \text{ is } R\text{-dimensional}, \\ \boldsymbol{S}_1, \ldots, \boldsymbol{S}_N \text{ have full column rank}, \end{cases} \tag{28}$$

then dim(Y) = R, Y = range$(\mathbf{A})$ and the GCCA solution (8a)–(8b) *has the property range$(\mathbf{A})$ = range$(\boldsymbol{X}_n\boldsymbol{\Phi}_n)$, $\forall n \in \{1, \ldots, N\}$.*

Proof. The result follows immediately from relations (11), (16), (21) and (22). □

In Theorem IV.3 we exploited the fact that relation (16) tells us that if $\dim(Y) = R$, then $C = \{\mathbf{0}\}$. In words, common subspace identifiability means that the noise terms $\mathbf{C}_1, \ldots, \mathbf{C}_N$ have been "cancelled out" by subspace intersection. Using relation (22), Theorem IV.3 expresses the identifiability condition in terms of the observed data. This is useful when we want to check how well a GCCA model fits to the data and when we want to develop algorithms for GCCA. (In Section VI we develop an algorithm for GCCA based on a constructive use of Theorem IV.3.) On the other hand, if the factor matrices $\boldsymbol{A}, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$ in (9) are given, then checking the dimension of $Z^{(N)}$ in (28) can be cumbersome and it is not obvious how it is related to the factor matrices $\boldsymbol{A}, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$ in (9). In order to obtain a simpler condition for the recovery of range$(\boldsymbol{A})$ via

[3]Recall that the dimension of a direct sum is the sum of the dimensions of its summands. This fact also explains that if Y is R-dimensional, then $C = \{\mathbf{0}\}$.

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, that is expressed in terms of $\boldsymbol{A}, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$, the following identity will be used [31]:

$$\begin{aligned}&\dim\left(\bigcap_{n=1}^{N} \text{range}(\boldsymbol{X}_n)\right)\\&= \sum_{n=1}^{N} \text{rank}(\boldsymbol{X}_n) - \text{rank}(\boldsymbol{\Gamma}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)),\end{aligned} \tag{29}$$

where the matrix $\boldsymbol{\Gamma}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ is defined as follows

$$\boldsymbol{\Gamma}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N) = \begin{bmatrix} \boldsymbol{X}_1 & -\boldsymbol{X}_2 & & \\ \vdots & & \ddots & \\ \boldsymbol{X}_1 & & & -\boldsymbol{X}_N \end{bmatrix},$$

in which $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ are matrices of conformable sizes. Theorem IV.4 below is a simplified version of the common subspace identifiability condition in Theorem IV.3. It makes use of the matrix $\boldsymbol{\Gamma}^{(N)} \in \mathbb{C}^{(N-1)I \times ((N-1)R + \sum_{n=1}^{N} L_n)}$ given by

$$\begin{aligned}\boldsymbol{\Gamma}^{(N)} &= [\boldsymbol{1}_{N-1} \otimes \mathbf{C}_1, -\text{Blkdiag}([\mathbf{A}, \mathbf{C}_2], \cdots, [\mathbf{A}, \mathbf{C}_N])]\\&= \begin{bmatrix} \boldsymbol{C}_1 & -\boldsymbol{A} & -\boldsymbol{C}_2 & & \\ \vdots & & \ddots & \ddots & \\ \boldsymbol{C}_1 & & & -\boldsymbol{A} & -\boldsymbol{C}_N \end{bmatrix},\end{aligned} \tag{30}$$

where $\boldsymbol{1}_{N-1} = [1, \ldots, 1]^T \in \mathbb{C}^{N-1}$ is an all-ones vector and $\text{Blkdiag}([\mathbf{A}, \mathbf{C}_2], \cdots, [\mathbf{A}, \mathbf{C}_N])$ is a block-diagonal matrix that holds the matrices $[\mathbf{A}, \mathbf{C}_2], \ldots, [\mathbf{A}, \mathbf{C}_N]$ on its block-diagonal.

Theorem IV.4. *Consider the multi-view factorization of* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ *given by* (9)*. If*

$$\begin{cases} \boldsymbol{\Gamma}^{(N)} \text{ has full column rank}, \\ \boldsymbol{S}_1, \ldots, \boldsymbol{S}_N \text{ have full column rank}, \end{cases} \tag{31}$$

then $\dim(Y) = R$*,* $Y = \text{range}(\mathbf{A})$ *and the GCCA solution* (8a)–(8b) *has the property* $\text{range}(\mathbf{A}) = \text{range}(\boldsymbol{X}_n \boldsymbol{\Phi}_n)$*,* $\forall n \in \{1, \ldots, N\}$*.*

Proof. Relation (29) together with the full column rank assumptions on $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N$ imply that

$$\begin{aligned}&\dim\left(\bigcap_{n=1}^{N} \text{range}(\boldsymbol{X}_n)\right) =\\&\sum_{n=1}^{N} \text{rank}(\boldsymbol{X}_n) - \text{rank}(\boldsymbol{\Gamma}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)) =\\&\sum_{n=1}^{N} \text{rank}([\boldsymbol{A}, \boldsymbol{C}_n]) - \text{rank}(\boldsymbol{\Gamma}([\boldsymbol{A}, \boldsymbol{C}_1], \ldots, [\boldsymbol{A}, \boldsymbol{C}_N])).\end{aligned} \tag{32}$$

The full column rank property of the matrices $[\boldsymbol{A}, \boldsymbol{C}_1], \ldots, [\boldsymbol{A}, \boldsymbol{C}_N]$ in turn implies that relation (32) simplifies to

$$\begin{aligned}&\dim\left(\bigcap_{n=1}^{N} \text{range}(\boldsymbol{X}_n)\right) =\\&NR + \sum_{n=1}^{N} L_n - \text{rank}(\boldsymbol{\Gamma}([\boldsymbol{A}, \boldsymbol{C}_1], \ldots, [\boldsymbol{A}, \boldsymbol{C}_N]))\\&NR + \sum_{n=1}^{N} L_n - \text{rank}\left(\boldsymbol{\Gamma}^{(N)}\right),\end{aligned} \tag{33}$$

where the matrix $\boldsymbol{\Gamma}^{(N)}$ is given by (30). From (33) it is clear that if $\boldsymbol{\Gamma}^{(N)}$ has full column rank, then $\dim\left(\bigcap_{n=1}^{N} \text{range}(\boldsymbol{X}_n)\right) = R$ and $\bigcap_{n=1}^{N} \text{range}(\boldsymbol{X}_n) = \text{range}(\boldsymbol{A})$. □

Note that if $\boldsymbol{\Gamma}^{(N)}$ has full column rank, then $I(N-1) \geq R(N-1) + L_1 + \cdots + L_N$ or equivalently the row dimension (I) satisfies the inequality

$$I \geq R + \left\lceil \frac{L_1 + \cdots + L_N}{N-1} \right\rceil. \tag{34}$$

The necessary condition for $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N$ to have full column rank is:

$$I \geq R + \max_{1 \leq n \leq N} L_n = R + L_N, \tag{35}$$

where convention (13) was used. Interestingly, the inequalities (34) and (35), which are necessary for Theorem IV.4, are also sufficient when $\boldsymbol{A}$, $\boldsymbol{C}_n$ and $\boldsymbol{S}_n$, $n \in \{1, \ldots, N\}$ are generic, i.e., the entries of the involved factor matrices can be assumed to have been drawn from an absolutely continuous joint probability distribution. The generic version of Theorems IV.3 and IV.4 is presented as Theorem IV.5 below.

Theorem IV.5. *Consider the multi-view factorization of* $\boldsymbol{X}_1 \ldots, \boldsymbol{X}_N$ *given by* (9)*. If*

$$\begin{cases} R + \left\lceil \dfrac{1}{N-1} \displaystyle\sum_{n=1}^{N} L_n \right\rceil \leq I, \\ R + L_n \leq J_n, \ \forall n \in \{1, \ldots, N\}. \end{cases} \tag{36}$$

then generically $\dim(Y) = R$*,* $Y = \text{range}(\mathbf{A})$ *and the GCCA solution* (8a)–(8b) *has the property* $\text{range}(\mathbf{A}) = \text{range}(\boldsymbol{X}_n \boldsymbol{\Phi}_n)$*,* $\forall n \in \{1, \ldots, N\}$*.*

Proof. See Appendix B. □

Note that in order for the GCCA model (9) to be well-defined, the inequality (12) also has to be satisfied. Theorem IV.5 provides us with an easy way to check identifiability for the special case where the factor matrices of the GCCA model (9) are generic. The latter can be used to quickly assess whether the common subspace $\text{range}(\mathbf{A})$ can be recovered.

V. Discussion

In this section we discuss the effect of processing more views (i.e., $N > 2$) using GCCA compared to the more commonly used CCA model in which $N = 2$. First we note that when $N = 2$, condition (31) boils down to the standard two-view CCA identifiability condition (25). However, when $N > 2$ condition (31) yields relaxed identifiability, as we will demonstrate next.

A. GCCA can relax the bound on I

Consider the case where $\boldsymbol{A}, \boldsymbol{C}_n, \boldsymbol{S}_n$, $n \in \{1, \ldots, N\}$ are generic and $R = L_n = 100$, $n \in \{1, \ldots, N\}$. Then condition (26) for two-view CCA requires $I \geq R + L_1 + L_2 = 300$, in order to recover range $(\boldsymbol{A})$. When $N = 3$, however, the condition in (36) is relaxed to $I \geq R + \frac{1}{2}(L_1 + L_2 + L_3) = 250$. Furthermore, in the latter case none of the matrices $[\boldsymbol{A}, \boldsymbol{C}_1, \boldsymbol{C}_2]$, $[\boldsymbol{A}, \boldsymbol{C}_1, \boldsymbol{C}_3]$ and $[\boldsymbol{A}, \boldsymbol{C}_2, \boldsymbol{C}_3]$ are required to have full column rank, which is necessary in the two-view case. The identifiability condition for GCCA can be further relaxed by increasing N. In the previous example, when $N = 5$ the condition in (36) reduces to $I \geq 225$ and as $N \to \infty$ to $I \geq R + L_n = 200$, which is also a necessary condition to identify range $(\boldsymbol{A})$.

B. GCCA can identify higher dimensional common subspaces

Note that multi-view GCCA allows that $\dim(\text{range}(\mathbf{A}) \cap \text{range}(\mathbf{C}_1) \cap \text{range}(\mathbf{C}_2)) > 0$, which is not permitted in the two-view CCA case. Consequently, GCCA can identify higher dimensional common subspaces, range$(\mathbf{A})$.

As an example, let $I = 200$, $N = 5$, and $L_n = 100$, $1 \leq n \leq 5$. The condition (26) for two-view CCA requires that $R \leq I - L_1 - L_2 = 0$, which means that it is impossible to identify range$(\mathbf{A})$. On the hand, the GCCA identifiability condition (36) only requires that $R \leq I - \frac{1}{N-1}(\sum_{n=1}^{5} L_n) = 75$.

C. GCCA can handle "leaky" noise subspaces

Another important difference between two-view and multi-view CCA is that when $N = 2$, range $(\boldsymbol{C}_1) \cap$ range $(\boldsymbol{C}_2) = \{\mathbf{0}\}$ is a necessary identifiability condition. On the contrary, for $N > 2$ it is possible that range $(\boldsymbol{C}_m) \cap$ range $(\boldsymbol{C}_n) \neq \{\mathbf{0}\}$ for some $m \neq n$, i.e., some views are allowed to share common subspaces not included in range $(\boldsymbol{A})$. In words, GCCA allow for "leaky" noise subspaces with property $\dim(\text{range}\,(\boldsymbol{C}_m) \cap \text{range}\,(\boldsymbol{C}_n)) > 0$ for some $m \neq n$, as long as $C = \dim(\cap_{n=1}^{N} \text{range}\,(\boldsymbol{C}_n)) = 0$.

As an example, let $I = 200$, $N = 5$, and $L_n = 140$, $1 \leq n \leq 5$. The condition (26) for two-view CCA requires that $R \leq I - L_1 - L_2 = -80$, which means that it is impossible to identify range$(\mathbf{A})$. On the hand, the GCCA identifiability condition (36) only requires that $R \leq I - \frac{1}{N-1}(\sum_{n=1}^{5} L_n) = 25$. Note that for generic $\mathbf{C}_1$ and $\mathbf{C}_2$, we have $\dim(\text{range}(\mathbf{C}_1) \cap \text{range}(\mathbf{C}_2)) = 80$.

D. Asymptotic results

To further elaborate on the identifiability properties of CCA ($N = 2$) and GCCA ($N > 2$), consider the case where R is fixed. Condition (36) implies that

$$\sum_{n=1}^{N} L_n \leq (N-1)(I-R) \tag{37}$$

is necessary for condition (31) to be satisfied. If additionally $L := L_1 = \cdots = L_N$, then (37) reduces to

$$L \leq \frac{N-1}{N}(I-R), \tag{38}$$

which yields the following relation, when $N = 2$:

$$L \leq \frac{1}{2}(I-R). \tag{39}$$

Furthermore, when $N \to \infty$, then (38) reduces to

$$L \leq I - R. \tag{40}$$

Comparing (39) with (40), we conclude that in the balanced case where $L := L_1 = \cdots = L_N$, GCCA can at most relax the CCA bound on L by a factor $\frac{I-R}{\frac{1}{2}(I-R)} = 2$.

Let us now consider the balanced case ($L := L_1 = \cdots = L_N$) where L is fixed, while R is varying. Relation (39) implies that CCA with $N = 2$ views is able to recover the common subspace range $(\boldsymbol{A})$ only if

$$R \leq I - 2L. \tag{41}$$

In other words, $L < \frac{I}{2}$ is a necessary recovery condition for CCA. On the contrary, employing more views ($N > 2$), allows GCCA to recover the common subspace range $(\boldsymbol{A})$, even if $L \geq \frac{I}{2}$. For instance, if $I = 200$, $N = 5$ and $L = 100$, then it can be verified that condition (36) is satisfied as a long as $R \leq 75$, regardless of the fact that $L = \frac{I}{2}$. Furthermore, as $N \to \infty$ we get from (40) that

$$R \leq I - L \tag{42}$$

is necessary to satisfy condition (31) in Theorem IV.4. Comparing (41) with (42), we conclude that when $L := L_1 = \cdots = L_N$ and $L < \frac{I}{2}$, GCCA can at most relax the bound on R by a factor $\frac{I-L}{I-2L}$. Moreover, when $L \geq \frac{I}{2}$, GCCA can still ensure the recovery of range $(\boldsymbol{A})$ while this is *never* possible when $N = 2$.

VI. Algorithmic framework

A. Subspace intersection algorithms for CCA and GCCA

Formulas for computing a basis for the intersection of subspaces have been proposed in the literature. We mention that in [32] it was shown that

$$Y = \bigcap_{n=1}^{N} \text{range}(\boldsymbol{X}_n) = \ker\left(\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp}\right), \tag{43}$$

where $\mathbf{P}_{\boldsymbol{X}_n}^{\perp} \in \mathbb{C}^{I \times I}$ denotes the orthogonal projector onto the orthogonal complement of range$(\boldsymbol{X}_n)$. In the exact case, where there exist R maximally correlated components between $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, a basis for Y can be obtained by solving

$$\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp} \mathbf{A} = \mathbf{0}, \quad \text{s.t.} \quad \mathbf{A}^H \mathbf{A} = \mathbf{I}_R, \tag{44}$$

where the matrix whose columns form an orthonormal basis for Y is denoted by $\mathbf{A}$. It corresponds to matrix $\mathbf{A}$ in (9) and it can for instance be obtained via the singular value decomposition (SVD) of $\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp}$, i.e., the columns of $\mathbf{A}$ correspond to the R right singular vectors associated with the R smallest singular values of $\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp}$. In the inexact case, where there do not exist R maximally correlated components between $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, a basis for Y can be estimated via

$$\min_{\mathbf{A}^H \mathbf{A} = \mathbf{I}_R} \left\| \sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp} \mathbf{A} \right\|_F^2 \geq 0. \tag{45}$$

Note that the lower bound in (45) can only be attained when range$(\boldsymbol{A}) \subseteq$ ker$(\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp})$. [4] Matrix $\mathbf{A}$ can for instance be obtained via the SVD of $\sum_{n_1=1}^{N}\sum_{n_2=1}^{N} \mathbf{P}_{\boldsymbol{X}_{n_1}}^{\perp}\mathbf{P}_{\boldsymbol{X}_{n_2}}^{\perp}$, i.e., the columns of $\mathbf{A}$ correspond to the R right singular vectors associated with the R smallest singular values of $\sum_{n_1=1}^{N}\sum_{n_2=1}^{N} \mathbf{P}_{\boldsymbol{X}_{n_1}}^{\perp}\mathbf{P}_{\boldsymbol{X}_{n_2}}^{\perp}$. However, when I is large, even computing $\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp}$, can be computationally challenging at $\mathcal{O}(I^2(N+\sum_{n=1}^{N}(R+L_n)))$. Furthermore, storing $\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp}$ can also be prohibitive, especially when I is large.[5] In this section we develop a different algebraic range subspace intersection method to tackle the GCCA problem which can handle large values of I.

We will now present an alternative algorithm for computing a basis for the common subspace range$(\boldsymbol{A})$ associated with the factor matrix in (9) and the matrices $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_N$ in the GCCA model (8). Since the algorithm is not directly based on (45), it does not require the construction of the projectors $\mathbf{P}_{\boldsymbol{X}_n}^{\perp}$. The algorithm follows the previous analysis and can be viewed as a constructive interpretation of Theorem IV.3. It can be described in 3 steps:

step 1: Compute $\boldsymbol{U}_n \in \mathbb{C}^{I\times(R+L_n)}$ and $\boldsymbol{V}_n \in \mathbb{C}^{J_n\times(R+L_n)}$ whose columns form an orthonormal basis for range $(\boldsymbol{X}_n)$ and range $\left(\boldsymbol{X}_n^T\right)$, $n \in \{1, \ldots, N\}$ respectively. In practice, the matrices $\boldsymbol{X}_1, \ldots \boldsymbol{X}_N$ are often perturbed by additive noise. For this reason, we use the SVD to compute $\boldsymbol{U}_n$ and $\boldsymbol{V}_n$.

step 2: Using the SVD, we compute an orthonormal basis for $Z^{(N)}$ in (22) and retrieve the matrices $\boldsymbol{\Phi}_n$, $n = 1, \ldots, N$. To do that we first construct matrix $\boldsymbol{\Theta} \in \mathbb{C}^{\binom{N}{2}I\times(NR+\sum_{n=1}^{N} L_n)}$ (see (23) for an example when $N = 3$) and compute a basis for its null space via the SVD, represented by matrix $\boldsymbol{Q} = [\boldsymbol{Q}_1^T, \ldots, \boldsymbol{Q}_N^T]^T$, with $\boldsymbol{Q}_n \in \mathbb{C}^{(R+L_n)\times R}$. The columns of $\boldsymbol{Q}$ correspond to the R right singular vectors associated with the R smallest singular vectors of $\boldsymbol{\Theta}$. Note that in the exact case the columns of $\boldsymbol{Q}$ form an orthonormal basis for $Z^{(N)}$. Since, in the ideal case, $\boldsymbol{A} = \boldsymbol{U}_n\boldsymbol{Q}_n = \boldsymbol{X}_n\boldsymbol{\Phi}_n$, the matrices $\boldsymbol{\Phi}_n$, $n = 1, \ldots, N$ can now be obtained as $\boldsymbol{\Phi}_n = \boldsymbol{X}_n^{\dagger}\boldsymbol{U}_n\boldsymbol{Q}_n = \boldsymbol{V}_n\boldsymbol{\Sigma}_n^{-1}\boldsymbol{U}_n^H\boldsymbol{U}_n\boldsymbol{Q}_n = \boldsymbol{V}_n\boldsymbol{\Sigma}_n^{-1}\boldsymbol{Q}_n$, where $\boldsymbol{X}_n^{\dagger}$ denotes the left-inverse of $\boldsymbol{X}_n$ and $\boldsymbol{\Sigma}_n \in \mathbb{C}^{(R+L_n)\times(R+L_n)}$ is the diagonal matrix with the singular values of $\boldsymbol{X}_n$ on its diagonal.

step 3: In the exact case we have range$(\boldsymbol{A})$ = range$(\boldsymbol{U}_n\boldsymbol{Q}_n)$ when condition (28) in Theorem IV.3 is satisfied. In the inexact case, a more robust estimate of range$(\boldsymbol{A})$ can be obtained via the matrix $\boldsymbol{G} = [\boldsymbol{U}_1\boldsymbol{Q}_1, \ldots, \boldsymbol{U}_N\boldsymbol{Q}_N]$. In short, let the columns of $\boldsymbol{A}$ correspond to the R left singular vectors of $\boldsymbol{G}$ associated with the R largest singular values of $\boldsymbol{G}$, then (in the exact case) the columns of $\boldsymbol{A}$ form an orthonormal basis for the common subspace. The detailed steps can be found in Algorithm 1. Note that in Algorithm 1 we use two versions of SVD. To be specific, svd computes the 'thin' SVD, whereas $\text{svd}_t(\cdot, r)$ computes the truncated SVD corresponding to the r largest singular values.

Algorithm 1: RAnge subspaCe INtersection for Gcca (RACING)

1: **Input:** $\{\boldsymbol{X}_n, L_n\}_{n=1}^{N}$, R.
2: **Output:** $\boldsymbol{A}, \{\boldsymbol{\Phi}_n\}_{n=1}^{N}$.
3: **step 1:**
4: **for** $n = 1$ **to** N **do**
5: $\quad \boldsymbol{U}_n\boldsymbol{\Sigma}_n\boldsymbol{V}_n^T \leftarrow \text{svd}_t\left(\boldsymbol{X}_n, R + L_n\right)$;
6: **end for**
7: **step 2:**
8: $\boldsymbol{\Theta} = [\cdot]$;
9: **for** $n_1 = 1$ **to** $N-1$ **do**
10: $\quad$ **for** $n_2 = n_1 + 1$ **to** N **do**
11: $\quad\quad k = (n_1 - 1)R + \sum_{n=1}^{n_1-1} L_n$, $l = (n_2 - n_1 - 1)R + \sum_{n=n_1+1}^{n_2-1} L_n$, $m = (N - n_2)R + \sum_{n_2+1}^{N} L_n$;
12: $\quad\quad \boldsymbol{\Theta}_{n_1n_2} = [\mathbf{0}_{I\times k}, \boldsymbol{U}_{n_1}, \mathbf{0}_{I\times l}, -\boldsymbol{U}_{n_2}, \mathbf{0}_{I\times m}]$;
13: $\quad\quad \boldsymbol{\Theta} \leftarrow \begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{\Theta}_{n_1n_2} \end{bmatrix}$;
14: $\quad$ **end for**
15: **end for**
16: $\boldsymbol{U}_\theta\boldsymbol{\Sigma}_\theta\boldsymbol{V}_\theta^T \leftarrow \text{svd}\left(\boldsymbol{\Theta}\right)$;
17: $\boldsymbol{Q} = [\boldsymbol{Q}_1^T, \ldots, \boldsymbol{Q}_N^T]^T = \boldsymbol{V}_\theta(:, \text{end} - R + 1 : \text{end})$;
18: **step 3:**
19: $\boldsymbol{G} = [\cdot]$;
20: **for** $n = 1$ **to** N **do**
21: $\quad \boldsymbol{G} \leftarrow [\boldsymbol{G}, \boldsymbol{U}_n\boldsymbol{Q}_n]$;
22: $\quad \boldsymbol{\Phi}_n = \boldsymbol{V}_n\boldsymbol{\Sigma}_n^{-1}\boldsymbol{Q}_n$;
23: **end for**
24: $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \leftarrow \text{svd}_t\left(\boldsymbol{G}, R\right)$;
25: $\boldsymbol{A} = \boldsymbol{U}$;

In terms of computational complexity and memory requirements, the main bottleneck of the proposed algorithm lies in computing the SVD in line 5 and 16. The column dimension (J_n) of each view $\boldsymbol{X}^{(n)}$ is usually large which makes the SVD computation very intensive. Traditional algorithms require $\mathcal{O}(IJ_n\min(I, J_n))$ flops to compute the SVD of line 5 and $\mathcal{O}(N(N-1)I(NR + \sum_1^N L_n)^2)$ flops to compute the SVD of line 16 and computation is prohibitive when big and high dimensional data are involved. To overcome this issue, we propose to employ Lanczos-type iterative algorithms [33] (e.g., Matlab's routine svds) to compute the truncated SVD in line 5. The complexity then is depending on the number of principal components $R + L_n$, therefore setting L_n to be relatively small compared to the dimensions $(R + L_n \ll I)$ markedly reduces the computational complexity, especially for sparse data. The reason is that these Lanzos-type approaches involve multiplications of $\mathbf{X}_n$ with a $(J_n \times R+L_n)$ matrix. When $\mathbf{X}_n$ is sparse this multiplication can be carried out significantly faster compared to the case of $\mathbf{X}_n$ being dense. In practice, noise makes $\mathbf{X}_n$ full rank $(I = R+L_n)$. However, $\mathbf{X}_n$ still typically admits a good low-rank approximation $\mathbf{X}_n \approx [\boldsymbol{A}, \boldsymbol{C}_n]\boldsymbol{S}_n^T$, where $I > R+L_n$. We say that the $R+L_n$ represents the useful signal rank, i.e., the dimension of the signal subspace, which

[4] Let $\mathbf{G} \in \mathbb{C}^{N\times N}$ denote the Grammian matrix with entries $g_{ij} = \text{vec}(\mathbf{P}_{\boldsymbol{X}_i}^{\perp})^H\text{vec}(\mathbf{P}_{\boldsymbol{X}_j}^{\perp})$, where $\text{vec}(\mathbf{P}_{\boldsymbol{X}_i}^{\perp})$ denotes the vectorized version of $\mathbf{P}_{\boldsymbol{X}_i}^{\perp}$. Then $\|\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}^{\perp}\mathbf{a}\|_F^2 = 0$ for some nonzero $\mathbf{a} \in \mathbb{C}^N$ if and only if the determinant of $\mathbf{G}$ is equal to zero, which means that the vectors $\text{vec}(\mathbf{P}_{\boldsymbol{X}_1}^{\perp}), \ldots, \text{vec}(\mathbf{P}_{\boldsymbol{X}_N}^{\perp})$ are linearly dependent.

[5] Define $\boldsymbol{Q} = \boldsymbol{I}_I - \frac{1}{N}\sum_{n=1}^{N} \mathbf{P}_{\boldsymbol{X}_n}$. Then $\boldsymbol{P}_Y = \boldsymbol{I}_I - \boldsymbol{Q}^{\dagger}\boldsymbol{Q}$ is a projector onto Y, where $\boldsymbol{Q}^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{Q}$; see [32] for a proof. Hence, as an alternative to (45), $\mathbf{A}$ could also be determined via the SVD of $\boldsymbol{P}_Y$. However, this approach also requires the construction of the $(I \times I)$ matrix $\boldsymbol{P}_Y$.

is small enough for every $n \in \{1, \ldots, N\}$. Small values of L_n also reduce the computations required in line 16 significantly. For example, choosing $R + L_n$ to be in the order of 500 will allow the proposed algorithm to work for very large and high-dimensional data.

We mention that a relaxed SUMCOR-type algorithm has also been proposed in [34], [35] that is similar to Algorithm 1. However, there are notable differences that we will now point out. First, the starting points of the derivations are very different. While Algorithm 1 follows immediately from the proposed subspace intersection interpretation of GCCA, the approach in [34], [35] is based on a relaxation of the SUMCOR cost function (6). Second, an important difference is that Algorithm 1 and the SUMCOR/SUMCOR-type methods fit different models. The former method looks for a "common subspace" range($\mathbf{A}$) while the latter methods look for individual subspaces range($\boldsymbol{X}_1\boldsymbol{\Phi}_1$), ..., range($\boldsymbol{X}_N\boldsymbol{\Phi}_N$). Third, a difference is that in Step 3 in Algorithm 1 SVDs are used to obtain more robust estimates of $\mathbf{A}$ and $\{\boldsymbol{\Phi}_n\}$ while this is not the case in [34], [35]. In particular, we are interested in $\mathbf{A}$ while in [34], [35], just as in SUMCOR, the focus is on the computation of $\{\boldsymbol{\Phi}_n\}$. Fourth, define $\mathbf{U}^{(N)} = [\mathbf{U}_1, \ldots, \mathbf{U}_N] \in \mathbb{C}^{I \times (NR + \sum_{n=1}^{N} L_n)}$. Then in [34], [35] matrices $\{\boldsymbol{\Phi}_n\}$ are computed via the right singular vectors of $\mathbf{U}^{(N)}$. In Algorithm 1, $\{\boldsymbol{\Phi}_n\}$ are computed via the right singular vectors of $\boldsymbol{\Theta}$. Note that this leads to a different weighting of the data. More precisely, the approach in [34], [35] corresponds to computing a basis for the kernel of $\mathbf{U}^{(N)H}\mathbf{U}^{(N)}$ with block matrices $\mathbf{U}_m^H\mathbf{U}_n$ while the approach in Algorithm 1 corresponds to computing a basis for the kernel of $\boldsymbol{\Theta}^H\boldsymbol{\Theta}$ with on-diagonal blocks $(N-1)\mathbf{U}_m^H\mathbf{U}_m$ and off-diagonal blocks $-\mathbf{U}_m^H\mathbf{U}_n$.

In the next sections we compare the popular MAXVAR method for GCCA computation with the proposed subspace intersection based approach.

B. Comparison between MAXVAR and subspace intersection.

The GCCA method MAXVAR aims to find range($\mathbf{A}$) by minimizing the cost function

$$\min_{\boldsymbol{\Phi}_1,\ldots,\boldsymbol{\Phi}_N,\mathbf{A}} \sum_{n=1}^{N} \|\mathbf{X}_n\boldsymbol{\Phi}_n - \mathbf{A}\|_F^2, \quad \text{s.t.} \quad \mathbf{A}^H\mathbf{A} = \mathbf{I}_R. \quad (46)$$

Let the columns of $\mathbf{U}_n \in \mathbb{C}^{I \times (R+L_n)}$ form a columnwise orthonormal basis for range($\mathbf{X}_n$). Then range($\mathbf{A}$) can be obtained via

$$\min_{\mathbf{V}_1,\ldots,\mathbf{V}_N,\mathbf{A}} \sum_{n=1}^{N} \|\mathbf{U}_n\mathbf{V}_n - \mathbf{A}\|_F^2, \quad \text{s.t.} \quad \mathbf{A}^H\mathbf{A} = \mathbf{I}_R, \quad (47)$$

where $\mathbf{V}_n \in \mathbb{C}^{(R+L_n)\times R}$ is an unknown full column rank matrix. However, it is not evident when range($\mathbf{A}$) is correctly computed via the cost function (47). We will now explain that in the exact case, where $\mathbf{X}_n$ admits the decomposition (9), MAXVAR correctly computes range($\mathbf{A}$) via subspace intersection when $\dim(Y) = R$. The minimizer of (47) corresponds to the maximizer of

$$\max_{\mathbf{A}^H\mathbf{A}=\mathbf{I}_R} \sum_{n=1}^{N} \left\|\mathbf{U}_n^H\mathbf{A}\right\|_F^2 = \max_{\mathbf{A}^H\mathbf{A}=\mathbf{I}_R} \sum_{n=1}^{N} \|\mathbf{P}_{\mathbf{X}_n}\mathbf{A}\|_F^2 \quad (48)$$

with $\mathbf{V}_n = \mathbf{U}_n^H\mathbf{A}$ and where $\mathbf{P}_{\mathbf{X}_n}$ denotes the orthogonal projector onto the subspace spanned by the columns of $\mathbf{X}_n$. The orthogonal decomposition theorem tells us the maximizer of (48) corresponds to the minimizer of

$$\min_{\mathbf{A}^H\mathbf{A}=\mathbf{I}_R} \sum_{n=1}^{N} \left\|\mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}\right\|_F^2 \geq 0, \quad (49)$$

where $\mathbf{P}_{\mathbf{X}_n}^{\perp}$ denotes the orthogonal projector onto the orthogonal complement of the subspace spanned by the columns of $\mathbf{X}_n$. Note that the lower bound in (49) can only be attained when range($\boldsymbol{A}$) $\subseteq$ ker($\sum_{n=1}^{N} \mathbf{P}_{\mathbf{X}_n}^{\perp}$). (This fact follows from inequality (45) and inequality (50) below.) In the exact case it means that we are looking for a columnwise orthonormal matrix $\mathbf{A}$ that satisfies relation (44), where we exploited that range($\mathbf{A}$) $\subseteq$ ker($\sum_{n=1}^{N} \mathbf{P}_{\mathbf{X}_n}^{\perp}$) $= Y$, in which the latter equality is due to (43). This means that if $\dim(Y) = R$, then the lower bound in (49) is attained if and only if the columns of $\mathbf{A}$ form a basis for Y. It is now evident that in the exact case, the MAXVAR solution corresponds to a basis for the intersecting subspace Y, i.e., $Y =$ range($\mathbf{A}$).

We note in passing that efficient implementations of the MAXVAR method have been proposed (e.g., [9], [35]). Briefly, the solution to the MAXVAR problem (48) can be obtained via the eigenvalue decomposition (EVD) of $\sum_{n=1}^{N} \mathbf{U}_n\mathbf{U}_n^H$. However, since $\sum_{n=1}^{N} \mathbf{U}_n\mathbf{U}_n^H = \mathbf{U}^{(N)}\mathbf{U}^{(N)H}$, where $\mathbf{U}^{(N)} = [\mathbf{U}_1, \ldots, \mathbf{U}_N] \in \mathbb{C}^{I \times (NR + \sum_{n=1}^{N} L_n)}$, and rank($\mathbf{U}^{(N)}\mathbf{U}^{(N)H}$) = rank($\mathbf{U}^{(N)}$), the solution to the MAXVAR problem (48) can be computed more efficiently via the SVD of $\mathbf{U}^{(N)}$, without first computing the orthogonal projectors $\{\mathbf{U}_n\mathbf{U}_n^H\}$. More precisely, if $\dim(\text{range}(\mathbf{A})) = R$, then the R left singular vectors of $\mathbf{U}^{(N)}$ associated with the R largest singular values of $\mathbf{U}^{(N)}$ form a basis for Y.

We will now argue that in the inexact case, MAXVAR and the proposed subspace intersection approach for GCCA can lead to different but related solutions. From (45) and (49) we observe that the difference between MAXVAR and subspace intersection is that the former method aims to minimize the term $\sum_{n=1}^{N} \left\|\mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}\right\|_F^2$ while the latter method aims to minimize the term $\|\sum_{n=1}^{N} \mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}\|_F^2$. The triangle inequality tells us that

$$\sum_{n=1}^{N} \left\|\mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}\right\|_F^2 \geq \left\|\sum_{n=1}^{N} \mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}\right\|_F^2. \quad (50)$$

Hence, from (50) we observe that in the inexact case, the MAXVAR solution can be interpreted as an approximate solution to the subspace intersection problem in which cross terms of the form trace($\mathbf{A}^H\mathbf{P}_{\mathbf{X}_m}^{\perp}\mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}$) with $m \neq n$ in $\|\sum_{n=1}^{N} \mathbf{P}_{\mathbf{X}_n}^{\perp}\mathbf{A}\|_F^2$ are ignored.

To summarize, using the identity in (43) we argued that in the exact case MAXVAR is performing subspace intersection. Consequently, the link between subspace intersection and GCCA presented in Section III tells us that if the identifiability condition (31) in Theorem IV.4 is satisfied, then in the exact case MAXVAR correctly computes range($\mathbf{A}$). In the inexact case, inequality (50) tells us that MAXVAR can be interpreted as an approximate method for subspace intersection.

VII. Experiments

In this section we demonstrate the performance of the proposed algorithmic framework and showcase its effectiveness in synthetic- and real-data experiments. All simulations are implemented in Matlab and are executed on a Linux server comprising 32 cores at 2GHz and 128GB RAM.

A. Synthetic-Data Experiments

First we test the proposed framework using experiments with synthetically generated data. The multiple views are generated according to equation (9). We assume that the views share a common latent factor $\boldsymbol{A} \in \mathbb{C}^{I \times R}$ with entries randomly and independently drawn from a zero-mean unit-variance Gaussian distribution. The individual matrices $\boldsymbol{C}_n \in \mathbb{C}^{I \times L_n}$ and $\boldsymbol{S}_n \in \mathbb{C}^{K_n \times (R+L_n)}$ are also generated with entries independently drawn from a zero-mean unit-variance Gaussian distribution and for simplicity we set $L_n = L$ and $K_n = K = L + R$ for every $n \in \{1, \ldots, N\}$.

We test the algorithm in a noisy setup. To be more specific, $\boldsymbol{X}_n,\ n = 1, \ldots, N$ are generated according to the model in (9), as previously described. However, instead of $\boldsymbol{X}_n$ we observe $\boldsymbol{Y}_n,\ n = 1, \ldots, N$ which are generated as: $\boldsymbol{Y}_n = \boldsymbol{X}_n + \boldsymbol{W}_n,\ n \in \{1, \ldots, N\}$, where $\boldsymbol{W}_n$ is an additive white Gaussian noise term. Note that if condition (31) in Theorem IV.4 is satisfied, then the proposed algorithm is guaranteed to find range$(\boldsymbol{A})$ in the exact case.

For baselines we use the exact solution of MAXVAR formulation, computed via eigenvalue decomposition and CSR, which solves the SUMCOR formulation, using a change of variables and a block coordinate descent (BCD) approach [10]. CSR is an iterative algorithm and is initialized randomly. We also include comparisons with CSR initialized with RACING which we refer to as RACING-CSR. To evaluate the performance, we measure the angle between the generated common subspace and the estimated one as defined in [36], [37], i.e.,

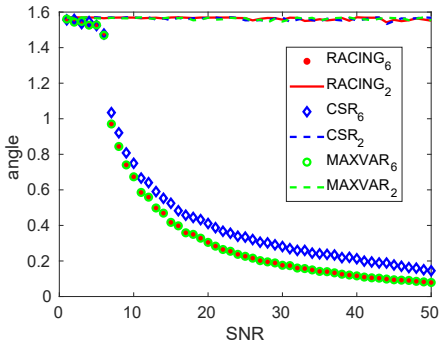
$$\texttt{angle}(\boldsymbol{A}, \widehat{\boldsymbol{A}}) = \sin^{-1}\left(\|\boldsymbol{P}_{\boldsymbol{A}} - \boldsymbol{P}_{\widehat{\boldsymbol{A}}}\|_2\right), \qquad (51)$$

where $\|\ \|_2$ denotes the Euclidean norm, $\boldsymbol{P}_{\boldsymbol{A}}$ is the orthogonal projector onto the subspace spanned by the columns of $\boldsymbol{A}$ and $\boldsymbol{P}_{\widehat{\boldsymbol{A}}}$ is the orthogonal projector onto the subspace spanned by the columns of $\widehat{\boldsymbol{A}}$.

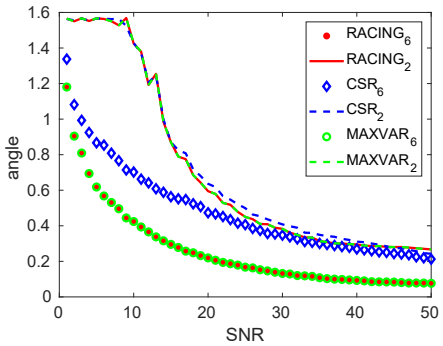
We consider $N = 6$ different views, that share a common subspace of dimension $R = 50$. Two scenarios are generated as follows. In the first each view consists of $I = 2000$ rows, and $L = 1000$ that leads to $K = 1050$ columns for each view, whereas in the second $I = 2000$, $L = 500$ that leads to $K = 550$ columns for each view. We test the algorithmic performance for different levels of *signal-to-noise-ratio* (SNR), which is defined as:

$$SNR = 20 \log \frac{\sum_{n=1}^{N} \|\boldsymbol{X}_n\|_F}{\sum_{n=1}^{N} \|\boldsymbol{W}_n\|_F}.$$

Fig. 1(a) shows the performance of the proposed RACING along the baselines for different levels of SNR in the first scenario. Note that each algorithm is implemented to utilize either all 6 views to identify the common space, or the first 2 views, denoted by the subscript next to the name of the algorithm. We observe that the proposed algorithm is able to identify the common subspace for a wide range of SNRs, when 6 views are utilized. On the contrary all algorithms fail to identify the correct subspace, when only 2 views are employed. Note that the identifiability condition in (36) yields $50 + \frac{N}{N-1}1000 \leq 2000$, which is satisfied for $N = 6$ but fails when $N = 2$.



(a) First scenario, $L = 1000$.

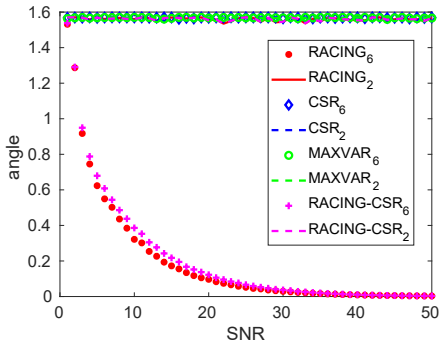


(b) Second scenario, $L = 500$.

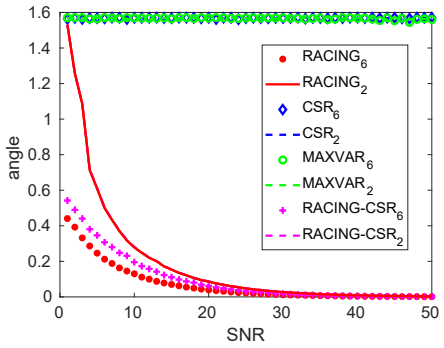
Fig. 1: Angle between true and recovered subspace.

In the second scenario we reduce the dimension of the columnspace of each view to $K = 550$. In this case the identifiability condition in (36) yields $50 + \frac{N}{N-1}500 \leq 2000$, which is satisfied for both $N = 6$ and $N = 2$. The results are illustrated in Fig. 1(b). We observe that although the identifiability condition in (36) is satisfied in both cases where 2 and 6 views are utilized, the algorithms perform better in the 6-view implementation. From both experiments we can also deduce that the proposed RACING works similarly to the MAXVAR solution and significantly outperforms CSR. This is a notable, considering that both MAXVAR and CSR are optimization approaches and are expected to perform better in the presence of noise. Note that RACING-CSR was omitted from Fig. 1 because it yielded the same performance as CSR.

Next we test the performance of the proposed approach and the baselines in the case where the signal rank $(R + L_n)$ of the views is smaller than the dimensions, i.e., $R + L_n < \min\{I, K_n\}$. This way we generate views that have full rank, but the signal part $\boldsymbol{X}_n$ has low rank. This is very often the case in practice, since although real data are typically full rank due to noise and measurement errors, the useful signal rank is often lower, and the remaining components are mostly

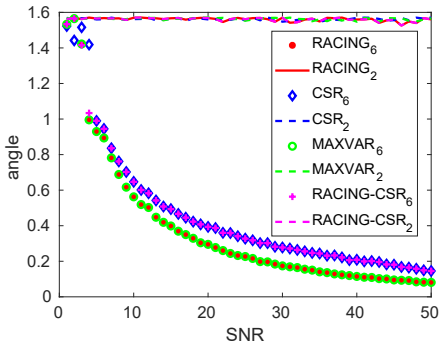


(a) First scenario with low-rank signal part.

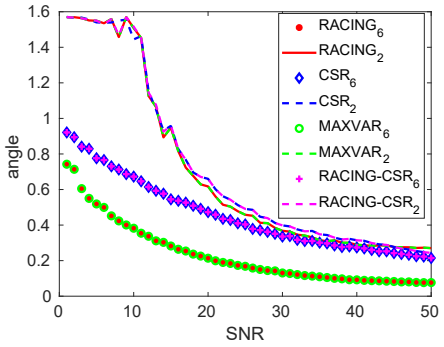


(b) Second scenario with low-rank signal part.

Fig. 2: Angle between true and recovered subspace.



(a) First scenario, $L = 1000$.



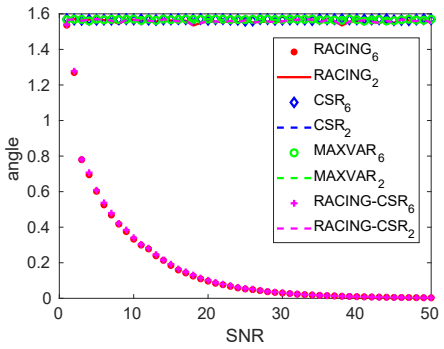
(b) Second scenario, $L = 500$.

Fig. 3: Angle between true and recovered subspace in the presence of outliers.

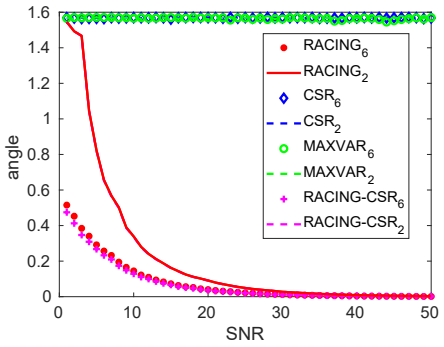
noise. To this end, we generate $\boldsymbol{A}$, $\boldsymbol{C}_n$ as before in scenario 1 and 2 ($I = 2000$, $L_n = L = 1000$, $R = 50$ and $I = 2000$, $L_n = L = 500$, $R = 50$ respectively), but this time we allow $\boldsymbol{X}_n$ to have low rank by letting $\boldsymbol{S}_n$ to be 'tall matrices', i.e., $\boldsymbol{S}_n \in \mathbb{C}^{K\times(R+L)}$, with $K = 1900$. We add noise as before, so the views are technically full rank, but when the noise is small they are 'approximately low-rank' – i.e., they can be well-approximated by low-rank matrices. The results are presented in Fig. 2.

It is clear from Fig. 2 that views with low-rank signal part do not affect the performance of the proposed RACING. However, MAXVAR and SUMCOR (CSR) formulations fail to identify the common subspace. This can be explained from the fact that the MAXVAR and SUMCOR analysis and algorithm assume that the views are effectively full rank as mentioned earlier. On the contrary, the proposed RACING allows prescribing the useful signal rank of each view. Furthermore, there is clear benefit when initializing CSR with RACING rather than randomly, which suggests that our proposed RACING can work as a great initialization for iterative approaches.

We also test the performance of the proposed approach and the baselines in the presence of outliers. To this end we generate the matrix views $\mathbf{Y}_n = \mathbf{X}_n + \mathbf{W}_n$, $n = 1, \ldots, N$ as before, but this time $\mathbf{W}_n$ is a sparse matrix with sparsity level in the order of 10^{-3} and non-zeros drawn from a Gaussian distribution. We again consider full column rank and low rank views $\mathbf{X}_n$ and assess the performance for $L = 500,\ 1000$. The results for different levels of SNR are presented in Figs. 3 and 4.



(a) First scenario, $L = 1000$.



(b) Second scenario, $L = 500$.

Fig. 4: Angle between true and recovered subspace in the presence of outliers.

Similar conclusions can be derived with the previous experiments. Our proposed RACING works the best in all experiments. When all the $\mathbf{X}_n$ have full column rank MAXVAR works similarly to RACING and CSR achieves worse but acceptable performance. When $\mathbf{X}_n$'s have low rank, MAXVAR is not working and CSR needs to be initialized by RACING to perform well. We also observe that employing more views is beneficial even in the case where the identifiability condition is satisfied for both 6 and 2 views. In particular when $L = 500$ and the SNR is small there is a clear benefit of using 6 views rather than 2 views of the data.

B. Cross Language Information Retrieval

Finally, we test the proposed approach on the task of cross language information retrieval (CLIR). CLIR is a natural language processing application, where given a set of sentences along with their translations in multiple languages the goal is to learn a low-dimensional subspace where the sentences and their translations are maximally correlated. Then, new high-dimensional sentences are mapped to the associated lower dimensional space in order to retrieve their translation from a database of possible choices. CLIR is essential to fast query and search across languages, which also benefits machine translation systems [38]–[40].

Data: The dataset employed is the Europarl parallel corpus [41]. It contains a collection of sentences translated in 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek. In the experiments we focus on the Germanic languages, i.e., English, Dutch, German, Danish and Swedish. Each sentence is represented by $J = 267,752$ feature vector of 'bag of words' composed with inner-product preserving hashing [42], [43]. In particular, we use 2^{19} hash slots as in [43] and remove features that are empty in all views. Then a set of sentences in a specific language can be represented as a matrix view, $\mathbf{X}_n \in \mathbb{R}^{I \times J_n}$, where I is the total number of sentences and $J_1, \ldots, J_N = J$ corresponds to the feature dimension of each sentence, which is described above.

Procedure: The objective of CLIR is to align sentences with their translations. In order to do that we apply GCCA on a set of $I = 153,403$ training sentences, learn their common low-dimensional subspace $\mathbf{A}$ along with the matrices $\mathbf{\Phi}_n \ n = 1, \ldots, N$ that map the sentences to the common subspace. The idea is that sentences and their translations in different languages have a common low-dimensional representation, which is not-language specific and capture the semantic meaning of the sentence. In other words, GCCA enforces sentences to be maximally correlated with their translations and less correlated with other sentences. In the training phase in addition to learning low-dimensional representations of the sentences in the common subspace, we also learn the mapping from the high-dimensional to the common low-dimensional space. This mapping is represented by matrices $\mathbf{\Phi}_n \ n = 1, \ldots, N$. These matrices are then used to map a testing set of $I_t = 38,351$ sentences and their translations to the common low-dimensional space (Subscript 't' denotes the testing set). Note that the testing sentences and their translations are not aligned. The mapping is performed by multiplying each sentence in the $n-$th language with $\mathbf{\Phi}_n$. This mapping provides a low-dimensional representation of each testing sentence in every language. Since GCCA was employed to learn these mappings we expect the query sentences to be maximally correlated with their translations in the testing set. The CLIR task is completed by matching the query sentences with their translations, according to their Euclidean distance in the low-dimensional subspace. For example, suppose we are interested in performing a CLIR task between sentences in Dutch and Danish. In the training phase we learn $\mathbf{\Phi}_{Dutch}$ and $\mathbf{\Phi}_{Danish}$. Then we embed each testing sentence in Dutch and Danish using $\mathbf{\Phi}_{Dutch}$ and $\mathbf{\Phi}_{Danish}$. For each testing sentence in Dutch (or Danish) we find the closest embedded sentence in Danish (or Dutch) according to the Euclidean distance in the embedding space. In simple words, the training phase learns how to embed high dimensional sentences from different languages and the testing phase aligns the embedded sentences with their translations.

We consider two scenarios. In the first one, training and testing are performed using only $N = 2$ languages, i.e., Dutch and Danish. In the second scenario, the training phase takes into account all Germanic languages (Dutch, Danish, English, German, Swedish) to learn $\mathbf{\Phi}_n, \ n = 1, \ldots, N$.

Evaluation: The baseline algorithms used for comparison are MVLSA [9], which is an approximate eigen-based solver for the MAXVAR criterion in large-scale settings and PDD−GCCA [11], which is a primal-dual algorithm that tackles the SUMCOR formulation for sparse large-scale data. We initialize PDD−GCCA, with RACING and ran for 25 iterations (total number of 5 inner and 5 outer iterations). Note that, RACING and MVLSA are primarily focused on learning the common subspace $\boldsymbol{A}$ and the solution for $\mathbf{\Phi}_n, \ n = 1, \ldots, N$ is suboptimal due to the low-rank assumption. Since CLIR is mainly concerned with effectively learning $\mathbf{\Phi}_n, \ n = 1, \ldots, N$, we add one more step in RACING and MVLSA which resulted in improved performance for both algorithms. In particular, after learning the common subspace $\boldsymbol{A}$, $\mathbf{\Phi}_n, \ n = 1, \ldots, N$ are computed by solving:

$$\min_{\{\mathbf{\Phi}_n\}_{n=1}^N} \sum_{n=1}^{N} \|\boldsymbol{X}_n \mathbf{\Phi}_n - \boldsymbol{A}\|_F^2. \tag{52}$$

The solution of (52) is efficiently obtained via 20 conjugate gradient iterations [44].

To assess the performance of the competing algorithms we measure the average recall@k for $k \in \{10, 20, 50\}$ and the average area under ROC curve (AUC). Recall is defined as the number of relevant sentences among the retrieved k divided by the number of total relevant sentences. Since there is only one relevant sentence (the translation) recall@k is equal to 1 if the translation is ranked among the top k sentences and 0 otherwise. As a result the average recall@k indicates the probability of the correct translation to be ranked among the top k hits. The ROC curve plots the true positive rate, which is the recall, against the false positive rate, which is defined as the

number of non-relevant sentences among the retrieved divided by the total number of nonrelevant sentences. Then the AUC is defined as $AUC = 1 - \frac{\text{pos}-1}{I_t-1}$, where pos is the position at which the correct translation is ranked. Therefore, the average AUC is a ranking metric indicative for the position of the correct translation. Detailed description of the two metrics can be found in [45, Chapter 8].

Results: Tables II and III show the performance of the competing algorithms for the two scenarios of CLIR. Mean and standard deviation are reported over 10 randomly drawn $80-20$ splits for training and testing. The dimension of the subspace where each sentence is mapped varies from $R = 1$ to $R = 50$ with $L_n + R = 300$ for all views.

One can see that the CLIR task significantly benefits from incorporating multiple languages, which is also justified from our theoretical analysis. To be more precise, all metrics show performance improvement when all Germanic languages are employed. For example `RACING` achieves an improvement of approximately $7-9\%$ in recall@k and 1% in AUC, when all Germanic languages are used during training and $R = 50$. Furthermore, we observe that $PDD-GCCA$ initialized by `RACING` outperforms the other two methods, whereas `RACING` works better than `MVLSA` for $R = 1, 5, 10, 20$ and comparably for $R = 50$. We also observe that there is a trade-off between the dimension of the embedding and the number of sentences one should retrieve in order to find the correct translation. On the one hand very low dimensional embeddings (small values of R) are desirable to reduce the complexity of computing similarity between sentences. On the other hand, larger values of R, result in better retrieval performance, i.e., the correct translation ranks higher.

VIII. Conclusion

In this paper we studied GCCA from a linear algebraic perspective. In particular, we showed that GCCA can be interpreted as subspace intersection and provided identifiability conditions for recovering the common subspace between the views, which are relaxed compared to the standard two-view CCA. We also developed a range subspace intersection algorithm to perform GCCA, which can also handle large and high-dimensional datasets. Numerical experiments demonstrated the effectiveness of the proposed approach in the context of multi-view learning.

Appendix A
Proof of equation (11)

In order to prove equation (11) we first need to prove that the following properties hold without loss of generality:

Property 1: $\boldsymbol{A}$, $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$ have full column rank. Without loss of generality we can assume that the matrices $\boldsymbol{A} \in \mathbb{C}^{I\times R}, \boldsymbol{C}_1 \in \mathbb{C}^{I\times L_1}, \ldots, \boldsymbol{C}_N \in \mathbb{C}^{I\times L_N}$ in (9) all have full column rank. Indeed, if the columns of $\boldsymbol{A}$ are linearly dependent, then $\boldsymbol{A}$ can replaced by any subset of its columns that form a basis for range($\boldsymbol{A}$) and the matrix $\mathbf{S}_n$ can be adjusted accordingly, without changing $\mathbf{X}_n$. (Similarly for $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_N$).

Property 2: $[\boldsymbol{A}, \boldsymbol{C}_1], \ldots, [\boldsymbol{A}, \boldsymbol{C}_N]$ have full column rank. First, note that the column dimension $(R + L_n)$ of $[\boldsymbol{A}, \boldsymbol{C}_n]$ should not exceed its row dimension I, i.e., $I \geq R + L_n$. Indeed, if $I < R + L_n$, then $R + L_n - I$ columns of $\boldsymbol{C}^{(n)}$ can be written as linear combinations of the other I columns in $[\boldsymbol{A}, \boldsymbol{C}_n]$. Therefore these $R + L_n - I$ columns of $\boldsymbol{C}_n$ could be discarded, while accordingly adjusting matrix $\mathbf{S}^{(n)}$, without changing $\mathbf{X}_n$. Furthermore, we can w.l.o.g. assume that range($\boldsymbol{A}$) $\cap$ range($\boldsymbol{C}_n$) $= \{\mathbf{0}\}$, $n \in \{1, \ldots, N\}$, i.e., we assume w.l.o.g. that $(\boldsymbol{c}_n)_q \notin$ range($\boldsymbol{A}$), $q \in \{1, \ldots, L_n\}$, $n \in \{1, \ldots, N\}$, where $(\boldsymbol{c}_n)_q$ denotes the q-th column of $\boldsymbol{C}_n$. Indeed, if $(\boldsymbol{c}_n)_t = \boldsymbol{A}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{C}^R$, then

$$\boldsymbol{X}_n = \sum_{r=1}^{R} \boldsymbol{a}_r(\boldsymbol{s}_n)_r^T + \sum_{q=1}^{L_n}(\boldsymbol{c}_n)_q(\boldsymbol{s}_n)_r^T = \sum_{r=1}^{R} \boldsymbol{a}_r((\boldsymbol{s}_n)_r^T + \beta_r(\boldsymbol{s}_n)_{R+t}^T) + \sum_{\substack{q=1\\q\neq t}}^{L_n}(\boldsymbol{c}_n)_q(\boldsymbol{s}_n)_r^T, \tag{53}$$

where $(\boldsymbol{s}_n)_r$ denotes the r-th column of $\boldsymbol{S}_n$. In other words, if $(\boldsymbol{c}_n)_q \in$ range($\boldsymbol{A}$), then we can simply consider a factorization of $\boldsymbol{X}_n$, as in (53), that only involves a smaller I-by-$(L_n - 1)$ matrix $\boldsymbol{C}_n$. Now since range($\boldsymbol{A}$) $\cap$ range($\boldsymbol{C}_n$) $= \{\mathbf{0}\}$, $n \in \{1, \ldots, N\}$ and $\boldsymbol{A}$, $\boldsymbol{C}_n$ have full column rank, we conclude that w.l.o.g. $[\boldsymbol{A}, \boldsymbol{C}_n]$ has full column rank.

Relation (11) now follows naturally from Property 2. i.e., the fact that matrix $[\boldsymbol{A}, \boldsymbol{C}_n]$ has full column rank and that the subspaces range($\boldsymbol{A}$) and range($\boldsymbol{C}_n$) are complementary.

Appendix B
Proof of Theorem IV.5

Using Lemma B.1 below we show that when condition (36) is satisfied, then $\mathbf{S}_1 \ldots, \mathbf{S}_N$ and $\boldsymbol{\Gamma}^{(N)}$ generically have full column rank, implying generic uniqueness of the GCCA factorization of $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

Lemma B.1. *[46] Let $f : \mathbb{C}^n \to \mathbb{C}$ be an analytic function. If there exists an element $\boldsymbol{x} \in \mathbb{C}^n$ such that $f(\boldsymbol{x}) \neq 0$, then the set $\{\, \boldsymbol{x} \mid f(\boldsymbol{x}) = 0 \,\}$ is of Lebesgue measure zero.*

Recall that an $m \times n$ matrix has full column rank n if it has a non-vanishing $n \times n$ minor. Since a minor is an analytic function, if it is nonzero at one point (one constructive example) then it is nonzero generically (at almost every point except for a set of measure zero). Lemma B.1 can now be used to verify whether the matrices in (31) generically have full column rank when condition (36) is satisfied.

Lemma B.2. *If $K_n \geq R + L_n$ for all $n \in \{1, \ldots, N\}$, then $\mathbf{S}_1 \in \mathbb{C}^{K_1\times(R+L_1)}, \ldots, \mathbf{S}_N \in \mathbb{C}^{K_1\times(R+L_N)}$ generically have full column rank.*

Proof. This is an immediate consequence of Lemma B.1, e.g., use $\mathbf{S}_n = [\mathbf{I}_{K_n}\mathbf{e}_1^{(K_n)}, \ldots, \mathbf{I}_{K_n}\mathbf{e}_{R+L_n}^{(K_n)}]$ as the generic example, where $\mathbf{e}_k^{(K_n)} \in \mathbb{C}^{K_n}$ denotes a unit vector with unit entry at position k. □

Lemma B.3. *If $I \geq R + \left\lceil \frac{L_1+\cdots+L_N}{N-1} \right\rceil$, then $\boldsymbol{\Gamma}^{(N)}$ given by (30) generically has full column rank.*

TABLE II: Average AUC and recall@k for the Dutch-Danish CLIR using Dutch and Danish views in training.

Algorithm	metric	Danish-Dutch				
		R=1	R=5	R=10	R=20	R=50
PDD-GCCA	avg. AUC	**0.8650** ± 0.0013	**0.9687** ± 0.0019	**0.9807** ± 0.0004	**0.9816** ± 0.0004	**0.9823** ± 0.0004
	avg. recall@1	**0.0013**±0.0005	**0.0790**±0.0051	**0.2997**±0.0071	**0.4729**±0.0032	**0.5886**±0.0026
	avg. recall@10	**0.0066**±0.0005	**0.2367**±0.0111	**0.5280**±0.0082	**0.6597**±0.0030	**0.7372**±0.0024
	avg. recall@20	**0.0104**±0.0008	**0.2980**±0.0131	**0.5871**±0.0073	**0.7022**±0.0029	**0.7687**±0.0023
	avg. recall@50	**0.0210**±0.0008	**0.3966**±0.0149	**0.6666**±0.0069	**0.7587**±0.0027	**0.8099**±0.0022
	time (sec)	447 ±11	533±17.5	635±10	940±16	1862±60
RACING	avg. AUC	0.8553±0.0013	0.9580±0.0018	0.9717±0.0005	0.9708±0.0007	0.9711±0.0004
	avg. recall@1	0.0004±0.0003	0.0596±0.0027	0.2383±0.0040	0.3617±0.0031	0.4481±0.0026
	avg. recall@10	0.0037 ±0.0008	0.1963±0.0052	0.4556±0.0029	0.5554±0.0034	0.6219±0.0024
	avg. recall@20	0.0072 ±0.0012	0.2552±0.0052	0.5161±0.0034	0.6038±0.0027	0.6610±0.0019
	avg. recall@50	0.0155 ±0.0013	0.3525±0.0055	0.5988±0.0032	0.6695±0.0030	0.7156±0.0020
	time (sec)	426 ±11	420±17	423±9	440±15	481±13
MVLSA	avg. AUC	0.7421 ±0.0021	0.9450±0.0016	0.9699±0.0005	0.9706±0.0006	0.9712±0.0004
	avg. recall@1	0.0016±0.0005	0.0396±0.0021	0.1982±0.0023	0.3327±0.0034	0.4462±0.0022
	avg. recall@10	0.0075±0.0007	0.1431±0.0031	0.4047±0.0023	0.5306±0.0021	0.6210±0.0025
	avg. recall@20	0.0114±0.0004	0.1889±0.0038	0.4648±0.0026	0.5811±0.0023	0.6609±0.0019
	avg. recall@50	0.0195±0.0010	0.2677±0.0050	0.5482±0.0028	0.6483±0.0027	0.7141±0.0020
	time (sec)	434 ±15	464±13	468±19	474±16	530±55

TABLE III: Average AUC and recall@k for the Dutch-Danish CLIR using all 5 Germanic views in training.

Algorithm	metric	5 Germanic languages				
		R=1	R=5	R=10	R=20	R=50
PDD-GCCA	avg. AUC	**0.8817**±0.0016	**0.9768**±0.0003	**0.9848**±0.0001	**0.9852**±0.0002	**0.9856**±0.0004
	avg. recall@1	**0.0023**±0.0002	**0.1105**±0.0038	**0.3725**±0.0028	**0.5472**±0.0025	**0.6671**±0.0009
	avg. recall@10	**0.0102**±0.0009	**0.3051**±0.0060	**0.6036**±0.0021	**0.7228**±0.0021	**0.7965**±0.0009
	avg. recall@20	**0.0152**±0.0010	**0.3735**±0.0063	**0.6593**±0.0020	**0.7606**±0.0028	**0.8230**±0.0010
	avg. recall@50	**0.0281**±0.0009	**0.4777**±0.0065	**0.7324**±0.0017	**0.8102**±0.0026	**0.8576**±0.0011
	time (sec)	1238±35	494±47	1754±33	2487±29	4777±43
RACING	avg. AUC	0.8740±0.0005	0.9675±0.0006	0.9773±0.0001	0.9776±0.0004	0.9818±0.0002
	avg. recall@1	0.0005±0.0001	0.0878±0.0027	0.3015±0.0022	0.4292±0.0006	0.5399±0.0013
	avg. recall@10	0.0046±0.0002	0.2684±0.0072	0.5295±0.0019	0.6227±0.0012	0.7017±0.0011
	avg. recall@20	0.0086±0.0004	0.3373±0.0071	0.5880±0.0016	0.6685±0.0012	0.7378±0.008
	avg. recall@50	0.0176±0.0005	0.4435±0.0077	0.6661±0.0014	0.7288±0.0015	0.7856±0.0010
	time (sec)	1184±34	1210±44	1218±32	1232±31	1345±31
MVLSA	avg. AUC	0.7903±0.0013	0.9645±0.0006	0.9755±0.0002	0.9773±0.0003	0.9820±0.0001
	avg. recall@1	0.0008±0.0001	0.0754±0.0032	0.2462±0.0020	0.3977±0.0018	0.5392±0.0011
	avg. recall@10	0.0037±0.0003	0.2381±0.0053	0.4639±0.0014	0.5982±0.0004	0.7026±0.0006
	avg. recall@20	0.0062±0.0003	0.3011±0.0050	0.5244±0.0010	0.6466±0.0008	0.7384±0.0004
	avg. recall@50	0.0130±0.0004	0.3988±0.0051	0.6080±0.0007	0.7106±0.0007	0.7851±0.0007
	time (sec)	1107±37	1186±55	1141±57	1181±41	1259±35

Proof. Based on Lemma B.1, the overall idea is to find a single set $\{\mathbf{A}, \mathbf{C}_1, \ldots, \mathbf{C}_N\}$ such that the matrix $\mathbf{\Gamma}^{(N)}$ has full column rank. Let us consider the extreme case where

$$I = R + \frac{L_1 + \cdots + L_N}{N-1}. \tag{54}$$

Cases where $I > R + (L_1 + \cdots + L_N)/(N-1)$ will follow by adding rows to the matrices $\mathbf{A}, \mathbf{C}_1, \ldots, \mathbf{C}_N$ constructed for the extremes case, these added rows will simply add rows to $\mathbf{\Gamma}^{(N)}$. In more detail, since we want to show that $\mathbf{\Gamma}^{(N)}$ has full column rank, we can w.l.o.g. limit the rows of $\mathbf{A}, \mathbf{C}_1, \ldots, \mathbf{C}_N$ to those needed for exact equality, and then set the remaining rows arbitrarily, because they will not affect the rank of $\mathbf{\Gamma}^{(N)}$ in our construction.

Observe that $\mathbf{\Gamma}^{(N)}$ can be seen as a matrix obtained by stacking $N-1$ blocks of the form

$$\begin{bmatrix}\mathbf{C}_1 \ \mathbf{0}_{I\times(\sum_{1<m<n} L_m + R)} \ -\mathbf{A} \ -\mathbf{C}_n \ \mathbf{0}_{I\times(\sum_{m>n} L_m + R)}\end{bmatrix} \in \mathbb{C}^{I\times(R(N-1)+\sum_{n=1}^{N} L_n)}. \tag{55}$$

We will select the columns of $\mathbf{A}, \mathbf{C}_1, \ldots, \mathbf{C}_N$ to be unit vectors, e.g., $\mathbf{a}_l = \mathbf{e}^{(R)}_{\sigma(l)}$, where $\mathbf{e}^{(R)}_{\sigma(l)} \in \mathbb{C}^R$ denotes a unit vector with unit entry at position $\sigma(l) \in \{1, \ldots, R\}$ and zero elsewhere. We first fix the first I columns of

$$[\mathbf{C}_1 \ -\mathbf{A} \ -\mathbf{C}_n] \in \mathbb{C}^{I\times(L_1+R+L_n)}, \quad 2 \le n \le N$$

as follows

$$\left[\mathbf{C}_1 \ -\mathbf{A} \ -\mathbf{C}_n\mathbf{e}^{(L_n)}_1 \ \ldots \ -\mathbf{C}_n\mathbf{e}^{(L_n)}_{I-L_1-R}\right] = \mathbf{I}_{I\times I}, \quad 2 \le n \le N. \tag{56}$$

The next step is to select the remaining columns $\mathbf{C}_n\mathbf{e}^{(L_n)}_{I-L_1-R+1}, \ldots, \mathbf{C}_n\mathbf{e}^{(L_n)}_{L_n}$, which will be referred to as the free vectors in the construction of $\mathbf{\Gamma}^{(N)}$. Due to inequality (35), each block $[\mathbf{C}_1 \ -\mathbf{A} \ -\mathbf{C}_n]$, and consequently also each block of the form (55) in $\mathbf{\Gamma}^{(N)}$, can at most contain L_1 free vectors. In more detail, since $I \ge \max_{1\le n\le N} R + L_n = R + L_N$ and $L_1 \le \cdots \le L_N$, the column dimension $(L_1 + R + L_n)$ of $[\mathbf{C}_1 \ -\mathbf{A} \ -\mathbf{C}_n]$ can at most be L_1 elements larger than its row dimension (I). This property is important in our construction, and for that reason we repeat it below as a statement

$$[\mathbf{C}_1 \ -\mathbf{A} \ -\mathbf{C}_n] \text{ contains at most } L_1 \text{ free vectors.} \tag{57}$$

Due to the construction (56), the n-th block $[\mathbf{C}_1 \ -\mathbf{A} \ -\mathbf{C}_n]$, and consequently also each block of the form (55) in $\mathbf{\Gamma}^{(N)}$, contains

$$R+L_1+L_n-I = R+L_1+L_n-\left(R+\frac{\sum_{n=1}^N L_n}{N-1}\right) \quad (58)$$

free vectors, where we recall from (54) that we assume that $I = R+(L_1+\cdots L_N)/(N-1)$, which is the extreme case. From (56) and (58) we can conclude that the total number of free vectors in the construction of $\mathbf{\Gamma}^{(N)}$ is equal to

$$\begin{aligned}&\sum_{n=2}^N (R+L_1+L_n-I) = \\ &(R+L_1)(N-1)+\sum_{n=2}^N L_n-\left(R+\frac{\sum_{n=1}^N L_n}{N-1}\right)(N-1) \\ &= L_1(N-2). \end{aligned} \quad (59)$$

We will now select the $L_1(N-2)$ free vectors in $\mathbf{C}_2,\ldots,\mathbf{C}_N$ in a way so that the matrix $\mathbf{\Gamma}^{(N)}$ has full column rank. Property (57) enables to further restrict $\mathbf{C}_n$ to the following:

$$\begin{aligned}&\mathbf{C}_n = \\ &-\begin{bmatrix} \mathbf{0}_{L_1\times(I-L_1-R)} & \mathbf{C}_{\text{free}}^{(n)} \\ \mathbf{0}_{R\times(I-L_1-R)} & \mathbf{0}_{R\times(L_n-I+L_1+R)} \\ \mathbf{I}_{(I-L_1-R)\times(I-L_1-R)} & \mathbf{0}_{(I-L_1-R)\times(L_n-I+L_1+R)} \end{bmatrix}, \\ &2\le n\le N, \end{aligned} \quad (60)$$

where

$$\begin{aligned}&\mathbf{C}_{\text{free}}^{(n)} = \mathbf{I}_{L_1\times(L_n-I+L_1+R)}\mathbf{\Pi}^{(n)} \in \mathbb{C}^{L_1\times(L_n-I+L_1+R)}, \\ &2\le n\le N, \end{aligned} \quad (61)$$

in which $\mathbf{\Pi}^{(n)} \in \mathbb{C}^{(L_n-I+L_1+R)\times(L_n-I+L_1+R)}$ is a column permutation matrix that still needs to be determined, and

$$\begin{aligned}\mathbf{I}_{L_1\times(L_n-I+L_1+R)} &= \left[\mathbf{I}_{L_1\times L_1}\mathbf{e}_1^{(L_1)},\ldots,\mathbf{I}_{L_1\times L_1}\mathbf{e}_{(L_n-I+L_1+R)}^{(L_1)}\right] \\ &= \begin{bmatrix} \mathbf{I}_{(L_n-I+L_1+R)\times(L_n-I+L_1+R)} \\ \mathbf{0}_{(L_n-I+R)\times(L_n-I+L_1+R)} \end{bmatrix}\end{aligned}$$

corresponds to the first (L_n-I+L_1+R) columns of the identity matrix $\mathbf{I}_{L_1\times L_1}$. (Note that $\mathbf{C}_n$ has at most L_1 free vectors.) Define $\mathbf{J}^{(N)} \in \mathbb{C}^{L_1\times L_1(N-2)}$ as follows

$$\mathbf{J}^{(N)} = \mathbf{1}_{N-2}^T \otimes \mathbf{I}_{L_1\times L_1} = [\mathbf{I}_{L_1\times L_1},\ldots,\ \mathbf{I}_{L_1\times L_1}], \quad (62)$$

where $\mathbf{1}_{N-2} = [1,\ldots,1]^T \in \mathbb{C}^{(N-2)}$ is an all-ones vector. We will now select $\mathbf{C}_{\text{free}}^{(2)},\ldots,\mathbf{C}_{\text{free}}^{(N)}$ as follows:

$$\left[\mathbf{C}_{\text{free}}^{(2)},\ldots,\mathbf{C}_{\text{free}}^{(N)}\right] = \mathbf{J}^{(N)}, \quad (63)$$

where relation (59) was exploited, i.e., $\sum_{n=2}^N (L_n-I+L_1+R) = L_1(N-2)$. Except for the first L_1 columns, the columns of $\mathbf{\Gamma}^{(N)}$ consists of distinct unit vectors. Note that each column of $\mathbf{\Gamma}^{(N)}$ contains at least one unit entry. The construction of $\mathbf{C}_{\text{free}}^{(2)},\ldots,\mathbf{C}_{\text{free}}^{(N)}$ allows us to "eliminate" up to $L_1(N-2)$ nonzero entries in $\mathbf{1}_{N-1}\otimes\mathbf{C}_1$ in $\mathbf{\Gamma}^{(N)}$. More formally, there exists a nonsingular matrix $\mathbf{F}$ such that all rows in $\mathbf{\Gamma}^{(N)}$ that contains two unit entries (one in a row of $\mathbf{1}_{N-1}\otimes\mathbf{C}_1$ and one in the corresponding row of $\text{Blkdiag}([\mathbf{A},\mathbf{C}_2],\cdots,[\mathbf{A},\mathbf{C}_N])$ that involves $\mathbf{C}_{\text{free}}^{(n)}$ for some $n\in\{2,\ldots,N\}$) are reduced to row-vectors with only one unit entry, in which the prior additional unit entry $\mathbf{1}_{N-1}\otimes\mathbf{C}_1$ has been deleted. Note that since this transform will at most "eliminate" up to $L_1(N-2)$ nonzero entries in $\mathbf{1}_{N-1}\otimes\mathbf{C}_1$ in $\mathbf{\Gamma}^{(N)}$, the first L_1 columns of the latter matrix will still contain L_1 unit vector after this elimination step. Let $J = R(N-1)+L_1+\cdots+L_N$, which is the column dimension of $\mathbf{\Gamma}^{(N)}$. Then this also means that there exists a row permutation matrix $\mathbf{P}_{\text{row}}$ and a column permutation matrix $\mathbf{P}_{\text{column}}$ such that the top $(J\times J)$ submatrix of $\mathbf{P}_{\text{row}}\mathbf{\Gamma}^{(N)}\mathbf{F}\mathbf{P}_{\text{column}}$ corresponds to the $(J\times J)$ identity matrix. Since $\mathbf{P}_{\text{row}}$, $\mathbf{F}$ and $\mathbf{P}_{\text{column}}$ are nonsingular, $\mathbf{\Gamma}^{(N)}$ has full column rank. □

Theorem IV.4 together with Lemmas B.2 and B.3 now implies that the GCCA factorization of $\mathbf{X}_1,\ldots,\mathbf{X}_N$ is generically unique. This proves the assertion that condition (36) in Theorem IV.5 generically guarantees the uniqueness of the GCCA factorization of $\mathbf{X}_1,\ldots,\mathbf{X}_N$.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variants," *Biometrika*, vol. 28, no. 3–4, pp. 321–377, 1936.
[2] G. H. Golub and H. Zha, "The canonical correlations of matrix pairs and their numerical computation," *IMA Vol. Math. Appl,*, vol. 69, pp. 27–49, 1995.
[3] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
[4] A. Vinokourov, J. J. Shawe-Taylor, and N. Cristianini, "Inferring a semantic representation of text via cross-language correlation analysis," in *Proc. NIPS 2003, December 11-13, 2003, Whistler, British Columbia, Canada*.
[5] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.
[6] P. Dhillon, D. Foster, and L. Ungar, "Multi-view learning of word embeddings via CCA," in *Proc. NIPS 2011, December 12-17, 2011, Granada, Spain*.
[7] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 17–19 Jun 2013, pp. 1247–1255.
[8] S. Bickel and T. Scheffer, "Multi-view clustering." in *ICDM*, vol. 4, 2004, pp. 19–26.
[9] P. Rastogi, B. Van Durme, and R. Arora, "Multiview lsa: Representation learning via generalized cca." in *HLT-NAACL*, 2015, pp. 556–566.
[10] X. Fu, K. Huang, E. E. Papalexakis, H.-A. Song, P. P. Talukdar, N. D. Sidiropoulos, C. Faloutsos, and T. Mitchell, "Efficient and distributed algorithms for large-scale generalized canonical correlations analysis," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 871–876.
[11] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and M. Hong, "Structured sumcor multiview canonical correlation analysis for large-scale data," *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 306–319, Jan 2019.
[12] J. Vía, I. Santamaría, and J. Pérez, "Deterministic CCA-based algorithms for blind equalization of FIR-MIMO channels," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, July 2007.
[13] S. Van Vaerenbergh, J. Vía, and I. Santamaría, "Blind identification of SIMO Wiener systems based on kernel canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 61, pp. 2219–2230, May 2013.
[14] R. Arora and K. Livescu, "Multi-view learning with supervision for transformed bottleneck features," in *Proc. ICASSP 2014*, 2014, pp. 2499–2503.
[15] J. Manco-Vásquez, S. Van Vaerenbergh, J. Vía, and I. Santamaría, "Kernel canonical correlation analysis for robust cooperative spectrum sensing in cognitive radio networks," *Transactions on Emerging Telecommunications Technologies*, vol. 28, 2014.

[16] M. Ibrahim and N. D. Sidiropoulos, "Reliable detection of unknown cell-edge users via canonical correlation analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4170–4182, Jun 2020.
[17] M. Borga and H. Knutsson, "A canonical correlation approach to blind source separation," Department of Biomedical Engineering, Linköping University, Linköping, Sweden, Tech. Rep. LiU-IMT-EX-0062, 2001.
[18] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, and S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2583–2587, Nov 2006.
[19] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, Oct 2009.
[20] N. M. Correa, T. Adali, Y.-O. Li, and V. D. Calhoun, "Canonical correlation analysis for data fusion and group inferences: Examining applications of medical imaging data," *IEEE Signal Process Mag.*, vol. 27, no. 4, pp. 39–50, Jul 2010.
[21] C. Campi, L. Parkkonen, R. Hari, and A. Hyvärinen, "Non-linear canonical correlation for joint analysis of meg signals from two subjects," *Frontiers in Brain Imaging Methods*, vol. 7, 2013. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2013.00107
[22] J. C. Vasquez-Correa, J. R. Orozco-Arroyave, R. Arora, E. Nöth, N. Dehak, H. Christensen, F. Rudzicz, T. Bocklet, M. Cernak, H. Chinaei *et al.*, "Multi-view representation learning via gcca for multimodal analysis of parkinson" s disease," in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017)*, no. EPFL-CONF-224545, 2017.
[23] E. Parkhomenko, D. Tritchler, J. Beyene *et al.*, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, pp. 1–34, 2009.
[24] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Statistical applications in genetics and molecular biology*, vol. 8, no. 1, pp. 1–27, 2009.
[25] Y. Lu and D. P. Foster, "Large scale canonical correlation analysis with iterative least squares," in *Advances in Neural Information Processing Systems*, 2014, pp. 91–99.
[26] R. Ge, C. Jin, P. Netrapalli, A. Sidford *et al.*, "Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis," in *International Conference on Machine Learning*, 2016, pp. 2741–2750.
[27] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *Proceedings of the 76th annual convention of the American Psychological Association*, vol. 3, 1968, pp. 227–228.
[28] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Dept. Statist., Univ. California, Berkeley, CA, USA, Tech. Rep. 688, 2005.
[29] A. Podosinnikova, F. Bach, and S. Lacoste-Julien, "Beyond CCA: Moment matching for multi-view models," in *Proc. of The 33rd International Conference on Machine Learning*, vol. 48, New York, New York, USA, 20–22 Jun 2016, pp. 458–467.
[30] V. Uurtio, J. Monteiro, J. Kandola, J. R. Shawe-Taylor, D. Fernandez-Reyes, and J. Rousu, "A tutorial on canonical correlation methods," *ACM Computing Surveys*, vol. 56, no. 6, pp. 95:33–95:31, 2017.
[31] Y. Tiann, "The dimension of intersection of k subspaces," *Missouri J. Math. Sci.*, vol. 14, no. 2, 2002.
[32] A. Ben-Israel, "Projectors on intersection of subspaces," in *Contemporary Mathematics*, 2015, vol. 636, pp. 41–50.
[33] R. M. Larsen, "Lanczos bidiagonalization with partial reorthogonalization," *DAIMI Report Series*, no. 537, 1998.
[34] J. Vía, I. Santamaría, and J. Pérez, "A learning algorithm for adaptive canonical correlation analysis of several data sets," *Neural Networks*, vol. 20, pp. 139–152, 2007.
[35] B. Draper, M. Kirby, J. Marks, T. Marrinan, and C. Peterson, "A flag representation for finite collections of subspaces of mixed dimensions," *Linear Algebra and its Applications*, vol. 451, pp. 15–32, 2014.
[36] P. Å. Wedin, "On angles between subspaces of a finite dimensional inner product space," in *Matrix Pencils*. Springer, 1983, pp. 263–285.
[37] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.
[38] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," in *ACM SIGIR Forum*, vol. 31, no. SI. ACM, 1997, pp. 84–91.
[39] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 74–81.
[40] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1393–1398.
[41] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005, pp. 79–86.
[42] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, "Feature hashing for large scale multitask learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1113–1120.
[43] P. Mineiro and N. Karampatziakis, "A randomized algorithm for cca," *arXiv preprint arXiv:1411.3409*, 2014.
[44] J. R. Shewchuk *et al.*, "An introduction to the conjugate gradient method without the agonizing pain," 1994.
[45] U. Cambridge, "Introduction to information retrieval," 2009.
[46] R. C. Gunning and H. Rossi, *Analytic Functions in Several Complex Variables*. Prentice-Hall, 1965.

ACKNOWLEDGMENT

This work was supported in part by NSF ECCS 1852831, ARO W911NF1910407, NSF ECCS-1807660.

Mikael Sørensen received the Master's degree from Aalborg University, Denmark, and the Ph.D. degree from University of Nice, France, in 2006 and 2010, respectively, both in electrical engineering. From 2010 to 2016 he was a Post-doctoral Fellow with the KU Leuven, Belgium. Since 2018 he has been a research scientist with the University of Virginia, Virginia, USA. His current research interests include tensor decompositions, machine learning and network analysis.

Charilaos I. Kanatsoulis is a postdoctoral researcher in the department of Electrical and Systems Engineering at the University of Pennsylvania. He received his Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2014 and his Ph.D. in electrical and computer engineering from the University of Minnesota (UMN), Twin Cities in 2020. His research interests include signal processing, machine learning, tensor analysis, and graph mining.

Nicholas D. Sidiropoulos (F'09) received the Diploma degree in electrical engineering from Aristotelian University of Thessaloniki, Thessaloniki, Greece, and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Maryland-College Park, College Park, MD, USA, in 1988, 1990, and 1992, respectively. He has served on the faculty of the University of Virginia (UVA), University of Minnesota, and the Technical University of Crete, Greece, prior to his current appointment as Louis T. Rader Professor and Chair of the Electrical and Computer Engineering Department at UVA. His research interests are in signal processing, communications, optimization, tensor decomposition, and factor analysis, with applications in machine learning and communications. He received the NSF/CAREER award in 1998, the IEEE Signal Processing Society (SPS) Best Paper Award in 2001, 2007, and 2011, served as IEEE SPS Distinguished Lecturer (2008-2009), and as Vice President - Membership (2017-2019) of IEEE SPS. He received the 2010 IEEE Signal Processing Society Meritorious Service Award, and the 2013 Distinguished Alumni Award from the University of Maryland, Dept. of ECE. He is a Fellow of IEEE (2009) and a Fellow of EURASIP (2014).