Efficient and Distributed Algorithms for Large-Scale Generalized Canonical Correlations Analysis

Xiao Fu†, Kejun Huang†, Evangelos E. Papalexakis*, Hyun-Ah Song‡, Partha Pratim Talukdar*, Nicholas D. Sidiropoulos†, Christos Faloutsos‡, Tom Mitchell§

†Department of Electrical and Computer Engineering, University of Minnesota
*Computer Science & Engineering Department, University of California Riverside
†Department of Computational and Data Sciences, Indian Institute of Science
‡Computer Science Department, Carnegie Mellon University
§Machine Learning Department, Carnegie Mellon University

Abstract—Generalized canonical correlation analysis (GCCA) aims at extracting common structure from multiple ‘views’, i.e., high-dimensional matrices representing the same objects in different feature domains – an extension of classical two-view CCA. Existing (G)CCA algorithms have serious scalability issues, since they involve square root factorization of the correlation matrices of the views. The memory and computational complexity associated with this step grow as a quadratic and cubic function of the problem dimension (the number of samples / features), respectively. To circumvent such difficulties, we propose a GCCA algorithm whose memory and computational costs scale linearly in the problem dimension and the number of nonzero data elements, respectively. Consequently, the proposed algorithm can easily handle very large sparse views whose sample and feature dimensions both exceed 100,000 – while the current approaches can only handle thousands of features / samples. Our second contribution is a distributed algorithm for GCCA, which computes the canonical components of different views in parallel and thus can further reduce the runtime significantly (by ≥ 30% in experiments) if multiple cores are available. Judiciously designed synthetic and real-data experiments using a multilingual dataset are employed to showcase the effectiveness of the proposed algorithms.

Keywords—Large-scale generalized canonical correlation analysis, distributed GCCA, multilingual word embeddings

I. INTRODUCTION

Canonical Correlation Analysis (CCA) [1] is used to extract low-dimensional representations from two different views of certain entities (e.g., speech and video features of the same object). CCA has a wide spectrum of applications, including clustering, regression, brain imaging, and natural language processing [2]–[4]. Classical CCA focuses on two views. Generalized CCA (GCCA) [5] aims to handle more than two views, and it has recently drawn renewed attention, motivated by modern applications [6]–[9].

Computationally, GCCA poses very challenging optimization problems. Starting from the 1960s, different GCCA formulations and algorithms have been considered [5], [10], such as the sum-of-correlations (SUMCOR), sum-of-squared-correlations (SSQCOR) and maximal-variance (MAX-VAR) formulations, to name a few. In recent years, along with the increasing ability of collecting more and more multiview data and the demand for integrating and analyzing such data, the study of GCCA algorithms has been gaining momentum [6], [7], [11]. Different approaches including power-method like algorithms [11], block coordinate descent (BCD) combined with local linearization [6], and semidefinite relaxation [7] were proposed to tackle the SUMCOR formulation. Recently, Rastogi et al [8] proposed an efficient approximation to the MAX-VAR problem and applied it to multilingual analytics.

Despite the long history of GCCA research, many challenges remain. First, the aforementioned algorithms mostly consider extracting the first canonical component of each view, and then use a deflation method to find the other ones; but deflation suffers from error propagation. Second, many (G)CCA algorithms have serious scalability issues since they employ a whitening process (i.e., multiplying the square roots of the correlation matrices of the views to the data) – which destroys sparsity, can easily exhaust memory resources, and also requires a large number of flops even for moderate-size views. There is a recent push towards designing scalable algorithms for CCA [12], [13], but mostly focusing on the two-view case. Third, to the best of our knowledge, no distributed GCCA algorithm has appeared in the literature. Designing distributed algorithms for GCCA is well-motivated by many reasons, including privacy, confidentiality, and data security/integrity concerns; in addition to better scaling, faster computation, and suitability for cloud computing.

Contributions In this work, our interest lies in the SUMCOR formulation where pairwise highly correlated reduced-dimension views are sought. Our design is under a setting where the views are large sparse matrices. Such a scenario is well-motivated in practice, e.g., in multilingual analytics. Our contributions are as follows:

• Scalable GCCA: We propose a highly scalable SUMCOR GCCA algorithm. The proposed algorithm does not need to explicitly instantiate the correlation matrices and whitening matrices, resulting in substantial memory savings. Computationally, the key components of the algorithms are sparse matrix-vector multiplications and singular value decomposition of (typically very) “thin” matrices, which are both...
lightweight and suitable for big data analytics. A sneak peek of the experimental results is provided in Fig. 1, where one can verify that our method (LasCCA) scales seamlessly to bigger views, while the competitor using whitening runs out of memory when the size of the views gets large. In addition, no deflation process is involved in the proposed algorithm.

- **Distributed GCCA:** We also propose a distributed GCCA algorithm for SUMCOR. To our best knowledge, this is the first distributed algorithm for GCCA. In the distributed algorithm, the views are stored in different nodes of a network, and only a small amount of information needs to be exchanged between the nodes and a coordinator. Most of the computations are carried out locally and in parallel at the nodes. By exploring more nodes/cores, running time can be further reduced by the distributed algorithm by at least 30% compared to that of the proposed scalable algorithm in our experiments (cf. the performance of DisCCA (multi-core) in Fig. 1).

**Reproducibility:** We open-source our code at the first authors’ homepage, where a MATLAB demo is available.

### II. Problem Statement

Let $Y_i \in \mathbb{R}^{L \times M_i}$ denote the $i$th (centered) view, where $Y_i(\ell,:)$ corresponds to the $\ell$th data point (or entity) in the $i$th view, $L$ is the number of entities, and $M_i$ denotes the number of features of the $i$th view. The SUMCOR GCCA problem can be expressed as [14]:

$$\max_{\{Q_i \in \mathbb{R}^{M_i \times K}\}_{i=1}^I} \sum_{i=1}^I \sum_{j \neq i} \text{Tr}(Q_i^T X_i^T X_j Q_j)$$  \hspace{1cm} (1a)

subject to $Q_i^T X_i Q_i = I_K$, $\forall i$;  \hspace{1cm} (1b)

where $X_i = (\sqrt{2}Y_i \in \mathbb{R}^{L \times M_i}$ denotes the $i$th normalized view, $Q_i \in \mathbb{R}^{M_i \times K}$, and $K$ is the number of canonical components that we seek. By this normalization, $(X_i^T X_i)$ is an estimate of the auto-correlation matrix of the $i$th view and $(X_i^T X_j)$ an estimate of the cross-correlation matrix between the $i$th and the $j$th views; i.e., we seek $K$-dimensional representations of the views such that they are pair-wise highly correlated, and we aim to extract $K$ canonical components simultaneously. Our approach builds upon the block coordinate descent (BCD) framework – i.e., updating each $Q_i$ for $i = 1, \ldots, I$ in turn while fixing the other $Q_j$’s for $j \neq i$. By doing so, the subproblem w.r.t. $Q_i$ can be expressed as

$$\max_{Q_i} \text{Tr} \left( Q_i^T X_i^T \sum_{j \neq i} X_j Q_j \right)$$  \hspace{1cm} (2a)

subject to $Q_i^T X_i Q_i = I_K$.  \hspace{1cm} (2b)

To solve the subproblem (2), it is tempting to do the following: Assume that $X_i^T X_i$ has full rank and thus admits a square root decomposition, i.e., $X_i^T X_i = (X_i^T X_i)^{1/2}(X_i^T X_i)^{1/2}$. Then, let $Q_i = (X_i^T X_i)^{-1/2} Z_i$. We can re-write (2) as

$$\max_{Z_i} Z_i^T (X_i^T X_i)^{-1/2} X_i^T \sum_{j \neq i} X_j Q_j$$  \hspace{1cm} (3)

In some existing approaches such as those in [6, 7], the key steps are essentially identical to the above change of variables – although they consider the simplest case where $K = 1$. The upshot of this reformulation is that Problem (3) is essentially a Procrustes projection problem [15] and can be solved to optimality by applying singular value decomposition (SVD) to $(X_i^T X_i)^{-1/2} X_i \sum_{j \neq i} X_j Q_j$. Such a method works well when $M_i$ is small. However, instantiating the whitening matrix $(X_i^T X_i)^{-1/2}$ poses serious complexity issues: First, the matrix $(X_i^T X_i)^{-1/2}$ is very likely to be dense even when $X_i$ is sparse. Consequently, computing its inverse requires $O(M_i^3)$ flops (e.g., when $M_i = 100,000$, this step requires $O(10^{15})$ flops). More importantly, storing a dense and large $M_i \times M_i$ matrix is almost prohibitive – when $M_i = 100,000$, storing $(X_i^T X_i)^{1/2}$ needs 75 GB memory. In this work, we will propose a new BCD approach that completely avoids instantiating $(X_i^T X_i)^{-1/2}$, and thus is highly scalable in terms of both memory and computational complexity.

### III. Large-Scale GCCA (LasCCA)

Instead of applying the change of variables $Q_i = (X_i^T X_i)^{-1/2} Z_i$ and reformulating (2) to (3), we let $G_i = X_i Q_i$, for all $i$. Note that $G_i \in \mathbb{R}^{L \times K}$ is a very “thin” matrix since in practice $L \gg K$ holds and $K$, i.e., the number of canonical components sought, is usually small and can be controlled by the designer. Assuming that rank($X_i$) = $M_i$, we have $Q_i = (X_i^T X_i)^{-1/2} X_i G_i$. Substituting the above form of $Q_i$ to Problem (2) leads to the following optimization problem:

$$\max_{G_i} \text{Tr} (G_i^T H_i)$$  \hspace{1cm} (4)

subject to $G_i^T G_i = I_K$, $G_i \in \mathcal{R}(X_i)$,

where $H_i = X_i (X_i^T X_i)^{-1} X_i^T \sum_{j \neq i} X_j (X_j^T X_j)^{-1} X_j G_j$, $\mathcal{R}(X)$ denotes the range space of $X$, and the constraints in (4) together ensure that Problem (4) and Problem (2) are equivalent. Note that one may also substitute $X_i Q_i = G_i$,
Algorithm 1: Large-Scale GCCA (LasCCA)

\[ \text{input : } \{X_i, C_i(t)\}_{t=1}^{K} \]
\[ r \leftarrow 1; \]
\[ \text{repeat} \]
\[ \text{for } i = 1: I \text{ do} \]
\[ G_i^{(r)} = \{G_i^{(r+1)}, \ldots, G_i^{(r+1)}, G_i^{(r+1)}, \ldots, G_i^{(r)}\} \]
\[ H_i^{(r)} \leftarrow \text{H\_Compute}(\{X_i\}_{t=1}^{K}, G_i^{(r)}); \]
\[ U_i^{(r)}(V_i^{(r)}(r)) \leftarrow \text{svd}(H_i^{(r)}); \]
\[ G_i^{(r+1)} \leftarrow U_i^{(r)}(V_i^{(r)}(r))^T; \]
\[ \text{end} \]
\[ r \leftarrow r + 1; \]
\[ \text{until some stopping criterion is reached;} \]
\[ \text{output: } \{G_i\} \]

into (2) and write \( H_i \) as \( H_i = \sum_{j \neq i} X_j(X_j^TX_j)^{-1}X_j^T G_j \); but our reformulation allows us to derive a simple solution as will be seen shortly. The key difference between our reformulation in (4) and that in (3) is that we avoided using the square root decomposition of the correlation matrices, i.e., \((X_i^TX_i)^{-1/2}\). As we will verify later, such a change will significantly reduce memory cost and computational complexity. At first glance, the solution to Problem (4) looks unclear because of the constraint \( G_i \in R(X_i) \). Nevertheless, the solution is in fact simple: Let us first relax the constraint \( G_i \in R(X_i) \) for the moment. Then, an optimal solution to the relaxed problem is the Procrustes projection; i.e.,

\[ G_i = U_i V_i^T, \]

where \( U_i \Sigma_i V_i = \text{svd}(H_i) \) denotes the economy-size SVD of \( H_i \). We note that such \( G_i \) satisfies \( G_i \in R(H_i) \) and so \( G_i \in R(X_i) \). Since the solution of the relaxed problem conforms to the constraint of the original problem, (5) is also an optimal solution of (4).

Under this assumption, one can see that the SVD step is not difficult in terms of both computational and memory costs. In fact, \( H_i \in \mathbb{R}^{L \times K} \) is a very “thin” matrix (recall that \( K \) is usually much smaller than \( L \)) — meaning that the memory burden is very light. In addition, computing SVD of \( H_i \) costs \( O(LK^2) \) flops, which is merely linear in \( L \). The overall BCD algorithm is presented in Algorithm 1, which we name large-scale generalized CCA (LasCCA). At this point, we have not elaborated how to compute \( H_i \) but only put an operator called \text{H\_Compute}(\cdot, \cdot) there (cf. line 2 in Algorithm 1); but computing \( H_i \) is the most difficult part as it involves inversions of the correlations. Next, we focus on this key computation.

A. Building Block: Iterative Least Squares

As we have mentioned, computing and storing \((X_j^TX_j)^{-1}\) may not be affordable when constructing \( H_i \) at each step. Fortunately, explicit computation of this term is not necessary when calculating \( H_i \). In fact, what has been used for constructing \( H_i \) is \((X_j^TX_j)^{-1}X_j G_j \) for different \( j \)'s, which is, again, a very thin matrix that only costs \( O(LK) \) memory. To obtain this thin matrix, consider the following unconstrained linear least squares (LS) problem

\[ \min_{R_j} \|X_j R_j - G_j\|^2_F. \]

Assume that \( L \geq M_j \) and \( \text{rank}(X_j) = M_j \). The optimal solution \( R_j^* \) can be written as \( R_j^* = (X_j^TX_j)^{-1}X_j G_j \). This means that to obtain the key ingredient of building up \( H_i \), what one needs to do is to solve a LS problem. The merit of looking at the problem of computing \((X_j^TX_j)^{-1}X_j G_j \) from a LS point of view is that the LS problem can be solved iteratively using very lightweight iterations that can also take advantage of sparsity of \( X_j \) — e.g., by stochastic gradient and accelerated gradient. In this work, we propose to employ a powerful tool for unconstrained LS — the conjugate gradient (CG) algorithm. There is rich literature on CG since it has been the workhorse for solving large-scale LS problems for decades; see [16] for details. The take-home point is that, to solve Problem (6), CG only requires multiplications of \( X_j \) and a matrix of size \( M_j \times K \) — if \( X_j \) is sparse, such multiplications can be easily carried out with a per-iteration complexity of \( O(\text{nnz}(X_j)) \) flops, where \( \text{nnz}(\cdot) \) counts the non-zero elements. Another advantage of using CG is that even in the worst case, CG is provably convergent within a finite number of iterations [16]; in practice, CG almost always converges to a very good accuracy level within 20 iterations.

Using CG, \( H_i \) can be computed efficiently. The corresponding algorithm is presented in Algorithm 2. In short, we first use CG to form \( R_j = (X_j^TX_j)^{-1}X_j G_j \) as we explained. Then, \( C_j = X_j^TX_j^{-1}X_j G_j \) can be formed by multiplying \( X_j \) to \( R_j \). If \( X_j \) is sparse, then this step is easy. Then, the matrix \( P_i = \sum_{j \neq i} X_j^TX_j^{-1}X_j G_j \) can be obtained by summation over the index set \( \{1, \ldots, I\} \setminus \{i\} \). By solving \( E_i = \arg \min_{E_i} \|P_i - X_i E_i\|^2_F \) using CG we obtain \( E_i = (X_i^TX_i)^{-1}X_i^T P_i \). Finally, by another multiplication of a sparse matrix \( X_i \) and the thin matrix \( E_i \), we get \( H_i \). We factor out each iteration’s processing details in the \text{H\_Compute} method and Algorithm 1. As one can see, the major computations all have the same complexity order — and it is linear in \( \text{nnz}(X_i) \).}

IV. DISTRIBUTED GCCA (DISCCA)

In practice, there are many scenarios in which distributed computing is desirable. For example, sometimes the views may be collected and stored at different nodes within a network, and directly exchanging the views may not be allowed for security and/or legal reasons. Another reason is multi-core parallel computing-based acceleration. Instead of computing \( G_i \) sequentially as in Algorithm 1, it may be more appealing
to compute $G_i$ for all $i$ in parallel in a distributed fashion – which may well expedite the whole process when $I$ is large.

Let us consider a network structure as in Fig. 2. There, the nodes can be machines that are located in different places or cores within one machine. The views are stored at the nodes and are not allowed to be exchanged. The master is a coordinator who collects and distributes some derivative information, and we wish to leave the heavy computations to the nodes. To see how we approach this problem, let us consider the following modification to LasCCA. Specifically, instead of updating $G_i$ sequentially and cyclically as in Algorithm 1, the nodes locally update $G_i$ simultaneously, based on the previous iterate, which results in the following $H_i$-update strategy:

$$ H_i^{(r)} \leftarrow \text{DH}_i \left( \{X_i\}_{i=1}^I, \{G_j^{(r-1)}\}_{j \neq i} \right) $$

(7)

where DH $i$ is a distributed version of Algorithm 2 (cf. Algorithm 3). The update in (7) is different from that in Algorithm 1: In Algorithm 1, the update of $G_i^{(r)}$ uses the information of $G_j^{(r-1)}$ for $j < i$ and $G_j^{(r)}$ for $j > i$, while this update in (7) only uses $G_j^{(r)}$ for all $j \neq i$. By this slight modification, a significant part of (7) can be computed in parallel at different nodes. A detailed implementation of the DH_Compute algorithm is presented in Algorithm 3, where one can see that the master only gathers matrices of size $L \times K$, i.e., $C_j$, and adds them over different sets of $j$’s. Using DH_Compute, the complete distributed implementation of GCCA is easy to derive. After all $G_i^{(r+1)}$’s are computed by different nodes, we update the $G_i$ that brings maximal improvement to the objective function of (1). Using this idea, the whole distributed CCA (DisCCA) algorithm is presented in Algorithm 4. One can see that the lines 6-11 implement the idea of selecting a block to update. In optimization theory, adding these lines makes the algorithm fall into the category of the maximum block improvement (MBI) framework [17] which is a greedy variant of the Gauss-Seidel (GS)-type BCD. Compared to GS-type BCD (cf. LasCCA), every update of MBI gives a larger improvement of the objective, which is clearly favorable. On the other hand, MBI “wastes” most of the computations since it computes all the blocks but updates only one. Hence, there is clearly an interesting resource-time trade-off between choosing LasCCA and DisCCA in practice. Another remark is that both LasCCA and DisCCA monotonically increase the objective value of (1) following the properties of generic BCD and MBI, which is desired in large-scale GCCA since every iteration makes progress towards the ultimate goal.

V. Experiments

A. Synthetic-Data Experiments

1) Baselines: Our first baseline is an algorithm that employs a change of variables strategy as proposed in [6], [7]. This follows a BCD approach where each subproblem is as implemented Algorithm 3: DH_Compute (Distributed Computation of $H_i$)

<table>
<thead>
<tr>
<th>input : ${X_i}, {G_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $R_i \leftarrow C_i(X_i, G_i)$;</td>
</tr>
<tr>
<td>2 $C_i \leftarrow X_i R_i$;</td>
</tr>
<tr>
<td>3 node $i$ for all $i$ sends $C_i$ to master;</td>
</tr>
<tr>
<td>4 $P_i \leftarrow \sum_{j \neq i} G_j$;</td>
</tr>
<tr>
<td>5 master sends $P_i$ to node $i$;</td>
</tr>
<tr>
<td>6 $E_i \leftarrow C_i(X_i, P_i)$;</td>
</tr>
<tr>
<td>7 $H_i \leftarrow X_i E_i$;</td>
</tr>
<tr>
<td>output: $H_i$;</td>
</tr>
</tbody>
</table>

Algorithm 4: Distributed GCCA (DisCCA)

<table>
<thead>
<tr>
<th>input : ${X_i, G_i^{(0)}}_{i=1}^I$; $K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $r \leftarrow 1$;</td>
</tr>
<tr>
<td>2 repeat</td>
</tr>
<tr>
<td>3 $H_{i}^{(r)} \leftarrow \text{DH}<em>i \left( {X_i}</em>{i=1}^I, {G_j^{(r)}}_{j \neq i} \right)$;</td>
</tr>
<tr>
<td>4 $U_i^{(r)} X_i^{(r)} V_i^{(r)} \leftarrow \text{svd}(H_{i}^{(r)})$;</td>
</tr>
<tr>
<td>5 $G_i^{(r+1)} \leftarrow U_i (V_i^{(r)})^T$;</td>
</tr>
<tr>
<td>6 $v_i^{(r)} \leftarrow \	ext{Tr}((G_i^{(r+1)})^T H_i^{(r)})$;</td>
</tr>
<tr>
<td>7 node $i$ sends $v_i^{(r)}$ to master;</td>
</tr>
<tr>
<td>8 $i^* \leftarrow \arg \max_{i \in {1, \ldots, I}} v_i^{(r)}$;</td>
</tr>
<tr>
<td>9 master sends an update command to node $i^*$;</td>
</tr>
<tr>
<td>10 $G_i^{(r+1)} \leftarrow G_i^{(r)}$;</td>
</tr>
<tr>
<td>11 $G_j^{(r+1)} \leftarrow G_j^{(r)}$ for $i \neq i^*$;</td>
</tr>
<tr>
<td>12 $r \leftarrow r + 1$;</td>
</tr>
<tr>
<td>until some stopping criterion is reached;</td>
</tr>
<tr>
<td>output: ${G_i}$</td>
</tr>
</tbody>
</table>

2) Data Generation: First, we let $Z \in \mathbb{R}^{L \times M}$ be a common latent factor of the different views, where $Z$ is a randomly generated sparse matrix whose non-zero entries follow the i.i.d. zero-mean unit-variance Gaussian distribution. Then, a sparse matrix $A_i \in \mathbb{R}^{M \times M}$ is multiplied to $Z$, resulting in
**Table I**

AVERAGE CORRELATIONS CAPTURED BY THE ALGORITHMS.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>L (entries)</th>
<th>MEAN_INNER_PRODUCT</th>
<th>MEAN_INNER_PRODUCT</th>
<th>MEAN_INNER_PRODUCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>1,000</td>
<td>5,000</td>
<td>99.30</td>
</tr>
<tr>
<td>LasCCA</td>
<td>99.87</td>
<td>99.30</td>
<td>99.05</td>
<td>99.10</td>
</tr>
<tr>
<td>Dist-BCD</td>
<td>99.60</td>
<td>98.73</td>
<td>98.35</td>
<td>98.24</td>
</tr>
<tr>
<td>CSR-BCD</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>MLSA</td>
<td>98.57</td>
<td>63.32</td>
<td>44.60</td>
<td>7.76</td>
</tr>
</tbody>
</table>

**Table II**

The average correlations captured after 20 iterations under different ρ’s: L = 120, 000; M = 100, 000; K = 5; I = 5; “†” means “OUT-OF-MEMORY” and “-” means not applicable; vopt = 100.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>LasCCA</td>
<td>corr. captured</td>
</tr>
<tr>
<td>Dist-BCD</td>
<td>corr. captured</td>
</tr>
<tr>
<td>CSR-BCD</td>
<td>corr. captured</td>
</tr>
<tr>
<td>MLSA</td>
<td>corr. captured</td>
</tr>
</tbody>
</table>

**Equation**

\[ X_i = Z A_i \]

The overall sparsity level of \( X_i \) is controlled to satisfy experiment specifications. This way, the views have a perfectly correlated common latent factor, i.e., \( Z \). We fix \( M_1 = \ldots = M_I = M \), and test the algorithms under different sizes and sparsity levels of \( X_i \), where the sparsity level \( \rho_i \) is defined as \( \rho_i = \max(X_{ik})/LM \), and we let \( \rho_i = \rho \) for all \( i \).

3) **Evaluation:** We evaluate the synthetic data experiment by observing the captured sum of pairwise correlations (“correlations captured” in short) between the views. Since the data are perfectly correlated in the \( Z \)-domain, the optimal correlation value is \( v_{opt} = I(I-1)K \). We also define a metric called “0.95_time”, which records the time that is needed for an algorithm to capture 95% of the optimal value of the total sum of correlations, i.e., \( I(I-1)K \).

4) **Results:** Table I shows the performance of the algorithms under various \( L \)’s (and \( M \)’s). Here, we fix \( \rho = 5 \times 10^{-3} \) and change \( L \) and \( M = 0.8 \times L \). We seek \( K = 5 \) canonical components of each of the \( I = 5 \) views. MLSA requires first truncating the range of views to a certain number, and thus we use the first 100 singular values and vectors to approximate each view. We let all the iterative algorithms run for 20 iterations in this experiment, and the results are averaged over 10 random trials. We initialize CSR-BCD, LasCCA and DisCCA using the same random initialization: As one can see from the table, in terms of captured correlations, the CSR-BCD algorithm, gives the best performance when \( L \leq 10,000 \) – it always reaches the optimal objective value, although the problem is non-convex and NP-hard. However, when \( L = 50,000 \), the algorithm exhausts the memory quota (32 GB RAM). The proposed algorithms, i.e., LasCCA and DisCCA, give slightly lower objective values compared to that of CSR-BCD, which are already very good – they both capture more than 98% of the total correlations, even under the case CSR-BCD cannot run, i.e., when \( L = 50,000 \). Another observation is that when \( L = 1,000 \), MLSA works very well, i.e., 98% of the correlations is captured. However, when \( L \) increases, the performance degrades rapidly since the approximation of the views becomes much coarser.

The corresponding 0.95_time performance of the algorithms in Table I can be found in Fig. 1, which we showed in the beginning of the paper. As one can see, the runtime of CSR-BCD grows rapidly with \( L \) increasing. On the other hand, the proposed algorithms scale well. In particular, DisCCA gives the best 0.95_time performance using a multi-core implementation, which is 104 seconds when \( L = 50,000 \). LasCCA is slower compared to DisCCA since it only uses one core, taking 199 seconds to capture 95% of the total sum of correlations. Obviously, if we implement DisCCA using a single core, it is slower than LasCCA since DisCCA needs more resources and computations before deciding which block to update. This result suggests that if only a single core is used, LasCCA may be more preferable.

Table II shows the captured correlations of a larger problem (i.e., \( L = 120,000 \) and \( M_i = 100,000 \)) under various sparsity levels of \( X_i \). In this simulation, CSR-BCD cannot start to run; it leads to out-of-memory under all the \( \rho \)'s since it destroys sparsity at the very first step. LasCCA and DisCCA, on the other hand, work well under different \( \rho \)'s. Both algorithms give objective values that are larger than 96 (recall that \( v_{opt} = 100 \)) under all \( \rho \)'s. When \( \rho \) increases, the captured correlations of both algorithms decrease very gracefully. The performance of MLSA also decreases along with \( \rho \) increasing, but more rapidly. In terms of 0.95_time, one can see that DisCCA (using multiple cores) still outperforms LasCCA as in the previous simulation. The runtimes of both DisCCA and LasCCA grows (roughly) linearly with \( \rho \), which is consistent with our analysis – the major computations in DisCCA and LasCCA have complexity order of \( O(\text{nnz}(X_i)K) \) flops.

### B. Real-Data Experiments

1) **Multilingual Data:** We also test our algorithms on a large-scale parallel corpus of multilingual data. The data were originally introduced in [18] and can be downloaded from http://wordvectors.org/eacl14-data.tar. We have \( I = 4 \) views, which are constructed from English, Spanish, German and French, respectively. The raw data contain a matrix per language that records the Point-wise Mutual Information (PMI) of the word co-occurrence for that language. We view the rows of the matrix as our data points and the columns as features. We use English words to form our first view, \( X_1 \), which contains \( L = 180,834 \) words defined by \( M_1 = 130,000 \) features. We set \( M_1 = \ldots = M_I = 130,000 \), and the features correspond to the columns in each view that have the \( M_i \) highest energies. There are 1.21% non-zero entries out of the 2.3508 \( \times 10^{10} \) entries of \( X_1 \). Using dictionaries, we pick out the corresponding words in French, German and Spanish to form \( X_2, X_3 \) and \( X_4 \), respectively. For those English terms which do not have a translation in \( X_i \) for \( i > 1 \), we simply let the corresponding row of \( X_i \) be zero. Consequently, the \( X_i \)'s are all 180,834 \( \times 130,000 \) sparse matrices.
2) Evaluation: In the real experiments, we evaluate the quality of the output embeddings of the English words. We employ the evaluation tool provided at wordvectors.org [19], which automatically evaluates the embedded English words (i.e., $X_1Q_1$) on several word embedding tasks by comparing the algorithm-learned embeddings with the judgment of humans. The outputs are scores between zero and one, and a score equal to one means a perfect alignment between the learned result and human judgment. In all the real experiments, we evaluate the results on the first 11 tasks of wordvectors.org.

3) Results: In Tables III and IV, we show the scores given by DisCCA, LasCCA and MLSA when $K = 100$ and $K = 300$, respectively. We run the proposed algorithms for 10 iterations and observe the results. For MLSA, we follow the procedure in [8] to truncate the rank of the views to 640 as pre-processing. We use the results given by MLSA to initialize our algorithms. The first observation is that all the algorithms work better using a larger $K$, which is reasonable since more dimensions are used. One can see that DisCCA gives the best evaluation scores on 6 and 8 tasks under the two cases, respectively. SVD also gives good scores and performs the best evaluation scores on 6 and 8 tasks under the two cases.

VI. Conclusion

In this paper, we investigated the problem of computing canonical components of large-scale sparse multiview data. A BCD-based algorithmic framework was proposed based on a judicious equivalent reformulation of the SUMCOR GCCA problem. Under this framework, two highly scalable algorithms, namely, LasCCA and DisCCA, based on different update strategies were proposed. DisCCA is also the first distributed algorithm for large-scale GCCA. Simulations and real experiments show that both of the proposed algorithms scale well to large-size problems and give promising results when applied to a word embedding problem.

ACKNOWLEDGMENT

This work is supported by the National Scientific Foundation (NSF) under Project NSF IIS-1247632, NSF IIS-1447788, and NSF IIS-1247489.

REFERENCES