Distributed Efficient Multimodal Data Clustering

Jia Chen
Dept. of ECE & Digital Technology Center,
Univ. of Minnesota
Minneapolis, MN, 55455
Email: chen5625@umn.edu

Ioannis D. Schizas
Dept. of EE
Univ. of Texas at Arlington
Arlington, TX 76010
Email: schizas@uta.edu

***Abstract*—Clustering of multimodal data according to their information content is considered in this paper. Statistical correlations present in data that contain similar information are exploited to perform the clustering task. Specifically, multiset canonical correlation analysis is equipped with norm-one regularization mechanisms to identify clusters within different types of data that share the same information content. A pertinent minimization formulation is put forth, while block coordinate descent is employed to derive a batch clustering algorithm which achieves better clustering performance than existing alternatives. Distributed implementations are also considered to cluster spatially clustered data utilizing the alternating direction method of multipliers. Relying on subgradient descent, an online clustering approach is derived which substantially lowers computational complexity compared to the batch approaches. Numerical tests demonstrate that the proposed schemes outperform existing alternatives.**

I. Introduction

In many applications such as sensor networks [6], genomic data integration [12] and remote sensing [15], different types of data are acquired corresponding to multiple sensing modes. Several techniques have been put forth for processing multimodal data including estimation [11], and regression [9]. Our focus here is clustering multimodal data according to their information content. Existing clustering schemes [2], [3], [13], [18] rely on pre-specified distance metrics to cluster unimodal data (single data type) according to their similarity in terms of magnitude.

The goal here is to cluster data according to the information they contain about some underlying sources/objects of interest. Statistical correlation between multimodal data that contain information about the same source of interest will be exploited. We build on multiset canonical correlation analysis (M-CCA) [8], [10] that is capable of uncovering maximally correlated features from multiple sets of data. Applications of M-CCA vary from information retrieval from different languages [16], to large scale biometric structure prediction [14], to remote sensing [15]. The novel regularized M-CCA framework proposed here generalizes our work in [5] that performs clustering of only two different data types. We equip M-CCA with norm-one regularization terms, which are capable to identify which multimodal data entries are correlated and cluster them in the same group. Block coordinate descent [1] will be utilized to obtain a batch algorithm that uses all the available data to perform multimodal clustering (Sec. III). Starting from the norm-one regularized multiset CCA formulation the alternating direction method of multipliers is employed to locally estimate global quantities and enable sensors to perform distributed clustering (Sec. IV). Trading-off clustering accuracy for computational complexity, an online multimodal clustering approach is put forth, employing subgradient descent techniques (Sec. V). Besides the significant computational savings, numerical tests demonstrate the small performance loss of online clustering with respect to the batch approach, as well as its better performance compared to existing clustering schemes (Sec. VI).

† Work in this paper is supported by the NSF Grant ECCS 1509780.

II. Problem Statement and Preliminaries

We consider a setting in which different kind of sensors measure different types of observations and the acquired multimodal sensor data contain information about q statistically uncorrelated stationary sources, which are denoted by $s_i(\tau)$ for $i = 1, \cdots, q$ while τ corresponds to time. Let $\mathcal{A}^m$ represent a set containing the mth type of sensors and the number of sensors in $\mathcal{A}^m$ is equal to p_m. Also, let $\mathcal{A}_j^m$ denote a sensor with index j in set $\mathcal{A}^m$, for $j = 1, \cdots, p_m$. It is assumed that each sensor senses at most one source which in practice happens when only one source is contained within the sensing range of each sensor. The measurements acquired by sensor $\mathcal{A}_j^m$ adhere to the generic (non)linear data model

$$x_{m,j}(\tau) = h_{m,j}(s_{f(m,j)}(\tau)) + w_{m,j}(\tau), \tag{1}$$

where $w_{m,j}(\tau)$ is zero-mean white sensing noise and $f(m,j) \in \{1, \cdots, q\}$ corresponds to the index of the source sensed by sensor $\mathcal{A}_j^m$. Further, $h_{m,j}(\cdot)$ denotes a random scalar mapping, which equals to zero when sensor $\mathcal{A}_j^m$ does not contain any information about the sources present in the field (placed far away).

Stacking the measurements from all the sensors belonging to $\mathcal{A}^m$ in vector $\mathbf{x}_{m,\tau} \in \mathbb{R}^{p_m \times 1}$, we have $\mathbf{x}_{m,\tau} := [x_{m,1}(\tau), x_{m,2}(\tau), \cdots, x_{m,p_m}(\tau)]^T$. Let $\mathcal{S}^i$ denote the subset of entries in $\{\mathbf{x}_{m,\tau}\}_{m=1}^M$ that contain information about source $s_i(\tau)$, while $\mathcal{S}^0$ is defined as the subset of entries that only have information about noise. The goal of this paper is to find the $q+1$ sets, namely $\{\mathcal{S}^i\}_{i=0}^q$, clustering in that way data according to their source information. Toward this end, a proper multiset canonical correlation analysis (M-CCA) [10] framework equipped with norm-one regularization is derived.

ISBN 978-0-9928626-7-1 © EURASIP 2017

After applying the ADMM framework and subgradient descent, the distributed and online algorithms are developed.

Given $M > 2$ data sets, $\{\mathbf{x}_{1,\tau}\} \in \mathbb{R}^{p_1\times 1}, \{\mathbf{x}_{2,\tau}\} \in \mathbb{R}^{p_2\times 1}, \cdots, \{\mathbf{x}_{M,\tau}\} \in \mathbb{R}^{p_M\times 1}$ with $\tau = 1, \cdots, T$, M-CCA searches for $q \times p_m$ matrices $\mathbf{D}_m$ so that the following cost is minimized

$$\{\check{\mathbf{D}}_m\}_{m=1}^M = \arg\min_{\mathbf{D}_1,\cdots,\mathbf{D}_M} 1/(2T) \sum_{\tau=1}^T \sum_{m=1}^M \sum_{n\neq m} \|\mathbf{D}_m(\mathbf{x}_{m,\tau} - \mathbf{u}_m) - \mathbf{D}_n(\mathbf{x}_{n,\tau} - \mathbf{u}_n)\|_2^2 \quad \text{s.to } \mathbf{D}_m\mathbf{\Sigma}_m\mathbf{D}_m^T = \mathbf{I} \text{ for } m = 1, \cdots, M, \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{q\times q}$ is the identity matrix, while the sample-average estimates of the expectation and covariance of $\mathbf{x}_{m,\tau}$ are represented by $\mathbf{u}_m$ and $\mathbf{\Sigma}_m$, respectively, i.e., $\mathbf{u}_m := T^{-1}\sum_{\tau=1}^T \mathbf{x}_{m,\tau}$ and $\mathbf{\Sigma}_m := T^{-1}\sum_{\tau=1}^T (\mathbf{x}_{m,\tau} - \mathbf{u}_m)(\mathbf{x}_{m,\tau} - \mathbf{u}_m)^T$.

Each row of the estimated vectors $\{\check{\mathbf{D}}_m(\mathbf{x}_{m,\tau} - \mathbf{u}_m)\}_{m=1}^M$, i.e., the ith row, can be viewed as an estimate of source signal $s_i(\tau)$, which is hidden in the measurements $\mathbf{x}_{1,\tau}, \cdots, \mathbf{x}_{M,\tau}$. Since only a subset of entries of $\mathbf{x}_{m,\tau}$ contain information about source $s_i(\tau)$, the goal is to find a matrix $\mathbf{D}_m$ whose corresponding entries in the ith row of $\mathbf{D}_m$ can be nonzero and the rest of the entries are zero. Thus, the sparsity structure of $\mathbf{D}_m$ can be utilized to reveal which sensors sense the same source. Ideally, the support (nonzero entries indices) of each row of $\mathbf{D}_m$ will coincide with the indices in set $\mathcal{S}^i$. Hence, in order to identify which entries of $\{\mathbf{x}_{m,\tau}\}_{m=1}^M$ acquire information about the same sources and perform multimodal data clustering according to their information content, we introduce a norm-one regularized M-CCA framework.

Notation: Operators $\|\cdot\|_F$ and $\|\cdot\|_1$ correspond to the Frobenius norm and norm-one, respectively. $\mathbf{A}(i,:)$ and $\mathbf{A}(:,j)$ ($\mathbf{a}(k)$) denote the ith row and the jth column of matrix $\mathbf{A}$ (the kth element of vector $\mathbf{a}$). Further, $[\mathbf{M}]_{b:}$ and $\mathbf{M}(b,:)$ refer to the bth row of matrix $\mathbf{M}$. Also, $\text{sgn}(\mathbf{v})$ represents the entry-wise sign operator applied to vector $\mathbf{v}$.

III. Batch Multimodal Clustering

In order to identify the source-based clusters of entries in $\{\mathbf{x}_{m,\tau}\}_{m=1}^M$, we combine the M-CCA formulation in (2) with norm-one penalization. Thus, the sparsity induced matrices $\{\mathbf{D}_m\}_{m=1}^M$ can be obtained by minimizing the following sparse M-CCA formulation

$$\{\hat{\mathbf{D}}_m\}_{m=1}^M = \arg\min_{\mathbf{D}_1,\cdots,\mathbf{D}_M} 1/(2T)\cdot \sum_{\tau=1}^T \sum_{m=1}^M \sum_{n\neq m} \|\mathbf{D}_m(\mathbf{x}_{m,\tau} - \mathbf{u}_m) - \mathbf{D}_n(\mathbf{x}_{n,\tau} - \mathbf{u}_n)\|_2^2 + \varepsilon \sum_{m=1}^q \|\mathbf{D}_m\mathbf{\Sigma}_m \mathbf{D}_m^T - \mathbf{I}\|_F^2 + \sum_{m=1}^M \sum_{\rho=1}^q \lambda_{m,\rho}\|\mathbf{D}_m(\rho,:)\|_1, \tag{3}$$

The sparsity-controlling coefficients $\lambda_{m,\rho} > 0$ control the number of zero entries in row $\mathbf{D}_m(\rho,:)$. Also, the middle term in (3) is used to impose the uncorrelated structure in the source estimates $\mathbf{D}_m(\mathbf{x}_{m,\tau} - \mathbf{u}_m)$. Next, block coordinate descent (BCD) techniques are applied to derive an iterative solution. Specifically, within each coordinate descent cycle, we minimize (3) with respect to (w.r.t.) one element of $\mathbf{D}_m$, while fixing the remaining entries in $\mathbf{D}_m$ and all the elements in matrices $\{\mathbf{D}_n\}_{n=1,n\neq m}^M$ to their most recent updates. In the t-th coordinate cycle, given $\{\hat{\mathbf{D}}_n^{t-1}\}_{n=1,n\neq m}^M$, we update $\hat{\mathbf{D}}_m^t$ via

$$\hat{\mathbf{D}}_m^t = \arg\min_{\mathbf{D}_m} \varepsilon\|\mathbf{D}_m\mathbf{\Sigma}_m\mathbf{D}_m^T - \mathbf{I}\|_F^2 + \frac{1}{T}\sum_{\tau=1}^T \sum_{n\neq m} \|\mathbf{D}_m(\mathbf{x}_{m,\tau} - \mathbf{u}_m) - \hat{\mathbf{D}}_n^{t-1}(\mathbf{x}_{n,\tau} - \mathbf{u}_n)\|_2^2 + \sum_{\rho=1}^q \lambda_{m,\rho}\|\mathbf{D}_m(\rho,:)\|_1. \tag{4}$$

To avoid fourth-order polynomial terms in the cost function when minimizing (4) w.r.t. a single entry of $\mathbf{D}_m$, the second $\mathbf{D}_m$ in the term $\varepsilon\|\mathbf{D}_m\mathbf{\Sigma}_m\mathbf{D}_m^T - \mathbf{I}\|_F^2$ is fixed to its up-to-date estimate. Under this approximation, $\hat{\mathbf{D}}_m^t$ can be obtained by

$$\hat{\mathbf{D}}_m^t = \arg\min_{\mathbf{D}_m} \frac{1}{T}\sum_{n\neq m} \|\mathbf{D}_m\bar{\mathbf{X}}_m - \hat{\mathbf{D}}_n^{t-1}\bar{\mathbf{X}}_n\|_F^2 + \varepsilon\|\mathbf{D}_m\mathbf{\Sigma}_m(\hat{\mathbf{D}}_m^{t-1})^T - \mathbf{I}\|_F^2 + \sum_{\rho=1}^q \lambda_{m,\rho}\|\mathbf{D}_m(\rho,:)\|_1 \tag{5}$$

where $\bar{\mathbf{X}}_m := [\mathbf{x}_{m,1} - \mathbf{u}_m, \mathbf{x}_{m,2} - \mathbf{u}_m, \cdots, \mathbf{x}_{m,T} - \mathbf{u}_m] \in \mathbb{R}^{p_m\times T}$ and $\bar{\mathbf{X}}_n := [\mathbf{x}_{n,1} - \mathbf{u}_n, \mathbf{x}_{n,2} - \mathbf{u}_n, \cdots, \mathbf{x}_{n,T} - \mathbf{u}_n] \in \mathbb{R}^{p_n\times T}$ for $n = 1, \cdots, m-1, m+1, \cdots, M$.

Utilizing the coordinate descent iterations, the problem in (5) will be split into $q \cdot p_m$ scalar minimization subproblems, each of which focuses on a corresponding entry in $\mathbf{D}_m$. In detail, the problem in (5) is minimized w.r.t. one entry of $\mathbf{D}_m$, i.e., $\mathbf{D}_m(a,b)$, while fixing the matrices $\{\mathbf{D}_n\}_{n=1,n\neq m}^M$, as well as the $qp_m - 1$ remaining entries of $\mathbf{D}_m$ to their most recent updates. Then, the updates $\hat{\mathbf{D}}_m^t(a,b)$ for $a = 1, \cdots, q$ and $b = 1, \cdots, p_m$, can be achieved by solving the following minimization problem

$$\hat{\mathbf{D}}_m^t(a,b) = \arg\min_d \sum_{n\neq m} \|\boldsymbol{\alpha}_{a,b}^{m,n,t} - d\cdot\boldsymbol{\beta}_{a,b}^{m,t}\|_2^2 + \lambda_{m,a}\cdot|d| + \|\boldsymbol{\psi}_{a,b}^{m,t} - d\cdot\boldsymbol{\eta}_{a,b}^{m,t}\|_2^2, \tag{6}$$

where

$$\boldsymbol{\alpha}_{a,b}^{m,n,t} := T^{-0.5}([\hat{\mathbf{D}}_n^{t-1}\cdot\bar{\mathbf{X}}_n]_{a:} - \sum_{\ell=1,\ell\neq b}^{p_m} \hat{\mathbf{D}}_m^{t-1}(a,\ell)\cdot\bar{\mathbf{X}}_m(\ell,:)),\ \boldsymbol{\beta}_{a,b}^{m,t} := T^{-0.5}\cdot\bar{\mathbf{X}}_m(b,:),$$

$$\boldsymbol{\psi}_{a,b}^{m,t} := \varepsilon^{0.5}(\mathbf{I}_{a:} - \sum_{\ell=1,\ell\neq b}^{p_m} \hat{\mathbf{D}}_m^{t-1}(a,\ell)\cdot[\mathbf{\Sigma}_m\cdot(\hat{\mathbf{D}}_m^{t-1})^T]_{\ell:}),$$

$$\text{and } \boldsymbol{\eta}_{a,b}^{m,t} := \varepsilon^{0.5}[\mathbf{\Sigma}_m(\hat{\mathbf{D}}_m^{t-1})^T]_{b:}. \tag{7}$$

Since the problem in (6) is a scalar sparse regression problem, the minimizer can be obtained as (details can be found in [17])

$$\hat{\mathbf{D}}_m^t(a,b) = \text{sgn}((\mathbf{p}_{a,b}^{m,t})^T\mathbf{q}_{a,b}^{m,t}) \times \left[\max\left(0, \left(\left|\frac{(\mathbf{p}_{a,b}^{m,t})^T\mathbf{q}_{a,b}^{m,t}}{\|\mathbf{q}_{a,b}^{m,t}\|_2^2}\right| - \frac{\lambda_{m,a}}{2\|\mathbf{q}_{a,b}^{m,t}\|_2^2}\right)\right)\right], \tag{8}$$

where $\mathbf{p}_{a,b}^{m,t} := [\boldsymbol{\alpha}_{a,b}^{m,1,t}, \cdots, \boldsymbol{\alpha}_{a,b}^{m,m-1,t}, \boldsymbol{\alpha}_{a,b}^{m,m+1,t}, \cdots, \boldsymbol{\alpha}_{a,b}^{m,M,t}, \boldsymbol{\psi}_{a,b}^{m,t}]^T$, and $\mathbf{q}_{a,b}^{m,t} := [\boldsymbol{\beta}_{a,b}^{m,t}, \cdots, \boldsymbol{\beta}_{a,b}^{m,t}, \boldsymbol{\eta}_{a,b}^{m,t}]^T$.

In summary, the batch M-CCA framework involves the following three steps:

Step 1: Initialize $\hat{\mathbf{D}}_m^0$ randomly for $m = 1, \cdots, M$.

Step 2: For the t-th coordinate descent cycle, update $\hat{\mathbf{D}}_m^t(a,b)$ via (8) for $a = 1, \cdots, q$, $b = 1, \cdots, p_m$ and $m = 1, \cdots, M$.

ISBN 978-0-9928626-7-1 © EURASIP 2017

Step 3: If the M-CCA cost reduction is larger than a prespecified threshold go back to **Step 2**, otherwise return $\hat{\mathbf{D}}_m = \hat{\mathbf{D}}_m^t$ for $m = 1, \cdots, M$, and exit.

IV. Distributed Implementation

When sensors are spatially scattered distributed techniques are essential to enable localized clustering of the acquired multimodal data. Thus, a distributed version of the centralized BM-CCA scheme is developed, abbreviated as BDM-CCA under the assumptions that: $\mathbf{a}_1$) sensor $\mathcal{A}_j^m$ is responsible for updating the jth column of $\mathbf{D}_m$ and it only has access to the jth element of data vector $\mathbf{x}_{m,\tau}$ for $\tau = 1, \cdots, T$; $\mathbf{a}_2$) each sensor can only exchange information with its single-hop neighboring sensors; and $\mathbf{a}_3$) there exists at least one sensor in $\mathcal{A}^i$ which communicates with sensor $\mathcal{A}_j^m$, for $i = 1, \cdots, M$ and $i \neq m$. Considering that $\mathbf{D}_m\bar{\mathbf{X}}_m = \sum_{i=1}^{p_m} \mathbf{D}_m(:,i)\bar{\mathbf{X}}_m(i,:)$, then (3) can be reformulated as

$$\{\hat{\mathbf{D}}_m\}_{m=1}^M = \arg\min_{\mathbf{D}_1,\cdots,\mathbf{D}_M} \frac{1}{2T}\sum_{m=1}^M \sum_{n=1,n\neq m}^M \| \sum_{i=1}^{p_m} \mathbf{D}_m(:,i)\bar{\mathbf{X}}_m(i,:) - \sum_{i=1}^{p_n} \mathbf{D}_n(:,i)\bar{\mathbf{X}}_n(i,:)\|_F^2 + \varepsilon \sum_{m=1}^M \|\frac{1}{T}(\sum_{i=1}^{p_m} \mathbf{D}_m(:,i)\bar{\mathbf{X}}_m(i,:))((\sum_{i=1}^{p_m} \mathbf{D}_m(:,i)\bar{\mathbf{X}}_m(i,:))^T - \mathbf{I}\|_F^2 + \sum_{m=1}^M \sum_{\rho=1}^q \lambda_{m,\rho}\|\mathbf{D}_m(\rho,:)\|_1 \quad (9)$$

Combing BCD with the ADMM framework, the problem in (9) can be solved in a distributed fashion. To be specific, BCD is utilized to split (9) into $\sum_{m=1}^M p_m$ subproblems, each of which focuses on updating one submatrix of $\mathbf{D}_m$, namely the jth column of $\mathbf{D}_m$, i.e., $\mathbf{D}_m(:,j)$, by sensor $\mathcal{A}_j^m$. At the same time, ADMM will be employed to estimate the global value $\sum_{i=1}^{p_m} \mathbf{D}_m(:,i)\bar{\mathbf{X}}_m(i,:)$ in a distributed way. Toward this end, during the tth coordinate descent, sensor $\mathcal{A}_j^m$ updates $\hat{\mathbf{D}}_m^t(:,j)$ by minimizing the following cost w.r.t. variable $\mathbf{d}$

$$\frac{1}{T}\sum_{n\neq m} \| \sum_{i=1}^{p_m} \hat{\mathbf{D}}_m^{t-1}(:,i)\bar{\mathbf{X}}_m(i,:) - \sum_{i=1}^{p_n} \hat{\mathbf{D}}_n^{t-1}(:,i)\bar{\mathbf{X}}_n(i,:) - \hat{\mathbf{D}}_m^{t-1}(:,j)\bar{\mathbf{X}}_m(j,:) - \mathbf{d}\bar{\mathbf{X}}_m(j,:)\|_F^2 + \varepsilon\|\frac{1}{T}[(\sum_{i=1}^{p_m} \hat{\mathbf{D}}_m^{t-1}(:,i)\bar{\mathbf{X}}_m(i,:) - \hat{\mathbf{D}}_m^{t-1}(:,j)\bar{\mathbf{X}}_m(j,:) + \mathbf{d}\bar{\mathbf{X}}_m(j,:))(\sum_{i=1}^{p_m} \hat{\mathbf{D}}_m^{t-1}(:,i)\bar{\mathbf{X}}_m(i,:))^T] - \mathbf{I}\|_F^2 + \sum_{\rho=1}^q \lambda_{m,\rho}|\mathbf{d}(\rho)|. \quad (10)$$

Next, ADMM is applied to estimate the global quantity $\sum_{i=1}^{p_m} \hat{\mathbf{D}}_m^{t-1}(:,i)\bar{\mathbf{X}}_m(i,:)$ which involves information from all sensors in $\mathcal{A}^m$. ADMM will enable a distributed estimation of the aforementioned quantity via local updating recursions that will be obtained by minimizing the following constrained minimization problem

$$\hat{\mathbf{U}}_{m,j}^t = \arg\min_{\mathbf{U}_{m,j}^t} \sum_{j=1}^{p_m} \|\mathbf{U}_{m,j}^t - p_m\hat{\mathbf{D}}_m^{t-1}(:,j)\bar{\mathbf{X}}_m(j,:)\|_F^2 \quad \text{s. to } \mathbf{U}_{m,j}^t = \mathbf{U}_{m,j'}^t, \;\; j' \in \mathcal{N}_j^m \quad (11)$$

that formulates $\sum_{i=1}^{p_m} \hat{\mathbf{D}}_m^{t-1}(:,i)\bar{\mathbf{X}}_m(i,:)$ as the optimal solution, namely $\hat{\mathbf{U}}_{m,j}^t$, of a separable minimization problem that is amenable to distributed implementation. Note that the optimization variables $\mathbf{U}_{m,j}^t \in \mathbb{R}^{q\times T}$ correspond to a local estimate of $\sum_{i=1}^{p_m} \mathbf{D}_m(:,i)\bar{\mathbf{X}}_m(i,:)$ at sensor $\mathcal{A}_j^m$, and $\mathcal{N}_j^m$ denotes the set of neighboring sensors of sensor $\mathcal{A}_j^m$ which belong to $\mathcal{A}^m$. The task of solving (11) by sensor $\mathcal{A}_j^m$ boils down to the updating recursions (details of the procedure can be found in [5])

$$\mathbf{V}_{j,j'}^{m,t}(\kappa) = \mathbf{V}_{j,j'}^{m,t}(\kappa-1) + 0.5c(\hat{\mathbf{U}}_{m,j}^t(\kappa) - \hat{\mathbf{U}}_{m,j'}^t(\kappa))$$
$$\hat{\mathbf{U}}_{m,j}^t(\kappa+1) = (2 + 2c|\mathcal{N}_j^m|\mathbf{I})^{-1}[2p_m\hat{\mathbf{D}}_m^{t-1}(:,j)\bar{\mathbf{X}}_m(j,:) - \sum_{j'\in\mathcal{N}_j^m} (\mathbf{V}_{j,j'}^{m,t}(\kappa) - \mathbf{V}_{j',j}^{m,t}(\kappa) + c\hat{\mathbf{U}}_{m,j}^t(\kappa) + c\hat{\mathbf{U}}_{m,j'}^t(\kappa))]$$

where $\mathbf{V}_{j,j'}^{m,t}(\kappa)$ correspond to Lagrange multipliers locally updated at sensor $\mathcal{A}_j^m$ making sure that the consensus constraint $\mathbf{U}_{m,j}^t = \mathbf{U}_{m,j'}^t$ is satisfied. Further, index κ denotes the ADMM iteration index and $c > 0$ corresponds to a stepsize. A finite number of iterations, saying K, are carried out to estimate the quantity $\sum_{i=1}^{p_m} \hat{\mathbf{D}}_m^{t-1}(:,i)\bar{\mathbf{X}}_m(i,:)$. To this end, the local estimate $\hat{\mathbf{U}}_{m,j}^t(K)$ will be denoted as $\hat{\mathbf{U}}_{m,j}^t$. These estimates locally obtained across the different sensors are substituted in the minimization problem in (10) to obtain the following local formulation

$$\arg\min_{\mathbf{d}} 1/T\sum_{n\neq m} \|\hat{\mathbf{U}}_{m,j}^t - \hat{\mathbf{U}}_{n,j'(n)}^t + \hat{\mathbf{D}}_m^{t-1}(:,j)\bar{\mathbf{X}}_m(j,:) - \mathbf{d}\bar{\mathbf{X}}_m(j,:)\|_F^2 + \varepsilon\|1/T[(\hat{\mathbf{U}}_{m,j}^t - \hat{\mathbf{D}}_m^{t-1}(:,j)\bar{\mathbf{X}}_m(j,:) + \mathbf{d}\bar{\mathbf{X}}_m(j,:))(\hat{\mathbf{U}}_{m,j}^t)^T] - \mathbf{I}\|_F^2 + \sum_{\rho=1}^q \lambda_{m,\rho}|\mathbf{d}(\rho)| \quad (12)$$

in which, $\hat{\mathbf{U}}_{n,j'(n)}^t$ corresponds to the local estimate of $\sum_{i=1}^{p_n} \hat{\mathbf{D}}_n^{t-1}(:,i)\bar{\mathbf{X}}_n(i,:)$ at sensor $\mathcal{A}_{j'(n)}^n$, which corresponds to one of the neighboring sensor of sensor $\mathcal{A}_j^m$ in set $\mathcal{A}^n$. Notice that, the quantities $\hat{\mathbf{U}}_{n,j'(n)}^t$ in (12) can be obtained at sensor $\mathcal{A}_m^j$ by communicating with his single hop neighbors $j'(n)$ in set $\mathcal{A}^n$, while the rest of the quantities are available locally at sensor $\mathcal{A}_j^m$ (cf. assumption $\mathbf{a}_3$). To minimize the cost in (12) coordinate descent is employed to divide the problem in (12) into q subproblems each of which focuses on one entry of $\mathbf{D}_m(:,j)$, which can be solved using a similar procedure as the one described in Sec. III to obtain recursions similar to (8).

To summarize BDM-CCA algorithm, at each coordinate descent cycle, every sensor, i.e., $\mathcal{A}_j^m$, relying on communication only with its neighboring sensors in sets $\mathcal{A}^n$, completes the following two steps: 1) carries out K ADMM iterations to find the estimate$\hat{\mathbf{U}}_{m,j}^t$; and 2) solves the minimization problem in (12) by splitting it into q scalar minimization subtasks.

V. Online Processing

To reduce computational complexity and enable processing of a large volume of data, we design an online framework (abbreviated OM-CCA) that entails regularized M-CCA to process efficiently large data sets. Online clustering is pertinent for a setting where data are constantly acquired, making

ISBN 978-0-9928626-7-1 © EURASIP 2017

necessary real-time processing. Starting from the batch cost in (3), we build the online formulation

$$\{\hat{\mathbf{D}}_m^\tau\}_{m=1}^M = \arg\min_{\mathbf{D}_1,\cdots,\mathbf{D}_M} \frac{1}{2}\sum_{m=1}^M \sum_{n=1,n\neq m}^M \|\mathbf{D}_m \cdot (\mathbf{x}_{m,\tau} - \hat{\mathbf{u}}_m^\tau) - \mathbf{D}_n \cdot (\mathbf{x}_{n,\tau} - \hat{\mathbf{u}}_n^\tau)\|_2^2 + \sum_{m=1}^M \sum_{\rho=1}^q \lambda_{m,\rho,\tau}\|\mathbf{D}_m(\rho,:)\|_1 + \varepsilon \sum_{m=1}^M \|\mathbf{D}_m \hat{\mathbf{\Sigma}}_m^\tau \cdot \mathbf{D}_m^T - \mathbf{I}\|_F^2 \quad (13)$$

which emphasizes only to the present data at time instant τ, while $\hat{\mathbf{u}}_m^\tau = \frac{1}{\tau}\sum_{t=1}^\tau \mathbf{x}_{m,t}$, and $\hat{\mathbf{\Sigma}}_m^\tau = \tau^{-1}\sum_{t=1}^\tau (\mathbf{x}_{m,t} - \hat{\mathbf{u}}_m^\tau)(\mathbf{x}_{m,t} - \hat{\mathbf{u}}_m^\tau)^T$. Note that the quantities $\hat{\mathbf{u}}_m^\tau$ and $\hat{\mathbf{\Sigma}}_m^\tau$ can be updated in an online recursive fashion that does not require storage of all the data acquired so far. To be specific, $\hat{\mathbf{u}}_m^\tau = \frac{\tau-1}{\tau}\hat{\mathbf{u}}_m^{\tau-1} + \frac{1}{\tau}\mathbf{x}_{m,\tau}$ and a similar updating formula can be derived for $\hat{\mathbf{\Sigma}}_m^\tau$. Further, time-decreasing $\lambda_{m,\rho,\tau}$ is introduced to induce entry-wise sparsity in $\mathbf{D}_m(\rho,:)$.

To facilitate applicability of block coordinate descent, the problem in (13) will be split into M subproblems, each of which updates $\mathbf{D}_m$ while fixing the matrices $\{\mathbf{D}_n\}_{n=1,n\neq m}^M$ to their latest updates. Further, subgradient descent [4] is employed to estimate $\hat{\mathbf{D}}_m^\tau$ at time instant τ according to the updating rule

$$\hat{\mathbf{D}}_m^\tau = \hat{\mathbf{D}}_m^{\tau-1} - c_\tau \cdot \nabla f(\hat{\mathbf{D}}_m^{\tau-1}), \quad (14)$$

where $c_\tau > 0$ represents the step-size used in the subgradient descent method, while $\nabla f(\hat{\mathbf{D}}_m^{\tau-1})$ denotes a sub-gradient of (13) w.r.t. $\mathbf{D}_m$ evaluated at $\hat{\mathbf{D}}_m^{\tau-1}$, which equals to

$$\nabla f(\hat{\mathbf{D}}_m^{\tau-1}) = 2\cdot\sum_{n=1,n\neq m}^M [\hat{\mathbf{D}}_m^{\tau-1}\cdot(\mathbf{x}_{m,\tau} - \hat{\mathbf{u}}_m^\tau) - \hat{\mathbf{D}}_n^{\tau-1} \cdot (\mathbf{x}_{n,\tau} - \hat{\mathbf{u}}_n^\tau)]\cdot(\mathbf{x}_{m,\tau} - \hat{\mathbf{u}}_m^\tau)^T + 4\cdot\varepsilon(\hat{\mathbf{D}}_m^{\tau-1}\cdot\hat{\mathbf{\Sigma}}_m^\tau\cdot(\hat{\mathbf{D}}_m^{\tau-1})^T - \mathbf{I})\cdot\hat{\mathbf{D}}_m^{\tau-1}\cdot\hat{\mathbf{\Sigma}}_m^\tau + \begin{bmatrix} \lambda_{m,1}\cdot\text{sgn}(\hat{\mathbf{D}}_m^{\tau-1}(1,:)) \\ \cdot \\ \cdot \\ \lambda_{m,q}\cdot\text{sgn}(\hat{\mathbf{D}}_m^{\tau-1}(q,:)) \end{bmatrix}.$$

VI. Numerical Tests

In this section, the clustering performance in terms of the probability of correctly clustering the sensor data based on their source content for BM-CCA, BDM-CCA, and OM-CCA is tested and compared with i) the traditional M-CCA (TM-CCA) which is obtained after applying $\{\lambda_{m,\rho} = 0\}_{m=1,\rho=1}^{M,q}$ to BM-CCA, ii) an online clustering algorithm (OCE) in [7], and iii) the traditional K-means algorithm [13].

The sparsity-controlling coefficients in BM-CCA are selected using **Alg. 1**, where a dominant entry refers to the entry with the largest absolute value with respect to the other entries in the same column of a matrix. A step-size of $\delta_\lambda = 0.01$ is used to increase or decrease the sparsity-controlling coefficients $\lambda_{m,\rho}$. Intuitively, **Alg. 1** increases or decreases the sparsity controlling coefficients such that at most one dominant entry is present in each column of $\mathbf{D}_m$, depending on whether a sensor observes a source or not. If a row of $\mathbf{D}_m$ does not have a dominant entry this implies that all entries are small in magnitude, thus λs are decreased to introduce stronger entries. On the other hand if there are nonzero columns λs are increased to identify noisy sensors with no information.

During the implementation of OM-CCA, two types of parameters are involved, i.e., $\{c_\tau\}_{\tau=1}^T$ and $\{\lambda_{m,\rho,\tau}\}_{m,\rho,\tau=1}^{M,q,T}$. To simplify the process of choosing those parameters, we assume that $c_\tau = c^0/\tau^{\omega_1}$ and $\lambda_{m,\rho,\tau} = \lambda_{m,\rho}^0/\tau^{\omega_2}$, where ω_1, ω_2, c^0, and $\lambda_{m,\rho}^0$ are positive scalars. Also, we make the assumption that $\omega_1 = \omega_2 = 1.05$ and $\omega_1 = 0.51$, while c^0 and $\lambda_{m,\rho}^0$ are set to a sufficiently small value to ensure convergence of OM-CCA. General selection rules that guarantee convergence of OM-CCA are currently under investigation. Furthermore, when testing the distributed BDM-CCA approach the number of ADMM iterations, namely K, utilized to estimate pertinent global quantities is set to $K = 20$.

Algorithm 1 Parameter Selection for BM-CCA

Initialize $\{\lambda_{m,\rho}\}_{m=1,\rho=1}^{M,q}$ to a small value (e.g., 0.1).
while(**true**)
Estimate $\{\hat{\mathbf{D}}_m\}_{m=1}^M$ via BM-CCA using $\{\lambda_{m,\rho}\}_{m=1,\rho=1}^{M,q}$.
If $\forall m$, and $\forall\rho$, $\hat{\mathbf{D}}_m(\rho,:)$ has at least one dominant entry, and $\forall m$ $\hat{\mathbf{D}}_m$ has some all-zeros columns.
Break while.
else if
If $\hat{\mathbf{D}}_m(\rho,:)$ does not have a dominant entry, for $m = 1,\cdots,M$ and $\rho = 1,\cdots,q$
$\lambda_{m,\rho} \leftarrow \lambda_{m,\rho} - \delta_\lambda$.
end if
If All the columns of $\hat{\mathbf{D}}_m$ are non-zero columns for $m = 1,\cdots,M$
$\lambda_{m,\rho} \leftarrow \lambda_{m,\rho} + \delta_\lambda$ for $\rho = 1,\cdots,q$.
end if
Estimate $\{\hat{\mathbf{D}}_m\}_{m=1}^M$ via BM-CCA.
end if
end while

Two sources ($q = 2$) are generated that adhere to a first-order autoregressive (AR) model. $M = 4$ types of sensors are considered with $\{p_m = 15\}_{m=1}^{15}$ sensors in each type. In the testing settings considered: i) $4, 5, 6$ and 5 sensors from the each of the 4 different sensing types, respectively, are randomly assigned to observe source $s_1(\tau)$; while ii) $5, 5, 3$ and 4 sensors from the remaining sensors in each of the different sets $\mathcal{A}^m$ for $m = 1, 2, 3, 4$ are randomly assigned to sense source $s_2(\tau)$. The remaining sensors in these four sets are noninformative sensors and observe just noise. Both a linear and nonlinear setting are considered for testing here. In the linear data setting, the mapping function in (1) is set as $h_{m,j}(s_{f(m,j)}(\tau)) = s_{f(m,j)}(\tau)$, while in the nonlinear data setting, $h_{m,j}(s_{f(m,j)}(\tau)) = (s_{f(m,j)}(\tau))^\phi$, where ϕ is randomly chosen from the set of values $\{1, 1.1, 1.3, 1.4, 1.5, 1.6\}$.

We compare the clustering performance of BM-CCA, BDM-CCA, OM-CCA, TM-CCA, OCE and K-means under both linear and nonlinear data settings, denoted by $(\cdot) - L$ and $(\cdot) - NL$ in Fig.1, respectively. Fig. 1 demonstrates that the novel BM-CCA, BDM-CCA, and OM-CCA yield better clustering performance as the number of data increases. Interestingly, BM-CCA, BDM-CCA, and OM-CCA achieve higher clustering probabilities than TM-CCA, K-means and

ISBN 978-0-9928626-7-1 © EURASIP 2017

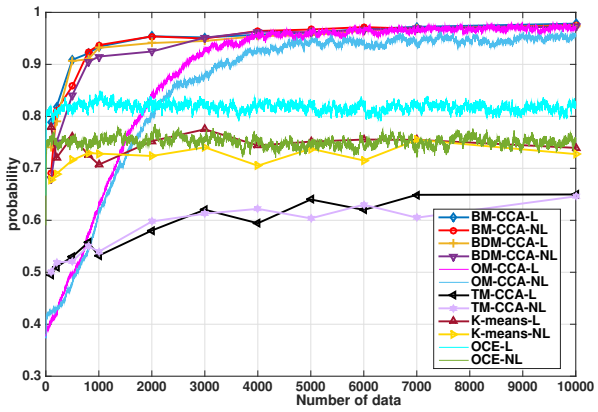

Fig. 1. Probability of correct data clustering vs. number of data τ.

OCE. Comparing BM-CCA with TM-CCA, it can be seen that introducing sparsity via norm-one regularization in M-CCA significantly improves the clustering performance by 46%. Also, it can be seen that the performance of the distributed approach BDM-CCA is approaching the performance of BM-CCA in both linear and nonlinear data settings, for a sufficiently large number of ADMM iterations. In fact, as K goes to infinity BDM-CCA will coincide with BM-CCA. Further, the online approach OM-CCA for small number of data has worse performance than the batch approaches BM-CCA and BDM-CCA. Nonetheless, performance gradually improves and reaches the batch performance as the number of data increases.

The real advantage of the OM-CCA is the computational savings introduced by utilizing the online formulation in (13). Specifically, for the linear setting we plot the average running time among 100 independent Monte Carlo tests for the algorithms BM-CCA, OM-CCA, OCE as well as K-means. **Table 1** depicts that the proposed OM-CCA runs (running time is in seconds) as fast as K-means, while achieving better clustering performance. Further, OM-CCA is much more computationally efficient than BM-CCA by achieving similar clustering performance in only 3 times less running time on average compared to BM-CCA.

T	BM-CCA	OM-CCA	OCE	K-means
1000	32.6	14.95	13.86	10.91
2000	33	16.14	18.14	12.14
5000	85.2	20.6	30.08	16.9
10000	96.5	28.7	58.75	28.67

TABLE I
AVERAGE RUNNING TIME (SEC.) FOR CLUSTERING. RESULTS ARE OBTAINED IN A 8GB RAM MACHINE WITH 3.0 GHZ PROCESSOR.

VII. CONCLUSION

The problem of clustering multimodal data according to their information content was explored. Both a batch and online implementation was considered, trading-off accuracy for computational complexity. The effectiveness of the online approach was demonstrated in both linear and nonlinear settings. The alternating direction method of multipliers was utilized to derive a distributed implementation of the centralized batch approach enabling localized multimodal data clustering across spatially scattered sensors. Numerical tests demonstrate the potential of the novel approaches over existing alternatives, while demonstrating the computational efficiency of the online approach in terms of significantly smaller running time.

REFERENCES

[1] D. P. Bertsekas, *Nonlinear Programming.* 2nd Edition, Athena Scientific, Massachussets, 1999.
[2] A. Banerjee, 1. Dhillon, J. Ghosh, and S. Sra, "Generative Model-Based Clustering of Directional Data," *Proc. of ACM SIGKDD Intl. Conf. on Knowledge Disc. and Data Mining*, Washington, DC, pp. 19–28, 2003.
[3] N. Bansal, A. Blum, and S. Chawla, "Correlation Clustering," *Machine Learning*, vol. 56, no.1-3, pp. 89–113, 2004.
[4] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient Methods," Lecture notes of EE392o, Stanford University, Autumn Quarter 2003–2004.
[5] J. Chen and I. D. Schizas, "Online Distributed Sparsity-Aware Canonical Correlation Analysis," *IEEE Trans. on Signal Processing*, vol. 64, no. 3, pp. 688–703, 2016.
[6] J. Chen and I. D. Schizas, "Distributed Information-Based Clustering of Heterogeneous Sensor Data," *Elsevier Signal Processing*, vol. 126, pp. 35–51, September 2016.
[7] A. Choromanska, and C. Monteleoni, "Online Clustering with Experts," *In Proc. Of 15th Int. Conf. on Artificial Intelligence and statistics (AISTATS)*, La Palma, Canary Islands, pp. 1–18, Apr. 2012.
[8] N. M. Correa, T. Adali, Y. O. Li, and V. D. Calhoun, "Canonical Correlation Analysis for Data Fusion and Group Inferences," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 39–50, 2010.
[9] S. M. Kakade and D. P. Foster, "Multi-view Regression Via Canonical Correlation Analysis," *Conf. Learning Thy*, pp. 82-96, 2007.
[10] J. R. Kettenring, " Canonical Analysis of Several Sets of Variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
[11] G. Lee, A. Singanamalli, H. Wang, M. D. Feldman, S. R. Master, N. N. C. Shih, E. Spangler, T. Rebbeck, J. E. Tomaszewski and A. Madabhushi "Supervised Multi-View Canonical Correlation Analysis (sMVCCA): Integrating Histologic and Proteomic Features for Predicting Recurrent Prostate Cancer," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, pp. 284-297, Jan. 2015.
[12] D. Lin, J. Zhang, J. Li, V. D. Calhoun, H.-W. Deng and Y. P. Wang, "Group Sparse Canonical Correlation Analysis for Genomic Data Integration," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–16, 2013.
[13] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Inf. Theory,* vol. 28, no. 2, pp. 129–137, Mar. 1982.
[14] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor Canonical Correlation Analysis for Multi-view Dimension Reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3111–3124, 2015.
[15] A. A. Nielsen, "Multiset Canonical Correlations Analysis and Multispectral Truly Multi-Temporal Remote Sensing Data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.
[16] J. Rupnik and J. S. Taylor, "Multi-view Canonical Correlation Analysis," *Proc. Slovenian KDD Conf. Data Mining and Data Warehouses (SiKDD)*, pp. 1–4, Ljubljana, Slovenia, Oct. 2010.
[17] I. D. Schizas and G. B. Giannakis, "Covariance Eigenvector Sparsity for Data Compression and Denoising," *IEEE Trans. on Signal Processing*, vol. 60, no. 5, pp. 2408–2421, May 2012.
[18] R. Xu and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

ISBN 978-0-9928626-7-1 © EURASIP 2017