REGULARIZED CANONICAL CORRELATIONS FOR SENSOR DATA CLUSTERING

Jia Chen and Ioannis D. Schizas *

*Department of EE, Univ. of Texas at Arlington, 416 Yates Street, Arlington, TX 76010, USA

ABSTRACT

The task of determining informative sensors and clustering the sensor measurements according to their information content is considered. To this end, the standard canonical correlation analysis (CCA) framework is equipped with norm-one and norm-two regularization terms to estimate the unknown number of field sources and identify informative groups of sensors. Coordinate descent techniques are combined with the alternating direction method of multipliers to derive an algorithm that minimizes the regularized CCA framework. An efficient scheme to properly select the regularization coefficients associated with the norm-one and norm-two terms is also developed. Numerical tests corroborate that the novel scheme outperforms existing alternatives.

Index Terms— Canonical correlation analysis, sparsity, clustering, optimization

1. INTRODUCTION

In sensor networks, the acquired sensor data contain information about multiple sources, unknown in number, placed at different spatial locations. Such sources could correspond to different e.g., thermal sources or transmitters located inside the sensed field. Before applying any statistical inference task, it is essential to determine which groups of sensor observations contain information about the same sources. This is essential to avoid 'mixing' observations that contain information about uncorrelated sources and isolate sensors that sense just noise.

In this paper a framework for grouping sensors based on their information content is put forth which is able to deal with nonlinear settings and unknown number of sources. Sensor measurements containing information about the same sources are statistically correlated. To exploit such spatial correlations, CCA, see e.g., [6, 11], is regularized with norm-one and norm-two terms [17, 19, 22] to obtain a framework that can extract correlated sensor data and cluster them in groups. Existing norm-one regularized CCA formulations [7, 10, 20] seek to maximize the correlation between two data sets, while performing variable selection, while [8] assumes the number of sources is available.

To this end, the alternating direction method of multipliers (ADMM) (see e.g., [4, 5]) and coordinate descent techniques [3, 18] are put forth to minimize the regularized CCA framework. The resulting iterative scheme involves simple updating recursions that perform the task of sensor clustering at a fusion center. A simple yet effective scheme is also put forth to appropriately select the regularization coefficients. A number of different approaches have been put forth to address the problem of clustering data into different groups that share similar properties. The K-means algorithm [13] is one of the major representatives when it comes to data clustering. Clusters are represented by centroid points and the idea is to allocate each data vector to the cluster that has the most similar centroid with respect to a distance metric. Variations that rely on underlying probabilistic models and/or pertinent similarity measures have been developed [2, 21], while clustering techniques for an unknown number of clusters have also been proposed [9, 14]. The challenge in our setting stems from the fact that the type of similarity between sensor data containing information about the same source is unknown due to the unavailability of the underlying data model. Thus, the fact that sensor measurements observe the same source does not necessarily make them similar in e.g., magnitude or with respect to distance metrics used by the above schemes. Numerical tests will demonstrate the advantages of our approach over existing sensor clustering alternatives.

This work is supported by the NSF grant CCF 1218079 and UTA.

2. PROBLEM FORMULATION

Consider p sensors that monitor a field formed by an unknown number of M zero-mean and spatially uncorrelated sources which are modeled as stationary random variables $s_m(t)$ for $t = 0, 2, ..., N-1$ and $m = 1, ..., M$. Sensor j acquires at time instant t scalar measurement $x_j(t)$ that adheres to the following *unknown* model

$$x_j(t) = \sum_{m=1}^{M} g_{m,j}(s_m(t)) + w_j(t), \quad j = 1, 2, ..., p \qquad (1)$$

where $g_{m,j}(\cdot)$ is a random nonlinear mapping (if sensor j does not sense source m then $g_{m,j}(\cdot) = 0$), while $w_j(t)$ is zero-mean white noise signal that is independent of the source signal $s_m(t)$. Denote $\boldsymbol{\chi}(t) := [x_1(t) \dots x_p(t)]^T \in \mathbb{R}^{p\times 1}$ as the measurements across the p sensors which are transmitted to a fusion center.

It is assumed that the field sources are quite localized and affect a small percentage of sensors in the network. This further implies that different subsets of entries in $\boldsymbol{\chi}(t)$ will contain information about different field sources. Let $\mathcal{S}^m$ denote the subset of entries of $\boldsymbol{\chi}(t)$ that contain information about source $s_m(t)$, and let $\mathcal{S}^0$ denote the subset of sensors whose measurements do not contain information about any of the sources, e.g., sensors that acquire sensing noise. For instance, in a network of $p = 12$ sensors that observe a field with $M = 2$ sources, namely $s_1(t)$ and $s_2(t)$, if sensors $\{1, 2, 3, 7\}$ observe source $s_1(t)$, while sensors $\{4, 5, 6, 9\}$ observe source $s_2(t)$ and the rest of the sensors $\{8, 10, 11, 12\}$ acquire just noise, then $\mathcal{S}^0 = \{8, 10, 11, 12\}$, $\mathcal{S}^1 = \{1, 2, 3, 7\}$ and $\mathcal{S}^2 = \{4, 5, 6, 9\}$. The goals here is to estimate the unknown number of sources, and cluster the entries in $\boldsymbol{\chi}(t)$ according to their unknown source content.

Entries in $\boldsymbol{\chi}(t)$ that contain information about the same source are correlated. CCA is an effective way to extract common features that are present in two data sequences and result correlations, see e.g., [6]. Given the data sequence $\boldsymbol{\chi}(t)$, the following two data sequences are build representing the past and present/future of the sensor measurements at time instant t,

$$\mathbf{x}(t) = \Big[\boldsymbol{\chi}^T(t-1), \boldsymbol{\chi}^T(t-2), \ldots, \boldsymbol{\chi}^T(t-f)\Big]^T \quad (2)$$

$$\mathbf{y}(t) = \Big[\boldsymbol{\chi}^T(t), \boldsymbol{\chi}^T(t+1), \ldots, \boldsymbol{\chi}^T(t+f-1)\Big]^T \quad (3)$$

where $f \geq 0$ controls the memory length.
Given the data pairs $\{\mathbf{x}(t), \mathbf{y}(t)\}_{\tau=1}^{T} \in \mathbb{R}^{pf\times 1}$ the traditional CCA is utilized to *linearly* extract common correlated features from them [6, Chpt. 10], [11]. The latter task is performed by searching for matrices $\mathbf{E} \in \mathbb{R}^{q\times pf}$ and $\mathbf{D} \in \mathbb{R}^{q\times pf}$ with $q \leq pf$ that minimize

$$(\hat{\mathbf{E}}, \hat{\mathbf{D}}) = \arg\min_{\mathbf{E},\mathbf{D}} N^{-1} \textstyle\sum_{t=0}^{N-1} \|\mathbf{E}\mathbf{y}(t) - \mathbf{D}\mathbf{x}(t) - \hat{\boldsymbol{\mu}}\|_2^2, \quad (4)$$

$$\text{s. to } \mathbf{E}\hat{\boldsymbol{\Sigma}}_y\mathbf{E}^T = \mathbf{I}, \;\; \mathbf{D}\hat{\boldsymbol{\Sigma}}_x\mathbf{D}^T = \mathbf{I}, \quad (5)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. The matrix $\hat{\boldsymbol{\Sigma}}_x := N^{-1}\sum_{t=0}^{N-1}(\mathbf{x}(t) - \hat{\boldsymbol{\mu}}_x)(\mathbf{x}(t) - \hat{\boldsymbol{\mu}}_x)^T$ denotes the sample-average estimate for the covariance matrix of measurements $\mathbf{x}(t)$, while $\hat{\boldsymbol{\mu}}_x$ is the sample-average estimate for the expectation of $\mathbf{x}(t)$. The covariance matrix $\hat{\boldsymbol{\Sigma}}_y$ is defined similarly, while $\hat{\boldsymbol{\mu}} := \mathbf{E}\hat{\boldsymbol{\mu}}_y - \mathbf{D}\hat{\boldsymbol{\mu}}_x$. The optimal matrices $\hat{\mathbf{E}}$ and $\hat{\mathbf{D}}$ can be found in, e.g., [6, pg. 370].

Intuitively, the reason for imposing the *canonical variates* $\hat{\mathbf{E}}\mathbf{y}(t)$ and $\hat{\mathbf{D}}\mathbf{x}(t)$ to be as similar as possible is the fact that each entry of these estimators try to uncover the common sources sensed in $\mathbf{x}(t)$ and $\mathbf{y}(t)$. However, in order for the latter task to be carried out in standard CCA the number of sources has to be known, i.e., $q = M$. Standard CCA has the ability to identify common components (sources) contained in both $\mathbf{x}(t)$ and $\mathbf{y}(t)$, however is not capable of identifying *which* entries in $\mathbf{x}(t)$ and $\mathbf{y}(t)$ contain information about the same source. A necessary ingredient is the proper introduction of zeros (sparsity) in the CCA matrices $\mathbf{E}$ and $\mathbf{D}$ in a way such that the nonzero entries in each row of $\mathbf{E}$ and $\mathbf{D}$ will point to these entries in $\mathbf{x}(t)$ and $\mathbf{y}(t)$ (and subsequently $\boldsymbol{\chi}(t)$) that contain information about a common source. For instance in the example presented earlier, there must be a row in $\mathbf{E}$ and $\mathbf{D}$ with nonzeros in those entries with indices $\{1, 2, 3, 7\}$ corresponding to the sensors acquiring information about source $s_1(t)$. Similarly, another row of $\mathbf{E}$ and $\mathbf{D}$ should have nonzeros in entries $\{4, 5, 6, 9\}$, corresponding to the sensors observing source $s_2(t)$. To this end, we enhance CCA with norm-one and norm-two regularization to properly induce sparsity in $\mathbf{E}$ and $\mathbf{D}$, and cope with the unknown number of sources M. In the following the ρth row of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{\rho:}$, or, $(\mathbf{A})(\rho:)$, while $\mathbf{A}(\alpha, \beta)$ denotes the (α, β) entry of $\mathbf{A}$. Also, $\mathbf{I}_q$ denotes the identity matrix of size $q \times q$.

3. REGULARIZED CCA

To deal with the unknown number of sources the number of rows q in $\mathbf{E}$ and $\mathbf{D}$ is selected larger than M. Then, norm-two regularization is utilized, e.g., see [1, 19], to zero-out the extra rows that are not needed. Further, to induce entry-wise sparsity in the rows of $\mathbf{E}$ and $\mathbf{D}$ and subsequently identify different subsets of entries in $\boldsymbol{\chi}(t)$ that contain information about a source, norm-one regularization terms, see e.g., [17, 20, 22], are introduced in the standard CCA in (4).
The following regularized CCA framework is considered

$$\begin{aligned}(\hat{\mathbf{D}}, \hat{\mathbf{E}}) &= \arg\min_{\mathbf{D},\mathbf{E}} N^{-1}\textstyle\sum_{t=0}^{N-1} \|\mathbf{E}\mathbf{y}(t) - \mathbf{D}\mathbf{x}(t) - \hat{\boldsymbol{\mu}}\|_2^2 \\ &+ \upsilon\|\mathbf{E}\hat{\boldsymbol{\Sigma}}_y\mathbf{E}^T - \mathbf{I}\|_F^2 + \varepsilon\|\mathbf{D}\hat{\boldsymbol{\Sigma}}_x\mathbf{D}^T - \mathbf{I}\|_F^2 \\ &+ \textstyle\sum_{\rho=1}^{q}\lambda_{E,\rho}\|\mathbf{E}_{\rho:}\|_1 + \sum_{\rho=1}^{q}\lambda_{D,\rho}\|\mathbf{D}_{\rho:}\|_1 \\ &+ \phi_D\textstyle\sum_{\rho=1}^{q}\|\mathbf{D}_{\rho:}\|_2 + \phi_E\sum_{\rho=1}^{q}\|\mathbf{E}_{\rho:}\|_2 \end{aligned} \quad (6)$$

where the second and third terms in (6) account for the constraints in (4), while $\|\cdot\|_1$ and $\|\cdot\|_F$ denote norm-one and Frobenius norm, respectively. The sparsity controlling coefficients $\lambda_{E,\rho}$ and $\lambda_{D,\rho}$ assume positive values and control the number of zero entries in $\mathbf{E}_{\rho:}$ and $\mathbf{D}_{\rho:}$, respectively. The last two Euclidean norm terms induce group sparsity, which is used to zero-out unnecessary rows in $\mathbf{E}$ and $\mathbf{D}$ by adjusting properly the coefficients $\phi_D > 0$ and $\phi_E > 0$. The number of nonzero rows in $\mathbf{E}$ and $\mathbf{D}$ will be an estimate of the number of field sources.

To tackle the problem in (6) its corresponding cost will be split into smaller minimization tasks involving minimization with respect to (wrt) one row of $\mathbf{E}$, say $\mathbf{E}_{\rho:}$ (or $\mathbf{D}$, say $\mathbf{D}_{\rho:}$) while fixing the rest entries in $\mathbf{E}$ and $\mathbf{D}$ to their most up-to-date values. Each of these minimization subtasks will be tackled using the alternating direction method of multipliers (ADMM), see e.g., [4, 5], to solve the problem in (6). ADMM is an efficient technique used to minimize costs having sparsity-inducing regularization terms, e.g., see [16].

3.1. Algorithm Implementation

The minimization problem in (6) is split into $2q$ minimization subproblems. Each of these minimization subtasks focuses on minimizing the cost in (6) wrt $\mathbf{D}_{\rho:}$ (or $\mathbf{E}_{\rho:}$) for $\rho = 1, 2, ..., q$, while fixing the remaining rows of $\mathbf{E}$ (or $\mathbf{D}$). Specifically, the minimization subtask wrt $\mathbf{D}_{\rho:}$ can be written as,

$$\begin{aligned}\hat{\mathbf{D}}_{\rho:} &= \arg\min_{\mathbf{D}_{\rho:}} N^{-1}\|\mathbf{E}_{\rho:}\mathbf{Y} - \mathbf{D}_{\rho:}\mathbf{X}\|_2^2 + \lambda_{D,\rho}\|\mathbf{D}_{\rho:}^T\|_1 \\ &+ \phi_D\|\mathbf{D}_{\rho:}\|_2 + \varepsilon\|\mathbf{I}_q(\rho,:) - \mathbf{D}_{\rho:}\boldsymbol{\Sigma}_x\mathbf{D}\|_2^2 \end{aligned} \quad (7)$$

where $\mathbf{X} := [\mathbf{x}(0) - \hat{\boldsymbol{\mu}}_x, \mathbf{x}(2) - \hat{\boldsymbol{\mu}}_x, ..., \mathbf{x}(N-1) - \hat{\boldsymbol{\mu}}_x] \in \mathbf{R}^{pf\times N}$ and $\mathbf{Y} := [\mathbf{y}(0) - \hat{\boldsymbol{\mu}}_y, \mathbf{y}(2) - \hat{\boldsymbol{\mu}}_y, ..., \mathbf{y}(N-1) - \hat{\boldsymbol{\mu}}_y] \in \mathbf{R}^{pf\times N}$. For notational simplicity, we denote $\mathbf{E}_{\rho:}\mathbf{Y}$, $\boldsymbol{\Sigma}_x\mathbf{D}^T$ and $\mathbf{I}_q(\rho,:)$ by $\mathbf{T}_{1,\rho}^E$, $\mathbf{T}_2^D$ and $\mathbf{T}_{3,\rho}$, respectively. First $\mathbf{D}_{\rho:}$ is updated while using the most recent updates from iteration τ and replacing $\mathbf{T}_{1,\rho}^E$ with $\mathbf{T}_{1,\rho}^{E,\tau} = \mathbf{E}_{\rho:}^{\tau}\mathbf{Y}$,and $\mathbf{T}_2^D$ with $\mathbf{T}_2^{D,\tau} = \boldsymbol{\Sigma}_x(\mathbf{D}^{\tau})^T$. Then, $\hat{\mathbf{D}}_{\rho:}^{\tau+1}$ can be obtained as

$$\begin{aligned}\hat{\mathbf{D}}_{\rho:}^{\tau+1} &= \arg\min_{\mathbf{D}_{\rho:}} N^{-1}\|\mathbf{T}_{1,\rho}^{E,\tau} - \mathbf{D}_{\rho:}\mathbf{X}\|_2^2 + \lambda_{D,\rho}\|\mathbf{D}_{\rho:}\|_1 \\ &+ \phi_D\|\mathbf{D}_{\rho:}\|_2 + \varepsilon\|\mathbf{T}_{3,\rho} - \mathbf{D}_{\rho:}\mathbf{T}_2^{D,\tau}\|_2^2 \end{aligned} \quad (8)$$

The problem in (8) can be converted into the equivalent constrained minimization problem

$$\begin{aligned}(\hat{\mathbf{D}}_{\rho:}^{\tau+1}, \mathbf{b}_{\rho}^{\tau+1}) &= \arg\min_{\mathbf{D}_{\rho:},\mathbf{b}_\rho} \frac{1}{N}\|\mathbf{T}_{1,\rho}^{E,\tau} - \mathbf{D}_{\rho:}\mathbf{X}\|_2^2 + \lambda_{D,\rho}\|\mathbf{D}_{\rho:}\|_1 \\ &+ \phi_D\|\mathbf{b}_\rho\|_2 + \varepsilon\|\mathbf{T}_{3,\rho} - \mathbf{D}_{\rho:}\mathbf{T}_2^{D,\tau}\|_2^2, \text{ subj. to } \mathbf{b}_\rho = \mathbf{D}_{\rho:}, \end{aligned} \quad (9)$$

where $\mathbf{b}_\rho$ is an extra minimization variable introduced to facilitate applicability of ADMM. ADMM involves updates for $\mathbf{D}_{\rho:}$, $\mathbf{b}_\rho$ and the Lagrange multiplier vector $\mathbf{p}_\rho$ accounting for the constraint $\mathbf{b}_\rho = \mathbf{D}_{\rho:}$ for multiple iterations. The latter updating iterates are denoted as $\hat{\mathbf{D}}_{\rho:}^{\tau,k}$, $\mathbf{b}_\rho^{\tau,k}$ and $\mathbf{p}_\rho^{\tau,k}$ respectively, while $k = 1, 2, \ldots, K$ corresponds to the ADMM iteration index. To obtain the updating recursions the augmented Lagrangian function of (9) is first formed

$$\begin{aligned}L^{\tau}(\mathbf{D}_{\rho:}, \mathbf{b}_\rho, \mathbf{p}_\rho) &= N^{-1}\|\mathbf{T}_{1,\rho}^{E,\tau} - \mathbf{D}_{\rho:}\mathbf{X}\|_2^2 + \lambda_{D,\rho}\|\mathbf{D}_{\rho:}\|_1 \\ &+ \phi_D\|\mathbf{b}_\rho\|_2 + \varepsilon\|\mathbf{T}_{3,\rho} - \mathbf{D}_{\rho:}\mathbf{T}_2^{D,\tau}\|_2^2 + (\mathbf{D}_{\rho:} - \mathbf{b}_\rho)\mathbf{p}_\rho \\ &+ \frac{c}{2}\|\mathbf{D}_{\rho:} - \mathbf{b}_\rho\|_2^2 \end{aligned} \quad (10)$$

where c is a positive penalty coefficient making (9) strictly convex. In every iteration, vector $\mathbf{D}_{\rho:}$ (or $\mathbf{b}_\rho$) is updated by minimizing the

augmented Lagrangian function in (10) while fixing the remaining vectors $\mathbf{b}_\rho$ and $\mathbf{p}_\rho$ (or $\mathbf{D}_{\rho:}$ and $\mathbf{p}_\rho$) to their most recent updates. During ADMM iteration $k+1$ and coordinate iteration $\tau+1$ the following three updating steps take place:

$$\hat{\mathbf{D}}_{\rho:}^{\tau,k+1} = \arg\min_{\mathbf{D}_{\rho:}} N^{-1}\|\mathbf{T}_{1,\rho}^{E,\tau} - \mathbf{D}_{\rho:}\mathbf{X}\|_2^2 + \lambda_{D,\rho}\|\mathbf{D}_{\rho:}\|_1 + \varepsilon\|\mathbf{T}_{3,\rho} - \mathbf{D}_{\rho:}\mathbf{T}_2^{D,\tau}\|_2^2 + \mathbf{D}_{\rho:}\mathbf{p}_\rho^{\tau,k} + \frac{c}{2}\|\mathbf{D}_{\rho:} - \mathbf{b}_\rho^{\tau,k}\|_2^2 \quad (11)$$

then $\mathbf{b}_\rho$ is updated while using the updates $\hat{\mathbf{D}}_{\rho:}^{\tau,k+1}$ and $\mathbf{p}_\rho^{\tau,k}$ as

$$\mathbf{b}_\rho^{\tau,k+1} = \text{argmin}_{\mathbf{b}_\rho} \phi_D\|\mathbf{b}_\rho\|_2 - \mathbf{b}_\rho\mathbf{p}_{\rho,k}^\tau + 0.5c\|\hat{\mathbf{D}}_{\rho:}^{\tau,k+1} - \mathbf{b}_\rho\|_2^2 \quad (12)$$

finally the multiplier is updated using the most recent updates $\hat{\mathbf{D}}_{\rho:}^{\tau,k+1}$ and $\mathbf{b}_\rho^{\tau,k+1}$ as

$$\mathbf{p}_\rho^{\tau,k+1} = \mathbf{p}_\rho^{\tau,k} + c(\hat{\mathbf{D}}_{\rho:}^{\tau,k+1} - \mathbf{b}_\rho^{\tau,k+1})^T. \quad (13)$$

Note that after finite K ADMM iterations $\hat{\mathbf{D}}_{\rho:}^{\tau,k+1}$ will correspond to an estimate for $\hat{\mathbf{D}}_{\rho:}^{\tau+1}$ in (9), while if $K \to \infty$ then ADMM iterates satisfy $\lim_{k\to\infty} \hat{\mathbf{D}}_{\rho:}^{\tau,k+1} = \hat{\mathbf{D}}_{\rho:}^{\tau+1}$ (convergence result in [4]), where $\hat{\mathbf{D}}_{\rho:}^{\tau+1}$ the minimizer in (8). The minimization in (11) can be split into pf subtasks each of which subtasks involves minimizing (11) wrt one entry of $\mathbf{D}_{\rho:}$, namely $\mathbf{D}(\rho,\beta)$, while letting the rest of the entries fixed. After some algebraic manipulations, we obtain

$$\hat{\mathbf{D}}^{\tau,k+1}(\rho,\beta) = \arg\min_d \|\boldsymbol{\zeta}_{\rho,\beta}^{\tau,k} - d\mathbf{h}_{\rho,\beta}^\tau\|_2^2 + \frac{c}{2}[d - \mathbf{b}_\rho^{\tau,k}(\beta)]^2 + d\mathbf{p}_\rho^{\tau,k}(\beta) + \lambda_{D,\rho}|d|, \text{ for } \beta = 1,2,...,pf,\ \rho = 1,\ldots,q \quad (14)$$

with $\boldsymbol{\zeta}_{\rho,\beta}^{\tau,k} := [\boldsymbol{\zeta}_{\rho,\beta}^{1,\tau,k}, \boldsymbol{\zeta}_{\rho,\beta}^{2,\tau,k}]^T$ and $\mathbf{h}_{\rho,\beta}^\tau := [N^{-0.5}\mathbf{X}_{\beta:}, \mathbf{T}_{2,\beta:}^{D,\tau}]^T$, while

$$\boldsymbol{\zeta}_{\rho,\beta}^{1,\tau,k} := N^{-\frac{1}{2}}[\mathbf{T}_{1,\rho}^{E,\tau} - \sum_{\ell=1}^{\beta}\hat{\mathbf{D}}^{\tau,k+1}(\rho,\ell)\mathbf{X}_{\ell:} - \sum_{\ell=\beta+1}^{pf}\hat{\mathbf{D}}^{\tau,k}(\rho,\ell)\mathbf{X}_{\ell:}]$$

$$\boldsymbol{\zeta}_{\rho,\beta}^{2,\tau,k} := \varepsilon^{0.5}[\mathbf{T}_{3,\rho} - \sum_{\ell=1}^{\beta}\hat{\mathbf{D}}^{\tau,k+1}(\rho,\ell)\mathbf{T}_{2,\ell:}^{D,\tau} - \sum_{\ell=\beta+1}^{pf}\hat{\mathbf{D}}^{\tau,k}(\rho,\ell)\mathbf{T}_{2,\ell:}^{D,\tau}]$$

The solution of the minimization problem in (14) can be expressed in closed form (e.g., see [15, 17]) as

$$\hat{\mathbf{D}}^{\tau,k+1}(\rho,\beta) = \text{sgn}((\boldsymbol{\zeta}_{\rho,\beta}^{\tau,k})^T\mathbf{h}_{\rho,\beta}^\tau + \frac{c}{2}(\mathbf{b}_\rho^{\tau,k}(\beta) - \frac{\mathbf{p}_\rho^{\tau,k}(\beta)}{c})) \times \left(\left|\frac{(\boldsymbol{\zeta}_{\rho,\beta}^{\tau,k})^T\mathbf{h}_{\rho,\beta}^\tau + \frac{c}{2}(\mathbf{b}_\rho^{\tau,k}(\beta) - \frac{\mathbf{p}_\rho^{\tau,k}(\beta)}{c})}{\|\mathbf{h}_{\rho,\beta}^\tau\|_2^2 + \frac{c}{2}}\right| - \frac{\lambda_{D,\rho}}{2\|\mathbf{h}_{\rho,\beta}^\tau\|_2^2 + c}\right)_+, \quad (15)$$

where $(x)_+ = \max(x,0)$.

Using the results in [19] the minimizer of (12) is given as

$$\mathbf{b}_\rho^{\tau,k} = c^{-1}\mathcal{S}_v((\mathbf{p}_\rho^{\tau,k})^T + c\mathbf{D}_{\rho:}^{\tau,k}, \phi_D) \quad (16)$$

where $\mathcal{S}_v(\mathbf{v},\phi) = [1 - \frac{\phi}{\|\mathbf{v}\|_2}]_+\mathbf{v}$.

A similar process can be followed starting from (6), to obtain updating recursions for $\hat{\mathbf{E}}^{\tau,k}(\rho,\beta)$. The corresponding quantities involved in forming $\hat{\mathbf{E}}^{\tau,k}(\rho,\beta)$ will be denoted as $\mathbf{b}_\rho^{E,\tau,k}$ and $\mathbf{p}_\rho^{E,\tau,k}$ that are the equivalents for $\mathbf{b}_\rho^{\tau,k}$ and $\mathbf{p}_\rho^{\tau,k}$.The updating process in the regularized (R-) CCA algorithm is tabulated as Algorithm 1.

Algorithm 1: RCCA

Initialize $\hat{\mathbf{D}}^0$, $\hat{\mathbf{E}}^0$ using the standard CCA solution [6, Chp. 10]. Initialize $\{\mathbf{b}_\rho^0 = \mathbf{b}_\rho^{E,0} = \mathbf{0}\}_{\rho=1}^q$ and $\{\mathbf{p}_\rho^0 = \mathbf{p}_\rho^{E,0} = \mathbf{0}\}_{\rho=1}^q$.
for $\tau = 1,2,\ldots,$ **do**
for $\rho = 1,2,\ldots,q$ **do**
for $k = 1,2,\ldots,K$ **do**
Update $\hat{\mathbf{D}}^{\tau,k}(\rho,\beta)$ via (15) for $\beta = 1,\ldots,pf$.
Update $\mathbf{b}_\rho^{\tau,k}$ via (16).
Update $\mathbf{p}_\rho^{\tau,k}$ via (13).
end for
end for
Similarly update $\{\hat{\mathbf{E}}^{\tau,k}(\rho,\beta)\}_{\beta=1}^{pf}$, $\{\mathbf{b}_\rho^{E,\tau,k}, \mathbf{p}_\rho^{E,\tau,k}\}$.
$\hat{\mathbf{D}}_{\rho:}^{\tau+1} = \hat{\mathbf{D}}_{\rho:}^{\tau,K}$ and $\hat{\mathbf{E}}_{\rho:}^{\tau+1} = \hat{\mathbf{E}}_{\rho:}^{\tau,K}$ for $\rho = 1,\ldots,q$.
If $\|\hat{\mathbf{D}}^{\tau+1} - \hat{\mathbf{D}}^\tau\|_F + \|\hat{\mathbf{E}}^{\tau+1} - \hat{\mathbf{E}}^\tau\|_F < \epsilon$ for a prescribed tolerance ϵ, **then** break.
end for

3.2. Selecting the Regularization Coefficients

The RCCA scheme utilizes two different types of regularization coefficients. The coefficients ϕ_D and ϕ_E control the number of zero rows in $\mathbf{E}$ and $\mathbf{D}$, while the coefficients $\lambda_{E,\rho}$ and $\lambda_{D,\rho}$ control the sparsity patterns inside the ρth row of matrix $\mathbf{E}$ and $\mathbf{D}$, respectively. Proper selection of these coefficients can ensure that RCCA will recover efficiently the correct number of different sensor groups acquiring information about the field sources which translates to keeping M out of the q rows in $\mathbf{E}$ and $\mathbf{D}$. A simple yet effective way to select the ϕ's and λ's is proposed here. To this end, we consider the case $\phi_D = \phi_E = \phi$ since both $\mathbf{D}$ and $\mathbf{E}$ should have the same number of nonzero rows corresponding to the number of sources M. Similarly, $\lambda_{E,\rho} = \lambda_{D,\rho} = \lambda_\rho$ since the support of $\mathbf{D}_{\rho:}$ and $\mathbf{E}_{\rho:}$ should coincide as explained in Sec. 2. Here it is assumed that $M \geq 1, q > M$ and there are at least two sensors sensing just noise.

We start by fixing λ's to a value λ_0, ϕ is initialized to a large value ϕ_0 which results all-zero matrices $\mathbf{E}$ and $\mathbf{D}$ when applying RCCA. By gradually decreasing ϕ by a small step size $\Delta\phi$, i.e., $\phi^n = \phi_0 - n\Delta\phi$, more non-zero rows will appear in $\mathbf{E}$ and $\mathbf{D}$. The ϕ is selected as the first ϕ^n resulting estimates $\hat{\mathbf{E}}$ and $\hat{\mathbf{D}}$ in RCCA whose rows contains at least two relatively small (close to zero) entries (in accordance with the assumption that at least two sensors sense noise), and at the same time, the following $\phi^{n+1}, \phi^{n+2}, \ldots$ result estimates $\hat{\mathbf{E}}$ and $\hat{\mathbf{D}}$ in RCCA that contain one or multiple rows which have less than two entries with close-to-zero magnitude.

After selecting ϕ, the sparsity-controlling coefficients λ_ρ are chosen. Let $\{\lambda_\rho^{\max}\}_{\rho=1}^q$ denote the smallest value of the sparsity controlling coefficients that result the ρth row of $\hat{\mathbf{D}}$ and $\hat{\mathbf{E}}$ obtained from RCCA to be zero. In the first step $\lambda_\rho^{\max}$ is estimated via estimates $\hat{\lambda}_\rho^m$. After randomly initializing $\hat{\lambda}_\rho^m$ and applying RCCA the support sets of the estimates $\hat{\mathbf{D}}_{\rho:}$ and $\hat{\mathbf{E}}_{\rho:}$ are checked. If the support sets are nonempty then $\hat{\lambda}_\rho^m$ is increased by a factor of $\omega_2 > 1$. The estimates $\hat{\lambda}_\rho^m$ will keep increasing until RCCA gives an empty support for $\hat{\mathbf{D}}_{\rho:}$ and $\hat{\mathbf{E}}_{\rho:}$ in which case it is certain that $\lambda_\rho^{\max}$ has been reached or exceeded. If the support sets $\hat{\mathbf{D}}_{\rho:}$ and/or $\hat{\mathbf{E}}_{\rho:}$ are empty then $\hat{\lambda}_\rho^m$ has exceeded $\lambda\rho^{\max}$ in which case $\hat{\lambda}_\rho^m$ is decreased by a factor of $\omega_1 \in [1-\epsilon, 1)$. The estimates $\hat{\lambda}_\rho^m$ will be decreased until RCCA gives a nonempty support for $\hat{\mathbf{D}}_{\rho:}$ and $\hat{\mathbf{E}}_{\rho:}$.

Given the estimate $\hat{\lambda}_\rho^m$ from earlier, the second step is recovering the indices of columns in $\mathbf{D}$ and $\mathbf{E}$ that are zero, denoted here as $\mathcal{C}$. Note that the index of a zero column indicates a sensor mea-

surement acquiring only sensing noise. The estimate $\hat{\lambda}_\rho^m$ is scaled with factors $\omega_3 < 1$ and $\omega_4 < 1$, where $\omega_4 << \omega_3$. Two different zero column index sets, namely $\mathcal{C}_1$ (using ω_3) and $\mathcal{C}_2$ (using ω_4), are obtained after applying RCCA. Since $\omega_4 << \omega_3$ it is expected that $\mathcal{C}_1 \supseteq \mathcal{C}_2$. The reason for getting two different sets $\mathcal{C}_1$ and $\mathcal{C}_2$ is to identify which columns (noisy sensors) in $\mathbf{D}$ and $\mathbf{E}$ will be zero for both different scalings of $\hat{\lambda}_\rho^m$ using ω_4 and ω_3. This way the columns of $\mathbf{E}, \mathbf{D}$ that match with entries in $\mathbf{x}(t), \mathbf{y}(t)$ that contain information about a source (nonzero columns) can be distinguished from the columns that correspond to entries in $\mathbf{x}(t), \mathbf{y}(t)$ with just sensing noise (zero columns). The last step is to select λ's that result estimates for $\mathbf{D}$ and $\mathbf{E}$ whose zero column index set coincides with $\mathcal{C}$ from the second step. To this end, starting from $\hat{\lambda}_\rho^m$ obtained in step one we gradually decrease their value by a factor $\omega_5 \in [1-\epsilon, 1)$ until the zero column index set of the $\mathbf{D}, \mathbf{E}$ estimates in RCCA coincides with $\mathcal{C}$. The selection scheme is summarized below as Alg. 2.

Algorithm 2: Coefficient selection

Initilization: $\phi = \phi_0$, $\lambda_{D,\rho} = \lambda_{E,\rho} = \lambda_0$ for $\rho = 1, 2, ..., q$
$-\phi-$ selection :
for $n = 1, 2, ...J = \lfloor \frac{\phi_0}{\Delta\phi} \rfloor$
$\phi^n = \phi_0 - n\Delta\phi$
endfor
Select ϕ^n such that $\{\phi^j\}_{j=n+1}^J$ (combined with λ_0) in RCCA gives $\hat{\mathbf{E}}, \hat{\mathbf{D}}$ with rows having less than two close-to-zero/zero entries.
$-\lambda-$selection:
Step 1): Find $\{\lambda_\rho^{\max}\}_{\rho=1}^q$.
Initialize $\{\hat{\lambda}_\rho^m > 0\}_{\rho=1}^q$ randomly.
while$(true)$
Find $\hat{\mathbf{D}}$ via RCCA using $\hat{\lambda}_\rho^m$
If $\hat{\mathbf{D}}_{\rho:} \neq \mathbf{0}$
Update $\hat{\lambda}_\rho^m = \omega_2 \hat{\lambda}_\rho^m$ where $\omega_2 > 1$
else if $\hat{\mathbf{D}}_{\rho:} = \mathbf{0}$
Update $\hat{\lambda}_\rho^m = \omega_1 \hat{\lambda}_\rho^m$, where $\omega_1 < 1$. Find $\check{\mathbf{D}}$ via RCCA with updated $\hat{\lambda}_\rho^m$
If $\check{\mathbf{D}}_{\rho:} = \mathbf{0}$
Update $\hat{\lambda}_\rho^m = \omega_1 \hat{\lambda}_\rho^m$.
else if $\check{\mathbf{D}}_{\rho:} \neq \mathbf{0}$
$\hat{\lambda}_\rho^m = \hat{\lambda}_\rho^m$
Break while
endIf
end If
end while
Step 2): Estimate zero column index set (denoted as $\mathcal{C}$) of $\mathbf{D}$
Find zero column index set $\mathcal{C}_1$ of $\hat{\mathbf{D}}$ via RCCA using $\omega_3 \hat{\lambda}_\rho^m$.
Find set $\mathcal{C}_2$ of $\hat{\mathbf{D}}$ via RCCA using $\omega_4 \hat{\lambda}_\rho^m$. Set $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$.
Step 3): Select $\{\hat{\lambda}_\rho\}_{\rho=1}^q$ to be used
Starting from $\hat{\lambda}_{\rho,0} = \hat{\lambda}_\rho^m$ decrease $\lambda_{\rho,n} = \omega_5 \lambda_{\rho,n-1}$ until RCCA gives $\hat{\mathbf{D}}$ whose zero column index set matches $\mathcal{C}$.

In the numerical results later on, we set $\phi_0 = 5$, $\Delta\phi = 0.05$, $\lambda_0 = 0.1, \omega_1 = 0.75, \omega_2 = 1.5, \omega_3 = 0.1, \omega_4 = 0.01, \omega_5 = 0.95$.

4. NUMERICAL RESULTS

The probability of correctly clustering sensor data in the right number of groups (M here) based on their source content is numerically evaluated here for: i) RCCA; ii) the K-means clustering scheme in [9, 14] that estimates the unknown number of clusters via sequential extraction of anomalous patterns in the data, abbreviated as ik-Means (intelligent K-Means); and iii) the sparse CCA approach in [20] abbreviated here as PMD.

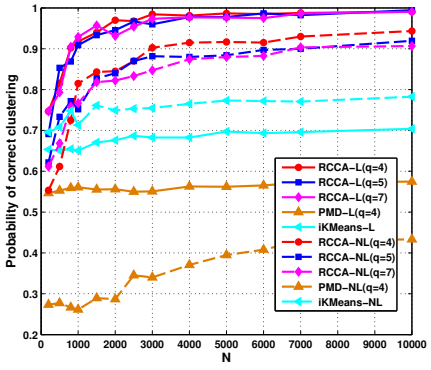

Fig. 1. Probability of correctly clustering sensors for the RCCA vs. number of data N in a linear (-L) and nonlinear (-NL) setting.

We consider a setting with $M = 2$ sources and $p = 15$ sensors. The corresponding mappings $h_{m,j}$ in (1) are summarized in the following matrices $\mathbf{G}_l = [g_{m,j}(\cdot)]$ and $\mathbf{G}_{nl} = [g_{m,j}(\cdot)]$ for a linear and nonlinear model in (1) respectively, while the (m, j)th entry of $\mathbf{G}_l$ (or $\mathbf{G}_{nl}$) corresponds to $g_{m,j}(s_m(t))$. Specifically, $\mathbf{G}_l := \mathbf{G} \odot \mathbf{A}_s$ and $\mathbf{G}_{nl} := \mathbf{G} \odot \mathbf{B}_s$, where $\odot$ denotes entry-wise product and

$$\mathbf{A}_s := \begin{bmatrix} s_1(t)\mathbf{1}_{1\times 5} & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times 5} \\ \mathbf{0}_{1\times 5} & s_2(t)\mathbf{1}_{1\times 5} & \mathbf{0}_{1\times 5} \end{bmatrix}$$

$$\mathbf{B}_s := \begin{bmatrix} s_1(t) & s_1^{1.1}(t) & s_1^{1.2}(t) & s_1^{1.3}(t) & s_1^{1.4}(t) & \mathbf{0}_{1\times 5} & \mathbf{0}_{1\times 5} \\ \mathbf{0}_{1\times 5} & s_2(t) & s_2^{1.1}(t) & s_2^{1.2}(t) & s_2^{1.3}(t) & s_2^{1.4}(t) & \mathbf{0}_{1\times 5} \end{bmatrix}$$

while $\mathbf{1}_{1\times 5}, \mathbf{0}_{1\times 5}$ denote the 1×5 all-ones and all-zeros vectors respectively and the entries of $\mathbf{G}$ are normally distributed. The matrices $\mathbf{G}_l$ and $\mathbf{G}_{nl}$ represent a setting where source $s_1(t)$ is observed by sensors $\{1, 2, 3, 4, 5\}$, source $s_2(t)$ is observed by sensor $\{6, 7, 8, 9, 10\}$ and the rest $\{11, 12, 13, 14, 15\}$ acquire just noise.

Fig. 1 depicts the probability of correctly clustering the sensor measurements in the right number of groups versus the number of data samples N acquired across each sensor. RCCA is applied for $K = 10$ ADMM iterations and selecting $q = 4, 5$ or 7, the parameters in PMD were selected on a trial and error basis to get the best observed performance while there are no parameters to set on iK-Means. Clearly, it can be seen that the probability achieved by RCCA is higher than that of PMD and iK-Means for both the linear (-L) and nonlinear (-NL) settings. Note also that the performance of RCCA is not really affected by how q, the number of nonzero rows of $\mathbf{E}, \mathbf{D}$, is selected as long as is larger than $M = 2$. This advocates the potential of RCCA to correctly cluster sensor data based on their information content even when M is unknown.

5. CONCLUDING REMARKS

The standard CCA framework was augmented with norm-one and norm-two regularization terms that facilitate estimation of the number of field sources and clustering of the sensor data based on their information content. ADMM and coordinate descent techniques were employed to tackle the associated minimization problem, while numerical tests demonstrate the advantages of correctly recovering different sensor groups over existing alternatives.

6. REFERENCES

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. " Convex optimization with sparsity-inducing norms," *Optimization for Machine Learning*, MIT press, 2011.

[2] A. Banerjee, I. 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] D. P. Bertsekas, *Nonlinear Programming.* 2nd Edition, Athena Scientific, Massachussets, 1999.

[4] D. P. Bertsekas and J. N. Tsitsiklis, "Parallel and distributed computation: numerical methods," Prentice-Hall, Inc., Upper Saddle River, NJ, 1989.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning,* vol. 3, no. 1, 2011.

[6] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Expanded Edition, Holden Day, 1981.

[7] X. Chen, H. Liu, and J. G. Carbonell, "Structured sparse canonical correlation analysis," *Proc. of Intl. Conf. on Artificial Intelligence and Stats. (AISTATS),* La Palma, Canary Islands, pp. 199–207, 2012.

[8] J. Chen and I. D. Schizas, "Distributed sparse canonical correlation analysis in clustering sensor data," *Proc. of the Asilomar Conf. on Signals, Systems and Comp.,* Pacific Grove, CA , Nov. 2013

[9] M. M. Chiang, B. Mirkin, "Intelligent choice of the number of clusters in k-Means clustering: an experimental study with different cluster spreads," *Journal of Classification,* vol. 27, no. 3, pp. 3–40, 2010.

[10] D. R. Hardoon and J. Taylor, "The double-barrelled Lasso," in *Learning from Multiple Sources Workshop, Advances on Neural Information Processing Systems*, Vancouver, Canada, 2008.

[11] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321-377, 1936.

[12] R. A. Johnson, and D. W. Wichern, *Applied Multivariate Analysis,* 4th Edition,. New Jersey: Prentice Hall, Englewood Cliffs, 1998.

[13] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory,* vol. 28, no. 2, pp. 129–137, Mar. 1982.

[14] B. Mirkin, *Clustering for Data Mining: A Data Recovery Approach*, Boca Raton, FL: Chapman and Hall/CRC, 2005.

[15] 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.

[16] P. Sprechmann , I. Ramirez , G. Sapiro , Y. C. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, p. 4183–4198, September 2011

[17] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[18] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Opt. Theory and Applications*, vol. 109, no. 3, pp. 475–494, Jun. 2001.

[19] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[20] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal Ccomponents and canonical correlation Analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[21] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.,* vol. 16, no. 3, pp. 645–678, May 2005.

[22] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, 2006.