REVISITING ROBUSTNESS OF THE UNION-OF-SUBSPACES MODEL FOR DATA-ADAPTIVE LEARNING OF NONLINEAR SIGNAL MODELS

Tong Wu and Waheed U. Bajwa

Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey
Emails: {tong.wu.ee, waheed.bajwa}@rutgers.edu

ABSTRACT

This paper revisits the problem of data-adaptive learning of geometric signal structures based on the Union-of-Subspaces (UoS) model. In contrast to prior work, it motivates and investigates an extension of the classical UoS model, termed the Metric-Constrained Union-of-Subspaces (MC-UoS) model. In this regard, it puts forth two iterative methods for data-adaptive learning of an MC-UoS in the presence of complete and missing data. The proposed methods outperform existing approaches to learning a UoS in numerical experiments involving both synthetic and real data, which demonstrates effectiveness of both an MC-UoS model and the proposed methods.

Index Terms— Nonlinear signal models, union of subspaces

1. INTRODUCTION

Much of machine learning, signal processing, and statistics literature is based on the premise that high-dimensional signals lie on or near low-dimensional structures embedded in higher-dimensional ambient spaces. Knowledge of the low-dimensional structure underlying a high-dimensional signal set helps reveal the information one is looking for, thereby greatly simplifying the sampling, processing, computational, and storage requirements [1]. But this knowledge is seldom, if ever, available to practitioners. Because of this, a significant fraction of the literature is devoted to study of plausible signal models and either characterization or learning of low-dimensional structures adhering to a prescribed signal model [2–5].

Two insights have emerged from decades of research on signal models. First, nonlinear signal models provide better abstractions of the real world signals than linear signal models [6]. Second, data-adaptive learning of signal models from training examples results in better performance than off-the-shelf characterization of low-dimensional structures describing the signal models of interest [7]. In recent years, a nonlinear signal model that has in particular helped advance the state-of-the-art in many applications is the *union-of-subspaces* (UoS) model [1, 8–10]. The UoS model, which is a generalization of the vanilla sparsity model, states that signals belong to a union of low-dimensional subspaces. While data-adaptive learning of subspaces underlying a UoS dates back more than a decade [11, 12], interest in this area has exploded in recent years because of our move toward a data-driven society. A necessarily incomplete list of works in this regard includes [5, 10, 13–21].

Our Contributions: The canonical UoS model does not impose any structure on the collection of subspaces underlying signals of interest. Intuitively, however, one expects that if a UoS describes similar signals (e.g., frontal face images of a single person) then the underlying collection of subspaces should be "related" to each other. In addition, it stands to reason that methods for learning UoS's would be better able to thwart errors caused by noise, outliers, missing data, etc., if they could explicitly account for any such relationship between subspaces describing a UoS. In order to capture this intuition of "related subspaces," we put forth a novel extension of the traditional UoS model, termed *metric-constrained union-of-subspaces* (MC-UoS) model. Heuristically, the MC-UoS model states that signals not only belong to a union of low-dimensional subspaces, but the individual subspaces are also close to each other with respect to a metric defined on the Grassmann manifold. The main challenge in this regard is formulation of methods for data-adaptive learning of subspaces underlying an MC-UoS. We address this challenge by presenting two iterative algorithms, termed MiCUSaL and rMiCUSaL, for data-adaptive learning of an MC-UoS using complete data and missing data, respectively. In order to demonstrate effectiveness of both the MC-UoS model and the proposed algorithms, we carry out extensive numerical experiments using synthetic and real data. Results of these experiments demonstrate that both MiCUSaL and rMiCUSaL outperform existing approaches to data-adaptive learning of a UoS in terms of robustness to noise, outliers, and missing data.

Notation: We use lower-case and upper-case letters for vectors and matrices, respectively. In addition, $(\cdot)^T$ and $\operatorname{tr}(\cdot)$ denote transpose and trace operations, respectively, while $\|\cdot\|_F$ and $\|\cdot\|_p$ denote Frobenius norm and ℓ_p norm of matrices and vectors, respectively. Given a set Ω, A_Ω (resp., v_Ω) denotes the submatrix of A (resp., subvector of v) corresponding to the rows of A (resp., entries of v) indexed by Ω. Finally, a_j^i denotes the (i, j)-th element of A and $v(i)$ denotes the i-th entry of a vector v.

2. PROBLEM FORMULATION

In this section, we mathematically formulate the problem studied in this paper. Recall the canonical UoS model [8], which asserts that signals of interest lie in a union of K low-dimensional subspaces: $\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{S}_k$, where $\mathcal{S}_k$ is a subspace of $\mathbb{R}^n$. In here, we make the simplified assumption that $\forall k, \dim(\mathcal{S}_k) = s \ll n$. The basic premise behind this paper is that the $\mathcal{S}_k$'s underlying similar signals likely do not correspond to arbitrary points on the Grassmann manifold $G_{n,s}$, defined as the collection of all s-dimensional subspaces of $\mathbb{R}^n$. In order to formalize this idea, we put forth the definition of a metric-constrained union-of-subspaces (MC-UoS).

Definition 1. ***(Metric-Constrained Union-of-Subspaces.)*** *A UoS $\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{S}_k$ is said to be constrained with respect to a metric $d_u : G_{n,s} \times G_{n,s} \to [0, \infty)$ if $\max_{j,k:j \neq k} d_u(\mathcal{S}_j, \mathcal{S}_k) \leq \epsilon$ for some positive constant ϵ.*

In words, the MC-UoS model asserts that signals of interest lie in a union of K low-dimensional subspaces *that are also "close" to each other*. Our goal in this paper is to learn an MC-UoS $\mathcal{M}$ in a data-adaptive manner. To this end, we assume access to a total

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

of N training samples, $Y = \{y_i \in \mathbb{R}^n\}_{i=1}^N$, that correspond to (possibly noisy) samples drawn from $\mathcal{M}$. The problem of learning $\mathcal{M}$ in this setting can be posed as learning K-subspaces, $\{\mathcal{S}_k\}_{k=1}^K$, such that ($i$) each y_i is well approximated by one of the $\mathcal{S}_k$'s, and (ii) the $\mathcal{S}_k$'s are close to each other. The metric we use in here to measure closeness of subspaces on $G_{n,s}$ is based on the Hausdorff distance, defined initially in [22] and proven to be a metric in [23]. Specifically, if $D_k \in \mathbb{R}^{n \times s}$ denotes an orthonormal basis of $\mathcal{S}_k$, then

$$d_u(\mathcal{S}_j, \mathcal{S}_k) = \sqrt{s - \mathrm{tr}(D_j^T D_k D_k^T D_j)} = \|D_k - P_{\mathcal{S}_j} D_k\|_F, \tag{1}$$

where $P_{\mathcal{S}_j}$ denotes the projection operator $P_{\mathcal{S}_j} = D_j D_j^T$. Note that being a distance on $G_{n,s}$, $d_u(\cdot, \cdot)$ in (1) is invariant under the choice of orthonormal bases of the two subspaces. While there exist other measures of subspace distances in the literature (see, e.g., [24]), we prefer (1) because of its ease of computation.

The preceding discussion helps us pose the problem of learning an MC-UoS $\mathcal{M}$ in terms of the following optimization problem:

$$\{\mathcal{S}_k\} = \operatorname*{arg\,min}_{\{\mathcal{S}_k\} \subset G_{n,s}} \sum_{\substack{j,k=1 \\ j \neq k}}^{K} d_u^2(\mathcal{S}_j, \mathcal{S}_k) + \lambda \sum_{i=1}^{N} \|y_i - P_{\mathcal{S}_{\hat{\imath}}} y_i\|_2^2, \tag{2}$$

where $\hat{\imath} = \arg\min_k \|y_i - P_{\mathcal{S}_k} y_i\|_2^2$ with $P_{\mathcal{S}_k} y_i$ denoting the projection of y_i onto $\mathcal{S}_k$. In words, the first term in (2) forces the learned subspaces to be close to each other, the second term in (2) forces the learned subspaces to provide reasonable approximations of the training data, and the parameter λ quantifies the desired trade-off between subspace closeness and data approximation. Our objective then is to provide fast computational methods for solving (2) for the cases of "complete" training data and "missing" training data.

3. MC-UOS LEARNING FROM COMPLETE DATA

In this section, we propose an algorithm for solving (2) for the case of complete training data. We begin by simplifying the expression in (2). To this end, we first define a $K \times N$ indicator matrix W as

$$W \overset{def}{=} \big[w_i^k \in \{0, 1\} : \forall i = 1, \dots, N, \sum_{k=1}^{K} w_i^k = 1\big]. \tag{3}$$

In words, W specifies memberships of the y_i's in different subspaces and $w_i^k = 1$ if and only if y_i belongs to the subspace $\mathcal{S}_k$. Next, notice from elementary manipulations that

$$\|y_i - P_{\mathcal{S}_k} y_i\|_2^2 = \|y_i - D_k D_k^T y_i\|_2^2 = \|y_i\|_2^2 - \|D_k^T y_i\|_2^2,$$

where once again D_k denotes an orthonormal basis of $\mathcal{S}_k$. Therefore, defining $D \overset{def}{=} \begin{bmatrix} D_1 & \dots & D_K \end{bmatrix}$, (2) can be equivalently expressed as $(D, W) = \arg\min_{D,W} F(D, W)$ such that

$$F(D, W) = \sum_{\substack{j,k=1 \\ j \neq k}}^{K} \|D_k - P_{\mathcal{S}_j} D_k\|_F^2 + \lambda \sum_{i=1}^{N} \sum_{k=1}^{K} w_i^k (\|y_i\|_2^2 - \|D_k^T y_i\|_2^2). \tag{4}$$

Instead of minimizing (4) simultaneously over (D, W), which will be computationally difficult, we will resort to minimizing it by alternating between minimizing $F(D, W)$ over W for a fixed D and minimizing $F(D, W)$ over D for a fixed W. In the following, we term minimization of $F(D, W)$ over W for a fixed D as the *subspace assignment* step and minimization of $F(D, W)$ over D for a fixed W as the *subspace update* step. In terms of performance measure, we are in particular interested in a *partial optimal solution*, defined as follows.

```
Algorithm 1 Metric-Constrained UoS Learning (MiCUSaL)
Input: Training data $Y$; problem parameters $K$, $s$, and $\lambda$.
Initialize: Orthonormal bases $D_k, k = 1, \dots, K$.
while stopping rule do
    for $i = 1$ to $N$ (Subspace Assignment) do
        $l_i \leftarrow \arg\max_k \|D_k^T y_i\|_2$, $w_i^{l_i} \leftarrow 1$, $\forall k \neq l_i, w_i^k \leftarrow 0$.
    end for
    for $k = 1$ to $K$ (Subspace Update) do
        $c_k \leftarrow \{1 \leq i \leq N : w_i^k = 1\}$, $Y_k \leftarrow \big[y_i : i \in c_k\big]$.
        $A_k \leftarrow \sum_{j \neq k} D_j D_j^T + \frac{\lambda}{2} Y_k Y_k^T$.
        Eigen decomposition of $A_k : U_k \Sigma_k U_k^T = A_k$.
        $D_k \leftarrow$ columns of $U_k$ corresponding to $s$-largest
                   diagonal elements in $\Sigma_k$.
    end for
end while
Output: Orthonormal bases $D_k, k = 1, \dots, K$.
```

Definition 2 ([25]). *A point (D^*, W^*) is a partial optimal solution of $\arg\min_{D,W} F(D, W)$ if $\forall D, F(D^*, W^*) \leq F(D, W^*)$ and $\forall W, F(D^*, W^*) \leq F(D^*, W)$.*

We refer the reader to [26] for a discussion of the significance of partial optimal solutions in the context of our problem.

In order to obtain a partial optimal solution, we begin with subspace assignment. When D is fixed, subspace assignment simply corresponds to solving $\forall i = 1, \dots, N$,

$$l_i = \operatorname*{arg\,min}_{k=1,\dots,K} \|y_i - P_{\mathcal{S}_k} y_i\|_2^2 = \operatorname*{arg\,max}_{k=1,\dots,K} \|D_k^T y_i\|_2^2 \tag{5}$$

and then setting $w_i^{l_i} = 1$. On the other hand, subspace update, which corresponds to a fixed W, is a more challenging task. In order to address this challenge, we resort to *block-coordinate descent* (BCD) [27], which results in sequential updates of the D_k's in D. Specifically, let $c_k = \{i \in \{1, \dots, N\} : w_i^k = 1\}$ denote the indices of all y_i's that are assigned to the k-th subspace and define $Y_k = \big[y_i : i \in c_k\big]$ to be the corresponding $n \times |c_k|$ matrix. Then, for a fixed W, the K subproblems corresponding to BCD of $\min_D F(D, W)$ can be expressed after some manipulations as

$$D_k = \operatorname*{arg\,min}_{D_k \in V_{n,s}} \sum_{j \neq k} \|D_k - P_{\mathcal{S}_j} D_k\|_F^2 + \frac{\lambda}{2}(\|Y_k\|_F^2 - \|D_k^T Y_k\|_F^2),$$

where $V_{n,s}$ denotes the Stiefel manifold (i.e., collection of all $n \times s$ orthonormal matrices). We can simplify the above expression to reduce each BCD subproblem to the following maximization problem:

$$D_k = \operatorname*{arg\,max}_{D_k \in V_{n,s}} \quad \mathrm{tr}\Big(D_k^T (\sum_{j \neq k} D_j D_j^T + \frac{\lambda}{2} Y_k Y_k^T) D_k\Big). \tag{6}$$

We now define $A_k = \sum_{j \neq k} D_j D_j^T + \frac{\lambda}{2} Y_k Y_k^T$. It then follows from [28] that $\max \mathrm{tr}(D_k^T A_k D_k)$ has a closed-form solution and the D_k that achieves the maximum is simply given by the s eigenvectors of A_k associated with its s-largest eigenvalues. This completes our description of the subspace update step. Combining the subspace assignment and subspace update steps, we can now formally describe our algorithm in Algorithm 1, which we term as *metric-constrained union-of-subspaces learning* (MiCUSaL). The following theorem, stated without proof because of space constraints, describes convergence behavior of MiCUSaL.

Theorem 1. *MiCUSaL is guaranteed to converge. It also returns a partial optimal solution if* $\forall k, \arg\max_{D_k} \operatorname{tr}(D_k^T A_k D_k)$ *during the subspace update step has a unique solution.*

We conclude by pointing out that a necessary and sufficient condition to have a unique solution to $\arg\max_{D_k} \operatorname{tr}(D_k^T A_k D_k)$ is to have distinct s-th and $(s+1)$-th largest eigenvalues of A_k.

4. MC-UOS LEARNING FROM MISSING DATA

In this section, we study MC-UoS learning for the case of training data with missing entries. The setup here corresponds to observing each y_i at locations $\Omega_i \subset \{1, \ldots, n\}$ with $|\Omega_i| \geq s$, denoted by $y_{\Omega_i} \in \mathbb{R}^{|\Omega_i|}$. Since we do not have access to the complete y_i's, the quantities $\|y_i - P_{\mathcal{S}_k} y_i\|_2^2$ in (2) cannot be computed directly. Instead, we leverage the results in [29] and replace $\|y_i - P_{\mathcal{S}_k} y_i\|_2^2$ by

$$\|y_{\Omega_i} - P_{\mathcal{S}_{k\Omega_i}} y_{\Omega_i}\|_2^2 = y_{\Omega_i}^T \Big(I - D_{k\Omega_i} (D_{k\Omega_i}^T D_{k\Omega_i})^{-1} D_{k\Omega_i}^T \Big) y_{\Omega_i},$$

where $P_{\mathcal{S}_{k\Omega_i}} \overset{def}{=} D_{k\Omega_i} (D_{k\Omega_i}^T D_{k\Omega_i})^{-1} D_{k\Omega_i}^T$. In this case, we can reformulate (2) as $(D, W) = \arg\min_{D,W} G(D, W)$, where

$$G(D,W) = \sum_{\substack{j,k=1 \\ j \neq k}}^{K} \|D_k - P_{\mathcal{S}_j} D_k\|_F^2 + \lambda \sum_{i=1}^{N} \sum_{k=1}^{K} w_i^k \|y_{\Omega_i} - P_{\mathcal{S}_{k\Omega_i}} y_{\Omega_i}\|_2^2. \quad (7)$$

In order to solve this problem, we once again make use of alternating minimization comprising subspace assignment and subspace update steps. When D is fixed, subspace assignment in here corresponds to solving $\forall i, l_i = \arg\min_{k=1,\ldots,K} \|y_{\Omega_i} - P_{\mathcal{S}_{k\Omega_i}} y_{\Omega_i}\|_2^2$. When W is fixed, we carry out subspace update using BCD again, in which case $\arg\min_D G(D, W)$ can be shown to be comprised of the following K subproblems: $D_k = \arg\min_{D_k \in V_{n,s}} g(D_k)$ with

$$g(D_k) \overset{def}{=} -\operatorname{tr}(D_k^T A_k D_k) + \frac{\lambda}{2} \sum_{i \in c_k} \|y_{\Omega_i} - P_{\mathcal{S}_{k\Omega_i}} y_{\Omega_i}\|_2^2. \quad (8)$$

Here, c_k is as defined in Section 3 and $A_k = \sum_{j \neq k} D_j D_j^T$. It can be shown that $g(D_k)$ is invariant to the choice of the orthonormal basis of $\mathcal{S}_k$. Hence $\min_{D_k \in V_{n,s}} g(D_k)$ is an optimization problem on the Grassmann manifold [30]. The cost function $g(D_k)$ consists of $1 + |c_k|$ terms. In order to minimize it, we employ incremental gradient descent procedure [31] and operate on a single component in each step. Inspired from algorithms in [30], we first compute the gradient of one term, and move along a short geodesic curve in the gradient direction. To be specific, the gradient of $g_1(D_k) \overset{def}{=} -\operatorname{tr}(D_k^T A_k D_k)$ is $\nabla g_1 = (I_n - D_k D_k^T) \frac{dg_1}{dD_k} = -2(I_n - D_k D_k^T) A_k D_k$, where I_n denotes the $n \times n$ identity matrix. The geodesic equation for a curve $D_k(\eta)$ in the direction $-\nabla g_1$ with a step length η is [30]

$$D_k(\eta) = D_k V_k \cos(\Sigma_k \eta) V_k^T + U_k \sin(\Sigma_k \eta) V_k^T, \quad (9)$$

Algorithm 2 Robust MC-UoS Learning (rMiCUSaL)

Input: Training data $\{y_{\Omega_i}\}_{i=1}^N$; parameters K, s, λ and η.
Initialize: Orthonormal bases $D_k, k = 1, \ldots, K$.

```
while stopping rule do
  for $i = 1$ to $N$ (Subspace Assignment) do
    $l_i \leftarrow \arg\min_k \|y_{\Omega_i} - P_{\mathcal{S}_{k\Omega_i}} y_{\Omega_i}\|_2^2$.
    $w_i^{l_i} \leftarrow 1, \forall k \neq l_i, w_i^k \leftarrow 0$.
  end for
  for $k = 1$ to $K$ (Subspace Update) do
    $c_k \leftarrow \{1 \leq i \leq N : w_i^k = 1\}$.
    while stopping rule do
      $A_k \leftarrow \sum_{j \neq k} D_j D_j^T, \Delta_k \leftarrow 2(I_n - D_k D_k^T) A_k D_k$.
      $D_k \leftarrow D_k V_k \cos(\Sigma_k \eta) V_k^T + U_k \sin(\Sigma_k \eta) V_k^T$
        where $U_k \Sigma_k V_k^T$ is the compact SVD of $\Delta_k$.
      for $p = 1$ to $|c_k|$ do
        $\theta \leftarrow (D_{k\Omega_{c_k(p)}}^T D_{k\Omega_{c_k(p)}})^{-1} D_{k\Omega_{c_k(p)}}^T y_{\Omega_{c_k(p)}}$.
        $q \leftarrow D_k \theta, \quad r \leftarrow y_{\Omega_{c_k(p)}} - q_{\Omega_{c_k(p)}}$.
        $\hat{r} \leftarrow \mathbf{0}, \quad \hat{r}_{\Omega_{c_k(p)}} \leftarrow r$.
        $D_k \leftarrow D_k + \Big( (\cos(\mu\lambda\eta) - 1) \frac{q}{\|q\|_2} +$
          $\sin(\mu\lambda\eta) \frac{\hat{r}}{\|\hat{r}\|_2} \Big) \frac{\theta^T}{\|\theta\|_2}$ where $\mu = \|\hat{r}\|_2 \|q\|_2$.
      end for
    end while
  end for
end while
```

Output: Orthonormal bases $D_k, k = 1, \ldots, K$.

where $U_k \Sigma_k V_k^T$ is the SVD decomposition of $-\nabla g_1$. Note that (8) in here differs from [32, Eqn (2)] only in terms of one additional term g_1 and a scaling factor $\frac{\lambda}{2}$ of all other terms in $g(D_k)$. It therefore follows that the optimization of other terms in $g(D_k)$ can be performed as in GROUSE [32] *but* with a constant step size $\frac{\lambda\eta}{2}$. We conclude this section by presenting our algorithm in Algorithm 2, termed *Robust MC-UoS Learning* (rMiCUSaL).

5. SIMULATION RESULTS AND DISCUSSION

In this section, we evaluate the performance of our proposed methods on both synthetic and real data. In the complete data setting, we compare the performance of MiCUSaL with Block-Sparse Dictionary Design (SAC+BK-SVD) [10], K-subspace clustering (K-sub) [33], Sparse Subspace Clustering (SSC) [15] and Robust Subspace Clustering (RSC) [17]. For the case of missing training data, we compare the results of rMiCUSaL with K-Subspaces with Missing Data (k-GROUSE) [21] and SSC [15]. In order to study the robustness of our methods, we assume every training and test sample y is noisy in the sense that $y = x + v$ where x belongs to one of the $\mathcal{S}_k$'s (also $\|x\|_2^2 = 1$) and v is additive white Gaussian noise with variance σ^2. We use X and X^{te} to denote "clean" training and test signals respectively, while the set of noisy test samples is denoted by Y^{te}. In our experiments, we add white Gaussian noise with different expected noise power ($E[\|v\|_2^2] = n\sigma^2$) ranging from 0.1 to 0.5 to the "clean" training and test samples. For the missing data experiments, we create training (but not test) data with different percentages of missing values ranging from 20% to 50% for every fixed noise power. Finally, we choose $\lambda = 1$ for all experiments.

Experiments with Synthetic Data: We first consider an experiment on synthetic data with parameters $K = 4$, $s = 15$, $n = 100$ and $N = 400$. We define the ground-truth $\mathcal{S}_k$'s by their orthonormal

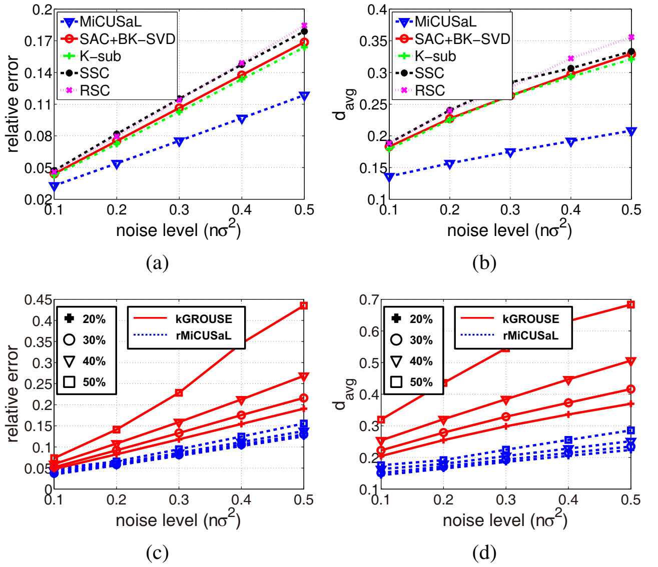

Fig. 1. Comparison of MC-UoS learning performance on synthetic data. (a) and (b) show relative errors and d_{avg}'s for complete data. (c) and (d) show relative errors and d_{avg}'s for missing data case.

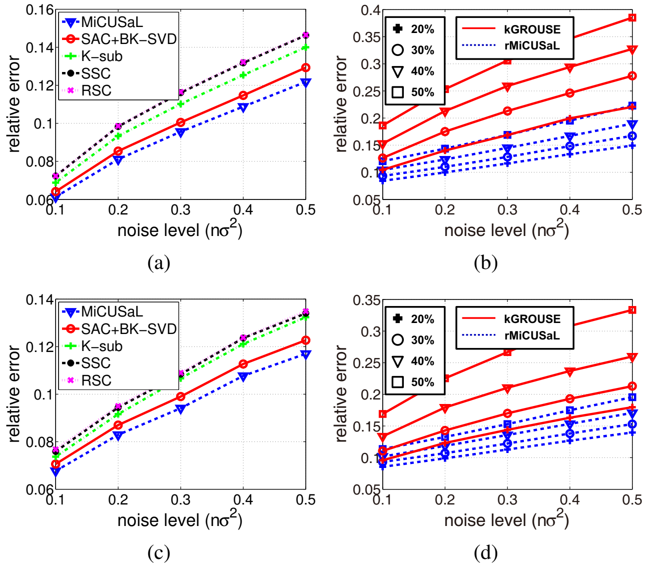

Fig. 2. Comparison of MC-UoS learning performance on face dataset. (a) and (b) show the relative errors for complete and missing data cases in the absence of outliers. (c) and (d) show the relative errors for complete and missing data cases in the presence of outliers.

bases $\{U_k \in \mathbb{R}^{n\times s}\}_{k=1}^K$. To generate the $\mathcal{S}_k$'s, we start with a random orthonormal basis $U \in \mathbb{R}^{n\times s}$ and let $\mathcal{S}_k = span(W_k)$ where $W_k = U + tB_k$ with B_k a random Gaussian $n \times s$ matrix and parameter t controlling the distance between subspaces. To make the $\mathcal{S}_k$'s close to each other, we set $t = 0.05$.

After generating the subspaces, we generate a set of points from $\mathcal{S}_k$ as $X_k = U_k C_k$, where $C_k \in \mathbb{R}^{s\times M}$ ($M = \frac{N}{K} = 100$) is a matrix whose entries are i.i.d. random variables with $\mathcal{N}(0,1)$ distribution. We then stack all the data into a matrix $X = [X_1, \ldots, X_4] = \{x_i\}_{i=1}^N$ and normalize all the points to unit ℓ_2 norms. Test data $X^{te} = \{x_i^{te}\}_{i=1}^N$ is produced using the same foregoing strategy.

Next, we make use of a collection of noisy samples, Y, to learn a union of K subspaces and stack the learned orthonormal bases $\{D_k\}_{k=1}^K$ into D. For MC-UoS learning performance analysis, we define d_{avg} as the average of normalized subspace distances between pairs of D_k's and U_k's as $d_{avg} \overset{def}{=} \frac{1}{K}\sum_{k=1}^K \sqrt{\frac{s-\text{tr}(D_k^T U_{\hat{k}} U_{\hat{k}}^T D_k)}{s}}$, where $\hat{k} = \arg\max_j \|D_k^T U_j\|_F$. We also ensure that no two D_k's are matched to the same U_k. A smaller d_{avg} indicates a better performance of MC-UoS learning. Also, if learned subspaces are closer to the ground truth, they are expected to have a good representation performance of complete test data. A good measure in this regard would be the mean of relative reconstruction errors of the test samples using learned subspaces. To be specific, we represent every signal $y_i^{te} \in Y^{te}$ such that $y_i^{te} \approx D_{\tilde{i}}\alpha_i^{te}$ where $\tilde{i} = \arg\max_k \|D_k^T y_i^{te}\|_2^2$ and $\alpha_i^{te} = D_{\tilde{i}}^T y_i^{te}$. The relative reconstruction error of $x_i^{te} \in X^{te}$ is then calculated as $\frac{\|x_i^{te} - D_{\tilde{i}}\alpha_i^{te}\|_2^2}{\|x_i^{te}\|_2^2}$.

It can be observed from Fig. 1(a) and Fig. 1(b) that (i) MiCUSaL learns a better MC-UoS in terms of smaller relative errors of test data and d_{avg}'s, and (ii) MiCUSaL degrades gracefully when noise power increases. Similarly, for rMiCUSaL, we can infer from Fig. 1(c) and Fig. 1(d) that (i) rMiCUSaL also outperforms k-GROUSE and (ii) for a fixed noise power, when the number of missing entries increases, the performance of rMiCUSaL degrades less compared to k-GROUSE. SSC fills in the missing entries with random values, which results in a poor performance in terms of large relative errors and d_{avg}'s (always above 0.75 and 0.90 respectively). Plots for SSC are omitted here because large gap exists between the performance of SSC and two other methods for the case of missing training data.

Experiments with Real Data: Finally, we study the performance of our methods on the Extended Yale B dataset [34]. By fixing the pose of one person and varying illumination, the set of images of one subject can be well represented by a union of 9-dimensional subspaces [35]. Here we assume the resulting images of a subject lie close to an MC-UoS with $K = 2$ and $s = 9$.

In our experiments, we focus on a collection of images of subject 8 and subsample the images to 48×42 pixels; thus $n = 2016$. We choose 54 images with good lighting conditions and all these samples are vectorized and normalized to have unit ℓ_2 norms. We randomly select half of them as X and the remaining ones belong to X^{te}. Fig. 2(a) and Fig. 2(b) show the relative reconstruction errors of test samples and we see both MiCUSaL and rMiCUSaL learn a better MC-UoS since they give rise to smaller relative errors.

We also study the scenario in which there exist some outliers in the training set. To do so, we randomly select 30 images from subject 8 and add another 10 unit ℓ_2-norm samples from subjects 5 and 13 (5 samples each), forming $X \in \mathbb{R}^{n\times 40}$. The test set X^{te} now consists of the remaining 24 samples of subject 8. We again provide evidence that our methods yield better representation performance in this setting and we refer the reader to Fig. 2(c) and Fig. 2(d) for a validation of this claim (plots for SSC are again omitted here because of its poor performance on missing data).

6. CONCLUSION AND FUTURE WORK

In this paper, we motivated and introduced a framework for data-adaptive learning of a metric-constrained union-of-subspaces model. Experimental results on both synthetic and real data indicate the effectiveness and robustness of our methods in the presence of noise, outliers and missing data. Our future work includes the estimation of the number and dimensions of the subspaces from the training data.

7. REFERENCES

[1] R. G. Baraniuk, V. Cevher, and M. B. Wakin, "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective," *Proc. IEEE*, vol. 98, no. 6, pp. 959–971, 2010.

[2] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psych.*, vol. 24, pp. 417–441, 498–520, 1933.

[3] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities," *Comm. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, 2004.

[4] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, 2005.

[5] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[6] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3655–3671, 2006.

[7] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[8] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2334–2345, 2008.

[9] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[10] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar, "Dictionary optimization for block-sparse representations," *IEEE Trans. Signal Processing*, vol. 60, no. 5, pp. 2386–2395, 2012.

[11] P. S. Bradley and O. L. Mangasarian, "k-Plane Clustering," *J. Global Optim.*, vol. 16, no. 1, pp. 23–32, 2000.

[12] P. Tseng, "Nearest q-flat to m points," *J. Optim. Theory Appl.*, vol. 105, no. 1, pp. 249–252, 2000.

[13] T. Zhang, A. Szlam, and G. Lerman, "Median K-flats for hybrid linear modeling with many outliers," in *Workshop on Subspace Methods*, 2009, pp. 234–241.

[14] B. V. Gowreesunker and A. H. Tewfik, "Learning sparse representation using iterative subspace identification," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3055–3065, 2010.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.

[16] E. L. Dyer, A. C. Sankaranarayanan, and R. G. Baraniuk, "Greedy feature selection for subspace clustering," *arXiv:1303.4778*, 2012.

[17] M. Soltanolkotabi, E. Elhamifar, and E. J. Candès, "Robust subspace clustering," *arXiv:1301.2603*, 2013.

[18] M. Soltanolkotabi and E. J. Candès, "A geometric analysis of subspace clustering with outliers," *Ann. Statist.*, vol. 40, no. 4, pp. 2195–2238, 2012.

[19] R. Heckel and H. Bölcskei, "Robust subspace clustering via thresholding," *arXiv:1307.4891*, 2013.

[20] B. Eriksson, L. Balzano, and R. Nowak, "High-rank matrix completion," in *Proc. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2012, pp. 373–381.

[21] L. Balzano, A. Szlam, B. Recht, and R. Nowak, "K-subspaces with missing data," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, 2012, pp. 612–615.

[22] L. Wang, X. Wang, and J. Feng, "Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition," *Pattern Recognition*, vol. 39, no. 3, pp. 456–464, 2006.

[23] X. Sun, L. Wang, and J. Feng, "Further results on the subspace distance," *Pattern Recognition*, vol. 40, no. 1, pp. 328–329, 2007.

[24] L. Wolf and A. Shashua, "Kernel principal angles for classification machines with applications to image sequence interpretation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2003, pp. 635–642.

[25] R. Wendel and A. Hurter, "Minimization of a non-separable objective function subject to disjoint constraints," *Oper. Res.*, vol. 24, pp. 643–657, 1976.

[26] S. Z. Selim and M. A. Ismail, "K-Means-Type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 1, pp. 81–87, 1984.

[27] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 2nd edition, 1999.

[28] E. Kokiopoulou, J. Chen, and Y. Saad, "Trace optimization and eigenproblems in dimension reduction methods," *Numer. Linear Algebra Appl.*, vol. 18, no. 3, pp. 565–602, 2011.

[29] L. Balzano, B. Recht, and R. Nowak, "High-dimensional matched subspace detection when data are missing," in *Proc. IEEE Intl. Symp. Inf. Theory (ISIT)*, 2010, pp. 1638–1642.

[30] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1999.

[31] A. Nedić and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

[32] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. of the Allerton Conference on Communication, Control and Computing*, 2010, pp. 704–711.

[33] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. J. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2003, pp. 11–18.

[34] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, 2005.

[35] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, 2003.