BEYOND UNION OF SUBSPACES: SUBSPACE PURSUIT ON GRASSMANN MANIFOLD FOR DATA REPRESENTATION

Xinyue Shen∗ Hamid Krim† Yuantao Gu∗

∗Tsinghua National Laboratory for Information Science and Technology Department of Electronic Engineering, Tsinghua University, Beijing 100084, CHINA
†Electrical and Computer Engineering Department, North Carolina State University Raleigh, NC 27695, U.S.A.

ABSTRACT

Discovering the underlying structure of a high-dimensional signal or big data has always been a challenging topic, and has become harder to tackle especially when the observations are exposed to arbitrary sparse perturbations. In this paper, built on the model of a union of subspaces (UoS) with sparse outliers and inspired by a basis pursuit strategy, we exploit the fundamental structure of a Grassmann manifold, and propose a new technique of pursuing the subspaces systematically by solving a non-convex optimization problem using the alternating direction method of multipliers. This problem as noted is further complicated by non-convex constraints on the Grassmann manifold, as well as the bilinearity in the penalty caused by the subspace bases and coefficients. Nevertheless, numerical experiments verify that the proposed algorithm, which provides elegant solutions to the sub-problems in each step, is able to de-couple the subspaces and pursue each of them under time-efficient parallel computation.

Index Terms— robust subspace pursuit, union of subspaces (UoS), Grassmann manifold constrained optimization, non-convex ADMM, parallel computation

1. INTRODUCTION

Recovering information or an underlying signal from noisy measurements has always been at the center of many signal processing problems. Many functional bases approaches have been proposed over the last two decades starting with a multiscale and nonlinear approximations of signals [1], including other robust reconstruction approaches [2], and more recent similarly inspired techniques [3, 4]. One approach which particularly caught the attention of researchers was that of a representation pursuit, such as the Matching Pursuit [5] and the Basis Pursuit [6], on account of their adaptive representation and hence recovery and of their high efficiency.

When the structure of the data deviates from the perennial vector subspace assumption, e.g., a curved space-manifold, or sets of subspaces, the recovery becomes challenging. Such subspace structures have also recently witnessed a resurgence, as in the recent works on subspace clustering [7–9] which show a particularly interesting and a promising potential of sparse models. In [7], a low-rank representation (LRR) also recovers subspace structures from sample-specific corruptions by jointly pursuing the lowest-rank representation of all data. The contaminated samples are sparse among all sampled data. Applications of such critical and important problem are available in signal processing, computer vision, and data mining [10]. Robust principal component pursuit is particularly successful in recovering low dimensional structures of high dimensional data from arbitrary sparse errors [11].

From another aspect, works in retrieving structural data, such as dictionary [12], low-rank matrix [13], and subspace [14–16], via optimization problems on manifolds have unveiled interesting geometry insights. In works [17–19], subspace projection matrix recovery from noisy partial information, as a subspace estimation problem, has been studied via non-convex optimization on a Grassmann manifold, and algorithms and theoretical recovery conditions are established by further exploiting the manifold structure.

Moreover, successful applications of sparse models in computer vision and machine learning [20] [21] have increasingly hinted at a more general model, namely that the underlying structure of high dimensional data looks more like a union of subspaces (UoS) [8, 9, 22–24] rather than one low dimensional subspace. In [22], the natural question about the feasibility of such an approach in high dimensional data modeling where the union of subspaces is further impacted by sparse errors, is addressed. This problem is intrinsically difficult, since the underlying subspace structure is also corrupted by unknown errors, which may lead to unreliable measurement of distance among data samples, and make data deviate from the original subspaces.

In this paper, built on the UoS with sparse outliers data model and inspired by a basis pursuit strategy, we exploit the fundamental structure of a Grassmann manifold, and propose a new technique where the subspaces of a UoS are systematically pursued by solving non-convex optimization problem on a Grassmann manifold. This problem as noted is further complicated by non-convex constraints as called for by the Grassmann manifold, as well as the bilinearity in the penalty caused by the subspace bases and coefficients. Nevertheless, the proposed algorithm gives an elegant solution to the Grassmann manifold-constraint part, and decouples the subspaces and admits parallel computation.

The following text is organized as follows. In Section 2, the background is reviewed and the problem is formulated as an optimization functional. In Section 3, the derivation is conducted to solve this optimization problem. In Section 4, we illustrate this approach by numerical experiments, and finally draw some concluding remarks in Section 5.
2. BACKGROUND AND PROBLEM FORMULATION

2.1. Grassmann Manifold

Definition 1. [25] A Grassmann manifold $\text{Gr}_{d,l}$ is the set of $l$-dimensional linear subspaces of $\mathbb{R}^d$, and is an $(l(d-l))$-dimensional compact manifold. It can be formally written as a quotient manifold

$$\text{Gr}_{N,s} := \frac{O(N)}{O(s) \times O(N-s)},$$

where $O(N)$ is the $N \times N$ orthogonal matrix group. A representative element of $\text{Gr}_{d,l}$ is a $d \times l$ orthonormal matrix $S$.

The projection distance defined on $\text{Gr}_{d,l}$ is as the following.

$$\forall S, V \in \text{Gr}_{d,l}, \quad d_p(S, V) := \frac{1}{2} \| SS^T - VV^T \|_F^2 = \sum_{i=1}^l \sin^2 \theta_i,$$

where $\{\theta_i\}_{i=1}^l$ are the principal angles between the two subspaces $S$ and $V$. In this work, we use $d_p$ to evaluate the distance between the target subspace and the estimated subspace.

2.2. Problem Statement and Formulation

We are given a $d \times n$ data matrix $X$ whose structure is a mixture of a noise-free component and perturbation of outliers, and hence admits a decomposition $X = L + E$. The matrix $E$ is thus a sparse matrix, and the columns of $L$ belong to $k$ unknown $l$-dimensional subspaces $\{S_1, \ldots, S_k\}$. We denote the orthogonal bases of $S_j$ by $S_j \in \mathbb{R}^{d \times l}$, then $L = \sum_{i=1}^k S_j \alpha_j$, where $\alpha_j \in \mathbb{R}^{d \times l}$. Our aim is to recover $E$ and $\{S_1, \ldots, S_k\}$ from $X$.

To that end, the recovery of $L$ in its assumed structure proceeds by a sparse constraint on $E$, which is in turn, reflected by its $\ell_1$ norm (the vector norm) as a convex relaxation of its $\ell_0$ norm. Each data point $X_t$, which is the $t$th column of $X$, is supposed to be only represented by the base vectors of its own subspace, i.e., for each $i$ the vector $\alpha_i := [\|\alpha_i\|_2, \ldots, \|\alpha_i\|_2]$ should only have one nonzero element, in which $\alpha_i$ is the $t$th column of $\alpha_i$. This, in turn, requires a sparse constraint on each $\alpha_i$, i.e., its $\ell_1$ norm should also form a penalty function.

In light of the above, and towards recovering $L$ and its associated orthogonal bases and $E$, we propose to solve the following optimization problem to find $E$ and $\{S_1, \ldots, S_k\}$ from $X$.

$$\min_{E, S_j, \alpha_j, j=1, \ldots, k} \lambda \| E \|_1 + \sum_{j=1}^k \sum_{i=1}^n \| \alpha_{ij} \|_2$$

s.t. $X = L + \sum_{j=1}^k S_j \alpha_j$, $S_j \in \text{Gr}_{d,l}, j = 1, \ldots, k$. (2)

For the given Grassmann manifold $\text{Gr}_{d,l}$, our search is for $l$ dimensional subspaces in $\mathbb{R}^d$. A brief and careful inspection of (2) reveals that the bilinearity of $\alpha$ and $S_j, j = 1, \ldots, k$, term results in the non-convexity of problem (2). This, in turn, suggests our adoption of ADMM [26] as a strategy towards a solution. Keep the $S_j$ constraints in (2), and the augmented Lagrangian is as follows,

$$L(E, S_j, \alpha_j, Y) = \lambda \| E \|_1 + \sum_{j=1}^k \sum_{i=1}^n \| \alpha_{ij} \|_2$$

$$+ \langle Y, X - \sum_{j=1}^k S_j \alpha_j \rangle + \frac{\mu}{2} \sum_{i=1}^n \| X - \sum_{j=1}^k S_j \alpha_j \|_F^2.$$
next.

Claim 3: The optimized subspace pursuit over \(G_{d,l} \) is achieved by,

\[
S_{j+1}^t = \arg \min_{S_j \in G_{d,l}} \langle Y^t, -S_j \alpha_{j+1}^t \rangle
\]

\[
\frac{\mu}{2} \left\| X - E^{t+1} - \sum_{p<j} \alpha_{p}^{t+1} - \sum_{p>j} S_p^t \alpha_p^{t+1} - S_j \alpha_j^{t+1} \right\|^2
\]

\[
= \arg \min_{S_j \in G_{d,l}} \left( Y^t + \mu \left( X - E^{t+1} - \sum_{p<j} \alpha_{p}^{t+1} - \sum_{p>j} S_p^t \alpha_p^{t+1} \right), -S_j \alpha_j^{t+1} \right)
\]

\[
= \arg \min_{S_j \in G_{d,l}} \left( -Y^t / \mu - X + E^{t+1} + \sum_{p<j} S_p^t \alpha_p^{t+1} + \sum_{p>j} S_p^t \alpha_p^{t+1} \right) \cdot (\alpha_{j+1}^T, S_j).
\]

The follow up closed form solution is obtained by first defining the singular value decomposition of

\[
\begin{pmatrix}
E^{t+1} & \sum_{p<j} S_p^t \alpha_p^{t+1} + \sum_{p>j} S_p^t \alpha_p^{t+1} - Y^t / \mu - X
\end{pmatrix} \cdot (\alpha_{j+1}^T)^T
\]

as \( U \Sigma V^T \), then \( \langle U \Sigma V^T, S_j \rangle = \langle V^T S_j^T U, \Sigma \rangle \). Since each entry of matrix \( V^T S_j^T U \) is not larger than 1, and \( \Sigma \) is diagonal, we have \( \langle V^T S_j^T U, \Sigma \rangle \geq \langle -I, \Sigma \rangle \), and the equality holds when

\[
S_j = -UV^T.
\]

The above calculation for \( S_{j+1}^t \) requiring a singular value decomposition, is not parallel for \( j \). This, as result, may entail a significant computational burden for large scale problems.

Claim 4: Towards lifting such a limitation, we propose to de-couple \( \{S_j\}_{j=1}^k \) by the following update of each \( \alpha_{ij} \) at the end of the second step.

\[
\alpha_{ij}^{t+1} = \begin{cases} \hat{\alpha}_{ij}^{t+1}, & j = \arg \max_{j=1, \ldots, k} \| \hat{\alpha}_{ij}^{t+1} \|_2, \\ 0, & \text{otherwise}, \end{cases}
\]

where \( \hat{\alpha}_{ij}^{t+1} \) is the output of the second step.

Since each data point in \( L \) should only belong to one subspace, for each fixed \( i \) we only preserve the column \( \alpha_{ij} \) with the largest \( \ell_2 \) norm, while the others are set to 0. For a fixed \( i \), denote the selected \( j \) as \( j(i) \). The benefit in so doing is that the objective function can be simplified, since if \( j = j(i) \), then \( \{\alpha_{ij}\}_{p \neq j} \) are all zeros, and if \( j \neq j(i) \), then \( \alpha_{ij} S_j \) vanishes. Therefore, (8) is simplified as follows.

\[
S_{j+1} = \arg \min_{S_j \in G_{d,l}} \sum_{i=1}^n \langle Y^t, X_i - E_i^{t+1} - S_{j(i)} \alpha_{j(i)}^{t+1} \rangle
\]

\[
+ \frac{\mu}{2} \sum_{i=1}^n \left\| X_i - E_i^{t+1} - S_{j(i)} \alpha_{j(i)}^{t+1} \right\|^2
\]

\[
= \arg \min_{S_j \in G_{d,l}} \sum_{i=1}^n I_{[j(i) = j]} \langle \mu E_i^{t+1} - \mu X_i - Y^t, S_j \alpha_{j(i)}^{t+1} \rangle
\]

\[
= -UV^T.
\]
each trial, the number of points $n = 24$, the dimension of the ambient space $d = 12$, the number of subspaces $k = 2$, the dimension of each subspace $l = 2$, and the sparsity of $E^*$ is 3. The simulation results are the dashed curves in Fig. 2. Note that for successful recovery, we mean that the error is less than 0.01. Firstly, the probability of recovering subspaces is higher than that of recovering $E$. When $\lambda = 1$ the gap is the largest, since $\lambda$ is too small to induce proper sparsity. Secondly, unlike some other non-convex ADMM algorithms [28], simply increasing $\mu$ does not ensure better convergence for the proposed algorithm. We can see that a relatively better choice is $\lambda = \mu = 2$, and the probability of recovery is about 0.93.

4.2. Adaptive $\mu$

For a better convergence behavior, we propose to use adaptive $\mu$ during the iterations. To be specific, we set $\mu^{t+1} = \rho^t \mu_0$ and $\rho > 1$.

In the following experiment, the optimal $\lambda$ and $\mu_0$ are tested for 100 random data, and $\rho = 1.00001$. The data settings are the same as the previous experiment, and the results are the solid curves in Fig. 2. Compared with the dashed curves, the convergence performance is improved, and with $\lambda = 2$ and $\mu_0 = 0.22$, the probability of subspace recovery reaches 1.

For a better understanding of the choice of parameters $\lambda$ and $\mu$, we test the subspace recovery probability with various $\lambda$ and $\mu_0$ in Fig. 3. In this experiment, $\rho = 1.00001$, $n = 24$, $d = 12$, $l = 2$, $k = 3$, and the sparsity level of $E^*$ is 3. Each probability is counted from 100 random trials. From the result, we first see that when $\mu_0 > \lambda$ the recovery is hardly achieved. Secondly, when $\mu_0 < \lambda$ the probability of recovery generally decreases as $\lambda$ increases from 1 to 1000. The optimal choice in this experiment is $\lambda = 1$ and $\mu_0 = 0.1$.

5. CONCLUSION

In this paper an optimization problem with constraints on Grassmann manifold is formulated to model the robust subspace pursuit problem of recovering a union of subspaces from a data matrix corrupted by sparse errors. Adopting the strategy of ADMM, the sparse errors, the representation coefficients, and the latent subspaces are iteratively solved in parallel by columns. Because of the non-convex constraint and the bilinearity in cost function, the convergence to a global optimum is difficult to be theoretically guaranteed. However, numerical simulations demonstrate that with elaborately selected parameter set, the union of subspaces can be successfully recovered in a large probability close to unitary.

This work is not exhausted, and there are further works to do on such problem. The optimal choice of $\lambda$ is to be theoretically analyzed for a given relative sparsity level between $E$ and the coefficients. For the algorithm, the impact of $\mu$ on the convergence behavior is to be theoretically studied.

6. REFERENCES


