Scalable Spectral k-Support Norm Regularization for Robust Low Rank Subspace Learning

Yiu-ming Cheung
Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
ymc@comp.hkbu.edu.hk

Jian Lou
Department of Computer Science
Hong Kong Baptist University
Hong Kong SAR, China
jianlou@comp.hkbu.edu.hk

ABSTRACT

As a fundamental tool in the fields of data mining and computer vision, robust low rank subspace learning is to recover a low rank matrix under gross corruptions that are often modeled by another sparse matrix. Within this learning, we investigate the spectral k-support norm, a more appealing convex relaxation than the popular nuclear norm, as a low rank penalty in this paper. Despite the better recovering performance, the spectral k-support norm entails the model difficult to be optimized efficiently, which severely limits its scalability from the practical perspective. Therefore, this paper proposes a scalable and efficient algorithm which considers the dual objective of the original problem that can take advantage of the more computational efficient linear oracle of the spectral k-support norm to be evaluated. Further, by studying the sub-gradient of the loss of the dual objective, a line-search strategy is adopted in the algorithm to enable it to adapt to the Hölder smoothness. Experiments on various tasks demonstrate the superior prediction performance and computation efficiency of the proposed algorithm.

Keywords

Robust Low Rank Subspace Learning, Spectral k-Support Norm, Conditional Gradient

1. INTRODUCTION

Recovering low rank matrix from gross corruptions has been a fundamental problem in machine learning, data mining and computer vision. Representative applications include collaborative filtering [27], background modeling [4], face clustering [17], among others. The gross corruption, also known as outliers, is often modeled by a sparse noise matrix. The robust low rank subspace learning tasks then aim to learn the low rank matrix with simultaneously minimizing the sparse noise matrix. In general, the low rank matrix and sparse matrix are required to satisfy certain linear constraints. With different designs of linear map, various tasks can be formulated by this linear constraint joint low rank and sparse matrix minimization problem, including robust principal component analysis (RPCA) [4] and low rank representation (LRR) [17].

Regarding the NP-hard rank minimization, nuclear norm is the most popular convex relaxation. As pointed out by [7], nuclear norm is actually the tightest convex relaxation of the nonconvex cardinality function (i.e. ℓ_0 norm function) of its singular values [7] under unit infinite norm ball. Recently, k-support norm [2], which seeks the tightest convex relaxation of the ℓ_0 norm (being value k) under unit ℓ_2-norm ball rather than infinite norm ball, has been studied. It has been shown that k-support norm outperforms the other convex relaxations such as ℓ_1 norm [24] and elastic net [33] for sparsity estimation, both theoretically and practically. Motivated by the success of k-support norm, spectral k-support norm [7, 18, 19] has been proposed to prompt low rankness of matrix by applying the k-support norm to the singular values of the matrix. Compared with nuclear norm, it provides tight relaxation of the rank k matrices under unit ℓ_2 norm ball of its singular values rather than infinite norm ball, which is often more preferred [7, 18]. Papers [18] and [19] have studied the spectral k-support norm in low rank matrix completion task and have reported the performance against the other convex penalties. [18] also shows the link of the spectral k-support norm between cluster norm used in the multi-task learning context. Furthermore, [19] extends it to spectral (k, p)-support norm to capture the decay of singular values of the underlying low rank matrix. Despite the superior recovery performance compared with other convex relaxations like nuclear norm, the spectral k-support norm is much more difficult to be optimized, which therefore severely limits its application domain, particularly for big data analysis. Although methods developed for k-support norm that relies on proximal map of the squared k-support norm [2, 7, 13] can be migrated to spectral k-support norm, its computation is laborious. A major reason is the full SVD decomposition involved in the proximal mapping computation. Furthermore, restricted by the property of the k-support norm, efficient approximation methods for nuclear norm (e.g. power method and Lanczos method) that requires leading singular values only are hardly applicable to spectral k-support norm. Further, a search operation that segments singular values into certain groups also needs additional computation.

In this paper, we will study the spectral k-support norm for robust low rank subspace learning task. Regarding optimization, it is apparently more challenging to design an

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983738

efficient and scalable algorithm than the previous research focusing on matrix completion [18, 19], given the additional linear constraint. We propose two variants for utilizing the spectral k-support norm, of which one uses the squared form as previous methods do. In the other variant, we show that we can also directly design an optimization algorithm for the original spectral k-support norm, which is hardly possible for most existing proximal gradient-based methods. We first follow the common practice to get rid of the linear constraint by Lagrangian dual. Next, instead of directly optimizing the Lagrangian dual alternatively as common ADM-based methods do, we further convert the problem by Fenchel conjugation [30]. The optimization of the resultant dual objective can then be solved via accelerated proximal gradient method (APG) [21], which only requires to evaluate the polar operator of spectral k-support norm, plus the proximal mapping related to infinity norm. Both of them are more computational efficient than the proximal map related to spectral k-support norm, in which the per-iteration cost is linear instead of superlinear. In principal, we follow the recently proposed primal-dual framework [22, 30] and recover the primal low rank variable along the dual APG iterations [30] By studying the (sub)gradient set of the loss function of the dual objective, we also incorporate the line-search strategy [30] that can adapt to the smoothness of the dual objective in the sense of Hölder continuity. Also, please note that line-search is possible in our method because the dual norm of the spectral k-support norm is more efficient to compute than itself, which is another advantage brought about by our dual conversion. Per-iteration complexity analysis shows that the time complexity of our method is linear with respect to the size of low rank matrix, whereas ADM-based methods would involve super-linear complexity.

In summary, we propose a polar operator-based algorithm featuring the following merits:

1. The proposed algorithm costs only linear per-iteration complexity rather than super-linear if proximal ADM method is adopted;

2. Our method is flexible to deal with both squared k-spectral norm and itself, whereas most of previous methods are confined with the former form; Also, our method is general so that it can be adapted to various choices of linear map, constant matrix and sparse norm to suit different model;

3. Our method converts to an equivalent dual form that deals with the dual spectral k-support norm, which is easier to compute than the primal norm. This further enables us to incorporate a line-search strategy to adapt to the degree and constant of the smoothness of the dual objective in the sense of Hölder smoothness.

2. PRELIMINARY

In this paper, we use the following notations. We denote a vector with a lowercase letter and a matrix with an uppercase letter. For a vector x, $||x||_1$, $||x||_2$ denote its ℓ_1 and ℓ_2 norm. For a matrix X, $||X||_1$, $||X||_{2,1}$, $||X||_F$, $||X||_*$ denote its ℓ_1, $\ell_{2,1}$ (sum of ℓ_2 norm of each column), Frobenius and nuclear norm (sum of singular values) correspondingly. For a particular singular value decomposition (SVD) of matrix $X \in \mathbb{R}^{m,n}$, we denote it as $X = Udiag(\sigma)V^T$, where $\sigma = (\sigma_1, ..., \sigma_{\min(m,n)})$ is the vector formed by singular values arranged in nonincreasing order and $diag(\sigma)$ is the diagonal matrix with its i-th diagonal element being σ_i. For a function f, we use $\nabla f(\Gamma)$ to denote its gradient or one of its subgradient at Γ, and use $\partial f(\Gamma)$ to denote the set of subgradient at Γ. The superscript $(\cdot)^T$ denotes the transpose for a matrix or the adjoint operation for a linear map.

2.1 Robust Low Rank Subspace Learning

In general, robust subspace learning methods seek a low rank component L plus a sparse component S capturing grossly corrupted outliers. L and S, together with a constant matrix M, are related by a linear constraint with constant linear map $\mathcal{B}$, which can be summarized into the following nonsmooth linear constraint problem,

$$\arg\min_{L,S} ||L||_r + \lambda||S||_s, \ s.t. \ \mathcal{B}(M - L) = S, \tag{1}$$

where the penalty $||\cdot||_r$ is used to promote low rankness of L, which is chosen as the spectral k-support norm $||\cdot||_{sp,k}$ [18],[19] in this paper. The second term $||\cdot||_s$ is the sparsity inducing penalty which can be ℓ_1 or $\ell_{2,1}$ norm [29]. λ is a constant parameter used to balance low rankness and sparsity.

In this paper, we focus on the RPCA [4] as a practical application, where M is the input data matrix D and $\mathcal{B}$ is identity matrix. With spectral k-support norm and ℓ_1 norm, the RPCA problem can be formulated as

$$\arg\min_{L,S} ||L||_{sp,k} + ||S||_1, \ s.t. \ D - L = S \ \textbf{(RPCA)}. \tag{2}$$

2.2 Spectral k-Support Norm

We first recall the k-support norm, which is introduced by [2] as a convex surrogate of the nonconvex cardinality function (a.k.a. ℓ_0 norm) for sparsity vector prediction. [2] observes that the most popular ℓ_1 norm is the convex hull of ℓ_0 norm on unit ℓ_∞ ball which assumes each entry to be bounded,

$$conv(x \in \mathbb{R}^d \big| ||x||_0 \leq k, \ ||x||_\infty \leq 1). \tag{3}$$

However, in many cases, we prefer the ℓ_2 norm of x to be bounded, i.e.

$$conv(x \in \mathbb{R}^d \big| ||x||_0 \leq k, \ ||x||_2 \leq 1), \tag{4}$$

which can help improve robustness and generalization. In this perspective, [2] proposes the k-support norm which can be calculated as follows,

$$||x||_{sp,k} = \Big(\sum_{i=1}^{k-t-1} (x_i)^2 + \frac{1}{t+1} \Big(\sum_{i=k-t}^{d} x_i \Big)^2 \Big)^{\frac{1}{2}}, \tag{5}$$

where t is an index satisfying the following relationship,

$$x_{k-t-1} > \frac{1}{t+1} \sum_{i=k-t}^{d} x_i \geq x_{k-t}. \tag{6}$$

[18],[19] then extend the k-support norm to low rank promoting purpose for matrices. Similar to the definition of nuclear norm, the spectral k-support norm (we use the same notation $||\cdot||_{sp,k}$ as the spectral form when the variable is matrix) is also defined in terms of the matrix singular values and is thus unitary invariant. In detail, for a matrix

$Z \in \mathbb{R}^{m\times n}$ and denoting a particular singular value decomposition (SVD) as $Z = Udiag(\sigma)V^T$, the spectral k-support norm can be computed by

$$||Z||_{sp,k} = \Big(\sum_{i=1}^{k-t-1} (\sigma_i)^2 + \frac{1}{t+1}\Big(\sum_{i=k-t}^{\min\{m,n\}} \sigma_i \Big)^2 \Big)^{\frac{1}{2}} \quad (7)$$

where index $t \in \{0, 1, ..., k-1\}$ is searched to satisfy $\sigma_{k-t-1} > \frac{1}{t+1}\sum_{i=k-t}^{\min\{m,n\}} \sigma_i \geq \sigma_{k-t}$. Apparently, the unit ball of spectral k-support norm is defined in terms of its singular values and can be expressed as the convex hull of vectors with at most k cardinality lying within the ℓ_2 norm ball, i.e.

$$\mathcal{D} = conv(\mathcal{A}),\ where \quad (8)$$

$$\mathcal{A} = \{A \in \mathbb{R}^{(m,n)} \big| A = Udiag(\sigma)V^T,\ ||\sigma||_0 \leq k, ||\sigma||_2 \leq 1\}. \quad (9)$$

When $k = 1$, the spectral k-support norm becomes nuclear norm, and when $k = \min\{m, n\}$, it coincides with Frobenius norm. Intuitively, it penalizes the largest $k - t - 1$ singular values with ℓ_2 norm while penalizing smaller $t + 1$ singular values with ℓ_1 norm. The k-support norm and spectral k-support norm are indeed norm functions [2, 18]. Also, denoting the dual norm by $||\cdot||^*_{sp,k}$, for any matrix Z with a particular SVD of $Z = Udiag(\sigma)V^T$, we have

$$||Z||^*_{sp,k} = \sqrt{\sum_{i=1}^{k} \sigma_i^2}. \quad (10)$$

It is obvious that the dual norm can be more efficient to compute because: 1) it only requires the first k singular values; 2) it avoids search for index t.

[19] also generalizes the spectral k-support norm to the so-called spectral (k,p)-support norm by using ℓ_p unit norm ball constraint in eq.(7) instead of the ℓ_2 unit norm ball. This extension can be denoted by $||\cdot||_{sp,(k,p)}$, under which the spectral k-support norm is $||\cdot||_{sp,(k,2)}$. [19] shows that by varying p, the generalized spectral (k,p)-norm can capture the decay of singular values of the desired low rank matrix in a low rank matrix completion task. Most computation of the spectral (k,p)-support norm is similar with spectral k-support norm. For example, to calculate the dual norm, we simply change 2 with q by

$$||Z||^*_{sp,(k,p)} = (\sum_{i=1}^{k} \sigma_i^q)^{\frac{1}{q}},\ where\ \frac{1}{p} + \frac{1}{q} = 1. \quad (11)$$

.

2.3 Scalable Algorithm with Spectral k-Support Norm

With the spectral k-support norm, the robust low rank subspace learning problem in eq.(1), featuring a nonsmooth and linear constraint optimization problem, is difficult to be solved in a scalable way, which severely limits the application of the spectral k-support norm from the practical perspective. In this subsection, we will explain that popular approaches to scaling nuclear norm regularization under this model is not applicable to spectral k-support norm. Specially, for nuclear norm, matrix factorization-based methods and ADM-type methods are both effective algorithms for solving eq.(1) efficiently, but none of them can be applied to solve spectral k-support norm regularized problem efficiently. The matrix factorization-based methods crucially rely on the following property of nuclear norm: $||Z||_* = \min_{P,Q} \frac{1}{2}||P||_F^2 + \frac{1}{2}||Q||_F^2,\ s.t. Z = PQ$, which is not applicable to spectral k-support norm. For the ADM-type methods, we argue that the proximal operator-based ADM method and its variants cannot optimize the spectral k-support norm regularized robust subspace learning problem in a scalable way by briefly deriving such an algorithm based on a particular linearized ADMM scheme [16] as follows:

$$||L||_{sp,k} + \lambda||S||_1 + \langle \Gamma, \mathcal{B}L + S - \mathcal{M}\rangle + \frac{\rho}{2}||\mathcal{B}L + S - \mathcal{B}M||_F^2. \quad (12)$$

Then, it will update L, S, Γ in an alternate fashion. In particular, to optimize L, we linearize the squared Frobenius norm term

$$\arg\min_L ||L||_{sp,k} + \langle \mathcal{B}^T\Gamma_t, L\rangle + \rho\mathcal{B}^T(\langle \mathcal{B}L_t + S_t - \mathcal{B}M, L\rangle) + \frac{\eta}{2}||L - L_t||_F^2. \quad (13)$$

This will require the proximal operator related to $||\cdot||_{sp,k}$,

$$L_{t+1} = \arg\min_L ||L||^2_{sp,k} + \frac{\eta}{2}||L - C_t||_F^2. \quad (14)$$

Please note that eq.(14) uses the squared spectral k-support norm instead, which has yet to know whether a closed-form solution exits for this norm in the literature. Actually, all existing methods resort to the squared k-support norm, which has closed-form solution. It is not difficult to adapt the proximal operators for the squared k-support norm [2, 13, 7] for spectral k-support norm. However, all existing proximal mappings cannot be computed in a scalable way. The main bottleneck is that proximal mapping would require a full SVD decomposition plus a searching step to segment the singular values into three different groups for different types of computation. [13] improves upon [2] by using binary search instead of the exhaustive search, and [7] proposes to solve the proximal mapping of the spectral k-support norm by computing the proximal mapping of its dual norm. However, none of these methods are able to avoid the full SVD because the search step and the subsequent computation both rely on all of the singular values. Nuclear norm-based ADM method is able to avoid such full SVD by an approximation technique that only requires to compute a few leading singular values, which is, unfortunately, not applicable here for spectral k-support norm. As a result, such ADM-based method would incur super-linear per-iteration cost that severely limits the scalability of spectral k-support norm's utilization under this model.

3. THE PROPOSED METHOD

In this section, we present our proposed method for learning robust low rank subspace with spectral k-support norm regularization in an efficient way. We begin with two reformulations and derive the corresponding equivalent problem based on Fenchel dual, one of which uses the squared spectral k-support norm and the other uses the original spectral k-support norm. The reformulated equivalent problems, referred as dual objectives, allow more efficient computation, in which the per-iteration cost hinges on solving a linear subproblem, referred as linear oracle evaluation of the spectral k-support norm. The linear oracle evaluation only needs to

compute the leading k-singular value decomposition (SVD), avoiding the full SVD computation otherwise required by proximal mapping-based ADM methods, is known to be more efficient to compute, especially with Lanczos method or power method techniques. Also, our method does not require the search step of the proximal mapping of spectral k-support norm. In addition, we study the smoothness of our loss function of the dual objective and incorporate a line-search strategy that can adapt to the smoothness change in the sense of Hölder continuity to further accelerate the algorithm.

3.1 Formulation I: Usage with Squared Spectral k-Support Norm

In our first formulation, we utilize the squared spectral k-support norm, which is adopted by almost all proximal mapping-based methods [2, 13, 7]. Let L denote the target low rank variable, we are solving the following constraint form of robust low rank subspace model:

$$\min_{L} \frac{1}{2}||L||^2_{sp,k},\ s.t.\ ||S||_s \leq \tau,\ \mathcal{B}(M-L)=S. \tag{15}$$

The above formulation amounts to the constraint $||\mathcal{B}(M-L)||_s$, which is considered more natural than regularization formulation because it directly signifies the tolerance on the misfit [1]. With a proper choice of τ, it is equivalent to the regularized form in eq.(1). Denoting the dual variable by Γ, by using the Lagrangian dual to handle the linear constraint $\mathcal{B}(M-L)=S$, we also get the following Lagrangian formulation,

$$\max_{\Gamma} \min_{S,L,||S||_s\leq\tau} \Big[\frac{1}{2}||L||^2_{sp,k} + \langle \Gamma, \mathcal{B}L + S - \mathcal{B}M\rangle\Big]. \tag{16}$$

However, instead of performing alternative updating strategy which would incur the usage of the expensive proximal map of the square form of the spectral k-support norm, we further convert eq.(16) by Fenchel conjugation, as summarized in the following proposition.

Proposition 1. *To solve the maximization problem in eq.(16), it is equivalent to solve the following minimization problem w.r.t the Lagrangian dual variable* Γ*,*

$$\min_{\Gamma} f(\Gamma) + r(\Gamma),\ where \tag{17}$$

$$f(\Gamma) = \frac{1}{2}(||-\mathcal{B}^T\Gamma||^*_{sp,k})^2 + \langle \Gamma, \mathcal{B}M\rangle, \tag{18}$$

$$r(\Gamma) = \tau||-\Gamma||^*_s. \tag{19}$$

In eq.(19), $||\cdot||^*_s$ denotes the dual norm of $||\cdot||_s$, e.g. $||\cdot||_\infty$ for $||\cdot||_1$ norm and $||\cdot||_{2,\infty}$ for $||\cdot||_{2,1}$ norm. Proposition 1 converts the optimization of eq.(16) to eq.(17) that is referred as dual objective in the sequel. To solve eq.(16) with respect to Lagrangian dual variable Γ, we can apply the proximal gradient descent algorithm [21]. The proximal map is now related to $||\cdot||^*_s$, which is essentially equivalent to projection onto $||\cdot||^*_s$ unit ball and is not expensive [6, 25]. Hence another major per-iteration cost would be the gradient evaluation of $f(\Gamma)$. Before proceeding to the computation of the gradient, we give a brief proof of Proposition 1, which would reveal a particular choice of (sub)gradient of the loss function $f(\Gamma)$.

Proof. To prove Proposition 1, we begin with the following sequence of equivalence relations:

$$\begin{aligned}
&\max_{\Gamma} \min_{S,L,||S||_s\leq\tau} \Big[\frac{1}{2}||L||^2_{sp,k} + \langle \Gamma, \mathcal{B}L + S - \mathcal{B}M\rangle\Big] \\
\Longleftrightarrow &\max_{\Gamma} \Big[\min_{L} \big(\frac{1}{2}||L||^2_{sp,k} + \langle \Gamma, \mathcal{B}L\rangle\big) - \langle \Gamma, \mathcal{B}M\rangle \\
&\qquad + \min_{||S||_s\leq\tau} \big(\langle \Gamma, S\rangle\big)\Big] \\
\Longleftrightarrow &\max_{\Gamma} \Big[\min_{L} -\big(\langle -\mathcal{B}^T\Gamma, L\rangle - \frac{1}{2}||L||^2_{sp,k}\big) - \langle \Gamma, \mathcal{B}M\rangle \\
&\qquad + \min_{||S||_s\leq\tau} -\big(\langle -\Gamma, S\rangle\big)\Big] \\
\Longleftrightarrow &\max_{\Gamma} -\Big[\max_{L} \big(\langle -\mathcal{B}^T\Gamma, L\rangle - \frac{1}{2}||L||^2_{sp,k}\big) + \langle \Gamma, \mathcal{B}M\rangle \\
&\qquad + \max_{||S||_s\leq\tau} \big(\langle -\Gamma, S\rangle\big)\Big].
\end{aligned} \tag{20}$$

The first and the second term in the square bracket can be combined and converted as follows:

$$\begin{aligned}
&\max_{L} \big(\langle -\mathcal{B}^T\Gamma, L\rangle - \frac{1}{2}||L||^2_{sp,k}\big) + \langle \Gamma, \mathcal{B}M\rangle \\
=&\frac{1}{2}(||-\mathcal{B}^T\Gamma||^*_{sp,k})^2 + \langle \Gamma, \mathcal{B}M\rangle := f(\Gamma)
\end{aligned} \tag{21}$$

The third term in the square bracket can be rewritten based on the definition of dual norm of $||\cdot||_s$, i.e.

$$\max_{||S||_s\leq\tau} \langle -\Gamma, S\rangle = \max_{||S/\tau||_s\leq 1} \big(\langle -\tau\Gamma, S/\tau\rangle\big) = \tau||-\Gamma||^*_s := r(\Gamma). \tag{22}$$

By combining the above derivation together, we can solve the right-hand side problem to equivalently solve the original Lagrangian dual problem on the left-hand side of the following equation:

$$\max_{\Gamma} \min_{S,L,||S||_s\leq\tau} \Big[\frac{1}{2}||L||^2_{sp,k} + \langle \Gamma, \mathcal{B}L + S - \mathcal{B}M\rangle\Big] \tag{23}$$

$$\Longleftrightarrow\ -\min_{\Gamma} \big(f(\Gamma) + r(\Gamma)\big). \tag{24}$$

□

Based on eq.(21) (i.e. taking the derivative of the first line in eq.(21) w.r.t. Γ), we have a particular choice of the (sub)gradient of the dual loss function $f(\Gamma)$, as shown in the following corollary.

Corollary 1. *Denote a particular subgradient of* $\partial f(\Gamma)$ *by* $g(\Gamma)$*, then it can be computed as*

$$g(\Gamma) = -\mathcal{B}L^{\#} + \mathcal{B}M,\ where\ L^{\#} = \arg\max_{||A||_{sp,k}\leq 1} \langle -\mathcal{B}^T\Gamma, A\rangle. \tag{25}$$

According to Proposition (1), the computation of computing the (sub)gradient of the dual objective comes from computing $L^{\#}$, which requires to solve a linear problem $\arg\max_{||A||_{sp,k}\leq 1}\langle -\mathcal{B}^T\Gamma, A\rangle$. The spectral k-support norm, as an gauge function [9] (i.e. nonnegative, positively homogeneous convex functions vanishing at the origin), allows the linear subproblem to be equivalently solved by the following polar operator:

$$\arg\max_{A\in\mathcal{A}} \langle -\mathcal{B}^T\Gamma, A\rangle. \tag{26}$$

Recall that $\mathcal{A}$ is the set of "atoms" of the spectral k-support norm defined in eq.(9) and also note that the structure of

the $A \in \mathcal{A}$ constraint set is much simpler to deal with than $||A||_{sp,k} \leq 1$. In fact, the polar operator has closed-form solution which only computes top k-SVD of matrix $\mathcal{B}^T\Gamma$ in eq.(25), as shown in the following lemma from [19]:

Lemma 1. *Denote a particular SVD of an arbitrary matrix $X \in \mathbb{R}^{(m,n)}$ by $X = Udiag(\sigma)V^T$. Then the polar operator of the spectral (k,p)-support norm, i.e. $L^\# = \arg\sup_{A\in\mathcal{A}}\langle X, A\rangle$ (recall that $\mathcal{A}$ is the "atomic" set in eq.(9)), admits the closed-form solution as $L^\# = Udiag(s)V^T$, where*

$$s_i = \begin{cases} (\frac{\sigma_i}{||\sigma||^*_{sp,(k,p)}})^{\frac{1}{p-1}}, & i = 1, ..., k \\ 0, & i = k+1, ... \min\{m, n\}. \end{cases} \tag{27}$$

Recall that $||\sigma||^*_{sp,(k,p)}$ is the dual spectral (k,p) support norm of X in eq.(11) and simply set $p = q = 2$ for spectral k-support norm. According to Lemma (1), the computation of the polar operator, and thus the gradient of the dual objective, only involves the top k-SVD, which is more efficient to evaluate than full SVD, especially with Lanczos [14] or perhaps power method [10] techniques. Please note that although [19] also utilizes the polar operator, their methods are based on vanilla Frank-Wolfe algorithm, which is not applicable when additional linear constraint is involved.

3.2 Formulation II: Usage with Spectral k-Support Norm

In this subsection, we propose our second formulation that utilizes the spectral k-support norm itself, which is impossible for proximal mapping-based approach due to the lack of known closed-form proximal mapping. Again, we begin with the following constraint formulation:

$$\min_L ||L||_{sp,k},\ s.t.\ ||S||_s \leq \tau,\ \mathcal{B}(M - L) = S. \tag{28}$$

Before converting it to Lagrangian dual form to get rid of the equality constraint, we introduce an auxiliary variable v_l with:

$$\min_{v_l}\ v_l,\ s.t.\ ||S||_1 \leq \tau,\ \mathcal{B}(M - L) = S,\ ||L||_{sp,k} \leq v_l \leq Q_l, \tag{29}$$

where Q_l is a constant estimation of the upper bound of $||L||_{sp,k}$. This technique has been previously introduced by [11] and later also adopted by [20] for extending Frank-Wolfe algorithms [12] to norm regularization problem. Again, denoting the Lagrangian dual variable by Γ, we have

$$\max_\Gamma \min_{L,v_l,S} \Big[v_l + \langle\Gamma, \mathcal{B}L + S - \mathcal{B}M\rangle \Big| ||L||_{sp,k} \leq v_l \leq Q_l,\ ||S||_s \leq \tau\Big]. \tag{30}$$

We then further transform the above formulation by Fenchel conjugation summarized by the following proposition.

Proposition 2. *To solve the maximization problem in eq.(30), it is equivalent to solve the following minimization problem with respect to the Lagrangian dual variable Γ:*

$$\min_\Gamma f(\Gamma) + r(\Gamma),\ where \tag{31}$$

$$f(\Gamma) = \max\{0, (Q_l||\ - \mathcal{B}^T\Gamma||^*_{sp.k} - 1)\} + \langle\Gamma, \mathcal{B}M\rangle, \tag{32}$$

$$r(\Gamma) = \tau||\ - \Gamma||^*_s. \tag{33}$$

Proof. To prove Proposition 2, we begin with the following equivalent relationship, which is related to the low rank component L and v_l:

$$\begin{aligned} &\min_{v_l,L}\Big[v_l + \langle\Gamma, \mathcal{B}L\rangle\Big| ||L||_{sp,k} \leq v_l \leq Q_l\Big] \\ =&\min_{v_l,A}\Big[v_l\big(1 + \langle\mathcal{B}^T\Gamma, A\rangle\big)\Big| ||A||_{sp,k} \leq 1,\ 0 \leq v_l \leq Q_l\Big] \\ =&\min_{0\leq v_l\leq Q_l}\Big[v_l\big(1 - \max_{A\in\mathcal{A}}\langle -\mathcal{B}^T\Gamma, A\rangle\big)\Big] \\ =&-\max_{0\leq v_l\leq Q_l}\Big[v_l(||\ - \mathcal{B}^T\Gamma||^*_{sp,k} - 1)\Big]. \end{aligned} \tag{34}$$

If $l(\Gamma) := (||\ - \mathcal{B}^T\Gamma||^*_{sp,k} - 1) > 0$, $\max_{0\leq v_l\leq Q_l}\big[v_l l(\Gamma)\big] = Q_l l(\Gamma)$ because the optimal $v_l^\# = Q_l$; Otherwise, $\max_{0\leq v_l\leq Q_l}\big[v_l l(\Gamma)\big] = 0$ because the optimal $v_l^\# = 0$. That is,

$$\min_{v_l,L}\Big[v_l + \langle\Gamma, \mathcal{B}L\rangle\Big| ||L||_{sp,k} \leq v_l \leq Q_l\Big] = -\max\{0, Q_l l(\Gamma)\}. \tag{35}$$

As for the sparse component S, we can obtain the reformulation similar to Formulation I in the previous subsection, i.e.

$$\min_{||S||_s\leq\tau}\langle\Gamma, S\rangle = -\max_{||S||_s\leq\tau}\langle -\Gamma, S\rangle = -\tau||\ - \Gamma||^*_s := -r(\Gamma). \tag{36}$$

Combining the above together, we have the following dual problem:

$$\begin{aligned} &\max_\Gamma -\Big[\max\{0, Q_l l(\Gamma)\} + \langle\Gamma, \mathcal{B}M\rangle + r(\Gamma)\Big] \\ =&-\min_\Gamma\Big[\max\{0, Q_l l(\Gamma)\} + \langle\Gamma, \mathcal{B}M\rangle + r(\Gamma)\Big]. \end{aligned} \tag{37}$$

Therefore, we can equivalently solve

$$\min_\Gamma\Big[\big(\max\{0, Q_l l(\Gamma)\} + \langle\Gamma, \mathcal{B}M\rangle\big) + r(\Gamma)\Big] := \min_\Gamma f(\Gamma) + r(\Gamma). \tag{38}$$

□

The next corollary shows a particular choice of (sub)gradient for $f(\Gamma)$.

Corollary 2. *A particular choice of the (sub)gradient for $f(\Gamma)$ is given by $g(\Gamma)$:*

$$g(\Gamma) = \begin{cases} \mathcal{B}M - Q_l\mathcal{B}L^\#, & (||\ - \mathcal{B}^T\Gamma||^*_{sp,k} - 1) > 0 \\ \mathcal{B}M - conv\{0, Q_l\mathcal{B}L^\#\}, & (||\ - \mathcal{B}^T\Gamma||^*_{sp,k} - 1) = 0 \\ \mathcal{B}M, & (||\ - \mathcal{B}^T\Gamma||^*_{sp,k} - 1) < 0 \end{cases}, \tag{39}$$

where $L^\# = \arg\max_{A\in\mathcal{A}}\langle -\mathcal{B}^T\Gamma, A\rangle$ can be computed according to Lemma 1.

Corollary 2 shows that the major computational cost of the (sub)gradient for $f(\Gamma)$ depends again on the linear optimization problem of evaluating the polar operator of spectral k-support norm. Compared with Formulation I in the previous subsection, to learn with the spectral k-support norm itself, we need to tune one more parameter Q_l, which is used in eq.(34).

3.3 Algorithm

3.3.1 APG for the Dual Objective

Following [22, 30], we can then solve the converted dual objective with the accelerated proximal gradient descent

(APG) [3, 21]. The gradient of each step can be evaluated according to Corollary 1 and Corollary 2. In detail, we keep two interpolation sequences $\hat{\Gamma}_t$ and Γ_t, which is typical for APG-type methods. Specifically, in each iteration, the algorithm updates the dual variable Γ_t by,

$$\Gamma_{t+1} = \arg\min_{\Gamma} f(\hat{\Gamma}_t) + \langle g(\hat{\Gamma}_t), \Gamma - \hat{\Gamma}_t \rangle + \frac{H_{t+1}}{2}||\Gamma - \hat{\Gamma}_t||_F^2 + r(\Gamma); \tag{40}$$

$$\hat{\Gamma}_{t+1} = \Gamma_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(\Gamma_{t+1} - \Gamma_t), \tag{41}$$

where λ_t is a scalar sequence updated iteratively as $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$ with the initial value 1. H_t is the reciprocal of the step size. Also, recall that $g(\hat{\Gamma}_t)$ is the gradient of f at $\hat{\Gamma}_t$ which can be evaluated by eq.(25) and eq.(39).

The subproblem eq.(40) is actually the proximal mapping related to the dual norm of the sparsity inducing norm, which is essentially to compute the projection onto ℓ_1 norm ball. In detail, eq.(40) is the proximal mapping corresponds to $||\cdot||_s^*$ that is denoted as $prox_{H_{t+1}^{-1} r(\Gamma)}\big(\hat{\Gamma}_t - g(\hat{\Gamma}_t)/H_{t+1}\big)$,

$$\Gamma_{t+1} = \arg\min_{\Gamma} \frac{1}{2}||\Gamma - (\hat{\Gamma}_t - g(\hat{\Gamma}_t)/H_{t+1})||_F^2 + \frac{\tau}{H_{t+1}}||\Gamma||_s^*, \tag{42}$$

which can be equivalently evaluated by the projection on the unit $||\cdot||_s$ norm ball according to

$$\Gamma_{t+1} = \big(\hat{\Gamma}_t - \frac{g(\hat{\Gamma}_t)}{H_{t+1}}\big) - \frac{\tau}{H_{t+1}} proj\Big(\frac{1}{\frac{\tau}{H_{t+1}}}\big(\hat{\Gamma}_t - \frac{g(\hat{\Gamma}_t)}{H_{t+1}}\big)\Big), \tag{43}$$

where $proj(X)$ denotes the projection operation, e.g. projects onto ℓ_1-ball or $\ell_{2,1}$-ball, both of which allow efficient computation that costs linear complexity with respect to the size of X, i.e. $O(mn)$ for $X \in \mathbb{R}^{(m,n)}$ [25].

3.3.2 Line-search

In the following, we study the (sub)gradient set of $f(\Gamma)$, which apparently depends on the structure of the (sub)gradient of the dual norm $||\cdot||_{sp,k}^*$ (see eq.(18) and eq.(32)). To keep the study more general, the following lemma shows the form of (sub)gradient of the dual norm of spectral (k,p)-norm $||\cdot||_{sp,(k,p)}^*$, which is generalized from Proposition 5 in [5] and also see [26].

Proposition 3. *For $\Gamma \neq 0$, denote a particular singular value decomposition of Γ by $\Gamma = U diag(\sigma) V^T$ and suppose the singular values satisfies $\sigma_1 \geq \sigma_2 \geq ... > \sigma_{k-a+1} = ... = \sigma_k = ... = \sigma_{k+b} > ... \geq \sigma_d$. q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. Then, the subgradient set of the dual norm of the spectral (k,p)-support norm at Γ is*

$$\frac{1}{||\Gamma||_{sp,(k,p)}^{*(q-1)}}\Big\{U_{[:,1:k-a]} diag(\sigma_{[1:k-a]}^{q-1}) V_{[:,1:k-a]}^T + U_{[:,k-a+1:k+b]} R V_{[:,k-b+1:k+b]}^T\Big\}, \tag{44}$$

where R is a symmetric matrix and satisfies $||R||_2 \leq 1$ and $||R||_* = a$. In particular, it is differentiable when $\sigma_k > \sigma_{k+1}$ or $\sigma_k = 0$ with the gradient equal to

$$\frac{1}{||\Gamma||_{sp,(k,p)}^{*(q-1)}}\Big\{U_{[:,1:k]} diag(\sigma_{[1:k]}^{q-1}) V_{[:,1:k]}^T\Big\}. \tag{45}$$

According to Proposition (3), we actually choose eq.(45) as the (sub)gradient in computing the gradient of $g(\Gamma)$. The conditions of the uniqueness of the subgradient set, i.e. whether $\sigma_k > \sigma_{k+1}$ or $\sigma_k = 0$ is satisfied, can be interpreted as whether the first k singular vales of Γ are well-separated with the remaining singular values. Proposition (3) indicates that, when the first k singular values are well-separated, $g(\Gamma)$ would be differentiable. In practice, initializing with a low rank matrix Γ (e.g. all-zero matrix), we would expect that the singular values of Γ change from satisfying the uniqueness condition (e.g. $\sigma_k = 0$) to dissatisfying across iterations.

Therefore, the smoothness of the dual objective loss $g(\Gamma)$ would change from differentiable to subdifferentiable across iterations, which corresponds to degree $\nu = 1$ to degree $\nu = 0$ in the sense of Hölder continuity, which guarantees the following relationship (for more detailed properties, please see [22]), $||\nabla f(x) - \nabla f(y)|| \leq H_\nu ||x - y||^\nu$, $\forall x, y$, where $\nu \in [0, 1]$ is referred as the degree of smoothness and H_ν is assumed finite that is defined by

$$H_\nu := H_\nu(f) = \sup_{x \neq y \in \mathcal{D}} \frac{||\nabla f(x) - \nabla f(y)||}{||x - y||^\nu}. \tag{46}$$

Next, we utilize a line-search scheme proposed recently by [30], which is able to automatically adapt to both the degree and constant of the Hölder continuity of the dual objective and thus chooses more optimal step size. We denote the reciprocal of the step size at iteration t by H_t. According to proximal gradient update related to $r(\Gamma)$, we have

$$Q_{H_t}(\Gamma; \hat{\Gamma}_t) = f(\hat{\Gamma}_t) + \langle g(\hat{\Gamma}_t), \Gamma - \hat{\Gamma}_t \rangle + \frac{H_{t+1}}{2}||\Gamma - \hat{\Gamma}_t||_F^2. \tag{47}$$

In essence, the line search aims to find the minimum H_{t+1} (corresponding to the largest step size) that satisfies the following criterion:

$$f(\Gamma_{t+1}) \leq Q_{H_{t+1}}(\Gamma_{t+1}; \hat{\Gamma}_t) + \frac{\epsilon}{2\lambda_t}, \tag{48}$$

where ϵ is the error tolerance and λ_t is the sequence kept by APG algorithm.

3.3.3 Primal Variable Recovery

Thus far, we have dealt with the dual objective and dual variable. However, our ultimate goal is the primal variable L. To do so, we follow [30] to simultaneously maintain the primal variable sequence L_t across dual variable updating procedure, i.e. $L_{t+1} = (1 - \gamma_t)L_t + \gamma_t L^\#$, where γ_t is the weighting parameter and $L^\#$ is the polar operator result computed during the gradient evaluation. According to [30], γ_t is constructed by also taking information from the adaptive step size

$$\gamma_t = \frac{\lambda_t / H_t}{\sum_{i=1}^t \lambda_i / H_i}. \tag{49}$$

This primal update step is similar to the Frank-Wolfe algorithm. With a constant step size, the weighting strategy would look even more similar to the "standard" Frank- Wolfe weighting strategy $\frac{2}{t+1}$. By combining the above parts, the complete procedure is summarized in Algorithm 1.

Algorithm 1 Proposed algorithm

Input: $\Gamma_0,\ \hat{\Gamma}_0,\ L_0 = \mathbf{0}_{(m,n)},\ \lambda_0 = 1,\ v_l,\ Q_l,\ \tau,\ \epsilon > 0,\ t_{max}$;

1: **for** $t = 0, 1, ..., t_{max}$ **do**
2: Compute $L^{\#}$ by evaluating the polar operator in eq. (26) at $\hat{\Gamma}_t$;
3: Compute the (sub)gradient $g(\hat{\Gamma}_t)$ of $f(\Gamma)$ at $\hat{\Gamma}_t$ by eq.(25) (for Formulation I) or eq.(39) (for Formulation II);
4: Compute $\Gamma_{t+1} = prox_{H_{t+1}^{-1} r(\Gamma)}\big(\hat{\Gamma}_t - H_{t+1}^{-1} g(\hat{\Gamma}_t)\big)$ by eq.(43), where H_{t+1} is decided by line-search subroutine: **line-search**$(\hat{\Gamma}_t, g(\hat{\Gamma}_t), H_t, \epsilon, \lambda_t)$;
5: Update the weight γ_t for primal recovery by eq.(49);
6: Update the sequence λ_{t+1}: $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$;
7: Update interpolation sequence $\hat{\Gamma}_{t+1} = \Gamma_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}(\Gamma_{t+1} - \Gamma_t)$;
8: Update the primal sequence $L_{t+1} = (1-\gamma_t)L_t + \gamma_t L^{\#}$.
9: **end for**
10: **Return:** $L_{t_{max}}$;

The following is the line-search subroutine.

Algorithm 2 line-search subroutine

Input: $\hat{\Gamma}, g(\hat{\Gamma}), H_0, \epsilon, \lambda$;

1: **for** $i = 0, 1, ..., i_{max}$ **do**
2: $\Gamma_{i+1} = prox_{H_i^{-1} r(\Gamma)}\big(\hat{\Gamma} - H_i^{-1} g(\hat{\Gamma})\big)$;
3: **if** $f(\Gamma_{i+1}) \leq f(\hat{\Gamma}) + \langle g(\hat{\Gamma}), \Gamma_{i+1} - \hat{\Gamma}\rangle + \frac{H_i}{2}||\Gamma_{i+1} - \hat{\Gamma}||_F^2 + \frac{\epsilon}{2\lambda}$ **then**
4: break;
5: **else**
6: $H_{i+1} = 2H_i$;
7: **end if**
8: **end for**
9: **Return:** Γ_i, H_i;

3.3.4 Algorithm Analysis

To recover an underlying low rank matrix of size (m, n), the time complexity of each part of Algorithm 1 is as follows: step 2 costs $O(kmn)$ to compute the top k SVD; step 3 is simply the point-wise multiplication and summation, which costs $O(mn)$; the proximal map of $r(\Gamma)$ in step 4 takes $O(mn)$ which mainly comes from projection onto sparse norm ball [25]; the line search in step 4 costs $O(i_{max}kmn)$ to compute at most i_{max} times dual loss value that requires top k SVD; step 8 costs $O(mn)$. Therefore, the per-iteration complexity is $O(i_{max}kmn)$, where i_{max} is 2 on average as observed by [30]. Recall that proximal map-based ADM methods would cost $O(\min\{m,n\}mn)$ to compute the full SVD and $\min\{m,n\}log(\min\{m,n\})$ to compute the proximal map of the singular values of the target matrix [18]. For practical applications, k is often much smaller than $\min\{m,n\}$, e.g. we set k=3 for tasks in subsection (4.2), (4.3). Hence, the proposed method enjoys much lower per-iteration cost.

Now we discuss the convergence behavior of the proposed method by Theorem 2 from [30], depicted by the following theorem.

Theorem 1. *The primal sequence L_t generated by Algorithm 1 converges with the worst case iteration number to achieve ϵ error with $t_{max} = O(\inf_{\nu\in[0,1]}(\frac{H_\nu}{\epsilon})^{\frac{2}{1+\nu}})$.*

With the smooth objective, i.e. $\nu = 1$, the worst-case iteration number is the same as the one of Frank-Wolfe type algorithms that trade off lower per-iteration complexity with slower convergence rate to scale to larger problem. Also, in practice we find the line-search condition is too conservative. Actually, more efficient implementation can be made by checking the line-search condition every 5 to 10 iterations instead of one per-iteration.

4. EXPERIMENT

In this section, we study the empirical performance of the proposed method on both synthetic and real datasets to test on the RPCA model in eq.(2). In our implementation, we solve the k-SVD by the lansvd function in the PROPACK package [10] [1]. We empirically set k to be equal or slightly larger than the desired rank of the low rank matrix, which can also be selected by cross-validation. All experiments are done on a laptop computer running MATLAB.

We compare with 1) IALM [15] uses nuclear norm as low rank penalty; 2) PSSV [2][23] uses partial sum of singular values, i.e. omits the leading singular values in the nuclear norm, which is nonconvex; 3) FWT [3] [20] also uses nuclear norm, but it is an FW-based method instead of proximal mapping. We use recommended or default parameter settings for these compared methods. We do not compare with neither Regℓ_1-ALM [32] which imposes additional assumption that $L = PZ$, where P is orthogonal and Z is low rank, nor the composition of nuclear norm with nonconvex functions [28] like SCAD [8] and MCP [31] functions. For the former, we can expect performance gain if we substitute the nuclear norm penalty with spectral k-support norm for the corresponding low rank part. For the latter, we omit them because this paper focuses only on studying whether the spectral k-support can be a better and computational feasible *convex* relaxation than nuclear norm for the robust subspace learning problems.

Also, we would like to point out that our algorithm can actually be applied to more general joint low rank and sparse minimization model by taking different linear map and constant matrix in eq.(1). A representative problem is the low rank representation problem (LRR) [17], where M can be identity matrix I and $\mathcal{B}$ equals input data matrix D. With $\ell_{2,1}$ norm [29] to promote column-wise sparsity, the problem becomes $\arg\min_{L,S} ||L||_{sp,k} + ||S||_{2,1},\ s.t.\ D - DL = S$. In this regard, the proposed method is more favorable than algorithms dedicated only to RPCA problem.

4.1 Synthetic Data

This subsection evaluates the performance of the proposed algorithm on synthetic data. We generated the ground truth $d \times d$ low rank matrix G by first generating random matrix uniformly sampled within 0 to 1, which was then truncated by *lansvd* with the various rank ratio r. We added random Gaussian noise $\mathcal{N}(0, 0.1)$ to G. Finally, we obtained the input matrix M for testing by randomly setting matrix elements to either -20 or +20 in G with a series of corruption ratio c, which are outliers.

Figure 1 reports the recovery performance under a series

[1] http://sun.stanford.edu/~rmunk/PROPACK/
[2] http://thoh.kaist.ac.kr/Research/PartialSum/PartialSum.htm
[3] https://sites.google.com/site/mucun1988/publi

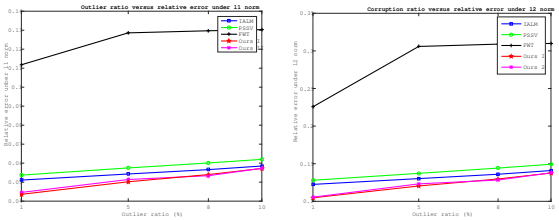

Figure 1: Corruption ratio versus relative error.

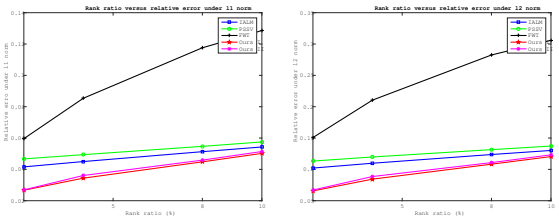

Figure 2: Rank ratio versus relative error.

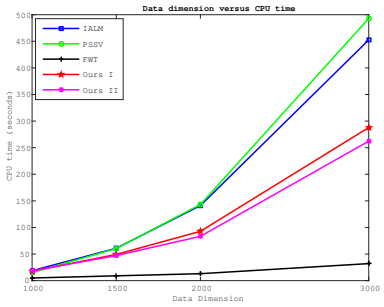

Figure 3: Data dimension versus CPU time.

Table 1: Videos used in the experiment

	campus	lobby
frame size	160×128	160×128
# of frames	$1,439$	$1,536$
M size	$20,480 \times 4,317$	$20,480 \times 4,608$

degree of corruption. We varied the corruption percentage from 1% to 10%. The data dimension was fixed with 1000×1000 and the rank ratio was set at 10%. We measured the reconstruction performance by ℓ_1 relative error $\frac{||L-G||_1}{d \times d}$ (left subplot) and ℓ_2 relative error $\frac{||L-G||_2}{||G||_F}$ (right subplot), where G is the ground truth matrix and L is the output of algorithms. From Figure 1, it can be seen that the recovery error increases with the more outliers. The two formulations of the spectral k-support norm regularized RPCA algorithm perform closely and are better than proximal mapping-based nuclear norm regularized (solved by IALM) and partial singular value sum regularizer algorithms (solved by PSSV). Note that the performance of PSSV is close to that of IALM, both of which are slightly worse. Among these algorithms, the performance of the FWT method is the worst. Although it uses proximal step for the sparse matrix update, the low rank part is still updated by pure Frank-Wolfe strategy, which is slow and cannot obtain enough decease of the objective compared to proximal algorithm for either primal (like ALM/ADMM) or for the dual form without further local refinement [12]. In Figure 2, we compared the algorithms with rank ratio varying from 1% to 10%, while fixing corruption ratio to 5% and data dimension to 1000×1000. Recovery performance under ℓ_1 (left) and ℓ_2 norm (right) are reported. Again, the proposed method with two formulations performs closely and are better than the counterparts. Therefore, the spectral k-support norm is superior to nuclear norm for RPCA task in terms of recovering performance.

Furthermore, we also studied the scalability of the proposed method, which is another key issue determining the feasibility for adopting spectral k-support norm in RPCA task. We generated the data with the sizes of 1000×1000 to 3000×3000 and set the corruption ratio to 1%, rank ratio to 10%. As shown in Figure 3, the proposed method is more efficient than IALM and PSSV. In the experiment, we also used the truncated SVD to approximately solve the SVT operator, i.e. proximal mapping of nuclear norm. Therefore, the proposed method costs comparable computation of per-iteration. Nevertheless, our method chooses optimal step size adaptive to the smoothness of the dual objective, which can explain why it is faster than IALM and PSSV. By contrast, FWT is much faster than all algorithms because it only computes the top singular value and corresponding vector that can be much more efficient than truncated SVD. As a result, our method, by avoiding full SVD if otherwise ADM is applied, makes the spectral k-support norm efficient enough to use compared to prevalent proximal mapping-based nuclear norm regularized methods. Although not as fast as FWT, the proposed method has the better recovery performance.

4.2 Background Modeling on Surveillance Videos

In this experiment, we considered modeling background in surveillance videos captured by a fixed camera. When stacking each frame as column vectors to form the input data matrix, the relative static background can be assumed to be low rank, while the foreground (e.g. human, car movements) can be modeled as sparse noise. Table 1 summarizes the dataset [4] used in this experiment. Since we dealt with color videos, in which each frame is described by three submatrices, we vectorized these matrices from each frame and stack them together to form a large matrix. Therefore, the row size of large input matrix equals length times width of the frame and the column size is three times of the frame number in the video.

Figure 4 shows the extraction performance of a sample frame of the campus dataset (left three columns) and lobby dataset (right three columns). In the raw image, the foreground mainly contains two pedestrians in the middle of the frame, one in dark shirts, while the other is in white, and a car at the left corner of the frame. The background extracted by spectral k-support norm is obviously better than that by (partial) nuclear norm, in which there are still vague contours of the pedestrian in white shirt in the middle and car in the left from the background extracted by (partial) nuclear norm. The rightmost three columns of Figure 4 present a pretty challenging sample frame from lobby dataset, where the two men stand in the middle of the frame for a moment leading them hard to be separate from background. In this

[4] http://perception.i2r.a-star.edu.sg/bkmodel/bkindex.html

Figure 4: Background modeling results on campus and lobby dataset, where the first, second and third row corresponds to nuclear norm by IALM (1.62e+04 seconds/1.62e+04 seconds), partial singular value by PSSV (1.18+e04 seconds/1.18+e04 seconds) and spectral k-support norm by proposed I (6.33e+03 seconds/2.91+e03 seconds) correspondingly. The first three columns show original sample frame (file:trees1831) from campus and the corresponding background and foreground; the last three columns show sample frame (file:SwitchLight2457) from lobby and the corresponding background and foreground.

case, (partial) nuclear norm is unable to remove these two men from the background. The spectral k-support norm is able to completely remove the man on the right and the man on the left only leaves with a vague contour. As a result, the spectral k-support norm has better recovery performance. Also, the running time indicates that our algorithm is efficient.

In conclusion, this experiment indicates that spectral k-support norm is superior than (partial) nuclear norm for recovering low rank matrix under sparse noise. Also, the proposed method is scalable to large scale tasks that makes the spectral k-support norm feasible to be applied for robust low rank subspace learning.

4.3 Face Reconstruction

In this experiment, we consider the face reconstruction task, where front face images are taken under varying conditions like changing illumination. When stacking all vectorized face together, the shadow and specularity caused by changing environment can be treated as sparse noise and the underling low rank matrix is the desired face image to be recovered. We used part of the Extended Yale-B Face Database-B (i.e. subjects 1 to 10 of 38 subjects in total), which contains 64 frontal face pictures of 192×168 pixels in each subject. When stacking them together, the input matrix is of size 32256×640, which is not very large compared to experiment in the previous subsection. In this case, IALM and PSSV are faster than the proposed method. A snapshot of the reconstruction result on sample images is illustrated in Figure 5. Visually, the spectral k-support norm outperforms both nuclear norm and partial sum of nuclear norm.

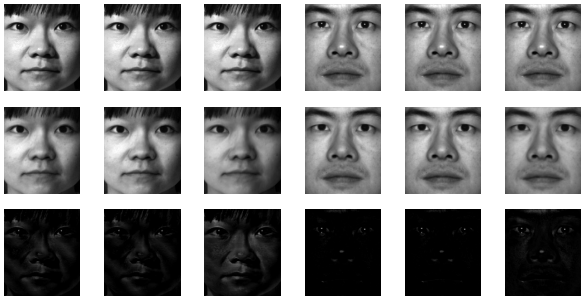

Figure 5: Face reconstruction result on Extended yale B face dataset. In each group of sample faces, the first, second and third column correspond to nuclear norm by IALM, partial sum of singular values by PSSV, and k-support norm by our Formulation I, respectively.

5. CONCLUSION

In this paper, we have studied robust low rank subspace learning problem with spectral k-support norm to promote the low rank property. Our method can utilize both the squared spectral k-support norm and itself. For both formulations, we consider a sparse norm fitting error ball constrained low rank optimization problem and transform it to the dual objective form. Solving the dual problem only involves a linear subproblem called polar operator and a projection onto the unit sparse ball, which allow us to avoid expensive proximal mapping of the spectral k-support norm. Furthermore, by studying the (sub)gradient of the dual norm of the more generalized spectral k-support norm, we have incorporated a line search strategy that is able to adapt to smoothness change. Experiment result on both synthetic and real datasets with background modeling and face reconstruction have successfully demonstrated the superiority of the proposed method in comparison with the existing counterparts.

6. ACKNOWLEDGMENT

This work was supported by the Faculty Research Grant of Hong Kong Baptist University under Project: FRG2/15-16/049, and National Natural Science Foundation of China under Grants: 61272366 and 61672444.

7. REFERENCES

[1] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy. Level-set methods for convex optimization. *arXiv:1602.01506*, 2016.

[2] A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1457–1465, 2012.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[5] X. V. Doan and S. Vavasis. Finding the largest low-rank clusters with ky fan 2-k-norm and l_1-norm. *arXiv:1403.5901*, 2014.

[6] J. C. Duchi, S. Shalev-shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning (ICML-08)*, pages 272–279, 2008.

[7] A. Eriksson, T. Thanh Pham, T.-J. Chin, and I. Reid. The k-support norm and convex envelopes of cardinality and rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3349–3357, 2015.

[8] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[9] R. M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.

[10] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[11] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[12] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 427–435, 2013.

[13] H. Lai, Y. Pan, C. Lu, Y. Tang, and S. Yan. Efficient k-support matrix pursuit. In *Computer Vision–ECCV 2014*, pages 617–631. Springer, 2014.

[14] R. M. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *DAIMI Report Series*, 27(537), 1998.

[15] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv:1009.5055*, 2010.

[16] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems*, pages 612–620, 2011.

[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184, 2013.

[18] A. M. McDonald, M. Pontil, and D. Stamos. Spectral k-support norm regularization. In *Advances in Neural Information Processing Systems*, pages 3644–3652, 2014.

[19] A. M. McDonald, M. Pontil, and D. Stamos. Fitting spectral decay with the k-support norm. *arXiv:1601.00449*, 2016.

[20] C. Mu, Y. Zhang, J. Wright, and D. Goldfarb. Scalable robust matrix recovery: Frank-wolfe meets proximal methods. *arXiv:1403.7588*, 2014.

[21] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[22] Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[23] T.-H. Oh, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Robust high dynamic range imaging by rank minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(6):1219–1232, 2015.

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

[25] E. Van, M. S. BERG, M. P. Friedlander, and K. Murphy. Group sparsity via linear-time projection. 2008.

[26] G. Watson. On matrix approximation problems with ky fank norms. *Numerical Algorithms*, 5(5):263–272, 1993.

[27] M. Weimer, A. Karatzoglou, Q. V. Le, and A. Smola. Maximum margin matrix factorization for collaborative ranking. In *Advances in neural information processing systems*, pages 1–8, 2007.

[28] Q. Yao, J. T. Kwok, and W. Zhong. Fast low-rank matrix learning with nonconvex regularization. In *Proceedings of the IEEE International Conference on Data Mining*, pages 539–548, 2015.

[29] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

[30] A. Yurtsever, Q. T. Dinh, and V. Cevher. A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*, pages 3132–3140, 2015.

[31] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010.

[32] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi. Practical low-rank matrix approximation under robust l 1-norm. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1410–1417, 2012.

[33] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.