Unmixing Incoherent Structures of Big Data by Randomized or Greedy Decomposition

Tianyi Zhou, Dacheng Tao

Abstract

Learning big data by matrix decomposition always suffers from expensive computation, mixing of complicated structures and noise. In this paper, we study more adaptive models and efficient algorithms that decompose a data matrix as the sum of semantic components with incoherent structures. We firstly introduce “GO decomposition (GoDec)”, an alternating projection method estimating the low-rank part $L$ and the sparse part $S$ from data matrix $X = L + S + G$ corrupted by noise $G$. Two acceleration strategies are proposed to obtain scalable unmixing algorithm on big data: 1) Bilateral random projection (BRP) is developed to speed up the update of $L$ in GoDec by a closed-form built from left and right random projections of $X \ominus S$ in lower dimensions; 2) Greedy bilateral (GreB) paradigm updates the left and right factors of $L$ in a mutually adaptive and greedy incremental manner, and achieve significant improvement in both time and sample complexities. Then we proposes three nontrivial variants of GoDec that generalizes GoDec to more general data type and whose fast algorithms can be derived from the two strategies: 1) for motion segmentation, we further decompose the sparse $S$ (moving objects) as the sum of multiple row-sparse matrices, each of which is a low-rank matrix after specific geometric transformation sequence and defines a motion shared by multiple objects; 2) for multi-label learning, we further decompose the low-rank $L$ into subcomponents with separable subspaces, each corresponds to the mapping a single label in feature space. Then the prediction can be effectively conducted by group lasso on the subspace ensemble; 3) for estimating scoring functions of each user in recommendation system, we further decompose the low-rank $L$ as $WZ^T$, where the rows of $W$ is the linear scoring functions and the rows of $Z$ are the items represented by available features. Empirical studies show the efficiency, robustness and effectiveness of the proposed methods in real applications.

Index Terms

Low-rank and sparse matrix decomposition, bilateral random projection, greedy bilateral paradigm, multi-label learning, background modeling, motion segmentation, recommendation systems

I. INTRODUCTION

Complex data is usually generated by mixing several components of different structures. These structures are often compressible, and are able to provide semantic interpretations of the data.
content. In addition, they can reveal the difference and similarity among data samples, and thus produce robust features playing vital roles in supervised or unsupervised learning tasks. Two types of structures have drawn lots of research attentions in recent years: 1) in compressed sensing [1], [2], a sparse signal can be exactly recovered from its linear measurements at a rate significant below the Nyquist rate, in sparse coding [3], [4], [5], an over-complete dictionary leads to sparse representations for dense signals of the same type; 2) in matrix completion [6], [7], [8], [9], [10], a low-rank matrix can be precisely rebuilt from a small portion of its entries by restricting the rows (samples) to lie in a subspace. In dimension reduction [11], [12], [13], [14], [15], [16], low-rank structure [17] has been broadly leveraged for exploring the geometry of point cloud. Although sparse and low-rank structures have been studied separately by a great number of researchers for years, the linear combination of them or their extensions is rarely explored until recently [18], [19], [20], [21]. Intuitively, fitting data with either sparse or low-rank structure is mature technique but is inevitably restricted by the limited data types they can model, while recent study shows that the linear mixture of them is more expressive in modeling complex data from different applications.

A motivating example is robust PCA [19] (RPCA), which decomposes the data matrix $X$ as $L + S$. The low-rank part $L$ summarizes a subspace that is shared by all the samples and thus reveals the global smoothness, while the sparse part $S$ captures the individual differences or abrupt changes among samples. A direct application of robust PCA is separating the sparse moving objects from the low-rank background in video sequence. Another interesting example is morphological component analysis (MCA) [22], which decompose the data into two parts that have sparse representations on two incoherent over-complete dictionaries, i.e., the first part has a very non-sparse representation on the dictionary of the second part, and vise versa. This requirement suggests that the two parts are separable on their sparse representations. Note that both RPCA and MCA can only work on data whose two building parts are incoherent, i.e., the content of one part cannot be moved to the other part without changing either of their structures (low-rank, sparse, dictionary, etc.). This incoherence condition could be viewed as a general extension of the statistical independence supporting independent component analysis (ICA) [23], [24] blindly separating non-Gaussian source signals. It leads to the identifiability of the structures in theory, and is demonstrated to be fulfilled on a wide class of real data.

However, new challenges arises when many recent studies tend to focus on big data with complex structures. Firstly, existing algorithms are computationally prohibitive to processing these data. For instance, the update of low-rank part in RPCA and in its extensions invoke a full singular value decomposition (SVD) per iterate, while MCA requires challenging $\ell_0$ or $\ell_1$ minimization per sample/feature and previously achieved incoherent dictionaries/transform.
operators encouraging sparse representations. Thus they suffer from a dramatic growth in time complexity when either feature dimensions or data samples increase. In previous methods, the structured information such as low-rank and sparse properties are always achieved at the price of time-consuming optimization, but are rarely leveraged for the purpose of improving the scalability. Recent progresses in randomized approximation and rank-revealing algorithms shed some light on the speedup of the robust PCA typed algorithms: the subspace of the low-rank part can be estimated from random sampling of its columns/rows or projections of its columns/rows on a random ensemble with bounded precision [25], [26], [27]. However, straightforward invoking this technique in RPCA problem needs to apply it to the updated residual matrix per iterate and thus may lead to costly computation. Besides, determining the rank of the low-rank part is not a trivial problem in practice.

Secondly, the simple low-rank, sparse and sparse representation assumptions cannot fully capture the sophisticated relation, individuality and sparsity of data samples with complex structures. While low-rank structure summarizes a global linear relationship between data points, the nonlinear relationship, local geometry and correlated functions are more common in big data and more expressive for a much wider class of structures. Moreover, the sparse matrix is simply explained by random noises on random positions in the past, but current studies reveal that it may have rich structured information that could be the central interests of various applications. For instance, the sparse motions captured by RPCA on video sequence data includes immense unexplored information favored by object tracking and behavior analysis. Furthermore, although the sparse representation is more general than sparse features, its quality largely relies on whether the given dictionary or transform operator fits the nature of data well. But this is difficult to evaluate when the data is of large volume and in general type.

Thirdly, two building parts are not sufficient to cover all the mixtures of incoherent structures in big data. One the one hand, dense noise is an extra component that has to be separated from the low-rank and sparse parts in many cases where the exact decomposition $X = L + S$ does not hold. This noisy assumption has been considered in stable PCP [28], DRMF [29] and other theoretical studies [20], and its robustness and adaptiveness to a broad class of data has also been verified. But efficient algorithm for the noisy model lacks. On the other hand, further decomposing the low-rank or sparse part to multiple distinguishable sub-components is potential to tell locally spatial or temporal relations within each identifiable structure and differences between them, which usually play pivot roles in supervised and unsupervised learning tasks. Although it appeals to be a natural extension to the two-part model in RPCA, how to formulate a proper decomposition model for learning problems and develop a practical algorithm are challenging.
A. Main Contributions

We start this paper by studying a novel low-rank and sparse matrix decomposition model “GO decomposition (GoDec)” \[ X = L + S + G \], which takes an extra dense noisy part \( G \) into account and casts the decomposition into alternating optimization of low-rank \( L \) and sparse \( S \). In order to overcome the computational burden caused by the large volume of data, we propose two acceleration strategies in designing the decomposition algorithms: the first is “bilateral random projection (BRP)” \[ \text{[31]} \] based fast low-rank approximation that results in a randomized update of the low-rank part or its nonlinear variant, this technique is based on recently developed random matrix theories that show a few random projections of a matrix is able to reveal its associated principle subspace \[ \text{[32], [25], [26], [27]} \]; the other is a Frank-Wolfe typed optimization scheme called “greedy bilateral (GreB)” paradigm \[ \text{[33]} \] that updates the left and right factors of the low-rank matrix variable in a mutually adaptive and greedy incremental manner. We show the two strategies generates considerably scalable algorithms for low-rank and sparse matrix decomposition. Moreover, both strategies have provable performance guarantee given by rigorous theoretical analysis (Appendix I and II).

In order to deal with the complicated structures that cannot be captured by the sum mixture of low-rank and sparse matrices, we proposes three variants of GoDec more expressive and general for learning from big data.

The first variant “shifted subspace tracking (SST)” \[ \text{[34]} \] is developed for motion segmentation \[ \text{[35], [36], [37], [38], [39]} \] from raw pixels of video sequence. SST further analyzes the unexplored rich structure of the sparse part \( S \) of GoDec, which could be seem as a sum mixture of several motions with distinct appearance and trajectories. SST unifies detection, tracking and segmenting multiple motions from complex scenes in a simple matrix factorization model.

The second variant “multi-label subspace ensemble (MSE)” \[ \text{[40]} \] extends the low-rank part \( L \) of GoDec to the sum of multiple low-rank matrices defined by distinguishable but correlated subspaces. MSE provides a novel insight into the multi-label learning (ML) problem \[ \text{[41], [42], [43], [44], [45]} \]. It addresses this problem by jointly learning inverse mappings that map each label to the feature space as a subspace, and formulating the prediction as finding the group sparse representation \[ \text{[46]} \] of a given sample on the ensemble of subspaces. There are only \( k \) subspaces needed to be learned, and the label correlations are fully used via considering correlation among subspaces.

The third variant “linear functional GoDec (LinGoDec)” learns scoring functions of users from their ratings matrix \( X \) and features of scored items \( Z \). It extends the low-rank part \( L \) of GoDec to \( WZ^T \), where \( W \) represents the linear functions and is constrained to be low-rank, while the rows of \( Z \) contain the features of items in the training set. In addition, the sparse
part $S$ is able to detect the advertising effects or anomaly of users’ ratings on specific items. LinGoDec formulates the collaborative filtering problem as supervised learning, and thus avoids time-consuming completion of the whole matrix when only a new item’s scores (a new row) are needed to be predicted.

The rest of this paper is organized as following: Section 2 introduces GoDec; Section 3 proposes the two acceleration strategies for processing large-scale data; Section 4 proposes the three variants of GoDec and their practical algorithms; Section 5 shows the experimental results of all the proposed algorithms on different application problems and justifies both the effectiveness and efficiency of them. The rows of all data matrices mentioned in this paper represents the samples and the columns denote the features.

II. GO DECOMPOSITION: UNMIXING LOW-RANK AND SPARSE STRUCTURES

In RPCA [19], PCP recovers $L$ and $S$ from $X$ by minimizing sum of the trace norm of $L$ and the $\ell_1$ norm of $S$. It can be proved that the solution to this convex relaxation is the exact recovery if $X = L + S$ indeed exists and $L$ and $S$ are sufficiently incoherent [18], [19]. That is, $L$ obeys the incoherence property and thus is not sparse, while $S$ has nonzero entries uniformly selected at random and thus is not low-rank. Popular optimization algorithms such as augmented Lagrangian multiplier, accelerated proximal gradient method and accelerated projected gradient method [21] have been applied. But full SVD as a costly subroutine is required to be repeatedly invoked in any of them.

Despite the strong theoretical guarantee of robust PCA, the exact decomposition $X = L + S$ does not always hold for real data matrix $X$ due to extra noise and complicated structure of $S$ that does not following Bernoulli-Gaussian distribution. Thus a more adaptive model $X = L + S + G$ is preferred, where $L + S$ approximates $X$ and $G$ is the dense noise. We then study the approximated “low-rank+sparse” decomposition of a matrix $X$, i.e.,

$$X = L + S + G, \text{rank}(L) \leq r, \text{card}(S) \leq k;$$

In this section, we develop “Go Decomposition” (GoDec) to estimate the low-rank part $L$ and the sparse part $S$ from $X$ by solving the following optimization problem, which aims at minimizing the decomposition error:

$$\min_{L,S} \|X - L - S\|_F^2,$$

s.t. \hspace{1em} \text{rank} (L) \leq r, \text{card} (S) \leq k.

(2)
A. Naïve GoDec

We propose the naïve GoDec algorithm at first and will study how to achieve an highly accelerated version in the next section. The optimization problem of GoDec (2) can be solved by alternatively solving the following two subproblems until convergence:

\[
\begin{align*}
L_t &= \arg \min_{\text{rank}(L) \leq r} \|X - L - S_{t-1}\|_F^2; \\
S_t &= \arg \min_{\text{card}(S) \leq k} \|X - L_t - S\|_F^2.
\end{align*}
\]  

(3)

Although both subproblems (3) have nonconvex constraints, their global solutions \(L_t\) and \(S_t\) exist. Let the SVD of a matrix \(X\) be \(U \Lambda V^T\) and \(\lambda_i(X)\) stands for the \(i^{th}\) largest singular value of \(X\); \(P_\Omega(\cdot)\) is the projection of a matrix to an entry set \(\Omega\).

In particular, the two subproblems in (3) can be solved by updating \(L_t\) via singular value hard thresholding of \(X - S_{t-1}\) and updating \(S_t\) via entry-wise hard thresholding of \(X - L_t\), respectively, i.e.,

\[
\begin{align*}
L_t &= \sum_{i=1}^{r} \lambda_i U_i V_i^T, \text{svd}(X - S_{t-1}) = U \Lambda V^T; \\
S_t &= P_\Omega(X - L_t), \Omega : \left| (X - L_t)_{i,j} \right| \neq 0 \\
&\quad \text{and } \geq \left| (X - L_t)_{i,j} \right|, |\Omega| \leq k.
\end{align*}
\]  

(4)

The main computation in the naïve GoDec algorithm (4) is the SVD of \(X - S_{t-1}\) in the updating \(L_t\) sequence. SVD requires \(\min(mn^2, m^2n)\) flops, so it is impractical when \(X\) is of large size, and more efficient algorithm is needed to be developed later.

GoDec alternatively assigns the \(r\)-rank approximation of \(X - S\) to \(L\) and assigns the sparse approximation with cardinality \(k\) of \(X - L\) to \(S\). The updating of \(L\) is obtained via singular value hard thresholding of \(X - S\), while the updating of \(S\) is obtained via entry-wise hard thresholding of \(X - L\). The term “GO” is owing to the similarities between \(L/S\) in the GoDec iteration rounds and the two players in the game of go.

Except the additional noisy part \(G\) and faster speed, the direct constraints to the rank of \(L\) and the cardinality \(S\) also makes GoDec different from RPCA minimizing their convex polytopes. This makes the rank and cardinality controllable, which is preferred in practice. Because prior information of these two parameters can be applied and lots of computations might be saved. In addition, GoDec introduces an efficient matrix completion algorithm [30], in which the cardinality constraint is replaced by a fixed support set. Convergence and robustness analysis of GoDec is given in Appendix I based on theory of alternating projection on two manifolds [48].
III. Two Acceleration Strategies for Unmixing Incoherent Structures

We firstly introduce the bilateral random projections (BRP) based low-rank approximation and its power scheme modification. BRP reduces the time consuming SVD in naïve GoDec to a closed-form approximation merely requiring small matrix multiplications. However, we need to invoke more expensive power scheme of BRP when the matrix spectrum does not have dramatic decreasing. Moreover, the rank needs to be estimated for saving unnecessary computations. Thus we propose greedy bilateral sketch (GreBske), which augments the matrix factors column/rows-wisely by selecting the best rank-one directions for approximation. It can adaptively determines the rank by stopping the augmenting when error is sufficiently small, and has accuracy closer to SVD.

A. Bilateral Random Projection based Strategy

1) Low-rank approximation with closed form: Given \( r \) bilateral random projections (BRP) of an \( m \times n \) dense matrix \( X \) (w.l.o.g, \( m \geq n \)), i.e., \( Y_1 = X A_1 \) and \( Y_2 = X^T A_2 \), wherein \( A_1 \in \mathbb{R}^{n \times r} \) and \( A_2 \in \mathbb{R}^{m \times r} \) are random matrices,

\[
L = Y_1 \left( A_1^T Y_1 \right)^{-1} Y_2^T
\]  

(5)

is a fast rank-\( r \) approximation of \( X \). The computation of \( L \) includes an inverse of an \( r \times r \) matrix and three matrix multiplications. Thus, for a dense \( X \), \( 2mn + r^2(2n + r) + mnr \) flops are required to obtain BRP, \( r^2(2n + r) + mnr \) flops are required to obtain \( L \). The computational cost is much less than SVD based approximation.

In order to improve the approximation precision of \( L \) in (5) when \( A_1 \) and \( A_2 \) are standard Gaussian matrices, we use the obtained right random projection \( Y_1 \) to build a better left projection matrix \( A_2 \), and use \( Y_2 \) to build a better \( A_1 \). In particular, after \( Y_1 = X A_1 \), we update \( A_2 = Y_1 \) and calculate the left random projection \( Y_2 = X^T A_2 \), then we update \( A_1 = Y_2 \) and calculate the right random projection \( Y_1 = X A_1 \). A better low-rank approximation \( L \) will be obtained if the new \( Y_1 \) and \( Y_2 \) are applied to (5). This improvement requires additional flops of \( mnr \) in BRP calculation.

2) Power scheme modification: When singular values of \( X \) decay slowly, (5) may perform poorly. We design a modification for this situation based on the power scheme [49]. In the power scheme modification, we instead calculate the BRP of a matrix \( \tilde{X} = (XX^T)^q X \), whose singular values decay faster than \( X \). In particular, \( \lambda_i(\tilde{X}) = \lambda_i(X)^{2q+1} \). Both \( X \) and \( \tilde{X} \) share the same singular vectors. The BRP of \( \tilde{X} \) is:

\[
Y_1 = \tilde{X} A_1, Y_2 = \tilde{X}^T A_2.
\]  

(6)
According to (5), the BRP based $r$ rank approximation of $\tilde{X}$ is:

$$\tilde{L} = Y_1 \left( A_2^T Y_1 \right)^{-1} Y_2^T. \tag{7}$$

In order to obtain the approximation of $X$ with rank $r$, we calculate the QR decomposition of $Y_1$ and $Y_2$, i.e.,

$$Y_1 = Q_1 R_1, Y_2 = Q_2 R_2. \tag{8}$$

The low-rank approximation of $X$ is then given by:

$$L = \left( \tilde{L} \right)^{\frac{1}{2q+1}} = Q_1 \left[ R_1 \left( A_2^T Y_1 \right)^{-1} R_2^T \right]^{\frac{1}{2q+1}} Q_2^T. \tag{9}$$

The power scheme modification (9) requires an inverse of an $r \times r$ matrix, an SVD of an $r \times r$ matrix and five matrix multiplications. Therefore, for dense $X$, $2(2q+1)mnr$ flops are required to obtain BRP, $r^2(m+n)$ flops are required to obtain the QR decompositions, $2r^2(n+2r) + mnr$ flops are required to obtain $L$. The power scheme modification reduces the error of (5) by increasing $q$. When the random matrices $A_1$ and $A_2$ are built from $Y_1$ and $Y_2$, $mnr$ additional flops are required in the BRP calculation. Thorough error bound analysis of BRP and its power scheme is given in Appendix II.

3) Fast GoDec by Bilateral Random Projection: Since BRP based low-rank approximation is near optimal and efficient, we replace SVD with BRP in naive GoDec in order to significantly reduce the time cost.

We summarize GoDec using BRP based low-rank approximation (5) and power scheme modification (9) in Algorithm 1. When $q = 0$, For dense $X$, (5) is applied. Thus the QR decomposition of $Y_1$ and $Y_2$ in Algorithm 1 are not performed, and $L_t$ is updated as $L_t = Y_1 \left( A_2^T Y_1 \right)^{-1} Y_2^T$. In this case, Algorithm 1 requires $r^2(2n+r) + 4mnr$ flops per iteration. When integer $q > 0$, (9) is applied and Algorithm 1 requires $r^2(m+3n+4r) + (4q+4)mnr$ flops per iteration.

B. Greedy Bilateral Factorization Strategy

The major computation in naive GoDec is the update of the low-rank part $L$, which requires at least a truncated SVD. Although the proposed randomized strategy provides a faster and SVD-free algorithm for GoDec, how to determine the rank of $L$ and the cardinality of $S$ is still an unsolved problem in real applications. In fact, these two parameters are not easy to determine and could lead to unstable solutions when estimated incorrectly. Noisy robust PCA methods such as stable PCP [28], GoDec [30] and DRMF [29] usually suffer from this problem. Another shortcoming of the randomized strategy is that the time complexity is dominated by matrix multiplications, which could be computationally slow on high-dimensional data.
Algorithm 1: GO Decomposition (GoDec) by BRP

Input: $X$, $r$, $k$, $\epsilon$, $q$
Output: $L$, $S$

Initialize $L_0 := X$, $S_0 := 0$, $t := 0$

while $\|X - L_t - S_t\|_F^2/\|X\|_F^2 > \epsilon$ do

$t := t + 1$;
$
\tilde{L} = \left( (X - S_{t-1}) (X - S_{t-1})^T \right)^q (X - S_{t-1});$

$Y_1 = \tilde{L} A_1$, $A_2 = Y_1$;

$Y_2 = \tilde{L}^T Y_1 = Q_2 R_2$, $Y_1 = \tilde{L} Y_2 = Q_1 R_1$;

If rank $\left( A_2^T Y_1 \right) < r$ then $r := \text{rank} \left( A_2^T Y_1 \right)$, go to the first step; end;

$L_t = Q_1 \left[ R_1 \left( A_2^T Y_1 \right)^{-1} R_2^T \right]^{1/(2q+1)} Q_2^T$;

$S_t = P_\Omega \left( X - L_t \right)$, $\Omega$ is the nonzero subset of the first $k$ largest entries of $|X - L_t|$;

end

In this part, we describe and analyze a general scheme called “greedy bilateral (GreB)” paradigm for solving optimizing low-rank matrix in mainstream problems. In GreB, the low-rank variable $L$ is modeled in a bilateral factorization form $UV$, where $U$ is a tall matrix and $V$ is a fat matrix. It starts from $U$ and $V$ respectively containing a very few (e.g., one) columns and rows, and optimizes them alternately. Their updates are based on observation that the object value is determined by the product $UV$ rather than individual $U$ or $V$. Thus we can choose a different pair $(U,V)$ producing the same $UV$ but computed faster than the one derived by alternating least squares like in IRLS-M [50] and ALS [51]. In GreB, the updates of $U$ and $V$ can be viewed as mutually adaptive update of the left and right sketches of the low-rank matrix. Such updates are repeated until the object convergence, then a few more columns (or rows) are concatenated to the obtained $U$ (or $V$), and the alternating updates are restarted on a higher rank. Here, the added columns (or rows) are selected in a greedy manner. Specifically, they are composed of the rank-1 column (or row) directions on which the object decreases fastest. GreB incrementally increases the rank until when $UV$ is adequately consistent with the observations.

GreB’s greedy strategy avoids the failures brought by possible biased rank estimation. Moreover, greedy selecting optimization directions from 1 to $r$ is faster than updating $r$ directions in all iterates like in LMaFit [52] and [30]. In addition, the lower rank solution before each rank increment is invoked as the “warm start” of the next higher rank optimization and thus speed up convergence. Furthermore, its mutually adaptive updates of $U$ and $V$ yields a simple yet efficient SVD-free implementation. Under GreB paradigm, the overall time complexity of matrix completion is $O(\max\{\|\Omega\|_0 r^2, (m + n) r^3\})$ ($\Omega$-sampling set, $m \times n$-matrix size, $r$-rank), while the overall complexities of low-rank approximation and noisy robust PCA are $O(mnr^2)$.
Algorithm 2: Greedy Bilateral (GreB) Paradigm

**Input:** Object function $f$; rank step size $\Delta r$; power $K$; tolerance $\tau$; observations of data matrix $X$

**Output:** low-rank matrix $UV$ and sparse $S$

**Initialize** $V \in \mathbb{R}^{r_0 \times n}$ (and $S$);

**while** residual error $\leq \tau$ **do**

**for** $k \leftarrow 1$ **to** $K$ **do**

Update $U$, $V$ and $S$ by alternating minimization rules, other faster $U$ and $V$ update rules can be applied if they produce equal $UV$;

Greedy Bilateral Smoothing: sequentially compute (13);

end

Calculate the top $\Delta r$ right singular vectors $v$ (or $\Delta r$-dimensional random projections) of $\partial f/\partial V$ (for GreBsmo compute (14)); Set $V := [V; v]$;

end

An improvement on sample complexity can also be justified. An theoretical analysis of GreB solution convergence based on the result of GECO [53] is given in Appendix III.

In the following, we present GreB by using it to derive a practical algorithm “greedy bilateral smoothing (GreBsmo)” for GoDec. It can also be directly applied to low-rank approximation and matrix completion [7]. We summarize general GreB paradigm in Algorithm 2 and then present the detailed GreBsmo algorithm.

1) Faster GoDec by Greedy Bilateral Smoothing: In particular, we formulate GoDec by replacing $L$ with its bilateral factorization $L = UV$ and regularizing the $\ell_1$ norm of $S$’s entries:

$$\min_{U,V,S} \|X - UV - S\|_F^2 + \lambda\|\text{vec}(S)\|_1$$

s.t. \hspace{1em} \text{rank}(U) = \text{rank}(V) \leq r. \hspace{1em} (10)

Note the $\ell_1$ regularization is a minor modification to the cardinality constraint in (2). It induces soft-thresholding in updating $S$, which is faster than sorting caused by cardinality constraint in GoDec and DRMF.

Alternately optimizing $U$, $V$ and $S$ in (10) immediately yields the following updating rules:

$$
\begin{align*}
U_k &= (X - S_{k-1}) V_{k-1}^T \left(V_{k-1} V_{k-1}^T\right)^\dagger, \\
V_k &= (U_k^T U_k)^\dagger U_k^T (X - S_{k-1}), \\
S_k &= S_\lambda(X - U_k V_k),
\end{align*}
$$

where $S_\lambda$ is an element-wise soft thresholding operator with threshold $\lambda$ such that

$$S_\lambda X = \{\text{sgn}(X_{ij}) \max(|X_{ij}| - \lambda, 0) : (i, j) \in [m] \times [n]\}. \hspace{1em} (12)$$
The same trick of replacing the \((U, V)\) pair with a faster computed one is applied and produce

\[
\begin{aligned}
U_k &= Q, \text{QR} \left( (X - S_{k-1}) V_{k-1}^T \right) = QR, \\
V_k &= Q^T (X - S_{k-1}), \\
S_k &= S_{\lambda} (X - U_k V_k),
\end{aligned}
\] (13)

The above procedure can be performed in \(3mnr_i + mr_i^2\) flops for \(U \in \mathbb{R}^{m \times r_i}\) and \(V \in \mathbb{R}^{r_i \times n}\).

In GreBsmo, (13) is iterated as a subroutine of GreB’s greedy incremental paradigm. In particular, the updates in (13) are iterated for \(K\) times or until the object converging, then \(\Delta r\) rows are added into \(V\) as the new directions for decreasing the object value. In order to achieve the fastest decreasing directions, we greedily select the added \(\Delta r\) rows as the top \(\Delta r\) right singular vectors of the partial derivative

\[
\frac{\partial \|X - UV - S\|_F^2}{\partial V} = X - UV - S.
\] (14)

We also allow to approximate row space of the singular vectors via random projections [25]. The selected \(\Delta r\) rows maximize the magnitude of the above partial derivative and thus lead to the most rapid decreasing of the object value, a.k.a., the decomposition error. GreBsmo repeatedly increases the rank until a sufficiently small decomposition error is achieved. So the rank of the low-rank component is adaptively estimated in GreBsmo and does not relies on initial estimation.

**IV. THREE VARIANTS OF GODEC**

Although the two strategies successfully generate efficient low-rank and sparse decomposition capable to tackle large volume problem of big data, the complicated structures widely existing in big data cannot be always expressed by the sum of low-rank and sparse matrices and thus may still lead to the failure of RPCA typed models. Therefore, we address this problem by developing several GoDec’s variants that unravel different combination of incoherent structures beyond low-rank and sparse matrices, where the two strategies can be still used to achieve scalable algorithms.

**A. Shifted Subspace Tracking (SST) for Motion Segmentation**

SST decomposes \(S\) of GoDec into the sum of several matrices, each of whose rows are generated by imposing a smooth geometric transformation sequence to the rows of a low-rank matrix. These rows store moving object in the same motion after aligning them across different frames, while the geometric transformation sequence defines the shared trajectories and deformations of those moving objects across frames. In the following, we develop an efficient randomized algorithm extracting the motions in sequel, where the low-rank matrix for each
motion is updated by BRP, and the geometric transformation sequence is updated in a piecewise linear approximation manner.

We consider the problem of motion segmentation from the raw video data. Given a data matrix \( X \in \mathbb{R}^{n \times p} \) that stores a video sequence of \( n \) frames, each of which has \( w \times h = p \) pixels and reshaped as a row vector in \( X \), the goal of SST framework is to separate the motions of different object flows, recover both their low-rank patterns and geometric transformation sequences. This task is decomposed as two steps, background modeling that separates all the moving objects from the static background, and flow tracking that recovers the information of each motion. In this problem, \( \cdot_i \) stands for the \( i \)th entry of a vector or the \( i \)th row of a matrix, while \( \cdot_{i,j} \) signifies the entry at the \( i \)th row and the \( j \)th column of a matrix.

The first step can be accomplished by either GoDec or GreBsmo. After obtaining the sparse outliers \( S \) storing multiple motions, SST treats the sparse matrix \( S \) as the new data matrix \( X \), and decomposes it as \( X = \sum_{i=1}^{k} \tilde{L}(i) + S + G \), wherein \( \tilde{L}(i) \) denotes the \( i \)th motion, \( S \) stands for the sparse outliers and \( G \) stands for the Gaussian noise.

The motion segmentation in SST is based on an observation to the implicit structures of the sparse matrix \( \tilde{L}(i) \). If the trajectory of the object flow \( \tilde{L}(i) \) is known and each frame (row) in \( \tilde{L}(i) \) is shifted to the position of a reference frame, due to the limited number of poses for the same object flow in different frames, it is reasonable to assume that the rows of the shifted \( \tilde{L}(i) \) exist in a subspace. In other words, \( \tilde{L}(i) \) after inverse geometric transformation is low-rank. Hence the sparse motion matrix \( \tilde{L}(i) \) has the following structured representation

\[
\tilde{L}(i) = \begin{bmatrix}
L(i)_1 \circ \tau(i)_1 \\
\vdots \\
L(i)_n \circ \tau(i)_n
\end{bmatrix} = L(i) \circ \tau(i).
\]

The invertible transformation \( \tau(i)_j : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \) denotes the 2-D geometric transformation (to the reference frame) associated with the \( i \)th motion in the \( j \)th frame, which is represented by \( L(i)_j \). To be specific, the \( j \)th row in \( \tilde{L}(i) \) is \( L(i)_j \) after certain permutation of its entries. The permutation results from applying the nonlinear transformation \( \tau(i)_j \) to each nonzero pixel in \( L(i)_j \) such that,

\[
\tau(i)_j(x, y) = (u, v),
\]

where \( \tau(i)_j \) could be one of the five geometric transformations \([54]\), i.e., translation, Euclidean, similarity, affine and homography, which are able to be represented by 2, 3, 4, 6 and 9 free parameters, respectively. For example, affine transformation is defined as

\[
\begin{bmatrix}
u \\
v
\end{bmatrix} = \begin{bmatrix}
\rho \cos \theta & \rho \sin \theta \\
-\rho \sin \theta & \rho \cos \theta
\end{bmatrix} \begin{bmatrix}
x \\
y
\end{bmatrix} + \begin{bmatrix}
t_x \\
t_y
\end{bmatrix},
\]

May 11, 2014 DRAFT
wherein $\theta$ is the rotation angle, $t_x$ and $t_y$ are the two translations and $\rho$ is the scaling ratio. It is worth to point out that $\tau(i)_j$ can be any other transformation beyond the geometric group. So SST can be applied to sparse structure in other applications if parametric form of $\tau(i)_j$ is known. We define the nonlinear operator $\circ$ as

$$ \tilde{L}(i)_{j,u+(v-1)h} = (L(i)_j \circ \tau(i)_j)_{u+(v-1)h} = L(i)_{j,x+(y-1)h}. $$

(18)

Therefore, the flow tracking in SST aims at decomposing the sparse matrix $X$ (obtained in the background modeling) as

$$ X = \sum_{i=1}^{k} L(i) \circ \tau(i) + S + G, $$

$$ \text{rank}(L(i)) \leq r_i, \text{card}(S) \leq s. $$

(19)

In SST, we iteratively invoke $k$ times of the following matrix decomposition to greedily construct the decomposition in (19):

$$ X = L \circ \tau + S + G, \text{rank}(L) \leq r, \text{card}(S) \leq s. $$

(20)

In each time of the matrix decomposition above, the data matrix $X$ is $S$ obtained by former decomposition. In order to save the computation and facilitate the parameter tuning, we cast the decomposition (20) into an optimization similar to (2),

$$ \min_{L,\tau,S} \|X - L \circ \tau - S\|_F^2 + \lambda\|S\|_1 $$

$$ \text{s.t.} \ \text{rank}(L) \leq r, $$

(21)

Flow tracking in SST solves a sequence of optimization problem of type (21). Thus we firstly apply alternating minimization to (21). This results in iterative update of the solutions to the following three subproblems,

$$ \begin{cases} 
\tau^t = \arg \min_{\tau} \|X - L^{t-1} \circ \tau - S^{t-1}\|_F^2; \\ 
L^t = \arg \min_{\text{rank}(L) \leq r} \|X - L \circ \tau^t - S^{t-1}\|_F^2; \\ 
S^t = \arg \min_{S} \|X - L^t \circ \tau^t - S\|_F^2 + \lambda\|S\|_1. 
\end{cases} $$

(22)

1) **Update of $\tau$:** The first subproblem aims at solving the following series of nonlinear equations of $\tau_j$,

$$ L^{t-1}_j \circ \tau_j = X_j - S^{t-1}_j, j = 1, \ldots, n. $$

(23)

Albeit directly solving the above equation is difficult due to its strong nonlinearity, we can approximate the geometric transformation $L^{t-1}_j \circ \tau_j$ by using piece-wise linear transformations, where each piece corresponds to a small change of $\tau_j$ defined by $\Delta \tau_j$. Thus the solution of (23)
can be approximated by accumulating a series of $\Delta \tau_j$. This can be viewed as an inner loop included in the update of $\tau$. Thus we have linear approximation

$$L_j^{t-1} \circ (\tau_j + \Delta \tau_j) \approx L_j^{t-1} \circ \tau_j + \Delta \tau_j J_j,$$

(24)

where $J_j$ is the Jacobian of $L_j^{t-1} \circ \tau_j$ with respect to the transformation parameters in $\tau_j$. Therefore, by substituting (24) into (23), $\Delta \tau_j$ in each linear piece can be solved as

$$\Delta \tau_j = (X_j - S_j^{t-1} - L_j^{t-1} \circ \tau_j) (J_j)^{+}.$$

(25)

The update of $\tau_j$ starts from some initial $\tau_j$, and iteratively solves the overdetermined linear equation (25) with update $\tau_j := \tau_j + \Delta \tau_j$ until the difference between the left hand side and the right hand side of (23) is sufficiently small. It is critical to emphasize that a well selected initial value of $\tau_j$ can significantly save computational time. Based on the between-frame affinity, we initialize $\tau_j$ by the transformation of its adjacent frame that is closer to the template frame $s$,

$$\tau_j := \begin{cases} 
\tau_{j+1}, & j < s; \\
\tau_{j-1}, & j > s.
\end{cases}$$

(26)

Another important support set constraint, $\text{supp}(L \circ \tau) \subseteq \text{supp}(X)$, needs to be considered in calculating $L_j^{t-1} \circ \tau_j$ during the update of $\tau$. This constraint ensures that the object flows or segmented motions obtained by SST always belong to the sparse part achieved from the background modeling, and thus rules out the noise in background. Hence, suppose the complement set of $\text{supp}(X_j)$ to be $\text{supp}_c(X_j)$, each calculation of $L_j^{t-1} \circ \tau_j$ follows a screening such that,

$$(L_j^{t-1} \circ \tau_j)_{\text{supp}_c(X_j)} = \overrightarrow{0}.$$

(27)

2) Update of $L$: The second subproblem has the following global solution that can be updated by BRP based low-rank approximation (5) and its power scheme modification,

$$L^t = \sum_{i=1}^r \lambda_i U_i V_i^T, \text{svd} \left( (X - S_t^{t-1}) \circ \tau^{-1} \right) = U A V^T,$$

(28)

wherein $\tau^{-1}$ denotes the inverse transformation towards $\tau$. The SVDs can be accelerated by BRP based low-rank approximation [2]. Another acceleration trick is based on the fact that most columns of $(X - S_t^{t-1}) \circ \tau^{-1}$ are nearly all-zeros. This is because the object flow or motion after transformation occupies a very small area of the whole frame. Therefore, The update of $L^t$ can be reduced to low-rank approximation of a submatrix of $(X - S_t^{t-1}) \circ \tau^{-1}$ that only includes dense columns. Since the number of dense columns is far less than $p$, the update of $L^t$ can become much faster.
3) Update of $S$: The third subproblem has a global solution that can be obtained via soft-thresholding $\mathcal{P}_\lambda(\cdot)$ similar to the update of $S$ in GreBsmo,

$$S^t = \mathcal{P}_\lambda \left( X - L^t \circ \tau^t \right).$$ (29)

**Algorithm 3:** Shifted Subspace Tracking (SST)

| Input: $X$, $\tau_i, \lambda_i(i = 1, \cdots, n)$, $k$ |
| Output: $L_i(i = 1, \cdots, n)$, $S$ |
| for $i \leftarrow 1$ to $k$ do |
| Initialize: $s = \arg \max_i \text{card}(X_i)$; $L = [X_1; \cdots; X_n]$, $S = 0$, $\tau = \overrightarrow{0}$; |
| while not converge do |
| for $j \leftarrow s - 1$ to 1 do |
| $\tau_j := \tau_j + 1$; |
| while not converge do |
| $\tilde{L}_{j}^{t-1} = L_{j}^{t-1} \circ \tau_j$, $\tilde{L}_{j,\text{supp}(X_j)}^{t-1} = \overrightarrow{0}$; |
| $\tau_j := \tau_j + \left( X_j - S_{j}^{t-1} - \tilde{L}_{j}^{t-1} \right) (J_j)^\dagger$; |
| end |
| end |
| for $j \leftarrow s + 1$ to $n$ do |
| $\tau_j := \tau_j - 1$; |
| while not converge do |
| $\tilde{L}_{j}^{t-1} = L_{j}^{t-1} \circ \tau_j$, $\tilde{L}_{j,\text{supp}(X_j)}^{t-1} = \overrightarrow{0}$; |
| $\tau_j := \tau_j + \left( X_j - S_{j}^{t-1} - \tilde{L}_{j}^{t-1} \right) (J_j)^\dagger$; |
| end |
| end |
| $\tau^t = \tau$; |
| $L^t = \text{BRP} \left( (X - S^{t-1}) \circ \tau^{-1} \right)$; |
| $S^t = \mathcal{P}_\lambda \left( X - L^t \circ \tau^t \right)$, $S_{j,\text{supp}(X_j)}^t = \overrightarrow{0}$; |
| end |
| $X := S^t$, $L(i) := L^t$, $\tau(i) = \tau^t$; |
| end |

A support set constraint $\text{supp}(S) \subseteq \text{supp}(X)$ should be considered in the update of $S$ as well. Hence the above update follows a postprocessing,

$$S_{j,\text{supp}(X_j)}^t = \overrightarrow{0}, j = 1, \cdots, n.$$ (30)

Note the transformation computation $\circ$ in the update can be accelerated by leveraging the sparsity of the motions. Specifically, the sparsity allows SST to only compute the transformed positions of the nonzero pixels. We summarize the SST algorithm in Algorithm 3.

B. Multi-label Subspace Ensemble

MSE provides a novel insight into the multi-label learning (ML) problem, which aims at predicting multiple labels of a data sample. Most previous ML methods [55], [56], [57], [58],...
[59], [60] focus on training effective classifiers that establishes a mapping from feature space to label space, and take the label correlation into account in the training process. Because it has been longly believed that label correlation is useful for improving prediction performance. However, in these methods, both the label space and the model complexity will grow rapidly when increasing the number of labels and simultaneously modeling their joint correlations. This usually makes the available training samples insufficient for learning a joint prediction model.

MSE eliminates this problem by jointly learning inverse mappings that map each label to the feature space as a subspace, and formulating the prediction as finding the group sparse representation [46] of a given sample on the ensemble of subspaces. In the training stage, the training data matrix $X$ is decomposed as the sum of several low-rank matrices and a sparse residual via a randomized optimization. Each low-rank part defines a subspace mapped by a label, and its rows are nonzero only when the corresponding samples are annotated by the label. The sparse part captures the rest contents in the features that cannot be explained by the labels.

1) MSE training: randomized decomposition: The training stage of MSE approximately decomposes the training data matrix $X \in \mathbb{R}^{n \times p}$ into $X = \sum_{i=1}^{k} L^i + S$. For the matrix $L^i$, the rows corresponding to the samples with label $i$ are nonzero, while the other rows are all-zero vectors. The nonzero rows denote the components explained by label $i$ in the feature space. We use $\Omega_i$ to denote the index set of samples with label $i$ in the matrix $X$ and $L^i$, and then the matrix composed of the nonzero rows in $L^i$ is represented by $L^i_{\Omega_i}$. In the decomposition, the rank of $L^i_{\Omega_i}$ is upper bounded, which indicates that all the components explained by label $i$ nearly lies in a linear subspace. The matrix $S$ is the residual of the samples that cannot be explained by the given labels. In the decomposition, the cardinality of $S$ is upper bounded, which makes $S$ sparse.

If the label matrix of $X$ is $Y \in \{0, 1\}^{n \times k}$, the rank of $L^i_{\Omega_i}$ is upper bounded by $r^i$ and the cardinality of $S$ is upper bounded by $K$, the decomposition can be written as solving the following constrained minimization problem:

$$\min_{L^i, S} \left\| X - \sum_{i=1}^{k} L^i - S \right\|_F^2$$

s.t. $\text{rank} (L^i_{\Omega_i}) \leq r^i, L^i_{\Omega_i} = 0, \forall i = 1, \ldots, k$

$$\text{card} (S) \leq K. \quad (31)$$

Therefore, each training sample in $X$ is decomposed as the sum of several components, which respectively correspond to multiple labels that the sample belongs to. MSE separates these components from the original sample by building the mapping from the labels to the feature space. For label $i$, we obtain its mapping in the feature space as the row space of $L^i_{\Omega_i}$.

Although the rank constraint to $L^i_{\Omega_i}$ and cardinality constraint to $S$ are not convex, the
optimization in (31) can be solved by alternating minimization that decomposes it as the following \( k + 1 \) subproblems, each of which has the global solution:

\[
\begin{aligned}
L_{\Omega_i}^i &= \arg \min_{\text{rank}(L_{\Omega_i}^i) \leq r^i} \left\| X - \sum_{j=1, j \neq i}^k L_j^i - S^i \right\|_F^2, \\
forall i &= 1, \ldots, k, \\
S &= \arg \min_{\text{card}(S) \leq K} \left\| X - \sum_{j=1}^k L_j^i - S \right\|_F^2.
\end{aligned}
\]  

The solutions of \( L_{\Omega_i}^i \) and \( S \) in the above subproblems can be obtained via hard thresholding of singular values and the matrix entries, respectively. Note that both SVD and matrix entry-wise hard thresholding have global solutions. In particular, \( L_{\Omega_i}^i \) is built from the first \( r^i \) largest singular values and the corresponding singular vectors of \( \left( X - \sum_{j=1, j \neq i}^k L_j^i - S \right)_{\Omega_i} \), while \( S \) is built from the \( K \) entries with the largest absolute value in \( X - \sum_{j=1}^k L_j^i \), i.e.,

\[
\begin{aligned}
L_{\Omega_i}^i &= \sum_{q=1}^{r^i} \lambda_q U_q V_q^T, i = 1, \ldots, k, \\
\text{svd} \left[ \left( X - \sum_{j=1, j \neq i}^k L_j^i - S \right)_{\Omega_i} \right] &= U \Lambda V^T; \\
S &= \mathcal{P}_{\Phi} \left( X - \sum_{j=1}^k L_j^i \right), \Phi : \left( X - \sum_{j=1}^k L_j^i \right)_{r,s \in \Phi} \neq 0 \\
\text{and } \geq \left| \left( X - \sum_{j=1}^k L_j^i \right)_{r,s \in \Phi} \right|, |\Phi| \leq K.
\end{aligned}
\]

The projection \( S = \mathcal{P}_{\Phi}(R) \) represents that the matrix \( S \) has the same entries as \( R \) on the index set \( \Phi \), while the other entries are all zeros.

The decomposition is then obtained by iteratively solving these \( k + 1 \) subproblems in \((32)\) according to \((33)\). In this problem, we initialize \( L_{\Omega_i}^i \) and \( S \) as

\[
\begin{aligned}
L_{\Omega_i}^i := Z_{\Omega_i}, i = 1, \ldots, k, \\
Z = D^{-1}X, D = \text{diag}(Y) \; ; \\
S := 0.
\end{aligned}
\]

In each subproblem, only one variable is optimized with the other variables fixed. Similar to GoDec, BRP based acceleration strategy can be applied to the above model and produces the practical training algorithm in Algorithm [4]

In the training, the label correlations is naturally preserved in the subspace ensemble, because all the subspaces are jointly learned. Since only \( k \) subspaces are learned in the training stage,
MSE explores label correlations without increasing the model complexity.

Algorithm 4: MSE Training

Input: $X, \Omega_i, r^i, i = 1, \ldots, k, K, \epsilon$
Output: $C^i, i = 1, \ldots, k$
Initialize $L^i$ and $S$ according to (34), $t := 0$;

while $\|X - \sum_{j=1}^{k} L^j - S\|_F^2 > \epsilon$ do

$t := t + 1$;

for $i \leftarrow 1$ to $k$ do

$N := (X - \sum_{j=1, j \neq i}^{k} L^j - S)_{\Omega_i}$;
Generate standard Gaussian matrix $A_1 \in \mathbb{R}^{p \times r^i}$;
$Y_1 := N A_1, Y_2 := N^T Y_1, Y_1 := N Y_2$;
$L_{\Omega_i} := Y_1 (A_2^T Y_1)^{-1} Y_2^T, L_{\Omega_i} := 0$;

end

$N := |X - \sum_{j=1}^{k} L^j|$;
$S := \mathcal{P}\Phi (N), \Phi$ is the index set of the first $K$ largest entries of $|N|$;

end

QR decomposition $(L_{\Omega_i})^T = Q^i R^i$ for $i = 1, \ldots, k, C^i := (Q^i)^T$;

2) MSE prediction: group sparsity: In the prediction stage of MSE, we use group lasso \cite{46, 61} to estimate the group sparse representation $\beta \in \mathbb{R}^{\sum r^i}$ of a test sample $x \in \mathbb{R}^p$ on the subspace ensemble $C = [C^1; \ldots; C^k]$, wherein the $k$ groups are defined as index sets of the coefficients corresponding to $C^1, \ldots, C^k$. Since group lasso selects nonzero coefficients groupwisely, nonzero coefficients in the group sparse representation will concentrate on the groups corresponding to the labels that the sample belongs to.

According to the above analysis, we solve the following group lasso problem in the prediction stage of MSE

$$\min_{\beta} \frac{1}{2} \|x - \beta C\|_F^2 + \lambda \sum_{i=1}^{k} \|\beta_{G_i}\|_2,$$  \hspace{1cm} (35)

where the index set $G_i$ includes all the integers between $1 + \sum_{j=1}^{i-1} r^j$ and $\sum_{j=1}^{i} r^j$ (including these two).

To obtain the final prediction of the label vector $y \in \{0, 1\}^k$ for a test sample $x$, we use a simple thresholding of the magnitude sum of coefficients in each group to test which groups that the sparse coefficients in $\beta$ concentrate on

$$y_{\Psi} = 1, y_{\Psi} = 0, \Psi = \{i : \|\beta_{G_i}\|_1 \geq \delta\}.$$  \hspace{1cm} (36)

Although $y$ can also be obtained via selecting the groups with nonzero coefficients when $\lambda$ in (35) is chosen properly, we set the threshold $\delta$ as a small positive value to guarantee the robustness to $\lambda$. 

May 11, 2014 DRAFT
C. Linear Functional GoDec for Learning Recommendation System

Although low-rank matrix completion provides an effective and simple mathematical model predicting a user’s rating to an item from her/his ratings to other items and the ratings of other users by exploring the user relationships, a primary problem of this model is that adding a new item or a new user to the model requires an new optimization of the whole low-rank rating matrix, which is not practical due to its expensive time cost. Moreover, although the attributes of users are always missing in real recommendation systems, features of the items have been proved to be helpful side information that is much easier to obtain. But previous matrix completion methods and GoDec cannot leverage this information in their models. Furthermore, robust rating prediction should allow advertising effects in known ratings.

In this part, we propose a variant of GoDec called “linear functional GoDec (LinGoDec)”. It formulates the collaborative filtering problem as supervised learning, and avoids time-consuming completion of the whole matrix when only a new item’s scores (a new row) are needed to be predicted. In particular, LinGoDec decomposes rating matrix $X$ whose rows index the users, columns index the items, and entries denote the scores of items given by different users. Given the features of some items, which are usually available, and the ratings of these items scored by all users, LinGoDec learns a scoring function for each user so that efficient prediction of ratings can be made item-wisely. It studies the case when the scoring functions of different users are linear and related to each other. In the mode, it replaces the low-rank part $L$ of GoDec with $WZ^T$, where $W$ represents the linear related functions and the rows of $Z$ are items represented by features. The sparse part $S$ is able to capture the advertising effects or anomaly of users’ ratings on specific items, which cannot be represented by the low-rank scoring functions. In the algorithm of LinGoDec, the update of low-rank $W$ is accomplished by invoking an elegant closed-form solution for least square rank minimization [51], which could be accelerated by BRP.

LinGoDec aims at solving the following optimization,

$$\min_{W,S} \|X - WZ^T - S\|^2_F + \lambda \|\text{vec}(S)\|_1$$

s.t. $\text{rank}(W) \leq r$. (37)

We constrain $W$ to be low-rank so that the functions of different users share the same small set of basis functions. In addition, we apply $\ell_1$ regularization to the entries of $S$ so that the advertising effects in training ratings can be captured and ruled out from the learning of $W$. By applying alternating minimization to (37), we have

$$\begin{cases} W_k = \arg\min_W \|X - WZ^T\|^2_F \quad \text{s.t.} \quad \text{rank}(W) \leq r, \\ S_k = S\lambda (X - W_kZ^T), \end{cases}$$

(38)
The update of $W_k$ in above procedures equals to solve a least squares rank minimization, which has been discovered owning closed-form solution that can be obtained by truncated SVD [1] when $X$ is singular (the most common case in our problem). By applying bilateral random projection based acceleration to the truncated SVD, we immediately achieve the final fast algorithm for LinGoDec. LinGoDec has a similar model as rank-regularized multi-task learning, but the major difference is that the sparse matrix in LinGoDec is a component of the data matrix rather than the linear functions $W$.

V. Experiments

This section evaluates both the effectiveness and the efficiency of all the algorithms proposed in this paper, and compares them with state-of-the-art rivals. We will show experimental results of GoDec and GreBsmo on both surveillance video sequences for background modeling and synthetic data. Then we will apply SST, MSE and LinGoDec to the problems of motion segmentation, multi-label learning and collaborative filtering. We run all the experiments in MatLab on a server with dual quad-core 3.33 GHz Intel Xeon processors and 32 GB RAM. The relative error $\frac{\|X - \hat{X}\|_F}{\|X\|_F}$ is used to evaluate the effectiveness, wherein $X$ is the original matrix and $\hat{X}$ is an estimate/approximation.

| size($X$) | rank($L$) | card($S$) | rel.error($X$) | rel.error($L$) | rel.error($S$) | time (seconds) |
|----------|-----------|-----------|----------------|----------------|----------------|----------------|
| (square) | (1)       | ($10^4$)  | ($10^{-8}$)    | ($10^{-8}$)    | ($10^{-6}$)    |                |
| 500      | 25        | 1.25      | 3.70/1.80      | 1.50/1.20      | 2.00/0.95      | 6.07/2.83      |
| 1000     | 50        | 5.00      | 4.98/4.56      | 1.82/1.85      | 5.16/4.90      | 20.96/12.71    |
| 2000     | 100       | 20.0      | 8.80/1.13      | 3.10/1.10      | 1.81/1.24      | 101.74/74.16   |
| 3000     | 250       | 45.0      | 6.29/4.98      | 5.09/5.05      | 33.9/55.3      | 562.09/266.11  |
| 5000     | 400       | 125       | 63.1/24.4      | 30.2/29.3      | 54.2/18.8      | 2495.31/840.39 |
| 10000    | 500       | 600       | 6.18/3.04      | 2.27/2.88      | 58.3/36.6      | 9560.74/3030.15|

A. GoDec on Synthetic Data

We compare the relative errors and time costs of Robust PCA and GoDec on square matrices with different sizes, different ranks of low-rank components and different cardinality of sparse components. For a matrix $X = L + S + G$, its low-rank component is built as $L = AB$, wherein both $A$ and $B$ are $n \times r$ standard Gaussian matrices. Its sparse part is built as $S = \mathcal{P}_\Omega(D)$, wherein $D$ is a standard Gaussian matrix and $\Omega$ is an entry set of size $k$ drawn uniformly at
random. Its noise part is built as \( G = 10^{-3} \cdot F \), wherein \( F \) is a standard Gaussian matrix. In our experiments, we compare RPCA \( \text{inexact alm rpca} \) with GoDec (Algorithm \( \text{1} \) with \( q = 2 \)). Since both algorithms adopt the relative error of \( X \) as the stopping criterion, we use the same tolerance \( \epsilon = 10^{-7} \). Table \( \text{1} \) shows the results and indicates that both algorithms are successful in recovering the correct “low-rank+sparse” decompositions with relative error less than \( 10^{-6} \). GoDec usually produces less relative error with much less CPU seconds than RPCA. The improvement of accuracy is due to that the model of GoDec in \( \text{1} \) is more general than that of RPCA by considering the noise part. The improvement of speed is due to that BRP based low-rank approximation significantly saves the computation of each iteration round.

B. GoDec for Background Modeling

![GoDec for Background Modeling](image)

Fig. 1. Background modeling results of four 200-frame surveillance video sequences in \( X = L + S \) mode. Top left: lobby in an office building (resolution 128 \( \times \) 160, learning time 39.75 seconds). Top right: shopping center (resolution 256 \( \times \) 320, learning time 203.72 seconds). Bottom left: Restaurant (resolution 120 \( \times \) 160, learning time 36.84 seconds). Bottom right: Hall of a business building (resolution 144 \( \times \) 176, learning time 47.38 seconds).

Background modeling \( \text{62} \) is a challenging task to reveal the correlation between video frames, model background variations and foreground moving objects. A video sequence satisfies the low-rank+sparse structure, because backgrounds of all the frames are related, while the variation and the moving objects are sparse and independent. We apply GoDec (Algorithm \( \text{1} \) with \( q = 2 \)) to four surveillance videos \( \text{1} \) respectively. The matrix \( X \) is composed of the first 200 frames of each

\( \text{1} \)http://watt.csl.illinois.edu/~perceive/matrix-rank

\( \text{2} \)http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html
video. For example, the second video is composed of 200 frames with the resolution $256 \times 320$, we convert each frame as a vector and thus the matrix $X$ is of size $81920 \times 200$. We show the decomposition result of one frame in each video sequence in Figure 1. The background and moving objects are precisely separated (the person in $L$ of the fourth sequence does not move throughout the video) without losing details. The results of the first sequence and the fourth sequence are comparable with those shown in [19]. However, compared with RPCA (36 minutes for the first sequence and 43 minutes for the fourth sequence) [19], GoDec requires around 50 seconds for each of both. Therefore, GoDec makes large-scale applications available.

C. GoDec for Shadow/Light removal

Shadow and light in training images always pull down the quality of learning in computer vision applications. GoDec can remove the shadow/light noises by assuming that they are sparse and the rest parts of the images are low-rank. We apply GoDec (Algorithm 1 with $q = 2$) to face images of four individuals in the Yale B database. Each individual has 64 images with resolution $192 \times 168$ captured under different illuminations. Thus the matrix $X$ for each individual is of size $32760 \times 64$. We show the GoDec of eight example images (2 per individual) in Figure 2. The real face of each individual are remained in the low rank component, while the shadow/light noises are successfully removed from the real face images and stored in the sparse component. The learning time of GoDec for each individual is less than 30 seconds, which encourages for large-scale applications, while RPCA requires around 685 seconds.

\begin{center}
\includegraphics[width=\textwidth]{fig2.png}
\end{center}

Fig. 2. Shadow/light removal of face images from four individuals in Yale B database in $X = L + S$ mode. Each individual has 64 images with resolution $192 \times 168$ and needs 24 seconds learning time.

D. GreBsmo on Synthetic Data

We report the phase diagram of GreBsmo in Figure 5 from results on randomly generated matrix that is the sum of a low-rank part and a sparse part. The low-rank part is generated

\footnote{http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html}
as the product of two Gaussian matrices and the sparse part has a Bernoulli model generated support set on which ±1 values are randomly assigned. The phase transition phenomenon is in consistency with existing low-rank and sparse decomposition algorithms. It also shows that GreBsmo is able to gain accurate separation of \( L \) even if its rank is close to \( 0.4n \), given the sparse part has an adequately sparse support set. This is competitive to published result [19]. Interestingly, the phase transition curve has a regular shape and implies a theoretical analysis to its behavior is highly possible in future studies.

### TABLE II

**Comparison of time costs in CPU seconds of PCP, GoDec and GreBsmo in low-rank and sparse matrix decomposition task on background modeling datasets.**

|        | PCP   | GoDec | GreBsmo |
|--------|-------|-------|---------|
| Hall   | 87s   | 56s   | 1.13s   |
| ShoppingMall | 351s | 266s  | 3.29s   |
| Bootstrap | 71s  | 49s   | 0.98s   |

### E. GreBsmo for Background Modeling

For real data, three robust PCA algorithms, i.e., inexact augmented Lagrangian multiplier method for PCP, GoDec and GreBsmo are applied to separate the low-rank background and
sparse moving objects in 3 video sequences from the same dataset used in GoDec experiment above. We show the robust PCA decomposition results of one frame for each video sequence obtained by GreBsmo in the left plot of Figure V-D. The time costs for all the three methods are listed in Table [II]. It shows GreBsmo considerably speed up the decomposition and performs 30-100 times faster than most existing algorithms.

F. SST for Motion Segmentation

We evaluate SST by using it to track object flows in four surveillance video sequences from the same dataset. In these experiments, the type of geometric transformation $\tau$ is simply selected as translation. The detection, tracking and segmentation results as well as associated time costs are shown in Figure V-F.

![Fig. 4. Background modeling and object flow tracking results of a 50-frame surveillance video sequence from Hall dataset with resolution $144 \times 176$ (left), and Shoppingmall dataset with resolution $256 \times 320$ (right).](image)

The results show SST can successfully recover both the low-rank patterns and the associated geometric transformations for motions of multiple object flows from the sparse component achieved by GoDec. The detection, tracking and segmentation are seamlessly unified in a matrix factorization framework and achieved with high accuracy. Moreover, it also verifies that SST performs significantly robust on complicated motions in complex scenes. This is attributed to...
their distinguishing shifted low-rank patterns, because different object flows can hardly share a subspace after the same geometric transformation. Since SST show stable and appealing performance in motion detection, tracking and segmentation for either crowd or individual, it provides a more semantic and intelligent analysis to the video content than existing methods.

G. MSE for Multi-label Learning

We evaluate MSE on 13 benchmark datasets from different domains and of different scales, including Corel5k (image), Scene (image), Mediamill (video), Enron (text), Genbase (genomics), Medical (text), Emotions (music), Slashdot (text) and 5 sub datasets selected in Yahoo dataset (web data). These datasets were obtained from Mulan’s website and MEKA’s website. They were collected from different practical problems.

We compare MSE with BR [43], ML-KNN [63] and MDDM [57] on four evaluation metrics for evaluating the effectiveness, as well as the CPU seconds for evaluating the efficiency. In multi-label prediction, four metrics, which are precision, recall, F1 score and accuracy, are used to measure the prediction performance. The detailed definitions of these metrics are given in Section 7.1.1 of [42]. A fair evaluation of prediction performance should include integrative consideration of all the four metrics, whose importances can be roughly given by $F_1, Acc > \{Prec, Rec\}$.

We show the prediction performance and time cost in CPU seconds of BR, ML-KNN, MDDM and MSE in Table IV and Table III. In BR, we use the MatLab interface of LIBSVM 3.0 to train the classic linear SVM classifiers for each label. The parameter $C \in \{10^{-3}, 10^{-2}, 0.1, 1, 10, 10^2, 10^3\}$ with the best performance on the training set was used. In ML-KNN, the number of neighbors was 30 for all the datasets.

In MDDM, the regularization parameter for uncorrelated subspace dimensionality reduction was selected as 0.12 and the dimension of the subspace was set as 20% of the dimension of the original data. In MSE, we selected $r^j$ as an integer in $[1, 6]$, $K \in [10^{-6}, 10^{-3}]$, $\lambda \in [0.2, 0.45]$ and $\delta \in [10^{-4}, 10^{-2}]$. We roughly selected 4 groups of parameters in the ranges for each dataset and chose the one with the best performance on the training data. Group lasso in MSE is solved by SLEP [61] in our experiments.

The experimental results show that MSE is competitive on both speed and prediction performance, because it explores label correlations and structure without increasing the problem size. In addition, the bilateral random projections further accelerate the computation. In particular,
its training time increases much more slowly than other methods, so it is more efficient when applied to large scale datasets such as Mediamill, Arts and Education. MDDM is faster than MSE on a few datasets because MDDM invokes ML-knn on the data after dimension reduction, while MSE is directly applicable to the original high dimensional data.

TABLE III
Prediction performances (%) and CPU seconds of BR [43], ML-KNN [63], MDDM [57] and MSE on Yahoo. Prec-precision, Rec-recall, F1-F1 score, Acc-accuracy

| Methods  | Prec | Rec | F1 | Acc | CPU sec. |
|----------|------|-----|----|-----|----------|
| BR       | 76   | 25  | 26 | 24  | 46.8     |
| ML-knn   | 62   | 7   | 25 | 6   | 77.6     |
| MDDM     | 68   | 6   | 21 | 5   | 37.4     |
| MSE      | 35   | 40  | 31 | 28  | 11.7     |
| BR       | 69   | 27  | 28 | 26  | 50.1     |
| ML-knn   | 58   | 6   | 31 | 5   | 99.8     |
| MDDM     | 59   | 5   | 26 | 5   | 45.2     |
| MSE      | 41   | 35  | 32 | 29  | 12.6     |
| BR       | 84   | 23  | 23 | 22  | 53.2     |
| ML-knn   | 70   | 9   | 23 | 8   | 112      |
| MDDM     | 66   | 7   | 18 | 6   | 41.9     |
| MSE      | 41   | 49  | 36 | 30  | 19.1     |
| BR       | 79   | 19  | 19 | 19  | 84.9     |
| ML-knn   | 59   | 4   | 20 | 4   | 139      |
| MDDM     | 66   | 4   | 19 | 4   | 53.0     |
| MSE      | 31   | 39  | 29 | 26  | 20.1     |
| BR       | 87   | 74  | 76 | 71  | 28.9     |
| ML-knn   | 68   | 9   | 70 | 8   | 93.2     |
| MDDM     | 66   | 7   | 69 | 7   | 42.7     |
| MSE      | 84   | 82  | 78 | 78  | 13.5     |

TABLE IV
Prediction performances (%) and CPU seconds of BR [43], ML-KNN [63], MDDM [57] and MSE on 8 datasets. Prec-precision, Rec-recall, F1-F1 score, Acc-accuracy

| Methods  | Prec | Rec | F1 | Acc | CPU sec. |
|----------|------|-----|----|-----|----------|
| BR       | 69   | 35  | 43 | 33  | 120141   |
| ML-knn   | 41   | 6   | 54 | 5   | 5713     |
| MDDM     | 36   | 5   | 53 | 4   | 48237    |
| MSE      | 58   | 78  | 53 | 37  | 1155     |
| BR       | 51   | 28  | 35 | 24  | 77.1     |
| ML-knn   | 51   | 7   | 46 | 5   | 527      |
| MDDM     | 50   | 8   | 49 | 7   | 29       |
| MSE      | 44   | 50  | 40 | 28  | 271      |
| BR       | 2    | 26  | 5  | 2   | 4.88     |
| ML-knn   | 75   | 7   | 48 | 6   | 22.8     |
| MDDM     | 74   | 3   | 30 | 2   | 32.3     |
| MSE      | 36   | 90  | 45 | 26  | 7.5      |
| BR       | 11   | 22  | 14 | 10  | 140      |
| ML-knn   | 71   | 10  | 31 | 8   | 708      |
| MDDM     | 39   | 1   | 4  | 1   | 114      |
| MSE      | 38   | 61  | 37 | 27  | 175      |
| BR       | 55   | 67  | 66 | 63  | 4.19     |
| ML-knn   | 78   | 62  | 69 | 54  | 14.3     |
| MDDM     | 75   | 64  | 69 | 53  | 7.59     |
| MSE      | 61   | 85  | 70 | 68  | 3.62     |
| BR       | 55   | 53  | 51 | 42  | 0.68     |
| ML-knn   | 68   | 28  | 41 | 22  | 0.66     |
| MDDM     | 54   | 28  | 41 | 22  | 0.66     |
| MSE      | 40   | 100 | 52 | 37  | 0.01     |
| BR       | 5    | 39  | 9  | 5   | 1.99     |
| ML-knn   | 100  | 50  | 92 | 50  | 9.38     |
| MDDM     | 98   | 51  | 92 | 51  | 6.09     |
| MSE      | 83   | 96  | 86 | 70  | 8.62     |
| BR       | 2    | 20  | 4  | 2   | 2240     |
| ML-knn   | 62   | 1   | 3  | 0.9 | 2106     |
| MDDM     | 62   | 1   | 7  | 1   | 458      |
| MSE      | 9    | 11  | 8  | 5   | 1054     |
In the comparison of performance via the four metrics, the F1 score and accuracy of MSE outperform those of other methods on most datasets. Moreover, MSE has smaller gaps between precision and recall on different tasks than other methods, and this implies it is robust to the imbalance between positive and negative samples. Note in multi-label prediction, only large values of all four metrics are sufficient to indicate the success of the prediction, while the combination of some large valued metrics and some small valued ones are always caused by the imbalance of the samples. Therefore, MSE provides better prediction performance than other methods on most datasets.

Fig. 5. Phase diagram (left) and corresponding CPU seconds (right) for LinGoDec on $750 \times 750$ matrices. Low-rank weight matrix $W$ is of size $750 \times 500$, and is generated by $W = UV$, where entries of $U$ and $V$ are sampled from $\mathcal{N}(0, 1/750)$ and $\mathcal{N}(0, 1/750)$, respectively. Features of items in $Z$ is sampled from $\mathcal{N}(0, 1/750)$. Entries of sparse anomaly $S$ are sampled as 1 or $-1$ with probability $\rho/2$ and 0 with probability $1 - \rho$. Noise $G$ has entries sampled from $\mathcal{N}(0, 10^{-3})$. On the $50 \times 30$ grid of sparsity-rank/n plane, 10 trials are performed for each $(\rho, r)$ pair. $W$ is said to be successfully recovered if its rel. err.$\leq 10^{-2}$. The phase diagram shows the successful recovery rate for each $(\rho, r)$ pair.

H. LinGoDec on Synthetic Data

Since most public available dataset for recommendation system rarely fulfill our demands for the training data in LinGoDec, we justify LinGoDec on synthetic data. Specifically, the rating matrix $X$ is generated by $WZ^T + S + G$. The weight matrix of linear functions $W$ is generated as the product of two Gaussian matrices. Entries in both the item feature matrix $Z$ and noise matrix $G$ are generated by i.i.d. Gaussian distribution. The sparse part has a Bernoulli model generated support set on which $\pm 1$ values are randomly assigned.

We show the phase diagram and the corresponding time cost in Figure V-G. It could be seem that LinGoDec has a slightly larger region (the white region) for successful recovery than both GreBsmo and robust PCA [19]. This is because side-information, i.e., the features of items, is
utilized in LinGoDec. Moreover, the time cost of LinGoDec is still small due to the closed-form update of $W$ and BRP based acceleration.

Therefore, LinGoDec is capable to achieve the scoring functions of users, which cannot be learned by previous matrix completion based methods, and is effective to rule out the advertising effects in user ratings. Its fast speed makes it very efficient when applied to practical systems.

REFERENCES

[1] D. L. Donoho, “Compressed sensing,” IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289–1306, 2006.
[2] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?” IEEE Transactions on Information Theory, vol. 52, no. 12, pp. 5406–5425, 2006.
[3] M. Aharon, M. Elad, and A. Bruckstein, “K-svd: An algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, 2006.
[4] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, “Dictionary learning algorithms for sparse representation,” Neural Computation, vol. 15, no. 2, pp. 349–396, 2003.
[5] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in Advances in Neural Information Processing Systems (NIPS), 2006.
[6] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Foundations of Computational Mathematics, vol. 9, pp. 717–772, 2008.
[7] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” arXiv: 0903.1476, 2009.
[8] R. Keshavan and S. Oh, “Optspace: A gradient descent algorithm on grassman manifold for matrix completion,” Submitted to IEEE Transactions on Signal Processing, 2009.
[9] S. Ji and J. Ye, “An accelerated gradient method for trace norm minimization,” in International Conference on Machine Learning (ICML), 2009.
[10] P. Jain, R. Meka, and I. S. Dhillon, “Guaranteed rank minimization via singular value projection,” in Advances in Neural Information Processing Systems (NIPS), 2010.
[11] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, vol. 24, pp. 417–441, 1936.
[12] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” Annals of Eugenics, vol. 7, no. 2, pp. 179–188, 1936.
[13] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” Science, vol. 290, no. 5500, pp. 2323–2326, 2000.
[14] J. B. Tenenbaum, V. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” Science, vol. 290, no. 5500, pp. 2319–2323, 2000.
[15] T. Zhang, D. Tao, X. Li, and J. Yang, “Patch alignment for dimensionality reduction,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1299–1313, 2009.
[16] T. Zhou, D. Tao, and X. Wu, “Manifold elastic net: a unified framework for sparse dimension reduction,” Data Mining and Knowledge Discovery (Springer), vol. 22, no. 3, pp. 340–371, 2011.
[17] J. Ye, “Generalized low rank approximations of matrices,” Machine Learning Journal, vol. 61, no. 1, pp. 167–191, 2005.
[18] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” SIAM Journal on Optimization, vol. 21, no. 2, pp. 572–596, 2011.
[19] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” Journal of the ACM, vol. 58, no. 3, pp. 11:1–11:37, 2011.
[20] D. Hsu, S. Kakade, and T. Zhang, “Robust matrix decomposition with sparse corruptions,” IEEE Transactions on Information Theory, 2011.

[21] J. Chen, J. Liu, and J. Ye, “Learning incoherent sparse and low-rank patterns from multiple tasks,” in ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2010.

[22] J. Bobin, J.-L. Starck, J. Fadili, Y. Moudden, and D. L. Donoho, “Morphological component analysis: An adaptive thresholding strategy,” IEEE Transactions on Image Processing, vol. 16, no. 11, pp. 2675–2681, 2007.

[23] P. Comon, “Independent component analysis, a new concept?” Signal Processing, vol. 36, no. 3, pp. 287–314, 1994.

[24] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” Neural Networks, vol. 13, no. 4-5, pp. 411–430, 2000.

[25] N. Halko, P. G. Martinsson, and J. A. Tropp, “Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions,” arXiv: 0909.4061, 2009.

[26] K. L. Clarkson and D. P. Woodruff, “Numerical linear algebra in the streaming model,” in ACM symposium on Theory of computing (STOC), 2009.

[27] N. Ailon and B. Chazelle, “Approximate nearest neighbors and the fast johnson-lindenstrauss transform,” in ACM symposium on Theory of computing (STOC), 2006, pp. 557–563.

[28] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, “Stable principal component pursuit,” in International Symposium on Information Theory (ISIT), 2010.

[29] L. Xiong, X. Chen, and J. Schneider, “Direct robust matrix factorization for anomaly detection,” in International Conference on Data Mining (ICDM), 2010.

[30] T. Zhou and D. Tao, “Godec: Randomized low-rank & sparse matrix decomposition in noisy case,” in International Conference on Machine Learning (ICML), 2011.

[31] ———, “Bilateral random projections,” in International Symposium on Information Theory (ISIT), 2012.

[32] S. S. Vempala, The Random Projection Method, ser. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, 2004, vol. 65.

[33] T. Zhou and D. Tao, “Greedy bilateral sketch, completion and smoothing,” in International Conference on Artificial Intelligence and Statistics (AISTATS), 2013.

[34] ———, “Shifted subspaces tracking on sparse outlier for motion segmentation,” in International Joint Conference on Artificial Intelligence (IJCAI), 2013.

[35] S. Wu, O. Oreifej, and M. Shah, “Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories,” in International Conference on Computer Vision (ICCV), 2011, pp. 1419–1426.

[36] S. Ali and M. Shah, “A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[37] K. Fragkiadaki and J. Shi, “Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2073–2080.

[38] R. Hess and A. Fern, “Discriminatively trained particle filters for complex multi-object tracking,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[39] C. Yang, R. Duraiswami, and L. S. Davis, “Fast multiple object tracking via a hierarchical particle filter,” in International Conference on Computer Vision (ICCV), 2005, pp. 212–219.

[40] T. Zhou and D. Tao, “Multi-label subspace ensemble,” in International Conference on Artificial Intelligence and Statistics (AISTATS), 2012.

[41] G. Tsoumakas, M.-L. Zhang, and Z.-H. Zhou, “Learning from multi-label data,” in The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2009.

[42] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Mining multi-label data,” Data Mining and Knowledge Discovery Handbook, 2010.

[43] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview,” International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1–13, 2007.
[44] J. Petterson and T. Caetano, “Reverse multi-label learning,” in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[45] G. Tsoumakas, I. Katakis, and I. Vlahavas, “Effective and efficient multilabel classification in domains with large number of labels,” in *ECML/PKDD Workshop on Mining Multidimensional Data*, 2008.

[46] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society, Series B*, vol. 68, pp. 49–67, 2006.

[47] K. Bredies and D. A. Lorenz, “Iterated hard shrinkage for minimization problems with sparsity constraints,” *SIAM Journal on Scientific Computing*, vol. 30, no. 2, pp. 657–683, 2008.

[48] A. S. Lewis and J. Malick, “Alternating projections on manifolds,” *Mathematics of Operations Research*, vol. 33, no. 1, pp. 216–234, 2008.

[49] S. Roweis, “Em algorithms for pca and spca,” in *Advances in Neural Information Processing Systems (NIPS)*, 1998.

[50] M. Fornasier, H. Rauhut, and R. Ward, “Low-rank matrix recovery via iteratively reweighted least squares minimization,” *SIAM Journal on Optimization*, vol. 21, no. 4, pp. 1614–1640, 2011.

[51] D. Zachariah, M. Sundin, M. Jansson, and S. Chatterjee, “Alternating least-squares for low-rank matrix reconstruction,” *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 231–234, 2012.

[52] Z. Wen, W. Yin, and Y. Zhang, “Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm,” *Mathematical Programming Computation*, vol. 4, no. 4, pp. 333–361, 2012.

[53] S. Shalev-Shwartz, A. Gonen, and O. Shamir, “Large-scale convex minimization with a low-rank constraint,” in *International Conference on Machine Learning (ICML)*, 2011.

[54] S. J. Prince, *Computer vision: models, learning and inference*. Cambridge University Press, 2011.

[55] J. Read, B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” *Machine Learning and Knowledge Discovery in Databases*, pp. 254–269, 2009.

[56] N. C. Bianchi, C. Gentile, and L. Zaniboni, “Incremental algorithms for hierarchical classification,” *Journal of Machine Learning Research*, vol. 7, pp. 31–54, 2006.

[57] Y. Zhang and Z. H. Zhou, “Multi-label dimensionality reduction via dependence maximization,” in *International conference on Artificial intelligence (AAAI)*, 2008, pp. 1503–1505.

[58] S. Ji, L. Tang, S. Yu, and J. Ye, “A shared-subspace learning framework for multi-label classification,” *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 1, 2010.

[59] G. Tsoumakas and I. Vlahavas, “Random k-labelsets: An ensemble method for multilabel classification,” in *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2007, pp. 406–417.

[60] W. Cheng and E. H"ullermeier, “Combining instance-based learning and logistic regression for multilabel classification,” *Machine Learning*, vol. 76, no. 2-3, pp. 211–225, 2009.

[61] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: [http://www.public.asu.edu/~jye02/Software/SLEP](http://www.public.asu.edu/~jye02/Software/SLEP)

[62] L. Cheng, M. Gong, D. Schuurmans, and T. Caelli, “Real-time discriminative background subtraction,” *to appear in IEEE Trans on Image Processing*, 2010.

[63] M. L. Zhang and Z. H. Zhou, “ML-knn: A lazy learning approach to multi-label learning,” *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[64] R. J. Muirhead, *Aspects of multivariate statistical theory*. New York: John Wiley & Sons Inc., 1982.
APPENDIX I: ANALYSIS OF GODEC

We theoretically analyze the convergence of GoDec. The objective value (decomposition error) \( \| X - L - S \|_F^2 \) monotonically decreases and converges to a local minimum. Since the updating of \( L \) and \( S \) in GoDec is equivalent to alternatively projecting \( L \) or \( S \) onto two smooth manifolds, we use the framework proposed in [48] to prove the asymptotical property and linear convergence of \( L \) and \( S \). The asymptotic and convergence speeds are mainly determined by the angle between the two manifolds. We discuss how \( L \), \( S \) and \( G \) influence these speeds via influencing the cosine of the angle. The analyses show the convergence of GoDec is robust to the noise \( G \).

In particular, we first prove that the objective value \( \| X - L - S \|_F^2 \) (decomposition error) converges to a local minimum. Then we demonstrate the asymptotic properties of GoDec and prove that the solutions \( L \) and \( S \) respectively converge to local optimums with linear rate less than 1. The influence of \( L \), \( S \) and \( G \) to the asymptotic/convergence speeds is analyzed. The speeds are slowed down by augmenting the magnitude of noise part \( \| G \|_F^2 \). However, the convergence still holds unless \( \| G \|_F^2 \gg \| L \|_F^2 \) or \( \| G \|_F^2 \gg \| S \|_F^2 \).

We have the following theorem about the convergence of the objective value \( \| X - L - S \|_F^2 \) in (2).

**Theorem 1:** (Convergence of objective value). The alternative optimization (3) produces a sequence of \( \| X - L - S \|_F^2 \) that converges to a local minimum.

**Proof:** Let the objective value \( \| X - L - S \|_F^2 \) after solving the two subproblems in (3) be \( E_t^1 \) and \( E_t^2 \), respectively, in the \( t \)th iteration. On the one hand, we have

\[
E_t^1 = \| X - L_t - S_{t-1} \|_F^2, \quad E_t^2 = \| X - L_t - S_t \|_F^2.
\]

The global optimality of \( S_t \) yields \( E_t^1 \geq E_t^2 \). On the other hand,

\[
E_t^2 = \| X - L_t - S_t \|_F^2, \quad E_{t+1}^1 = \| X - L_{t+1} - S_t \|_F^2.
\]

The global optimality of \( L_{t+1} \) yields \( E_t^2 \geq E_{t+1}^1 \). Therefore, the objective values (decomposition errors) \( \| X - L - S \|_F^2 \) keep decreasing throughout GoDec (3):

\[
E_t^1 \geq E_t^2 \geq E_{t+1}^1 \geq \cdots \geq E_t^1 \geq E_t^2 \geq E_{t+1}^1 \geq \cdots
\]

Since the objective of (2) is monotonically decreasing and the constraints are satisfied all the time, (3) produces a sequence of objective values that converge to a local minimum. This completes the proof.

The asymptotic property and the linear convergence of \( L \) and \( S \) in GoDec are demonstrated based on the framework proposed in [48]. We firstly consider \( L \). From a different prospective, GoDec algorithm shown in [41] is equivalent to iteratively projecting \( L \) onto one manifold \( M \) and then onto another manifold \( N \). This kind of optimization method is the so called “alternating projections on manifolds”. To see this, in (4), by substituting \( S_t \) into the next updating of \( L_{t+1} \), we have:

\[
L_{t+1} = \mathcal{P}_M (X - \mathcal{P}_N (X - L_t)) = \mathcal{P}_M \mathcal{P}_N (L_t),
\]

Both \( M \) and \( N \) are two \( C^k \)-manifolds around a point \( \bar{L} \in M \cap N \): 

\[
\mathcal{M} = \{ H \in \mathbb{R}^{m \times n} : \text{rank} \ (H) = r \}, \quad \mathcal{N} = \{ X - \mathcal{P}_N (X - H) : H \in \mathbb{R}^{m \times n} \}.
\]
According to the above definitions, any point \( L \in \mathcal{M} \cap \mathcal{N} \) satisfies:

\[
L = \mathcal{P}_{\mathcal{M} \cap \mathcal{N}}(L) \quad \Rightarrow \\
L = X - \mathcal{P}_{\Omega}(X - L), \ \text{rank}(L) = r.
\]

(44)

(45)

Thus any point \( L \in \mathcal{M} \cap \mathcal{N} \) is a local solution of \( L \) in \( (\mathcal{O}) \).

We define the angle between two manifolds \( \mathcal{M} \) and \( \mathcal{N} \) at point \( L \) as the angle between the corresponding tangent spaces \( T_{\mathcal{M}}(L) \) and \( T_{\mathcal{N}}(L) \). The angle is between 0 and \( \pi/2 \) with cosine:

\[
c(\mathcal{M}, \mathcal{N}, L) = c(T_{\mathcal{M}}(L), T_{\mathcal{N}}(L)).
\]

(46)

In addition, if \( S \) is the unit sphere in \( \mathbb{R}^{m \times n} \), the angle between two subspaces \( M \) and \( N \) in \( \mathbb{R}^{m \times n} \) is defined as the angle between 0 and \( \pi/2 \) with cosine:

\[
c(M, N) = \max \left\{ \langle x, y \rangle : x \in S \cap M \cap (M \cap N)\perp, \ y \in S \cap N \cap (M \cap N)\perp \right\}.
\]

We give the following proposition about the angle between two subspaces \( M \) and \( N \):

**Proposition 1:** Following the above definition of the angle between two subspaces \( M \) and \( N \), we have

\[
c(M, N) = \max \left\{ \langle x, y \rangle : x \in S \cap M \cap N\perp, \ y \in S \cap N \cap M\perp \right\}
\]

The angle between \( \mathcal{M} \) and \( \mathcal{N} \) is used in the asymptotical property and the linear convergence rate of “alternating projections on manifolds” algorithms.

**Theorem 2:** (Asymptotic property \([48]\)). Let \( \mathcal{M} \) and \( \mathcal{N} \) be two transverse \( C^2 \)-manifolds around a point \( \mathcal{T} \in \mathcal{M} \cap \mathcal{N} \). Then

\[
\limsup_{L \to \mathcal{T}, L \notin \mathcal{M} \cap \mathcal{N}} \frac{\|\mathcal{P}_{\mathcal{M}} \mathcal{P}_{\mathcal{N}}(L) - \mathcal{P}_{\mathcal{M} \cap \mathcal{N}}(L)\|}{\|L - \mathcal{P}_{\mathcal{M} \cap \mathcal{N}}(L)\|} \leq c(\mathcal{M}, \mathcal{N}, \mathcal{T}).
\]

A refinement of the above argument is

\[
\limsup_{L \to \mathcal{T}, L \notin \mathcal{M} \cap \mathcal{N}} \frac{\|\left(\mathcal{P}_{\mathcal{M}} \mathcal{P}_{\mathcal{N}}\right)^n(L) - \mathcal{P}_{\mathcal{M} \cap \mathcal{N}}(L)\|}{\|L - \mathcal{P}_{\mathcal{M} \cap \mathcal{N}}(L)\|} \leq c^{2^{n-1}}
\]

for \( n = 1, 2, \ldots \) and \( c = c(\mathcal{M}, \mathcal{N}, \mathcal{T}) \).

**Theorem 3:** (Linear convergence of variables \([48]\)). In \( \mathbb{R}^{m \times n} \), let \( \mathcal{M} \) and \( \mathcal{N} \) be two transverse manifolds around a point \( \mathcal{T} \in \mathcal{M} \cap \mathcal{N} \). If the initial point \( L_0 \in \mathbb{R}^{m \times n} \) is close to \( \mathcal{T} \), then the method of alternating projections

\[
L_{t+1} = \mathcal{P}_{\mathcal{M}} \mathcal{P}_{\mathcal{N}}(L_t), \ (t = 0, 1, 2, \ldots)
\]

is well-defined, and the distance \( d_{\mathcal{M} \cap \mathcal{N}}(L_t) \) from the iterate \( L_t \) to the intersection \( \mathcal{M} \cap \mathcal{N} \) decreases Q-linearly to zero. More precisely, given any constant \( c \) strictly larger than the cosine of the angle of the intersection between the manifolds, \( \langle \mathcal{M}, \mathcal{N}, \mathcal{T} \rangle \), if \( L_0 \) is close to \( \mathcal{T} \), then the iterates satisfy

\[
d_{\mathcal{M} \cap \mathcal{N}}(L_{t+1}) \leq c \cdot d_{\mathcal{M} \cap \mathcal{N}}(L_t), \ (t = 0, 1, 2, \ldots)
\]

Furthermore, \( L_t \) converges linearly to some point \( L^* \in \mathcal{M} \cap \mathcal{N} \), i.e., for some constant \( \alpha > 0 \),

\[
\|L_t - L^*\| \leq \alpha c^t, \ (t = 0, 1, 2, \ldots).
\]
Since GoDec algorithm can be written as the form of alternating projections on two manifolds \( \mathcal{M} \) and \( \mathcal{N} \) given in (43) and they satisfy the assumptions of Theorem 2 and Theorem 3, \( L \) in GoDec converges to a local optimum with linear rate. Similarly, we can prove the linear convergence of \( S \).

Since cosine \( c(\mathcal{M}, \mathcal{N}, \mathcal{L}) \) in Theorem 2 and Theorem 3 determines the asymptotic and convergence speeds of the algorithm. We discuss how \( L, S \) and \( G \) influence the asymptotic and convergence speeds via analyzing the relationship between \( L, S, G \) and \( c(\mathcal{M}, \mathcal{N}, \mathcal{L}) \).

**Theorem 4:** (Asymptotic and convergence speed). In GoDec, the asymptotical improvement and the linear convergence of \( L \) and \( S \) stated in Theorem 2 and Theorem 3 will be slowed by augmenting

\[
\frac{\| \Delta_L \|_F}{\| L + \Delta_L \|_F}, \Delta_L = (S + G) - \mathcal{P}_\Omega (S + G),
\]

\[
\frac{\| \Delta_S \|_F}{\| S + \Delta_S \|_F}, \Delta_S = (L + G) - \mathcal{P}_\mathcal{M} (L + G).
\]

However, the asymptotical improvement and the linear convergence will not be harmed and is robust to the noise \( G \) unless when \( \| G \|_F \gg \| S \|_F \) and \( \| G \|_F \gg \| L \|_F \), which lead the two terms increasing to 1.

**Proof:** GoDec approximately decomposes a matrix \( X = L + S + G \) into the low-rank part \( L \) and the sparse part \( S \). According to the above analysis, GoDec is equivalent to alternating projections of \( L \) on \( \mathcal{M} \) and \( \mathcal{N} \), which are given in (43). According to Theorem 2 and Theorem 3, smaller \( c(\mathcal{M}, \mathcal{N}, \mathcal{L}) \) produces faster asymptotic and convergence speeds, while \( c(\mathcal{M}, \mathcal{N}, \mathcal{L}) = 1 \) is possible to make \( L \) and \( S \) stopping converging. Below we discuss how \( L, S \) and \( G \) influence \( c(\mathcal{M}, \mathcal{N}, \mathcal{L}) \) and further influence the asymptotic and convergence speeds of GeDec.

According to (46), we have

\[
c(\mathcal{M}, \mathcal{N}, \mathcal{L}) = c(\mathcal{T}_\mathcal{M}(\mathcal{L}), \mathcal{T}_\mathcal{N}(\mathcal{L})).
\]  

Substituting the equation given in Proposition 1 into the right-hand side of the above equation yields

\[
c(\mathcal{M}, \mathcal{N}, \mathcal{L}) = \max \{ \langle x, y \rangle : x \in \mathcal{S} \cap \mathcal{T}_\mathcal{M}(\mathcal{L}) \cap \mathcal{N}_\mathcal{N}(\mathcal{L}), y \in \mathcal{S} \cap \mathcal{T}_\mathcal{N}(\mathcal{L}) \cap \mathcal{N}_\mathcal{M}(\mathcal{L}) \}.
\]  

The normal spaces of manifolds \( \mathcal{M} \) and \( \mathcal{N} \) on point \( \mathcal{L} \) is respectively given by

\[
\mathcal{N}_\mathcal{M}(\mathcal{L}) = \{ y \in \mathbb{R}^{m \times n} : u_i^T y v_j = 0, \mathcal{L} = UDV^T \},
\]

\[
\mathcal{N}_\mathcal{N}(\mathcal{L}) = \{ X - \mathcal{P}_\Omega (X - \mathcal{L}) \},
\]

where \( \mathcal{L} = UDV^T \) represents the eigenvalue decomposition of \( \mathcal{L} \), \( U = [u_1, ..., u_r] \) and \( V = [v_1, ..., v_r] \). Assume \( X = \mathcal{L} + \mathcal{S} + \mathcal{G} \), wherein \( \mathcal{G} \) is the noise corresponding to \( \mathcal{L} \), we have

\[
\hat{\mathcal{L}} = X - (\mathcal{S} + \mathcal{G}),
\]

\[
\hat{\mathcal{L}} = X - \mathcal{P}_\Omega (\mathcal{S} + \mathcal{G}) \Rightarrow
\]

\[
\hat{\mathcal{L}} = \mathcal{L} + [(\mathcal{S} + \mathcal{G}) - \mathcal{P}_\Omega (\mathcal{S} + \mathcal{G})] = \mathcal{L} + \Delta.
\]

Thus the normal space of manifold \( \mathcal{N} \) is

\[
\mathcal{N}_\mathcal{N}(\mathcal{L}) = \{ \mathcal{L} + \Delta \}.
\]  

Since the tangent space is the complement space of the normal space, by using the normal space of \( \mathcal{M} \) in (49) and the normal space of \( \mathcal{N} \) given in (51), we can verify

\[
\mathcal{N}_\mathcal{N}(\mathcal{L}) \subseteq \mathcal{T}_\mathcal{M}(\mathcal{L}), \mathcal{N}_\mathcal{M}(\mathcal{L}) \subseteq \mathcal{T}_\mathcal{N}(\mathcal{L}).
\]
By substituting the above results into (53), we obtain
\[
c(M, \mathcal{N}, \mathcal{L}) = \max \{ \langle x, y \rangle : x \in \mathcal{S} \cap N_M(\mathcal{L}), y \in \mathcal{S} \cap N_M(\mathcal{L}) \}.
\] (53)

Hence we have
\[
\langle x, y \rangle = \text{tr} (VDU^T y + \Delta^T y) = \text{tr} (DU^T yV) + \text{tr} (\Delta^T y) = \text{tr} (\Delta^T y) .
\] (54)

The last equivalence is due to \( u_i^T yv_j = 0 \) in (49). Thus
\[
c(M, \mathcal{N}, \mathcal{L}) = \max \{ \langle x, y \rangle \} \leq \max \{ \langle D\Delta, D_y \rangle \} ,
\] (55)

where the diagonal entries of \( D\Delta \) and \( D_y \) are composed by eigenvalues of \( \Delta \) and \( y \), respectively. The last inequality is obtained by considering the case when \( x \) and \( y \) have identical left and right singular vectors. Because \( \mathcal{L} + \Delta, y \in \mathcal{S} \) infers \( \| \mathcal{L} + \Delta \|_F^2 = \| y \|_F^2 = 1 \), we have
\[
c(M, \mathcal{N}, \mathcal{L}) \leq \max \{ \langle D\Delta, D_y \rangle \}
\leq \| D\Delta \|_F \| D_y \|_F \leq \| D\Delta \|_F .
\] (56)

Since \( c \) in Theorem 3 can be selected as any constant that is strictly larger than \( c(M, \mathcal{N}, \mathcal{L}) \leq \| D\Delta \|_F \), we can choose \( c = c(M, \mathcal{N}, \mathcal{L}) + \Delta_c \leq \| D\Delta \|_F \). In Theorem 2, the cosine \( c(M, \mathcal{N}, \mathcal{L}) \) is directly used.

Therefore, the asymptotic and convergence speeds of \( L \) will be slowed by augmenting \( \| \Delta \|_F \), and vice versa. However, the asymptotical improvement and the linear convergence will not be jeopardized unless \( \| \Delta \|_F = 1 \). For general \( L + \Delta \) that is not normalized onto the sphere \( \mathcal{S} \), \( \| \Delta \|_F \) should be replaced by \( \| \Delta \|_F / \| L + \Delta \|_F \).

For the variable \( S \), we can obtain an analogous result via an analysis in a similar style as above. For general \( L + \Delta \) without normalization, the asymptotic/convergence speed of \( S \) will be slowed by augmenting \( \| \Delta \|_F / \| S + \Delta \|_F \), and vice versa, wherein
\[
\Delta = (L + G) - \mathcal{P}_M (L + G) .
\] (57)

The asymptotical improvement and the linear convergence will not be jeopardized unless \( \| \Delta \|_F / \| S + \Delta \|_F = 1 \).

This completes the proof. \( \blacksquare \)

Theorem 4 reveals the influence of the low-rank part \( L \), the sparse part \( S \) and the noise part \( G \) to the asymptotic/convergence speeds of \( L \) and \( S \) in GoDec. Both \( \Delta_L \) and \( \Delta_S \) are the element-wise hard thresholding error of \( S + G \) and the singular value hard thresholding error of \( L + G \), respectively. Large errors will slow the asymptotic and convergence speeds of GoDec. Since \( S - \mathcal{P}_M(S) = 0 \) and \( L - \mathcal{P}_M(L) = 0 \), the noise part \( G \) in \( \Delta_L \) and \( \Delta_S \) can be interpreted as the perturbations to \( S \) and \( L \) and deviates the two errors from 0. Thus noise \( G \) with large magnitude will decelerate the asymptotical improvement and the linear convergence, but it will not ruin the convergence unless \( \| G \|_F \gg \| S \|_F \) or \( \| G \|_F \gg \| L \|_F \). Therefore, GoDec is robust to the additive noise in \( X \) and is able to find the approximated \( L + S \) decomposition when noise \( G \) is not overwhelming.
APPENDIX II: APPROXIMATION ERROR BOUND OF BRP

VI. APPROXIMATION ERROR BOUNDS

We analyze the error bounds of the BRP based low-rank approximation \([5]\) and its power scheme modification \([9]\).

The SVD of an \(m \times n\) (w.l.o.g, \(m \geq n\)) matrix \(X\) takes the form:

\[
X = U \Lambda V^T = U_1 \Lambda_1 V_1^T + U_2 \Lambda_2 V_2^T,
\]

where \(\Lambda_1\) is an \(r \times r\) diagonal matrix which diagonal elements are the first largest \(r\) singular values, \(U_1\) and \(V_1\) are the corresponding singular vectors, \(A_2\), \(U_2\) and \(V_2\) forms the rest part of SVD. Assume that \(r\) is the target rank, \(A_1\) and \(A_2\) have \(r+p\) columns for oversampling. We consider the spectral norm of the approximation error \(E\) for \([5]\):

\[
\|X - L\| = \left\| X - Y_1 (A_2^T Y_1)^{-1} Y_2 \right\|
= \left\| I - X A_1 (A_2^T X A_1)^{-1} A_2^T \right\| X. \tag{59}
\]

The unitary invariance of the spectral norm leads to

\[
\|X - L\| = \left\| U^T \left[ I - X A_1 (A_2^T X A_1)^{-1} A_2^T \right] X \right\|
= \left\| \Lambda \left[ I - V^T A_1 (A_2^T X A_1)^{-1} A_2^T U A \right] \right\|. \tag{60}
\]

In low-rank approximation, the left random projection matrix \(A_2\) is built from the left random projection \(Y_1 = X A_1\), and then the right random projection matrix \(A_1\) is built from the left random projection \(Y_2 = X^T A_2\). Thus \(A_2 = Y_1 = X A_1 = U \Lambda V^T A_1\) and \(A_1 = Y_2 = X^T A_2 = X^T X A_1 = V \Lambda^2 V^T A_1\). Hence the approximation error given in (60) has the following form:

\[
\left\| \Lambda \left[ I - \Lambda^2 V^T A_1 (A_2^T V A^4 V^T A_1)^{-1} A_2^T V A^2 \right] \right\|. \tag{61}
\]

The following Theorem 5 gives the bound for the spectral norm of the deterministic error \(\|X - L\|\).

**Theorem 5:** (Deterministic error bound) Given an \(m \times n\) (\(m \geq n\)) real matrix \(X\) with singular value decomposition \(X = U \Lambda V^T = U_1 \Lambda_1 V_1^T + U_2 \Lambda_2 V_2^T\), and chosen a target rank \(r \leq n - 1\) and an \(n \times (r + p)\) (\(p \geq 2\)) standard Gaussian matrix \(A_1\), the BRP based low-rank approximation \([5]\) approximates \(X\) with the error upper bounded by

\[
\|X - L\|^2 \leq \left\| A_2^2 (V_2^T A_1) (V_1^T A_1)^{\dagger} \Lambda_1^{-1} \right\|^2 + \| A_2 \|^2.
\]

See Section VI-A for the proof of Theorem 5.

If the singular values of \(X\) decay fast, the first term in the deterministic error bound will be very small. The last term is the rank-\(r\) SVD approximation error. Therefore, the BRP based low-rank approximation \([5]\) is nearly optimal.

**Theorem 6:** (Deterministic error bound, power scheme) Frame the hypotheses of Theorem 5 the power scheme modification \([9]\) approximates \(X\) with the error upper bounded by

\[
\|X - L\|^2 \leq \left( \left\| A_2^{2(2q+1)} (V_2^T A_1) (V_1^T A_1)^{\dagger} \Lambda_1^{-2q+1} \right\|^2 + \left\| A_2^{2q+1} \right\|^{1/(2q+1)} \right).
\]
If the singular values of $X$ decay slowly, the error produced by the power scheme modification (9) is less than the BRP based low-rank approximation (5) and decreasing with the increasing of $q$.

The average error bound of BRP based low-rank approximation is obtained by analyzing the statistical properties of the random matrices that appear in the deterministic error bound in Theorem 5.

**Theorem 7: (Average error bound)** Frame the hypotheses of Theorem 5.

$$E\|X - L\| \leq \left( \frac{1}{p - 1} \sum_{i=1}^{r} \frac{\lambda_{r+1}^2 + 1}{\lambda_{i}^2} \right) |\lambda_{r+1}|$$

$$+ \frac{e^{r+p}}{p} \sqrt{\sum_{i=r+1}^{n} \frac{\lambda_{i}^2}{\lambda_{r}^2}}.$$  

See Section VI-A for the proof of Theorem 7.

The average error bound will approach to the SVD approximation error $|\lambda_{r+1}|$ if $|\lambda_{r+1}| \ll |\lambda_{i}; i=1,\ldots,r|$ and $|\lambda_{i}| \gg |\lambda_{i}; i=r+1,\ldots,n|$.

The average error bound for the power scheme modification is then obtained from the result of Theorem 7.

**Theorem 8: (Average error bound, power scheme)** Frame the hypotheses of Theorem 5 the power scheme modification (9) approximates $X$ with the expected error upper bounded by

$$E\|X - L\| \leq \left( \frac{1}{p - 1} \sum_{i=1}^{r} \frac{\lambda_{r+1}^{2q+1} + 1}{\lambda_{i}^{2q+1}} \right) |\lambda_{r+1}|$$

$$+ \frac{e^{r+p}}{p} \sqrt{\sum_{i=r+1}^{n} \frac{\lambda_{i}^{2q+1}}{\lambda_{r}^{2q+1}}}.$$  

See Section VI-A for the proof of Theorem 8.

Compared the average error bounds of the BRP based low-rank approximation with its power scheme modification, the latter produces less error than the former, and the error can be further decreased by increasing $q$.

The deviation bound for the spectral norm of the approximation error can be obtained by analyzing the deviation bound of $\|A_2^2 (V_2^T A_1) (V_1^T A_1)^\dagger A_1^{-1}\|$ in the deterministic error bound and by applying the concentration inequality for Lipschitz functions of a Gaussian matrix.

**Theorem 9: (Deviation bound)** Frame the hypotheses of Theorem 5. Assume that $p \geq 4$. For all $u, t \geq 1$, it holds that

$$\|X - L\| \leq \left( \frac{1}{p} + t \frac{\sqrt{12r}}{p} \left( \sum_{i=1}^{r} \lambda_{i}^{-1} \right)^{1/2} + \frac{e^{r+p}}{p+1} \right).$$

$$tu\lambda_{r+1}^{-1} \lambda_{r+1}^{2} + \frac{e^{r+p}}{p+1} \cdot t\lambda_{r+1}^{-1} \left( \sum_{i=r+1}^{n} \lambda_{i}^{2} \right)^{1/2}.$$  

except with probability $e^{-u^2/2} + 4t^{-p} + t^{-(p+1)}$.

See Section VI-A for the proof of Theorem 9.
A. Proofs of error bounds

1) Proof of Theorem 5. The following lemma and propositions from [26] will be used in the proof.

Lemma 1: Suppose that \( M \geq 0 \). For every \( A \), the matrix \( A^T MA \geq 0 \). In particular,
\[
M \preceq N \Rightarrow A^T MA \preceq A^T NA.
\] (62)

Proposition 2: Suppose \( \text{range}(N) \subseteq \text{range}(M) \). Then, for each matrix \( A \), it holds that \( \|P_N A\| \leq \|P_M A\| \) and that \( \|(I - P_M)A\| \leq \|(I - P_N)A\| \).

Proposition 3: Suppose that \( M \geq 0 \). Then
\[
I - (I + M)^{-1} \preceq M.
\] (63)

Proposition 4: We have \( \|M\| \leq \|A\| + \|C\| \) for each partitioned positive semidefinite matrix
\[
M = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}.
\] (64)

The proof of Theorem 5 is given below.

Proof: Since an orthogonal projector projects a given matrix to the range (column space) of a matrix \( M \) is defined as \( P_M = M(M^T M)^{-1}M^T \), the deterministic error (61) can be written as
\[
\|E\| = \|\Lambda (I - P_M)\|, \quad M = \Lambda^2 V^T A_1.
\] (65)

By applying Proposition 2 to the error (65), because \( \text{range}(M(V_1^T A_1)^{-1}A_1^{-2}) \subseteq \text{range}(M) \), we have
\[
\|E\| = \|\Lambda (I - P_M)\| \leq \|\Lambda (I - P_N)\|,
\] (66)

where
\[
N = \begin{bmatrix} \Lambda^2 V_1^T A_1 \\ \Lambda^2 V_2^T A_1 \end{bmatrix} (V_1^T A_1)^{-1}A_1^{-2} = \begin{bmatrix} I \\ H \end{bmatrix}.
\] (67)

Thus \( (I - P_N) \) can be written as
\[
I - P_N = \begin{bmatrix} I - (I + H^T H)^{-1} - (I + H^T H)^{-1} H^T \\ -H (I + H^T H)^{-1} I - H (I + H^T H)^{-1} H^T \end{bmatrix}
\]

For the top-left block in (68), Proposition 3 leads to \( I - (I + H^T H)^{-1} \preceq H^T H \). For the bottom-right block in (68), Lemma 1 leads to \( I - H (I + H^T H)^{-1} H^T \preceq I \). Therefore,
\[
I - P_N \preceq \begin{bmatrix} H^T H & - (I + H^T H)^{-1} H^T \\ -H (I + H^T H)^{-1} I \end{bmatrix}
\]

By applying Lemma 1, we have
\[
\Lambda (I - P_N) \Lambda \preceq \begin{bmatrix} \Lambda^2 I^T H^T H A_1 & -\Lambda^2 I^T (I + H^T H)^{-1} H^T A_2 \\ -\Lambda^2 H (I + H^T H)^{-1} A_1 & \Lambda^2 A_2 \end{bmatrix}
\]

According to Proposition 4, the spectral norm of \( \Lambda(I - P_N) \) is bounded by
\[
\|\Lambda (I - P_N)\| \leq \|\Lambda (I - P_N)\| \leq \|\Lambda (I - P_N)\| \leq \|\Lambda^2 (V_1^T A_1) (V_1^T A_1)^{-1} A_1^{-2}\|^2 + \|A_2\|^2.
\] (68)

By substituting (68) into (66), we obtain the deterministic error bound. This completes the proof.
2) Proof of Theorem 6. The following proposition from [25] will be used in the proof.

**Proposition 5:** Let $P$ be an orthogonal projector, and let $A$ be a matrix. For each nonnegative $q$,

$$
\|PA\| \leq \left\|P \left( AA^T \right)^q A \right\|^{1/(2q+1)}.
$$

(69)

The proof of Theorem 6 is given below.

**Proof:** The power scheme modification applies the BRP based low-rank approximation to $\tilde{X} = (XX^T)^qX = U\Lambda^{2q+1}V^T$ rather than $X$. In this case, the approximation error is

$$
\|\tilde{X} - \tilde{L}\|^2 \
\leq \left\|\Lambda_2^{2(2q+1)} (V_2^T A_1) (V_1^T A_1)^\dagger \Lambda_1^{-(2q+1)}\right\|^2 + \left\|\Lambda_2^{2q+1}\right\|^2.
$$

(71)

The deterministic error bound for the power scheme modification is obtained by applying Proposition 5 to (71).

This completes the proof.

3) Proof of Theorem 7. The following propositions from [25] will be used in the proof.

**Proposition 6:** Fix matrices $S$, $T$, and draw a standard Gaussian matrix $G$. Then it holds that

$$
\mathbb{E}\left\|SGT^T\right\| \leq \|S\|\|T\|_F + \|S\|_F\|T\|.
$$

(72)

**Proposition 7:** Draw an $r \times (r+p)$ standard Gaussian matrix $G$ with $p \geq 2$. Then it holds that

$$
\mathbb{E}\|G^1\|^2_F = \frac{r}{p-1}, \mathbb{E}\|G^1\| \leq \frac{e\sqrt{T + p}}{p}.
$$

(73)

The proof of Theorem 7 is given below.

**Proof:** The distribution of a standard Gaussian matrix is rotational invariant. Since 1) $A_1$ is a standard Gaussian matrix and 2) $V$ is an orthogonal matrix, $V^T A_1$ is a standard Gaussian matrix, and its disjoint submatrices $V_1^T A_1$ and $V_2^T A_1$ are standard Gaussian matrices as well.

Theorem 5 and the Hölder’s inequality imply that

$$
\mathbb{E}\|X - L\| \leq \mathbb{E}\left(\left\|\Lambda_2^2 (V_2^T A_1) (V_1^T A_1)^\dagger \Lambda_1^{-1}\right\|^2 + \|A_2\|^2\right)^{1/2}
\leq \mathbb{E}\left\|\Lambda_2^2 (V_2^T A_1) (V_1^T A_1)^\dagger \Lambda_1^{-1}\right\| + \|A_2\|.
$$

(74)

We condition on $V_1^T A_1$ and apply Proposition 6 to bound the expectation w.r.t. $V_2^T A_1$, i.e.,

$$
\mathbb{E}\|\Lambda_2^2 (V_2^T A_1) (V_1^T A_1)^\dagger \Lambda_1^{-1}\| \
\leq \mathbb{E}\left(\|\Lambda_2^2\| \| (V_1^T A_1)^\dagger \Lambda_1^{-1}\|_F + \|\Lambda_2^2\|_F \| (V_1^T A_1)^\dagger \Lambda_1^{-1}\|\right)
\leq \|\Lambda_2^2\| \left(\mathbb{E}\| (V_1^T A_1)^\dagger \Lambda_1^{-1}\|_F^2\right)^{1/2} + \|\Lambda_2^2\|_F \mathbb{E} \| (V_1^T A_1)^\dagger \| \| \Lambda_1^{-1} \| .
$$

(75)
The Frobenius norm of \((V_1^T A_1)^\dagger \Lambda_1^{-1}\) can be calculated as
\[
\|(V_1^T A_1)^\dagger \Lambda_1^{-1}\|_F^2 = \text{trace} \left( \Lambda_1^{-1} \left( (V_1^T A_1)^\dagger \right)^T (V_1^T A_1)^\dagger \right) \Lambda_1^{-1} \\
= \text{trace} \left( \left( (\Lambda_1 V_1^T A_1) (\Lambda_1 V_1^T A_1)^T \right)^{-1} \right) .
\]

Since 1) \(V_1^T A_1\) is a standard Gaussian matrix and 2) \(\Lambda_1\) is a diagonal matrix, each column of \(\Lambda_1 V_1^T A_1\) follows \(r\)-variate Gaussian distribution \(\mathcal{N}(\mathbf{0}, \Lambda_1^2)\). Thus the random matrix \(\left( (\Lambda_1 V_1^T A_1) (\Lambda_1 V_1^T A_1)^T \right)^{-1}\) follows the inverted Wishart distribution \(\mathcal{W}_r^{-1}(\Lambda_1^{-2}, r + p)\). According to the expectation of inverted Wishart distribution \(\mathcal{W}_r^{-1}\), we have
\[
E \| (V_1^T A_1)^\dagger \Lambda_1^{-1} \|^2_F = \text{trace} \left( \left( (\Lambda_1 V_1^T A_1) (\Lambda_1 V_1^T A_1)^T \right)^{-1} \right) \\
= \frac{1}{p - 1} \sum_{i=1}^r \lambda_i^{-2} .
\] (76)

We apply Proposition 7 to the standard Gaussian matrix \(V_1^T A_1\) and obtain
\[
E \| (V_1^T A_1)^\dagger \| \leq \frac{e^{\sqrt{r + p}}}{p} .
\] (77)

Therefore, (75) can be further derived as
\[
E \| \Lambda_2^2 (V_2^T A_1) (V_1^T A_1)^\dagger \Lambda_1^{-1} \| \\
\leq \lambda_{r+1}^2 \cdot \sqrt{\frac{1}{p - 1} \sum_{i=1}^r \lambda_i^{-2} + \frac{\sum_{i=r+1}^n \lambda_i^2}{p} \cdot |\lambda_r^{-1}|} \\
= |\lambda_{r+1}| \sqrt{\frac{1}{p - 1} \sum_{i=1}^r \frac{\lambda_{r+1}^2}{\lambda_i^2} + \frac{e^{\sqrt{r + p}}}{p} \cdot \sum_{i=r+1}^n \frac{\lambda_i^2}{\lambda_r^2} .}
\] (78)

By substituting (78) into (74), we obtain the average error bound
\[
E \| X - L \| \leq \left( \sqrt{\frac{1}{p - 1} \sum_{i=1}^r \frac{\lambda_{r+1}^2}{\lambda_i^2} + 1} \right) |\lambda_{r+1}| + \frac{e^{\sqrt{r + p}}}{p} \sqrt{\sum_{i=r+1}^n \frac{\lambda_i^2}{\lambda_r^2} .}
\] (79)

This completes the proof.

4) Proof of Theorem 8: The proof of Theorem 8 is given below.

Proof: By using Hölder’s inequality and Theorem 6 we have
\[
E \| X - L \| \leq \left( E \| X - L \|^{2q+1} \right)^{1/(2q+1)} \\
\leq \left( E \| \tilde{X} - \tilde{L} \|^{1/(2q+1)} \right) .
\] (80)
We apply Theorem 7 to \( \tilde{X} \) and \( \tilde{L} \) and obtain the bound of \( \mathbb{E}\|\tilde{X} - \tilde{L}\| \), noting that \( \lambda_i(\tilde{X}) = \lambda_i(X)^{2q+1} \).

\[
\mathbb{E}\|\tilde{X} - \tilde{L}\| = \left( \frac{1}{p-1} \sum_{i=1}^{r} \frac{\lambda_i^{2(2q+1)} + 1}{\lambda_i^{2(2q+1)}} \right) |\lambda_r^{2q+1}| + \frac{e^{\sqrt{r} + p}}{p} \sum_{i=r+1}^{n} \frac{\lambda_i^{2(2q+1)}}{\lambda_i^{2(2q+1)}}. \tag{81}
\]

By substituting (81) into (80), we obtain the average error bound of the power scheme modification shown in Theorem 8. This completes the proof. \( \Box \)

5) Proof of Theorem 9: The following propositions from \cite{25} will be used in the proof.

**Proposition 8:** Suppose that \( h \) is a Lipschitz function on matrices:

\[
|h(X) - h(Y)| \leq L\|X - F\|_F \quad \text{for all } X, Y. \tag{82}
\]

Draw a standard Gaussian matrix \( G \). Then

\[
\Pr \{ h(G) \geq \mathbb{E}h(G) + Lt \} \leq e^{-t^2/2}. \tag{83}
\]

**Proposition 9:** Let \( G \) be a \( r \times (r+p) \) standard Gaussian matrix where \( p \geq 4 \). For all \( t \geq 1 \),

\[
\Pr \left\{ \|G^\dagger\|_F \geq \sqrt{\frac{12r}{p}} \cdot t \right\} \leq 4t^{-p} \quad \text{and}
\]

\[
\Pr \left\{ \|G^\dagger\| \geq \frac{e^{\sqrt{p} + p}}{p + 1} \cdot t \right\} \leq t^{-(p+1)}. \tag{84}
\]

The proof of Theorem 9 is given below.

**Proof:** According to the deterministic error bound in Theorem 5, we study the deviation of \( \|\Lambda_2^2 (V_2^T A_1) (V_1^T A_1)^\dagger \Lambda_1^{-1}\| \).

Consider the Lipschitz function \( h(X) = \|\Lambda_2^2 (V_1^T A_1)^\dagger \Lambda_1^{-1}\| \), its Lipschitz constant \( L \) can be estimated by using the triangle inequality:

\[
|h(X) - h(Y)| \leq \|\Lambda_2^2 (X - Y) (V_1^T A_1)^\dagger \Lambda_1^{-1}\| \\
\leq \|\Lambda_2^2\| \|X - Y\| \| (V_1^T A_1)^\dagger \| \|\Lambda_1^{-1}\| \\
\leq \|\Lambda_2^2\| \| (V_1^T A_1)^\dagger \| \|\Lambda_1^{-1}\| \|X - Y\|_F. \tag{85}
\]

Hence the Lipschitz constant satisfies \( L \leq \|\Lambda_2^2\| \| (V_1^T A_1)^\dagger \| \|\Lambda_1^{-1}\| \). We condition on \( V_1^T A_1 \) and then Proposition 6 implies that

\[
\mathbb{E} [h(V_2^T A_1) | V_1^T A_1] \leq \|\Lambda_2^2\| \| (V_1^T A_1)^\dagger \|_F \|\Lambda_1^{-1}\| + \|\Lambda_2^2\|_F \| (V_1^T A_1)^\dagger \| \|\Lambda_1^{-1}\|. \]

We define an event \( T \) as

\[
T = \left\{ \| (V_1^T A_1)^\dagger \|_F \leq \sqrt{\frac{12r}{p}} \cdot t \quad \text{and} \quad \| (V_1^T A_1)^\dagger \| \leq \frac{e^{\sqrt{p} + p}}{p + 1} \cdot t \right\}. \tag{86}
\]

May 11, 2014 DRAFT
According to Proposition 9, the event $T$ happens except with probability
\[ \Pr\{T\} \leq 4t^{-p} + t^{-(p+1)}. \] (87)

Applying Proposition 8 to the function $h(V_2^TA_1)$, given the event $T$, we have
\[
\Pr\left\{ \| \Lambda_2^2 (V_2^TA_1) (V_1^TA_1)^\dagger \Lambda_1^{-1} \| > \| \Lambda_2^2 \| \| \Lambda_1^{-1} \| \frac{\sqrt{2r}}{p} \cdot t + \| \Lambda_2^2 \| \| \Lambda_1^{-1} \| \frac{e\sqrt{r + p}}{p + 1} \cdot t \right. \\
\left. + \| \Lambda_2^2 \| \| \Lambda_1^{-1} \| \frac{e\sqrt{r + p}}{p + 1} \cdot tu \right\} \leq e^{-u^2/2}. \] (88)

According to the definition of the event $T$ and the probability of $T$, we obtain
\[
\Pr\left\{ \| \Lambda_2^2 (V_2^TA_1) (V_1^TA_1)^\dagger \Lambda_1^{-1} \| + \| \Lambda_2 \| > \right. \\
\left. \left( 1 + t \sqrt{\frac{12r}{p}} \left( \sum_{i=1}^{r} \lambda_i^{-1} \right)^{1/2} + \frac{e\sqrt{r + p}}{p + 1} \cdot tu \lambda_i^{-1} \right) \lambda_{r+1}^2 + \\
\frac{e\sqrt{r + p}}{p + 1} \cdot t \lambda_r^{-1} \left( \sum_{i=r+1}^{n} \lambda_i^2 \right)^{1/2} \right\} \leq \\
e^{-u^2/2} + 4t^{-p} + t^{-(p+1)}. \] (89)

Since Theorem 5 implies $\| X - L \| \leq \| \Lambda_2^2 (V_2^TA_1) (V_1^TA_1)^\dagger \Lambda_1^{-1} \| + \| \Lambda_2 \|$, we obtain the deviation bound in Theorem 9. This completes the proof.
APPENDIX III: ANALYSIS OF GreB

It is not direct to analyze the theoretical guarantee of GreB due to its combination of alternating minimization and greedy forward selection. Hence, we consider analyzing its convergence behavior by leveraging the results from GECO [53] analysis. This is reasonable because they share the same objective function yet different optimization variables. In particular, the risk function in GECO is $R_{\text{GECO}}[53]$ analysis. This is reasonable because they share the same objective function yet different optimization and greedy forward selection. Hence, we consider analyzing its convergence behavior by leveraging the results from the beginning of iteration $i$. It can be seen that the variable $A$ in GECO is able to be written as $A = UV$ without any loss of generality. Therefore, for the same selection of $R(A)$, we can compare the objective value of GECO and GreB at arbitrary step of their algorithm. This results in the following theorem.

Theorem 10: Assume $R(A)$ is a $\beta$-smooth function according to GECO [53] and $\epsilon > 0$, and $F(U, V) = R(UV)$ is the objective function of GreB. Given a rank constraint $r$ to $A$ and a tolerance parameter $\tau \in [0, 1)$. Let $A^* = U^*V^*$ is the solution of GreB. Then for all matrices $A = UV$ with

$$\|UV\|_F^2 \leq \frac{\epsilon(r + 1)(1 - \tau)^2}{2\beta}$$

we have $F(U^*, V^*) \leq F(U, V) + \epsilon$.

Proof: According to Lemma 3 in GECO [53], let $\epsilon_i = f(\lambda^{(i)}) - f(\bar{\lambda})$, where $\lambda^{(i)}$ is the value of $\lambda$ at the beginning of iteration $i$ and $\bar{\lambda}$ fulfills $f(\lambda) > f(\bar{\lambda})$, we have

$$f(\lambda^{(i)}) - \min_{\eta} f(\lambda^{(i)} + \eta v) \geq \frac{\epsilon_i^2 (1 - \tau)^2}{2\beta \|A\|_F^2}.$$  

(91)

At the end of iteration $i$, the objective value of GreB equals $R(UV)$, while GECO optimizes $\lambda$ over the support of $\text{span}(U) \times \text{span}(V)$ (i.e., optimizes $S$ when fixing $U$ and $V$). We use the same notation $\lambda^{(i)}$ to denote the variable in iteration $i$. This yields

$$F(U^{(i)}, V^{(i)}) = R(U^{(i)}V^{(i)}) \geq \min_S R(U^{(i)}SV^{(i)}) = f(\lambda^{(i)}).$$

(92)

At the beginning of iteration $i + 1$, both GECO and GreB computes the direction $(u, v)$ along which the objective declines fastest. However, GECO adds both $u$ and $v$ to the ranges of $U$ and $V$, while GreB only adds $v$ to $V$ and then optimizes $U$ when fixing $V$. Because the range of $U$ in GreB is optimized rather than previously fixed, we have

$$F(U^{(i+1)}, V^{(i+1)}) = \min_{\eta} F(U, [V^{(i+1)}; v]) \leq \min_{\eta} f(\lambda^{(i)} + \eta v).$$

(93)

Plug (92) and (93) into (91), we gain a similar result:

$$F(U, V) - \min_U F(U, [V; v]) \geq \frac{\epsilon_i^2 (1 - \tau)^2}{2\beta \|A\|_F^2}.$$  

(94)

Following the analysis after Lemma 3 in GECO [53], we can immediately obtain the results of the theorem. ■

The theorem states that GreB solution is at least close to optimum as GECO. Note when sparse $S$ is alternatively optimized with $UV$ in GreB scheme, such as GreBcom, the theorem can still holds. This is because after optimizing $S$ in each iteration of GreBcom, we have $P_{\Omega^c}(S + UV) = 0$, which enforces the objective function $\|M - UV - S\|_F^2$ degenerates to that of GECO, which is $\|P_{\Omega^c}(M - UV)\|_F^2$. 

May 11, 2014 DRAFT