LETTER Communicated by Guangcan Liu

Relations Among Some Low-Rank Subspace Recovery Models

Hongyang Zhang
hy_zh@pku.edu.cn
Zhouchen Lin
zlin@pku.edu.cn
Chao Zhang
chzhang@cis.pku.edu.cn
Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China, and Cooperative Medianet Center, Shanghai Jiaotong University, Shanghai 200240, China

Junbin Gao
jbgao@csu.edu.au
School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia

Recovering intrinsic low-dimensional subspaces from data distributed on them is a key preprocessing step to many applications. In recent years, a lot of work has modeled subspace recovery as low-rank minimization problems. We find that some representative models, such as robust principal component analysis (R-PCA), robust low-rank representation (R-LRR), and robust latent low-rank representation (R-LatLRR), are actually deeply connected. More specifically, we discover that once a solution to one of the models is obtained, we can obtain the solutions to other models in closed-form formulations. Since R-PCA is the simplest, our discovery makes it the center of low-rank subspace recovery models. Our work has two important implications. First, R-PCA has a solid theoretical foundation. Under certain conditions, we could find globally optimal solutions to these low-rank models at an overwhelming probability, although these models are nonconvex. Second, we can obtain significantly faster algorithms for these models by solving R-PCA first. The computation cost can be further cut by applying low-complexity randomized algorithms, for example, our novel $\ell_{2,1}$ filtering algorithm, to R-PCA. Although for the moment the formal proof of our $\ell_{2,1}$ filtering algorithm is not yet available, experiments verify the advantages of our algorithm over other state-of-the-art methods based on the alternating direction method.

Neural Computation **27, 1915–1950** (2015)
doi:10.1162/NECO_a_00762

1 Introduction

Subspaces are commonly assumed structures for high-dimensional data due to their simplicity and effectiveness. For example, motion (Tomasi & Kanade, 1992), face (Belhumeur, Hespanha, & Kriegman, 1997; Belhumeur & Kriegman, 1998; Basri & Jacobs, 2003), and texture (Ma, Derksen, Hong, & Wright, 2007) data have been known to be well characterized by low-dimensional subspaces. A lot of effort has been devoted to robustly recovering the underlying subspaces of data. The most widely adopted approach is principal component analysis (PCA). Unfortunately, PCA is known to be fragile to large noise or outliers in the data, and much work has been devoted to improving its robustness (Gnanadesikan & Kettenring, 1972; Huber, 2011; Fischler & Bolles, 1981; De La Torre & Black, 2003; Ke & Kanade, 2005; McCoy & Tropp, 2011; Zhang & Lerman, 2014; Lerman, McCoy, Tropp, & Zhang, 2014; Hardt & Moitra, 2012), among which the robust PCA (R-PCA) (Wright, Ganesh, Rao, Peng, & Ma, 2009; Chandrasekaran, Sanghavi, Parrilo, & Willsky, 2011; Candès, Li, Ma, & Wright, 2011) is one of the models with theoretical guarantee. Candès et al. (2011), Chandrasekaran et al. (2011) and Wright et al. (2009) proved that under certain conditions, the ground truth subspace can be exactly recovered with overwhelming probability. Later work (Hsu, Kakade, & Zhang, 2011) gave a justification of R-PCA in the case where the spatial pattern of the corruptions is deterministic.

Although R-PCA has found wide application, such as video denoising, background modeling, image alignment, photometric stereo, and texture representation (see e.g., Wright et al., 2009; De La Torre & Black, 2003, Ji, Liu, Shen, & Xu, 2010; Peng, Ganesh, Wright, Xu, & Ma, 2010; Zhang, Ganesh, Liang, & Ma, 2012), it only aims at recovering a single subspace that spans the entire data. To identify finer structures of data, the multiple subspaces recovery problem is considered, which aims at clustering data according to the subspaces they lie in. This problem has attracted a lot of attention in recent years (Vidal, 2011), and much work has provided a strong theoretical guarantee for the problem (see e.g., Soltanolkotabi & Candès, 2012). Rank minimization methods account for a large class of subspace clustering algorithms, where rank is connected to the dimensions of subspaces. Representative rank minimization–based methods include low-rank representation (LRR) (Liu & Yan, 2011; Liu, Lin, Yan, Sun, & Ma, 2013), robust low-rank representation (R-LRR) (Wei & Lin, 2010; Vidal & Favaro, 2014),[1] latent low-rank representation (LatLRR) (Liu, Lin, & Yu, 2010; Zhang, Lin, & Zhang, 2013), and its robust version (R-LatLRR) (Zhang, Lin, Zhang, & Gao, 2014). Subspace clustering algorithms, including these

[1] Note that Wei and Lin (2010) and Vidal and Favaro (2014) called R-LRR "robust shape interaction" (RSI) and low-rank subspace clustering (LRSC), respectively. The two models are essentially the same, differing only in the optimization algorithms. In order to remind readers that they are both robust versions of LRR by using a denoised dictionary, in this letter, we call them "robust low-rank representation (R-LRR)."

low-rank methods, have been widely applied to motion segmentation (Gear, 1998; Costeira & Kanade, 1998; Vidal & Hartley, 2004; Yan & Pollefeys, 2006; Rao, Tron, Vidal, & Ma, 2010), image segmentation (Yang, Wright, Ma, & Sastry, 2008; Cheng, Liu, Wang, Li, & Yan, 2011), face classification (Ho, Yang, Lim, Lee, & Kriegman, 2003; Vidal, Ma, & Sastry, 2005; Liu & Yan, 2011; Liu et al., 2013), and system identification (Vidal, Soatto, Ma, & Sastry, 2003; Zhang & Bitmead, 2005; Paoletti, Juloski, Ferrari-Trecate, & Vidal, 2007).

1.1 Our Contributions. In this letter, we show that some of the low-rank subspace recovery models are actually deeply connected, even though they were proposed independently and targeted different problems (single or multiple subspaces recovery). Our discoveries are based on a characteristic of low-rank recovery models: they may have closed-form solutions. Such a characteristic has not been found in sparsity-based models for subspace recovery, such as sparse subspace clustering (Elhamifar & Vidal, 2009).

There are two main contributions of this letter. First, we find a close relation between R-LRR (Wei & Lin, 2010; Vidal & Favaro, 2014) and R-PCA (Wright et al., 2009; Candès et al., 2011), showing that, surprisingly, their solutions are mutually expressible. Similarly, R-LatLRR (Zhang et al., 2014) and R-PCA are closely connected too: their solutions are also mutually expressible. Our analysis allows an arbitrary regularizer for the noise term.

Second, since R-PCA is the simplest low-rank recovery model, our analysis naturally positions it at the center of existing low-rank recovery models. In particular, we propose to first apply R-PCA to the data and then use the solution of R-PCA to obtain the solution for other models. This approach has two important implications. First, although R-LRR and R-LatLRR are nonconvex problems, under certain conditions we can obtain globally optimal solutions with an overwhelming probability (see remark 3). Namely, if the noiseless data are sampled from a union of independent subspaces and the dimension of the subspace containing the union of subspaces is much smaller than the dimension of the ambient space, we are able to recover exact subspace structure as long as the noise is sparse (even if the magnitudes of the noise are arbitrarily large). Second, solving R-PCA is much faster than solving other models. The computation cost could be further cut if we solve R-PCA by randomized algorithms. For example, we propose the $\ell_{2,1}$ filtering algorithm to solve R-PCA when the noise term uses $\ell_{2,1}$ norm (see Table 1 for a definition). Experiments verify the significant advantages of our algorithms.

The remainder of this letter is organized as follows. Section 2 reviews the representative low-rank models for subspace recovery. Section 3 gives our theoretical results—the interexpressibility among the solutions of R-PCA, R-LRR, and R-LatLRR. In section 4, we present detailed proofs of our theoretical results. Section 5 gives two implications of our theoretical analysis: better solutions and faster algorithms. We show the experimental results on both synthetic and real data in section 6. Section 7 concludes the paper.

Table 1: Summary of Main Notations.

Notations	Meanings
Capital letter	A matrix
m, n	Size of the data matrix M
$n_{(1)}, n_{(2)}$	$n_{(1)} = \max\{m, n\}$, $n_{(2)} = \min\{m, n\}$
log	Natural logarithm
I, 0, 1	The identity matrix, all-zero matrix, and all-one vector
e_i	Vector whose ith entry is 1 and others are 0
$M_{:j}$	The jth column of matrix M
M_{ij}	The entry at the ith row and jth column of matrix M
M^T	Transpose of matrix M
$M^\dagger$	Moore-Penrose pseudo-inverse of matrix M
$\|M\|$	$\|M\|_{ij} = \|M_{ij}\|, i = 1, \ldots, m, j = 1, \ldots, n$
$\|\|\cdot\|\|_2$	Euclidean norm for a vector, $\|\|v\|\|_2 = \sqrt{\sum_i v_i^2}$
$\|\|\cdot\|\|_*$	Nuclear norm of a matrix (the sum of its singular values)
$\|\|\cdot\|\|_{\ell_0}$	ℓ_0 norm of a matrix (the number of nonzero entries)
$\|\|\cdot\|\|_{\ell_{2,0}}$	$\ell_{2,0}$ norm of a matrix (the number of nonzero columns)
$\|\|\cdot\|\|_{\ell_1}$	ℓ_1 norm of a matrix, $\|\|M\|\|_{\ell_1} = \sum_{i,j} \|M_{ij}\|$
$\|\|\cdot\|\|_{\ell_{2,1}}$	$\ell_{2,1}$ norm of a matrix, $\|\|M\|\|_{\ell_{2,1}} = \sum_j \|\|M_{:j}\|\|_2$
$\|\|\cdot\|\|_{\ell_\infty}$	ℓ_∞ norm of a matrix, $\|\|M\|\|_{\ell_\infty} = \max_{ij} \|M_{ij}\|$
$\|\|\cdot\|\|_{\ell_{2,\infty}}$	$\ell_{2,\infty}$ norm of a matrix, $\|\|M\|\|_{\ell_{2,\infty}} = \max_j \|\|M_{:j}\|\|_2$
$\|\|\cdot\|\|_F$	Frobenius norm of a matrix, $\|\|M\|\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$
$\|\|\cdot\|\|$	Matrix operator norm, the largest singular value of a matrix

2 Related Work

In this section, we review a number of existing low rank models for subspace recovery.

2.1 Notations and Naming Conventions. We first define some notations. Table 1 summarizes the main notations that appear in this letter.

Since this letter involves multiple subspace recovery models, to minimize confusion, we name the models that minimize rank functions and nuclear norms as the *original* model and the *relaxed* model, respectively. We also name the models that use the denoised data matrices for dictionaries as *robust* models, with the prefix "R-".

2.2 Robust Principal Component Analysis. Robust principal component analysis (R-PCA) (Wright et al., 2009; Candès et al., 2011) is a robust version of PCA. It aims at recovering a hidden low-dimensional subspace from the observed high-dimensional data that have unknown sparse corruptions. The low-dimensional subspace and sparse corruptions

correspond to a low-rank matrix A_0 and a sparse matrix E_0, respectively. The mathematical formulation of R-PCA is as follows:

$$\min_{A,E} \operatorname{rank}(A) + \lambda ||E||_{\ell_0}, \quad \text{s.t. } X = A + E, \tag{2.1}$$

where $X = A_0 + E_0 \in \mathbb{R}^{m \times n}$ is the observation with data samples being its columns and $\lambda > 0$ is a regularization parameter.

Since solving the original R-PCA is NP-hard, which prevents the practical use of R-PCA, Candès et al. (2011) proposed solving its convex surrogate, called principal component pursuit or relaxed R-PCA by our naming conventions, defined as:

$$\min_{A,E} ||A||_* + \lambda ||E||_{\ell_1}, \quad \text{s.t. } X = A + E. \tag{2.2}$$

This relaxation makes use of two facts. First, the nuclear norm is the convex envelope of rank within the unit ball of matrix operator norm. Second, the ℓ_1 norm is the convex envelope of the ℓ_0 norm within the unit ball of the ℓ_∞ norm. Candès et al. (2011) further proved that when the rank of the structure component A_0 is $O(n_{(2)}/\log^2 n_{(1)})$, A_0 is nonsparse (see the incoherent conditions, equations 5.1 and 5.2), and the nonzeros of the noise matrix E_0 are uniformly distributed whose number is $O(mn)$ (it is remarkable that the magnitudes of noise could be arbitrarily large), then with a particular regularization parameter $\lambda = 1/\sqrt{n_{(1)}}$ the solution of the convex relaxed R-PCA problem, equation 2.2, perfectly recovers the ground truth data matrix A_0 and the noise matrix E_0 with an overwhelming probability.

2.3 Low-Rank Representation. While R-PCA works well for a single subspace with sparse corruptions, it is unable to identify multiple subspaces, the main target of the subspace clustering problem. To overcome this drawback, Liu et al. (2010, 2013) proposed the following (noiseless) low-rank representation (LRR) model:

$$\min_{Z} \operatorname{rank}(Z), \text{ s.t. } X = XZ. \tag{2.3}$$

The idea of LRR is to self-express the data, that is, using data as the dictionary, and then find the lowest-rank representation matrix. The rank measures the dimension of the sum of the subspaces, and the pattern in the optimal Z (i.e., block diagonal structure), can help identify the subspaces. To make the model robust to outliers, Liu et al. (2010, 2013) added a regularization term to the LRR model,

$$\min_{Z,E} \operatorname{rank}(Z) + \lambda ||E||_{\ell_{2,0}}, \text{ s.t. } X = XZ + E, \tag{2.4}$$

supposing that the corruptions are column sparse.

Again, due to the NP-hardness of the original LRR, Liu et al. (2010, 2013) proposed solving the relaxed LRR instead:

$$\min_{Z,E} ||Z||_* + \lambda ||E||_{\ell_{2,1}}, \text{ s.t. } X = XZ + E, \tag{2.5}$$

where the $\ell_{2,1}$ norm is the convex envelope of the $\ell_{2,0}$ norm within the unit ball of the $\ell_{2,\infty}$ norm. They proved that if the fraction of corruptions does not exceed a threshold, the row space of the ground truth Z and the indices of nonzero columns of the ground truth E can be exactly recovered (Liu et al., 2013).

2.4 Robust Low-Rank Representation (Robust Shape Interaction and Low-Rank Subspace Clustering). LRR uses the data matrix itself as the dictionary to represent data samples. This is not reasonable when the data contain severe noise or outliers. To remedy this issue, Wei and Lin (2010) suggested using denoised data as the dictionary to express itself, resulting in the following model:

$$\min_{Z,E} \text{rank}(Z) + \lambda ||E||_{\ell_{2,0}}, \text{ s.t. } X - E = (X - E)Z. \tag{2.6}$$

It is called the original robust shape interaction (RSI) model. Again, it has a relaxed version,

$$\min_{Z,E} ||Z||_* + \lambda ||E||_{\ell_{2,1}}, \text{ s.t. } X - E = (X - E)Z, \tag{2.7}$$

by replacing rank and $\ell_{2,0}$ with their respective convex envelopes.

Note that the relaxed RSI is still nonconvex due to its bilinear constraint, which may cause difficulty in finding its globally optimal solution. Wei and Lin (2010) first proved the following result on the relaxed noiseless LRR, which is also the noiseless version of the relaxed RSI.

Proposition 1. *The solution to relaxed noiseless LRR (RSI),*

$$\min_{Z} \|Z\|_*, \quad s.t. \quad A = AZ, \tag{2.8}$$

is unique and given by $Z^* = V_A V_A^T$, *where* $U_A \Sigma_A V_A^T$ *is the skinny SVD of* A.

Remark 1. $V_A V_A^T$ can also be written as $A^\dagger A$. Equation 2.8 is a relaxed version of the original noiseless LRR:

$$\min_{Z} \text{rank}(Z), \quad \text{s.t. } A = AZ. \tag{2.9}$$

$V_A V_A^T$ is called the shape interaction matrix in the field of structure from motion (Costeira & Kanade, 1998). Hence model 2.6 is named robust shape

interaction. $V_A V_A^T$ is block diagonal when the column vectors of A lie strictly on independent subspaces. The block diagonal pattern reveals the structure of each subspace and therefore offers the possibility of subspace clustering.

Wei and Lin (2010) proposed to solve the optimal A^* and E^* from

$$\min_{A,E} ||A||_* + \lambda ||E||_{\ell_{2,1}}, \quad \text{s.t. } X = A + E, \tag{2.10}$$

which we call the column sparse relaxed R-PCA since it is the convex relaxation of the original problem:

$$\min_{A,E} \operatorname{rank}(A) + \lambda ||E||_{\ell_{2,0}}, \quad \text{s.t. } X = A + E. \tag{2.11}$$

Then Wei and Lin (2010) used (Z^*, E^*) as the solution to the relaxed RSI problem (see equation 2.7), where Z^* is the shape interaction matrix of A^* by proposition 1. In this way, they deduced the optimal solution. We prove in section 4 that this is indeed true and actually holds for arbitrary functions on E.

It is worth noting that Xu, Caramanis, and Sanghavi (2012) proved that problem 2.10 is capable of exactly recognizing the sparse outliers and simultaneously recovering the column space of the ground truth data under rather broad conditions. Our previous work (Zhang, Lin, Zhang, & Chang, 2015) further showed that the parameter $\lambda = 1/\sqrt{\log n}$ guarantees the success of the model, even when the rank of the intrinsic matrix and the number of nonzero columns of the noise matrix are almost $O(n)$, where n is the column number of the input.

As a closely connected work of RSI, Favaro, Vidal, and Ravichandran (2011) and Vidal and Favaro (2014) proposed a similar model, low-rank subspace clustering (LRSC):

$$\min_{Z,A,E} ||Z||_* + \lambda ||E||_{\ell_1}, \text{ s.t. } X = A + E, A = AZ. \tag{2.12}$$

LRSC differs from RSI, equation (2.7), only by the norm of E. The other difference is that LRSC adopts the alternating direction method (ADMM) (Lin, Liu, & Su, 2011) to solve equation 2.12.

In order not to confuse readers and to highlight that both RSI and LRSC are robust versions of LRR, we call them robust LRR (R-LRR) instead as our theoretical analysis allows for arbitrary functions on E.

2.5 Robust Latent Low-Rank Representation. Although LRR and R-LRR have been successful in applications such as face recognition (Liu et al., 2010, 2013; Wei & Lin, 2010), motion segmentation (Liu et al., 2010, 2013; Favaro et al., 2011), and image classification (Bull & Gao, 2012; Zhang, Jiang, & Davis, 2013), they break down when the samples are insufficient,

especially when the number of samples is fewer than the dimensions of subspaces. Liu and Yan (2011) addressed this small sample problem by introducing hidden data X_H into the dictionary:

$$\min_R \|R\|_*, \quad \text{s.t. } X = [X, X_H]R. \tag{2.13}$$

Obviously it is impossible to solve problem 2.13 because X_H is unobserved. Nevertheless, by utilizing proposition 1, Liu and Yan (2011) proved that X can be written as $X = XZ + LX$, where both Z and L are low rank, resulting in the following latent low-rank representation (LatLRR) model,

$$\min_{Z,L,E} \text{rank}(Z) + \text{rank}(L) + \lambda||E||_{\ell_0}, \text{ s.t. } X = XZ + LX + E, \tag{2.14}$$

where sparse corruptions are considered. As a common practice, its relaxed version is solved instead:

$$\min_{Z,L,E} ||Z||_* + ||L||_* + \lambda||E||_{\ell_1}, \text{ s.t. } X = XZ + LX + E. \tag{2.15}$$

As in the case of LRR, when the data are very noisy or highly corrupted, it is inappropriate to use X itself as the dictionary. So Zhang et al. (2014) borrowed the idea of R-LRR to use denoised data as the dictionary, giving rise to the following robust latent LRR (R-LatLRR) model,

$$\begin{aligned} &\min_{Z,L,E} \text{rank}(Z) + \text{rank}(L) + \lambda||E||_{\ell_0}, \\ &\text{s.t. } X - E = (X - E)Z + L(X - E), \end{aligned} \tag{2.16}$$

and its relaxed version,

$$\begin{aligned} &\min_{Z,L,E} ||Z||_* + ||L||_* + \lambda||E||_{\ell_1}, \\ &\text{s.t. } X - E = (X - E)Z + L(X - E). \end{aligned} \tag{2.17}$$

Again the relaxed R-LatLRR model is nonconvex. Zhang, Lin et al. (2013) proved that when there is no noise, both the original R-LatLRR and relaxed R-LatLRR have nonunique closed-form solutions, and they described the complete solution sets, among which there are a large number of inappropriate solutions for subspace clustering. In order to choose the appropriate solution, according to Wright, Ma, Mairal, Sapiro, and Huang (2010), an informative similarity matrix should have an adaptive neighborhood, high discriminating power, and high sparsity (Zhuang et al., 2012). As the graphs constructed by LatLRR have had high discriminating power and

adaptive neighborhood (Liu & Yan, 2011), Zhang, Lin et al. (2013) considered using high sparsity to choose the optimal solution Z^*. In particular, like RSI, Zhang, Lin et al. (2013) proposed applying R-PCA to separate X into $X = A^* + E^*$. Next, they found the sparsest solution among the solution set of relaxed noiseless R-LatLRR,

$$\min_{Z,L} ||Z||_* + ||L||_*, \quad \text{s.t. } A = AZ + LA, \tag{2.18}$$

with A being A^*. Equation 2.18 is a relaxed version of the original noiseless R-LatLRR model:

$$\min_{Z,L} \text{rank}(Z) + \text{rank}(L), \quad \text{s.t. } A = AZ + LA. \tag{2.19}$$

In section 4, we prove that the above two-step procedure solves equation 2.17 correctly. More in-depth analysis will also be provided.

2.6 Other Low-Rank Models for Subspace Clustering. In this section, we mention more low-rank subspace recovery models, although they are not our focus in this letter. Also aiming at addressing the small sample issue, Liu et al. (2012) proposed fixed rank representation by requiring that the representation matrix be as close to a rank r matrix as possible, where r is a prescribed rank. Then the best rank r matrix, which still has a block diagonal structure, is used for subspace clustering. Wang, Saligrama, and Castañón (2011) extended LRR to address nonlinear multimanifold segmentation, where the error E is regularized by the square of Frobenius norm so that the kernel trick can be used. Ni, Sun, Yuan, Yan, and Cheong (2010) augmented the LRR model with a semidefiniteness constraint on the representation matrix Z. In contrast, the representation matrices by R-LRR and R-LatLRR are both naturally semidefinite as they are shape interaction matrices.

3 Main Results: Relations Among Low-Rank Models

In this section, we present the hidden connections among representative low-rank recovery models—R-PCA, R-LRR, and R-LatLRR—although they appear different and have been proposed for different purposes. Actually, our analysis holds for more general models where the regularization on noise term E can be arbitrary. More specifically, the generalized models are:

$$\min_{A,E} \text{rank}(A) + \lambda f(E), \quad \text{s.t. } X = A + E, \tag{3.1}$$

$$\min_{A,E} \|A\|_* + \lambda f(E), \quad \text{s.t. } X = A + E, \tag{3.2}$$

$$\min_{Z,E} \text{rank}(Z) + \lambda f(E), \text{ s.t. } X - E = (X - E)Z, \tag{3.3}$$

$$\min_{Z,E} ||Z||_* + \lambda f(E), \text{ s.t. } X - E = (X - E)Z, \tag{3.4}$$

$$\min_{Z,L,E} \text{rank}(Z) + \text{rank}(L) + \lambda f(E),$$
$$\text{s.t. } X - E = (X - E)Z + L(X - E), \tag{3.5}$$

$$\min_{Z,L,E} ||Z||_* + ||L||_* + \lambda f(E), \text{ s.t. } X - E = (X - E)Z + L(X - E), \tag{3.6}$$

where f is any function. For brevity, we still call equations 3.1 to 3.6 the original R-PCA, relaxed R-PCA, original R-LRR, relaxed R-LRR, original R-LatLRR, and relaxed R-LatLRR, respectively, without mentioning "generalized."

We show that the solutions to the above models are mutually expressible; if we have a solution to one of the models, we will obtain the solutions to other models in closed-form formulations. We further show in section 5 that such mutual expressibility is useful.

It suffices to show that the solutions of the original R-PCA and those of other models are mutually expressible (i.e., letting the original R-PCA hinge all the above models). We summarize our results as the following theorems.

Theorem 1 *(connection between the original R-PCA and the original R-LRR). For any minimizer (A^*, E^*) of the original R-PCA problem, equation 3.1, suppose $U_{A^*}\Sigma_{A^*}V_{A^*}^T$ is the skinny SVD of the matrix A^*. Then $((A^*)^\dagger A^* + SV_{A^*}^T, E^*)$ is the optimal solution to the original R-LRR problem, equation 3.3, where S is any matrix such that $V_{A^*}^T S = 0$. Conversely, provided that (Z^*, E^*) is an optimal solution to the original R-LRR problem, equation 3.3, $(X - E^*, E^*)$ is a minimizer of the original R-PCA problem, equation 3.1.*

Theorem 2 *(connection between the original R-PCA and the relaxed R-LRR) . For any minimizer (A^*, E^*) of the original R-PCA problem, equation 3.1, the relaxed R-LRR problem, equation 3.4, has an optimal solution $((A^*)^\dagger A^*, E^*)$. Conversely, suppose that the relaxed R-LRR problem, equation 3.4, has a minimizer (Z^*, E^*); then $(X - E^*, E^*)$ is an optimal solution to the original R-PCA problem, equation 3.1.*

Remark 2. According to theorem 2, the relaxed R-LRR can be viewed as denoising the data first by the original R-PCA and then adopting the shape interaction matrix of the denoised data matrix as the affinity matrix. Such a procedure is exactly the same as that in Wei and Lin (2010), which was proposed out of heuristics and for which no proof was provided.

Theorem 3 *(connection between the original R-PCA and the original R-LatLRR). Let the pair (A^*, E^*) be any optimal solution to the original R-PCA problem, equation 3.1. Then the original R-LatLRR model, equation 3.5, has minimizers*

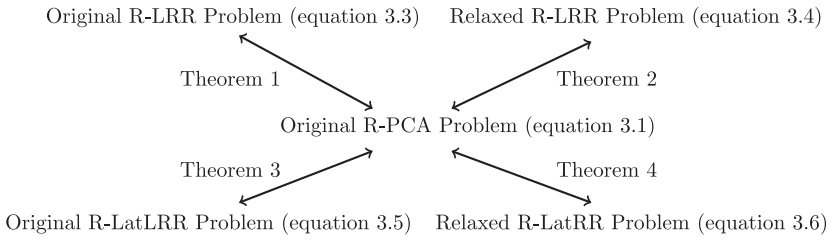

Figure 1: Visualization of the relationship among problems 3.3, 3.4, 3.5, 3.6, and 3.1, where an arrow means that a solution to one problem could be used to express a solution (or solutions) to the other problem in a closed form.

(Z^*, L^*, E^*), *where*

$$\begin{aligned} Z^* &= V_{A^*}\widetilde{W}V_{A^*}^T + S_1\widetilde{W}V_{A^*}^T, \\ L^* &= U_{A^*}\Sigma_{A^*}(I-\widetilde{W})\Sigma_{A^*}^{-1}U_{A^*}^T + U_{A^*}\Sigma_{A^*}(I-\widetilde{W})S_2. \end{aligned} \tag{3.7}$$

$\widetilde{W}$ *is any idempotent matrix and* S_1 *and* S_2 *are any matrices satisfying:*

1. $V_A^T S_1 = 0$ *and* $S_2 U_A = 0$
2. $Rank(S_1) \leq rank(\widetilde{W})$ *and* $rank(S_2) \leq rank(I-\widetilde{W})$.

Conversely, let (Z^*, L^*, E^*) *be any optimal solution to the original R-LatLRR, equation 3.5. Then* $(X-E^*, E^*)$ *is a minimizer of the original R-PCA problem, equation 3.1.*

Theorem 4 *(connection between the original R-PCA and the relaxed R-LatLRR). Let the pair* (A^*, E^*) *be any optimal solution to the original R-PCA problem, equation 3.1. Then the relaxed R-LatLRR model, equation 3.6, has minimizers* (Z^*, L^*, E^*), *where*

$$Z^* = V_{A^*}\widehat{W}V_{A^*}^T, \quad L^* = U_{A^*}(I-\widehat{W})U_{A^*}^T, \tag{3.8}$$

and $\widehat{W}$ *is any block diagonal matrix satisfying:*

1. *Its blocks are compatible with* Σ_{A^*}, *i.e., if* $[\Sigma_{A^*}]_{ii} \neq [\Sigma_{A^*}]_{jj}$ *then* $[\widehat{W}]_{ij}=0$
2. *Both* $\widehat{W}$ *and* $I-\widehat{W}$ *are positive semidefinite.*

Conversely, let (Z^*, L^*, E^*) *be any optimal solution to the relaxed R-LatLRR, equation 3.6. Then* $(X-E^*, E^*)$ *is a minimizer of the original R-PCA problem, equation 3.1.*

Figure 1 illustrates our theorems by putting the original R-PCA at the center of the low-rank subspace clustering models under consideration.

By the above theorems, we easily have the following corollary:

Corollary 1. *The solutions to the original R-PCA, equation 3.1, original R-LRR, equation 3.3, relaxed R-LRR, equation 3.4, original R-LatLRR, equation 3.5, and relaxed R-LatLRR, equation 3.6 are all mutually expressible.*

Remark 3. According to the above results, once we obtain a globally optimal solution to the original R-PCA, equation 3.1, we can obtain globally optimal solutions to the original and relaxed R-LRR and R-LatLRR problems. Although in general solving the original R-PCA is NP hard, under certain conditions (see section 5.1), its globally optimal solution can be obtained with an overwhelming probability by solving the relaxed R-PCA, equation 3.2. If one solves the original and relaxed R-LRR or R-LatLRR directly (e.g., by ADMM), there is no analysis on whether their globally optimal solutions can be attained due to their nonconvex nature. In this sense, we say that we can obtain a better solution for the original and relaxed R-LRR and R-LatLRR if we reduce them to the original R-PCA. Our numerical experiments in section 6.1 testify to our claims.

4 Proofs of Main Results

In this section, we provide detailed proofs of the four theorems in the previous section.

4.1 Connection between R-PCA and R-LRR.

The following lemma is useful throughout the proof of theorem 1.

Lemma 1 *(Zhang, Lin et al., 2013). Suppose $U_A\Sigma_A V_A^T$ is the skinny SVD of A. Then the complete solutions to equation 2.9 are $Z^* = A^\dagger A + SV_A^T$, where S is any matrix such that $V_A^T S = 0$.*

Using lemma 1, we can prove theorem 1.

Proof of Theorem 1. We first prove the first part of the theorem. Since (A^*, E^*) is a feasible solution to problem 3.1, it is easy to check that $((A^*)^\dagger A^* + SV_{A^*}^T, E^*)$ is also feasible for equation 3.3 by using a fundamental property of Moore-Penrose pseudo-inverse: $YY^\dagger Y = Y$. Now suppose that $((A^*)^\dagger A^* + SV_{A^*}^T, E^*)$ is not an optimal solution to equation 3.3. Then there exists an optimal solution to it, denoted by $(\widetilde{Z}, \widetilde{E})$, such that

$$\operatorname{rank}(\widetilde{Z}) + \lambda f(\widetilde{E}) < \operatorname{rank}((A^*)^\dagger A^* + SV_{A^*}^T) + \lambda f(E^*). \tag{4.1}$$

Meanwhile $(\widetilde{Z}, \widetilde{E})$ is feasible: $X - \widetilde{E} = (X - \widetilde{E})\widetilde{Z}$. Since $(\widetilde{Z}, \widetilde{E})$ is optimal for problem 3.3, by lemma 1, we fix $\widetilde{E}$ and have

$$\begin{aligned}\operatorname{rank}(\widetilde{Z}) + \lambda f(\widetilde{E}) &= \operatorname{rank}((X - \widetilde{E})^\dagger (X - \widetilde{E})) + \lambda f(\widetilde{E}) \\ &= \operatorname{rank}(X - \widetilde{E}) + \lambda f(\widetilde{E}).\end{aligned} \tag{4.2}$$

On the other hand,

$$\text{rank}((A^*)^\dagger A^* + SV_{A^*}^T) + \lambda f(E^*) = \text{rank}(A^*) + \lambda f(E^*). \tag{4.3}$$

From equations 4.1 to 4.3, we have

$$\text{rank}(X - \widetilde{E}) + \lambda f(\widetilde{E}) < \text{rank}(A^*) + \lambda f(E^*), \tag{4.4}$$

which leads to a contradiction with the optimality of (A^*, E^*) to R-PCA, equation 3.1.

We then prove the converse, also by contradiction. Suppose that (Z^*, E^*) is a minimizer to the original R-LRR problem, equation 3.3, while $(X - E^*, E^*)$ is not a minimizer to the R-PCA problem, equation 3.1. Then there will be a better solution to problem 3.1, termed $(\widetilde{A}, \widetilde{E})$, which satisfies

$$\text{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) < \text{rank}(X - E^*) + \lambda f(E^*). \tag{4.5}$$

Fixing E as E^* in equation 3.3, by lemma 1 and the optimality of Z^*, we infer that

$$\begin{aligned}\text{rank}(X - E^*) + \lambda f(E^*) &= \text{rank}((X - E^*)^\dagger (X - E^*)) + \lambda f(E^*) \\ &= \text{rank}(Z^*) + \lambda f(E^*).\end{aligned} \tag{4.6}$$

On the other hand,

$$\text{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) = \text{rank}(\widetilde{A}^\dagger \widetilde{A}) + \lambda f(\widetilde{E}), \tag{4.7}$$

where we have utilized another property of Moore-Penrose pseudo-inverse: rank($Y^\dagger Y$)=rank(Y). Combining equations 4.5 to 4.7, we have

$$\text{rank}(\widetilde{A}^\dagger \widetilde{A}) + \lambda f(\widetilde{E}) < \text{rank}(Z^*) + \lambda f(E^*). \tag{4.8}$$

Notice that $(\widetilde{A}^\dagger \widetilde{A}, \widetilde{E})$ satisfies the constraint of the original R-LRR problem, equation 3.3, due to $\widetilde{A} + \widetilde{E} = X$ and $\widetilde{A}(\widetilde{A}^\dagger \widetilde{A}) = \widetilde{A}$. The inequality, equation 4.8 leads to a contradiction with the optimality of the pair (Z^*, E^*) for R-LRR.

Thus we finish the proof.

Now we prove theorem 2. Proposition 1 is critical for the proof.

Proof of Theorem 2. We first prove the first part of the theorem. Obviously, according to the conditions of the theorem, $((A^*)^\dagger A^*, E^*)$ is a feasible solution to problem 3.4. Now suppose it is not optimal, and the optimal solution to problem 3.4 is $(\widetilde{Z}, \widetilde{E})$. So we have

$$||\widetilde{Z}||_* + \lambda f(\widetilde{E}) < ||(A^*)^\dagger A^*||_* + \lambda f(E^*). \tag{4.9}$$

Viewing the noise E as a fixed matrix, by proposition 1 we have

$$\begin{aligned}||\widetilde{Z}||_* + \lambda f(\widetilde{E}) &= ||(X-\widetilde{E})^\dagger (X-\widetilde{E})||_* + \lambda f(\widetilde{E}) \\ &= \text{rank}(X-\widetilde{E}) + \lambda f(\widetilde{E}). \end{aligned} \tag{4.10}$$

On the other hand, $||(A^*)^\dagger A^*||_* + \lambda f(E^*) = \text{rank}(A^*) + \lambda f(E^*)$. So we derive

$$\text{rank}(X-\widetilde{E}) + \lambda f(\widetilde{E}) < \text{rank}(A^*) + \lambda f(E^*). \tag{4.11}$$

This is a contradiction because (A^*, E^*) has been an optimal solution to the R-PCA problem 3.1, thus proving the first part of the theorem.

Next, we prove the second part of the theorem. Similarly, suppose $(X - E^*, E^*)$ is not the optimal solution to the R-PCA problem, equation 3.1. Then there exists a pair $(\widetilde{A}, \widetilde{E})$ that is better;

$$\text{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) < \text{rank}(X-E^*) + \lambda f(E^*). \tag{4.12}$$

On one hand, $\text{rank}(X-E^*) + \lambda f(E^*) = ||(X-E^*)^\dagger (X-E^*)||_* + \lambda f(E^*)$. On the other hand, $\text{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) = ||\widetilde{A}^\dagger \widetilde{A}||_* + \lambda f(\widetilde{E})$. Notice that the pair $(\widetilde{A}^\dagger \widetilde{A}, \widetilde{E})$ is feasible for the relaxed R-LRR, equation 3.4. Thus we have a contradiction.

4.2 Connection between R-PCA and R-LatLRR. Now we prove the mutual expressibility between the solutions of R-PCA and R-LatLRR. Our previous work (Zhang, Lin et al., 2013) gives the complete closed-form solutions to noiseless R-LatLRR problems 2.19 and 2.18, which are both critical to our proofs.

Lemma 2 *(Zhang, Lin et al., 2013). Suppose $U_A \Sigma_A V_A^T$ is the skinny SVD of a denoised data matrix A. Then the complete solutions to the original noiseless R-LatLRR problem, equation 2.19, are as follows:*

$$\begin{aligned} &Z^* = V_A \widetilde{W} V_A^T + S_1 \widetilde{W} V_A^T \text{ and} \\ &L^* = U_A \Sigma_A (I - \widetilde{W}) \Sigma_A^{-1} U_A^T + U_A \Sigma_A (I - \widetilde{W}) S_2, \end{aligned} \tag{4.13}$$

where $\widetilde{W}$ is any idempotent matrix and S_1 and S_2 are any matrices satisfying:

1. *$V_A^T S_1 = 0$ and $S_2 U_A = 0$*
2. *$Rank(S_1) \leq rank(\widetilde{W})$, and $rank(S_2) \leq rank(I - \widetilde{W})$.*

Now we are ready to prove theorem 3.

Proof of Theorem 3. We first prove the first part of the theorem. Since equation 4.13 is the minimizer to problem 2.19 with $A = A^*$, it naturally satisfies

the constraint: $A^* = A^*Z^* + L^*A^*$. Together with the fact that $A^* = X - E^*$ based on the assumption of the theorem, we conclude that (Z^*, L^*, E^*) satisfies the constraint of the original R-LatLRR, equation 3.5.

Now suppose that there exists a better solution, termed $(\widetilde{Z}, \widetilde{L}, \widetilde{E})$, than (Z^*, L^*, E^*) for equation 3.5, which satisfies the constraint

$$X - \widetilde{E} = (X - \widetilde{E})\widetilde{Z} + \widetilde{L}(X - \widetilde{E})$$

and has a lower objective function value:

$$\operatorname{rank}(\widetilde{Z}) + \operatorname{rank}(\widetilde{L}) + \lambda f(\widetilde{E}) < \operatorname{rank}(Z^*) + \operatorname{rank}(L^*) + \lambda f(E^*). \quad (4.14)$$

Without loss of generality, we assume that $(\widetilde{Z}, \widetilde{L}, \widetilde{E})$ is optimal to equation 3.5. Then according to lemma 2, by fixing $\widetilde{E}$ and E^*, respectively, we have

$$\operatorname{rank}(\widetilde{Z}) + \operatorname{rank}(\widetilde{L}) + \lambda f(\widetilde{E}) = \operatorname{rank}(X - \widetilde{E}) + \lambda f(\widetilde{E}), \quad (4.15)$$

$$\operatorname{rank}(Z^*) + \operatorname{rank}(L^*) + \lambda f(E^*) = \operatorname{rank}(X - E^*) + \lambda f(E^*). \quad (4.16)$$

From equations 4.14 to 4.16, we finally obtain

$$\operatorname{rank}(X - \widetilde{E}) + \lambda f(\widetilde{E}) < \operatorname{rank}(X - E^*) + \lambda f(E^*), \quad (4.17)$$

which leads to a contradiction with our assumption that (A^*, E^*) is optimal for R-PCA.

We then prove the converse. Similarly, suppose that $(\widetilde{A}, \widetilde{E})$ is a better solution than $(X - E^*, E^*)$ for R-PCA, equation 3.1. Then

$$\begin{aligned} \operatorname{rank}(\widetilde{A}^\dagger\widetilde{A}) + \operatorname{rank}(0) + \lambda f(\widetilde{E}) &= \operatorname{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) \\ &< \operatorname{rank}(X - E^*) + \lambda f(E^*) \quad (4.18) \\ &= \operatorname{rank}(Z^*) + \operatorname{rank}(L^*) + \lambda f(E^*), \end{aligned}$$

where the last equality holds since (Z^*, L^*, E^*) is optimal to equation 3.5 and its corresponding minimum objective function value is $\operatorname{rank}(X - E^*) + \lambda f(E^*)$. Since $(\widetilde{A}^\dagger\widetilde{A}, 0, \widetilde{E})$ is feasible for the original R-LatLRR, equation 3.5, we obtain a contradiction with the optimality of (Z^*, L^*, E^*) for R-LatLRR.

The following lemma is helpful for proving the connection between the R-PCA, equation 3.1, and the relaxed R-LatLRR, equation 3.6.

Lemma 3 *(Zhang, Lin et al., 2013). Suppose $U_A\Sigma_A V_A^T$ is the skinny SVD of a denoised data matrix A. Then the complete optimal solutions to the relaxed noiseless R-LatLRR problem, equation 2.18 are as follows:*

$$Z^* = V_A\widehat{W}V_A^T \text{ and } L^* = U_A(I - \widehat{W})U_A^T, \quad (4.19)$$

where $\widehat{W}$ is any block diagonal matrix satisfying:

1. *Its blocks are compatible with Σ_A, that is, if $[\Sigma_A]_{ii} \neq [\Sigma_A]_{jj}$ then $[\widehat{W}]_{ij}$=0.*
2. *Both $\widehat{W}$ and $I - \widehat{W}$ are positive semidefinite.*

Now we are ready to prove theorem 4.

Proof of Theorem 4. Suppose $(\widetilde{Z}, \widetilde{L}, \widetilde{E})$ is a better solution than (Z^*, L^*, E^*) to the relaxed R-LatLRR, equation 3.6:

$$||\widetilde{Z}||_* + ||\widetilde{L}||_* + \lambda f(\widetilde{E}) < ||Z^*||_* + ||L^*||_* + \lambda f(E^*). \tag{4.20}$$

Without loss of generality, we assume $(\widetilde{Z}, \widetilde{L}, \widetilde{E})$ is the optimal solution to equation 3.6. So according to lemma 3, $(\widetilde{Z}, \widetilde{L}, \widetilde{E})$ can be written as the form 4.19:

$$\widetilde{Z} = V_A \widehat{W} V_A^T \text{ and } \widetilde{L} = U_A (I - \widehat{W}) U_A^T, \tag{4.21}$$

where $A = X - \widetilde{E}$ and $\widehat{W}$ satisfies all the conditions in lemma 3. Taking equation 4.21 into the objective function of problem 3.6, we have

$$||\widetilde{Z}||_* + ||\widetilde{L}||_* + \lambda f(\widetilde{E}) = \text{rank}(X - \widetilde{E}) + \lambda f(\widetilde{E}), \tag{4.22}$$

where conditions 1 and 2 in lemma 3 guarantee $||\widetilde{Z}||_* + ||\widetilde{L}||_* = \text{rank}(A) = \text{rank}(X - \widetilde{E})$. On the other hand, taking equation 3.8 into the objective function of problem 3.6 and using conditions 1 and 2 in the theorem, we have

$$||Z^*||_* + ||L^*||_* + \lambda f(E^*) = \text{rank}(X - E^*) + \lambda f(E^*). \tag{4.23}$$

Thus we obtain a contradiction by considering equations 4.20, 4.22, and 4.23.

Conversely, suppose the R-PCA problem, equation 3.1, has a better solution $(\widetilde{A}, \widetilde{E})$ than $(X - E^*, E^*)$:

$$\text{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) < \text{rank}(X - E^*) + \lambda f(E^*). \tag{4.24}$$

On one hand, we have

$$\text{rank}(\widetilde{A}) + \lambda f(\widetilde{E}) = ||\widetilde{A}^\dagger \widetilde{A}||_* + ||0||_* + \lambda f(\widetilde{E}). \tag{4.25}$$

On the other hand, since (Z^*, L^*, E^*) is optimal to the relaxed R-LatLRR, equation 3.6, it can be written as

$$Z^* = V_A \widehat{W} V_A^T \text{ and } L^* = U_A (I - \widehat{W}) U_A^T, \tag{4.26}$$

with conditions 1 and 2 in lemma 4.2 satisfied, where $A = X - E^*$. Taking equation 4.26 into the objective function of problem 3.6, we have

$$\text{rank}(X - E^*) + \lambda f(E^*) = ||Z^*||_* + ||L^*||_* + \lambda f(E^*), \tag{4.27}$$

where conditions 1 and 2 in lemma 4.2 guarantee that equation 4.27 holds. So the inequality follows:

$$||\widetilde{A}^\dagger \widetilde{A}||_* + ||0||_* + \lambda f(\widetilde{E}) < ||Z^*||_* + ||L^*||_* + \lambda f(E^*), \tag{4.28}$$

which is contradictory to the optimality of the (Z^*, L^*, E^*) to the relaxed R-LatLRR, equation 3.6.

Finally, viewing R-PCA as a hinge, we connect all the models considered in section 3. We now prove corollary 1.

Proof of Corollary 1. According to theorems 1, 2, 3, and 4, the solutions to R-PCA and those of other models are mutually expressible. Next, we build the relationships among equations 3.3 to 3.6. For simplicity, we take only equations 3.3 and 3.4 as example. The proofs of the remaining connections are similar.

Suppose (Z^*, E^*) is optimal to the original R-LRR problem, equation 3.3. Then, based on theorem 1, $(X - E^*, E^*)$ is an optimal solution to the R-PCA problem, equation 3.1. Then theorem 2 concludes that $((X - E^*)^\dagger (X - E^*), E^*)$ is a minimizer of the relaxed R-LRR problem, equation 3.4. Conversely, suppose that (Z^*, E^*) is optimal to the relaxed R-LRR problem. By theorems 1 and 2, we conclude that $((X - E^*)^\dagger (X - E^*) + SV_{X-E^*}^T, E^*)$ is an optimal solution to the original R-LRR problem, equation 3.3, where V_{X-E^*} is the matrix of right singular vectors in the skinny SVD of $X - E^*$ and S is any matrix satisfying $V_{X-E^*}^T S = 0$.

5 Applications of the Theoretical Analysis

In this section, we discuss pragmatic values of our theoretical results in section 3. As one can see in Figure 1, we put R-PCA at the center of all the low-rank models under consideration because it is the simplest one, which implies that we prefer deriving solutions of other models from that of R-PCA. For simplicity, we turn to our two-step approach, first reducing to R-PCA and then expressing the desired solution by the solution of R-PCA, as REDU-EXPR method. There are two advantages of REDU-EXPR. First, we could obtain better solutions to other low-rank models (see remark 3). R-PCA has a solid theoretical foundation. Candès et al. (2011) proved that under certain conditions, solving the relaxed R-PCA, equation 2.2, which is convex, can recover the ground truth solution at an overwhelming probability (see section 5.1.1). Xu et al. (2012) and Zhang et al. (2015) also

proved similar results for column sparse relaxed R-PCA, equation 2.10 (see section 5.1.2). Then by the mutualexpressibility of solutions, we could also obtain globally optimal solutions to other models. In contrast, the optimality of a solution is uncertain if we solve other models using specific algorithms, such as, ADMM (Lin et al., 2011), due to their nonconvex nature.

The second advantage is that we could have much faster algorithms for other low-rank models. Due to the simplicity of R-PCA, solving R-PCA is much faster than other models. In particular, the expensive $O(mn^2)$ complexity of matrix-matrix multiplication (between X and Z or L) could be avoided. Moreover, there are low-complexity randomized algorithms for solving R-PCA, making the computational cost of solving other models even lower. In particular, we propose an $\ell_{2,1}$ filtering algorithm for column sparse relaxed R-PCA (equation 3.2 with $f(E) = \|E\|_{\ell_{2,1}}$). If one is directly faced with other models, it is nontrivial to design low-complexity algorithms (either deterministic or randomized).[2]

In summary, based on our analysis, we could achieve low rankness–based subspace clustering with better performance and faster speed.

5.1 Better Solution for Subspace Recovery. Reducing to R-PCA could help overcome the nonconvexity issue of the low-rank recovery models we consider (see remark 3). We defer the numerical verification of this claim until section 6.1. In this section, we discuss the theoretical conditions under which reducing to R-PCA succeeds for the subspace clustering problem.

We focus on the application of theorem 2, which shows that given the solution (A^*, E^*) to R-PCA problem 3.1, the optimal solution to the relaxed R-LRR problem 3.4, is presented by $((A^*)^\dagger A^*, E^*)$. Note that $(A^*)^\dagger A^*$ is called the shape interaction matrix in the field of structure from motion and has been proven to be block diagonal by Costeira and Kanade (1998) when the column vectors of A^* lie strictly on independent subspaces and the sampling number of A^* from each subspace is larger than the subspace dimension (Liu et al., 2013). The block diagonal pattern reveals the structure of each subspace and hence offers the possibility of subspace clustering. Thus, to illustrate the success of our approach, we show under which conditions the R-PCA problem exactly recovers the noiseless data matrix or correctly recognizes the indices of noise. We discuss the cases where the corruptions are sparse element-wise noise, sparse column-wise noise, and dense gaussian noise, respectively. As for the application of theorems 1, 3, and 4, given that the solutions to problems 3.3, 3.5, and 3.6 are all nonunique and thus

[2]We emphasize that although there is a linear time SVD algorithm (Avron, Maymounkov, & Toledo, 2010; Mahoney, 2011) for computing SVD at low cost, which is typically needed in the existing solvers for all models, linear time SVD is known to have relative error. Moreover, even adopting linear time SVD, the whole complexity could still be $O(mn^2)$ due to matrix-matrix multiplications outside the SVD, computation in each iteration if there is no careful treatment.

it is possible for an optimal solution to have better or worse performance (e.g., clustering accuracy) than another optimal solution, one should adopt another criterion (e.g., the sparsity constraint) to select the most suitable solution for specific application tasks (see e.g., Zhang et al., 2014).

5.1.1 Sparse Element-Wise Noises. Suppose each column of the data matrix is an observation. In the case where the corruptions are sparse element-wise noise, we assume that the positions of the corrupted elements sparsely and uniformly distribute on the input matrix. In this case, we consider the use of the ℓ_1 norm, model 2.2, to remove the corruption.

Candès et al. (2011) gave certain conditions under which model 2.2 exactly recovers the noiseless data A_0 from the corrupted observations $X = A_0 + E_0 \in \mathbb{R}^{m\times n}$. We apply them to the success conditions of our approach. First, to avoid the possibility that the low-rank part A_0 is sparse, A_0 needs to satisfy the following incoherent conditions:

$$\max_i ||V_0^T e_i||_2 \leq \sqrt{\frac{\mu r}{n}}, \tag{5.1}$$

$$\max_i ||U_0^T e_i||_2 \leq \sqrt{\frac{\mu r}{m}}, \quad ||U_0 V_0^T||_\infty \leq \sqrt{\frac{\mu r}{mn}}, \tag{5.2}$$

where $U_0\Sigma_0 V_0^T$ is the skinny SVD of A_0, $r = \text{rank}(A_0)$, and μ is a constant. The second assumption for the success of the algorithm is that the dimension of the sum of the subspaces is sufficiently low and the support number s of the noise matrix E_0 is not too large, namely,

$$\text{rank}(A_0) \leq \rho_r \frac{n_{(2)}}{\mu(\log n_{(1)})^2} \quad \text{and} \quad s \leq \rho_s mn, \tag{5.3}$$

where ρ_r and ρ_s are numerical constants, $n_{(1)} = \max\{m, n\}$ and $n_{(2)} = \min\{m, n\}$. Under these conditions, Candès et al. (2011) justified that relaxed R-PCA, equation 2.2, with $\lambda = 1/\sqrt{n_{(1)}}$ exactly recovers the noiseless data A_0. Thus, the algorithm of reducing to R-PCA succeeds as long as the subspaces are independent and the sampling number from each subspace is larger than the subspace dimension (Liu et al., 2013).

5.1.2 Sparse Column-Wise Noise. In the more general case, the noise exists in a small number of columns, each nonzero column of E_0 corresponds to a corruption. In this case, we consider the use of the $\ell_{2,1}$ norm, model 2.10 to remove the corruption.

Several articles have investigated the theoretical conditions under which column sparse relaxed R-PCA, equation 2.10, succeeds (Xu et al., 2012; Chen, Xu, Caramanis, & Sanghavi, 2011; Zhang et al., 2015). With slightly stronger

conditions, our discovery in Zhang et al. (2015) gave tight recovery bounds under which model 2.10 exactly identifies the indices of noise. Notice that it is impossible to recover a corrupted sample into its right subspace, since the magnitude of noise here can be arbitrarily large. Moreover, for observations like

$$M = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{5.4}$$

where the first column is the corrupted sample while others are noiseless, it is even harder to identify that the ground truth of the first column of M belongs to the space Range(e_1) or the space Range(e_2). So we remove the corrupted observation identified by the algorithm rather than exactly recovering its ground truth and use the remaining noiseless data to reveal the real structure of the subspaces.

According to our discovery (Zhang et al., 2015), the success of model 2.10 requires incoherence as well. However, only condition 5.1 is needed, which is sufficient to guarantee that the low-rank part cannot be column sparse. Similarly, to avoid the column-sparse part being low rank when the number of its nonzero columns is comparable to n, we assume $||\mathcal{B}(E_0)|| \leq \sqrt{\log n}/4$, where $\mathcal{B}(E_0) = \{H : \mathcal{P}_{\mathcal{I}^\perp}(H) = \mathbf{0}; H_{:j} = [E_0]_{:j}/||[E_0]_{:j}||_2, [E_0]_{:j} \in \mathcal{I}\}$, $\mathcal{I} = \{j : [E_0]_{:j} \neq \mathbf{0}\}$, and $\mathcal{P}_{\mathcal{I}^\perp}$ is a projection onto the complement of $\mathcal{I}$. Note that, for example, when the columns of E_0 are independent subgaussian isotropic random vectors, the constraint holds. So the constraint is feasible, even though the number of the nonzero columns of E_0 is comparable to n. The dimension of the sum of the subspaces is also required to be low and the column support number s of the noise matrix E_0 to not be too large. More specifically,

$$\text{rank}(A_0) \leq \rho_r \frac{n_{(2)}}{\mu \log n} \quad \text{and} \quad s \leq \rho_s n, \tag{5.5}$$

where ρ_r and ρ_s are numerical constants. Note that the range of the successful rank(A_0) in equation 5.5 is broader than that of equation 5.3, and has been proved to be tight (Zhang et al., 2015). Moreover, to avoid $[A_0 + E_0]_{:j}$ lying in an incorrect subspace, we assume $[E_0]_{:j} \notin \text{Range}(A_0)$ for $\forall j \in \mathcal{I}$. Under these conditions, our theorem justifies that column-sparse relaxed R-PCA, equation 2.10, with $\lambda = 1/\sqrt{\log n}$ exactly recognizes the indices of noises. Thus our approach succeeds.

5.1.3 Dense Gaussian Noises. Assume that the data A_0 lie in an r-dimension subspace where r is relatively small. For dense gaussian noises, we consider the use of squared Frobenius norm, leading to the following relaxed R-LRR problem:

$$\min_{A,Z,E} ||Z||_* + \lambda||E||_F^2, \quad \text{s.t. } A = AZ, \quad X = A + E. \tag{5.6}$$

We quote the following result from Favaro et al. (2011), which gave the closed-form solution to problem 5.6. Based on our results in section 3, we give a new proof:

Corollary 2 *(Favaro et al., 2011). Let $X = U\Sigma V^T$ be the SVD of the data matrix X. Then the optimal solution to equation 5.6 is given by $A^* = U_1\Sigma_1 V_1^T$, $Z^* = V_1 V_1^T$, and $E^* = X - U_1\Sigma_1 V_1^T$, where Σ_1, U_1, and V_1 correspond to the top $r = \arg\min_k (k + \lambda \sum_{i>k} \sigma_i^2)$ singular values and singular vectors of X, respectively.*

Proof. The optimal solution to problem

$$\min_A \operatorname{rank}(A) + \lambda||X - A||_F^2 \tag{5.7}$$

is $A^* = U_1\Sigma_1 V_1^T$, where Σ_1, U_1, and V_1 correspond to the top $r = \arg\min_k (k + \lambda \sum_{i>k} \sigma_i^2)$ singular values and singular vectors of X, respectively. This can be easily seen by probing the different rank k of A and observing that $\min_{\operatorname{rank}(A)\le k} ||X - A||_F^2 = \sum_{i>k} \sigma_i^2$.

Next, according to theorem 2, where f is chosen as the squared Frobenius norm, the optimal solution to problem 5.6 is given by $A^* = U_1\Sigma_1 V_1^T$, $Z^* = (A^*)^\dagger A^* = V_1 V_1^T$, and $E^* = X - A^*$ as claimed.

Corollary 2 offers insight into the relaxed R-LRR, equation 5.6. We can first solve the classical PCA problem with parameter $r = \arg\min_k k + \lambda \sum_{i>k} \sigma_i^2$ and then adopt the shape interaction matrix of the denoised data matrix as the affinity matrix for subspace clustering. This is consistent with the well-known fact that, empirically and theoretically, PCA is capable of dealing effectively with small, dense gaussian noise. Note that one needs to tune the parameter λ in problem 5.6 in order to obtain a suitable parameter r for the PCA problem.

5.1.4 Other Cases. Although our approach works well under rather broad conditions, it might fail in some cases (e.g., the noiseless data matrix is not low rank). However, for certain data structures, the following numerical experiment shows that reducing to R-PCA correctly identifies the indices of noise even though the ground truth data matrix is of full rank. The synthetic data are generated as follows. In the linear space $\mathbb{R}^{5D}$, we construct five independent D-dimensional subspaces $\{S_i\}_{i=1}^5$, whose bases $\{U_i\}_{i=1}^5$ are

Table 2: Exact Identification of Indices of Noise on the Matrix $M \in \mathbb{R}^{5D\times 100D}$.

D	$\text{dist}(\mathcal{I}^*, \mathcal{I}_0)$	D	$\text{dist}(\mathcal{I}^*, \mathcal{I}_0)$	D	$\text{dist}(\mathcal{I}^*, \mathcal{I}_0)$	D	$\text{dist}(\mathcal{I}^*, \mathcal{I}_0)$
5	0	10	0	50	0	100	0

Note: $\text{Rank}(A_0) = 5D$, $||E_0||_{2,0} = 15D$, and $\lambda = 1/\sqrt{\log(100D)}$. $\mathcal{I}^*$ refers to the indices obtained by solving model 2.10, and $\mathcal{I}_0$ refers to the ground truth indices of noise.

randomly generated column orthonormal matrices. Then $20D$ points are sampled from each subspace by multiplying its basis matrix with a $D \times 20D$ gaussian distribution matrix, whose entries are independent and identically distributed (i.i.d.) $\mathcal{N}(0, 1)$. Thus, we obtain a $5D \times 100D$ structured sample matrix without noise, and the noiseless data matrix is of rank $5D$. We then add 15% column-wise gaussian noises whose entries are i.i.d. $\mathcal{N}(0, 1)$ on the noiseless matrix and solve model 2.10 with $\lambda = 1/\sqrt{\log(100D)}$. Table 2 reports the Hamming distance between the ground truth indices and the identified indices by model 2.10, under different input sizes. It shows that reducing to R-PCA succeeds for structured data distributions even when the dimension of the sum of the subspaces is equal to that of the ambient space. In contrast, the algorithm fails for unstructured data distributions; for example, the noiseless data are $5D \times 100D$ gaussian matrix whose element is totally random, obeying i.i.d. $\mathcal{N}(0, 1)$. Since the main focus of this letter is the relations among several low-rank models and the success conditions are within the research of R-PCA, the theoretical analysis on how data distribution influences the success of R-PCA will be our future work.

5.2 Fast Algorithms for Subspace Recovery. Representative low-rank subspace recovery models like LRR and LatLRR are solved by ADMM (Lin et al., 2011) and the overall complexity[3] is $O(mn^2)$ (Liu et al., 2010; Liu & Yan, 2011; Liu et al., 2013). For LRR, by employing linearized ADMM (LADMM) and some advanced tricks for computing partial SVD, the resulting algorithm is of $O(rn^2)$ overall complexity, where r is the rank of optimal Z. We show that our REDU-EXPR approach can be much faster.

We take a real experiment for an example. We test face image clustering on the extended YaleB database, which consists of 38 persons with 64 different illuminations for each person. All the faces are frontal, and thus images of each person lie in a low-dimensional subspace (Belhumeur et al., 1997). We generate the input data as follows. We reshape each image into a 32,256-dimensional column vector. Then the data matrix X is $32,256 \times 2,432$. We record the running times and the clustering accuracies[4] of relaxed LRR (Liu

[3] All complexity appearing in our letter refers to the overall complexity, that is, taking the iteration complexity into account.

[4] Liu et al. (2010) reported an accuracy of 62.53% by LRR, but there were only 10 classes in their data set. In contrast, there are 38 classes in our data set.

Table 3: Unsupervised Face Image Clustering Results on the Extended YaleB Database.

Model	Method	Accuracy	CPU Time (h)
LRR	ADMM	-	>10
R-LRR	ADMM	-	Did not converge
R-LRR	Partial ADMM	-	>10
R-LRR	REDU-EXPR	61.6365%	0.4603

et al., 2010, 2013) and relaxed R-LRR (Favaro et al., 2011; Wei & Lin, 2010). LRR is solved by ADMM. For R-LRR, we test three algorithms. The first one is traditional ADMM: updating A, E, and Z alternately by minimizing the augmented Lagrangian function of relaxed R-LRR:

$$L(A,E,Z) = ||Z||_* + \lambda f(E) + \langle X - E - A, Y_1\rangle + \langle A - AZ, Y_2\rangle + \frac{\mu}{2}(||X - E - A||_F^2 + ||A - AZ||_F^2). \tag{5.8}$$

The second algorithm is partial ADMM, which updates A, E, and Z by minimizing the partial augmented Lagrangian function:

$$L(A,E,Z) = ||Z||_* + \lambda f(E) + \langle X - E - A, Y\rangle + \frac{\mu}{2}||X - E - A||_F^2, \tag{5.9}$$

subject to $A = AZ$. This method is adopted by Favaro et al. (2011). A key difference between partial ADMM and traditional ADMM is that the former updates A and Z simultaneously by using corollary 2. (For more details, refer to Favaro et al., 2011.) The third method is REDU-EXPR, adopted by Wei and Lin (2010). Except the ADMM method for solving R-LRR, we run the codes provided by their respective authors.

One can see from Table 3 that REDU-EXPR is significantly faster than the ADMM-based method. Actually, solving R-LRR by ADMM did not converge. We want to point out that the partial ADMM method used the closed-form solution shown in corollary 2. However, its speed is still much inferior to that of REDU-EXPR.

For large-scale data, neither $O(mn^2)$ nor $O(rn^2)$ is fast enough. Fortunately, for R-PCA, it is relatively easy to design low-complexity randomized algorithms to further reduce its computational load. Liu, Lin, Su, and Gao (2014) has reported an efficient randomized algorithm, ℓ_1 filtering, to solve R-PCA when $f(E) = \|E\|_{\ell_1}$. The ℓ_1 filtering is completely parallel, and its complexity is only $O(r^2(m+n))$—linear to the matrix size. In the following, we sketch the ℓ_1 filtering algorithm (Liu et al., 2014), and in the same spirit

propose a novel $\ell_{2,1}$ filtering algorithm for solving column-sparse R-PCA, equation 2.11, that is, R-PCA with $f(E) = \|E\|_{\ell_{2,1}}$.

5.2.1 Outline of ℓ_1 Filtering Algorithm (Liu et al., 2014). The ℓ_1 filtering algorithm aims at solving the R-PCA problem, equation 3.1, with $f(E) = \|E\|_{\ell_1}$. There are two main steps. The first step is to recover a seed matrix. The second is to process the rest of the data matrix by ℓ_1-norm-based linear regression.

Step 1: Recovery of a Seed Matrix. Assume that the target rank r of the low-rank component A is very small compared with the size of the data matrix: $r \ll \min\{m, n\}$. By randomly sampling an $s_r r \times s_c r$ submatrix X^s from X, where $s_r, s_c > 1$ are oversampling rates, we partition the data matrix X, together with the underlying matrix A and the noise E, into four parts (for simplicity, we assume that X^s is at the top left corner of X):

$$X = \begin{bmatrix} X^s & X^c \\ X^r & \widetilde{X}^s \end{bmatrix}, \quad A = \begin{bmatrix} A^s & A^c \\ A^r & \widetilde{A}^s \end{bmatrix}, \quad E = \begin{bmatrix} E^s & E^c \\ E^r & \widetilde{E}^s \end{bmatrix}. \tag{5.10}$$

We first recover the seed matrix A^s of the underlying matrix A from X^s by solving a small-scale relaxed R-PCA problem,

$$\min_{A^s, E^s} ||A^s||_* + \lambda^s ||E^s||_{\ell_1}, \ \text{ s.t. } \ X^s = A^s + E^s, \tag{5.11}$$

where $\lambda^s = 1/\sqrt{\max\{s_r r, s_c r\}}$, suggested in (Candès et al., 2011), for exact recovery of the underlying A^s. This problem can be efficiently solved by ADMM (Lin et al., 2011).

Step 2: ℓ_1 Filtering. Since $\text{rank}(A) = r$ and A^s is a randomly sampled $s_r r \times s_c r$ submatrix of A, with an overwhelming probability $\text{rank}(A^s) = r$, so A^c and A^r must be represented as linear combinations of the columns or rows in A^s. Thus, we obtain the following ℓ_1-norm-based linear regression problems:

$$\min_{Q, E^c} ||E^c||_{\ell_1}, \ \text{ s.t. } \ X^c = A^s Q + E^c, \tag{5.12}$$

$$\min_{P, E^r} ||E^r||_{\ell_1}, \ \text{ s.t. } \ X^r = P^T A^s + E^r. \tag{5.13}$$

As soon as $A^c = A^s Q$ and $A^r = P^T A^s$ are computed, the generalized Nyström method (Wang, Dong, Tong, Lin, & Guo, 2009) gives

$$\widetilde{A}^s = P^T A^s Q. \tag{5.14}$$

Thus we recover all the submatrices in A. As shown in Liu et al. (2014), the complexity of this algorithm is only $O(r^2(m+n))$ without considering the reading and writing time.

5.2.2 $\ell_{2,1}$ Filtering Algorithm. ℓ_1 filtering is for entry-sparse R-PCA. For R-LRR, we need to solve column sparse R-PCA. Unlike the ℓ_1 case, which breaks the full matrix into four blocks, the $\ell_{2,1}$ norm requires viewing each column in a holistic way, so we can only partition the whole matrix into two blocks. We inherit the idea of ℓ_1 filtering to propose a randomized algorithm, called $\ell_{2,1}$ filtering, to solve column-sparse R-PCA. It also consists of two steps. We first recover a seed matrix and then process the remaining columns via ℓ_2 norm-based linear regression, which turns out to be a least square problem.

Recovery of a Seed Matrix. The step of recovering a seed matrix is nearly the same as that of the ℓ_1 filtering method, except that we partition the whole matrix into only two blocks. Suppose the rank of A is $r \ll \min\{m, n\}$. We randomly sample sr columns of X, where $s > 1$ is an oversampling rate. These sr columns form a submatrix X_l. For brevity, we assume that X_l is the leftmost submatrix of X. Then we may partition X, A, and E as follows:

$$X = [X_l, X_r], \quad E = [E_l, E_r], \quad A = [A_l, A_r],$$

respectively. We could first recover A_l from X_l by a small-scale relaxed column-sparse R-PCA problem,

$$\min_{A_l, E_l} ||A_l||_* + \lambda_l ||E_l||_{\ell_{2,1}}, \quad \text{s.t. } X_l = A_l + E_l, \tag{5.15}$$

where $\lambda_l = 1/\sqrt{\log(sr)}$ (Zhang et al., 2015).

$\ell_{2,1}$ Filtering. After the seed matrix A_l is obtained, since $\text{rank}(A) = r$ and with an overwhelming probability $\text{rank}(A_l) = r$, the columns of A_r must be linear combinations of A_l. So there exists a representation matrix $Q \in \mathbb{R}^{sr \times (n-sr)}$ such that

$$A_r = A_l Q. \tag{5.16}$$

The part E_r of noise should still be column sparse, however, so we have the following $\ell_{2,1}$ norm-based linear regression problem:

$$\min_{Q, E_r} ||E_r||_{\ell_{2,1}}, \quad \text{s.t. } X_r = A_l Q + E_r. \tag{5.17}$$

If equation 5.17 is solved directly by using ADMM (Liu et al., 2012), the complexity of our algorithm will be nearly the same as that of solving the

Algorithm 1: $\ell_{2,1}$ Filtering Algorithm for Column-Sparse R-PCA.

Input: Observed data matrix X and estimated rank r.

1. Randomly sample sr columns from X to form X_l.

2. Solve small-scale relaxed R-PCA, equation 5.15, by ADMM and obtain SVD of A_l: $U_{A_l}\Sigma_{A_l}V_{A_l}^T$.

3. Recover $A_r = A_l Q$ by solving, equation 5.17, whose solution is $A_r = U_{A_l}(U_{A_l}^T X_r)$.

Output: Low-rank component $A = [A_l, A_r]$ and column sparse matrix $E = X - A$.

whole original problem. Fortunately, we can solve equation 5.17 column-wise independently due to the separability of $\ell_{2,1}$ norms.

Let $x_r^{(i)}$, $q^{(i)}$, and $e_r^{(i)}$ represent the ith column of X_r, Q, and E_r, respectively ($i = 1, 2, \cdots, n - sr$). Then problem 5.17 could be decomposed into $n - sr$ subproblems:

$$\min_{q^{(i)}, e_r^{(i)}} ||e_r^{(i)}||_2, \quad \text{s.t. } x_r^{(i)} = A_l q^{(i)} + e_r^{(i)}, \ i = 1, \cdots, n - sr. \tag{5.18}$$

As least square problems, equation 5.18 has closed-form solutions $q^{(i)} = A_l^\dagger x_r^{(i)}$, $i = 1, \cdots, n - sr$. Then $Q^* = A_l^\dagger X_l$ and the solution to the original problem, equation 5.17, is $(A_l^\dagger X_r, X_r - A_l A_l^\dagger X_r)$. Interestingly, it is the same solution if replacing the $\ell_{2,1}$ norm in equation 5.17 with the Frobenius norm.

Note that our target is to recover the right patch $A_r = A_l Q^*$. Let $U_{A_l}\Sigma_{A_l}V_{A_l}^T$ be the skinny SVD of A_l, which is available when solving equation 5.15. Then A_r could be written as

$$A_r = A_l Q^* = A_l A_l^\dagger X_r = U_{A_l} U_{A_l}^T X_r. \tag{5.19}$$

We may first compute $U_{A_l}^T X_r$ and then $U_{A_l}(U_{A_l}^T X_r)$. This little trick reduces the complexity of computing A_r.

The Complete Algorithm. Algorithm 1 summarizes our $\ell_{2,1}$ filtering algorithm for solving column-sparse R-PCA.

As soon as the solution (A, E) to column-sparse R-PCA is solved, we can obtain the representation matrix of R-LRR Z by $Z = A^\dagger A$. Note that we should not compute Z naively as it is written, whose complexity will be more than $O(mn^2)$. A more clever way is as follows. Suppose $U_A\Sigma_A V_A^T$ is the skinny SVD of A; then $Z = A^\dagger A = V_A V_A^T$. On the other hand, $A = U_{A_l}[\Sigma_{A_l}V_{A_l}^T, U_{A_l}^T X_r]$. So we have only to compute the row space

Algorithm 2: Subspace Clustering Based on the Relaxed R-LRR Model, Equation 2.7.

Input: Observed data matrix X, estimated rank r.

1. Solve relaxed column sparse R-PCA, equation 3.2, with $f(E) = \|E\|_{\ell_{2,1}}$ by algorithm 1.

2. Conduct LQ decomposition on the matrix $\hat{A} = [\Sigma_{A_l} V_{A_l}^T, U_{A_l}^T X_r]$ as $\hat{A} = LV^T$.

3. Obtain the affinity matrix by $|Z| = |VV^T|$ and conduct spectral clustering.

Output: Label for each data point.

of $\hat{A} = [\Sigma_{A_l} V_{A_l}^T, U_{A_l}^T X_r]$, where $U_{A_l}^T X_r$ has been saved in step 3 of algorithm 1. This can be easily done by doing LQ decomposition (Golub & Van Loan, 2012) of $\hat{A}$: $\hat{A} = LV^T$, where L is lower triangular and $V^T V = I$. Then $Z = VV^T$. Since LQ decomposition is much cheaper than SVD, the above trick is very efficient and all the matrix-matrix multiplications are $O(r^2 n)$. The complete procedure for solving the R-LRR problem, equation 2.7 is described in algorithm 2.

Unlike LRR, the optimal solution to R-LRR problem, 2.7 is symmetric, and thus we could directly use $|Z|$ as the affinity matrix instead of $|Z| + |Z^T|$. After that, we can apply spectral clustering algorithms, such as normalized cut, to cluster each data point into its corresponding subspace.

Although for the moment the formal proof of our $\ell_{2,1}$ filtering algorithm is not yet available, our algorithm intuitively works well. To this end, we assume two conditions to guarantee the exact recoverability of our algorithm. First, to guarantee the exact recovery of the ground-truth subspace from the whole matrix X, we need to ensure that the same subspace can be fully recovered from the seed matrix X_l. So applying the result of Zhang et al. (2015) to the seed matrix, we assume that $r \leq O(sr \log sr)$ and E_0 has $O(n)$ nonzero columns, which guarantees that E_l has $O(n)$ column support at an overwhelming probability. Also, the incoherence conditions and the regularization parameter $\lambda = 1/\sqrt{\log sr}$ are required. Second, as for the $\ell_{2,1}$ filtering step, to represent the rest of the whole matrix by A_l, it seems that A should be low rank: Range(A_l) = Range(A). Under these conditions, our $\ell_{2,1}$ filtering algorithm intuitively succeeds with overwhelming probabilities. We leave the rigorous analysis as our future work.

Complexity Analysis. In algorithm 1, step 2 requires $O(r^2 m)$ time, and step 3 requires $2rmn$ time. Thus the whole complexity of the $\ell_{2,1}$ filtering algorithm for solving column-sparse R-PCA is $O(r^2 m) + 2rmn$. In algorithm 2, for solving the relaxed R-LRR problem, equation 2.7, as just analyzed, step 1 requires $O(r^2 m) + 2rmn$ time. The LQ decomposition in step 2 requires $6r^2 n$ time at most (Golub & Van Loan, 2012). Computing VV^T in

step 3 requires rn^2 time. Thus, the whole complexity for solving equation 2.7 is $O(r^2m) + 6r^2n + 2rmn + rn^2$.[5] As most of the low-rank subspace clustering models require $O(mn^2)$ time to solve, due to SVD or matrix-matrix multiplication in every iteration, our algorithm is significantly faster than state-of-the-art methods.

6 Experiments

In this section, we use experiments to illustrate the applications of our theoretical analysis.

6.1 Comparison of Optimality on Synthetic Data. We compare the two algorithms, partial ADMM[6] (Favaro et al., 2011) and REDU-EXPR (Wei & Lin, 2010), which we mentioned in section 5.2, for solving the nonconvex-relaxed R-LRR problem, equation 2.7. Since the traditional ADMM is not-convergent, we do not compare with it. Because we want to compare only the quality of solutions produced by the two methods, for REDU-EXPR we temporarily do not use the $\ell_{2,1}$ filtering algorithm introduced in section 5 to solve column-sparse R-PCA.

The synthetic data are generated as follows. In the linear space $\mathbb{R}^{1000}$, we construct five independent four-dimensional subspaces $\{S_i\}_{i=1}^5$, whose bases $\{U_i\}_{i=1}^5$ are randomly generated column orthonormal matrices. Then 200 points are uniformly sampled from each subspace by multiplying its basis matrix with a 4×200 gaussian distribution matrix, whose entries are independent and identically distributed (i.i.d.). $\mathcal{N}(0, 1)$. Thus, we obtain a $1{,}000 \times 1{,}000$ sample matrix without noise.

We compare the clustering accuracies[7] as the percentage of corruption increases, where noise uniformly distributed on $(-0.6, 0.6)$ is added at uniformly distributed positions. We run the test 10 times and compute the mean clustering accuracy. Figure 2 presents the comparison on the accuracy, where all the parameters are tuned to be the same: $\lambda = 1/\sqrt{\log 1000}$. One can see that R-LRR solved by REDU-EXPR is much more robust to column-sparse corruptions than by partial ADMM.

To further compare the optimality, we also record the objective function values computed by the two algorithms. Since both algorithms aim to

[5] Here we highlight the difference between $2rmn + rn^2$ and $O(rmn + rn^2)$. The former is independent of numerical precision. It is due to the three matrix-matrix multiplications to form $\hat{A}$ and Z, respectively. In contrast, $O(rmn + rn^2)$ usually grows with numerical precision. The more iterations there are, the larger the constant in the big O is.

[6] The partial ADMM method of Favaro et al. (2011) was designed for the ℓ_1 norm on the noise matrix E, while here we have adapted it for the $\ell_{2,1}$ norm.

[7] Just as Liu et al. (2010) did, given the ground truth labeling, we set the label of a cluster to be the index of the ground truth that contributes the maximum number of samples to the cluster. Then all these labels are used to compute the clustering accuracy after comparing with the ground truth.

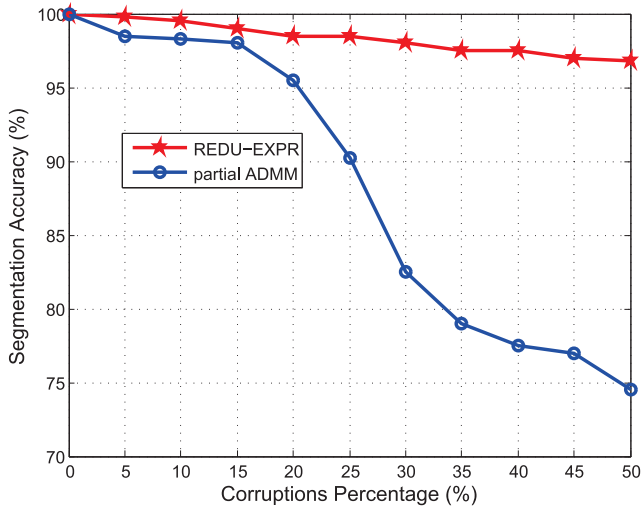

Figure 2: Comparison of accuracies of solutions to relaxed R-LRR, equation 2.7, computed by REDU-EXPR (Wei & Lin, 2010) and partial ADMM (Favaro et al., 2011), where the parameter λ is adopted as $1/\sqrt{\log n}$ and n is the input size. The program is run 10 times, and the average accuracies are reported.

achieve the low rankness of the affinity matrix and the column sparsity of the noise matrix, we compare the objective function of the original R-LRR, equation 2.6,

$$\mathcal{F}(Z, E) = \text{rank}(Z) + \lambda ||E||_{\ell_{2,0}}. \tag{6.1}$$

As shown in Table 4, R-LRR by REDU-EXPR could obtain smaller rank(Z) and objective function than those of partial ADMM. Table 4 also shows the CPU times (in seconds). One can see that REDU-EXPR is significantly faster than partial ADMM when solving the same model.

6.2 Comparison of Speed on Synthetic Data. In this section, we show the great speed advantage of our REDU-EXPR algorithm in solving low-rank recovery models. We compare the algorithms to solve relaxed R-LRR, equation 2.7. We also present the results of solving LRR by ADMM for reference, although it is a slightly different model. Except our $\ell_{2,1}$ filtering algorithm, all the codes run in this test are offered by Liu et al. (2013), Liu and Yan (2011), and Favaro et al. (2011).

The parameter λ is set for each method so that the highest accuracy is obtained. We generate clean data as we did in section 6.1. The only differences are the choice of the dimension of the ambient space and the number

Table 4: Comparison of Robustness and Speed Between Partial ADMM (LRSC) (Favaro et al., 2011) and REDU-EXPR (RSI) (Wei & Lin, 2010) Methods for Solving R-LRR When the Percentage of Corruptions Increases.

Noise Percentage (%)	0	10	20	30	40	50
Rank(Z) (partial ADMM)	20	30	30	30	30	30
Rank(Z) (REDU-EXPR)	20	20	20	20	20	20
$\|\|E\|\|_{\ell_{2,0}}$ (partial ADMM)	0	99	200	300	400	500
$\|\|E\|\|_{\ell_{2,0}}$ (REDU-EXPR)	0	100	200	300	400	500
Objective (partial ADMM)	**20.00**	67.67	106.10	144.14	182.19	220.24
Objective (REDU-EXPR)	**20.00**	**58.05**	**96.10**	**134.14**	**172.19**	**210.24**
Time (s, partial ADMM)	**4.89**	124.33	126.34	119.12	115.20	113.94
Time (s, REDU-EXPR)	10.67	**9.60**	**8.34**	**8.60**	**9.00**	**12.86**

Notes: All the experiments are run 10 times, and the λ is set to be the same: $\lambda = 1/\sqrt{\log n}$, where n is the data size. The numbers in bold refer to the better results between the two methods: partial ADMM and REDU-EXPR.

of points sampled from subspaces. We compare the speed of different algorithms on corrupted data, where the noises are added in the same way as in Liu et al. (2010) and Liu et al. (2013). Namely, the noises are added by submitting to 5% column-wise gaussian noises with zero means and $0.1||x||_2$ standard deviation, where x indicates corresponding vector in the subspace. For REDU-EXPR, with or without using $\ell_{2,1}$ filtering, the rank is estimated at its exact value, 20, and the oversampling parameter s_c is set to be 10. As the data size goes up, the CPU times are shown in Table 5. When the corruptions are not heavy, all the methods in this test achieve 100% accuracy. We can see that REDU-EXPR consistently outperforms ADMM-based methods. By $\ell_{2,1}$ filtering, the computation time is further reduced. The advantage of $\ell_{2,1}$ filtering is more salient when the data size is larger.

6.3 Test on Real Data: AR Face Database. Now we test different algorithms on real data, the AR face database, to classify face images. The AR face database contains 2574 color images of 99 frontal faces. All the faces have different facial expressions, illumination conditions, and occlusions (e.g., sunglasses or scarf; see Figure 3). Thus the AR database is much harder than the YaleB database for face clustering. We replace the spectral clustering (step 3 in algorithm 2) with a linear classifier. The classification is as follows,

$$\min_W ||H - WF||_F^2 + \gamma ||W||_F^2, \tag{6.2}$$

which is simply a ridge regression, and the regularization parameter γ is fixed at 0.8, where F is the feature data and H is the label matrix. The

Table 5: Comparison of CPU Time (Seconds) Between LRR (Liu et al., 2010, 2013) Solved by ADMM, R-LRR Solved by Partial ADMM (LRSC) (Favaro et al., 2011), R-LRR Solved by REDU-EXPR without Using $\ell_{2,1}$ Filtering (RSI) (Wei & Lin, 2010), and R-LRR Solved by REDU-EXPR Using $\ell_{2,1}$ Filtering as Data Size Increases.

Data Size	LRR (ADMM)	R-LRR (partial ADMM)	R-LRR (REDU-EXPR)	R-LRR (filtering REDU-EXPR)
250×250	33.0879	4.9581	1.4315	**0.6843**
500×500	58.9177	7.2029	1.8383	**1.0917**
1000×1000	370.1058	24.5236	6.1054	**1.5429**
2000×2000	>3600	124.3417	28.3048	**2.4426**
4000×4000	>3600	411.8664	115.7095	**3.4253**

Note: In this test, REDU-EXPR with $\ell_{2,1}$ filtering is significantly faster than other methods, and its computation time grows at most linearly with the data size.

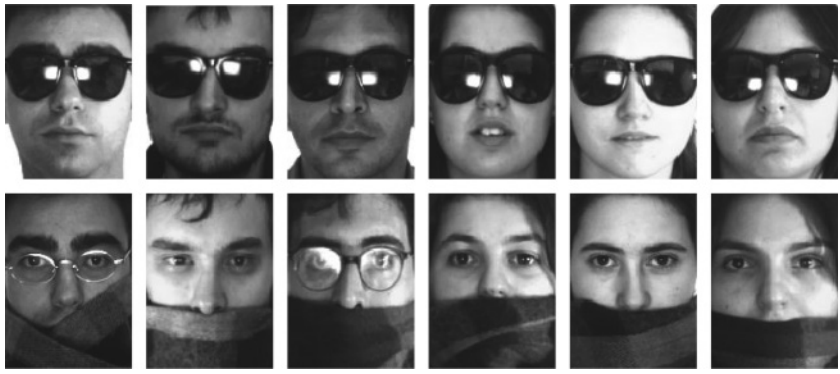

Figure 3: Examples of images with severe occlusions in the AR database. The images in the same column belong to the same person.

classifier is trained as follows. We first run LRR or R-LRR on the original input data $X \in \mathbb{R}^{m \times n}$ and obtain an approximately block diagonal matrix $Z \in \mathbb{R}^{n \times n}$. View each column of Z as a new observation,[8] and separate the columns of Z into two parts, where one part corresponds to the training data and the other to the test data. We train the ridge regression model by the training samples and use the obtained W to classify the test samples.

Unlike the existing literature, Liu et al. (2010, 2013), which manually removed severely corrupted images and shrank the input images to small-sized ones in order to reduce the computation load, our experiment uses all

[8]Since Z is approximately block diagonal, each column of Z has few nonzero coefficients, and thus the new observations are suitable for classification.

Table 6: Comparison of Classification Accuracy and Speed on the AR Database with the Task of Face Image Classification.

Model	Method	Accuracy	CPU Time (h)
LRR	ADMM	-	>10
R-LRR	partial ADMM	86.3371%	53.5165
R-LRR	REDU-EXPR	90.1648%	0.5639
R-LRR	REDU-EXPR with $\ell_{2,1}$ filtering	**90.5901%**	**0.1542**

Notes: For fair comparison of both the accuracy and the speed for different algorithms, the parameters are tuned to be the best according to the classification accuracy, and we observe the CPU time. The figures in bold refer to the best results.

the full-sized face images. So the size of our data matrix is $19{,}800 \times 2574$, where each image is reshaped as a column of the matrix, 19,800 is the number of pixels in each image, and 2574 is the total number of face images. We test LRR (Liu et al., 2010, 2013), solved by ADMM, and relaxed R-LRR, solved by partial ADMM (Favaro et al., 2011), REDU-EXPR (Wei & Lin, 2010), and REDU-EXPR with $\ell_{2,1}$ filtering) for both classification accuracy and speed. Table 6 shows the results, where the parameters have been tuned to be the best. Since the ADMM-based method requires too much time to converge, we terminate it after 60 hours. This experiment testifies to the great speed advantage of REDU-EXPR and $\ell_{2,1}$ filtering. Note that with $\ell_{2,1}$ filtering, the speed of REDU-EXPR is three times faster than that without $\ell_{2,1}$ filtering, and the accuracy is not compromised.

7 Conclusion and Future Work

In this letter, we investigate the connections among solutions of some representative low-rank subspace recovery models: R-PCA, R-LRR, R-LatLRR, and their convex relaxations. We show that their solutions can be mutually expressed in closed forms. Since R-PCA is the simplest model, it naturally becomes a hinge for all low-rank subspace recovery models. Based on our theoretical findings, under certain conditions we are able to find better solutions to low-rank subspace recovery models and also significantly speed up finding their solutions numerically by solving R-PCA first, and then express their solutions by that of R-PCA in closed forms. Since there are randomized algorithms for R-PCA, for example, we propose the $\ell_{2,1}$ filtering algorithm for column-sparse R-PCA, the computation complexities for solving existing low-rank subspace recovery models can be much lower than the existing algorithms. Extensive experiments on both synthetic and real-world data testify to the utility of our theories.

As shown in section 5.1.4, our approach may succeed even when the conditions of sections 5.1.1, 5.1.2, and 5.1.3 do not hold. The theoretical analysis on how data distribution influences the success of our approach,

together with the theoretical guarantee of our $\ell_{2,1}$ filtering algorithm, will be our future work.

Acknowledgments

We thank Rene Vidal for valuable discussions. H. Zhang and C. Zhang are supported by National Key Basic Research Project of China (973 Program) (nos. 2015CB352303 and 2011CB302400) and National Nature Science Foundation (NSF) of China (nos. 61071156 and 61131003). Z. Lin is supported by NSF China (nos. 61231002 and 61272341), 973 Program of China (no. 2015CB352502), and Microsoft Research Asia Collaborative Research Program. J. Gao is partially supported under Australian Research Council's Discovery Projects funding scheme (project DP130100364).

References

Avron, H., Maymounkov, P., & Toledo, S. (2010). Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing*, *32*(3), 1217–1236.

Basri, R., & Jacobs, D. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(2), 218–233.

Belhumeur, P., Hespanha, J., & Kriegman, D. (1997). Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Belhumeur, P., & Kriegman, D. (1998). What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, *28*(3), 245–260.

Bull, G., & Gao, J. (2012). Transposed low rank representation for image classification. In *IEEE International Conference on Digital Image Computing Techniques and Application* (pp. 1–7). Piscataway, NJ: IEEE.

Candès, E., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(3), 11.

Chandrasekaran, V., Sanghavi, S., Parrilo, P., & Willsky, A. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, *21*(2), 572–596.

Chen, Y., Xu, H., Caramanis, C., & Sanghavi, S. (2011). Robust matrix completion and corrupted columns. In *Proceedings of the International Conference on Machine Learning* (pp. 873–880). New York: ACM.

Cheng, B., Liu, G., Wang, J., Li, H., & Yan, S. (2011). Multi-task low-rank affinity pursuit for image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2439–2446). Piscataway, NJ: IEEE.

Costeira, J., & Kanade, T. (1998). A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, *29*(3), 159–179.

De La Torre, F., & Black, M. (2003). A framework for robust subspace learning. *International Journal of Computer Vision*, *54*(1), 117–142.

Elhamifar, E., & Vidal, R. (2009). Sparse subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2790–2797). Piscataway, NJ: IEEE.

Favaro, P., Vidal, R., & Ravichandran, A. (2011). A closed form solution to robust subspace estimation and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1801–1807). Piscataway, NJ: IEEE.

Fischler, M., & Bolles, R. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Gear, W. (1998). Multibody grouping from motion images. *International Journal of Computer Vision*, *29*(2), 133–150.

Gnanadesikan, R., & Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*(1), 81–124.

Golub, G., & Van Loan, C. (2012). *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

Hardt, M., & Moitra, A. (2012). *Algorithms and hardness for robust subspace recovery*. arXiv preprint: 1211.1041.

Ho, J., Yang, M., Lim, J., Lee, K., & Kriegman, D. (2003). Clustering appearances of objects under varying illumination conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 313–320). Piscataway, NJ: IEEE.

Hsu, D., Kakade, S. M., & Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, *57*(11), 7221–7234.

Huber, P. (2011*). Robust statistics*. New York: Springer.

Ji, H., Liu, C., Shen, Z., & Xu, Y. (2010). Robust video denoising using low-rank matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1791–1798). Piscataway, NJ: IEEE.

Ke, Q., & Kanade, T. (2005). Robust ℓ_1-norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 739–746). Piscataway, NJ: IEEE.

Lerman, G., McCoy, M. B., Tropp, J. A., & Zhang, T. (2014). Robust computation of linear models by convex relaxation. *Foundations of Computational Mathematics 15*, 363–410.

Lin, Z., Liu, R., & Su, Z. (2011). Linearized alternating direction method with adaptive penalty for low-rank representation. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 24* (pp. 612–620). Red Hook, NY: Curran.

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

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the International Conference on Machine Learning* (pp. 663–670). Madison, WI: Omnipress.

Liu, G., & Yan, S. (2011). Latent low-rank representation for subspace segmentation and feature extraction. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1615–1622). Piscataway, NJ: IEEE.

Liu, R., Lin, Z., De la Torre, F., & Su, Z. (2012). Fixed-rank representation for unsupervised visual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 598–605). Piscataway, NJ: IEEE.

Liu, R., Lin, Z., Su, Z., & Gao, J. (2014). Linear time principal component pursuit and its extensions using ℓ_1 filtering. *Neurocomputing*, *142*, 529–541.

Ma, Y., Derksen, H., Hong, W., & Wright, J. (2007). Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(9), 1546–1562.

Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, *3*(2), 123–224.

McCoy, M., & Tropp, J. A. (2011). Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, *5*, 1123–1160.

Ni, Y., Sun, J., Yuan, X., Yan, S., & Cheong, L. (2010). Robust low-rank subspace segmentation with semidefinite guarantees. In *Proceedings of the IEEE International Conference on Data Mining Workshops*. Piscataway, NJ: IEEE.

Paoletti, S., Juloski, A., Ferrari-Trecate, G., & Vidal, R. (2007). Identification of hybrid systems—a tutorial. *European Journal of Control*, *13*(2–3), 242–260.

Peng, Y., Ganesh, A., Wright, J., Xu, W., & Ma, Y. (2010). RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 763–770). Piscataway, NJ: IEEE.

Rao, S., Tron, R., Vidal, R., & Ma, Y. (2010). Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(10), 1832–1845.

Soltanolkotabi, M., & Candès, E. (2012). A geometric analysis of subspace clustering with outliers. *Annals of Statistics*, *40*(4), 2195–2238.

Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography—a factorization method. *International Journal of Computer Vision*, *9*(2), 137–154.

Vidal, R. (2011). Subspace clustering. *IEEE Signal Processing Magazine*, *28*(2), 52–68.

Vidal, R., & Favaro, P. (2014). Low rank subspace clustering. *Pattern Recognition Letters*, *43*, 47–61.

Vidal, R., & Hartley, R. (2004). Motion segmentation with missing data using PowerFactorization and GPCA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 85–105). Piscataway, NJ: IEEE.

Vidal, R., Ma, Y., & Sastry, S. (2005). Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1945–1959.

Vidal, R., Soatto, S., Ma, Y., & Sastry, S. (2003). An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proceedings of the IEEE International Conference on Decision and Control* (pp. 167–172). Piscataway, NJ: IEEE.

Wang, J., Dong, Y., Tong, X., Lin, Z., & Guo, B. (2009). Kernel Nyström method for light transport. In *Proceedings of the ACM SIGGRAPH, 28* (pp. 1–10). New York: ACM.

Wang, J., Saligrama, V., & Castañón, D. (2011). Structural similarity and distance in learning. In *Proceedings of the Annual Allerton Conference on Communication, Control, and Computing* (pp. 744–751). Piscataway, NJ: IEEE.

Wei, S., & Lin, Z. (2010). *Analysis and improvement of low rank representation for subspace segmentation*. arXiv preprint: 1107.1561.

Wright, J., Ganesh, A., Rao, S., Peng, Y., & Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 22* (pp. 2080–2088). Red Hook, NY: Curran.

Wright, J., Ma, Y., Mairal, J., Sapiro, G., & Huang, T. S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, *98*(6), 1031–1044.

Xu, H., Caramanis, C., & Sanghavi, S. (2012). Robust PCA via outlier pursuit. *IEEE Transaction on Information Theory*, *58*(5), 3047–3064.

Yan, J., & Pollefeys, M. (2006). A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and nondegenerate. In *Proceedings of the European Conference on Computer Vision* (vol. 3954, pp. 94–106). New York: Springer-Verlag.

Yang, A., Wright, J., Ma, Y., & Sastry, S. (2008). Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, *110*(2), 212–225.

Zhang, C., & Bitmead, R. (2005). Subspace system identification for training-based MIMO channel estimation. *Automatica*, *41*(9), 1623–1632.

Zhang, H., Lin, Z., & Zhang, C. (2013). A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (vol. 8189, pp. 226–241). New York: IEEE.

Zhang, H., Lin, Z., Zhang, C., & Chang, E. (2015). Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 3143–3149). Cambridge, MA: AAAI Press.

Zhang, H., Lin, Z., Zhang, C., & Gao, J. (2014). Robust latent low rank representation for subspace clustering. *Neurocomputing*, *145*, 369–373.

Zhang, T., & Lerman, G. (2014). A novel *m*-estimator for robust PCA. *Journal of Machine Learning Research*, *15*(1), 749–808.

Zhang, Y., Jiang, Z., & Davis, L. (2013). Learning structured low-rank representations for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 676–683). Piscataway, NJ: IEEE.

Zhang, Z., Ganesh, A., Liang, X., & Ma, Y. (2012). TILT: Transform-invariant low-rank textures. *International Journal of Computer Vision*, *99*(1), 1–24.

Zhuang, L., Gao, H., Lin, Z., Ma, Y., Zhang, X., & Yu, N. (2012). Non-negative low rank and sparse graph for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2328–2335). Piscataway, NJ: IEEE.

Received November 11, 2014; accepted May 1, 2015.