ROUTE: Robust Outlier Estimation for Low Rank Matrix Recovery

Xiaojie Guo[†‡] **Zhouchen Lin**[§*]
[†]State Key Laboratory of Information Security, IIE, Chinese Academy of Sciences
[‡]University of Chinese Academy of Sciences
[§]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[*]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
xj.max.guo@gmail.com zlin@pku.edu.cn

Abstract

In practice, even very high-dimensional data are typically sampled from low-dimensional subspaces but with intrusion of outliers and/or noises. Recovering the underlying structure and the pollution from the observations is key to understanding and processing such data. Besides properly modeling the low-rank structure of subspace, how to handle the pollution is core regarding the performance of recovery. Often, the observed data is posed as a superimposition of the clean data and residual, while the residual can be roughly divided into two groups, including small dense noises and gross sparse outliers. Compared with small noises, outliers more likely ruin the recovery, as they can be arbitrarily large. By considering the above, this paper designs a method for recovering the low rank matrix with robust outlier estimation, termed as ROUTE, in a unified manner. Theoretical analysis on convergence and optimality, and experimental results on both synthetic and real data are provided to demonstrate the efficacy of our proposed method and show its superiority over other state-of-the-arts.

1 Introduction

Low rank matrix recovery (LRMR) is a process of discovering underlying structures from given measurements, the inspiration and motivation of which are both that, in real cases, even very high-dimensional observations should be from a low-dimensional subspace but unfortunately with intrusion of outliers and/or noises. As a theoretic foundation in computer vision, pattern recognition and machine learning, the effectiveness of LRMR has been confirmed by numerous fundamental tasks, such as principal component analysis [Pearson, 1901; Candès *et al.*, 2011], collaborative filtering [Zhang and Wang, 2016; Chen *et al.*, 2014] and subspace clustering [Nie and Huang, 2016; Liu *et al.*, 2013], as well as a wide spectrum of applications, like image denoising [Gu *et al.*, 2014], reflection separation [Guo *et al.*, 2014] and super-resolution [Jing *et al.*, 2015], to name just a few.

Formally, the LRMR problem can be directly or indirectly written in the following form:

$$\min_{\mathbf{L},\mathbf{E}} \ \mathrm{rank}(\mathbf{L}) + \alpha\Psi(\mathbf{E}) \ \mathrm{s.t.} \quad \mathcal{P}_\Omega(\mathbf{Y}) = \mathcal{P}_\Omega(\mathbf{L}+\mathbf{E}), \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times n}$, $\mathbf{L} \in \mathbb{R}^{m\times n}$ and $\mathbf{E} \in \mathbb{R}^{m\times n}$ designate the given data, the desired structure and the error residue, respectively. The function $\Psi(\cdot)$ is a penalty on the residual between the observed and recovered signals, $\mathrm{rank}(\cdot)$ stands for the low-rank constraint, and α is a non-negative parameter that provides a trade-off between the recovery fidelity and the low-rank promoting regularizer. Furthermore, $\mathcal{P}_\Omega(\cdot)$ is the orthogonal projection operator on the support $\Omega \in \{0,1\}^{m\times n}$. From Eq. (1), we can find that the quality of recovery depends on both the models of $\mathrm{rank}(\mathbf{L})$ and $\Psi(\mathbf{E})$.

As one of the two pivotal factors in LRMR, a proper low-rank promoting constraint on $\mathbf{L}$ is required to advocate the expected structure. It is computationally intractable (NP-hard) to directly minimize the rank function, say $\mathrm{rank}(\mathbf{L})$, due to its non-convexity and discontinuity. A widely used scheme is employing its tightest convex proxy, *i.e.* the nuclear norm $\|\mathbf{L}\|_*$ [Recht *et al.*, 2010; Candès *et al.*, 2011; Zhou *et al.*, 2013]. Nuclear norm minimization (NNM) based approaches can perform stably without knowing the target rank of recovery in advance. But, their applicability is often limited by the necessity of executing expensive singular value decomposition (SVD) for multiple times. At (much) less expense, bilinear factorization (BF) [Eriksson and van den Hengel, 2010; Srebro and T.Jaakkola, 2003; Meng and De la Torre, 2013; Salakhutdinov and Mnih, 2008; Lakshminarayanan *et al.*, 2011] is an alternative by replacing $\mathbf{L}$ with $\mathbf{UV}$, where the product of two factor matrices $\mathbf{U} \in \mathbb{R}^{m\times r}$ and $\mathbf{V} \in \mathbb{R}^{r\times n}$ implicitly guarantees that the rank of $\mathbf{UV}$ is never over r, typically $r \ll \min(m,n)$. This factorization strategy, through getting rid of SVDs, can greatly release the pressure of computation and provide accurate results when the target rank is given. Unfortunately, in some tasks, the target rank is unknown beforehand. In such a situation, the performance of BF would sharply degrade because of its sensitivity to the guess of target rank, especially when the data are severely contaminated. For bridging NNM and BF, and inheriting their respective merits, some bridges are recently built [Zheng *et al.*, 2012; Wang *et al.*, 2012]. One representative is adding $\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2$ into the objective of the BF [Cabral *et al.*, 2013]. Although the techniques above have made great progresses, the tolerance to dirty data is expected to be further improved.

In practical scenarios, acquiring perfect data is never the case. Furthermore, "*a little gall spoils a great deal of honey*" is quite a common issue. This is to say, without an effective strategy to reduce the negative effect from outliers and/or noises, the low rank matrix recovery is very likely prevented from reasonable solutions. Hence, besides properly modeling the low-rank structure, how to handle the pollution, especially gross outliers, is core to the performance of recovery. Arguably, the square loss (*a.k.a.* ℓ^2 loss) is the most commonly used penalty, which is optimal to Gaussian noises, like PCA [Pearson, 1901]. But, the square loss lacks robustness to outliers that are not unusual to find in real data. To be robust against gross corruptions, the ℓ^1 loss becomes popular, *e.g.* Robust PCA [Candès *et al.*, 2011]. For better fitting the residual, combining the ℓ^1 loss for the outlier component and the ℓ^2 for the small noise one is considered in [Zhou *et al.*, 2010]. Although the ℓ^1 loss can perform better than the ℓ^2 in dealing with the outlier, it still suffers from the scale issue. The ideal option to model outliers is the ℓ^0 loss, due to its scale invariance. The non-convexity and discreteness of the ℓ^0 penalty make it not so preferred by the community, although many works have proven that the ℓ^0 loss can improve the performance on different tasks, like [Pan *et al.*, 2017] for image deblurring, [Zhou *et al.*, 2013] for foreground detection and [Guo *et al.*, 2013] for video editing.

Back to the general formulation (1), if the support of both outliers and missing elements is given, the problem turns out to be a simpler version, *i.e.* the low rank matrix completion (LRMC). Compared with LRMC, the difficulty of LRMR, because of the unknown outlier support, significantly increases, which corroborates the intuition and theoretical fact that knowing the corruption location is beneficial. Therefore, it is natural to ask that if we can connect the LRMR to the LRMC via robustly estimating outliers, since by doing so the LRMR will be conquered more easily.

To answer the question, this paper proposes a Robust OUTlier Estimation method, called ROUTE, for recovering low rank matrices. More concretely, the contributions of this work are summarized as follows:

1. We design a method to jointly estimate outliers and recover the low rank matrix, namely ROUTE-LRMR, which unifies the LRMR and LRMC by treating both the missing and estimated outliers as weights;
2. Compared with the hard binary support, our weighting scheme assigns real-valued weights $[0, 1]$, which can be viewed as classification with confidence/probability;
3. Our design employs a maximum entropy regularization term to minimize the prediction bias, which behaves like a sigmoid function, arguably the most suitable classification function;
4. To seek the optimal solution for ROUTE-LRMR, we customize an Alternating Direction Minimization based algorithm. Theoretical analysis together with experimental results on both synthetic and real data are provided to show the efficacy of our ROUTE and reveal its superiority over other state-of-the-arts.

2 Methodology

2.1 Problem Formulation

In the simplest case, the support of observed elements is at hand, and the data are clean or just with slight noises. An option for recovering the low rank component (LRMC) is to optimize the following problem:

$$\min_{\mathbf{L}} \ \|\mathbf{L}\|_* + \frac{\alpha}{2}\|\Omega \circ (\mathbf{Y} - \mathbf{L})\|_F^2, \qquad (2)$$

where $\circ$ is the Hadamard product operator. As mentioned, the nuclear norm minimization requires to execute expensive SVDs on the whole data. To mitigate the computational pressure, Theorem 1 provides a bridge between NNM and BF models.

Theorem 1. *For any matrix* $\mathbf{L} \in \mathbb{R}^{m\times n}$*, the following relationship holds [Mazumder* et al.*, 2010]:*

$$\|\mathbf{L}\|_* = \min_{\mathbf{U},\mathbf{V}} \frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{1}{2}\|\mathbf{V}\|_F^2 \ \ \mathrm{s.t.}\ \ \mathbf{L} = \mathbf{UV}.$$

If $\mathrm{rank}(\mathbf{L}) = r \leq \min(m, n)$*, then the minimum solution above is attained at a factor decomposition* $\mathbf{L} = \mathbf{UV}$*, where* $\mathbf{U} \in \mathbb{R}^{m\times r}$ *and* $\mathbf{V} \in \mathbb{R}^{r\times n}$*.*

In the sequel, applying Theorem 1 on (2) reads:

$$\min_{\mathbf{U},\mathbf{V}} \ \frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{1}{2}\|\mathbf{V}\|_F^2 + \frac{\alpha}{2}\|\Omega \circ (\mathbf{Y} - \mathbf{UV})\|_F^2. \qquad (3)$$

Compared with directly minimizing $\|\Omega \circ (\mathbf{Y} - \mathbf{UV})\|_F^2$, the model (3) inherits the advantage of (2), which avoids overfitting when r is larger than the intrinsic rank.

In the real world, however, the data are frequently polluted by, besides small noises, gross corruptions, which may prevent the recovery from reasonable results. Hence, some steps should be taken for reducing the negative effect of such pollution. Recall that the ℓ^0 loss is ideal to host outliers, with slight modification, we have:

$$\min \frac{1}{2}\|\mathbf{U}\|_F^2 + \frac{1}{2}\|\mathbf{V}\|_F^2 + \frac{\alpha}{2}\|\mathbf{W} \circ (\mathbf{Y} - \mathbf{UV})\|_F^2 + \beta\|\overline{\mathbf{W}}\|_1$$
$$\mathrm{s.t.}\ \ \mathbf{W} + \overline{\mathbf{W}} = \mathbf{1};\ \ \mathbf{W} \text{ and } \overline{\mathbf{W}} \in \{0, 1\}^{m\times n}, \qquad (4)$$

where β is a weight to the corresponding term and $\mathbf{1}$ represents an all-one matrix with comparable size. We can see that from Eq. (4), the support Ω is replaced by a weight matrix $\mathbf{W}$ that can contain both the given support and the estimated outlier support. Please note that, under the binary weighting, $\|\mathbf{W} \circ (\mathbf{Y} - \mathbf{UV})\|_F^2 = \sum_{i,j} w_{ij}[\mathbf{Y} - \mathbf{UV}]_{ij}^2$ and $\|\overline{\mathbf{W}}\|_1$ equals to $\|\overline{\mathbf{W}}\|_0$ that imposes the sparsity on the outliers.

The hard weighting, for one thing, frequently leads the optimization to be stuck into bad local minima. For another thing, the pollution in data is often non-homogeneously distributed. To address the discreteness issue and reflect the importance of elements more faithfully, we employ an entropy term. The definition of entropy is $-\sum_{c=1}^{k} p_c \log p_c$ with $\sum_{c=1}^{k} p_c = 1$. The principle of maximum entropy tells that, the probability distribution which best represents the current state of knowledge is the one with largest entropy subject to accurately stated prior data. In other words, it is able to minimize the prediction bias. Return to our problem, the weighting variable w_{ij} can be equally viewed as the probability of the corresponding entry being classified as an outlier. It is instructive to note that maximizing the entropy (concave) is equivalent to minimizing its negative (convex). Consequently, we have the final formulation of ROUTE-LRMR as follows:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\mathbf{W}}\ & \frac{1}{2}\|\mathbf{U}\|_F^2+\frac{1}{2}\|\mathbf{V}\|_F^2+\frac{\alpha}{2}\|\sqrt{\mathbf{W}}\circ(\mathbf{Y}-\mathbf{U}\mathbf{V})\|_F^2\\ & +\beta\|\overline{\mathbf{W}}\|_1+\gamma\sum_{i,j}(w_{ij}\log w_{ij}+\bar{w}_{ij}\log\bar{w}_{ij})\\ \text{s.t.}\ & \mathbf{W}+\overline{\mathbf{W}}=\mathbf{1};\ \ \mathbf{W}\text{ and }\overline{\mathbf{W}}\in[0,1]^{m\times n},\end{aligned}\tag{5}$$

where γ is a non-negative coefficient controlling the importance of the corresponding term. Further, due to the relaxation, $\sqrt{\mathbf{W}}$ with entries $\sqrt{w_{ij}}$ is used to hold the equivalence: $\sum_{i,j}w_{ij}[\mathbf{Y}-\mathbf{U}\mathbf{V}]_{ij}^2=\|\sqrt{\mathbf{W}}\circ(\mathbf{Y}-\mathbf{U}\mathbf{V})\|_F^2$. For (4), $\|\sqrt{\mathbf{W}}\circ(\mathbf{Y}-\mathbf{U}\mathbf{V})\|_F^2=\|\mathbf{W}\circ(\mathbf{Y}-\mathbf{U}\mathbf{V})\|_F^2$.

2.2 Optimization

As we have seen in (5), it has embraced all the aforementioned concerns for simultaneously pursuing outliers and recovering the low rank matrix. The Augmented Lagrange Multiplier (ALM) with Alternating Direction Minimization (ADM) scheme [Lin *et al.*, 2011] has proven to be an efficient and effective solver for problems like (5). To apply ALM-ADM on our problem, the objective is required to be separable. To this end, an auxiliary variable $\mathbf{L}$ is introduced to replace $\mathbf{U}\mathbf{V}$ in the third term. Accordingly, $\mathbf{L}=\mathbf{U}\mathbf{V}$ performs as an additional constraint. It is worth noting that the constraints on $\mathbf{W}$ and $\overline{\mathbf{W}}$ are enforced as hard constraints. The augmented Lagrangian function of (5) is defined as:

$$\begin{aligned}&\mathcal{L}^{\mu}_{\{\mathbf{W}+\overline{\mathbf{W}}=\mathbf{1};\mathbf{W},\overline{\mathbf{W}}\in[0,1]^{m\times n}\}}(\mathbf{U},\mathbf{V},\mathbf{L},\mathbf{W},\mathbf{Z}):=\\ &\frac{1}{2}\|\mathbf{U}\|_F^2+\frac{1}{2}\|\mathbf{V}\|_F^2+\frac{\alpha}{2}\|\sqrt{\mathbf{W}}\circ(\mathbf{Y}-\mathbf{L})\|_F^2+\\ &\beta\|\overline{\mathbf{W}}\|_1+\gamma\sum_{i,j}(w_{ij}\log w_{ij}+\bar{w}_{ij}\log\bar{w}_{ij})+\\ &\frac{\mu}{2}\|\mathbf{L}-\mathbf{U}\mathbf{V}\|_F^2+\langle\mathbf{Z},\mathbf{L}-\mathbf{U}\mathbf{V}\rangle,\end{aligned}\tag{6}$$

where $\langle\cdot,\cdot\rangle$ designates the inner product, μ is a positive penalty and $\mathbf{Z}$ is a Lagrangian multiplier. The solver updates the variables in an iterative manner. For ease of exposition, we split the variables, except for $\mathbf{Z}$ and μ, into two groups, including Group I: $\{\mathbf{U},\mathbf{V},\mathbf{L}\}$ and Group II: $\{\mathbf{W},\overline{\mathbf{W}}\}$, connected by $\mathbf{L}$.

Group I – Dropping the unrelated terms yields:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\mathbf{L}}\ & \frac{1}{2}\|\mathbf{U}\|_F^2+\frac{1}{2}\|\mathbf{V}\|_F^2+\frac{\alpha}{2}\|\sqrt{\mathbf{W}}_{(p)}\circ(\mathbf{Y}-\mathbf{L})\|_F^2\\ & +\frac{\mu^{(t)}}{2}\|\mathbf{L}-\mathbf{U}\mathbf{V}\|_F^2+\langle\mathbf{Z}^{(t)},\mathbf{L}-\mathbf{U}\mathbf{V}\rangle.\end{aligned}\tag{7}$$

For all of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{L}$, their respective solutions in closed-form are calculated via equating the derivatives of (7) in $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{L}$ to zero:

$$\begin{aligned}\mathbf{U}_{(p+1)}&\leftarrow(\mu^{(t)}\mathbf{L}_{(p)}+\mathbf{Z}^{(t)})\mathbf{V}^T_{(p)}(\mathbf{I}+\mu^{(t)}\mathbf{V}_{(p)}\mathbf{V}^T_{(p)})^{-1};\\ \mathbf{V}_{(p+1)}&\leftarrow\mathbf{C}^{-1}_{(p+1)}\mathbf{U}^T_{(p+1)}(\mu^{(t)}\mathbf{L}_{(p)}+\mathbf{Z}^{(t)});\\ \mathbf{L}_{(p+1)}&\leftarrow\frac{\alpha\mathbf{W}_{(p)}\circ\mathbf{Y}+\mu^{(t)}\mathbf{U}_{(p+1)}\mathbf{V}_{(p+1)}-\mathbf{Z}^{(t)}}{\alpha\mathbf{W}_{(p)}+\mu^{(t)}\mathbf{1}},\end{aligned}\tag{8}$$

where $\mathbf{I}$ means the identity matrix with proper size, $\mathbf{C}_{(p+1)}$ stands for $\mathbf{I}+\mu^{(t)}\mathbf{U}^T_{(p+1)}\mathbf{U}_{(p+1)}$, and the division in updating $\mathbf{L}$ is element-wise.

Group II – Picking out the terms relevant to $\mathbf{W}$ and $\overline{\mathbf{W}}$ results in the following optimization problem:

$$\begin{aligned}\min_{\mathbf{W},\overline{\mathbf{W}}}\ & \frac{\alpha}{2}\|\sqrt{\mathbf{W}}\circ(\mathbf{Y}-\mathbf{L}_{(p+1)})\|_F^2+\beta\|\overline{\mathbf{W}}\|_1\\ & +\gamma\sum_{i,j}(w_{ij}\log w_{ij}+\bar{w}_{ij}\log\bar{w}_{ij})\\ \text{s.t.}\ & \mathbf{W}+\overline{\mathbf{W}}=\mathbf{1};\ \ \mathbf{W}\text{ and }\overline{\mathbf{W}}\in[0,1]^{m\times n}.\end{aligned}\tag{9}$$

From the objective of (9), we find that the problem can be decomposed into a set of independent sub-problems. Now, without any loss of generality, let us take the (i,j)-th element for example. Casting the problem into the Lagrange Multiplier framework gives the following Lagrange function:

$$\begin{aligned}\mathcal{Q}(w_i,\overline{w}_i,\lambda_i):=\frac{\alpha}{2}w_{ij}[\mathbf{Y}-\mathbf{L}_{(p+1)}]_{ij}^2+\beta\overline{w}_{ij}+\\ \gamma(w_{ij}\log w_{ij}+\overline{w}_{ij}\log\overline{w}_{ij})+\lambda_i(w_{ij}+\overline{w}_{ij}-1),\end{aligned}\tag{10}$$

where λ_i is a Lagrange multiplier. Taking the derivative of $\mathcal{Q}(w_i,\overline{w}_i,\lambda_i)$ to w_i, $\overline{w}_i$ and λ_i respectively and setting them to zero gives the optimal solution to w_i as:

$$\begin{aligned}w_{ij(p+1)}\leftarrow&\frac{\exp(-\alpha[\mathbf{Y}-\mathbf{L}_{(p+1)}]_{ij}^2/2\gamma)}{\exp(-\alpha[\mathbf{Y}-\mathbf{L}_{(p+1)}]_{ij}^2/2\gamma)+\exp(-\beta/\gamma)}\\ =&\frac{1}{1+\exp((\alpha[\mathbf{Y}-\mathbf{L}_{(p+1)}]_{ij}^2/2-\beta)/\gamma)},\end{aligned}\tag{11}$$

which is in a standard sigmoid form. And its complementary $\overline{w}_{ij(p+1)}\leftarrow 1-w_{ij(p+1)}$.

Remarks (a) When $w_{ij}\in\{0,1\}$ adopted and the entropy term disabled (hard weighting), the solution to Eq. (9) is: if $\frac{\alpha}{2}[\mathbf{Y}-\mathbf{L}]_{ij}^2<\beta$, then $w_{ij}\leftarrow 1$; if $\frac{\alpha}{2}[\mathbf{Y}-\mathbf{L}]_{ij}^2=\beta$, then w_{ij} could be either of $\{0,1\}$; otherwise $w_{ij}\leftarrow 0$. (b) When $w_{ij}\in[0,1]$ adopted and the entropy term disabled (relaxed version), the solution to Eq. (9) is: if $\frac{\alpha}{2}[\mathbf{Y}-\mathbf{L}]_{ij}^2<\beta$, then $w_{ij}\leftarrow 1$; if $\frac{\alpha}{2}[\mathbf{Y}-\mathbf{L}]_{ij}^2=\beta$, then w_{ij} could be any value in $[0,1]$; otherwise $w_{ij}\leftarrow 0$.

Multiplier and μ – The Lagrange multiplier $\mathbf{Z}$ and μ are updated via:

$$\begin{aligned}&\mathbf{Z}^{(t+1)}\leftarrow\mathbf{Z}^{(t)}+\mu^{(t)}(\mathbf{L}^{(t+1)}-\mathbf{U}^{(t+1)}\mathbf{V}^{(t+1)});\\ &\mu^{(t+1)}\leftarrow\mu^{(t)}\rho,\ \ \rho>1.\end{aligned}\tag{12}$$

The parameter μ is monotonically increased by ρ during iterations, leading the solution to the feasible region.

Algorithm 1: ROUTE-LRMR: Solver to Eq.(5)

```
Input: Observation matrix $\mathbf{Y} \in \mathbb{R}^{m\times n}$; support
       $\Omega \in \{0,1\}^{m\times n}$; a guess/target rank $r$;
       non-negative parameters $\alpha$, $\beta$ and $\gamma$.
Init.: $\mu^{(0)} \leftarrow 1$ and $\rho \leftarrow 1.1$; $\mathbf{W}_{(0)} \in \mathbb{R}^{m\times n} \leftarrow \Omega \circ \mathbf{1}$;
  $\mathbf{L}_{(0)} \in \mathbb{R}^{m\times n}$, $\mathbf{U}_{(0)} \in \mathbb{R}^{m\times r}$ and $\mathbf{V}_{(0)} \in \mathbb{R}^{r\times n}$ are all
  initialized randomly; $\mathbf{Z}^{(0)} \in \mathbb{R}^{m\times n} \leftarrow \mathbf{0}$; $t \leftarrow 0$.
while not converged do
    $p \leftarrow 0$;
    while not converged do
        Update $\mathbf{U}_{(p+1)}$, $\mathbf{V}_{(p+1)}$ and $\mathbf{L}_{(p+1)}$ via (8);
        for $\forall\ (i,j)\ \&\ \Omega_{ij}$ do
            Update $w_{ij(p+1)}$ via (11);
        end
        $p \leftarrow p+1$;
    end
    $\{\mathbf{L}^{(t+1)}, \mathbf{U}^{(t+1)}, \mathbf{V}^{(t+1)}\} \leftarrow \{\mathbf{L}_{(p)}, \mathbf{U}_{(p)}, \mathbf{V}_{(p)}\}$;
    Update $\mathbf{Z}^{(t+1)}$ and $\mu^{(t+1)}$ via (12); $t \leftarrow t+1$
end
Output: Optimal $\mathbf{W}^*$ and $\mathbf{L}^*$
```

For clarity and completeness, the procedure of solving (5) is outlined in Algorithm 1. The algorithm should not be terminated until the equality constraint $\mathbf{L} = \mathbf{UV}$ is satisfied up to a given tolerance, that is $\|\mathbf{L} - \mathbf{UV}\|_F \leq \varsigma\|\mathbf{Y}\|_F$, or the maximal number of iterations is reached. In all our experiments, the tolerance factor ς is chosen as $1e{-}7$. Please refer to the complete Algorithm 1 for other details that we can not cover in the text.

3 Theoretical Analysis

We first provide some useful theoretical results, including Lemma 1 and Proposition 1 for Group II, as well as Theorem 2 for the inner loop of Algorithm 1.

Lemma 1. *At stage p with $\mathbf{L}_{(p)}$ fixed, the solution to Eq.* (9) *(Group II), i.e. w_{ij} given in Eqn.*(11)*, is global optimal to the corresponding intermediary problem.*

Proof. First, having $\mathbf{L}_{(t)}$ fixed, the objective function in (9) is convex with respect to $w_{ij} \in [0,1]$. The solution in Eq.(11) is computed by the Lagrange multiplier method, which guarantees that the obtained solution is feasible and satisfies the KKT conditions for the convex problem (9). Thus, we reach the conclusion. □

Proposition 1. *The function defined in Eq.* (9)*, containing three parameters including $\tilde{\beta} := \beta/\alpha$, $\tilde{\gamma} := \gamma/\alpha$ and $\varepsilon_{ij} := [\mathbf{Y} - \mathbf{L}]^2_{ij}$, has the following properties:*

1. *$w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij})$ is monotonically decreasing with respect to ε_{ij}, which holds $\lim_{\varepsilon_{ij}\to 0} w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij}) = \frac{1}{1+\exp(-\tilde{\beta}/\tilde{\gamma})}$ and $\lim_{\varepsilon_{ij}\to +\infty} w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij}) = 0$;*
2. *$w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij})$ is monotonically increasing with respect to $\tilde{\beta}$, which holds that $\lim_{\tilde{\beta}\to 0} w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij}) = \frac{1}{1+\exp(\varepsilon_{ij}/\tilde{\gamma})}$ and $\lim_{\tilde{\beta}\to +\infty} w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij}) = 1$;*
3. *$w_{ij}(\tilde{\beta}, \tilde{\gamma}, \varepsilon_{ij})$ is an inverse-'S' shaped function, which approaches a binary function when $\tilde{\gamma} \to 0$ and the constant $1/2$ when $\tilde{\gamma} \to +\infty$.*

Each statement takes care of one target parameter with the others fixed to be constants.

Proof. It can be easily verified by the definition. □

Theorem 2. *At stage t with $\mathbf{Z}^{(t-1)}$ fixed, the inner loop of Algorithm 1 is guaranteed to converge to a partial minimum (either a stationary point or a local minimum) of the corresponding intermediary problem.*

Proof. The updating of $\mathbf{U}$, $\mathbf{V}$, $\mathbf{L}$ and $\mathbf{W}$ in the inner loop follows the manner of alternate convex search (ACS) [Bazaraa *et al.*, 1993]. As can be seen from Eq. (8) together with Lemma 1, the KKT conditions to all the involved variables are satisfied, which is sufficient to draw the conclusion [Gorski *et al.*, 2007]. □

Next, we shall consider the following lemmas required by analysis on convergence and optimality of the designed ROUTE-LRMR algorithm.

Lemma 2. *Let $\{(\mathbf{U}^{(t)}, \mathbf{V}^{(t)}, \mathbf{L}^{(t)}, \mathbf{W}^{(t)})\}$ be a sequence generated by Algorithm 1. Then the sequence approaches to a feasible solution.*

Proof. First, we come to prove the boundedness of $\{\mathbf{Z}^{(t)}\}$. According to Theorem 1 and the optimality condition for (5) with respect to $\hat{\mathbf{L}} := \mathbf{UV}$, we have:

$$\mathbf{Z}^{(t-1)} + \mu^{(t-1)}(\mathbf{L}^{(t)} - \mathbf{U}^{(t)}\mathbf{V}^{(t)}) = \mathbf{Z}^{(t)} \in \partial\|\hat{\mathbf{L}}^{(t)})\|_*.$$

Through applying Lemma 3 on the above:

Lemma 3. *[Fazel, 2002] Let $\mathcal{H}$ be a real Hilbert space endowed with an inner product $\langle\cdot,\cdot\rangle$ and a corresponding norm $\|\cdot\|$, and any $\mathbf{y} \in \partial\|\mathbf{x}\|$, where $\partial\|\cdot\|$ denotes the subgradient. Then $\|\mathbf{y}\|^* = 1$ if $\mathbf{x} \neq 0$, and $\|\mathbf{y}\|^* \leq 1$ if $\mathbf{x} = 0$, where $\|\cdot\|^*$ is the dual norm of the norm $\|\cdot\|$.*

we obtain that the sequence $\{\mathbf{Z}^{(t)}\}$ is bounded via observing the fact that the dual norm of $\|\cdot\|_*$ is the spectral norm. Together with the boundedness of $\{\mathbf{Z}^{(t)}\}$ and $\lim_{t\to\infty} \mu^{(t)} = \infty$, the relationship $\mathbf{L}^{(t)} - \mathbf{U}^{(t)}\mathbf{V}^{(t)} = \frac{\mathbf{Z}^{(t)} - \mathbf{Z}^{(t-1)}}{\mu^{(t-1)}}$ gives $\lim_{t\to\infty} \mathbf{L}^{(t)} - \mathbf{U}^{(t)}\mathbf{V}^{(t)} = \mathbf{0}$. Further, the constraints of $\mathbf{W} + \overline{\mathbf{W}} = \mathbf{1}$ and $\mathbf{W}, \overline{\mathbf{W}} \in [0,1]^{m\times n}$ are immediately satisfied at each update, please see Lemma 1. Thus the statement holds. □

Having the above theoretical results, we finally come to the convergence and optimality of ROUTE-LRMR.

Theorem 3. *The proposed Algorithm 1 converges to a partial minimum to the optimization problem* (5)*.*

Proof. By Lemmas 1 and 2, Theorem 2, and the updating rule of multiplier $\mathbf{Z}$, the KKT conditions for the constraints as well as the solutions to the variables $\mathbf{U}$, $\mathbf{V}$, $\mathbf{L}$, $\mathbf{W}$ and $\mathbf{Z}$ are all satisfied. The first order optimality of (5) is sufficient to guarantee that the ROUTE-LRMR converges to a partial optimum to the problem (5). □

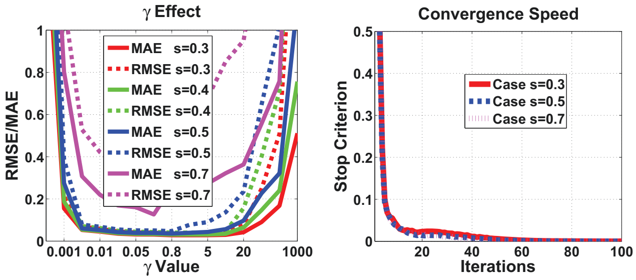

Figure 1: Parameter effect of γ and convergence speed

Method	PRMF	MoG	L1Reg	factEN	Unifying	PSMSV	HW	Ours
RMSE$_{s=0.3}$	0.1106	0.7440	1.2789	0.2121	0.0608	1.0214	0.2731	**0.0523**
MAE$_{s=0.3}$	0.0573	0.0984	0.2844	0.1069	0.0457	0.2813	0.2084	**0.0445**
RMSE$_{s=0.4}$	0.5803	0.9796	1.6146	0.3731	0.0762	1.9496	0.4089	**0.0624**
MAE$_{s=0.4}$	0.1716	0.1430	0.4432	0.2030	0.0565	0.7361	0.3114	**0.0480**
RMSE$_{s=0.5}$	1.0920	1.3344	1.9255	0.5181	0.0975	2.7295	0.7953	**0.0676**
MAE$_{s=0.5}$	0.3719	0.2201	0.6576	0.2868	0.0719	1.4310	0.5944	**0.0520**
RMSE$_{s=0.6}$	1.6075	1.6381	2.4863	0.6861	0.1492	3.7988	0.9730	**0.1092**
MAE$_{s=0.6}$	0.6287	0.3607	1.0438	0.4048	0.1093	2.3843	0.7236	**0.0651**
RMSE$_{s=0.7}$	2.2421	1.9513	3.2555	0.8734	0.3566	4.8753	1.7483	**0.3294**
MAE$_{s=0.7}$	1.0660	0.4956	1.7142	0.5677	0.2352	3.3894	1.2933	**0.2088**

Table 1: Performance comparison in terms of RMSE and MAE with different outlier ratios s. The numbers are averaged over 10 runs. The best results are highlighted in bold.

Further, based on Theorem 1, the problem (5) is equivalent to the following one:

$$\begin{aligned}\min_{\mathbf{L},\mathbf{W}}\ &\|\mathbf{L}\|_* + \frac{\alpha}{2}\|\sqrt{\mathbf{W}}\circ(\mathbf{Y}-\mathbf{L})\|_F^2 + \beta\|\overline{\mathbf{W}}\|_1\\ &+\gamma\sum_{i,j}(w_{ij}\log w_{ij} + \bar{w}_{ij}\log\bar{w}_{ij}) \qquad (13)\\ \text{s. t.}\ \ &\mathbf{W}+\overline{\mathbf{W}}=\mathbf{1};\ \ \mathbf{W}\text{ and }\overline{\mathbf{W}}\in[0,1]^{m\times n},\end{aligned}$$

which is biconvex in $\mathbf{W}$ and $\mathbf{L}$. The convergence to a partial optimum holds for the problem (13) too. Our ROUTE-LRMR is free to switch modes between NNM and BF. Concretely, instead of separately refreshing $\mathbf{U}$ and $\mathbf{V}$, the updating of $\hat{\mathbf{L}} := \mathbf{U}\mathbf{V}$ in (7) can be done by minimizing the problem:

$$\min_{\hat{\mathbf{L}}}\ \|\hat{\mathbf{L}}\|_* + \frac{\mu^{(t)}}{2}\|\mathbf{L}_{(p)}-\hat{\mathbf{L}}\|_F^2 + \langle\mathbf{Z}^{(t)},\mathbf{L}_{(p)}-\hat{\mathbf{L}}\rangle, \quad (14)$$

which can be solved in closed-form by the singular value thresholding [Cai *et al.*, 2010]. Except for this step, no other changes happen in Algorithm 1.

4 Experimental Verification

In this section, we assess the performance of ROUTE-LRMR in comparison with several state-of-the-art methods including RegL1 [Zheng *et al.*, 2012], PRMF [Wang *et al.*, 2012], MoG [Meng and De la Torre, 2013], factEN [Kim *et al.*, 2015], PSMSV [Oh *et al.*, 2016] and Unifying [Cabral *et al.*, 2013], the codes of which are either downloaded from the authors' websites or provided by the authors. Their settings follow the suggestions by the authors or the given parameters.

4.1 Synthetic Data

Data Preparation and Quantitative Metrics Similar to [Candès *et al.*, 2011; Cabral *et al.*, 2013], we generate a matrix $\mathbf{Y}_0$ as a product $\mathbf{Y}_0 = \mathbf{U}_0\mathbf{V}_0$. The $\mathbf{U}_0$ and $\mathbf{V}_0$ are of size $m \times r$ and $r \times n$ respectively, both of which are randomly produced by sampling each entry from the Gaussian distribution $\mathcal{N}(0, 1)$, leading to a ground truth rank-r matrix. Then we corrupt the entries via replacing a fraction s of $\mathbf{Y}_0$ with errors drawn from a uniform distribution over $[-20, 20]$ at random, and the rest entries are polluted by Gaussian noise $\mathcal{N}(0, 0.1^2)$. To quantitatively measure the recovery performance, we employ 1) *root mean square error* (RMSE): $\frac{1}{\sqrt{mn}}\|\mathbf{Y}_0 - \hat{\mathbf{U}}\hat{\mathbf{V}}\|_F$ and 2) *mean absolute error* (MAE): $\frac{1}{mn}\|\mathbf{Y}_0 - \hat{\mathbf{U}}\hat{\mathbf{V}}\|_1$.

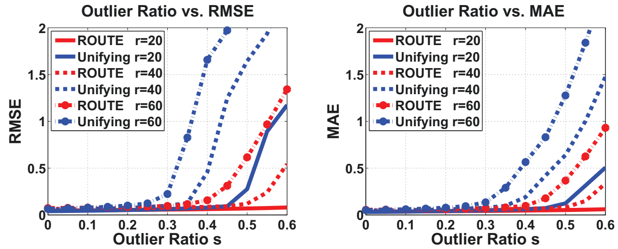

Figure 2: Outlier ratio s versus RMSE and MAE

Parameter Effect We here focus on the parameter γ that controls the entropy term, the other two parameters α and β are empirically set to 50 and 1 throughout this paper. In this experiment, without loss of generality, square matrices of dimension $m = n = 100$ and rank $r = 4$ are considered. The left picture in Fig. 1 depicts RMSE and MAE curves (averaged over 10 trials) with respect to different outlier ratios. From the plots, we see that when γ approaches to 0, the errors rapidly go up. This is because, as analyzed in Sec. 3, the smaller γ is, the harder the weighting carries out, say the risk of being stuck into bad minima gets higher. It is also the evidence to prove the soft weighting is beneficial. In opposite, if γ gets too large, the performance also drops. The reason is that, in this situation, the weighting becomes almost constant (0.5 for each entry), which degenerates ROUTE-LRMR to PCA. Although the work range of γ shrinks as s grows, γ in $[0.005, 0.8]$ can perform stably and sufficiently well. For the rest experiments unless stated otherwise, we set $\gamma = 0.01$. To better reveal the advantage of our method over the competitors especially on heavily ruined data, Table 1 reports the numerical comparison. As can be seen from Tab. 1, ROUTE-LRMR wins for all the cases, and the closest performance to ours is from Unifying. Note that the method HW is ROUTE with $\gamma = 0.001$ for mimicking the hard weighting strategy.

Convergence As regards convergence speed, for different cases, the right graph in Fig. 1 shows that the stop criterion quickly declines within 20 iterations, while the algorithm converges within 60 $\sim$ 80 iterations. Moreover, experimental findings here and follow-up tell that our algorithm has very stable convergence behavior even with respect to random initializations.

Tolerance to Outliers To more thoroughly show the tolerance to outliers, we fix $m = n = 400$ and test the tendency by varying outlier ratio $s \in [0, 0.6]$ and rank $r \in \{20, 40, 60\}$. According to the results in Tab. 1, Unifying is the method chosen to compare. From the left picture of Fig. 2, we see that at the beginning, Unifying and our method are close in

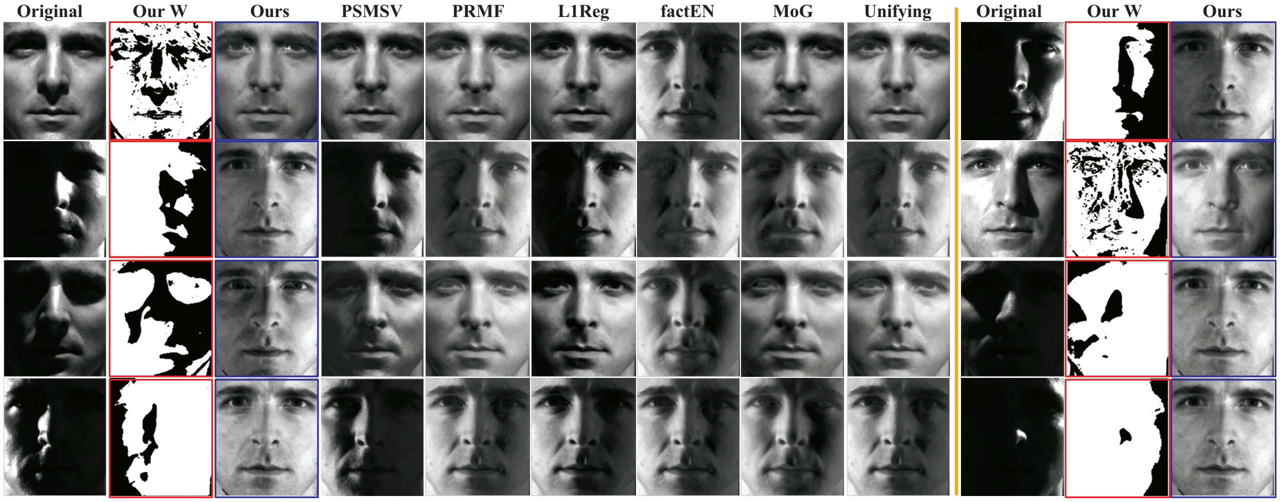

Figure 3: Visual comparison on the task of photometric stereo. ROUTE-LRMR adopts $\gamma = 0.001$ in this experiment.

terms of RMSE, but as s increases, the margin between them enlarges. The second graph in Fig. 2 further confirms the first one. In the case of $r = 20$, both the RMSE and MAE of ROUTE-LRMF stay very low even when s reaches 0.6. The tolerance to outliers becomes weaker when r gets larger, not just for our method and Unifying but also for all the methods. The reason is that a higher-dimensional space requires more data to accomplish the recovery.

4.2 Real Data

Photometric Stereo Images of a static Lambertian object sensed by a fixed camera under a varying but distant point lighting source lie in a rank-3 subspace [Hayakawa, 1994]. This experiment aims to evaluate the effectiveness of the LRMR techniques on modeling the face under different illuminations. The cropped Extended YaleB-10 sequence, containing 64 faces of one subject with size 192×168, is adopted as the dataset. The light imbalance including shadows and highlights on the face significantly breaks the low-rank structure (please see the 1^{st} column in Fig. 3 for example). In this part, we set the guess rank r to 5 for all the competitors.

Comparison Figure 3 gives several comparison. We can observe that PRMF, factEN, MoG and Unifying perform reasonably well, which are superior to PSMSV and L1Reg but inferior to ours. As shown in the 2^{nd} and 4^{th} rows of Fig. 3, PSMSV and L1Reg fail to remove shadows. The results by PRMF, factEN, MoG and Unifying, although recalling some details previously hidden in the dark, look unreal in the 2^{nd} and 3^{rd} cases. Our ROUTE-LRMR[1] provides visually pleasant and real results for all the given cases, the benefit of which mainly comes from the effective outlier detection. The 2^{nd} column in Fig. 3 displays the estimated weights $\mathbf{W}$ (brighter regions indicate closer values to 1, while darker ones stand for those to 0), from which we can find our strategy successfully detects and thus eliminates outliers. On the right side of Fig. 3, we further provide several results by our method (only, due to space limit). One may wonder if the weights can be formed by treating as outliers the pixels with intensity greater (highlights) or lower (shadows) than predefined thresholds like [Zheng *et al.*, 2012]. This way can reduce the problem to LRMC, but is too heuristic, at high risk of sacrificing much useful information for recovery. Taking the bottom-right original for example, the thresholding may determine all the pixels as outliers, while our strategy can finish the job wisely and nicely. Moreover, in many real-world applications, manually seeking appropriate thresholds is, if not impossible, very difficult. Being able to adaptively assign weights to data is definitely desired, which is the goal and motivation of our design.

5 Conclusion

This paper has shown a method for jointly detecting outliers and recovering the underlying low-rank matrix, called ROUTE-LRMR. Our weighting strategy employs an entropy regularization term to minimize the prediction bias, which behaves like a sigmoid function. To seek the optimal solution for ROUTE-LRMR, we have developed an Alternating Direction Minimization based algorithm. The theoretical analysis and the experimental results compared to the state-of-the-arts, have demonstrated the advantages of the proposed ROUTE-LRMR. Our strategy can be applied to numerous tasks such as regression, clustering, inpainting and foreground detection. It is also ready to embrace specific domain knowledge, like graph regularizer on the weight, for further boosting the performance on different applications.

Acknowledgments

X. Guo is supported by National Natural Science Foundation of China (grant no. 61402467). Z. Lin is supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502) and National Natural Science Foundation of China (grant nos. 61625301 and 61231002).

[1] In image/video data, the outliers, such as shadows and foregrounds, often appear coherently. Considering this, in this experiment, we employ a 2×2 median filter on $\mathbf{W}$.

References

[Bazaraa *et al.*, 1993] M. Bazaraa, H. Sherali, and C. Shetty. *Nonlinear Programming-Theory and Algorithms*. John Wiley and Sons Inc., 1993.

[Cabral *et al.*, 2013] R. Cabral, F. De la Torre, J. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *ICCV*, 2013.

[Cai *et al.*, 2010] J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[Candès *et al.*, 2011] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[Chen *et al.*, 2014] C. Chen, X. Zheng, Y. Wang, F. Hong, and Z. Lin. Context-aware collaborative topic regression with social matrix factorization for recommender systems. In *AAAI*, 2014.

[Eriksson and van den Hengel, 2010] A. Eriksson and A. van den Hengel. Efficient computation of robust low-rank matrix approximation in the presence of missing data using the ℓ_1 norm. In *CVPR*, 2010.

[Fazel, 2002] M. Fazel. Matrix rank minimization with applications. *PhD Thesis, Stanford University*, 2002.

[Gorski *et al.*, 2007] J. Gorski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math Meth Oper Res*, 66(3):373–407, 2007.

[Gu *et al.*, 2014] S. Gu, L. Zhang, W. Zuo, and X. Feng. Weighted nuclear norm minimization with applications to image denoising. In *CVPR*, 2014.

[Guo *et al.*, 2013] X. Guo, X. Cao, X. Chen, and Y. Ma. Video editing with temporal, spatial and appearance consistency. In *CVPR*, 2013.

[Guo *et al.*, 2014] X. Guo, X. Cao, and Y. Ma. Robust separation of reflection from multiple images. In *CVPR*, 2014.

[Hayakawa, 1994] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *JOSA*, 11:3079–3089, 1994.

[Jing *et al.*, 2015] X. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 2015.

[Kim *et al.*, 2015] E. Kim, M. Lee, and S. Oh. Elastic-net regularization of singular values for robust subspace learning. In *CVPR*, 2015.

[Lakshminarayanan *et al.*, 2011] B. Lakshminarayanan, G. Bouchard, and C. Archambeau. Robust bayesian matrix factorisation. In *AISTATS*, 2011.

[Lin *et al.*, 2011] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low rank representation. In *NIPS*, 2011.

[Liu *et al.*, 2013] G. Liu, Z. Lin, S. Yan, J. Sun, and Y. Yu. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[Mazumder *et al.*, 2010] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *JMLR*, 99:2287–2322, 2010.

[Meng and De la Torre, 2013] D Meng and F. De la Torre. Robust matrix factorization with unknown noise. In *ICCV*, 2013.

[Nie and Huang, 2016] F. Nie and H. Huang. Subspace clustering via new discrete group structure constrained low-rank model. In *IJCAI*, 2016.

[Oh *et al.*, 2016] T. Oh, Y. Tai, J. Bazin, H. Kim, and I. Kweon. Partial sum minimization of singular values in robust pca: Algorithm and applications. *TPAMI*, 4(3):744–758, 2016.

[Pan *et al.*, 2017] J. Pan, Z. Hu, Z. Su, and M.-Y. Yang. L0-regularized intensity and gradient prior for deblurring text images and beyond. *TPAMI*, 39(2):342–355, 2017.

[Pearson, 1901] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901.

[Recht *et al.*, 2010] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[Salakhutdinov and Mnih, 2008] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, 2008.

[Srebro and T.Jaakkola, 2003] N. Srebro and T.Jaakkola. Weighted low-rank approximations. In *ICML*, 2003.

[Wang *et al.*, 2012] N. Wang, T. Yao, J. Wang, and D. Yeung. A probabilistic approach to robust matrix factorization. In *ECCV*, 2012.

[Zhang and Wang, 2016] Q. Zhang and H. Wang. Collaborative filtering with generalized laplacian constraint via overlapping decomposition. In *IJCAI*, 2016.

[Zheng *et al.*, 2012] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi. Practical low-rank matrix approximation under robust l1-norm. In *CVPR*, 2012.

[Zhou *et al.*, 2010] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma. Stable principal component pursuit. In *ISIT*, 2010.

[Zhou *et al.*, 2013] X. Zhou, C. Yang, and W. Yu. Moving object detection by detecting contiguous outliers in the low-rank representation. *TPAMI*, 35(3):597–610, 2013.