IMPROVED ROBUST PCA USING LOW-RANK DENOISING WITH OPTIMAL SINGULAR VALUE SHRINKAGE

Brian E. Moore, Raj Rao Nadakuditi, and Jeffrey A. Fessler

University of Michigan, Dept. of EECS, 1301 Beal Avenue, Ann Arbor, MI 48109, USA

ABSTRACT

We study the robust PCA problem of reliably recovering a low-rank signal matrix from a signal-plus-noise-plus-outliers matrix. We analytically characterize the extent to which the singular vectors of the signal-plus-noise-plus-outliers matrix can be degraded by outliers and discuss why a recently proposed method for robust PCA that exploits outlier sparsity to improve low-rank estimation will produce suboptimal low-rank matrix estimates in the presence of noise. Next, we propose a new iterative algorithm for robust PCA that utilizes an optimal, data-driven low-rank matrix estimator (OptShrink). Finally, we show that the proposed approach yields superior background subtraction on a computer vision dataset.

Index Terms— low-rank plus sparse decomposition, robust PCA, random matrix theory

1. INTRODUCTION

Principal component analysis (PCA) is a powerful technique for uncovering latent low-rank structure in high dimensional datasets. It is ubiquitous in statistical signal processing theory and practice and is the first step in many inferential procedures for detection, estimation and classification. It is well-known, however, that PCA is relatively brittle in the sense that relatively few outliers can severely degrade the quality of low-rank components estimated from noisy data. This, in turn, degrades the performance of inferential tasks that utilize these estimated low-rank components. Robust PCA aims to mitigate such problems by producing the 'best' (with respect to squared error) low-rank estimates that are robust to outlier contamination.

Recent breakthroughs [1, 2, 3] have established that, in the noise-free setting, convex optimization-based approaches can reliably recover low-rank components in the presence of outliers. Indeed, sufficient conditions for error-free recovery of the low-rank matrix are given in [1, 2, 3] for the noiseless case. Much less is known, however, about the noisy setting, except the unsurprising fact that one cannot expect error-free recovery. There is no theoretical reason to expect convex optimization-based algorithms for robust PCA that have been designed for the noise-free setting to also be optimal in the noisy setting.

In [7] it is shown that, in the noisy outlier-less setting, the low-rank components produced by solving any convex optimization problem are provably suboptimal. Indeed, in [7], an algorithm (OptShrink) for optimal low-rank matrix denoising is proposed that provably outperforms convex optimization-based methods. The primary contribution of this paper is the development of an iterative algorithm for robust PCA that utilizes OptShrink to improve low-rank component estimation.

The paper is organized as follows. In Section 2, we formulate the robust PCA problem, and, in Section 3, we analytically characterize the effect of outliers on the low-rank estimation problem. We describe a convex optimization-based approach to robust PCA in Section 4, discuss why it is suboptimal in Section 5, and propose an improved algorithm in Section 6. Finally, in Section 7, we demonstrate the performance gains of our proposed approach on a background subtraction task, and we offer some conclusions in Section 8.

2. PROBLEM FORMULATION

Consider the setting where the $n \times m$ observed signal-plus-noise-plus-outliers matrix $\widetilde{X}$ is modeled as

$$\widetilde{X} = \underbrace{\sum_{i=1}^{r} \theta_i u_i v_i^H}_{=:L} + S + X. \tag{1}$$

In (1), L represents the rank-r low-rank signal matrix that we are interested in reliably recovering, where, for $i = 1, \dots, r$, u_i and v_i are the left and right singular vectors associated with singular value θ_i. The matrix S is modeled as

$$S_{ij} = \begin{cases} q_{ij} & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases}$$

where q_{ij} are elements drawn from an unknown distribution $q(s)$ with $\mathbb{E}[q_{ij}] = 0$, $\mathbb{E}[|q_{ij}|^2] = \sigma_q^2$, and $\mathbb{E}[|q_{ij}|^4] < \infty$. The S matrix represents a sparse matrix of outliers (relative to L). The matrix X has elements that are independently and identically distributed with zero mean, variance σ^2/m, and bounded fourth moment. In (1), we assume the outliers are sparse with respect to the standard Euclidean basis. If they are sparse with respect to some other basis (e.g. Fourier or wavelet), we can, without loss of generality, assume that (1) holds after an appropriate sparsifying transformation has been applied to the vectorized elements of the observed matrix.

Work supported by ONR YIP Award N000141110660, NSF award CCF-1116115, AFOSR YI Award FA9550-12-1-0266, ARO MURI grant W911NF-11-1-0391, and NIH grant 1P01-CA87634.

978-1-4799-4975-5/14/$31.00 ©2014 IEEE

3. ROBUST PCA MOTIVATION

Our goal is to estimate, as accurately as possible, the low-rank component L from the matrix $\widetilde{X}$ under the model (1). This objective is complicated by the presence of the outlier matrix S, which we assume is not low-rank. Let $\widetilde{X} = \sum_{i=1}^{q} \widetilde{\sigma}_i \widetilde{u}_i \widetilde{v}_i^H$ be the singular value decomposition (SVD) of $\widetilde{X}$, where $\widetilde{\sigma}_1 \geq \ldots \widetilde{\sigma}_q$ and $q = \min(n, m)$ is the rank of $\widetilde{X}$. Recent results from random matrix theory [4] can be extended to quantify the degradation incurred when estimating the singular vectors of L in the presence of outliers. We omit the proof here due to space considerations.

Theorem 1. *Assume that the singular vectors of L satisfy a 'low-coherence' condition, i.e.,*

$$\max_{i=1,\ldots r} ||u_i||_\infty \leq C_u \frac{\log^{\eta_u} n}{\sqrt{n}} \;\&\; \max_{i=1,\ldots r} ||v_i||_\infty \leq C_v \frac{\log^{\eta_v} m}{\sqrt{m}}$$

for some universal constants $C_u, C_v, \eta_u, \eta_v > 0$. Suppose that $\theta_1 > \ldots > \theta_r > 0$ in (1) and that r is fixed. Then, as $n, m_n \to \infty$ with $n/m_n \to c \in (0, \infty)$, we have

$$|\langle u_i, \widetilde{u}_i \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \dfrac{c\left(1 + \bar{\theta}_i^2\right)}{\bar{\theta}_i^2 \left(\bar{\theta}_i^2 + c\right)} & \text{if } \bar{\theta}_i > c^{1/4} \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

and

$$|\langle v_i, \widetilde{v}_i \rangle|^2 \xrightarrow{a.s.} \begin{cases} 1 - \dfrac{\left(c + \bar{\theta}_i^2\right)}{\bar{\theta}_i^2 \left(\bar{\theta}_i^2 + 1\right)} & \text{if } \bar{\theta}_i > c^{1/4} \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where

$$\bar{\theta}_i = \lim_{m \to \infty} \frac{\theta_i}{\sqrt{p\, m\, \sigma_q^2 + \sigma^2}}. \tag{4}$$

Theorem 1 brings into sharp focus the detrimental effect of sparse outliers on the estimation of low-rank components. For example, suppose that $\sigma_q^2 = O(1)$ and $p = O(1)$. Then, by (4), $\bar{\theta}_i = 0$ so that $|\langle u_i, \widetilde{u}_i \rangle|^2 \to 0$ and $|\langle v_i, \widetilde{v}_i \rangle|^2 \to 0$ as $m \to \infty$, irrespective of the magnitude of θ_i relative to the noise variance σ^2 (assuming θ_i is bounded with m). Consequently, the singular vectors of $\widetilde{X}$ will be poor estimates of the singular vectors of L. In contrast, when $p = 0$ so that we are in the outlier-free setting and $\theta_i / \sigma \gg c^{1/4}$, we can expect the singular vectors of $\widetilde{X}$ to be good estimates of the singular vectors of L.

More generally, from Theorem 1 and (4), we conclude that the singular vectors of $\widetilde{X}$ will be very poor estimates of the singular vectors of L whenever $\sigma_q^2 = O(1)$ and $pm \to \infty$. The latter condition includes the few-outlier setting when $p = O(\log m/m)$, so that an average of just $O(m \log m)$ corrupted entries out of the mn total entries will suffice to severely degrade the eigen-structure of the matrix $\widetilde{X}$. This motivates the development of robust PCA methods for reliably extracting low-rank structure in the presence of outliers.

4. ROBUST PCA VIA CONVEX OPTIMIZATION

This section briefly reviews existing convex optimization-based approaches to robust PCA as a precursor to our proposed algorithm. Consider the optimization problem

$$\{\widehat{L}_{\text{svt}}, \widehat{S}_{\text{svt}}\} = \arg\min_{L,S} \frac{1}{2} \|\widetilde{X} - (L+S)\|_2^2 + \lambda_L \|L\|_\star + \lambda_S \|S\|_1, \tag{5}$$

where $\|.\|_\star$ denotes the nuclear norm (*i.e.,* sum of singular values) and $\|.\|_1$ denotes the ℓ_1 norm. It was shown [1, 2] that, under the hypotheses in Theorem 1, the solution to (5) in the noise-free setting yields $\widehat{L}_{\text{svt}} = L$ and $\widehat{S}_{\text{svt}} = S$ with very high probability. Thus (5) can accurately recover low-rank structure in the presence the outliers.

One can solve (5) iteratively using the proximal gradient method [5], where, at the k-th iteration, one computes the updates

$$\begin{aligned} L_{k,\text{svt}} &:= \mathbf{SVT}_{\tau_k \lambda_L} \left(M_{k-1,\text{svt}} - S_{k-1,\text{svt}}\right) \\ S_{k,\text{svt}} &:= \mathbf{soft}_{\tau_k \lambda_S} \left(M_{k-1,\text{svt}} - L_{k-1,\text{svt}}\right) \\ M_{k,\text{svt}} &:= L_{k,\text{svt}} + S_{k,\text{svt}} - \tau_k (L_{k,\text{svt}} + S_{k,\text{svt}} - \widetilde{X}), \end{aligned} \tag{6}$$

where τ_k denotes the step size at the k-th iteration. In (6), $\mathbf{SVT}(.)$ is the singular value thresholding operator [2]

$$\mathbf{SVT}_\lambda(Y) = \sum_{i=1}^{q} (\sigma_i - \lambda)_+ u_i v_i^H, \tag{7}$$

$Y = U \Sigma V^H$ is the SVD of Y, $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_q)$, and $\mathbf{soft}(.)$ is the element-wise soft thresholding operator

$$[\mathbf{soft}_\lambda(Y)]_{ij} = \mathbf{sign}(Y_{ij})(|Y_{ij}| - \lambda)_+, \tag{8}$$

where $(y)_+ := \max(y, 0)$. Because $\|.\|_\star$ and $\|.\|_1$ are convex, standard convergence results [6] can be used to show that the iterates $L_{k,\text{svt}}$ and $S_{k,\text{svt}}$ will eventually minimize the cost function of (6) provided that $\tau_k = \tau < 1$. The matrix $M_{k,\text{svt}}$ from the updates (6) can be interpreted as a low-rank plus sparse matrix minus a residual (*i.e.,* noise-only) term. Therefore, the matrix $M_{k-1,\text{svt}} - S_{k-1,\text{svt}}$ from the L-update of (6) is approximately a low-rank plus noise matrix, and one can interpret $\mathbf{SVT}(M_{k-1,\text{svt}} - S_{k-1,\text{svt}})$ as a low-rank matrix denoising operation with singular value shrinkage defined by (7). Because our goal is to optimally recover the low-rank component, we now argue, following [7], why singular value thresholding will produce suboptimal low-rank estimates.

5. SUBOPTIMALITY OF SVT

Consider the oracle low-rank denoising problem

$$\boldsymbol{w}^{\text{opt}} = \underset{[w_1, \ldots, w_r]^T \in \mathbb{R}^r}{\arg\min} || \sum_{i=1}^{r} \theta_i u_i v_i^H - \sum_{i=1}^{r} w_i \widetilde{u}_i \widetilde{v}_i^H ||_F, \tag{9}$$

which seeks the best approximation of a low-rank signal matrix (here L) by an optimally weighted combination of estimates of its left and right singular vectors.[1] Equation (9) can

[1] In (9), it is not necessary to specify how the singular vectors were estimated.

be solved in closed-form (see [7]) and yields the expression

$$w_i^{\text{opt}} = \sum_{j=1}^{r} \theta_i \left(\widetilde{u}_i^H u_j\right) \cdot \left(v_j^H \widetilde{v}_i\right) \qquad \text{for } i = 1, \ldots, r. \tag{10}$$

When $\widetilde{u}_i$ and $\widetilde{v}_i$ are good estimates of u_i and v_i, respectively, we expect $(\widetilde{u}_i^H u_i)$ and $(v_i^H \widetilde{v}_i)$ to be close to 1. Consequently, from (10), we expect $w_i^{\text{opt}} \approx \theta_i$. Conversely, when $\widetilde{u}_i$ and $\widetilde{v}_i$ are poor estimates of u_i and v_i, respectively, then we expect $(\widetilde{u}_i^H u_i)$ and $(v_i^H \widetilde{v}_i)$ to be closer to 0 and for w_i^{opt} to be smaller than θ_i. This is why shrinking the singular values of the data improves low-rank matrix denoising; it accounts for the errors in estimating the singular vectors. Intuitively, we expect well-estimated subspaces to be shrunk less than poorly estimated subspaces. Indeed, from Theorem 1, we see that, if $\theta_i \to \infty$, then $(\widetilde{u}_i^H u_i) \to 1$ and $(v_i^H \widetilde{v}_i) \to 1$. Thus, the optimal shrinkage function should ensure that the *absolute* shrinkage applied should vanish as $\theta_i \to \infty$. Consequently, any shrinkage function for which the absolute shrinkage does not vanish as $\theta_i \to \infty$ will necessarily be suboptimal. Applying this reasoning allows us to conclude that the SVT shrinkage operator in (7) will necessarily produce suboptimal low-rank estimates, since, for every $\lambda_L > 0$, the absolute shrinkage does not vanish as $\sigma_i \to \infty$. See [7] for more details.

6. OPTSHRINK AND A NEW ALGORITHM

With the insight that the SVT-based low-rank denoising in (6) produces suboptimal low-rank estimates, we now have reason to believe that an algorithm that can improve low-rank estimation should, in turn, improve the performance of robust PCA. Towards this end, we now specify a concrete algorithm for optimal low-rank matrix denoising.

Theorem 2 (OptShrink [7])**.** *Assume that the hypotheses in Theorem 1 are satisfied and S is known. Let $\widetilde{\sigma}_i$ denote the i-th largest singular value of $\widetilde{X} - S$. Consider the solution to* (9) *given by* (10) *with $\widetilde{u}_i$ and $\widetilde{v}_i$ equal to the left and right singular vectors of $\widetilde{X} - S$, and let $\widetilde{\sigma}_i$ denote the corresponding singular value. Then as $n, m_n \to \infty$ with $n/m_n \to c \in (0, \infty)$, we have that*

$$w_i^{\text{opt}} \xrightarrow{a.s.} -2 \, \frac{D_{\widehat{\mu}_X}(\widetilde{\sigma}_i)}{D'_{\widehat{\mu}_X}(\widetilde{\sigma}_i)} \qquad \textit{for } \ i = 1, \ldots, r, \tag{11}$$

where $q = \min(m, n)$, $\mathrm{d}\widehat{\mu}_X(t) = \frac{1}{q-r} \sum_{i=r+1}^{q} \delta\left(t - \widetilde{\sigma}_i\right)$, and the D-transform is defined as

$$\begin{aligned} D_{\widehat{\mu}_X}(z) := \quad & \left[\int \frac{z}{z^2 - t^2} \, \mathrm{d}\widehat{\mu}_X(t) \right] \times \\ & \left[c \int \frac{z}{z^2 - t^2} \, \mathrm{d}\widehat{\mu}_X(t) + \frac{1 - c}{z} \right], \end{aligned} \tag{12}$$

with $D'_{\widehat{\mu}_X}(z) = \frac{\partial}{\partial z} D_{\widehat{\mu}_X}(z)$.

Theorem 2 provides a concrete algorithm (OptShrink) that was shown in [7] to asymptotically solve (9). This motivates the following proposed robust PCA algorithm where, at the k-th iteration, one computes the updates

$$\begin{aligned} L_k \quad &:= \quad \mathbf{OptShrink}_r \left(M_{k-1} - S_{k-1} \right) \\ S_k \quad &:= \quad \mathbf{soft}_{\tau_k \lambda_S} \left(M_{k-1} - L_{k-1} \right) \\ M_k \quad &:= \quad L_k + S_k - \tau_k (L_k + S_k - \widetilde{X}). \end{aligned} \tag{13}$$

In (13), τ_k denotes the step size at the k-th iteration, and we have replaced the $\mathbf{SVT}(.)$ operator from (6) with

$$\mathbf{OptShrink}_r(Y) = \sum_{i=1}^{r} \left(-2 \, \frac{D_{\mu_Y}(\sigma_i)}{D'_{\mu_Y}(\sigma_i)} \right) u_i v_i^H, \tag{14}$$

where $Y = U \Sigma V^H$ is the SVD of Y, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_q)$, and

$$\mathrm{d}\mu_Y(t) = \frac{1}{q - r} \sum_{i=r+1}^{q} \delta\left(t - \sigma_i\right), \tag{15}$$

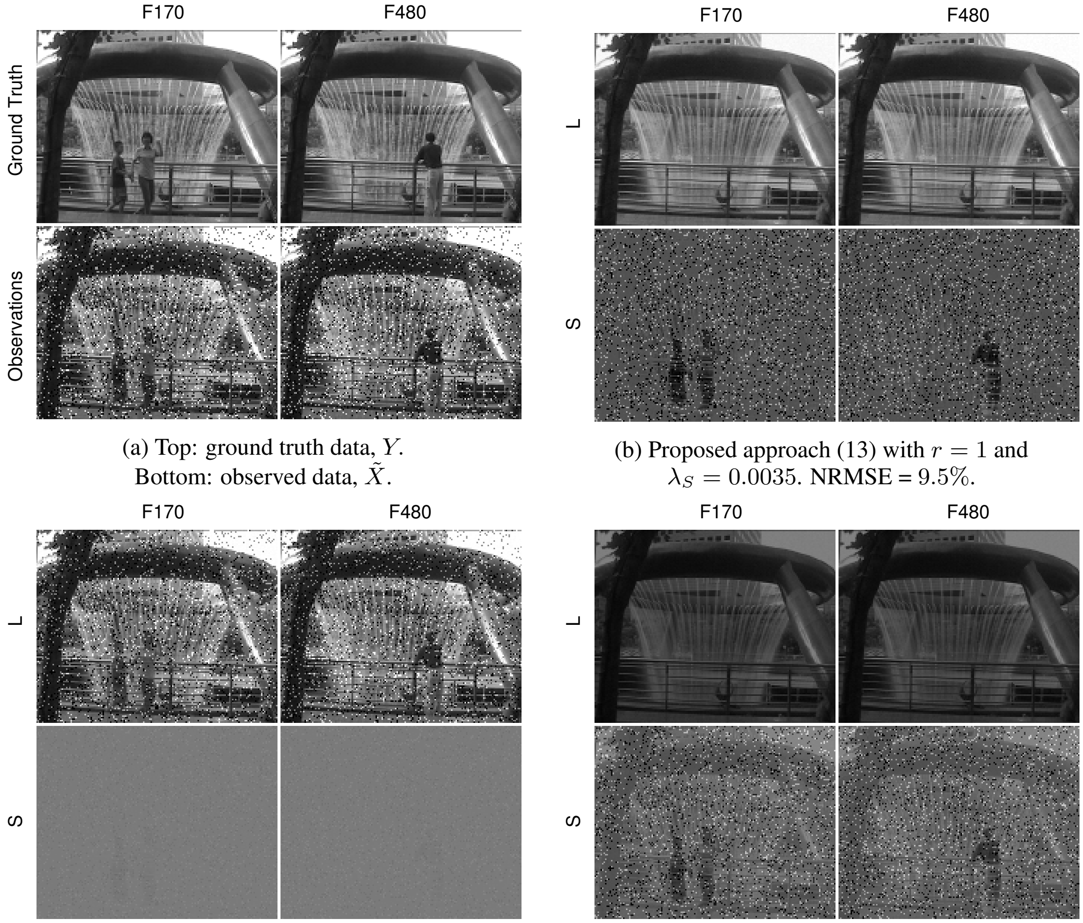
as suggested by Theorem 2. Because the computational cost of the L-updates in both (6) and (13) are dominated by the cost of computing the SVD, the additional complexity of the singular value shrinkage in (14) is negligible.

7. RESULTS

We now compare the low-rank reconstruction quality of the proposed update scheme (13) with the SVT-based updates from (6) on a fountain dataset.[2] The data contains $m = 523$ images, each with resolution $n_x \times n_y = 128 \times 160$, of a scene with a fountain in the background and people walking intermittently in the foreground. We arranged the raw data in an $n \times m$ matrix Y whose columns contain the $n = n_x n_y$ vectorized pixels of each image. In the language of model (1), one can think of $Y = X + L$ as the low-rank plus noise component of the observed data. As such, we generated observations $\widetilde{X} = Y + S$ with S_{ij} drawn as in (1) with $q_{ij} = \pm K$ equiprobably for a given $K > 0$. Finally, we packaged each update scheme into an algorithm by initializing $L_0 = M_0 = \widetilde{X}$ and $S_0 = 0$ and terminating when $\|M_k - M_{k-1}\|_F < \delta \|M_{k-1}\|_F$ for a given stopping tolerance $\delta > 0$.

Figure 1 displays the algorithm outputs for a representative pair of image frames. Figure 1c shows that applying (6) with small λ_L fails to recover the low-rank background; indeed, $\hat{L}_{\text{svt}}$ was full-rank. On the other hand, Figure 1d shows that setting λ_L large enough to force $\mathbf{rank}(\hat{L}_{\text{svt}}) = 1$ is also problematic because the SVT-based L-update in (6) uniformly shrinks all singular values, which, for large λ_L, results in suboptimal degradation of large singular values (cf. Section 5). This effect manifests in Figure 1d through the dimness of the low-rank components and the leakage of the remaining background intensity into the sparse component. In contrast, the proposed approach successfully isolates the fountain background in its low-rank component.

[2] Obtained from `perception.i2r.a-star.edu.sg/bk_model/bk_index.html`.



(a) Top: ground truth data, Y.
Bottom: observed data, $\tilde{X}$.

(b) Proposed approach (13) with $r = 1$ and $\lambda_S = 0.0035$. NRMSE = 9.5%.

(c) SVT-based updates (6) with $\lambda_L = 6.5$ and $\lambda_S = 0.0035$. NRMSE = 38.4%.

(d) SVT-based updates (6) with $\lambda_L = 300$ and $\lambda_S = 0.0035$. NRMSE = 42.3%.

Fig. 1: Simulation results for the fountain dataset with parameters $\delta = 0.0025$, $\tau_k = 0.5$, $p = 0.15$, and $K = 0.5$. The row labels L and S denote the low-rank and sparse components, respectively, returned by each algorithm. The column labels denote the frame number (*i.e.,* column of L and S) that is displayed. Each subfigure has the same intensity scale. NRMSE values are reported for the low-rank components using output of the SVT-based updates with $p = 0$ as ground truth.

8. CONCLUSIONS

We proposed a new iterative robust PCA method that utilizes an optimal low-rank matrix estimator (OptShrink), and showed that it yields improved background subtraction with respect to existing robust PCA methods on a real-world dataset. Analysis of the proposed algorithm's theoretical properties is ongoing.

9. REFERENCES

[1] R. Otazo, E. J. Candès, and D. K. Sodickson, "Low-rank and sparse matrix decomposition for accelerated dynamic mri with seperation of background and dynamic components," Sept. 2013.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[3] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optm.*, vol. 21, no. 2, pp. 572–596, 2011.

[4] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[5] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[6] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.

[7] R. R. Nadakuditi, "OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage," *IEEE Trans. of IT*, vol. 60, no. 5, pp. 3002–3018, May 2013.