

Connections Between Nuclear Norm and Frobenius Norm Based Representations

Xi Peng, Canyi Lu, Zhang Yi, *Fellow IEEE*, Huajin Tang, *Member IEEE*

Abstract—A lot of works have shown that Frobenius-norm based representation (FNR) is competitive to sparse representation and nuclear-norm based representation (NNR) in numerous tasks such as subspace clustering. Despite the success of FNR in experimental studies, less theoretical analysis is provided to understand its working mechanism. In this paper, we fill this gap by building the theoretical connections between FNR and NNR. More specially, we prove that: 1) when the dictionary can provide enough representative capacity, FNR is exactly NNR even though the data set contains the Gaussian noise, Laplacian noise, or sample-specified corruption; 2) otherwise, FNR and NNR are two solutions on the column space of the dictionary.

Index Terms—Equivalence, low rank representation, least square regression, ℓ_2 -minimization, rank-minimization.

I. INTRODUCTION

MANY problems in machine learning and computer vision begin with the processing of linearly inseparable data. The goal of processing is to distinct linearly inseparable data with linear methods. To achieve this, the inputs are always projected from the original space into another space. This is so-called representation learning and three methods have been extensively investigated in the community of computer vision, *i.e.*, sparse representation (SR), low rank representation (LRR), and Frobenius-norm based representation (FNR).

During the past decade, sparse representation [1], [2] has been one of the most popular representation learning methods. It linearly reconstructs each sample using a few of basis and has shown the effectiveness in a lot of applications, *e.g.*, image repairing [3], face recognition [4], online learning control [5], dimension reduction [6], and subspace clustering [7], [8].

As another popular method, low rank representation [9]–[14] has been proposed for subspace learning and subspace clustering. Different from SR, LRR computes the representation of a data set rather than a data point by solving a nuclear norm minimization problem. Thus, LRR is also known as nuclear norm based representation (NNR). Both LRR and SR benefit from the compressive sensing theory [15] which establishes the equivalence between ℓ_0 - (rank-minimization *w.r.t.* matrix space) and ℓ_1 -norm (nuclear-norm *w.r.t.* matrix space)

This work was supported by A*STAR Industrial Robotics Programme - Distributed Sensing and Perception under SERC grant 1225100002, the National Natural Science Foundation of China under Grant 61432012 and 61673283. Corresponding author: H. Tang.

X. Peng is with Institute for Infocomm Research, A*STAR, Singapore 138632 (E-mail: pangsaai@gmail.com).

C. Lu is with Department of Electrical and Computer Engineering at National University of Singapore, Singapore 119077. (E-mail: canyilu@gmail.com).

Z. Yi and H. Tang are with College of Computer Science, Sichuan University, Chengdu, China 610065 (E-mail: {zhangyi,htang}@scu.edu.cn).

based optimization problems. More specifically, compressive sensing provides the theoretical foundation to transform the non-convex problem caused by ℓ_0 -norm into a convex problem using ℓ_1 -norm.

Recently, several works have shown that the Frobenius norm based representation (FNR) is competitive to SR and NNR in face recognition [16]–[18], subspace learning [19], [20], feature selection [21], and subspace clustering [22], [23]. The advantage of FNR is that the objective only involves a strictly convex problem and thus the trap of local minimal is avoided.

Although more and more experimental evidences have been provided to show the effectiveness of FNR, the success of FNR is counter-intuitive as FNR is generally considered to be inferior to SR and NNR. Furthermore, fewer theoretical studies have been done to explore what makes FNR effective. Motivated by two NNR works [10], [11], this paper provides a novel theoretical explanation by bridging FNR and NNR. In other words, we show that under some mild conditions, the convex problem caused by nuclear norm can be converted to a strictly convex problem based on the Frobenius norm. More specifically, we prove that: 1) when the dictionary has enough representative capacity, FNR is equivalent to the NNR [10], [11] even though the data set contains the Gaussian noise, Laplacian noise, or sample-specified corruption; 2) when the dictionary has limited representative capacity, FNR and NNR are two solutions of the column space spanned by inputs. Our theoretical results unify FNR and NNR into a framework, *i.e.*, FNR and NNR are in the form of $\mathbf{V}\mathcal{P}(\Sigma)\mathbf{V}^T$, where $\mathbf{U}\Sigma\mathbf{V}^T$ is the singular value decomposition (SVD) of a given data matrix and $\mathcal{P}(\cdot)$ denotes the shrinkage-thresholding operator. The difference between FNR and NNR lies in the different choices of the shrinkage-thresholding operator. To the best of our knowledge, this is one of the first several works to establish the connections between FNR and NNR.

II. BACKGROUND

For a given data set $\mathbf{X} \in \mathbb{R}^{m \times n'}$ (each column denotes a data point), it can be decomposed as the linear combination of $\mathbf{D} \in \mathbb{R}^{m \times n}$ by

$$\min_{\mathbf{C}} f(\mathbf{C}) \quad \text{s.t. } \mathbf{X} = \mathbf{D}\mathbf{C}, \quad (1)$$

where $f(\mathbf{C})$ denotes the constraint enforced over the representation $\mathbf{C} \in \mathbb{R}^{n \times n'}$. The main difference among most existing works is their objective functions, basically, the choice of $f(\mathbf{C})$. Different assumptions motivate different $f(\cdot)$ and this work focuses on the discussion of two popular objective functions, *i.e.*, nuclear-norm and Frobenius-norm.

By assuming \mathbf{C} is low rank and the input contains noise, Liu *et al.* [9] propose solving the following nuclear norm based minimization problem:

$$\min_{\mathbf{C}, \mathbf{E}} \underbrace{\|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_p}_{f(\mathbf{C})} \quad \text{s.t.} \quad \underbrace{\mathbf{D} = \mathbf{D}\mathbf{C} + \mathbf{E}}_{\text{Noisy Case}}, \quad (2)$$

where $\|\mathbf{C}\|_* = \sum \sigma_i(\mathbf{C})$, $\sigma_i(\mathbf{C})$ is the i th singular value of \mathbf{C} , and $\|\cdot\|_p$ could be chosen as $\ell_{2,1}$ -, ℓ_1 -, or Frobenius-norm. $\ell_{2,1}$ -norm is usually adopted to depict the sample-specific corruptions such as outliers, ℓ_1 -norm is used to characterize the Laplacian noise, and Frobenius norm is used to describe the Gaussian noise.

Although Eq.(2) can be easily solved by the Augmented Lagrangian method (ALM) [24], its computational complexity is still very high. Recently, Favaro and Vidal [10], [11] proposed a new formulation of LRR which can be calculated very fast. The proposed objective function is as follows:

$$\min_{\mathbf{C}, \mathbf{D}_0} \|\mathbf{C}\|_* + \lambda \|\mathbf{D} - \mathbf{D}_0\|_F \quad \text{s.t.} \quad \mathbf{D} = \mathbf{D}_0\mathbf{C} + \mathbf{E}, \quad (3)$$

where \mathbf{D}_0 denotes the clean dictionary and $\|\cdot\|_F$ denotes the Frobenius-norm of a given data matrix. Different from Eq.(2), Eq.(3) calculates the low rank representation using a clean dictionary \mathbf{D}_0 instead of the original data \mathbf{D} . Moreover, Eq.(3) has a closed-form solution. In this paper, we mainly investigate this formulation of NNR.

Another popular representation is based on ℓ_2 -norm or its induced matrix norm (*i.e.*, the Frobenius norm). The basic formulation of FNR is as follows:

$$\min \|\mathbf{C}\|_F \quad \text{s.t.} \quad \mathbf{X} = \mathbf{D}\mathbf{C}. \quad (4)$$

In our previous work [25], we have shown that the optimal solution to Eq.(4) is also the lowest rank solution, *i.e.*,

Theorem 1 ([25]). *Assume $\mathbf{D} \neq \mathbf{0}$ and $\mathbf{X} = \mathbf{D}\mathbf{C}$ has feasible solution(s), *i.e.*, $\mathbf{X} \in \text{span}(\mathbf{D})$. Then*

$$\mathbf{C}^* = \mathbf{D}^\dagger \mathbf{X} \quad (5)$$

is the unique minimizer to Eq.(4), where \mathbf{D}^\dagger is the pseudo-inverse of \mathbf{D} .

Considering nuclear norm based minimization problem, Liu *et al.* [9] have shown that

Theorem 2 ([9]). *Assume $\mathbf{D} \neq \mathbf{0}$ and $\mathbf{X} = \mathbf{D}\mathbf{C}$ has feasible solution(s), *i.e.*, $\mathbf{X} \in \text{span}(\mathbf{D})$. Then*

$$\mathbf{C}^* = \mathbf{D}^\dagger \mathbf{X} \quad (6)$$

is the unique minimizer to

$$\min \|\mathbf{C}\|_* \quad \text{s.t.} \quad \mathbf{X} = \mathbf{D}\mathbf{C}, \quad (7)$$

where \mathbf{D}^\dagger is the pseudo-inverse of \mathbf{D} .

Theorems 1 and 2 actually imply the equivalence between NNR and FNR when the dictionary can exactly reconstruct inputs and the data set is immune from corruptions. In this paper, we will further investigate the connections between NNR and FNR by considering more complex situations, *e.g.*, the data set is corrupted by Gaussian noise.

$\min_{\mathbf{D}_0, \mathbf{E}, \mathbf{C}} \frac{1}{2} \ \mathbf{C}\ _q + \frac{\lambda}{2} \ \mathbf{E}\ _p \quad \text{s.t.} \quad \mathcal{J}(\mathbf{X})$		
Conditions	Exact Constraint $\mathcal{J}(\mathbf{X}) : \mathbf{X} = \mathbf{D}\mathbf{C} + \mathbf{E}$	Relax Constraint $\mathcal{J}(\mathbf{X}) : \mathbf{D} \neq \mathbf{D}_0\mathbf{C} + \mathbf{E}$
Noiseless $\mathbf{E} = \mathbf{0}$	Equivalent, <i>i.e.</i>, $\mathbf{C}^* = \mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	Two solutions on the column space of \mathbf{D} are in the form of $\mathbf{C}^* = \mathbf{V}\mathcal{P}_\gamma(\mathbf{\Sigma})\mathbf{V}^T$
Noisy $\mathbf{E} \neq \mathbf{0}$		

Fig. 1. An overview of the connections between FNR ($\|\cdot\|_q = \|\cdot\|_F$) and NNR ($\|\cdot\|_q = \|\cdot\|_*$), where ℓ_p can be chosen as ℓ_1 -, $\ell_{2,1}$ -, and ℓ_2 -norm corresponding to the Laplacian noise, Gaussian noise, and outliers, respectively. With the relax constraint, the major difference between NNR and FNR is the value of γ . More details are summarized in Tables I and II.

III. CONNECTIONS BETWEEN NUCLEAR NORM AND FROBENIUS NORM BASED REPRESENTATION

For a data matrix $\mathbf{D} \in \mathbb{R}^{m \times n}$, let $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ and $\mathbf{D} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T$ be the full SVD and skinny SVD of \mathbf{D} , where $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_r$ are in descending order and r denotes the rank of \mathbf{D} . \mathbf{U}_r , \mathbf{V}_r and $\mathbf{\Sigma}_r$ consist of the top (*i.e.*, largest) r singular vectors and singular values of \mathbf{D} . Similar to [4], [7], [9]–[11], we assume $\mathbf{D} = \mathbf{D}_0 + \mathbf{E}$, where \mathbf{D}_0 denotes the clean data set and \mathbf{E} denotes the errors.

Our theoretical results will show that the optimal solutions of Frobenius-norm and nuclear-norm based objective functions are in the form of $\mathbf{C}^* = \mathbf{V}\mathcal{P}(\mathbf{\Sigma})\mathbf{V}^T$, where $\mathcal{P}(\cdot)$ denotes the shrinkage-thresholding operator. In other words, FNR and NNR are two solutions on the column space of \mathbf{D} and they are identical in some situations. This provides a unified framework to understand FNR and NNR. The analysis will be performed considering several popular cases including exact/relax constraint and non-corrupted/corrupted data. When the dictionary has enough representative capacity, the objective function can be formulated with the exact constraint. Otherwise, the objective function is with the relax constraint. Noticed that, the exact constraint is considerably mild since most of data sets can be naturally reconstructed by itself in practice. With the exact constraint, many methods [10], [11] have been proposed and shown competitive performance comparing with the relax case. Besides the situation of noise-free, we will also investigate the connections between FNR and NNR when the data set contains the Gaussian noise, the Laplacian noise, or sample-specified corruption. Fig. 1, Tables I and II summary our results. Noticed that, in another independent work [26], Pan *et al.* proposed a subspace clustering method based on Frobenius norm and reported some similar conclusions with this work. Different from this work, we mainly devote to build the theoretical connections between NNR and FNR involving different settings rather than developing new algorithm.

A. Exact Constraint and Uncorrupted Data

In the following analysis, we mainly focus on the case of self-expression because almost all works on NNR are carried out under such settings.

When the data set is uncorrupted and the dictionary has enough representative capacity, Liu *et al.* [9] have shown that:

TABLE I
 CONNECTIONS BETWEEN NUCLEAR NORM ($\|\mathbf{C}\|_* \triangleq \sum_i \sigma_i(\mathbf{C})$) AND FROBENIUS NORM ($\|\mathbf{C}\|_F^2 \triangleq \sum_i \sigma_i^2(\mathbf{C})$) BASED REPRESENTATION IN THE CASE OF *the noise-free* AND *the Gaussian noise* SITUATIONS, WHERE $\sigma_i(\mathbf{C})$ DENOTES THE i TH SINGULAR VALUE OF \mathbf{C} . $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ IS THE FULL SVD OF THE DICTIONARY \mathbf{D} AND $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots)$.

Objective Function	\mathbf{C}^*	$\mathcal{P}_k(\sigma_i)$ or $\mathcal{P}_\gamma(\sigma_i)$	k or ω_i
$\min \ \mathbf{C}\ _F$ s.t. $\mathbf{X} = \mathbf{DC}$	$\mathbf{D}^\dagger \mathbf{X}$	Nil	Nil
$\min \ \mathbf{C}\ _*$ s.t. $\mathbf{X} = \mathbf{DC}$	$\mathbf{D}^\dagger \mathbf{X}$	Nil	Nil
$\min \ \mathbf{C}\ _F$ s.t. $\mathbf{D} = \mathbf{DC}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$k = \text{rank}(\mathbf{D})$
$\min \ \mathbf{C}\ _*$ s.t. $\mathbf{D} = \mathbf{DC}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$k = \text{rank}(\mathbf{D})$
$\min \frac{1}{2} \ \mathbf{C}\ _F^2 + \frac{\lambda}{2} \ \mathbf{D} - \mathbf{D}_0\ _F^2$ s.t. $\mathbf{D}_0 = \mathbf{D}_0\mathbf{C}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$\begin{cases} k = \text{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2 \\ r = \text{rank}(\mathbf{D}_0) \end{cases}$
$\min \ \mathbf{C}\ _* + \frac{\lambda}{2} \ \mathbf{D} - \mathbf{D}_0\ _F^2$ s.t. $\mathbf{D}_0 = \mathbf{D}_0\mathbf{C}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$\begin{cases} k = \text{argmin}_r r + \frac{\lambda}{2} \sum_{i>r} \sigma_i^2 \\ r = \text{rank}(\mathbf{D}_0) \end{cases}$
$\min \frac{1}{2} \ \mathbf{C}\ _F^2 + \frac{\gamma}{2} \ \mathbf{D} - \mathbf{DC}\ _F^2$	$\mathbf{V}\mathcal{P}_\gamma(\mathbf{\Sigma})\mathbf{V}^T$	$\frac{\gamma\sigma_i^2}{1+\gamma\sigma_i^2}$	Nil
$\min \ \mathbf{C}\ _* + \frac{\gamma}{2} \ \mathbf{D} - \mathbf{DC}\ _F^2$	$\mathbf{V}\mathcal{P}_\gamma(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 - \frac{1}{\gamma\sigma_i^2} & \sigma_i > 1/\sqrt{\gamma} \\ 0 & \sigma_i \leq 1/\sqrt{\gamma} \end{cases}$	Nil
$\min \ \mathbf{C}\ _F + \frac{\lambda}{2} \ \mathbf{D} - \mathbf{D}_0\ _F^2 + \frac{\gamma}{2} \ \mathbf{D}_0 - \mathbf{D}_0\mathbf{C}\ _F^2$	$\mathbf{V}\mathcal{P}_\gamma(\mathbf{\Sigma})\mathbf{V}^T$	$\frac{\gamma\sigma_i^2}{1+\gamma\sigma_i^2}$	$\sigma_i = \omega_i + \frac{\gamma\omega_i}{\lambda(1+\gamma\omega_i^2)^2}$
$\min \ \mathbf{C}\ _* + \frac{\lambda}{2} \ \mathbf{D} - \mathbf{D}_0\ _F^2 + \frac{\gamma}{2} \ \mathbf{D}_0 - \mathbf{D}_0\mathbf{C}\ _F^2$	$\mathbf{V}\mathcal{P}_\gamma(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 - \frac{1}{\gamma\omega_i^2} & \omega_i > 1/\sqrt{\gamma} \\ 0 & \omega_i \leq 1/\sqrt{\gamma} \end{cases}$	$\sigma_i = \begin{cases} \omega_i + \frac{1}{\lambda\gamma}\omega_i^{-3} & \omega_i > 1/\sqrt{\gamma} \\ \omega_i + \frac{\gamma}{\lambda}\omega_i & \omega_i \leq 1/\sqrt{\gamma} \end{cases}$

TABLE II

CONNECTIONS BETWEEN NNR AND FNR IN THE CASE OF *the Laplacian noise* AND *the sample-specified corruption*. $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{D} - \mathbf{E}_t + \alpha_t^{-1}\mathbf{Y}_t$ IS THE FULL SVD OF $\mathbf{D} - \mathbf{E}_t + \alpha_t^{-1}\mathbf{Y}_t$, $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots)$, \mathbf{E}_t IS CALCULATED USING THE AUGMENTED LAGRANGE MULTIPLIER METHOD, AND α_t AND \mathbf{Y} ARE ALM PARAMETERS. NOTE THAT, THE LAPLACIAN NOISE AND THE SAMPLE-SPECIFIED CORRUPTION WILL LEAD TO DIFFERENT \mathbf{E}_t .

Objective Function	\mathbf{C}^*	$\mathcal{P}_k(\sigma_i)$ or $\mathcal{P}_\gamma(\sigma_i)$	k or ω_i
$\min \frac{1}{2} \ \mathbf{C}\ _F^2 + \lambda \ \mathbf{D} - \mathbf{D}_0\ _1$ s.t. $\mathbf{D}_0 = \mathbf{D}_0\mathbf{C}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$\begin{cases} k = \text{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2 \\ r = \text{rank}(\mathbf{D} - \mathbf{E}_t + \alpha_t^{-1}\mathbf{Y}_t) \end{cases}$
$\min \ \mathbf{C}\ _* + \lambda \ \mathbf{D} - \mathbf{D}_0\ _1$ s.t. $\mathbf{D}_0 = \mathbf{D}_0\mathbf{C}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$\begin{cases} k = \text{argmin}_r r + \frac{\lambda}{2} \sum_{i>r} \sigma_i^2 \\ r = \text{rank}(\mathbf{D} - \mathbf{E}_t + \alpha_t^{-1}\mathbf{Y}_t) \end{cases}$
$\min \frac{1}{2} \ \mathbf{C}\ _F^2 + \lambda \ \mathbf{D} - \mathbf{D}_0\ _{2,1}$ s.t. $\mathbf{D}_0 = \mathbf{D}_0\mathbf{C}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$\begin{cases} k = \text{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2 \\ r = \text{rank}(\mathbf{D} - \mathbf{E}_t + \alpha_t^{-1}\mathbf{Y}_t) \end{cases}$
$\min \ \mathbf{C}\ _* + \lambda \ \mathbf{D} - \mathbf{D}_0\ _{2,1}$ s.t. $\mathbf{D}_0 = \mathbf{D}_0\mathbf{C}$	$\mathbf{V}\mathcal{P}_k(\mathbf{\Sigma})\mathbf{V}^T$	$\begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases}$	$\begin{cases} k = \text{argmin}_r r + \frac{\lambda}{2} \sum_{i>r} \sigma_i^2 \\ r = \text{rank}(\mathbf{D} - \mathbf{E}_t + \alpha_t^{-1}\mathbf{Y}_t) \end{cases}$

Corollary 1 ([9]). Assume $\mathbf{D} \neq \mathbf{0}$ and $\mathbf{D} = \mathbf{DC}$ have feasible solution(s), i.e., $\mathbf{D} \in \text{span}(\mathbf{D})$. Then

$$\mathbf{C}^* = \mathbf{V}_r \mathbf{V}_r^T \quad (8)$$

is the unique minimizer to

$$\min \|\mathbf{C}\|_* \quad \text{s.t. } \mathbf{X} = \mathbf{DC}, \quad (9)$$

where $\mathbf{D} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$ is the skinny SVD of \mathbf{D} .

Considering the Frobenius norm, we can obtain the following result:

Corollary 2. Let $\mathbf{D} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$ be the skinny SVD of the data matrix $\mathbf{D} \neq \mathbf{0}$. The unique solution to

$$\min \|\mathbf{C}\|_F \quad \text{s.t. } \mathbf{D} = \mathbf{DC}. \quad (10)$$

is given by $\mathbf{C}^* = \mathbf{V}_r \mathbf{V}_r^T$, where r is the rank of \mathbf{D} and \mathbf{D} denotes a given data set without corruptions.

Proof. Let $\mathbf{D} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T$ be the skinny SVD of \mathbf{D} . The pseudo-inverse of \mathbf{D} is $\mathbf{D}^\dagger = \mathbf{V}_r \mathbf{\Sigma}_r^{-1} \mathbf{U}_r^T$. By Theorem 1, we obtain $\mathbf{C}^* = \mathbf{V}_r \mathbf{\Sigma}_r^{-1} \mathbf{U}_r^T \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T = \mathbf{V}_r \mathbf{V}_r^T$, as desired. \square

From Corollaries 1 and 2, ones can find that NNR and FNR have the same optimal solution $\mathbf{V}_r \mathbf{V}_r^T$. This solution is also

known as the shape interaction matrix [27].

B. Exact Constraint and Data Corrupted by Gaussian Noise

When the data set contains Gaussian noises (i.e., $\mathbf{E} \neq \mathbf{0}$ and \mathbf{E} is characterized by the Frobenius norm), we prove that

$$\min_{\mathbf{C}, \mathbf{D}_0, \mathbf{E}} \|\mathbf{C}\|_* + \frac{\lambda}{2} \|\mathbf{E}\|_F^2 \quad \text{s.t. } \mathbf{D} = \mathbf{D}_0 + \mathbf{E}, \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (11)$$

and

$$\min_{\mathbf{C}, \mathbf{D}_0, \mathbf{E}} \frac{1}{2} \|\mathbf{C}\|_F^2 + \frac{\lambda}{2} \|\mathbf{E}\|_F^2 \quad \text{s.t. } \mathbf{D} = \mathbf{D}_0 + \mathbf{E}, \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (12)$$

have the same minimizer in the form of $\mathbf{V}_k \mathbf{V}_k^T$, where k is a parameter. By a simple transformation, we have the following results.

Theorem 3 ([10]). *Let $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ be the SVD of the data matrix \mathbf{D} . The optimal solution to*

$$\min_{\mathbf{C}, \mathbf{D}_0} \|\mathbf{C}\|_* + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 \quad \text{s.t. } \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C}, \quad (13)$$

is given by $\mathbf{C}^* = \mathbf{V}_k \mathbf{V}_k^T$, where Σ_k , \mathbf{U}_k , and \mathbf{V}_k correspond to the top $k = \text{argmin}_r r + \frac{\lambda}{2} \sum_{i>r} \sigma_i^2$ singular values and singular vectors of \mathbf{D} , respectively.

Theorem 4. *Let $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ be the full SVD of $\mathbf{D} \in \mathbb{R}^{m \times n}$, where the diagonal entries of Σ are in descending order, \mathbf{U} and \mathbf{V} are the left and right singular vectors of \mathbf{D} , respectively. Suppose there exists a clean data set and errors, denoted by \mathbf{D}_0 and \mathbf{E} , respectively. The optimal \mathbf{C} to*

$$\min_{\mathbf{D}_0, \mathbf{C}} \frac{1}{2} \|\mathbf{C}\|_F^2 + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 \quad \text{s.t. } \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (14)$$

is given by

$$\mathbf{C}^* = \mathbf{V} \mathcal{P}_k(\Sigma) \mathbf{V} = \mathbf{V}_k \mathbf{V}_k^T, \quad (15)$$

where the operator $\mathcal{P}_k(\Sigma)$ performs hard thresholding on the diagonal entries of Σ by

$$\mathcal{P}_k(\sigma_i) = \begin{cases} 1 & i \leq k \\ 0 & i > k \end{cases} \quad (16)$$

λ is a balanced parameter, $k = \text{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2$, and σ_i denotes the i -th diagonal entry of Σ . i.e., \mathbf{V}_k consists of the first k column vectors of \mathbf{V} .

Proof. Let \mathbf{D}_0^* be the optimal solution to Eq.(14) and its skinny SVD be $\mathbf{D}_0^* = \mathbf{U}_r \Sigma_r \mathbf{V}_r^T$, where r is the rank of \mathbf{D}_0^* . Let \mathbf{U}_c and \mathbf{V}_c be the basis that orthogonal to \mathbf{U}_r and \mathbf{V}_r , respectively. Clearly, $\mathbf{I} = \mathbf{V}_r \mathbf{V}_r^T + \mathbf{V}_c \mathbf{V}_c^T$. By Corollary 1, we have $\mathbf{C}^* = \mathbf{V}_r \mathbf{V}_r^T$. Next, we will bridge \mathbf{V}_r and \mathbf{V} .

Use the method of Lagrange multipliers, we obtain

$$\mathcal{L} = \frac{1}{2} \|\mathbf{C}\|_F^2 + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 + \langle \beta, \mathbf{D}_0 - \mathbf{D}_0 \mathbf{C} \rangle, \quad (17)$$

where β is the Lagrange multiplier.

Letting $\frac{\partial \mathcal{L}}{\partial \mathbf{D}_0} = 0$, it gives that

$$\beta \mathbf{V}_c \mathbf{V}_c^T = \lambda (\mathbf{D} - \mathbf{D}_0). \quad (18)$$

Letting $\frac{\partial \mathcal{L}}{\partial \mathbf{C}} = 0$, it gives that

$$\mathbf{V}_r \mathbf{V}_r^T = \mathbf{V}_r \Sigma_r \mathbf{U}_r^T \beta. \quad (19)$$

Thus, β must be in the form of $\beta = \mathbf{U}_r \Sigma_r^{-1} \mathbf{V}_r^T + \mathbf{U}_c \mathbf{M}$ for some \mathbf{M} . Substituting this into (18), it given that

$$\mathbf{U}_c \mathbf{M} \mathbf{V}_c \mathbf{V}_c^T = \lambda (\mathbf{D} - \mathbf{D}_0). \quad (20)$$

Then, we have $\|\mathbf{D} - \mathbf{D}_0\|_F^2 = \frac{1}{\lambda^2} \|\mathbf{M} \mathbf{V}_c\|_F^2$. Since $\mathbf{V}_c^T \mathbf{V}_c = \mathbf{I}$, $\|\mathbf{D} - \mathbf{D}_0\|_F^2$ is minimized when $\mathbf{M} \mathbf{V}_c$ is a diagonal matrix and can be chosen as $\mathbf{M} \mathbf{V}_c = \Sigma_c$. Then, $\mathbf{D} - \mathbf{D}_0 = \frac{1}{\lambda} \mathbf{U}_c \Sigma_c \mathbf{V}_c^T$. Consequently, the SVD of \mathbf{D} can be rewritten as

$$\mathbf{D} = \mathbf{U} \Sigma \mathbf{V}^T = [\mathbf{U}_r \mathbf{U}_c] \begin{bmatrix} \Sigma_r & \mathbf{0} \\ \mathbf{0} & \frac{1}{\lambda} \Sigma_c \end{bmatrix} [\mathbf{V}_r \mathbf{V}_c]^T. \quad (21)$$

Thus, the minimal cost of (14) is given by

$$\begin{aligned} \mathcal{L}_{\min} &= \frac{1}{2} \|\mathbf{V}_r \mathbf{V}_r^T\|_F^2 + \frac{\lambda}{2} \left\| \frac{1}{\lambda} \Sigma_c \right\|_F^2 \\ &= \frac{1}{2} r + \frac{\lambda}{2} \sum_{i=r+1}^{\min\{m,n\}} \sigma_i^2, \end{aligned}$$

where σ_i is the i -th largest singular value of \mathbf{D} . Let k be the optimal r , then, $k = \text{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2$. \square

From Theorems 3 and 4, ones can find that the values of k are slightly different. However, such difference cannot affect the equivalence conclusion because k depends on the user-specified parameter λ . Moreover, this difference actually results from the constant term in our objective function.

C. Relaxed Constraint and Uncorrupted Data

In this section, we discuss the connections between FNR and NNR when the dictionary is uncorrupted and has limited representative capacity. The objective functions are

$$\min_{\mathbf{C}} \frac{1}{2} \|\mathbf{C}\|_F^2 + \frac{\gamma}{2} \|\mathbf{D} - \mathbf{D} \mathbf{C}\|_F^2, \quad (22)$$

and

$$\min_{\mathbf{C}} \|\mathbf{C}\|_* + \frac{\gamma}{2} \|\mathbf{D} - \mathbf{D} \mathbf{C}\|_F^2. \quad (23)$$

In a lot works such as [18], [22], (22) is minimized at $\mathbf{C}^* = (\mathbf{D}^T \mathbf{D} + \gamma \mathbf{I})^{-1} \mathbf{D}^T \mathbf{D}$. In this paper, we will give another form of the solution to (22) and the new solution is performing shrinkage operation on the right eigenvectors of \mathbf{D} , like NNR.

Theorem 5 ([10]). *Let $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ be the SVD of a given matrix \mathbf{D} . The optimal solution to*

$$\min_{\mathbf{C}} \|\mathbf{C}\|_* + \frac{\gamma}{2} \|\mathbf{D} - \mathbf{D} \mathbf{C}\|_F^2 \quad (24)$$

is

$$\mathbf{C}^* = \mathbf{V}_1 \left(\mathbf{I} - \frac{1}{\gamma} \Sigma_1^{-2} \right) \mathbf{V}_1^T, \quad (25)$$

where $\mathbf{U} = [\mathbf{U}_1 \mathbf{U}_2]$, $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$, and $\mathbf{V} = [\mathbf{V}_1 \mathbf{V}_2]$ are partitioned according to the sets $\mathbf{I}_1 = \{i : \sigma_i > 1/\sqrt{\gamma}\}$ and $\mathbf{I}_2 = \{i : \sigma_i \leq 1/\sqrt{\gamma}\}$.

Theorem 6. *Let $\mathbf{D} = \mathbf{U}\Sigma\mathbf{V}^T$ be the full SVD of $\mathbf{D} \in \mathbb{R}^{m \times n}$, where the diagonal entries of Σ are in descending order,*

\mathbf{U} and \mathbf{V} are corresponding left and right singular vectors, respectively. The optimal \mathbf{C} to

$$\min_{\mathbf{C}} \frac{1}{2} \|\mathbf{C}\|_F^2 + \frac{\gamma}{2} \|\mathbf{D} - \mathbf{D}\mathbf{C}\|_F^2, \quad (26)$$

is given by

$$\mathbf{C}^* = \mathbf{V}\mathcal{P}_\gamma(\boldsymbol{\Sigma})\mathbf{V}^T = \mathbf{V}_r (\mathbf{I} - (\mathbf{I} + \gamma\boldsymbol{\Sigma}_r^2)^{-1}) \mathbf{V}_r^T, \quad (27)$$

where γ is a balanced factor and the operator $\mathcal{P}_\gamma(\boldsymbol{\Sigma})$ performs shrinkage-thresholding on the diagonal entries of $\boldsymbol{\Sigma}$ by

$$\mathcal{P}_\gamma(\sigma_i) = \begin{cases} 1 - \frac{1}{1 + \gamma\sigma_i^2} & i \leq r \\ 0 & i > r \end{cases}, \quad (28)$$

and r is the rank of \mathbf{D} and σ_i denotes the i th diagonal entry of $\boldsymbol{\Sigma}$.

Proof. Letting \mathcal{L} denote the loss, and then we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{C}} = \mathbf{C} - \gamma \mathbf{D}^T \mathbf{D} (\mathbf{I} - \mathbf{C}). \quad (29)$$

Next, we will show that (27) is the minimizer of \mathcal{L} since $\frac{\partial \mathcal{L}}{\partial \mathbf{C}^*} = 0$. Letting $\mathbf{M} = (\mathbf{I} + \gamma\boldsymbol{\Sigma}_r^2)^{-1}$ and substituting (27) into (29), we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{C}^*} = \mathbf{V}_r (\mathbf{I} - \mathbf{M}) \mathbf{V}_r^T - \gamma \mathbf{D}^T \mathbf{D} (\mathbf{I} - \mathbf{V}_r (\mathbf{I} - \mathbf{M}) \mathbf{V}_r^T). \quad (30)$$

Let \mathbf{V}_r and \mathbf{V}_c be mutually orthogonal, then $\mathbf{I} = \mathbf{V}_r \mathbf{V}_r^T + \mathbf{V}_c \mathbf{V}_c^T$. Moreover, let the skinny SVD of \mathbf{D} be $\mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^T$, we obtain

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{C}^*} &= \mathbf{V}_r \mathbf{V}_r^T - \mathbf{V}_r \mathbf{M} \mathbf{V}_r^T - \gamma \mathbf{V}_r \boldsymbol{\Sigma}_r^2 \mathbf{V}_r^T (\mathbf{V}_c \mathbf{V}_c^T + \mathbf{V}_r \mathbf{M} \mathbf{V}_r^T) \\ &= \mathbf{V}_r \mathbf{V}_r^T - \mathbf{V}_r \mathbf{M} \mathbf{V}_r^T - \gamma \mathbf{V}_r \boldsymbol{\Sigma}_r^2 \mathbf{M} \mathbf{V}_r^T \\ &= 0 \end{aligned} \quad (31)$$

as desired. \square

D. Relax Constraint and Data Corrupted by Gaussian Noise

Suppose the data set is corrupted by \mathbf{E} and has limited representative capacity, the problems can be formulated as follows:

$$\min_{\mathbf{C}, \mathbf{D}_0} \|\mathbf{C}\|_F + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 + \frac{\gamma}{2} \|\mathbf{D}_0 - \mathbf{D}_0 \mathbf{C}\|_F^2, \quad (32)$$

and

$$\min_{\mathbf{C}, \mathbf{D}_0} \|\mathbf{C}\|_* + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 + \frac{\gamma}{2} \|\mathbf{D}_0 - \mathbf{D}_0 \mathbf{C}\|_F^2. \quad (33)$$

Theorem 7 ([10]). Let $\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the SVD of the data matrix \mathbf{D} . The optimal solution to

$$\min_{\mathbf{C}, \mathbf{D}_0} \|\mathbf{C}\|_* + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 + \frac{\gamma}{2} \|\mathbf{D}_0 - \mathbf{D}_0 \mathbf{C}\|_F^2. \quad (34)$$

is given by

$$\mathbf{C}^* = \mathbf{V}_1 (\mathbf{I} - \frac{1}{\gamma} \boldsymbol{\Omega}_1^{-2}) \mathbf{V}_1^T, \quad (35)$$

where each entry of $\boldsymbol{\Omega} = \text{diag}(\omega_1, \dots, \omega_n)$ is obtained from one entry of $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_n)$ as the solution to

$$\sigma_i = \begin{cases} \omega_i + \frac{1}{\lambda\gamma} \omega_i^{-3} & \omega_i > 1/\sqrt{\gamma} \\ \omega_i + \frac{\gamma}{\lambda} \omega_i & \omega_i \leq 1/\sqrt{\gamma} \end{cases}, \quad (36)$$

that minimizes the cost, and the matrices $\mathbf{U} = [\mathbf{U}_1 \ \mathbf{U}_2]$, $\boldsymbol{\Omega} = \text{diag}(\omega_1, \omega_2)$, and $\mathbf{V} = [\mathbf{V}_1 \ \mathbf{V}_2]$ are partitioned according to the sets $\mathbf{I}_1 = \{i : \omega_i > 1/\sqrt{\gamma}\}$ and $\mathbf{I}_2 = \{i : \omega_i \leq 1/\sqrt{\gamma}\}$.

Theorem 8. Let $\mathbf{D} = \mathbf{U}_r \boldsymbol{\Sigma}_r \mathbf{V}_r^T$ be the skinny SVD of $\mathbf{D} \in \mathbb{R}^{m \times n}$, where r denotes the rank of \mathbf{D} and the diagonal entries of $\boldsymbol{\Omega}_r$ is in descending order. The optimal solutions to

$$\min_{\mathbf{C}, \mathbf{D}_0} \|\mathbf{C}\|_F + \frac{\lambda}{2} \|\mathbf{D} - \mathbf{D}_0\|_F^2 + \frac{\gamma}{2} \|\mathbf{D}_0 - \mathbf{D}_0 \mathbf{C}\|_F^2, \quad (37)$$

are given by

$$\mathbf{D}_0^* = \mathbf{U}_r \boldsymbol{\Omega}_r \mathbf{V}_r^T, \quad (38)$$

and

$$\mathbf{C}^* = \mathbf{V}_r (\mathbf{I} - (\mathbf{I} + \gamma\boldsymbol{\Omega}_r^2)^{-1}) \mathbf{V}_r^T, \quad (39)$$

where σ_i and ω_i are the diagonal entries on $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Omega}_r$, respectively.

$$\sigma_i = \omega_i + \frac{\gamma\omega_i}{\lambda(1 + \gamma\omega_i^2)^2}. \quad (40)$$

Proof. Letting \mathcal{L} denote the loss, we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{D}_0} = -\lambda(\mathbf{D} - \mathbf{D}_0) + \gamma \mathbf{D}_0 (\mathbf{I} - \mathbf{C}^*) (\mathbf{I} - \mathbf{C}^*)^T. \quad (41)$$

Next, we will bridge \mathbf{D} and \mathbf{D}_0 . Let $\mathbf{D}_0 = \mathbf{U}_r \boldsymbol{\Omega}_r \mathbf{V}_r^T$ be the skinny SVD of \mathbf{D}_0 . From Theorem 4, we have $\mathbf{C}^* = \mathbf{V}_r (\mathbf{I} - (\mathbf{I} + \gamma\boldsymbol{\Omega}_r^2)^{-1}) \mathbf{V}_r^T$. Substituting this into (41) and letting $\frac{\partial \mathcal{L}}{\partial \mathbf{D}_0} = 0$, we obtain

$$\mathbf{D} = \mathbf{U}_r \left(\boldsymbol{\Omega}_r + \frac{\gamma}{\lambda} \boldsymbol{\Omega}_r (\mathbf{I} + \gamma\boldsymbol{\Omega}_r^2)^{-2} \right) \mathbf{V}_r^T, \quad (42)$$

which is a valid SVD of \mathbf{D} . \square

Theorems 5–8 establish the relationships between FNR and NNR in the case of the limited representative capacity. Although FNR and NNR are not identical in such settings, they can be unified into a framework, *i.e.*, both two methods obtain a solution from the column space of \mathbf{D} . The major difference between them is the adopted scaling factor. Moreover, NNR and FNR will truncates the trivial entries of coefficients in the case of uncorrupted data. With respect to corrupted case, two methods only scales the self-expressive coefficients by performing shrinkage.

E. Exact Constraint and Data Corrupted by Laplacian Noise

The above analysis are based on the noise-free or the Gaussian noise assumptions. In this section, we investigate the Laplacian noise situation with the exact constraint. More specifically, we will prove that the optimal solutions to

$$\min_{\mathbf{C}, \mathbf{D}_0, \mathbf{E}} \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_1 \quad \text{s.t. } \mathbf{D} = \mathbf{D}_0 + \mathbf{E}, \ \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (43)$$

and

$$\min_{\mathbf{C}, \mathbf{D}_0, \mathbf{E}} \frac{1}{2} \|\mathbf{C}\|_F^2 + \lambda \|\mathbf{E}\|_1 \quad \text{s.t. } \mathbf{D} = \mathbf{D}_0 + \mathbf{E}, \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (44)$$

have the same form.

As \mathbf{D}_0 is unknown and ℓ_1 -norm has no closed-form solution, we can solve Eqs.(43) and (44) using the augmented Lagrange multiplier method (ALM) [28].

Proposition 1 ([10]). *The optimal solution to Eq.(43) is given by*

$$\mathbf{C}^* = \mathbf{V}_k \mathbf{V}_k^T, \quad (45)$$

where $k = \operatorname{argmin}_r r + \frac{\lambda}{2} \sum_{i>r} \sigma_i^2$, \mathbf{V}_k consists of the first k column vectors of \mathbf{V} , and \mathbf{V} is iteratively computed via the following updated rules:

$$\mathbf{U} \Sigma \mathbf{V}^T = \mathbf{D} - \mathbf{E}_t + \alpha_t^{-1} \mathbf{Y}_t \quad (46)$$

$$\mathbf{D}_{0_{t+1}} = \mathcal{U} \mathcal{P}_k(\Sigma) \mathbf{V}^T \quad (47)$$

$$\mathbf{E}_{t+1} = \mathcal{S}_{\gamma \alpha^{-1}}(\mathbf{D} - \mathbf{D}_{0_{t+1}} + \alpha_t^{-1} \mathbf{Y}_t) \quad (48)$$

$$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \alpha_k (\mathbf{D} - \mathbf{D}_{0_{t+1}} - \mathbf{E}_{t+1}) \quad (49)$$

$$\alpha_{t+1} = \rho \alpha_t, \quad (50)$$

where $\rho > 1$ is the learning rate of ALM and \mathcal{S} is a shrinkage-thresholding operator

$$\mathcal{S}_\epsilon(x) = \begin{cases} x - \epsilon & x > \epsilon \\ x + \epsilon & x < -\epsilon \\ 0 & \text{else} \end{cases} \quad (51)$$

Proposition 2. *The optimal solution to Eq.(44) is given by*

$$\mathbf{C}^* = \mathbf{V}_k \mathbf{V}_k^T, \quad (52)$$

where \mathbf{V}_k consists of the first k column vectors of \mathbf{V} , and the updated rules are

$$\mathbf{U} \Sigma \mathbf{V}^T = \mathbf{D} - \mathbf{E}_t + \alpha_t^{-1} \mathbf{Y}_t \quad (53)$$

$$\mathbf{D}_{0_{t+1}} = \mathcal{U} \mathcal{P}_k(\Sigma) \mathbf{V}^T \quad (54)$$

$$\mathbf{E}_{t+1} = \mathcal{S}_{\lambda \alpha^{-1}}(\mathbf{D} - \mathbf{D}_{0_{t+1}} + \alpha_t^{-1} \mathbf{Y}_t) \quad (55)$$

$$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \alpha_k (\mathbf{D} - \mathbf{D}_{0_{t+1}} - \mathbf{E}_{t+1}) \quad (56)$$

$$\alpha_{t+1} = \rho \alpha_t, \quad (57)$$

Proof. Using the augmented Lagrangian formulation, Eq.(44) can be rewritten as

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{C}\|_F^2 + \lambda \|\mathbf{E}\|_1 + \frac{\alpha}{2} \|\mathbf{D} - \mathbf{D}_0 - \mathbf{E}\|_F^2 \\ & + \langle \mathbf{Y}, \mathbf{D} - \mathbf{D}_0 - \mathbf{E} \rangle \\ \text{s.t.} \quad & \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C}. \end{aligned} \quad (58)$$

By fixing others, we obtain \mathbf{D}_0^* by solving

$$\min \frac{\alpha}{2} \|\mathbf{D} - \mathbf{E} + \alpha^{-1} \mathbf{Y} - \mathbf{D}_0\|_F^2 \quad \text{s.t. } \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (59)$$

According to Theorem 4, the optimal solutions to Eq.(59) is given by $\mathbf{D}_0^* = \mathbf{U}_k \Sigma_k \mathbf{V}_k^k$ and $\mathbf{C}^* = \mathbf{V}_k \mathbf{V}_k^T$, where \mathbf{V}_k consists of the first k right singular vectors of $\mathbf{D} - \mathbf{E} + \alpha^{-1} \mathbf{Y}$,

$k = \operatorname{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2$. Therefore, the optimal solutions to Eq.(44) can be iteratively computed via Eqs.(53)–(57). \square

From Propositions 1–2, ones can find that the updated rules of NNR and FNR are identical under the framework of ALM. This would lead to the same minimizer to NNR and FNR.

F. Exact Constraint and Data Corrupted by Sample-specified Noise

Besides the Gaussian noise and the Laplacian noise, we investigate sample-specified corruptions such as outliers [9], [15], [29] by adopting the $\ell_{2,1}$ norm. The formulations are as follows:

$$\min_{\mathbf{C}, \mathbf{D}_0, \mathbf{E}} \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_{2,1} \quad \text{s.t. } \mathbf{D} = \mathbf{D}_0 + \mathbf{E}, \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (60)$$

and

$$\min_{\mathbf{C}, \mathbf{D}_0, \mathbf{E}} \frac{1}{2} \|\mathbf{C}\|_F^2 + \lambda \|\mathbf{E}\|_{2,1} \quad \text{s.t. } \mathbf{D} = \mathbf{D}_0 + \mathbf{E}, \mathbf{D}_0 = \mathbf{D}_0 \mathbf{C} \quad (61)$$

Similar to Propositions 1 and 2, it is easy to show that the optimal solutions to Eqs.(60)–(61) can be calculated via

$$\mathbf{U} \Sigma \mathbf{V}^T = \mathbf{D} - \mathbf{E}_t + \alpha_t^{-1} \mathbf{Y}_t \quad (62)$$

$$\mathbf{D}_{0_{t+1}} = \mathcal{U} \mathcal{P}_k(\Sigma) \mathbf{V}^T \quad (63)$$

$$\mathbf{E}_{t+1} = \mathcal{Q}_{\lambda \alpha^{-1}}(\mathbf{D} - \mathbf{D}_{0_{t+1}} + \alpha_t^{-1} \mathbf{Y}_t) \quad (64)$$

$$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \alpha_k (\mathbf{D} - \mathbf{D}_{0_{t+1}} - \mathbf{E}_{t+1}) \quad (65)$$

$$\alpha_{t+1} = \rho \alpha_t, \quad (66)$$

where the operator $\mathcal{Q}_\epsilon(\mathbf{X})$ is defined on the column of \mathbf{X} , i.e.,

$$\mathcal{Q}_\epsilon([\mathbf{X}]_{:,i}) = \begin{cases} \frac{\|[\mathbf{X}]_{:,i}\|_2 - \epsilon}{\|[\mathbf{X}]_{:,i}\|_2} & \|[\mathbf{X}]_{:,i}\|_2 > \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (67)$$

where $[\mathbf{X}]_{:,i}$ denotes the i th column of \mathbf{X} .

Thus, ones can find that the optimal solutions of FNR and NNR are with the same form. The only one difference between them is the value of k , i.e., $k = \operatorname{argmin}_r r + \frac{\lambda}{2} \sum_{i>r} \sigma_i^2$ for NNR and $k = \operatorname{argmin}_r r + \lambda \sum_{i>r} \sigma_i^2$ for FNR.

With respect to the relax constraint, ones can also establish the connections between FNR and NNR considering the Laplacian noise and sample-specified corruption. The analysis will be based on Theorem 4 and the form of \mathbf{C}^* is similar to the case of the exact constraint.

IV. DISCUSSIONS

In this section, we first give the computational complexity analysis for FNR in different settings and then discuss the advantages of FNR over NNR in application scenario.

The above analysis shows that FNR and NNR are with the same form of solution. Thus, we can easily conclude that their computational complexity are the same under the same setting. More specifically, 1) when the input is free to corruption or contaminated by Gaussian noise, FNR and NNR will take $O(m^2 n + n^3)$ to perform SVD on the input and then use

nk^2 to obtain the representation; 2) when the input contains Laplacian noise or sample-specified corruption, FNR and NNR will take $O(tnm^2 + tn^3)$ to iteratively obtain the SVD of the input and then use nk^2 to obtain the representation.

Our analysis explicitly gives the connections between NNR and FNR in theory. Thus, ones may hope to further understand them in the context of application scenario based on the theoretical analysis. Referring to experimental studies in existing works, we could conclude that: 1) for face recognition task, FNR would be more competitive since it could achieve comparable performance with over hundred times speedup as shown in [16]–[18], [30]; 2) when dictionary can exactly reconstruct the input, both our theoretical and experimental analysis show that FNR and NNR perform comparable in feature extraction [20], image clustering and motion segmentation [10], [11]; 3) otherwise, FNR is better than NNR for feature extraction [19], image clustering and motion segmentation [9], [22], [23].

V. CONCLUSION

In this paper, we investigated the connections between FNR and NNR in the case of the exact and the relax constraint. When the objective function is with the exact constraint, FNR is exactly NNR even though the data set contains the Gaussian noise, Laplacian noise, or sample-specified corruption. In the case of the relax constraint, FNR and NNR are two solutions on the column space of inputs. Under such a setting, the only one difference between FNR and NNR is the value of the thresholding parameter γ . Our theoretical results is complementary and a small step forward to existing compressive sensing. The major difference is that this work establishes the connections between the convex problem caused by ℓ_1 -norm and the strictly convex problem caused by ℓ_2 -norm in matrix space, while compressive sensing focuses on the equivalence between the non-convex problem caused by ℓ_0 -norm and convex problem caused by ℓ_1 -norm.

REFERENCES

- [1] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [2] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.
- [3] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE T. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.
- [4] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2009.
- [5] X. Xu, Z. Hou, C. Lian, and H. He, "Online learning control using adaptive critic designs with sparse kernel machines," *IEEE T. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 762–775, May 2013.
- [6] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, "Learning with L1-graph for image analysis," *IEEE T. Image Process.*, vol. 19, no. 4, pp. 858–866, 2010.
- [7] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, 2013.
- [8] X. Peng, H. Tang, L. Zhang, Z. Yi, and S. Xiao, "A unified framework for representation-based subspace clustering of out-of-sample and large-scale data," *IEEE T. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–14, 2015.
- [9] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, 2013.
- [10] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. of 24th IEEE Conf. Comput. Vis. and Pattern Recognit.*, Colorado Springs, CO, Jun. 2011, pp. 1801–1807.
- [11] R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *Pattern Recognit. Lett.*, vol. 43, no. 0, pp. 47 – 61, 2014.
- [12] P. Sprechmann, A. Bronstein, and G. Sapiro, "Learning efficient sparse and low rank models," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1821–1833, Sept 2015.
- [13] S. Xiao, M. Tan, and D. Xu, "Robust kernel low rank representation," *IEEE T. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–1, 2015.
- [14] S. Xiao, D. Xu, and J. Wu, "Automatic face naming by learning discriminative affinity matrices from weakly labeled images," *IEEE T. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2440–2452, Oct. 2015.
- [15] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.
- [16] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.
- [17] Q. Shi, A. Eriksson, A. Van Den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. of 24th IEEE Conf. Comput. Vis. and Pattern Recognit.*, Colorado, Springs, Jun. 2011, pp. 553–560.
- [18] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. the 13th Int. Conf. on Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 471–478.
- [19] X. Peng, Z. Yu, Z. Yi, and H. Tang, "Constructing the l2-graph for robust subspace learning and subspace clustering," *IEEE T. Cybern.*, vol. PP, no. 99, pp. 1–14, 2016.
- [20] X. Peng, J. Lu, Z. Yi, and R. Yan, "Automatic subspace learning via principal coefficients embedding," *IEEE T. Cybern.*, vol. PP, no. 99, pp. 1–14, 2016.
- [21] X. Xu, Z. Huang, L. Zuo, and H. He, "Manifold-based reinforcement learning via locally linear reconstruction," *IEEE T. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–14, 2016.
- [22] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan, "Robust and efficient subspace segmentation via least squares regression," in *Proc. of 12th Eur. Conf. Computer Vis.*, Firenze, Italy, Oct. 2012, pp. 347–360.
- [23] X. Peng, Z. Yi, and H. Tang, "Robust subspace clustering via thresholding ridge regression," in *Proc. of 29th AAAI Conf. Artif. Intell.*, Austin Texas, USA, Jan. 2015, pp. 3827–3833.
- [24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. and Trends in Machine Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
- [25] H. Zhang, Z. Yi, and X. Peng, "fLRR: fast low-rank representation using frobenius-norm," *Electron. Lett.*, vol. 50, no. 13, pp. 936–938, June 2014.
- [26] P. Ji, M. Salzmann, and H. Li, "Efficient dense subspace clustering," in *Proc. of 14th IEEE Winter Conf. Appl. of Computer Vis.*, Springs, CO, Mar. 2014, pp. 461–468.
- [27] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vis.*, vol. 29, no. 3, pp. 159–179, 1998.
- [28] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. of 24th Adv. in Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 612–620.
- [29] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$ -norms minimization," in *Proc. of 23th Adv. in Neural Inf. Process. Syst.*, Harrahs and Harveys, Lake Tahoe, Dec. 2010, pp. 1813–1821.
- [30] X. Peng, L. Zhang, Z. Yi, and K. K. Tan, "Learning locality-constrained collaborative representation for robust face recognition," *Pattern Recognition*, vol. 47, no. 9, pp. 2794–2806, 2014.