Globally Optimal Robust Matrix Completion Based on M-Estimation

Conference Paper · January 2020

DOI: 10.1109/MLSP49062.2020.9231573

Michael Muma
Technical University of Darmstadt

Abdelhak M Zoubir
Technical University of Darmstadt

GLOBALLY OPTIMAL ROBUST MATRIX COMPLETION BASED ON M-ESTIMATION

Felicia Ruppel, Michael Muma, Abdelhak M. Zoubir

Signal Processing Group
Technische Universität Darmstadt
Merckstraße 25, 64283 Darmstadt, Germany
feliciaruppel@gmail.com, {muma, zoubir}@spg.tu-darmstadt.de

ABSTRACT

Robust matrix completion allows for estimating a low-rank matrix based on a subset of its entries, even in presence of impulsive noise and outliers. We explore recent progress in the theoretical analysis of non-convex low-rank factorization problems to develop a robust approach that is based on a fast factorization method. We propose an algorithm that uses joint regression and scale estimation to compute the estimates. We prove that our algorithm converges to a global minimum with random initialization. An example function for which the guarantees hold is the pseudo-Huber function. In simulations, the proposed approach is compared to state-of the art robust and non-robust methods. In addition, its applicability to image inpainting and occlusion removal is demonstrated.

Index Terms— Matrix completion, optimality, low-rank factorization, robustness, image inpainting, occlusion removal

1. INTRODUCTION

Recently, substantial progress has been made in the analysis of non-convex problems [1–5], which is beneficial for matrix completion, where a low rank matrix is recovered from a subset Ω of its entries. This task arises in many areas, for example in recommender systems, such as the well known Netflix problem, compressive sensing, and in medical applications.

The original formulation of the matrix completion problem [6] is non-convex due to the minimization of the rank:

$$\begin{aligned} \min_{\boldsymbol{A}} \quad & \operatorname{rank}(\boldsymbol{A}) \\ \text{s.t.} \quad & \boldsymbol{A}_\Omega = \boldsymbol{M}_\Omega, \end{aligned} \tag{1}$$

where $\boldsymbol{M}_\Omega$ denotes the observed entries of $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$, the matrix for which the missing entries should be estimated. To make this problem convex, the rank can be replaced by the nuclear norm, which is the corresponding convex hull and therefore provides the same results under some conditions [6]. However, algorithms based on the nuclear norm may be computationally demanding. Therefore, an alternative problem formulation emerged, implicitly enforcing the rank constraint through factorization [7]:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \|(\boldsymbol{U}\boldsymbol{V})_\Omega - \boldsymbol{M}_\Omega\|_{\mathrm{F}}^2, \tag{2}$$

where $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{r \times n_2}$ and r is the desired rank. This problem, in turn, is non-convex due to the factorization, which yields infinite possible optimal solutions. It can, however, be solved very fast with alternating minimization [1]. In spite of its non-convexity, researchers were able to prove convergence to an optimal solution, first with certain initialization schemes [1], later even with random initialization [2].

The occurrence of impulsive noise and outliers has been reported in numerous applications [8,9]. Therefore, robust matrix completion methods have been developed that allow the observed matrix $\boldsymbol{X}$ to be contaminated by impulsive noise and outliers. For example recommender systems are affected by fraud and by impulsive noise in measurements. Robustification using M-estimation of an optimization problem based on the nuclear norm has been proposed in [10], which comes with strong guarantees, such as sharp and non-sharp oracle inequalities, at the cost of computational efficiency, because of the required Singular Value Decomposition (SVD) of an $n_1 \times n_2$ matrix at each iteration. Algorithms with an increased computational efficiency that robustify the problem stated in (2) have been proposed, by replacing the quadratic function in the Frobenius norm with a different, robust, loss function [11–13]. Of these approaches, only [11], which reformulates the robust matrix completion problem into a set of regression M-estimation problems, is guaranteed to converge to a stationary point of the non-convex optimization problem.

In this paper, we will present a novel, modified version of the aforementioned robust factorized matrix completion approaches, for which we will show that the global convergence guarantees derived for Problem (2) still hold. We can show that our robust problem formulation converges to a global minimum with any random initialization. We also show that this cannot be achieved with Huber's [11] or the l_p [13] approach. We give the conditions on the robust loss function to be used and show that the result is not strongly affected by outliers in the observations.

The paper is structured as follows: Section 2 formulates the problem while Section 3 details the proposed robust matrix completion algorithm. Convergence guarantees are derived in Section 4, with the main results stated in Theorems 2 and 4. Numerical experiments follow in Section 5 and conclusions are drawn in Section 6.

Notation: Bold uppercase and lowercase letters are used to denote matrices and column vectors, respectively, while scalars are written in regular font. An element of a matrix $\boldsymbol{X}$ is either written as x_{ij} or $[\boldsymbol{X}]_{ij}$, the jth column of $\boldsymbol{X}$ is denoted as $\boldsymbol{x}_j$ and its ith row as $\boldsymbol{x}_{\underline{i}}^\top$. A bilinear form of the Hessian is used, which is defined as $\left[\nabla^2 f(\boldsymbol{X})\right](\boldsymbol{A},\boldsymbol{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\boldsymbol{X})}{\partial x_{ij} \partial x_{kl}} a_{ij} b_{kl}$.

2. PROBLEM FORMULATION

2.1. Robust matrix norm

Let us define an outlier-robust matrix "norm" as follows, which we will refer to as ρ-norm :

The work of M. Muma is supported by the 'Athene Young Investigator Programme' of Technische Universität Darmstadt.

978-1-7281-6662-9/20/$31.00 ©2020 IEEE

$$\|\boldsymbol{X}\|_{\rho} = \sqrt[q]{\sigma^q \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \rho\left(\frac{x_{ij}}{\sigma}\right)}, \tag{3}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, $\sigma > 0$ denotes the scale parameter, $q \in \mathbb{R}$ and $q \geq 1$, and $\rho : \mathbb{R} \to \mathbb{R}_0^+$ is a convex loss function that fulfills

L1. $\rho(0) = 0$ **L2.** $\rho(x) > 0$ for any $x \neq 0$.

The parameter q is chosen so that it is compatible with $\rho(x)$. For example, to be efficient at a Gaussian model, many robust loss functions behave like a quadratic function for clean data, so that $q = 2$ is a good choice. Note that the ρ-norm is a generalization of the ℓ_p-norm, which is defined as $\|\boldsymbol{X}\|_{\ell_p} = \sqrt[p]{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} |x_{ij}|^p}$ with $1 \leq p \leq 2$ [14]. The Frobenius norm $\|\cdot\|_{\mathrm{F}}$, in turn, is a special case of the ℓ_p-norm with $p = 2$.

Whether or not the ρ-norm is a proper norm depends on the choice of loss function and of the parameter q. The following properties are fulfilled with any loss function as defined above, and $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$:

N1. $\|\boldsymbol{X}\|_{\rho} \geq 0$ **N2.** $\|\boldsymbol{X}\|_{\rho} = 0 \Rightarrow \boldsymbol{X} = 0$

The properties that only hold for certain loss functions and $\alpha \in \mathbb{R}$, $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$ are:

N3. $\|\boldsymbol{X} + \boldsymbol{Y}\|_{\rho} \leq \|\boldsymbol{X}\|_{\rho} + \|\boldsymbol{Y}\|_{\rho}$

N4. $\|\alpha\boldsymbol{X}\|_{\rho} = |\alpha| \, \|\boldsymbol{X}\|_{\rho}$

For our purposes, properties N1 and N2 are sufficient. It is interesting to note that violating property N4 is actually beneficial to achieve robustness to outliers while still performing as well as possible on clean data, since this can be achieved by downweighting large entries of $\boldsymbol{X}$.

2.2. Optimization Problem

The aim of this paper is to estimate a low rank matrix $\boldsymbol{M}$ from a partly observed matrix $\boldsymbol{X}$ which is assumed to be generated as

$$\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{N} + \boldsymbol{S}, \tag{4}$$

where $\boldsymbol{N}$ contains additive Gaussian noise and $\boldsymbol{S}$ is a sparse outlier matrix. Alternatively, the observation matrix could be modeled as

$$\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{N}_{\mathrm{imp}}, \tag{5}$$

where $\boldsymbol{N}_{\mathrm{imp}}$ represents impulsive noise, i.e. noise that follows a heavy-tailed distribution.

The set Ω contains all indices (i, j) of entries that were observed. Let us introduce the notation

$$[\boldsymbol{X}_{\Omega}]_{ij} = \begin{cases} x_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega. \end{cases} \tag{6}$$

The factorization problem to be optimized is

$$\min_{\boldsymbol{U},\boldsymbol{V}} \|(\boldsymbol{U}\boldsymbol{V})_{\Omega} - \boldsymbol{X}_{\Omega}\|_{\rho}, \tag{7}$$

where $\boldsymbol{X}_{\Omega} \in \mathbb{R}^{n_1 \times n_2}$ is the observation matrix, $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{r \times n_2}$ are factors that ensure the estimation result to be of rank r. Note that, in this minimization context, the choice of parameter q in the ρ-norm does not affect the optimal solution, therefore, we set $q = 1$.

In Section 4, global optimality guarantees are derived for this problem, under the condition that a robust loss functions $\rho(x)$ with certain properties (see Theorem 4) is used. An example function for which the guarantees hold is the pseudo-Huber function [15]

$$\rho(x) = \frac{1}{2}(|x| + (1 + x)^{-1} - 1). \tag{8}$$

Alternating minimization is used to solve Equation (7). This means that $\boldsymbol{V}$ is updated while $\boldsymbol{U}$ is fixed and vice-versa.

$$\begin{aligned} \boldsymbol{V}^{(k+1)} &= \arg\min_{\boldsymbol{V}} \left\|(\boldsymbol{U}^{(k)}\boldsymbol{V})_{\Omega} - \boldsymbol{X}_{\Omega}\right\|_{\rho} \\ \boldsymbol{U}^{(k+1)} &= \arg\min_{\boldsymbol{U}} \left\|(\boldsymbol{U}\boldsymbol{V}^{(k+1)})_{\Omega} - \boldsymbol{X}_{\Omega}\right\|_{\rho} \end{aligned} \tag{9}$$

Each updating equation is convex since $\rho(x)$ is convex, so that alternating minimization finds a stationary point of the original problem in Equation (7) [11]. We can show that this point is equal to the global minimum for certain loss functions. This will be discussed in detail in Section 4.

Due to the nature of matrix multiplication and because the ρ-norm with parameter $q = 1$ is additive, Problem (9) can be written in vector-wise form. Let

$$\Omega_j := \{k \mid (k, j) \in \Omega\}$$

denote the (row) indices of available observations in the jth column of $\boldsymbol{X}_{\Omega}$. $\Omega_{\underline{i}}$ is defined analogously for available observations in the ith row: $\Omega_{\underline{i}} := \{k \mid (i, k) \in \Omega\}$.

For a vector $\boldsymbol{y}$, let

$$\boldsymbol{y}_{\Omega_j} = (y_k)_{k \in \Omega_j} \qquad \boldsymbol{y}_{\Omega_{\underline{i}}} = (y_k)_{k \in \Omega_{\underline{i}}}. \tag{10}$$

Note that $(\boldsymbol{x}_j)_{\Omega_j}$ contains the same observed values as the jth column of $\boldsymbol{X}_{\Omega}$, but the missing cells are removed instead of filled with zero. The same holds for $\left(\boldsymbol{x}_{\underline{i}}^{\top}\right)_{\Omega_{\underline{i}}}$ compared to the ith row of $\boldsymbol{X}_{\Omega}$.

From Equation (3), it is clear that the ρ-norm of a vector is defined as $\|\boldsymbol{x}\|_{\rho} = \sqrt[q]{\sigma^q \sum_{i=1}^{n} \rho\left(\frac{x_i}{\sigma}\right)}$. We formulate the alternating optimization problem

$$\begin{aligned} \boldsymbol{v}_j^{(k+1)} &= \arg\min_{\boldsymbol{v}_j} \left\|\left(\boldsymbol{U}^{(k)}\boldsymbol{v}_j\right)_{\Omega_j} - (\boldsymbol{x}_j)_{\Omega_j}\right\|_{\rho} \\ \boldsymbol{u}_{\underline{i}}^{(k+1)} &= \arg\min_{\boldsymbol{u}_{\underline{i}}} \left\|\left(\boldsymbol{u}_{\underline{i}}^{\top}\boldsymbol{V}^{(k+1)}\right)_{\Omega_{\underline{i}}} - \left(\boldsymbol{x}_{\underline{i}}^{\top}\right)_{\Omega_{\underline{i}}}\right\|_{\rho} \\ &= \arg\min_{\boldsymbol{u}_{\underline{i}}} \left\|\left(\left(\boldsymbol{V}^{(k+1)}\right)^{\top}\boldsymbol{u}_{\underline{i}}\right)_{\Omega_{\underline{i}}} - (\boldsymbol{x}_{\underline{i}})_{\Omega_{\underline{i}}}\right\|_{\rho}, \end{aligned} \tag{11}$$

where $j = 1, 2, \ldots, n_2$ and $i = 1, 2, \ldots, n_1$.

This optimization problem is equivalent to Problem (9), except for one difference: While there exists only one scale parameter σ for all residuals per update equation in Problem (9), we have an individual scale parameter as part of the ρ-norm for each column of $\boldsymbol{V}$ during its update in Problem (11) (and for each row during the update of $\boldsymbol{U}$). This is negligible if the residuals are iid., a consistent estimator is used for the estimation of the scale parameter and the sample size is sufficient, since the column-wise scale estimates will be very close to the matrix-wide scale estimate. If the estimates differ slightly, this should still not be problematic. However, if a large amount of outliers is concentrated in one column or row they will dominate the scale estimate.

3. PROPOSED ALGORITHM

In this section, an algorithm to solve Problem (11) is presented. Due to its vectorized form, the individual sub-problems are smaller than in matrix form and the computation can be parallelized. An overview of the robust matrix completion algorithm is detailed in Algorithm 1.

The updates of the columns of $\boldsymbol{V}$ and rows of $\boldsymbol{U}$ are robust regression problems of the form

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{y}\|_{\rho}, \tag{12}$$

Algorithm 1: Globally optimal robust matrix completion

Input: set Ω containing the indices of available data, data matrix $\boldsymbol{X}_\Omega \in \mathbb{R}^{n_1 \times n_2}$, rank r

Output: factorization $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{r \times n_2}$

Initialize : $\boldsymbol{U}^{(0)} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V}^{(0)} \in \mathbb{R}^{r \times n_2}$, $\epsilon_1, \epsilon_2 > 0$, $N_{\text{maxiter}} \in \mathbb{N}$, loss function $\rho(x)$

for $k = 0, 1, \ldots, N_{\text{maxiter}} - 1$ **do**

  for $j = 0, 1, \ldots, n_2 - 1$ **do**

    $\boldsymbol{v}_j^{(k+1)} = \arg\min_{\boldsymbol{v}_j} \left\| (\boldsymbol{U}^{(k)} \boldsymbol{v}_j)_{\Omega_j} - (\boldsymbol{x}_j)_{\Omega_j} \right\|_\rho$

    /* update each column of $\boldsymbol{V}$ */

  for $i = 0, 1, \ldots, n_1 - 1$ **do**

    $\boldsymbol{u}_{\underline{i}}^{(k+1)} =$

    $\arg\min_{\boldsymbol{u}_{\underline{i}}} \left\| \left(\left(\boldsymbol{V}^{(k+1)} \right)^\top \boldsymbol{u}_{\underline{i}} \right)_{\Omega_{\underline{i}}} - (\boldsymbol{x}_{\underline{i}})_{\Omega_{\underline{i}}} \right\|_\rho$

    /* update each row of $\boldsymbol{U}$ */

  if $\frac{\|\boldsymbol{V}^{(k+1)} - \boldsymbol{V}^{(k)}\|}{\|\boldsymbol{V}^{(k)}\|} < \epsilon_1$ **and** $\frac{\|\boldsymbol{U}^{(k+1)} - \boldsymbol{U}^{(k)}\|}{\|\boldsymbol{U}^{(k)}\|} < \epsilon_2$ **then**

    break

return $\left(\boldsymbol{V}^{(k+1)}, \boldsymbol{U}^{(k+1)} \right)$

where $\boldsymbol{y} \in \mathbb{R}^{n_1 \times 1}$ and $\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}$ are the dependent and independent variables, respectively, and $\boldsymbol{\beta} \in \mathbb{R}^{n_2 \times 1}$ is the parameter vector to be estimated.

To update $\boldsymbol{v}_j$, we will set $\boldsymbol{y} = (\boldsymbol{x}_j)_{\Omega_j}$, $\boldsymbol{\beta} = \boldsymbol{v}_j$ and $\boldsymbol{Z} = (u_{kl})_{k \in \Omega_j}$. To update $\boldsymbol{u}_{\underline{i}}$, set $\boldsymbol{y} = (\boldsymbol{x}_{\underline{i}})_{\Omega_{\underline{i}}}$, $\boldsymbol{\beta} = \boldsymbol{u}_{\underline{i}}$ and $\boldsymbol{Z} = \left((v_{kl})_{l \in \Omega_{\underline{i}}} \right)^\top$. There exist many algorithms that can solve Problem (12), such as gradient descent or iteratively reweighted least squares (IRWLS). Both of them require an auxiliary scale estimate in every iteration, which can be computationally demanding.

For these reasons, we recommend to use the computationally more efficient joint regression and scale estimation, which was originally proposed by Huber and Ronchetti [16] and during which Huber's criterion

$$\min_{\beta, \sigma > 0} n_1 \alpha \sigma + \sum_{i=1}^{n_1} \rho \left(\frac{y_i - \boldsymbol{z}_{\underline{i}}^\top \boldsymbol{\beta}}{\sigma} \right) \sigma \tag{13}$$

is minimized, where α is a constant to obtain Fisher consistency depending on the loss function ρ . For convex ρ, (13) is jointly convex in (β, σ), see [9, pp. 57-60] for further discussion and a minimization-maximization algorithm to solve this joint regression and scale problem for any loss function that obeys

<u>**L3.**</u> $\rho(x)/|x|$ is a concave function for $x \geq 0$

<u>**L4.**</u> ρ is twice differentiable and $0 \leq \frac{\partial^2 \rho(x)}{\partial x^2} \leq 1$.

Our implementation of robust matrix completion with joint regression and scale estimation is available online. [1]

4. CONVERGENCE GUARANTEES

The optimization problem in Equation (7) is naturally non-convex due to the factorization $\boldsymbol{UV}$, even if the ρ-norm of choice is convex. This can easily be shown: Let $\boldsymbol{U}^*$, $\boldsymbol{V}^*$ denote an optimal solution. Then, $\tilde{\boldsymbol{U}} = \boldsymbol{U}^* \boldsymbol{G}_1$ and $\tilde{\boldsymbol{V}} = \boldsymbol{G}_2 \boldsymbol{V}^*$ is also an optimal solution for any $\boldsymbol{G}_1, \boldsymbol{G}_2 \in \mathbb{R}^{r \times r}$, for which $\boldsymbol{G}_1 \boldsymbol{G}_2 = \boldsymbol{I}_r$ holds.

[1] github.com/FeliciaRu/robust-matrix-completion

As mentioned in the introduction Zhu *et al.* [2] were able to guarantee convergence of the non-robust Problem (2) to a global optimum with a random initialization of $\boldsymbol{U}$ and $\boldsymbol{V}$. The optimization problem they studied is very general, therefore we will show that their results are also applicable to robust matrix completion. They introduce the optimization problem

$$\min_{\boldsymbol{A}} f(\boldsymbol{A}) \qquad \text{s.t. } \operatorname{rank}(\boldsymbol{A}) \leq r \tag{14}$$

and then consider the corresponding factorization

$$\min_{\boldsymbol{U}, \boldsymbol{V}} f(\boldsymbol{UV}), \tag{15}$$

where $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{r \times n_2}$ and $\boldsymbol{A} = \boldsymbol{UV}$. Note that we adapt their notation to be consistent with ours. They introduce the regularized problem

$$\min_{\boldsymbol{U}, \boldsymbol{V}} f_R(\boldsymbol{U}, \boldsymbol{V}) = f(\boldsymbol{UV}) + g(\boldsymbol{U}, \boldsymbol{V}) \tag{16}$$

with

$$g(\boldsymbol{U}, \boldsymbol{V}) = \frac{\mu}{4} \|\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}}^2, \tag{17}$$

where $\mu > 0$. Besides this, the $(2r, 4r)$-restricted strong convexity and smoothness condition is introduced that $f(\boldsymbol{A})$ is supposed to fulfill:

$$\alpha \|\boldsymbol{G}\|_{\mathrm{F}}^2 \leq [\nabla^2 f(\boldsymbol{A})](\boldsymbol{G}, \boldsymbol{G}) \leq \beta \|\boldsymbol{G}\|_{\mathrm{F}}^2 \tag{18}$$

for any matrices $\boldsymbol{A}, \boldsymbol{G} \in \mathbb{R}^{n_1 \times n_2}$ with $\operatorname{rank}(\boldsymbol{A}) \leq 2r$ and $\operatorname{rank}(\boldsymbol{G}) \leq 4r$ and positive α, β. Now, we can summarize their main result in Theorem 1.

Theorem 1 (Global optimality [2])**.** *With a function $f(\boldsymbol{A})$ that satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition, any local minimum $\boldsymbol{U}, \boldsymbol{V}$ of the regularized optimization problem in Equation (16) is also a global minimum and the strict saddle property is fulfilled. Since any local minimum of Problem (16) obeys $g(\boldsymbol{U}, \boldsymbol{V}) = 0$, it is also a global minimum $\boldsymbol{A}^* = \boldsymbol{UV}$ of the original problem in Equation (14).*

This guarantee is very strong because any algorithm that is able to find a local minimum, such as gradient descent, will be able to find the global minimum of Problem (14) with random initialization, even though it is a non-convex optimization problem. In their paper, the authors did however not explore its applicability in the context of robust statistics, therefore we will show under which circumstances Theorem 1 is applicable to robust matrix completion and robust low-rank approximation.

4.1. Robust low-rank approximation

The term "robust low-rank approximation" refers to a scenario in which the matrix $\boldsymbol{X}$ is fully observed, i.e. $\Omega = [n_1] \times [n_2]$. We will therefore examine the optimization problem

$$\min_{\boldsymbol{U}, \boldsymbol{V}} \|\boldsymbol{UV} - \boldsymbol{X}\|_\rho . \tag{19}$$

in this context, omitting Ω, resulting in

$$f(\boldsymbol{A}) = \|\boldsymbol{A} - \boldsymbol{X}\|_\rho = \sigma \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \rho \left(\frac{a_{ij} - x_{ij}}{\sigma} \right) \tag{20}$$

and

$$\begin{aligned} \left[\nabla^2 f(\boldsymbol{A}) \right] (\boldsymbol{G}, \boldsymbol{G}) &= \sum_{i,j} \frac{\partial f(\boldsymbol{A})}{\partial a_{ij} \partial a_{ij}} g_{ij}^2 \\ &= \frac{1}{\sigma} \sum_{i,j} \rho'' \left(\frac{a_{ij} - x_{ij}}{\sigma} \right) g_{ij}^2, \end{aligned} \tag{21}$$

where $\boldsymbol{A}, \boldsymbol{G} \in \mathbb{R}^{n_1 \times n_2}$, $\operatorname{rank}(\boldsymbol{A}) \leq 2r$ and $\operatorname{rank}(\boldsymbol{G}) \leq 4r$. Here, we used the fact that $\frac{\partial f(\boldsymbol{A})}{\partial a_{ij} \partial a_{kl}} = 0$ for $(i,j) \neq (k,l)$.

For the pseudo-Huber function $\rho(x) = \frac{1}{2}(|x| + (1+x)^{-1} - 1)$ in particular, we obtain

$$[\nabla^2 f(\boldsymbol{A})](\boldsymbol{G}, \boldsymbol{G}) = \frac{1}{\sigma} \sum_{i,j} \left(1 + \left|\frac{a_{ij} - x_{ij}}{\sigma}\right|\right)^{-3} g_{ij}^2, \tag{22}$$

with $\rho''(x) = (1 + |x|)^{-3}$. The property $[\nabla^2 f(\boldsymbol{A})](\boldsymbol{G}, \boldsymbol{G}) \leq \beta \|\boldsymbol{G}\|_{\mathrm{F}}^2$ holds for $\beta = \frac{1}{\sigma}$. There exists a parameter $\alpha > 0$ so that

$$\alpha \|\boldsymbol{G}\|_{\mathrm{F}}^2 \leq [\nabla^2 f(\boldsymbol{A})](\boldsymbol{G}, \boldsymbol{G}), \tag{23}$$

only if we can define an upper bound for $\left|\frac{a_{ij} - x_{ij}}{\sigma}\right|$, i.e.

$$\left|\frac{a_{ij} - x_{ij}}{\sigma}\right| \leq \gamma \tag{24}$$

for any $(i,j) \in [n_1] \times [n_2]$. The triangle inequality leads to

$$\left|\frac{a_{ij} - x_{ij}}{\sigma}\right| \leq \frac{1}{\sigma}(|a_{ij}| + |x_{ij}|). \tag{25}$$

With $\frac{\|\boldsymbol{A}\|_\infty}{\sigma} + \frac{\|\boldsymbol{X}\|_\infty}{\sigma} \leq \gamma$, it follows that

$$\frac{1}{\sigma}(|a_{ij}| + |x_{ij}|) \leq \gamma \tag{26}$$

holds for any $(i,j) \in [n_1] \times [n_2]$ and Equation (23) is satisfied with $\alpha = \frac{1}{\sigma(1+\gamma)^3}$. This requirement can be ensured at the beginning of the algorithm with a bounded initialization of $\boldsymbol{A}$ and by clipping extremely large values in $\boldsymbol{X}$, if necessary. During the algorithm, the residual in Equation (24) should generally become smaller as we minimize a loss function that depends on the residual.

The above result for $\rho(x) = \frac{1}{2}(|x| + (1+|x|)^{-1} - 1)$ can easily be generalized to many other convex robust loss functions, which have in common that their second derivative can be expressed as a function of the absolute value of the residual, i.e.

$$\rho''\left(\frac{a_{ij} - x_{ij}}{\sigma}\right) = \zeta\left(\left|\frac{a_{ij} - x_{ij}}{\sigma}\right|\right) \tag{27}$$

with $\zeta : \mathbb{R}^+ \to \mathbb{R}, t \mapsto \zeta(t)$. The result is formalized in Theorem 2 that states our first main result.

Theorem 2 (Robust low-rank approximation)**.** *The function $f(\boldsymbol{A}) = \|\boldsymbol{A} - \boldsymbol{X}\|_\rho$ fulfills the $(2r, 4r)$-restricted strong convexity and smoothness condition with parameters $\alpha = \frac{\zeta(\gamma)}{\sigma}$ and $\beta = \frac{\zeta(0)}{\sigma}$ if a loss function $\rho(x)$ that can be expressed as in Equation (27) with $0 < \zeta(t) \leq 1$ and $\zeta'(t) \leq 0$ is chosen and if $\frac{\|\boldsymbol{A}\|_\infty}{\sigma} + \frac{\|\boldsymbol{X}\|_\infty}{\sigma} \leq \gamma$ holds.*
In this case, Theorem 1 is applicable to $f(\boldsymbol{A})$ and the corresponding factorized robust low-rank approximation problem.

Now, consider the Huber loss function $\rho_{\text{Hub}}(x)$ as described in [11]. Since $\rho''_{\text{Hub}}(x) = 0$ for $|x| > c$, it is not possible to guarantee that $f(\boldsymbol{A}) = \|\boldsymbol{A} - \boldsymbol{X}\|_{\rho_{\text{Hub}}}$ fulfills the $(2r, 4r)$-restricted strong convexity and smoothness condition without prior assumptions on $\boldsymbol{G}$ (if any entry of $\frac{\boldsymbol{A}-\boldsymbol{X}}{\sigma}$ is allowed to be larger than c), while for other loss functions, as mentioned beforehand, assumptions on $\boldsymbol{X}$ and $\boldsymbol{A}$ are sufficient. To understand why, let

$$[\boldsymbol{G}]_{ij} = \begin{cases} 1 & \text{if } (i,j) = (1,1) \\ 0 & \text{else,} \end{cases} \tag{28}$$

so that $\boldsymbol{G}$ is a matrix of rank 1. Now, if $|a_{11} - x_{11}| > c$, we have

$$\left[\nabla^2 f(\boldsymbol{A})\right](\boldsymbol{G}, \boldsymbol{G}) = 0. \tag{29}$$

In Theorem 1, a regularization term $g(\boldsymbol{U}, \boldsymbol{V}) = \frac{\mu}{4}\|\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}}^2$ with $\mu > 0$ is introduced. We found that in our particular case, this regularization term does not necessarily need to be enforced during the algorithm: Any solution (i.e. local minimum) $\boldsymbol{U}, \boldsymbol{V}$ the algorithm finds can be transformed to $\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}$ which obey $g\left(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}\right) = 0$ and are a local minimum of both the regularized and not regularized optimization problem. With Theorem 1, $\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}$ is therefore a global minimum of the original optimization problem. To transform a solution $\boldsymbol{U}, \boldsymbol{V}$ after the proposed algorithm converged, compact singular value decomposition (compact SVD) is performed on $\hat{\boldsymbol{M}} = \boldsymbol{U}\boldsymbol{V}$ so that

$$\hat{\boldsymbol{M}} = \boldsymbol{Q}_U \boldsymbol{\Sigma} \boldsymbol{Q}_V, \tag{30}$$

where $\boldsymbol{Q}_U \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{Q}_V \in \mathbb{R}^{r \times n_2}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix. By definition of the compact SVD,

$$\boldsymbol{Q}_U^\top \boldsymbol{Q}_U = \boldsymbol{Q}_V \boldsymbol{Q}_V^\top = \boldsymbol{I}_r \tag{31}$$

holds. Now, let $\tilde{\boldsymbol{U}} = \boldsymbol{Q}_U \boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\tilde{\boldsymbol{V}} = \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Q}_V$, so that

$$g\left(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}\right) = \frac{\mu}{4}\left\|\tilde{\boldsymbol{U}}^\top \tilde{\boldsymbol{U}} - \tilde{\boldsymbol{V}}\tilde{\boldsymbol{V}}^\top\right\|_{\mathrm{F}}^2 = 0. \tag{32}$$

4.2. Partially observed robust matrix completion

With partially observed robust matrix completion, we have $\Omega \subset [n_1] \times [n_2]$. To transfer the the results from Theorem 2 to this case, let us introduce the restricted isometry property (RIP) for a measurement operator $\mathcal{M}(\boldsymbol{A})$, which is widely used.

Definition 1 (Restricted isometry property [17])**.** *A measurement operator $\mathcal{M} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ obeys the r-RIP if*

$$(1 - \delta_r)\|\boldsymbol{A}\|_{\mathrm{F}}^2 \leq \|\mathcal{M}(\boldsymbol{A})\|_{\mathrm{F}}^2 \leq (1 + \delta_r)\|\boldsymbol{A}\|_{\mathrm{F}}^2 \tag{33}$$

holds for any matrix $\boldsymbol{A}$ with $\operatorname{rank}(\boldsymbol{A}) \leq r$, where δ_r is a constant.

In the case of matrix completion, the relevant measurement operator is defined as

$$\mathcal{M}(\boldsymbol{A}) = \mathcal{P}_\Omega(\boldsymbol{A}), \tag{34}$$

with $\mathcal{P}_\Omega(\boldsymbol{A}) = \boldsymbol{A}_\Omega$. In [18], Jain *et al.* derive conditions under which $\mathcal{P}_\Omega(\boldsymbol{A})$ fulfills the r-RIP, which are cited in Theorem 3. They introduce the following incoherence condition: A matrix $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$ with SVD $\boldsymbol{A} = \boldsymbol{Q}_U \boldsymbol{\Sigma} \boldsymbol{Q}_V$ is μ-incoherent if

$$\max_{i,j}\left|[\boldsymbol{Q}_U]_{i,j}\right| \leq \frac{\sqrt{\mu}}{\sqrt{n_1}}, \qquad \max_{i,j}\left|[\boldsymbol{Q}_V]_{i,j}\right| \leq \frac{\sqrt{\mu}}{\sqrt{n_2}}. \tag{35}$$

Theorem 3 (Refined RIP [18])**.** *With $0 < \delta_k < 1$, $\mu \geq 1$, $n = \max(n_1, n_2) \geq 3$ and $m = \min(n_1, n_2)$, there exists a constant C independent of m, n and k so that*

$$(1 - \delta_k)p\|\boldsymbol{A}\|_{\mathrm{F}}^2 \leq \|\mathcal{P}_\Omega(\boldsymbol{A})\|_{\mathrm{F}}^2 \leq (1 + \delta_k)p\|\boldsymbol{A}\|_{\mathrm{F}}^2 \tag{36}$$

holds with probability at least $1 - \exp(-n \log n)$ for all μ-incoherent matrices $\boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}$ with $\operatorname{rank}(\boldsymbol{A}) \leq k$ if $\Omega \subseteq [n_1] \times [n_2]$ is drawn according to Bernoulli sampling with inclusion probability $p \geq C\mu^2 k^2 \log n / \left(\delta_k^2 m\right)$.

In the following, assume $\mathcal{P}_\Omega(\boldsymbol{A})$ obeys the refined RIP with $k = 4r$. Let $f_\Omega(\boldsymbol{A}) = \|\mathcal{P}_\Omega(\boldsymbol{A} - \boldsymbol{X})\|_\rho$ with gradient

$$[\nabla f_\Omega(\boldsymbol{A})]_{ij} = \frac{\partial f_\Omega(\boldsymbol{A})}{\partial a_{ij}} = \begin{cases} \frac{\partial f(\boldsymbol{A})}{\partial a_{ij}} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega, \end{cases} \tag{37}$$

where $f(\boldsymbol{A})$ is as defined in Equation (20). Its bilinear Hessian form in terms of $(\boldsymbol{G}, \boldsymbol{G})$ is

$$\left[\nabla^2 f_\Omega(\boldsymbol{A})\right](\boldsymbol{G}, \boldsymbol{G}) = \frac{1}{\sigma} \sum_{(i,j)\in\Omega} \rho''\left(\frac{a_{ij} - x_{ij}}{\sigma}\right) g_{ij}^2. \tag{38}$$

If $f(\boldsymbol{A})$ fulfills the $(2r, 4r)$-restricted strong convexity and smoothness condition in Equation (18), then $f_\Omega(\boldsymbol{A})$ fulfills

$$\alpha\|\mathcal{P}_\Omega(\boldsymbol{G})\|_\mathrm{F}^2 \le [\nabla^2 f_\Omega(\boldsymbol{A})](\boldsymbol{G}, \boldsymbol{G}) \le \beta\|\mathcal{P}_\Omega(\boldsymbol{G})\|_\mathrm{F}^2 \tag{39}$$

for any matrices $\boldsymbol{A}, \boldsymbol{G} \in \mathbb{R}^{n_1\times n_2}$ with $\operatorname{rank}(\boldsymbol{A}) \le 2r$ and $\operatorname{rank}(\boldsymbol{G}) \le 4r$. From the refined RIP property with $k = 4r$ we can conclude that

$$\alpha(1 - \delta_{4r})p\|\boldsymbol{G}\|_\mathrm{F}^2 \le \alpha\|\mathcal{P}_\Omega(\boldsymbol{G})\|_\mathrm{F}^2 \tag{40}$$

and

$$\beta\|\mathcal{P}_\Omega(\boldsymbol{G})\|_\mathrm{F}^2 \le \beta(1 + \delta_{4r})p\|\boldsymbol{G}\|_\mathrm{F}^2 \tag{41}$$

hold for all μ-incoherent matrices $\boldsymbol{G}$ with $\operatorname{rank}(\boldsymbol{G}) \le 4r$. Therefore,

$$\alpha(1 - \delta_{4r})p\|\boldsymbol{G}\|_\mathrm{F}^2 \le [\nabla^2 f_\Omega(\boldsymbol{A})](\boldsymbol{G}, \boldsymbol{G}) \le \beta(1 + \delta_{4r})p\|\boldsymbol{G}\|_\mathrm{F}^2, \tag{42}$$

i.e. $f_\Omega(\boldsymbol{A})$ fulfills the $(2r, 4r)$-restricted strong convexity and smoothness condition with parameters $\alpha' = \alpha(1 - \delta_{4r})p$ and $\beta' = \beta(1 + \delta_{4r})p$. With these results, we are able to formulate our second main result that is stated in Theorem 4.

Theorem 4 (Robust matrix completion)**.** *The function*

$$f_\Omega(\boldsymbol{A}) = \|\mathcal{P}_\Omega(\boldsymbol{A} - \boldsymbol{X})\|_\rho, \tag{43}$$

where $\boldsymbol{A} \in \mathbb{R}^{n_1\times n_2}$, fulfills the $(2r, 4r)$-restricted strong convexity and smoothness condition with parameters $\alpha' = \frac{\zeta(\gamma)}{\sigma}(1-\delta_{4r})p$ and $\beta' = \frac{\zeta(0)}{\sigma}(1+\delta_{4r})p$ if the conditions from Theorem 2 hold and if Ω is drawn according to Bernoulli sampling with inclusion probability $p \ge 16C\mu^2r^2\log n/\left(\delta_{4r}^2 m\right)$ so that the refined RIP holds, where $n = \max(n_1, n_2) \ge 3$, $m = \min(n_1, n_2)$.
In this case, Theorem 1 is applicable to $f_\Omega(\boldsymbol{A})$ and the corresponding factorized robust matrix completion problem.

Some authors propose a regularization term to enforce the incoherence of the iterates $\boldsymbol{U}, \boldsymbol{V}$, such as [19]. However, since we found that the iterates remain incoherent, even without the regularizer, and to allow for faster computation time, we decided to omit the regularization term in the proposed robust matrix completion algorithm.

5. EXPERIMENTAL RESULTS

This section reports Monte Carlo simulations to benchmark our robust matrix completion algorithm using the pseudo-Huber loss function (denoted as *psHub*) against the robust approach proposed in [11], denoted as *Hub*, the SVT algorithm [20] and the non-robust factorization approach using the ℓ_2-loss function [7].

The low-rank matrices $\boldsymbol{M}$ with $r = 10$ are generated from factors $\boldsymbol{U} \in \mathbb{R}^{150\times 10}$ and $\boldsymbol{V} \in \mathbb{R}^{10\times 300}$ drawn from a standard normal distribution. Then, noise is added according to a Gaussian mixture model, which consists of a dense component plus a sparse outlying component with a much larger variance, resulting in the observation matrix $\boldsymbol{X}$. The results are an average over 100 Monte Carlo runs. In addition to the different algorithm results, we plotted $\mathrm{RMSE}(\boldsymbol{M} - \boldsymbol{X})$ for comparison, denoted *naive*.

Unless mentioned otherwise, the outlier contamination proportion was set to $c = 0.1$, the inclusion probability of an entry was $p = 45\%$ during Bernoulli sampling, and the SNR was set to 9 dB. One by one, each of these parameters was varied, resulting in the plots that are shown in Figures 2, 3 and 4. In the latter, c was varied, while not changing the variances of the dense noise and sparse outlier components in the GMM model, which were selected so that $\mathrm{SNR} = 9$ dB holds for $c = 0.1$. While the non-robust approaches can hardly achieve better results than the naive approach, both robust algorithms significantly improve the SNR while completing the matrix. They can achieve good results with a sampling probability as low as $p = 0.2$. It is also observed that our approach performs similarly well compared to the *Hub*-algorithm, while slightly outperforming it in challenging scenarios such as a very high outlier contamination or a very low sampling probability.

Fig. 1. *Left to right:* Original image [21], contaminated image with salt and pepper noise and missing entries, ℓ_2-estimate, pseudo-Huber estimate. *Top to bottom:* The salt and pepper noise contamination percentage c is varied, $c \in \{0, 0.3, 0.5\}$

In addition to these simulations, we showcase a comparison between our approach with the non-robust factorization approach using the ℓ_2-loss function in the context of image inpainting and occlusion removal. An image of a tree in front of a building facade is used. Since the tree does not fit into the low-rank nature in the image, we expect it to be removed. Text is written over the image, which is interpreted as missing entries. After this, salt and pepper noise was added. From top to bottom in Figure 1, the salt and pepper contamination was varied from $c \in \{0, 0.3, 0.5\}$. Without contamination, both algorithms perform similarly well, while the robust approach clearly outperforms the other when outliers are added.

6. CONCLUSION

A robust low-rank matrix completion approach was proposed. It was supported by convergence guarantees to the global optimum, while offering the computational advantages that factorization-based matrix completion entails. The experiments demonstrate satisfactory performance compared to state-of-the-art robust approaches. The method can be readily used since its implementation will be published.

7. REFERENCES

[1] M. Hardt, “Understanding alternating minimization for matrix completion,” in *2014 IEEE 55th Annu. Symp. on Found. of Comput. Sci.*, pp. 651–660.

[2] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, “Global optimality in low-rank matrix optimization,” *IEEE Trans. on Signal Process.*, vol. 66, no. 13, pp. 3614–3628, 2018.

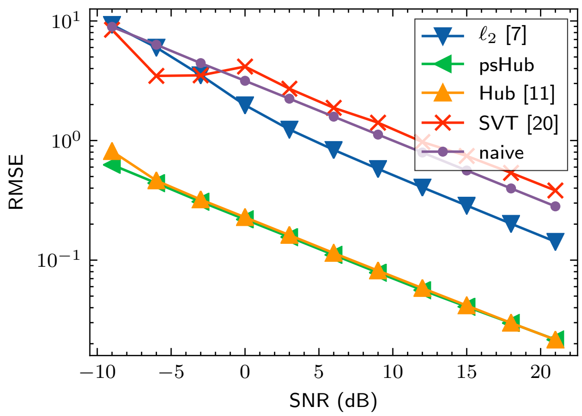

Fig. 2. The SNR is varied in Gaussian mixture model noise.

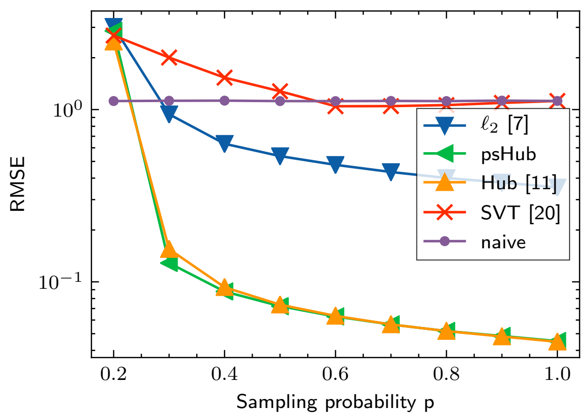

Fig. 3. The inclusion probability p during Bernoulli sampling is varied, while the SNR is constant.

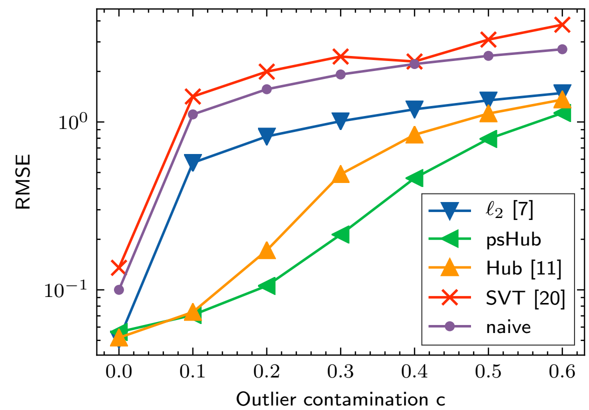

Fig. 4. RMSE over the percentage of contaminated matrix cells c. At $c = 0.1$, SNR $= 9$ dB was enforced. Then, c was varied while keeping the other parameters constant.

[3] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. of the 34th Int. Conf. on Mach. Learn. (ICML)-Volume 70*, pp. 1233–1242, 2017.

[4] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via nonconvex factorization," *IEEE Trans. on Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[5] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Adv. in Neural Inf. Process. Syst.*, pp. 3873–3881, 2016.

[6] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. of Comput. Math.*, vol. 9, no. 6, p. 717, 2009.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[8] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, 2012.

[9] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust statistics for signal processing*. Cambridge University Press, 2018.

[10] A. Elsener and S. van de Geer, "Robust low-rank matrix estimation," *The Ann. of Statist.*, vol. 46, no. 6B, pp. 3481–3509, 2018.

[11] M. Muma, W. Zeng, and A. M. Zoubir, "Robust M-estimation Based Matrix Completion," in *2019 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 5476–5480, 2019.

[12] L. Zhao, P. Babu, and D. P. Palomar, "Efficient algorithms on robust low-rank matrix completion against outliers," *IEEE Trans. on Signal Process.*, vol. 64, no. 18, pp. 4767–4780, 2016.

[13] W.-J. Zeng and H. C. So, "Outlier-robust matrix completion via ℓ_p-minimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1125–1140, 2017.

[14] W. Zeng, *Robust Low-Rank Approximation of Matrices in lp-Space*. PhD thesis, Technische Universität, Darmstadt.

[15] K. Lange, "Convergence of EM image algorithms with Gibbs reconstruction smoothing," *IEEE Trans. on Med. Imag.*, vol. 9, no. 4, pp. 439–446, 1990.

[16] P. J. Huber and E. M. Ronchetti, "Robust Statistics, John Wiley & Sons, Inc," 2009.

[17] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[18] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Adv. in Neural Inf. Process. Syst.*, pp. 937–945, 2010.

[19] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Adv. in Neural Inf. Process. Syst.*, pp. 2973–2981, 2016.

[20] J.-F. Cai, E. J. Candès, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM J. on Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[21] D. Dai, H. Riemenschneider, G. Schmitt, and L. van Gool, "Example-based facade texture synthesis," in *Proc. of the IEEE Int. Conf. on Comput. Vis.*, pp. 1065–1072, 2013.