Robust Low Rank Tensor Recovery with Regularized Redescending M-Estimator

Yuning Yang, Yunlong Feng, and Johan A. K. Suykens, *Fellow, IEEE*

***Abstract*—This paper addresses the robust low rank tensor recovery problems. Tensor recovery aims at reconstructing a low rank tensor from some linear measurements, which finds applications in image processing, pattern recognition, multitask learning and so on. In real-world applications, data might be contaminated by sparse gross errors. However, existing approaches may not be very robust to outliers. To resolve this problem, this paper proposes approaches based on the regularized redescending M-estimators, which have been introduced in robust statistics. The robustness of the proposed approaches is achieved by the regularized redescending M-estimators. However, the nonconvexity also leads to computational difficulty. To handle this problem, we develop algorithms based on proximal and linearized block coordinate descent methods. By explicitly deriving the Lipschitz constant of the gradient of the data-fitting risk, the descent property of the algorithms is present. Moreover, we verify that the objective functions of the proposed approaches satisfy the Kurdyka-Łojasiewicz property, which establishes the global convergence of the algorithms. Numerical experiments on synthetic data as well as real data verify that our approaches are robust in the presence of outliers and still effective in the absence of outliers.**

***Index Terms*—robust tensor recovery, redescending M-estimator, nonconvexity, block coordinate descent, global convergence**

I. Introduction

Tensors, appearing as the higher order generalization of vectors and matrices, make it possible to represent data that have intrinsically many dimensions, and give a better understanding of the relationship behind the information from a higher order perspective. Similar to matrices, some higher order data may naturally have sparse structure in some sense, e.g., the low rank property, or include dominant information in a few factors. To explore the underlying structure of higher order tensors, two decomposition strategies, namely the Tucker decomposition and the CP decomposition [1]–[4] are commonly used.

Although the above two decomposition strategies can decompose or approximate the underlying structure of tensors, the fact that they are built under the framework of unsupervised learning limits their use when some information is known a priori. On the other hand, the strategies may be sensitive to outliers. Recently, within the supervised learning and convex optimization frameworks, [5], [6] independently proposed approaches for recovering a low rank tensor from some linear measurements, among which recovering a low rank tensor from incomplete observations is of particular interest. These new approaches take some prior knowledge into consideration, and moreover, they introduce the tensor nuclear norm to model the low rank property of tensors. Applications including image processing [7], pattern recognition [8], multitask learning [9], spectral data [10] and intelligent transportation systems [11] demonstrate the effectiveness of the new approaches. Apart from these applications, other researches are carried out on understanding the theoretical behaviors and improving computational efficiency of the new approaches and their variants, see e.g., [12]–[17].

In the presence of noise, the low rank tensor recovery approaches incorporate the learning framework of "loss + regularization", where the least squares loss is commonly adopted, see e.g., [6], [14]. It is known that the least squares loss possesses the interpretation of the maximum likelihood estimation when the distribution of the noise is Gaussian. However, in real-world applications such as image processing [18], the distribution of the noise is usually unknown, and data might be grossly corrupted by outliers, whereas in this case, the least squares loss based approaches behave poorly due to their non-robustness.

To address tensor recovery problems with outliers, robust approaches were proposed in [19]–[21]. Specifically, [19], [20] were focused on the tensor PCA problems, whereas [21] took tensor completion as well as tensor PCA into account, the robustness of which benefits from a robust loss function — the least absolute deviations loss (LAD) $|\cdot|$. The LAD loss shares the ability of being relatively insensitive to large deviations in outlying observations, and its convexity also leads to the global solution to the above approaches. However, the LAD loss still penalizes outliers linearly, which may not be very robust [22]. [23] considered robust multilinear PCA by using the Welsch loss, which is in fact a redescending M-estimator. Proposed in robust statistics [22], redescending M-estimators have been widely used in robust regression problems. The redescending property makes the redescending M-estimators be particularly resistant to extreme outliers and may outperform bounded M-estimators [24], such as those induced by the LAD loss.

Inspired by the success of applications of redescending M-estimators, in this paper, the Welsch loss and the Cauchy loss based redescending M-estimators are employed to seek robustness when recovering a low rank tensor from observations contaminated by non-Gaussian noise or outliers. Our considerations of using these two losses are also motivated by the theoretical investigations presented in [25]–[27] and empirical successes reported in [23], [28]–[30]. To explore the underlying low rank tensor data, the mixture tensor model [13], [14], which is the sum of some factor tensors and each of

Y. Yang, Y. Feng and J. A. K. Suykens are with the Department of Electrical Engineering ESAT-STADIUS, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (email: yuning.yang, yunlong.feng, johan.suykens@esat.kuleuven.be).

which is low rank in only a specific mode, is adopted. Then, the nuclear norm regularization as well as the rank constrained strategies are incorporated to ensure the low rank property of each factor.

By exploiting the separable structure of the proposed models, proximal and linearized block coordinate descent methods are developed. To show the descent property of the algorithms, the Lipschitz constant of the gradient of the data-fitting risk is derived. Moreover, we verify the analytic property of the data-fitting risk and the semi-algebraic property of the regularization schemes, which, within the framework of [31], show the global convergence of the developed algorithms.

The remainder of this paper is organized as follows. In Section II, we introduce preliminaries on basic tensor algebra and tensor recovery problems. We propose our robust tensor recovery approaches based on the regularized redescending M-estimators in Section III and give some discussions. We then provide a proximal and linearized block coordinate descent method in Section IV, and verify their convergence properties in Section V. The performance of our methods are tested in Section VI on synthetic data as well as real data. We end this paper in Section VII with concluding remarks.

II. Preliminaries

A. Basic tensor algebra

A tensor is a multiway array. An N-th order tensor is denoted as $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_N}$. We use $\mathbb{T} := \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_N}$ to denote an N-th order tensor space throughout this paper. Below we list some concepts of tensors. The readers can be referred to the survey paper [32] for more details.

Tensor-matrix unfolding. A fiber of a tensor $\mathcal{X}$ is a column vector obtained by fixing every index of $\mathcal{X}$ but one. The mode-d unfolding, or matricization, denoted by $\mathbf{X}_{(d)}$, is a matrix of size $n_d \times \prod_{i \neq d}^{N} n_i$, whose columns are the mode-d fibers of $\mathcal{X}$ in the lexicographical order.

Inner product and tensor-matrix multiplication. For $\mathcal{A}, \mathcal{B} \in \mathbb{T}$, their inner product is given by

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1=1}^{n_1} \cdots \sum_{i_N=1}^{n_N} \mathcal{A}_{i_1 \cdots i_N} \mathcal{B}_{i_1 \cdots i_N}.$$

The Frobenius norm of $\mathcal{A}$ is defined by $\|\mathcal{A}\|_F = \langle \mathcal{A}, \mathcal{A} \rangle^{1/2}$. The mode-$d$ tensor-matrix multiplication of a tensor $\mathcal{X} \in \mathbb{T}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J_d \times n_d}$, denoted by $\mathcal{Y} = \mathcal{X} \times_d \mathbf{U}$, is of size $n_1 \times \cdots \times n_{d-1} \times J_d \times n_{d+1} \times \cdots \times n_N$, and can be expressed in terms of unfolding as $\mathbf{Y}_{(d)} = \mathbf{U}\mathbf{X}_{(d)}$.

Tensor rank and decompositions. There are mainly two types of tensor rank, namely the CP-rank and the mode-d rank. The CP-rank is defined as the minimum positive integer R such that for a tensor $\mathcal{X} \in \mathbb{T}$, it can be factorized as a sum of R rank-one tensors. This factorization is called CP decomposition. Finding such a decomposition exactly is NP-hard for higher order tensors [32]. The mode-d rank of a tensor $\mathcal{X} \in \mathbb{T}$, also known as the Tucker rank, is defined as the rank of the mode-d unfolding matrix $\mathbf{X}_{(d)}$. $\mathcal{X}$ is said to be rank-$(R_1, \ldots, R_N)$ if the mode-d rank of $\mathcal{X}$ is $(R_1, \ldots, R_N)$. Any rank-$(R_1, \ldots, R_N)$ tensor $\mathcal{X}$ can be decomposed as $\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}^1 \times_2 \cdots \times_N \mathbf{U}^N$, where $\mathcal{G} \in \mathbb{R}^{R_1 \times \cdots \times R_N}$ is called the core tensor and $\mathbf{U}^i \in \mathbb{R}^{n_i \times R_i}$ are factor matrices. This decomposition is called the Tucker decomposition.

Tuple of tensors. Given a set of tensors $\mathcal{X}_1, \ldots, \mathcal{X}_N \in \mathbb{T}$, we denote the tuple of tensors $\boldsymbol{\mathcal{X}}$ as

$$\boldsymbol{\mathcal{X}} := \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N\}.$$

Similar to the notation for vector spaces, we denote the space of tuples of tensors as $\mathbb{T}^N$. For ease of notation, the mode-l unfolding of the k-th tensor is denoted as $\mathbf{X}_{k,(l)}$. For convenience, we also define the summation operator $\sum : \mathbb{T}^N \to \mathbb{T}$ such that $\sum(\boldsymbol{\mathcal{X}}) = \sum_{i=1}^{N} \mathcal{X}_i$.

B. Tensor recovery problems

The goal of tensor recovery is to find a tensor $\mathcal{X} \in \mathbb{T}$ satisfying $\boldsymbol{\mathcal{A}}(\mathcal{X}) = \mathbf{b}$, where $\mathbf{b} \in \mathbb{R}^p$ is a vector, and $\boldsymbol{\mathcal{A}} : \mathbb{T} \to \mathbb{R}^p$ with $p \leq \prod_{i=1}^{N} n_i$ is a linear map defined as

$$\boldsymbol{\mathcal{A}}(\cdot) = \left[\langle \mathcal{A}_1, \cdot \rangle, \langle \mathcal{A}_2, \cdot \rangle, \ldots, \langle \mathcal{A}_p, \cdot \rangle\right]^T,$$

with $\mathcal{A}_i \in \mathbb{T}, 1 \leq i \leq p$. Along with the low rank requirement on $\mathcal{X}$, the low rank tensor recovery problem reads as follows

$$\text{finding a low rank tensor } \mathcal{X} \quad \text{s.t.} \quad \boldsymbol{\mathcal{A}}(\mathcal{X}) = \mathbf{b}.$$

Since the mode-d rank is easier to compute than the CP-rank, the problem boils down to [6]

$$\min_{\mathcal{X} \in \mathbb{T}} \sum_{i=1}^{N} \operatorname{rank}(\mathbf{X}_{(i)}) \quad \text{s.t.} \quad \boldsymbol{\mathcal{A}}(\mathcal{X}) = \mathbf{b}. \tag{1}$$

A special and important case of (1) is the tensor completion problem, which aims at recovering a low rank tensor from partial observations, i.e.,

$$\min_{\mathcal{X} \in \mathbb{T}} \sum_{i=1}^{N} \operatorname{rank}(\mathbf{X}_{(i)}) \text{ s.t. } \mathcal{X}_{i_1 \cdots i_N} = \mathcal{B}_{i_1 \cdots i_N}, (i_1 \cdots i_N) \in \Omega,$$

where $\mathcal{B} \in \mathbb{T}$ represents the observed tensor, and Ω denotes the set of multi-indices that correspond to the observed entries. To get a tractable problem, the objective functions in the above two problems are replaced by the sum of nuclear norms, which is also called the tensor nuclear norm, see e.g, [5], [6], [12].

So far we have reviewed the noiseless tensor recovery problems. Suppose now the observation is given by $\mathbf{b} = \boldsymbol{\mathcal{A}}(\mathcal{X}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^p$ represents noise or outliers. A common approach to handle the above problem takes the form $\min_{\mathcal{X} \in \mathbb{T}} f(\mathcal{X}) + g(\mathcal{X})$, where f refers to the empirical risk while g is the regularization term, such as the tensor nuclear norm regularization. In the presence of Gaussian noise, the least squares loss based approaches have been frequently adopted, see e.g, [6], [14], [33]. With the presence of gross errors or outliers, robust techniques should be adopted. Based on the LAD loss, [19]–[21] proposed approaches for the tensor PCA and tensor completion problems. Particularly, [21] introduces the LAD loss into different tensor models and provides a variety of approaches. An important advantage of the LAD loss based approaches is the convexity. However, as we have pointed out, the LAD loss based approaches may not be very robust, as they penalize outliers linearly.

III. Tensor Recovery with Regularized Redescending M-Estimator and Discussions

A. The proposed approaches

Similar to [13], [14], we assume that the tensor to be recovered can be approximated by a sum of N factor tensors of the same size, each being low rank in only one mode. Specifically, we let $\mathcal{X} := \sum(\boldsymbol{\mathcal{X}}) = \sum_{j=1}^{N} \mathcal{X}_j$, with $\mathbf{X}_{j,(j)}$ being a low rank matrix, $1 \leq j \leq N$, where $\mathbf{X}_{j,(j)}$ represents the mode-j unfolding matrix of the j-th factor tensor. This is called the mixture model [13], [14]. Based on this representation, our data-fitting risk takes the following form

$$J_\sigma(\sum(\boldsymbol{\mathcal{X}})) := \sum_{i=1}^{p} \rho_\sigma\left(\langle \mathcal{A}_i, \sum(\boldsymbol{\mathcal{X}})\rangle - \mathbf{b}_i\right),$$

where $\sigma > 0$ is a scale parameter, and ρ_σ is a robust loss whose influence function has the redescending property. Specifically, let $\psi_\sigma(t) = \rho'_\sigma(t)$ be the empirical influence function. According to [22], ψ_σ of a redescending M-estimator satisfies the property that $\lim_{|t|\to+\infty} \psi_\sigma(t) = 0$. In this paper, we employ the following two losses:

- Cauchy loss: $\rho_\sigma(t) = 0.5\sigma^2 \log\left(1 + t^2/\sigma^2\right)$;
- Welsch loss: $\rho_\sigma(t) = 0.5\sigma^2\left(1 - \exp(-t^2/\sigma^2)\right)$.

For tensor completion problems, J_σ can be simplified as

$$J_\sigma(\sum(\boldsymbol{\mathcal{X}})) = \sum_{(i_1\cdots i_N)\in\Omega} \rho_\sigma\left(\sum(\boldsymbol{\mathcal{X}})_{i_1\cdots i_N} - \mathcal{B}_{i_1\cdots i_N}\right).$$

Concerning the regularization terms and considering the low rank property of $\mathbf{X}_{j,(j)}$, motivating by [13], [14], we can use the nuclear norm to penalize these unfolding matrices. Thus, the approach based on the regularized redescending M-estimator is formulated as

$$\min_{\boldsymbol{\mathcal{X}}\in\mathbb{T}^N} J_\sigma\left(\sum\nolimits_{j=1}^{N} \mathcal{X}_j\right) + \lambda \sum_{j=1}^{N} \|\mathbf{X}_{j,(j)}\|_*, \tag{2}$$

where $\lambda > 0$ is a regularization parameter.

We also propose the following rank constrained redescending M-estimator based approach, when the rank information of the tensor to be recovered is known in advance

$$\min_{\boldsymbol{\mathcal{X}}\in\mathbb{T}^N} J_\sigma\left(\sum\nolimits_{j=1}^{N} \mathcal{X}_j\right) \text{ s.t. } \operatorname{rank}(\mathbf{X}_{j,(j)}) \leq R_j,\ 1 \leq j \leq N. \tag{3}$$

It is known that (3) can be reformulated as an unconstrained problem by introducing the indicator functions. Let C be a nonempty and closed set. The indicator function of C, denoted as δ_C, is defined by

$$\delta_C(x) = \begin{cases} 0, & \text{if } x \in C, \\ +\infty, & \text{otherwise.} \end{cases}$$

For each $j = 1, \ldots, N$, denote

$$M_{R_j} = \{\mathbf{X} \in \mathbb{R}^{n_j \times \prod_{k\neq j}^{N} n_k} \,|\, \operatorname{rank}(\mathbf{X}) \leq R_j\}.$$

Then M_{R_j} is a nonempty and closed set for each j. By using the indicator functions, (3) can be reformulated as

$$\min_{\boldsymbol{\mathcal{X}}\in\mathbb{T}^N} J_\sigma\left(\sum\nolimits_{j=1}^{N} \mathcal{X}_j\right) + \sum_{j=1}^{N} \delta_{M_{R_j}}(\mathbf{X}_{j,(j)}). \tag{4}$$

For convenience, in the rest of this paper, if necessary, we unify the regularizations $\lambda\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$ by the notation $g(\cdot)$. We also denote

$$\Phi(\boldsymbol{\mathcal{X}}) = \Phi(\mathcal{X}_1, \ldots, \mathcal{X}_N) := J_\sigma\left(\sum\nolimits_{j=1}^{N} \mathcal{X}_j\right) + \sum\nolimits_{j=1}^{N} g(\mathbf{X}_{j,(j)}), \tag{5}$$

and then the models (2) and (4) can be unified as

$$\min_{\boldsymbol{\mathcal{X}}\in\mathbb{T}^N} \Phi(\boldsymbol{\mathcal{X}}). \tag{6}$$

Remark 1: The matrix Schatten-p norm regularization has been used for matrix recovery [34]. The advantage of the Schatten-p norm is that can approximate the rank function when $p \to 0$. Thus it will be also interesting to replace the nuclear norm in (2) by the Schatten-p norm.

B. Tuning the regularization parameters

In the proposed approaches, the regularization parameters are important in controlling the low rank property of the resulting tensors. We first consider the λ value of (2). In the literature [21], [35], the value $\sqrt{n_{\max}}$ is suggested for the regularization parameter of the LAD loss based matrix and tensor completion approaches, where $n_{\max}$ denotes the largest dimension of a tensor. Motivated by this, in our study, we empirically find that $\lambda = \alpha n_{\min}/\sqrt{n_{\max}}$ is a proper choice, where $\alpha > 0$ is to be tuned, usually in the interval $(0, 0.5)$, and $n_{\min}$ denotes the smallest dimension of a tensor.

The ranks $R_1, \ldots, R_N$ in (3) or (4) can be regarded as regularization parameters. If we know the ranks of the tensor to be recovered, then we can set the values of $R_1, \ldots, R_N$ accordingly. Otherwise, the values of the ranks have also to be tuned. A possible and simple heuristic is to apply some algorithms to solve (2) at first with finite steps, and then use the generated rank information in (3), possibly rescaled by a factor smaller than 1. In the PCA setting, i.e., $\boldsymbol{\mathcal{A}}$ is an identity, a strategy similar to the Q-based method for multilinear PCA [36] can be applied. In our setting, we define $Q^{(d)} = \sum_{i_d=1}^{R_d} \lambda_{i_d}^{(d)} / \sum_{i_d=1}^{n_d} \lambda_{i_d}^{(d)}$, where $\lambda_{i_d}^{(d)}$ is the i-th singular value of the mode-d unfolding matrix $B_{(d)}$, with $\lambda_{1_d}^{(d)} \geq \lambda_{2_d}^{(d)} \geq \cdots \geq \lambda_{n_d}^{(d)}$, $1 \leq i_d \leq n_d$, $1 \leq d \leq N$. One then chooses some value $\epsilon \in (0, 1)$ and selects R_d such that $Q^{(d)} \approx \epsilon$ for $d = 1, \ldots, N$. Thus by choosing a suitable ϵ, one ignores the less important eigenvectors of the unfoldings of $\mathcal{B}$. Other heuristics are also available in the literature, see e.g., [37].

C. Discussions on the employed loss functions

This part presents some discussions on the employed loss functions. To make the discussions convenient, we first distinguish the two employed losses by $\rho_\sigma^{\text{Welsh}}$ and $\rho_\sigma^{\text{Cauchy}}$, respectively. We have the following observations:

1) From their corresponding influence functions
   - $\psi_\sigma^{\text{Welsh}}(t) = \exp(-t^2/\sigma^2)t$,
   - $\psi_\sigma^{\text{Cauchy}}(t) = t/(1 + t^2/\sigma^2)$,

   one can see that they both satisfy $\lim_{|t|\to+\infty} \psi_\sigma(t) = 0$. As a result, associated with the robust loss functions

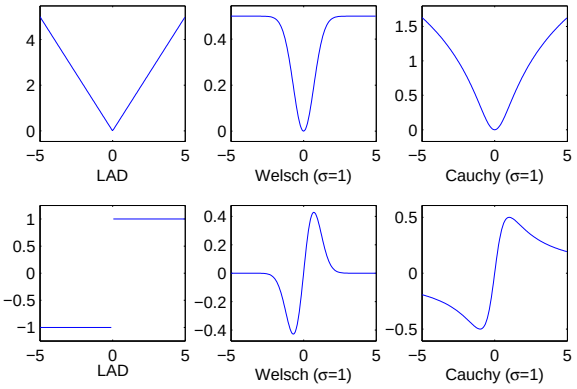


Fig. 1: Plots of different losses (top) and their influence functions (bottom).

$\rho_\sigma^{\text{Welsch}}(t)$ and $\rho_\sigma^{\text{Cauchy}}(t)$, the proposed approaches (2) and (4) are essentially regularized redescending M-estimators. In fact, the value $\psi(t)/t$ can be regarded as a weight of t. One observes that for redescending M-estimators, as t increases, $\psi(t)/t$ decreases sharply, which gives a small weight to the value t. On the other hand, the influence function of the LAD loss is only bounded instead of redescending, and the LAD loss penalizes the large deviation linearly. This leads to the result that (2) and (4) are more robust than the LAD based approaches. To give an intuitive impression, we plot the three losses and their influence functions in Fig. 1, from which we can easily see the redescending property of ψ^{Welsh} and ψ^{Cauchy}.

2) Another nice property of the influence functions of Welsh loss and Cauchy loss is the Lipschitz continuity, i.e., for any $t_1, t_2 \in \mathbb{R}$, $|\psi_\sigma(t_1) - \psi_\sigma(t_2)| \leq |t_1 - t_2|$. This property is very important and serves as a basic for the convergence of the algorithms presented later.
3) As $t \to 0$, $\rho_\sigma^{\text{Welsh}}$ and $\rho_\sigma^{\text{Cauchy}}$ approximate the least squares loss, which can be seen from their Taylor series. Given a fixed $\sigma > 0$, both of the two losses possess the form $\rho_\sigma(t) = t^2/2 + o((t/\sigma)^2)$. Therefore, $\rho_\sigma(t) \approx t^2/2$ provided $t/\sigma \to 0$. This also reminds us that a large σ can lead to the closeness between ρ_σ and the least squares loss. Such a property gives more flexibility to the two employed losses than LAD. We also mention that a comparative study on the parameter σ of the Welsch loss can be found in [27].
4) Although the employed losses enjoy nice robustness property, their nonconvexity seems to be a barrier in practical use. Nevertheless, from experiments we find that approaches based on the employed losses can yield satisfying results, as will be shown later.
5) Other redescending type losses such as the Tukey's biweight loss can also be considered.

IV. A Proximal and Linearized Block Coordinate Descent Method

This part concerns with algorithms for (2) and (4). Classical robust M-estimators can be solved by the iteratively reweighted least squares (IRLS) algorithms. However, due to the existence of regularization terms, IRLS cannot be directly applied to solve (2) and (4), and the special structure of Φ requires us to seek other proper methods. In the following, a block coordinate descent type method is developed. First we discuss some properties of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$.

A. Properties of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$

First we show that $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$ are *proper* and *lower semi-continuous*, which are essential in deriving the Moreau envelopes and are important in the convergence analysis of the developed algorithms. The definitions are given as follows.

Definition 1 (c.f. [38]): A function f is called *lower semi-continuous* at $\mathbf{x}_0 \in \mathbb{R}^n$, if for every $\epsilon > 0$, there is a neighborhood $\mathbb{U}(\mathbf{x}_0, \epsilon)$ of $\mathbf{x}_0$ such that for all $\mathbf{x} \in \mathbb{U}(\mathbf{x}_0, \epsilon)$, there holds $f(\mathbf{x}) \geq f(\mathbf{x}_0) - \epsilon$. Equivalently, $\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0)$. f is called *proper* if $f(\mathbf{x}) < \infty$ for at least one $\mathbf{x} \in \mathbb{R}^n$, and $f(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in \mathbb{R}^n$.

Proposition 1: Both of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$ are proper and lower semi-continuous functions.

Proof: From their definitions, we can verify that they are proper functions. On the other hand, since $\|\cdot\|_*$ is a norm, it is continuous and hence lower semi-continuous. $\delta_{M_R}(\cdot)$ is also lower semi-continuous due to the fact that the $\delta_{M_R}(\cdot)$ is an indicator function of a closed set M_R [39]. ■

Since the sum of lower semi-continuous functions is also lower semi-continuous, we see that Φ in (2) and (4) are both lower semi-continuous.

The lower semi-continuity plays an important role in deriving the proximal maps. For a proper, lower semi-continuous and possibly nonconvex function $g : \mathbb{R}^n \to \mathbb{R}$ and parameter $\tau > 0$, the proximal map $\text{prox}_{g,\tau}$ and Moreau envelope $m_{g,\tau}$ [38] are defined by

$$\begin{aligned} \text{prox}_{g,\tau}(\mathbf{x}) &:= \arg\min \left\{ g(\mathbf{u}) + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{u}\|_F^2 \mid \mathbf{u} \in \mathbb{R}^n \right\}, \\ m_{g,\tau}(\mathbf{x}) &:= \inf \left\{ g(\mathbf{u}) + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{u}\|_F^2 \mid \mathbf{u} \in \mathbb{R}^n \right\}. \end{aligned}$$

The nonconvexity of g implies that $\text{prox}_{g,\tau}(\mathbf{x})$ may be a set-valued mapping.

The above definitions together with Proposition 1 give the proximal maps for $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$. $\text{prox}_{\|\cdot\|_*,\tau}(\mathbf{X})$ is the matrix shrinkage/soft thresholding operator [40], [41]

$$\text{prox}_{\|\cdot\|_*,\lambda}(\mathbf{X}) = \mathbf{U}\text{diag}\left(\max\{\sigma_i - \tau, 0\}\right)\mathbf{V}^T,$$

where $\mathbf{X} = \mathbf{U}\text{diag}\left(\{\sigma_i\}_{1\leq i\leq r}\right)\mathbf{V}^T$ is the SVD of $\mathbf{X}$. Correspondingly, $\text{prox}_{\delta_{M_R},\tau}(\mathbf{X})$ is the best rank-R approximation to $\mathbf{X}$, and is also called the matrix hard thresholding operator [37]. Noticing that the parameter τ does not effect $\text{prox}_{\delta_{M_R},\tau}(\mathbf{X})$, we can write it by $\text{prox}_{\delta_{M_R}}(\mathbf{X})$ for short.

B. The algorithm

At first glance, (2) and (4) have a separable structure with respect to $\mathcal{X}_j, 1 \leq j \leq N$, which can be solved via the conventional block coordinated descent (BCD) method.

BCD solves successively for $j = 1, \ldots, N$ the following subproblems

$$\mathcal{X}_j^{(k+1)} \in \arg\min_{\mathcal{Z}\in\mathbb{T}} \Phi(\mathcal{X}_1^{(k+1)}, \ldots, \mathcal{X}_{j-1}^{(k+1)}, \mathcal{Z}, \mathcal{X}_{j+1}^{(k)}, \mathcal{X}_N^{(k)}).$$

However, the above subproblems do not admit closed-form solutions. In view of this, and noticing that J_σ is differentiable everywhere, we can use the linearized and proximal algorithm to approximately solve the subproblems. At the $(k+1)$-iteration, suppose we already have $\mathcal{X}_1^{(k+1)}, \ldots, \mathcal{X}_{j-1}^{(k+1)}$. To obtain $\mathcal{X}_j^{(k+1)}$, we can solve the following subproblem

$$\begin{aligned}\mathcal{X}_j^{(k+1)} \in \arg\min_{\mathcal{X}\in\mathbb{T}}\ & g\left(\mathbf{X}_{(j)}\right) + \frac{\alpha}{2}\|\mathcal{X} - \mathcal{X}_j^{(k)}\|_F^2 \\ & + \left\langle \nabla_{\mathcal{X}_j} J_\sigma \left(\sum_{i=1}^{j-1} \mathcal{X}_i^{(k+1)} + \sum_{i=j}^{N} \mathcal{X}_i^{(k)} \right), \mathcal{X} - \mathcal{X}_j^{(k)} \right\rangle. \end{aligned} \tag{7}$$

Here $\alpha > 0$ is a parameter, and $\nabla_{\mathcal{X}_j} J_\sigma(\sum(\boldsymbol{\mathcal{X}}))$ is the partial derivative of J_σ with respect to $\mathcal{X}_j$ that will be specified later. By denoting

$$\mathcal{Y}_j^{(k+1)} = \mathcal{X}_j^{(k)} - \frac{1}{\alpha}\nabla_{\mathcal{X}_j} J_\sigma \left(\sum_{i=1}^{j-1} \mathcal{X}_i^{(k+1)} + \sum_{i=j}^{N} \mathcal{X}_i^{(k)} \right), \tag{8}$$

(7) can be concisely written as

$$\mathcal{X}_j^{(k+1)} \in \arg\min_{\mathcal{X}\in\mathbb{T}} \frac{1}{\alpha} g\left(\mathbf{X}_{(j)}\right) + \frac{1}{2}\|\mathcal{X} - \mathcal{Y}_j^{(k+1)}\|_F^2, \tag{9}$$

where by the definition of $g(\cdot)$, we have

$$\mathcal{X}_j^{(k+1)} \in \begin{cases} \operatorname{prox}_{\|\cdot\|_*, \lambda/\alpha}\left(\mathcal{Y}_j^{(k+1)}\right), & \text{if } g(\cdot) = \|\cdot\|_*, \\ \operatorname{prox}_{\delta_{M_{R_j}}}\left(\mathcal{Y}_j^{(k+1)}\right), & \text{if } g(\cdot) = \delta_{M_{R_j}}(\cdot). \end{cases} \tag{10}$$

Now we come to our algorithm. Given an initial guess $\boldsymbol{\mathcal{X}}^{(0)} = \{\mathcal{X}_1^{(0)}, \ldots, \mathcal{X}_N^{(0)}\} \in \mathbb{T}^N$, at each iteration, we compute successively $\mathcal{X}_1^{(k+1)}, \ldots, \mathcal{X}_N^{(k+1)}$ via solving (9). The algorithm stops whenever some stopping criterion is satisfied. This procedure is summarized in Algorithm 1.

Algorithm 1 Proximal and Linearized BCD with Gauss-Seidel update rule (PLiBCD-GS) for (2) and (4)

Input: linear operator $\mathcal{A} : \mathbb{T} \to \mathbb{R}^p$, initial guess $\boldsymbol{\mathcal{X}}^{(0)} \in \mathbb{T}^N$, parameters $R_j, 1 \le j \le N, \sigma > 0, \lambda > 0, \alpha > 0, \epsilon > 0$.

Output: the recovered tensors $\boldsymbol{\mathcal{X}}^{(k+1)} = \{\mathcal{X}_1^{(k+1)}, \ldots, \mathcal{X}_N^{(k+1)}\}$.

while certain stopping criterion is not satisfied **do**

  for $j = 1$ **to** N **do**

  • Compute $\mathcal{Y}_j^{(k+1)}$ via the gradient descent step (8).

  • Compute $\mathcal{X}_j^{(k+1)}$ via the soft/hard thresholding (10).

  end for

  Set $k := k + 1$.

end while

Remark 2: The algorithm is called the Gauss-Seidel update rule, as the computation of $\mathcal{Y}_j^{(k+1)}$ uses the information of the new trial $\mathcal{X}_1^{(k+1)}, \ldots, \mathcal{X}_{j-1}^{(k+1)}$ immediately. To ensure the convergence of the algorithm, a suitable step-size α^{-1} in (8) is preferred. A typical choice is $\alpha > L$, where L is the Lipschitz constant of partial derivative of J_σ that will be given explicitly later. One can also compute α^{-1} by certain line-search rule, such as the Armijo search rule [42]. Note that a Jacobi update rule can also be employed. That is, in (8) we only use $\mathcal{X}_1^{(k)}, \ldots, \mathcal{X}_N^{(k)}$ to compute $\mathcal{Y}_j^{(k+1)}$ for each j. The advantage is in each step, the computation of $\mathcal{X}_j^{(k+1)}$ can be parallel. However, an evident drawback is that it cannot utilize the latest information. The stopping criterion can be $\|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_F \le \epsilon$ where ϵ is a prescribed number. This will be explained in the next section.

Computing the gradients[1]**.** Now we specify the partial derivative of J_σ with respective to $\mathcal{X}_j$. At $\hat{\boldsymbol{\mathcal{X}}} = \{\hat{\mathcal{X}}_1, \ldots, \hat{\mathcal{X}}_N\}$, the partial derivative $\nabla_{\mathcal{X}_j} J_\sigma(\sum(\hat{\boldsymbol{\mathcal{X}}}))$ associated with the Cauchy loss is given by

$$\sum_{i=1}^{p} \left(1 + (\langle \mathcal{A}_i, \sum(\hat{\boldsymbol{\mathcal{X}}})\rangle - \mathbf{b}_i)^2/\sigma^2\right)^{-1} (\langle \mathcal{A}_i, \sum(\hat{\boldsymbol{\mathcal{X}}})\rangle - \mathbf{b}_i)\mathcal{A}_i,$$

which can also be further formulated into the following concise form

$$\nabla_{\mathcal{X}_j} J_\sigma \left(\sum(\hat{\boldsymbol{\mathcal{X}}})\right) = \boldsymbol{\mathcal{A}}^\star \boldsymbol{\Lambda} \left(\boldsymbol{\mathcal{A}}(\sum(\hat{\boldsymbol{\mathcal{X}}})) - \mathbf{b}\right), \tag{11}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ is a diagonal matrix, with its i-th diagonal entry $\boldsymbol{\Lambda}_{ii} = \left(1 + (\langle \mathcal{A}_i, \sum(\hat{\boldsymbol{\mathcal{X}}})\rangle - \mathbf{b}_i)^2/\sigma^2\right)^{-1}$, and $\boldsymbol{\mathcal{A}}^\star$ denotes the adjoint of $\boldsymbol{\mathcal{A}}$. It turns out that for tensor completion problems, the partial derivative enjoys a simpler form, i.e.,

$$\nabla_{\mathcal{X}_j} J_\sigma \left(\sum(\hat{\boldsymbol{\mathcal{X}}})\right) = \mathcal{W} * \boldsymbol{\Lambda} * \left(\sum(\hat{\boldsymbol{\mathcal{X}}}) - \mathcal{B}\right), \tag{12}$$

where "$*$" denotes the Hadamard operator, i.e., the entry-wise product, $\boldsymbol{\Lambda} \in \mathbb{T}$ with $\boldsymbol{\Lambda}_{i_1\cdots i_N} = \left(1 + (\sum(\hat{\boldsymbol{\mathcal{X}}})_{i_1\cdots i_N} - \mathcal{B}_{i_1\cdots i_N})^2/\sigma^2\right)^{-1}$, and $\mathcal{W} \in \mathbb{T}$ denotes the mask tensor, i.e., $\mathcal{W}_{i_1\cdots i_N} = 1$ if $(i_1 \cdots i_N) \in \Omega$ and $\mathcal{W}_{i_1\cdots i_N} = 0$ otherwise. Similarly, for the Welsch loss, $\nabla_{\mathcal{X}_j} J_\sigma(\sum(\hat{\boldsymbol{\mathcal{X}}}))$ takes the following form

$$\sum_{i=1}^{p} \exp\left(-(\langle \mathcal{A}_i, \sum(\hat{\boldsymbol{\mathcal{X}}})\rangle - \mathbf{b}_i)^2/\sigma^2\right) (\langle \mathcal{A}_i, \sum(\hat{\boldsymbol{\mathcal{X}}})\rangle - \mathbf{b}_i)\mathcal{A}_i.$$

Remark 3: The above gradients are similar with the one obtained by using the least squares estimator, except with the introduction of a weight matrix $\boldsymbol{\Lambda}$. We also observe that $\boldsymbol{\Lambda}_{ii}$ gives a small weight if the corresponding error goes large, which is essentially the key to reduce the influence of outliers.

V. Convergence Results

In this section, we verify the convergence results of PLiBCD-GS (Algorithm 1) step by step. First, we show that their associated partial derivative are Lipschitz continuous, based on which, the descent property of the algorithms can be derived. Then the convergence results are strengthened

[1]Similar to the vector and matrix cases, the gradient of a function with respect to a tensor is defined entry-wise, e.g., $\frac{d\|\mathcal{X}\|_F^2}{d\mathcal{X}} = 2\mathcal{X}$.

by verifying that the function Φ is essentially a Kurdyka-Łojasiewicz function and using the results of [31].

Lipschitz continuity of the gradients. First we show the Lipschitz continuity of the partial derivative $\nabla_{\mathcal{X}_j} J_\sigma(\sum(\boldsymbol{\mathcal{X}}))$, where $\nabla_{\mathcal{X}_j}$ refers to the j-th partial derivative, $1 \leq j \leq N$. Denote $L = \|\boldsymbol{\mathcal{A}}\|_2^2$, where $\|\boldsymbol{\mathcal{A}}\|_2$ is the spectral norm of $\boldsymbol{\mathcal{A}}$.

Proposition 2: For each $j = 1, \ldots, N$ and for any $\mathcal{X}_j, \mathcal{X}_j^+ \in \mathbb{T}$, it holds that

$$\left\| \nabla_{\mathcal{X}_j} J_\sigma(\sum_{i \neq j}^{N} \mathcal{X}_i + \mathcal{X}_j^+) - \nabla_{\mathcal{X}_j} J_\sigma(\sum_{i=1}^{N} \mathcal{X}_i) \right\|_F \leq L \left\| \mathcal{X}_j^+ - \mathcal{X}_j \right\|_F.$$

The proof is left to Appendix A.

Descent property. Based on the above propositions, we can derive the descent property for the developed algorithms.

Proposition 3 (Descent property): For PLiBCD-GS, if the step-size α^{-1} satisfies $\alpha > L$, then it holds that

$$\Phi(\boldsymbol{\mathcal{X}}^{(k+1)}) \leq \Phi(\boldsymbol{\mathcal{X}}^{(k)}) - \frac{\alpha - L}{2} \|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_F^2. \quad (13)$$

The proof is left to Appendix A.

The above property shows that $\Phi(\boldsymbol{\mathcal{X}}^{(k)})$ is a non-increasing sequence. Since $\Phi(\cdot)$ is lower bounded by zero, we can deduce that $\|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_F \to 0$. Thus $\|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_F \leq \epsilon$ can serve as a stopping criterion where ϵ is a prescribed number.

Global convergence. By verifying that Φ satisfies the assumptions in [31], we have

Theorem 1 (Global convergence): For PLiBCD-GS, assume that the step-size α^{-1} satisfies $\alpha > L$. If $\{\boldsymbol{\mathcal{X}}^{(k)}\}$ is a sequence generated by PLiBCD-GS for solving the nuclear norm regularized problem (2), then

1) $\sum_{k=1}^{\infty} \|\boldsymbol{\mathcal{X}}^{(k+1)} - \boldsymbol{\mathcal{X}}^{(k)}\|_F < \infty$;
2) $\{\boldsymbol{\mathcal{X}}^{(k)}\}$ converges to a critical point of Φ.

If $\{\boldsymbol{\mathcal{X}}^{(k)}\}$ is a bounded sequence generated by PLiBCD-GS for solving the rank constrained problem (4), then the above two conclusions also hold.

Detailed discussions are left to Appendix B.

Remark 4: We follow [31] in using the words "global convergence". In fact, they do not mean that the sequence converges to a global optimizer of the optimization problem. Instead, it states that the whole sequence converges to a critical point, which strengthens the conventional convergence results "every limit point of the sequence is a critical point".

VI. Numerical Experiments

In this section, we present some numerical experiments on synthetic data as well as real data to illustrate the effectiveness of our methods for tensor completion problems. All the numerical computations are conducted on an Intel i7-3770 CPU desktop computer with 16 GB of RAM. The supporting software is MATLAB R2013a. The MATLAB toolbox Tensorlab [43] is employed to perform tensor computations. The following methods are mainly compared:

1) Our proposed method PLiBCD-GS (W, ST): The PLiBCD-GS method is used to solve (2) with the Welsch loss. "ST" is short for "soft thresholding". This method is abbreviated as **W-ST** without any confusion. Similar notation **C-ST** is short for solving (2) with the Cauchy loss.

2) Our proposed method PLiBCD-GS (W, HT): The PLiBCD-GS method is used to solve (4) with the Welsch loss. "HT" is short for "hard thresholding". **W-HT** is short for PLiBCD-GS (W, HT). Similar notation **C-HT** is short for solving (4) with the Cauchy loss.

3) Four LAD loss based methods [21]: **S-ST**, **S-HT**, **M-ST** and **M-HT**. They are defined as follows. Denote $L(\mathcal{X}) := \sum_{(i_1 \cdots i_N) \in \Omega} |\mathcal{X}_{i_1 \cdots i_N} - \mathcal{B}_{i_1 \cdots i_N}|$ and $L(\sum(\boldsymbol{\mathcal{X}})) := \sum_{(i_1 \cdots i_N) \in \Omega} |\sum(\boldsymbol{\mathcal{X}})_{i_1 \cdots i_N} - \mathcal{B}_{i_1 \cdots i_N}|$. Then S-HT and M-ST respectively refer to the singleton models

$$\min_{\mathcal{X} \in \mathbb{T}} L(\mathcal{X}) + \lambda \sum_{j=1}^{N} \|\mathbf{X}_{(j)}\|_* \text{ and } \min_{\mathcal{X} \in \mathbb{T}} L(\mathcal{X}) + \sum_{j=1}^{N} \delta_{M_{R_j}}(\mathbf{X}_{(j)});$$

M-ST and M-HT respectively refer to the mixture models

$$\min_{\mathcal{X} \in \mathbb{T}^N} L(\sum(\boldsymbol{\mathcal{X}})) + \lambda \sum_{j=1}^{N} \|\mathbf{X}_{j,(j)}\|_* \text{ and}$$

$$\min_{\mathcal{X} \in \mathbb{T}^N} L(\sum(\boldsymbol{\mathcal{X}})) + \sum_{j=1}^{N} \delta_{M_{R_j}}(\mathbf{X}_{j,(j)}).$$

4) The least squares loss based method with nuclear norm heuristic [6] **LS-ST**:

$$\min_{\mathcal{X} \in \mathbb{T}} \lambda/2 \sum_{(i_1 \cdots i_N) \in \Omega} (\mathcal{X}_{i_1 \cdots i_N} - \mathcal{B}_{i_1 \cdots i_N})^2 + \sum_{j=1}^{N} \|\mathbf{X}_{(j)}\|_*.$$

Some frequently used notations are introduced in Table I.

TABLE I: Notations in the experiments

Notation	Description
ρ_r	the ratio of the rank to the dimensionality of a mode
ρ_o	the ratio of outliers to the number of entries of a tensor
ρ_m	the ratio of missing entries
s	the factor of scale of outliers

A. Synthetic data

We first generate tensors of size $(50, 50, 50)$ with entries drawn i.i.d. from normal distribution, and then truncate the rank of each mode to yield tensors of the Tucker rank $(50\rho_r, 50\rho_r, 50\rho_r)$. The low rank tensors are denoted as $\mathcal{T}$. The error tensor is denoted by $\mathcal{E}$, with ρ_o entries drawn from some given distributions while other $(1 - \rho_o)$ entries being zero. Then $\rho_m \times 100\%$ of the entries of $\mathcal{T} + s\mathcal{E}$ are randomly missing. The following three situations are considered:

1) We fix missing ratio $\rho_m = 0.3$, rank ratio $\rho_r = 0.1$, scale factor $s = 1$, and vary outliers ratio ρ_o from 0 to 1.The outliers are drawn from the Chi-Square distribution.
2) We fix $\rho_r = 0.1$, $\rho_o = 0.2$, $s = 2$, and vary ρ_m from 0 to 1. Outliers are drawn from the Chi-Square distribution.
3) We fix $\rho_m = 0.3$, $\rho_o = 0.2$, $s = 2$, and vary ρ_r from 0.05 to 1.

The nine methods mentioned in the beginning of this section are tested. The scale parameters σ of the Welsch loss and the Cauchy loss are empirically set to 0.1 and 0.05, respectively. The regularization parameters λ of W-ST and C-ST are set to $0.05 n_{\min}/\sqrt{n_{\max}}$. Parameters of S-ST and M-ST are set following the suggestions in [21]. The ranks for the HT type methods are set to the Tucker ranks of $\mathcal{T}$. The initial guess for all the methods is the zero tensor. It seems that since our

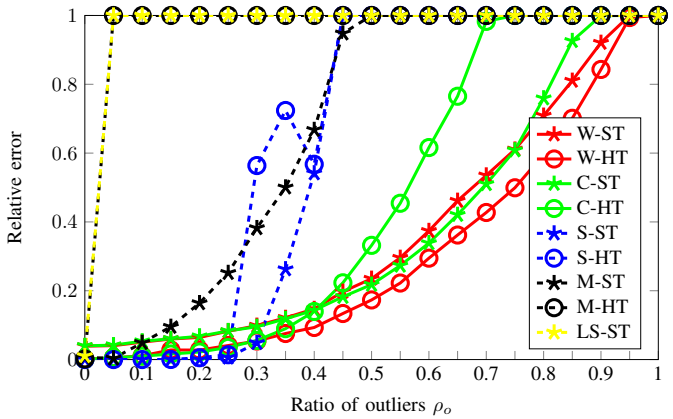


Fig. 2: Recovery results of third-order tensors with $\rho_m = 0.3$, $\rho_r = 0.1$, $s = 1$ and varying ρ_o. The results show that our methods are more resistant to outliers.

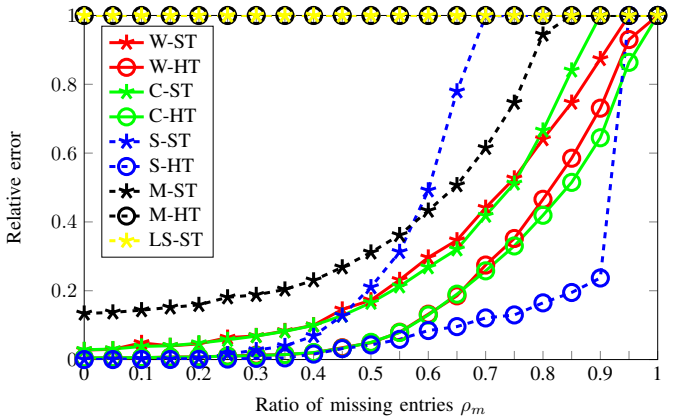


Fig. 3: Recovery results of third-order tensors with $\rho_o = 0.2$, $\rho_r = 0.1$, $s = 2$ and varying ρ_m, using different methods.

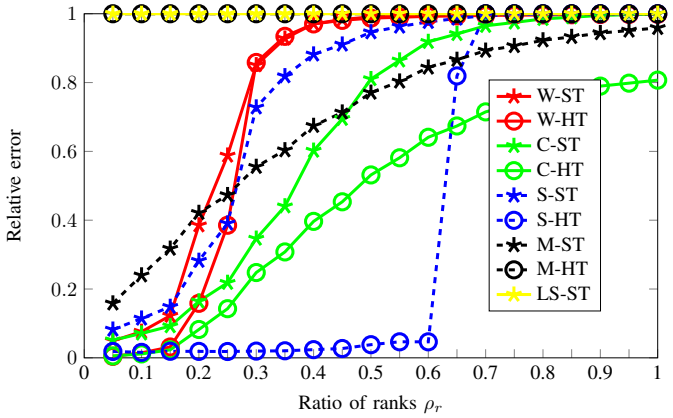


Fig. 4: Recovery results of third-order tensors with $\rho_m = 0.3$, $\rho_o = 0.2$, $s = 2$ and varying ρ_r, using different methods.

methods are nonconvex, the results may be potentially influenced by the initial value. However, we have tried different initial guesses and find that the zero tensor gives better or comparable results. Thus for simplicity and reproducibility, we choose zero as our initial guess. The stopping criterion is $\|\mathcal{X}^{(k+1)} - \mathcal{X}^{(k)}\|_F \leq 10^{-4}$ or the iterations exceed 200. The relative error $\|\mathcal{T} - \mathcal{X}\|_F / \|\mathcal{T}\|_F$ is used to measure the recovery ability of the methods. If the error is larger than 1, then we set it to 1. All the results are averaged by 10 instances.

Recovery results of the four situations in terms of the relative error are reported in Figs. 2, 3 and 4, respectively. In these figures, plots of W-ST and W-HT are in red; plots of C-ST and C-HT are in green; S-ST and S-HT are in blue, M-ST and M-HT are in black and LS-ST is in yellow. Plots of the ST type methods are marked with circles "$\circ$", while the HT type methods with five-point star "$\star$". From all the plots, we observe that LS-ST performs poor when there exist outliers. M-HT also gives poor results. In fact, the relative error of M-HT is larger than one in almost all the tests. The reason may be that the convergence of the algorithm is not guaranteed [21]. From Fig. 2, we observe that our methods have significantly higher recovery accuracy than those in [21]. From Fig. 3, we see that when the level of missing entries increases, our ST type methods perform more stable when $\rho_m > 0.45$. Comparing with the HT type methods, S-HT performs better when $\rho_m > 0.6$. This might be due to that when the true ranks are known, the Tucker model used in S-HT is more suitable to approximate the original tensor. Considering the cases of varying ρ_r and the ST type methods, Fig. 4 shows that C-ST outperforms other methods when $\rho_r \leq 0.45$; for HT type methods, S-HT gets a higher recovery accuracy when $\rho_r \leq 0.6$, which might be also due to the same reason, but the accuracy decreases sharply when $\rho_r > 0.6$. In summary, our methods are more robust in most situations.

B. Traffic flow data

In intelligent transportation systems, traffic data such as traffic flow volume are collected to support applications in traffic prediction, optimal route choice and others. Due to some hardware or software problems, the data might be contaminated by outliers, or having missing values. Therefore, some techniques should be performed to recover the original data. The data used here records traffic flow volume from a detector in San Diego and can be downloaded from http://pems.dot.ca.gov. The data consist of 8928 records, each of which represents the traffic flow volume per 5 minutes, and are collected from Jan. 1st to Jan. 31st 2014. The data can be mapped into a third-order tensor of size $12 \times 24 \times 31$, with modes indicating samples per hour, hours per day and days per month, respectively. The tensor may have low rank structure due to the periodicity. We plot the vectorizations of the tensor along each mode in Fig. 5, from which we can see that the periodicity of the records confirms the low rank property of the tensor. The data can also be modeled as a fourth-order tensor by imposing a week mode since a week trend is also available. In this case, the size of the tensor is $12 \times 24 \times 7 \times 5$, where the data of the last four days of the last week can be seen as missing.

In order to assess the performance of our approaches, randomly chosen entries are corrupted by outliers, with scale being the largest magnitude of the entries. 0% or 20% of the entries are randomly missing. W-ST and C-ST are applied to this problem and compared with S-ST, M-ST and LS-ST. The scale parameters σ of the Welsch loss and the

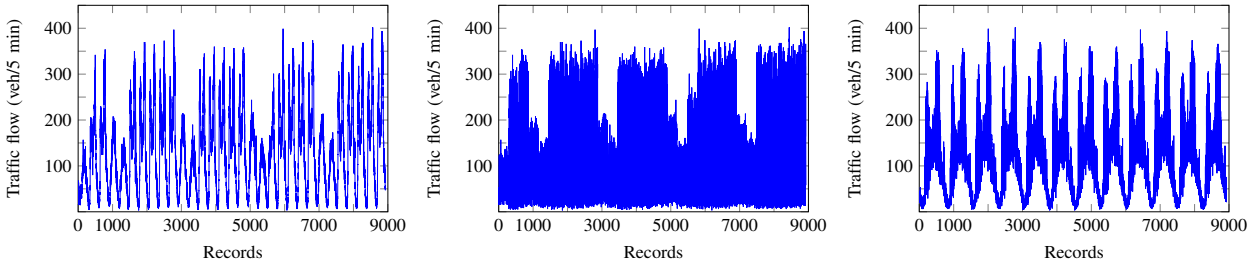
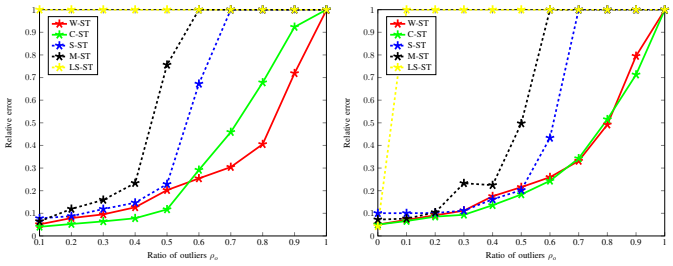


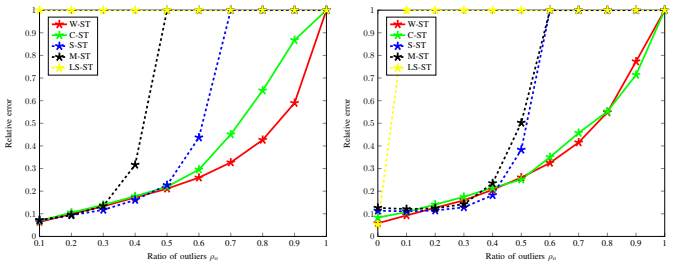
Fig. 5: Plots of vectorizations of the traffic data tensor along mode-1, mode-2 and mode-3, respectively.



(a) Fully observed and varying outliers

(b) 20% missing entries and varying outliers

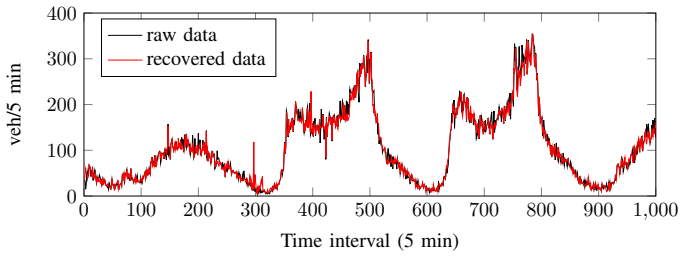
Fig. 6: Comparison between W-ST, C-ST, S-ST, M-ST and LS-ST on traffic data corrupted by outliers and missing entries. The data are represented by a third-order tensor of size $12 \times 24 \times 31$.



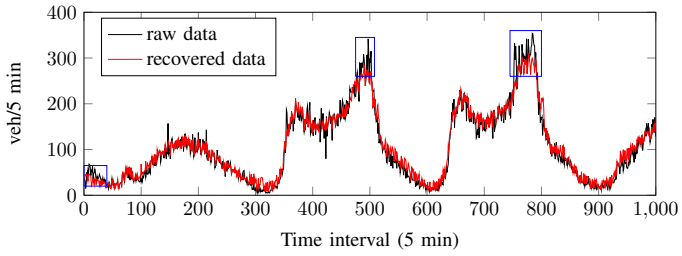
(a) Fully observed and varying outliers

(b) 20% missing entries and varying outliers

Fig. 7: Comparison between W-ST, C-ST, S-ST, M-ST and LS-ST on traffic data corrupted by outliers and missing entries. The data are represented by a fourth-order tensor of size $12 \times 24 \times 7 \times 5$.



(a) C-ST



(b) S-ST

Fig. 8: Plots of the first 1000 records and the recovery results of C-ST (top), S-ST (bottom). The data are corrupted by 30% outliers and 20% missing entries.

Cauchy loss are empirically set to 1 and 0.2, respectively. The regularization parameters of our methods are both set to $\lambda = 0.1 n_{\min}/\sqrt{n_{\max}}$. Parameters of S-ST and M-ST are set according to the suggestions in [21]. The initial guesses for all the methods are the zero tensors. The algorithms stop whenever $\|\mathcal{X}^{(k+1)} - \mathcal{X}^{(k)}\|_F \leq 10^{-4}$ or the iterations exceed 500. All the results are averaged over 10 instances.

Relative error results of the algorithms evaluated on the generated third-order and fourth-order tensors are plotted in Fig. 6 and Fig. 7, respectively. From the figures we again observe that LS-ST cannot be resistant with outliers. We can also see that our methods perform better than S-ST and M-ST in most situations. As ρ_o increases, the performances of S-ST and M-ST decrease sharply. Comparing between W-ST and C-ST, it seems that they performs similar when ρ_0 is small, while W-ST is better when ρ_0 increases. This indicates that W-ST is more robust. All the methods seem to have slightly better performance on the third-order tensor model than on the fourth-order tensor model when $\rho_0 \leq 0.3$. The reason may be that the data of the last four days of the last week is missing in the fourth-order tensor model, which weakens the low rank property of the tensor. Finally, we report that in the third-order tensors case, the ranks of the unfoldings of the three recovered factor tensors by our methods are $(\text{rank}(\mathbf{X}_{1,(1)}), \text{rank}(\mathbf{X}_{2,(2)}), \text{rank}(\mathbf{X}_{3,(3)})) = (12, 4, 1)$, respectively, which indicates that the original tensor can be approximated by the sum of three low rank tensors.

To intuitively illustrate the recovered results, we plot the first 1000 records and the recovered data of C-ST and S-ST in Fig. 8. In each sub-figure, the black line represents the raw data while the red one denotes the recovered data. We can observe that S-ST (Fig. 8b) does not perform well in the intervals near 0, 500 and 800, which may be due to the reason that S-ST treats these records as outliers. On the contrary, the data recovered by our method C-ST fit the raw data better, as we observe from Fig. 8a.

C. Image inpainting

A color image of size $m \times n$ can be represented as a third-order tensor of size $m \times n \times 3$ by considering the three channels

(red, green and blue) of the image. In real life, images may be corrupted by outliers or partially masked by text. To recover the damaged information, robust tensor completion approaches can be applied. By noticing that the data tensors may be only low rank in the first two modes, the tensor models in our approaches (2) and (4) only consist of two factors, which are low rank in the first two modes, respectively. The experiments are conducted as follows:

1) 4 color images named "House" (256×256), "Lena" (225×225), "Lake" (512×512), and "Pepper" (512×512) are tested. Randomly chosen entries are corrupted by outliers with magnitude restricted in $[-1, 1]$. The ratio of outliers ρ_o varies form 0 to 0.4. The images are then covered by some text. Some corrupted images are shown in Fig. 9.

2) The eight robust methods mentioned in the beginning of this section are compared in this experiment. The scale parameters σ of W-ST and C-ST are set to 0.3 and 0.1, respectively, with the regularization parameters being $\lambda = 0.05 \min\{m, n\}/\sqrt{\max\{m, n\}}$. To obtain the ranks for W-HT and C-HT, we use the ranks of the factor tensors generated by W-ST and C-ST and rescale them by 0.15, respectively. The parameters of other methods are set following the suggestions in [21]. In accordance with our methods, the tensor models of M-ST and M-HT also contain two factors only. We also compare with two of the state-of-the-art methods in [44], [45], where the method in [44] is based on tensor decomposition and uses prior information of decomposition factors, and is denoted by FP; the second method is based on the PDE, Cahn-Hilliard equation, and is denoted as CH.

3) The initial guesses for all the methods are the zero tensors. The stopping criterion is $\|\mathcal{X}^{(k+1)} - \mathcal{X}^{(k)}\|_F \leq 10^{-4}$, or the iterations exceed 200. The Peak-Signal-to-Noise-Ratio (PSNR) value is employed to measure the recovery results.

The PSNR values of images recovered by different methods are reported in Table II. From the table, we see that when there are no outliers, the FP method outperforms other methods. When there exist outliers, the performance of FP decreases sharply. In this setting, our method W-ST achieves the highest values among all the images and all the situations, followed by C-ST. This may be because the Welsch loss is more robust than the Cauchy loss, which can be implied from the discussions about the weight of outliers in Section III-C and seen from Fig. 1. LAD loss based methods are more sensitive to outliers than ours. Considering the methods with the hard thresholding, namely W-HT, we observe that the HT type methods do not perform as well as the ST type methods, which is due to the reason that the HT type methods cannot completely remove the text from the images. To intuitively illustrate the effectiveness of our methods, recovered images of "Pepper" are presented in Fig. 10. We can see that the image recovered by W-ST can retain more details and remove the text, followed by C-ST. S-ST and M-ST can also remove the text but lose more details than our methods, which may be that they are not as robust as our methods. W-HT also retains more details than other HT type methods, but some text cannot be totally removed, as masked by the ellipsoid in (10b). S-HT and M-HT cannot remove the outliers. In fact, we have tried different choices of ranks while get similar results. We also observe that CH is influenced by outliers seriously.

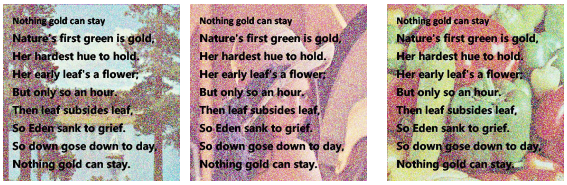
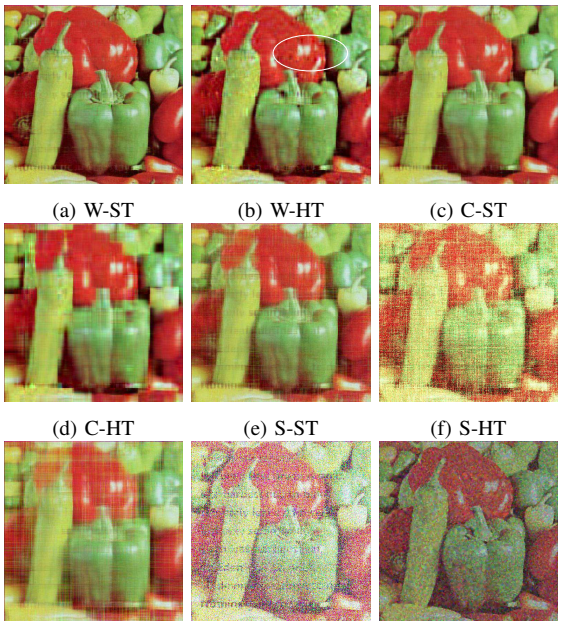


Fig. 9: Examples of images corrupted by outliers and masked by text.



(a) W-ST (b) W-HT (c) C-ST

(d) C-HT (e) S-ST (f) S-HT

(g) M-ST (h) M-HT (i) CH

Fig. 10: Recovery results of image "Pepper" contaminated by 40% outliers.

D. MRI data

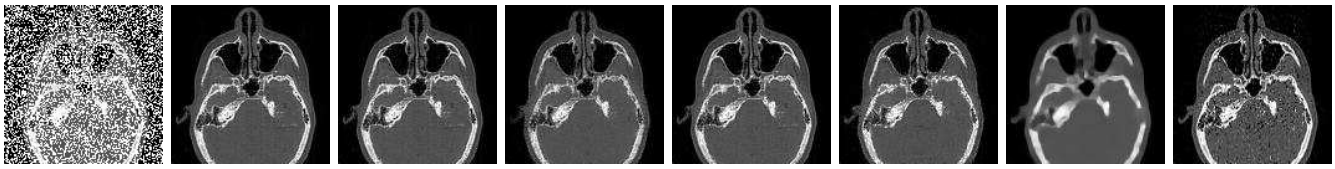
We test different methods on the MRI dataset INCISIX ($166 \times 128 \times 128$) and KNIX from http://www.osirix-viewer.com/datasets/. From KNIX, we choose two datasets, each of which is of size $22 \times 128 \times 128$. We try W-ST, C-ST, S-ST, M-ST, LS-ST, CH and FP on the data with different ratios of missing entries. The parameters are set the same as the previous experiment. The model LS-ST is solved iteratively with an increasing λ value because of the absence of noise. The relative errors are reported in Table III, along with the recovered results of one slice of INCISIX presented in Fig. 11. The results show that in most cases, W-ST, C-ST perform better than other methods. The reason may be that the raw data can be better approximated by the mixture tensor model. CH performs worst, which may be that the PDE type methods do not explore the low rank structure of the tensor to be recovered. The results also indicate that our methods are safe when there is no noise or outliers.

E. Removing shadows and specularities from face images

We test our methods on the YaleB dataset. The goal is to remove shadows and specularities from face images. The dataset consists of 64 face images of size 192×168, which

TABLE II: PSNR values of the recovery images by different methods. The original images are contaminated by different level of outliers and masked by text. The outliers ratio ρ_o varies from 10% to 40%.

Image	ρ_o	W-ST	W-HT	C-ST	C-HT	S-ST [21]	S-HT [21]	M-ST [21]	M-HT [21]	CH [45]	FP [44]
House	0	29.86	24.25	27.47	22.75	26.22	14.47	27.13	15.70	32.24	**32.93**
	10%	**29.33**	22.78	26.88	22.42	25.10	15.98	26.54	10.59	11.93	10.28
	20%	**28.70**	23.63	26.20	22.24	24.86	12.89	25.45	7.02	13.40	6.99
	30%	**27.90**	24.13	25.31	21.51	24.11	9.51	23.27	4.95	14.98	5.30
	40%	**26.99**	23.20	24.28	19.93	21.48	6.14	19.22	3.56	16.49	4.15
Lake	0	27.78	23.43	26.22	22.61	23.90	15.16	25.22	16.59	27.87	**29.70**
	10%	**27.31**	23.23	25.52	22.42	23.68	15.03	24.27	10.34	12.50	10.40
	20%	**26.72**	23.20	24.67	22.18	23.30	11.33	23.41	6.79	13.79	7.34
	30%	**26.00**	22.85	23.62	21.60	22.10	8.08	20.94	4.82	15.08	5.58
	40%	**25.07**	22.46	22.23	20.63	19.71	5.58	19.48	3.57	15.97	4.33
Lena	0	29.23	25.29	27.88	24.71	26.61	15.30	26.95	16.26	31.07	**31.77**
	10%	**28.83**	25.22	27.33	24.47	25.80	15.40	26.45	10.65	12.49	10.48
	20%	**28.33**	25.08	26.65	23.95	25.41	11.91	25.45	6.99	13.95	7.39
	30%	**27.73**	24.86	25.85	23.14	23.97	8.63	23.11	4.97	15.41	5.62
	40%	**26.97**	24.09	24.73	22.30	21.32	5.74	19.45	3.61	16.68	4.35
Pepper	0	29.67	25.00	28.20	23.33	26.52	16.30	27.99	17.52	32.71	**32.91**
	10%	**29.22**	25.03	27.50	22.87	26.03	15.20	27.06	10.60	13.29	10.42
	20%	**28.67**	24.85	26.66	22.49	25.32	11.21	25.50	6.98	14.67	7.31
	30%	**27.97**	24.69	25.61	21.84	23.48	8.12	21.66	4.97	15.87	5.56
	40%	**27.11**	24.37	24.30	20.86	20.15	5.48	18.28	3.62	16.57	4.32



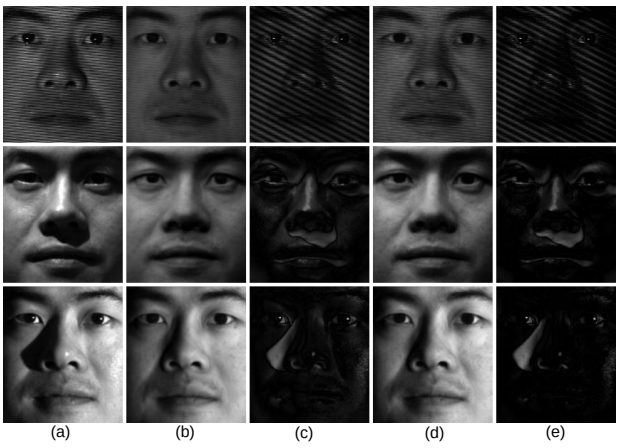
(a) One slice (b) W-ST (c) C-ST (d) S-ST [21] (e) M-ST [21] (f) LS-ST [6] (g) CH [45] (h) FP [44]

Fig. 11: Recovery results of one slice of the INCISIX dataset. (11a) One slice with 40% entries missing. (11b)-(11h) Recovered results of different methods.

TABLE III: Relative errors of different methods on recovering the INCISIX (D1), KNIX1 (D2) and KNIX2 (D3) datasets.

	ρ_m	W-ST	C-ST	S-ST [21]	M-ST [21]	LS-ST [6]	CH [45]	FP [44]
D1	0.4	**0.1075**	0.1095	0.1489	0.1205	0.1386	0.2078	0.1282
	0.6	0.1765	**0.1738**	0.2738	0.2008	0.2057	0.2398	0.1913
	0.8	0.2940	**0.2906**	0.4495	0.3654	0.3361	0.2917	0.3766
D2	0.4	**0.0948**	0.0960	0.1403	0.1194	0.1009	0.4101	0.1169
	0.6	0.1719	0.1740	0.2005	0.1788	**0.1581**	0.4595	0.1653
	0.8	0.2441	**0.2418**	0.3440	0.2726	0.2511	0.5689	0.2608
D3	0.4	0.0969	0.0979	0.1634	0.1326	0.0930	0.2683	**0.0937**
	0.6	0.1394	0.1382	0.1983	0.1692	0.1635	0.3137	**0.1355**
	0.8	0.2357	**0.2322**	0.3805	0.3066	0.2863	0.3863	0.2486

gives a tensor of size $64 \times 192 \times 168$. The shadows and specularities can be seen as outliers [35]. We apply W-ST and C-ST on the dataset and the methods can successfully process the images. We compare our methods with the robust PCA (RPCA) approach [35], which is solved by the algorithm proposed in [46]. To save space, we only present results generated by W-ST and RPCA in Fig. 12. It seems that our method may better remove the shadows and specularities, particularly from those presented in the first row of Fig. 12.



(a) (b) (c) (d) (e)

Fig. 12: Results of some face images. (a) Three original images. (b) Faces recovered by W-ST. (c) Shadows and specularities of W-ST. (d) Faces recovered by RPCA. (e) Shadows and specularities of RPCA. The results show that our method may better remove shadows and specularities from faces images.

F. Video surveillance

For the last experiment, we test our methods on the video surveillance data which can be downloaded from http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html. The aim is to detect the moving objects from static background. The dataset used here consists of 100 color frames of size $144 \times 176 \times 3$, which forms a tensor of size $100 \times 144 \times 176 \times 3$. The moving objects can be regarded as outliers. We apply W-ST on this dataset and compare our method with RPCA. Some results are illustrated in Fig. 13. The results show that our method is comparable with RPCA. Particularly, from the first figures of column (b) and column (d), it seems that our method can better extract the moving objects from the background.

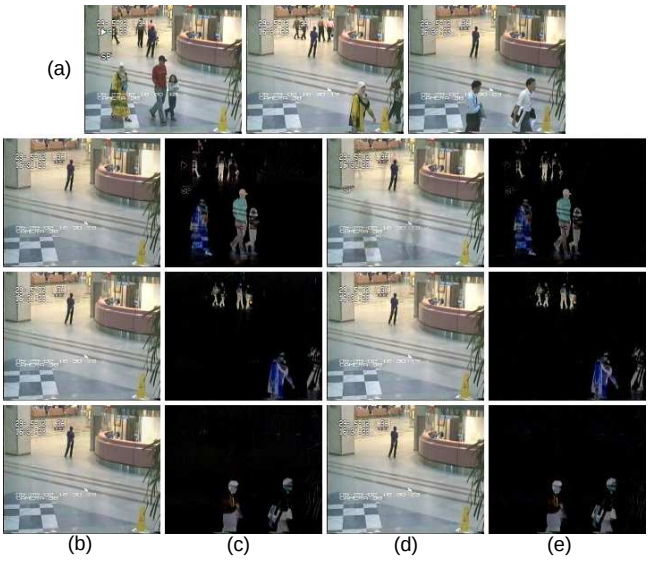


Fig. 13: Extracting results of some frames. Row (a). Three original frames. Column (b)–(c). W-ST: Background and moving objects. Column (d)–(e). RPCA: Background and moving objects.

VII. Concluding Remarks

This paper addressed the robust tensor recovery problems by using regularized redescending M-estimators, where the Welsch loss and the Cauchy loss were employed. To solve the nonconvex problems, a linearized and proximal block coordinate methods was developed and global convergence was verified. Numerical experiments on synthetic data and real data under various circumstances were conducted. Empirically we find that our models are more robust in the presence of outliers and can give at least comparable performance in the absence of outliers. Thus it is safe to use our methods in tensor recovery problems under various circumstances.

Appendix A
Proof of Proposition 2 and Proposition 3

Proof of Proposition 2: To prove the conclusion, it suffices to show that for any $\mathcal{X}^+, \mathcal{X} \in \mathbb{T}$

$$\left\|\nabla J_\sigma(\mathcal{X}^+) - \nabla J_\sigma(\mathcal{X})\right\|_F \leq L\left\|\mathcal{X}^+ - \mathcal{X}\right\|_F. \tag{14}$$

We first observe that

$$\begin{aligned}&\left\|\nabla J_\sigma(\mathcal{X}^+) - \nabla J_\sigma(\mathcal{X})\right\|_F\\ =\ &\left\|\boldsymbol{\mathcal{A}}^*\boldsymbol{\Lambda}_+\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}^+\right) - b\right) - \boldsymbol{\mathcal{A}}^*\boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}\right) - b\right)\right\|_F\\ \leq\ &\|\boldsymbol{\mathcal{A}}\|_2\left\|\boldsymbol{\Lambda}_+\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}^+\right) - b\right) - \boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}\right) - b\right)\right\|_F,\end{aligned} \tag{15}$$

where $\boldsymbol{\Lambda}_+$ and $\boldsymbol{\Lambda}$ are respectively the diagonal matrices corresponding to $\nabla J_\sigma(\mathcal{X}^+)$ and $\nabla J_\sigma(\mathcal{X})$. Let $\mathbf{z}^+ = \boldsymbol{\mathcal{A}}(\mathcal{X}^+) - b$ and $\mathbf{z} = \boldsymbol{\mathcal{A}}(\mathcal{X}) - b$. Associated with the Cauchy loss we have

$$\begin{aligned}&\left\|\boldsymbol{\Lambda}_+\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}^+\right) - b\right) - \boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}\right) - b\right)\right\|_F^2\\ =\ &\sum\nolimits_{i=1}^{p}\left(\left(1+(\mathbf{z}_i^+)^2/\sigma^2\right)^{-1}\mathbf{z}_i^+ - \left(1+\mathbf{z}_i^2/\sigma^2\right)^{-1}\mathbf{z}_i\right)^2.\end{aligned}$$

Associated with the Welsch loss we have

$$\begin{aligned}&\left\|\boldsymbol{\Lambda}_+\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}^+\right) - b\right) - \boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}\right) - b\right)\right\|_F^2\\ =\ &\sum\nolimits_{i=1}^{p}\left(\exp\left(-(\mathbf{z}_i^+)^2/\sigma^2\right)\mathbf{z}_i^+ - \exp\left(-\mathbf{z}_i^2/\sigma^2\right)\mathbf{z}_i\right)^2.\end{aligned}$$

For the Cauchy loss, it can be verified that the magnitude of the derivative of function $(1+t^2/\sigma^2)^{-1}t$ is not larger than 1 for any $t \in \mathbb{R}$ and $\sigma > 0$. Therefore, it follows from the mean value theorem that for any $t_1, t_2 \in \mathbb{R}$ and $\sigma > 0$,

$$|(1+t_1^2/\sigma^2)^{-1}t_1 - (1+t_2^2/\sigma^2)^{-1}t_2| \leq |t_1 - t_2|.$$

A similar inequality holds for the Welsch loss. Thus we obtain

$$\begin{aligned}&\left\|\boldsymbol{\Lambda}_+\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}^+\right) - b\right) - \boldsymbol{\Lambda}\left(\boldsymbol{\mathcal{A}}\left(\mathcal{X}\right) - b\right)\right\|_F\\ \leq\ &\|\boldsymbol{\mathcal{A}}\mathcal{X}^+ - \boldsymbol{\mathcal{A}}\mathcal{X}\|_F \leq \|\boldsymbol{\mathcal{A}}\|_2\|\mathcal{X}^+ - \mathcal{X}\|_F.\end{aligned}$$

This in connection with (15) implies (14). The proof is completed. ■

Proof of Proposition 3:

First, since $\mathcal{X}_j^{(k+1)}$ is an optimal solution of (7), we get

$$\begin{aligned}\frac{\alpha}{2}\|\mathcal{X}_j^{(k+1)} - \mathcal{X}_j^{(k)}\|_F^2 + \left\langle\nabla_{\mathcal{X}_j}J_\sigma, \mathcal{X}_j^{(k+1)} - \mathcal{X}_j^{(k)}\right\rangle\\ +g(X_{j,(j)}^{(k+1)}) \leq g(X_{j,(j)}^{(k)}),\end{aligned} \tag{16}$$

where $\nabla_{\mathcal{X}_j}J_\sigma$ is short for the partial derivative of J_σ, $\nabla_{\mathcal{X}_j}J_\sigma\left(\sum_{i=1}^{j-1}\mathcal{X}_i^{(k+1)} + \sum_{i=j}^{N}\mathcal{X}_i^{(k)}\right)$. Then, Proposition 2 implies that

$$\begin{aligned}&J_\sigma(\sum_{i=1}^{j}\mathcal{X}_i^{(k+1)} + \sum_{i=j+1}^{N}\mathcal{X}_i^{(k)}) \leq J_\sigma(\sum_{i=1}^{j-1}\mathcal{X}_i^{(k+1)} + \sum_{i=j}^{N}\mathcal{X}_i^{(k)})\\ &+\left\langle\nabla_{\mathcal{X}_j}J_\sigma, \mathcal{X}_j^{(k+1)} - \mathcal{X}_j^{(k)}\right\rangle + \frac{L}{2}\|\mathcal{X}_j^{(k+1)} - \mathcal{X}_j^{(k)}\|_F^2.\end{aligned} \tag{17}$$

Combining (16) and (17) yields

$$\begin{aligned}&J_\sigma(\sum_{i=1}^{j}\mathcal{X}_i^{(k+1)} + \sum_{i=j+1}^{N}\mathcal{X}_i^{(k)}) + g(X_{j,(j)}^{(k+1)}) \leq\\ &J_\sigma(\sum_{i=1}^{j-1}\mathcal{X}_i^{(k+1)} + \sum_{i=j}^{N}\mathcal{X}_i^{(k)}) + g(X_{j,(j)}^{(k)}) - \frac{a-L}{2}\|\mathcal{X}_j^{(k+1)} - \mathcal{X}_j^{(k)}\|_F^2.\end{aligned} \tag{18}$$

Summing (18) from $j = 1$ to N thus gives (13). ■

Appendix B
Kurdyka-Łojasiewicz property and the global convergence

Suppose $\{\mathbf{x}^{(k)}\}$ is a sequence generated by an algorithm applied to a nonconvex optimization problem. Recently, the promising work in [31] ([31, Theorem 1] and [31, Sec. 3.6]) shows that, if the objective function of the problem satisfies the Kurdyka-Łojasiewicz (KL) property and some additional assumptions, and if the proximal and linearized type methods are applied to the problem, then the sequence $\{\mathbf{x}^{(k)}\}$ converges to a critical point.

We first verify that the objective function Φ defined in (5) is exactly a KL function, and then verify that Φ meets assumptions required in [31], hence showing the global convergence of PLiBCD-GS. Here we do not go into the detail of the KL

functions, instead, we only identify the KL property of Φ by using some properties provided in [31].

Proposition 4 (c.f. [31]): A function is KL if it is subanalytic or it is both lower semi-continuous and semi-algebraic.

We first have the following results.

Proposition 5: The data-fitting risk J_σ is analytic.

Proof: By the additivity of the analytic functions, it suffices to show the Cauchy loss and the Welsch loss are analytic. Because they are respectively the composition of logarithm and quadratic function, and composition of exponential and quadratic function, by the properties of the analytic functions, it follows that both of them are analytic functions. ■

Semi-algebraic property of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$.

Proposition 6: Both of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$ are semi-algebraic functions.

To investigate the semi-algebraic property of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$, some definitions and properties concerning the semi-algebraic sets and functions will be introduced briefly.

Definition 2 (Semi-algebraic sets and functions, see e.g., [47]): A set $C \subset \mathbb{R}^n$ is called semi-algebraic if there exists a finite number of real polynomials $p_{ij}, q_{ij} : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq s, 1 \leq j \leq t$ such that

$$C = \bigcup_{i=1}^{s} \bigcap_{j=1}^{t} \{\mathbf{x} \in \mathbb{R}^n \mid p_{ij}(\mathbf{x}) = 0,\ q_{ij}(\mathbf{x}) > 0\}.$$

A function $\phi : \mathbb{R}^n \to \mathbb{R} \setminus \{-\infty\}$ is called semi-algebraic if its graph

$$\operatorname{graph} \phi := \big\{(\mathbf{x}, t) \in \mathbb{R}^{n+1} \mid \phi(\mathbf{x}) = t\big\}$$

is a semi-algebraic set.

Operations that keep the semi-algebraic property of sets include: finite unions, finite intersections, complementation and Cartesian products. The following Tarski-Seidenberg theorem is also helpful for identifying semi-algebraic sets.

Theorem 2 (Tarski-Seidenberg theorem, c.f. [48]): Let C be a semi-algebraic set of $\mathbb{R}^n$. Then the image $\pi(C)$ is a semi-algebraic set if

1) $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ is the projection on the first $(n-1)$ coordinates.
2) $\pi : \mathbb{R}^n \to \mathbb{R}^m$ is a polynomial mapping.

Some common examples of semi-algebraic functions are listed in the following, which are provided in [31].

1) Real polynomial functions.
2) Characteristic functions of semi-algebraic sets.
3) Finite sums, product and composition of semi-algebraic functions.
4) supreme type functions: $\sup\{g(\mathbf{x}, \mathbf{y}) \mid \mathbf{y} \in C\}$ is semi-algebraic if g and C are semi-algebraic.

Based on the above definitions and properties, we are able to prove the semi-algebraic property of $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$.

Proof of Proposition 6: . We first prove the semi-algebraic property of $\|\cdot\|_*$. Recall that for any matrix $\mathbf{X} \in \mathbb{R}^{m_1 \times m_2}$, the dual formulation of $\|\mathbf{X}\|_*$ is given by

$$\|\mathbf{X}\|_* = \sup\{\langle \mathbf{X}, \mathbf{Y}\rangle \mid \|\mathbf{Y}\|_2 \leq 1\} = \sup\{\langle \mathbf{X}, \mathbf{Y}\rangle \mid \|\mathbf{Y}\|_2 = 1\}.$$

Since $\langle \mathbf{X}, \mathbf{Y}\rangle$ is a polynomial, it suffices to show the level set of the spectral norm $\{\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2} \mid \|\mathbf{Y}\|_2 = 1\}$ is a semi-algebraic set. To prove this, we need to first show that $\|\cdot\|_2$ is a semi-algebraic function. This can be verified by adopting the following representation

$$\|\mathbf{Y}\|_2 = \max\{\mathbf{z}_1^T \mathbf{Y} \mathbf{z}_2 \mid \mathbf{z}_i \in \mathbb{R}^{m_i}, \|\mathbf{z}_i\| = 1, i = 1, 2\},$$

and the semi-algebraic property of the set $\{\mathbf{z}_i \in \mathbb{R}^{m_i} \mid \|\mathbf{z}_1\| = 1\}$, $i = 1, 2$. Since $\|\cdot\|_2$ is semi-algebraic, by definition, the graph of the spectral norm

$$\operatorname{graph} \|\cdot\|_2 = \{(\mathbf{Y}, t) \mid \|\mathbf{Y}\|_2 = t\}$$

is a semi-algebraic set. Applying Tarski-Seidenberg theorem to $\operatorname{graph}\|\cdot\|_2$ thus yields the semi-algebraic property of the level set of $\|\cdot\|_2$. As a result, we have proved $\|\cdot\|_*$ is a semi-algebraic function.

We then show the semi-algebraic property for $\delta_{M_R}(\cdot)$. This can be verified by showing that the same property holds for set M_R. Define $\varphi : \mathbb{R}^{m_1 \times R} \times \mathbb{R}^{m_2 \times R} \to \mathbb{R}^{m_1 \times m_2}$ by $\varphi(\mathbf{U}, \mathbf{V}) = \mathbf{U}\mathbf{V}^T$. It is clear that φ is a polynomial, and the image of φ is exactly the set M_R. Therefore, the semi-algebraic property of M_R follows again from Tarski-Seidenberg theorem. The proof has been completed. ■

Propositions 1, 4 and 6 show that $\|\cdot\|_*$ and $\delta_{M_R}(\cdot)$ are KL functions. This in connection with Proposition 5 and the additivity of KL functions yields

Proposition 7: The function Φ defined in (5) is a KL function.

We then can prove Theorem 1.

Proof of Theorem 1: According to [31, Theorem 1] and the discussions in [31, Sec. 3.6], to prove Theorem 1, besides the KL property of Φ we also need to verify that 1) the regularization term is proper and lower semi-continuous; 2) the loss term is continuously differentiable; 3) the gradient of the loss term is Lipschitz continuous. 4) $\{\boldsymbol{\mathcal{X}}^{(k)}\}$ is bounded. Item 1) has been shown in Proposition 1. For 2) and 3), we have for any $\boldsymbol{\mathcal{X}}^+, \boldsymbol{\mathcal{X}} \in \mathbb{T}^N$

$$\begin{aligned} & \|\nabla J_\sigma(\textstyle\sum(\boldsymbol{\mathcal{X}}^+)) - \nabla J_\sigma(\sum(\boldsymbol{\mathcal{X}}))\|_F^2 \\ = & \textstyle\sum_{j=1}^{N} \|\nabla_{\mathcal{X}_j} J_\sigma(\sum(\boldsymbol{\mathcal{X}}^+)) - \nabla_{\mathcal{X}_j} J_\sigma(\sum(\boldsymbol{\mathcal{X}}))\|_F^2 \\ \leq & N\|\boldsymbol{\mathcal{A}}\|_2^4 \|\textstyle\sum(\boldsymbol{\mathcal{X}}^+) - \sum(\boldsymbol{\mathcal{X}})\|_F^2 \leq N^2 \|\boldsymbol{\mathcal{A}}\|_2^4 \|\boldsymbol{\mathcal{X}}^+ - \boldsymbol{\mathcal{X}}\|_F^2, \end{aligned}$$

where $\nabla_{\mathcal{X}_j}$ refers to the j-th partial derivative, and the first inequality is similar to the proof of Proposition 2 in Appendix A. Therefore, 2) and 3) are verified simultaneously. If $g(\cdot)$ takes the nuclear norm, then the boundedness of $\{\boldsymbol{\mathcal{X}}^{(k)}\}$ follows from the coercivity of $J(\cdot) + \sum \|\cdot\|_*$; if $g(\cdot)$ is the indicator function of the rank constraint, then the boundedness follows from the assumption of Theorem 1, and this gives 4). The above results in connection with Proposition 7, [31, Theorem 1] and [31, Sec. 3.6] gives Theorem 1. ■

Remark 5: The boundedness of $\{\boldsymbol{\mathcal{X}}^{(k)}\}$ is assumed when $g(\cdot)$ is the indicator function of the rank constraint, because $J(\cdot) + \sum \delta_{M_{R_j}}(\cdot)$ is not coercive due to the nullspace of $\boldsymbol{\mathcal{A}}$ may be nontrivial.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This

paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information; Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants; Flemish Government: FWO: PhD/Postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); IWT: PhD/-Postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017). Johan Suykens is a professor at KU Leuven, Belgium.

REFERENCES

[1] L. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[2] J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of eckart-young decomposition," *Psychometrika*, vol. 35, pp. 283–319, 1970.

[3] R. A. Harshman, "Foundations of the PARAFAC procedure: models and conditions for an" explanatory" multimodal factor analysis," *UCLA working papers in phonetics, 16 (1970), pp. 1–84.*

[4] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, pp. 1253–1278, 2000.

[5] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 208–220, 2013.

[6] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[7] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 517–520.

[8] X. Geng, K. Smith-Miles, Z.-H. Zhou, and L. Wang, "Face image modeling by multilinear subspace analysis with missing values," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 3, pp. 881–892, 2011.

[9] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. K. Suykens, "Learning with tensors: a framework based on convex optimization and spectral regularization," *Machine Learning*, vol. 94, pp. 303–351, 2014.

[10] M. Signoretto, R. Van de Plas, B. De Moor, and J. A. K. Suykens, "Tensor versus matrix completion: a comparison with application to spectral data," *Signal Processing Letters, IEEE*, vol. 18, no. 7, pp. 403–406, 2011.

[11] H. Tan, J. Feng, G. Feng, W. Wang, and Y.-J. Zhang, "Traffic volume data outlier recovery via tensor model," *Mathematical Problems in Engineering*, vol. vol. 2013, Article ID 164810, 2013.

[12] M. Signoretto, L. De Lathauwer, and J. A. K. Suykens, "Nuclear norms for tensors and their use for convex multilinear estimation," *Internal Report 10-186, ESAT-SISTA, KU Leuven, Leuven, Belgium.*, vol. 43, 2010.

[13] R. Tomioka, K. Hayashi, and H. Kashima, "Estimation of low-rank tensors via convex optimization," *arXiv preprint arXiv:1010.0789*, 2010.

[14] R. Tomioka and T. Suzuki, "Convex tensor decomposition via structured schatten norm regularization," in *Advances in Neural Information Processing Systems*, 2013, pp. 1331–1339.

[15] B. Huang, C. Mu, D. Goldfarb, and J. Wright, "Provable low-rank tensor recovery," 2014. [Online]. Available: http://www.optimization-online.org/DB_HTML/2014/02/4252.html

[16] H. Rauhut, R. Schneider, and Z. Stojanac, "Low rank tensor recovery via iterative hard thresholding," in *Proc. 10th International Conference on Sampling Theory and Applications*, 2013.

[17] L. Yang, Z.-H. Huang, and X. Shi, "A fixed point iterative method for low n-rank tensor pursuit," *Signal Processing, IEEE Transactions on*, vol. 61, no. 11, pp. 2952–2962, 2013.

[18] X. Zhang, D. Wang, Z. Zhou, and Y. Ma, "Simultaneous rectification and alignment via robust recovery of low-rank tensors," in *Advances in Neural Information Processing Systems*, 2013, pp. 1637–1645.

[19] Y. Li, J. Yan, Y. Zhou, and J. Yang, "Optimum subspace learning and error correction for tensors," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 790–803.

[20] H. Tan, B. Cheng, J. Feng, G. Feng, and Y. Zhang, "Tensor recovery via multi-linear augmented lagrange multiplier method," in *Image and Graphics (ICIG), 2011 Sixth International Conference on*. IEEE, 2011, pp. 141–146.

[21] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 1, pp. 225–253, 2014.

[22] P. J. Huber, *Robust statistics*. Springer, 2011.

[23] K. Inoue, K. Hara, and K. Urahama, "Robust multilinear principal component analysis," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 591–597.

[24] G. Shevlyakov, S. Morgenthaler, and A. Shurygin, "Redescending M-estimators," *Journal of Statistical Planning and Inference*, vol. 138, no. 10, pp. 2906 – 2917, 2008.

[25] I. Mizera and C. H. Müller, "Breakdown points of cauchy regression-scale estimators," *Statistics & Probability Letters*, vol. 57, no. 1, pp. 79 – 89, 2002.

[26] X. Wang, Y. Jiang, M. Huang, and H. Zhang, "Robust variable selection with exponential squared loss," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 632–643, 2013.

[27] Y. Feng, X. Huang, L. Shi, Y. Yang, and J. A. K. Suykens, "Learning with the maximum correntropy criterion induced losses for regression," *Internal Report 13-244, ESAT-SISTA, KU Leuven, Leuven, Belgium.*, 2013.

[28] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: properties and applications in non-gaussian signal processing," *Signal Processing, IEEE Transactions on*, vol. 55, no. 11, pp. 5286–5298, 2007.

[29] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1561–1576, 2011.

[30] Y. Feng, Y. Yang, and J. A. K. Suykens, "Robust gradient learning with applications," *Internal Report 14-62, ESAT-SISTA, KU Leuven, Leuven, Belgium.*, 2014.

[31] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, pp. 1–36, 2013.

[32] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, pp. 455–500, 2009.

[33] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, "Statistical performance of convex tensor decomposition." in *NIPS*, 2011, pp. 972–980.

[34] D. Kong, M. Zhang, and C. Ding, "Minimal shrinkage for noisy data recovery using Schatten-p norm objective," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 177–193.

[35] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[36] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *Neural Networks, IEEE Transactions on*, vol. 19, no. 1, pp. 18–39, 2008.

[37] D. Goldfarb and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization," *Foundations of Computational Mathematics*, vol. 11, no. 2, pp. 183–210, 2011.

[38] R. T. Rockafellar, R. J.-B. Wets, and M. Wets, *Variational Analysis*. Springer, 1998, vol. 317.

[39] J.-B. Hiriart-Urruty and H. Y. Le, "A variational approach of the rank function," *Top*, vol. 21, no. 2, pp. 207–240, 2013.

[40] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[41] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[42] A. A. Goldstein, "Convex programming in Hilbert space," *Bulletin of the American Mathematical Society*, vol. 70, no. 5, pp. 709–710, 1964.

[43] L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab v2.0, available online, january 2014." [Online]. Available: http://www.tensorlab.net/

[44] Y.-L. Chen, C.-T. Hsu, and H.-Y. Liao, "Simultaneous tensor decomposition and completion using factor priors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 3, pp. 577–591, 2014.

[45] M. Burger, L. He, and C.-B. Schönlieb, "Cahn-Hilliard inpainting and a generalization for grayvalue images," *SIAM Journal on Imaging Sciences*, vol. 2, no. 4, pp. 1129–1167, 2009.

[46] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[47] J. Bochnak, M. Coste, M.-F. Roy *et al.*, *Real algebraic geometry*. Springer Berlin, 1998, vol. 95.

[48] M. Coste, "An introduction to semi-algebraic geometry," *RAAG network school*, vol. 145, 2002.