

# Non-convex Penalty for Tensor Completion and Robust PCA

Tao Li Jinwen Ma

Department of Information Science, School of Mathematical Sciences and LMAM  
Peking University, Beijing, China

jwma@math.pku.edu.cn

## Abstract

In this paper, we propose a novel non-convex tensor rank surrogate function and a novel non-convex sparsity measure for tensor. The basic idea is to sidestep the bias of  $\ell_1$ -norm by introducing concavity. Furthermore, we employ the proposed non-convex penalties in tensor recovery problems such as tensor completion and tensor robust principal component analysis, which has various real applications such as image inpainting and denoising. Due to the concavity, the models are difficult to solve. To tackle this problem, we devise majorization minimization algorithms, which optimize upper bounds of original functions in each iteration, and every sub-problem is solved by alternating direction multiplier method. Finally, experimental results on natural images and hyperspectral images demonstrate the effectiveness and efficiency of the proposed methods.

## 1. Introduction

Tensors [16], or multi-way arrays, have been extensively used in computer vision [18, 35], signal processing and machine learning [8, 23]. For example, in image processing, a color image is a 3-order tensor of  $height \times weight \times channel$ , a multispectral image is a 3-order tensor of  $height \times weight \times band$ . Due to technical reasons, tensors in most applications are incomplete or polluted. Generically, recovering a tensor from corrupted observations is an inverse problem, which is ill-posed without prior knowledge. However, in real applications, entries in a tensor are usually highly correlated, which means a high-dimensional tensor is intrinsically determined by low-dimensional factors. Exploiting such low-dimensional structures makes it possible to restore tensors from limited or corrupted observations. Mathematically, this prior knowledge is equivalent to assume the tensors are low-rank.

In this work, we mainly consider two tensor recovery problems: tensor completion and tensor robust principal component analysis (TRPCA). The tensor completion problem is to estimate the missing values in tensors from par-

tially observed data, while TRPCA aims to decompose a tensor into a low-rank tensor and sparse tensor, see Figure 1. In image processing, tensor completion corresponds to image inpainting, and TRPCA corresponds to image denoising. In the case of matrix, both problems have been investigated thoroughly [6, 4, 5, 27]. Since the concept of a tensor is an extension of a matrix, it's natural to employ matrix recovery methods to tensors. Most matrix recovery methods are optimization based, penalizing rank surrogate function or/and certain sparsity measure. Similar methods have been developed for tensors. Tensor  $\ell_1$ -norm is often used as sparsity measure. However, the concept of tensor rank is far more complicated than matrix rank, thus there are various surrogate functions for tensor rank, such as the sum of nuclear norms (SNN) [18], tensor nuclear norm (TNN) [19] and twisted tensor nuclear norm (t-TNN) [10].

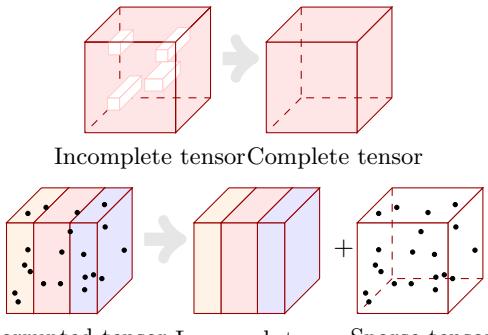


Figure 1. An illustration of tensor completion and TRPCA.

As in the matrix case, the choices of rank surrogate function and sparsity measure substantially influence the final results. The nuclear norm of a matrix is equivalent to the  $\ell_1$ -norm of its singular value. However, as indicated by Fan and Li [9],  $\ell_1$ -norm over-penalizes large entries of vectors. Smoothly clipped absolute deviation (SCAD) penalty [9] and min-max concave plus (MCP) penalty [30] were proposed as ideal penalty functions, and their superiority over  $\ell_1$ -norm has been demonstrated in [30, 9, 22, 33]. This inspires us that nuclear norm based tensor rank surrogate functions and  $\ell_1$ -norm based tensor sparsity measure

may suffer from similar problem. To alleviate such phenomena, we propose to use non-convex penalties (SCAD and MCP) instead of  $\ell_1$ -norm in TNN and tensor sparsity measure.

However, the introduction of non-convex penalties makes optimization problems even harder to solve. For example, TNN based TRPCA [19] is a convex optimization model, thus can be efficiently solved by alternating direction multiplier method (ADMM) [3]. Once we replace  $\ell_1$ -norm by SCAD or MCP, the problem is not convex anymore, and ADMM is not guaranteed to converge [2] in such circumstance. Therefore, we propose to apply majorization-minimization algorithm [11, 24] for solving the non-convex optimization problems. Based on the proposed non-convex tensor completion and TRPCA model and their corresponding MM algorithms, we conduct experiments on natural images and multispectral images to validate the effectiveness of the proposed methods.

The remainder of this paper is organized as follows. In Section 2, we review related work on tensor recovery and non-convex penalties. Section 3 introduces notations and preliminaries. We propose new tensor rank surrogate function and sparsity measure together with some theoretical properties in Section 4. Then, we formulate the non-convex models for TC and TRPCA, devise corresponding MM algorithms in Section 5 and Section 6 respectively. Extensive experimental results and analysis are reported in Section 7. Finally, we give concluding remarks in Section 8.

## 2. Related Work

**Tensor recovery.** For tensor completion, one seminal work is [18], in which SNN was proposed and three different algorithms for solving SNN based TC were devised. Zhao *et al.* proposed Bayesian CANDECOMP/PARAFAC (CP) tensor factorization model in [35]. The highlights of [35] include automatic rank determination property, full Bayesian treatment, and uncertainty quantification. Kilmer and Martin proposed a new tensor singular value decomposition (t-SVD) based on discrete Fourier transform for 3-order tensors in [15, 14]. The key point is that t-SVD offers an efficient way to define tensor nuclear norm (TNN), which has been extensively used in tensor recovery recently [32, 19, 36, 20]. Furthermore, Lu *et al.* proved exact recovery property of their proposed TRPCA model under certain suitable assumptions [19, 20].

**Non-convex penalties.** Wang and Zhang [26] developed a non-convex optimization model for low-rank matrix recovery problem. Cao *et al.* [7] applied folded-concave penalties in SNN, while Ji *et al.* [12] used log determinant penalty instead. Zhao *et al.* [34] proposed to use the product of nuclear norm instead of the sum of nuclear norms, which has a natural physical meaning. In addition, they also considered non-convex penalties such as SCAD and MCP. One

major difference between our work and [7, 34] is that our methods are based on t-SVD, while in [7, 34] they transform a tensor to matrices via simply unfolding. Recently, Jiang *et al.* [13] and Xu *et al.* [28] introduced non-convex penalties to TNN, but MCP or SCAD was not considered. Besides, our work not only improves the tensor rank surrogate function but also modifies the tensor sparsity measure.

## 3. Notations and Preliminaries

### 3.1. Notations

Throughout this paper, we use calligraphic letters to denote 3-way tensors, *e.g.*,  $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ . The  $(i, j, k)$ -th element of  $\mathcal{A}$  may be denoted by  $\mathcal{A}(i, j, k)$  or  $\mathcal{A}_{ijk}$  alternatively. The  $k$ -th frontal slice of  $\mathcal{A}$  is defined as  $\mathcal{A}(:, :, k)$ , which is an  $n_1 \times n_2$  matrix. For brevity, we use  $\mathcal{A}^{(k)}$  to denote  $\mathcal{A}(:, :, k)$ . The  $(i, j)$ -th tube of  $\mathcal{A}$  is defined as  $\mathcal{A}(i, j, :)$ , which is a vector of length  $n_3$ . The inner product of two 3-way tensors  $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$  is defined as  $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_k \text{Tr}((\mathcal{A}^{(k)})^* \mathcal{B}^{(k)})$ . We use  $|\mathcal{A}|$  to denote the tensor with  $(i, j, k)$ -th element equals to  $|\mathcal{A}_{ijk}|$ . Similar to vectors and matrices, we can also define various norms of tensors. We denote  $\ell_1$ -norm by  $\|\mathcal{A}\|_1 = \sum_{ijk} |\mathcal{A}_{ijk}|$ ,  $\ell_\infty$ -norm by  $\|\mathcal{A}\|_\infty = \max_{ijk} |\mathcal{A}_{ijk}|$  and Frobenius norm by  $\|\mathcal{A}\|_F = \sqrt{\sum_{ijk} |\mathcal{A}_{ijk}|^2}$ . We can transform a tensor to a matrix along the third dimension by `unfold`. Suppose  $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ , then  $\text{unfold}(\mathcal{A}) = [A^{(1)}; A^{(2)}; \dots; A^{(n_3)}]$ . The inverse transformation is denoted by `fold`, which transforms an  $(n_1 n_3) \times n_2$  matrix to an  $n_1 \times n_2 \times n_3$  tensor satisfying  $\text{fold}(\text{unfold}(\mathcal{A})) = \mathcal{A}$ .

Discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) are essential to the definitions in Section 3.2. We use the Matlab command `fft` and `ifft` to denote DFT and FFT applying to each tube of a 3-way tensor. We define  $\bar{\mathcal{A}} = \text{fft}(\mathcal{A}, [], 3)$ , and it is obvious that  $\mathcal{A} = \text{ifft}(\bar{\mathcal{A}}, [], 3)$ . Furthermore, we use  $\bar{\mathcal{A}} \in \mathbb{C}^{(n_1 n_2) \times (n_1 n_2)}$  to denote the block diagonal matrix whose blocks are frontal slices of  $\bar{\mathcal{A}}$ . With a little abuse of terminology, we say  $\mathcal{A}$  is in original domain and  $\bar{\mathcal{A}}$  (or equivalently  $\bar{\mathcal{A}}$ ) is in transformation domain or Fourier domain.

### 3.2. T-Product and T-SVD

**Definition 3.1 (T-product).** [15, 20] Suppose  $\mathcal{A} \in \mathbb{C}^{n_1 \times m \times n_3}$  and  $\mathcal{B} \in \mathbb{C}^{m \times n_2 \times n_3}$ , then the t-product  $\mathcal{A} * \mathcal{B} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$  is defined as

$$\mathcal{A} * \mathcal{B} = \text{ifft}(\text{fold}(\bar{\mathcal{A}} \bar{\mathcal{B}}), [], 3) \quad (1)$$

Note that Definition 3.1 is different from [15, 20] in form, but it is equivalent to the standard definitions. The reason why we choose this form is to avoid some cumbersome notations and better reveal the relationship between original domain and transformation domain. We may regard

$t$ -product as transforming the tensors by DFT, then multiplying corresponding frontal slices in Fourier domain, and finally transforming the result back to original domain by IDFT. It has been proved in [15, 20] that if  $\mathcal{C} = \mathcal{A} * \mathcal{B}$  then  $\overline{\mathcal{C}} = \overline{\mathcal{A}} \overline{\mathcal{B}}$ .

Before we introduce T-SVD, we need some further definitions, which are direct extensions of the corresponding definitions in the matrix case.

**Definition 3.2 (Conjugate transpose).** [15, 20] Suppose  $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ , the conjugate transpose of  $\mathcal{A}$  is denoted by  $\mathcal{A}^* \in \mathbb{C}^{n_2 \times n_1 \times n_3}$  whose first frontal slice equals to  $(\mathcal{A}^{(1)})^*$  and whose  $k$ -th frontal slice ( $k = 2, 3, \dots, n_3$ ) equals to  $(\mathcal{A}^{(n_3+2-k)})^*$ .

**Definition 3.3 (Identity tensor).** [15, 20] The identity tensor  $\mathcal{I} \in \mathbb{R}^{n \times n \times n_3}$  is the tensor whose first frontal slice is the  $n \times n$  identity matrix and whose other slices are zero matrices.

**Definition 3.4 (Orthogonal tensor).** [15, 20] A tensor  $\mathcal{Q} \in \mathbb{R}^{n \times n \times n_3}$  is said to be orthogonal if  $\mathcal{Q} * \mathcal{Q}^* = \mathcal{Q}^* * \mathcal{Q} = \mathcal{I}$ .

**Definition 3.5 (F-diagonal).** [15, 20] A tensor is said to be f-diagonal if its every frontal slice is a diagonal matrix.

**Theorem 3.1 (T-SVD).** [15, 20] Suppose  $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ . Then there exists  $\mathcal{U} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ ,  $\mathcal{V} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$  and  $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$  such that  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ . Furthermore,  $\mathcal{U}$  and  $\mathcal{V}$  are orthogonal, while  $\mathcal{S}$  is f-diagonal.

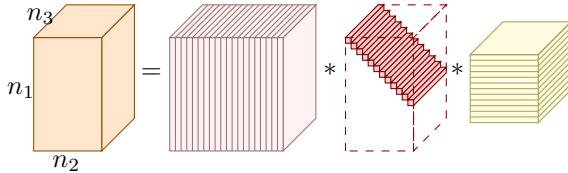


Figure 2. An illustration of the t-SVD of an  $n_1 \times n_2 \times n_3$  tensor.

An illustration of t-SVD is shown in Figure 2. Note that  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$  in original domain is equivalent to  $\overline{\mathcal{A}} = \overline{\mathcal{U}} \overline{\mathcal{S}} \overline{\mathcal{V}}^*$  in Fourier domain. Intuitively, we can obtain the T-SVD of  $\mathcal{A}$  by calculating SVD of each frontal slice  $\overline{\mathcal{A}}^{(k)}$  in frequency domain, i.e.,  $\overline{\mathcal{A}}^{(k)} = \overline{\mathcal{U}}^{(k)} \overline{\mathcal{S}}^{(k)} (\overline{\mathcal{V}}^{(k)})^*$ , then transforming  $\overline{\mathcal{U}}, \overline{\mathcal{S}}, \overline{\mathcal{V}}$  to original domain by IDFT. However, as indicated in [20], this method may result in complex entries due to non-uniqueness of matrix SVD. We omit the detailed algorithm for calculating T-SVD and refer to [20] for further discussions.

The concept of rank for tensors is very complicated. In fact, there are various different definitions of tensor rank [16, 8, 23]. The rank of a matrix is equivalent to the number of its non-zero singular values, and we often use nuclear norm (the sum of all singular values) as a surrogate function

for matrix rank. Intuitively, we may extend the concept of the nuclear norm to the tensor case, and the extension may be a reasonable surrogate for tensor rank.

**Definition 3.6 (Tensor nuclear norm).** [15, 20] Let  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$  be the t-SVD of  $\mathcal{A}$ , the nuclear norm of  $\mathcal{A}$  is defined as  $\|\mathcal{A}\|_* = \sum_i \mathcal{S}(i, i, 1)$ .

It has been proved in [20] that Definition 3.6 is the convex envelope of tensor average rank. Besides, the tensor nuclear norm is the dual norm of the tensor spectral norm, which is consistent with the matrix case. At first glance, the definition above may be a little amazing since only the first frontal slice of  $\mathcal{S}$  is used. According to the definition of IDFT, we have  $\mathcal{S}(i, i, 1) = \frac{1}{n_3} \sum_k \overline{\mathcal{S}}(i, i, k)$ . Thus, in the transformation domain, the tensor nuclear norm is equal to the sum of all singular values of all frontal slices up to a constant factor.

### 3.3. Non-convex Penalties: SCAD and MCP

As indicated in [9], an ideal penalty function should result in an estimator with three properties: unbiasedness, sparsity and continuity. Smoothly clipped absolute deviation (SCAD) was proposed in [9] to improve the properties of the  $\ell_1$  penalty, which does not satisfy the three properties simultaneously.

**Definition 3.7 (SCAD).** [9] For some  $\gamma > 1$  and  $\lambda > 0$ , the SCAD function is given by

$$\varphi_{\lambda, \gamma}^{\text{SCAD}}(t) = \begin{cases} \lambda|t| & \text{if } |t| \leq \lambda, \\ \frac{\gamma\lambda|t|-0.5(t^2+\lambda^2)}{\gamma-1} & \text{if } \lambda < |t| \leq \gamma\lambda, \\ \frac{\gamma+1}{2}\lambda^2 & \text{if } |t| > \gamma\lambda. \end{cases} \quad (2)$$

A continuous, nearly unbiased and accurate variable selection penalty called minimax concave penalty (MCP) was proposed in [30]. The precise definition is given as follows.

**Definition 3.8 (MCP).** [30] For some  $\gamma > 1$  and  $\lambda > 0$ , the MCP function is given by

$$\varphi_{\lambda, \gamma}^{\text{MCP}}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\gamma} & \text{if } |t| < \gamma\lambda, \\ \frac{\gamma\lambda^2}{2} & \text{if } |t| \geq \gamma\lambda. \end{cases} \quad (3)$$

It is well known that  $\ell_1$ -norm penalty over-penalizes large components. However, in SCAD and MCP, the penalty remains constant once the variable is larger than a threshold. Besides, we point out that as  $\gamma \rightarrow \infty$ , we have  $\varphi_{\lambda, \gamma}^{\text{SCAD}}(t) \rightarrow \lambda|t|$  and  $\varphi_{\lambda, \gamma}^{\text{MCP}}(t) \rightarrow \lambda|t|$  pointwisely. Last but not least, if we restrict  $t \geq 0$ , or equivalently view SCAD and MCP as functions of  $|t|$ , then they are concave functions. In the following, we use  $\varphi_{\lambda, \gamma}(t)$  to denote SCAD or MCP alternatively.

There are two parameters in SCAD and MCP:  $\lambda$  and  $\gamma$ . The effects of  $\lambda$  and  $\gamma$  can be intuitively understood by considering  $\varphi_{\lambda, \gamma}(t) \rightarrow \lambda|t|$ . Roughly speaking,  $\lambda$  controls the relative importance of the penalty, and  $\gamma$  controls how similar is  $\varphi_{\lambda, \gamma}(t)$  compared with  $\lambda|t|$ .

## 4. Theoretical Foundations

### 4.1. A Novel Tensor Sparsity Measure

The  $\ell_1$ -norm has been widely used as a sparsity measure in statistics, machine learning and computer vision. For tensors, the tensor  $\ell_1$ -norm plays a vital role in TRPCA [34, 19, 20]. However,  $\ell_1$ -norm penalty over-penalizes larger entries and may result in biased estimator. Therefore, we propose to use SCAD or MCP instead of the  $\ell_1$ -norm penalty. The novel tensor sparsity measure is defined as

$$\Phi_{\lambda,\gamma}(\mathcal{A}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \varphi_{\lambda,\gamma}(|\mathcal{A}_{ijk}|). \quad (4)$$

Here, we may set  $\varphi_{\lambda,\gamma}$  to be  $\varphi_{\lambda,\gamma}^{\text{SCAD}}$  or  $\varphi_{\lambda,\gamma}^{\text{MCP}}$ . It is easy to verify the following properties.

**Proposition 4.1.** For  $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ ,  $\Phi_{\lambda,\gamma}(\mathcal{A})$  satisfies:

- (i)  $\Phi_{\lambda,\gamma}(\mathcal{A}) \geq 0$  with the equality holds iff  $\mathcal{A} = 0$ ;
- (ii)  $\Phi_{\lambda,\gamma}(\mathcal{A})$  is concave with respect to  $|\mathcal{A}|$ ;
- (iii)  $\Phi_{\lambda,\gamma}(\mathcal{A})$  is increasing in  $\gamma$ ,  $\Phi_{\lambda,\gamma}(\mathcal{A}) \leq \lambda \|\mathcal{A}\|_1$  and  $\lim_{\gamma \rightarrow \infty} \Phi_{\lambda,\gamma}(\mathcal{A}) = \lambda \|\mathcal{A}\|_1$ .

### 4.2. A Novel Tensor Rank Penalty

In this part, we always assume  $\lambda = 1$ . Similar to tensor nuclear norm, we can apply SCAD or MCP to the singular values of a tensor. However, this may result in difficulty in optimization algorithms. Instead, we propose to apply penalty function to all singular values in Fourier domain. More precisely, suppose  $\mathcal{A}$  has t-SVD  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ , we define the  $\gamma$ -norm of  $\mathcal{A}$  as

$$\|\mathcal{A}\|_\gamma = \frac{1}{n_3} \sum_{i,k} \varphi_{1,\gamma}(\overline{\mathcal{S}}(i,i,k)). \quad (5)$$

The tensor  $\gamma$ -norm enjoys the following properties.

**Proposition 4.2.** For  $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ , suppose  $\mathcal{A}$  has t-SVD  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ , then  $\|\mathcal{A}\|_\gamma$  satisfies:

- (i)  $\|\mathcal{A}\|_\gamma \geq 0$  with equality holds iff  $\mathcal{A} = 0$ ;
- (ii)  $\|\mathcal{A}\|_\gamma$  is increasing in  $\gamma$ ,  $\|\mathcal{A}\|_\gamma \leq \|\mathcal{A}\|_*$  and  $\lim_{\gamma \rightarrow \infty} \|\mathcal{A}\|_\gamma = \|\mathcal{A}\|_*$ ;
- (iii)  $\|\mathcal{A}\|_\gamma$  is concave with respect to  $\{\overline{\mathcal{S}}(i,i,k)\}_{i,k}$ ;
- (iv)  $\|\mathcal{A}\|_\gamma$  is orthogonal invariant, i.e., for any orthogonal tensors  $\mathcal{P} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ ,  $\mathcal{Q} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ , we have  $\|\mathcal{P} * \mathcal{A} * \mathcal{Q}\|_\gamma = \|\mathcal{A}\|_\gamma$ .

### 4.3. Generalized Thresholding Operators

We will use majorizatoin minimization algorithm in Section 5.2 and 6.2. In this part, we derive some properties that are vital to MM algorithm based on the concavity of SCAD and MCP. As mentioned in Section 3.3, SCAP and MCP are continuous differentiable concave functions restricted

on  $[0, \infty)$ , thus we can bound  $\varphi_{\lambda,\gamma}(t)$  by its first-order Taylor expansion  $\varphi_{\lambda,\gamma}(t_0) + \varphi'_{\lambda,\gamma}(t_0)(t - t_0)$ . This observation leads to the following theorem.

**Theorem 4.3.** We can view  $\Phi_{\lambda,\mu}(\mathcal{X})$  as a function of  $|\mathcal{X}|$ , and  $\|\mathcal{X}\|_\gamma$  as a function of  $\{\overline{\mathcal{S}}(i,i,k)\}_{i,k}$ . For any  $\mathcal{X}^{\text{old}}$ , let

$$\begin{aligned} Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}) &= \Phi_{\lambda,\gamma}(\mathcal{X}^{\text{old}}) + \\ &\quad \sum_{i,j,k} \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)(|\mathcal{X}_{ijk}| - |\mathcal{X}_{ijk}^{\text{old}}|), \\ Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}}) &= \|\mathcal{X}^{\text{old}}\|_\gamma + \frac{1}{n_3} \sum_{i,k} \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}})(\overline{\mathcal{S}}_{iik} - \overline{\mathcal{S}}_{iik}^{\text{old}}), \end{aligned} \quad (6)$$

then  $\Phi_{\lambda,\gamma}(\mathcal{X}) \leq Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}})$ ,  $\|\mathcal{X}\|_\gamma \leq Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}})$ .

Due to the concavity of  $\Phi_{\lambda,\gamma}(\mathcal{X})$  and  $\|\mathcal{X}\|_\gamma$ , optimization problems involving  $\Phi_{\lambda,\gamma}(\mathcal{X})$  and  $\|\mathcal{X}\|_\gamma$  are generally extremely difficult to solve. However, optimizing upper bounds given in Theorem D.1 instead is relatively easy. It's well-known that soft thresholding operator  $\mathcal{T}_\lambda(z) = \text{sgn}(z)[|z| - \lambda]_+$  is the proximal operator of  $\ell_1$ -norm. In the following, we introduce generalized thresholding operators based on  $\mathcal{T}_\lambda$ , then derive the proximal operators of  $Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}})$  and  $Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}})$ .

**Definition 4.1 (Generalized soft thresholding).** Suppose  $\mathcal{X}, \mathcal{W} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ , the generalized soft thresholding operator is defined as

$$[\mathcal{T}_\mathcal{W}(\mathcal{X})]_{ijk} = \mathcal{T}_{\mathcal{W}_{ijk}}(\mathcal{X}_{ijk}). \quad (7)$$

**Theorem 4.4.** For  $\forall \mu > 0$ , let  $\mathcal{W}_{ijk} = \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)/\mu$ , then

$$\mathcal{T}_\mathcal{W}(\mathcal{Y}) = \arg \min_{\mathcal{X}} Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2. \quad (8)$$

**Definition 4.2 (Generalized t-SVT).** Suppose a 3-way tensor  $\mathcal{Y}$  has t-SVD  $\mathcal{Y} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ ,  $\mathcal{W}$  is a tensor with the same shape of  $\mathcal{Y}$ , the generalized tensor singular value thresholding operator is defined as

$$\mathcal{D}_\mathcal{W}(\mathcal{Y}) = \mathcal{U} * \tilde{\mathcal{S}} * \mathcal{V}^*, \quad (9)$$

where  $\tilde{\mathcal{S}} = \text{fft}(\mathcal{T}_\mathcal{W}(\mathcal{S}), [], 3)$ .

**Theorem 4.5.** For  $\forall \mu > 0$ , let  $\mathcal{W}_{ijk} = \delta_i^j \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}})/\mu$  where  $\delta_i^j$  is the Kronecker symbol, then

$$\mathcal{D}_\mathcal{W}(\mathcal{Y}) = \arg \min_{\mathcal{X}} Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2. \quad (10)$$

## 5. Non-convex Tensor Completion

### 5.1. Basic Model

Given a partially observed tensor  $\mathcal{O} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ , tensor completion task aims to recover the full tensor  $\mathcal{X}$  which coincides with  $\mathcal{O}$  in the observed positions. Suppose the observed positions are indexed by  $\Omega$ , i.e.,  $\Omega_{ijk} = 1$  denotes the  $(i, j, k)$ -th element is observed while  $\Omega_{ijk} = 0$  denotes

the  $(i, j, k)$ -th element is unknown. Based on low rank assumption, tensor completion can be modeled as

$$\min_{\mathcal{X}} \text{rank}(\mathcal{X}) \quad \text{s.t. } \mathcal{O}_{\Omega} = \mathcal{X}_{\Omega}. \quad (11)$$

Since the concept of rank is very complicated for tensors, many types of tensor rank or surrogate functions can be used in Equation 11. Here, we use the proposed tensor  $\gamma$ -norm,

$$\min_{\mathcal{X}} \|\mathcal{X}\|_{\gamma} \quad \text{s.t. } \mathcal{O}_{\Omega} = \mathcal{X}_{\Omega}. \quad (12)$$

Note that we can set  $\varphi_{1,\gamma}$  in Equation (12) to be SCAD or MCP. In the following we refer these non-convex low-rank tensor completion models as LRTC<sub>scad</sub> and LRTC<sub>mcp</sub> respectively.

## 5.2. MM Optimization

We apply majorization minimization algorithm to solve problem (12). Given  $\mathcal{X}^{\text{old}}$ , we minimize the upper bound of  $\|\mathcal{X}\|_{\gamma}$  given in Theorem D.1,

$$\min_{\mathcal{X}} Q_{\gamma}(\mathcal{X} | \mathcal{X}^{\text{old}}) \quad \text{s.t. } \mathcal{O}_{\Omega} = \mathcal{X}_{\Omega}. \quad (13)$$

Problem (13) is convex, thus we can use ADMM to solve it. Introducing auxiliary variable  $\mathcal{M}$  and let  $D$  be the feasible domain  $\{\mathcal{X} | \mathcal{O}_{\Omega} = \mathcal{X}_{\Omega}\}$ , then Equation (12) is equivalent to

$$\min_{\mathcal{X} \in D} Q_{\gamma}(\mathcal{M} | \mathcal{X}^{\text{old}}) \quad \text{s.t. } \mathcal{M} = \mathcal{X}. \quad (14)$$

The augmented Lagrangian function is given by

$$L(\mathcal{M}, \mathcal{X}, \mathcal{Y}) = Q_{\gamma}(\mathcal{M} | \mathcal{X}^{\text{old}}) + \langle \mathcal{M} - \mathcal{X}, \mathcal{Y} \rangle + \frac{\mu}{2} \|\mathcal{M} - \mathcal{X}\|_F^2. \quad (15)$$

According to the ADMM algorithm, we have the following iteration scheme:

$$\begin{aligned} \mathcal{M}_{k+1} &= \arg \min_{\mathcal{M}} Q(\mathcal{M} | \mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{M} - (\mathcal{X}_k - \frac{1}{\mu} \mathcal{Y}_k)\|_F^2, \\ \mathcal{X}_{k+1} &= \arg \min_{\mathcal{X} \in D} \|\mathcal{X} - (\mathcal{M}_{k+1} + \frac{1}{\mu} \mathcal{Y}_k)\|_F^2, \\ \mathcal{Y}_{k+1} &= \mathcal{Y}_k + \mu(\mathcal{M}_{k+1} - \mathcal{X}_{k+1}). \end{aligned} \quad (16)$$

The sub-problem of updating  $\mathcal{M}_{k+1}$  can be solved by generalized t-SVT as indicated in Theorem D.3. The sub-problem of updating  $\mathcal{X}_{k+1}$  has a closed-form solution:  $\mathcal{X}_{k+1} = (\mathcal{M}_{k+1} + \frac{1}{\mu} \mathcal{Y}_k) \circledast (1 - \Omega) + \mathcal{O} \circledast \Omega$ , where  $\circledast$  is elementwise product. Note that ADMM is the inner loop, after the ADMM converges we should update  $\mathcal{X}^{\text{old}}$  and repeat ADMM iterations again. Detailed algorithm is described in Algorithm 1.

## 6. Non-convex Tensor Robust PCA

### 6.1. Basic Model

Given a tensor  $\mathcal{X}$ , the goal of robust PCA is to decompose  $\mathcal{X}$  into two parts: low-rank tensor  $\mathcal{L}$  and sparse tensor  $\mathcal{E}$ . This problem can be formulated as

$$\min_{\mathcal{L}, \mathcal{E}} \text{rank}(\mathcal{L}) + \|\mathcal{E}\|_0 \quad \text{s.t. } \mathcal{L} + \mathcal{E} = \mathcal{X}. \quad (17)$$

---

**Algorithm 1** MM algorithm for non-convex low-rank tensor completion

---

**Input:**  $\Omega, \mathcal{O}$

**Hyper parameters:**  $\gamma, \mu_0, \rho, \mu_{\max}$

```

1: Initialize  $\mathcal{X}^0 = \mathcal{X}^{\text{old}}$ 
2: while not converged do
3:   Calculate  $\bar{\mathcal{S}}^{\text{old}}$  and set  $\mathcal{W}_{ijk}^t = \delta_i^j \varphi'_{1,\gamma}(\bar{\mathcal{S}}_{ikk}^{\text{old}}) / \mu$ 
4:   while not converged do
5:      $\mathcal{M}_{k+1}^t = \mathcal{D}_{\mathcal{W}^t}(\mathcal{X}_k^t - \frac{1}{\mu_k} \mathcal{Y}_k^t)$ 
6:      $\mathcal{X}_{k+1}^t = (\mathcal{M}_{k+1}^t + \frac{1}{\mu_k} \mathcal{Y}_k^t) \circledast (1 - \Omega) + \mathcal{O} \circledast \Omega$ 
7:      $\mathcal{Y}_{k+1}^t = \mathcal{Y}_k^t + \mu_k(\mathcal{M}_{k+1}^t - \mathcal{X}_{k+1}^t)$ 
8:      $\mu_{k+1} = \min(\rho \mu_k, \mu_{\max})$ 
9:   end while
10:  Update  $\mathcal{X}^{t+1}$  by the result of inner iteration
11:  Set  $\mathcal{X}^{\text{old}} = \mathcal{X}^{t+1}$ 
12: end while

```

---

Apply the proposed novel sparsity measure and tensor  $\gamma$ -norm, we obtain

$$\min_{\mathcal{L}, \mathcal{E}} \|\mathcal{L}\|_{\gamma_1} + \Phi_{\lambda, \gamma_2}(\mathcal{E}) \quad \text{s.t. } \mathcal{L} + \mathcal{E} = \mathcal{X}. \quad (18)$$

We may set  $\varphi_{\lambda, \gamma}$  to be SCAP or MCP in Equation (18), and refer them as TRPCA<sub>scad</sub> and TRPCA<sub>mcp</sub> respectively.

## 6.2. MM Optimization

We apply majorization minimization algorithm to solve problem (18). Given  $\mathcal{L}^{\text{old}}, \mathcal{E}^{\text{old}}$ , we minimize the upper bound of  $\|\mathcal{L}\|_{\gamma_1} + \Phi_{\lambda, \gamma_2}(\mathcal{E})$  given in Theorem D.1,

$$\min_{\mathcal{L}, \mathcal{E}} Q_{\gamma_1}(\mathcal{L} | \mathcal{L}^{\text{old}}) + Q_{\lambda, \gamma_2}(\mathcal{E} | \mathcal{E}^{\text{old}}) \quad \text{s.t. } \mathcal{L} + \mathcal{E} = \mathcal{X}. \quad (19)$$

The augmented Lagrangian function is

$$\begin{aligned} L(\mathcal{L}, \mathcal{E}, \mathcal{Y}) &= Q_{\gamma_1}(\mathcal{L} | \mathcal{L}^{\text{old}}) + Q_{\lambda, \gamma_2}(\mathcal{E} | \mathcal{E}^{\text{old}}) \\ &\quad + \langle \mathcal{Y}, \mathcal{L} + \mathcal{E} - \mathcal{X} \rangle + \frac{\mu}{2} \|\mathcal{L} + \mathcal{E} - \mathcal{X}\|_F^2. \end{aligned} \quad (20)$$

According to the ADMM algorithm, we may iterate variables as following:

$$\begin{aligned} \mathcal{L}_{k+1} &= \arg \min_{\mathcal{L}} Q_{\gamma_1}(\mathcal{L} | \mathcal{L}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{L} - (\mathcal{X} - (\mathcal{E}_k + \frac{1}{\mu} \mathcal{Y}_k))\|_F^2, \\ \mathcal{E}_{k+1} &= \arg \min_{\mathcal{E}} Q_{\lambda, \gamma_2}(\mathcal{E} | \mathcal{E}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{E} - (\mathcal{X} - (\mathcal{L}_{k+1} + \frac{1}{\mu} \mathcal{Y}_k))\|_F^2, \\ \mathcal{Y}_{k+1} &= \mathcal{Y}_k + \mu(\mathcal{L}_{k+1} + \mathcal{E}_{k+1} - \mathcal{X}). \end{aligned} \quad (21)$$

The sub-problem of updating  $\mathcal{L}$  and  $\mathcal{E}$  has closed-form solutions using Theorem D.2 and Theorem D.3. We describe the detailed algorithm in Algorithm 2.

## 7. Experiments

### 7.1. Datasets and Experimental Settings

We evaluate the effectiveness of the proposed non-convex tensor completion and tensor RPCA algorithms on



Figure 3. Tensor completion performance comparison on example images.

---

**Algorithm 2** MM algorithm for tensor RPCA

---

**Input:**  $\mathcal{X}$

**Hyper parameters:**  $\gamma_1, \gamma_2, \mu_0, \rho, \mu_{\max}$

- 1: Initialize  $\mathcal{L}^0, \mathcal{E}^0$  by other tensor RPCA algorithm
  - 2: Initialize  $\mathcal{Y}^0$  by random guess
  - 3: **while** not converged **do**
  - 4: Calculate t-SVD of  $\mathcal{L}^{\text{old}} = \mathcal{U} * \mathcal{S}^{\text{old}} * \mathcal{V}^*$
  - 5: Set  $\mathcal{Z}_{ijk}^t = \delta_i^j \varphi'_{1,\gamma_1}(\mathcal{S}_{ik})/\mu$
  - 6: Set  $\mathcal{W}_{ijk}^t = \varphi'_{\lambda,\gamma_2}(\mathcal{E}_{ijk}^{\text{old}})/\mu$
  - 7: **while** not converged **do**
  - 8:  $\mathcal{L}_{k+1}^t = \mathcal{D}_{\mathcal{Z}^t}(\mathcal{X} - (\mathcal{E}_k^t + \frac{1}{\mu_k} \mathcal{Y}_k^t))$
  - 9:  $\mathcal{E}_{k+1}^t = \mathcal{T}_{\mathcal{W}^t}(\mathcal{X} - (\mathcal{L}_{k+1}^t + \frac{1}{\mu_k} \mathcal{Y}_k^t))$
  - 10:  $\mathcal{Y}_{k+1}^t = \mathcal{Y}_k^t + \mu_k(\mathcal{L}_{k+1}^t + \mathcal{E}_{k+1}^t - \mathcal{X})$
  - 11:  $\mu_{k+1} = \min(\rho\mu_k, \mu_{\max})$
  - 12: **end while**
  - 13: Update  $\mathcal{L}^{t+1}, \mathcal{E}^{t+1}$  by the result of inner iteration
  - 14: Set  $\mathcal{L}^{\text{old}} = \mathcal{L}^{t+1}, \mathcal{E}^{\text{old}} = \mathcal{E}^{t+1}$
  - 15: **end while**
- 

Berkeley Segmentation 500 Dataset (BSD 500) [1]<sup>1</sup> and Natural Scenes 2002 Dataset (NS 2002) [21]<sup>2</sup>. Berkeley Segmentation 500 Dataset consists of 500 natural images, and Natural Scenes 2002 Dataset contains 8 hyperspectral images with 31 bands sampled from 410nm to 710nm at 10nm intervals. All the hyperspectral images are downsampled by factor 2. We employ Mean Square

<sup>1</sup><https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html#bsds500>

<sup>2</sup>[https://personalpages.manchester.ac.uk/staff/d.h.foster/Hyperspectral\\_images\\_of\\_natural\\_scenes\\_02.html](https://personalpages.manchester.ac.uk/staff/d.h.foster/Hyperspectral_images_of_natural_scenes_02.html)

Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Feature SIMilarity (FSIM) [31], Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS)[25] and Spectral Angle Mapper (SAM) [17, 29] as performance evaluation indexes. Smaller MSE, ERGAS, SAM and larger PSNR, FSIM indicates the result is better.

There are some practical issues to clarify about Algorithm 1 and Algorithm 2. First, the hyper-parameters  $\mu_0, \rho, \mu_{\max}$  are introduced to accelerate the convergence speed. The inner ADMM iteration is always convergent regardless of the settings of these parameters, but the speed of convergence is different. In practice, we find setting  $\mu_0 = 1, \rho = 1.1, \mu_{\max} = 1e10$  results in fast convergence. Second, the initialization of  $\mathcal{X}^0$  and  $\mathcal{L}, \mathcal{E}$  is very important, since a good starting position usually leads to better final result in non-convex optimization problems. We suggest initializing  $\mathcal{X}^0$  or  $\mathcal{L}^0, \mathcal{E}^0$  by other tensor completion or tensor RPCA methods. Last but not least, it usually takes a long time for the outer iteration to converge. In practice, it's not necessary to wait for convergence. Instead, we can iterate the outer loop for fixed times.

## 7.2. Tensor Completion Experiments

We conduct tensor completion experiments on BSD 500 and NS 2002 to test the performances of  $\text{LRTC}_{\text{mcp}}$  and  $\text{LRTC}_{\text{scad}}$ . For comparison, we also consider five competing tensor completion methods: Bayesian CP Factorization (FBCP) [35], Simple Low-Rank Tensor Completion (SiLRTC) [18], High Accuracy Low-Rank Tensor Completion (HaLRTC) [18], tensor-SVD based method (t-SVD) [32], twist Tensor Nuclear Norm based method (t-TNN) [10].

**Natural image inpainting.** We randomly select 200 images in BSD 500 for evaluation. For each image, pixels are randomly sampled with sampling rate ranging from

Method	20%			40%			60%			80%			time (s)
	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	PSNR	SSIM	FSIM	
SiLRTC [18]	23.59	0.798	0.822	27.987	0.899	0.915	32.24	0.951	0.964	37.47	0.977	0.988	19.95
HaLRTC [18]	23.82	0.797	0.828	28.39	0.902	0.920	33.038	0.953	0.968	39.27	0.978	0.991	31.32
FBCP [35]	24.08	0.668	0.794	26.40	0.753	0.837	27.35	0.799	0.857	27.71	0.82	0.865	103.68
t-SVD [32]	24.13	0.764	0.835	29.703	0.893	0.931	36.03	0.950	0.977	45.04	0.969	0.992	33.47
t-TNN [10]	25.30	0.841	0.864	30.50	0.923	0.943	36.27	0.952	0.978	44.14	0.967	0.991	3.03
LRTC <sub>mcp</sub>	25.70	0.845	0.869	31.06	0.927	0.946	36.87	0.959	0.980	45.46	0.973	0.993	3.79
LRTC <sub>scad</sub>	25.70	0.844	0.869	31.04	0.926	0.946	36.84	0.959	0.980	45.45	0.973	0.993	3.83

Table 1. Tensor completion performances evaluation on natural images under varying sampling rates.

Method	20%			40%			60%			80%			time (s)
	PSNR	MSE	ERGAS	PSNR	MSE	ERGAS	PSNR	MSE	ERGAS	PSNR	MSE	ERGAS	
SiLRTC [18]	41.71	4.70	30.912	45.46	1.95	21.524	49.14	0.827	14.412	52.86	0.354	10.341	42.21
HaLRTC [18]	42.11	4.53	29.626	45.95	1.90	20.556	49.79	0.801	13.514	53.67	0.342	9.660	53.11
FBCP [35]	37.09	14.47	52.931	43.25	3.85	29.318	46.00	2.225	22.344	46.67	2.011	20.688	210.18
t-SVD [32]	41.64	5.10	31.835	45.52	2.12	22.142	49.42	0.886	14.685	53.49	0.365	10.157	224.21
t-TNN [10]	42.46	3.81	28.702	46.07	1.60	20.135	49.82	0.667	13.272	53.61	0.290	9.515	47.29
LRTC <sub>mcp</sub>	42.91	3.58	27.799	46.75	1.51	19.684	50.47	6.65	13.293	54.11	3.16	9.642	70.95
LRTC <sub>scad</sub>	42.91	3.58	27.804	46.76	1.51	19.684	50.46	6.66	13.298	54.11	3.16	9.642	71.14

Table 2. Tensor completion performances evaluation on hyperspectral images under varying sampling rates. The unit is  $10^{-4}$  for MSE.

20% to 80%. For our LRTC<sub>mcp</sub> and LRTC<sub>scad</sub> models, we set  $\gamma = 25$ , and use the result of t-TNN as initialization. To alleviate redundant computations, we apply the one-step LLA strategy [37, 26], *i.e.*, we run the outer loop only once instead of waiting for convergence. The average performances over selected 200 images under different sampling rates are summarized in Table 1. From this table, we can see that our proposed LRTC<sub>mcp</sub> and LRTC<sub>scad</sub> outperform other competing methods in terms of all performance evaluation indexes. As for efficiency, the proposed methods are significantly faster than FBCP, SiLRTC, HaLRTC, and t-SVD. Since the proposed methods are initialized by t-TNN, the running time is always slightly longer than t-TNN. However, the performances are improved by only one MM iteration and the extra running times are marginal. Therefore, we claim that it's necessary to introduce non-convexity in tensor completion task. These observations demonstrate LRTC<sub>mcp</sub> and LRTC<sub>scad</sub> are both effective and efficient. We also give visual comparisons in Figure 3.

**Hyperspectral image inpainting.** We use all 8 hyperspectral images in this experiment. For each hyperspectral image, we randomly sample its elements with sampling rate ranging from 0.2 to 0.8. Since the sizes of hyperspectral images are relatively large, we run the outer loop in Algorithm 1 for 10 iterations based on t-TNN initialization. The performance comparison are shown in Table 2. We have similar observations as in the natural image case: the results obtained by LRTC<sub>mcp</sub> and LRTC<sub>scad</sub> have lower MSE, ERGAS, and higher PSNR, indicating that the proposed methods outperform competing methods.

$p_n$	Index	RPCA [27]	TRPCA [19]	TRPCA <sub>mcp</sub>	TRPCA <sub>scad</sub>
		MSE	PSNR	ERGAS	SAM
0.1	MSE	20.113	3.680	3.332	3.329
	PSNR	41.89	46.11	46.57	46.57
	ERGAS	23.224	13.203	12.898	12.893
	SAM	0.0887	0.0894	0.0880	0.0879
0.2	MSE	21.298	4.353	3.837	3.834
	PSNR	41.31	45.63	46.08	46.09
	ERGAS	24.374	14.473	13.703	13.706
	SAM	0.1191	0.1003	0.0961	0.0961
0.3	MSE	26.430	6.637	4.731	4.726
	PSNR	39.68	44.13	45.28	45.27
	ERGAS	28.956	18.668	15.177	15.205
	SAM	0.1710	0.1291	0.1076	0.1079
0.4	MSE	49.858	28.275	6.979	6.964
	PSNR	36.13	38.51	43.16	43.11
	ERGAS	49.877	50.354	20.126	20.298
	SAM	0.2602	0.2483	0.1428	0.1440

Table 3. Tensor Robust Principal Component Analysis performances evaluation on hyperspectral images.

### 7.3. Tensor RPCA Experiments

We compare our proposed TRPCA<sub>mcp</sub> and TRPCA<sub>scad</sub> with matrix RPCA [27, 5] and TNN based TRPCA [19, 20] on both natural images and multispectral images. To apply matrix RPCA in tensor RPCA task, we simply apply matrix RPCA to each frontal slices of the corrupted tensor.

**Natural image restoration by Tensor RPCA.** We first test TRPCA<sub>mcp</sub> and TRPCA<sub>scad</sub> on BSD 500. Each image is corrupted by salt-and-pepper noise with probability  $p_n = 0.1$ . We set  $\gamma_1 = \gamma_2 = 20$  and run the outer loop of Algorithm 2 for 10 iterations. Performance comparison on randomly selected 50 images are shown in Figure 5. From this figure, we have following observations. First, the results of TRPCA, TRPCA<sub>mcp</sub> and TRPCA<sub>scad</sub> are significantly better than results of RPCA. This indicates that con-

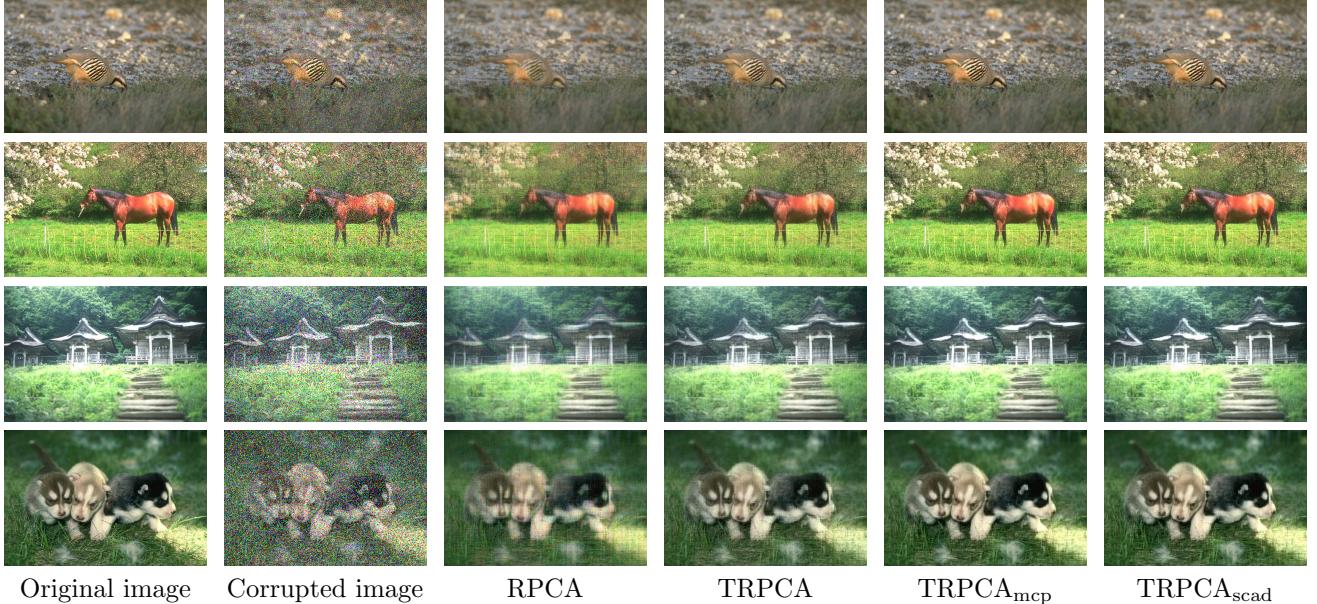


Figure 4. Tensor RPCA performance comparison on example images. From top to bottom:  $p_n = 0.1, 0.2, 0.3, 0.4$ .

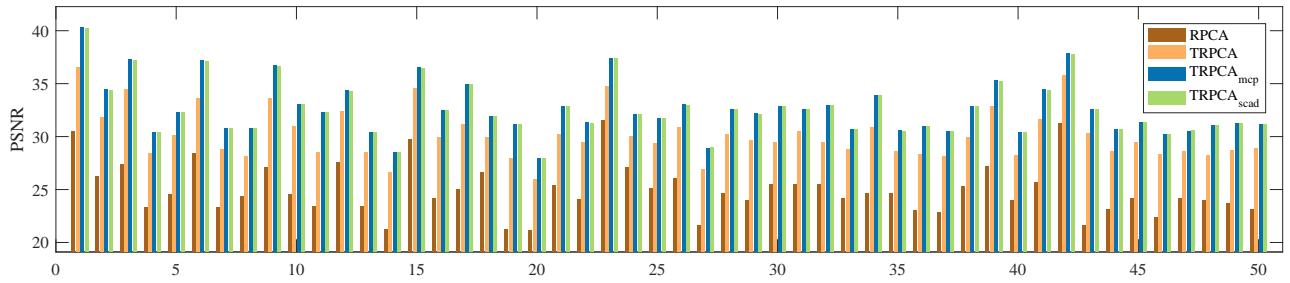


Figure 5. Comparison of PSNR values obtained by RPCA, TRPCA, TRPCA<sub>mcp</sub>,TRPCA<sub>scad</sub> on randomly selected 50 images.

sidering tensor structure helps to improve recovery quality compared to consider each channel individually. Second, TRPCA<sub>mcp</sub> and TRPCA<sub>scad</sub> obtained better performance than TRPCA, which means introducing concavity to Tensor RPCA tasks are necessary. Third, the PSNR values of TRPCA<sub>mcp</sub> and TRPCA<sub>scad</sub> are very similar, indicating the final result is not sensitive to the choice of non-convex penalty. We also give visual comparisons in Figure 4. Note that for the noise proportion ranging from 0.1 to 0.4 in Figure 4.

**Hyperspectral image restoration.** In this part, we test the proposed models on NS 2002. We add random noise to each hyperspectral image with probability  $p_n$  ranging from 0.1 to 0.4. Here, the noise is uniformly distributed in  $[0, 0.1 * M]$  where  $M$  is the maximum absolute value of the original image. We set  $\gamma_1 = \gamma_2 = 50$ , and run the outer loop for 10 iterations. The results of TRPCA are used as initialization for the proposed methods. We employ MSE, PSNR, ERGAS, and SAM as quality indexes. The results are reported in Table 3. It's easy to see that

TRPCA<sub>mcp</sub> and TRPCA<sub>scad</sub> outperform competing methods in terms of all quality indexes. Specifically, we note that when  $p_n = 0.4$ , *i.e.*, the noise proportion is rather large, the proposed methods improve the results of TRPCA significantly. In this circumstance, the sparse assumption on noise may not hold. Although RPCA and TRPCA have nice exact recovery property under certain conditions, these conditions are rather strict and sometimes not satisfied. However, the proposed methods still recover the images successfully.

## 8. Conclusions

In this paper, we have presented a new non-convex tensor rank surrogate function and a new non-convex sparsity measure. Then, we have analyzed some theoretical properties of the proposed penalties. In particular, we applied the non-convex penalties in tensor completion and tensor robust principal component analysis tasks, and devised optimization algorithms based on majorization minimization. Experimental results on natural images and hyperspectral images substantiate the proposed methods outperform com-

peting methods.

## References

- [1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, May 2011.
- [2] D. P. Bertsekas and A. Scientific. *Convex optimization algorithms*. Athena Scientific Belmont, 2015.
- [3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [4] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- [5] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
- [6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.
- [7] W. Cao, Y. Wang, C. Yang, X. Chang, Z. Han, and Z. Xu. Folded-concave penalization approaches to tensor completion. *Neurocomputing*, 152:261–273, 2015.
- [8] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multi-way component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163, 2015.
- [9] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- [10] W. Hu, D. Tao, W. Zhang, Y. Xie, and Y. Yang. A new low-rank tensor model for video completion. *arXiv:1509.02027*, 2015.
- [11] D. R. Hunter and R. Li. Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617, 2005.
- [12] T.-Y. Ji, T.-Z. Huang, X.-L. Zhao, T.-H. Ma, and L.-J. Deng. A non-convex tensor rank approximation for tensor completion. *Applied Mathematical Modelling*, 48:410–422, 2017.
- [13] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, and L.-J. Deng. A novel nonconvex approach to recover the low-tubal-rank tensor data: when t-svd meets pssv. *arXiv:1712.05870*, 2017.
- [14] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM Journal on Matrix Analysis and Applications*, 34(1):148–172, 2013.
- [15] M. E. Kilmer and C. D. Martin. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658, 2011.
- [16] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [17] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz. The spectral image processing system (sips)interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3):145–163, 1993.
- [18] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE TPAMI*, 35(1):208–220, 2013.
- [19] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *IEEE CVPR*, 2016.
- [20] C. Lu, J. Feng, W. Liu, Z. Lin, S. Yan, et al. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE TPAMI*, 2019.
- [21] S. M. Nascimento, F. P. Ferreira, and D. H. Foster. Statistics of spatial cone-excitation ratios in natural scenes. *JOSA A*, 19(8):1484–1490, 2002.
- [22] J. Shi, X. Ren, G. Dai, J. Wang, and Z. Zhang. A non-convex relaxation approach to sparse dictionary learning. In *IEEE CVPR*, 2011.
- [23] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE TSP*, 65(13):3551–3582, 2017.
- [24] Y. Sun, P. Babu, and D. P. Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE TSP*, 65(3):794–816, 2017.
- [25] L. Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002.
- [26] S. Wang, D. Liu, and Z. Zhang. Nonconvex relaxation approaches to robust matrix recovery. In *IJCAI*, 2013.
- [27] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NeurIPS*, 2009.
- [28] W.-H. Xu, X.-L. Zhao, T.-Y. Ji, J.-Q. Miao, T.-H. Ma, S. Wang, and T.-Z. Huang. Laplace function based nonconvex surrogate for low-rank tensor completion. *Signal Processing: Image Communication*, 2018.
- [29] R. H. Yuhas, A. F. Goetz, and J. W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. 1992.
- [30] C.-H. Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.
- [31] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011.
- [32] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-svd. In *IEEE CVPR*, 2014.
- [33] Z. Zhang and B. Tu. Nonconvex penalization using laplace exponents and concave conjugates. In *NeurIPS*, 2012.
- [34] Q. Zhao, D. Meng, X. Kong, Q. Xie, W. Cao, Y. Wang, and Z. Xu. A novel sparsity measure for tensor recovery. In *IEEE ICCV*, 2015.
- [35] Q. Zhao, L. Zhang, and A. Cichocki. Bayesian cp factorization of incomplete tensors with automatic rank determination. *IEEE TPAMI*, 37(9):1751–1763, 2015.
- [36] P. Zhou and J. Feng. Outlier-robust tensor pca. In *IEEE CVPR*, 2017.
- [37] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*,

36(4):1509, 2008.

## Appendix

### A. SCAD and MCP

**Definition A.1 (SCAD).** For some  $\gamma > 1$  and  $\lambda > 0$ , the SCAD function is given by

$$\varphi_{\lambda,\gamma}^{\text{SCAD}}(t) = \begin{cases} \lambda|t| & \text{if } |t| \leq \lambda, \\ \frac{\gamma\lambda|t|-0.5(t^2+\lambda^2)}{\gamma-1} & \text{if } \lambda < |t| \leq \gamma\lambda, \\ \frac{\gamma+1}{2}\lambda^2 & \text{if } |t| > \gamma\lambda. \end{cases} \quad (22)$$

**Definition A.2 (MCP).** For some  $\gamma > 1$  and  $\lambda > 0$ , the MCP function is given by

$$\varphi_{\lambda,\gamma}^{\text{MCP}}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\gamma} & \text{if } |t| < \gamma\lambda, \\ \frac{\gamma\lambda^2}{2} & \text{if } |t| \geq \gamma\lambda. \end{cases} \quad (23)$$

We use  $\varphi_{\lambda,\gamma}(t)$  to denote  $\varphi_{\lambda,\gamma}^{\text{SCAD}}(t)$  or  $\varphi_{\lambda,\gamma}^{\text{MCP}}(t)$  alternatively, then we have the following properties:

**Proposition A.1.** (i)  $\varphi_{\lambda,\gamma}(t) \geq 0$  and  $\varphi_{\lambda,\gamma}(t) = 0$  if and only if  $t = 0$ .

(ii) For fixed  $t$  and  $\lambda$ ,  $\varphi_{\lambda,\gamma}(t)$  is increasing in  $\gamma$ .

(iii) As  $\gamma \rightarrow \infty$ ,  $\varphi_{\lambda,\gamma}(t) \rightarrow \lambda|t|$ .

(iv) When restricted on  $t \geq 0$ ,  $\varphi_{\lambda,\gamma}(t)$  is concave.

*Proof.* Note that  $\varphi_{\lambda,\gamma}(t)$  is an even function, thus we only consider the case  $t \geq 0$ .

(i) For SCAD, if  $0 \leq t \leq \lambda$  or  $t > \gamma\lambda$ ,  $\varphi_{\lambda,\gamma}^{\text{SCAD}}(t) \geq 0$  since  $\lambda > 0$  and  $\gamma > 0$ . If  $\lambda < t \leq \gamma\lambda$ , the minimum of  $\varphi_{\lambda,\gamma}^{\text{SCAD}}(t)$  is attained at  $\lambda$ , which equals to  $\lambda^2$ . Therefore,  $\varphi_{\lambda,\gamma}^{\text{SCAD}}(t) \geq 0$ .

For MCP, if  $t > \lambda\gamma$ ,  $\varphi_{\lambda,\gamma}^{\text{MCP}}(t) \geq 0$  since  $\lambda > 0, \gamma > 0$ . If  $t \leq \gamma\lambda$ , the minimum of  $\varphi_{\lambda,\gamma}^{\text{MCP}}(t)$  is attained at 0, which equals to 0. Therefore,  $\varphi_{\lambda,\gamma}^{\text{MCP}}(t) \geq 0$ .

Obviously,  $\varphi_{\lambda,\gamma}(t) = 0$  if and only if  $t = 0$ .

(ii) Suppose we have  $\gamma_2 > \gamma_1 > 1$ .

For SCAD, increasing  $\gamma$  has no influence for  $0 \leq t \leq \lambda$ . If  $\lambda < t < \gamma_1\lambda$ , then  $\lambda < t < \gamma_2\lambda$ . Note that

$$\frac{d}{d\gamma} \frac{\gamma\lambda t - 0.5(t^2 + \lambda^2)}{\gamma - 1} = \frac{(t - \lambda)^2}{2(\gamma - 1)^2} \geq 0.$$

Therefore,  $\varphi_{\lambda,\gamma_2}^{\text{SCAD}}(t) \geq \varphi_{\lambda,\gamma_1}^{\text{SCAD}}(t)$ . If  $t > \gamma_1\lambda$ , then we have two cases:  $t > \gamma_2\lambda$  or  $\lambda < t \leq \gamma_2\lambda$ . If  $t > \gamma_2\lambda$ , then  $\varphi_{\lambda,\gamma_2}^{\text{SCAD}}(t) = \frac{\gamma_2+1}{2}\lambda^2 > \frac{\gamma_1+1}{2}\lambda^2 = \varphi_{\lambda,\gamma_1}^{\text{SCAD}}(t)$ . If  $\lambda < t \leq \gamma_2\lambda$ , then

$$\varphi_{\lambda,\gamma_2}^{\text{SCAD}}(t) - \varphi_{\lambda,\gamma_1}^{\text{SCAD}}(t) = \frac{\gamma_2\lambda t - 0.5(t^2 + \lambda^2)}{\gamma_2 - 1} - \frac{\gamma_1+1}{2}\lambda^2 = \frac{-t^2 + 2\gamma_2\lambda t - \lambda^2 - (\gamma_1+1)(\gamma_2-1)\lambda^2}{2(\gamma_2-1)}.$$

Note that in this case  $\gamma_1\lambda \leq t \leq \gamma_2\lambda$ , the above equation attains its minimum at  $\gamma_1\lambda$ , thus

$$\varphi_{\lambda,\gamma_2}^{\text{SCAD}}(t) - \varphi_{\lambda,\gamma_1}^{\text{SCAD}}(t) \geq \frac{(\gamma_2 - \gamma_1)(\gamma_1 - 1)\lambda^2}{2(\gamma_2 - 1)} > 0.$$

The last inequality follows from the condition  $\gamma_2 > \gamma_1 > 1$ . Therefore, we conclude that  $\varphi_{\lambda,\gamma}^{\text{SCAD}}(t)$  is increasing with respect to  $\gamma$ .

For MCP, if  $0 \leq t < \gamma_1\lambda$ , then  $0 \leq t < \gamma_2\lambda$ , thus  $\varphi_{\lambda,\gamma_2}^{\text{MCP}}(t) = \lambda t - \frac{t^2}{2\gamma_2} > \lambda t - \frac{t^2}{2\gamma_1} = \varphi_{\lambda,\gamma_1}^{\text{MCP}}(t)$ . If  $t \geq \gamma_1\lambda$ , we have two cases:  $t \geq \gamma_2\lambda$  or  $0 < t < \gamma_2\lambda$ . If  $t \geq \gamma_2\lambda$ , then  $\varphi_{\lambda,\gamma_2}^{\text{MCP}}(t) = \frac{\gamma_2\lambda^2}{2} > \frac{\gamma_1\lambda^2}{2} = \varphi_{\lambda,\gamma_1}^{\text{MCP}}(t)$ . If  $0 \leq t \leq \gamma_2\lambda$ , then

$$\varphi_{\lambda,\gamma_2}^{\text{MCP}}(t) - \varphi_{\lambda,\gamma_1}^{\text{MCP}}(t) = \lambda t - \frac{t^2}{2\gamma_2} - \frac{\gamma_1\lambda^2}{2} > \frac{(\gamma_2 - \gamma_1)\gamma_1\lambda^2}{2\gamma_2} > 0.$$

Therefore, we conclude that  $\varphi_{\lambda,\gamma}^{\text{MCP}}(t)$  is increasing with respect to  $\gamma$ .

(iii) Note that as  $\gamma \rightarrow \infty$ ,

$$\frac{\gamma\lambda|t| - 0.5(t^2 + \lambda^2)}{\gamma - 1} \rightarrow \lambda|t| \quad , \quad \lambda|t| - \frac{t^2}{2\gamma} \rightarrow \lambda|t|.$$

The result follows easily.

(iv) Consider the second order derivative of  $\varphi_{\lambda,\gamma}(t)$ .

For SCAD,

$$\frac{d^2}{dt^2}\varphi_{\lambda,\gamma}^{\text{SCAD}}(t) = \begin{cases} 0, & \text{if } 0 \leq t \leq \lambda \text{ or } t > \gamma\lambda, \\ -\frac{1}{\gamma-1}, & \text{if } t \leq \gamma\lambda. \end{cases}$$

For MCP,

$$\frac{d^2}{dt^2}\varphi_{\lambda,\gamma}^{\text{MCP}}(t) = \begin{cases} -\frac{1}{\gamma}, & \text{if } 0 \leq t \leq \gamma\lambda, \\ 0, & \text{if } t \geq \gamma\lambda. \end{cases}$$

Therefore,  $\varphi_{\lambda,\gamma}(t)$  is concave over  $[0, \infty)$  since the second order derivative is non-positive. Besides, by the definition of concave function we know that  $\varphi_{\lambda,\gamma}(t) \leq \varphi_{\lambda,\gamma}(s) + \varphi'_{\lambda,\gamma}(s)(t - s)$  for  $s, t \geq 0$ .

□

## B. A Novel Tensor Sparsity Measure

Recall that the novel tensor sparsity measure is defined as

$$\Phi_{\lambda,\gamma}(\mathcal{A}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \varphi_{\lambda,\gamma}(\mathcal{A}_{ijk}). \quad (24)$$

Here, we may set  $\varphi_{\lambda,\gamma}$  to be  $\varphi_{\lambda,\gamma}^{\text{SCAD}}$  or  $\varphi_{\lambda,\gamma}^{\text{MCP}}$ . We have the following proposition:

**Proposition B.1.** For  $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ ,  $\Phi_{\lambda,\gamma}(\mathcal{A})$  satisfies:

- (i)  $\Phi_{\lambda,\gamma}(\mathcal{A}) \geq 0$  with the equality holds iff  $\mathcal{A} = 0$ ;
- (ii)  $\Phi_{\lambda,\gamma}(\mathcal{A})$  is concave with respect to  $|\mathcal{A}|$ ;
- (iii)  $\Phi_{\lambda,\gamma}(\mathcal{A})$  is increasing in  $\gamma$ ,  $\Phi_{\lambda,\gamma}(\mathcal{A}) \leq \lambda\|\mathcal{A}\|_1$  and  $\lim_{\gamma \rightarrow \infty} \Phi_{\lambda,\gamma}(\mathcal{A}) = \lambda\|\mathcal{A}\|_1$ .

*Proof.* Note that  $\Phi_{\lambda,\gamma}(\mathcal{A})$  is separable with respect to each entry of  $\mathcal{A}$ . Thus, applying the related properties of  $\varphi_{\lambda,\gamma}(t)$  to each entry  $\mathcal{A}_{ijk}$ , we immediately get the result. □

## C. A Novel Tensor Rank Penalty

Suppose  $\mathcal{A}$  has t-SVD  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ , we define the  $\gamma$ -norm of  $\mathcal{A}$  as

$$\|\mathcal{A}\|_\gamma = \frac{1}{n_3} \sum_{i,k} \varphi_{1,\gamma}(\bar{\mathcal{S}}(i, i, k)). \quad (25)$$

The tensor  $\gamma$ -norm enjoys the following properties.

**Proposition C.1.** For  $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ , suppose  $\mathcal{A}$  has t-SVD  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ , then  $\|\mathcal{A}\|_\gamma$  satisfies:

- (i)  $\|\mathcal{A}\|_\gamma \geq 0$  with equality holds iff  $\mathcal{A} = 0$ ;
- (ii)  $\|\mathcal{A}\|_\gamma$  is increasing in  $\gamma$ ,  $\|\mathcal{A}\|_\gamma \leq \|\mathcal{A}\|_*$  and  $\lim_{\gamma \rightarrow \infty} \|\mathcal{A}\|_\gamma = \|\mathcal{A}\|_*$ ;
- (iii)  $\|\mathcal{A}\|_\gamma$  is concave with respect to  $\{\bar{\mathcal{S}}(i, i, k)\}_{i,k}$ ;
- (iv)  $\|\mathcal{A}\|_\gamma$  is orthogonal invariant, i.e., for any orthogonal tensors  $\mathcal{P} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ ,  $\mathcal{Q} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ ,  $\|\mathcal{P} * \mathcal{A} * \mathcal{Q}\|_\gamma = \|\mathcal{A}\|_\gamma$ .

*Proof.* (i) Since  $\varphi_{1,\gamma}(t) \geq 0$  and  $\|\mathcal{A}\|_\gamma$  is the sum of  $\varphi_{1,\gamma}(t)$ , we immediately know  $\|\mathcal{A}\|_\gamma \geq 0$ . If  $\mathcal{A} = 0$ , then obviously  $\mathcal{S} = 0$  and  $\bar{\mathcal{S}} = 0$ , which implies  $\|\mathcal{A}\|_\gamma = 0$ . On the other hand,  $\|\mathcal{A}\|_\gamma = 0$  implies  $\bar{\mathcal{S}}(i, i, k) = 0$ . However,  $\bar{\mathcal{S}}$  is f-diagonal, thus  $\bar{\mathcal{S}} = 0$ , and  $\mathcal{A} = \mathcal{U} * \text{iff}t(\bar{\mathcal{S}}, [], 3) * \mathcal{V}^* = 0$ .

- (ii) Since  $\varphi_{1,\gamma}(t)$  is increasing with respect to  $\gamma$  and  $\|\mathcal{A}\|_\gamma$  is the sum of  $\varphi_{1,\gamma}(t)$ , using the properties of  $\varphi_{\lambda,\gamma}(t)$  we know  $\|\mathcal{A}\|_\gamma$  is increasing with respect to  $\gamma$  and

$$\lim_{\gamma \rightarrow \infty} \|\mathcal{A}\|_\gamma = \lim_{\gamma \rightarrow \infty} \frac{1}{n_3} \sum_{i,k} \varphi_{1,\gamma}(\bar{\mathcal{S}}(i,i,k)) = \frac{1}{n_3} \sum_{i,k} |\bar{\mathcal{S}}(i,i,k)| = \|\mathcal{A}\|_*.$$

Combining the above facts, we also get  $\|\mathcal{A}\|_\gamma \leq \|\mathcal{A}\|_*$ .

- (iii) This follows from the fact that  $\varphi_{\lambda,\gamma}(t)$  is concave over  $[0, \infty)$ .

- (iv) Since  $\mathcal{A}$  has t-SVD  $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ , we claim that  $\mathcal{P} * \mathcal{U} * \mathcal{S} * \mathcal{V}^* * \mathcal{Q}^*$  is the t-SVD of  $\mathcal{P} * \mathcal{A} * \mathcal{Q}^*$ .  $\mathcal{S}$  is already f-diagonal, so we only need to verify  $\mathcal{P} * \mathcal{U}$  and  $(\mathcal{V}^* * \mathcal{Q}^*)^* = \mathcal{Q} * \mathcal{V}$  are orthogonal.

$$(\mathcal{P} * \mathcal{U}) * (\mathcal{P} * \mathcal{U})^* = \mathcal{P} * \mathcal{U} * \mathcal{U}^* * \mathcal{P} = \mathcal{I}.$$

Other equalities are similar to verify. Therefore,  $\|\mathcal{P} * \mathcal{A} * \mathcal{Q}^*\|_\gamma = \frac{1}{n_3} \sum_{i,k} \varphi_{1,\gamma}(\bar{\mathcal{S}}(i,i,k)) = \|\mathcal{A}\|_\gamma$ .  $\square$

## D. Generalized Thresholding Operators

**Theorem D.1.** We can view  $\Phi_{\lambda,\mu}(\mathcal{X})$  as a function of  $|\mathcal{X}|$ , and  $\|\mathcal{X}\|_\gamma$  as a function of  $\{\bar{\mathcal{S}}(i,i,k)\}_{i,k}$ . For any  $\mathcal{X}^{\text{old}}$ , let

$$Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}) = \Phi_{\lambda,\gamma}(\mathcal{X}^{\text{old}}) + \sum_{i,j,k} \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)(|\mathcal{X}_{ijk}| - |\mathcal{X}_{ijk}^{\text{old}}|),$$

$$Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}}) = \|\mathcal{X}^{\text{old}}\|_\gamma + \frac{1}{n_3} \sum_{i,k} \varphi'_{1,\gamma}(\bar{\mathcal{S}}_{iik}^{\text{old}})(\bar{\mathcal{S}}_{iik} - \bar{\mathcal{S}}_{iik}^{\text{old}}),$$

then  $\Phi_{\lambda,\gamma}(\mathcal{X}) \leq Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}})$ ,  $\|\mathcal{X}\|_\gamma \leq Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}})$ .

*Proof.* Recall that for any  $s, t \geq 0$  we have  $\varphi_{\lambda,\gamma}(t) \leq \varphi_{\lambda,\gamma}(s) + \varphi_{\lambda,\gamma}(s)(t-s)$ .

$$\begin{aligned} \Phi_{\lambda,\gamma}(\mathcal{X}) &= \sum_{i,j,k} \varphi_{\lambda,\gamma}(\mathcal{X}_{ijk}) = \sum_{i,j,k} \varphi_{\lambda,\gamma}(|\mathcal{X}_{ijk}|) \leq \sum_{i,j,k} (\varphi_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|) + \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)(|\mathcal{X}_{ijk}| - |\mathcal{X}_{ijk}^{\text{old}}|)) = Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}). \\ \|\mathcal{X}\|_\gamma &= \frac{1}{n_3} \sum_{i,k} \varphi_{1,\gamma}(\bar{\mathcal{S}}_{ijk}) \leq \frac{1}{n_3} \sum_{i,k} \left( \varphi_{1,\gamma}(\bar{\mathcal{S}}_{ijk}^{\text{old}}) + \varphi'_{1,\gamma}(\bar{\mathcal{S}}_{ijk}^{\text{old}})(\bar{\mathcal{S}}_{ijk} - \bar{\mathcal{S}}_{ijk}^{\text{old}}) \right) = Q_\gamma(\mathcal{X}|\mathcal{X}^{\text{old}}). \end{aligned}$$

$\square$

In the following soft thresholding operator is defined as  $\mathcal{T}_\lambda(z) = \text{sgn}(z)[|z| - \lambda]_+$ , which is the proximal operator of  $\ell_1$ -norm.

**Definition D.1 (Generalized soft thresholding).** Suppose  $\mathcal{X}, \mathcal{W} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ , the generalized soft thresholding operator is defined as

$$[\mathcal{T}_\lambda(\mathcal{X})]_{ijk} = \mathcal{T}_{\mathcal{W}_{ijk}}(\mathcal{X}_{ijk}). \quad (26)$$

**Theorem D.2.** For  $\forall \mu > 0$ , let  $\mathcal{W}_{ijk} = \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)/\mu$ , then

$$\mathcal{T}_\lambda(\mathcal{Y}) = \arg \min_{\mathcal{X}} Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2. \quad (27)$$

*Proof.* In fact,

$$\begin{aligned} \arg \min_{\mathcal{X}} Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2 &= \arg \min_{\mathcal{X}} \sum_{ijk} \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|) |\mathcal{X}_{ijk}| + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2 \\ &= \arg \min_{\mathcal{X}} \sum_{ijk} \left( \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|) |\mathcal{X}_{ijk}| + \frac{\mu}{2} (\mathcal{X}_{ijk} - \mathcal{Y}_{ijk})^2 \right), \end{aligned}$$

which is separable to each entries of  $\mathcal{X}$ . Consider  $\mathcal{X}_{ijk}$ , according to the property of soft thresholding operator, the minimum of  $\varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|) |\mathcal{X}_{ijk}| + \frac{\mu}{2} (\mathcal{X}_{ijk} - \mathcal{Y}_{ijk})^2$  is attained at  $\mathcal{T}_{\varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)/\mu}(\mathcal{Y}_{ijk})$ . Therefore, let  $\mathcal{W}_{ijk} = \varphi'_{\lambda,\gamma}(|\mathcal{X}_{ijk}^{\text{old}}|)/\mu$ , by the definition of generalized soft thresholding, we immediately get

$$\mathcal{T}_\lambda(\mathcal{Y}) = \arg \min_{\mathcal{X}} Q_{\lambda,\gamma}(\mathcal{X}|\mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2.$$

$\square$

**Definition D.2 (Generalized t-SVT).** Suppose a 3-way tensor  $\mathcal{Y}$  has t-SVD  $\mathcal{Y} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ ,  $\mathcal{W}$  is a tensor with the same shape of  $\mathcal{Y}$ , the generalized tensor singular value thresholding operator is defined as

$$\mathcal{D}_{\mathcal{W}}(\mathcal{Y}) = \mathcal{U} * \tilde{\mathcal{S}} * \mathcal{V}^*, \quad (28)$$

where  $\tilde{\mathcal{S}} = \text{ifft}(\mathcal{T}_{\mathcal{W}}(\mathcal{S}), [], 3)$ .

**Theorem D.3.** For  $\forall \mu > 0$ , let  $\mathcal{W}_{ijk} = \delta_i^j \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}})/\mu$  where  $\delta_i^j$  is the Kronecker symbol, then

$$\mathcal{D}_{\mathcal{W}}(\mathcal{Y}) = \arg \min_{\mathcal{X}} Q_{\gamma}(\mathcal{X} | \mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2. \quad (29)$$

*Proof.* In fact,

$$\begin{aligned} \arg \min_{\mathcal{X}} Q_{\gamma}(\mathcal{X} | \mathcal{X}^{\text{old}}) + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2 &= \arg \min_{\mathcal{X}} \frac{1}{n_3} \sum_{i,k} \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{\mathcal{S}}_{iik} + \frac{\mu}{2} \|\mathcal{X} - \mathcal{Y}\|_F^2 \\ &= \arg \min_{\mathcal{X}} \frac{1}{n_3} \sum_{i,k} \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{\mathcal{S}}_{iik} + \frac{\mu}{2n_3} \sum_k \|\overline{X}^{(k)} - \overline{Y}^{(k)}\|_F^2 \\ &= \arg \min_{\mathcal{X}} \frac{1}{n_3} \sum_k \left( \sum_i \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{S}_{i,i}^{(k)} + \frac{\mu}{2} \|\overline{X}^{(k)} - \overline{Y}^{(k)}\|_F^2 \right). \end{aligned}$$

This optimization problem is separable in transformation domain with respect to each slice. Consider the subproblem of minimizing  $\sum_k \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{S}_{i,i}^{(k)} + \frac{\mu}{2} \|\overline{X}^{(k)} - \overline{Y}^{(k)}\|_F^2$ , suppose  $\mathcal{Y}$  has t-SVD  $\mathcal{Y} = \mathcal{U} * \mathcal{R} * \mathcal{V}$  or equivalently  $\overline{Y} = \overline{U} \overline{R} \overline{V}^*$ , we have

$$\begin{aligned} \sum_k \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{S}_{i,i}^{(k)} + \frac{\mu}{2} \|\overline{X}^{(k)} - \overline{Y}^{(k)}\|_F^2 &= \sum_k \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{S}_{i,i}^{(k)} + \frac{\mu}{2} \|\overline{X}^{(k)} - \overline{U}^{(k)} \overline{R}^{(k)} (\overline{V}^{(k)})^*\|_F^2 \\ &= \sum_k \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}}) \overline{S}_{i,i}^{(k)} + \frac{\mu}{2} \|(\overline{U}^{(k)})^* \overline{X}^{(k)} \overline{V}^{(k)} - \overline{R}^{(k)}\|_F^2. \end{aligned}$$

However, note that  $\overline{S}_{i,i}^{(k)}$  is still the singular values of  $(\overline{U}^{(k)})^* \overline{X}^{(k)} \overline{V}^{(k)}$ , simple calculation reveals that we obtain the minimum if  $(\overline{U}^{(i)})^* \overline{X}^{(k)} \overline{V}^{(k)}$  is diagonal with  $i$ -th diagonal element equals to  $\mathcal{T}_{\varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}})/\mu}(\overline{R}_{i,i}^k)$ . That is, let  $W$  be a matrix with  $W_{i,j}^{(k)} = \delta_i^j \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}})/\mu$ , then  $\overline{X} = \overline{U} \mathcal{T}_W \overline{R} \overline{V}$ . Transform back to original domain, by the definition of generalized t-SVT, we get  $\mathcal{X} = \mathcal{U} * \tilde{\mathcal{S}} * \mathcal{V}^*$  where  $\tilde{\mathcal{S}} = \text{ifft}(\mathcal{T}_{\mathcal{W}}(\mathcal{R}), [], 3)$  and  $\mathcal{W}_{ijk} = \delta_i^j \varphi'_{1,\gamma}(\overline{\mathcal{S}}_{iik}^{\text{old}})/\mu$ .  $\square$