Tensor Decomposition with Smoothness

Masaaki Imaizumi [1] **Kohei Hayashi** [2] [3]

Abstract

Real data tensors are typically high dimensional; however, their intrinsic information is preserved in low-dimensional space, which motivates the use of tensor decompositions such as Tucker decomposition. Frequently, real data tensors *smooth* in addition to being low dimensional, which implies that adjacent elements are similar or continuously changing. These elements typically appear as spatial or temporal data. We propose *smoothed Tucker decomposition (STD)* to incorporate the smoothness property. STD leverages smoothness using the sum of a few basis functions; this reduces the number of parameters. An objective function is formulated as a convex problem, and an algorithm based on the alternating direction method of multipliers is derived to solve the problem. We theoretically show that, under the smoothness assumption, STD achieves a better error bound. The theoretical result and performances of STD are numerically verified.

1. Introduction

A tensor (i.e., a multi-way array) is a data structure that is a generalization of a matrix, and it can represent higher-order relationships. Tensors appear in various applications such as image analysis (Jia et al., 2014), data mining (Kolda & Sun, 2008), and medical analysis (Zhou et al., 2013). For instance, functional magnetic resonance imaging (fMRI) records brain activities in each time period as voxels, which are represented as 4-way tensors (X-axis × Y-axis × Z-axis × time). Frequently, data tensors in the real world contain several missing elements and/or are corrupted by noise, which leads to the *tensor completion* problem for predicting missing elements and the *tensor recovery* problem for removing noise.

[1]Institute of Statistical Mathematics [2]National Institute of Advanced Industrial Science and Technology [3]RIKEN. Correspondence to: Masaaki Imaizumi <insou11@hotmail.com>.

To solve these problems, the *low-rank assumption*, i.e., given tensor is generated from a small number of latent factors, is widely used. If the number of observed elements is sufficiently larger than the number of latent factors (i.e., rank) and noise level, we can estimate latent factors and reconstruct the entire structure. The methods of estimating latent factors are collectively referred to as *tensor decompositions*. There are several formulations of tensor decompositions such as Tucker decomposition (Tucker, 1966) and the CANDECOMP/PARAFAC(CP) decomposition (Harshman, 1970). While these methods were originally formulated as nonconvex problems, several authors have studied their convex relaxations in recent years (Liu et al., 2009; Tomioka et al., 2010; Signoretto et al., 2011; Gandy et al., 2011).

Another important, yet less explored, assumption is the *smoothness* property. Consider fMRI data as a tensor X. As fMRI data are spatiotemporal, each element of X is expected to be similar to its adjacent elements with every way, i.e., $x_{i,j,k,t}$ should be close to $x_{i\pm1,j,k,t}, x_{i,j\pm1,k,t}, x_{i,j,k\pm1,t}$, and $x_{i,j,k,t\pm1}$. In statistics, this kind of smoothness property has been studied through functional data analysis (Ramsay, 2006; Hsing & Eubank, 2015). Studies show that the smoothness assumption increases *sample efficiency*, i.e., estimation is more accurate with small sample size. Another advantage is that the smoothness assumption makes *interpolation* possible, i.e., we can impute an unobserved value using its adjacent observed values. This interpolation ability is particularly useful for solving a specific tensor completion problem referred to as the *tensor interpolation* problem, as known as the "cold-start" problem (Gantner et al., 2010). Suppose a case in which fMRI tensor X is completely missing at $t = t'$. In this case, standard tensor decompositions cannot predict missing elements because there is no information to estimate the latent factor at $t = t'$. However, using the smoothness property, we can estimate the missing elements from the elements at $t = t' - 1$ and $t = t' + 1$.

A fundamental challenge of tensor completion and recovery methods is to analyze their performance. Tomioka et al. (2011) extensively studied the statistical performance of low-rank tensor decompositions. On the contrary, the performance of tensor decompositions incorporating smoothness (Yokota et al., 2015b;a; Amini et al., 2013) has never

been addressed. The most important barrier is that all methods are formulated as nonconvex problems, which hinders the use of the tools developed in the convex tensor decompositions (Liu et al., 2009; Signoretto et al., 2010; Tomioka et al., 2010).

Contributions   In this paper, we propose a simple tensor decomposition model incorporating the smoothness property, which we refer to as *Smoothed Tucker Decomposition* (STD). Following the notions of functional data analysis, STD approximates an observed tensor by a small number of basis functions, such as Fourier series, and decomposes them through Tucker decomposition (Figure 1). STD is formulated as a convex optimization problem that regularizes the ranks and degree of smoothness. To solve this problem, we derive an algorithm based on the alternating direction method of multipliers (ADMM), which always finds the global optimum.

Based on the convex formulation, we provide a few theoretical guarantees of STD, namely, we derive error bounds for tensor recovery and interpolation problems. We show that the error bounds for smooth tensors are improved and better than those for other methods. In addition, to the best of our knowledge, this is the first analysis that establishes an error bound for tensor interpolation. These results are empirically confirmed through experiments using synthetic and real data.

To summarize, STD has the following advantages.

- Sample efficiency: STD achieves the same error with less sample size.
- Interplation ability: STD can solve the tensor interpolation problem.
- Convex formulation: STD ensures that a global solution is obtained.

Related Works   A few authors have investigated the smoothness property for tensor decompositions. Amini et al. (2013) proposed a kernel method, and Yokota et al. (2015a;b) developed a smooth decomposition method for matrices and tensors using basis functions. These studies demonstrated that the smoothness assumption significantly improves the performance of tensor decompositions for actual applications such as noisy image reconstruction (Yokota et al., 2015b). However, these performance gains were confirmed only in an empirical manner.

Several authors have addressed the tensor interpolation problem by extending tensor decompositions; however, instead of smoothness, these methods utilize additional information such as network structures (Hu et al., 2015) or side information (Gantner et al., 2010; Narita et al., 2011). Moreover, the performance of the tensor interpolation problem has never been analyzed theoretically.
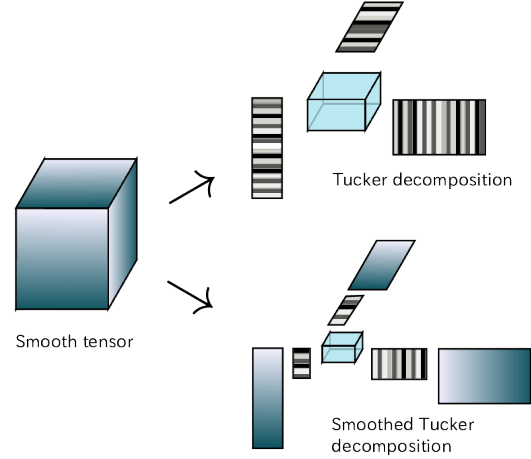

Figure 1. Comparison of Tucker decomposition and STD. In STD, the mode-wise smoothness of the observed tensor is preserved via basis functions.

2. Preliminaries

Given $K \in \mathbb{N}$ natural numbers, $I_1, \ldots, I_K \in \mathbb{N}$, let $\mathcal{X} \subset \mathbb{R}^{I_1 \times \ldots \times I_K}$ be the space of a K-way tensor, and $X \in \mathcal{X}$ be the K-way tensor that belongs to $\mathcal{X}$. For practical use, we define $I_{\backslash k} := \prod_{k' = k} I_k$. Each way of a tensor is referred to as *mode*; I_k is the dimensionality of the k-th mode for $k = 1, \ldots, K$. For vector $Y \in \mathbb{R}^d$, $[Y]_j$ denotes its j-th element. Similarly, $[X]_{j_1 j_2 \ldots j_K}$ denotes the $(i_1, i_2, \ldots, i_K)$-th element of X. The inner product in $\mathcal{X}$ is defined as $\langle X, X' \rangle = \sum_{j_1, j_2, \ldots, j_K=1}^{I_1, I_2, \ldots, I_K} [X]_{j_1, j_2, \ldots, j_K} [X']_{j_1, j_2, \ldots, j_K}$ for $X, X' \in \mathcal{X}$. This induces the Frobenius norm, $\||X\||_F = \sqrt{\langle X, X \rangle}$. For vectors $Z \in \mathbb{R}^d$, let $\|Z\| = \sqrt{Z^T Z}$ denote the norm. In addition, we introduce the L_2 norm for functions as $\|f\|_2^2 = \int_I f(t)^2 dt$ for function $f : I \to \mathbb{R}$ with some domain $I \subset \mathbb{R}$. $\mathcal{C}^\alpha(I)$ denotes a set of an α-times differentiable function on I.

2.1. Tucker Decomposition

With a set of finite positive integers $(R_1, \ldots, R_K)$, the Tucker decomposition of X is defined as

$$X = \sum_{r_1, \ldots, r_K=1}^{R_1, \ldots, R_K} g_{r_1 \ldots r_K} u_{r_1}^{(k)} \otimes u_{r_2}^{(k)} \otimes \ldots u_{r_K}^{(k)}, \quad (1)$$

where $g_{r_1 \ldots r_K} \in \mathbb{R}$ is a coefficient for each r_k, $\otimes$ denotes the tensor product, and $u_{r_k} \in \mathbb{R}^{I_k}$ denotes vector for each $r_k (k = 1, \ldots, K)$, which are orthogonal to each other for $r_k = 1, \ldots, R_k$. Here, we refer to $(R_1, \ldots, R_K)$ as the *Tucker rank*, and X is an $(R_1, \ldots, R_K)$-*rank tensor*. In addition, we let tensor $G \in \mathbb{R}^{R_1 \times \ldots \times R_K}$ with $[G]_{r_1, \ldots, r_K} = g_{r_1 \ldots r_K}$ be a *core tensor* and matrix $U_{(k)} = (u_{r_1}^{(k)} \ldots u_{r_K}^{(k)}) \in \mathbb{R}^{I_k \times R_k}$ is a set of the vectors for all $k = 1, \ldots, K$. Using this notation, Tucker decomposition

(1) can be written as

$$X = G \times_1 U_{(1)} \times_1 U_{(2)} \times_2 \ldots \times_K U_{(K)}, \qquad (2)$$

where $\times_k$ denotes the k-mode matrix product (see Kolda & Bader (2009) for more details).

2.2. Application Problems for Tensors

Let $S \subset \{(j_1, j_2, \ldots, j_K)\}_{j_1, j_2, \ldots, j_K=1}^{I_1, I_2, \ldots, I_K}$ be an index set and $n := |S|$. Let $j(i)$ be the i-th element of S for $i = 1, \ldots, n$. Then, we consider the following observation model:

$$y_i = [X^*]_{j(i)} + \epsilon_i, \qquad (3)$$

where $X^* \in \mathcal{X}$ is an unobserved true tensor, $\{y_i\}_{i=1}^n$ is the set of observed values, and ϵ_i is noise, where the mean is zero and the variance is σ^2. We define an observation vector $Y := (y_1 \ldots y_n)^T \in \mathbb{R}^n$, and a noise vector $\mathcal{E} := (\epsilon_1 \ldots \epsilon_n)^T \in \mathbb{R}^n$. Additionally, we define a rearranging operator $\mathfrak{X} : \mathcal{X} \to \mathbb{R}^n$ via $[\mathfrak{X}(X)]_i = [X]_{j(i)}$. Using this notation, observation model (3) is written as

$$Y = \mathfrak{X}(X^*) + \mathcal{E}. \qquad (4)$$

When all the elements are observed, i.e., $n = \prod_k I_k$, the problem of estimating X^* is referred to as the *tensor recovery problem*. When a few elements of X are missing, i.e., $n < \prod_k I_k$, the problem is referred to as the *tensor completion problem*. Specifically, for any mode k, if there exists an index $j'_k \in [I_k]$ that S does not contain, we refer to the problem of estimating $[X^*]_{j_1=1}^{I_1} \cdots_{j_k=j'_k} \cdots_{j_K=1}^{I_K}$ as the *tensor interpolation* problem.

Using observation model (4), we provide an estimator for the unknown true tensor X^*. The estimator of $\hat{X}$ is obtained by solving the following optimization problem:

$$\min_{X \in \Theta} \left[\frac{1}{2n} \|Y - \mathfrak{X}(X)\|^2 + \Omega(X) \right], \qquad (5)$$

where $\Theta \subset \mathcal{X}$ is a convex subset of $\mathcal{X}$, and $\Omega : \Theta \to \mathbb{R}_+$ is a regularization term. For the regularization $\Omega(X)$, the *overlapped Schatten 1-norm* is frequently used (Liu et al., 2009; Tomioka et al., 2010; Signoretto et al., 2011; Gandy et al., 2011); is is defined as

$$\|X\|_s := \frac{1}{K} \sum_{k=1}^K \|X_{(k)}\|_s := \frac{1}{K} \sum_{k=1}^K \sum_{r_k=1}^{R_k} \sigma_{r_k}(X_{(k)}),$$

where $X_{(k)} \in \mathbb{R}^{I_k \times \prod_{k' \neq k} I_k}$ denotes the unfolding matrix obtained by concatenating the mode-k fibers of X as column vectors and $\sigma_{r_k}(X_{(k)})$ denotes the r_k-th largest eigenvalue of $X_{(k)}$. This penalty term regularizes the Tucker rank of X (Negahban & Wainwright, 2011; Tomioka et al., 2011).

To solve the problem (5) using the Schatten regularization, ADMM is frequently employed (Boyd et al., 2011; Tomioka et al., 2010). ADMM generates a sequence of variables and Lagrangian multipliers by iteratively minimizing the augmented Lagrangian function. It is known that ADMM can easily solve an optimization problem with a non-differentiable regularization term such as $\|\cdot\|_s$.

3. STD: Smoothed Tucker Decomposition

3.1. Smoothness on Tensors

Before explaining the proposed approach, we introduce the notion of smoothness on tensors. We start with the idea that a data tensor is obtained as a result of the discretization of a multivariate function. For example, consider an observation model of the wind power on a land surface. Suppose that the land surface is described by a plain $[0,1]^2$ (i.e., longitude and latitude) and the observation model is given by a function $f : [0,1]^2 \to \mathbb{R}$. Assume that we have infinite memory space so that we can record the wind power $y = f(a,b)$ for any points $a, b \in [0,1]$. In such an unrealistic case, it is possible to handle the entire information about f. However, only finite memory space is available; we resort to retain finite observations $\{f(a_i, b_i)\}_{i=1}^n$. If the points (a_i, b_i) are considered as a grid, the observations can be considered as a matrix.

This idea is generalized to tensors as follows. Consider a K-variate function $f_X : [0,1]^K \to \mathbb{R}$, and a set of points $\{(j_1, \ldots, j_K) \in [0,1]^K\}_{j_1 \ldots j_K=1}^{I_1 \ldots I_K}$ as grid points in $[0,1]^K$. Then, each element of X is represented as

$$[X]_{j_1 \ldots j_K} = f_X(g_{j_1}, \ldots, g_{j_K}), \qquad (6)$$

for $j_k = 1, 2, \ldots$ with each $k = 1, \ldots, K$.

As the smoothness of the function, we assume that f_X is differentiable with respect to all K arguments, which allows for the expansion of the basis function to a few useful basis functions (for example, Tsybakov (2008)) and the decomposition of multivariate functions by the basis (see Hackbusch (2012) for detail). Let $\{\phi_m^{(k)} : [0,1] \to \mathbb{R}\}_m$ be a set of orthonormal basis functions, such as Fourier series or wavelet series, and $\{w_{m_1, \ldots, m_K} \in \mathbb{R}\}_{m_1, \ldots, m_K}$ be a set of coefficients. Because of the differentiability, f_X is written as the weighted sum of the basis functions as

$$f_X = \sum_{m_1=1}^{\infty} \cdots \sum_{m_K=1}^{\infty} w_{m_1 \ldots m_K} \phi_{m_1}^{(1)} \cdots \phi_{m_K}^{(K)}. \qquad (7)$$

Combining (6) and (7) yields a formulation of the elements of the smooth tensor as

$$[X]_{j_1, \ldots, j_K} \qquad (8)$$

$$= \sum_{m_1=1}^{\infty} \cdots \sum_{m_K=1}^{\infty} w_{m_1 \ldots m_K} \phi_{m_1}^{(1)}(g_{j_1}) \cdots \phi_{m_K}^{(K)}(g_{j_K}).$$

Hereafter, we say that X *is smooth* if it follows (8).

3.2. Objective Function

Model (8) is not directly applicable because it requires an infinite number of basis functions. To prevent this, we consider their truncation. Let $M^{(k)} < \infty$ be the basis functions for mode k, which represents a degree of smoothness of X in terms of mode k. For example, when $M^{(k)}$ is large, such as $M^{(k)} = I_k$ for all k, the basis function formulation can represent any X, which implies that it neglects the smooth structure. Then, we consider X such that it satisfies the following relation:

$$[X]_{j_1,\dots,j_K} \qquad (9)$$
$$= \sum_{m_1=1}^{M^{(1)}} \cdots \sum_{m_K=1}^{M^{(K)}} w_{m_1 \cdots m_K} \phi_{m_1}^{(1)}(g_{j_1}) \cdots \phi_{m_K}^{(K)}(g_{j_K}).$$

For practical use, let $W_X \in \mathbb{R}^{M^{(1)} \times \cdots \times M^{(K)}}$ be a coefficient tensor that satisfies

$$[W_X]_{m_1 \dots m_K} = w_{m_1 \dots m_K},$$

with given X.

Using the representation, we propose an objective function of STD. Based on the same convex optimization approach as (5), we define the objective function as

$$\min_{X \in \Theta} \left[\frac{1}{2n} \|Y - \mathfrak{X}(X)\|^2 + \lambda_n \|\|W_X\|\|_s + \mu_n \|\|W_X\|\|_F^2 \right],$$
$$(10)$$

where $\lambda_n, \mu_n \geq 0$ are regularization coefficients that depend on n and $\lambda_n, \mu_n \to 0$ as $n \to \infty$. Here, regularization terms $\|\|W_X\|\|_s$ and $\|\|W_X\|\|_F^2$ are employed.

There are three primary advantages of the formulation given by (10). Firstly, (10) is written as a convex optimization problem. Thus, it is ensured that to obtain the global solution of W_X will be obtained. Secondly, regularization term $\|\|W_X\|\|_s$ determines the Tucker rank of W_X appropriately. Even though we must select λ_n, this is considerably easier than selecting the values of K. Thirdly, the regularization $\|\|W_X\|\|_F^2$ penalizes the smoothness of f_X. Note that the smoothness of f_X is related to $M^{(k)}$, and we introduce $\|\|W_X\|\|_F^2$ to select an appropriate degree of smoothness.

3.3. Algorithm

To optimize (10), we first reformulate it through vectorization and matricization. We define $x \in \mathbb{R}^{\prod_k I_k}$ as the vectorized tensor of X, and Q is a $\prod_k I_k \times n$ matrix, which is the matricized version of the rearranging operator $\mathfrak{X}$. We define $M^{(\backslash k)} := \prod_{k' \neq k} M^{(k)}$, and $Z_k \in \mathbb{R}^{M_k \times M^{(\backslash k)}}$ is the mode-k unfolding matrix of W_X. Let $w \in \mathbb{R}^{\prod_k M^{(k)}}$ be the

vectorized tensor of W_X. Let $\Phi : \mathbb{R}^{\prod_k M^{(1)} \times \cdots \times M^{(K)}} \to \mathbb{R}^{I_1 \times \cdots \times I_K}$ be an operator that converts W_X to X as given by (9) using $\{\phi_m(g_j)\}_{m,j}$. Let Γ be a $\prod_k I_k \times \prod_k M^{(k)}$ matrix that satisfies $\Phi(w) = \Gamma w$. As $\Phi(w)$ is a linear mapping by (9), the existence of Γ is ensured. Then, (10) is rewritten as follows:

$$\min_{x,w,\{Z_k\}_{k=1}^K} \frac{1}{2n}\|Y - Qx\|^2 + \frac{\lambda_n}{K}\sum_{k=1}^K \|Z_k\|_s + \mu_n\|w\|^2,$$
$$\text{s.t. } x = \Phi(w), P_k(w) = Z_k, \forall k \qquad (11)$$

where $P_k : \mathbb{R}^{\prod_k M^{(k)}} \to \mathbb{R}^{M^{(k)} \times M^{(\backslash k)}}$ is a rearranging operator from the vector to the unfolding matrix. Note that $\|\|W\|\|_F = \|w\|_F$ holds by the definition of w.

We use the ADMM approach to solve (11). Maximizing the augmented Lagrangian function for (11), we obtain the following iteration steps. Here, $\eta > 0$ is a step size and $\{\alpha_k \in \mathbb{R}^{\prod_k M^{(k)}}\}_{k=1}^K$ and $\beta \in \mathbb{R}^{\prod_k I_k}$ are Lagrangian multipliers. Let $(x_0, w_0, \{Z_{k,0}\}_{k=1}^K, \{\alpha_{k,0}\}_{k=1}^K, \beta_0)$ be an initial point. The ADMM step at the ℓ-th iteration is written as follows:

$$x^{\ell+1} = \left(Q^T Q + n\eta I\right)^{-1}\left(Q^T Y - n\beta^\ell + n\eta\Gamma w^\ell\right)$$
$$w^{\ell+1} = (2\mu_n I + \eta K I + \eta \Gamma^T \Gamma)^{-1}$$
$$\times \left(\sum_{k=1}^K \left\{\eta P_k^{-1}(Z_k^\ell) - \alpha_k^\ell\right\} + \Gamma^T \beta^\ell + \eta \Gamma^T x^{\ell+1}\right)$$
$$Z_k^{\ell+1} = \text{prox}_{\lambda_n/\eta}(P_k(w^{\ell+1} + \alpha_k^\ell))$$
$$\alpha_k^{\ell+1} = \alpha_k^\ell + (w^{\ell+1} - P_k^{-1}(Z_k^{\ell+1}))$$
$$\beta^{\ell+1} = \beta^\ell + (x^{\ell+1} - \Gamma w^{\ell+1}),$$

where $\text{prox}_{\lambda_n/\eta}(\cdot)$ denotes the shrinkage operation of the singular values, which is defined as $\text{prox}_\eta(Z) = U \max(S - \eta I, 0)V^T$, where $S, U,$ and V are obtained through the singular value decomposition as $Z = USV^T$. Note that the ADMM steps for Z_k and α_k are required for every $k = 1, \dots, K$. For η, Tomioka et al. (2010) suggest setting $\eta = \eta_0/\sqrt{\text{Var}(y_i)}$ with some constant η_0. As the regularization terms are convex, the sequence of the variables of ADMM is ensured to converge to the optimal solution of (10) (Gandy et al., 2011, Theorem 5.1).

3.4. Practical Issues

STD has several hyperparameters, i.e., $\lambda_n, \mu_n, \{M^{(k)}\}$, and $\{\phi_m\}$. λ_n and μ_n can be determined through cross validation. $\{M^{(k)}\}$ is initialized as a large value and reduced during the algorithm depending on μ_n. As $M^{(k)}$ does not exceed I_k because of an identification reason, the initial value of $M^{(k)}$ is bounded. Practically, $M^{(k)}$ is considerably less than I_k. Thus, we can start the iteration with small values.

One can criticize that a few data tensors are smooth with one mode, but not with others. We emphasize that STD can address such a situation by controlling $M^{(k)}$ for each k. As STD can represent tensors without the smooth structure when $M^{(k)} = I_k$, setting $M^{(k)} = I_k$ for some mode k and $M^{(k)} \leq I_k$ for other modes is sufficient to address the situation.

In this study, selecting the form of the basis functions $\{\phi_m\}_m$ is not our primary interest because it does not specifically affect the theoretical result. However, there are a few typical choices. For instance, when the data tensor is periodic, such as audio, the Fourier basis is appropriate. Even through other functions, such as wavelet or spline functions, provide the theoretical guarantees of approximating f.

4. Theoretical Analysis

We introduce a few notations for convenience. Let $\||X|\|_m := \frac{1}{K} \sum_{k=1}^{K} \max_r \sigma_r(X_{(k)})$ be a norm of a tensor, which is necessary for evaluating the penalty parameters. Let $\mathfrak{X}^*$ be an adjoint operator of $\mathfrak{X}$, namely, $\langle \mathfrak{X}(z), z' \rangle = \langle z, \mathfrak{X}^*(z') \rangle$ holds for all $z, z' \in \mathcal{X}$. For theoretical requirement, we let the basis functions $\{\phi_j : [0,1] \to \mathbb{R}\}_j$ be uniformly bounded for all $j \geq 1$. All proofs of this section are provided in the supplemental material.

4.1. Error Bound with X^*

First, we impose the following assumption on $\mathfrak{X}$.

Assumption 1 (Restricted Strong Convexity (RSC) condition). *A finite constant $C_X > 0$ depending on $\{I_k\}_k$ exists, then the rearranging operator $\mathfrak{X}$ satisfies*

$$\frac{1}{2n} \|\mathfrak{X}(X)\|^2 \geq C_X \||X|\|_F^2,$$

for all $X \in \Theta$.

Intuitively, this assumption requires that $\mathfrak{X}$ is sufficiently sensitive to the perturbation of X. A similar type of condition has been used in previous studies on sparse regression, such as LASSO (Bickel et al., 2009; Raskutti et al., 2010), i.e., the restricted isometry condition. The RSC condition is weaker than the isometry condition because the RSC condition requires only the lower bound.

We provide the following lemma regarding the error bound when true tensor X^* can be neither smooth nor low-rank. Let $(R_1^W, \ldots, R_K^W)$ be the Tucker rank of W_X.

Lemma 2. *Consider $X^* \in \Theta$, and the rearranging operator $\mathfrak{X}$ that satisfies the RSC condition. Suppose there exist sequences λ_n, μ_n, and κ_n that satisfy $\kappa_n \geq c(\frac{2}{n}\||\mathfrak{X}^*(\mathcal{E})|\|_m + \lambda_n)$ and $1 - c\mu_n < 3/4$ with a constant*

$c > 0$. *Then, with some constants $C_1, C_2, C_3 > 0$, we have*

$$\||\hat{X} - X^*|\|_F \leq \max\{I, II, III\},$$

where I, II, and III are:

$$I = \frac{C_1 \kappa_n}{K} \sum_{k=1}^{K} \sqrt{R_k^W},$$

$$II = \left(\frac{C_2 \lambda_n}{K} \sum_{k=1}^{K} \sum_{r > R_k^W} \sigma_r(W_{X^*(k)})\right)^{\frac{1}{2}},$$

$$III = \left(\frac{C_3 \mu_n}{K} \sum_{\{m_k > M^{(k)}\}_k} (w_{m_1 \ldots m_k}^*)^2\right)^{\frac{1}{2}},$$

respectively; $w_{m_1 \ldots m_k}^*$ is the coefficient of X^*.

Lemma 2 states that the estimation error of $\hat{X}$ is bounded by three types of values, where (I) indicates the error resulting from estimating a tensor that is smooth and low-rank; (II) indicates the error resulting from introducing the low-rank; and (III) indicates the error resulting from approximating by the smooth tensor.

From Lemma 2, we see that (II) and (III) disappear and (I) remains when X^* is low rank and smooth, which we show in the next proposition. Here, we define $\tilde{\Theta} \subset \Theta$ as the set of the tensors represented by (9) with $\{M^{(k)}\}_k$ and the coefficient tensor W_X with its rank $(R_1^W, \ldots, R_K^W)$.

Proposition 3. *Suppose the same conditions of Lemma 1 hold and $X^* \in \tilde{\Theta}$ is smooth and low-rank. Then, with some constant $C_f > 0$ we have*

$$\||\hat{X} - X^*|\|_F \leq \frac{C_f \kappa_n}{K} \sum_{k=1}^{K} \sqrt{R_k^W}.$$

4.2. Error Bound with f_{X^*}

One of the advantages of STD is that it can estimate X^* and the smooth function f_X defined in (6), which allows for the interpolation of X^*. Here, we evaluate the estimation error of STD with respect to the norm $\| \cdot \|_{L^2}$ for the functional space. Let us define $f_{X^*} := \sum_{m_1, \ldots, m_K}^{M^{(1)}, \ldots, M^{(K)}} w_{m_1, \ldots, m_K}^* \phi_{m_1}^{(1)} \cdots \phi_{m_K}^{(K)}$ which is one of the smooth function as the limit of X^* as $I_k \to \infty$ for all $k \in \{1, \ldots, K\}$. We define the estimator of f_{X^*} by the following:

$$f_{\hat{X}} := \sum_{m_1 \ldots m_K = 1}^{M^{(1)} \ldots M^{(K)}} \hat{w}_{m_1 \ldots m_K} \phi_{m_1}^{(1)} \cdots \phi_{m_K}^{(K)},$$

where $\hat{w}_{m_1 \ldots m_K}$ is an element of $W_{\hat{X}}$. Estimation error is provided as follows.

METHOD	RECOVERY	INTERPOLATION	ASSUMPTION
TOMIOKA ET AL. (2011)	$O(R^X I^{-2})$	N/A	LOW-RANK
TOMIOKA & SUZUKI (2013)	$O(R^X I^{-2})$	N/A	LOW-RANK
WIMALAWARNE ET AL. (2014)	$O(R^X I^{-2})$	N/A	LOW-RANK
STD	$O(R^W I^{-2})$	$O\left(n^{-1/2} I^{K-1} \sqrt{R^W}\right)$	LOW-RANK & SMOOTH

Table 1. Comparison of error bounds under the low-rank and smoothness assumptions. For clarity, the special cases of the error bounds are shown, where the shape of X is symmetric, i.e., $I := I_1 = \cdots = I_K$, $R^X := R_1^X = \cdots = R_K^X$, and $R^W := R_1^W = \cdots = R_K^W$.

Lemma 4. *Suppose that the rearranging operator $\mathfrak{X}$ that satisfies the RSC condition and the true tensor $X^* \in \tilde{\Theta}$. Then, with some constant $C_F > 0$, we have*

$$\sup_{g \in [0,1]^K} |f_{\hat{X}}(g) - f_{X^*}(g)| \leq \frac{C_F \kappa_n}{K} \sum_{k=1}^{K} \sqrt{R_k^W}.$$

When $\kappa_n \to 0$ by the setting, Lemma 4 shows that $f_{\hat{X}}$ estimated by STD uniformly converges to f_{X^*}.

4.3. Applications and Comparison

To discuss the result of Lemma 2 more precisely, we consider the following two practical settings: tensor recovery and tensor interpolation. For each setting, we derive rigorous error bounds.

4.3.1. TENSOR RECOVERY

We consider that all the elements of X are observed, and they are affected by noise, i.e., we set $n = \prod_{k=1}^{K} I_k$ and $\mathfrak{X}$ is a vectorization operator. Then, by applying Lemma 2, we obtain the following result.

Theorem 5. *Suppose that $X^* \in \tilde{\Theta}$ and the rearranging operator $\mathfrak{X}$ that satisfies the RSC condition, and the noise is i.i.d. Gaussian. Let $C_{10}, C_{11} > 0$ be some finite constants. By setting $\kappa_n = C_{10} \sigma \frac{1}{nK} \sum_{k=1}^{K} (\sqrt{I_k} + \sqrt{I_{\setminus k}})$, with high probability, we have*

$$\||\hat{X} - X^*\||_F^2$$
$$\leq C_{11} \sigma^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{I_k} + \sqrt{I_{\setminus k}} \right)^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{R_k^W} \right)^2.$$

Note that, in Theorem 5, the first part $\left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{I_k} + \sqrt{I_{\setminus k}} \right)^2$ comes from the noise and the second part $\left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{R_k^W} \right)^2$ comes from the Tucker rank of W_X.

4.3.2. TENSOR INTERPOLATION

Lemma 4 shows that STD can estimate the value of f_{X^*} for all in $g \in [0,1]^K$, and not only on the given grids $\{(g_{j_1}, \ldots, g_{j_K})\} \subset [0,1]^K$. By tuning κ_n in Lemma 4, we obtain the error bound for the tensor interpolation problem.

Theorem 6. *Suppose that $X^* \in \tilde{\Theta}$ and the rearranging operator $\mathfrak{X}$ that satisfies the RSC condition, and the noise is i.i.d. Gaussian. By setting $C_{20}, C_{21} > 0$ and $\kappa_n = \frac{C_{20}}{\sqrt{n}K} \sigma \sum_{k=1}^{K} (\sqrt{I_k} + \sqrt{I_{\setminus k}})$, with high probability, we have*

$$\sup_{g \in [0,1]^K} |f_{\hat{X}}(g) - f_{X^*}(g)|$$
$$\leq \frac{C_{21} \sigma}{\sqrt{n}} \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{I_k} + \sqrt{I_{\setminus k}} \right) \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{R_k^W} \right).$$

4.4. Comparison to Related Studies

Several studies have derived an error bound for $\||\hat{X} - X^*\||_F^2 / n$ in each situation. Tomioka et al. (2011) investigated the tensor decomposition problem with an overlapped Schatten-norm and derived the error bound as

$$\||\hat{X} - X^*\||_F =$$
$$O\left(\left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{I_k} + \sqrt{I_{\setminus k}} \right) \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{R_k^X} \right) \right), \quad (12)$$

where $(R_1^X, \ldots, R_K^X)$ is the Tucker of X^*. The error bound in Proposition 3 is obtained by replacing the part R_k^X in (12) by R_k^W. Tomioka & Suzuki (2013) and Wimalawarne et al. (2014) introduced modified norms with the Schatten-norm and derived other error bounds.

Table 1 compares the main coefficients for the convergence. When the tensor is sufficiently smooth, i.e., $R_k^W < R_k^X$ case, the bound of STD is tighter than those of the other methods.

5. Experiment

5.1. Theoretical Validation

Firstly, we verify the theoretical bound derived in Section 4 through experiments for the tensor recovery problem. We generate data tensors by following data generating processes and investigate the relation between a mean squared error (MSE) and other factors. We set $K = 3$ and prepare two different sizes: $(I_1, I_2, I_3) = (10, 10, 20)$ and $(50, 50, 20)$. We set the Tucker rank as (R_1, R_2, R_3) and select R_k from $\{2, 3, 4\}$ for each $k = 1, 2, 3$. In addition, we generate the core tensor and its elements are obtained
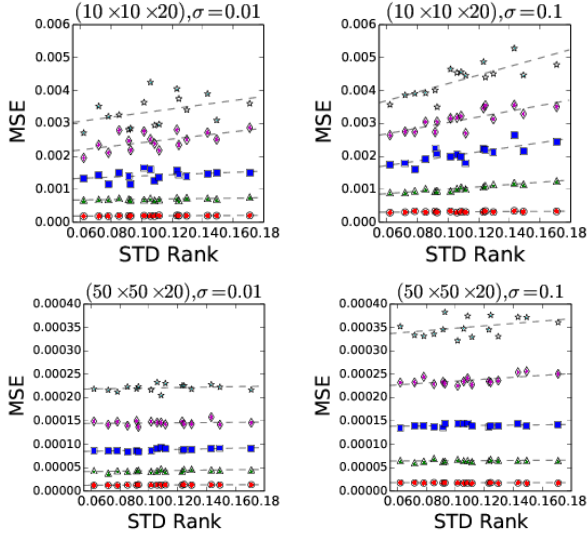
Figure 2. Result for tensor recovery problem. λ_n and μ_n are 0.2 (red circle), 0.4 (green triangle), 0.6 (blue square), 0.8 (purple diamond), and 1.0 (yellow star). Each dashed line is the linear fitting to the errors.

Figure 3. Plotted MSEs with the tensor recovery problem against the noise level. Each symbol shows Schatten (blue circle), STD (red diamond), latent Schatten (purple left triangle), and matrix completion (green triangle).

using the standard normal distribution. Then, we generate vectors $u_{r_k}^{(k)}$ in the following manner and obtain X using (1). To make X smooth, we set $u_{r_k}^{(k)}$ as a discretized smooth functions f_k, i.e., $[u_{r_k}^{(k)}]_i = f_k(i/I_k)$: $\{f_k\}_{k=1,2,3}$ is defined as follows: $f_1(z) = \theta_1 z, f_2(z) = \theta_2 z^2$, and $f_3(z) = \theta_3 z^{0.5}$ with random parameters $(\theta_1, \theta_2, \theta_3)$. The scale of noise is varied as $\sigma \in \{0.01, 0.1\}$.

To investigate the MSE $\|\hat{X} - X^*\|_F^2 / n$, we define the STD rank as follows: STD rank $:= \left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sqrt{I_k}}\right)^2 \left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{R_k^W}\right)^2$. According to the theoretical result, the upper bounds for the MSEs for STD have a linear relation with the STD rank (see Theorem 5).

Figure 2 shows a lot of the MSEs against the STD rank. The results show that the MSE and STD rank have a linear relationship for each panel and each value of the penalty parameters. This result supports Theorem 5; the bound for MSE in Theorem 5 varies linearly with the STD rank. In addition, we can see that the increment in the MSEs against the STD rank increases with the regularization parameter, and it decreases as the size of the tensor increases. This result is explained by the theoretical results, as the MSE is scaled by the regularization parameter and divided by n.

5.2. Comparison with Other Convex Methods

We compare the performances of convex tensor decompositions with the tensor recovery problem. To investigate the performance with smoothness, we generate two types of
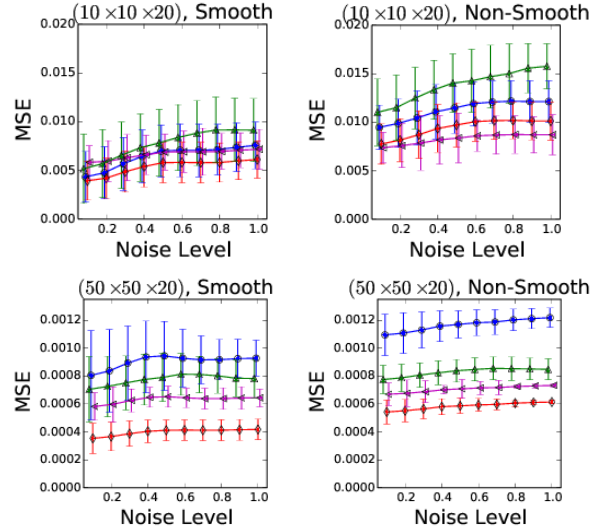
tensors, i.e., a smooth tensor and a non-smooth tensor. The smooth tensor, which is a discretized smooth function, is generated using basis vectors. For the non-smooth tensor, we generate vectors $u_{r_k}^{(k)}$ using a multivariate normal distribution, and make X through (1). The scale of noise is varied as $\sigma \in \{0.1, 0.2 \dots, 1.0\}$.

In the experiment, we compare the following four methods: STD, Tucker decomposition with Schatten regularization, Tucker decomposition with latent Schatten regularization, and matrix decomposition with unfolding X, where the last three methods were proposed by Tomioka et al. (2011). For each method, regularization parameters are selected such that they minimize the generalization error $\||\hat{X} - X^*\||_F$ with grid search in an interval $[0.1, 8.0]$.

The MSEs and their standard deviations for 100 replications are shown in Figure 3. For a small tensor size $(10 \times 10 \times 20)$, STD performs better when the tensor is smooth, and the latent Schatten approach is better when the tensor is not smooth. With the large tensor $(50 \times 50 \times 20)$ with the smooth structure, STD outperforms other methods. For a large tensor is non-smooth, the advantages of STD reduce, even though it exhibits good performance. When the tensor is small, the optimization of STD is close to that of Tucker decomposition, as $M^{(k)}$ and I_k are similar. Thus, the performances of the methods are similar for the small tensor. In contrast, when the tensor is large, STD can provide a different estimator by letting $M^{(k)} \ll I_k$, and STD successfully reduces the MSE. This difference becomes evident when the tensor has the smooth structure.

5.3. Analysis of Real Data

5.3.1. AMINO ACID DATA

We conduct tensor completion and interpolation using amino acids data (Kiers, 1998). The dataset contains amounts of tyrosine, dissolved in phosphate water, which are measured using a spectrofluorometer for each $1\ nm$ interval, and the data are represented by 201×61 matrices.

We make a few elements of the dataset missing, and complete them using the Schatten method (Tomioka et al., 2011) and STD. We consider the following four missing patterns: (A) element-wise missing (20%), (B) element-wise missing (50%), (C) element-wise missing (80%), and (D) column-wise and row-wise missing (50%). We employ the trigonometric basis functions for STD. The hyperparameters are determined through the cross-validation.

Figure 4 shows the result. For random missing cases (A), (B), and (C), the Schatten method and STD can complete the missing elements. In contrast, when rows and columns are completely missing (D), only STD can interpolate missing values and achieve the data.
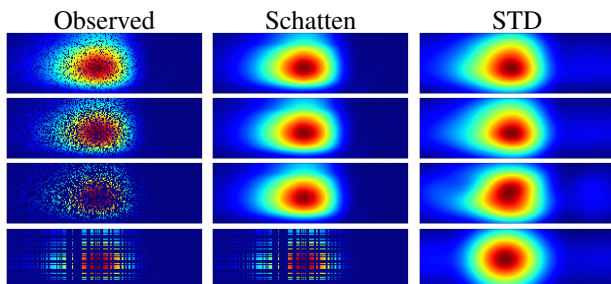

Figure 4. Completion of missing elements in amino acid data.

5.3.2. HUMAN ACTIVITY VIDEO

We conduct an experiment with a human activity video dataset (Schuldt et al., 2004). It contains the human running action; it has a resolution of 160×120 pixels and an average length for 4 seconds length for average with a frame rate of 25 fps. First, we downscale the resolution of each frame to one-fourth so that a $40 \times 30 \times 100$ tensor is obtained. We make a few pixels (50% or 100%) of the video at $t = 0.18\ s$ missing, and complete them from other frames. The experiments are conducted using matrix completion, the Schatten method (Tomioka et al., 2011), and STD. For STD, we set the basis functions as trigonometric series and $\lambda_n = \mu_n = 0.1$.

Figure 5 shows the results of the experiments. We observe that matrix completion and the Schatten method work well when the 50% missing case. However, they recover nothing when 100% pixels are missing. On the contrary, STD successfully recovers the background. In addition, the man's

body and shadow are interpolated at the correct position at t, even though they are blurred. Note that the completion result for STD for 50% missing pixels contains block-noise. This is possibly because of over-fitting by the basis functions.
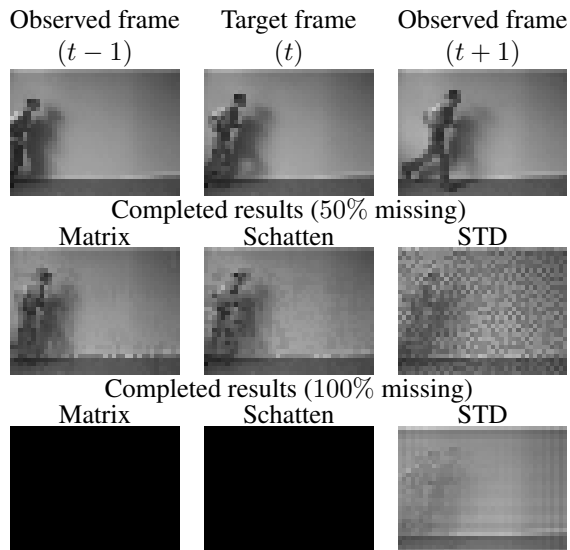

Figure 5. Completion of missing pixels in the human activity video.

6. Discussion

The smoothness we focus on is closely related to the studies of matrix completion. When a tensor is expanded by the basis functions that are close to independent with each other, this implies the tensor satisfies the *incoherence property*, which is frequently used as requirement for the matrix completion problem (Candes & Plan, 2010; Candes & Recht, 2012). Using the similarity, we may apply the formulation of STD to the tensor completion problem and analyze some properties such as the sample complexity.

Note that the focus of this study is primary on the theoretical aspect, which provides scope for addressing more practical requirements. First, the ADMM algorithm is not scalable when the size of the tensor is large. The primary computational burden is caused by a matrix of size $\prod_k I_k \times \prod_k M^{(k)}$, which is essential for the convex formulation. We may use a reduction technique as proposed by Cheng et al. (2016); this is an important challenge for future work. Second, the assumed smoothness (i.e., differentiability) can be extremely general for a few actual applications. For example, images possibly contain solid edges, such as the boundaries of objects, which do not fit the smoothness assumption. Exploring more domain-specific smoothness is an open problem.

References

Amini, Arash A, Levina, Elizaveta, and Shedden, Kerby A. Structured functional regression models for high-dimensional spatial spectroscopy data. *arXiv preprint arXiv:1311.0416*, 2013.

Bickel, Peter J, Ritov, Ya'acov, and Tsybakov, Alexandre B. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pp. 1705–1732, 2009.

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Candes, Emmanuel and Recht, Benjamin. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Candes, Emmanuel J and Plan, Yaniv. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Cheng, Hao, Yu, Yaoliang, Zhang, Xinhua, Xing, Eric, and Schuurmans, Dale. Scalable and sound low-rank tensor learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 1114–1123, 2016.

Gandy, Silvia, Recht, Benjamin, and Yamada, Isao. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

Gantner, Zeno, Drumond, Lucas, Freudenthaler, Christoph, Rendle, Steffen, and Schmidt-Thieme, Lars. Learning attribute-to-feature mappings for cold-start recommendations. In *2010 IEEE International Conference on Data Mining*, pp. 176–185. IEEE, 2010.

Hackbusch, Wolfgang. *Tensor spaces and numerical tensor calculus*, volume 42. Springer Science & Business Media, 2012.

Harshman, Richard A. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 316:1–84, 1970.

Hsing, Tailen and Eubank, Randall. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

Hu, Changwei, Rai, Piyush, and Carin, Lawrence. Zero-truncated poisson tensor factorization for massive binary tensors. *arXiv preprint arXiv:1508.04210*, 2015.

Jia, Chengcheng, Zhong, Guoqiang, and Fu, Yun. Low-rank tensor learning with discriminant analysis for action classification and image recovery. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1228–1234. AAAI Press, 2014.

Kiers, Henk AL. A three-step algorithm for candecomp/parafac analysis of large data sets with multicollinearity. *Journal of Chemometrics*, 12(3):155–171, 1998.

Kolda, Tamara G and Bader, Brett W. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Kolda, Tamara G and Sun, Jimeng. Scalable tensor decompositions for multi-aspect data mining. In *2008 Eighth IEEE international conference on data mining*, pp. 363–372. IEEE, 2008.

Liu, Ji, Musialski, Przemyslaw, Wonka, Peter, and Ye, Jieping. Tensor completion for estimating missing values in visual data. In *ICCV*, 2009.

Narita, Atsuhiro, Hayashi, Kohei, Tomioka, Ryota, and Kashima, Hisashi. Tensor factorization using auxiliary information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 501–516. Springer, 2011.

Negahban, Sahand and Wainwright, Martin J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2): 1069–1097, 2011.

Ramsay, James O. *Functional data analysis*. Wiley Online Library, 2006.

Raskutti, Garvesh, Wainwright, Martin J, and Yu, Bin. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug): 2241–2259, 2010.

Schuldt, Christian, Laptev, Ivan, and Caputo, Barbara. Recognizing human actions: A local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pp. 32–36. IEEE, 2004.

Signoretto, M., Van de Plas, R., De Moor, B., and Suykens, J. A. K. Tensor Versus Matrix Completion: A Comparison With Application to Spectral Data. *Signal Processing Letters, IEEE*, 18(7):403–406, 2011.

Signoretto, Marco, De Lathauwer, Lieven, and Suykens, Johan AK. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications*, 43, 2010.

Tomioka, Ryota and Suzuki, Taiji. Convex tensor decomposition via structured schatten norm regularization. In *Advances in neural information processing systems*, pp. 1331–1339, 2013.

Tomioka, Ryota, Hayashi, Kohei, and Kashima, Hisashi. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

Tomioka, Ryota, Suzuki, Taiji, Hayashi, Kohei, and Kashima, Hisashi. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems*, pp. 972–980, 2011.

Tsybakov, Alexandre B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

Tucker, Ledyard R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Wimalawarne, Kishan, Sugiyama, Masashi, and Tomioka, Ryota. Multitask learning meets tensor factorization: task imputation via convex optimization. In *Advances in neural information processing systems*, pp. 2825–2833, 2014.

Yokota, Tatsuya, Zdunek, Rafal, Cichocki, Andrzej, and Yamashita, Yukihiko. Smooth nonnegative matrix and tensor factorizations for robust multi-way data analysis. *Signal Processing*, 113:234–249, 2015a.

Yokota, Tatsuya, Zhao, Qibin, and Cichocki, Andrzej. Smooth parafac decomposition for tensor completion. *arXiv preprint arXiv:1505.06611*, 2015b.

Zhou, Hua, Li, Lexin, and Zhu, Hongtu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.