Low-Rank Tensor Completion: A Pseudo-Bayesian Learning Approach

Wei Chen Nan Song
State Key Laboratory of Rail Traffic Control and Safety
Beijing Jiaotong University, Beijing, China
{weich,15120129}@bjtu.edu.cn

Abstract

Low rank tensor completion, which solves a linear inverse problem with the principle of parsimony, is a powerful technique used in many application domains in computer vision and pattern recognition. As a surrogate function of the matrix rank that is non-convex and discontinuous, the nuclear norm is often used instead to derive efficient algorithms for recovering missing information in matrices and higher order tensors. However, the nuclear norm is a loose approximation of the matrix rank, and what is more, the tensor nuclear norm is not guaranteed to be the tightest convex envelope of a multilinear rank. Alternative algorithms either require specifying/tuning several parameters (e.g., the tensor rank), and/or have a performance far from reaching the theoretical limit where the number of observed elements equals the degree of freedom in the unknown low-rank tensor. In this paper, we propose a pseudo-Bayesian approach, where a Bayesian-inspired cost function is adjusted using appropriate approximations that lead to desirable attributes including concavity and symmetry. Although deviating from the original Bayesian model, the resulting non-convex cost function is proved to have the ability to recover the true tensor with a low multilinear rank. A computational efficient algorithm is derived to solve the resulting non-convex optimization problem. We demonstrate the superior performance of the proposed algorithm in comparison with state-of-the-art alternatives by conducting extensive experiments on both synthetic data and several visual data recovery tasks.

1. Introduction

In many application domains in computer vision and pattern recognition, we often have to deal with incomplete or noisy data. In addition, multidimensional data arrays, i.e., tensors, are becoming more and more common. For instance, serial image stacks in diffusion magnetic resonance imaging constitute a three-order tensor, and a color video sequence in gait/gesture recognition for surveillance or human-computer interaction is a four-order tensor. Tensor completion aims to recover missing values in a tensor from observed elements, and has many applications such as image/video in-painting, identigram/watermark removal, scan completion, and appearance acquisition completion. However, as a higher order generalization of matrices, the study on tensor completion is far from mature in comparison to matrix completion, partly because most tensor analogues of many efficiently computable problems in numerical linear algebra are NP-hard [7].

To handle high dimensional data that has a large amount of redundancy, one usually resorts to sparsity models, meaning that the model is described by relatively few parameters. One popular model used to capture the parsimony characteristics is the low multilinear rank, where a mode-i rank of a tensor is the matrix rank via unfolding the tensor along the ith mode. Mathematically, the tensor completion problem with a low multilinear rank can be described as

$$\begin{aligned} \min_{\boldsymbol{\mathcal{X}}} \quad & \text{rank}[\boldsymbol{\mathcal{X}}] \\ \text{s.t.} \quad & \boldsymbol{\mathcal{X}}_\Omega = \boldsymbol{\mathcal{Y}}_\Omega, \end{aligned} \tag{1}$$

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times \ldots \times n_k}$, $\text{rank}[\boldsymbol{\mathcal{X}}] = [r_1, \ldots, r_k]$ denotes the multilinear rank of $\boldsymbol{\mathcal{X}}$, and the m observed elements of $\boldsymbol{\mathcal{X}}$ in the set Ω are given by $\boldsymbol{\mathcal{Y}}_\Omega$.

Unfortunately, the non-convexity and discontinuous nature of the rank function make the problem challenging to solve. A widely used approach is to replace matrix rank function by the nuclear norm (i.e., the sum of singular values), which is a convex surrogate of the non-convex matrix rank function. In [13], Liu et al. define a tensor nuclear norm, which is a convex combination of the nuclear norms of all matrices unfolded along different modes. Based on the tensor nuclear norm, a number of methods have been derived [14, 17, 15, 19]. However, the solution of the convex relaxed problem is usually suboptimal as the nuclear norm is a loose approximation of the matrix rank. Furthermore, there is no theoretical guarantee that the tensor nuclear norm is the tightest convex envelope of a tensor rank.

Low-rank tensor completion can also be solved by factorization-based methods, where Tucker decomposition

and Candecomp/Parafac (CP) decomposition have attracted considerable interest. Tucker decomposition represents a k-order tensor by multilinear operations between k factor matrices and a core tensor. In contrast, CP decomposition enforces a strict super-diagonal core tensor. Various factorization-based methods have been developed in literature [23, 8, 9, 27]. For example, Kasai and Mishra propose a Riemannian preconditioning approach based on Tucker decomposition for the tensor completion problem with rank constraints [9]. Zhao et al. develop a Bayesian probabilistic CP factorization method in [27] that uses a hierarchical model and applies variational Bayesian inference for determining the true tensor. Other approximations of the multilinear rank function and pre-defined tensor structural assumptions can be found in [25, 3, 12, 4, 11, 5].

1.1. Motivations

For the existing state-of-the-art algorithms, there is still a large gap between the number of degrees of freedom in the unknown low-rank tensor and the number of observed elements for successful reconstruction. Interestingly, recent work [16] proves that for a low multilinear rank tensor, using nuclear norms along different modes, can do no better, order-wise, than only using a nuclear norm for one of the tensor unfoldings. This result reveals a fundamental limitation imposed by using convex relaxation, and gives the motivation to develop non-convex algorithms. In addition, owing to the multilinear characteristics, it is difficult to derive a fully Bayesian model that leads to an efficient algorithm capable of reaching the theoretical reconstruction limit.

The multilinear rank is usually seen as a natural extension of the matrix rank from a matrix to a higher order tensor. Xin and Wipf propose a low-rank matrix recovery algorithm, namely BARM, that uses a probabilistic PCA-like model [21], resulting in solving a non-convex optimization problem. Surprisingly, BARM is capable to successful recovery at the theoretical limit where the number of observations equals the degrees of freedom in the low-rank matrix. However, the extension of BARM to the tensor case is far from trivial for two reasons. Firstly, the symmetric model of BARM leads to the formulation of the reconstructed matrix as the sum of two matrices, whose ranks are penalized from the column space and row space, respectively. However, a higher order tensor with a low multilinear rank cannot be decomposed in this way. For instance, a three-order tensor $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_1 + \boldsymbol{\mathcal{X}}_2 + \boldsymbol{\mathcal{X}}_3$ may not be a low-rank tensor, even if the unfolding of $\boldsymbol{\mathcal{X}}_i$ $(i = 1, 2, 3)$ along the ith mode is a low rank matrix. Secondly, BARM has a high computational cost owing to the computation of a matrix inverse in a size $m \times m$, where m is the number of observed elements.

In this paper, we propose a pseudo-Bayesian approach, in which a Bayesian-inspired cost function is adjusted using appropriate approximations that lead to desirable attributes.

1.2. Contributions

The contributions of this work are summarized as:

- We propose a pseudo-Bayesian approach, where the cost function is inspired from the probabilistic PCA-like model used in BARM [21] and symmetrized in the form of the Kronecker product to accommodate the higher order tensor. The resulting non-convex cost function is adjusted using appropriate approximations to facilitate the derivation of efficient algorithms. The newly proposed cost function requires no tuning parameters, except for a single standard trade-off parameter to balance data-fit and minimal rank in the noisy case.
- By analysing the cost function, we provide the rationale why it has the ability to recover the true tensor with a low multilinear rank, although the cost function deviates from the original Bayesian model. For solving this optimization problem, we develop an effective and computational efficient algorithm that bypasses the computation of a matrix inverse in a size $m \times m$ and is guaranteed to reduce the cost function or leave it unchanged in each iteration.
- We demonstrate the superior performance of the proposed algorithm in comparison with state-of-the-art alternatives by conducting extensive experiments on both synthetic data and several visual data recovery tasks.

2. Methodology

2.1. Problem Formulation

Bayesian Model: To begin with, we consider a Gaussian likelihood model

$$p(\boldsymbol{\mathcal{Y}}_\Omega|\boldsymbol{\mathcal{X}}) \propto \exp\left[-\frac{1}{2\nu}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2\right], \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^m$ is the observed data vector corresponding to $\boldsymbol{\mathcal{Y}}_\Omega$, $\mathbf{x} = \text{vec}[\boldsymbol{\mathcal{X}}] \in \mathbb{R}^n$ $(n = \prod_{i=1}^k n_i)$ is the vectored data corresponding to the tensor $\boldsymbol{\mathcal{X}}$, and $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a matrix corresponding to the sampling process. The rows of $\mathbf{A}$ can be regarded as m rows of a permuted identity matrix of a size $n \times n$. $\nu \to 0$ will enforce the same constraint as in (1). We adopt a zero-mean multilinear Gaussian prior distribution $p(\mathbf{x}; \mathbf{0}, \bar{\boldsymbol{\Psi}})$ with the covariance constructed by

$$\bar{\boldsymbol{\Psi}} = \boldsymbol{\Psi}_k \otimes \ldots \otimes \boldsymbol{\Psi}_1. \tag{3}$$

where $\boldsymbol{\Psi}_i$ is a positive semi-definite and symmetric matrix, and $\otimes$ denotes the Kronecker product. The covariance in (3) is a symmetric function in the form of Kronecker product and involvs hyperparameters associated to different modes.

Given both likelihood and prior are Gaussian, the posterior $p(\mathbf{x}|\mathbf{y};\bar{\mathbf{\Psi}})$ is also Gaussian, with mean

$$\hat{\mathbf{x}} = \bar{\mathbf{\Psi}}\mathbf{A}^T(\nu\mathbf{I} + \mathbf{A}\bar{\mathbf{\Psi}}\mathbf{A}^T)^{-1}\mathbf{y}. \tag{4}$$

Therefore, the problem boils down to the estimation of $\bar{\mathbf{\Psi}}$. By employing empirical Bayesian approach that treats $\mathbf{x}$ as hidden variables, integrates them out, and conducts maximum a posteriori (MAP) estimation on the hyperparameters, we have

$$\begin{aligned}&\arg\max_{\{\mathbf{\Psi}_i\}} \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x};\bar{\mathbf{\Psi}})d\mathbf{x}\\ =&\arg\min_{\{\mathbf{\Psi}_i\}}\ \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \log|\mathbf{\Sigma}|,\end{aligned} \tag{5}$$

where $\mathbf{\Sigma} = \nu\mathbf{I} + \mathbf{A}\bar{\mathbf{\Psi}}\mathbf{A}^T \in \mathbb{R}^{m\times m}$.

Note that in the case $k = 2$ where the general tensor completion problem collapses to a matrix completion problem, the cost function of this optimization problem differs with the cost function in BARM [21]. The construction of the covariance via (3) is $\bar{\mathbf{\Psi}} = \mathbf{\Psi}_2 \otimes \mathbf{\Psi}_1$, while BARM applies a zero-mean multilinear Gaussian prior distribution with a covariance[1] $\bar{\mathbf{\Psi}} = \mathbf{\Psi}_2 \otimes \mathbf{I}_{n_1} + \mathbf{I}_{n_2} \otimes \mathbf{\Psi}_1$. The newly proposed model exploits the multilinear characteristics, while brings new challenges in the derivation of efficient algorithms.

As the log-determinant function is a concave non-decreasing function of the singular values of symmetric positive definite matrices, and thus the term $\log|\mathbf{\Sigma}|$ in (5) favors minimal rank of $\bar{\mathbf{\Psi}}$. According to (4) and properties of Kronecker product, the rank of the mode i unfolding of $\hat{\mathcal{X}}$ is equal to the rank of $\mathbf{\Psi}_i$. Therefore, the empirical Bayesian estimate of the tensor is likely to be low-rank, where the mode i rank is determined by $\mathbf{\Psi}_i$. However, solving the problem in (5) is difficult in part because the Kronecker structure in $\bar{\mathbf{\Psi}}$, together with the fact that the dimensions of $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ are huge even for reasonably sized problems. To alleviate this problem, we will need certain approximations that lead to affordable update rules.

Pseudo-Bayesian Objective: By using determinant identities, the log-determinant term in the cost function of (5) can be rewritten as

$$\log|\mathbf{\Sigma}| = \log|\bar{\mathbf{\Psi}}| + \log|\mathbf{A}^T\mathbf{A} + \nu\bar{\mathbf{\Psi}}^{-1}|. \tag{6}$$

To alleviate the difficulties in solving (5), we first use a full-rank diagonal approximation to $\mathbf{A}^T\mathbf{A}$, which leads to the penalty function

$$\begin{aligned}&\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \log|\bar{\mathbf{\Psi}}| + \log|\beta\mathbf{I}_n + \nu\bar{\mathbf{\Psi}}^{-1}|\\ =&\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \log|\nu\mathbf{I}_n + \beta\bar{\mathbf{\Psi}}|,\end{aligned} \tag{7}$$

where the constant $\beta = \frac{m}{n}$ so that the approximation satisfies $\mathbb{E}(\mathbf{A}^T\mathbf{A}) = \beta\mathbf{I}_n$ (by assuming $\mathbf{A}$ as a random sampling matrix). Furthermore, we approximate the matrix determinant $|\nu\mathbf{I} + \beta\bar{\mathbf{\Psi}}|$ by

$$\lim_{\nu\to 0}|\nu\mathbf{I}_n + \beta\bar{\mathbf{\Psi}}| = \lim_{\nu\to 0}\left|\prod_{i=1}^{k}\otimes\left(\nu^{\frac{1}{k}}\mathbf{I}_{n_i} + \beta^{\frac{1}{k}}\mathbf{\Psi}_i\right)\right|, \tag{8}$$

where $\prod_{i=1}^{k}\otimes$ denotes the Kronecker product of k components. This approximation is also accurate in the noisy case with $\nu \to \infty$. Now, we obtain the simplified cost function as

$$\begin{aligned}\mathcal{L}(\{\mathbf{\Psi}_i\}) &= \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \log\left|\prod_{i=1}^{k}\otimes(\nu^{\frac{1}{k}}\mathbf{I}_{n_i} + \beta^{\frac{1}{k}}\mathbf{\Psi}_i)\right|\\ &= \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \sum_{i=1}^{k}\frac{n}{n_i}\log|\nu^{\frac{1}{k}}\mathbf{I}_{n_i} + \beta^{\frac{1}{k}}\mathbf{\Psi}_i|,\end{aligned} \tag{9}$$

where the second equality is obtained by using properties of Kronecker product.

Although this simplified cost function deviates from the original Bayesian model owing to the approximation and cannot be justified from formal probabilistic terms, we will explain shortly that it is a viable cost function that has the ability to recover the true tensor. In addition, with this approximation, the Kronecker product in the log-determinant $\log|\mathbf{\Sigma}|$ is dissolved, and dimensionality of the matrix inside the log-determinant is reduced significantly in this simplified cost function.

2.2. Optimization

To solve the non-convex optimization problem in (9), we employ several upper bounds and the cost function can be minimized using coordinate descent method, which leads to an algorithm with computational efficient update rules.

2.2.1 Update $\mathbf{x}$

The first term in (9) can be upper bounded by

$$\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} \le \frac{1}{\nu}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mathbf{x}^T\bar{\mathbf{\Psi}}^{-1}\mathbf{x}, \tag{10}$$

where the equality in (10) holds if

$$\mathbf{x} = \bar{\mathbf{\Psi}}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{y}. \tag{11}$$

This standard upper bound given in (10) has been used in deriving update rules for various Bayesian learning algorithms [2, 21, 20]. However, the resulting update rule (11) involves a matrix inverse operation which has a computational complexity of $\mathcal{O}(m^3)$ and prohibits its application to recover data of high dimensionality, e.g., tensors.

[1] Generalizing the BARM for a three-order tensor by using an additive component covariance model ($\bar{\mathbf{\Psi}} = \mathbf{\Psi}_3 \otimes \mathbf{I}_{n_2} \otimes \mathbf{I}_{n_1} + \mathbf{I}_{n_3} \otimes \mathbf{\Psi}_2 \otimes \mathbf{I}_{n_1} + \mathbf{I}_{n_3} \otimes \mathbf{I}_{n_2} \otimes \mathbf{\Psi}_1$) fails, as the reconstructed tensor given in (5) is not necessary low-rank, even if all $\mathbf{\Psi}_i$ have low ranks.

To overcome this drawback, we consider to further bound the righthand side of (10) using the fundamental property of continuously differentiable function [1]. From Taylor's theorem, for some $\mathbf{b}$ we have

$$\begin{aligned}\mathcal{F}(\mathbf{x}) &= \frac{1}{\nu}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 \\ &= \mathcal{F}(\mathbf{z}) + (\mathbf{x}-\mathbf{z})^T\nabla\mathcal{F}(\mathbf{z}) + \frac{1}{2}(\mathbf{x}-\mathbf{z})^T\nabla^2\mathcal{F}(\mathbf{b})(\mathbf{x}-\mathbf{z}).\end{aligned} \tag{12}$$

As the function $\mathcal{F}(\mathbf{x})$ is strongly-convex, its gradient is Lipschitz-continuous, i.e.,

$$\nabla^2\mathcal{F}(\mathbf{x}) \leq L_F, \tag{13}$$

for some L_F. Then the righthand side of (10) can be upper bounded as

$$\begin{aligned}&\frac{1}{\nu}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \mathbf{x}^T\bar{\mathbf{\Psi}}^{-1}\mathbf{x} \\ \leq&\mathcal{F}(\mathbf{z}) + (\mathbf{x}-\mathbf{z})^T\nabla\mathcal{F}(\mathbf{z}) + \frac{L}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 + \mathbf{x}^T\bar{\mathbf{\Psi}}^{-1}\mathbf{x} \\ =&\frac{1}{\nu}\big(\mathbf{y}^T\mathbf{y} + 2\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{z} - \mathbf{z}^T\mathbf{A}^T\mathbf{A}\mathbf{z} - 2\mathbf{y}^T\mathbf{A}\mathbf{x}\big) \\ &+ \frac{L}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 + \mathbf{x}^T\bar{\mathbf{\Psi}}^{-1}\mathbf{x},\end{aligned} \tag{14}$$

where $L \leq L_F$. Obviously, the equality in (14) holds with $\mathbf{x}=\mathbf{z}$. The largest L, i.e., the (smallest) Lipschitz constant of the gradient $\nabla\mathcal{F}$, is $L_F = 2\lambda_{\max}[\frac{1}{\nu}\mathbf{A}^T\mathbf{A}] = \frac{2}{\nu}$.

Insert the upper bounds (10) and (14) into the cost function (9). With irrelevant terms omitted, the estimate of $\mathbf{x}$ can be updated by solving the following optimization problem

$$\min_{\mathbf{x}} \frac{1}{\nu}\big(2\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{z} - 2\mathbf{y}^T\mathbf{A}\mathbf{x}\big) + \frac{1}{\nu}\|\mathbf{x}-\mathbf{z}\|_2^2 + \mathbf{x}^T\bar{\mathbf{\Psi}}^{-1}\mathbf{x}, \tag{15}$$

which has a closed-form expression

$$\mathbf{x} = \left(\bar{\mathbf{\Psi}}^{-1} + \frac{1}{\nu}\mathbf{I}\right)^{-1}\left(\frac{1}{\nu}\mathbf{z} + \frac{1}{\nu}(\mathbf{A}^T\mathbf{y} - \mathbf{A}^T\mathbf{A}\mathbf{z})\right). \tag{16}$$

The matrix inverse $(\bar{\mathbf{\Psi}} + \frac{1}{\nu}\mathbf{I})^{-1}$ can be computed efficiently by

$$\left(\bar{\mathbf{\Psi}}^{-1} + \frac{1}{\nu}\mathbf{I}\right)^{-1} = \mathbf{U}\left(\frac{1}{\nu}\mathbf{I} + \mathbf{D}_k^{-1}\otimes\ldots\otimes\mathbf{D}_1^{-1}\right)^{-1}\mathbf{U}^T, \tag{17}$$

where $\mathbf{U} = \mathbf{U}_k\otimes\ldots\otimes\mathbf{U}_1$, $\mathbf{U}_i$ and $\mathbf{D}_i$ $(i=1,\ldots,k)$ come from the eigendecomposition of the positive semi-definite and symmetric matrix $\mathbf{\Psi}_i$, i.e., $\mathbf{U}_i\mathbf{D}_i\mathbf{U}_i^T = \mathbf{\Psi}_i$, and the matrix $\mathbf{D}_i$ is diagonal. Therefore the term $\frac{1}{\nu}\mathbf{I}+\mathbf{D}_k^{-1}\otimes\ldots\otimes\mathbf{D}_1^{-1}$ is also diagonal and its inverse can be easily obtained. The computational complexity is now reduced to that of the eigendecompositions, i.e., $\mathcal{O}(\sum_{i=1}^k n_i^3)$.

2.2.2 Update $\mathbf{\Psi}_i$

Now we consider upper bounds for the log-determinant terms of the cost function (9). As the log-determinant terms are concave non-decreasing functions of the singular values of symmetric positive definite matrices, we define the concave conjugate functions

$$g_i(\mathbf{W}_i) = \min_{\mathbf{\Psi}_i} \mathrm{Tr}[\mathbf{W}_i^T\mathbf{\Psi}_i^{-1}] - \log|\mathbf{\Psi}_i^{-1} + (\beta/\nu)^{\frac{1}{k}}\mathbf{I}_{n_i}|, \tag{18}$$

where $i=1,\ldots,k$. According to the duality relationship of concave conjugate functions, we have the following upper bound:

$$\log|\mathbf{\Psi}_i^{-1} + (\beta/\nu)^{\frac{1}{k}}\mathbf{I}_{n_i}| = \min_{\mathbf{W}_i} \mathrm{Tr}[\mathbf{W}_i^T\mathbf{\Psi}_i^{-1}] - g_i(\mathbf{W}_i), \tag{19}$$

where the bound is tight when

$$\mathbf{W}_i = (\mathbf{\Psi}_i^{-1} + (\beta/\nu)^{\frac{1}{k}}\mathbf{I}_{n_i})^{-1}. \tag{20}$$

Inserting the upper bound (19) into the cost function (9) and considering $\mathbf{\Psi}_i$ related terms in the upper bounds, we arrive at the following approximation

$$\min_{\mathbf{\Psi}_i} \mathbf{x}^T\bar{\mathbf{\Psi}}^{-1}\mathbf{x} + \frac{n}{n_i}\big(\mathrm{Tr}[\mathbf{W}_i^T\mathbf{\Psi}_i^{-1}] + \log|\mathbf{\Psi}_i|\big), \tag{21}$$

and its solution is

$$\mathbf{\Psi}_i = \mathbf{W}_i + \frac{n_i}{n}\mathbf{Q}_{(i)}\mathbf{X}_{(i)}^T, \tag{22}$$

where $\mathbf{X}_{(i)}$ is the ith mode unfolding of the tensor $\boldsymbol{\mathcal{X}} = \mathrm{Fold}[\mathbf{x}, n_1, \ldots, n_k]$. Here $\mathrm{Fold}[\mathbf{x}, n_1, \ldots, n_k]$ denotes the transform of a vector into a tensor of a size $n_1\times\ldots\times n_k$. $\mathbf{Q}_{(i)}$ is the ith mode unfolding of the tensor

$$\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{X}}\times_1\mathbf{\Psi}_1^{-1}\ldots\times_{i-1}\mathbf{\Psi}_{i-1}^{-1}\times_i\mathbf{I}\times_{i+1}\mathbf{\Psi}_{i+1}^{-1}\ldots\times_k\mathbf{\Psi}_k^{-1}, \tag{23}$$

where $\times_i$ denotes the multiplication at mode i. Note that all the $\mathbf{\Psi}_i$ iterates are positive semi-definite, if all $\mathbf{\Psi}_i$ are initialized as positive semi-definite symmetric matrices.

2.2.3 Convergence

By iteratively cycling through each of the above subproblems, we arrive at Algorithm 1. Although each iteration of the proposed algorithm is guaranteed to reduce or leave the cost function (9) unchanged, it is insufficient to guarantee formal convergence to a stationary point. Deriving a theoretical guarantee for the proposed algorithm is difficult, as it requires, for example, that the additional conditions of the Zangwills Global Convergence Theorem hold [24]. Instead, we provide empirical evidence to demonstrate the feasibility and applicability of the proposed method in Section 3.

Algorithm 1 The proposed algorithm for tensor recovery with low multilinear ranks

Step 1: Initialize $\mathbf{\Psi}_i = \mathbf{I}\ \forall i$ and $\mathbf{z} = \mathbf{0}$;

Step 2: Update $\mathbf{x}$ using (16) and let $\mathbf{z} = \mathbf{x}$ (this step could be repeated for a certain number of times);

Step 3: For each mode i, compute $\mathbf{W}_i$ using (20) and update $\mathbf{\Psi}_i$ using (22);

Step 4: Iterate steps 2 and 3 until convergence.

2.3. Analysis of the Cost Function

In this subsection, we provide the rationale why the cost function (9) has the ability to recover the true tensor with a low multilinear rank, although the cost function deviates from the original Bayesian model.

Firstly, the log-determinant function is a concave non-decreasing function of the singular values of symmetric positive definite matrices, and thus the cost function (9) favors minimal rank of $\mathbf{\Psi}_i$. The log-determinant function is strongly concave, which avoids over-penalize large singular values in comparison to the nuclear norm. In addition, by rewriting the update rule (16) as

$$\mathbf{x} = \bar{\mathbf{\Psi}} \left(\mathbf{I} + \frac{1}{\nu} \bar{\mathbf{\Psi}} \right)^{-1} \left(\frac{1}{\nu} \mathbf{z} + \frac{1}{\nu} (\mathbf{A}^T \mathbf{y} - \mathbf{A}^T \mathbf{A} \mathbf{z}) \right), \tag{24}$$

we note that $\mathbf{X}_{(i)}$, i.e., the ith mode folding matrix of $\boldsymbol{\mathcal{X}}$, results from a left-multiplication with $\mathbf{\Psi}_i$ (according to properties of the Kronecker product). Thus, if $\mathbf{\Psi}_i$ is a low rank matrix, $\mathbf{X}_{(i)}$ must be low-rank as well.

Ideally, for a tensor that has a low multilinear rank, it is expected that the one with the lowest multilinear rank should be the solution that minimizes the cost function (at least in the noiseless case). In the following result (Theorem 1), we show that the global minima of the cost function (9) produces the solution with the lowest multilinear rank. Note that as the original low rank tensor completion problem involves multiple objectives, i.e., ranks along different modes, the global minima of the cost function (9) is optimal for the scalarized problem, i.e., minimum $\sum_{i=1}^{k} \frac{r_i}{n_i}$, where r_i denotes the mode i rank of any feasible solution.

Theorem 1 *Let* $\mathbf{y} = \mathbf{A}\mathbf{x}$*, and define* r_i *as the mode* i *rank of any feasible solution that leads to the smallest* $\sum_{i=1}^{k} \frac{r_i}{n_i}$*. Then the global minima of the cost function* $\lim_{\nu \to 0} \mathbf{y}^T \mathbf{\Sigma}^{-1} \mathbf{y} + \sum_{i=1}^{k} \frac{n}{n_i} \log |\nu^{\frac{1}{k}} \mathbf{I}_{n_i} + \beta^{\frac{1}{k}} \mathbf{\Psi}_i|$ *is achieved at* $\{\hat{\mathbf{\Psi}}_i\}_{i=1}^{k}$ *such that* $\hat{\mathbf{x}} = \hat{\bar{\mathbf{\Psi}}} \mathbf{A}^T (\mathbf{A} \hat{\bar{\mathbf{\Psi}}} \mathbf{A}^T)^{-1} \mathbf{y}$ *and rank*$[\hat{\mathbf{X}}_{(i)}] = r_i$.

Proof: The following proof is based on the result of Lemma 1 in [21], which considers a low-rank matrix recovery problem. However, for tensor completion with a low multilinear rank, some modifications are required.

In the limit $\nu \to 0$, a minimizer of the cost function in (9) must satisfy $\mathbf{y} \in \text{span}[\mathbf{\Sigma}^{\frac{1}{2}}]$, otherwise the cost function would diverge to infinity as $\mathbf{y}\mathbf{\Sigma}^{-1}\mathbf{y}$ tends to be infinity with a faster rate than the log-determinant terms approaching minus infinity. The constraint $\mathbf{y} \in \text{span}[\mathbf{\Sigma}^{\frac{1}{2}}]$ is equivalent to requiring

$$\mathbf{y}^T (\nu \mathbf{I} + \mathbf{A} \bar{\mathbf{\Psi}} \mathbf{A}^T)^{-1} \mathbf{y} \leq \rho,$$

where $\rho > 0$ denotes some finite bound.

While $\mathbf{y}(\nu \mathbf{I} + \mathbf{A} \bar{\mathbf{\Psi}} \mathbf{A}^T)^{-1} \mathbf{y}$ is bounded, the minimum occurs when the log-determinant terms are approaching minus infinity. The sum of the log-determinant terms is given by

$$\begin{aligned} &\sum_{i=1}^{k} \frac{n}{n_i} \log |\nu^{\frac{1}{k}} \mathbf{I}_{n_i} + \beta^{\frac{1}{k}} \mathbf{\Psi}_i| \\ =&\sum_{i=1}^{k} \frac{n}{n_i} \left(\sum_{h=1}^{r_i} \log \left(\nu^{\frac{1}{k}} + \beta^{\frac{1}{k}} \sigma_h[\mathbf{\Psi}_i] \right) + \frac{n_i - r_i}{k} \log |\nu| \right), \end{aligned} \tag{25}$$

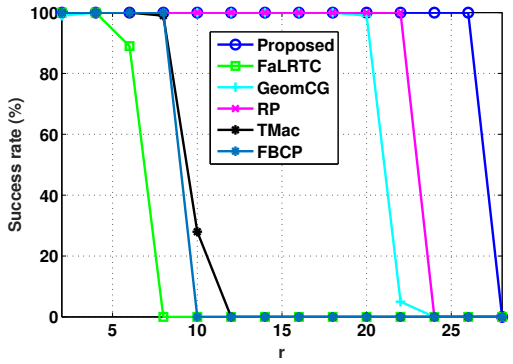
where $\sigma_h[\cdot]$ denotes the hth singular value of a matrix. Consequently, when $\nu \to 0$, the sum of log-determinant terms scales as $\frac{n}{k} \sum_{i=1}^{k} (1 - \frac{r_i}{n_i}) \log |\nu|$, and hence the overall cost function is minimized when $\sum_{i=1}^{k} \frac{r_i}{n_i}$ achieves its minimum. Now we complete the proof.

3. Experimental Validation

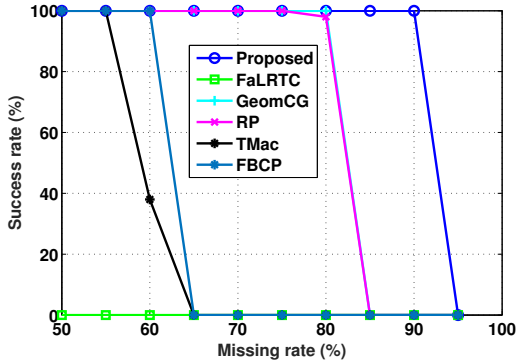
In this section we compare the proposed algorithm with existing state-of-the-art algorithms, which include GomCG [10], TMac [22], FBCP [27] and RP [9], i.e., decomposition based algorithms, and FaLRTC [14], i.e., a nuclear norm minimization algorithm. All the code of compared algorithms is available from the original authors. Note that our focus here is on algorithms that are based on the low multilinear rank model and hence we do not include comparison with many algorithms that exploit structural knowledge from specific applications. For instance, GTV [6] considers a total variation model for visual tensor to enforce piece-wise smooth. We validate the proposed algorithm by both synthetic data and real visual data. Our simulations are performed in MATLAB R2014a environment on a system with a dual-core 3.4 GHz CPU and 16 GB RAM, running under the Microsoft Windows 7 operating system.

3.1. Experiments With Synthetic Data

In the experiments with synthetic data, we use the Tucker model, i.e., $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \mathbf{V}_1 \ldots \times_k \mathbf{V}_k$, to generate



(a) With different ranks (missing rate 80%)



(b) With different missing rates ($r = 18$)

Figure 1. Comparison of reconstruction success rate in the noiseless case ($r_1 = r_2 = r_3 = r$).

the ground truth tensor. First, elements of the core tensor $\mathcal{C} \in \mathbb{R}^{r_1 \times \ldots \times r_k}$ and all $\mathbf{V}_i \in \mathbb{R}^{r_i \times n_i}$ $(i = 1, \ldots, k)$ are generated independently from $\mathcal{N}(0, 1)$. Then the tensor is normalized so that $\|\boldsymbol{\mathcal{X}}\|_F^2 = \prod_{i=1}^{k} n_i$. A portion of elements are randomly chosen as observed data while the rest are left as missing components. In each iteration of the proposed algorithm, the step 2 is repeated with 20 times.

3.1.1 Noiseless Case

We begin with the noiseless case where the tensor is exactly low-rank. We set $n_i = 50$ $(i = 1, 2, 3)$, resulting in the ground truth tensor with size $50 \times 50 \times 50$. The recovery performance is evaluated via relative recovery error defined by $\frac{\|\hat{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}\|_F}{\|\boldsymbol{\mathcal{X}}\|_F}$, and averaged over 100 trials. If the relative recover error is smaller than 10^{-3}, $\hat{\boldsymbol{\mathcal{X}}}$ is regarded as a successful recovery of $\boldsymbol{\mathcal{X}}$. The proposed algorithm is stopped when either the change of $\boldsymbol{\mathcal{X}}$ in an iteration is below 10^{-8} or the number of iterations exceeds 1000.

In the first experiment, we consider a symmetric rank setting, i.e., $r_1 = r_2 = r_3 = r$, for the rank parameter r_i along each mode. Fig. 1 (a) and fig. 1 (b) show how the proposed algorithm perform in different ranks and different missing rates, respectively. As shown in the figure, our proposed algorithm outperforms the existing state-of-the-art alternatives. It is the only algorithm that is able to recover the tensor with 85% missing data. FaLRTC uses tensor nuclear norm as an approximation of the multilinear rank. Although FaLRTC is computational efficient, its performance is poor owing to the approximation. As a tuning parameter-free approach that uses Bayesian inference, FBCP has little performance improvement in comparison to FaLRTC.

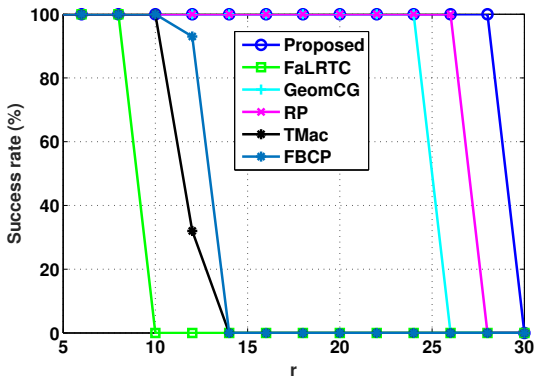


Figure 2. Comparison of reconstruction success rate in the noiseless case ($2r_1 = r_2 = r_3 = r$, and missing rate 80%).

To investigate instances where the ranks along certain modes are different than others, we set $2r_1 = r_2 = r_3 = r$. As shown in fig. 2, the proposed algorithm has the highest reconstruction accuracy among all the algorithms.

3.1.2 Noisy Case

In this experiment we investigate how the proposed algorithm performs if the data is corrupted by noise. The ground truth tensor with size $50 \times 50 \times 50$ and rank $r_1 = r_2 = r_3 = 18$ is randomly generated as the previous experiments. Then an additive noise tensor is produced, where elements are generated following a zero-mean Gaussian distribution with variance adjusted to have a desired value of the signal to noise ratio (SNR). In this experiment, we follow a heuristic strategy introduced in [26] to adaptively setting ν. We simply set $\nu = 0.1$ and reduce the value by $\nu = 0.98\nu$ for each iteration. The proposed algorithm is stopped when either the change of $\boldsymbol{\mathcal{X}}$ in an iteration is below 10^{-8} or the number of iterations exceeds a threshold. The relative recovery error is averaged over 100 trials. Results are shown in fig. 2, where the proposed algorithm exhibits superior reconstruction accuracy in comparison to all the competitors.

3.1.3 Computing Time

The averaged quantitative results in terms of recovery performance and runtime is given in Table 1, where the ground truth tensor is $50 \times 50 \times 50$ with rank $r_1 = r_2 = r_3 =$

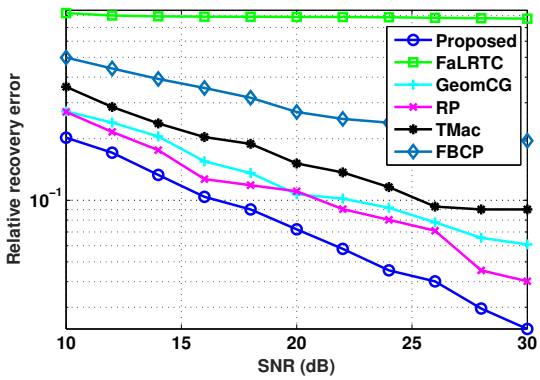


Figure 3. Comparison of reconstruction accuracy in the noisy case.

15, and is corrupted by additive zero-mean Gaussian noise yielding an SNR of 20 dB. For different data missing rates, the proposed algorithm consistently achieves the best recovery accuracy, although it is not the one with the shortest runtime. Among all the compared algorithms, FBCP has the worst runtime performance according to our investigation.

3.2. Experiments With Real Data

Now we evaluate the proposed algorithm by considering two real-world applications including hyperspectral image inpainting and facial image synthesis. Although the main focus here is on the comparison of low-rank based algorithms, we still include one additional algorithm, i.e., S-DTC [3], for comparison, which takes into account the local similarity in addition to the low rank assumption. Furthermore, as suggested in [27], FBCP is also adjusted to integrate a Gaussian mixture factor prior that takes into account structures in visual data.

3.2.1 Hyperspectral Image

In the first experiment with real data, we complete missing information in hyperspectral images, which are three-order tensors with each slice corresponding to an image of a particular scene measured at a different wavelength. As suggested in [9], all hyperspectral images are resized to $204 \times 268 \times 33$. We randomly remove 80% data and using the remaining 20% for completion. To enhance visualization, hyperspectral images and the reconstructed ones are transformed to RGB colour images. As shown in fig. 4, our approach outperforms the other algorithms in terms of the performance of the relative recovery error. SDTC and F-BCP produce images that are smooth owing to the use of more complex models, and thus the visual quality of the recovered images is not too bad, although their relative recovery errors are much higher than our algorithm. We envisage that by incorporating additional visual structures, our algorithm may have a better visual quality, while doing so is out

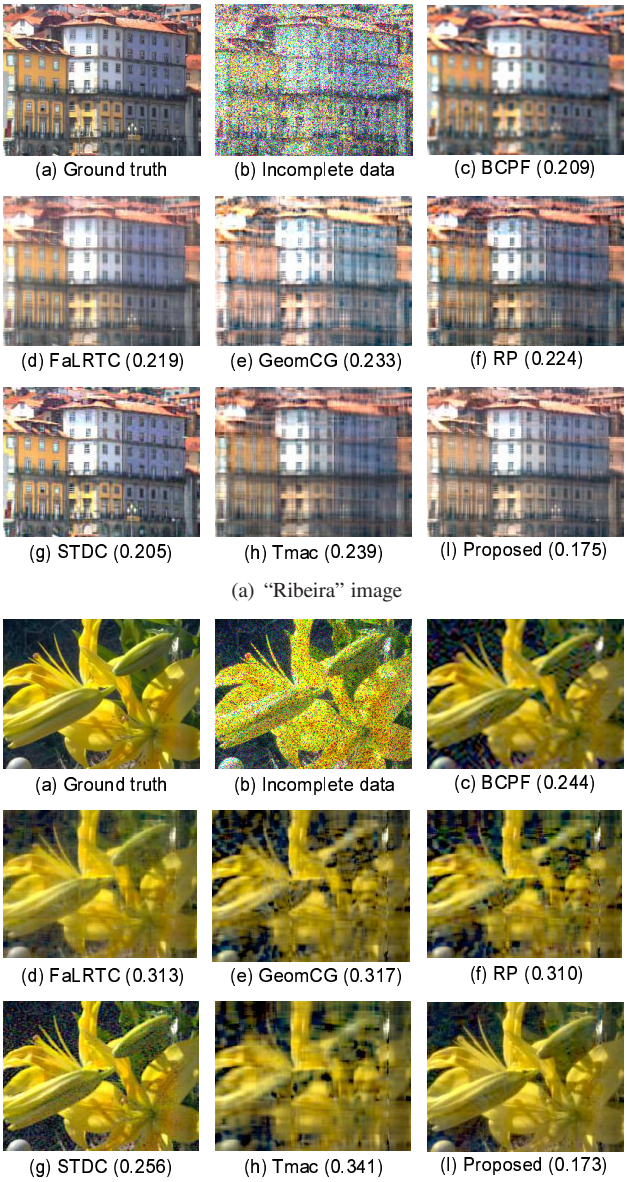


Figure 4. Visual effects of hyperspectral image inpainting (the relative recovery error is given in brackets).

of the scope this paper.

3.2.2 CMU-PIE Face Database

In this experiment, we aim to generate novel facial images under multiple conditions (e.g., poses and illumination changes) given images under other conditions. The recovered facial images can be used to create a robust classifier in applications such as face recognition from surveillance videos, where a complete training set is not available. We use the CMU-PIE Face Database [18], where the facial images are aligned by eye positions and cropped to size

Missing rates	Performance	FaLRTC	geomCG	TMac	RP	FBCP	Proposed
70%	Relative recovery error	0.4955	0.0490	0.0552	0.0570	0.2753	0.0533
	Computing time	0.9	24.3	0.3	24.0	88.9	7.3
80%	Relative recovery error	0.7455	0.0667	0.1541	0.0579	0.0819	0.0604
	Computing time	0.6	17.2	0.3	12.6	17.4	8.6
90%	Relative recovery error	0.9160	1.9103	0.5367	0.1897	2.3944	0.1025
	Computing time	0.5	11.3	0.3	8.0	11.1	10.8

Table 1. The averaged recovery performance and computing time (seconds) with missing rates of 70, 80 and 90 Percent.

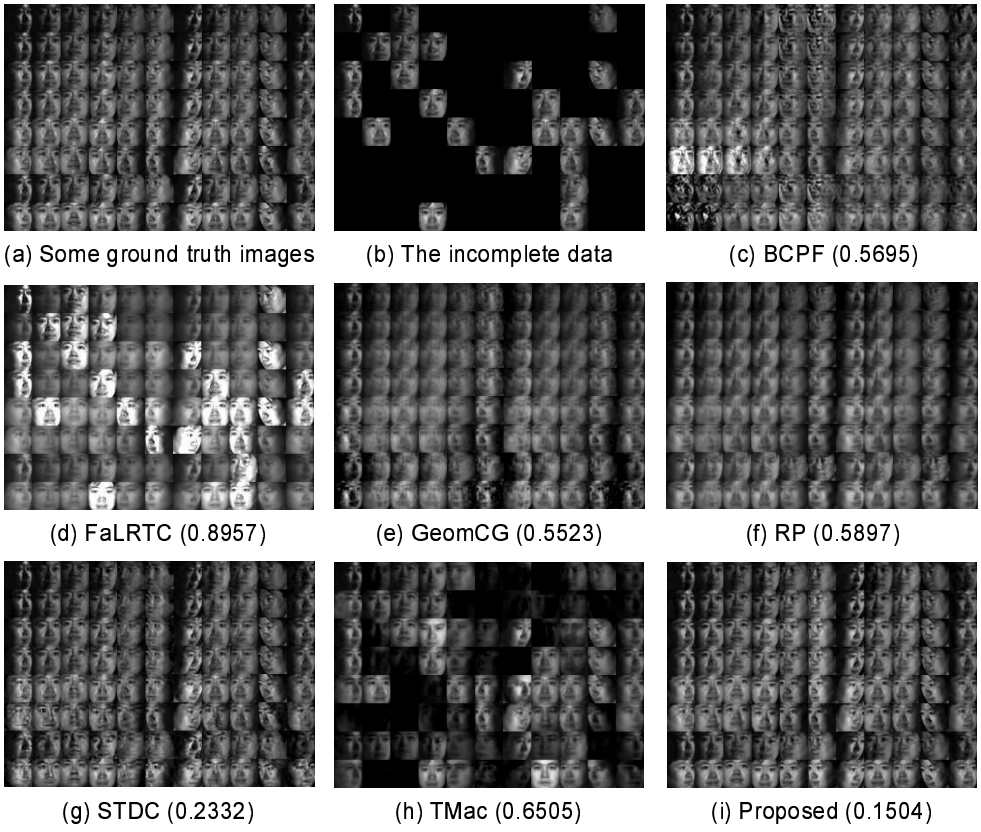


Figure 5. CMU-PIE face completion (the relative recovery error is given in brackets).

32×32. We use a subset selected from the first 30 people, each rendered in 11 different poses under 21 different illumination changes. Each image is vectored, as facial images do not possess an intrinsic low-rank structure. Thus, we construct a fourth order tensor $\mathcal{X} \in \mathbb{R}^{30\times11\times21\times1024}$. We assume 80% facial images are fully missing, and compare the proposed algorithm with FBCP, FaLRTC, GeomCG, R-P, STDC and TMac. Results are reported in fig. 5, where the visual quality of image synthesis obtained by the proposed algorithm is significantly superior to those by other methods.

4. Conclusion

In this paper, we present a pseudo-Bayesian learning approach for low-rank tensor completion, where a Bayesian-inspired cost function is adjusted using appropriate approximations that lead to desirable attributes, i.e., global minima, concavity and symmetry. We demonstrate the superior performance of the proposed algorithm in comparison with state-of-the-art alternatives by conducting extensive experiments on both synthetic data and several visual data recovery tasks.

5. Acknowledgement

This work is supported by the Natural Science Foundation of China (61671046, 61401018).

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[2] W. Chen, D. Wipf, Y. Wang, Y. Liu, and I. J. Wassell. Simultaneous bayesian sparse approximation with structured sparse models. *IEEE Transactions on Signal Processing*, 64(23):6145–6159, Dec 2016.

[3] Y. L. Chen, C. T. Hsu, and H. Y. M. Liao. Simultaneous tensor decomposition and completion using factor priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):577–591, March 2014.

[4] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[5] D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.

[6] X. Guo and Y. Ma. Generalized tensor total variation minimization for visual data recovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3603–3611, June 2015.

[7] C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM*, 60(6):45, 2013.

[8] P. Jain and S. Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

[9] H. Kasai and B. Mishra. Low-rank tensor completion: a riemannian manifold preconditioning approach. In *The 33rd International Conference on Machine Learning*, pages 1012–1021, 2016.

[10] D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by riemannian optimization. *Bit Numerical Mathematics*, 54(2):447–468, 2014.

[11] A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 836–844, 2013.

[12] M. Li, J. Liu, Z. Xiong, X. Sun, and Z. Guo. Marlow: A joint multiplanar autoregressive and low-rank approach for image completion. In *European Conference on Computer Vision*, pages 819–834, 2016.

[13] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. In *IEEE International Conference on Computer Vision*, pages 2114–2121, Sept 2009.

[14] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, Jan 2013.

[15] 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 Conference on Computer Vision and Pattern Recognition*, pages 5249–5257, June 2016.

[16] S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, May 2015.

[17] H. Rauhut and Z. Stojanac. Recovery of third order tensors via convex optimization. In *2015 International Conference on Sampling Theory and Applications*, pages 397–401, May 2015.

[18] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, Dec 2003.

[19] R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. *Mathematics*, 2010.

[20] D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, April 2010.

[21] B. Xin and D. Wipf. Pushing the limits of affine rank minimization by adapting probabilistic PCA. pages 419–427, 2015.

[22] Y. Xu, R. Hao, W. Yin, and Z. Su. Parallel matrix factorization for low-rank tensor completion. *Inverse Problems & Imaging*, 9(2), 2013.

[23] Z. Xu, F. Yan, and Y. Qi. Bayesian nonparametric models for multiway data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):475–487, Feb 2015.

[24] W. I. Zangwill. *Nonlinear programming: a unified approach*, volume 196. Prentice-Hall Englewood Cliffs, NJ, 1969.

[25] D. Zhang, Y. Hu, J. Ye, X. Li, and X. He. Matrix completion by truncated nuclear norm regularization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2192–2199, June 2012.

[26] Q. Zhao, D. Meng, X. Kong, Q. Xie, W. Cao, Y. Wang, and Z. Xu. A novel sparsity measure for tensor recovery. In *IEEE International Conference on Computer Vision*, pages 271–279, Dec 2015.

[27] Q. Zhao, L. Zhang, and A. Cichocki. Bayesian cp factorization of incomplete tensors with automatic rank determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1751–1763, Sept 2015.