A Fused CP Factorization Method for Incomplete Tensors

Yuankai Wu, Huachun Tan, *Member, IEEE*, Yong Li, Jian Zhang, *Member, IEEE*, and Xiaoxuan Chen

***Abstract*— Low-rank tensor completion methods have been advanced recently for modeling sparsely observed data with a multimode structure. However, low-rank priors may fail to interpret the model factors of general tensor objects. The most common method to address this drawback is to use regularizations together with the low-rank priors. However, due to the complex nature and diverse characteristics of real-world multiway data, the use of a single or a few regularizations remains far from efficient, and there are limited systematic experimental reports on the advantages of these regularizations for tensor completion. To fill these gaps, we propose a modified CP tensor factorization framework that fuses the l_2 norm constraint, sparseness (l_1 norm), manifold, and smooth information simultaneously. The factorization problem is addressed through a combination of Nesterov's optimal gradient descent method and block coordinate descent. Here, we construct a smooth approximation to the l_1 norm and TV norm regularizations, and then, the tensor factor is updated using the projected gradient method, where the step size is determined by the Lipschitz constant. Extensive experiments on simulation data, visual data completion, intelligent transportation systems, and GPS data of user involvement are conducted, and the efficiency of our method is confirmed by the results. Moreover, the obtained results reveal the characteristics of these commonly used regularizations for tensor completion in a certain sense and give experimental guidance concerning how to use them.**

***Index Terms*— CP factorization, incomplete data analysis, regularizations, tensor completion.**

I. Introduction

With the rapid development of data sensing, collecting, and networking techniques, sparsely observed data are now routinely and increasingly collected in various areas, such as intelligent transportation systems [1], social networks [2], recommender systems [3], [4], biomedicine [5], smart urban systems [6], learning systems [7], [8], and computer vision [9]–[12]. The accelerating growth of sparsely observed data has made the completion of incomplete data a critical task for both academia and industry. Information collected from complex processes and systems is naturally characterized by various sources, dimensions, connections, correlations, and other types of properties. Due to the plentiful characteristics and complexity of these natural processes and/or phenomena, it is difficult for a completion method with a simple structure and limited capacity to provide a reliable completion performance. Thus, we require new methods to recover and analyze incomplete data.

Tensor completion, which is the generalization of matrix completion, provides a useful tool for such existing sparsely observed data with natural multidimensional, multirelation, multimode, and multicorrelation structures. The increasing dimension N ($N \geq 3$) allows the tensor structure to represent more latent properties of data and reduce the computational complexity of certain data processing methods, such as the discrete cosine transform [13], which makes tensorial approaches superior to traditional methods. Signoretto *et al.* [14] examined the performance of matrix completion and tensor completion on three-way hyperspectral data. They found that tensor completion is more advantageous when there is correlation along the added dimension. In another similar study, Ran *et al.* [15] reported the similar results from comparisons on traffic volume data. Lahat *et al.* [16] provided an in-depth analysis of the advantages of tensor patterns over matrix patterns, and they claimed that the N-way array ($N \geq 3$) provides the additional types of diversity over matrix patterns and thus enrich the observational domain of data. These superiorities obtained from both theoretical and experimental findings have made tensor completion a hot topic.

The rank of a tensor [17] and its generations, such as the n-rank [18], TT-rank [19], nuclear norm [9], square norm [20], and TT-norm [21], provide sparsity measures for tensor objects; therefore, they are the most popular tools for tensor object analysis. A number of researchers have since adopted low-rank constraints as a subcomponent in incomplete tensor completion. Acar *et al.* [22] developed a tensor completion method based on a low-rank CP decomposition model. In addition, similar works based on the low-n-rank Tucker decomposition [23]–[26] have been proposed. Tensor decompositions are always formulated as nonconvex optimization problems, which hinder the performance of tensor completion.

Manuscript received December 31, 2016; revised August 16, 2017, January 11, 2018, June 2, 2018, and June 23, 2018; accepted June 25, 2018. Date of publication July 26, 2018; date of current version February 19, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61620106002 and Grant 61271376. *(Corresponding author: Huachun Tan.)*

Y. Wu is with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: kaimaogege@gmail.com).

H. Tan was with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China. He is now with Southeast University, Nanjing 210096, China (e-mail: tanhc@bit.edu.cn).

Y. Li is with the School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: yli@bupt.edu.cn).

J. Zhang is with the Research Center for Internet of Mobility, School of Transportation, Southeast University, Nanjing 210096, China (e-mail: jianzhang@seu.edu.cn).

X. Chen is with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: xiaoxuan.chen@wisc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2851612

Liu *et al.* [9] first defined the nuclear norm of a tensor and translated tensor completion into a convex optimization problem. Following their works, a series of modified and similar methods has been proposed [27]–[29]. The nuclear norm, which is defined as the weighted sum of the nuclear norms of mode matrices, is not significantly powerful when the dimension N of the tensor is 4 or larger. Therefore, the square norm [20] and the TT-norm [21] are proposed to recover high-dimensional tensors. All the previously mentioned approaches are powerful and useful methods for low-rank tensor completion to a certain extent; however, the general tensor objects are always not exactly low rank. Numerous published studies in the literature [30], [31] have reported that only using low-rank constraints is not sufficient in cases where the tensors and/or matrices do not have an explicit low-rank structure.

Several attempts have been made to address the drawback of low-rank completion methods on general tensor objects. One of the most popular methods is to add an l_2 norm regularizer on the factor matrices of the tensor decomposition [32], [33]. In addition, several studies of other regularizations have been performed on tensor completion. Narita *et al.* [34] proposed using graph regularization approaches as data manifold information in addition to the low-rank assumption to improve the performance of low-rank CP-decomposition-based tensor completion. Chen *et al.* [30] used both nuclear norm and manifold information to regularize the Tucker decomposition model and proposed a simultaneous tensor decomposition and completion method, which significantly outperforms conventional low-rank completion approaches on visual data completion. A similar work on a low-rank CP decomposition has been proposed [35]. Zhao *et al.* [36] developed a Bayesian inference for low-rank CP decomposition and incorporated a sparse prior for the factor matrices and an automatic rank determination procedure; this method is better than low-rank decompositions that do not incorporate priors, and a robust version of this paper in the presence of noise has been proposed [37]. Yokota *et al.* [38] emphasized that smoothness is a quite important factor for tensor completion and proposed two low-rank tensor completion methods with optional smoothness constraints called smooth variation Parafac and total variation smooth PARAFAC.

Each of the aforementioned priors that had been incorporated into tensor completion is helpful to a certain extent in certain circumstances. However, the data generated from natural systems generally consist of multiple latent properties, each with a different sense, distribution, and meaning. It is difficult for a model with a single factor prior to accurately characterize such complex data. For example, models with a single sparse prior have shown limitations in certain situations; for example, when the latent correlation is very high [39], models with smooth priors and graph priors cannot be easily adapted to data without such properties. As for tensorial data, it is believed that each type of constraint, structural (i.e., on the factor matrices) or observational, that contribute to the tensor decomposition can enhance its performance [16]. Previous experimental findings also indicate that a tensor completion model with one or more specific priors together with low-rank priors outperforms a model with only a low-rank prior. Thus, one can easily extrapolate that a tensor completion framework including more than one factor, such as a manifold, smoothness, distribution, and sparsity, together with the low-rank prior can achieve a better performance. However, previous works concentrated on the introduction of new priors for tensor completion and their evaluation on visual data completion. There is a gap on developing a framework that includes numerous priors all together and a systematic experimental study about those priors on many different data sets.

TABLE I
NOTATIONS INVOLVING TENSOR ALGEBRA

Notation	Description
$\underline{\mathbf{X}}$	N-way tensor
$\overline{X}$	matrix
$X_{(n)}$	mode-n matricization of tensor $\underline{\mathbf{X}}$
A_n	mode-n matrix in CP model
$\odot$	Khatri-Rao product
$\circ$	vector outer product

To fill this gap, we proposed a CP-decomposition-based method named fused CP (FCP) decomposition that fuses numerous factors simultaneously. We introduce the l_2 norm penalty, the l_1 norm penalty, the total variation penalty, and the graph penalty on the factor matrices of the CP decomposition and conduct extensive experiments for tensor completion on simulation data, visual data completion, intelligent transportation systems, and GPS data of user involvement to provide a systematic study of those tensor factor priors. To optimize the proposed tensor decomposition problem, we approximate the nonsmooth l_1 norm penalty and total variation penalty as a smooth function using the technique from [40]. We also apply Nesterov's optimal gradient method to optimize each subproblem. In addition, an automatic rank determination approach is used to avoid overfitting. The experimental results confirm the efficiencies of the proposed method and provide the experimental guidance concerning how to use the aforementioned priors.

The remainder of this paper is structured as follows. In Section II, we introduce the notations and preliminaries in this paper. In Section III, we propose novel algorithms for the FCP-decomposition-based tensor completion. In Section IV, we evaluate the performance of our methods on extensive experiments, and we compare them with some state-of-the-art methods. The conclusions are drawn in Section V.

II. PRELIMINARIES AND NOTATIONS

A brief overview of tensor decomposition and its application can be found in [41]. The important notations of this paper are provided in Table I.

Multiway arrays, also referred to as tensors, are higher order generalizations of vectors and matrices. Higher order arrays are represented as $\underline{\mathbf{X}} \in R^{I_\Delta \times I_\Theta \times \cdots I_N}$, where the order of $\underline{\mathbf{X}}$ is N. Each dimension of a multiway array is called a mode. The mode-n unfolding (also called matricization or flattening) of a tensor $\underline{\mathbf{X}} \in R^{I_\Delta \times I_\Theta \times \cdots \times I_N}$ is defined as unfolding $(\underline{\mathbf{X}}, n) = X_{(n)}$, where the tensor element $(i_\Delta, i_\Theta, \ldots, i_n)$ is

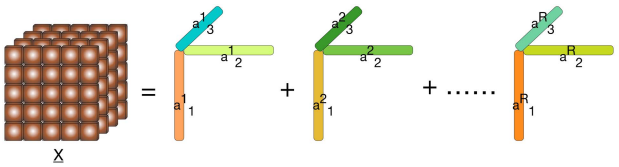


Fig. 1. CP decomposition for a 3-D tensor of rank R. A three-way tensor $\underline{\mathbf{X}}$ can be decomposed as $\sum_{r=1}^{R} a_1^r \circ a_2^r \circ a_3^r$.

mapped to the matrix element (i_n, j), in which

$$j = 1 + \sum_{\substack{k=1 \\ k \neq n}}^{N} (i_k - 1) J_k \quad \text{with } J_k = \prod_{m=1.m\neq n}^{k-1} I_m. \tag{1}$$

Therefore, $X_{(n)} \in R^{I_N \times J}$, where $J = \prod_{k=1.k\neq n}^{k-1} I_k$.

The CP decomposition decomposes a tensor into a sum of component rank-one tensors, which can be concisely expressed as

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} a_1^r \circ a_2^r \cdots \circ a_N^r = [\![A_1, A_2, \ldots, A_N]\!] \tag{2}$$

where R is the rank of the tensor, $\circ$ denotes the vector outer product, and $A_{(i)} = [a_{(i)}^1, a_{(i)}^2, \ldots, a_{(i)}^R] \in R^{I_i \times R}$ denotes the factor matrix of the ith mode. It should be noted that some papers use the number of columns to denote the rank of tensor. In this paper, we follow the work of [36], and use the number of rows to denote the rank of a tensor. Fig. 1 shows a CP decomposition of a three-way tensor.

Let $\underline{\mathbf{X}}_\Omega$ be the observed tensor that stores all the observed values such that

$$(\underline{\mathbf{X}}_\Omega)_{i_1,i_2,\ldots,i_n} = \begin{cases} \underline{\mathbf{X}}_{i_1,i_2,\ldots,i_n} & \text{if } (i_1, i_2, \ldots, i_n) \in \Omega \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

III. Methodology

In this section, we present our FCP tensor factorization model for incomplete tensors.

A. Model

As mentioned in Section I, tensor structures admit additional types of priors that can further contribute to interpretability, robustness, uniqueness, and other desired properties [16]. Thus, the FCP tensor factorization model is introduced in the following attempts to combine a set of priors to improve model accuracy and flexibility. Our algorithm consists of the following well-known priors.

1) *CP factorization Model Over the Incomplete Tensor* $\underline{\mathbf{T}} \in R^{I_1 \times I_2 \cdots \times I_N}$*:* The CP factorization model is the foundation of this paper.
2) *l_1 and l_2 Norm Regularization Terms Over the Factor Matrices of the CP Model:* The l_1 and l_2 norms are the main regularization terms used for vectorial, matrix, and tensorial data. l_2 regularization increases the completion accuracy by preventing overfitting. However, l_2 regularization cannot produce a successful model selection, thus reducing the interpretability of the factor matrices. Due to the nature of l_1 norm regularization, the l_1 regularized model produces sparse factors, and it achieves a better performance with respect to variable selection and model explanation [42]. In vectorial cases, it is found that neither l_1 nor l_2 norms uniformly outperform the other on different data sets [43], [44]. It is shown that using both l_1 and l_2 norms for vector objects can provide the advantages of both norms [39]. The completion accuracy and interpretability are equally important for incomplete tensor analysis; thus, both l_1 and l_2 norms are used in our model.
3) *Graph Regularization Terms Over Factor Matrices:* Typically, we generally have some knowledge and understanding about the local geometric structure of the targeted data. Recent studies on tensor completion have demonstrated that the graph regularized tensor factorization model can effectively model the data structure and improve the tensor completion performance [30], [34]. Thus, the FCP model naturally takes the local geometric structure into account by imposing a graph regularization term.
4) *Smooth Regularization Terms Over Factor Matrices:* Smoothness is a common property characterizing numerous data sets, such as computer vision, traffic, and gene expression data sets. A smooth regularization on the factor matrix helps to preserve the smooth property of multiway data and, thus, is very useful for multiway data analysis.

Given an incomplete tensor $\underline{\mathbf{T}} \in R^{I_1 \times I_2 \cdots \times I_N}$, we define a general fused tensor factorization model as follows:

$$\begin{aligned} \min_{\underline{\mathbf{X}}, A_1, \ldots, A_N} & \|\underline{\mathbf{X}} - [\![A_1, \ldots, A_N]\!]\|_F^2 + \sum_{n=1}^{N} \lambda_1^n \|A_n\|_1 \\ & + \sum_{n=1}^{N} \lambda_2^n \|A_n\|_F^2 + \sum_{n=1}^{N} \beta_1^n \Phi(A_n) + \sum_{i=1}^{N} \beta_2^n \|L_n A_n\|_1 \\ \text{s.t.: } & \underline{\mathbf{X}}_\Omega = \underline{\mathbf{T}}_\Omega \end{aligned} \tag{4}$$

where $[\![A_1, \ldots, A_N]\!]$ is the shorthand of CP factorization ($A_n \in R^{I_n \times R_n}$), $\Phi(A_n)$ are the graph constraints over the factor matrices, Ω is the set of observed entries, $\underline{\mathbf{X}}_\Omega$ is a tensor that stores all the values within the set Ω, λ_1^n, λ_2^n, β_1^n, and λ_2^n are the weight parameters for the l_1 norm, the l_2 norm, the Graph regularization, and the smooth term of factor matrix A_n, respectively, and $L_n \in R^{I_n - 1 \times I_n}$ is a smooth constraint matrix defined as follows:

$$L_n = \begin{vmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & 1 & -1 \end{vmatrix}. \tag{5}$$

It can be found that the term $\beta_2^n \|L_n A_n\|_1$ is equivalent to the total variation norm constraint on the factor matrices A_n; thus, $\beta_2^n \|L_n A_n\|_1 = \beta_2^n \|A_n\|_{TV}$. The total variation was first introduced in the computer vision field as a regularizing criterion [45]; it has been one of the most successful smooth constraints in image restoration and denoising. Therefore, we apply the TV constraint to FCP.

The graph term $\Phi(A_n)$ is defined as $\mathrm{Tr}(A_n^T L_n^g A_n)$. $L_n^g \in R^{I_n \times I_n}$ is a Laplacian matrix induced from the similarity matrix $W_n^g \in R^{I_n \times I_n}$, which is defined as $L_n^g = D_n^g - W_n^g$ (D_n^g is the diagonal matrix whose ith diagonal element is the sum of all the elements in the ith row of W_n^g). The similarity matrix W_n^g is a graphical representation used to measure similarity in the nth tensor mode; its elements represent the similarity of pairs in the nth mode of the target tensor.

B. Learning Factor Matrix A_n

The solver of the objective function (4) is based on a combination of Nesterov's optimal gradient method and block coordinate descent (BCD). Nesterov's optimal gradient method has been widely used in smooth convex optimization. It is shown that Nesterov's optimal gradient for smooth optimization achieves the optimal convergence rate $O(1/k^2)$ [46]. In each iteration, Nesterov's optimal gradient uses additional past information to take an extragradient step via an auxiliary sequence, and the step size is determined by the Lipchitz constant. Since (4) is a nonconvex minimization problem, it is usually infeasible to obtain the optimal solution, and thus, we employ the BCD method to obtain a local solution of (4). Given the mode-n matrix of (4), the BCD method alternatively solves

$$\min_{A_n} F^n(A_n) = \left\| X_{(n)} - A_n S_{-n}^A \right\|_F^2 + \lambda_1^n \|A_n\|_1 + \lambda_2^n \|A_n\|_F^2 + \beta_1^n \Phi(A_n) + \beta_2^n \|L_n A_n\|_1 \quad (6)$$

and

$$\min_{\underline{\mathbf{X}}} \|\underline{\mathbf{X}} - (\underline{\mathbf{T}}_{\Omega} + [\![A_1, \ldots, A_N]\!]_{\bar{\Omega}})\|_F^2 \quad (7)$$

where S_{-n}^A is expressed by the Khatri–Rao products of the component matrices, except the matrix A_n, $S_{-n}^A = A_{n-1} \odot \cdots A_2 \odot A_1 \odot A_N \cdots \odot A_{n+2} \odot A_{n+1}$, and $\bar{\Omega}$ is the complement set of the observed set Ω.

In this section, we mainly introduce how to solve subproblem (6). We first prove that (6) is convex with some positive semidefinite matrices L_n^g. Then, we translate the nonsmooth parts of the objective function into smooth parts. Finally, we propose using Nesterov's optimal gradient descent to optimize the factor matrices. In the BCD, we always hope that the subproblem is convex. For the problem in (6), we have Lemma 1.

Lemma 1: If L_n^g is positive semidefinite, then the objective function $F^n(A_n)$ is convex.

The proof for Lemma 1 is given in Appendix A. Lemma 1 guarantees that we can always find a global optimum for the subproblem in (6) with some positive semidefinite L_n^g, but the experimental results in Section IV also show that our method is efficient for optimizing (6) when L_n^g is indefinite.

The l_1 norm term $\|A_n\|_1$ in $F^n(A_n)$ is a nonsmooth function of A_n, which makes the optimization challenging. To address this problem, we use Nesterov's technique to construct a smooth approximation to $\|A_n\|_1$ as follows:

$$f_\mu(A_n) = \max_{\|B\|_{\max} \le 1} (\mathrm{Tr}(B^T A_n) - \mu d(B)) \quad (8)$$

where μ is a positive smooth parameter, $d(B)$ is a smooth function defined as $(1/2)\|B\|_F^2$, and $\|B\|_{\max}$ is the max norm of the matrix $\|B\|$ and its value is the max value of the matrix B. The l_1 term can be viewed as $f_\mu(A_n)$ with $\mu = 0$. For $f_\mu(A_n)$ in (8), we have Lemma 2.

Lemma 2: For any $\mu > 0$, $f_\mu(A_n)$ is a convex function in A_n, and the gradient of $f_\mu(A_n)$ takes the following form:

$$\nabla f_\mu(A_n) = B^* \quad (9)$$

where B^* is the optimal solution to (8). Moreover, the gradient of $\nabla f_\mu(A_n)$ is Lipschitz continuous with the Lipschitz constant $L_\mu = (1/\mu)$.

With Lemma 2, we can easily obtain the gradient of $f_\mu(A_n)$ by Proposition 1 given as follows.

Proposition 1: Let B^* be the optimal solution of (8). Then, we have

$$B^* = S\left(\frac{A_n}{\mu}\right)$$
$$S(x) = \begin{cases} x & \text{if } -1 < x < 1 \\ -1 & \text{if } x < -1 \\ 1 & \text{if } x > 1. \end{cases} \quad (10)$$

The proofs for Lemma 2 and Proposition 1 are given in Appendix B. Using the same strategy, the non-smooth term $\|A_n\|_{TV}$ can also be approximated as a smooth function $f_{\mu_s}^s = \max_{\|B\|_{\max} \le 1}(\mathrm{Tr}(B^T L_n A_n) - \mu d(B))$, and its gradient takes the following form:

$$\nabla f_{\mu_s}^s(A_n) = S\left(\frac{L_n A_n}{\mu_s}\right). \quad (11)$$

Moreover, the gradient of $f_{\mu_s}^s$ is Lipschitz continuous with the Lipschitz constant $(1/\mu_s)\|L_n\|_2^2$. The proofs are easily extended by the proofs of Lemma 2 and Proposition 1, they can also be found in [47, Appendix].

Using the property of $f_\mu(A_n)$ in Lemma 2, Proposition 1, and the properties of $f_{\mu_s}^s$, we substitute the l_1-norm and TV-norm regularizers in (6) to obtain the following optimization problem:

$$\min_{A_n} h^n(A_n) = \frac{1}{2}\left\| X_{(n)} - A_n S_{-n}^A \right\|_F^2 + \lambda_1^n f_\mu(A_n) + \frac{1}{2}\lambda_2^n \|A_n\|_F^2 + \frac{1}{2}\beta_1^n \mathrm{Tr}(A_n^T L_n^g A_n) + \frac{1}{2}\beta_2^n f_{\mu_s}^s(A_n). \quad (12)$$

The gradient of $h^n(A_n)$ is given as

$$\nabla h^n(A_n) = A_n (S_{-n}^A)^T S_{-n}^A - X_{(n)} (S_{-n}^A)^T + \lambda_2^n A_n + \lambda_1^n S\left(\frac{A_n}{\mu}\right) + \beta_1^n L_n^g A_n + \beta_2^n S\left(\frac{L_n A_n}{\mu_s}\right). \quad (13)$$

Proposition 2: The gradient of $\nabla h^n(A_n)$ is Lipschitz continuous with the Lipschitz constant $\|S_{-n}^A (S_{-n}^A)^T + \lambda_2^n E_{-n}^A\|_2 + \beta_1^n \|L_n^g\|_2 + (\lambda_1^n/\mu) + (\beta_2^n/\mu_s)\|L_n\|_2^2$, where $E_{-n}^A \in R^{r \times r}$ is an identity matrix.

The proof for Proposition 2 is given in Appendix C. With Proposition 2, Nesterov's method can be used to efficiently

optimize (12). The proximal function of $h^n(A_n)$ on the auxiliary sequence Y_n^k is

$$\varphi\left(A_n, Y_n^k\right) = h^n\left(Y_n^k\right) + \left\langle \nabla_{A_n} h^n\left(Y_n^k\right), A_n - Y_n^k \right\rangle + \frac{L}{2}\left\| A_n, -Y_n^k \right\|_F^2. \quad (14)$$

L is the Lipschitz constant given in Proposition 2. $Y_k = A_{nk-1} + ((\alpha_{k-1}-1)/\alpha_k)(A_{nk-1} - A_{nk-2})$. According to [46] and [48], we choose $\alpha_{k+1} \approx ((1+(4\alpha_k^2+1)^{1/2})/2)$. Ignoring the nonnegative limit of the factor matrices in [48], the optimal gradient method proposed for nonnegative matrix factorization can be easily adopted in this paper. We summarize the optimal gradient method for optimizing $h^n(A_n)$ in Algorithm 1.

Algorithm 1: Optimal Gradient Method

input : A_n;
$S_{-n}^A = A_{n-1} \odot \cdots A_2 \odot A_1 \odot A_N \cdots \odot A_{n+2} \odot A_{n+1}$;
μ; β_1^n; β_2^n; λ_1^n; λ_2^n
output: A_n

initialization: set $Y_0 = A_n, \alpha_0 = 1, l = \|S_{-n}^A (S_{-n}^A)^T + \lambda_2^n E_{-n}^A\|_2 + \beta_1^n \|L_n^g\|_2 + (\lambda_1^n/\mu) + (\beta_2^n/\mu_s)\|L_n\|_2^2$;
Repeat:;
Update $A_{nk}, Y_{k+1}, \alpha_{k+1}$;
$A_{nk} = Y_k - \frac{\nabla h^n(Y_k)}{L}$;
$\alpha_{k+1} = \frac{1+\sqrt{4\alpha_k^2+1}}{2}$;
$Y_{k+1} = A_n{}^k + \frac{\alpha_k - 1}{\alpha_{k+1}}(A_n{}^k - A_n{}^{k-1})$;
$k = k+1$;
Until stopping criterion $\nabla h^n(Y_k) < \epsilon_n$ is satisfied

C. Supplementary Information

A proper estimation of the tensor rank R for the CP decomposition is essential for the success of tensor completion. Our strategy is to increase tensor rank R to $R + R_u$ when the updated changes in the factor matrices $A_1, A_2, \ldots, A_n$ stagnate; R_u is the number of components that the tensor rank changes. Specifically, we propose to increase R if both R has not reached the set maximum and the following condition is met:

$$\sum_n^N \frac{\left\| A_n^t - A_n^{t-1} \right\|_F}{\left\| A_n^{t-1} \right\|_F} < \eta \quad (15)$$

where A_n^t denotes the factor matrix A_n at iteration step t. This condition means that the change speed of the factor matrices becomes substantially slow. When the condition in (15) is reached, we update the tensor rank R to $R + R_u$ and add R_u rows random data to A_n. In this paper, the added element $A_n(:, R+1 : R_u) \in R^{I_n \times R_u}$ is drawn from the Gaussian distribution $N(0, 0.1)$.

For the graph regularizer term $\mathrm{Tr}(A_n^T L_n^g A_n)$ in our model, the most important step is to determine the edge weight $w_n^g(i, j)$, which should correctly maintain the similarity between different objects in a tensor pattern. Many studies on learning and/or choosing edge weights for visual images/videos [30], [49], [50] and other domains [51], [52] have been conducted. In this paper, the edge weight is chosen based on the specific data, and we will give the edge weight $w_n^g(i, j)$ in Section IV.

For the update strategy of $\underline{\mathbf{X}}$ in (7), we can easily obtain

$$\underline{\mathbf{X}}_{i_1,i_2,\ldots,i_n} = \begin{cases} \underline{\mathbf{T}}_{i_1,i_2,\ldots,i_n} & \text{if } i_1, i_2, \ldots, i_n \in \Omega \\ [\![A_1, A_2, \ldots, A_N]\!]_{i_1,i_2,\ldots,i_n} & \text{otherwise} \end{cases} \quad (16)$$

where Ω is the set of observed entries. After the tensor $\underline{\mathbf{X}}$ is updated using (16), the factor matrices $A_1, A_2, \ldots, A_N$ are updated using Algorithm 1. Combining all the above-mentioned terms, we summarize the method for optimizing FCP in Algorithm 2.

Algorithm 2: FCP

input : N-way incomplete tensor $\underline{\mathbf{T}}$; observed set Ω; $N \times 1$ vectors λ_1, λ_2, β_1, β_2; $maxiter$; η; R; R_u; $Rmax$; N graph matrices L_n^g; N smooth constraints matrices L_n; θ; μ
output: complete tensor $\underline{\mathbf{X}}$; CP factor matrices $A_1, A_2, \ldots, A_N$

initialization: Randomly set A_n; $\underline{\mathbf{X}}_{\bar{\Omega}}$; $\underline{\mathbf{X}}_{\Omega} = \underline{\mathbf{T}}_{\Omega}$;
for iter=1 to $maxiter$:;
for i=1 to N;
update A_n using algorithm 1;
end for;
update R to $R + R_u$ if (15) is satisfied and $R + R_u \le Rmax$;
update $\underline{\mathbf{X}}$ using (16);
break if $\|\underline{\mathbf{X}}_{\Omega}^{iter} - \underline{\mathbf{X}}_{\Omega}^{iter-1}\|_F < \eta$

The main time cost of FCP results from the computation for the gradient in (13). Since $(S_{-n}^A)^T S_{-n}^A$ and $X_{(n)}(S_{-n}^A)^T$ can be calculated before the iterations of Algorithm 1, the complexity of one iteration in Algorithm 1 is $O(I^{N-1}R^2 + I^N R) + K \times O(IR^2)$ for an order N tensor $\underline{\mathbf{T}} \in R^{I \times I \times \cdots I}$. According to [48], Algorithm 1 typically stops after a small number of iterations, generally $K < R$. An Nth-order tensor can be factorized into N factor matrices; therefore, the total complexity of one iteration in Algorithm 2 is $O(NI^{N-1}R^2 + NI^N R) + K \times O(NIR^2)$.

IV. Experimental Results and Analyses

We use a number of experiments to evaluate the proposed FCP factorization method based on simulation data and numerous real-world applications, such as image, traffic, and GPS data. We compare our FCP with several state-of-the-art methods. HaLRTC[1] [9] (no prior), STDC[2] [30] (manifold prior), FBCP-MP[3] [36] (sparsity-inducing Gaussian prior and adjacent similarity prior), SPC[4] [38] (smooth prior), and

[1] http://www.cs.rochester.edu/u/jliu/publications.html
[2] http://mp.cs.nthu.edu.tw/project_STDC
[3] http://www.bsp.brain.riken.jp/ q̃ibin/homepage/BayesTensorFactorization.html
[4] https://sites.google.com/site/yokotatsuya/home/software

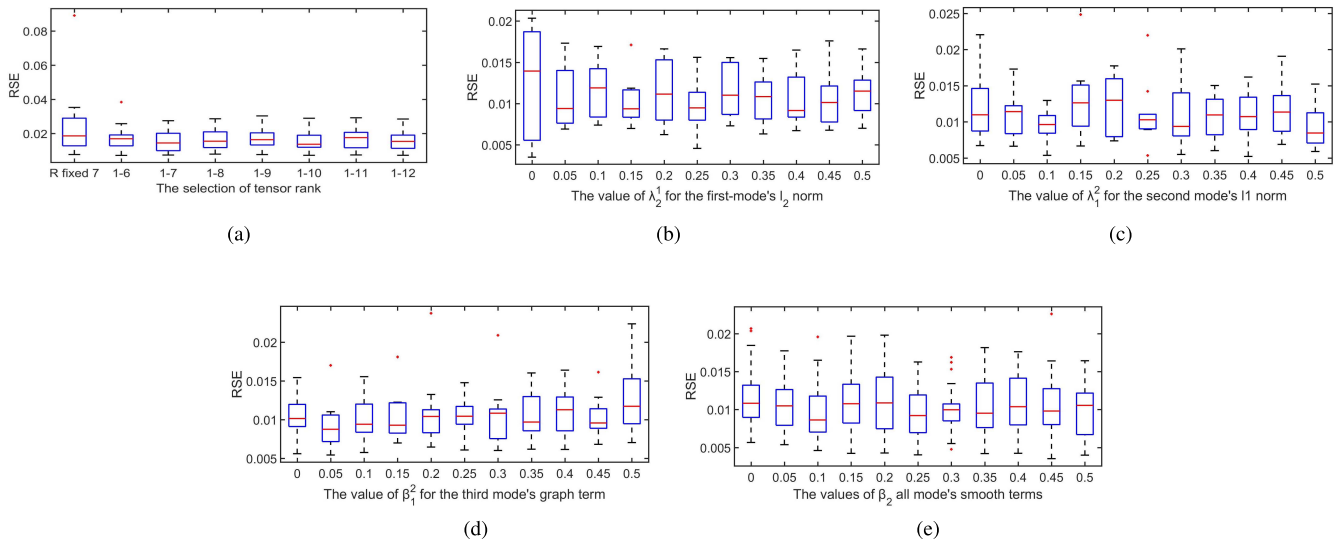


Fig. 2. Completion results of different FCP models. (a) Models with different tensor rank setups. (b) Models with different λ_2 values from (0, 0, 0) to (0.5, 0, 0); the initial rank is set to 1, and the maximum rank is set to 7. (c) Models with different λ_1 values from (0, 0, 0) to (0, 0.5, 0), where λ_2 is fixed as (0.15, 0,0). (d) Models with different β_1 values from (0, 0, 0) to (0, 0, 0.5), where λ_1 is fixed as (0, 0.5, 0). (e) Models with different β_2 values from (0, 0, 0) to (0.5, 0.5, 0.5), where β_1 is fixed as (0, 0, 0.05).

APG-TC[5] [32] (l_2 norm and nonnegative constraints on factors). All experiments are performed on a PC (2.1 GHz Intel Xeon E5-2620 and 64 GB of memory).

A. Test on Simulation Data

The purpose of this simulation is to show that the FCP can accurately impute missing data in tensor objects and to study the role of the factor priors and rank determination procedure in FCP. We simulate data under the following model:

$$\underline{\mathbf{Y}}_s = [\![B_1, B_2, \ldots, B_N]\!] + \alpha\underline{\mathbf{Z}} \tag{17}$$

where $[\![B_1, B_2, \ldots, B_N]\!]$ is the CP tensor model and $\underline{\mathbf{Z}}$ is a noise tensor whose entries $z_{i_1,i_2,\ldots,i_N} \sim N(0,1)$. We randomly simulated 50 three-way tensors $\underline{\mathbf{Y}}_s \in R^{40\times40\times40}$ with 80% missing data. The noise parameter α is set to be 0.1. $B_1 \in R^{40\times5}$ and its elements obey the Gauss distribution $N(0,1)$, $B_2 \in R^{40\times5}$ and its elements obey the standard Laplace distribution, and $B_3 \in R^{40\times5}$ is generated by the linear formulae [34] $[B_3]_{ir} = i\xi_r + \xi'_r$, where $[\xi_r, \xi'_r]_{r=1}^5$ are generated using the standard Gaussian distribution.

To achieve better selections for the model parameters, numerous FCP models with different parameters are evaluated. For all the models, the update rank parameter η is set to 0.006, and the maximum number of iterations is set to 1000. The models all use the function $\|\underline{\mathbf{X}}_\Omega^{\text{iter}} - \underline{\mathbf{X}}_\Omega^{\text{iter}-1}\|_F$ to set the stopping rule $(10^{-2.5}\|\underline{\mathbf{X}}_\Omega^{\text{iter}}\|_F)$. Since the columns of each factor matrix are generated by linear functions, the similarity weight matrix $W_3 \in R^{40\times40}$ for the third mode is naturally defined as the tridiagonal matrix

$$W_3 = \begin{vmatrix} 0 & 1 & 0 & . & . & . & \\ 1 & 0 & 1 & . & . & . & \\ 0 & 1 & 0 & . & . & . & \\ . & . & . & . & & & \\ . & . & . & & . & & \\ . & . & . & & & . & . \end{vmatrix}. \tag{18}$$

[5]https://www.ima.umn.edu/~yangyang/publication.html

Since there are numerous types of regularizer parameters in our model, it is not feasible to simulate all combinations of these parameters. The experimental procedure is as follows. We first set all the regularization terms to zero and find a suitable tensor rank setup. Then, we fixed the tensor rank setup and changed the value of the l_2 norm parameter λ_2 to study the effect of the l_2-norm term. Next, we fixed the l_2-norm parameter and changed the value of the l_1-norm parameter λ_1. Finally, we set the graph term parameter β_1 and the smooth term parameter β_2. The simulation was repeated 30 times each round.

Fig. 2 shows the completion results in terms of the relative square error. In Fig. 2(a), the models with automatic rank determination are more accurate than the model with a fixed tensor rank. In order to guarantee the fairness of comparison, the initial rank is set to 1 and the maximum rank is set to 7 for later comparison. It should be noted that the performance of FCP is better when its rank is $\geq$5 compared with its rank is $\leq$5. Rank 7 is chosen because it achieves a relatively stable performance. However, the best fixed rank is difficult to determine, and it is related to many factors, e.g., the magnitude of noise in our simulation. Therefore, we did not provide this result in this paper. From Fig. 2(b)–(e), we see that: 1) a lower average error can be achieved by adding new regularization terms in the proposed model if proper parameters are used and 2) although the data generation process does not contain a smooth factor for all three modes of the simulated tensor, the models with the TV norm term still can achieve lower errors than those without the TV norm regularization. The results agree with the hypothesis "structures admit additional types of priors that can further contribute to interpretability, robustness, uniqueness, and other desired properties [16]. Fig. 3 shows the plots of the errors $\|\underline{\mathbf{T}}_{\bar{\Omega}} - [\![A_1, \ldots, A_N]\!]_{\bar{\Omega}})\|_F$ and the estimated tensor rank with respect to the number of iterations of FCP1, FCP2 (with automatic rank determination procedure), and FCP3 (with both automatic rank determination procedure and regularization).

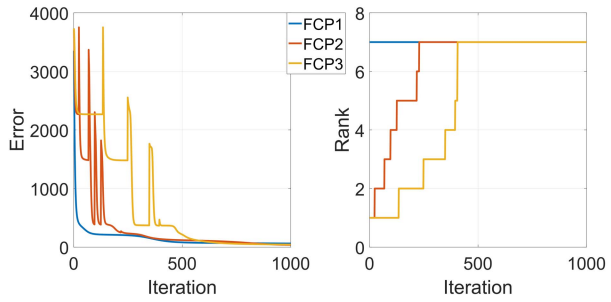


Fig. 3. Convergence of stopping rule function $\|Z_\Omega - [A_1, \ldots, A_N]\|_F$ and estimated rank in terms of the number of iterations in FCP1, FCP2, and FCP3 for three-way cases.

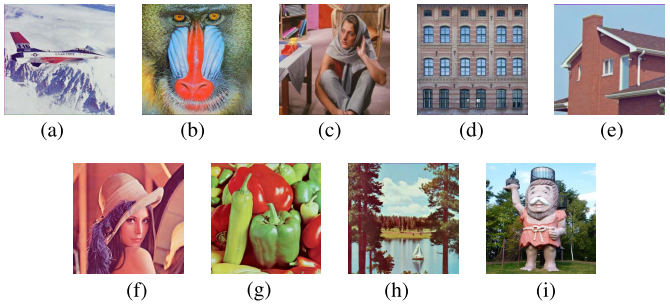


Fig. 4. Test images of size $256 \times 256 \times 3$. (a) Airplane. (b) Baboon. (c) Barbara. (d) Facade. (e) House. (f) Lena. (g) Peppers. (h) Sailboat. (i) Giant.

The regularization parameters for FCP3 are set as follows. λ_2 of the l_2-norm regularization is set to be $[0.15, 0, 0]$, λ_1 of the l_1-norm regularization is set to be $[0, 0.5, 0]$, β_2 of the TV norm regularization is set to be $[0.1, 0.1, 0.1]$, and β_1 of the graph regularization is set to be $[0, 0, 0.05]$. Since the rank is updated when the change speed of the factor matrices becomes low, the overall error functions of FCP2 and FCP3 increase when the rank is updated. However, as observed from Fig. 3, FCP1 and FCP2 achieve higher convergence speeds and lower stopping rule functions when incorporating the automatic rank determination procedure. Noting that CP factorization is a nonconvex optimization problem, the automatic rank determination procedure, which helps to find a more optimal solution, might achieve a more accurate completion.

B. Test on Color Images

In practice, target images always suffer from missing data, e.g., in certain overwritten images, some pixels of images can be found to be missing. Therefore, an image inpainting method is very functional. To investigate the performance of the proposed FCP on image inpainting, several standard images, shown in Fig. 4, are used to test the performance of tensor-completion-based image inpainting methods, including the proposed FCP, HaLRTC [9], STDC [30], FBCP-MP [36], SPC [38], and APG-TC [32]. The size of the tested RGB images is $256 \times 256 \times 3$. We conducted two groups of experiments: 1) images with random missing pixels, where we randomly discard entries from nine image tensors, and the missing ratio varies from 10% to 95% and 2) images overwritten by two types of texts (Chinese poems and English poems) with the result being missing entries in the images, where the missing ratio is approximately 25%. There is a considerable variety of missing data situations in reality. It is noteworthy that tensor-completion-based image inpainting methods are sensitive to the parameters and distribution of the missing data. To provide a fair comparison, we adopt a fixed set of parameters chosen from the literature [9], [30], [32], [36], [38], and our experimental results for all image inpainting experiments. Some key parameters are chosen as follows.

For the proposed FCP, the initial rank was set to 1, the maximum rank was set as 270, λ_2 of the l_2-norm regularization was set as $[0.5, 0.5, 0.5]$, λ_1 of the l_1-norm regularization was set as $[0.1, 0.1, 0.1]$, β_2 of the TV-norm regularization was set as $[10, 10, 0.5]$, β_1 of graph regularization was set as $[60, 60, 10]$, and the elements w_{ij} of the graph matrix were defined as $w_{ij} = e^{i-j}$. For SPC, the smooth parameter ρ is set as $[0.5, 0.5, 0]$. For FBCP-MP, the initial rank is 100. For STDC, the same regularization terms for FCP are applied, the terms k and w in [30] are set as $10^{0.2}$ and $10^{0.2}$, respectively, and the input tensors are mapped into $[0, 1]$. For HaLRTC, the weighted parameters α of the nuclear norm of the mode-n matricizations are set as $[1, 1, 1\mathrm{e}-3]$, and the parameter ρ for controlling the optimization step size is $1\mathrm{e}-5$. For APG-TC, the estimated rank is 270. We use the PSNR and RMSE [RMSE $= (\|\underline{\mathbf{Y}}^o_{\bar{\Omega}} - \underline{\mathbf{Y}}_{\bar{\Omega}}\|_F)/\sqrt{n_{\bar{\Omega}}}$), where n_Ω is the number of missing data points] to evaluate the performance. PSNR is most commonly used to measure the quality of image inpainting, and therefore, it is used in this experiment.

Parts of qualitative results of the image inpainting are shown in Fig. 5, and the quantitative results for two situations are shown in Fig. 6 and Table II. From the visual qualities presented in Fig. 5, we can see that the tensor completion approaches HaLRTC and APG-TC using low-rank priors and l_2-norm priors can only achieve a satisfactory performance under low ratios of missing data. The FCP, FBCP-MP, STDC, and SPC approaches using "strong" priors, including manifold information and total variation constraints, achieve a higher visual quality when the missing ratio is very large (note that the manifold information in this experiments is similar to the adjacent similarity prior of FBCP-MP). It can be observed from the PSNRs and RMSEs in Fig. 6 that HaLRTC only using low-rank priors achieves a desirable performance when the missing ratio is below 50%, but its effectiveness is low when a large number of entries are missing. APG-TC with the l_2 norm and nonnegative constraints on the factor matrices in addition to the low-rank CP assumption gives worse results than the other methods, particularly under high ratios of missing data. The reasons for these results may be twofold. First, the method uses a fixed rank without utilizing an automatic rank determination method, and thus, it is prone to overfitting and locally optimal solutions resulting from an incorrect rank estimation. Second, although the nonnegative constraints on the factor matrices increase the interpretability of tensor methods in certain applications, the nonnegativity of the data does not mean a nonnegativity of the causal factors in the data. The nonnegative assumption on the causal factors of APG-TC may result in a severe deterioration of the

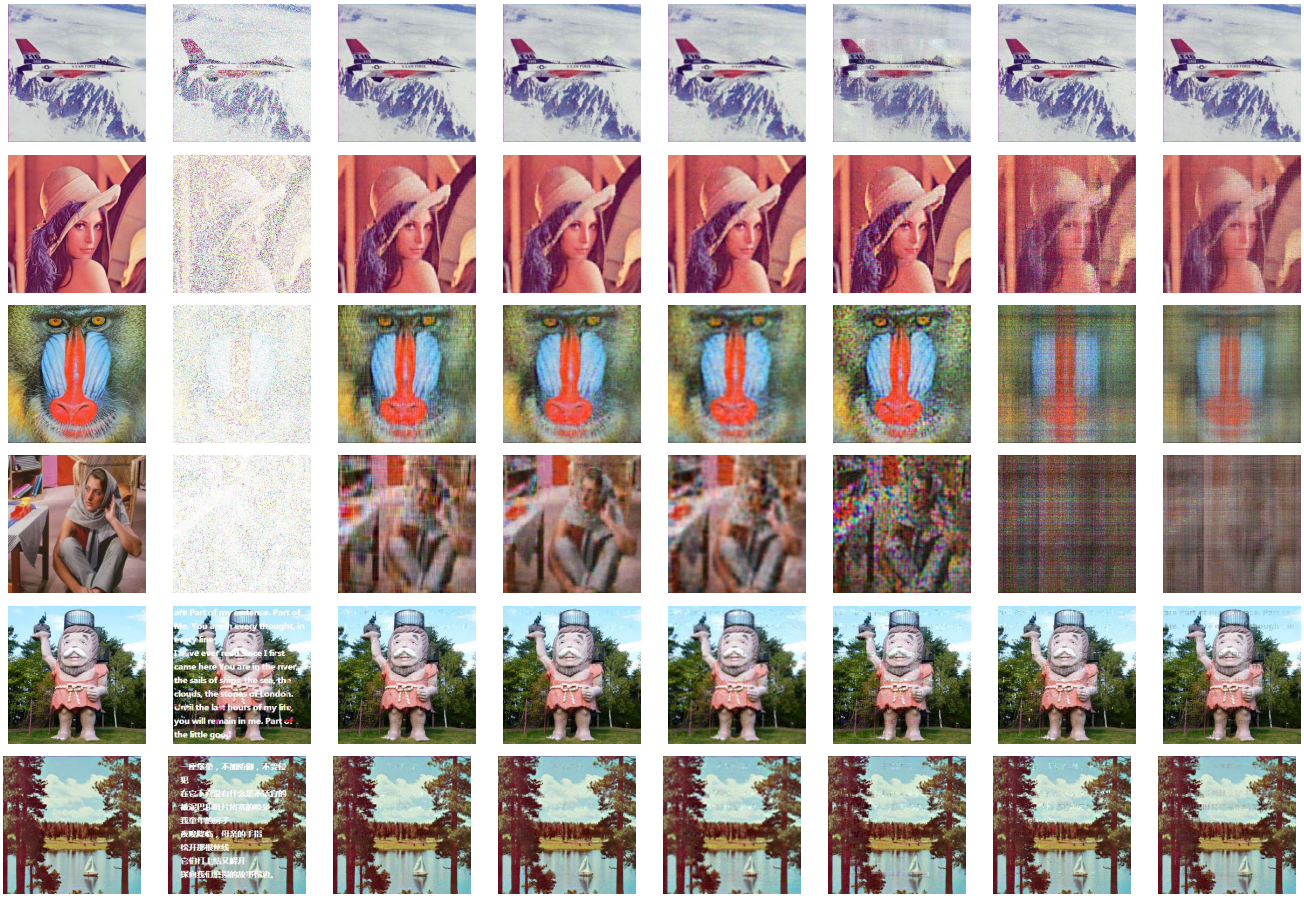


Fig. 5. Original images, the incomplete images, and the tensor completion results for the parts of the test images obtained by FCP, SPC, FBCP-MP, STDC, HaLRTC, and APG-TC (from left to right) under the missing data rates of 40%, 80%, 90%, and 95%, and images covered by English poetry and Chinese poetry (from top to bottom).

Fig. 6. (a) Average PSNRs and (b) average RMSEs for all test images under different random missing ratios.

completion results due to incorrect point estimations of the latent factors. STDC does not give a desirable performance under low missing ratios, but it gives a fairly satisfactory completion under high missing ratios. This indicates that the "strong prior" graph regularizers are less effective under a small amount of missing data. The proposed method, FCP, with the l_1 norm, l_2 norm, TV norm, and graph regularizations achieves a slightly better performance than the other methods when the missing ratio is below 80%, and its performance is comparable with STDC, FBCP-MP, and SPC when the missing ratio is extremely high. As shown in Table II, the proposed FCP outperforms the other methods on most images, except for "giant" and "baboon" under case (2). Interestingly, for plain image, "facade," APG-TC achieves the lowest RMSE, which

TABLE II
AVERAGE PSNRS AND RMSES OF DIFFERENT METHODS ON IMAGES COVERED BY POETRY

methods		airplane	baboon	barbara	facade	giant	house	lena	peppers	sailboat
FCP	PSNR	**32.1663**	28.9575	**35.4613**	**37.0784**	29.0736	**36.4185**	**32.9153**	**34.4089**	**31.4123**
	RMSE	**17.1328**	24.7897	**13.0041**	9.7325	24.4605	**10.5007**	**15.7174**	**13.2342**	**18.6885**
SPC	PSNR	30.7214	**29.9548**	33.4236	35.2773	**29.3357**	33.6187	32.2947	32.8054	31.0537
	RMSE	20.2357	**22.1006**	14.8240	12.0442	**23.7334**	14.4963	16.8762	15.9175	19.4713
FBCP-MP	PSNR	30.7634	29.0811	32.6850	31.0238	28.4268	35.1761	32.0543	33.0867	30.5234
	RMSE	20.1360	24.4392	16.1397	19.5413	26.3514	12.1154	17.3552	15.4103	20.7001
STDC	PSNR	28.6339	26.9302	29.5243	34.7749	25.8373	31.5728	29.3395	28.6233	27.3504
	RMSE	25.7306	31.3778	23.2235	12.6882	35.5042	18.3443	23.7229	25.7620	29.8279
HaLRTC	PSNR	29.9043	28.2996	31.1946	35.5673	27.1810	32.1885	30.9000	30.9902	28.7548
	RMSE	22.2295	26.7402	19.1608	11.5819	30.4154	17.0892	19.8219	19.6170	25.3750
APG-TC	PSNR	29,3003	25.4575	30.0342	34.3400	25.5005	32.4261	28.7587	28.6373	26.6793
	RMSE	23.8082	26.4243	15.6016	**9.5035**	26.2939	11.8461	18.0695	18.2471	22.9569

means that the nonnegative tensor factorization scheme can successfully characterize some tensor objects with a simple structure, although it always produces undesirable completion results with general tensor objects. Note that images masked by writing are more common than images with randomly missing data; these results demonstrate the superiority and effectiveness of the proposed FCP. Summarizing, first, FCP uses "weak priors," such as the low-rank CP model and l_2- and l_1-norm regularizers, which are particularly useful for tensor objects with a simple structure under low missing ratios. Second, the method also uses "strong priors," such as TV norm constraints and manifold information, which are powerful under high missing ratios. Third, the method incorporates an automatic rank determination approach to avoid overfitting and locally optimal solutions obtained under nonconvex CP factorization. Hence, it produces desirable performances under most cases.

C. Test on Traffic Data

In practice, there are various multivariate time series with plentiful multimode information [53]–[55], and these time series can be naturally adapted to tensor formats. It is natural to adopt tensor completion to impute missing data within such time series [56], [57], and it is necessary to test new tensor completion methods on such types of data. In this section, traffic data, a typical example of a multivariate time series, are used to test the proposed method. The peculiar traffic flow data from PeMS[6] for a north-bound I-405 trip are used for our experiments. The numbers of used locations are 716, 663; 717, 742; 717, 744; 716, 670; 716, 826; 717, 752; 717, 755; 717, 758; 717, 763; and 717, 769 (from upstream to downstream). Each location has four lanes. The distances between each location vary from 0.15 to 2.27 mile. The time period used in this paper is from April 01, 2014 to May 20, 2014. The traffic volume is aggregated every 5 min. Thus, one detector preserves 288 data points per day. These traffic data are constructed into a four-way tensor $\underline{\mathbf{A}} \in R^{\Delta\Gamma\times\Theta\Phi\Phi\times\Upsilon\times\Upsilon}$, whose modes represent the location, time, day, and week.

To characterize the intrinsic structure of the constructed traffic tensor model, the similarity weight matrices used by FCP and STDC for the traffic tensor should be well defined. The details about the design of the similarity weight matrices are given in Appendix D. For the proposed FCP, the initial rank is set as 1, the maximum rank is 300, λ_2 of the l_2-norm regularization is $[0.15, 0.15, 0.15, 0.15]$, λ_1 of the l_1-norm regularization is $[0.08, 0.08, 0.08, 0.08]$, β_2 of the TV-norm regularization is $[10, 10, 10, 10]$, and β_1 of the graph regularization is $[8, 1, 5, 50]$. For SPC, the smooth parameter ρ is $[0.5, 0.5, 0.5, 0.5]$. For FBCP-MP, the initial rank is set to be 100. For STDC, the terms k and w should be selected dynamically based on the missing ratio; otherwise, the performance of STDC will be poor. For HaLRTC, the weighted parameters α for the nuclear norm of the mode-n matricizations are $[100, 100, 100, 1]$, and the parameter ρ for controlling the optimization step size is $5e-4$. For APG-TC, the estimated rank is 300. We use the MAE [MAE $= (\sum(\underline{\mathbf{Y}}^o_{\bar{\Omega}} - \underline{\mathbf{Y}}_{\bar{\Omega}})/n_{\bar{\Omega}})$, where n_Ω is the number of missing data] to evaluate the performance. The reason why we use the MAE measure is that we care about the errors on the number of vehicles. Fig. 7 shows the MAE with randomly missing data ratios varying from 10% to 90%. It can be found that the results on traffic data are similar to those on image inpainting; an exception is that the FBCP-MP gives a worse performance on traffic data compared with image inpainting. This may be because FBCP-MP assumes a high correlation between adjacent rows in each mode, but this assumption is not applicable to traffic data.

Since the proposed method is based on tensor decomposition, the factor matrices generated from our algorithm can be utilized to understand the traffic flow with missing data from a "spectral" perspective [6]. The vectors in the factor matrices can be viewed as the location basis, time basis, day basis, and week basis of the traffic flow. Considering that the traffic flow basis in each mode cannot be negative, we add

[6]http://pems.dot.ca.gov/

Fig. 7. Average MAEs for the constructed traffic tensor under different random missing data ratios.

nonnegative constraints on the factor matrices to our model in this application. However, our experiments have shown that the nonnegative constraints on the factor matrices are usually not helpful in missing data imputation. The nonnegative factorization version of FCP (NFCP) can be correspondingly obtained by adding a nonnegative projection at step 2.1 in Algorithm 1. Fig. 8 shows the most significant basis group obtained from NFCP. From Fig. 8, we can see that there is a distinct peak representing the morning peak of traffic flow in the time mode, and there is a day mode basis that shows weak strength corresponding to weekends. These two bases are quite revealing, as it is known that traffic flow data exhibit stronger morning peaks on workdays than on weekends. However, there is no significant evening peak in the first time base. The reason is that the traffic flow of these locations exhibits stronger morning peaks in workdays than evening peaks. The first time base only captures the principle variation of traffic flow in time mode. The week mode and space mode bases show that traffic is similar on these weeks and at these locations. These results indicate that NFCP can successfully understand the traffic flow from a multimode "spectral" perspective.

D. Test on Activity Recommendation

In practice, there are many incomplete binary data sets that require completion, some of which include multimode information. Thus, evaluations of tensor completion on such data sets are very meaningful. In this section, we test tensor completion on activity participation data, in which 0 denotes no participation and 1 denotes participation. In particular, the experiments are conducted on location data generated by mobile devices. The data can be accumulated in the form of location trajectories and user activities, which always include multimode information such as space, user, and activity modes [58]. Thus, a three-way user × location × activity tensor can be constructed to represent such data, and tensor completion methods can be used to both infer the participation of users in certain location-specific activities and make corresponding recommendations [59]. To evaluate FCP on recommendation systems, a GPS data set obtained from Microsoft Research[7] is used. Similar to [59], a binary

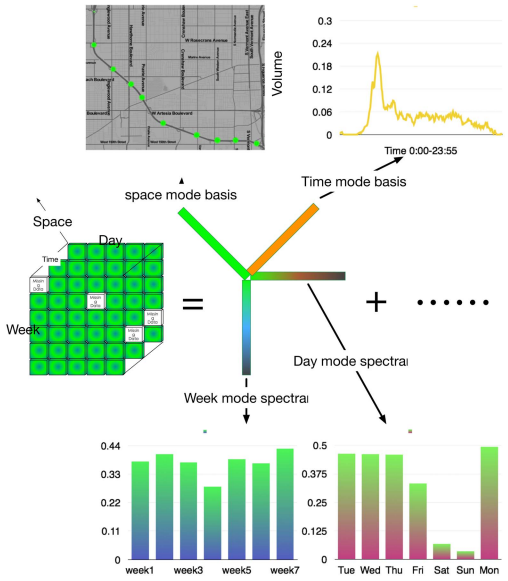


Fig. 8. Illustration of basic traffic pattern in the space mode, time mode, week mode, and day mode learned by NFCP.

three-way user × location × activity tensor $\underline{\mathbf{T}}$ of size $146 \times 168 \times 5$ is constructed, and the auxiliary user × user and activity × activity matrices in this data sets are used to construct the similarity weight matrices for FCP and STDC (more details[8] about these data sets can be found in [59]).

In this experiment, we consider a case with randomly missing data whereby 80% of the entries of $\underline{\mathbf{T}}$ are randomly removed. In addition, we compare FCP to not only the aforementioned tensor completion methods but also SDF—a data fusion model proposed in [59]. As with the experiments in [59], the receiver operating characteristic (ROC) curves and the area under the curve (AUC) are used to measure the performance of all the methods. To enhance the performance of FCP and STDC which use manifold information, the similarity weight matrices in the user mode and the activity mode are directly deduced from the user × user and activity × activity matrices within this data set. For the proposed FCP, the initial rank is set as 1, the maximum rank is 180, λ_2 of the l_2-norm regularization is $[0.15, 0.15, 0.15]$, λ_1 of the l_1-norm regularization is $[0.08, 0.08, 0.08]$, β_2 of the TV-norm regularization is $[0.1, 0.1, 0.1]$, and β_1 of graph regularization is $[5, 0, 5]$. For SPC, the smooth parameter ρ is set as $[1, 1, 1]$. For FBCP-MP, the initial rank is 100. For STDC, the terms k and w are $10^{-0.2}$ and $10^{0.4}$, respectively. For HaLRTC, the weighted parameters α for the nuclear norm of the mode-n matrices are set as $[100, 100, 10]$, and the parameter ρ for controlling the optimization step size is $2e - 2$. For APG-TC, the estimated rank is 180. In addition, the same parameters in [59] are used for SDF.

Fig. 9 shows the ROC curves of all these methods and compares the AUC of these methods. The ROC curves are drawn by the method proposed in [59]. It can be found that the

[7] http://research.microsoft.com/pubs/143146/aaai10.uclaf.data.zip

[8] http://www.tensorlab.net/demos/demo-gps.html#demo-gps

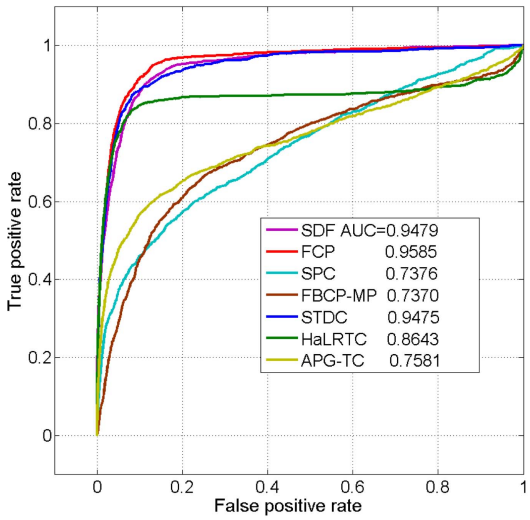


Fig. 9. ROC curves for the prediction of the 80% missing data user-location-activity link in the tensor $\underline{\mathbf{T}}$. The proposed FCP outperforms the other methods with an AUC of 0.9585.

TABLE III
AUCs WITH DIFFERENT l_1 NORM WEIGHTS

λ_1	AUCs
(0,0,0)	0.8210
(0.1,0.1,0.1)	0.9312
(0.2,0.2,0.2)	0.9333
(0.3,0.3,0.3)	0.9343
(0.4,0.4,0.4)	0.9314
(0.5,0.5,0.5)	0.9383
(0.6,0.6,0.6)	0.9478
(0.7,0.7,0.7)	0.9479

FCP model gives the best ROC curve with an AUC of 95.85%. Its performance is slightly better than SDF, which exploits more information with a data fusion approach, and STDC, which also utilizes the constructed manifold information. This indicates that the graph regularizer can successfully incorporate the auxiliary information and enhance the performance of tensor completion with a proper manifold design under the random missing data scenario. Interestingly, SPC, FBCP-MP, and APG-TC perform even worse than HaLRTC using only the low-rank prior. This may be because priors, such as smooth variation and nonnegative constraints, are not very useful to the user-location-activity link data.

Different from the traditional tensor completion models, an l_1 norm regularizer is incorporated in the proposed FCP. It is known that the l_1 norm encourages the sparsity of data, while tensor completion itself might destroy sparsity. Someone might be afraid of the contradiction between the l_1 norm and tensor completion. From the above experiments on simulation data, we have shown that a sparse prior with proper weight parameter achieves a better performance than without it. In this experiment, the user-location-activity link is naturally sparse, which only contains 10% nonzero values. Therefore, it provides a good resource to discuss the usage of the l_1 regularizer.

In order to better show the influence of l_1 terms of FCP, the same λ_2, β_1, and β_2 of the above experiments are used, and the l_1 norm is increased from $(0,0,0)$ to $(0.7,0.7,0.7)$ with unit $(0.1,0.1,0.1)$.

Table III shows the results of these experiments. It can be found that FCP with l_1 regularizers significantly outperforms that without l_1 norm. This result highlights the value of l_1 norm on real data sets that are naturally sparse. Because the data sets are very sparse, the best performance is achieved when the larger l_1 norm values are adopted.

V. CONCLUSION

In the context of sparsely observed multiway data generated by very complex systems, the underlying structure of such data sets is not always sufficiently understood because of their complexity. However, we can use simple priors such as sparsity, nonnegativity, low-rank, smoothness, manifold information, and statistical independence to represent common features of such data. As a result, we can enrich the observational domain of the data and explore the nature of very complex systems. To complete and analyze the sparsely observed multiway data, we propose a CP tensor factorization approach to fuse the l_2-norm constraint, sparseness (l_1 norm), manifold, smooth information, and low-rank properties simultaneously. The proposed method is evaluated on many complex data sets, including image, traffic, and user involvement data. The experiments demonstrate that the proposed method performs better than or equal to the other tensor completion methods depending on the problems being addressed. In addition, the experiments can be viewed as a systematical study on the used priors for tensor completion. In the experiments, we find that "strong priors," such as manifold information and smooth information, are very powerful under high ratios of missing data and "weak priors," such as low rank, the l_2-norm constraint, and sparseness, are useful for tensor objects with low ratios of missing data. Moreover, the graph regularizer with manifold information can successfully incorporate different types of auxiliary information with the proper manifold design, whose advantage in terms of flexibility makes it applicable in various situations. We also find that a proper automatic rank determination procedure can be used to avoid overfitting and locally optimal solutions of nonconvex CP factorization.

In the future, we expect that the proposed methods will be able to be used in various applications. The parameters of the FCP in current version are a little difficult to tune. The automatic regularization parameters tuning methods will be a valuable future direction. In addition, the joint analysis and fusion of multiple data sets based on the proposed framework represents a future direction. Moreover, we hope to investigate automatically choosing the nuisance parameters in our model, where the Bayesian optimization framework would be a desirable choice. The time complexity of FCP grows exponentially with the order of tensor and quadratic with tensor rank. It makes the proposed method FCP become slower when the tensor order is very high and the rank of tensor is very large. The research for a more efficient tensor completion model will be a valuable future direction.

APPENDIX

A. Proof of Lemma 1

Given any two matrices $A_1.A_2 \in \Re^{m\times r}$ and a positive $\theta \in (0, 1)$ and $G(A) = \|X - AS\|_F^2$, we have

$$\begin{aligned} &G(\theta A_1 + (1-\theta)A_2) - (\theta G(A_1) + (1-\theta)G(A_2)) \\ &\quad = \mathrm{Tr}(S^T(\theta A_1 + (1-\theta)A_2)^T(\theta A_1 + (1-\theta)A_2)S) \\ &\qquad - \big(\theta \mathrm{Tr}\big(S^T A_1^T A_1 S\big) + (1-\theta)\mathrm{Tr}\big(S^T A_2^T A_2 S\big)\big) \\ &\quad = \theta(\theta-1)\mathrm{Tr}\big(A_1^T A_1 S S^T\big) + 2\theta(1-\theta)\mathrm{Tr}\big(A_2^T A_1 S S^T\big) \\ &\qquad + \theta(\theta-1)\mathrm{Tr}\big(A_2^T A_2 S S^T\big) \\ &\quad = -(\theta(1-\theta))\|(A_1 - A_2)S\|_F^2 \le 0. \end{aligned} \tag{19}$$

Therefore, according to the definition of a convex function, we know that $G(A)$ is convex. Given $h_2(A) = \mathrm{Tr}(A^T L_g A)$ and a positive semidefinite matrix L_g, we have

$$H(h_2(A)) = L_g \tag{20}$$

where $H(h_2(A))$ is the Hessian matrix of $h_2(A)$ and L_g is positive semidefinite; thus, $h_2(A)$ is convex. By considering the convexity of $\|A\|_F^2$, $\|A\|_1$, $\|A\|_{TV}$ [60], $G(A)$, and $h_2(A)$, we know that $F^n(A_n)$ is convex.

B. Proof of Lemma 2 and Proposition 1

$F_\mu(A_n)$ is convex as a maximum of functions that are linear in A_n. The conjugate of $d()$ at (A_n/μ) is $d^*(A_n/\mu) = \sup_{\|B\|_{\max}\le 1}(\mathrm{Tr}(B^T(A_n/\mu)) - d(B))$, and hence, $f_\mu(A_n) = \mu d^*(A_n/\mu)$. Consider that "a closed proper convex function is essentially strictly convex if and only if its conjugate is essentially smooth." [61]. Since $d(B)$ is a closely proper strictly convex function, its conjugate is smooth. Therefore, $F_\mu(A_n)$ is a smooth function.

Let $\phi(A_n, B) = \mathrm{Tr}(B^T A_n) - \mu d(B)$. Since $d()$ is a strongly convex function, $\arg\max_{\|B\|_{\max}\le 1}\phi(A_n, B)$ has a unique optimal solution according to Danskin's theorem [62], and we denote it as B^*. Then, $\Delta f_\mu(A_n) = \nabla_{A_n}\phi(A_n, B^*) = B^*$. In addition, $B^* = \arg\max_{\|B\|_{\max}\le 1}(\mathrm{Tr}(B^T A) - (\mu/2)\|B\|_F^2) = \arg\min_{\|B\|_{\max}\le 1}\|B - (A_n/\mu)\|_F^2$. Hence, the optimal solution B^* can be obtained by projecting (A_n/μ) onto the max-norm ball where

$$B^* = S\left(\frac{A_n}{\mu}\right)$$
$$S(x) = \begin{cases} x & \text{if } -1 < x < 1 \\ -1 & \text{if } x < -1 \\ 1 & \text{if } x > 1. \end{cases} \tag{21}$$

Furthermore, for any two matrices A_1 and A_2, we have

$$\left\|S\left(\frac{A_1}{\mu}\right) - S\left(\frac{A_2}{\mu}\right)\right\|_F \le \left\|\frac{A_1}{\mu} - \frac{A_2}{\mu}\right\|_F = \frac{1}{\mu}\|A_1 - A_2\|_F. \tag{22}$$

Thus, $S(A_n/\mu)$ is Lipschitz continuous with Lipschitz constant $L_\mu = (1/\mu)$.

C. Proof of Proposition 2

We divided $h^n(A_n)$ into four parts, $h_1(A_n) = (1/2)\|X_{(n)} - A_n S_{-n}^A\|_F^2 + (1/2)\lambda_2^n\|A_n\|_F^2$, $h_2(A_n) = (1/2)\mathrm{Tr}(A_n^T L_n^g A_n)$, $\beta_2^n F_{\mu_s}^s(A_n)$, and $\lambda_1^n F_\mu(A_n)$, where the Lipschitz constant of $\bigtriangledown h^n(A_n)$ can be easily calculated as a linear combination of the Lipschitz constants of $\bigtriangledown h_1(A_n)$, $\bigtriangledown h_2(A_n)$, $\bigtriangledown F_{\mu_s}^s(A_n)$, and $\bigtriangledown F_\mu(A_n)$.

For $\bigtriangledown h_1(A_n)$, given two matrices A_1 and A_2, we have

$$\begin{aligned} &\left\|\bigtriangledown h_1(A_1) - \bigtriangledown h_1(A_2)\right\|_F^2 \\ &\quad = \left\|(A_1 - A_2)\left(\left(S_{-n}^A\right)^T S_{-n}^A + \lambda_2^n E_{-n}^A\right)\right\|_F^2 \\ &\quad = \mathrm{Tr}(((A_1 - A_2)U\Sigma U^T)^T((A_1 - A_2)U\Sigma U^T)) \end{aligned} \tag{23}$$

where $U\Sigma U^T$ is the singular value decomposition of $(S_{-n}^A)^T S_{-n}^A + \lambda_2^n E_{-n}^A$. By the properties of the trace of matrices, (23) is equivalent to

$$\begin{aligned} &\mathrm{Tr}(U^T(A_1 - A_2)^T(A_1 - A_2)U\Sigma^2) \\ &\le \left\|\left(S_{-n}^A\right)^T S_{-n}^A + \lambda_2^n E_{-n}^A\right\|_2^2 \mathrm{Tr}(U^T(A_1 - A_2)^T(A_1 - A_2)U) \\ &= \left\|\left(S_{-n}^A\right)^T S_{-n}^A + \lambda_2^n E_{-n}^A\right\|_2^2 \|A_1 - A_2\|_F^2. \end{aligned} \tag{24}$$

Therefore, $\bigtriangledown h_1(A_n)$ is Lipschitz continuous, and the Lipschitz constant is $\|(S_{-n}^A)^T S_{-n}^A + \lambda_2^n E_{-n}^A\|_2$.

For $\bigtriangledown h_2(A_n)$, given two matrices A_1 and A_2, we have

$$\left\|\bigtriangledown h_2(A_1) - \bigtriangledown h_2(A_2)\right\|_F^2 = \left\|L_n^g(A_1 - A_2)\right\|_F^2. \tag{25}$$

As with $\bigtriangledown h_1(A_n)$, we find that $\bigtriangledown h_1(A_n)$ is Lipschitz continuous, and the Lipschitz constant is $\|L_n^g\|_2$.

From (23) and Lemma 2, we have

$$\begin{aligned} &\left\|\bigtriangledown h^n(A_1) - \bigtriangledown h^n(A_2)\right\|_F \\ &\quad \le \left\|\bigtriangledown h_1(A_1) - \bigtriangledown h_1(A_2)\right\|_F \\ &\qquad + \beta_1^n\left\|\bigtriangledown h_2(A_1) - \bigtriangledown h_2(A_2)\right\|_F \\ &\qquad + \lambda_1^n\left\|\bigtriangledown F_\mu(A_1) - \bigtriangledown F_\mu(A_2)\right\|_F \\ &\qquad + \beta_2^n\left\|\bigtriangledown F_{\mu_s}^s(A_1) - \bigtriangledown F_{\mu_s}^s(A_2)\right\|_F \\ &\quad \le \left(\left\|S_{-n}^A\left(S_{-n}^A\right)^T + \lambda_2^n E_{-n}^A\right\|_2 + \beta_1^n\left\|L_n^g\right\|_2\right. \\ &\qquad \left. + \lambda_1^n\frac{1}{\mu} + \lambda_2^n\frac{1}{\mu_s}\|L_n\|_2^2\right)\|A_1 - A_2\|_F. \end{aligned} \tag{26}$$

This completes the proof.

D. Design of Graph for Traffic Data

In the location mode, it is easy to deduce that the similarity between traffic volumes declines with increasing distances between locations; thus, we define

$$w_{ij}^{\text{space}} = (\exp(-|\text{distance}_{ij}|))^2. \tag{27}$$

In the time mode, since the traffic volumes at different time points are only correlated in a certain horizon, we define

$$w_{ij}^{\text{time}} = \begin{cases} (\exp(-|i-j|))^2 & \text{if } |i-j| \le 3 \\ 0 & \text{if } |i-j| > 3. \end{cases} \tag{28}$$

In the day mode, it is obvious that the traffic volumes are better correlated between the same types of days. The correlations

between weekends and workdays are relatively weak, and we define

$$w_{ij}^{\text{day}} = \begin{cases} 1 & \text{if } i = j \\ 0.9 & \text{if } i \in N(j) \quad \text{or } j \in N(i) \\ 0.3 & \text{otherwise} \end{cases} \tag{29}$$

where $N()$ denotes the type of day (workday or weekend). In the week mode, the traffic data for different weeks are all highly correlated. This is because the traffic flow exhibits strong weekly periodicity. We set the entries of $W^{(\text{week})}$ as

$$w_{ij}^{\text{week}} = \begin{cases} 1 & \text{if } i = j \\ 0.9 & \text{otherwise.} \end{cases} \tag{30}$$

ACKNOWLEDGMENTS

The authors would like to thank H. Zhang from the University of California at Davis for his help.

REFERENCES

[1] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 108–120, Sep. 2013.

[2] B. Ermiş, E. Acar, and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Mining Knowl. Discovery*, vol. 29, no. 1, pp. 203–236, 2015.

[3] Y. Gao *et al.*, "Camera constraint-free view-based 3-D object retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2269–2281, Apr. 2012.

[4] E. Frolov and I. Oseledets. (2016). "Tensor methods and recommender systems." [Online]. Available: https://arxiv.org/abs/1603.06038

[5] Z. Xu, F. Yan, and Y. Qi, "Bayesian nonparametric models for multiway data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 475–487, Feb. 2015.

[6] Z. Fan, X. Song, and R. Shibasaki, "CitySpectrum: A non-negative tensor factorization approach," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 213–223.

[7] J.-T. Chien and Y.-T. Bao, "Tensor-factorized neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1998–2011, May 2018.

[8] Z. Chen, K. Batselier, J. A. K. Suykens, and N. Wong, "Parallelized tensor train learning of polynomial classifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[9] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.

[10] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[11] X. Li, M. K. Ng, G. Cong, Y. Ye, and Q. Wu, "MR-NTD: Manifold regularization nonnegative tucker decomposition for tensor data dimension reduction and representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1787–1800, Aug. 2017.

[12] W. Hu, D. Tao, W. Zhang, Y. Xie, and Y. Yang, "The twist tensor nuclear norm for video completion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2961–2973, Dec. 2016.

[13] Q. Dai, X. Chen, and C. Lin, "Fast algorithms for multidimensional DCT-to-DCT computation between a block and its associated subblocks," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3219–3225, Aug. 2005.

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

[15] B. Ran, H. Tan, Y. Wu, and P. J. Jin, "Tensor based missing traffic data completion with spatial–temporal correlation," *Phys. A, Stat. Mech. Appl.*, vol. 446, pp. 54–63, Mar. 2016.

[16] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[17] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Stud. Appl. Math.*, vol. 6, nos. 1–4, pp. 164–189, 1927.

[18] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[19] I. V. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, no. 5, pp. 2295–2317, Jan. 2011.

[20] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. II-73–II-81.

[21] H. N. Phien, H. D. Tuan, J. A. Bengua, and M. N. Do. (2016). "Efficient tensor completion: Low-rank tensor train." [Online]. Available: https://arxiv.org/abs/1601.01083

[22] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics Intell. Lab. Syst.*, vol. 106, no. 1, pp. 41–56, 2011.

[23] M. Filipović and A. Jukić, "Tucker factorization with missing data with application to low-n-rank tensor completion," *Multidimensional Syst. Signal Process.*, vol. 26, no. 3, pp. 677–692, 2015.

[24] H. Tan *et al.*, "A tensor-based method for missing traffic data completion," *Transp. Res. C, Emerg. Technol.*, vol. 28, pp. 15–27, Mar. 2013.

[25] H. Tan, B. Cheng, W. Wang, Y.-J. Zhang, and B. Ran, "Tensor completion via a multi-linear low-n-rank factorization model," *Neurocomputing*, vol. 133, pp. 161–169, Jun. 2014.

[26] Y. Wu, H. Tan, Y. Li, F. Li, and H. He, "Robust tensor decomposition based on Cauchy distribution and its applications," *Neurocomputing*, vol. 223, pp. 107–117, Feb. 2017.

[27] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. K. Suykens, "Learning with tensors: A framework based on convex optimization and spectral regularization," *Mach. Learn.*, vol. 94, no. 3, pp. 303–351, 2014.

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

[29] W. Cao, Y. Wang, C. Yang, X. Chang, Z. Han, and Z. Xu, "Folded-concave penalization approaches to tensor completion," *Neurocomputing*, vol. 152, pp. 261–273, Mar. 2015.

[30] Y.-L. Chen, C.-T. Hsu, and H.-Y. M. Liao, "Simultaneous tensor decomposition and completion using factor priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 577–591, Mar. 2014.

[31] T. Zhou, H. Shan, A. Banerjee, and G. Sapiro, "Kernelized probabilistic matrix factorization: Exploiting graphs and side information," in *Proc. SDM*, vol. 12, 2012, pp. 403–414.

[32] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[33] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Rank regularization and Bayesian inference for tensor completion and extrapolation," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5689–5703, Nov. 2013.

[34] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, "Tensor factorization using auxiliary information," *Data Mining Knowl. Discovery*, vol. 25, no. 2, pp. 298–324, 2012.

[35] Y. Liu, F. Shang, W. Fan, J. Cheng, and H. Cheng, "Generalized higher order orthogonal iteration for tensor learning and decomposition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2551–2563, Dec. 2016.

[36] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian CP factorization of incomplete tensors with automatic rank determination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1751–1763, Sep. 2015.

[37] Q. Zhao, G. Zhou, L. Zhang, A. Cichocki, and S.-I. Amari, "Bayesian robust tensor factorization for incomplete multiway data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 736–748, Apr. 2016.

[38] T. Yokota, Q. Zhao, and A. Cichocki, "Smooth PARAFAC decomposition for tensor completion," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5423–5436, Oct. 2016.

[39] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[40] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[41] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[42] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[43] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *J. Comput. Graph. Statist.*, vol. 7, no. 3, pp. 397–416, 1998.

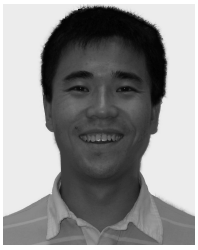
[44] M.-X. Hou, Y.-L. Gao, J.-X. Liu, J.-L. Shang, and C.-H. Zheng, "Comparison of non-negative matrix factorization methods for clustering genomic data," in *Proc. Int. Conf. Intell. Comput.*, Lanzhou, China, Springer, 2016, pp. 290–299.
[45] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.
[46] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Sov. Math. Dokl.*, vol. 27, no. 2, pp. 372–376, 1983.
[47] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, "Smoothing proximal gradient method for general structured sparse regression," *Ann. Appl. Statist.*, vol. 6, no. 2, pp. 719–752, 2012.
[48] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: An optimal gradient method for nonnegative matrix factorization," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2882–2898, Jun. 2012.
[49] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[50] X. Wang, Z. Li, and D. Tao, "Subspaces indexing model on Grassmann manifold for image search," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2627–2635, Sep. 2011.
[51] Y. Han and F. Moutarde, "Analysis of network-level traffic states using locality preservative non-negative matrix factorization," in *Proc. IEEE 14th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 501–506.
[52] Q. Gu, J. Zhou, and C. Ding, "Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs," in *Proc. SDM*, 2010, pp. 199–210.
[53] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 166–180, May 2018.
[54] L. Li, S. He, J. Zhang, and B. Ran, "Short-term highway traffic flow prediction based on a hybrid strategy considering temporal–spatial information," *J. Adv. Transp. Banner*, vol. 50, no. 8, pp. 2029–2040, 2016.
[55] X.-Y. Xu, J. Liu, H.-Y. Li, and M. Jiang, "Capacity-oriented passenger flow control under uncertain demand: Algorithm development and real-world case study," *Transp. Res. E, Log. Transp. Rev.*, vol. 87, pp. 130–148, Mar. 2016.
[56] B. Ran, L. Song, J. Zhang, Y. Cheng, and H. Tan, "Using tensor completion method to achieving better coverage of traffic state estimation from sparse floating car data," *PLoS ONE*, vol. 11, no. 7, p. e0157420, 2016.
[57] H. Tan, Y. Wu, B. Shen, P. J. Jin, and B. Ran, "Short-term traffic prediction based on dynamic tensor completion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2123–2133, Aug. 2016.
[58] V. W. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang, "Collaborative filtering meets mobile recommendation: A user-centered approach," in *Proc. AAAI*, 2010, pp. 236–241.
[59] L. Sorber, M. V. Barel, and L. D. Lathauwer, "Structured data fusion," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 586–600, Jun. 2015.
[60] Q. Dai and W. Sha. (2009). "The physics of compressive sensing and the gradient-based recovery algorithms." [Online]. Available: https://arxiv.org/abs/0906.1487
[61] R. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 2015.
[62] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

Yuankai Wu received the master's degree from the Department of Transportation Engineering, Beijing Institute of Technology, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Mechanical Engineering.

He has been a Visiting Ph.D. Student with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA, since 2018. He is currently an Assistant Researcher with the Institute on Internet of Mobility, Southeast University, Nanjing, China, and the University of Wisconsin-Madison. His current research interests include intelligent transportation systems and machine learning.

Huachun Tan (M'07) received the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2006.

He is currently a Professor with the School of Transportation Engineering, Southeast University, Nanjing, China. His current research interests include image engineering, pattern recognition, and intelligent transportation systems.



Yong Li received the M.Sc. degree in applied mathematics under the supervision of Prof. G. Misiolek and the Ph.D. degree under the supervision of Prof. R. L. Stevenson from the University of Notre Dame, Notre Dame, IN, USA.

He is currently an Associate Professor with the School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include multispectral image registration, object detection, deep learning, and super-resolution.

Jian Zhang (M'11) received the Ph.D. degree in transportation engineering from Southeast University (SEU), Nanjing, China, in 2011.

He is currently an Assistant Professor with the School of Transportation, SEU, and the Executive Vice Director of the Research Center for Internet of Mobility, SEU. His current research interests include intelligent transportation systems, transportation applications of big data, and connected vehicles.

Dr. Zhang is a member of the American Society of Civil Engineers.

Xiaoxuan Chen received the master's degree from the Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA, USA, in 2010. He is currently pursuing the Ph.D. degree with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA.

His current research interests include intelligent transportation engineering, cellular probe technology, and big data analysis.