Low-Rank Hankel Tensor Completion for Traffic Speed Estimation

Xudong Wang®, Yuankai Wu, Dingyi Zhuang®, and Lijun Sun®, Senior Member, IEEE

Abstract—This paper studies the traffic state estimation (TSE) problem using sparse observations from mobile sensors. Most existing TSE methods either rely on well-defined physical traffic flow models or require large amounts of simulation data as input to train learning algorithms. Different from previous studies, in this paper we propose a purely data-driven and model-free solution. We consider TSE as a spatiotemporal matrix completion/interpolation problem and apply spatiotemporal delay embedding to transform the original incomplete matrix into a fourth-order Hankel structured tensor. By imposing a low-rank assumption on this tensor structure, we can approximate and characterize both global patterns and local patterns in a data-driven manner. We use a truncated nuclear norm of a balanced spatiotemporal unfolding to approximate the tensor rank and develop an efficient algorithm based on the Alternating Direction Method of Multipliers (ADMM) to solve the problem. The proposed framework only involves two hyperparameters, spatial and temporal window lengths, which are easy to set given the degree of data sparsity. To validate the effectiveness of our proposed method, we conducted numerical experiments on real-world high-resolution trajectory data, which demonstrated its superiority in some challenging scenarios. The proposed method shows great potential for solving the TSE problem using sparse observations from mobile sensors and can be applied in various traffic applications.

Index Terms—Spatio–temporal traffic data, traffic state estimation, missing data imputation, low-rank tensor completion, delay embedding transform.

I. INTRODUCTION

Understanding and modeling the evolution of traffic state over space and time is a critical problem in traffic flow research and traffic control and management. For example, ramp metering on the highway requires accurate traffic state information to control the traffic flow and mitigate traffic congestion [1]. Traffic state information, such as flow, speed and density, is typically collected from stationary/fixed sensors (e.g., loop detector/video camera) and moving sensors (e.g., floating cars). These sensors are referred to as either Eulerian sensors or Lagrangian sensors. However, neither type of sensor can provide a comprehensive picture of how the traffic state evolves in space and time. On one hand, Eulerian sensors, such as loop detectors, can continuously gather data over time, but they are only capable of covering a specific point or a limited segment of the road. As loop detectors can be expensive to install and maintain, they are often installed sparsely for a large-scale transportation network. On the other hand, Lagrangian sensors—such as floating cars equipped with GPS—can provide real-time traffic state variables (e.g., speed and spacing) along their trajectories [2]. In theory, with information from all vehicles on a road segment, it is possible to build the full/complete time-space diagram from the trajectory data and estimate traffic state accordingly. However, in practice, data is only available from a small fraction of vehicles (e.g., penetration rate less than 5%). As a result, the traffic state information collected from sparse Lagrangian sensors is also limited in space and time [3], [4].

Traffic State Estimation (TSE) refers to the task that estimates the complete spatiotemporal profiles of traffic variables based on the partial observations collected from sensing systems [5], [6]. There are two types of TSE approaches in the literature: model-based and data-driven methods. Model-based methods based on macroscopic traffic flow models—such as the first-order Lighthill-Whitham-Richards (LWR) model and its extensions [7], [8], [9]—to estimate traffic state. This type of approach combines fundamental diagram (FD) and partial differential equations (PDE) with other techniques (e.g., Kalman filter) to depict traffic dynamics (see e.g., [5]). The advantages of model-based methods are that they can reflect the basic traffic principles and provide insights to explain the estimation process. However, model-based methods highly depend on various theoretical assumptions of traffic physics, which might be too abstract and biased to model the noisy and stochastic real-world data. As a result, model-based methods often require reliable prior information to guide the estimation.

Thanks to the availability of large-scale traffic data and recent advances in machine learning, data-driven methods have become increasingly popular in TSE. The key idea of the data-driven approach is to learn and leverage the spatiotemporal correlation/dependency from the traffic data to achieve the estimation task. For example, the characteristics of traffic speed data can be captured by spatiotemporal kernels and then integrated into learning frameworks such as Gaussian processes [10], convolutional neural networks [4], and graph neural networks [11]. Essentially, such learning-based and data-driven methods require a large amount of training data to
effectively learn and characterize the complex spatiotemporal patterns in traffic data.

To deal with scenarios where data is limited and sparse, some recent studies have also tried to integrate the physics of traffic flow to guide the learning process in data-driven methods to achieve higher accuracy. For example, Yuan et al. [12] developed a physics-regularized Gaussian process (PRGP) model, and Huang and Agarwal [13] and Shi et al. [14], [15] developed a physics-informed deep learning (PIDL) model individually. We refer interested readers to [4], [12], and [14] for a summary of existing methods and representative studies. These physics-guided and physics-informed models provide an effective solution in case where observed data is sparse. In these models, the underlying physics of traffic flow, whether encoded as a prior or represented by learnable parameters, is a critical factor in determining the performance of traffic state estimation. In principle, the exact form of the physics model is assumed to be known in advance and consistent over the spatiotemporal domain of interest. These assumptions, however, may not be applicable in real-world scenarios and it cannot be guaranteed that a reliable and consistent macroscopic model can be derived from real-world data [12], [14].

In this paper, we propose a data-driven model to solve the high-resolution traffic speed estimation problem. Specially, we use highly sparse speed observations collected from floating car data to achieve this goal. We model the spatiotemporal traffic speed data as a multivariate time series matrix (location×time) and treat the estimation of the incomplete traffic speed as a matrix/tensor completion problem. Low-rank models are one of the most widely used approaches for matrix/tensor completion. The idea behind such models is that a small number of latent factors determine the traffic state matrix; namely, the spatiotemporal traffic state matrix has low-rank characteristics [16]. A variety of matrix/tensor completion models have been proposed, however, few of them can handle high missing rate scenarios, particular in cases where there is continuous column/row missing data.

Inspired by the Hankel matrix/tensor applied in diverse applications such as seismic processing [17], image inpainting [18], [19], time series forecasting [20], and signal processing/reconstruction [21], we propose a novel traffic speed estimation model, called Spatiotemporal Hankel Low-Rank Tensor Completion (STH-LRTC), to address the TSE problem with a high missing rate. The fundamental assumption of STH-LRTC is that the traffic state data has low-rank characteristics [16], [22], and the transformed Hankel tensor also shows smooth manifolds in the low-rank space [19]. STH-LRTC differs from traditional approaches by not imposing predefined physical priors, but rather by using a purely data-driven approach to learn and approximate the unknown spatiotemporal patterns, thus filling in the missing values.

The proposed STH-LRTC model has three steps. Firstly, we transform the original incomplete traffic state matrix into a fourth-order Hankel tensor by performing both spatial and temporal delay embedding (i.e., Hankelization). The principle idea of delay embedding is to recursively augment the multivariate time series. By doing so, the local information around missing values, such as adjacent locations/timestamps, can be extended in higher dimensions. Secondly, we complete the fourth-order Hankel tensor using a low-rank tensor completion technique. The sum-of-nuclear norm (SNN) [23] is widely used in approximating tensor rank; however, it shows high computation cost and suboptimal solution for a fourth-order tensor [24]. To accelerate the algorithm for the fourth-order tensor and have better completion performance, we unfold the tensor to a balanced spatiotemporal matrix [25] and apply truncated nuclear norm (TNN) [26], [27] to approximate the tensor rank. Finally, we obtain the estimated matrix by performing inverse Hankelization on the completed tensor. The main contribution of this paper is fourfold:

- We model spatiotemporal traffic state data (only one variable–speed) as a matrix and apply spatiotemporal delay embedding (i.e., Hankelization) to transform the traffic state matrix into a fourth-order Hankel tensor structure. As a result, the spatiotemporal correlation/dependency between observations and missing values is naturally augmented without any other additional constraints.
- We characterize the higher-order correlations/dependency in the traffic data by imposing a low-rank assumption on the spatiotemporal Hankel tensor. Thus, we can achieve TSE by performing low-rank completion on the Hankel tensor, and this approach is purely data-driven.
- We incorporate the spatiotemporal tensor unfolding strategy and truncated nuclear norm to improve the computation efficiency and performance for fourth-order tensor.
- We conduct a challenging TSE, on a real-world dataset, by estimating all traffic speed trajectories using only 5% of them. The results show that the proposed STH-LRTC offers superior performance than baseline models.

The remainder of this paper is organized as follows. In Section II, we introduce the preliminaries of this study. In Section III, we introduce the STH-LRTC model in detail and develop an Alternating Direction Method of Multipliers (ADMM) algorithm for model estimation. In Section IV, we present a case study using real-world traffic data to evaluate the performance of STH-LRTC. Section V concludes this study and discusses some directions for future research.

II. Preliminaries

A. Notations

We follow [23], [28] to define all notations. We use lowercase letters to denote scalars, e.g., $$x \in \mathbb{R}$$, boldface lowercase letters to denote vectors, e.g., $$x \in \mathbb{R}^N$$, boldface capital letters to denote matrices, e.g., $$X \in \mathbb{R}^{N \times T}$$, and boldface Euler script letters to denote higher-order tensors, e.g., $$\mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_R}$$. We denote the $$(i,j)$$th entry of a matrix by $$X_{ij}$$, and the $$(a,b,c,d)$$th entry of a fourth-order tensor by $$X_{abcd}$$. We use $$X_{ij} \in \mathbb{R}^{p \times T}$$ and $$X_{ij;j+p-1} \in \mathbb{R}^{N \times p}$$ to denote the sub-matrix that consists of all columns from $$i$$th row to $$(i+p-1)$$th row and all rows from $$j$$th column to $$(j+p-1)$$th column, respectively.

The Frobenius norm of a tensor is defined as

$$\|X\|_F = \sqrt{\sum_{i=1}^{I_1} \cdots \sum_{i=1}^{I_R} x_{i_1 \cdots i_R}^2}$$

The inner product of two tensors of
Fig. 1. Matrix representation of traffic state variables collected from floating cars. Traffic state data are discretized and averaged as a matrix. This representation transforms TSE into a matrix completion/interpolation problem. The dark blue dots represent raw spatiotemporal readings from floating cars. The green dot and the red arrows illustrate TV regularization defined in (2).

the same size is \( \langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_4=1}^{I_4} x_{i_1 \ldots i_4} y_{i_1 \ldots i_4} \). The tensor unfolding is the process of reordering the elements of an tensor into a matrix. Take a fourth-order tensor for example, the \( k \)-th mode \((k = 1, \ldots, 4)\) unfolding denoted as \( \mathbf{X}^{(k)} \) is reshape the tensor to a \( I_k \times (\prod_{i=1,i\neq k}^{4} I_i) \) matrix.

B. Problem Description

Traffic state variables are collected from floating cars within the spatiotemporal domain of \( L \times W \), where \( L \) represents the road length and \( W \) represents the time window length. Using a pre-defined spatial resolution of \( l_s \) and temporal resolution of \( T_s \), the original trajectories are transformed into an incomplete traffic variable matrix \( \mathbf{Y} \in \mathbb{R}^{N \times T} \), where \( N = L/l_s \) (spatial dimension) and \( T = W/l_t \) (temporal dimension). The value of each entry corresponds to the average traffic state variable of that entry.

This study defines TSE as a matrix completion problem using trajectories collected from floating cars. Specifically, we focus on highly sparse matrix completion, e.g., using 5% trajectories to infer the traffic speed of other unobserved locations or timestamps. In the rest of this paper, we refer to traffic state as the traffic speed variable only, but the model can be expanded to other traffic state variables. Fig. 1 shows an example of using three trajectories to estimate the missing traffic speed in both spatial and temporal domains. In this case, a high proportion of missing values (120/140 \( \approx \) 86%) is observed in a few select observations. Additionally, several consecutive periods exhibit a complete lack of observations.

C. Low-Rank-Based Matrix Completion

Let \( Z \in \mathbb{R}^{N \times T} \) be the desired underlying matrix and \( \Omega \) be the binary index set to indicate whether the value is observed or not: 0 represents missing value, 1 represents observed value. The low-rank-based matrix completion aims to solve the following optimization problem

\[
\min_\mathbf{Y} \text{ rank}(\mathbf{Y}) + \gamma \mathcal{R}, \quad \text{s.t. } Z_\Omega = Y_\Omega, \tag{1}
\]

where \( \text{rank}(\mathbf{Y}) \) denotes the rank of \( \mathbf{Y} \) which can reflect the global consistency, and \( \mathcal{R} \) denotes regularization term used to incorporate the spatiotemporal constraint (local consistency) underlying the data to avoid overfitting. The hyper-parameter \( \gamma \) is used to trade-off the global consistency and the local consistency.

The matrix rank can be approximated by convex approximation, such as matrix nuclear norm. The regularization term is crucial to matrix completion as the correlation between unknown values and observations can be captured. For example, the total variation (TV) regularization is widely used in matrix/tensor completion problem [29]:

\[
\mathcal{R} = \sum_{i=1}^{N-1} \sum_{j=1}^{T-1} |Y_{i,j} - Y_{i+1,j}| + |Y_{i,j} - Y_{i,j+1}| + \sum_{i=1}^{N-1} |Y_{i,T} - Y_{i+1,T}| + \sum_{j=1}^{T-1} |Y_{N,j} - Y_{N,j+1}|. \tag{2}
\]

The TV regularization assumes that neighbor locations (same row) and timestamps (same column) are likely to have similar values, e.g., using the known information from the red arrow’s direction to estimate the green dot in Fig. 1. However, it should be noted that the performance of matrix completion models may degrade when entire consecutive rows or columns of data are missing, as there is no neighbor information available to inform the estimation of missing values. In such scenarios, the low-rank structure and regularization term may not be able to fully capture the information contained in the limited number of observations.

In order to address the challenge of high missing rates in the context of consecutive missing data, we propose a novel approach that leverages the spatiotemporal Hankel structure of the data matrix. Specifically, we convert the matrix completion problem into a Hankel tensor completion problem by incorporating the spatiotemporal Hankel structure on the matrix. This transformation enables us to exploit the underlying spatiotemporal correlations in the data, which are often useful in modeling dynamic processes. The technical details of our approach are provided in the following section, where we present a rigorous mathematical formulation of the Hankel tensor completion problem and describe the algorithmic implementation of the proposed method.

III. METHODOLOGY

This section introduces a spatiotemporal Hankel low-rank tensor completion (STH-LRTC) model for estimating the traffic variables. Firstly, we present a spatiotemporal Hankel tensor transformation in Section III-A and a balanced spatiotemporal unfolding in III-B. Subsequently, we propose an efficient low-rank tensor completion algorithm in Section III-C. Finally, measurements and convergence criteria are given in Section III-D. Overall, the proposed STH-LRTC model consists of three main steps: (1) tensor Hankelization, (2) Hankel tensor completion, and (3) inverse tensor Hankelization, as illustrated in Fig. 2.

A. Spatio–Temporal Hankel Tensor Transformation

We define a spatiotemporal Hankelization operation \( \mathcal{H}_{t_s, t_t} \) with spatial embedding length \( t_s \) and temporal embedding...
The Hankel tensor. For example, for the embedding length \( \tau \), we expect the corresponding inverse Hankelization operation, denoted as \( \mathcal{H}_{\tau_1, \tau_2}^{-1} (\mathcal{X}) \), transforms a Hankel tensor \( \mathcal{X} = \mathcal{H}_{\tau_1, \tau_2} (X) \subseteq \mathbb{R}^{t \times \tau_1 \times (N-\tau_1+1) \times (T-\tau_2+1)} \) to a matrix \( \hat{X} \) [19] by computing the average of the corresponding entries in the Hankel tensor. For example, for the \((i, j)\)th entry in \( \hat{X} \), we estimate it by taking the average of all the entries in the tensor that are generated from \( X_{i, j} \) during the Hankelization process.

Figure 2 demonstrates an example of tensor Hankelization and inverse tensor Hankelization with spatial delay embedding length \( \tau_s = 3 \) and temporal delay embedding length \( \tau_t = 4 \), shown in Step 1 and Step 3, respectively. It is worth noting that the delay embedding process can be extended to multi-dimensional tensors by following similar operations [19].

### B. Spatio–Temporal Tensor Unfolding and Folding

As the underlying true Hankel tensor is low-rank, we develop the following optimization problem to complete the missing values:

\[
\min_{\mathcal{X}} \text{rank}(\mathcal{X})
\]

subject to:

\[
\mathcal{X} = \mathcal{H}_{\tau_1, \tau_2} (Z),
\]

\[
\text{s.t. } \quad Z_{\Omega} = Y_{\Omega},
\]

where \( Y \) and \( Z \) are the incomplete matrix and desired underlying matrix, respectively.

Computing the tensor rank is an NP-hard problem. One common approach is applying a convex surrogate to approximate tensor rank, such as the sum-of-nuclear norms (SNN) of all tensor unfoldings (matrization): \( \| \mathcal{X} \|_* := \sum_{k=1}^{K} \alpha_k \| \mathcal{X}^{(k)} \|_* \), where \( \alpha_k \geq 0 \) and \( \sum_{k=1}^{K} \alpha_k = 1 \) [23]. The matrix nuclear norm (NN) is defined as \( \| \mathcal{X}^{(k)} \|_* = \sum_{i} \sigma_i (\mathcal{X}^{(k)}) \) is the \( i \)th largest singular value of \( \mathcal{X}^{(k)} \). Another approach is based on heuristic models, such as Tucker decomposition and CANDECOMP/PARAFAC (CP) decomposition. However, such models require prior knowledge to tune the tensor rank, which significantly affects the completion performance. Liu et al. [23] showed that the convex formulation SNN outperforms Tucker decomposition and CP decomposition in completing the missing values in their third-order tensor scenarios.

While the sum-of-nuclear norms (SNN) has been successfully applied in various applications, its computation cost is high when dealing with high-dimensional tensors because it requires applying singular value decomposition (SVD) for \( K \) unfolding matrices. To improve the efficiency of tensor completion, Mu et al. [25] proposed a square norm as an approximation of the tensor rank. The square norm only calculates the nuclear norm of one tensor unfolding, a more balanced matrix, instead of calculating the nuclear norm of all tensor unfoldings. Inspired by this work, we propose a spatiotemporal unfolding operation that transforms the Hankel tensor \( \mathcal{X} \) to a matrix \( \hat{X} \subseteq \mathbb{R}^{p \times q} \). Following MATLAB notation, the spatiotemporal unfolding operation can be
written as:

\[ X_{\square} = \text{reshape}(X, [p, q]), \]  

(5)

where \( p = r_s \times r_t \) and \( q = (N - r_s + 1) \times (T - r_t + 1) \). Correspondingly, the spatiotemporal folding operation that converts the matrix \( X_{\square} \) to the Hankel tensor \( X \) can be denoted by the following MATLAB operation:

\[ X = \text{reshape}(X_{\square}, [r_s, r_t, N - r_s + 1, T - r_t + 1]). \]  

(6)

The spatiotemporal tensor unfolding and folding processes are shown in Fig. 2 Step 2. Instead of seeking a more balanced/square matrix as in the square norm approach [25], we organize the fourth-order Hankel tensor as a \( p \times q \) matrix for two reasons. First, the computation cost of calculating matrix rank is determined by the smaller dimension of the matrix (see variable \( X \) update process). Therefore, \( X_{\square} \) is faster to compute than any other reshaped matrices, since \( r_s \ll N, r_t \ll T \). Second, each \( r_s \times r_t \) patch captures local spatiotemporal correlations simultaneously, and this local consistency is reflected in a single column of the spatiotemporal unfolding matrix.

Combining Hankelization and spatiotemporal unfolding operations provides an effective data-driven and model-free solution to characterize complex spatiotemporal dependencies. For example, consider the missing value marked by the red dot in Fig. 2. Since there are no observations at the same column/row as the red dot in \( X \), it is difficult to estimate its value. However, in \( X_{\square} \), the Hankelization and spatiotemporal unfolding operations relate the red dot to other known values, i.e., the neighbors of the red dot in \( X \). The two hyperparameters, \( r_s \) and \( r_t \), can be easily adjusted to construct the correlation/dependency structure of the data. When the data becomes highly sparse, one can increase the values of \( r_s \) and/or \( r_t \) to capture and encode correlations in an expanded spatiotemporal domain.

C. Low-Rank Tensor Completion With Truncated Square Norm Minimization

We use the truncated nuclear norm (TNN) of the spatiotemporal unfolding matrix, i.e., \( \|X_{\square}\|_{r_s} \), as a more accurate approximation to the rank of tensor \( X \) [26]. The optimization problem (4) can be transformed into:

\[
\begin{align*}
\min_{X} & \quad \|X_{\square}\|_{r_s} \\
\text{s.t.} & \quad X = H_{r_s,r_t}(Z), \\
& \quad Z_{\Omega} = Y_{\Omega}.
\end{align*}
\]  

(7)

where \( \|X_{\square}\|_{r_s} = \sum_{i=r_s+1}^{p,q} \sigma_i (X_{\square}) \) with a positive integer \( r < \min\{p, q\} \) [27].

The Alternating Direction Method of Multipliers (ADMM) method is excellent for solving the separable convex optimization problems [30], and it has been applied in most existing convex-based matrix/tensor completion models [18], [23], [25], [26], [27]. In our work, we also use the ADMM to solve the problem (7). The augmented Lagrangian function of (7) is given to set up an ADMM framework:

\[
\begin{align*}
\mathcal{L}(X, Z, E) = & \|X_{\square}\|_{r_s} + \frac{\rho}{2} \|X - H_{r_s,r_t}(Z)\|_F^2 \\
& + \frac{\rho}{2} \|H_{r_s,r_t}(Z) - E\|_F^2,
\end{align*}
\]  

(8)

where \( E \) is a dual variable and \( \rho > 0 \) is a penalty parameter. The ADMM framework divides the optimization problem (7) into several suboptimal problems; therefore, the variables \( X, Z, \) and \( E \) are updated alternatively:

\[
\begin{align*}
X^{\ell+1} = & \arg \min_X \mathcal{L}(X, Z^\ell, E^\ell), \\
Z^{\ell+1} = & \arg \min_Z \mathcal{L}(X^{\ell+1}, Z, E^\ell), \\
E^{\ell+1} = & E^\ell + \rho \left( X^{\ell+1} - H_{r_s,r_t}(Z^{\ell+1}) \right).
\end{align*}
\]  

(9)

The convergence rate becomes slow when applying ADMM to solve problems with many variables and constraints; therefore, we use an adaptive penalty parameter, \( \rho^{\ell+1} = \beta \rho^\ell \) with \( \beta \in [1.0, 1.2] \) to accelerate the convergence [31]. The inference of updating \( X \) and \( Z \) are shown below.

1) Update Variable \( X \):

\[
\begin{align*}
X^{\ell+1} = & \arg \min_X \frac{1}{\rho} \|X_{\square}\|_{r_s} \\
+ & \frac{1}{2} \|X - H_{r_s,r_t}(Z^\ell) - \frac{1}{\rho} E^\ell\|_F^2, \\
= & \text{reshape} \left( D_1/\rho \left( H_{r_s,r_t}(Z^\ell) - \frac{1}{\rho} E^\ell \right) \right),
\end{align*}
\]  

(10)

where \( \text{dim} = [r_s, r_t, N - r_s + 1, T - r_t + 1] \) and \( D_1(\cdot) \) denotes the singular value shrinkage operator described in the following Lemma 1.

Lemma 1 [22]: Let \( A = U \Sigma V^T \) be the SVD of a matrix \( A \in \mathbb{R}^{p \times q} \). For any \( \lambda \geq 0 \) and \( X \in \mathbb{R}^{p \times q} \), we have

\[
D_1(\lambda) = \arg \min_Z \lambda \|X\|_{r_s} + \frac{1}{2} \|X - A\|_F^2,
\]  

(11)

where \( D_1(\lambda) = U \Sigma_\lambda V^T \) and \( \Sigma_\lambda = \text{diag}(\sigma_1, \cdots, \sigma_r, \max(\sigma_{r+1} - \lambda, 0), \cdots, \max(\sigma_{\text{min}(p,q)} - \lambda, 0)) \).

2) Update Variable \( Z \):

\[
\begin{align*}
Z^{\ell+1} = & \arg \min_Z \rho \|X^{\ell+1} - H_{r_s,r_t}(Z^\ell)\|_F^2 \\
& + \frac{\rho}{2} \|X^{\ell+1} - H_{r_s,r_t}(Z^{\ell+1})\|_F^2.
\end{align*}
\]  

(12)

We denote the objective function in (12) by \( L_\rho(Z) \). The optimization can be solved by letting the partial gradient

\[
\frac{\partial L_\rho(Z)}{\partial Z} = 0.
\]

Therefore, the update equation of \( Z^{\ell+1} \) is:

\[
\begin{align*}
Z^{\ell+1}_{\Omega} = & \left( H_{r_s,r_t}^{-1} \left( X^{\ell+1} - \frac{1}{\rho} E^\ell \right) \right)_{\Omega}, \\
Z^{\ell+1} = & Y_{\Omega}.
\end{align*}
\]  

(13)

The equations above ensure that only missing values are estimated at each iteration while the observed values remain fixed.
D. Implementation

In the following imputation experiments, we utilize the partially observed speed data from floating cars to recover the full traffic speed. To accomplish this, we randomly select trajectories and use them to generate a training index set \( \Omega \). We use mean absolute error (MAE) and root mean squared error (RMSE) to measure the performance of the model:

\[
\text{MAE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |x_i - \hat{x}_i|,
\]

\[
\text{RMSE} = \sqrt{\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (x_i - \hat{x}_i)^2},
\]

where \( N_{\text{test}} \) is the number of test data, \( x_i \) is the observation and \( \hat{x}_i \) is the estimation.

The following convergence criterion is applied to stop algorithm:

\[
\frac{||Z^{\ell+1} - Z^{\ell}||_F}{||Y_\Omega||_F} < \epsilon,
\]

where \( Z^{\ell+1} \) and \( Z^\ell \) denote the recovered matrices at two consecutive iterations. Algorithm 1 summarizes the proposed method STH-LRTC.

**Algorithm 1** Spatiotemporal Hankel Low Rank Tensor Completion (STH-LRTC)

**Input:** \( Y_\Omega \in \mathbb{R}^{N \times T}, \tau_s, \tau_t, \epsilon, \rho, \rho_{\text{max}}, \beta \) and \( r \)

**Output:** \( Z \)

**Initialize:** \( \ell = 1 \)

1. \( Z^{\ell}_\Omega = Y_\Omega \) and \( Z^{\ell}_\bar{\Omega} = 0 \)
2. \( E^{\ell} = Z^\ell \)
3. **while** not converged **do**
   4. \( Z^{\ell}_\Omega = H_{\tau_s,\tau_t}(Z^\ell) \) and \( E^{\ell} = H_{\tau_s,\tau_t}(E^\ell) \)
   5. \( \chi^{\ell+1} = \text{reshape}(D_{1/\rho^\ell} \left( Z^{\ell}_{\bar{\Omega}} - \frac{1}{\rho^\ell} E^{\ell}_{\bar{\Omega}} \right), \text{dim}) \)
   6. \( Z^{\ell+1}_\Omega = H_{\tau_s,\tau_t}(\chi^{\ell+1} - \frac{1}{\rho^\ell} E^{\ell}) \) and \( Z^{\ell+1}_{\bar{\Omega}} = Y_\Omega \)
   7. \( E^{\ell+1} = E^\ell + \rho^\ell (H_{\tau_s,\tau_t}(\chi^{\ell+1}) - Z^{\ell+1}) \)
   8. \( \rho^{\ell+1} = \min(\beta \rho^\ell, \rho_{\text{max}}) \)
   9. \( \ell = \ell + 1 \)
10. **end while**

IV. CASE STUDY

In this section, we apply the proposed STH-LRTC on a real-world traffic dataset to evaluate the TSE performance. Our scenario settings follow the real-world data collection process of floating cars (see e.g., [2]). Our aim is to recover the full traffic speed with high-resolution trajectories observed from a small number of floating cars. We compare the proposed STH-LRTC with several baseline models. The code for algorithm and experiment is available at github https://github.com/mcgill-smart-transport/traffic_state_estimation.

A. NGSIM Traffic Speed Data

The real-world mobile traffic speed dataset used in this paper is an open-source dataset from the Next Generation simulation (NGSIM) program [32] on southbound US highway 101 of lane 2. The traffic trajectories were collected by digital video cameras. We select the vehicle trajectories with a global time of fewer than 2700 seconds along a 1500 ft long section, resulting in a total of 1239 vehicles. To organize the trajectories as a matrix, we average the traffic speed with the resolution of \( l_t = 10 \) ft and \( l_i = 5 \) s. We remove several rows and columns at the beginning and end of the aggregated traffic speed data, which contained all-zero values. After preprocessing, the size of the traffic speed matrix is \( 130 \times 480 \), representing a spatiotemporal domain of 1300 ft and 2400 s.

The proposed model is based on the assumption that the data has low-rank characteristics. To confirm this, we use SVD to calculate the cumulative eigenvalue percentage (CEP) of the NGSIM data. As shown in Fig. 3, a small number of leading singular values make a significant contribution, with the first 42 singular values covering 90% of all singular values. This indicates the low-rank feature of the NGSIM data. It has been shown in [19] that the Hankel tensor of a low-rank matrix also exhibits smooth manifolds in the low-rank space.

B. Baseline Models

We compare the proposed STH-LRTC method with matrix-based models and Hankel tensor-based models:

- Matrix factorization with TV regularization (MFTV) defined in (1) and (2).
- Adaptive smoothing interpolation method (ASM) [10]: a spatiotemporal kernel weighted method that considers different traffic state (i.e., free-flow/congested) characteristics.
- Spatiotemporal Hankel tensor with Tucker decomposition (STH-Tucker) [19]: using Tucker-based tensor completion model with the spatiotemporal Hankelization to complete the missing values.
- Spatiotemporal Hankel tensor with sum-of-nuclear-norm (STH-SNN): using SNN [23] to approximate the tensor rank of spatiotemporal Hankel tensor.

The hyper-parameters in each model greatly affect the imputation performance. For a fair comparison, the baseline models are fine-tuned. We set \( \gamma = 1 \) to balance the low-rank
part and the regularization part in MFTV model. For the ASM model, we set the parameters according to the suggested values in [10] and the characteristics of data: free-flow speed $c_{\text{free}} = 54.6$ ft/s, propagation speed of congestion $c_{\text{cong}} = -10.0$ ft/s, crossover from congested to free traffic $V_{\text{th}} = 40.0$ ft/s, and transition width between congested and free traffic $\Delta V = 9.1$ ft/s. The spatial smoothing width $s$ and the temporal smoothing width $\tau$ are 200 ft and 10 s, respectively.

The choice of embedding lengths $\tau_x$ and $\tau_t$ is crucial for the Hankel tensor-based model. If the embedding length is too small, it may not capture enough spatiotemporal information, especially when the missing rate is high. However, a large embedding length can capture more information, leading to heavy computation. In the experiment, we set $\tau_x = 40$ and $\tau_t = 30$ to obtain the Hankel tensor for STH-Tucker, STH-SNN and STH-LRTC. We set the core tensor rank $R = [4, 4, 8, 16]$ and $\alpha = [0.1, 0.4, 0.1, 0.4]$ for the STH-Tucker model and the STH-SNN model, respectively. The truncated parameter $r = \lfloor 0.05N \rfloor$ [22] for the proposed model.

We set the convergence $\epsilon = 1 \times 10^{-3}$ for all low-rank based models, the ADMM penalty parameter $\rho = 5 \times 10^{-6}$ and the adaptive parameter $\beta = 1.1$ for all ADMM-based models, i.e., MFTV, STH-SNN and STH-LRTC.

### C. Traffic Speed Estimation Results

Our scenario settings follow the experimental implementation in [4], we randomly select 5% trajectories/vehicles (62 out of 1239) to construct the incomplete traffic speed matrix. It has to be noted that the training data is slightly different from the ground truth in the same data entries since we only use the average speed of the selected 5% vehicles to create the incomplete matrix. We repeat the experiment 10 times by randomly selecting different training trajectories, resulting in an average missing rate of 88%, meaning that we use 12% observations to estimate the other 88% data.

Table I presents the average MAE and RMSE obtained from the baseline models and STH-LRTC, numerically demonstrating the missing value completion performance of these models. Despite the highly sparse traffic speed data, we see that both ASM and the proposed model outperform the other models. However, the ASM implementation has a data leakage issue since the hyper-parameters are selected/estimated based on the ground truth. When the ground truth is not accessible, selecting and tuning the hyper-parameters is challenging, resulting in the failure to estimate the traffic state. Both STH-Tucker and STH-SNN exhibit larger average MAE and RMSE than the proposed truncated square norm method, even with the same spatiotemporal Hankel structure. In particular, the STH-Tucker model has the highest average MAE and RMSE, with a high standard deviation, as it is challenging to set the proper rank for a fourth-order tensor in practice.

To provide a better understanding of the results, we have chosen one training data and corresponding results, except for the STH-Tucker model in Fig. 4. The aggregated traffic speed from all the trajectories (ground truth) is shown in Fig. 4 (a), in which we can observe several shockwaves backpropagate along the highway. Fig. 4 (b) shows one of the selected training data. Several consecutive periods have either entirely missing values or only a few values, such as from 300 s to 340 s and from 1185 s to 1265 s. The high portion of missing grids makes it challenging for models to recover the entire piece of the road traffic speed.

Fig. 4 (c) shows the TSE result of the MFTV model. Due to the TV regularization, close timestamps and locations have similar values. The MFTV model can estimate the missing values but cannot fully capture details when the missing rate is high. Thanks to the data augmentation power offered by delay embedding. STH-SNN (Fig. 4 (d)) shows slightly better performance than MFTV, but it fails to depict complicated traffic states. The completion results of ASM and the proposed model are shown in Fig. 4 (e) and Fig. 4 (f), respectively. Both models can estimate the shockwaves, but we can see that STH-LRTC shows superior imputation performance. For example, ASM provides oversmoothed results so that the transitions between waves are blurred, but STH-LRTC can recover the traffic states well, even in the transition state. Additionally, ASM fails to recover small shock waves (highlighted with black rectangular) while STH-LRTC can estimate these small shock waves successfully.

### D. Interpretation of STH-LRTC

In this subsection, we explain why STH-LRTC works for our traffic data. In fact, we believe this model could work for many types of spatiotemporal data. Figure 5(a) displays the square transformation of the true traffic speed data, where each column represents the vectorization of a $\tau_x \times \tau_t$ patch. We randomly select 1000 columns out of 41041 (91 $\times$ 451) columns and quantify the 1000 $\times$ 1000 covariance matrix, as shown in Figure 5(b). The singular values of the covariance matrix (Figure 5(c)) indicate that the covariance matrix is a highly low-rank matrix. Therefore, the complete picture of the traffic data can be captured by a few spatiotemporal patches. The patch-based constraint breaks the limitation of TV, making local spatiotemporal patterns (sub-matrices) similar regardless of whether they are adjacent or not. Additionally, after obtaining the completed data in the spatiotemporal unfolding matrix, the inverse Hankelization guarantees that the corresponding entries generated from the Hankelization process are identical by averaging the estimations in the delay embedding space.

In summary, STH-LRTC can complete the scenario of high missing value rate for two reasons. First, the low-rank

| Method          | MAE      | RMSE     |
|-----------------|----------|----------|
| Matrix          |          |          |
| MFTV            | 6.45 (0.58) | 8.79 (0.90) |
| ASM             | 5.05 (0.19) | 6.62 (0.31) |
| Hankel tensor   |          |          |
| STH-Tucker      | 7.18 (1.01) | 9.66 (1.20) |
| STH-SNN         | 6.60 (0.83) | 8.58 (1.06) |
| Proposed        | **4.69 (0.25)** | **6.27 (0.40)** |
characteristic calculated by the truncated nuclear norm can exploit the global information behind the data, allowing for the discovery of distinct shapes of stop waves. Second, the delay embedding enhances the spatial and temporal information to capture the local consistency in the data, which is particularly helpful in tackling the problem of consecutive missing values.

V. DISCUSSION AND CONCLUSION

This study focuses on recovering spatiotemporal traffic state data (average traffic speed in our case study) from sparse trajectory-based observations, such as a limited number of floating cars. Unlike previous TSE studies, which require either physics models of traffic flow or large amounts of simulation data, we propose a purely data-driven and model-free approach based on tensor completion in a spatiotemporal Hankel delay embedded space. To achieve this, we treat the spatiotemporal traffic state diagram as a matrix with one dimension representing time and the other representing space. We obtain a fourth-order Hankel tensor structure by applying a two-way delay embedding transform (i.e., spatiotemporal Hankelization). As a powerful data augmentation solution, spatiotemporal delay embedding not only preserves the global pattern of the original data but also introduces a higher-order dependency/correlation structure within a local spatiotemporal domain.

The proposed framework involves only two hyperparameters: the spatial window length \( \tau_s \) and the temporal window length \( \tau_t \), which affect the size of the encoded local correlations. In practice, these two parameters can be determined based on the degree of data sparsity and computational cost. We propose an efficient algorithm, STH-LRTC, to complete the Hankel tensor by minimizing the truncated nuclear norm of the spatiotemporal unfolding matrix. We conduct a TSE experiment on real-world high-resolution trajectory data.
and our results verify the effectiveness and superiority of the proposed framework, even with an extremely sparse dataset (88% missing values).

There are several potential directions for future research. Firstly, the transductive nature of matrix/tensor completion means that the model needs to be trained from scratch for each case, which could result in high computational costs for real-time applications, particularly when the data is sparse and requires a large embedding size of $t_r \times t_r$. One possible solution is to leverage model updating methods and inductive models, such as graph convolutional neural networks [11], to achieve real-time estimation. Secondly, our current method uses the truncated nuclear norm on balanced tensor unfolding to approximate the rank of the Hankel-structured tensor; however, this may not be the best approach considering the strong spectral information encoded in spatiotemporal traffic data. One possible direction for future research is to explore alternative rank approximation techniques, such as tensor tubal nuclear norm [33], [34], to estimate the tensor rank. Thirdly, our model does not explicitly incorporate local smoothness. For instance, if two patches are close in space and time, they should have similar spatiotemporal characteristics. Therefore, incorporating local smoothness into the model could also be a promising direction for future research.

REFERENCES

[1] M. Papageorgiou and A. Kotsialos, “Freeway ramp metering: An overview,” IEEE Trans. Intell. Transp. Syst., vol. 3, no. 4, pp. 271–281, Dec. 2002.

[2] T. Seo, T. Kusakabe, and Y. Asakura, “Estimation of flow and density using probe vehicles with spacing measurement equipment,” Transp. Res. C, Emerg. Technol., vol. 53, no. 6, pp. 134–150, 2015.

[3] W. Ma and S. Qian, “High-resolution traffic sensing with probe autonomous vehicles: A data-driven approach,” Sensors, vol. 21, no. 2, p. 464, Jan. 2021.

[4] B. T. Thodi, Z. S. Khan, S. E. Jabari, and M. Menendez, “Incorporating kinematic wave theory into a deep learning method for high-resolution traffic speed estimation,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 10, pp. 17849–17862, Oct. 2021.

[5] Y. Wang and M. Papageorgiou, “Real-time freeway traffic state estimation based on extended Kalman filter: A general approach,” Transp. Res. B, Methodol., vol. 39, no. 2, pp. 141–167, 2005.

[6] T. Seo, A. M. Bayen, T. Kusakabe, and Y. Asakura, “Traffic state estimation on highway: A comprehensive survey,” Ann. Rev. Control, vol. 43, pp. 128–151, Jan. 2017.

[7] J. C. Herrera and A. M. Bayen, “Incorporation of Lagrangian measurements in freeway traffic state estimation,” Transp. Res. B, Methodol., vol. 44, no. 4, pp. 460–481, May 2010.

[8] Y. Yuan, J. W. C. van Lint, R. E. Wilson, F. van Wageningen-Kessels, and S. P. Hoogendoorn, “Real-time Lagrangian traffic state estimator for freeways,” IEEE Trans. Intell. Transp. Syst., vol. 13, no. 1, pp. 59–70, Mar. 2012.

[9] A. Duret and Y. Yuan, “Traffic state estimation based on Eulerian and Lagrangian observations in a mesoscopic modeling framework,” Transp. Res. B, Methodol., vol. 101, pp. 51–71, Jul. 2017.

[10] M. Treiber, A. Kesting, and R. E. Wilson, “Reconstructing the traffic state by fusion of heterogeneous data,” Comput. Aided Civil Infrastruct. Eng., vol. 26, no. 6, pp. 408–419, Aug. 2011.

[11] Y. Wu, D. Zhang, A. Labbe, and L. Sun, “Inductive graph neural networks for spatiotemporal kriging,” 2020, arXiv:2006.07527.

[12] Y. Yuan, Z. Zhang, X. T. Yang, and S. Zhe, “Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation,” Transp. Res. B, Methodol., vol. 146, pp. 88–110, Apr. 2021.

[13] A. J. Huang and S. Agarwal, “Physics informed deep learning for traffic state estimation,” in Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC), Sep. 2020, pp. 1–6.

[14] R. Shi, Z. Mo, K. Huang, X. Di, and Q. Du, “A physics-informed deep learning paradigm for traffic state and fundamental diagram estimation,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8, pp. 11688–11698, Aug. 2022.

[15] R. Shi, Z. Mo, and X. Di, “Physics-informed deep learning for traffic state estimation: A hybrid paradigm informed by second-order traffic models,” in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 1, 2021, pp. 540–547.

[16] X. Wang and L. Sun, “Diagnosing spatiotemporal traffic anomalies with low-rank tensor autoregression,” IEEE Trans. Intell. Transp. Syst., vol. 22, no. 12, pp. 7904–7913, Dec. 2021.

[17] S. Trickert, L. Burroughs, and A. Milton, “Interpolation using Hankel tensor completion,” in SEG Technical Program Expanded Abstracts. Houston, TX, USA: Society of Exploration Geophysicists, 2013, pp. 3634–3638.

[18] K. H. Jin and J. C. Ye, “Anihilating filter-based low-rank Hankel matrix approach for image inpainting,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 3498–3511, Nov. 2015.

[19] T. Yokota, B. Erem, S. Guler, S. K. Warfield, and H. Hontani, “Missing slice recovery for tensors using a low-rank model in embedded space,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8251–8259.

[20] Q. Shi et al., “Block Hankel tensor ARIMA for multiple short time series forecasting,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 2, 2020, pp. 5758–5766.

[21] F. Sedighin, A. Cichocki, T. Yokota, and Q. Shi, “Matrix and tensor completion in multiway delay embedded space using tensor train, with application to signal reconstruction,” IEEE Signal Process. Lett., vol. 27, pp. 810–814, 2020.

[22] X. Chen, J. Yang, and L. Sun, “A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation,” Transp. Res. C, Emerg. Technol., vol. 117, Aug. 2020, Art. no. 102673.

[23] 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.

[24] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, “Tensor robust principal component analysis with a new tensor nuclear norm,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 925–938, Jan. 2020.

[25] C. Mu, B. Huang, J. Wright, and D. Goldfarb, “Square deal: Lower bounds and improved relaxations for tensor recovery,” in Proc. Int. Conf. Mach. Learn., 2014, pp. 73–81.

[26] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, “Fast and accurate matrix completion via truncated nuclear norm regularization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 9, pp. 2117–2130, Sep. 2013.

[27] D. Zhang, Y. Hu, J. Ye, X. Li, and X. He, “Matrix completion by truncated nuclear norm regularization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 2192–2199.

[28] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” SIAM Rev., vol. 51, no. 3, pp. 455–500, Sep. 2009.

[29] W. He, H. Zhang, L. Zhang, and H. Shen, “Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 1, pp. 178–188, Jan. 2016.

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Found. Trends Mach. Learn., vol. 3, no. 1, pp. 1–122, 2011.

[31] J. C. Herrera and A. Milton, “Interpolation using Hankel tensor completion,” in Proc. Adv. Neural Inf. Process. Syst., vol. 22, no. 12, pp. 7904–7913, Dec. 2011.

[32] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low rank representation,” in Proc. Adv. Neural Inf. Process. Syst., 2011, pp. 612–620.

[33] Z. Zhang and S. Aeron, “Exact tensor completion using t-SVD,” IEEE Trans. Signal Process., vol. 65, no. 6, pp. 1511–1520, Mar. 2017.

[34] A. Wang, C. Li, Z. Jin, and Q. Zhao, “Robust tensor decomposition via orientation invariant tubal nuclear norms,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 4, 2020, pp. 6102–6109.
Xudong Wang received the B.S. degree in automation from Sichuan University, Sichuan, China, in 2014, and the M.S. degree in automation from Beihang University, Beijing, China, in 2017. She is currently pursuing the Ph.D. degree with the Department of Civil Engineering, McGill University, Montreal, QC, Canada. Her research interests include spatio-temporal traffic data mining and anomaly detection.

Yuankai Wu received the Ph.D. degree from the School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China, in 2019. He was a Visiting Ph.D. Student with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA, from November 2016 to November 2017. He is currently a Post-Doctoral Researcher with the Department of Civil Engineering, McGill University, supported by the Institute For Data Valorization (IVADO). His research interests include intelligent transportation systems, intelligent energy management, and machine learning.

Dingyi Zhuang received the B.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently pursuing the M.Eng. degree in transportation engineering. He was a Visiting Research Assistant with the Department of Civil and Environmental Engineering, National University of Singapore. His research interests include intelligent transportation systems, machine learning, and travel behavior modeling.

Lijun Sun (Senior Member, IEEE) received the B.S. degree in civil engineering from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in civil engineering (transportation) from the National University of Singapore in 2015. He is currently an Assistant Professor with the Department of Civil Engineering, McGill University, Montreal, QC, Canada. His research interests include intelligent transportation systems, machine learning, spatio-temporal modeling, travel behavior, and agent-based simulation.