Diagnosing Spatiotemporal Traffic Anomalies With Low-Rank Tensor Autoregression

Xudong Wang, *Student Member, IEEE*, and Lijun Sun

***Abstract*— Traffic data collected from sensor networks often exhibit strong spatial correlations and recurrent temporal patterns. Learning these patterns and diagnosing anomalies in such spatiotemporal traffic data is critical to improving transportation systems and services. This paper proposes a dynamic framework to model spatiotemporal traffic data, with a particular application on diagnosing anomalies. Within the framework, we focus on characterizing the variation in system dynamics with a time-varying vector autoregressive model. We impose a low-rank tensor structure to model the collection of time-varying system matrices. As the temporal factor matrix captures the principal patterns/signatures across all time-varying system matrices, it is a useful tool to diagnose abnormal generative mechanisms and unexpected temporal patterns. We demonstrate the proposed tensor learning framework's effectiveness by experimenting with a synthetic data set and real-world spatiotemporal traffic speed data set. The results show the superiority of the proposed model in uncovering anomalous traffic network dynamics.**

***Index Terms*— Spatiotemporal traffic data, anomaly detection, tensor learning.**

I. Introduction

WITH recent advances in information and communication technologies (ICT) and sensor networks, large-scale spatiotemporal traffic data sets—such as time series of traffic speed and volume collected from different detectors in a network—are becoming ubiquitous. These data sets encode essential information about traffic conditions and often show strong spatiotemporal structures and dependencies on urban traffic's inherent patterns. For example, traffic speed data gathered from multiple loop detectors on an arterial/highway can reflect the real-time traffic states (congestion or not) and exhibit salient recurrent day-to-day temporal patterns and spatial correlations between adjacent upstream and downstream observations [1]. How to extract valuable information from the massive spatiotemporal data set remains a challenging problem.

One fundamental task of modeling spatiotemporal traffic data is anomaly detection [2]. The purpose of anomaly detection is to identify and diagnose abrupt changes in transportation systems caused by incidents and system failures and provide proper and reliable decision-making accordingly [1]. In general, there are two significant ways to detect anomaly: offline anomaly detection and online anomaly detection [3]. Essentially, offline fashion analyzes full historical data to learn and diagnose abrupt changes and non-recurrent patterns in traffic conditions/dynamics. The results can be used to identify systematic issues in network design and operation. On the other hand, an online framework focuses on detecting anomalies from real-time traffic state time-series data, providing early-warning before large operational issues happen. In this paper, we focus on the offline diagnosing framework with the entire data set in Section V and discuss the expanded online manner of the proposed method in Section VI.

As pointed out in [4], the key in detecting anomaly is to build an appropriate time series model to characterize the inherent spatiotemporal patterns, dependencies, and the generative mechanism of urban traffic data. There are two challenges in practice. The first challenge is finding a way to set a global anomaly definition and measure anomaly's effect from the complex spatiotemporal dependencies. In other words, the definition of anomaly may vary depending on the time in a day, day of the week, or locations. The situation would be more complicated with hundreds/thousands of sensors. So it is elusive to detect anomaly directly from original data. The second is the scalability issue since traffic data is large-scale multivariate time-series. Although one can build univariate modeling frameworks (i.e., analyzing the time series from each sensor individually), this approach essentially overlooks the shared spatiotemporal traffic information. It becomes inefficient and even unreliable anomaly detection, given the coupled relations and strong dependencies among different sensors.

To tackle these two challenges, we develop an anomaly detection framework—low-rank dynamic tensor learning (LRDTL)—focusing on identifying anomalous dynamics in spatiotemporal traffic data in this paper. The dynamics reflect the transportation system's inherent generative mechanisms, which depicts a smoother variation in temporal space. Namely, the transportation system dynamics are relatively more stable in continuous time compared with original data. Therefore, the variation of dynamics becomes more prominent, and it is easier to locate the anomaly. A low-rank framework can significantly reduce the number of parameters in the model. Also, the model can include the spatiotemporal correlations of traffic data by adding regularizers.

Manuscript received March 12, 2020; revised September 3, 2020 and November 3, 2020; accepted December 3, 2020. Date of publication January 1, 2021; date of current version November 29, 2021. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Foundation for Innovation (CFI). The work of Xudong Wang was supported by the Fonds de Recherche du Quebec–Nature et Technologies (FRQNT) for providing scholarship. The Associate Editor for this article was Y. Lv. *(Corresponding author: Lijun Sun.)*

The authors are with the Department of Civil Engineering, McGill University, Montreal, QC H3A 0C3, Canada (e-mail: lijun.sun@mcgill.ca).

Digital Object Identifier 10.1109/TITS.2020.3044466

The main idea of the proposed LRDTL is to model the multivariate traffic state time-series data as a time-varying vector autoregressive model: $\boldsymbol{x}_{t+1} \approx A_t \boldsymbol{x}_t$ $(t = 1, \ldots, T-1)$, where $\boldsymbol{x}_t \in \mathbb{R}^N$ is the observation at time t from N traffic sensors. We assume that the system matrices A_t $(N \times N)$ exhibit strong consistency in normal conditions over time, and thus we define **anomalous dynamics** as abrupt changes in A_t comparing with that in adjacent time instances. In order to model the consistency of system dynamics, we follow [5] to organize the time-varying system matrices as a third-order (i.e., $N \times N \times (T-1)$) tensor $\mathcal{A}$ and impose a low-rank assumption on $\mathcal{A}$. In other words, the system coefficient matrices have certain base patterns/signatures, from which anomalous dynamics can be detected. Specifically, we use a low-rank CANDECOMP/PARAFAC (CP) structure to characterize the base patterns and formulate our framework as a tensor learning problem [5], [6]. We insert an additional temporal smoothing regularizer to avoid overfitting and eliminate minor anomalies caused by the strong noise in traffic data. Because there is no anomaly label of traffic data, we conduct an experimental analysis on synthetic data set at first to verify the effectiveness of this framework. Then the framework is applied to a 4-week traffic speed data set collected from the highways of Seattle, US.

The remainder of this paper is organized as follows. In Section II, we review some relevant literature on anomaly detection in multivariate time series data, especially factorization-based models. The temporal matrix factorization method is introduced in Section III. In Section IV, we introduce the LRDTL model and present an alternative projection method for model parameters inference. In Section V, we present two case studies to evaluate the proposed method. Section VI concludes this study and discusses some directions for future research.

II. Related Work

The commonly used models to detect anomaly and change-points are time-varying autoregressive (TVAR) [7], [8] and switching Kalman filtering/smoothing (SKF/SKS) [9]–[11], which are both built on traditional dynamic linear models (DLMs). However, as the number of sensors increases, scalability becomes a critical issue in traditional DLMs, a critical challenge for large systems.

One efficient method to address the large-scale problem is dimensionality reduction. For example, low-rank models such as principle component analysis (PCA) and matrix/tensor factorization (MF/TF) have been applied to model large-scale spatiotemporal data these years (see, e.g., [1], [12]–[16]). Essentially, these factorization-based models project the original spatiotemporal data into a low-dimensional latent space, in which the expected data with better spatial and temporal consistency can be recovered. The “denoised” factors of subspace can be used to define certain anomaly score functions (e.g., quantifying the deviation in the latent space over time). Based on the anomaly score, the anomalous data generated by some events, like an accident or bad weather, can be identified by examining the latent factors. For example, Yang *et al.* [1] proposed a Bayesian PCA model to capture both normal traffic patterns and anomalies in an integrated framework. Tonnelier *et al.* [17] applied matrix factorization to reveal meaningful latent passenger demand patterns shared across train stations and defined anomalous score on temporal demand data by these patterns. As an extension of traditional matrix-based time series model, Xu *et al.* [18] proposed a sliding-window tensor factorization scheme to detect anomalies. In the work of [19], the authors combined traffic flow tensor and topology tensor to address the problem of event anomaly detection. ACS-Tucker decomposes the hybrid model to factor matrices, on which using statistical tests to detect anomalies. Wang *et al.* [20] expanded the traditional Tucker decomposition in a probabilistic manner to detect abnormal activity behaviors.

The factorization-based models provide us with a powerful data-driven tool to identify abnormal traffic observations (anomalous data) over time. However, as the traditional factorization model is invariant to the permutation of timestamps [14], it ignores the strong temporal dynamics/dependencies in urban traffic data. As a result, this model may overfit the noise in the data, which undermines our ability to detect meaningful anomalies. To address this issue, additional temporal regularizers have been introduced into time-series factorization models. For example, the most commonly used temporal regularizer assumes the adjacent time slots are similar by adding a Toeplitz matrix, in which the central diagonal given by ones, and the first upper diagonal given by negative ones [21]. Chen *et al.* proposed a temporal regularizer by considering the past several temporal points with the forgetting factor [22]. The work of [14] and [23] empowered factorization-based models in terms of prediction capability by integrating vector autoregressive (VAR) to characterize generative temporal dynamics. The temporal regularizers are very effective in eliminating insignificant changes caused by the intense noise in traffic data. In addition to the overfitting problem, there remains another issue: these factorization-based models can only characterize abrupt changes in the observations, while they cannot capture the abrupt changes in temporal dynamics/dependencies. To this end, the work of [6], [24] and [5] recently proposed time-varying autoregressive models for multivariate time series. In short, these models first increase their capacity in modeling temporal dynamics with a time-varying VAR framework and then efficiently learn the system tensor (the collection of time-varying coefficient matrices) with a parsimonious approach such as tensor regression.

Inspired by these works, we develop a new anomaly detection framework for urban transportation data in this paper. We follow the work of [5] in their approach to applying the time-varying VAR model. However, instead of assuming the system state is unchanged in a L-length time window, we focus on each time stamp’s system matrix to detect dynamic system anomaly more precisely. In short, the goal of this framework is to detect anomalies in *system dynamics* instead of that in *observations* based on a tensor learning framework.

III. Temporal Matrix Factorization Model

We denote by $X \in \mathbb{R}^{N \times T}$ the multivariate time series (i.e., spatiotemporal traffic data) collected from N sensors over T timestamps and $\boldsymbol{x}_t \in \mathbb{R}^N$ the column vector of X at time step t. The temporal matrix factorization model is a powerful tool to detect anomalies by comparing the distance between observations and estimations. An anomaly occurs when the observation is far away from its expected value produced by the factorization model. Essentially, the model estimates two latent factor matrices $U \in \mathbb{R}^{N \times R}$ and $W \in \mathbb{R}^{T \times R}$ that can recover the original data with $\hat{X} = UW^\top$, where R is the rank of the model. Matrix factorization aims to minimize the difference between observations X and the estimations $UW^\top$. In general, to avoid overfitting and capture temporal relations/similarities [25], the regularizer of U and W is added to the objective function described in the following:

$$\underset{U,W}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| X - UW^\top \right\|_F^2 + \mathcal{R}_1(U, W) + \mathcal{R}_2(W) \right\}, \quad (1)$$

where $\mathcal{R}_1(U, W) = \frac{1}{2\eta} \left(\|U\|_F^2 + \|W\|_F^2 \right)$ with regularization parameter η. Temporal smoothing is achieved by adding a smoothing regularization term, for example:

$$\mathcal{R}_2(W) = \frac{\beta}{2} \sum_{r=1}^{R} \sum_{t=2}^{T-1} \left(w_{t-1,r} - 2w_{t,r} + w_{t+1,r} \right)^2, \quad (2)$$

where β is the temporal regularization parameter.

However, even with temporal regularization, matrix factorization still cannot fully capture anomalies in the generative mechanism or temporal dynamics of the data because the goal of matrix factorization is to reproduce the observations instead of capturing the temporal dynamics. The following subsection presents a dynamic tensor learning model for anomaly detection to fill this gap.

IV. Low-Rank Dynamic Tensor Learning Model

A. Model Framework

To model the generative temporal dynamics, we assume the data follows a time-varying VAR(p) with order $p = 1$ (i.e., first-order):

$$\boldsymbol{x}_{t+1} = A_t \boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \ \text{for } t = 1, \ldots, T-1, \quad (3)$$

where $A_t \in \mathbb{R}^{N \times N}$ is a time-varying coefficient matrix capturing the temporal dynamics/dependencies in spatiotemporal data and $\boldsymbol{\epsilon}_t$ is a zero-mean Gaussian noise vector. As A_t is time-varying and we expect adjacent time instances to have little difference for a steady system, so the problem of estimating A can be formulated as the following optimization problem with system matrix smoothing:

$$\underset{A}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{t=1}^{T-1} \|\boldsymbol{x}_{t+1} - A_t \boldsymbol{x}_t\|_F^2 + \mathcal{R}(A) \right\}, \quad (4)$$

where $\mathcal{R}(A)$ is a regularizer defined as $(A_{t-1} - A_t) - (A_t - A_{t+1})$ constraining the adjacent time.

To solve problem (4), the collection of system coefficient matrices along the whole period can be organized as a system tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times T-1}$ with the tth frontal slice $\mathcal{A}_{::t} = A_t$. However, estimating A in (4) is practically infeasible, given a large number of parameters $N^2(T-1)$. Therefore, we need to seek alternative approaches to parameterize the model. Recent work [5] and [6] consider imposing a low-rank structure in modeling the system tensor $\mathcal{A}$, which have shown promising results in capturing complex dependencies and dynamics among a set of time series. Following this method, we assume the system tensor also exhibits strong spatiotemporal patterns/signatures following a low-rank structure for spatiotemporal traffic data. In this paper, we use a CP structure to model $\mathcal{A}$:

$$\mathcal{A} = \sum_{r=1}^{R} \boldsymbol{u}_r \circ \boldsymbol{v}_r \circ \boldsymbol{w}_r, \quad (5)$$

where $\boldsymbol{u}_r$, $\boldsymbol{v}_r$, and $\boldsymbol{w}_r$ are the rth column of factor matrices $U \in \mathbb{R}^{N \times R}$, $V \in \mathbb{R}^{N \times R}$, and $W \in \mathbb{R}^{(T-1) \times R}$, respectively, the symbol $\circ$ represents the outer product.

The CP assumption reduces the number of parameters significantly from $N^2(T-1)$ to $R(2N + T - 1)$ due to $R \ll N$ and T, thus providing us with a parsimonious solution. Figure 1 gives a graphical illustration of the tensor learning model for multivariate time series data. Specifically, U and V are two spatial factor matrices, which determine the loadings of A_t onto the spatial dimensions in the data, and W is the temporal factor matrix that characterizes the time-varying patterns in system dynamics. Therefore, we have a new optimization problem with factor matrices U, V and W as variables:

$$\underset{U,V,W}{\operatorname{argmin}} \ \frac{1}{2} \sum_{t=1}^{T-1} \|\boldsymbol{x}_{t+1} - A_t \boldsymbol{x}_t\|_F^2 + \mathcal{R}_1(U, V, W) + \mathcal{R}_2(W). \quad (6)$$

Similar to the temporal matrix factorization model in (1), here we add a standard Frobenius norm regularizer $\mathcal{R}_1(U, V, W) = \frac{1}{2\eta} \left(\|U\|_F^2 + \|V\|_F^2 + \|W\|_F^2 \right)$ and also an identical temporal regularizer $\mathcal{R}_2(W)$ as in Equation (2) to avoid false alarms and increase the generalization power of the model. Parameters η and β control the importance of $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. One critical challenge in estimating this model is to tune the two regularization parameters. A common approach to obtain the appropriate regularization parameters is to apply grid search and perform cross-validation. We will cover this in the following section.

B. Model Inference

The optimization problem in Equation (6) is multi-convex, which can be solved using an alternating projection method that takes each block as convex. In other words, we update the parameters in one block at a time and keep the rest parameters fixed. To simplify the model, we first decompose the frontal slices of $\mathcal{A}$ at the outset as $A_t = UD_tV^\top$, where $D_t = \operatorname{diag}(W_{t:})$. With this decomposition, the overall loss

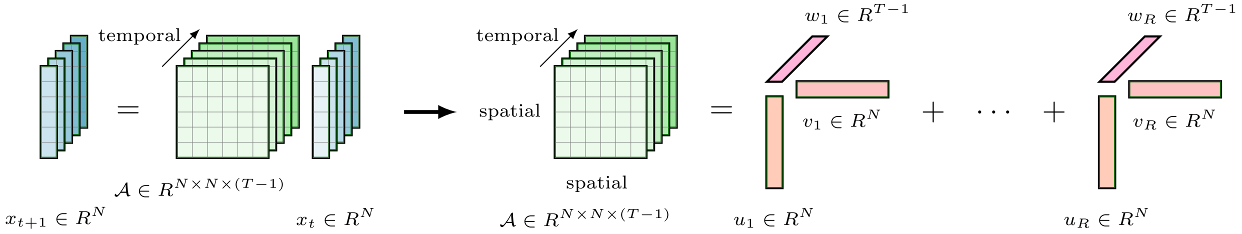

Fig. 1. A graphical illustration for the LRDTL model for spatiotemporal traffic data (blue: observations $\boldsymbol{x}_t$, green: system tensor $\mathcal{A}$, red: latent factor vectors).

function can be written as:

$$\begin{aligned} &L\left(U, V, W\right) \\ &= \frac{1}{2}\sum_{t=1}^{T-1}\left\|\boldsymbol{x}_{t+1} - UD_tV^\top\boldsymbol{x}_t\right\|_F^2 \\ &\quad + \frac{1}{2\eta}\left(\|U\|_F^2 + \|V\|_F^2 + \|W\|_F^2\right) + \frac{\beta}{2}\|QW\|_F^2, \end{aligned} \quad (7)$$

where Q is the coefficient matrix of Equation (2).

Harris *et al.* [5] has provided the full derivations of updating rules for solving $\nabla_U L = 0$, $\nabla_V L = 0$, and $\nabla_W L = 0$ in similar tensor regression model. In the following, we provide the gradient of the loss function with respect to each variable. Algorithm 1 summarizes the alternating projection method to get optimal U, V, and W.

1) Gradient w.r.t. the Left Spatial Modes U:

$$\begin{aligned} \nabla_U L = U&\left(\sum_{t=1}^{T-1} D_t V^\top \boldsymbol{x}_t\boldsymbol{x}_t^\top V D_t + \frac{1}{\eta}I_R\right) \\ &- \left(\sum_{t=1}^{T-1}\boldsymbol{x}_{t+1}\boldsymbol{x}_t^\top V D_t\right), \end{aligned} \quad (8)$$

where I_R is a $R \times R$ unit matrix. Set $\nabla_U L = 0$, we can obtain the update function of U solving by solving a linear system:

$$U = \left(\sum_{t=1}^{T-1}\boldsymbol{x}_{t+1}\boldsymbol{x}_t^\top V D_t\right)\left(\sum_{t=1}^{T-1} D_t V^\top \boldsymbol{x}_t\boldsymbol{x}_t^\top V D_t + \frac{1}{\eta}I_R\right)^{-1}. \quad (9)$$

2) Gradient w.r.t. the Right Spatial Modes V:

$$\begin{aligned} \nabla_V L &= \left(\sum_{t=1}^{T-1}\boldsymbol{x}_t\boldsymbol{x}_t^\top V D_t U^\top U D_t\right) \\ &\quad - \left(\sum_{t=1}^{T-1}\boldsymbol{x}_t\boldsymbol{x}_{t+1}^\top U D_t\right) + \frac{1}{\eta}V \\ &\equiv \sum_{t=1}^{T-1} L_t V R_t + \frac{1}{\eta}V - H. \end{aligned} \quad (10)$$

Let $\nabla_V L = 0$, we get a Sylvester equation [26] for V:

$$\sum_{t=1}^{T-1} L_t V R_t = H - \frac{1}{\eta}V. \quad (11)$$

For small N and R, we can use the Kronecker product and vectorization operation to solve Equation (11); however, it is impractical for large N. One efficient method for solving the Sylvester equation is conjugate gradients (CG) [27], and we apply a modified CG named preconditioned conjugate gradients (PCG) to speed up the convergence.

3) Gradient w.r.t. the Temporal Modes W: The partial derivatives of L with respect to W can be taken for each diagonal matrix D_t:

$$\begin{aligned} \nabla_{D_t} L &= V^\top \boldsymbol{x}_t\boldsymbol{x}_t^\top V D_t U^\top U + \frac{1}{\eta}D_t \\ &\quad + \beta Q^\top Q D_t - V^\top \boldsymbol{x}_t\boldsymbol{x}_{t+1}^\top U \\ &\equiv \left(L'{}_t D_t R'{}_t + \frac{1}{\eta}D_t + \beta Q^\top Q D_t - J\right) * I, \end{aligned} \quad (12)$$

where $*$ is the Hadamard product to constraint $\nabla_{D_t} L$ is diagonal. The Theorem 2.5 in [28] is applied to simplified the Equation (12) and we also applied PCG algorithm to obtain the solution of temporal pattern W:

$$\begin{aligned} \mathrm{vec}(\nabla_W L) = \mathrm{vec}&\left(\sum_{t=1}^{T-1}\left(L'_t * R'_t + \frac{1}{\eta}I_R\right)W_{t:}^\top + \beta Q^\top Q W\right) \\ &- \mathrm{vec}(J). \end{aligned} \quad (13)$$

C. Implementation Details

1) Determine the Rank of CP Decomposition: For all the low-rank models, the rank of the model is a pre-determined parameter [5]. A large rank will better fit the data, but in the meanwhile, it involves more parameters that may increase the risk of overfitting and time-consuming. On the other hand, a small rank may be insufficient in capturing complex interdependence in the data [29]. Here, we apply a sensitivity-driven rank selection considering the ratio threshold of singular values to determine the size of rank [30]. Specifically, the singular values $\lambda_{1:N}$ of system matrices are obtained from singular value decomposition (SVD) and the truncation ratio threshold is defined as cumulative eigenvalue percentage (CEP) $\sum_{i=1}^{R}\lambda_i / \sum_{i=1}^{N}\lambda_i$. By doing so, the low-rank model can capture enough information from data with a proper rank.

2) Determine the Regularization Parameters: The two regularization parameters η and β play an essential role in the model, enhancing the model's generalization power and avoiding overfitting. To tune these two parameters, we perform a grid search on η and β and apply K-fold cross-validation to find the best combination of regularization parameters. This paper defines K based on the input's periodic nature to keep the spatiotemporal correlation. For example, if we have a data set of K weeks, we may consider each week as a group. The root-mean-square error (RMSE) is used for model evaluation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N(T-1)} \sum_{t=1}^{T-1} \|\boldsymbol{x}_{t+1} - A_t \boldsymbol{x}_t\|_F^2}. \tag{14}$$

In training the model, we first estimate the matrix $A = X_{2:T} X_{1:T-1}^{\dagger}$ which can be decomposed by SVD as $A = U_0 \Sigma V_0^{\top}$. Because $R \ll N$, we truncate the $U_0 \in \mathbb{R}^{N \times R}$ and $V_0 \in \mathbb{R}^{N \times R}$ to be the initial of left spatial matrix and right spatial matrix, respectively. The initial temporal matrix W_0 is set to a matrix of all ones. In the testing model, the optimal $U^{\star}$ and $V^{\star}$ obtained from the training model are used as initial U_0 and V_0. The initial W_0 is calculated by $W_0 = \mathcal{A}_{(3)}(U_0 \odot V_0)(U_0^{\top} U_0 * V_0^{\top} V_0)^{\dagger}$ [31].

The absolute tolerance γ_a and relative tolerance γ_r of the decrease in the loss function (7) are both applied in Algorithm 1 as a stopping criteria. We set $\gamma_a = 10^{-6}$, $\gamma_r = 10^{-4}$, and the Algorithm 1 stops when reaching either criterion.

Algorithm 1 Alternating Minimization Algorithm

```
Input: Raw data $X$, tensor rank $R$, regularization parameters
    $\eta$ and $\beta$, relative tolerance $\gamma_r$ and absolute tolerance $\gamma_a$.
Output: Optimal factor matrices $U^\star$, $V^\star$ and $W^\star$.
Initialize $U$, $V$ and $W$.
repeat
    $U^{(i+1)} \leftarrow \arg\min_U L\left(U^{(i)}, V^{(i)}, W^{(i)}\right)$ in Equation (9)
    (by solving a linear system);
    $V^{(i+1)} \leftarrow \arg\min_V L\left(U^{(i+1)}, V^{(i)}, W^{(i)}\right)$ in Equation (11) (by preconditioned conjugate gradient);
    $W^{(i+1)} \leftarrow \arg\min_W L\left(U^{(i+1)}, V^{(i+1)}, W^{(i)}\right)$ in Equation (13) (by preconditioned conjugate gradient);
    rmse$\leftarrow$ RMSE in Equation (14).
until convergence
return $U^\star$, $V^\star$, and $W^\star$.
```

The general procedure of K-fold cross-validation to search optimal regularization parameters is summarized in Algorithm 2.

3) Anomaly Detection Indicator: After obtaining the optimal regularizer parameters $\eta^\star$ and $\beta^\star$, we apply the Algorithm 1 on the whole data set to obtain the final latent factor matrices and analyze abnormal patterns on the temporal latent factor W. As mentioned before, the system matrix A_t should keep relatively steady over time, so a dynamic anomaly occurs if there is an abrupt change in W when fixing spatial latent factors. Therefore, we define anomaly score S_t as the absolute value of the sum of the first difference of $W_{t:}$ across all the latent modes:

$$S_t = \sum_{r=1}^{R} \left| W_{t+1,r} - W_{t,r} \right|, \quad t \in [1, T-1] \tag{15}$$

V. Case Study

In this section, we apply the tensor learning model on synthetic vector autoregressive data set (case 1) at first to measure the efficiency of the proposed unsupervised method and then apply on the traffic speed data set collected in Seattle, US (case 2). As a comparison, we also show the anomaly detection results obtained from the switching Kalman filter (SKF) and switch Kalman smoother (SKS) and temporal matrix factorization (MF) model.

Algorithm 2 K-Fold Cross-Validation for Searching Optimal Regularization Parameters

```
Input: Raw data $X$, tensor rank $R$, stopping criteria $\gamma_a$ and
    $\gamma_r$, regularization parameters set $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, and number of
    data groups (weeks) $K$.
Output: Optimal regularization parameters $\eta^\star$ and $\beta^\star$.
Split the data set into $K$ groups.
for each $\eta$ and $\beta$ do
    for each $k = 1, \ldots, K$ do
        $\mathcal{D}_{\text{Test}}$ = the $k$th week's data of $X$;
        $\mathcal{D}_{\text{Train}}$ = remaining data of $X$;
        Apply Algorithm 1 on $\mathcal{D}_{\text{Train}}$ to get $U^\star$, $V^\star$ and $W^\star$;
        repeat
            $W \leftarrow \arg\min_W L(U^\star, V^\star, W)$ on $\mathcal{D}_{\text{Test}}$
        until convergence.
        $\epsilon_k \leftarrow$ RMSE in Equation (14).
    end for
    $E_{\eta,\beta} \leftarrow \sum_{k=1}^{K} \epsilon_k / K$.
end for
$\eta^\star, \beta^\star \leftarrow \text{argmin}_{\eta,\beta}\{E_{\eta,\beta}\}$;
return $\eta^\star, \beta^\star$.
```

A. Case Study 1: Synthetic Data Set

To examine the proposed LRDTL framework's effectiveness, we first consider an artificial case using synthetically generated data. The data set is generated from a time-varying vector autoregressive model with a Gaussian noise included, as described in Equation (3). Specifically, we consider two system matrices, and each element of them are randomly drawn from two uniform distribution as $A_1 : a_{ij} \sim \mathrm{U}[-0.4, 0.4]$ and $A_2 : a_{ij} \sim \mathrm{U}[-0.8, 0.8]$, for $i, j \in [1, N]$ [10]. A Gaussian noise $\mathcal{N}(0, 1)$ is added to each element of the system matrix. We set $N = 10$, $T = 400$ to generate data by concatenating two VAR(1) processes dominated by A_1 and A_2 alternatively. The generated data includes four time-blocks with a fixed length $T' = 100$ of each and three change points at $t = 101$, $t = 201$ and $t = 301$ which can be regarded as anomalies. The synthetic data with the two varying states are depicted in different background colors is illustrated in Figure 2 (a).

We compared the LRDTL model with the traditional temporal MF model described in (1) and state-switching models, including SKF and SKS [10]. SKF and SKS turn the nonlinear system dynamics to linear dynamics as discrete modes by specifying a process model for each mode and then estimating the probability of switching from one mode to another at each time points [32], which are widely used in detecting change points. There are two designed system matrices in generating the synthetic data; therefore, the number of states is 2 in both SKF and SKS model.

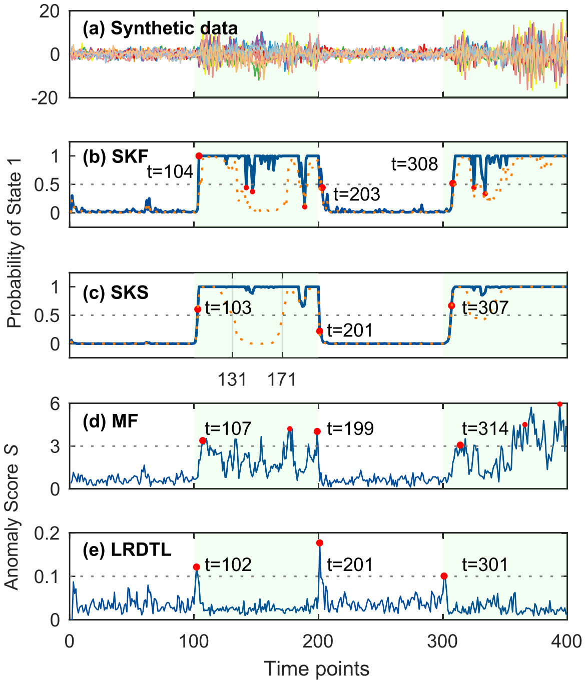

Fig. 2. The results of the synthetic data set. (a): Synthetic data generating from VAR(1) process by two system matrices; (b): Probability of state 1 by SKF (solid line for 2 states, dot line for 3 states); (b): Probability of state 1 by SKS (solid line for 2 states, dot line for 3 states); (c): Anomaly score by MF; (d): Anomaly score by LRDTL.

We set the truncation ratio threshold set as 0.75 to determine rank R, and we get $R = 4$ for both MF and LRDTL. We set the optimal regularization parameters $\eta = 10, \beta = 1$ in MF and $\eta = 1$, $\beta = 1000$ in LRDTL model, which are obtained by the cross-validation in Algorithm 2.

The inferred state probability sequences by SKF and SKS are illustrated in Figure 2 (b) and (c), respectively. It can be seen that SKF shows several state-switching points (anomalies) within a specific state, where SKS performs much smoother. Because SKF only includes forward inference given the available observations up to time t, while SKS considers all observations [10]. In doing so, the uncertainty will be significantly reduced by conditioning on past and future observations [33].

We apply MF model on synthetic data directly and the anomaly score S_t in Equation (15) at time t is shown in Figure 2 (d). We find that the magnitude of the temporal pattern affects the anomaly score significantly. A larger magnitude of observation may easier to get a higher anomaly score, which is difficult to set a proper threshold to recognize anomalies. The anomaly threshold is 3 in the MF model to avoid numerous anomalies. Though the state switching can be found by comparing temporal pattern differences for a period using MF, it has two main drawbacks here: (i) it cannot detect switching points immediately; (ii) it is not a straightforward way to display the switching points due to magnificent impact. However, the LRDTL model can overcome these two shortcomings.

Figure. 2 (e) shows the anomaly score S by the proposed method LRDTL model, and we find the score is relatively low except for three extremely high points at time $t = 102, 201$ and 301. The anomaly threshold is 0.1 in this case. Compared with the other three methods, especially the MF model, all the anomalies can be detected instantly by LRDTL. Though SKS can detect the second anomaly correctly, it exhibits time delay when detecting the other two anomalies. Another limitation of SKF and SKS is the number of states is a pre-determined parameter, and an improper state number might lead to a false alarm. For instance, the dotted line presenting three states of SKF/SKS is also shown in Figure 2 (b) and (c). It can be seen that the probability of state 1 will drop under 0.5 between $t = 131$ and $t = 171$, suggesting that the model detects two incorrect anomalies. It is hard to determine the number of states in practice due to the dynamic correlation with real traffic data. In the following, we will only compare the effectiveness of MF and LRDTL using real-world traffic speed data.

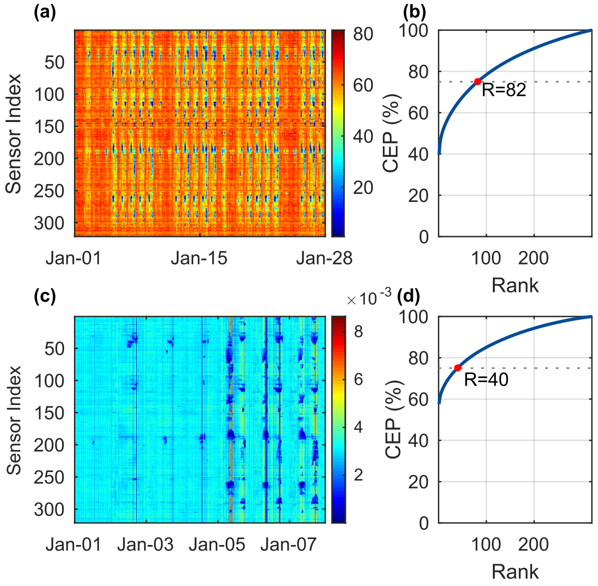

Fig. 3. The traffic speed data set. (a): The average speed of sensors, (b): The cumulative eigenvalue percentage of traffic speed data, (c): The system matrices of sensors (the first week), (d): The cumulative eigenvalue percentage of system matrices (the first week).

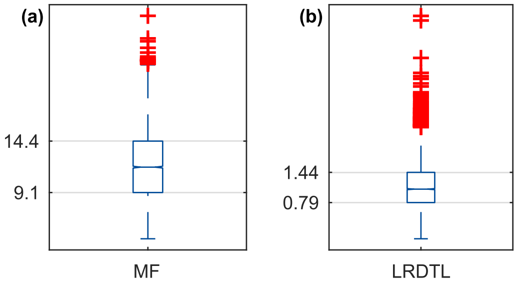

Fig. 4. The boxplot of anomaly score S by (a): MF model and (b): LRDTL model.

B. Case Study 2: Seattle Traffic Speed Data Set

1) Traffic Speed Data: Our numerical experiment is based on a traffic speed data set collected by inductive loop detectors

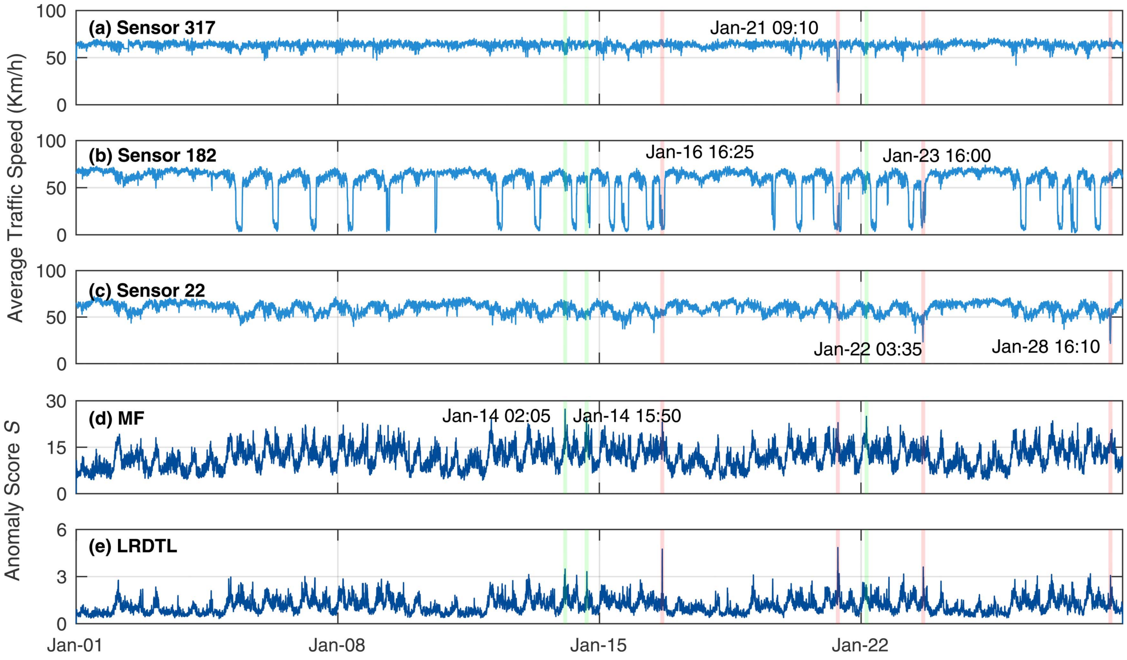

Fig. 5. The raw traffic speed data at three sensors (sensor 317, sensor 182 and sensor 22) and the anomaly score by MF and LRDTL.

located on the highway of Seattle [34]. The data measures from 321 sensor stations (156 North to South or East to West sensor stations and 165 South to North or West to East sensor stations) on four freeways (I-5, I-405, I-90, and SR-520) of Seattle at an interval of 5 minutes (288 time-slots per day) in 2015. In the experiment, we use the data of January ($M = 321$ and $N = 8064$) to demonstrate the effectiveness of the method (4 weeks in total).

The traffic speed data set and its dynamic matrices are shown in Figure 3. Specifically, Figure 3 (a) shows the average speed of sensors over four weeks, and its cumulative eigenvalue percentage (CEP) is shown in Figure 3 (b). The system matrices for the first week and its corresponding CEP are shown in Figure 3 (c) and (d), respectively.

From Figure 3 (a), we can see that the speed data set follows a periodic time-varying pattern with a clear difference from weekdays to weekends. The difference between peak hours and off-peak hours is also significant within a day. On the other hand, the CEP of traffic data obtained from SVD also demonstrates the first few latent factors can well capture the low-rank property of traffic data. We use the first week of system matrices to illustrate its low-rank characteristics. The system matrix A_t is calculated by $\boldsymbol{x}_{t+1}\boldsymbol{x}_t^{\dagger}$ at each time t and concatenated along temporal dimension (Figure 3 (c)). Like the CEP of traffic data, the system matrices can also be estimated by several latent factors shown in Figure 3 (d).

2) Anomaly Detection: We apply the temporal matrix factorization model on the traffic speed data directly with the same k-fold cross-validation process by setting rank $R = 82$ (CEP $=$ 0.75). The temporal matrix factorization model is estimated using a gradient descent algorithm proposed in [35] by introducing a Laplacian regularization term on the temporal factor matrix W, which has the same effect as Equation (2).

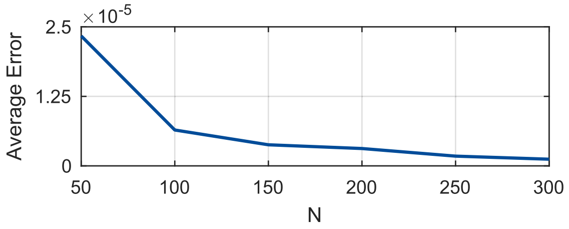

Fig. 6. The average Frobenius norm of system matrices reconstruction error under different sensor number N.

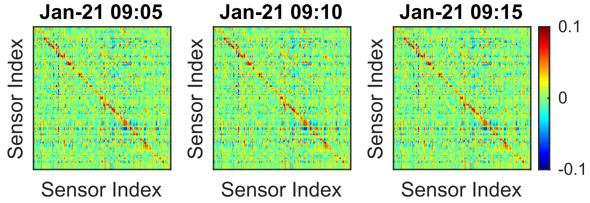

Fig. 7. The estimated system matrices $\hat{A}_t$ from 9:05 - 9:15 on January 21st.

We next apply the proposed LRDTL model on the traffic speed data X and obtain the CP decomposition for the system dynamics tensor $\mathcal{A}$. Note that W characterizes the temporal evolution of system dynamics (i.e., VAR coefficient matrix A_t of $\boldsymbol{x}_{t+1} = A_t\boldsymbol{x}_t$) instead of the observation $\boldsymbol{x}_t$ itself. In other words, A_t reflects the dynamic of the traffic system, and the abrupt changes of A_t is regarded as anomalous dynamics measured by S. Specifically, we apply LRDTL with $R = 40$ (CEP $= 0.75$), $\eta = \{0.001, 0.01, 0.1\}$ and $\beta = \{1, 10, 100\}$ to find the optimal regularization parameters. The best model is obtained with $\eta = 0.01$ and $\beta = 10$, with an average RMSE of 3.7936. The spatiotemporal latent patterns of the one-month traffic data set is obtained by Algorithm 1 based on the optimal parameters.

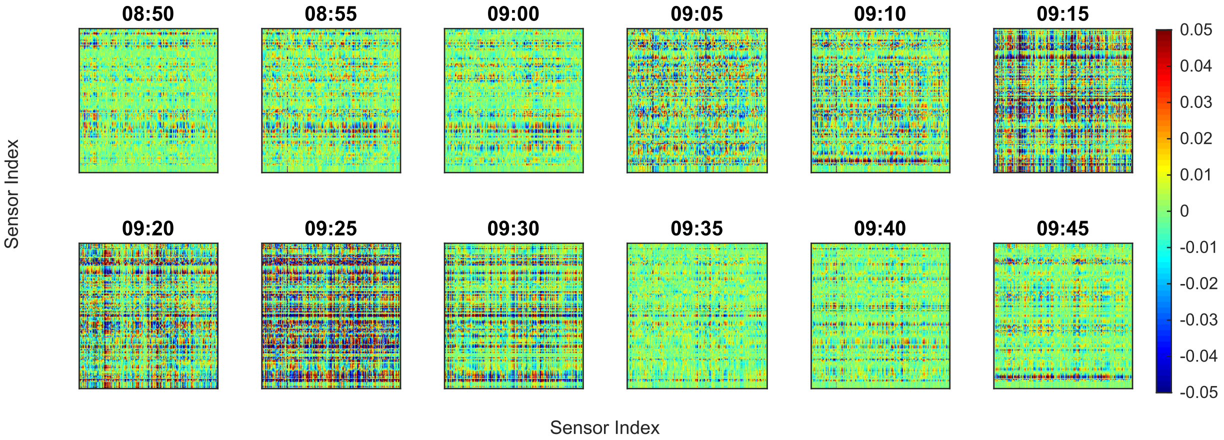

Fig. 8. The difference of the estimated system matrices $\Delta\hat{A}_t = \hat{A}_t - \hat{A}_{t-1}$ between 8:50 a.m. to 9:45 a.m. on January 21st.

Figure 4 (a) and Figure 4 (b) show the boxplots of anomaly scores S obtained from MF and LRDTL, respectively. The original S are also shown in Figure 5 (d) and (e), respectively. The outliers (red symbol) are measured by a distance of 1.5 times interquartile range (IQR). We can see that the LRDTL model has a smaller IQR but more outliers than those in the MF model, which means that the LRDTL model performs much better in distinguishing normal time points and abnormal time points.

Since there is no label in traffic time-series point to illustrate if it is an anomaly or not, we compare the raw speed data and anomaly score to demonstrate the proposed model's efficiency in Figure 5. It can be seen that some sensors in Figure 5 (a) - (c) may not show similar patterns due to the complexity of the spatiotemporal traffic network, causing the difficulty of analyzing the anomaly detection problem directly from the raw data. For example, there are two slow-speed (around 25 km/h) periods from Sensor 22, which can be regarded as anomalies comparing with other time instances. However, there is no such unusual behavior from Sensor 317 at the same time. Slow speed is a period phenomenon in a specific period of a day in Sensor 182. The apparent spike signals can be found in the difference of dynamical temporal modes shown in Figure 5 (e), providing an intuitive way to detect anomalies.

We select three time-points with the highest anomaly score for the MF model (green shaded windows) and the LRDTL model (red shaded windows), respectively. Several obvious spikes that reveal anomalies in the typical traffic speed time-series data in Figure 5 (a) - (c) can be detected by the proposed model according to S. Although the MF model can also detect anomalies, it is difficult to set a threshold to determine if it is an anomaly or not due to the similar S, which can lead to false alarm (also occurs in case study 1).

3) Dynamic Anomaly Analysis: To measure how the performance of the estimated system matrix $\hat{A}$ varies with the number of sensors, we depict the reconstruction error between real system matrices A and $\hat{A}$ as Frobenius norm $||A - \hat{A}||_F^2$ averaged across the first week in Figure 6, where the numbers of sensors are selected as {50,100,150,200,250,300}. Under each number of sensors, the rank threshold is also set as 0.75 to determine the rank. As we can see in Figure 6, the LRDTL model works stable with an increasing number of sensors.

To better understand how LRDTL works in detecting dynamics anomaly, we show a detailed example at 9:10 a.m. on January 21st in Figure 5. The recovered system matrices $\hat{A}_t$ obtained from two spatial latent matrices U and V with the specific temporal $\boldsymbol{w}$ during 9:05 to 9:15 on January 21st are shown in Figure 7. We can see the estimated system matrices show the sub-block and diagonal characteristics, which means that the last observation mainly determines the travel speed on the same sensor and the off-diagonals mainly captures the effects of neighbor sensors.

The difference system matrices $\Delta\hat{A}_t = \hat{A}_t - \hat{A}_{t-1}$ near 9:10 a.m. are shown in Figure 8. From the matrix difference, we can see that the system matrix dramatically changes between 9:10 and 9:15. Specifically, the system matrix keeps relatively stable before the anomaly occurs, which leads to almost all the entries of the system difference matrix being zero before 9:00. When an anomaly occurs, the system matrix changes in one direction and then changes in the opposite direction to recover from the anomaly impact (positive and negative values in the matrix means direction). After the anomaly, the system matrix returns to relative stability again. Besides, we can also find the anomaly effect lasts around 30 minutes from 9:00 to 9:30.

VI. Conclusion and Discussion

In this paper, we propose a new modeling framework to detect anomalous dynamics in multivariate spatiotemporal traffic time series data. Unlike previous factorization-based approaches—such as PCA and temporal matrix factorization—which model the data X directly, we use a time-varying VAR model to characterize the generative mechanism and temporal dependencies (i.e., system tensor $\mathcal{A}$). To efficiently and effectively model the system tensor, we transform the time-varying VAR problem to a low-rank tensor learning problem. As a result, we can characterize the system dynamics using latent factor matrices in a parsimonious way. This model not only uncovers the expected evolving and recurrent dynamics of traffic data but also serves as a new anomaly detection

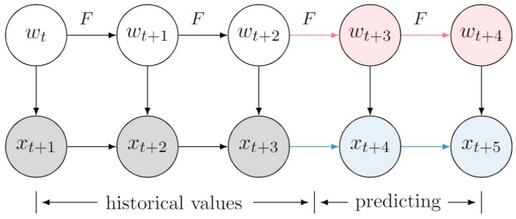

Fig. 9. The structure of online anomaly detection.

framework to identify and diagnose abnormal dynamics in spatiotemporal traffic data, which differs from its expected behavior recovered by the low-rank representation.

We assume the system dynamics tensor of the autoregressive model follows a simple structure with a low CP rank. Thus, the optimization problem can be efficiently solved by the alternating projection method derived in [5], in which different approaches such as linear systems and conjugate descent are used for different variables. We develop two case studies based on a synthetic data set and a large-scale spatiotemporal traffic speed data set to demonstrate the effectiveness of the proposed framework by comparing it with the traditional temporal matrix factorization model with identical regularization terms. Our results show that the tensor learning framework can identify more interesting **anomalous dynamics** that cannot be detected from traditional factorization-based models.

One of the limitations of the proposed method is that the model cannot detect anomalies in real-time. This is because the temporal pattern $\boldsymbol{w}_t$ does not have the forecasting ability under the time-varying constraint $||QW||_F^2$. To overcome the problem, the time-varying constraint can be set as $\boldsymbol{w}_{t+1} = F\boldsymbol{w}_t$ [14] to capture dynamic of temporal pattern, where F is the coefficient matrix of temporal pattern. Therefore, the online anomaly detection structure is illustrated in Figure 9, which is also equivalent to the state-space model. In the historical part, we can first obtain the temporal dynamic $\boldsymbol{w}_t$ from the system matrix A_t based on CP decomposition shown in Figure 1, and then update the time-varying temporal coefficient matrix F by $\boldsymbol{w}_{t+1} = F\boldsymbol{w}_t$. In the prediction part, we can obtain $\boldsymbol{w}_{t+3} = F\boldsymbol{w}_{t+2}$ to detect anomalies in real-time and the system matrix can also be acquired by $A_{t+3} = \sum_{r=1}^{R} \boldsymbol{u}_r \circ \boldsymbol{v}_r \circ \boldsymbol{w}_{r,t+3}$. In the end, the prediction $\boldsymbol{x}_{t+4}$ can be obtained from $\boldsymbol{x}_{t+4} = A_{t+3}\boldsymbol{x}_{t+3}$.

There are several directions for future research. (i) The current model requires careful tuning of regularization parameters, which is computationally expensive. The tuning process is data-specific, and there exist no universal solutions. To address this issue, we are interested in developing the Bayesian counterpart of this model to avoid the tuning of regularization parameters and achieve automatic learning for different applications/data sets. (ii) This proposed model focuses on multivariate time series data. However, we often encounter high-dimensional data for transportation applications. For example, traffic demand can be organized as a matrix-valued time series with two separate spatial dimensions for both origin and destination. It is a challenging question to model and capture the complex system dynamics in such high-dimensional data sets. (iii) The current model uses a third-order system tensor to capture temporal dynamics. However, the third-order representation may not be very efficient in modeling the periodic and recurrent nature of traffic data sets (i.e., short-term and long-term trends [4]). One potential approach is to reorganize the system dynamics tensor (spatial×spatial×timestamp) with a four-order (spatial×spatial×timestamp×day). For the 28-day traffic speed data in our case study, the fourth-order representation can further reduce the number parameters from $(321 + 321 + 288 \times 28) \times R$ to $(321 + 321 + 288 + 28) \times R$ and this will allow us to use a larger rank R to capture temporal dynamics better.

References

[1] S. Yang, K. Kalpakis, and A. Biem, "Detecting road traffic events by coupling multiple timeseries with a nonparametric Bayesian method," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 1936–1946, Oct. 2014.

[2] Z. Zhang, Q. He, H. Tong, J. Gou, and X. Li, "Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network," *Transp. Res. Part C, Emerg. Technol.*, vol. 71, pp. 284–302, Oct. 2016.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.

[4] L. Li, X. Su, Y. Zhang, Y. Lin, and Z. Li, "Trend modeling for traffic time series analysis: An integrated study," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3430–3439, Dec. 2015.

[5] K. D. Harris, A. Aravkin, R. Rao, and B. Wen Brunton, "Time-varying autoregression with low rank tensors," 2019, *arXiv:1905.08389*. [Online]. Available: http://arxiv.org/abs/1905.08389

[6] M. T. Bahadori, Q. R. Yu, and Y. Liu, "Fast multivariate spatio-temporal analysis via low rank tensor learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3491–3499.

[7] L. F. Bringmann, E. L. Hamaker, D. E. Vigo, A. Aubert, D. Borsboom, and F. Tuerlinckx, "Changing dynamics: Time-varying autoregressive models using generalized additive modeling," *Psychol. Methods*, vol. 22, no. 3, p. 409, 2017.

[8] L. F. Bringmann, E. Ferrer, E. L. Hamaker, D. Borsboom, and F. Tuerlinckx, "Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model," *Multivariate Behav. Res.*, vol. 53, no. 3, pp. 293–314, May 2018.

[9] K. P. Murphy, "Switching Kalman filters," Dept. Comput. Sci., UC Berkeley, Berkeley, CA, USA, Tech. Rep., 1998.

[10] C.-M. Ting, H. Ombao, S. B. Samdin, and S.-H. Salleh, "Estimating time-varying effective connectivity in high-dimensional fMRI data using regime-switching factor models," 2017, *arXiv:1701.06754*. [Online]. Available: http://arxiv.org/abs/1701.06754

[11] L. H. Nguyen and J.-A. Goulet, "Anomaly detection with the switching Kalman filter for structural health monitoring," *Struct. Control Health Monitor.*, vol. 25, no. 4, p. e2136, Apr. 2018.

[12] Y. Han and F. Moutarde, "Statistical traffic state analysis in large-scale transportation networks using locality-preserving non-negative matrix factorisation," *IET Intell. Transp. Syst.*, vol. 7, no. 3, pp. 283–295, Sep. 2013.

[13] N. Takeishi and T. Yairi, "Anomaly detection from multivariate time-series with sparse representation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2014, pp. 2651–2656.

[14] H.-F. Yu, N. Rao, and I. S. Dhillon, "Temporal regularized matrix factorization for high-dimensional time series prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 847–855.

[15] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet, "Matrix and tensor based methods for missing data estimation in large traffic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1816–1825, Jul. 2016.

[16] X. Chen, Z. He, and L. Sun, "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation," *Transp. Res. Part C, Emerg. Technol.*, vol. 98, pp. 73–84, Jan. 2019.

[17] E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari, "Anomaly detection in smart card logs and distant evaluation with Twitter: A robust framework," *Neurocomputing*, vol. 298, pp. 109–121, Jul. 2018.

[18] M. Xu, J. Wu, H. Wang, and M. Cao, "Anomaly detection in road networks using sliding-window tensor factorization," 2018, *arXiv:1803.04534*. [Online]. Available: http://arxiv.org/abs/1803.04534

[19] H. Fanaee-T and J. Gama, "Event detection from traffic tensors: A hybrid model," *Neurocomputing*, vol. 203, pp. 22–33, Aug. 2016.

[20] X. Wang, A. Fagette, P. Sartelet, and L. Sun, "A probabilistic tensor factorization approach to detect anomalies in spatiotemporal traffic activities," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 1658–1663.

[21] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and Internet traffic matrices (extended version)," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 662–676, Jun. 2012.

[22] Z. Chen, A. Cichocki, and T. M. Rutkowski, "Constrained non-negative matrix factorization method for EEG analysis in early detection of alzheimer disease," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 5, Dec. 2006, p. 5.

[23] D. Deng, C. Shahabi, U. Demiryurek, L. Zhu, R. Yu, and Y. Liu, "Latent space model for road networks to predict time-varying traffic," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1525–1534.

[24] R. Yu, G. Li, and Y. Liu, "Tensor regression meets Gaussian processes," 2017, *arXiv:1710.11345*. [Online]. Available: http://arxiv.org/abs/1710.11345

[25] N. Rao, H.-F. Yu, P. K. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2107–2115.

[26] J. D. Gardiner, A. J. Laub, J. J. Amato, and C. B. Moler, "Solution of the Sylvester matrix equation $\mathrm{AXB}^T + \mathrm{CXD}^T = \mathrm{E}$," *ACM Trans. Math. Softw.*, vol. 18, no. 2, pp. 223–231, 1992.

[27] S. Karimi, "Global conjugate gradient method for solving large general Sylvester matrix equation," *J. Math. Model.*, vol. 1, no. 1, pp. 15–27, 2013.

[28] E. Million, "The Hadamard product," *Course Notes*, vol. 3, p. 6, Apr. 2007.

[29] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transp. Res. Part B, Methodol.*, vol. 91, pp. 511–524, Sep. 2016.

[30] X. Chen, Z. He, and J. Wang, "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition," *Transp. Res. Part C, Emerg. Technol.*, vol. 86, pp. 59–77, Jan. 2018.

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

[32] I. Jonsen, R. Myers, and M. James, "Identifying leatherback turtle foraging behaviour from satellite telemetry using a switching state-space model," *Mar. Ecol. Prog. Ser.*, vol. 337, pp. 255–264, May 2007.

[33] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[34] Y. Wang, W. Zhang, K. Henrickson, R. Ke, and Z. Cui, "Digital roadway interactive visualization and evaluation network applications to WSDOT operational data usage," Dept. Transp., Washington, DC, USA, Tech. Rep. WA-RD 854.1, 2016.

[35] W.-J. Li and D.-Y. Yeung, "Relation regularized matrix factorization," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, p. 1126.

Xudong Wang (Student Member, IEEE) 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.

Lijun Sun 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.