ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448399

SPATIOTEMPORAL GROUP ANOMALY DETECTION VIA GRAPH TOTAL VARIATION ON TENSORS

Mert Indibi and Selin Aviyente

Department of Electrical and Computer Engineering, Michigan State University
East Lansing, 48824
indibimu@msu.edu, aviyente@egr.msu.edu

ABSTRACT

Anomaly detection in spatiotemporal data is a challenging problem encountered in a variety of applications including urban traffic monitoring. Existing anomaly detection methods mostly focus on point anomalies and cannot deal with temporal and spatial dependencies that arise in spatiotemporal data. Tensor-based anomaly detection methods have been proposed to address this problem. While these methods are able to capture the dependencies across the different modes, they are mostly supervised and do not take the particular nature of anomalies into account. In this paper, we introduce an unsupervised tensor-based anomaly detection method that simultaneously considers the sparse and spatiotemporally smooth nature of anomalies. The anomaly detection problem is formulated as a regularized robust low-rank + sparse tensor decomposition where the spatiotemporal smoothness of the anomalies is quantified by the graph total variation with respect to the underlying spatial and temporal graphs. This minimization ensures that the extracted anomalies are temporally persistent and spatially smooth. The proposed framework is evaluated on both synthetic and real spatiotemporal urban traffic data.

***Index Terms*—** Anomaly Detection, Tensor Decomposition, Spatiotemporal Smoothness, Graph Total Variation, Urban Spatiotemporal Data.

1. INTRODUCTION

Large volumes of spatiotemporal (ST) data are commonly encountered in a diverse range of applications including climate science, social sciences, neuroscience, epidemiology [1], and transportation systems [2]. Detecting anomalies from these large data volumes is important for identifying interesting but rare phenomena, e.g., traffic congestion or irregular crowd movement in urban areas.

While various techniques have been developed for identifying anomalies in time-series [3] and spatial data [4], the joint presence of spatial and temporal aspects introduces novel ways of describing anomalies in ST data, resulting in unique anomaly detection problems. The most commonly addressed types of anomalies in ST data are point anomalies, trajectory anomalies and group anomalies. In this paper, we focus on group anomalies, which are defined as anomalies that appear as spatially contiguous groups of locations (regions) that show anomalous values consistently for a short duration of time. Some examples of such group anomalies include rare events such as traffic accidents that result in abnormally high traffic volume in a given region for a certain duration of time or an abnormal number of tweets from a spatial region in a small time window. Most approaches for detecting group anomalies decompose the anomaly detection problem by first treating the spatial and temporal properties of the outliers independently, which are then merged together in a post-processing step [5]. Recently, low-rank tensor decomposition methods have been proposed as they can capture the higher order correlations across spatial and temporal modes and recover the low-dimensional structure of data as the normal activity [6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

Relationship to Existing Work: Existing tensor decomposition based anomaly detection methods have multiple shortcomings. First, they [16] are mostly supervised or semi-supervised relying on historical data. Unsupervised tensor-based anomaly detection methods [12], on the other hand, aim to learn spatiotemporal features within a representation learning framework [17, 12, 18]. The learned features, *i.e.*, factor matrices or core tensors, are then used to detect anomalies by monitoring the reconstruction error at each time point [19, 20, 21, 12] or by applying well-known statistical tests to the extracted multivariate features [17, 8]. Second, current methods rely on well-known low-rank tensor approximation models such as Tucker [17, 8, 12] CP [11], higher order RPCA (HoRPCA) [6, 7], and do not explicitly consider the particular structure of anomalies.

In this paper, we simultaneously take into account the characteristics of normal and anomalous data where normal activity is low-rank; anomalies are sparse, temporally persistent, *i.e.*, the local changes last for a reasonably long time period, and spatially smooth, *i.e.*, neighboring regions are likely to exhibit similar anomalous activity. In prior work, we proposed Low-rank plus Temporally Smooth Sparse Decomposition (LOSS) [13] to incorporate temporal persistence into robust low-rank + sparse tensor decomposition. In the current paper, we generalize this framework by formulating the anomaly detection problem as a low-rank+sparse tensor decomposition with additional geometric structure on the sparse part. The temporal persistence and spatial contiguity of the sparse part, i.e., anomalies, are quantified by minimizing the graph signal variation with respect to both temporal and spatial graphs. This formulation also introduces a generalization of the conventional total variation norm to non-Euclidean domains by considering the corresponding graph Laplacian operators.

2. BACKGROUND

2.1. Notation and Tensor Operations

We use the calligraphic letters e.g., $\mathcal{X}$, to denote a multiway array (tensors), bold capital letters for 2-way arrays (matrices), e.g., $\mathbf{X}$, bold lowercase letters for 1-way arrays (vectors) e.g. $\mathbf{x}$, lowercase letters for scalars e.g. x, $\mathbf{I}$ to denote the identity matrix with appropriate dimensions and $x_{i_1,\ldots,i_n}$ to denote the entry of the tensor $\mathcal{X}$

This work was supported in part by the NSF under CCF-2006800.

indexed by $(i_1, \ldots, i_n)$

Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ be an N-mode tensor. We call vectors obtained by fixing all but the n-th index, mode-n fibers of the tensor. The mode-n unfolding of the tensor $\mathcal{X}$ is the matrix denoted as $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times \prod_{n' \neq n} I_{n'}}$ where the mode-n fibers of $\mathcal{X}$ are the columns of $\mathbf{X}_{(n)}$.

The matricized tensor is defined as $\mathbf{X}_{(\mathcal{R} \times \mathcal{C})} \in \mathbb{R}^{J \times K}$, with $J = \prod_{n \in \mathcal{R}} I_n$ and $K = \prod_{n \in \mathcal{C}} I_n$, where the ordered sets $\mathcal{R} = \{r_1, \ldots, r_L\}$ and $\mathcal{C} = \{c_1, \ldots, c_M\}$ correspond to a partitioning of the modes $\mathcal{N} = \{1, \ldots, N\}$. The indices in $\mathcal{R}$ are mapped to the rows and the indices in $\mathcal{C}$ are mapped to the columns. Specifically, $(\mathbf{X}_{(\mathcal{R} \times \mathcal{C})})_{i,j} = x_{i_1, i_2, \ldots, i_N}$ [22].

The mode-n product of $\mathcal{X}$ and matrix a $\mathbf{U} \in \mathbb{R}^{J \times I_n}$ is denoted as $\mathcal{Y} = \mathcal{X} \times_n \mathbf{U}$ where $\mathcal{Y} \in \mathbb{R}^{I_1 \times \ldots \times I_{n-1} \times J \times I_{n+1} \times \ldots \times I_N}$. The mode-$n$ product can also be expressed in matrix form as $\mathbf{Y}_{(n)} = \mathbf{U}\mathbf{X}_{(n)}$. We use the symbols, $\circ$ for Hadamard product and $\otimes$ for the Kronecker product operations.

2.2. Graph Signal Model

An undirected graph is denoted by $G = (V, E, W)$ where V is the node set with $|V| = n$, $E \subseteq V \times V$ is the edge set with $|E| = m$. An edge between nodes i and j is shown by e_{ij} and is associated with a weight W_{ij}. Algebraically, G can be represented by an $n \times n$ symmetric adjacency matrix $\mathbf{W}$ where $W_{ij} = W_{ji} = w_{ij}$ if $e_{ij} \in E$ and 0, otherwise. The Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal degree matrix, i.e., $D_{ii} = \sum_{j=1}^{n} W_{ij}$. The symmetrically normalized Laplacian is defined as $\mathbf{L}_n = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. A graph signal $\mathbf{x} \in \mathbb{R}^n$ is a vector whose entries reside on the nodes of an unsigned graph G.

Signal smoothness is a qualitative characteristic that expresses how much the signal samples vary with respect to the underlying graph. The notion of smoothness for a graph signal, $\mathbf{x}$, has been quantified using different definitions of total variation. In this paper, we use the graph total variation based on the ℓ_p norm used in [23]:

$$S_p(\mathbf{x}) = \|\mathbf{x} - \mathbf{A}_n\mathbf{x}\|_p^p, \tag{1}$$

where $\mathbf{A}_n$ is the normalized adjacency matrix to ensure that the shifted signal is properly scaled with respect to the original one.

2.3. Higher Order RPCA

For two-way data, robust PCA (RPCA) was introduced to address the limitations of PCA against outliers and non-Gaussian errors [24]. In this approach, a given data matrix is decomposed into a low-rank plus a sparse model:

$$\text{minimize}_{\mathbf{X},\mathbf{S}} \{ \text{ rank}(\mathbf{X}) + \lambda||\mathbf{S}||_0 \mid \mathbf{X} + \mathbf{S} = \mathbf{Y} \}. \tag{2}$$

In [25] higher-order RPCA (HoRPCA) is defined for tensors with (2) modified by replacing the rank of a matrix by the Tucker rank (Trank) of a tensor. Similar to RPCA, Trank and l_0 norms are replaced with their convex counterparts CTrank, i.e., the sum of the nuclear norms of each mode unfoldings, and l_1 norm yielding

$$\text{minimize}_{\mathcal{X},\mathcal{S}} \{ \text{ CTrank}(\mathcal{X}) + \lambda||\mathcal{S}||_1 \mid \mathcal{X} + \mathcal{S} = \mathcal{Y} \}. \tag{3}$$

Goldfarb and Qin [25] proposed various models to solve this optimization problem. One such model is the Singleton model, which estimates the nuclear norm of the tensor as the weighted sum of the nuclear norms of the mode-n unfoldings of the tensor yielding (4), where the nuclear norm is the Schatten-1 norm of the matrix.

$$\text{minimize}_{\mathcal{X},\mathcal{S}} \{ \sum_{i=1}^{N} \psi_i||\mathbf{X}_{(n)}||_* + \lambda||\mathcal{S}||_1 \mid \mathcal{X} + \mathcal{S} = \mathcal{Y} \}. \tag{4}$$

3. METHOD

In our method, we model the spatiotemporal data as a multi-way tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ where different modes correspond to spatial and temporal domains as well as different types of features. In order to extract the anomalies, we make three assumptions. First, normal activity can be modeled as a low-rank tensor while the anomalies are modeled as the sparse part, i.e., $\mathcal{Y} = \mathcal{X} + \mathcal{S}$. Second, the anomalies are assumed to last for periods of time, i.e., have strong short-term dependencies. This assumption ensures that instantaneous changes in the data, which may be due to errors in sensing, are not mistaken for actual anomalies. This assumption can be quantified as $\|\mathcal{S} \times_t \Delta\|_1$, where Δ is the first order discrete time differentiation operator which can be expressed as the Laplacian of the time domain line graph, i.e. $\Delta = \mathbf{I} - \mathbf{A}_t$, with $\mathbf{A}_t$ being the cyclic time shift operator defined in [26]. The graph total variation in the time domain with $p = 1$ can then be expressed as

$$S_1(\mathbf{S}_{(t)}) = \|\mathbf{S}_{(t)} - \mathbf{A}\mathbf{S}_{(t)}\|_1^1 = \|\mathbf{\Delta}\mathbf{S}_{(t)}\|_1 = \|\mathcal{S} \times_t \mathbf{\Delta}\|_1. \tag{5}$$

Finally, we assume that group anomalies within spatiotemporal data exhibit themselves as spatially contiguous groups of locations, i.e., the local variation of the anomalies is sparse. This assumption can be quantified by minimizing the variation of the anomaly with respect to the spatial graph, $\mathbf{A}_s$, with $p = 1$. In our case, the graph signals with respect to the spatial domain are the column vectors of the unfolding of $\mathcal{S}$ across the location mode, l.

$$S_1(\mathbf{S}_{(l)}) = \|\mathbf{S}_{(l)} - \mathbf{A}_n\mathbf{S}_{(l)}\|_1 = \|\mathbf{L}_n\mathbf{S}_{(l)}\|_1 = \|\mathcal{S} \times_l \mathbf{L}_n\|_1. \tag{6}$$

3.1. Optimization Formulation and Algorithm

Taking into account the three assumptions, we formulate the following optimization problem for ST anomaly detection.

$$\begin{aligned} \underset{\mathcal{X},\mathcal{S}}{\text{minimize}} \quad & \sum_{i=1}^{N} (\psi_i \|\mathbf{X}_{(i)}\|_*) + \lambda_1 \|\mathcal{S}\|_1 \\ & + \lambda_l \|\mathcal{S} \times_l \mathbf{L}_n\|_1 + \lambda_t \|\mathcal{S} \times_t \Delta\|_1 \\ \text{subject to} \quad & \mathcal{X} + \mathcal{S} = \mathcal{Y}, \end{aligned} \tag{7}$$

where λ_1, λ_l and λ_t are the regularization parameters that control the level of sparsity, amount of smoothing in the spatial domain and amount of smoothing in the temporal domain, respectively.

In order to separate the dependencies between the different terms containing $\mathcal{S}$, we introduce auxiliary variables $\mathcal{W} = \mathcal{S}, \mathcal{W}_l = \mathcal{S} \times_l \mathbf{L}_n$, $\mathcal{W}_t = \mathcal{S} \times_t \Delta$. Similarly, we define the auxiliary variables $\mathcal{X}_1, \ldots, \mathcal{X}_N = \mathcal{X}$ to separate the dependencies between the sum of nuclear norms containing the unfoldings of $\mathcal{X}$. In addition, the constraint $\mathcal{X} + \mathcal{S} = \mathcal{Y}$ can be replaced by a fidelity term with regularization parameter λ_2. The problem can be solved with a two block ADMM algorithm where the update blocks are, $\{\mathcal{X}, \mathcal{W}, \mathcal{W}_t, \mathcal{W}_l\}$ and $\{\mathcal{S}, \mathcal{X}_1, \ldots, \mathcal{X}_N\}$. The optimization problem in ADMM form becomes

$$\begin{aligned} \underset{\substack{\mathcal{X},\mathcal{W},\mathcal{W}_t,\mathcal{W}_l, \\ \{\mathcal{X}_i\}_{i=1,\ldots,N},\mathcal{S}}}{\text{minimize}} \quad & \sum_{i=1}^{N} \left(\psi_i \|\mathbf{X}_{i_{(i)}}\|_*\right) + \lambda_1 \|\mathcal{W}\|_1 + \lambda_l \|\mathcal{W}_l\|_1 \\ & + \lambda_t \|\mathcal{W}_t\|_1 + \frac{\lambda_2}{2} \|\mathcal{X} + \mathcal{S} - \mathcal{Y}\|_F^2 \\ \text{subject to} \quad & \mathcal{X} = \mathcal{X}_i, \ i = 1, ..., N, \\ & \mathcal{W} = \mathcal{S}, \quad \mathcal{W}_t = \mathcal{S} \times_t \mathbf{\Delta}, \mathcal{W}_l = \mathcal{S} \times_l \mathbf{L}_n. \end{aligned} \tag{8}$$

The convergence of two-block ADMM algorithms has been proven in [27]. Since (8) is a convex problem with two-block ADMM form, global convergence is guaranteed. A sketch of the proof of convergence is given in [13] for a similar model.

Augmented Lagrangian of the optimization problem in (8) is

$$\begin{aligned}\mathcal{L}_\rho = &\sum_{i=1}^{N} \big(\psi_i \|\mathbf{X}_{i_{(i)}}\|_* + \frac{\rho}{2}\|\mathcal{X} - \mathcal{X}_i + \frac{\Lambda_i}{\rho}\|_F^2\big) + \frac{\lambda_2}{2}\|\mathcal{X} + \mathcal{S} - \mathcal{Y}\|_F^2 \\ &+ \lambda_1\|\mathcal{W}\|_1 + \frac{\rho}{2}\|\mathcal{W} - \mathcal{S} + \frac{\Lambda}{\rho}\|_F^2 \\ &+ \lambda_l\|\mathcal{W}_l\|_1 + \frac{\rho}{2}\|\mathcal{W}_l - \mathcal{S} \times_l \mathbf{L}_n + \frac{\Lambda_l}{\rho}\|_F^2 \\ &+ \lambda_t\|\mathcal{W}_t\|_1 + \frac{\rho}{2}\|\mathcal{W}_t - \mathcal{S} \times_t \mathbf{\Delta} + \frac{\Lambda_t}{\rho}\|_F^2, \end{aligned} \tag{9}$$

where ρ is the step size and $\{\Lambda_i\}_{i=1,\dots,N}$,$\Lambda$, Λ_t, Λ_l are the dual variables for the constraints $\{\mathcal{X} = \mathcal{X}_i\}_{i=1,\dots,N}$, $\mathcal{W} = \mathcal{S}$, $\mathcal{W}_t = \mathcal{S} \times_t \mathbf{\Delta}$, $\mathcal{W}_l \times_l \mathbf{L}_n$, respectively.

Iterative update scheme for (8) is given in Algorithm 1. The update steps for the variables $\mathcal{W}, \mathcal{W}_l, \mathcal{W}_t$ are the proximal mappings of ℓ_1 norm defined as $\mathbf{prox}_{\lambda\|.\|_1}(\mathcal{B}) = \text{sign}(\mathcal{B}) \circ \max(|\mathcal{B}| - \lambda, 0)$, and update steps for the variables $\mathcal{X}_1, \dots, \mathcal{X}_N$ are obtained via the proximal operator of the Schatten-1 norm denoted as $\mathbf{prox}_{\psi\|.\|_*}(\mathbf{B})$ which corresponds to soft thresholding the singular value matrix of $\mathbf{B}$. The update step for the variable $\mathcal{S}$ involves the solution of (10), which is a quadratic problem with an analytical solution.

$$\begin{aligned}&\text{minimize}_{\mathcal{S}}\ \frac{\lambda}{2}\|\mathcal{X}^{k+1} + \mathcal{S} - \mathcal{Y}\|_F^2 + \frac{\rho}{2}\|\mathcal{W}_t^{k+1} - \mathcal{S} \times_t \mathbf{\Delta} + \frac{\Lambda_t^k}{\rho}\|_F^2 \\ &+\frac{\rho}{2}\|\mathcal{W}_l^{k+1} - \mathcal{S} \times \mathbf{L}_n + \frac{\Lambda_l^k}{\rho}\|_F^2. + \frac{\rho}{2}\|\mathcal{W}^{k+1} - \mathcal{S} + \frac{\Lambda^k}{\rho}\|_F^2. \end{aligned} \tag{10}$$

Let $\tilde{\mathcal{B}}^k = \rho\big(\mathcal{W}_t^{k+1} - \frac{\Lambda_t^k}{\rho} \times_t \mathbf{\Delta}^T + \mathcal{W}_l^{k+1} - \frac{\Lambda_l^k}{\rho} \times_l \mathbf{L}_n^T + \mathcal{W}^{k+1} - \frac{\Lambda^k}{\rho}\big) + \lambda_2\big(\mathcal{Y} - \mathcal{X}^{k+1}\big)$ and $\mathbf{G} = \rho\big[\mathbf{I}_l \otimes (\mathbf{\Delta}^T\mathbf{\Delta} + \frac{\rho+\lambda_2}{2\rho}\mathbf{I}_t) + (\mathbf{L}_n^T\mathbf{L}_n + \frac{\rho+\lambda_2}{2\rho}\mathbf{I}_l) \otimes \mathbf{I}_t\big]$. Solution for $\mathcal{S}$ matricized across the location (l) and time (t) modes, $\mathbf{S}_{(l,t)}$, becomes $\mathbf{G}^{-1}\tilde{\mathbf{B}}_{(l,t)}$. For large $\mathbf{L}_n$, $\mathcal{S}$ can be updated approximately using conjugate gradient method. We calculate the exact solution by caching the eigendecomposition of matrices $(\mathbf{L}_n^T\mathbf{L}_n + \frac{\rho+\lambda}{2\rho}\mathbf{I}_l)$ and $(\mathbf{\Delta}^T\mathbf{\Delta} + \frac{\rho+\lambda_2}{2\rho}\mathbf{I}_t)$ and using them to avoid inverting a matrix of size $n_l.n_t \times n_l.n_t$. The details of the implementation are given in [1].

4. RESULTS

In this section, we evaluate the performance of LR-STSS on both synthetic and real datasets and compare it with HoRPCA, low rank temporally smooth (LR-TS), i.e., $\lambda_l = 0$ in (8), and low rank spatially smooth (LR-SS), i.e., $\lambda_t = 0$ in (8). Absolute value of the recovered sparse part is used as the anomaly score. These anomaly scores are used to evaluate the area under the curve for the ROC curves (AUC-ROC). For real data, the number of detected events were reported for varying percentage of top K anomalies. For hyperparameter selection, we used the python library Optuna [28] with tree Parzen estimator [29]. In the synthetic experiment setting, the hyperparameters were selected to maximize the AUC score. In the real experiment setting, the hyperparameters were selected to maximize the number of total detected events.

[1] https://github.com/indibi/TensorAnomalyDetection

Algorithm 1 Low-rank and spatiotemporally smooth anomaly separation (LR-STSS) pseudocode

procedure LR-STSS($\mathcal{Y}$, $\mathbf{L}_n$, λ_1, λ_2, λ_t, λ_l, ψ_i, ρ)
  $\mathcal{X}, \mathcal{S}, \mathcal{W}, \mathcal{W}_t, \mathcal{W}_l \leftarrow \mathbf{0}$
  $\mathcal{X}_i \leftarrow \mathbf{0}, \quad i = 1 \cdots N$ ▷ Initialize primal variables
  $\Lambda, \Lambda_t, \Lambda_l \leftarrow \mathbf{0}$
  $\Lambda_i \leftarrow \mathbf{0}, \quad i = 1 \cdots N$ ▷ Initialize dual variables
  for $k = 0, \dots,$ maximum_iteration **do**
    ▷ First ADMM block updates
    $\mathcal{X}^{k+1} \leftarrow \frac{1}{\lambda_2 + N\rho}\big(\sum_i(\mathcal{X}_i^k\rho + \Lambda_i^k) + \lambda_2(\mathcal{Y} - \mathcal{S}^k)\big)$
    $\mathcal{W}_t^{k+1} \leftarrow \mathbf{prox}_{\frac{\lambda_t}{\rho}\|.\|_1}\Big(\mathcal{S}^k \times_t \mathbf{\Delta} - \frac{\Lambda_t^k}{\rho}\Big)$
    $\mathcal{W}_l^{k+1} \leftarrow \mathbf{prox}_{\frac{\lambda_l}{\rho}\|.\|_1}\Big(\mathcal{S}^k \times_l \mathbf{L}_n - \frac{\Lambda_l^k}{\rho}\Big)$
    $\mathcal{W}^{k+1} \leftarrow \mathbf{prox}_{\frac{\lambda_1}{\rho}\|.\|_1}\Big(\mathcal{S}^k - \frac{\Lambda_l^k}{\rho}\Big)$
    ▷ Second ADMM block updates
    $\mathbf{X}_{i_{(i)}}^{k+1} \leftarrow \mathbf{prox}_{\frac{\psi_i}{\rho}\|.\|_*}\Big(\mathbf{X}_{(i)}^{k+1} + \frac{\Lambda_i^k}{\rho}\Big), \quad i = 1, ..., N$
    $\mathbf{S}_{(l,t)}^{k+1} \leftarrow \mathbf{G}^{-1}\tilde{\mathbf{B}}_{(l,t)}^k$
    ▷ Dual variable updates
    $\Lambda_i^{k+1} \leftarrow \Lambda_l^k + \rho(\mathcal{W}^{k+1} - \mathcal{S}^{k+1})$
    $\Lambda_t^{k+1} \leftarrow \Lambda_t^k + \rho(\mathcal{W}_t^{k+1} - \mathcal{S}^{k+1} \times_t \mathbf{\Delta})$
    $\Lambda_i^{k+1} \leftarrow \Lambda_l^k + \rho(\mathcal{W}_l^{k+1} - \mathcal{S}^{k+1} \times_l \mathbf{L}_n)$
    $\Lambda_i^{k+1} \leftarrow \Lambda_i^k + \rho(\mathcal{X}^{k+1} - \mathcal{X}_i^{k+1}) \quad i = 1, ..., N$
  end for
end procedure

4.1. Synthetic Data Experiments

For synthetic data experiments, we generate the anomalous data $\mathcal{Y} = \mathcal{X} + \mathcal{S}$ with dimensions $n_1 \times n_2 \times n_3 \times n_4 = 40 \times 24 \times 7 \times 20$ representing the locations, hours in a day, days in a week and the number of subsequent weeks, respectively. Low-rank $\mathcal{X}$ corresponding to the normal activity is constructed with $\text{TRank}(\mathcal{X}) = (8, 8, 5, 5) = (m_1, m_2, m_3, m_4)$ by first generating a core tensor $\mathcal{C} \in \mathbb{R}^{8\times8\times5\times5}$ whose entries are drawn from the standard normal distribution then taking the mode-n products of the core tensor with random orthonormal matrices $\mathbf{U}_i \in \mathbb{R}^{n_i \times m_i}$ with appropriate conditioning [25], i.e., $\mathcal{X} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4$. We then normalize $\mathcal{X}$ by the standard deviation $std(\mathcal{X})$. Synthetic anomaly groups in $\mathcal{S}$ are randomly generated to be locally contiguous over an 8×5 Cartesian grid graph with a nonzero temporal duration. This is achieved by first randomly selecting an element indexed by (i_1, i_2, i_3, i_4) per anomaly group in the tensor where i_1 is the local center of the anomaly and setting all of the vertices within a r-hop distance to 1, where r is the spatial radius of the anomaly. A rectangular pulse with duration d centered around index i_4 is applied to each of these locations to generate the ST anomaly with temporal persistence. In our experiments, we study the performance of the algorithms by varying the number of anomaly groups, d and r. The experiments were repeated 10 times and the mean AUC scores are reported.

4.1.1. Synthetic Experiment 1: Varying the Number of Anomalies

In the first experiment, we evaluate the performance of the different methods for varying number of anomalies when $r = 1$ and $d = 1$. Fig. 1 shows that the performance of all methods drops as the number of anomalies increase since the sparsity of the anomaly is no longer a valid assumption. LR-SSTS performs the best followed

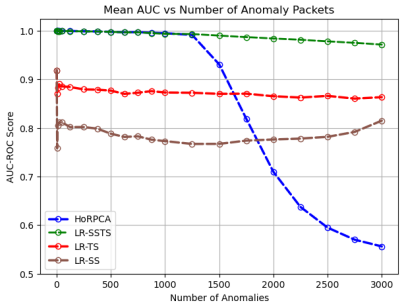

Fig. 1: Mean AUC vs. Number of Anomalies for $r = 1$ and $d = 1$

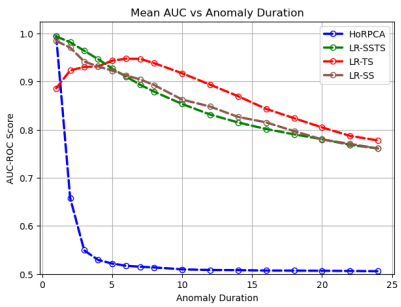

Fig. 2: Mean AUC vs. Varying Temporal Duration of Anomalies

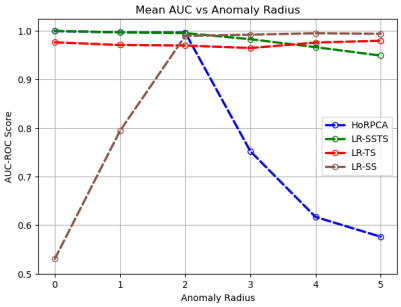

Fig. 3: Mean AUC vs. Varying Local Radius of Anomalies

by LR-TS. It's also interesting to note that LR-SSTS is robust to increasing number of anomalies as the performance drop is much smaller compared to HoRPCA.

4.1.2. Synthetic Experiment 2: Varying Anomaly Duration

In the second experiment, we evaluate the performance of the different methods for varying anomaly duration, d. In this case, the number of anomaly groups is set to 450 and the anomaly radius is set to 2. For short anomaly durations, LR-STSS performs the best. However, as the duration of the anomaly increases ($d > 5$), LR-TS becomes the best performing method since temporal persistence of the anomaly becomes important. When the duration becomes larger than 8, the performance starts to drop for LR-TS since the number of anomalies increases with the increase in duration.

4.1.3. Synthetic Experiment 3: Varying Local Anomaly Radius

In the third experiment, we evaluate the performance of the different methods for varying local anomaly radius, r. In this case, the number of anomaly groups is set to 100 and the anomaly duration is set to 4. From Fig. 3, it can be seen that as r increases the performance of LR-SS increases since the spatial contiguity assumption becomes more dominant. The performance of LR-STSS is better than LR-SS for $r < 3$, after that point their performances become comparable with LR-SS having slightly higher AUC.

4.2. Real Data: 2018 NYC Taxi Data Anomaly Detection

We use the NYC yellow taxi trip records for 2018 as our real spatiotemporal data. This dataset consists of trip information such as the departure zone and time, arrival zone and time, number of passengers, tips for each yellow taxi trip in NYC. In the following experiments, we only use the arrival zone and time to collect the number of arrivals for each zone aggregated over one hour time intervals. We selected 81 central zones to avoid zones with very low traffic. Thus, we created a tensor $\mathcal{Y}$ of size $24 \times 7 \times 53 \times 81$, where the first mode corresponds to hours within a day, the second mode corresponds to days of a week, the third mode corresponds to weeks of a year and the last mode corresponds to the zones. The adjacency matrix of the physical locations graphs is constructed using a k-NN binary graph with $k = 2$ where the Euclidean distance between location pairs is computed between the vectors corresponding to the number of hourly arrivals.

To evaluate the performance of the proposed methods on real data, we compiled a list of 20 urban events that took place in the important urban activity centers such as city squares, concert halls etc. during 2018 [13]. To detect the events, top-K percent, with varying K, of the highest anomaly scores of the extracted sparse tensors are selected as anomalies and compared against the compiled list. In previous work, similar case studies were presented for experiments on real data [13, 30, 31].

%	0.014	0.07	0.14	0.3	0.7	1	2	3
LR-STSS	**3**	**4**	**7**	**12**	**15**	**17**	**19**	**19**
LR-TS	3	4	5	6	13	13	18	19
LR-SS	1	1	2	3	5	6	13	16
HoRPCA	0	0	2	2	2	3	7	10

Table 1: Number of detected events among 20 compiled events in NYC for varying top-$K\%$ of the anomaly scores

From Table 1, it can be seen that LR-STSS can detect more events by using smaller percentage of the top anomaly scores. This is followed by LR-TS which implies that for real urban traffic temporal persistence in the hours mode plays an important role in identifying anomalies. HoRPCA, on the other hand, does not perform well as it cannot detect most of the events. This result shows the importance of taking the spatiotemporal characteristics of anomalies into account as sparsity by itself is not sufficient to model them.

5. CONCLUSIONS

In this paper, we introduced a new tensor based anomaly detection method for group anomaly detection in spatiotemporal data. The proposed method models the group anomalies as sparse, spatially contiguous and temporally persistent entries of the observed tensor. These assumptions are quantified by graph total variation of the sparse part of the tensor with respect to the underlying temporal and spatial adjacency graphs. The performance of the resulting algorithm is evaluated for both synthetic and real ST data and shown to be superior compared to HoRPCA, LR-SS and LR-TS. While the current paper focuses on the smoothness of the anomaly across the temporal and spatial modes, the proposed framework can be generalized to other types of higher order data including video and functional MRI (fMRI) data where the smoothness can be defined for more than two domains. The current formulation can also be extended by increasing the order of the highpass graph filtering, i.e., powers of $\mathbf{L}$, to consider higher order neighborhoods for different anomaly profiles across time and space.

6. REFERENCES

[1] G. S. Bhunia, S. Kesari, N. Chatterjee, V. Kumar, and P. Das, "Spatial and temporal variation and hotspot detection of kala-azar disease in vaishali district (bihar), india," *BMC infectious diseases*, vol. 13, no. 1, p. 64, 2013.

[2] Y. Djenouri, A. Belhadi, J. C.-W. Lin, D. Djenouri, and A. Cano, "A survey on urban traffic anomalies detection algorithms," *IEEE Access*, vol. 7, pp. 12 192–12 205, 2019.

[3] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2013.

[4] C. C. Aggarwal and C. C. Aggarwal, *An introduction to outlier analysis*. Springer, 2017.

[5] J. H. Faghmous, M. Le, M. Uluyol, V. Kumar, and S. Chatterjee, "A parameter-free spatio-temporal pattern mining model to catalog global ocean dynamics," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 151–160.

[6] S. Li, W. Wang, H. Qi, B. Ayhan, C. Kwan, and S. Vance, "Low-rank tensor decomposition based anomaly detection for hyperspectral imagery," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4525–4529.

[7] X. Geng, K. Sun, L. Ji, and Y. Zhao, "A high-order statistical tensor based algorithm for anomaly detection in hyperspectral imagery," *Scientific reports*, vol. 4, p. 6869, 2014.

[8] X. Zhang, G. Wen, and W. Dai, "A tensor decomposition-based anomaly detection algorithm for hyperspectral image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5801–5820, 2016.

[9] L. Chen, J. Jakubowicz, D. Yang, D. Zhang, and G. Pan, "Fine-grained urban event detection and characterization based on tensor cofactorization," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 3, pp. 380–391, 2016.

[10] C. Lin, Q. Zhu, S. Guo, Z. Jin, Y.-R. Lin, and N. Cao, "Anomaly detection in spatiotemporal data via regularized non-negative tensor analysis," *Data Mining and Knowledge Discovery*, vol. 32, no. 4, pp. 1056–1073, 2018.

[11] Z. Li, N. D. Sergin, H. Yan, C. Zhang, and F. Tsung, "Tensor completion for weakly-dependent data on graph for metro passenger flow prediction," *arXiv preprint arXiv:1912.05693*, 2019.

[12] M. Xu, J. Wu, H. Wang, and M. Cao, "Anomaly detection in road networks using sliding-window tensor factorization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4704–4713, 2019.

[13] S. E. Sofuoglu and S. Aviyente, "Gloss: Tensor-based anomaly detection in spatiotemporal urban traffic data," *Signal Processing*, vol. 192, p. 108370, 2022.

[14] ——, "Low-rank on graphs plus temporally smooth sparse decomposition for anomaly detection in spatiotemporal data," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5614–5618.

[15] T. Cheng and B. Wang, "Graph and total variation regularized low-rank representation for hyperspectral anomaly detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 391–406, 2019.

[16] H. Fanaee-T and J. Gama, "Tensor-based anomaly detection: An interdisciplinary survey," *Knowledge-Based Systems*, vol. 98, pp. 130–147, 2016.

[17] ——, "Event detection from traffic tensors: A hybrid model," *Neurocomputing*, vol. 203, pp. 22–33, 2016.

[18] L. Shi, A. Gangopadhyay, and V. P. Janeja, "Stensr: Spatio-temporal tensor streams for anomaly detection and pattern discovery," *Knowledge and Information Systems*, vol. 43, no. 2, pp. 333–353, 2015.

[19] E. E. Papalexakis, A. Beutel, and P. Steenkiste, "Network anomaly detection using co-clustering," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2012, pp. 403–410.

[20] E. Papalexakis, K. Pelechrinis, and C. Faloutsos, "Spotting misbehaviors in location-based social networks using tensors," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 551–552.

[21] J. Sun, D. Tao, and C. Faloutsos, "Beyond streams and graphs: dynamic tensor analysis," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 374–383.

[22] T. G. Kolda, "Multilinear operators for higher-order decompositions." Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA . . . , Tech. Rep., 2006.

[23] S. Chen, A. Sandryhaila, J. M. Moura, and J. Kovačević, "Signal recovery on graphs: Variation minimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 17, pp. 4609–4624, 2015.

[24] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

[25] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 1, pp. 225–253, 2014.

[26] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[27] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," *Journal of Scientific Computing*, vol. 66, no. 3, pp. 889–916, 2016.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[29] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*. PMLR, 2013, pp. 115–123.

[30] M. Zhang, T. Li, Y. Yu, Y. Li, P. Hui, and Y. Zheng, "Urban anomaly analytics: Description, detection and prediction," *IEEE Transactions on Big Data*, 2020.

[31] M. Zhang, T. Li, H. Shi, Y. Li, and P. Hui, "A decomposition approach for urban anomaly detection across spatiotemporal data," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 6043–6049.