IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 10, 2019, accepted May 6, 2019, date of publication May 14, 2019, date of current version May 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916567

Graph Learning Based on Spatiotemporal Smoothness for Time-Varying Graph Signal

YUELIANG LIU[1,2], (Student Member, IEEE), LISHAN YANG[1], (Student Member, IEEE), KANGYONG YOU[1], (Student Member, IEEE), WENBIN GUO[1,2], (Member, IEEE), AND WENBO WANG[1], (Senior Member, IEEE)

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, Shijiazhuang 050000, China

Corresponding author: Wenbin Guo (gwb@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61271181 and Grant 61571054, in part by the Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory Foundation, and in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2018101.

ABSTRACT Graph learning often boils down to uncovering the hidden structure of data, which has been applied in various fields such as biology, sociology, and environmental studies. However, distributed sensing in realistic application often gives rise to spatiotemporal signals, which can be characterized through new tools of graph signal processing as a time-varying graph signal. It calls upon the development from static graph signal studies to the joint space-time analysis. In this paper, we study the problem of learning graphs from time-varying graph signals. Based on the correlated properties in observed signals, a dynamic graph-based model is first presented, which particularly takes into account space-time interactions in signal representation. Considering the case that the time correlation pattern is unavailable, the graph learning problem is cast as a joint correlation detecting and graph refining problem. Then it is solved by the proposed correlation-aware and spatiotemporal smoothness-based graph learning method (CASTS), which novelly introduces the spatiotemporal smooth prior to the field of time-vertex signal analysis. By promoting such smoothness in each learning steps, the graph learning accuracy can be efficiently improved. The experiments on both synthetic and real-world datasets demonstrate the improvement of the proposed CASTS over current state-of-the-art graph learning methods, and meanwhile show the capability of dynamic prediction in climate analysis.

INDEX TERMS Graph learning, time-varying graph signal, spatiotemporal smoothness, correlation, space-time interaction.

I. INTRODUCTION

With the explosive growth of dataset in a variety of applications, from finance and biology to social and sensor networks, spatiotemporal data often emerges as long time series measured and stored over a certain spatial region. For instance, environmental sensor networks are often deployed in a climate zone. The geographical location of distributed sensors leads to spatial correlation. Meanwhile, real-time distributed sensing in days, months and years, collects a large number of time series which exhibit time dependence of sensors in dynamic evolution. Nevertheless, though the analysis of spatiotemporal data has been successful in sociology [1], brain imaging [2] and climate researches [3], it is still a challenging problem due to the correlation properties and complex space-time interactions.

A flexible way to characterize the spatiotemporal data living on an irregular domain is to use a graph [4], referred to as time-varying graph signals. In recent years, graph signal processing (GSP) [5], [6] provides a new engineering paradigm for processing and analyzing signals on graphs, by utilizing graph Laplacian matrices to deal with multiple tasks such as graph filtering [7], [8], graph signal compression [9], [10] and sampling and reconstruction on graph signals [11], [12], etcetera. These researches in GSP have taken full advantage of graph structure which is prior known or pre-defined, e.g., road connection of transportation network [5], [13] or geographical k nearest neighbor (KNN) models [11], [14].

However, in many cases the graph structure itself may be unknown, which significantly raises demands for efficient

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tan.

methods to reveal the underlying mechanism in real-world complex systems. For example, air temperature provides valuable information for monitoring climate dynamics. Due to a potential relationship between geographical locations, the temperature in neighboring area interacts with each other and exhibits similar variation tendency when affected by atmospheric circulation or ocean currents. Hence, it is critical to extract the structural information from collected data for weather condition forecasting and advanced decision making. For another example, in water evaporation sensor network, graph structure is not readily available nor easy to define. In order to determine the optimal deployment of sensor nodes and complete further processing such as adaptable sampling and fast inpainting, it also calls for graph learning method to provide some useful information. Therefore, this paper focus on learning hidden graph from time-varying graph signals, i.e., to recover the graph Laplacian matrix from the observations.

A. RELATED WORKS

In the literature, since the correlation among signals may fail to capture the causality relations, most efforts in early researches have been dedicated to graphical model estimation [15]–[18], where a sparse precision matrix (SPM) is recovered to reflect the relationship between signals. In [15], Dempster firstly proposes a method that regulates the SPM with sparsity by forcing the off-diagonal entries with zeroes. Then a widely-used method of recovering SPM is Graphical Lasso [16]. Improvements of the Graphical Lasso in computational efficiency are gradually discussed in [17] and [18].

Nowadays the emerging GSP provides a strong impulse to discover new techniques for inferring the graph topology. Several GSP-based studies aim to recover the adjacency or Laplacian matrix by postulating the sparsity [19] or smoothness of signals. They extend the classical methods on SPM to combinatorial graph Laplacian (CGL) learning and achieve satisfying results. Lake and Tenenbaum [20] address the adjacency matrix learning by introducing logarithmic function in optimization to force sparse connectivity. To avoid the above full-rank restriction of graph Laplacian matrix, Dong *et al.* [21] relax the constraints and propose to learn a valid CGL under the smoothness prior. Particularly, they present a smooth signal representation that is consistent with the given statistical prior to latent variables. Later, a generalized graph learning framework is discussed in [22], where the optimization problem is reformulated as an l-1 minimization with an additional penalty term on node degrees and is solved by GSP Toolbox [23]. Then, alternative approaches [25], [26] make an assumption that the observation is a Gaussian Markov Random Field (GMRF) [24] whose precision matrix is graph Laplacian. Egilmez *et al.* [25] formulate the graph learning problem as recovering the precision matrix in different types of graph Laplacian given some structural constraints. Rabbat [26] studies a threshold-based estimator for inferring a graph, and meanwhile provides theoretical results on the reconstruction error for the case that the graph is sparse. Since real-world data tend to be smooth on graph rather than strictly bandlimited, smoothness-based graph learning methods continue to receive a lot of interest. Chepuri *et al.* [27] simplify the graph learning problem to propose an edge selection mechanism based on the smoothness assumption, while Kalofolias *et al.* [28] learn a time-varying graph by imposing a new prior that graph edges change smoothly in time.

There are a few recent researches that aim to learn graph topology from signals based on the diffused model [29]–[32] and casual model [33]–[35]. Segarra *et al.* [29] focus on identifying graph shift operators given the only eigenvectors of shift operators. The eigenvalues are estimated from the covariance matrix of stationary signals that are postulated a diffusion process on the graph. Pasdeloup *et al.* [30] describe the graph learning problem in a similar way as [29], yet through a different matrix selection strategy. To overcome the stationary limitation, Shafipour *et al.* [31] explore the problem of graph inference from non-stationary graph signals. Thanou *et al.* [32] propose a graph learning method, under the assumption that graph signals are generated from heat diffusion processes, by imposing a Laplace prior to control the sparsity. Another works in [33]–[35] concentrate on estimating asymmetric adjacency matrix which corresponds to directed graphs. In [33], Mei and Moura propose an algorithm for estimating the adjacent matrix that describes the dependence among time series. Authors in [34] utilize a structural equation model (SEM) to capture causal relationships, and meanwhile jointly track the signal state and graph structure through recursive least-squares estimator. Similar to the SEM, Shen *et al.* [35] propose to model nonlinear dependencies of signals based on a structural vector autoregressive model (SVARM), in which an efficient regularized estimator is developed to infer a sparse graph topology. However, the aforementioned graph learning methods are not specifically designed for time-varying graph signals, and hence do not explore space-time interaction and correlated property of spatiotemporal signals to facilitate graph learning procedure.

B. CONTRIBUTIONS

In this paper, to learn the underlying structure from time-varying graph signals that are prevalent in real-world applications, a correlation-aware and spatiotemporal smoothness-based graph learning method is proposed. The main contributions of this paper are summarized as follows.

1) To the best of our knowledge, we first present a dynamic graph-based model that integrates the property of space-time interactions into a linear dynamic system for comprehensive signal representation. Specifically, by exploiting the correlated properties in both space and time dimension, spatiotemporal smoothness is novelly introduced to the field of time-vertex signal analysis.
2) Under the dynamic model, graph learning problem is formulated as an optimization problem based on correlation and spatiotemporal smoothness with respect to the graph structure, which is then solved by the proposed

Correlation-aware and Spatiotemporal Smoothness-based graph learning method (CASTS) as an application of the block coordinate descent scheme. Through simultaneously capturing the dependencies among time series and enforcing the spatiotemporal smoothness property of graph signal in each iteration steps, it brings improvement on graph learning accuracy.

3) We provide performance analysis of the proposed method on both synthetic and real-world datasets. Specifically, we perform the visual and quantitative comparison for assessing the accuracy of the graph topology estimation. In addition, extensive tasks on real-world datasets demonstrate the effectiveness of dynamic graph-based model and the superior learning performance of the proposed CASTS over the state-of-the-art graph learning methods.

The remainder of this paper is organized as follows. In Section II, a brief overview of the notation and the basics including graph Laplacian and smooth graph signals are reviewed. In Section III, a dynamic graph-based model is proposed and spatiotemporal smoothness of time-varying graph signal is introduced. In Section IV, we formulate the graph learning problem as an optimization problem, and propose CASTS to alternatively solve the optimization problem. The performance of the proposed CASTS is evaluated and compared with baseline methods on both synthetic and real-world datasets in Section V. Section VI concludes the whole paper.

II. NOTATION AND PRELIMINARIES

A. NOTATIONS

Throughout the paper, the lowercase boldface letters, e.g., $\mathbf{x}$ and the uppercase boldface letters, e.g., $\mathbf{X}$, denote vectors and matrices, respectively. Given a vector $\mathbf{x}$, x_i denotes the ith entry of $\mathbf{x}$, and $\mathbf{1}$ and $\mathbf{0}$ denote the constant one and zero vectors. For a matrix $\mathbf{X}$, X_{ij} denotes the element on the ith row and jth column. An undirected and weighted graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V}$ is the set of vertices with $|\mathcal{V}| = N$ and $\mathcal{E}$ is the edge set of the graph. The matrix $\mathbf{W}$ is the weighted adjacency matrix, with W_{ij} denoting the positive weight of an edge connecting vertices i and j. Otherwise, $W_{ij} = 0$ if there is no edge. For a vector $\mathbf{x} \in \mathbb{R}^N$, diag (x) denotes the diagonal matrix with its diagonal elements $\{x_1, \ldots, x_N\}$. For a matrix $\mathbf{X} \in \mathbb{R}^{N\times M}$, the vectorization, the trace, and the Frobenius norm of matrix are denoted as vec $(\mathbf{X})$, tr $(\mathbf{X})$, and $\|\mathbf{X}\|_F$, respectively. In addition, $\otimes$ is the Kronecker product operator.

B. GRAPH LAPLACIAN

The graph-based model in this paper focuses on an N-vertex, undirected, weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ with positive edge weights. The CGL of $\mathcal{G}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ denotes the degree matrix of $\mathcal{G}$ which is a diagonal matrix with entries $\text{diag}(\mathbf{D})_i = \sum_{j=1}^{N} W_{ij}$ and $(\mathbf{D})_{ij} = 0$ for $i \neq j$. Based on the definition above, the set of CGL matrices can also be written as

$$\mathcal{L}^N = \left\{ \mathbf{L} \in \mathbb{R}^{N\times N} | \mathbf{L} \succeq 0, L_{ij} = L_{ji} \leq 0, i \neq j, \text{ and } \mathbf{L} \cdot \mathbf{1} = \mathbf{0} \right\}. \tag{1}$$

The CGL in (1) is a real symmetric positive semidefinite matrix. Thus, the eigendecomposition of CGL can be represented as $\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N]$ are the matrix of nonnegative eigenvalues and orthogonal eigenvectors, respectively. The ascending array of eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ are called as graph frequencies in the graph spectral domain. The eigenvectors associated with low frequencies indicate the slow variation of graph signals across the graph, while the ones associated with high frequencies imply a rapid variation of signals on graph. In particular, zero value appears as an eigenvalue (i.e., $\lambda_1 = 0$), whose multiplicity equals to one showing that the graph contains only one connected component.

C. GRAPH LEARNING FROM SMOOTH GRAPH SIGNALS

With the eigenvalues of CGL indicating the frequency, the graph Fourier transform of graph signal $\mathbf{x}$ is defined as $\hat{\mathbf{x}} = \mathbf{U}^T\mathbf{x}$. The frequency components corresponding to smaller eigenvalues (i.e., frequency within $[0, w)$ with cut-off frequency w) are low frequency components, which brings the concept of bandlimited graph signals in GSP. However, the real-world data tend to be smooth, rather than strictly bandlimited. Thus, smoothness-based theory extends the studies on bandlimited graph signals to the smooth graph signals.

A static graph signal x over a $\mathcal{G}$ is defined as a vector mapping from the graph vertex domain to the real number field, i.e., $x : \mathcal{V} \to \mathbb{R}^N$, such that $x(i)$ denotes the value of graph signal on the vertex i. When it comes to graph learning problem, the smooth property is widely used as prior information. The smoothness of graph signals is a qualitative characteristic that expresses how frequently a graph signal varies with respect to the underlying graph [6]. In GSP, the smoothness of graph signal x is quantified by graph Laplacian quadratic form [5] as

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \sum_{(i,j)\in\mathcal{E}} W_{i,j}[x(j) - x(i)]^2, \tag{2}$$

which measures the total variation of the connecting vertices. The $\mathbf{x}^T\mathbf{L}\mathbf{x}$ is small when a large weight of edge connects the two vertices whose values are similar. In general, the smaller value of $\mathbf{x}^T\mathbf{L}\mathbf{x}$, the smoother the signal on graph.

For a smooth graph signal x that is assumed to yield an attractive GMRF, smoothness-based (static) graph learning problem can be expressed as the following general formulation

$$\min_{\mathbf{W}\in\mathcal{W}} \|\mathbf{W} \circ \mathbf{Z}\|_{1,1} + S(\mathbf{W}), \tag{3}$$

where $Z_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ and $\circ$ is the Hadamard product. The first term corresponds to tr $(\mathbf{X}^T\mathbf{L}\mathbf{X})$ in a matrix form of (2) and $\mathcal{W}$ denotes the set of valid adjacency matrices

(symmetric and positive). The second term, penalty function $S(\mathbf{W})$, determines the sparsity of learned graph structure. For example, Kalofolias [22] defines

$$S(\mathbf{W}) = c\mathbf{1}^T \log(\mathbf{W} \cdot \mathbf{1}) + d\|\mathbf{W}\|_F^2, \tag{4}$$

with regularization parameters c and d. While Dong *et al.* [21] impose $\|\mathbf{W}\|_{1,1} = \mathrm{tr}(\mathbf{W}) = N$ and propose

$$S(\mathbf{W}) = c\|\mathbf{W} \cdot \mathbf{1}\|^2 + c\|\mathbf{W}\|_F^2. \tag{5}$$

The sparsity of graph in the former method is controlled by the logarithmic term in (4) which ensures the degree of each vertex is not empty, while the latter one penalizes the degree of vertices in (5). A suitable choice depends on the features of data and the application.

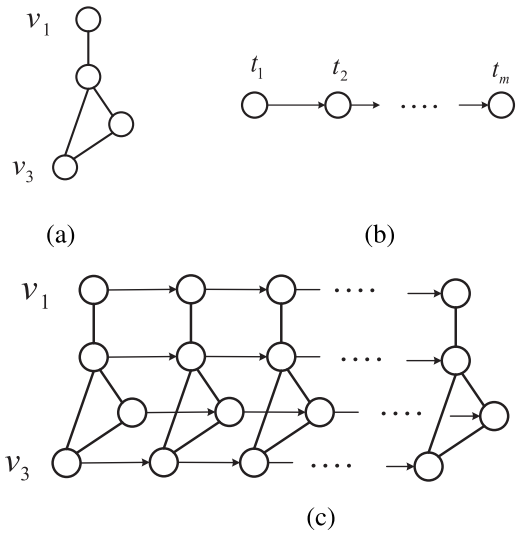

FIGURE 1. The time-vertex analysis of time-varying graph signal. (a) Spatial correlation described by a graph. (b) Time correlation in one graph vertex. (c) Time-varying graph signal. Edges connecting the vertices at each time instant denote the spatial correlation represented by solid lines. Time-dependent processes where edges connecting at different time instants are shown by arrow lines.

D. CORRELATION OF TIME-VARYING GRAPH SIGNALS

As depicted in Fig. 1, spatiotemporal data can be viewed as time-varying graph signals living on a graph of the observations with edges labeling the intrinsic relationship. It is pointed out in [36] that nearby values in both space and time directions tend to be more similar than those far apart. This means the existence of a strong correlation in time-varying graph signals which are not just the static graph signals stacked into a sequence. These prevalent properties in time-varying graph signal are stated as follows.

- *Spatial smoothness:* At each time instant in the spatial dimension, time-varying graph signals in geographical neighbors are close to each other.
- *Temporal smoothness:* For each observation site in the temporal dimension, influenced by the temporal correlation, the observed value varies smoothly over time.

To be noted, spatial smoothness prior has been applied in many graph learning studies including [21], [22] and [27]. There are a few works, such as [11], [28], learning graphs based on temporal smoothness. By combining the above two types of smoothness together, we introduce spatiotemporal smoothness in the following assumption.

Assumption 1 (Spatiotemporal smoothness): The temporal evolution of time-varying graph signal is smooth on the graph topology.

The detail of spatiotemporal smoothness is described in *Definition* 2, which can bring benefit to graph learning.

III. SPACE-TIME REPRESENTATION FOR TIME-VARYING GRAPH SIGNALS

A. DYNAMIC GRAPH-BASED MODEL

A time-varying graph signal can be expressed by a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] \in \mathbb{R}^{N\times M}$, where N is the number of the observing sites and M is the number of the time instants. We consider the following dynamic model for the measured signal

$$\mathbf{x}_t = \mathbf{U}\mathbf{h}_t + \mathbf{n}_t, \tag{6}$$
$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{v}_t. \tag{7}$$

For the graph vertex domain, the Eq. (6) is an observation model which maps the latent state space into the observed space. The observation and latent variable at time instant t are respectively denoted as $\mathbf{x}_t \in \mathbb{R}^N$ and $\mathbf{h}_t \in \mathbb{R}^N$. The eigenvector matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$, which can be interpreted as the graph Fourier basis for representing graph signal [37], is selected as representation matrix. It linearly relates these two variables with graph structural information. In addition, we assume that the observation noise $\mathbf{n}_t$ follows a multivariate Gaussian distribution with zero mean and covariance ${\sigma_n}^2\mathbf{I}_N$, which can be expressed as

$$\mathbf{n}_t \sim \mathcal{N}\left(\mathbf{0}, {\sigma_n}^2\mathbf{I}_N\right), \tag{8}$$

For the time domain, we add Eq. (7) to characterize the temporal evolution by imposing a first-order Gaussian Markov autoregressive process on latent variable. Concretely speaking, the state transition matrix that is applied to the previous state $\mathbf{h}_{t-1}$ is denoted as $\mathbf{A} = \mathbf{U}^T\mathbf{R}\mathbf{U}$, where $\mathbf{R} = \mathrm{diag}(c_1, c_2, \ldots, c_N) \in \mathbb{R}^{N\times N}$ is the time correlation matrix and the diagonal element c_i is the correlation coefficient in ith observation site with value ranging from 0 to 1. The parameter c is similar to the Pearson correlation coefficient that describes the correlation of data with a delayed copy (one-time lag in our model) of itself. Besides, in real-world applications such as sensor networks, due to different geographical locations of observation sites, the collected data may exhibit different correlation property. To generalize our model that can characterize multiple types of time-correlated data, we set distinct c_i in $\mathbf{R}$. Observing that the transition matrix $\mathbf{A}$ is a space-time coupling term which encodes both spatial and temporal information through $\mathbf{U}$ and $\mathbf{R}$, respectively. Furthermore, the process variable $\mathbf{v}_t$ is assumed to

follow a multivariate Gaussian distribution with precision matrix defined as the eigenvalue matrix Λ of the graph Laplacian $\mathbf{L}$, i.e.,

$$\mathbf{v}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Lambda}^{\dagger}\right), \tag{9}$$

where $\mathbf{\Lambda}^{\dagger}$ is the Moore-Penrose pseudoinverse of $\mathbf{\Lambda}$. Together with the definition of $\mathbf{U}$ as the representation matrix, the above assumption on $\mathbf{v}_t$ directly links the graph structure to the temporal evolution of graph signals, which offers a way to model the structured data. Meanwhile, the assumption implies that the energy of weighted difference signal tends to lie mainly in the low frequency and thus it promotes the spatiotemporal smoothness for the signal on the graph, which is shown in the next section.

Notice that the above linear dynamic model in (6) and (7) may be regarded as analogous to the equations of the Kalman Filter. The difference between the two models is twofold. First, at each discrete time increment, we postulate a Gaussian Markov autoregressive on latent variable, which links the dynamics of time-varying graph signals in the time domain with their spatial structure in the graph vertex domain. That is to say, spatial and temporal processes closely interact with each other, which is shown in the following derivation. Second, the Kalman filter works with prior known observation model and state transition model, while the representation matrix $\mathbf{U}$ in (6) and transition matrix $\mathbf{A}$ in (7) are unknown, leading to an additional estimation procedure for further processing.

To intuitively understand the space-time process interacting dynamics in the proposed model, we first introduce weighted difference signal. Utilizing the orthogonality of $\mathbf{U}$, the tth component is given as

$$\begin{aligned}\mathbf{d}_t &= \mathbf{x}_t - \mathbf{R}\mathbf{x}_{t-1} = \mathbf{U}\mathbf{A}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{v}_t + \mathbf{n}_t - \mathbf{R}\mathbf{U}\mathbf{h}_{t-1} - \mathbf{R}\mathbf{n}_{t-1} \\ &= \mathbf{U}\mathbf{v}_t + \mathbf{n}_t - \mathbf{R}\mathbf{n}_{t-1},\end{aligned} \tag{10}$$

with the first term defined as $\mathbf{d}_1 = \mathbf{x}_1$. Based on (8) and (9), the conditional probability of $\mathbf{d}_t$ given $\mathbf{v}_t$ and the probability of $\mathbf{d}_t$ are given as

$$\mathbf{d}_t | \mathbf{v}_t \sim \mathcal{N}\left(\mathbf{U}\mathbf{v}_t, {\sigma_n}^2\left(\mathbf{I}_N + \mathbf{R}\mathbf{R}^T\right)\right), \tag{11}$$

$$\mathbf{d}_t \sim \mathcal{N}\left(\mathbf{0}, \mathbf{L}^{\dagger} + {\sigma_n}^2\left(\mathbf{I}_N + \mathbf{R}\mathbf{R}^T\right)\right), \tag{12}$$

where $\mathbf{L}^{\dagger}$ is the pseudo-inverse of $\mathbf{L}$ and it admits the following eigendecomposition $\mathbf{L}^{\dagger} = \mathbf{U}\mathbf{\Lambda}^{\dagger}\mathbf{U}^T$.

As shown in Eq. (12), in a noise-free scenario where $\sigma_n = 0$, the weighted difference signal $\mathbf{d}_t$ follows a degenerate multivariate Gaussian distribution where the precision matrix is simplified as graph Laplacian. Thus, $\mathbf{d}_t$ can be viewed as GMRF with respect to graph $\mathcal{G}$, which establishes a connection between temporal dynamics and the spatial structure. In other words, a probabilistic interpretation of the graph learning problem can be summarized as recovering the GMRF model from temporal evolution of graph signals.

In the presence of noise, we see from Eq. (12) that, graph topology recovery can be viewed as learning the principal components of the covariance of $\mathbf{d}_t$ under a Gaussian prior of $\mathbf{v}_t$. Different from smooth graph-based model, the proposed model characterizes the space-time property, which is proved to favor the spatiotemporal smoothness in Section IV, thereby extending the static signal analysis to the time-vertex one. More importantly, such a dynamic representation has the capability to deal with prediction tasks in the realistic sensor network, which will be shown in the simulation part.

B. SPATIOTEMPORAL SMOOTHNESS PROPERTY

With special consideration of space-time interactions in signal representation, we now describe and evaluate the temporal variation of graph signals.

Definition 1 (Weighted Difference Operator): Temporal evolution of observation $\mathbf{X}$ *can be expressed by a weighted difference operator* $\mathcal{D}(\mathbf{X}) = \mathbf{X} - \mathbf{R}\mathbf{X}\mathbf{B}$, *where* $\mathbf{R}$ *is the time correlation matrix and* $\mathbf{B}$ *is the shift operator defined as*

$$B = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix}_{M \times M}, \tag{13}$$

and weighted difference signal equals to

$$\mathcal{D}(X) = [x_1, x_2 - Rx_1, x_3 - Rx_2, \ldots, x_M - Rx_{M-1}].$$

It can be seen in *Definition* 1, there exists multiple interpretations for given different $\mathbf{R}$, e.g., when $\mathbf{R} = \mathbf{I}_N$, the temporal difference of graph signals $\mathbf{x}_t - \mathbf{x}_{t-1}$ can be obtained. As an extreme case, when given time correlation $\mathbf{R} = \mathbf{0}$, weighted difference signal degenerates to the original signal. Incorporating the notion of time dependence, the smoothness property evaluated in both graph vertex and time domain can be mathematically generalized from the Eq. (2).

Definition 2 (Spatiotemporal Smoothness): It measures the total variation in weighted difference of graph signal $\mathbf{X}$ *with respect to the graph structure.*

$$\sum_{t=1}^{M} \mathbf{d}_t{}^T \mathbf{L}\mathbf{d}_t = \mathrm{tr}\left(\mathcal{D}(\mathbf{X})^T \mathbf{L}\mathcal{D}(\mathbf{X})\right). \tag{14}$$

Spatiotemporal smoothness in (14) generalizes the graph Laplacian quadratic form (2) via defining the weighted difference operator for measuring both temporal and spatial smoothness. The small value of (14) implies that graph signal varies smoothly on the graph over time. To be noted, the trace term encodes the information on spatial and temporal structures of $\mathbf{X}$ in graph Laplacian $\mathbf{L}$ and $\mathbf{R}$, respectively. Although spatial smoothness has been widely used as a meaningful prior, it does not emphasize the temporal dynamics. This is where the time correlation matrix compensates. Moreover, according to the finding in [11], weighted difference signals, instead of signals themselves, exhibit better smoothness property which can be promoted through learning procedure for an accurate graph topology inference. The improvement

in graph learning accuracy is presented in the experimental section.

IV. GRAPH LEARNING METHOD BASED ON CORRELATION AND SPATIOTEMPORAL SMOOTHNESS (CASTS)

In this section, we propose a graph learning method by introducing the spatiotemporal smoothness prior to the field of time-vertex signal analysis and jointly applying the correlated property of time-varying graph signals. In Subsection A, we formulate the graph learning problem as a nonconvex optimization problem. After which, an optimization algorithm to the proposed problem, CASTS, is presented in Subsection B based on block coordinate decent scheme. We also provide convergence analysis to the proposed CASTS.

A. PROBLEM FORMULATION

As mentioned in Section III-A, time correlation of graph signals is represented by the matrix $\mathbf{R}$ with correlation coefficient of the observation site on its corresponding diagonal position. In practice, prior information of correlation pattern is unknown or not accurate enough due to the significant data loss. For example, data from adjacent sensor nodes in adjacent time slots are dropped together due to the communication congestion or hardware conditions (e.g., sensors are damaged or run out of energy) [38]. Hence, in this section, we study the case that the correlation matrix $\mathbf{R}$ is unknown.

According to (10), given the weighted difference signal $\mathbf{d}_t$ and the distribution of $\mathbf{v}_t$, the maximum a posteriori (MAP) estimation of $\mathbf{v}_t$ by applying Bayes' rule is written as

$$\max_{\mathbf{v}_t} p\left(\mathbf{v}_t|\mathbf{d}_t\right) \propto p\left(\mathbf{d}_t|\mathbf{v}_t\right) p\left(\mathbf{v}_t\right). \tag{15}$$

Specifically, from the multivariate Gaussian distribution shown in (9) and (11), the posterior probability of process variable $\mathbf{v}_t$ is proportional to (16)

$$\log\left(p\left(\mathbf{d}_t|\mathbf{v}_t\right) p\left(\mathbf{v}_t\right)\right) = \log p_E\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right) + \log p_V\left(\mathbf{v}_t\right), \tag{16}$$

where $p_V\left(\mathbf{v}_t\right)$ represent probability density function (p.d.f.) of $\mathbf{v}_t$, $p\left(\mathbf{d}_t|\mathbf{v}_t\right)$ is the conditional p.d.f. of $\mathbf{d}_t$ given $\mathbf{v}_t$, and $p_E\left(\mathbf{d}_t - \mathbf{U}\boldsymbol{\delta}_t\right) = p_E\left(\mathbf{n}_t - \mathbf{R}\mathbf{n}_{t-1}\right)$. According to (16) and Gaussian probability distributions in (9) and (11), the MAP estimate of (15) can be expressed as

$$\begin{aligned}
&\mathbf{v}_{t\,MAP}\left(\mathbf{d}_t\right): \\
&= \arg\min_{\mathbf{v}_t} 2\left(-\log p_E\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right) - \log p_V\left(\mathbf{v}_t\right)\right) \\
&= \arg\min_{\mathbf{v}_t} \left(-2\log e^{-\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right)^T \mathbf{W}^{-1}\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right)} - \alpha \log e^{-\mathbf{v}_t{}^T \Lambda \mathbf{v}_t}\right) \\
&= \arg\min_{\mathbf{v}_t} 2\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right)^T \mathbf{W}^{-1}\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right) + \alpha \mathbf{v}_t{}^T \Lambda \mathbf{v}_t
\end{aligned} \tag{17}$$

where scalable value (i.e., 2 in this case) is introduced for the following approximation, $\mathbf{W} = \mathbf{I}_N + \mathbf{R}\mathbf{R}^T$ and α is a constant parameter proportional to the variance of noise $\sigma_n{}^2$.

As we can see, (17) is a difficult problem as the objective function involves the inverse of $\mathbf{W}$, which is hard to process in the case of unknown $\mathbf{R}$. To obtain an approximation of such problem that is easier to solve, we introduce the following inequality as

$$\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right)^T \mathbf{W}^{-1}\left(\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right) \geq \lambda_{min}\left(\mathbf{W}^{-1}\right) \left\|\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right\|_2^2, \tag{18}$$

where λ_{min} indicates the minimum eigenvalue of matrix. As $0 \leq c \leq 1$, the inequality $0 \leq c^2 \leq 1$ is satisfied. Due to the diagonal matrix of $\mathbf{W}$, the minimum eigenvalue of $\mathbf{W}^{-1}$ is $\lambda_{min} = \frac{1}{2}$ when c^2 equals to one. Next, taking advantage of the above property, the relaxation of the minimization problem in (17) can be denoted as

$$\min_{\mathbf{v}_t} \left\|\mathbf{d}_t - \mathbf{U}\mathbf{v}_t\right\|_2^2 + \alpha \mathbf{v}_t{}^T \Lambda \mathbf{v}_t. \tag{19}$$

Notice that in (19) the representation matrix $\mathbf{U}$ and the precision matrix Λ of the Gaussian distribution on $\mathbf{v}_t$ come from the graph Laplacian matrix. When the graph and correlation matrix are unknown, Eq. (19) can be written as a joint optimization problem of $\mathbf{L}$, Ω and $\mathbf{R}$ in a matrix form

$$\min_{\mathbf{L},\boldsymbol{\Omega},\mathbf{R}} \left\|\mathcal{D}(\mathbf{X} - \boldsymbol{\Omega})\right\|_F^2 + \alpha \mathrm{tr}\left(\mathcal{D}\left(\boldsymbol{\Omega}\right)^T \mathbf{L} \mathcal{D}\left(\boldsymbol{\Omega}\right)\right), \tag{20}$$

where $\boldsymbol{\Omega} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_M]$ is the main component of graph signal $\mathbf{X}$, with mth element denoted as $\boldsymbol{\omega}_m = \mathbf{U}\mathbf{h}_m$. The first term can be read as denoising inference which finds Ω close to observation $\mathbf{X}$. As proved in the following *Theorem*, the second term is considered as the spatiotemporal smoothness of Ω which is promoted by the proposed model.

Theorem 1: Assume graph signal $\mathbf{x}_t$ *and latent variable* $\boldsymbol{h}_t$ *follow the dynamic graph-based model in (6) and (7) with process variable* $\mathbf{v}_t \sim \mathcal{N}\left(0, \boldsymbol{\Lambda}^{\dagger}\right)$*, the proposed model favors spatiotemporal smoothness of the main component* $\boldsymbol{\Omega}$ *in graph signals, which can be expressed as*

$$\sum_{t=1}^{M} \mathbf{v}_t{}^T \Lambda \mathbf{v}_t = \mathrm{tr}\left(\mathcal{D}\left(\boldsymbol{\Omega}\right)^T \mathbf{L} \mathcal{D}\left(\boldsymbol{\Omega}\right)\right).$$

Proof: In a noisy case, where $\mathbf{x}_t = \boldsymbol{\omega}_t + \mathbf{n}_t$, the second term in minimization problem of (19) can be formulated as

$$\begin{aligned}
\sum_{t=1}^{M} \mathbf{v}_t{}^T \Lambda \mathbf{v}_t &= \sum_{t=1}^{M} \left(\mathbf{h}_t - \mathbf{A}\mathbf{h}_{t-1}\right)^T \Lambda \left(\mathbf{h}_t - \mathbf{A}\mathbf{h}_{t-1}\right) \\
&= \sum_{t=1}^{M} \left(\boldsymbol{\omega}_t - \mathbf{R}\boldsymbol{\omega}_{t-1}\right)^T \mathbf{U}\Lambda\mathbf{U}^T \left(\boldsymbol{\omega}_t - \mathbf{R}\boldsymbol{\omega}_{t-1}\right) \\
&= \mathrm{tr}\left(\mathcal{D}\left(\boldsymbol{\Omega}\right)^T \mathbf{L} \mathcal{D}\left(\boldsymbol{\Omega}\right)\right).
\end{aligned}$$

Based on the *Definition* 2, the above expression denotes the spatiotemporal smoothness of the main component $\boldsymbol{\Omega}$ in observation $\mathbf{X}$, which is enforced by minimization procedure. A similar observation can be derived as well in a noiseless scenario. Recalling the smooth property of graph signal introduced in Section III-B, we have that the proposed model promotes spatiotemporal smoothness of graph signal in both noisy and noiseless cases. Hence, we complete the proof. ■

Having given the above analysis, we will now formally state the problem of interest.

Problem 1: Given the observation $\mathbf{X}$ *that are related to the unknown main component* $\mathbf{\Omega}$*, determine the valid graph Laplacian* $\mathbf{L}$ *and temporal correlation* $\mathbf{R}$ *such that the estimation* $\hat{\mathbf{\Omega}}$ *has the smallest estimation error, and meanwhile* $\hat{\mathbf{\Omega}}$ *is spatiotemporally smooth on the recovered graph.*

Mathematically, by adding additional constraints on Laplacian and time correlation, we propose to solve the problem (20) with the following objective function $Q_1(L, \Omega, R)$:

$$\begin{aligned}(\text{P1}) \quad & \min_{L,\Omega,R} Q_1(L,\Omega,R) \\ & \text{s.t. } Q_1(L,\Omega,R) = \|\mathcal{D}(\mathbf{\Omega}-X)\|_F^2 \\ & \quad + \alpha \text{tr}\left(\mathcal{D}(\mathbf{\Omega})^T L \mathcal{D}(\mathbf{\Omega})\right) \\ & \quad + \beta \|L\|_F^2 + \gamma \|R\|_F^2, \\ & \quad L \in \mathcal{L}^N, \quad \text{tr}(L) = N, \\ & \quad \|R\|_1 \le 1, \quad \text{std}(\text{diag}(R)) \le \kappa,\end{aligned}$$

where α, β and γ are positive regularization parameters corresponding to the three regularization terms. The trace term induces the spatiotemporal smoothness discussed in *Definition* 2, and the first Frobenius norm term controls the off-diagonal entries in $\mathbf{L}$, which affects the edge weight of graph. The second Frobenius norm controls the value in diagonal position of $\mathbf{R}$. These three terms coincide with each other since they both favor a sparse graph structure where weighted difference signals are smooth. Furthermore, the first constraint on $\mathbf{L}$ guarantees a valid Laplacian matrix and the other trace constraint is imposed to avoid the trivial solution, while the second constraints on $\mathbf{R}$ limit the value and the variation of diagonal entries in $\mathbf{R}$. The threshold κ is set to control the variation of correlation coefficients among different observations and the value depends on the property of application data.

B. OPTIMIZATION ALGORITHM

The optimization problem (P1) is not jointly convex in $\mathbf{L}$, $\mathbf{\Omega}$ and $\mathbf{R}$. Therefore, CASTS is proposed to solve the above non-convex problem through a block coordinate descent scheme where, at each step, we optimize one block of coordinate directions while holding all other blocks constant. The iteration is shown as follows

$$\begin{aligned}1.\ & \hat{\mathbf{L}}^k \triangleq \arg\min_{\mathbf{L}} Q_1\left(\mathbf{L}, \hat{\Omega}^{k-1}, \hat{\mathbf{R}}^{k-1}\right), \quad (\mathcal{S}_L) \\ & \quad \text{s.t.} \mathbf{L} \in \mathcal{L}^N, \text{tr}(\mathbf{L}) = N. \\ 2.\ & \hat{\Omega}^k \triangleq \arg\min_{\mathbf{\Omega}} Q_1\left(\hat{\mathbf{L}}^k, \Omega, \hat{\mathbf{R}}^{k-1}\right). \quad (\mathcal{S}_\Omega) \\ 3.\ & \hat{\mathbf{R}}^k \triangleq \arg\min_{\mathbf{R}} Q_1\left(\hat{\mathbf{L}}^k, \hat{\Omega}^k, \mathbf{R}\right), \quad (\mathcal{S}_R) \\ & \quad \text{s.t.} \|\mathbf{R}\|_1 \le 1, \quad \text{std}(\text{diag}(\mathbf{R})) \le \kappa.\end{aligned}$$

It is interesting to find that the problems of $(\mathcal{S}_L)$ and $(\mathcal{S}_R)$ can be cast as constrained convex optimization problems, and $(\mathcal{S}_\Omega)$ is an unconstrained one. By alternating among the three steps, we can get the final solution of (P1). The details are summarized in *Algorithm* 1.

Algorithm 1 : The Procedure of CASTS

Input: $\mathbf{X}$, α, β, κ, stopping criterion.

1: Initialization: $\mathbf{\Omega}^0 = \mathbf{X}$, $\mathbf{R}^0 = \mathbf{I}$, $k = 1$;
2: **repeat**
3: 1) Update $\mathbf{L}^k$ by (24)
4: 2) Update $\mathbf{\Omega}^k$ by (25) or *Algorithm* 2
5: 3) Update $\mathbf{R}^k$ by [42]
6: $k = k + 1$;
7: **until** Stopping criterion satisfied.

Output: Recovered signal $\mathbf{\Omega}$, learned graph $\mathbf{L}$ and $\mathbf{R}$.

1) COMPUTATION OF THE GRAPH TOPOLOGY $\hat{\mathbf{L}}^k$

As we can see, the $(\mathcal{S}_L)$ is a strictly convex optimization problem, since the Hessian matrix of the objective function is $2\beta\mathbf{I}_N$ which is positive definite. Besides, it is also a quadratic program of $\mathbf{L}$ subject to Laplacian constraints, which can be solved via alternating direction method of multipliers (ADMM) [39]. We reformulate the problem $(\mathcal{S}_L)$ as

$$\begin{aligned}& \min_{\mathbf{L}} \alpha \text{tr}\left(\mathcal{D}(\mathbf{\Omega})^T \mathbf{L} \mathcal{D}(\mathbf{\Omega})\right) + \beta \|\mathbf{L}\|_F^2, \\ & \text{s.t. } \mathbf{L} - \mathbf{Z} = 0, \\ & \quad \mathbf{Z} \in \mathcal{L}^*\end{aligned} \tag{21}$$

where $\mathbf{Z}$ is the auxiliary variables and $\mathcal{L}^*$ is denoted as

$$\mathcal{L}^* = \left\{\mathbf{L} | \mathbf{L} \succeq 0, L_{ji} = L_{ij} \le 0, i \ne j, \text{and} \mathbf{L} \cdot \mathbf{1} = \mathbf{0}, \text{tr}(\mathbf{L}) = N\right\}. \tag{22}$$

Since the graph Laplacian in (1) form a convex set and $\text{tr}(\cdot)$ is a linear function, the overall $\mathcal{L}^*$ is also a convex set. Then the augmented Lagrangian of (21) is

$$\begin{aligned}L_\rho(\mathbf{L}, \mathbf{Z}, \mathbf{P}) = & \alpha \text{tr}\left(\mathcal{D}(\mathbf{\Omega})^T \mathbf{L} \mathcal{D}(\mathbf{\Omega})\right) + \beta \|\mathbf{L}\|_F^2, \\ & + \frac{\rho}{2} \|\mathbf{Z} - \mathbf{L}\|_F^2 + \langle \mathbf{P}, \mathbf{Z} - \mathbf{L} \rangle,\end{aligned} \tag{23}$$

where $\mathbf{P}$ is the Lagrange multiplier and $\langle \cdot, \cdot \rangle$ is the inner product of matrices. As proved in [39], we use the following formulas to update $\mathbf{L}$, $\mathbf{Z}$ and $\mathbf{P}$ to find a saddle point for (23)

$$\begin{aligned}\mathbf{L}^{k+1} & := \frac{\rho \mathbf{Z}^k + \mathbf{P}^k - \alpha \mathcal{D}(\mathbf{\Omega}) \mathcal{D}(\mathbf{\Omega})^T}{2\beta + \rho}, \\ \mathbf{Z}^{k+1} & := \prod_{\mathcal{L}^*} \left(\mathbf{L}^{k+1} - \frac{1}{\rho} \mathbf{P}^k\right), \\ \mathbf{P}^{k+1} & := \mathbf{P}^k + \rho\left(\mathbf{Z}^{k+1} - \mathbf{L}^{k+1}\right),\end{aligned} \tag{24}$$

where $\rho > 0$ is the Lagrangian parameter and $\prod_{\mathcal{L}^*}$ is the Euclidean projection onto the set $\mathcal{L}^*$.

2) COMPUTATION OF THE RECOVERED SIGNAL $\hat{\omega}^k$

The optimal update for matrix Ω^k is provided in the following proposition.

Proposition 1: For the given graph Laplacian and time correlation matrix, $(\mathcal{S}_\Omega)$ *is an unconstrained and strictly*

convex optimization problem that admits a closed-form solution

$$\mathrm{vec}\,(\mathbf{\Omega}) = \left[\mathbf{T}_d\mathbf{T}_d{}^T + \alpha\mathbf{T}_d\,(\mathbf{I}_M \otimes \mathbf{L})\,\mathbf{T}_d{}^T\right]^{-1}\mathbf{T}_d\mathbf{T}_d{}^T\,\mathrm{vec}\,(\mathbf{X})\,. \tag{25}$$

with

$$\mathbf{T}_d = \begin{bmatrix} \mathbf{I}_N & -\mathbf{R} & & & \\ & \mathbf{I}_N & -\mathbf{R} & & \\ & & \mathbf{I}_N & \ddots & \\ & & & \ddots & -\mathbf{R} \\ & & & & \mathbf{I}_N \end{bmatrix}_{NM\times NM}.$$

Proof: Denoting the objective function in $(\mathcal{S}_\Omega)$ as $f_2(\cdot)$, we first verify the convexity by judging whether it satisfies second-order convexity conditions. For the convenience of analysis in the following derivation, we introduce the property of the vectorization operator, that is

$$\mathrm{vec}\,(\mathbf{AXB}) = \left(\mathbf{B}^T \otimes \mathbf{A}\right)\mathrm{vec}\,(\mathbf{X})\,.$$

Then we have

$$\begin{aligned}
&\mathrm{tr}\left(\mathcal{D}(\mathbf{\Omega})^T\mathbf{L}\mathcal{D}(\mathbf{\Omega})\right)\\
&= \mathrm{vec}\,(\mathbf{\Omega}-\mathbf{R\Omega B})^T\,\mathrm{vec}\,[\mathbf{L}(\mathbf{\Omega}-\mathbf{R\Omega B})]\\
&= \left[\mathrm{vec}\,(\mathbf{\Omega})^T - \mathrm{vec}\,(\mathbf{\Omega})^T(\mathbf{B}\otimes\mathbf{R})\right]\cdot\\
&\quad\times\left[(\mathbf{I}_M\otimes\mathbf{L})\,\mathrm{vec}\,(\mathbf{\Omega}) - \left(\mathbf{B}^T\otimes\mathbf{LR}\right)\mathrm{vec}\,(\mathbf{\Omega})\right]\\
&= \mathrm{vec}\,(\mathbf{\Omega})^T\left[(\mathbf{I}_M\otimes\mathbf{I}_N) - (\mathbf{B}\otimes\mathbf{R})\right]\cdot\\
&\quad\times\left[(\mathbf{I}_M\otimes\mathbf{L}) - \left(\mathbf{B}^T\otimes\mathbf{LR}\right)\right]\mathrm{vec}\,(\mathbf{\Omega})\\
&= \mathrm{vec}\,(\mathbf{\Omega})^T\,\mathbf{T}_d\,(\mathbf{I}_M\otimes\mathbf{L})\left[(\mathbf{I}_M\otimes\mathbf{I}_N) - \left(\mathbf{B}^T\otimes\mathbf{R}\right)\right]\mathrm{vec}\,(\mathbf{\Omega})\\
&= \mathrm{vec}\,(\mathbf{\Omega})^T\,\mathbf{T}_d\,(\mathbf{I}_M\otimes\mathbf{L})\,\mathbf{T}_d{}^T\mathrm{vec}\,(\mathbf{\Omega})\,.
\end{aligned}$$

Similarly, the first term in f_2 can be denoted as

$$\begin{aligned}
\|\mathcal{D}(\mathbf{\Omega}-\mathbf{X})\|_F^2 &= \mathrm{tr}\left(\mathcal{D}(\mathbf{\Omega}-\mathbf{X})^T\mathcal{D}(\mathbf{\Omega}-\mathbf{X})\right)\\
&= \mathrm{vec}\,(\mathbf{\Omega}-\mathbf{X})^T\,\mathbf{T}_d\mathbf{T}_d{}^T\mathrm{vec}\,(\mathbf{\Omega}-\mathbf{X})\,,
\end{aligned}$$

and problem $(\mathcal{S}_\Omega)$ can be equivalently written as

$$\min_{\mathbf{y}}\left(\mathbf{y}^T - \mathrm{vec}(\mathbf{X})^T\right)\mathbf{T}_d\mathbf{T}_d{}^T\,(\mathbf{y}-\mathrm{vec}\,(\mathbf{X})) + \alpha\mathbf{y}^T\mathbf{G}\mathbf{y} = \tilde{f}_2(\mathbf{y})\,,$$

where $\mathbf{G} = \mathbf{T}_d\,(\mathbf{I}_M\otimes\mathbf{L})\,\mathbf{T}_d{}^T \in \mathbb{R}^{NM\times NM}$, and $\mathbf{y} = \mathrm{vec}\,(\mathbf{\Omega})$. The gradient of $\tilde{f}_2(\mathbf{y})$ can be deduced as

$$\nabla\tilde{f}_2(\mathbf{y}) = 2\mathbf{T}_d\mathbf{T}_d{}^T\mathbf{y} - 2\mathbf{T}_d\mathbf{T}_d{}^T\mathrm{vec}\,(\mathbf{X}) + 2\alpha\mathbf{G}\mathbf{y}.$$

Then the Hessian matrix of function $\tilde{f}_2(\mathbf{y})$ can be derived as

$$\nabla^2\tilde{f}_2 = 2\mathbf{T}_d\mathbf{T}_d{}^T + 2\alpha\mathbf{G}.$$

Observing that $\mathbf{G}$ is positive semidefinite due to the positive semidefinite matrix $\mathbf{L}$ and $\mathbf{I}_M\otimes\mathbf{L}$. Since $\mathbf{T}_d{}^T$ is a invertible matrix according to its definition, for any nonzero matrix $\mathbf{x}$, $\mathbf{T}_d{}^T\mathbf{x} \neq 0$. Then, $\mathbf{T}_d\mathbf{T}_d{}^T$ is positive definite, and thus Hessian matrix of function $\tilde{f}_2$ is positive definite, which confirms the strictly convexity of the problem $(\mathcal{S}_\Omega)$. Next, by setting $\nabla\tilde{f}_2(\mathbf{y})$ to zero, the unique optimal solution $\mathrm{vec}\,(\mathbf{\Omega})$ can be obtained as (25). Then, the Proposition 1 is proved. ■

However, the unique solution involves calculating the inverse of a matrix, which is computationally expensive. The conjugate gradient method [40] can be applied to deal with such problem iteratively. In each iteration, it determines the dynamic stepsize and updates the next searching directions. In particular, due to the quadratic property of the function $f_2(\mathbf{\Omega})$, the optimal stepsize at the mth step can be decided by exact line search [41] given as

$$\min_\tau f_2\left(\mathbf{\Omega}^m + \tau\Delta\mathbf{\Omega}^m\right).$$

where τ and $\Delta\mathbf{\Omega}^m$ are the stepsize and the search direction of the mth step, respectively. By taking derivative of τ and set to zero, we have

$$0 = \frac{\partial f_2\left(\Omega^m + \tau\Delta\Omega^m\right)}{\partial\tau} = \mathrm{tr}\left[\left(\Delta\Omega^m\right)^T\nabla f_2\left(\Omega^m + \tau\Delta\Omega^m\right)\right],$$

where the gradient of the function $f_2(\mathbf{\Omega})$ is

$$\begin{aligned}
\nabla f_2(\Omega) &= \gamma\Omega - \mathbf{X} - (\gamma\mathbf{R}\Omega - \mathbf{RX})\,\mathbf{B}\\
&\quad - (\mathbf{R}\gamma\Omega - \mathbf{RX})\,\mathbf{B}^T + (\mathbf{R}\gamma\mathbf{R}\Omega - \mathbf{RRX})\,\mathbf{B}\mathbf{B}^T,
\end{aligned} \tag{26}$$

where $\gamma = \alpha\mathbf{L} + \mathbf{I}_N$. As a consequence, the optical stepsize can be obtained with the Fletcher-Reeves parameter given as $\theta = \left\|\nabla f_2\left(\mathbf{\Omega}^{m+1}\right)\right\|_F^2 \Big/ \left\|\nabla f_2\left(\mathbf{\Omega}^m\right)\right\|_F^2$. The detailed procedure of the algorithm is listed in *Algorithm* 2.

Algorithm 2 : Method for Solving Subproblem $(\mathcal{S}_\Omega)$

Input: $\mathbf{X}$, $\mathbf{R}^{k-1}$, $\mathbf{L}^k$, α, β, K, error tolerance δ.

1: Initialization: $\mathbf{\Omega}^0 = \mathbf{0}$; $\Delta\mathbf{\Omega}^0 = -\nabla f_2\left(\mathbf{\Omega}^0\right)$; $m = 0$;
2: **repeat**
3: 1) Dynamic stepsize selection:
4: $\tau = -\dfrac{\mathrm{tr}\left\{(\Delta\mathbf{\Omega}^m)^T\nabla f_2(\mathbf{\Omega}^m)\right\}}{\mathrm{tr}\left\{(\Delta\mathbf{\Omega}^m)^T\left[\nabla f_2(\Delta\mathbf{\Omega}^m) + 2\mathcal{D}(\mathbf{X}) - 2\mathbf{R}\mathcal{D}(\mathbf{X})\mathbf{B}^T\right]\right\}}$;
5: 2) Conjugate direction update:
6: $\mathbf{\Omega}^{m+1} = \mathbf{\Omega}^m + \tau\Delta\mathbf{\Omega}^m$;
7: $\theta = \left\|\nabla f_2\left(\mathbf{\Omega}^{m+1}\right)\right\|_F^2 \Big/ \left\|\nabla f_2\left(\mathbf{\Omega}^m\right)\right\|_F^2$;
8: $\Delta\mathbf{\Omega}^{m+1} = -\nabla f_2\left(\mathbf{\Omega}^{m+1}\right) + \theta\Delta\mathbf{\Omega}^m$;
9: $m = m + 1$;
10: **until** $m = K$ or $\left\|\Delta\mathbf{\Omega}^m\right\|_F \leq \delta$

Output: Recovered signal $\mathbf{\Omega}$.

3) COMPUTATION OF THE TIME CORRELATION $\hat{\mathbf{R}}^k$

Having obtained the $\mathbf{L}$ and Ω in previous steps, we now detect the correlation coefficient of each observation site, denoted by diagonal elements in $\mathbf{R}$, based on Proposition 2.

Proposition 2: For detecting the correlation matrix $\mathbf{R}$*, the problem* $(\mathcal{S}_R)$ *is strictly convex with two convex constraints.*

Proof: Being prepared for the following analysis, we utilize the function $\tilde{f}_3(\mathrm{vec}\,(\mathbf{R})) = f_3(\mathbf{R})$ with $\mathbf{z} = \mathrm{vec}\,(\mathbf{R})$. Similar to the derivation in Proposition 1 by

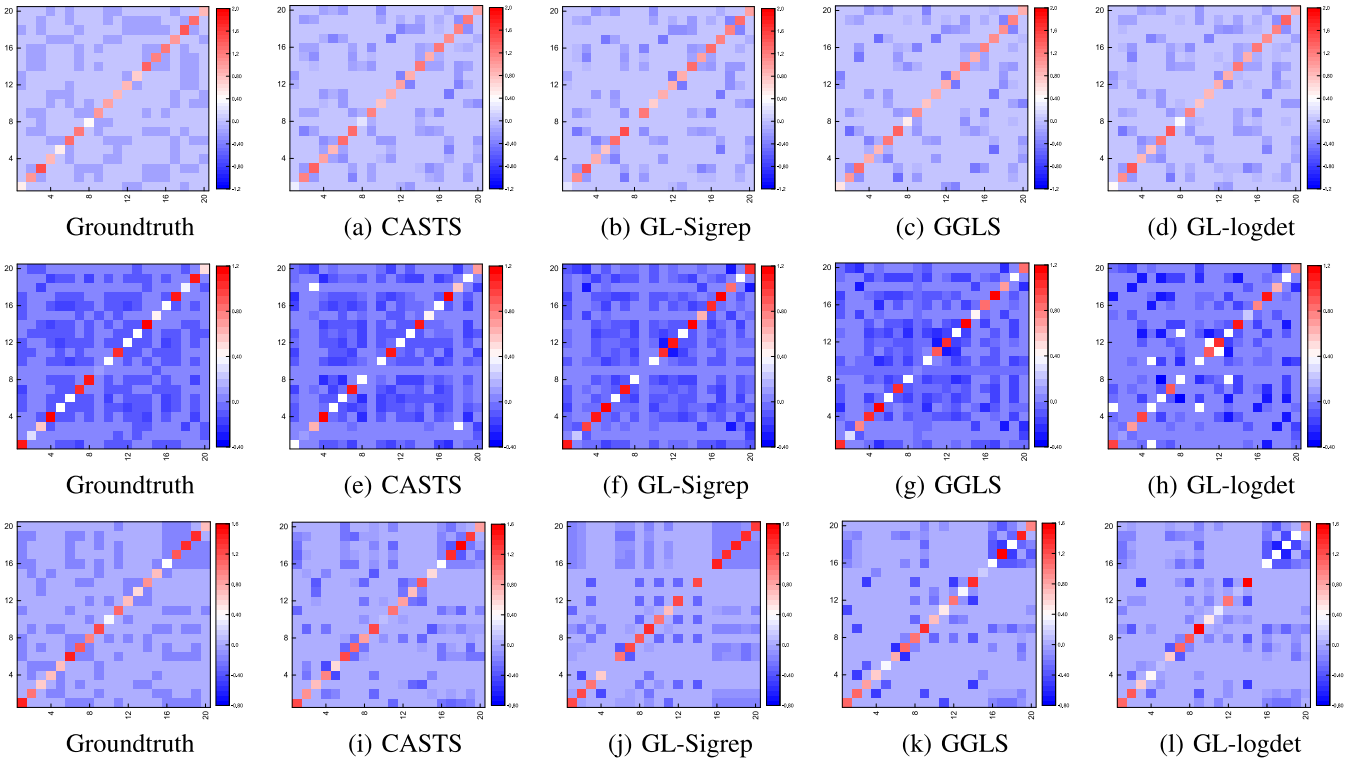

FIGURE 2. The learned graph Laplacian matrices for the random geometric graph $\mathcal{G}_{RGG}$. The columns from the left to the right are the groundtruth Laplacian, the Laplacians recovered by CASTS, GL-Sigrep, GGLS and GL-logdet. The rows from the top to the bottom are the learning results for different types of graph signal in case(i), case(ii) and case(iii), respectively.

exploiting the properties of vec-operator, the gradient of $\tilde{f}_3$ can be easily written as

$$\begin{aligned}\nabla\tilde{f}_3(\mathbf{z}) &= 2\left(\mathbf{\Omega}\mathbf{B}\otimes\mathbf{I}_N - \mathbf{X}\mathbf{B}\otimes\mathbf{I}_N\right)\left(\mathrm{vec}\left(\mathbf{X}\right)-\mathrm{vec}\left(\mathbf{\Omega}\right)\right)\\ &\quad+2\left[\left(\left(\mathbf{\Omega}-\mathbf{X}\right)\mathbf{B}\mathbf{B}^T\left(\mathbf{\Omega}-\mathbf{X}\right)^T\right)\otimes\mathbf{I}_N\right]\mathbf{z}\\ &\quad-\alpha\left(\mathbf{\Omega}\mathbf{B}\otimes\mathbf{L}\right)\mathrm{vec}\left(\mathbf{\Omega}\right)\\ &\quad-\alpha\left(\mathbf{\Omega}\mathbf{B}\otimes\mathbf{I}_N\right)\mathrm{vec}\left(\mathbf{L}\mathbf{\Omega}\right)+2\alpha\left(\mathbf{\Omega}\mathbf{B}(\mathbf{\Omega}\mathbf{B})^T\otimes\mathbf{L}\right)\mathbf{z}+2\gamma\mathbf{z}.\end{aligned}$$

Further, the Hessian matrix of objective function $\tilde{f}_3$ is

$$\begin{aligned}\nabla^2\tilde{f}_3 = 2\left(\left(\mathbf{\Omega}-\mathbf{X}\right)\mathbf{B}\mathbf{B}^T\left(\mathbf{\Omega}-\mathbf{X}\right)^T\right)\otimes\mathbf{I}_N\\ +2\alpha\left(\mathbf{\Omega}\mathbf{B}(\mathbf{\Omega}\mathbf{B})^T\otimes\mathbf{L}\right)+2\gamma\mathbf{I}_{N^2}.\end{aligned}$$

According to the definition and properties of positive semidefinite matrix, we have that$(\mathbf{\Omega}-\mathbf{X})\,\mathbf{B}\mathbf{B}^T(\mathbf{\Omega}-\mathbf{X})^T$ and $\mathbf{\Omega}\mathbf{B}(\mathbf{\Omega}\mathbf{B})^T\otimes\mathbf{L}$ are both positive semi-definite. As the last term $2\gamma\mathbf{I}_{N^2}$ is positive definite, the Hessian matrix of $\tilde{f}_3$ is positive definite, then we have $\tilde{f}_3$ is strictly convex, so it true for objective function f_3. ■

Since the estimation matrix $\mathbf{R}$ is diagonal in a constrained convex optimization problem $(\mathcal{S}_R)$ proved above, it is efficient to recover the diagonal elements in $\mathbf{R}$ via the convex optimization package CVX [42]. Finally, by solving the three subproblems in (P1) alternatively, we can get the final optimal solution within a few iterations. The stopping criterion could be a maximum number of iterations K, or the change of the objective function Q_1 less than a threshold.

Convergence analysis: In minimization of a differentiable function over a convex set and each subproblem attains a unique minimization, a block coordinate descent algorithm guarantees convergence to a stationary point [43], [44]. As shown in (P1), both set of $\mathbf{L}$ and $\mathbf{R}$ are convex set. Also, the objective function is continuously differentiable over such convex set. In addition, converge conditions (see Proposition 2.7.1 in [44]) require that the minimization of each block-coordinate update is uniquely attained. All the subproblems of (P1) are proved to be strictly convex, hence there exists at most one global minimum in each subproblems (see Proposition 3.1.1 in [43]). Since the whole convergence conditions are satisfied, the CASTS guarantees the convergence to the final solution. Moreover, due to the existence of three subproblems in the optimization procedure, it takes more iterations in our algorithm to achieve overall stopping criterion, which remains a future study to further reduce the time complexity.

V. EXPERIMENTAL RESULTS

To test the graph learning performance of the proposed method, we conduct experiments on three synthetic datasets with different correlation patterns and two real-world datasets, including the China daily temperature dataset

from National Oceanic and Atmospheric Administration (NOAA) [50] and the California daily evaporation dataset from California Department of Water Resources [51].

The proposed CASTS is compared with four baseline graph learning algorithms, including GL-Sigrep [21], GGLS [22], SpecTemp [29] and GL-logdet [20], to identify the graph Laplacian matrix. To make a fair comparison, we perform Monte-Carlo simulations for each method to find the best combination of regularization parameters that maximize the performance. More specifically, α, β and γ in (P1) are selected from powers of 10 ranging from -1 to -3 with a stepsize of 0.1, 0 to -2 with a stepsize of 0.05 and 0 to -2 with a stepsize of 0.1, respectively. Since temperature and evapotranspiration data vary based on consecutive changes on physics or chemistry, we heuristically take $\kappa = 0.1$.

Synthetic Datasets: In each experiment, we create several synthetic datasets based on different graph topologies and time correlation patterns. First, we consider a 20-vertex undirected weighted graph in which graph model is chosen from the following three options:

1) K-neighbor graph, $\mathcal{G}_{KN}$, with each vertex connected to its three nearest neighbors, where the coordinate is created randomly and the weight of each edge is inversely proportional to the distance between two vertices.
2) Random geometric graph, $\mathcal{G}_{RGG}$, with coordinate of vertex generated uniformly at random in the unit square and the edge weight is determined by Gaussian function $W(i,j) = \exp\left(-\frac{d(i,j)^2}{2\sigma^2}\right)$ where $\sigma = 0.5$, then threshold weights< 0.7.
3) Random scale-free graph, $\mathcal{G}_{RSF}$, with the probability of the new vertex connected to an existing vertex follows a preferential attachment criterion proposed in [45].

Given a specific graph structure, the CGL matrix is calculated and normalized according to constraints in (P1). Then, we generate 100 time series based on the proposed model shown in (12) with random initialization of $\mathbf{x}_1$ and standard deviation of noise $\sigma_n = 0.5$. Without loss of the generality, we create three types of graph signals with varying correlation pattern i) low correlated graph signal with correlation matrix $\mathbf{R} = 0.2 \times \mathbf{I}_{20}$, ii) high correlated graph signals with $\mathbf{R} = 0.9 \times \mathbf{I}_{20}$. iii) graph signal where diag $(\mathbf{R})$ is generated from Gaussian distribution with 0.5 mean and 0.01 variance.

Experimental settings: To show the graph learning performance, we create various experimental scenarios by choosing a time correlation pattern under different graph models, and provide both visual and quantitative comparison. Particularly, we average the results over 10 random instances of three graphs with the selected type of graph signal $\mathbf{X}$. For quantitative comparison, we evaluate the performance of all algorithms in terms of the accuracy of edge weight and the recovery of the edge position in the groundtruth graph by using the following metrics,

$$RWE\left(\hat{\mathbf{L}}, \mathbf{L}\right) = \frac{\left\|\hat{\mathbf{L}} - \mathbf{L}\right\|_F}{\|\mathbf{L}\|_F}, \tag{27}$$

which is the relative weight error between the groundtruth graph $\mathbf{L}$ and learned graph $\hat{\mathbf{L}}$, and

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{28}$$

tests the overall accuracy of the learned edge set $\widehat{c}$ with respect to the groundtruth edge set c, which involves both *Precision* and *Recall* parameters denoted as $\#\left\{(i,j) : \widehat{c}_{i,j} \neq 0, c_{i,j} \neq 0\right\} \Big/ \#\left\{(i,j) : \widehat{c}_{i,j} \neq 0\right\}$ and $\#\left\{(i,j) : \widehat{c}_{i,j} \neq 0, c_{i,j} \neq 0\right\} \Big/ \#\left\{(i,j) : c_{i,j} \neq 0\right\}$, respectively. Finally, *Normalized Mutual Information (NMI)* [46] is used to measure the mutual dependence from an information theoretic perspective. Both *F-measure* and *NMI* take values between 0 and 1, where the value 1 implies the perfect recovered accuracy. Besides, the *RWE* more close to 0 indicates that a more accurate graph is learned. For a fair comparison, we normalize all learned Laplacians with the same scale as the groundtruth CGL and remove the edges in the learned graph whose magnitude is smaller than 10^{-4}.

A. PERFORMANCE COMPARISON

In this subsection, we compare the performance of four graph learning methods in both visual and quantitative form. Fig. 2 provides the visual comparison in three types of graph signals under the random instance of $\mathcal{G}_{RGG}$. The first row shows the graph learning performance from graph signals in case (i), which from left to right denotes the groundtruth graph Laplacian, the Laplacian matrices learned by CASTS, GL-Sigrep, GGSL and GL-logdet. The second and the third row denote the other results that are learned from graph signals in case (ii) and case (iii), respectively. As we can see, in each type of graph signal, Laplacian matrices learned by the proposed CASTS are visually more consistent with the groundtruth one than other baseline methods. One possible reason for this is that GL-Sigrep and GGLS only exploit the spatial smoothness in graph learning, in comparison, we learn a graph where signals are regularized to be spatiotemporally smooth. Besides, among the three cases, stable performance is achieved by the proposed CASTS, while the baseline methods are affected by time correlation, as we will see later. It is typically because the proper modeling of space-time interactions in the proposed model which takes into account time correlation in dynamic evolution.

For quantitative evaluation, we choose *Precision*, *Recall*, *F-measure*, *NMI* and *RWE* as evaluation metrics. Table 1 shows the comparison results in three types of graph signal with varying graph models. As shown in case (i), the proposed CASTS is superior to the others in all graph models, which achieves higher average *F-measure*, *NMI* scores and lower *RWE* scores, especially in $\mathcal{G}_{RSF}$ reaching the *F-measure* at 0.8612, *NMI* at 0.6722 and *RWE* score at 0.4489. Similar results can be also seen in case (ii) and case (iii) as expected. In addition, when it comes to a certain graph model, the performance of the baseline methods significantly improve with the decrease of correlation coefficient. Taking GL-Sigrep in

TABLE 1. Graph learning performance from different types of time-varying graph signal in the proposed and baseline methods.

	Graph signal in case (i)				Graph signal in case (ii)				Graph signal in case (iii)			
	CASTS	GL-Sigrep	GGLS	GL-logdet	CASTS	GL-Sigrep	GGLS	GL-logdet	CASTS	GL-Sigrep	GGLS	GL-logdet
$\mathcal{G}_{RGG}$												
F-measure	**0.8542**	0.8293	0.8121	0.7971	**0.8439**	0.6749	0.6225	0.5920	**0.8347**	0.7226	0.7365	0.7079
Precision	**0.8931**	0.8902	0.8865	0.8459	**0.8061**	0.5725	0.5536	0.7212	**0.7889**	0.6783	0.7377	0.6863
Recall	**0.8205**	0.7783	0.7632	0.7893	**0.9093**	0.8786	0.7911	0.5094	**0.8921**	0.7710	0.7252	0.7435
NMI	**0.5285**	0.5182	0.5094	0.4627	**0.5111**	0.1997	0.1674	0.1454	**0.5138**	0.2719	0.2928	0.2511
RWE	**0.2894**	0.3487	0.3511	0.3595	**0.2897**	0.3609	0.3345	0.5964	**0.3033**	0.4240	0.4218	0.4516
$\mathcal{G}_{RSF}$												
F-measure	**0.8612**	0.8353	0.7823	0.7697	**0.8367**	0.6552	0.6246	0.6009	**0.8362**	0.7246	0.7045	0.6667
Precision	**0.8884**	0.8192	0.7900	0.8689	**0.8985**	0.6178	0.6911	0.7848	**0.8352**	0.7813	0.6078	0.6136
Recall	0.8421	**0.8631**	0.7894	0.6842	**0.7879**	0.7268	0.5789	0.5247	**0.8421**	0.6757	0.8378	0.7297
NMI	**0.6722**	0.6285	0.6032	0.5719	**0.6502**	0.3606	0.3501	0.3476	**0.6278**	0.3826	0.3608	0.3408
RWE	**0.4489**	0.4550	0.4851	0.7886	**0.5257**	0.5537	0.5448	0.8473	**0.4949**	0.5778	0.5885	0.7164
$\mathcal{G}_{KN}$												
F-measure	**0.8278**	0.8151	0.7981	0.8023	**0.8192**	0.7298	0.6602	0.7186	**0.8235**	0.7462	0.7045	0.7246
Precision	0.7704	0.7643	**0.9073**	0.7362	**0.8058**	0.7028	0.6270	0.7014	0.7981	**0.8333**	0.6078	0.7812
Recall	**0.9119**	0.8784	0.7147	0.8889	**0.8596**	0.7756	0.7867	0.7768	**0.8918**	0.6756	0.8378	0.6527
NMI	**0.5548**	0.5247	0.5070	0.5117	**0.5436**	0.3764	0.3081	0.3748	**0.5539**	0.4052	0.3407	0.3826
RWE	**0.2564**	0.5261	0.4213	0.5345	0.5373	0.5809	**0.5137**	0.5517	**0.5106**	0.5715	0.5672	0.5590

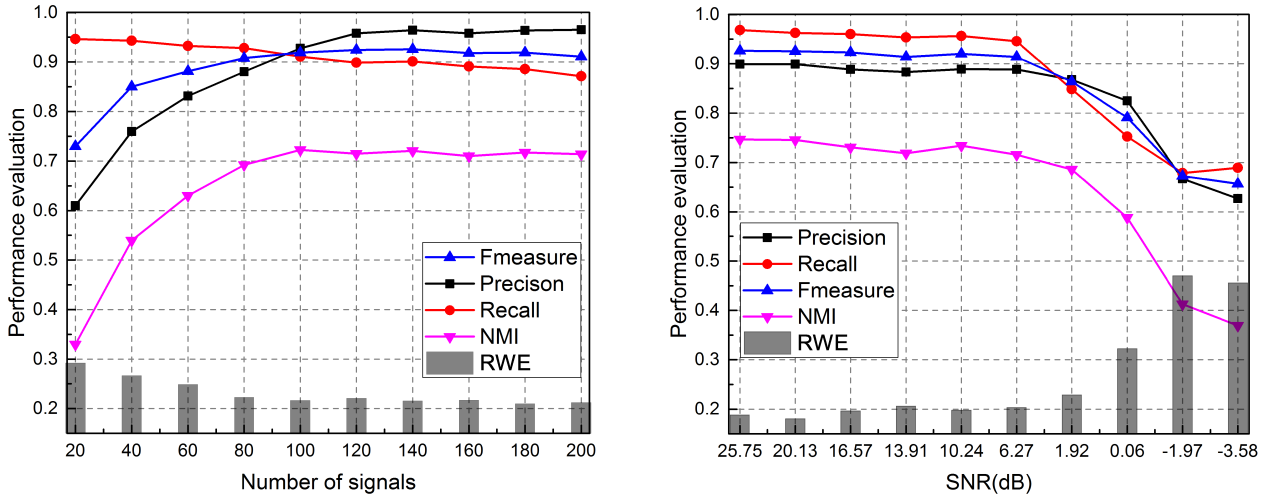

FIGURE 3. (a) Performance of the CASTS under different number of signals, and (b) Performance of the CASTS under different input SNR, for two random instances of $\mathcal{G}_{RGG}$ with graph signals in case (iii).

$\mathcal{G}_{KN}$ as an example, the score of *F-measure* in case (i) and case (ii) increases from 0.7298 to 0.8151 and the *RWE* from 0.5809 to 0.5261. One possible reason is that spatiotemporal signals degenerate to the static (i.e., time independent) graph signals under a small time correlation, which is suitable for static signal analysis (e.g., GL-Sigrep, GGLS and GL-logdet). In contrast, the performance of the proposed CASTS is quite robust across different graph and signal models. The improvement of CASTS over the baseline methods comes from the proper modeling of both the spatial and temporal correlation, which also shows the benefits of applying spatiotemporal smoothness property in graph learning.

B. GRAPH LEARNING WITH RESPECT TO NUMBER OF OBSERVED SIGNAL AND MEASUREMENT NOISE

To investigate the effect of the number of observed signals available for learning and measurement noise level, we consider graph signal in case (iii) under two random instances of $\mathcal{G}_{RGG}$, respectively. We show the learning accuracy of edge position as well as edge weight in both Fig. 3(a) and Fig. 3(b).

First, the performance of CASTS versus a different number of signals is shown in Fig. 3(a). We see that, as more signals are available, the *Recall* keeps a high value and the *Precision* gradually increases leading to an increasing value of *F-measure*. Meanwhile, *RWE* score decreases towards 0. These results indicate the improvement in graph learning performance. But the performance remains quite stable when the number of signals is more than 100. Next, We test the graph learning performance under different values of the signal-to-noise-ratio (SNR) given the variance of noise $\sigma_n{}^2$. From Fig. 3(b) one can read that the performance improves with the increase of input SNR. Specifically, when SNR is higher than 1.92dB, *F-measure* and *NMI* keeps the high

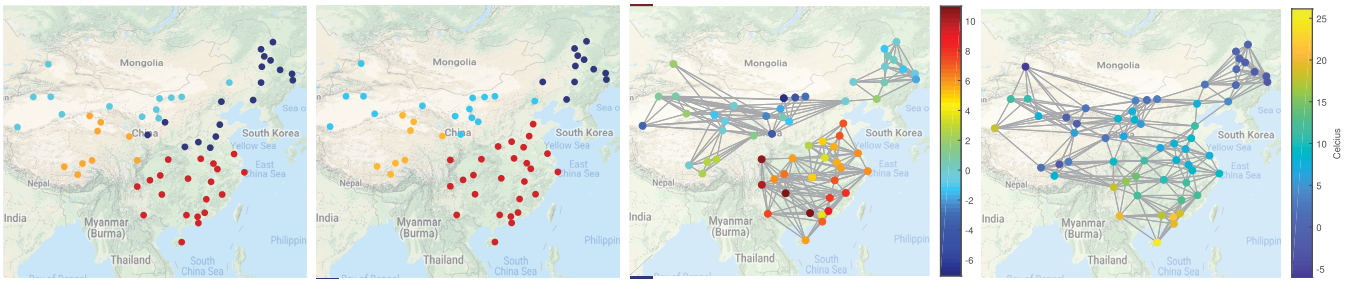

FIGURE 4. (a) The locations of 60 measuring stations in China. Different colors represent the groundtruth 4 clusters that correspond to 4 geographical regions. (b) The clustering results utilizing learned graph Laplacian obtained by the CASTS. (c) Graph structure learned by the CASTS, which achieves the best *RI* score in clustering performance. (d) Graph structure established by *KNN* with $K = 8$ from real world temperature signals. The color code in (c) and (d) respectively represent the temporal evolution of the temperature and the realistic temperature in Celcius scale on the 20th day.

scores around 0.9 and 0.7, respectively. As the level of noise continues to increase, both *F-measure* and *NMI* scores drop rapidly. Similar results can be observed in terms of *RWE* scores as well. This shows that CASTS is able to learn a graph that is very close to the groundtruth one until the SNR becomes very low.

C. GRAPH LEARNING FROM TEMPERATURE DATA

The China daily temperature dataset [50] is published by National Oceanic and Atmospheric Administration (NOAA). It is collected by 60 measuring stations in China over the first five months in 2017, thus in total there are 150 continuous samples for each measuring stations. We aim to learn a graph structure to explore the inherent relations in different areas.

However, the groundtruth graph structure is not available nor easy to define. According to the geographical location and climate condition, the land of China can be divided into four regions, including northern, southern, northwest and Qinghai-Tibet, shown by different colors in Fig. 4(a). We evaluate the performance of graph learning by applying spectral clustering [47] for the learned graph to partition the vertex set into four disjoint clusters. Metrics including *Purity*, *Rand Index (RI)* [48] and *NMI* score are applied for quantitative evaluation.

The results by the proposed CASTS are visually shown in Fig. 4. First, compared to the groundtruth clusters in Fig. 4(a), the clustering results in Fig. 4(b) are similar to the groundtruth ones with slight differences. Second, graph topology learned by the proposed CASTS and graph constructed by *KNN* scheme with $K = 8$ are depicted in Fig. 4(c) and Fig. 4(d), respectively. Notice that the colors in Fig. 4(c) and Fig. 4(d) have different meanings which represent the weighted time difference of temperature and the realistic temperature on the 20th day, respectively. Looking at Fig. 4(c) and Fig. 4(d) together, we find that though signals themselves in Fig. 4(d) exhibit similar values, their temporal evolution in Fig. 4(c) may have a big difference, and meanwhile graph learning from temporal evolution is more consistent with the meteorological features. One possible explanation could be that observation sites which are geometrically close may be geographically separated.

TABLE 2. The performance of graph learning methods in recovering groundtruth clusters of temperature measuring stations.

	CASTS	GL-Sigrep	GGLS	SpecTemp	GL-logdet	KNN
RI	0.8411	0.7900	0.7833	0.7832	0.7411	0.7567
Purity	0.8333	0.7167	0.75	0.5833	0.6667	0.6667
NMI	0.6712	0.5397	0.5236	0.5201	0.4701	0.4855

For further prove the effectiveness of the proposed method, we quantitatively compare clustering performance between the proposed and baseline methods. As we can see in Table 2, the performance of the proposed CASTS is better than the others in terms of clustering with 0.8411 for *RI*, 0.8333 for *Purity* and 0.6712 for *NMI*. There are two main reasons for the improvement of the proposed CASTS. Firstly, we learn a graph to describe the relation among temperatures, which adequately exploits the space-time properties of graph signals. Secondly, we utilize the smoothness of weighted difference signal rather than the smoothness of the signal itself, and the former one is more reasonable for this time-varying graph signal.

D. GRAPH LEARNING FROM EVAPOTRANSPIRATION DATA

The California daily evapotranspiration (ETo) dataset is published by California Department of Water Resources [51]. It is collected by 62 active observation stations over 150 days starting from February 1, 2018, and the size of data is 62×150. The selected data ranges from 0.02mm to 9.94mm, and the average is 3.838mm. Here, we would like to infer a graph that captures the similarities between these observation sites in the daily variation of evapotranspiration.

In this experiment, we do not have a groundtruth graph as well. Fortunately, a ETo Zone Map [52] divides the 62 stations into four ETo zones, which can be viewed as groundtruth clusters shown in Fig. 5. As depicted in this figure, though observation sites are geographically nearby, they may come from different clusters. By classifying the observation stations similar to the previous application, we evaluate the graph learning performance indirectly through clustering metrics.

Table 3 shows the clustering performance of the proposed and baseline methods. The *RI* scores of the comparison

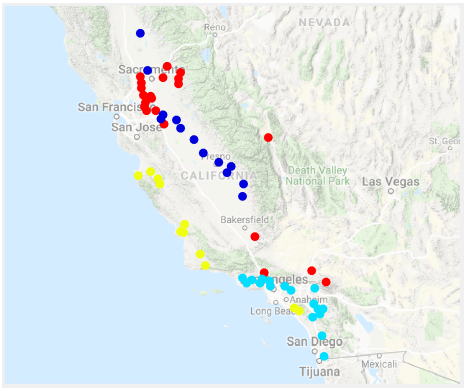

FIGURE 5. The geographical location of 62 measuring stations in California. The four colors from red, dark blue, yellow to light blue represent ETo zone 14, zone 12, zone 6 and zone 9, respectively.

TABLE 3. The performance of graph learning methods in recovering groundtruth clusters of ETo measuring stations.

	CASTS	GL-Sigrep	GGLS	SpecTemp	GL-logdet
RI	0.8461	0.8065	0.7820	0.7612	0.7653
Purity	0.8065	0.7419	0.6290	0.6451	0.6290
NMI	0.6462	0.5865	0.5613	0.4799	0.4613

methods GL-Sigrep, GGLS, SpecTemp, GL-logdet and the proposed CASTS are 0.8065, 0.7820, 0.7612, 0.7653 and 0.8461, respectively. The performance evaluation for all the methods in terms of *Purity* and *NMI* are also displayed in this table. Since the higher scores are achieved by the CASTS in all metrics, we can draw a similar conclusion as the previous experiment that the proposed CASTS is superior to the other graph learning methods on this ETo dataset.

E. PREDICTION PERFORMANCE OF THE PROPOSED METHOD

In this experiment, we test the prediction performance of the CASTS under the dynamic graph-based model on two real-world datasets in Section V-C and Section V-D. Specifically, the first 120 time series are training, by the proposed CASTS, to learn the graph structure L and temporal structure R,[1] and the remaining for testing. As discussed in Section III-A, with a known space-time structure, the proposed model can be regarded as a Kalman filter. Thus, for each testing time instant, we use the Kalman filter method in [49] to implement one-step prediction in above two datasets. In Fig. 6(a), we illustrate the true temperature signal presented at node 12 (black line) and the estimated one (red dotted line) in the next 30 time index. We repeat this operation for the ETo signal at node 7 and the results are displayed in Fig. 6(b). As we can see in Fig. 6(a) and Fig. 6(b), the estimation in both datasets is very close to its true signal. It means that the inferred graph

[1]We show the correlation coefficients of selected five observation sites in two real-world data as an example. The correlation coefficients are 0.6896, 0.6336, 0.7306, 0.5950, 0.6602, for temperature dataset; and 0.2244, 0.2797, 0.2427, 0.1858, 0.2172, for ETo dataset.

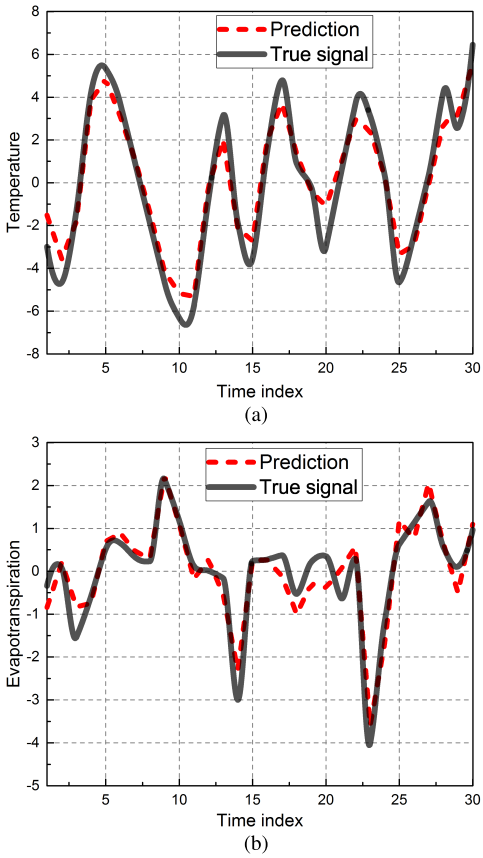

FIGURE 6. The true and predicted signals in two real-world datasets. (a) Temperature data in section V-C, and (b) ETo data in section V-D.

is close to reality, otherwise the temporal evolution based on a wrong graph topology could not lead to such a good prediction.

VI. CONCLUSION

This paper studies the problem of learning graphs from time-varying graph signal. Under the dynamic graph-based model for time-vertex representation, we novelly introduce spatiotemporal smoothness prior that accommodates time and graph setting in graph learning procedures. By exploiting the correlation and such smoothness property, we formulate the graph learning problem as a multi-convex optimization problem. A new graph learning method, CASTS, is proposed by applying the block coordinate descent scheme to simultaneously detect correlation and recover the graph. Simulations on both synthetic and real-world datasets demonstrate the improvement in graph learning performance over the state-of-the-art methods, and the effectiveness of the proposed model is also verified via prediction tasks in climate analysis.

REFERENCES

[1] H. Pham, C. Shahabi, and Y. Liu, "EBM: An entropy-based model to infer social strength from spatiotemporal data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 265–276.
[2] A. R. McIntosh, W. K. Chau, and A. B. Protzner, "Spatiotemporal analysis of event-related fMRI data using partial least squares," *NeuroImage*, vol. 23, no. 2, pp. 764–775, 2004.

[3] N. Eckert, E. Parent, R. Kies, and H. Baya, "A spatio-temporal modelling framework for assessing the fluctuations of avalanche occurrence resulting from climate change: Application to 60 years of data in the northern French Alps," *Climatic Change*, vol. 101, no. 3, pp. 515–553, 2010.

[4] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sep. 2014.

[5] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[6] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[7] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Graph filters," in *Proc. 38th IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 6163–6166.

[8] H. E. Egilmez and A. Ortega, "Spectral anomaly detection using graph-based filtering for wireless sensor networks," in *Proc. 39th IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 3892–3896.

[9] X. Zhu and M. Rabbat, "Approximating signals supported on graphs," in *Proc. 37th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 3921–3924.

[10] S. Celik, B. Logsdon, and S. I. Lee, "Efficient dimensionality reduction for high-dimensional network estimation," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1953–1961.

[11] K. Qiu, X. Mao, X. Shen, X. Wang, T. Li, and Y. Gu, "Time-varying graph signal reconstruction," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 870–882, Sep. 2017.

[12] D. Romero, M. Ma, and G. B. Giannakis, "Kernel-based reconstruction of graph signals," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 764–778, Feb. 2017.

[13] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3426–3477, Jul. 2016.

[14] A. Loukas and N. Perraudin. (2016). "Stationary time-vertex signal processing." [Online]. Available: https://arxiv.org/abs/1611.00255

[15] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, Mar. 1972.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.

[17] D. M. Witten, J. H. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *J. Comput. Graph. Statist.*, vol. 20, no. 4, pp. 892–900, 2011.

[18] R. Mazumder and T. Hastie, "Exact covariance thresholding into connected components for large-scale graphical lasso," *J. Mach. Learn. Res.*, vol. 13, pp. 781–794, Mar. 2012.

[19] H. P. Maretic, D. Thanou, and P. Frossard, "Graph learning under sparsity priors," in *Proc. 42th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 6523–6527.

[20] B. M. Lake and J. B. Tenenbaum, "Discovering structure by learning sparse graphs," in *Proc. 33rd Annu. Cogn. Sci. Conf.*, 2010, pp. 778–783.

[21] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, Dec. 2016.

[22] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, pp. 920–929, May 2016.

[23] N. Perraudin *et al.* (2014). "GSPBOX: A toolbox for signal processing on graphs." [Online]. Available: https://arxiv.org/abs/1408.5781

[24] C. Zhang, D. Florêncio, and P. A. Chou, "Graph signal processing—A probabilistic framework," Microsoft Res., WA, USA, Tech. Rep. MSR-TR-2015-31, 2015.

[25] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, Sep. 2017.

[26] M. G. Rabbat, "Inferring sparse graphs from smooth signals with theoretical guarantees," in *Proc. 42th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 6533–6537.

[27] S. P. Chepuri, S. Liu, G. Leus, and A. O. Hero, "Learning sparse graphs under smoothness prior," in *Proc. 42th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 6508–6512.

[28] V. Kalofolias, A. Loukas, D. Thanou, and P. Frossard, "Learning time varying graphs," in *Proc. 42th IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2826–2830.

[29] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, Sep. 2017.

[30] B. Pasdeloup, V. Gripon, G. Mercier, D. Pastor, and M. G. Rabbat, "Characterization and inference of graph diffusion processes from observations of stationary signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 3, pp. 481–496, Sep. 2018.

[31] R. Shafipour, S. Segarra, A. G. Marques, and G. Mateos, "Network topology inference from non-stationary graph signals," in *Proc. 42th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 5870–5874.

[32] D. Thanou, X. Dong, D. Kressner, and P. Frossard, "Learning heat diffusion graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 484–499, Sep. 2017.

[33] J. Mei and J. M. F. Moura, "Signal processing on graphs: Causal modeling of unstructured data," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 2077–2092, Apr. 2017.

[34] B. Baingana and G. B. Giannakis, "Tracking switched dynamic network topologies from information cascades," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 985–997, Feb. 2017.

[35] Y. Shen, B. Baingana, and G. B. Giannakis, "Topology inference of directed graphs using nonlinear structural vector autoregressive models," in *Proc. 42th IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 6513–6517.

[36] N. Cressie and C. K. Wikle, *Statistics for Spatio-Temporal Data*. Hoboken, NJ, USA: Wiley, 2011.

[37] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014.

[38] L. Kong *et al.*, "Data loss and reconstruction in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2818–2828, Nov. 2014.

[39] 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, Jan. 2011.

[40] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Pittsburgh, PA, USA, Tech. Rep., 1994.

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[42] M. Grant and S. Boyd. (Dec. 2017). *CVX: Matlab Software for Disciplined Convex Programming Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[43] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.

[44] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[45] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[46] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Infromation Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[47] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[48] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Amer. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[49] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Malaysia, Asia: Pearson, 2016.

[50] (2018). *National Centers for Environmental Information*. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datasets

[51] (2018). *California Irrigation Management Infromation System*. [Online]. Available: https://cimis.water.ca.gov

[52] (2018). *Evapotranspiration Zones California*. [Online]. Available: https://www.cimis.water.ca.gov/App-Themes/images/etozonemap.jpg

YUELIANG LIU received the B.S. degree in electronic engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communication, Ministry of Education. His research interests include graph signal processing and complex networks.

LISHAN YANG received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2018. He joined the Physics Science and Information Engineering College, Liaocheng University, Liaocheng, China, in 2018. His research interests include compressive sensing, signal processing on graphs, and machine learning.

KANGYONG YOU received the B.S. degree in communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communication, Ministry of Education. His current interests include wireless communication theory, array signal processing, compressive sensing, Bayesian statistical modeling, and source localization.

WENBIN GUO (M'06) received the B.S. degree from Tsinghua University, China, in 1994, and the M.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 1997 and 2005, respectively, all in electrical engineering. Since 1997, he has been with the Wireless Research Center, School of Telecommunication, Beijing University of Posts and Telecommunications, where he is currently a Professor. From 2006 to 2007, he was visiting the Department of Electrical and Computer Engineering, The University of Arizona, Tucson. His research interests include the various aspects of communication theory and signal processing, including blind signal processing, multi-terminal source-channel coding for wireless communication systems, and cognitive radio networks.

WENBO WANG received the B.S. degree in communication engineering and the M.S. and Ph.D. degrees in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1986, 1989, and 1992, respectively.

From 1992 to 1993, he was a Researcher with ICON Communication, Inc., Dallas, TX, USA. He has been a Chair Professor with the School of Information and Communications Engineering, BUPT, since 2000, where he is currently the Executive Dean of the Graduate School. He has been the Director of the Beijing Institute of Communication, since 2002, and an Assistant Director of the National Defense Communication Committee, China Institute of Communication. He has published more than 200 papers and six books, and holds 12 patents. His research interests include transmission and wireless access technology, wireless network theory, digital signal processing, mutiple-input multiple-output, cooperative and cognitive communications, and software radio technology.