Probabilistic Reconstruction of Spatio-Temporal Processes Over Multi-Relational Graphs

Qin Lu and Georgios B. Giannakis, Fellow, IEEE

Abstract—Given nodal observations that can be limited due to sampling costs or privacy concerns, several network-science-related applications entail reconstruction of values on all network nodes by leveraging topology information. Such a semi-supervised learning (SSL) task has been tackled mainly for graphs capturing a single class of inter-dependencies (or relations) among nodal variables. Faced with multi-relational graphs (MRGs), which emerge in various real-world networks, the present work introduces a principled framework to extrapolate spatio-temporal nodal processes that could be stationary or nonstationary. Broadening the scope of graph kernel-based approaches to MRGs, stationary graph processes are modeled first using a Gaussian mixture (GM) prior, where the covariance matrix of each Gaussian component describes one of the relations in the MRG. To further cope with nonstationary nodal processes, a first-order topology-dependent Gaussian transition prior is considered per relation, what gives rise to a GM transition density that accounts for all relations. In both cases, adapting the expectation-maximization (EM) algorithm yields two novel graph-adaptive solvers that not only reconstructs nodal features over unobserved nodes, but also quantifies the contribution of each relation. To enrich expressiveness of these novel EM-based approaches, multiple kernels per relation are also explored. Experiments with real data showcase the merits of the proposed methods relative to the existing alternatives.

Index Terms—Spatio-temporal process, multi-relational graphs, EM, semi-supervised learning, multi-kernel learning.

I. INTRODUCTION

OVER the last decade we have witnessed the emergence of massive data, a huge amount of which is collected by inter-networked agents (sensors or nodes) whose connectivity is captured by graphs. Capable of capturing relations among nodes, graph structures find widespread applications in fields as diverse as sociology, biology, neuroscience and economy [14], [30]. These so-termed network science applications rely on features captured by graphs. Nonstationary nodal processes, which cannot be all observable because sampling graph processes can be costly or impossible due to privacy considerations. With nodes and edges representing users and their friendships for example, individuals in social networks may be reluctant to share personal information, including age, income, and education level. In brain networks where nodes capture brain regions and edges denote connectivity patterns, acquiring nodal samples may require invasive procedures such as electrocorticography.

Given nodal observations over a subset of nodes, the semi-supervised learning (SSL) task of extrapolating nodal features to unobserved nodes can be addressed with the aid of the underlying graph topology that captures nodal inter-dependencies [10], [14]. Existing approaches to reconstructing time-invariant (TI) graph processes often rely on the smoothness of graph processes [15], [31], which asserts that connected vertices have similar features. In Facebook for instance, one can infer the income of a specific user from her/his friends’ income. Other than smoothness, inference from limited nodal observations can rely on e.g., ‘graph bandlimitedness’ [29], [32], [33], sparsity, overcomplete dictionaries [9], and neighbourhood similarities [23]. Most of these approaches can be unified under the framework of learning using graph kernels; see e.g., [27].

Temporal dynamics in the nodal processes add challenges to the SSL task, which has been tackled in several works. Capitalizing on the so-termed graph bandlimited model, inference of slow-varying processes over graphs has been pursued in [7], [8], [34]; while graph kernel-based estimators have been leveraged in [26] to reconstruct general dynamic processes. Aided by a topology-dependent dynamical model, semi-supervised tracking of dynamic processes over switching graphs is also possible using an online Bayesian approach [18], [19]. However, all these existing approaches to reconstructing stationary and nonstationary nodal processes are pertinent for graphs capturing a single relation characterizing the connectivity pattern among nodes.

In many contemporary applications, multiple types of relations over the same set of nodes coexist, giving rise to a multi-relational (a.k.a. multi-layer) graph (MRG) [13], [22]. In the particular case of multi-relational social networks (see Fig. 1), each relation accounts for a specific form of social interaction,
such as friendship, family bonds, or coworker-ties [35]. With a different type of tissue, e.g., brain or muscle, contributing to one connectivity pattern among the proteins (nodes), the protein interaction network can also be characterized as an MRG [37]. Albeit nontrivial, generalizing network analysis toolkits for single-relational graphs to MRGs is fortunately made possible by seeking a single connectivity pattern that summarizes the relational information in an MRG; see, e.g., [25] and [5]. While graph convolutional network based semi-supervised classification can incorporate multiple topologies [11], SSL of nodal processes over MRGs has not been addressed so far.

Relative to the aforementioned prior works, this paper deals with SSL of both static and dynamic spatio-temporal processes over MRGs. Its contributions are summarized next.

1) Building on kernel-based learning for single-relation graphs, a probabilistic generative perspective is introduced for stationary processes over MRGs. Given the MRG and partially observed nodal observations, the unobserved nodal values are reconstructed by adapting the celebrated expectation-maximization (EM) algorithm, which also outputs the weight measuring the contribution of each relation.

2) To further accommodate temporal dynamics in the nodal processes, a novel dynamical model is constructed to account for the spatio-temporal information in a principled manner. Following such a model, an EM-based solver is put forth to jointly estimate the nodal values and the per-relation contribution.

3) To bypass extra kernel selection and endow learning with richer expressiveness, the novel EM-based approaches are extended to incorporate multiple kernels per relation.

4) Real data tests confirm the merits of the proposed methods in extrapolating unobserved nodal values relative to state-of-the-art alternatives.

The rest of the paper is organized as follows. Section II presents preliminaries on kernel-based learning over single-relation graphs. Section III deals with SSL of stationary processes over MRGs, while Section IV introduces a novel dynamical model for nonstationary graph processes. Section IV-C broadens the proposed approaches to incorporate multiple kernels. Experimental tests and conclusions are presented in Sections V and VI, respectively.

Notation: Scalars are denoted by lowercase fonts, and column vectors (matrices) by bold lower (upper) case fonts. Superscripts $^\top$ and $^{-1}$ denote transpose and inverse, respectively; while $1_N$ stands for the $N \times 1$ all-one vector; and $\mathcal{N}(\mu, \Sigma)$ for the probability density function (pdf) of a Gaussian random vector $x$ with mean $\mu$, and covariance matrix $\Sigma$. Finally, if $A$ is a matrix and $x$ a vector, then $||x||_A^2 := x^\top A^{-1}x$.

II. FROM KERNEL-BASED LEARNING TO GENERATIVE MODELING OVER SINGLE-RELATION GRAPHS

To elicit the ensuing setup on learning over MRGs, preliminaries on kernel-based inference over single-relation graphs will be recapitulated first. Consider a network with $N$ nodes constituting the vertex set $\mathcal{V} := \{1, \ldots, N\}$. The relational information among the nodes in $\mathcal{V}$ is encoded in the $N \times N$ adjacency matrix $A$, whose $(n, n')$th entry $A(n, n')$ represents the weight of the edge connecting nodes $n$ and $n'$. The time-invariant nodal attribute is given by the image $\mathbf{x}$ of the mapping: $\mathcal{V} \rightarrow \mathbb{R}$. For instance, $x(n)$ may denote the income of user $n$ in Facebook. The values over all the nodes are collected in $\mathbf{x} := [x(1), \ldots, x(N)]^\top$.

In several applications, sampling costs or privacy considerations allow attributes over only a subset of nodes to be observed, and the resultant observation model is

$$y = Hx + e$$

where the sampling matrix $H \in \{0, 1\}^{N \times M}$ selects the $M$ observed nodes via the column indices of the only “1”s per row. Since $M < N$, finding $\mathbf{x}$ given $\mathbf{y}$ without any other constraints is under-determined, leading itself to multiple possible solutions. To uniquely reconstruct $\mathbf{x}$, one can rely on the regularized least-squares formulation

$$\hat{x} = \underset{x}{\text{arg min}} \|y - Hx\|_2^2 + \mu \Omega(x)$$

where $\Omega(\cdot)$ is a chosen regularizer promoting a certain structural in the estimate, and $\mu \geq 0$ is the regularization coefficient that controls the importance of the regularization term with respect to the fitting error over the observed nodes.

With the Laplacian matrix $L := D - A$, where $D := \text{diag}(A1_N)$, the well-known Laplacian $\text{regularizer}$

$$\Omega_{LR}(x) := x^\top L x = \frac{1}{2} \sum_{n=1}^{N} \sum_{n'=1}^{N} A(n, n') |x(n) - x(n')|^2$$

encourages smoothness of the estimated signal on the graph as vertices connected by strong links (large $A(n, n')$) will have similar signal estimates in order to minimize (3). To promote other properties such as diffusion or graph bandlimitedness, $L$ in (3) is replaced by $r(L)$, where the scalar energy mapping $r : \mathbb{R} \rightarrow \mathbb{R}_+$ is applied on the eigenvalues of $L$ to promote desired properties (cf. Table I). The pseudo-inverse of $r(L)$ then yields the graph Laplacian kernel, see e.g., [27]

$$K := r^2(L).$$

With $K$ as in (4), and $\Omega(x) := ||x||_K^2$, (2) yields the family of kernel ridge regression (KRR) estimators, which enjoys well-documented reconstruction performance [31].

So far, we outlined SSL over single-relational graphs using a deterministic kernel-based framework. It is however instructive to identify its Bayesian generative counterpart. To this end, we view $x$ as random with prior probability density function (pdf)
$p(x) = \mathcal{N}(\mathbf{x}; 0, \mathbf{K})$, and conditional data likelihood $p(y|x) = \mathcal{N}(y; \mathbf{Hx}, \mathbf{R})$ with $\mathbf{R} = \mathbf{R}_M$. Under these Gaussian densities, the maximum a posteriori (MAP) estimator of $\mathbf{x}$ given $y$ is equivalent to the KRR estimator, which amounts to the linear minimum mean-square error (LMMSE) estimator

$$\hat{x} = \arg\max_{x} p(x|y) = \arg\max_{x} p(y|x)p(x)$$
$$= \arg\min_{x} \|y - \mathbf{Hx}\|^2_{\mathbf{R}} + \|x\|^2_{\mathbf{K}}. \tag{5}$$

Aided by probabilistic modeling, we are ready to broaden our scope to graphs with multiple relations, as well as nonstationary nodal processes in the remaining of the paper.

III. SSL OF STATIONARY PROCESSES OVER MRGs

Although attractive and with documented performance, the kernel-based approach of Section II cannot accommodate multi-relational graphs, which arise in several applications. In Fig. 1 for instance, Facebook users can also be connected through their LinkedIn and Twitter social networks. To account for such multiple relations of nodes in $\mathcal{V}$, an MRG is defined as $\mathcal{G} := \{\mathcal{V}, \mathcal{A}\}$, where the topology set $\mathcal{A} := \{\mathbf{A}^s, s = 1, \ldots, S\}$, with $\mathbf{A}^s$ denoting the adjacency matrix associated with relation $s$. The Laplacian matrix associated with $\mathbf{A}^s$ is $\mathbf{L}^s := \mathbf{D}^s - \mathbf{A}^s$, where $\mathbf{D}^s := \text{diag}(\mathbf{A}^s \mathbf{1}_N)$.

To model time-varying nodal features, consider the spatio-temporal graph process $x$ given by the map $\mathcal{V} \times T \rightarrow \mathbb{R}$, where $T := \{1, 2, \ldots\}$ is the set of slot indices. Here, $x_t(n)$ represents the feature of node $n$ at slot $t$. Features across all the nodes per slot $t$ are collected in the vector $\mathbf{x}_t := [x_t(1), \ldots, x_t(N)]^\top$. With the slot index $t$ as subscript, the dynamic counterpart of (1) in the semi-supervised setting is

$$y_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{e}_t \tag{6}$$

where $\mathbf{H}_t \in \{0, 1\}^{M_t \times N}$ denotes a generally time-varying sampling matrix, and $\mathbf{e}_t$ is noise that accounts for unmodeled uncertainties, and it is assumed white and Gaussian distributed with zero mean and covariance $\mathbf{R}_t$. Correspondingly, (6) gives rise to the data likelihood $p(y_t|\mathbf{x}_t) = \mathcal{N}(y_t; \mathbf{H}_t \mathbf{x}_t, \mathbf{R}_t)$.

Given the MRG with topologies in $\mathcal{A}$, and the observations over $T$ slots $\mathbf{Y} := [y_1, \ldots, y_T]^\top$, our goal is to estimate $\mathbf{X} := [x_1, \ldots, x_T]^\top$. To this end, suppose first that $\mathbf{x}_t$ is independent and identically distributed (iid) across slots $t$, and thus a stationary vector process. Inference of nonstationary processes will be investigated in the next section. Compared with the SSL task over a single-relation graph in the previous section, the problem here faces the additional challenge of synthesizing contributions from the multiple relations of the MRG, which we will demonstrate possible to handle in a principled manner via probabilistic modeling.

First, similar to a single-relation graph, each relation $s$ ‘contributes’ to the description of $x_t$ via the pdf $p(x_t|s) = \mathcal{N}(x_t; 0, \mathbf{K}^s)$, where $\mathbf{K}^s = \mathbf{P}(\mathbf{L}^s)$ is the Laplacian kernel associated with relation $s$. To further account for the multiple relations, the ensemble prior pdf of $\mathbf{x}_t$ is given by the following Gaussian mixture (GM) generative model

$$p(x_t; \pi) = \sum_{s=1}^{S} \pi_s p(x_t|s) = \sum_{s=1}^{S} \pi_s \mathcal{N}(x_t; 0, \mathbf{K}^s) \tag{7}$$

where $\pi := [\pi_1, \ldots, \pi_S]^\top$ holds the unknown probabilities $\pi_s \in [0, 1]$ that satisfy $\sum_{s=1}^{S} \pi_s = 1$. In words, $\pi_s$ denotes the probability $x_t$ adheres to relation $s$, and has thus been generated from $\mathcal{N}(x_t; 0, \mathbf{K}^s)$. The GM prior in (7) captures data multi-modality arising from heterogeneous relations that are possibly present in the MRG, thus offering a more expressive model than their single-mode counterparts.

Building on (6) and (7), we will rely on the Bayes formulation to jointly seek the maximum a posteriori (MAP) estimate of $\mathbf{X}$ and the maximum likelihood (ML) estimate of $\pi$ as $\{\hat{\mathbf{X}}, \hat{\pi}\} = \arg\max_{\mathbf{X}, \pi} p(\mathbf{X}|\mathbf{Y}; \pi)$. Bayes’ rule allows us to factorize the posterior as $p(\mathbf{X}|\mathbf{Y}; \pi) \propto p(\mathbf{Y}|\mathbf{X}) p(\mathbf{X}^s; \pi)$. Further, conditional independence of the variables in (6) and (7) across slots yields factorization of the batch data likelihood and prior as $p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} p(y_t|x_t)$ and $p(\mathbf{X}^s; \pi) = \prod_{t=1}^{T} p(x_t^s; \pi)$. Consequently, the sought optimization problem in log form is cast as

$$\{\hat{\mathbf{X}}, \hat{\pi}\} = \arg\max_{\mathbf{X}, \pi} \sum_{t=1}^{T} \left[ \log p(y_t|x_t) + \log p(x_t; \pi) \right]. \tag{8}$$

Although the non-convex objective in (8) can be in principle solved using a gradient-ascent based iteration, its convergence to local optima entails judicious selection of the step size per iteration. Also, gradient-based approaches do not fully leverage the potential benefits arising from a data generative model. These considerations prompted us to adopt the celebrated expectation-maximization (EM) algorithm [6], which is a ‘workhorse’ iterative optimization technique to find the ML [6] or MAP [17], [36] solutions for probabilistic models with latent variables [4].

Next, an EM-based solver is developed to jointly seek the MAP and ML estimates in (8) with provable convergence at least to a local optimum.

A. EM for Joint MAP and ML Estimation

Aiming to adapt the EM as a solver of (8), we start by rewriting the GM prior (7) as

$$p(x_t; \pi) = \sum_{s=1}^{S} \Pr(z_t = s)p(x_t|z_t = s) \tag{9}$$

where the per-slot $t$ latent relation variable $z_t \in \{1, \ldots, S\}$ is assumed to be iid with probability mass function (pmf) $\pi$. Fig. 2(a) shows how the probabilistic graphical model depicts the links among random variables $\{x_t, y_t, z_t\}$.

Next, we will express the objective in (8) using the so-called Kullback-Leibler (KL) divergence between two pdfs $p(x)$ and $q(x)$ that is given by KL($q(x)||p(x)$) := $\int q(x) \log \frac{q(x)}{p(x)} dx \geq 0$; and the entropy of $q(x)$ defined as $H(q(x)) := -\int q(x) \log q(x) dx$. For any pdf $q(x)$, it can be readily verified that, see e.g., [4, Chap. 9.4]

$$\log p(X|Y;\pi) = KL(q(z)||p(z|X,Y;\pi)) + L(q(X,\pi)) \tag{10}$$
Algorithm 1: EM Algorithm for Static Processes Over MRGs.

1: Input: \( Y, K\), \( s = 1, \ldots, S, R_t, H_t, t = 1, \ldots, T \)
2: Initialization: \( \hat{X}_t^{(0)}, \hat{\pi}^{(0)} \)
3: \( n = 0 \)
4: while not converged do
5: \( n = n + 1 \)
6: E-Step:
7: Evaluate \( w_{t,t}^{(n)}, s = 1, \ldots, S, t = 1, \ldots, T, \) via (18)
8: M-Step:
9: M1: Update \( \hat{\pi}_s^{(n)} \) via (19) for \( s = 1, \ldots, S \)
10: M2: Update \( \hat{X}_t^{(n)} \) via (20) for \( t = 1, \ldots, T \)
11: end while
12: Output: \( \hat{X}_t^{(n)}, \hat{\pi}_s^{(n)} \)

where \( z := [z_1 \ldots z_T]^T \), and
\[
\mathcal{L}(q, X, \pi) := \mathbb{E}_{q(z)} \left[ \log p(X, Y, z; \pi) - \log p(Y) + H(q(z)) \right].
\]

It is evident that (11), which is a functional of \( q(z) \) and a function of \( X \) and \( \pi \), amounts to a lower bound for \( \log p(X, Y, \pi) \). As we shall see later, the EM algorithm maximizes the lower bound (11) iteratively by alternating between E- and M-steps per iteration.

Let \( \hat{X}_t^{(n-1)} \) and \( \hat{\pi}_s^{(n-1)} \) denote estimates of \( X \) and \( \pi \) at the end of iteration \( n - 1 \). The E-Step at iteration \( n \) maximizes (11) with respect to \( q(z) \) and \( \hat{\pi}_s^{(n-1)} \) fixed, yielding (12). The joint pdf \( q(z) \) maximizes (11) and the auxiliary function \( \mathcal{L}(q(z), X, \pi) \) with respect to \( \pi \), which boils down to seeking the maximum of the following auxiliary function
\[
Q(X, \pi; \hat{X}_t^{(n-1)}, \hat{\pi}_s^{(n-1)}):= \mathbb{E}_{q(z)} \left[ \log p(X, Z, \pi) \right].
\]

Thus, the updates of \( X \) and \( \pi \) at iteration \( n \) are given by
\[
\hat{X}_t^{(n)} = \arg\max_{X} Q(X, \pi; \hat{X}_t^{(n-1)}, \hat{\pi}_s^{(n-1)}).
\]

Since \( \mathcal{L}(q(z), \hat{X}_t^{(n)}, \hat{\pi}_s^{(n)}), \hat{\pi}_s^{(n)} \geq \mathcal{L}(q(z), \hat{X}_t^{(n)}, \hat{\pi}_s^{(n)}), \hat{\pi}_s^{(n)} \) and \( KL(q(z)||p(z|X_t, Y, \hat{\pi}_s^{(n)})) \geq 0 \), it holds that
\[
\log p(X_t, Y; \hat{\pi}_s^{(n)}) \geq \log p(\hat{X}_t^{(n-1)}|Y, \hat{\pi}_s^{(n-1)})
\]

thereby proving the convergence of the EM algorithm at least to the local maxima of \( \log p(X, Y, \pi) \).

B. Graph-Guided EM Solver

Our graph-aware solver of (8), abbreviated hereafter as “EM-Static,” entails alternating between the following E- and M-steps per iteration.

1) E-Step: With \( \hat{X}_t^{(n-1)} \) and \( \hat{\pi}_s^{(n-1)} \) available from iteration \( n - 1 \), we will rely on \( q(z) = p(z|X_t^{(n-1)}, Y, \hat{\pi}_s^{(n-1)} \) to evaluate the Q-function (12). Towards this end, it is crucial to express the joint pdf \( p(X, Y, z; \pi) \), which, based on the probabilistic graphical model in Fig. 2(a), can be factored as
\[
p(X, Y, Z; \pi) = p(Y|X)p(X, Z; \pi)
\]

where the indicator function \( I(z_t = s) \) is given
\[
I(z_t = s) = \begin{cases} 1 & z_t = s \\ 0 & z_t \neq s \end{cases}
\]

Accordingly, the Q-function (12) boils down to
\[
Q(X, \pi; \hat{X}_t^{(n-1)}, \hat{\pi}_s^{(n-1)})
\]

where terms unrelated with \( X \) and \( \pi \) are ignored, and
\[
w_{t,s}^{(n)} := \mathbb{E}_{q(z)}[I(z_t = s)] = Pr(z_t = s|X_t^{(n-1)}, Y; \hat{\pi}_s^{(n-1)})
\]

Thus, the updates of \( X \) and \( \pi \) at iteration \( n \) are given by
\[
\hat{X}_t^{(n)} = \arg\max_{X} Q(X, \pi; \hat{X}_t^{(n-1)}, \hat{\pi}_s^{(n-1)}).
\]

On the other hand, the update of \( x_t, t = 1, \ldots, T, \) is given by
\[
x_t^{(n)} = \arg\max_{x_t} - \frac{1}{2} ||y_t - H_t x_t ||^2 - \sum_{s=1}^{S} w_{t,s}^{(n)} ||x_t ||^2 K_s
\]

The E- and M-steps are repeated until at least one of the convergence conditions is met; namely, either the maximum number of iterations is reached, or, the increment of the objective between successive iterations falls below a preselected threshold; see also Alg. 1. Interestingly, for a single-relation graph \( (S = 1) \), the closed-form reconstruction in (20) reduces to the LMMSE estimator obtained as a solution of (5) with information matrix \( K^{-1} \) obtained from the underlying graph kernel. While for MRGs, the reconstruction in (20) involves
an aggregate information matrix formed as a weighted convex combination of all S single-relation graph kernels.

So far, SSL of nodal processes over MRGs has been handled assuming no temporal dynamics, which may not hold in several practical settings, such as stock prices or climate conditions. The next section will target at the more challenging task of reconstructing nonstationary nodal processes.

IV. SSL OF DYNAMIC PROCESSES OVER MRGs

Graph processes with arbitrary dynamics render the SSL task impossible to tackle, in general. Fortunately, building structured dynamical models, as the one dealt with in the ensuing subsection, allows us to exploit temporal correlations across slots, thus making the best of the available observations.

A. Probabilistic Modeling of Dynamic Processes Over MRGs

To capture the spatio-temporal structure of the dynamic graph processes, we model the evolution from \( x_t \) to \( x_t \) pertinent to relation \( s \) using a first-order Markov process as

\[
\begin{align*}
    x_t = F^s x_{t-1} + v_t^s
\end{align*}
\]  

(21)

where the state transition matrix is a known function \( f \) of the associated adjacency matrix given by

\[
F^s := f (A^s)^{1};
\]  

(22)

while the process noise \( v_t^s \) is assumed uncorrelated with \( x_t \), temporally white and Gaussian distributed with zero mean and covariance \( K^s \) corresponding to the Laplacian kernel (4). Temporal correlation is accounted through the topology-dependent term \( F^s x_{t-1} \), and spatial correlation per slot via the Laplacian kernel \( K^s \). This explains why (21) offers a valuable dynamical model for graph processes.

Along the lines we followed for stationary graph processes, probabilistic weighting the contributions of the \( S \) relations in the MRG to the dynamics from slot \( t - 1 \) to \( t \), yields the marginal state transition pdf as (cf. (7))

\[
\begin{align*}
    p(x_t|x_{t-1}; \pi_t) = \sum_{s=1}^{S} \Pr(z_t = s)p(x_t|x_{t-1}, z_t = s)
\end{align*}
\]  

(23a)

\[
= \sum_{s=1}^{S} \pi_{t,s} N(x_t; F^s x_{t-1}, K^s)
\]  

(23b)

where \( \pi_t := [\pi_{t,1} \ldots \pi_{t,S}]^T \) is the unknown pmf for the latent variables \( \{ z_t \} \) that model the ‘contribution’ of different relations and are independent across slots. Different from the stationary case in Sec. III, \( z_t \) here is not iid in order to account for the general nonstationarity in \( x_t \). The probabilistic graphical model in Fig. 2(b) showcases the generative process for random variables \( \{ x_t, y_t, z_t \} \) based on (6) and (23).

B. Graph-Guided EM Solver

Estimating \( \mathbf{X} \) and \( \mathbf{\Pi} := [\pi_1 \ldots \pi_T]^T \) relies on the posterior \( p(\mathbf{X}|\mathbf{Y}; \mathbf{\Pi}) \), which according to Bayes’ rule, is given by

\[
p(\mathbf{X}|\mathbf{Y}; \mathbf{\Pi}) \propto p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}; \mathbf{\Pi}).
\]

Hence, seeking the MAP estimate of \( \mathbf{X} \) and the MLE of \( \mathbf{\Pi} \) boils down to solving

\[
\left\{ \hat{\mathbf{X}}, \hat{\mathbf{\Pi}} \right\} = \arg \max_{\mathbf{X}, \mathbf{\Pi}} \log p(\mathbf{X}|\mathbf{Y}) + \log p(\mathbf{Y}|\mathbf{X})
\]

s.to \( 1^T \pi_t = 1 \ \forall t \)  

(24)

where the temporal independence in (6) and (23) allow for factorizing the likelihood and prior, respectively, as

\[
p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} p(y_t|x_t)
\]  

(25)

\[
p(\mathbf{X}; \mathbf{\Pi}) = \prod_{t=1}^{T} p(x_t|x_{t-1}; \pi_t)
\]  

(26)

Towards solving the nonconvex problem (24) with convergence to local optima, we will once again leverage the EM solver by first adapting the auxiliary function (12) as

\[
Q(\mathbf{X}, \mathbf{\Pi}; \mathbf{\hat{X}}^{(n-1)}, \mathbf{\hat{\Pi}}^{(n-1)}) = \mathbb{E} \left[ q^{(n)}(z|x, \mathbf{Y}; \mathbf{\Pi}) \log p(\mathbf{X}, \mathbf{Y}, z; \mathbf{\Pi}) \right]
\]  

(27)

where \( q^{(n)}(z) = \mathbb{P}(z|x^{(n-1)}, \mathbf{Y}, \mathbf{\Pi}^{(n-1)}) \). Next, we will elaborate on the E- and M-steps in the so-termed “EM-Dynamic” in the following subsections.

1) E-Step: Given \( \mathbf{X}^{(n-1)} \) and \( \mathbf{\Pi}^{(n-1)} \) the goal is to evaluate the Q-function, where the joint pdf in (27) is

\[
p(\mathbf{X}, \mathbf{Y}, z; \mathbf{\Pi}) = p(\mathbf{X}, z; \mathbf{\Pi})p(\mathbf{Y}|\mathbf{X})
\]

\[
= \prod_{t=1}^{T} p(z_t)p(x_t|x_{t-1}, z_t)p(y_t|x_t)
\]

\[
= \prod_{t=1}^{T} \left( \prod_{s=1}^{S} \pi_{t,s} N(x_t; F^s x_{t-1}, K^s) I(z_t = s) \right)
\]

(28)

\[
\times N(y_t; H_t x_t, R_t)
\]

Algorithm 2: EM Algorithm for Dynamic Processes Over MRGs.

1: Input: \( \mathbf{Y}, \mathbf{F}^s, \mathbf{K}^s, \mathbf{R}_t, \mathbf{H}_t, t = 1, \ldots, T, s = 1, \ldots, S \)
2: Initialization: \( \{ \hat{x}_1^{(0)}, \hat{\pi}_1^{(0)} \} \)
3: \( n = 0 \)
4: while not converged do
5: \( n = n + 1 \)
6: E-Step:
7: Evaluate \( w_{t,s}^{(n)} : s = 1, \ldots, S, t = 1, \ldots, T \), via (30)
8: M-Step:
9: M1 Update \( \hat{\pi}_t^{(n)}, t = 1, \ldots, T, \) via (32)
10: M2.1 Obtain \( \{ \hat{x}_t^{(n)}, \hat{\pi}_t^{(n)} \}, t = 1, \ldots, T \) via Alg.3
11: M2.2 Obtain \( \{ \hat{x}_t^{(n)}, \hat{\Pi}_t^{(n)} \}, t = 1, \ldots, T \) via Alg.4
12: end while
13: Output: \( \hat{x}_t^{(n)}, \hat{\Pi}_t^{(n)}, t = 1, \ldots, T, s = 1, \ldots, S \)
whose logarithm after dropping irrelevant constants is

\[
\log(p(X, Y, z; \Pi)) = \sum_{t=1}^{T} \left( \sum_{s=1}^{S} I(z_t = s) \left( \log \pi_{t,s} - \frac{\|x_t - F^s x_{t-1}\|_{K,s}^2}{2} - \frac{\|y_t - H \tilde{x}_t\|_{R_s}^2}{2} \right) \right).
\]

(29)

With \(\hat{X}^{(n-1)}\) and \(\hat{\Pi}^{(n-1)}\) available, \(u_{t,s}^{(n)} := \mathbb{E}_{q_{t,s}^{(n)}}[I(z_t = s)]\) can be obtained as

\[
u_{t,s}^{(n)} = \Pr(z_t = s|\hat{X}^{(n-1)}, Y; \hat{\Pi}^{(n-1)}) = \Pr(z_t = s|\hat{X}^{(n-1)}) p(\hat{X}^{(n-1)}|z_t = s)p(Y|\hat{X}^{(n-1)})
\]

(30)

With \(\tilde{\pi}_{t,s}^{(n-1)}\) and \(\tilde{\pi}_{t,s}^{(n-1)}\) available, \(\tilde{\pi}_{t,s}^{(n)} := (\tilde{\pi}_{t,s}^{(n-1)})^{1/\alpha} \pi_{t,s}^{(n-1)}\) yields the new iterate as

\[
Q(X, \Pi, X^{(n)}, \Pi^{(n)}) = \mathbb{E}_{q_{t,s}^{(n)}}[\log(p(X, z, Y; \Pi))]
\]

\[
= \sum_{t=1}^{T} \left( \sum_{s=1}^{S} \nu_{t,s}^{(n)} \log \tilde{\pi}_{t,s} - \frac{\|x_t - F^s x_{t-1}\|_{K,s}^2}{2} - \frac{\|y_t - H \tilde{x}_t\|_{R_s}^2}{2} \right).
\]

(31)

2) M-Step: Maximizing (31) over \(\pi_t, t = 1, \ldots, T\), yields the new iterate as

\[
\tilde{\pi}_{t,s}^{(n)} = \arg \max_{\pi_t, \tilde{\pi}_{t,s}} \sum_{s=1}^{S} u_{t,s}^{(n)} \log \pi_{t,s} = w_{t,s}^{(n)}
\]

(32)

where \(w_{t,s}^{(n)} := \left[u_{t,s}^{(n)} \cdots u_{t,S}^{(n)}\right]^T\).

As for the update of \(\tilde{x}_t\), we first establish that

\[
\sum_{t=1}^{T} \sum_{s=1}^{S} u_{t,s}^{(n)} \|x_t - F^s x_{t-1}\|_{K}^2 = \sum_{t=1}^{T} \|x_t - F^s x_{t-1}\|_{K}^2
\]

(33)

where \(F^s x_{t-1}\) and \(K^{(n)}\) are obtained as detailed in the Appendix. Using (33), the objective to optimize over \(X\) is

\[
Q_X = \sum_{t=1}^{T} \left( \|y_t - H \tilde{x}_t\|_{R_1}^2 + \|x_t - F^s x_{t-1}\|_{K}^2 \right)
\]

(34)

which corresponds to an aggregate dynamical model evolving as \(x_t = \tilde{F}_t^{(n)} x_{t-1} + \tilde{v}_t\), with \(\tilde{v}_t \sim \mathcal{N}(\tilde{v}_t; 0, \tilde{K}_t^{(n)})\). Apparently, optimizing (34) with respect to \(X\) can be implemented efficiently using the celebrated Kalman smoother [24], whose detailed steps are summarized in Alg. 4; see also e.g. [13], [21]. In addition to the point estimates \(\{\tilde{x}_t^{(n)}\}\), the Kalman smoother outputs the accompanying covariance matrices \(\{P_t^{(n)}\}\) that can self-assess the estimation accuracy. The implementation details of our EM-Dynamic are summarized in Alg. 2.

C. Multi-Kernel Adaptivity

So far, SSL of both stationary and nonstationary processes has been handled with a single preselected Laplacian kernel per relation. Besides requiring extra tuning efforts, the single kernel may limit expressiveness. To enrich learning expressiveness, we will associate each topology with a dictionary of energy mappings \(\{r^1, \ldots, r^K\}\) to yield an expanded list of \(S = SK\) Laplacian kernels \(\{\mathbf{K}_{s,k}\}_{s=1}^{S} = r^k(L_s)\). Having such a kernel list available, the aforementioned EM-based solvers can be readily adopted to infer the unobserved nodal values, and select the kernel combination in a data-driven fashion.

When only a single relation is available to characterize the connectivity pattern among the nodes, the novel EM-based approaches can still be leveraged for probabilistic multi-kernel learning, nicely complementing their KRR-based multi-kernel counterparts from the deterministic viewpoint in [27]. It is worth pointing out that this special case with a single topology \(A\) will simplify (33) in EM-Dynamic as

\[
\sum_{k=1}^{K} w_{t,k}^{(n)} \|x_t - F x_{t-1}\|_{K_t} = \|x_t - F x_{t-1}\|_{K_t}^2
\]

(35)

where \(F := f(A)\) and

\[
K_t = \left( \sum_{k=1}^{K} w_{t,k}^{(n)} (K_t^{(1,k)})^{-1} \right)^{-1}
\]

(36)

The following remarks are now in order.

Remark 1: (Dynamic graphs). Although the setup so far has dealt with a time-invariant adjacency matrix per relation, the novel approaches can be readily broadened to accommodate time-varying topologies, if present, through a time-dependent Laplacian \(K_t = r^k(I_t^s)\) and \(F_t = f(A_t)\) in (7) and (23).

Remark 2: (Computational complexity). Each iteration of EM-Static involves inversion of the \(N \times N\) matrix in (20) per slot \(t\), thus incurring complexity of \(O((S + TN_{max}) N^3)\) over \(N_{max}\) iterations with precomputed \(\{K_t\}^{S}_{s=1}\). On the other hand, \(N_{max}\) iterations in EM-Dynamic incur complexity \(O((S + 2TN_{max}) N^3)\), since one pass of the operation in (33), as well as the Kalman smoother, require computations of order \(O(TN^3)\). Such cubic (in \(N\)) complexities limit their ability to handle large graphs without distributed solvers. To distribute computations along the lines of [28], the network can be divided in \(N_g\) subnetworks, each with at most \([N/N_g]\) nodes. As the complexity per subgraph scales as \(O([N/N_g]^3)\) (ignoring the factor not dependent on \(N\)), the overall complexity is of order \(O(N_g N_g^2 [N/N_g]^3)\), and scalability for large graphs can be effected by adjusting \(N_g\). Future research also includes optimal choice of \(N_g\), and graph partitioning based on the multiple relations.

V. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed EM-based approaches using real datasets, and compare them with existing algorithms including those based
Fig. 2. Probabilistic graphical models for (a) static and (b) dynamic processes over MRGs.

The first dataset, provided by the National Climatic Data Center [1], records hourly temperature measurements at $N = 109$ measuring stations across the continental United States in 2010. Here, $x_t(n)$ represents the daily temperature at the $n$th station. Relying on the geographical distances of the measuring stations, we construct an MRG with two relations following the approach in [26], where 7- and 11-nearest neighbors are considered, respectively.

For EM-Static, the per-relation $K_s$ in (7) is modeled by the diffusion kernel with parameter $a = 2$ (cf. Table I), where the value of $a$, chosen from the set $\{1, 1.2, 1.4, \ldots, 3\}$, attains the lowest NMSE. On the other hand, EM-Dynamic postulates that the true temperature values evolve based on (23), where $F_s = 0.05(A_s + I_N)$, and process noise covariance $K_s$ is that of a diffusion kernel with $a = 2$. Here, the value 0.05 and $a = 2$ are selected to yield the lowest NMSE from the sets $\{0.01, 0.02, 0.03, \ldots, 0.2\}$ and $\{1, 1.2, 1.4, \ldots, 3\}$.

With $M = 20$ observed nodes, the superior reconstruction performance of the novel EM-based methods relative to the competing alternatives is illustrated in Fig. 3, where EM-Dynamic slightly outperforms EM-Static by exploiting temporal correlations. Fig. 4 further corroborates this assertion by displaying the true and estimated temperature values over an unobserved node. Moreover, it is evident from Fig. 5 that increasing the number of observed nodes yields lower overall NMSE values.
Fig. 5. Overall NMSE versus $M$ for the temperature dataset ($\mu_{\text{LMS}} = 1.5$, $B_{\text{LMS}} = 10$, $\mu_{\text{DLSR}} = 1.2$, $B_{\text{DLSR}} = 4$, $\beta_{\text{DLSR}} = 0.5$).

Fig. 6. NMSE versus $t$ with $M = 20$ for the network delay dataset ($\mu_{\text{LMS}} = 1.5$, $B_{\text{LMS}} = 12$).

Fig. 7. True and estimated delay over an unobserved path ($\mu_{\text{LMS}} = 1.5$, $B_{\text{LMS}} = 12$).

Fig. 8. NMSE versus $t$ with $M = 53$ for the ECoG data ($\mu_{\text{LMS}} = 0.6$, $B_{\text{LMS}} = 2$, $\mu_{\text{DLSR}} = 1.2$, $B_{\text{DLSR}} = 6$, $\beta_{\text{DLSR}} = 0.5$).

Fig. 9. True and estimated brain signals over an unobserved node on the ECoG data ($\mu_{\text{LMS}} = 0.6$, $B_{\text{LMS}} = 2$, $\mu_{\text{DLSR}} = 1.2$, $B_{\text{DLSR}} = 6$, $\beta_{\text{DLSR}} = 0.5$).

for all approaches, among which the proposed EM-Static and EM-Dynamic stand out by handling multiple relations in a principled manner.

B. Network Delay Prediction

The second dataset we tested comprises measurements of delays over $N = 70$ paths on the Internet2backbone [2]. Each path connects 2 out of the 9 origin-destination nodes by a subset of the 26 directed links. The nodal process $x_t(n)$ represents the delay of path $n$ at slot $t$ in minutes. An MRG with two adjacency matrices $A^1$ and $A^2$ is built to characterize the interdependencies of different paths. While $A^1$ is given by the correlation
Graph using samples \( \{x_t\} \) that are not used to configure the experiment, \( A^2 \) is constructed based on the path-link routing matrix \( B \in \{0, 1\}^{70 \times 26} \), whose \((n, l)\)th entry \( B_{n,l} \) is 1, if path \( n \) traverses link \( l \), and 0 otherwise. Specifically, the \((n, n')\)th entry \((n \neq n')\) of \( A^2 \) is given by

\[
A^2(n, n') = \frac{\sum_{l=1}^{26} B_{n,l} B_{n',l}}{\sum_{l=1}^{26} B_{n,l} + \sum_{l=1}^{26} B_{n',l} - \sum_{l=1}^{26} B_{n,l} B_{n',l}}
\]

which places large weights for vertices (paths) with a large number of common links.

The dimension of the observation vector per slot is selected to be \( M = 20 \), and the total number of observations is \( T = 60 \). Assuming no temporal dynamics, the per-topology graph kernel in EM-Static is chosen to be a diffusion one with parameter \( a = 1 \). Accounting for temporal evolution, EM-Dynamic models \( x_t \) via (23) with \( F^s = 0.17(A^s + I_N) \) and \( K^s \) chosen as a diffusion kernel with \( a = 0.8 \). Fig. 6 demonstrates that EM-Dynamic markedly outperforms the competing alternatives in terms of NMSE. DSLR is not included here because it does not provide comparable estimation performance relative to the rest. It is worth highlights that without accounting for temporal dynamics EM-Static achieves comparable NMSE relative to KKF that relies on the structured spatio-temporal kernel matrix [26]. The estimated and true network delays from an unobserved path reported in Fig. 7 further validates the superior reconstruction performance of EM-Dynamic. For

C. Brain ECoG Dataset

The last experiment is designated to showcase the performance of the proposed methods for probabilistic multi-kernel learning over a single-relation graph. We experimented with brain ECoG data obtained from an epilepsy study [16]. The ECoG time series were obtained from \( N = 76 \) electrodes implanted in a patient’s brain before and after a seizure, where the onset of the seizure was identified by a neurophysiologist. We extracted 250 samples from the pre-ictal stage of the dataset, and then preprocess them by subtracting the sample mean and normalizing by the sample standard deviation. A time-invariant symmetric correlation graph was constructed.

![Algorithm 3: Kalman Smoother Parameters.](image)

The observations are generated as in (6) with \( M = 53 \), and \( R_t = 10^{-2} I_M \). To enrich expressiveness, EM-Static and EM-Dynamic associate the only adjacency matrix with a dictionary of diffusion kernels with parameter \( a \in \{0.6, 0.7, \ldots, 1\} \). Further, the state transition matrix in EM-Dynamic is chosen as \( F = 0.15(A + I_N) \). As shown in Figs. 8–9, EM-Dynamic enjoys the best extrapolation performance. Even without exploiting temporal dynamics, EM-Static outperforms the rest of the competing alternatives.

Discussion: Fig. 10 showcases the average running times of the competing methods using Matlab (R2019b) on a machine with an Intel Core-i5 CPU and 8 GB of RAM. It is worth mentioning that the CPU times of EM-Dynamic are almost twice those of EM-Static in all three datasets, in agreement with the complexity analysis in Remark 2. Although faster than DSLR, the proposed EM-based approaches are slower than LMS and KKF. As also pointed out in Remark 2, distributed operations are well motivated as a future research direction to scale up the advocated methods. Regarding the transition function in EM-Dynamic, the choice of \( F^s = c(A^s + I_N) \) used in the experiments promotes smoothness of nodal values across time and space. Future work will further explore other types of dynamics, as well as the problem of jointly identifying the transition model, and reconstructing the nodal values.
Algorithm 4: Kalman Smoother.

1: Input: $Y, \tilde{F}^{(n)}_t, \tilde{K}^{(n)}_t, R_t, H_t, t = 1, \ldots, T; \tilde{x}_{0|0}, P_{0|0}$
2: Step 1: Forward filtering
3: for $t = 1$ to $T$ do
4: $\tilde{x}_{t|t-1} = \tilde{F}^{(n)}_t \tilde{x}_{t-1|t-1}$
5: $P_{t|t-1} = \tilde{F}^{(n)}_t P_{t-1|t-1} \tilde{F}^{(n)}_t^\top + \tilde{K}^{(n)}_t$
6: $\hat{y}_{t|t-1} = H_t \tilde{x}_{t|t-1}$
7: $S_{t|t-1} = H_t P_{t|t-1} H_t^\top + R_t$
8: $G_t = P_{t|t-1} S_{t|t-1}^{-1}$
9: $\tilde{x}_t = \tilde{x}_{t|t-1} + G_t (y_t - \hat{y}_{t|t-1})$
10: $P_t = P_{t|t-1} - G_t S_{t|t-1}^{-1} G_t^\top$
end for

12: Step 2: Backward smoothing
13: $x_T^{(n)} = \tilde{x}_{T|T}$
14: for $t = T - 1$ to 1 do
15: $P_{t+1|t} = H_t^\top P_{t+1|t+1} H_t + S_t$
16: $x_t^{(n)} = \tilde{x}_t + M_t (x_{t+1|t}^{(n)} - \tilde{x}_{t+1|t})$
17: $\tilde{P}_t^{(n)} = P_t + M_t (P_{t+1|t} - P_{t+1|t+1}) M_t^\top$
18: end for
19: Output: $\tilde{x}_t^{(n)}, \tilde{P}_t^{(n)}, t = 1, \ldots, T$

VI. CONCLUSIONS

This contribution dealt with semi-supervised learning of both stationary and nonstationary signal information processes over multi-relational graphs. A novel spatio-temporal model was developed for the latter, and was leveraged for probabilistic reconstruction of space-time varying processes over graph nodes. To jointly infer unobserved graph signals and decipher the contributions from multiple relations of the MRG in both cases, two EM-based solvers were developed that enjoy closed-form updates per iteration that is guaranteed to converge at least to a local optimum. Further, the proposed methods are flexible to accommodate multiple kernels per relation that lend themselves to richer expressiveness. Real data experiments validated the superior performance of the novel approaches.

APPENDIX DERIVATION

This Appendix deals with the derivation of $\{\hat{F}_t, \hat{K}_t\}_{t=1}^T$ in the right-hand side (RHS) of (33). For notional brevity, the iteration index $n$ will be henceforth omitted. To start, let us rewrite the left-hand side (LHS) of (33) as

$$
\sum_{t=1}^{T} \sum_{s=1}^{S} w_{t,s} ||x_t - F^s x_{t-1}||^2_{K^s}
$$

$$
= \sum_{t=1}^{T-1} x_t^\top \left( \sum_{s=1}^{S} w_{t,s} K_{s}^{-1} + w_{t+1,s} F^s K_{s}^{-1} F^s \right) x_t
$$

$$
- 2 \sum_{t=1}^{T} x_t^\top \left( \sum_{s=1}^{S} w_{t,s} K_{s}^{-1} F^s \right) x_{t-1}
$$

where

$$
C_t := -K_t^{-1} \hat{F}_t
$$

$$
D_t := \left\{ \begin{array}{ll}
K_t^{-1} + \hat{F}_t K_t^{-1} \hat{F}_t & t = T \\
K_t^{-1} + \hat{F}_t K_t^{-1} \hat{F}_{t+1} & t \neq T
\end{array} \right.
$$

Equating $C_t$ with $C_t'$ and $D_t$ with $D_t'$, we can uniquely identify $\hat{F}_t$ and $\hat{K}_t$ backwards from $t = T$ to $t = 1$ via Alg. 3.

REFERENCES


