STELAR: Spatio-temporal Tensor Factorization with Latent Epidemiological Regularization

Nikos Kargas[1,2,*]**, Cheng Qian**[2,*]**, Nicholas D. Sidiropoulos**[3]**, Cao Xiao**[2]**,**
Lucas M. Glass[2,4]**, Jimeng Sun**[5]

[1] Dept. of ECE, University of Minnesota, [2] Analytics Center of Excellence, IQVIA, [3] Dept. of ECE, University of Virginia,
[4] Dept. of Statistics, Temple University [5] Dept. of CS, University of Illinois Urbana-Champaign

Abstract

Accurate prediction of the transmission of epidemic diseases such as COVID-19 is crucial for implementing effective mitigation measures. In this work, we develop a tensor method to predict the evolution of epidemic trends for many regions simultaneously. We construct a 3-way spatio-temporal tensor (location, attribute, time) of case counts and propose a nonnegative tensor factorization with latent epidemiological model regularization named STELAR. Unlike standard tensor factorization methods which cannot predict slabs ahead, STELAR enables long-term prediction by incorporating latent temporal regularization through a system of discrete-time difference equations of a widely adopted epidemiological model. We use *latent* instead of location/attribute-level epidemiological dynamics to capture common epidemic profile sub-types and improve collaborative learning and prediction. We conduct experiments using both county- and state-level COVID-19 data and show that our model can identify interesting latent patterns of the epidemic. Finally, we evaluate the predictive ability of our method and show superior performance compared to the baselines, achieving up to 21% lower root mean square error and 25% lower mean absolute error for county-level prediction.

Introduction

Pandemic diseases such as the novel coronavirus disease (COVID-19) pose a serious threat to global public health, economy, and daily life. Accurate epidemiological measurement, modeling, and tracking are needed to inform public health officials, government executives, policy makers, emergency responders, and the public at large. Two types of epidemiological modeling methods are popular today:

- **Mechanistic models** define a set of ordinary differential equations (ODEs) which capture the epidemic transmission patterns and predict the long-term trajectory of the outbreak. These models include the Susceptible-Infected-Recovered (SIR) model (Kermack and McKendrick 1927), the Susceptible-Exposed-Infected-Recovered (SEIR) (Cooke and Van Den Driessche 1996) and their variants e.g, SIS, SIRS, and delayed SIR. These models have a small number of parameters

which are determined via curve fitting. These types of models do not require much training data, but they are quite restrictive and cannot leverage rich information.

- **Machine learning models** such as deep learning, model the epidemiological trends as a set of regression problems (Yang et al. 2020; Toda 2020; He, Peng, and Sun 2020; Chimmula and Zhang 2020; Tomar and Gupta 2020). These models are able to learn the trajectory of the outbreak only from data and can perform very well especially in short-term prediction. However, they usually require a large amount of training data.

This paper extends Canonical Polyadic Decomposition (CPD) (Harshman 1970) via epidemiological model (e.g., SIR and SEIR) regularization, which integrates spatio-temporal real-world case data and exploits the correlation between different regions for long-term epidemic prediction. CPD is a powerful tensor model with successful applications in many fields (Papalexakis, Faloutsos, and Sidiropoulos 2016; Sidiropoulos et al. 2017). Compared to traditional epidemiological models which cannot incorporate fine-grain observations or any kind of side information, tensors offer a natural way of representing multidimensional time evolving data and incorporate additional information (Acar, Dunlavy, and Kolda 2009; Araujo, Ribeiro, and Faloutsos 2017). CPD can capture the inherent correlations between the different modes and thanks to its uniqueness properties it can extract interpretable latent components rendering it an appealing solution for modeling and analyzing epidemic dynamics. Additionally, it can parsimoniously represent multidimensional data and can therefore learn from limited data. These two important properties differentiate CPD from neural networks and deep learning models which typically require a lot of training data and are often treated as black box models.

In this work, we propose a Spatio-temporal Tensor factorization with EpidemioLogicAl Regularization (STELAR). STELAR combines a nonnegative CPD model with a system of discrete-time difference equations to capture the epidemic transmission patterns. Unlike standard tensor factorization methods which cannot predict slabs ahead, STELAR can simultaneously forecast the evolution of the epidemic for a list of regions, and it can also perform long-term prediction.

In the experiments, we build a spatio-temporal tensor, location $\times$ attribute $\times$ time of case counts based on large

---

*equal contribution

real-world medical claims datasets, where the first dimension corresponds to different counties/states, the second dimension corresponds to different attributes or signals that evolve over time, such as daily new infections, deaths, number of hospitalized patients and other COVID-19 related signals, and the third is the time-window of the available signals. This spatio-temporal tensor is factorized using the proposed low-rank nonnegative tensor factorization with epidemiological model regularization STELAR. We observed that the extracted latent time components provide intuitive interpretation of different epidemic transmission patterns which traditional epidemiological models such as SIR and SEIR are lacking.

Our main contributions are summarized as follows:

- We propose STELAR, a new data-efficient tensor factorization method regularized by a disease transmission model in the latent domain to predict future slabs. We show that by jointly fitting a low-rank nonnegative CPD model with an SIR model on the time factor matrix we are able to accurately predict the evolution of epidemic trends.

- Thanks to the uniqueness properties of the CPD, our method produces interpretable prediction results. Specifically, the tensor is approximated via K rank-1 components, each of which is associated with a sub-type of the epidemic transmission patterns in a given list of regions. The latent time factor matrix includes K different patterns of the epidemic evolution and using the latent location and signal factor matrices we can identify the corresponding locations and signals associated with each pattern.

- We perform extensive experimental evaluation on both county- and state-level COVID-19 case data for 10 and 15 days-ahead prediction and show that our method outperforms standard epidemiological and machine learning models in COVID-19 pandemic prediction. Our method achieves up to 21% lower root mean square error and 25% lower mean absolute error for county-level prediction.

Related Work

Many epidemiological and deep learning models have been applied for modeling the COVID-19 pandemic evolution. Methods based on traditional epidemic prediction models such as SIR (Kermack and McKendrick 1927) and SEIR and their variants (Cooke and Van Den Driessche 1996), rely on a system of differential equations which describe the dynamics of the pandemic (Yang et al. 2020; Toda 2020; He, Peng, and Sun 2020). These models are trained for each location independently by curve fitting.

Machine learning approaches and specifically deep learning models treat the problem as a time series prediction problem. Given as input a time series of length L_w, $x_{t-L_w-1}, \ldots, x_{t-1}, x_t$, the goal is to output L_o future time points $x_{t+1}, x_{t+2}, \ldots, x_{L_o}$. A popular and very successful method for time series prediction is the Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber 1997). Recent studies have applied LSTMs for COVID-19 prediction (Chimmula and Zhang 2020; Tomar and Gupta

Notation	Description
$\underline{\mathbf{X}} \in \mathbb{R}^{M \times N \times L}$	spatio-temporal tensor
$\mathbf{A} \in \mathbb{R}^{M \times K}$	location factor matrix
$\mathbf{B} \in \mathbb{R}^{N \times K}$	signal factor matrix
$\mathbf{C} \in \mathbb{R}^{L \times K}$	temporal factor matrix
$\underline{\mathbf{X}}^{(i)}$	mode-i unfolding of $\underline{\mathbf{X}}$
$\beta; \gamma;$	contact rate; recovery rate
K	# of components
M	# of locations
N	# of signals
L	# of time points
T	transpose
$\circ$	outer product
$\otimes$	Kronecker product
$\odot$	Khatri-Rao product
$\circledast$	Hadamard product
$\| \cdot \|_F$	Frobenius norm
$\mathrm{diag}(\mathbf{x})$	diagonal matrix

Table 1: Notation

2020; Yang et al. 2020). To increase the data for LSTM training, Yang et al. used the 2003 SARS-CoV data to pretrain the model before using it for COVID-19 prediction. DEFSI (Wang, Chen, and Marathe 2019) is a deep learning method based on LSTMs for influenza like illness forecasting which combines synthetic epidemic data with real data to improve epidemic forecasts.

Graph Neural Network (GNN) (Kipf and Welling 2017) is a type of neural network which operates on a graph. Kapoor et al. proposed creating a graph with spatial and temporal edges. They leverage human mobility data to improve the prediction of daily new infections (Kapoor et al. 2020). STAN (Gao et al. 2020) is an attention-based graph convolutional network which constructs edges based on geographical proximity of the different regions and regularizes the model predictions based on an epidemiological model.

Methods based on tensor factorization have been applied for various time series prediction tasks. CP Forecasting (Dunlavy, Kolda, and Acar 2011), is a tensor method which computes a low-rank CPD model and uses the temporal factor to capture periodic patterns in data. TENSORCAST (Araujo, Ribeiro, and Faloutsos 2017) is a method that forecasts time-evolving networks using coupled tensors. Both methods are not suitable for COVID-19 pandemic prediction and rely on a two-step procedure – fitting a low-rank model and then performing forecasting based on the temporal factor matrix – which as we will see later leads to performance degradation relative to our joint optimization formulation.

Background

In this section, we review necessary background on epidemiological models and tensor decomposition that will prove useful for developing our method. Table 1 contains the notation used throughout the paper.

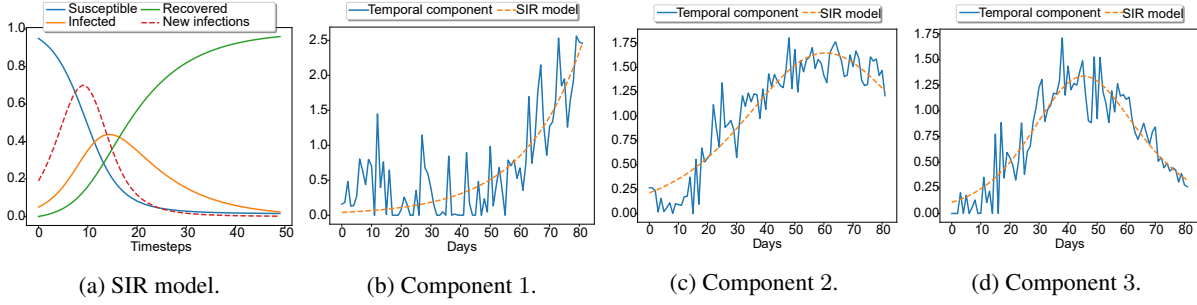(a) SIR model.　　(b) Component 1.　　(c) Component 2.　　(d) Component 3.

Figure 1: An example of an SIR model with parameters $N = 1, S(0) = 0.95, I(0) = 0.05, R(0) = 0, \beta = 0.4, \gamma = 0.1$ is shown in Figure 1a. The new infections curve has been scaled by 10 for visualization purposes. Figures 1b, 1c, 1d show 3 latent time components i.e., columns of the factor matrix $\mathbf{C}$ when we compute a plain nonnegative CPD of the spatio-temporal tensor.

Epidemiological Models

Epidemiological models have been popular solutions for pandemic modeling. For example, The SIR model (Kermack and McKendrick 1927) is one of the most famous and paradigmatic models in mathematical epidemiology. In this model, a population is divided into susceptible, infected and recovered subpopulations. The evolution of these quantities is described by the following equations:

$$S(t) - S(t-1) = -\beta S(t-1)I(t-1)/N, \tag{1a}$$

$$I(t) - I(t-1) = \frac{\beta S(t-1)I(t-1)}{N} - \gamma I(t-1), \tag{1b}$$

$$R(t) - R(t-1) = \gamma I(t-1), \tag{1c}$$

where $S(t)$, $I(t)$ and $R(t)$ stand for the size of susceptible, infected and recovered populations at time t, respectively and $N = S(t) + I(t) + R(t)$ is the total population. The parameter β controls the rate of spread and the parameter γ is the recovery rate, so its inverse $1/\gamma$ represents the average time period that an infected individual remains infectious. In Equation (1a), the quantity $\beta S(t)/N$ is the fraction of susceptible individuals that will contact an infected individual at time t. Therefore,

$$C(t) := \beta S(t)I(t)/N, \tag{2}$$

denotes the new infections at time t. In Equation (1b), $\gamma I(t)$ denotes the number of individuals recovered. Therefore, the change of the infected cases at time t will be the difference between new infected and recovered cases as shown in Equation (1b). Given $S(0), I(0)$ and $R(0)$, we can compute $S(t), I(t)$ and $R(t)$ for $t = 1, 2, \ldots, L$, where L is the size of the prediction window we are interested in. An example is shown in Figure (1a) where we have set $N = 1, S(0) = 0.95, I(0) = 0.05, R(0) = 0, \beta = 0.4, \gamma = 0.1$ i.e., we have normalized the different populations by the total population number N and set the window size to be $L = 50$.

The SEIR model is a variant of the SIR model. In this model, a population is divided into susceptible, exposed, infected and recovered subpopulations. The exposed population consists of individuals who have been infected but are not yet infectious. The SEIR model includes an additional parameter σ which is the rate at which the exposed individuals becoming infectious.

Canonical Polyadic Decomposition

The Canonical Polyadic Decomposition (CPD) expresses a 3-way tensor $\underline{\mathbf{X}} \in \mathbb{R}^{M \times N \times L}$ as a sum of rank-1 components, i.e.,

$$\underline{\mathbf{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{k=1}^{K} \mathbf{a}_k \circ \mathbf{b}_k \circ \mathbf{c}_k, \tag{3}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K] \in \mathbb{R}^{M \times K}, \mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_K] \in \mathbb{R}^{N \times K}, \mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K] \in \mathbb{R}^{L \times K}$. The rank K is the minimum number of components needed to synthesize $\underline{\mathbf{X}}$. We can express the CPD of a tensor in many different ways. For example, $\underline{\mathbf{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ can be represented by the matrix unfolding $\underline{\mathbf{X}}^{(1)} = (\mathbf{C} \odot \mathbf{B})\mathbf{A}^T$ where the mode-1 fibers are the rows of the resulting matrix. Using 'role symmetry', the mode-2 and mode-3 matrix unfoldings are given by $\underline{\mathbf{X}}^{(2)} = (\mathbf{C} \odot \mathbf{A})\mathbf{B}^T, \underline{\mathbf{X}}^{(3)} = (\mathbf{B} \odot \mathbf{A})\mathbf{C}^T$ respectively. CPD can parsimoniously represent tensors of size $M \times N \times L$ using only $(M + N + L) \times K$ parameters. The CPD model has two very important properties that make it a very powerful tool for data analysis 1) it is universal, i.e., every tensor admits a CPD of finite rank, and 2) it is unique under mild conditions i.e., it is possible to extract the true latent factors that synthesize $\underline{\mathbf{X}}$.

Method

Problem Formulation

Consider a location where we monitor N signals related to a pandemic over time, e.g., number of new infections, hospitalized patients, intensive care unit (ICU) patients, etc. At time t, the value of the nth signal at location m is denoted by $x_{m,n,t}$. Assuming that there are N signals, M locations and L time points, then, the dataset can be naturally described by a 3-way spatio-temporal tensor $\underline{\mathbf{X}} \in \mathbb{R}^{M \times N \times L}$, where $\underline{\mathbf{X}}(m, n, t) := x_{m,n,t}$. Tensor $\underline{\mathbf{X}}$ includes the evolution of all signals and regions for times 1 through L and we are interested in estimating the signals at time $L+1, \ldots, L+L_o$. In other words, we want to predict the frontal slabs $\underline{\mathbf{X}}(:, :, t)$ for L_o timesteps ahead. However, standard tensor factorization methods cannot predict slabs ahead. It is evident from Equation 3 and the mode-3 unfolding that is impossible to impute

when an entire slab $\underline{\mathbf{X}}(:,:,t)$ is missing since we have no information regarding the corresponding row of $\mathbf{C}$. To address this challenge, we take into consideration the transmission law dynamics of the disease. The key idea is to decompose the tensor using a CPD model and impose SIR constraints on the latent time factor.

To illustrate the key idea, we perform a preliminary experiment using a rank-5 plain nonnegative CPD on a spatio-temporal tensor $\underline{\mathbf{X}} \in \mathbb{R}^{140 \times 15 \times 82}$ with case counts, constructed from real COVID-19 data. As shown in Figures 1b, 1c, 1d, CPD is able to extract meaningful latent components from the data. Specifically, the figures show 3 columns of the latent time factor $\mathbf{C}$ where an SIR model has been fitted after obtaining the decomposition. We observe that each columns depicts a curve similar to the one in Figure 1a i.e., the CPD can unveil the principal patterns of the epidemic evolution, and each pattern corresponds to a different pandemic phase.

Therefore, we propose solving the following constrained nonnegative CPD problem

$$
\min_{\substack{\mathbf{A},\mathbf{B},\mathbf{C}, \\ \boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{s},\mathbf{i}}} \|\underline{\mathbf{X}} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]\|_F^2 + \mu \left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2 \right)
$$

$$
+ \nu \sum_{k=1}^{K} \sum_{t=1}^{L} \left(c_{t,k} - \beta_k S_k(t-1) I_k(t-1) \right)^2
$$

$$
\begin{aligned}
\text{s. t. } & \mathbf{A} \geq \mathbf{0}, \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \\
& \boldsymbol{\beta} \geq \mathbf{0}, \boldsymbol{\gamma} \geq \mathbf{0}, \mathbf{s} \geq \mathbf{0}, \mathbf{i} \geq \mathbf{0}, \qquad (4)\\
& S_k(t) = S_k(t-1) - \beta_k S_k(t-1) I_k(t-1), \\
& I_k(t) = I_k(t-1) + \beta_k S_k(t-1) I_k(t-1) \\
& \qquad - \gamma_k I_k(t-1), \\
& s_k = S_k(0), i_k = I_k(0).
\end{aligned}
$$

The first term is the data fitting term. We fit a CPD model of rank-K with nonnegativity constraints on the factors. The second term is Frobenius norm regularization which is typically used to avoid overfitting and improve generalization of the model. We introduce a third term which regularizes each column of the factor matrix $\mathbf{C}$ according to Equation (2) i.e., we learn K different SIR models, each of which is fully described by parameters $S_k(0), I_k(0), R_k(0)$ and β_k, γ_k according to Equations (1a), (1b), (1c). We aim at estimating these parameters such that each column of $\mathbf{C}$ follows the new infections curve of an SIR model. Vectors $\boldsymbol{\beta} \in \mathbb{R}^K$ and $\boldsymbol{\gamma} \in \mathbb{R}^K$ hold the β and γ parameters of each SIR model and vectors $\mathbf{s} = \in \mathbb{R}^K, \mathbf{i} \in \mathbb{R}^K$ hold the initial conditions of each SIR model. Note that the recovered cases curve can be safely removed because it does not affect the optimization cost and that the parameter N_k corresponding to the total population has been absorbed in the vector $\boldsymbol{\beta}$.

Prediction

After the convergence of the above optimization algorithm, we have some estimates of $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ and parameters $\{\beta_1, \cdots, \beta_K\}, \{\gamma_1, \cdots, \gamma_K\}, \{s_1, \cdots, s_K\}, \{i_1, \cdots, i_K\}$ where the pair of β_k and γ_k describes the epidemic transmission of the kth component in the time factor matrix and

s_k, i_k the initial values of the subpopulations. Using Equations (1a),(1b) and (2) we can predict "future" values for the kth column of $\mathbf{C}$. We repeat the same procedure for all columns of $\mathbf{C}$ such that we can predict the entire "future" rows. Let $\hat{\mathbf{C}}(t,:) \in \mathbb{R}^K$ be the prediction of the temporal information at a future time point t using estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\mathbf{s}}, \hat{\mathbf{i}}$. Since $\mathbf{A}$ and $\mathbf{B}$ do not depend on t, the prediction of all signals in the tensor at time t is given by

$$
\widehat{\underline{\mathbf{X}}}(:,:,t) = \hat{\mathbf{A}} \text{diag}(\hat{\mathbf{C}}(t,:)) \hat{\mathbf{B}}^T. \qquad (5)
$$

Adding latent temporal regularization through an SIR model offers significant advantages compared to having separate SIR models for each location. Our model can capture correlations between different locations and signals through their latent representations and therefore improve the prediction accuracy. Additionally, it enables expressing the evolution of a signal as weighted sum of K separate SIR models e.g., the prediction of the nth signal for the mth location for time point t, is given by

$$
\widehat{\underline{\mathbf{X}}}(m,n,t) = \sum_{k=1}^{K} \hat{a}_{m,k} \hat{b}_{n,k} \hat{\beta}_k S_k(t-1) I_k(t-1), \qquad (6)
$$

which makes it much more flexible and expressive.

Optimization

The optimization problem (4) is a nonconvex and very challenging optimization problem. To update factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ we rely on alternating optimization. Note that by fixing all variables except for $\mathbf{A}$, the resulting subproblem w.r.t. $\mathbf{A}$, is a nonnegative least squares problem, which is convex. Similarly for the factor matrices $\mathbf{B}$ and $\mathbf{C}$. We choose to solve each factor matrix subproblem via the Alternating Direction Method of Multipliers (ADMM) (Gabay and Mercier 1976), which is a very efficient algorithm that has been successfully applied to nonnegative tensor factorization problems (Huang, Sidiropoulos, and Liavas 2016).

Let us first consider the subproblem w.r.t. $\mathbf{A}$. Assume that at the ℓth iteration, we have some estimates of all the variables available. Fixing all variables except for $\mathbf{A}$, we have

$$
\min_{\mathbf{A} \geq \mathbf{0}} \|\underline{\mathbf{X}}^{(1)} - \boldsymbol{\Phi}_A^{(\ell)} \mathbf{A}^T\|_F^2 + \mu \|\mathbf{A}\|_F^2 \qquad (7)
$$

where $\boldsymbol{\Phi}_A^{(\ell)} = \mathbf{C}^{(\ell)} \odot \mathbf{B}^{(\ell)}$. The ADMM updates for optimization problem (7) are the following

$$
\widehat{\mathbf{A}} = \arg\min \|\underline{\mathbf{X}}^{(1)} - \boldsymbol{\Phi}_A^{(\ell)} \widehat{\mathbf{A}}\|_F^2 + \mu \|\widehat{\mathbf{A}}\|_F^2 \qquad (8a)
$$

$$
+ \rho \|\mathbf{A} - \widehat{\mathbf{A}}^T + \mathbf{A}_d\|_F^2,
$$

$$
\mathbf{A} = \arg\min_{\mathbf{A} \geq \mathbf{0}} \rho \|\mathbf{A} - \widehat{\mathbf{A}}^T + \mathbf{A}_d\|_F^2, \qquad (8b)
$$

$$
\mathbf{A}_d = \mathbf{A}_d + \mathbf{A} - \widehat{\mathbf{A}}^T. \qquad (8c)
$$

Equation (8a) is a least squares problem. Because ADMM is an iterative algorithm and the update is performed multiple times, we save computations by caching (Huang, Sidiropoulos, and Liavas 2016). Equation (8b) is a simple elementwise nonnegative projection operator and Equation (8c) is

Algorithm 1 `STELAR` Method

Input: Tensor $\underline{\mathbf{X}}$, rank K, max. outer iterations iters$_{\text{outer}}$, max. inner iterations iters$_{\text{inner}}$, gradient steps iters$_{\text{grad}}$, prediction window L_o

repeat

    Update $\mathbf{A}$ using (8) for iters$_{\text{inner}}$ iterations

    Update $\mathbf{B}$ using (9) for iters$_{\text{inner}}$ iterations

    Update $\mathbf{C}$ using (11) for iters$_{\text{inner}}$ iterations

    Perform iters$_{\text{grad}}$ projected gradient steps for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\mathbf{s}$, $\mathbf{i}$, using (14), (15), (16), (17)

until iters$_{\text{outer}}$ or validation RMSE increases

Predict L_o future slabs using Equation (5)

---

the dual variable update. The updates for $\mathbf{B}$ are similar.

$$\widehat{\mathbf{B}} = \arg\min \|\underline{\mathbf{X}}^{(2)} - \boldsymbol{\Phi}_B^{(\ell)}\widehat{\mathbf{B}}\|_F^2 + \mu\|\widehat{\mathbf{B}}\|_F^2 \quad (9a)$$
$$+ \rho\|\mathbf{B} - \widehat{\mathbf{B}}^T + \mathbf{B}_d\|_F^2,$$

$$\mathbf{B} = \arg\min_{\mathbf{B} \geq \mathbf{0}} \rho\|\mathbf{B} - \widehat{\mathbf{B}}^T + \mathbf{B}_d\|_F^2, \quad (9b)$$

$$\mathbf{B}_d = \mathbf{B}_d + \mathbf{B} - \widehat{\mathbf{B}}^T. \quad (9c)$$

where $\boldsymbol{\Phi}_B^{(\ell)} = \mathbf{C}^{(\ell)} \odot \mathbf{A}^{(\ell)}$. Now let us consider the update of $\mathbf{C}$. The related optimization problem takes the form of

$$\min_{\mathbf{C} \geq 0} \|\underline{\mathbf{X}}^{(3)} - \boldsymbol{\Phi}_C^{(\ell)}\mathbf{C}^T\|_F^2 + \mu\|\mathbf{C}\|_F^2 + \nu\|\mathbf{C} - \bar{\mathbf{C}}^{(\ell)}\|_F^2, \quad (10)$$

where $\boldsymbol{\Phi}_C^{(\ell)} = \mathbf{B}^{(\ell)} \odot \mathbf{A}^{(\ell)}$, $\bar{\mathbf{C}}^{(\ell)} = (\mathbf{P}^{(\ell)} \circledast \mathbf{Q}^{(\ell)})\text{diag}(\boldsymbol{\beta}^{(\ell)})$ and we define $\mathbf{P}^{(\ell)}(t,k) := S_k(t-1)$ and $\mathbf{Q}^{(\ell)}(t,k) := I_k(t-1)$. Optimization problem (10) is also a nonnegative least squares problem. Therefore the updates for $\mathbf{C}$ are

$$\widehat{\mathbf{C}} = \arg\min \|\underline{\mathbf{X}}^{(3)} - \boldsymbol{\Phi}_C^{(\ell)}\widehat{\mathbf{C}}\|_2^2 + \mu\|\widehat{\mathbf{C}}\|_F^2 \quad (11a)$$
$$+ \nu\|\mathbf{C} - \bar{\mathbf{C}}^{(\ell)}\|_F^2 + \rho\|\mathbf{C} - \widehat{\mathbf{C}}^T + \mathbf{C}_d\|_F^2$$

$$\mathbf{C} = \arg\min_{\mathbf{C} \geq \mathbf{0}} \rho\|\mathbf{C} - \widehat{\mathbf{C}}^T + \mathbf{C}_d\|_F^2 \quad (11b)$$

$$\mathbf{C}_d = \mathbf{C}_d + \mathbf{C} - \widehat{\mathbf{C}}^T \quad (11c)$$

We observed that running a few ADMM inner iterations (~ 10) for each factor suffices for the algorithm to produce satisfactory results. By fixing the factor matrices, we have

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{s}, \mathbf{i}} \sum_{k=1}^{K}\sum_{t=1}^{L} (c_{t,k} - \beta_k S_k(t-1)I_k(t-1))^2$$

$$\text{s. t. } \boldsymbol{\beta} \geq \mathbf{0}, \boldsymbol{\gamma} \geq \mathbf{0}, \mathbf{s} \geq \mathbf{0}, \mathbf{i} \geq \mathbf{0},$$
$$S_k(t) = S_k(t-1) - \beta_k S_k(t-1)I_k(t-1), \quad (12)$$
$$I_k(t) = I_k(t-1) + \beta_k S_k(t-1)I_k(t-1)$$
$$- \gamma_k I_k(t-1),$$
$$s_k = S_k(0), i_k = I_k(0).$$

Both $S_k(t)$ and $I_k(t)$ are functions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and are calculated in a recursive manner. Therefore the optimization problem w.r.t. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is nonconvex. Optimization problem (12) corresponds to K independent one-dimensional curve fitting problems and since there are only 4 optimization variables for each problem, we can use off-the-shelf curve fitting methods. Alternatively, we can perform a few projected

gradient descent steps. Focusing on the kth subproblem we have

$$f(\beta_k) = \nu \sum_{t=1}^{L} (c_{t,k} - \beta_k S_k(t-1)I_k(t-1))^2. \quad (13)$$

The derivative of $f(\beta_k)$ w.r.t. β_k is

$$\frac{\partial f}{\partial \beta_k} = -2\nu \sum_{t=1}^{L} (c_{t,k} - \beta_k S_k(t-1)I_k(t-1)) \times$$
$$\left(S_k(t-1)I_k(t-1) + \beta_k \frac{\partial S_k(t-1)}{\partial \beta_k}I_k(t-1) + \right. \quad (14)$$
$$\left. \beta_k S_{t-1,k}\frac{\partial I_{t-1,k}}{\partial \beta_k} \right).$$

Note that both $S_k(t)$ and $I_k(t)$ are recursive functions. Thus, their respective derivatives w.r.t. β are computed recursively in L steps. Similarly, for γ_k we have

$$\frac{\partial f}{\partial \gamma_k} = -2\nu \sum_{t=1}^{L} (c_{t,k} - \beta_k S_k(t-1)I_k(t-1)) \times$$
$$\left(\beta_k \frac{\partial S_{t-1}}{\partial \gamma}I_{t-1} + \beta_k S_k(t-1)\frac{\partial I_k(t-1)}{\partial \gamma} \right). \quad (15)$$

Finally for $\mathbf{s}, \mathbf{i}$

$$\frac{\partial f}{\partial s_1} = -2\nu \sum_{t=1}^{L} (c_{t,k} - \beta_k S_k(t-1)I_k(t-1)) \times$$
$$\left(\beta_k \frac{\partial S_k(t-1)}{\partial s_1}I_k(t-1) + \beta_k S_k(t-1)\frac{\partial I_k(t-1)}{\partial s_1} \right),$$
$$\quad (16)$$
$$\frac{\partial f}{\partial i_1} = -2\nu \sum_{t=1}^{L} (c_{t,k} - \beta_k S_k(t-1)I_k(t-1)) \times$$
$$\left(\beta_k \frac{\partial S_k(t-1)}{\partial i_1}I_k(t-1) + \beta_k S_k(t-1)\frac{\partial I_k(t-1)}{\partial i_1} \right).$$
$$\quad (17)$$

The overall procedure is summarized in Algorithm 1.

Experiments

Dataset and Baselines

We use US county-level data from the Johns Hopkins University (JHU) (Dong, Du, and Gardner 2020) and a large IQVIA patient claims dataset, which can be publicly accessible upon request. JHU data includes the number of active cases, confirmed cases and deaths for different counties in the US. The total number of counties was 133. We use the reported active cases to compute the daily new infections. The claims dataset was created from 582,2748 claims from 732,269 COVID-19 patients from 03-24-2020 to 06-26-2020 (95 days). It contains the daily counts of 12 International Classification of Diseases ICD-10 codes observed in each county and the Current Procedural Terminology (CPT) codes related to hospitalization and utilization of intensive care unit (ICU). The size of the constructed tensor is $133 \times 15 \times 95$. Using this data, we also construct a tensor

Model	$L_o = 10$		$L_o = 15$	
	RMSE	MAE	RMSE	MAE
Mean	304.1	122.0	269.5	108.5
SIR	156.2	62.2	159.1	63.6
SEIR	177.1	72.9	163.2	69.7
LSTM (w/o feat.)	203.6	77.1	191.0	81.7
LSTM (w/ feat.)	162.3	68.2	187.6	78.3
STAN	164.2	61.1	152.6	61.8
STELAR ($\nu = 0$)	149.2	61.5	152.8	66.9
STELAR	**127.5**	**55.6**	**136.1**	**61.7**

Model	$L_o = 10$		$L_o = 15$	
	RMSE	MAE	RMSE	MAE
Mean	125.0	77.0	123.3	77.1
SIR	46.5	27.2	48.7	27.7
SEIR	39.1	23.9	41.1	25.7
LSTM (w/o feat.)	45.6	23.6	54.8	31.2
LSTM (w/ feat.)	42.5	23.3	47.5	26.8
STAN	30.6	17.3	42.8	24.2
STELAR ($\nu = 0$)	28.6	16.6	46.8	21.1
STELAR	**24.0**	**15.1**	**36.0**	**18.0**

Table 2: County-level prediction for new infections (left) and hospitalized patients (right).

Model	$L_o = 10$		$L_o = 15$	
	RMSE	MAE	RMSE	MAE
Mean	309.0	258.7	325.8	273.1
SIR	186.1	133.8	186.9	134.5
SEIR	162.4	127.0	162.6	130.2
LSTM (w/o feat.)	187.5	138.1	419.7	356.0
LSTM (w/ feat.)	197.9	151.6	359.2	286.5
STAN	**74.1**	**60.1**	**100.5**	79.6
STELAR ($\nu = 0$)	140.8	104.0	127.8	95.0
STELAR	117.8	89.8	107.3	**79.4**

Model	$L_o = 10$		$L_o = 15$	
	RMSE	MAE	RMSE	MAE
Mean	685.5	553.9	729.2	586.0
SIR	343.4	252.4	367.8	266.0
SEIR	109.0	97.6	192.5	154.3
LSTM (w/o feat.)	280.9	187.5	416.1	308.7
LSTM (w/ feat.)	295.3	182.3	276.0	208.5
STAN	100.3	73.6	177.7	144.7
STELAR ($\nu = 0$)	118.3	84.4	126.6	**75.2**
STELAR	**56.8**	**43.9**	**113.6**	83.8

Table 3: State-level prediction for new infections (left) and hospitalized patients (right).

which includes the aggregated data for 5 states, New York, New Jersey, Massachusetts, Connecticut, Pennsylvania and is of size $5 \times 15 \times 95$.

We perform two different experiments. Initially we use 85 days for training and validation and use the remaining $L_o = 10$ as the test set. In the second experiment we use 80 days for training and validation and $L_o = 15$ days for test. We compare our method against the following baselines:

1. **Mean**. We use the mean of the last 5 days of the training set as our prediction.

2. **SIR**. The susceptible-infected-removed model.

3. **SEIR**. The susceptible-exposed-infected-removed epidemiological model.

4. **LSTM (w/o feat.)** LSTM model without additional features. We use one type of time series as our input.

5. **LSTM (w feat.)** LSTM with additional features. We use 15 different time series as input.

6. **STAN** (Gao et al. 2020) A recently proposed GNN with attention mechanism.

7. **STELAR** ($\nu = 0$) The proposed method without regularization (two-step approach).

We use the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as our evaluation metrics.

Results

Table 2 shows the county-level results for 10 and 15 day prediction for daily new infections and hospitalized patients. For new infections and $L_o = 10$, our method achieves 18%

lower RMSE and 9% lower MAE compared to the best performing baselines which are the SIR and STAN model respectively. When $L_o = 15$ our method achieves 10% lower RMSE and the same MAE compared to the STAN model. For hospitalized patients and $L_o = 10$, our method achieves 21% lower RMSE and 12% lower MAE compared to the best performing baseline which is STAN. When $L_o = 15$ our method achieves 12% lower RMSE and 25% lower MAE compared to the best baseline. For state-level prediction of new infections, the best performing model is STAN but for hospitalized patients our method again outperforms all the baselines. Note that in all cases except one, joint optimization and SIR model fitting always improves the performance of our method compared to the two-step procedure.

Figure 2 shows some examples of county-level prediction. Figure 2a shows simple case where one can observe an increasing pattern of the new infections. Almost all models are able to capture this trend except for the LSTM (w/feat.) model. Figure 2b shows a more challenging scenario where it is not obvious if the curve will continue increasing but our method makes an accurate prediction, and is better than the baselines. Finally in Figure 2c we observe an example where a peak was already observed in the past and therefore SIR and SEIR models fail. On the other hand, our method is again able to make reasonable predictions.

Finally, we demonstrate the ability of our model to produce interpretable results. We train our model on county-level data using $K = 30$. After the algorithm converges we normalize each factor matrix such that each column has unit norm and absorb the scaling to a vector w. Using this vector, we extract the 3 rank-1 components with the highest weights
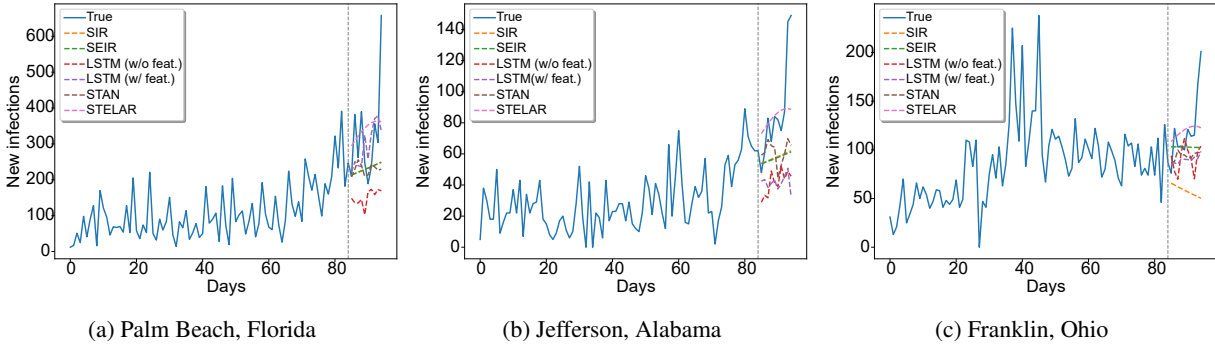
(a) Palm Beach, Florida   (b) Jefferson, Alabama   (c) Franklin, Ohio

Figure 2: 10 days county-level prediction for new infections.

Component 1	Component 2	Component 3
New York (NY)	L.A (CA)	Nassau (NY)
Westchester (NY)	Cook (IL)	L.A (CA)
Nassau (NY)	Milwaukee (WI)	Essex (NJ)
Bergen (NJ)	Fairfax (VA)	Wayne (MI)
Miami-Dade (FL)	Hennepin (MN)	Oakland (MI)
Hudson (NJ)	Montg. (MD)	Middlesex (NJ)
Union (NJ)	P. George's (MD)	New York (NY)
Phila. (PA)	Dallas (TX)	Phila. (PA)
Passaic (NJ)	Orange (CA)	Cook (IL)
Essex (NJ)	Harris (TX)	Bergen (NJ)

Table 4: Counties that contribute more to each of the strongest 3 rank-1 components of a rank-30 STELAR model.

Figure 3: The strongest 3 temporal components of a rank-30 STELAR model.

i.e., by selecting the corresponding columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$. Figure 3 depicts the 3 temporal profiles (columns of factor $\mathbf{C}$) that corresponds to the selected rank-1 components. We observe that the different curves depict different stages of the disease transmission. For example, the 1st component has captured an initial wave of the disease, the 2nd has captured a more recent increasing trend and the 3rd component is very similar to the first but shifted to the right. To gain some intuition regarding the results, we sort each column of factor matrices $\mathbf{A}$ (counties) and $\mathbf{B}$ (signals) and present the top 10 counties and top 5 signals that contribute to each one of the 3 strongest components in Tables 4, 5. We can clearly see that the counties which correspond to the early wave (1st component) are from the states of New York and New Jersey as

Component 1	Component 2	Component 3
New infections	New infections	Hosp. patients
Hosp. patients	Hosp. patients	ICU patients
ICU patients	ICU patients	J96
J96	J96	N17
R09	R05	R06

Table 5: Signals that contribute more to each of the strongest 3 rank-1 components of a rank-30 STELAR model. Definitions of ICD-10 codes: J96–Respiratory failure, not elsewhere classified; N17–Acute kidney failure; R05–Cough. R06–Abnormalities of breathing; R09–Other symptoms and signs involving the circulatory and respiratory system.
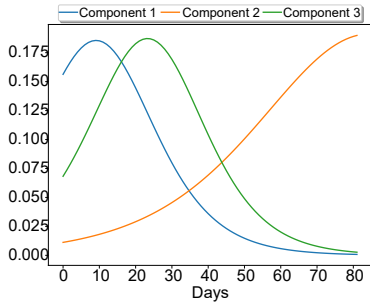
we would expect. Also the 1st component is mostly associated with new infections, hospitalized patients and ICU. On the other hand, the 3rd component which is very similar to the 1st is associated with hospitalized patients and ICU but not with new infections. Some counties appear in both the 1st and 3rd component which means that hospitalized patients and ICU cases started to increase (or getting reported) slightly after the new infections were reported. Finally, the second component depicts a later increase to the number of infections and hospitalized patients for some counties.

Conclusion

In this paper, we propose STELAR a data efficient and interpretable method based on constrained nonnegative tensor factorization. Unlike standard tensor factorization methods, our method enables long-term prediction of future slabs by incorporating latent epidemiological regularization. We demonstrated the ability of our method to make accurate predictions on real county- and state-level COVID-19 data. Our method achieves 18% and 10% lower RMSE compared to the best baseline, when predicting county-level daily new infections for 10 and 15 days-ahead respectively and 21% and 12% lower RMSE when predicting county-level hospitalized patients.

Acknowledgments

References

Acar, E.; Dunlavy, D. M.; and Kolda, T. G. 2009. Link prediction on evolving data using matrix and tensor factorizations. In *2009 IEEE International Conference on Data Mining Workshops (ICDMW)*, 262–269.

Araujo, M. R.; Ribeiro, P. M. P.; and Faloutsos, C. 2017. TensorCast: Forecasting with context using coupled tensors. In *2017 IEEE International Conference on Data Mining (ICDM)*, 71–80.

Chimmula, V. K. R.; and Zhang, L. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals* 135.

Cooke, K. L.; and Van Den Driessche, P. 1996. Analysis of an SEIRS epidemic model with two delays. *Journal of Mathematical Biology* 35(2): 240–260.

Dong, E.; Du, H.; and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20(5): 533–534.

Dunlavy, D. M.; Kolda, T. G.; and Acar, E. 2011. Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5(2): 1–27.

Gabay, D.; and Mercier, B. 1976. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* 2(1): 17–40.

Gao, J.; Sharma, R.; Qian, C.; Glass, L. M.; Spaeder, J.; Romberg, J.; Sun, J.; and Xiao, C. 2020. STAN: Spatio-Temporal Attention Network for Pandemic Prediction Using Real World Evidence. *arXiv preprint arXiv:2008.04215* .

Harshman, R. A. 1970. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers Phonetics* 16: 1–84.

He, S.; Peng, Y.; and Sun, K. 2020. SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics* 1–14.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Huang, K.; Sidiropoulos, N. D.; and Liavas, A. P. 2016. A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Transactions on Signal Processing* 64(19): 5052–5065.

Kapoor, A.; Ben, X.; Liu, L.; Perozzi, B.; Barnes, M.; Blais, M.; and O'Banion, S. 2020. Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks. *arXiv preprint arXiv:2007.03113* .

Kermack, W. O.; and McKendrick, A. G. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115(772): 700–721.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *2017 International Conference on Learning Representations (ICLR)*.

Papalexakis, E. E.; Faloutsos, C.; and Sidiropoulos, N. D. 2016. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8(2): 1–44.

Sidiropoulos, N. D.; De Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E. E.; and Faloutsos, C. 2017. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing* 65(13): 3551–3582.

Toda, A. A. 2020. Susceptible-infected-recovered (SIR) dynamics of COVID-19 and economic impact. *arXiv preprint arXiv:2003.11221* .

Tomar, A.; and Gupta, N. 2020. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment* 18.

Wang, L.; Chen, J.; and Marathe, M. 2019. DEFSI: Deep learning based epidemic forecasting with synthetic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9607–9612.

Yang, Z.; Zeng, Z.; Wang, K.; Wong, S.-S.; Liang, W.; Zanin, M.; Liu, P.; Cao, X.; Gao, Z.; Mai, Z.; et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease* 12(3): 165.