STELAR: Spatio-temporal Tensor Factorization with Latent Epidemiological Regularization

Nikos Kargas\textsuperscript{1,2}, Cheng Qian\textsuperscript{2}, Nicholas D. Sidiropoulos\textsuperscript{3}, Cao Xiao\textsuperscript{2}, Lucas M. Glass\textsuperscript{2}, Jimeng Sun\textsuperscript{4}

\textsuperscript{1}University of Minnesota, \textsuperscript{2}Analytics Center of Excellence, IQVIA, \textsuperscript{3}University of Virginia, \textsuperscript{4}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 EpidemioloLogicAl 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 real-world medical claims datasets, where the first dimen-
sion 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 of the pandemic (Yang et al. 2020; Toda 2020; He, Peng, and Sun 2020). They leverage human mobility 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.

| Notation | Description |
|---|---|
| $X \in \mathbb{R}^{M \times N \times L}$ | spatio-temporal tensor |
| $A \in \mathbb{R}^{M \times K}$ | location factor matrix |
| $B \in \mathbb{R}^{N \times K}$ | signal factor matrix |
| $C \in \mathbb{R}^{L \times K}$ | temporal factor matrix |
| $X^{(i)}$ | mode-$i$ unfolding of $X$ |
| $\beta; \gamma; \alpha$ | contact rate; recovery rate |
| $K$ | # of components |
| $M$ | # of locations |
| $N$ | # of signals |
| $L$ | # of time points |
| $T$ | transpose |
| $\odot$ | outer product |
| $\otimes$ | Kronecker product |
| $\circ$ | Khatri-Rao product |
| $\otimes$ | Hadamard product |
| $\| \cdot \|_F$ | Frobenius norm |
| $\text{diag}(x)$ | diagonal matrix |
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 exposed population 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. The SEIR model is a variant of the SIR model. In this example, the mode-

Canonical Polyadic Decomposition

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

$$X = [A, B, C] = \sum_{k=1}^{K} a_k \circ b_k \circ c_k,$$

where $A = [a_1, \ldots, a_K] \in \mathbb{R}^{M \times K}$, $B = [b_1, \ldots, b_K] \in \mathbb{R}^{N \times K}$, $C = [c_1, \ldots, c_K] \in \mathbb{R}^{L \times K}$. The rank $K$ is the minimum number of components needed to synthesize $X$. We can express the CPD of a tensor in many different ways. For example, $X = [A, B, C]$ can be represented by the matrix unfolding $X^{(1)} = (C \circ B)A^T$ 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 $X^{(2)} = (C \circ A)B^T$, $X^{(3)} = (B \circ A)C^T$ 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 $X$.

**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 $n$th 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 $X \in \mathbb{R}^{M \times N \times L}$, where $X(m,n,t) := x_{m,n,t}$. Tensor $X$ 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 $\tilde{X}(i+1, \ldots, L)$ 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 $X(:, :, t)$ is missing since we have no information regarding the corresponding row of $C$. 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 $X \in \mathbb{R}^{140 \times 15 \times 82}$ 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 $C$ where an SIR model has been fitted after obtaining the decomposition. We observe that each column 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

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

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

s.t. $A \geq 0, B \geq 0, C \geq 0, \beta \geq 0, \gamma \geq 0, s \geq 0, i \geq 0,$

$S_k(t) = S_k(t-1) - \beta_k S_k(t-1) I_k(t-1),$

$I_k(t) = I_k(t-1) + \beta_k S_k(t-1) I_k(t-1) - \gamma_k I_k(t-1),$

$s_k = S_k(0), i_k = I_k(0).$

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 $C$ 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 $\beta_k, \gamma_k$ according to Equations (1a), (1b), (1c). We aim at estimating these parameters such that each column of $C$ follows the new infections curve of an SIR model.

**Prediction**

After the convergence of the above optimization algorithm, we have some estimates of $\hat{A}, \hat{B}, \hat{C}$ 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 $\beta_k$ and $\gamma_k$ describes the epidemic transmission of the $k$th component in the time factor matrix and $s_k, i_k$ the initial values of the subpopulations. Using Equations (1a), (1b) and (1c) we can predict “future” values for the $k$th column of $C$. We repeat the same procedure for all columns of $C$ such that we can predict the entire “future” rows. Let $\hat{C}(t, i) \in \mathbb{R}^K$ be the prediction of the temporal information at a future time point $t$ using estimates $\beta, \gamma, \hat{s}, \hat{i}$. Since $A$ and $B$ do not depend on $t$, the prediction of all signals in the tensor at time $t$ is given by

$$
\hat{X}(\cdot, \cdot, t) = \hat{A} \text{diag}(\hat{C}(t, \cdot)) \hat{B}^T.
$$

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 $n$th signal for the $m$th location for time point $t$, is given by

$$
\hat{X}(m, n, t) = \sum_{k=1}^K \hat{a}_{m,k} \hat{b}_{n,k} \hat{\beta}_k \hat{S}_k (t-1) \hat{I}_k (t-1),
$$

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 $A, B, C$ we rely on alternating optimization. Note that by fixing all variables except for $A$, the resulting subproblem w.r.t. $A$ is a nonnegative least squares problem, which is convex. Similarly for the factor matrices $B$ and $C$. 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. $A$. Assume that at the $\ell$th iteration, we have some estimates of all the variables available. Fixing all variables except for $A$, we have

$$
\min_{A \geq 0} \| X^{(1)} - \Phi_A^{(\ell)} A^T \|^2_F + \mu \| A \|^2_F
$$

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

$$
\hat{A} = \arg \min_{A \geq 0} \| X^{(1)} - \Phi_A^{(\ell)} A^T \|^2_F + \mu \| A \|^2_F
$$

$$
+ \rho \| A - \hat{A}^T + A_d \|^2_F,
$$

$$
A_d = A_d + A - \hat{A}^T.
$$

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 element-wise nonnegative projection operator and Equation (8c) is
Algorithm 1 STELAR Method

**Input:** Tensor $X$, rank $K$, max. outer iterations $\text{iter}_{\text{souter}}$, max. inner iterations $\text{iter}_{\text{sinner}}$, gradient steps $\text{iter}_{\text{grad}}$, prediction window $L_o$.

**repeat**

- Update $A$ using (8) for $\text{iter}_{\text{sinner}}$ iterations
- Update $B$ using (9) for $\text{iter}_{\text{sinner}}$ iterations
- Update $C$ using (11) for $\text{iter}_{\text{sinner}}$ iterations
- Perform $\text{iter}_{\text{grad}}$ projected gradient steps for $\beta, \gamma, s_i$ using (14), (15), (16), (17)

**until** $\text{iter}_{\text{souter}}$ or validation RMSE increases

Predict $L_o$ future slabs using Equation (3).

The dual variable update. The updates for $B$ are similar:

\[
\hat{B} = \arg \min_{B \geq 0} \|X^{(3)} - \Phi_B^{(t)} \hat{B}\|_F^2 + \mu \|\hat{B}\|_F^2,
\]

(9a)

\[
\hat{B} = \arg \min_{B \geq 0} \|B - \hat{B}^T + B_d\|_F^2,
\]

(9b)

\[
B_d = B_d + B - \hat{B}^T,
\]

(9c)

where $\Phi_B^{(t)} = C^{(t)} \odot A^{(t)}$. Now let us consider the update of $C$. The related optimization problem takes the form of

\[
\min_{C \geq 0} \|X^{(3)} - \Phi_C^{(t)} C\|_F^2 + \mu \|C\|_F^2 + \nu \|C - C^{(t)}\|_F^2,
\]

(10)

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

\[
\hat{C} = \arg \min_{C \geq 0} \|X^{(3)} - \Phi_C^{(t)} C\|_2^2 + \mu \|\hat{C}\|_2^2
\]

(11a)

\[
+ \nu \|C - C^{(t)}\|_2^2 + \rho \|C - \hat{C}^T + C_d\|_2^2,
\]

(11b)

\[
C = \arg \min_{C \geq 0} \|C - \hat{C}^T + C_d\|_F^2,
\]

(11c)

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

\[
\min_{\beta,\gamma,s_i} \sum_{k=1}^K \sum_{t=1}^L (c_{t,k} - \beta_k S_k(t-1) I_k(t-1))^2
\]

s. t. $\beta \geq 0, \gamma \geq 0, s \geq 0, i \geq 0,

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)

- \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 $\beta$ and $\gamma$, and are calculated in a recursive manner. Therefore the optimization problem w.r.t. $\beta$ and $\gamma$ 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 $k$th subproblem, we have

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

(13)

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

(S_k(t-1) I_k(t-1) + \beta_k \frac{\partial S_k(t-1)}{\partial \beta_k} I_k(t-1) + \beta_k S_{k-1} \frac{\partial I_{k-1}}{\partial \beta_k}).
\]

(14)

Note that both $S_k(t)$ and $I_k(t)$ are recursive functions. Thus, their respective derivatives w.r.t. $\beta$ are computed recursively in $L$ steps. Similarly, for $\gamma_k$ 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_{k-1}}{\partial \gamma} I_{k-1} + \beta_k S_{k-1} \frac{\partial I_{k-1}}{\partial \gamma}\right).
\]

(15)

Finally for $s_i$

\[
\frac{\partial f}{\partial s_i} = -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_i} I_k(t-1) + \beta_k S_k(t-1) \frac{\partial I_k(t-1)}{\partial s_i}\right).
\]

(16)

\[
\frac{\partial f}{\partial i_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_k(t-1)}{\partial i_k} I_k(t-1) + \beta_k S_k(t-1) \frac{\partial I_k(t-1)}{\partial i_k}\right).
\]

(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 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 which includes data from the JHU dataset.
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      | 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  | 157.6 | 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  |

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

| Model     | $L_o = 10$ |       | $L_o = 15$ |       |
|-----------|------------|-------|------------|-------|
|           | RMSE | MAE | RMSE | MAE  |
| Mean      | 307.0 | 258.7 | 325.8 | 273.1 |
| SIR       | 163.1 | 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  |

Massachusetts, Connecticut, Pennsylvania and is of size 5 × 15 × 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-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.

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

| 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)   | Monong. (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 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.

References

Acar, E.; Dunlavy, D. M.; and Kolda, T. G. 2009. Link prediction on evolving data using matrix and tensor factoriza-
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 Workshops (ICDMW), 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” multimodal 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.

Sidirooulos, 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.

Tomas, 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.