Simple Epidemic Models with Segmentation Can Be Better than Complex Ones

Geon Lee\textsuperscript{1} \quad Se-eun Yoon\textsuperscript{2} \quad Kijung Shin\textsuperscript{1, 2}

Graduate School of AI\textsuperscript{1}, School of Electrical Engineering\textsuperscript{2}
Korea Advanced Institute of Science and Technology
Overview

1. Introduction
2. Backgrounds
3. Method
4. Experiments
5. Conclusion
Problem Definition

• **Understanding and predicting epidemic spreads** are important for prediction and effective decision making.
  - *How many people will be infected within a week?*
  - *How will lockdowns affect the spread?*

• To answer these questions, we require a method that is **simple & expressive** method to model and predict the spread of infectious diseases.
Is a Single Model Enough?

• Epidemic models describe dynamics of epidemic spreads.
• However, describing **long-term dynamics of epidemics** is challenging.
  ▪ Unpredictability & abruptness of real-world events
  ▪ E.g., lockdowns or the capability to perform tests
Is a Single Model Enough? (cont.)

• *Can a single model describe the long-term epidemic event sequence?*
  - The data is extremely complex.
  - The model can overfit the data.

**Our Idea**

Properly divide an epidemic event sequence into multiple segments and fit a simple epidemic model to each segment.
Many epidemic models are based on ordinary differential equations (ODEs). Some of the earliest epidemic models are SIS, SIR, and SEIR. They are based on human knowledge. Other data-driven models: LLD and NLLD. They model time-series data without relying on human knowledge.
Simple Epidemic Models with Segmentation Can Be Better than Complex Ones

SIR: Susceptible-Infectious-Recovered Model

- A closed population $P$ is divided into three states:
  - $S$ (susceptible), $I$ (infectious), and $R$ (recovered)
- At timestamp $t$, the number of individuals of each state is $S(t)$, $I(t)$, and $R(t)$.

\[
\begin{align*}
\frac{dS(t)}{dt} &= -\beta S(t)I(t) \\
\frac{dI(t)}{dt} &= \beta S(t)I(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t)
\end{align*}
\]
LLD: Linear Latent Dynamics Model [Matsubara and Sakurai, KDD 2016]

- $v(t)$: $d$-dimensional observable event sequence
  - In our case, we use 2-dimensional data $v(t) = [I(t), R(t)]$.
- $w(t)$: $k$-dimensional latent event sequence

\[
\frac{dw(t)}{dt} = p + Qw(t)
\]

\[
v(t) = u + Vw(t)
\]

- $p$ and $Q$ describe dynamics between latent factors
  - $p \ (\in \mathbb{R}^k)$: linear dynamics
  - $Q \ (\in \mathbb{R}^{k \times k})$: exponential dynamics
- $u \ (\in \mathbb{R}^d)$ and $V \ (\in \mathbb{R}^{k \times d})$ project latent factors to the observed events
NLLD: Non-Linear Latent Dynamics Model [Matsubara and Sakurai, KDD 2016]

- \( v(t) \): \( d \)-dimensional **observable** event sequence
  - In our case, we use 2-dimensional data \( v(t) = [I(t), R(t)] \).
- \( w(t) \): \( k \)-dimensional **latent** event sequence

\[
\frac{dw(t)}{dt} = p + Qw(t) + Aw(t)^2
\]

\[ v(t) = u + Vw(t) \]

- \( p, Q, \) and \( A \) describe dynamics between latent factors
  - \( p \) (\( \in \mathbb{R}^k \)): linear dynamics
  - \( Q \) (\( \in \mathbb{R}^{k \times k} \)): exponential dynamics
  - \( A \) (\( \in \mathbb{R}^k \)): **non-linear** dynamics
- \( u \) (\( \in \mathbb{R}^d \)) and \( V \) (\( \in \mathbb{R}^{k \times d} \)) project latent factors to the observed events
Our Method

Our Idea

Properly divide an epidemic event sequence into multiple segments and fit a simple epidemic model to each segment.

Q1. How to divide the epidemic event sequence?

Q2. How to automatically find the best segmentation?

Q3. How to control the trade-offs between the model complexity and fitness?
Description Length

- Given a sequence $X$ and a model $M$, the description length (in bits) of $X$ is:

\[
\text{Cost}(X) := \text{Cost}(M) + \text{Cost}(X|M)
\]

- **Model cost** $\text{Cost}(M)$: the number of bits required to describe the model $M$
- **Data cost** $\text{Cost}(X|M)$: the number of bits to encode $X$ given $M$
### Model Cost

<table>
<thead>
<tr>
<th>SIR Model</th>
<th>LLD Model</th>
<th>NLLD Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 \cdot C_F \text{ bits}</td>
<td>(k^2 + (2 + d) \cdot k + d) \cdot C_F \text{ bits}</td>
<td>(k^2 + (3 + d) \cdot k + d) \cdot C_F \text{ bits}</td>
</tr>
<tr>
<td>\cdot \beta: C_F \text{ bits}</td>
<td>\cdot w_0 \in \mathbb{R}^k: k \cdot C_F \text{ bits}</td>
<td>\cdot w_0 \in \mathbb{R}^k: k \cdot C_F \text{ bits}</td>
</tr>
<tr>
<td>\cdot \gamma: C_F \text{ bits}</td>
<td>\cdot p \in \mathbb{R}^k: k \cdot C_F \text{ bits}</td>
<td>\cdot p \in \mathbb{R}^k: k \cdot C_F \text{ bits}</td>
</tr>
<tr>
<td></td>
<td>\cdot Q \in \mathbb{R}^{k \times k}: k^2 \cdot C_F \text{ bits}</td>
<td>\cdot Q \in \mathbb{R}^{k \times k}: k^2 \cdot C_F \text{ bits}</td>
</tr>
<tr>
<td></td>
<td>\cdot u \in \mathbb{R}^d: d \cdot C_F \text{ bits}</td>
<td>\cdot A \in \mathbb{R}^k: k \cdot C_F \text{ bits}</td>
</tr>
<tr>
<td></td>
<td>\cdot V \in \mathbb{R}^{k \times d}: k \cdot d \cdot C_F \text{ bits}</td>
<td>\cdot u \in \mathbb{R}^d: d \cdot C_F \text{ bits}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>\cdot V \in \mathbb{R}^{k \times d}: k \cdot d \cdot C_F \text{ bits}</td>
</tr>
</tbody>
</table>

$C_F$ is the number of bits to encode a real number.
Data Cost

- **Data cost** is the number of bits to encode $X$ given $M$.
  - $X$: observed event sequence
  - $V$: estimated event sequence by the model $M$

- The number of bits required is the negative log-likelihood under a Gaussian distribution $\mathcal{N}(0, \sigma^2)$:

$$\text{Cost}(X|M) = -\log P(X - V) = -\log \prod_{t=1}^{n} \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i(t)-v_i(t))^2}{2\sigma^2}}$$
**Simple Epidemic Models with Segmentation Can Be Better than Complex Ones**

---

**Method**

**Trade-off: Model Complexity vs. Fitness**

- **Active Cases (True)**
- **Recovery/Death (True)**
- **Active Cases (Estimated)**
- **Recovery/Death (Estimated)**

- **Simple model**
  - Low model cost 🙄
  - High data cost 😞

- **Complex model**
  - High model cost 😞
  - Low data cost 😁

---

**Introduction**

**Backgrounds**

**Experiments**

**Conclusion**

---

RMSE: 247,856 × 1,000

RMSE: 189,228 × 1,000

---

epiDAMIK 2021
Simple Epidemic Models with Segmentation Can Be Better than Complex Ones
We divide the sequence into multiple segments and fit a simple model to each segment.

**Q.** How to **automatically** divide the sequence that leads to better trade-offs between model complexity and fitting error?
An epidemic event sequence $X (= X_{1:n})$ of length $n$ is divided into $r$ segments:

$$X_{s_1:e_1} \oplus \cdots \oplus X_{s_r:e_r}$$

Apply model $f_i$ to each segment $X_{s_i:e_i}$.
Minimum Description Length (MDL) Principle (cont.)

• The description length in bits of $X_{s_1:e_1} \oplus \cdots \oplus X_{s_r:e_r}$ is:

$$\text{Cost}(X_{s_1:e_1} \oplus \cdots \oplus X_{s_r:e_r}) = (r - 1) \cdot \log_2(n) + \sum_{i=1}^{r} \left( \text{Cost}(f_i) + \text{Cost}(X_{s_i:e_i:f_i}) \right)$$

1. Bits required to encode $r - 1$ splitting points.
2. The description length of each segment.
Segmentation Search

• Given an event sequence $X_{1:n}$, there are $2^n$ ways of segmentation.

• We propose a greedy segmentation scheme.

Find a splitting point where the description length is minimized.

Splitting the segment is no more beneficial. Stop splitting.
Experimental Settings

- We consider **70 countries** with the most confirmed cases of **COVID-19** as of the end of March 2021.

Argentina, Armenia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Bolivia, Brazil, Bulgaria, Canada, Chile, Colombia, Costa Rica, Croatia, Czech, Denmark, Dominican Republic, Ecuador, Egypt, France, Georgia, Germany, Greece, Guatemala, Honduras, Hungary, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Kuwait, Lebanon, Lithuania, Malaysia, Mexico, Moldova, Morocco, Nepal, Netherlands, Pakistan, Panama, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Saudi Arabia, Serbia, Slovakia, Slovenia, South Africa, Spain, Sweden, Switzerland, Tunisia, Turkey, United Arab Emirates, United Kingdom, Ukraine, United States

Q1. Effectiveness of Segmentation

Simple epidemic models with segmentation provide more concise and accurate description of the spread of COVID-19 than complex models without segmentation.

![Graph showing fitting error vs. model cost (bits) for NLLD (India). The graph includes a red arrow indicating a 11.54X improvement in fitting with segmentation compared to fitting without segmentation.]
Simple epidemic models with segmentation provide more concise and accurate description of the spread of COVID-19 than complex models without segmentation.
Q2. Effectiveness of Our Segmentation Scheme

Our segmentation scheme yields better segmentation scheme than in the incremental method [Matsubara and Sakurai, KDD 2016].

![Graph showing comparison between Greedy Segmentation (Proposed) and Incremental Segmentation]

- Greedy Segmentation (Proposed)
- Incremental Segmentation

NLLD (India)
Q3. Accuracy of Forecasting

**Segmentation** is helpful to accurate prediction of the spread of COVID-19.

- **Simple models** do not fit the long-term dynamics well.
- **Complex models** do not predict well due to overfitting.
- **Our model** accurately predicts the future dynamics.

---

Graph showing that simple models with segmentation (NLLD (India)) have lower RMSE compared to complex models without segmentation for different values of $k$.
Conclusion

• We propose to divide an epidemic event sequence into multiple segments and fit a simple model to each segment.

Our methodology is:

✓ **Automatic**: All parameters are tuned automatically based on MDL principle.

✓ **Model-agnostic**: Any ODE-based epidemic models can be used.

✓ **Effective**: It fits and predicts well in COVID-19 datasets.

Code & datasets: [https://github.com/geonlee0325/covid_segmentation](https://github.com/geonlee0325/covid_segmentation)