

Simple Epidemic Models with Segmentation Can Be Better than Complex Ones

Geon Lee¹

Se-eun Yoon²

Kijung Shin^{1,2}

Graduate School of AI¹, School of Electrical Engineering²
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

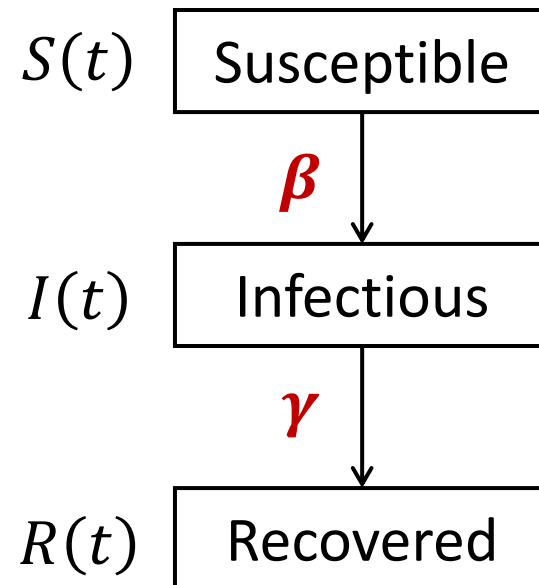
Properly divide an epidemic event sequence into **multiple segments** and fit a **simple epidemic model** to each segment.

Ordinary Differential Equations (ODEs)

- 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.

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 $R(t)$.



$$\frac{S(t)}{dt} = -\beta S(t)I(t)$$
$$\frac{I(t)}{dt} = \beta S(t)I(t) - \gamma I(t)$$
$$\frac{R(t)}{dt} = \gamma I(t)$$

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

$$\frac{dw(t)}{dt} = \mathbf{p} + \mathbf{Q}w(t)$$

$$v(t) = \mathbf{u} + \mathbf{V}w(t)$$

- \mathbf{p} and \mathbf{Q} describe dynamics between latent factors
 - $\mathbf{p} (\in \mathbb{R}^k)$: linear dynamics
 - $\mathbf{Q} (\in \mathbb{R}^{k \times k})$: exponential dynamics
- $\mathbf{u} (\in \mathbb{R}^d)$ and $\mathbf{V} (\in \mathbb{R}^{k \times d})$ 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

$$\frac{dw(t)}{dt} = \mathbf{p} + \mathbf{Q}w(t) + \mathbf{A}w(t)^2$$

$$v(t) = \mathbf{u} + \mathbf{V}w(t)$$

- \mathbf{p} , \mathbf{Q} , and \mathbf{A} describe dynamics between latent factors
 - $\mathbf{p} (\in \mathbb{R}^k)$: linear dynamics
 - $\mathbf{Q} (\in \mathbb{R}^{k \times k})$: exponential dynamics
 - $\mathbf{A} (\in \mathbb{R}^k)$: **non-linear** dynamics
- $\mathbf{u} (\in \mathbb{R}^d)$ and $\mathbf{V} (\in \mathbb{R}^{k \times d})$ 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:

$$\mathit{Cost}(X) := \mathit{Cost}(M) + \mathit{Cost}(X|M)$$

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

Model Cost

SIR Model

$$2 \cdot C_F \text{ bits}$$

- $\beta: C_F$ bits
- $\gamma: C_F$ bits

LLD Model

$$(k^2 + (2 + d) \cdot k + d) \cdot C_F \text{ bits}$$

- $w_0 \in \mathbb{R}^k: k \cdot C_F$ bits
- $p \in \mathbb{R}^k: k \cdot C_F$ bits
- $Q \in \mathbb{R}^{k \times k}: k^2 \cdot C_F$ bits
- $u \in \mathbb{R}^d: d \cdot C_F$ bits
- $V \in \mathbb{R}^{k \times d}: k \cdot d \cdot C_F$ bits

NLLD Model

$$(k^2 + (3 + d) \cdot k + d) \cdot C_F \text{ bits}$$

- $w_0 \in \mathbb{R}^k: k \cdot C_F$ bits
- $p \in \mathbb{R}^k: k \cdot C_F$ bits
- $Q \in \mathbb{R}^{k \times k}: k^2 \cdot C_F$ bits
- $A \in \mathbb{R}^k: k \cdot C_F$ bits
- $u \in \mathbb{R}^d: d \cdot C_F$ bits
- $V \in \mathbb{R}^{k \times d}: k \cdot d \cdot C_F$ bits

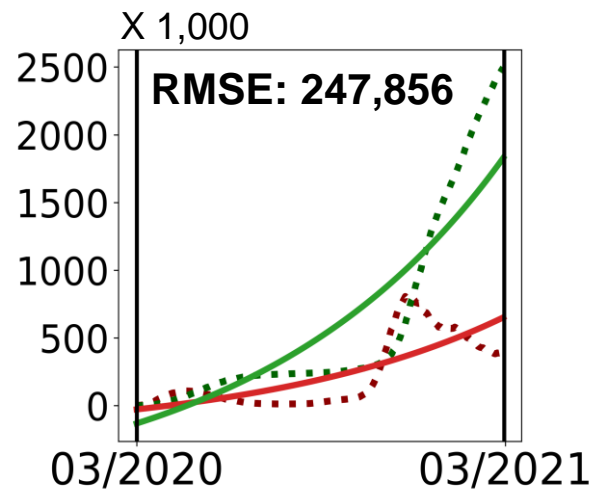
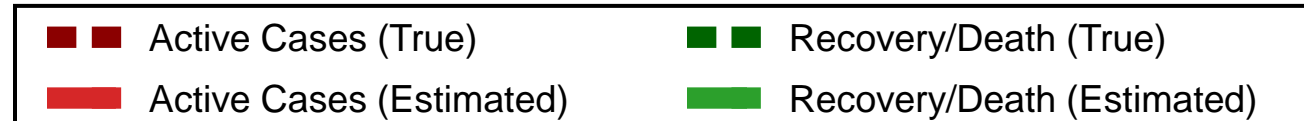
C_F is the number of bits to encode a real number.

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 $\mathcal{N}(0, \sigma^2)$:

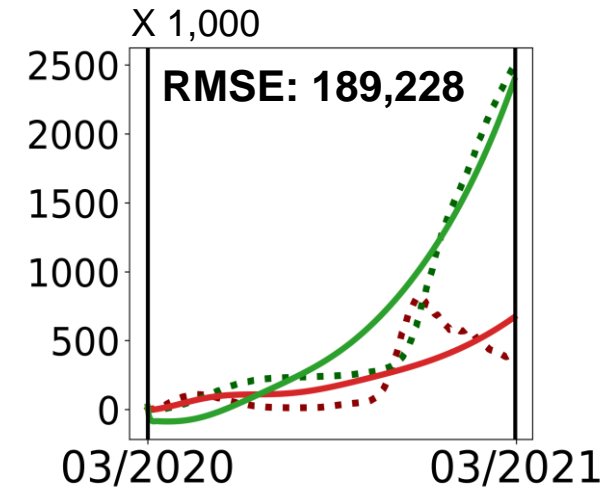
$$\mathit{Cost}(X|M) = -\log P(\mathbf{X} - \mathbf{V}) = -\log \prod_{t=1}^n \prod_{i=1}^d \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i(t)-v_i(t))^2}{2\sigma^2}}$$

Trade-off: Model Complexity vs. Fitness





Simple model

- Low model cost 
- High data cost 

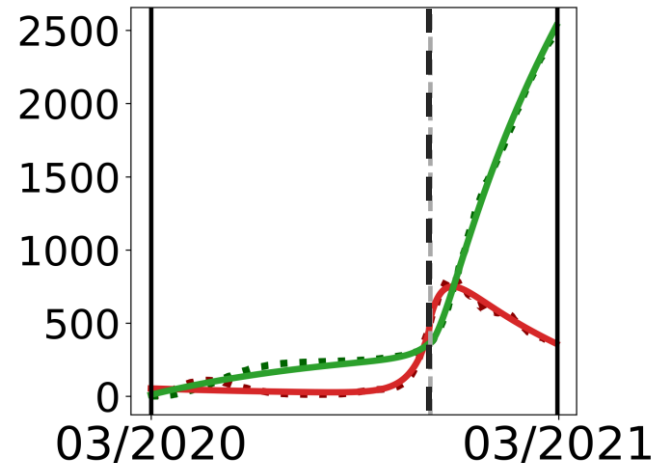


Complex model

- High model cost 
- Low data cost 

Trade-off: Model Complexity vs. Fitness (cont.)

- We divide the sequence into **multiple segments** and fit a **simple model** to each segment.



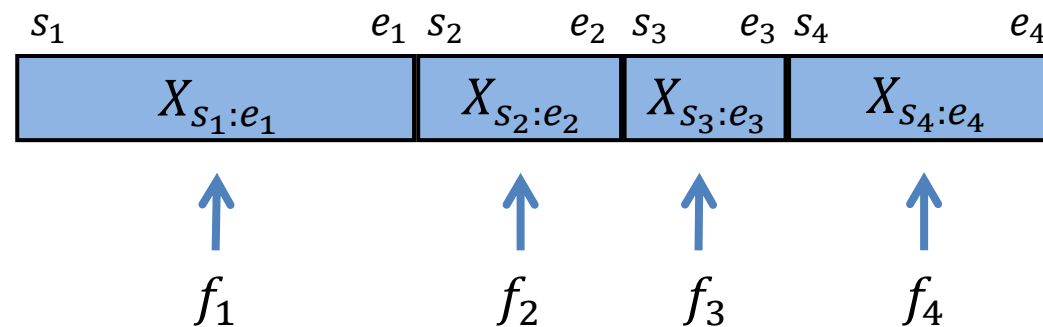
Q. How to **automatically** divide the sequence that leads to better trade-offs between model complexity and fitting error?

Minimum Description Length (MDL) Principle

- 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.)

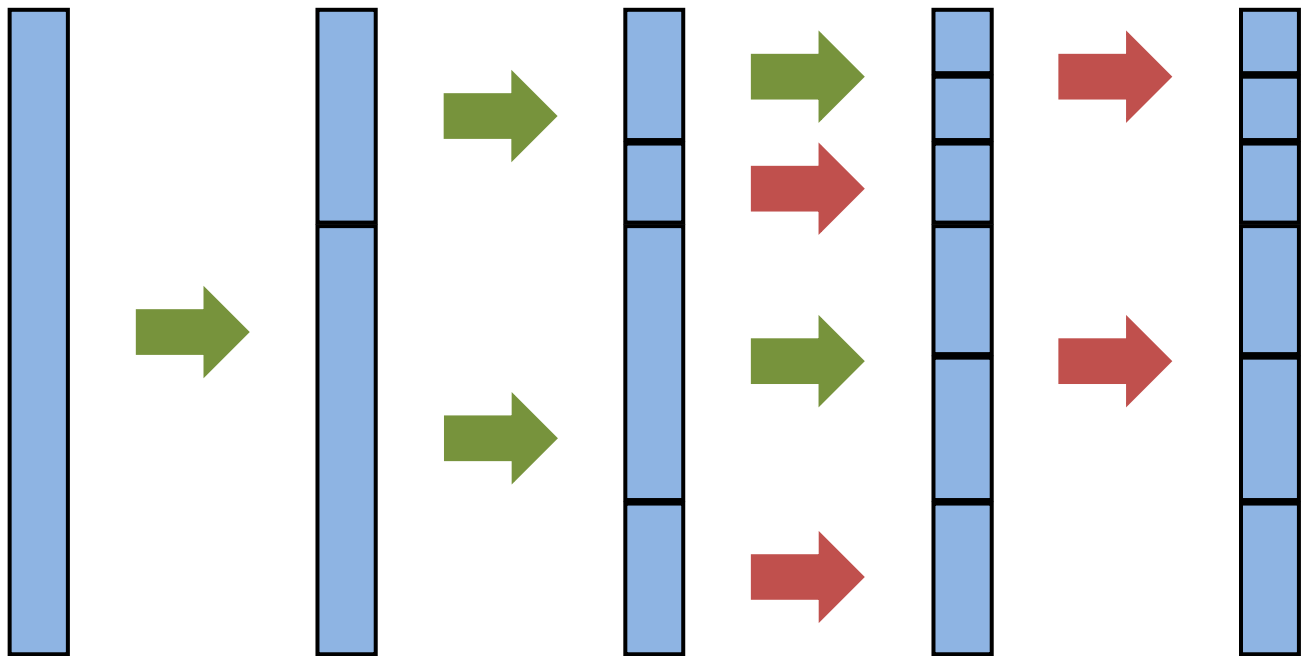
- The description length in bits of $X_{s_1:e_1} \oplus \dots \oplus X_{s_r:e_r}$ is:

$$\text{Cost}(X_{s_1:e_1} \oplus \dots \oplus X_{s_r:e_r}) = \underbrace{(r - 1) \cdot \log_2(n)}_{\textcircled{1}} + \underbrace{\sum_{i=1}^r \left(\overset{\text{Model cost}}{\downarrow} \text{Cost}(f_i) + \overset{\text{Data cost}}{\downarrow} \text{Cost}(X_{s_i:e_i} \cdot f_i) \right)}_{\textcircled{2}}$$

- 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

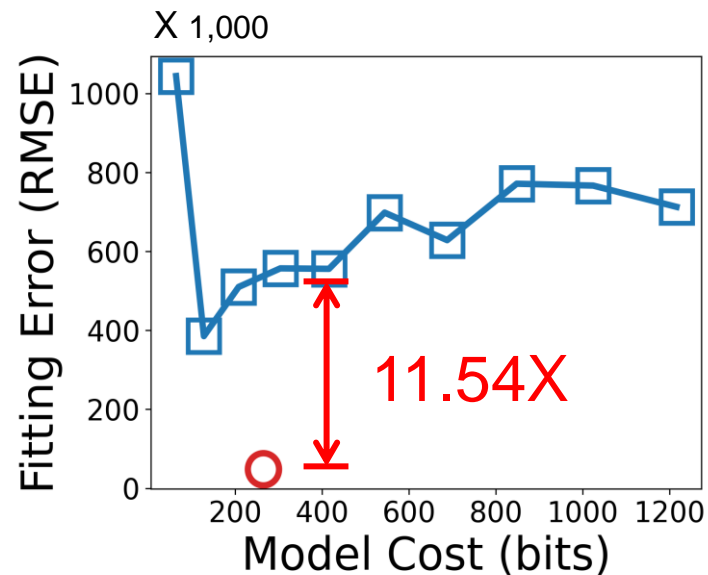
- We consider **70 countries** with the most confirmed cases of **COVID-19** as of the end of March 2021.

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

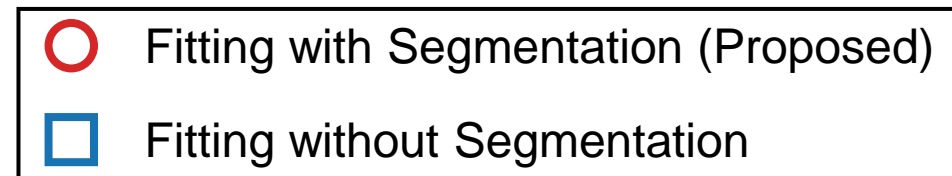
Public datasets: <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>

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**.

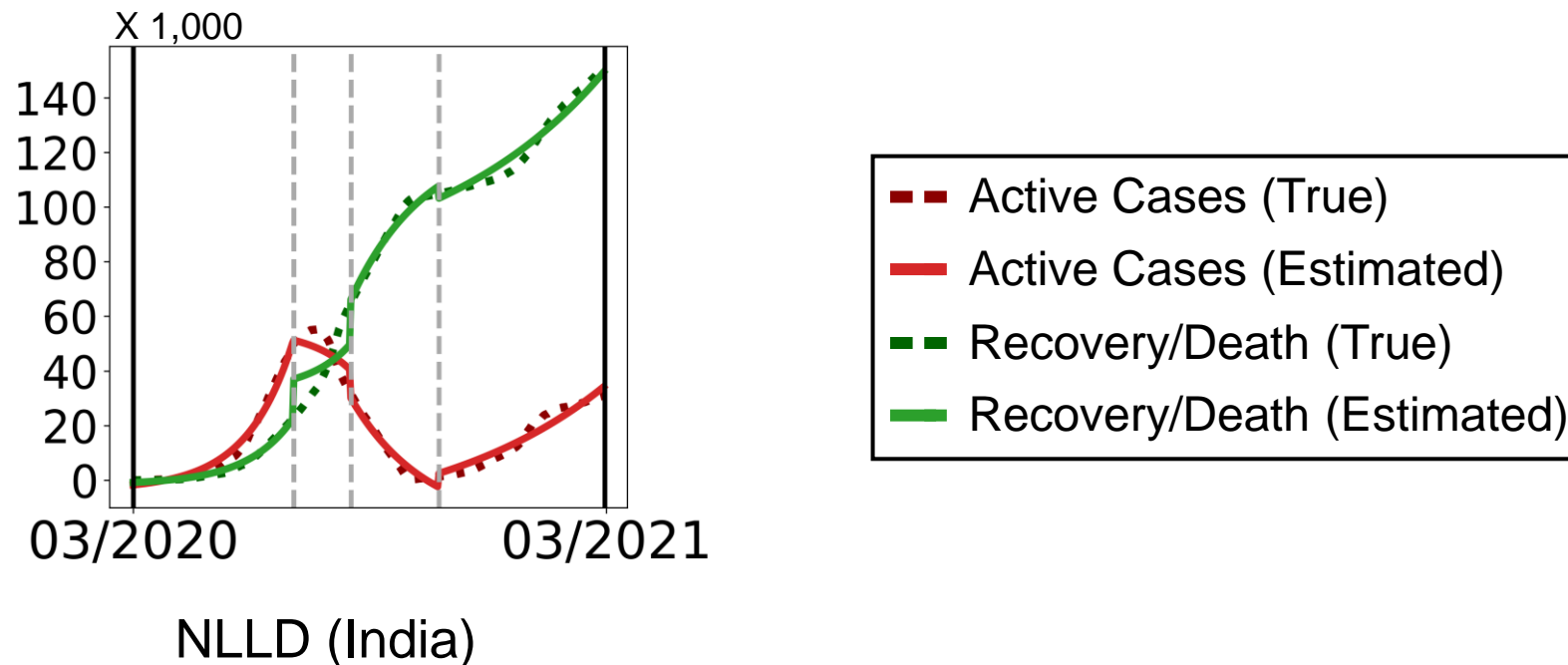


NLLD (India)



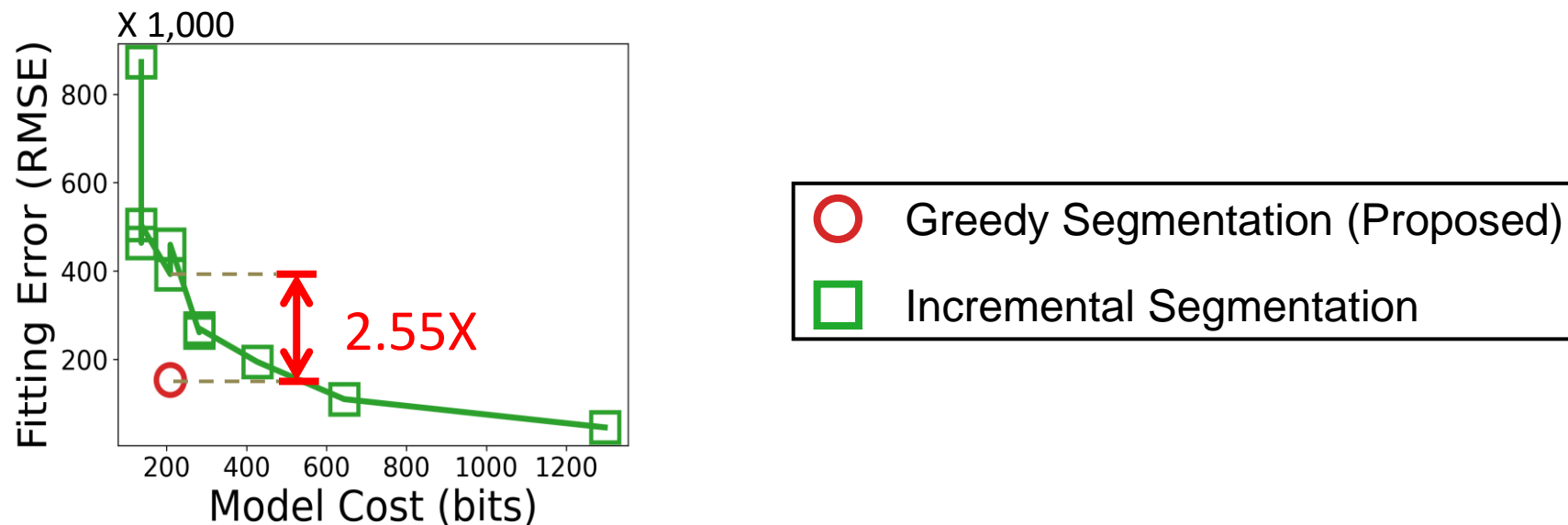
Q1. Effectiveness of Segmentation (cont.)

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

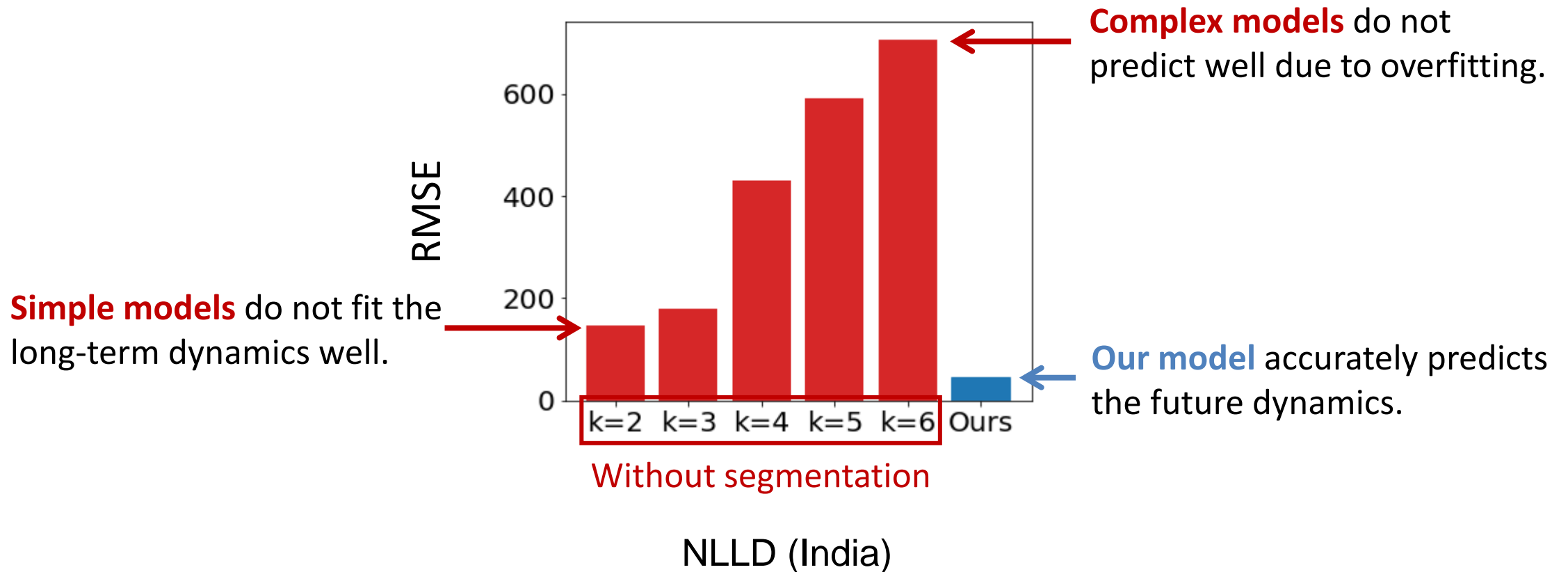
Our segmentation scheme yields better segmentation scheme than in the **incremental method** [Matsubara and Sakurai, KDD 2016].



NLLD (India)

Q3. Accuracy of Forecasting

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



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