PATHOGENESIS SIGNAL PROCESSING USING TOEPLITZ PROPAGATION OPERATORS

MARISA GAETZ

The Institute for Disease Modeling
Global Health — Bill & Melinda Gates Foundation

Abstract. In this paper, we present substantial progress towards formalizing and generalizing the biologically-motivated *pathogenesis signal processing* method introduced in 2022 by Thakkar and Famulare for COVID-19 modeling. In particular, we describe a generalization of this signal processing method, the accuracy of which relies on having the right notion of the *effective rank* of a matrix. Rather than adopting the *Roy-Vetterli effective rank* (which was used by Thakkar and Famulare for COVID-19), we introduce an alternative notion of effective rank and argue that it is the correct notion of effective rank for the purposes of pathogenesis signal processing. This generalized approach has the potential to be used in the analysis of other infectious diseases (besides COVID-19) as well as for multipathogen analysis.

1. Introduction and Motivation – Situational Awareness

Situational awareness (sometimes even at the level of deciding if trends are rising or falling) is a consistent challenge in infectious disease modeling. Part of the challenge is that the available data typically address the situation on a variety of levels: individual-level data like the progression of symptoms are separated in scale from population-level data like time-series of the number of cases. The aspiration for situational awareness is understanding how all these pieces fit together in a consistent epidemiology.

In [7], Thakkar and Famulare provide some guiding principles for situational awareness in the case of COVID-19. One of the key tenets of their paper is the importance of incorporating data from a variety of levels in our disease models. As they explain, it is not enough to only understand the data at one level. If we, for example, look at population-level data alone, we might get a noisy model that ignores or is even incompatible with important individual-level biology. However, relating these different levels of data is easier said than done: It's sometimes conceptually appealing to build from the ground up (i.e. to try to obtain a population-level understanding by using individual-level data in conjunction with an interaction model), but population dynamics are usually poorly constrained by tractable interaction models. Instead, Thakkar and Famulare develop formal mathematical relationships between population-level data and individual-level data directly, ultimately building a *pathogenesis signal processing* approach for COVID-19 modeling.

While Thakkar and Famulare's work was focused on developing a fast, real-time model specific to COVID-19, their core *pathogenesis signal processing* idea has much more general promise. The goal of this paper is to formalize and generalize the approach introduced in [7] so that it can be applied to: (1) other infectious diseases (besides COVID-19), and (2)

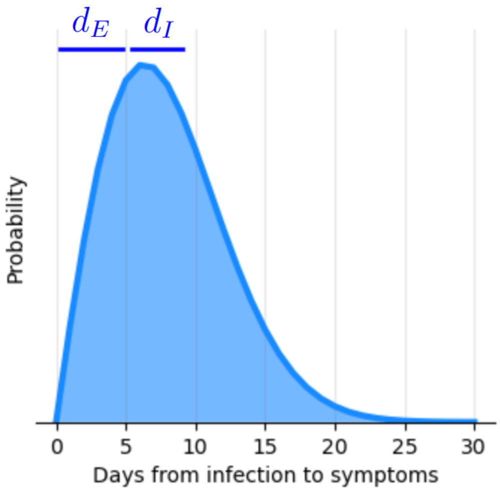

FIGURE 1. The *pathogenesis distribution* of the time from COVID-19 infection to symptom onset, characterizing individual-level pathogenesis, estimated in [4] using data on travel from Wuhan.

situations where we have data for multiple pathogens collected from the same locations, times, and/or people, where we expect there to be some commonality in the noise observed.

In Section 2, we introduce Thakkar and Famulare's mathematical relationships between population- and individual-level data and define a *Toeplitz matrix* $\mathbf{L}_\pi$ that can be associated to a given *pathogenesis distribution* $\pi(\tau)$. In Section 3, we describe the idea behind Thakkar and Famulare's *pathogenesis signal processing* approach, which relies on a notion of the *effective rank* of the matrix $\mathbf{L}_\pi$. In Section 4, we define an alternative notion of *effective rank* (different than the one used by Thakkar and Famulare for COVID-19 signal processing) and argue that this alternative notion is the correct notion of effective rank for the purposes of pathogenesis signal processing. We conclude in Section 5 with some open questions and directions for future work.

2. Mathematical Relationships Between Individual and Population Levels

In this section, we will start by explaining the intuitive mathematical relationships between individual- and population-level data as derived in [7]. While derived with COVID-19 in mind, these relationships hold for any infectious disease. We will then combine these relationships and will analyze the explicit form of the primary operator involved in the resulting combined relationship.

For individual-level data, we will consider a *pathogenesis distribution* $\pi(\tau)$, where $\pi(\tau)$ is the probability that, for an infected individual, the duration between the time of infection and the time of symptom onset equals τ. This is a typical measurable input; Figure 1 shows the distribution for COVID-19, as estimated by [4]. We will also consider d_E and d_I, the expected durations of the latent and infectious states, respectively. For COVID-19, these are estimated as $d_E = 5$ and $d_I = 4$ in [6].

For population-level data, we will consider the following values:

- $\hat{N}_t$: the expected number of new infections on day t.

- $\hat{E}_t$: the expected number of people who, on day t, have been exposed and infected but aren't yet infectious.
- $\hat{I}_t$: the expected number of people who, on day t, are infectious.

To illustrate the distinction between the individual- and population-level data, we will use blue to indicate individual-level information and red to indicate population-level information in the offset equations that follow.

2.1. **First mathematical relationship.** With hats indicating expected values, we have the following familiar relationships:

$$\hat{E}_t = \left(1 - \frac{1}{d_E}\right)\hat{E}_{t-1} + \hat{N}_{t-1} \tag{1}$$

$$\hat{I}_t = \left(1 - \frac{1}{d_I}\right)\hat{I}_{t-1} + \frac{1}{d_E}\hat{E}_{t-1} \tag{2}$$

Rearranging these equations and combining the results into vector formulas (where we the time period of interest ranges over days $t = 0, \ldots, T$), we get

$$\hat{N}_{t-1} = \hat{E}_t - \left(1 - \frac{1}{d_E}\right)\hat{E}_{t-1} \implies \hat{\mathbf{N}}_{T-2} = \mathbf{D}_E\hat{\mathbf{E}}_{T-1} \tag{3}$$

$$\hat{E}_{t-1} = d_E\left[\hat{I}_t - \left(1 - \frac{1}{d_I}\right)\hat{I}_{t-1}\right] \implies \hat{\mathbf{E}}_{T-1} = \mathbf{D}_I\hat{\mathbf{I}}_T, \tag{4}$$

where the boldface indicates matrices or column vectors, where the subscripts of the column vectors indicates dimension, and where $\mathbf{D}_E$ and $\mathbf{D}_I$ are weighted differencing matrices of sizes $(T-2) \times (T-1)$ and $(T-1) \times T$, respectively. Combining these expressions, we get that

$$\boxed{\hat{\mathbf{N}}_{T-2} = \mathbf{D}_E\mathbf{D}_I\hat{\mathbf{I}}_T} \tag{5}$$

2.2. **Second mathematical relationship.** Moving forward, we make the assumption that symptom onset marks the mid-point of the infectious period.[1] (In particular, we assume that everyone who becomes infected becomes symptomatic.) The midpoint assumption implies that

$$\boxed{\hat{I}_t = d_I\sum_{s=1}^{t}\pi(t-s)\hat{N}_s \implies \hat{\mathbf{I}}_{T-1} = \mathbf{P}_\pi\hat{\mathbf{N}}_{T-2}} \tag{6}$$

where the $(T-1)\times(T-2)$ matrix $\mathbf{P}_\pi$ has entries $(\mathbf{P}_\pi)_{t,s} = d_I\pi(t-s)$, and where $\pi(t-s) = 0$ for all $t \leq s$ (since infection has to happen before infectiousness). To see why this relationship holds, notice that from a point in time t, $\pi(t-s)$ represents the probability that a person who was exposed on day s becomes symptomatic (or, equivalently, is at the mid-point of their infectious period) on day t. The number $d_I\pi(t-s)\hat{N}_s$ therefore estimates the number

[1] For diseases in which this assumption is not reasonable, the distribution $\pi(\tau)$ should instead be defined directly as the probability distribution of the duration between the time of infection and the midpoint of the infectious period, rather than the duration between the time of infection and the time of symptom onset. For the purposes of this paper, we will stick with the original definition of $\pi(\tau)$ so that we can easily cross-reference and pull examples from [7].

of people who were exposed on day s and are infectious on day t.[2] Summing over s therefore gives an estimate for the total number of people who are infectious on day t.

We will assume that $\pi(\tau) \to 0$ as τ becomes large. In particular, we will assume that $\pi(\tau) = 0$ for $\tau \geq T-3$. We choose $T-3$ specifically merely because it will be a convenient choice in some calculations later on in the paper, but all we are assuming here is that someone exposed on day 1 will reach the midpoint of their infectious period on or before day $T-3$ (where T is chosen to be plenty large).

2.3. Combining mathematical relationships.

Combining (5) and (6), we get that

$$\boxed{\hat{\mathbf{I}}_{T-1} = \mathbf{P}_\pi \mathbf{D}_E \mathbf{D}_I \hat{\mathbf{I}}_T} \tag{7}$$

where $\mathbf{A}_\pi := \mathbf{P}_\pi \mathbf{D}_E \mathbf{D}_I$ is a $(T-1)\times T$ matrix. At first this may look like a boring relationship: If we apply the weighted differencing matrices $\mathbf{D}_I$ and $\mathbf{D}_E$ to $\hat{\mathbf{I}}_T$ and then aggregate with $\mathbf{P}_\pi$, we recover $\hat{\mathbf{I}}_{T-1}$. But for a given pathogenesis distribution $\pi(\tau)$ and consistent latent and infectious durations d_E and d_I, only certain time series $\hat{\mathbf{I}}_T$ satisfy this equation. In other words, the individual-level data $\{\pi(\tau), d_E, d_I\}$ *constrains* the population-level data $\hat{\mathbf{I}}_t$ via the intuitive mathematical relationship (7).

2.4. Mathematical details: writing $\mathbf{A}_\pi$ explicitly and approximating with a Toeplitz matrix.

The goal of this subsection is to derive from $\hat{\mathbf{I}}_{T-1} = \mathbf{A}_\pi \hat{\mathbf{I}}_T$ an approximate relationship

$$\hat{\mathbf{I}}'_T \approx \mathbf{A}''_\pi \hat{\mathbf{I}}'_T, \tag{8}$$

where $\hat{\mathbf{I}}'_T := [\hat{I}_3 \ \hat{I}_4 \ \cdots \ \hat{I}_{T-1}]$, and where $\mathbf{A}''_\pi$ is a $(T-3)\times(T-3)$ *Toeplitz* matrix (i.e. a matrix in which the values along any given diagonal are equal). This implies that any reasonable time series $\hat{\mathbf{I}}'_T$ is necessarily approximately in the nullspace of the Toeplitz operator $\mathbf{A}''_\pi - \mathbb{1}$, where $\mathbb{1}$ denotes the $(T-3)\times(T-3)$ identity matrix. This will inform a *pathogenesis signal processing* method for reducing noise in models of the population-level time series $\hat{\mathbf{I}}'_T$, defined by the constraints imposed by the individual-level data $\{\pi(\tau), d_E, d_I\}$. In particular, this signal processing method will utilize the *singular value decomposition* of the operator $\mathbf{A}''_\pi - \mathbb{1}$ (see Section 3 for more details). (The reason we approximate $\mathbf{A}_\pi$ with the Toeplitz operator $\mathbf{A}''_\pi$ is that a lot more can be said about the singular values of Toeplitz matrices than the singular values of non-Toeplitz matrices.)

To start, let us explicitly write out the $(T-1)\times T$ matrix $\mathbf{A}_\pi$. For notational convenience, set

$$D_1 := \left(1 - \frac{1}{d_E}\right)\left(1 - \frac{1}{d_I}\right) \quad \text{and} \quad D_2 := 2 - \frac{1}{d_E} - \frac{1}{d_I}. \tag{9}$$

Then $\mathbf{A}_\pi$ can be written as follows:

[2] This estimate might not work well for diseases with long duration of infection d_I.

$$d_E d_I \cdot \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
\pi(1)D_1 & -\pi(1)D_2 & \pi(1) & 0 & 0 & \cdots & 0 & 0 \\
\pi(2)D_1 & \begin{array}{c} -\pi(2)D_2 \\ +\pi(1)D_1 \end{array} & \begin{array}{c} \pi(2) \\ -\pi(1)D_2 \end{array} & \pi(1) & 0 & \cdots & 0 & 0 \\
\pi(3)D_1 & \begin{array}{c} -\pi(3)D_2 \\ +\pi(2)D_1 \end{array} & \begin{array}{c} \pi(3) \\ -\pi(2)D_2 \\ +\pi(1)D_1 \end{array} & \begin{array}{c} \pi(2) \\ -\pi(1)D_2 \end{array} & \pi(1) & \cdots & 0 & 0 \\
\pi(4)D_1 & \begin{array}{c} -\pi(4)D_2 \\ +\pi(3)D_1 \end{array} & \begin{array}{c} \pi(4) \\ -\pi(3)D_2 \\ +\pi(2)D_1 \end{array} & \begin{array}{c} \pi(3) \\ -\pi(2)D_2 \\ +\pi(1)D_1 \end{array} & \begin{array}{c} \pi(2) \\ -\pi(1)D_2 \end{array} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\pi(T-2)D_1 & \begin{array}{c} -\pi(T-2)D_2 \\ +\pi(T-3)D_1 \end{array} & \begin{array}{c} \pi(T-2) \\ -\pi(T-3)D_2 \\ +\pi(T-4)D_1 \end{array} & \begin{array}{c} \pi(T-3) \\ -\pi(T-4)D_2 \\ +\pi(T-5)D_1 \end{array} & \begin{array}{c} \pi(T-4) \\ -\pi(T-5)D_2 \\ +\pi(T-6)D_1 \end{array} & \cdots & \begin{array}{c} \pi(2) \\ -\pi(1)D_2 \end{array} & \pi(1)
\end{pmatrix}$$

Here, since $\pi(0) = 0$, all of the entries in the first row of $\mathbf{A}_\pi$ are zero. Additionally, for any time series, we have $\hat{I}_1 = \pi(0)\hat{N}_1 = 0$, meaning the first column of $\mathbf{A}_\pi$ does not provide any meaningful information. It follows that the first row and column of $\mathbf{A}_\pi$ can be removed without losing any information.

Let us therefore remove the first row and column of $\mathbf{A}_\pi$; in addition, let us remove the last column $[0\ \cdots\ 0\ \pi(1)]^t$ from $\mathbf{A}_\pi$. Call the resulting $(T-2) \times (T-2)$ matrix $\mathbf{A}'_\pi$. Then $\mathbf{A}'_\pi$ satisfies the approximate relationship

$$[\hat{I}_2\ \hat{I}_3\ \cdots\ \hat{I}_{T-1}]^t \approx \mathbf{A}'_\pi \cdot [\hat{I}_2\ \hat{I}_3\ \cdots\ \hat{I}_{T-1}]^t. \tag{10}$$

Here, by removing the last column from $\mathbf{A}_\pi$, we are approximating the equation

$$\hat{I}_{T-1} = [-\pi(T-2)D_2 + \pi(T-3)D_1]\hat{I}_1 + \cdots + [\pi(2) - \pi(1)D_2]\hat{I}_{T-1} + \pi(1)\hat{I}_T \tag{11}$$

from the last row of $\hat{\mathbf{I}}_{T-1} = \mathbf{A}_\pi \hat{\mathbf{I}}_T$ by

$$\hat{I}_{T-1} \approx [-\pi(T-2)D_2 + \pi(T-3)D_1]\hat{I}_1 + \cdots + [\pi(2) - \pi(1)D_2]\hat{I}_{T-1}. \tag{12}$$

For most pathogenesis distributions, $\pi(1)$ is very small (since it is typically rare for an individual to experience a latent infection period of only one day), so this is a reasonable approximation.

To obtain a square Toeplitz matrix, we will further remove the first row and column of $\mathbf{A}'_\pi$ (equivalently the second row and column of $\mathbf{A}_\pi$), yielding a $(T-3) \times (T-3)$ matrix which we'll call $\mathbf{A}''_\pi$. The resulting matrix $\mathbf{A}''_\pi$ satisfies the approximate relationship

$$\boxed{[\hat{I}_3\ \hat{I}_4\ \cdots\ \hat{I}_{T-1}]^t \approx \mathbf{A}''_\pi \cdot [\hat{I}_3\ \hat{I}_4\ \cdots\ \hat{I}_{T-1}]^t} \tag{13}$$

Notice that this approximation might not be great for the early values of the time series (since the removed column adds a more significant term to these equations). However, later in the time series, the removed terms become negligible (since $\pi(\tau)$ quickly approaches zero as τ becomes large). The matrix $\mathbf{A}''_\pi$ can easily be seen to be Toeplitz:

$$d_E d_I \cdot \begin{pmatrix}
\begin{array}{c}\pi(2)\\ -\pi(1)D_2\end{array} & \pi(1) & 0 & 0 & \cdots & 0 \\[1em]
\begin{array}{c}\pi(3)\\ -\pi(2)D_2\\ +\pi(1)D_1\end{array} & \begin{array}{c}\pi(2)\\ -\pi(1)D_2\end{array} & \pi(1) & 0 & \cdots & 0 \\[1em]
\begin{array}{c}\pi(4)\\ -\pi(3)D_2\\ +\pi(2)D_1\end{array} & \begin{array}{c}\pi(3)\\ -\pi(2)D_2\\ +\pi(1)D_1\end{array} & \begin{array}{c}\pi(2)\\ -\pi(1)D_2\end{array} & \pi(1) & \cdots & 0 \\[1em]
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\[1em]
\begin{array}{c}\pi(T-3)\\ -\pi(T-4)D_2\\ +\pi(T-5)D_1\end{array} & \begin{array}{c}\pi(T-4)\\ -\pi(T-5)D_2\\ +\pi(T-6)D_1\end{array} & \begin{array}{c}\pi(T-5)\\ -\pi(T-6)D_2\\ +\pi(T-7)D_1\end{array} & \cdots & \cdots & \pi(1) \\[1em]
\begin{array}{c}\pi(T-2)\\ -\pi(T-3)D_2\\ +\pi(T-4)D_1\end{array} & \begin{array}{c}\pi(T-3)\\ -\pi(T-4)D_2\\ +\pi(T-5)D_1\end{array} & \begin{array}{c}\pi(T-4)\\ -\pi(T-5)D_2\\ +\pi(T-6)D_1\end{array} & \begin{array}{c}\pi(T-5)\\ -\pi(T-6)D_2\\ +\pi(T-7)D_1\end{array} & \cdots & \begin{array}{c}\pi(2)\\ -\pi(1)D_2\end{array}
\end{pmatrix}$$

Here, the entries of $\mathbf{A}''_\pi$ are given by

$$(\mathbf{A}''_\pi)_{s,t} = d_E d_E(\pi(s-t+2) - \pi(s-t+1)D_2 + \pi(s-t)D_1). \tag{14}$$

Finally, define

$$\mathbf{L}_\pi := \mathbf{A}''_\pi - \mathbb{1} \tag{15}$$

where $\mathbb{1}$ denotes the $(T-3)\times(T-3)$ identity matrix, and set

$$\hat{\mathbf{I}}'_T := [\hat{I}_3 \ \hat{I}_4 \ \cdots \ \hat{I}_{T-1}]. \tag{16}$$

Then $\mathbf{L}_\pi$ is a $(T-3)\times(T-3)$ Toeplitz matrix such that

$$\boxed{\mathbf{L}_\pi \hat{\mathbf{I}}'_T \approx \mathbf{0}} \tag{17}$$

3. Signal Processing Using the Effective Rank of $\mathbf{L}_\pi$

What we have established above is that any realistic time series $\hat{\mathbf{I}}'_T$ should approximately be in the nullspace of the Toeplitz operator $\mathbf{L}_\pi$, which is defined in terms of the pathogenesis distribution $\pi(\tau)$ of durations between the time of infection and the time of symptom onset. In other words, a "reasonable" time series $\hat{\mathbf{I}}_T$ (population-level data) should be in some sense "compatible" with the pathogenesis distribution $\pi(\tau)$ (individual-level), and the relationship $\mathbf{L}_\pi \hat{\mathbf{I}}'_T \approx 0$ encodes this compatibility condition.

This suggests a powerful *pathogenesis signal processing* approach for noise reduction in our models of $\hat{\mathbf{I}}'_T$. In particular, suppose we have an *epi-curve* modeling $\hat{\mathbf{I}}'_T$, built from some population-level data (e.g. reported case and hospitalization numbers). (A preliminary model of $\hat{\mathbf{I}}'_T$ using population-level data is typically relatively straightforward to obtain, since $\hat{\mathbf{I}}'_T$ lives on the population level.) However, an epi-curve built from population-level data alone will likely have a lot of noise that is not compatible with individual-level biology. However, from the relationship $\mathbf{L}_\pi \hat{\mathbf{I}}'_T \approx 0$, we know that any reasonable model of $\hat{\mathbf{I}}'_T$ must approximately live in the nullspace of $\mathbf{L}_\pi$. By projecting our epi-curve onto the nullspace of $\mathbf{L}_\pi$, we should therefore expect to obtain a new "smoothed out" curve that both models the population-level data and is compatible with the individual-level pathogenesis.

An issue remains: $\mathbf{L}_\pi$ is typically, strictly speaking, a full-rank (or close to full-rank) matrix, so there's no well-defined nullspace on which to project the raw epi-curve. This is because projecting onto the exact nullspace of $\mathbf{L}_\pi$ is too strict of a requirement: instead of just checking reasonable compatibility with the individual-level biology, projecting onto the exact nullspace would remove anything that doesn't fit *exactly* with our pathogenesis distribution. What we need are the right notions of *effective rank* and *effective nullspace* that help us understand when a time series model is close enough to being in the nullspace of $\mathbf{L}_\pi$ that it is still reasonably compatible with individual-level biology.

One possible notion of effective rank – introduced by Roy and Vetterli in [5] – entails viewing the symmetric positive-definite matrix $\mathbf{L}_\pi^t \mathbf{L}_\pi$ as the covariance matrix of a Gaussian process and calculating the exponential entropy of the resulting Gaussian process:

Definition 1 (Roy-Vetterli effective rank [5])**.** Let $\sigma_j(\mathbf{L}_\pi)$ denote the j-th singular value of $\mathbf{L}_\pi$, where the singular values are arranged in descending order. Then the *Roy-Vetterli effective rank* of $\mathbf{L}_\pi$ is defined as

$$r_{\text{eff}}(\mathbf{L}_\pi) := \exp\left(\sum_{j=1}^{T-2}\left(\frac{-\sigma_j(\mathbf{L}_\pi)}{\sum_{j=1}^{T-2}\sigma_j(\mathbf{L}_\pi)} \cdot \ln\left(\frac{\sigma_j(\mathbf{L}_\pi)}{\sum_{j=1}^{T-2}\sigma_j(\mathbf{L}_\pi)}\right)\right)\right). \tag{18}$$

The idea behind this definition is as follows: since the *rank* of $\mathbf{L}_\pi$ equals the number of *nonzero* singular values of $\mathbf{L}_\pi$, the *effective rank* of $\mathbf{L}_\pi$ should equal the number of singular of $\mathbf{L}_\pi$ that are *statistically distinguishable from zero.* The exponential entropy characterizes the number of uncorrelated degrees of freedom, and therefore gives one way to define/estimate the number of singular values that are statistically distinguishable from zero.

Example. In [7], Thakkar and Famulare use COVID-19 data on reported cases C_t and hospitalizations H_t from Washington state between January 2020 and March 2021 (spanning $T = 437$ days) to construct a compartmental, stochastic process model of COVID-19 infection rates. This results in a raw epi-curve that is proportional to the number of infectious people over time (see [7] for the details of the COVID-19 epi-curve construction). Since this model does not incorporate any individual-level biology, the resulting curve has a lot of noise arising from phenomena such as weekend effects.

Using the Roy-Vetterli notion of effective rank and the pathogenesis distribution from [4], Thakkar and Famulare get that $r_{\text{eff}}(\mathbf{L}_\pi) = 371$. With this, they project their raw epi-curve onto the resulting 66-dimensional effective nullspace, yielding the following smoothed curve:

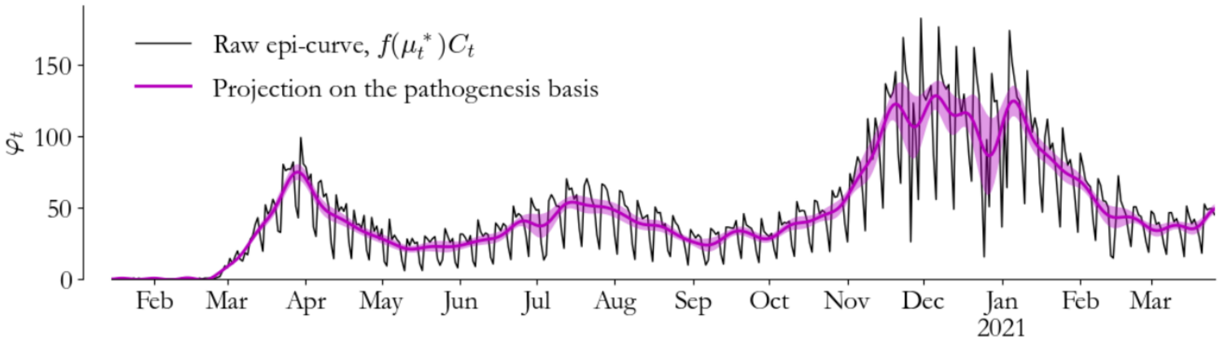

Figure 2. The raw epi-curve (black) and the projection of the raw epi-curve onto the effective nullspace of $\mathbf{L}_\pi$ (purple), as computed in [7].

While this clearly accomplishes a reduction of noise and demonstrates the power of this signal processing approach, a concern remains: The Roy-Vetterli effective rank definition proposed by [5] is very general and the application of the Roy-Vetterli effective rank to this setting is not very well-motivated. As a result, in doing this projection onto the Roy-Vetterli effective nullspace, Thakkar and Famulare might be classifying some parts of the raw epi-curve as noise that should in fact be considered trend (or vice versa). Moreover, as we will soon see, the singular value distributions of the possible $\mathbf{L}_\pi$ operators are extremely structured. This suggests that there should be a way to make a more informed estimate of the effective rank of $\mathbf{L}_\pi$ that is more specific to our setting. The next section of this paper is dedicated to proposing such an alternative notion of effective rank.

4. An Alternative Notion of Effective Rank

In Subsection 4.1, we establish a property that we should expect an appropriate notion of *effective rank* to satisfy. In Subsection 4.2, we show that the Roy-Vetterli effective rank does not satisfy this desired property. Motivated by this discussion, we define an alternative notion of effective rank in Subsection 4.3; this alternative notion is essentially defined in terms of an "elbow test" on the graph of the singular value decomposition of $\mathbf{L}_\pi$. To help provide some mathematical rigor and justification for this alternative notion, we discuss the singular values of Toeplitz matrices (and specifically the singular values of $\mathbf{L}_\pi$) in Subsection 4.4. Based on results and observations about these singular values, we formulate our primary conjecture in Subsection 4.5; if proved, this conjecture would show that this "elbow test" notion of effective rank has an equivalent mathematically rigorous formulation, and that this notion of effective rank satisfies the desired property outlined in Subsection 4.1. In Subsection 4.6, we conclude the section with a discussion of some potential drawbacks to this alternative notion of effective rank.

4.1. **A desired property of the effective rank.** Since $\mathbf{L}_\pi$ is a $(T-3)\times(T-3)$ matrix, the last $T-3-r_{\text{eff}}(\mathbf{L}_\pi)$ right singular vectors of $\mathbf{L}_\pi$ form a basis for the *Roy-Vetterli effective nullspace* of $\mathbf{L}_\pi$. Since any "reasonable" time series $\hat{\mathbf{I}}'_T$ is in the effective nullspace of $\mathbf{L}_\pi$, any such time series can be written as a linear combination of these right singular vectors. We should therefore expect that the right singular vectors in the effective nullspace of $\mathbf{L}_\pi$ are relatively "smooth," since they are coming from a signal in our biological data; otherwise, they will contribute to noise in our model. Likewise, we should expect the right singular vectors in the effective range of $\mathbf{L}_\pi$ to look "noisy"; otherwise, we might be omitting meaningful data. Putting this all together, we should expect / hope for the following: **the effective rank of $\mathbf{L}_\pi$ should delineate between "noisy" and "smooth" *right singular vectors* (more commonly called *modes*)**.

Example (continued)**.** To see what this looks like in practice, let's revisit the example above from [7], where we saw a raw epi-curve modeling COVID-19 cases and its projection onto $\mathbf{L}_\pi$'s Roy-Vetterli effective nullspace. Below we've plotted again this epi-curve and projection, along with some of the modes in the Roy-Vetterli effective nullspace (purple) and in the Roy-Vetterli effective range (black). As we can see, the purple modes are relatively "smooth," while the black modes are relatively "noisy" (as expected).

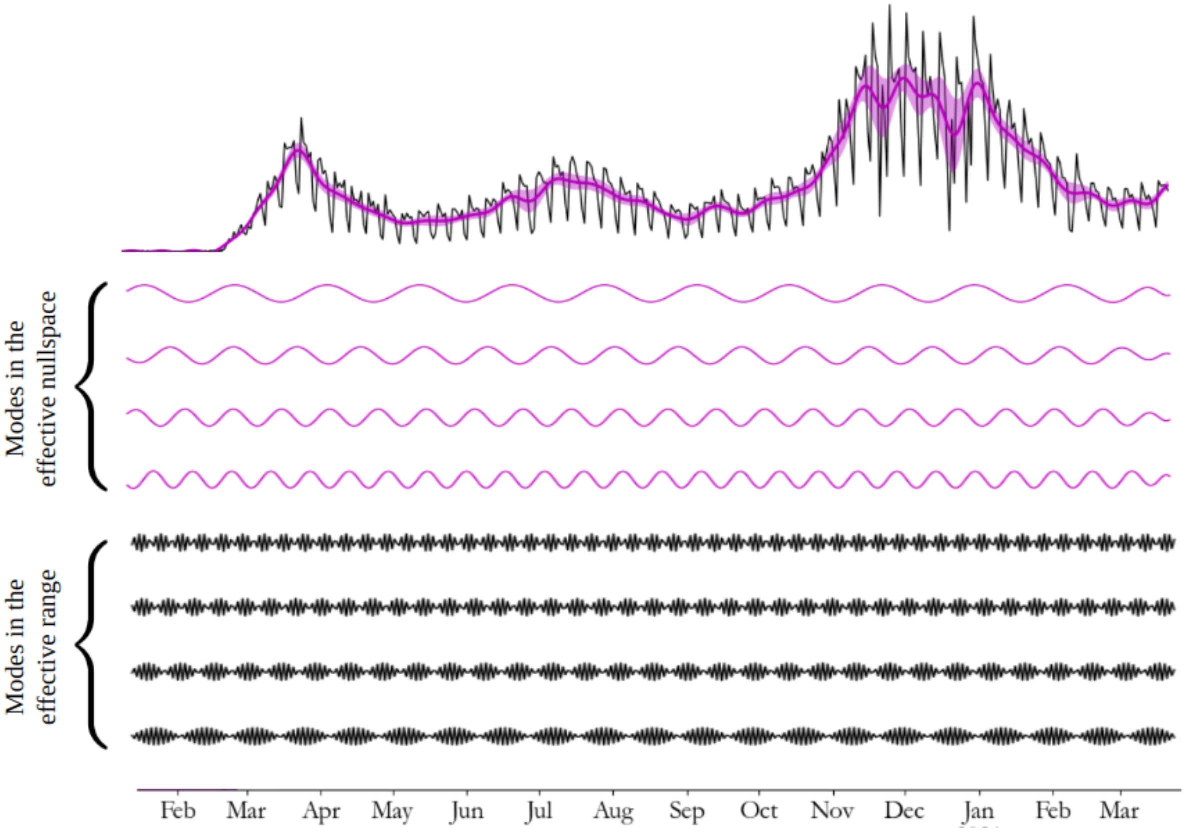

Figure 3. A COVID-19 epi-curve and projection onto the Roy-Vetterli effective nullspace of $\mathbf{L}_\pi$ [7] (top), modes in the effective nullspace of $\mathbf{L}_\pi$ (middle), and modes in the effective range of $\mathbf{L}_\pi$ (bottom).

While these specific modes look as expected, they were sampled from singular value indices comfortably above and comfortably below where we expect the effective rank to be. There

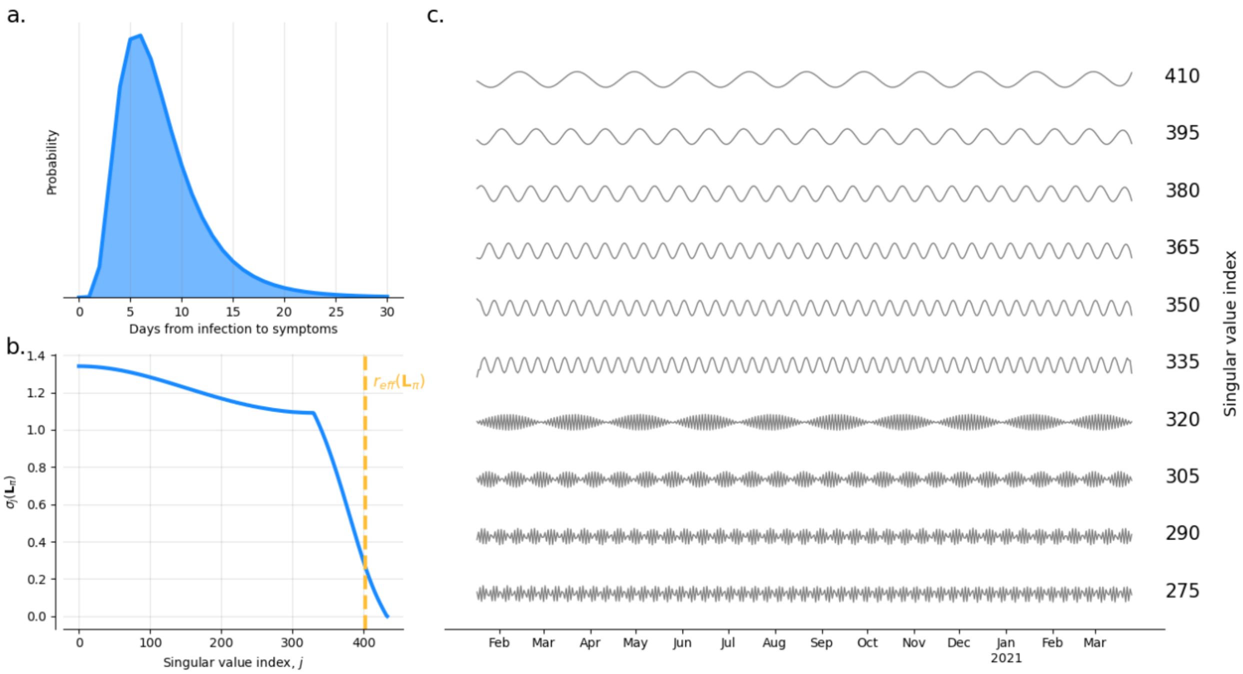

Figure 4. (a) the log-normal pathogenesis distribution $\pi(\tau)$ with mean $\mu = 8.06$ and variance $\sigma^2 = 18$, (b) the singular value distribution of $\mathbf{L}_\pi$, and (c) some of the modes of $\mathbf{L}_\pi$.

is a potential "problem area" at singular value indices near where we expect the effective rank to be. The following subsection aims to answer the following question: How good of a job does the Roy-Vetterli effective rank do at delineating between "noisy" and "smooth" modes?

4.2. **Testing the Roy-Vetterli effective rank against this desired property.** In this subsection, we look at some example pathogenesis distributions, and see whether the Roy-Vetterli effective rank provides a good delineation between "noisy" and "smooth" modes. To start, let's consider a log-normal pathogenesis distribution $\pi(\tau)$ with mean $\mu = 8.06$ and variance $\sigma^2 = 18$ (roughly the mean and variance of the distribution proposed in [4] for COVID-19), shown in Figure 4. In this case, the Roy-Vetterli effective rank of $\mathbf{L}_\pi$ is $r_{\text{eff}}(\mathbf{L}_\pi) = 402$. However, we see that there is a clear delineation between "noisy" and "smooth" modes that occurs around index 330.

Continuing with the log-normal pathogenesis distribution with mean $\mu = 8.06$ and increasing the variance to $\sigma^2 = 40$, we obtain the plots shown in Figure 5. In this case, $r_{\text{eff}}(\mathbf{L}_\pi) = 392$. However, there is again a clear delineation between "noisy" and "smooth" modes that occurs around index 420.

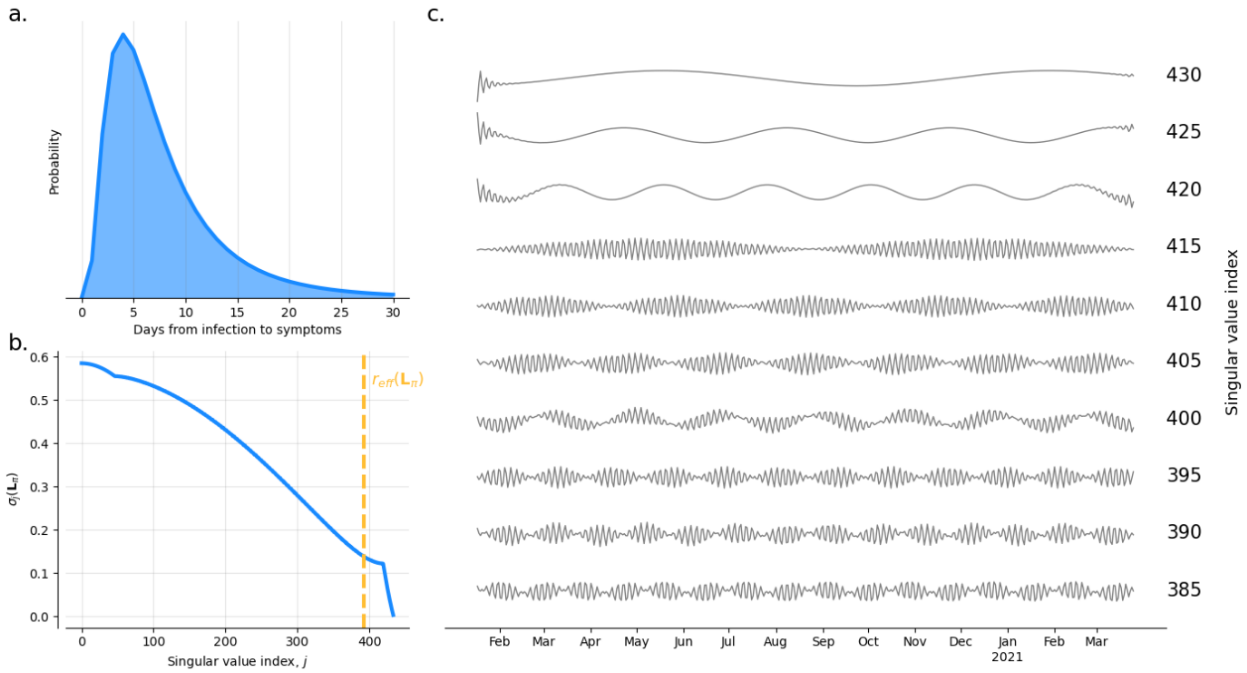

Figure 5. (a) the log-normal pathogenesis distribution $\pi(\tau)$ with mean $\mu = 8.06$ and variance $\sigma^2 = 40$, (b) the singular value distribution of $\mathbf{L}_\pi$, and (c) some of the modes of $\mathbf{L}_\pi$.

This phenomenon is not unique to log-normal pathogenesis distributions. For example, we can consider the following mixture sensitivity forward time model proposed in [4]:

$$\pi(\tau) = \alpha\lambda\left(\mu(\tau\lambda)^{\alpha-1} + \frac{1-\mu}{\Gamma(1/\alpha)}\right)e^{-(\tau\lambda)^\alpha}. \tag{19}$$

Taking $\mu = 1$, $\alpha = 2.5$, and $\lambda = .11$, we obtain the plots shown in Figure 6. In this case, $r_{\text{eff}}(\mathbf{L}_\pi) = 410$. However, there again appears to be a clear delineation between "noisy" and "smooth" modes, this time occurring around index 375.

In all of these examples (as well as in plenty more that are not included here), it appears that the Roy-Vetterli effective rank does not do a good job of delineating between "noisy" and "smooth" modes, suggesting that it might not be a great notion of effective rank for our purposes. Fortunately, these examples suggest an alternative notion that might be more appropriate.

4.3. **Defining an alternative notion of effective rank.** In all of the examples from the previous subsection, notice that the plots of the singular value distributions have sharp "elbows" after which point the singular values quickly drop off. Moreover, this elbow appears to occur exactly at the point of delineation between "noisy" and "smooth" modes. These observations suggest an alternative possibility for a definition of effective rank.

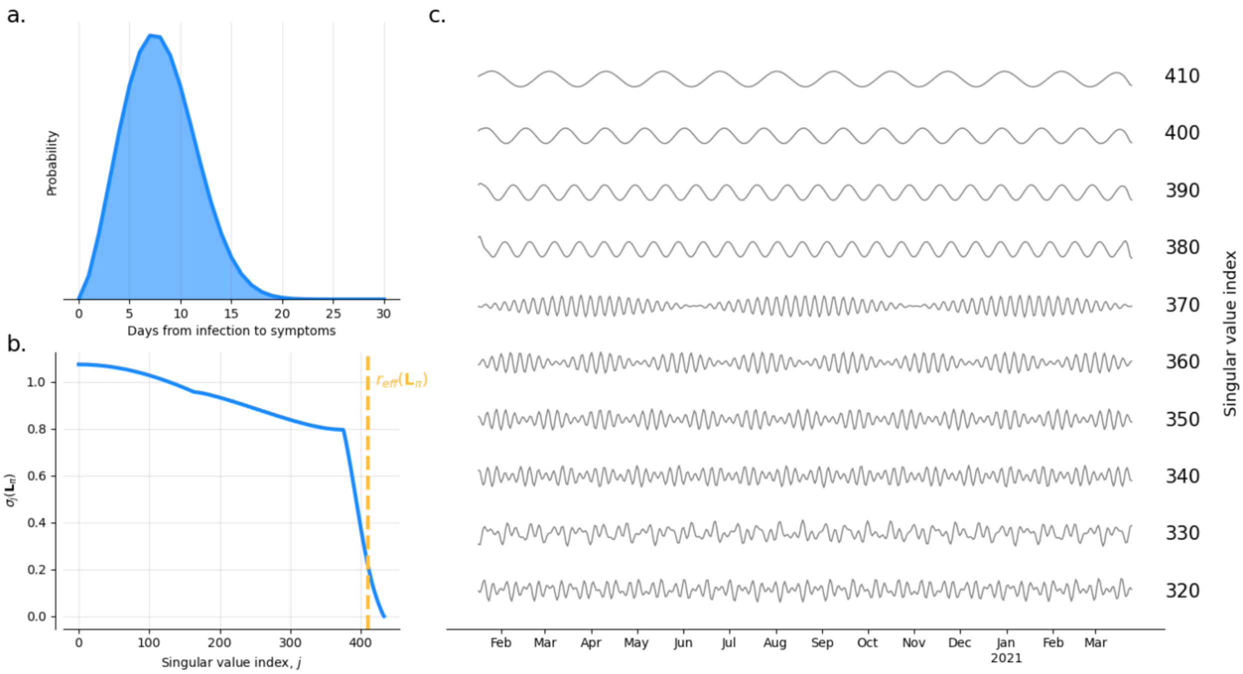

Figure 6. (a) the mixture sensitivity pathogenesis distribution $\pi(\tau)$ with mean $\mu = 1$, $\alpha = 2.5$, and $\lambda = .11$, (b) the singular value distribution of $\mathbf{L}_\pi$, and (c) some of the modes of $\mathbf{L}_\pi$.

Definition 2 (Pathogenesis effective rank)**.** Let π be a pathogenesis distribution with corresponding Toeplitz matrix $\mathbf{L}_\pi$. Suppose that the singular value distribution of $\mathbf{L}_\pi$ has an "elbow" at singular value index j. Then j is the *pathogenesis effective rank* of $\mathbf{L}_\pi$.[3]

There is a strong precedent for using "elbow" tests in mathematical optimization, cluster analysis, and multivariate statistics. In cluster analysis, the "elbow method" is used to determine a number of clusters at which including another cluster doesn't provide much better modeling of the data [1]. In multivariate statistics, "elbows" arising in "Scree plots" are used to determine the number of factors to retain in an exploratory factor analysis (FA) or the number of principal components to keep in a principal component analysis (PCA) [2].

However, elbow methods are widely criticized as being subjective and unreliable, especially in cases where the plot does not appear to have a sharp elbow. Therefore, in order to justify this "elbow" notion of effective rank, it would be extremely helpful to connect this approach to something more mathematically rigorous. This is where we will make use of results about the singular value decompositions of Toeplitz matrices.

4.4. **Singular values of Toeplitz matrices.** While the Roy-Vetterli effective rank can be defined for any matrix, recall that $\mathbf{L}_\pi$ is a Toeplitz matrix (i.e. the values along any given diagonal of $\mathbf{L}_\pi$ are equal). The singular values of Toeplitz matrices are very well-studied due to the wide range of applications of Toeplitz matrices to mathematical modelling problems where some sort of shift invariance occurs. For example, Toeplitz matrices are used in

[3] Note that the *pathogenesis effective rank* is not defined for π such that there is no "elbow" present in the graph of the corresponding singular value distribution.

modeling the numerical solutions of certain differential equations, time series analysis, signal and image processing, Markov chains and queueing theory, and polynomial and power series computations.

In this subsection, we largely follow [3], which describes a method for studying the singular values of a Toeplitz matrix in terms of the *generating function* associated to the Toeplitz matrix. Given an $n \times n$ Toeplitz matrix A_n, one can define $\hat{f}_k$ to be the value along the k-th diagonal of A_n (for $k = 1-n, \ldots, n-1$) and $\hat{f}_k$ to be 0 for $|k| > n-1$. Then $(A_n)_{s,t} = \hat{f}_{s-t}$ (for $s, t = 1, \ldots, n$). The Fourier series

$$f(x) := \sum_{-\infty}^{\infty} \hat{f}_k \cdot e^{ikx} = \sum_{k=1-n}^{n-1} \hat{f}_k \cdot e^{ikx} \tag{20}$$

is called the *generating function* of the Toeplitz matrix A_n. Representing a Toeplitz matrix by its generating function allows us to use results from calculus and Fourier analysis while studying its spectrum and singular values.

Remark 1. Note that mathematicians typically define this correspondence in the opposite order, beginning with a generating function $f \in L^1(-\pi, \pi)$, and *defining* the associated Toeplitz matrix $A_n(f)$ so that $(A_n(f))_{s,t} = \hat{f}_{s-t}$ (for $s, t = 1, \ldots, n$), where $\hat{f}_k$ is the k-th Fourier coefficient of f:

$$\hat{f}_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) e^{-ik\theta}\, d\theta, \quad k \in \mathbb{Z}. \tag{21}$$

This relationship between a Toeplitz matrix and its associated generating function is a very natural one-to-one correspondence: truncating a Toeplitz matrix corresponds to truncating the associated Fourier series, and taking the product of two Toeplitz matrices also has an analog on the function side. If the generating function f is real-valued, then the corresponding Toeplitz matrix is symmetric and several spectral properties are known. If f is complex-valued (which will be the case for our Toeplitz matrix $\mathbf{L}_\pi$), certain properties about the singular values are known, while the eigenvalues have “wild” behavior in some cases and regular behavior in other cases.

The paper [3] establishes upper and lower bounds for the singular values of a Toeplitz matrix in terms of properties of its associated generating function:

Proposition 1 ([3, Corollary 2])**.** *Let $f \in L^1(-\pi, \pi)$ be such that $\hat{f}_k \in \mathbb{R}$ for any integer k. Let d be the distance of the essential range of f from the complex zero, and let M be the essential supremum of $|f|$. Then all of the singular values of $A_n(f)$ belong to $[d, M]$ for any size n.*

For $\mathbf{L}_\pi$, the associated generating function f_π is a Fourier series with coefficients

$$\begin{array}{rcll} (\hat{f}_\pi)_0 & = & d_E d_I(\pi(2) - \pi(1)D_2) - 1 & \\ (\hat{f}_\pi)_k & = & d_E d_I(\pi(k+2) - \pi(k+1)D_2 + \pi(k)D_1) & (0 < |k| \leq T-4) \\ (\hat{f}_\pi)_k & = & 0 & (|k| > T-4) \end{array} \tag{22}$$

Since $\pi(\tau) = 0$ for $\tau \leq 0$, we also get that $(\hat{f}_\pi)_k = 0$ for $k \leq -2$, and hence that

$$f_\pi(x) = \sum_{-\infty}^{\infty} (\hat{f}_\pi)_k \cdot e^{ikx} = \sum_{k=-1}^{T-4} (\hat{f}_\pi)_k \cdot e^{ikx}. \tag{23}$$

Therefore, we have that

$$|f_\pi(x)| = \left(\left[\sum_{k=-1}^{T-4} (\hat{f}_\pi)_k \cdot \sin(kx) \right]^2 + \left[\sum_{k=-1}^{T-4} (\hat{f}_\pi)_k \cdot \cos(kx) \right]^2 \right)^{1/2}. \tag{24}$$

It is not hard to see that this is a continuous 2π-periodic function. Therefore, by the extreme value theorem, $|f_\pi(x)|$ has a minimum and maximum value in the interval $[0, 2\pi]$, so the phrase "essential supremum" in Proposition 1 can be replaced with "maximum" for our purposes.

We can prove that the distance d of the essential range of f_π from the complex zero is always 0. For this, it suffices to show that $f_\pi(0) = 0$. To this end, note that we have

$$\begin{aligned} f_\pi(0) = \sum_{k=-1}^{T-4} (\hat{f}_\pi)_k = -1 + d_E d_I \cdot \big[&\pi(1) + \pi(2) - \pi(1) D_2 \\ &+ \pi(3) - \pi(2) D_2 + \pi(1) D_1 \\ &+ \pi(4) - \pi(3) D_2 + \pi(2) D_1 \\ &\vdots \\ &+ \pi(T-3) - \pi(T-4) D_2 + \pi(T-5) D_1 \\ &+ \pi(T-2) - \pi(T-3) D_2 + \pi(T-4) D_1 \big]. \end{aligned}$$

Noting that $d_E d_I (1 - D_2 + D_1) = 1$ and recalling that $\pi(\tau) = 0$ for $\tau \geq T - 3$, we see that

$$f_\pi(0) = -1 + \sum_{\tau=1}^{T-4} \pi(\tau) = -1 + 1 = 0, \tag{25}$$

where we have used that $\pi(\tau)$ is a probability distribution. With this, we see that Proposition 1 tells us the following about the singular values of $\mathbf{L}_\pi$:

Corollary 1. *Let* $M = \max_{[0,2\pi]}(|f_\pi(x)|)$. *Then all of the singular values of* $\mathbf{L}_\pi$ *belong to the interval* $[0, M]$.

We can verify this by plotting the singular value distribution of $\mathbf{L}_\pi$ along side the graph of $|f_\pi(x)|$ for various pathogenesis distributions π. For consistency, let's consider again the examples considered above.

To start, let's consider the log-normal pathogenesis distribution with mean $\mu = 8.06$ and variance $\sigma^2 = 18$, shown in Figure 7. As we can see, the singular values of $\mathbf{L}_\pi$ indeed fall between the minimum of $|f_\pi(x)|$ (i.e. zero) and the maximum of $|f_\pi(x)|$. Notice also that the singular value at which the "elbow" in the graph of the singular value distribution of $\mathbf{L}_\pi$ occurs appears to be equal to the value of $|f_\pi(x)|$ at its local minimum ($x = \pi$).[4]

Next, let's consider the plots for a log-normal pathogenesis distribution with mean $\mu = 8.06$ and variance $\sigma^2 = 40$, shown in Figure 8. We again see that the singular values of $\mathbf{L}_\pi$ fall between the minimum of $|f_\pi(x)|$ (i.e. zero) and the maximum of $|f_\pi(x)|$. We also again see

[4] Note that π is being used in two different ways here: as the pathogenesis distribution $\pi(\tau)$ and as the value $\pi = 3.14...$. To stay consistent with convention, we'll maintain this potentially confusing notation and hope that the meaning is always clear from context.

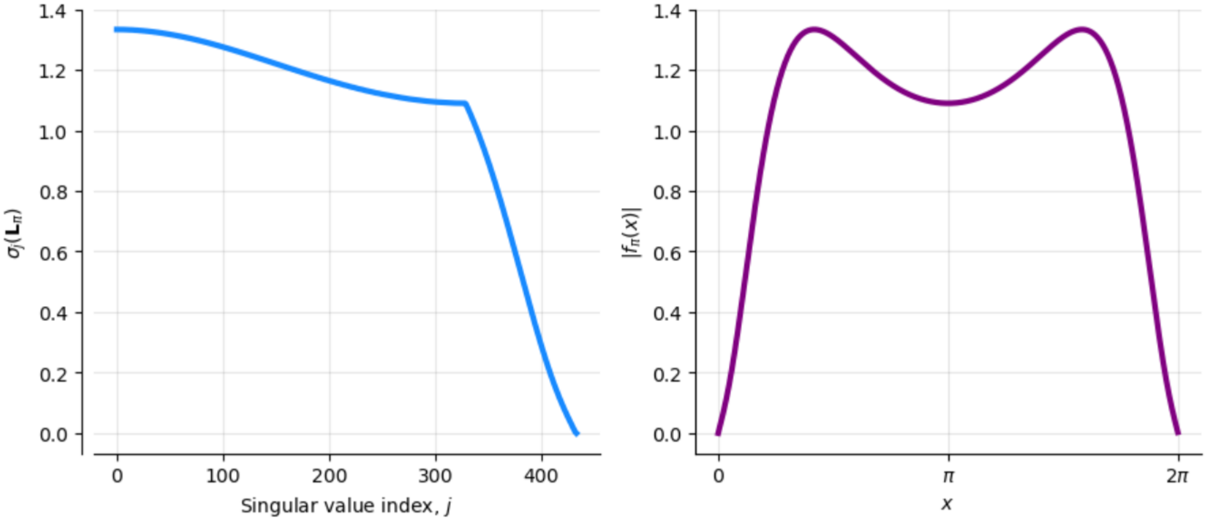

FIGURE 7. The singular value decomposition of $\mathbf{L}_\pi$ (left) and the generating function $|f_\pi(x)|$ (right) for the log-normal pathogenesis distribution $\pi(\tau)$ with mean $\mu = 8.06$ and variance $\sigma^2 = 18$.

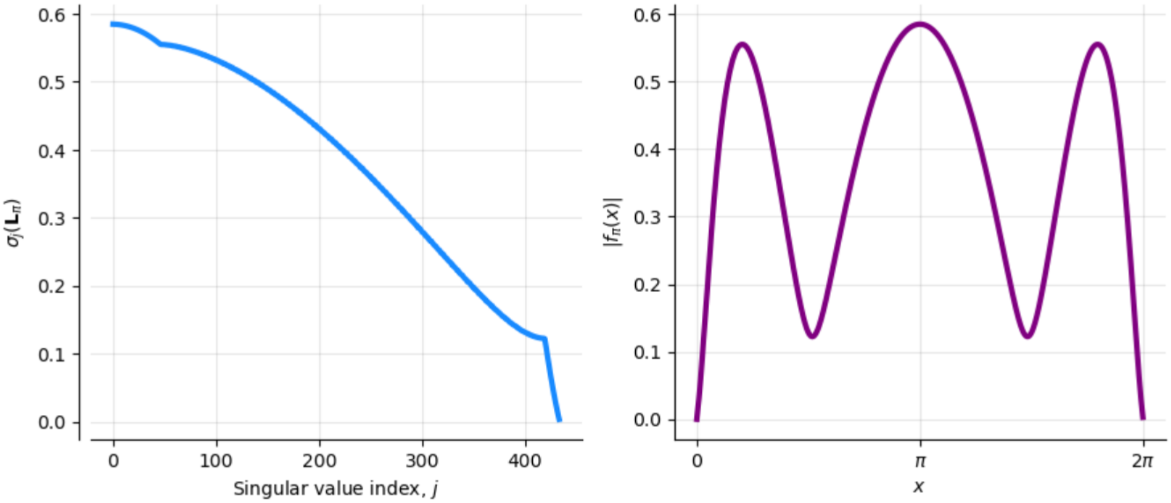

FIGURE 8. The singular value decomposition of $\mathbf{L}_\pi$ (left) and the generating function $|f_\pi(x)|$ (right) for the log-normal pathogenesis distribution $\pi(\tau)$ with mean $\mu = 8.06$ and variance $\sigma^2 = 40$.

that the singular value at which the "elbow" in the graph of the singular value distribution of $\mathbf{L}_\pi$ occurs appears to be equal to the value of $|f_\pi(x)|$ at its local minimum. Moreover, there appears to be another sharp point in the graph of the singular value distribution, roughly around singular value index 40; the singular value at this index appears to be equal to the value of $|f_\pi(x)|$ at one of its local maxima (the one slightly below its global maximum).

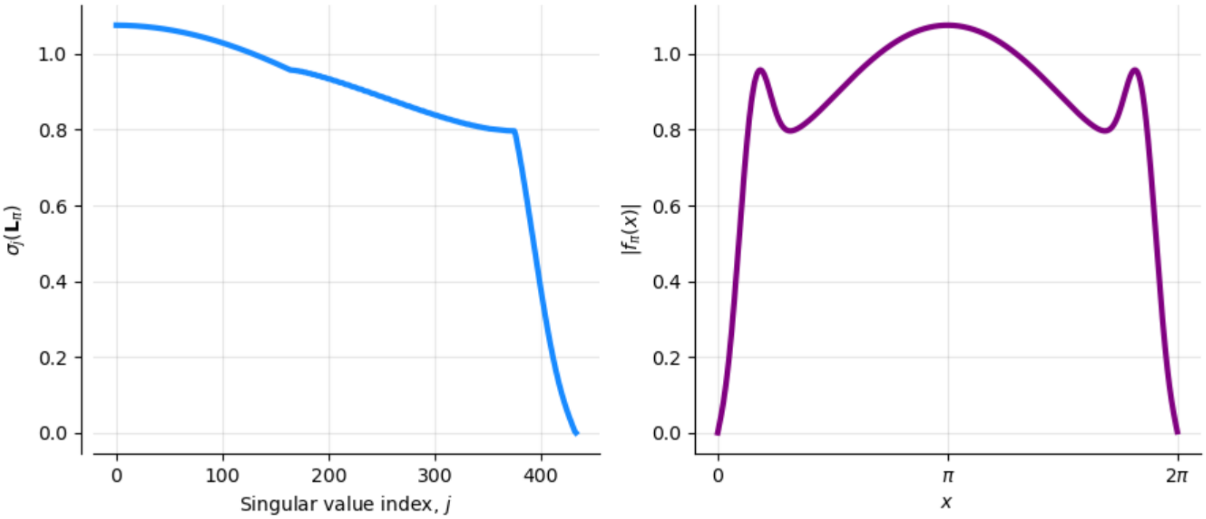

FIGURE 9. The singular value decomposition of $\mathbf{L}_\pi$ (left) and the generating function $|f_\pi(x)|$ (right) for the mixture sensitivity pathogenesis distribution $\pi(\tau)$ with $\mu = 1$, $\alpha = 2.5$, and $\lambda = .11$.

Finally, let's consider the plots for the mixure sensitivity forward time model (19) with $\mu = 1$, $\alpha = 2.5$, and $\lambda = .11$, shown in Figure 9. Again, the singular values of $\mathbf{L}_\pi$ fall within the range of $|f_\pi(x)|$, and the "elbow" in the graph of the singular value distribution of $\mathbf{L}_\pi$ coincides with the local minimum of $|f_\pi(x)|$. There also again appears to be an additional sharp point in the graph of the singular value distribution, which appears to coincide with one of the local maxima of $|f_\pi(x)|$.

4.5. **Our primary conjecture.** Based on the discussion in the previous subsection (as well as numerous other examples not included in this paper), we can formulate the following conjecture:

Conjecture 1. *Let π be a pathogenesis distribution with corresponding Toeplitz matrix $\mathbf{L}_\pi$ such that the graph of the singular value distribution of $\mathbf{L}_\pi$ has an "elbow" at singular value index j. Then we have the following:*

(i) the y-value of the smallest nonzero local minimum of $|f_\pi|$ equals $\sigma_j(\mathbf{L}_\pi)$;
(ii) at singular value index j, the modes of $\mathbf{L}_\pi$ switch from "noisy" to "smooth."

Now, if proved, why would this conjecture be useful? As things stand, our definition of pathogenesis effective rank (Definition 2) is somewhat subjective: it relies on visually identifying an "elbow" in the graph of the singular value distribution of $\mathbf{L}_\pi$ and estimating (by "eye-balling") the singular value index at which it occurs. If part (i) of the above conjecture was proved, we would have the following precise (and not at all subjective) method for calculating the effective rank of $\mathbf{L}_\pi$:

(a) Calculate the roots r_i of the derivative of $|f_\pi(x)|$.
(b) For each $0 < r_i < 2\pi$, calculate $|f_\pi(r_i)|$, and take the minimum resulting value, m.
(c) The effective rank of $\mathbf{L}_\pi$ is the cardinality of the set $\{\sigma_j(\mathbf{L}_\pi) \ : \ \sigma_j(\mathbf{L}_\pi) > m\}$.

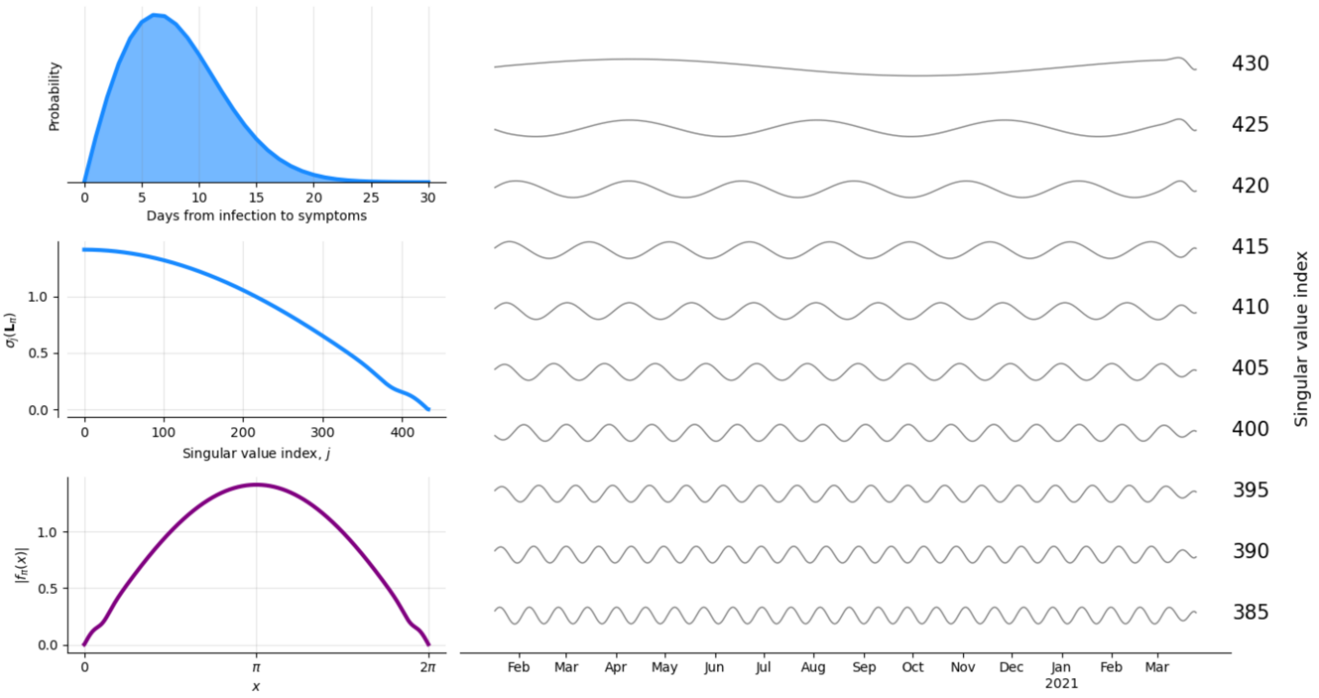

FIGURE 10. The mixture sensitivity pathogenesis distribution $\pi(\tau)$ with $\mu = 1$, $\alpha = 1.97$, and $\lambda = .11$ (top left), the singular value decomposition of $\mathbf{L}_\pi$ (middle left), the generating function $|f_\pi(x)|$ (bottom left), and some modes of $\mathbf{L}_\pi$ (right).

This provides some mathematical rigor that is often lacking in "elbow" test methods.

Additionally, if part *(ii)* of the above conjecture was proved, we would know for certain that the pathogenesis effective rank satisfies the desired property that we outlined at the beginning of this section for a notion of effective rank. This would give strong justification for using the pathogensis effective rank instead of the Roy-Vetterli effective rank for the purposes of pathogenesis signal processing (when an "elbow" occurs in the graph of the singular value distribution of $\mathbf{L}_\pi$).

4.6. **Potential drawbacks of the pathogenesis effective rank.** While the conjecture in the previous subsection would give strong justification for the use of the pathogenesis effective rank in pathogenesis signal processing, there are some potential drawbacks to this notion of effective rank.

4.6.1. *"Edge cases" in which the pathogenesis effective rank is not defined.* The most significant potential drawback is that the pathogenesis effective rank is only defined for pathogenesis distributions $\pi(\tau)$ such that the graph of the singular value distribution of $\mathbf{L}_\pi$ has an "elbow" (or, equivalently – if Conjecture 1 is true – such that $|f_\pi(x)|$ has a local minimum). While most of the log-normal distributions and mixture sensitivity distributions do seem to yield an "elbow," there are some examples that lack "elbows." These examples seem to be "edge cases" in the sense that slightly modifying the parameters in any direction causes "elbows" to appear.

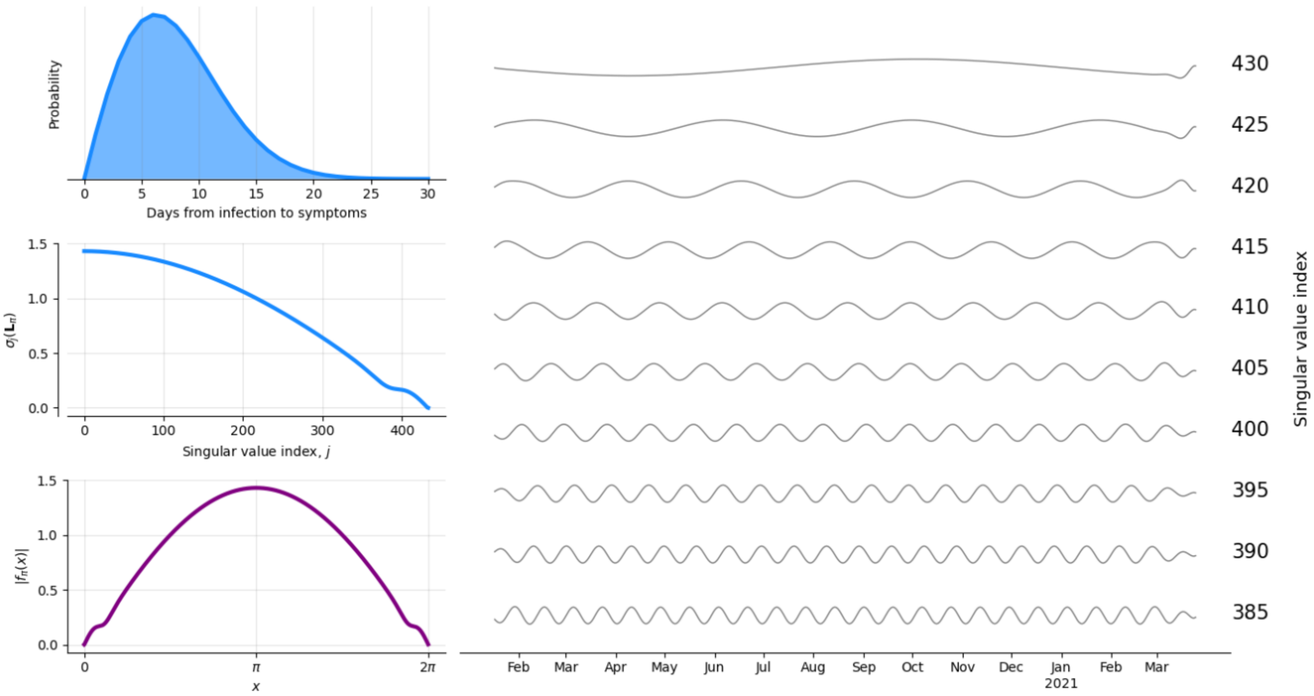

FIGURE 11. The mixture sensitivity pathogenesis distribution $\pi(\tau)$ with $\mu = 1$, $\alpha = 1.97$, and $\lambda = .112$ (top left), the singular value decomposition of $\mathbf{L}_\pi$ (middle left), the generating function $|f_\pi(x)|$ (bottom left), and some modes of $\mathbf{L}_\pi$ (right).

The pathogenesis distribution for COVID-19 introduced in [4] and used in [7] is one of these "edge case" examples. In particular, for modeling COVID-19, [4] suggests the mixture sensitivity model (19) with $\mu = 1$, $\alpha = 1.97$, and $\lambda = .11$, shown in Figure 10. As we can see, the graph of the singular value distribution of $\mathbf{L}_\pi$ does not contain an "elbow," and $|f_\pi(x)|$ does not have any local extrema besides the global minima at even multiples of π and the global maxima at odd multiples of π. As we might expect in this situation, there is also no clear delineation between "noisy" and "smooth" modes.

Next, let's consider the mixture sensitivity pathogenesis distribution with μ and α unchanged, and with λ slightly increased to .112, shown in Figure 11. In this case, we still do not observe any "elbows," local extrema, or "noisy" modes. However, compared to the previous set of plots, this graph of the singular value distribution of $\mathbf{L}_\pi$ appears slightly closer to having an "elbow," and the graph of $|f_\pi(x)|$ appears slightly closer to having local extrema.

Next, let's consider what happens when we again leave μ and α unchanged and slightly increase λ to .114, shown in Figure 12. In this case, we see that there now is something more "elbow"-like in the graph of the singular value distribution of $\mathbf{L}_\pi$ near singular value index

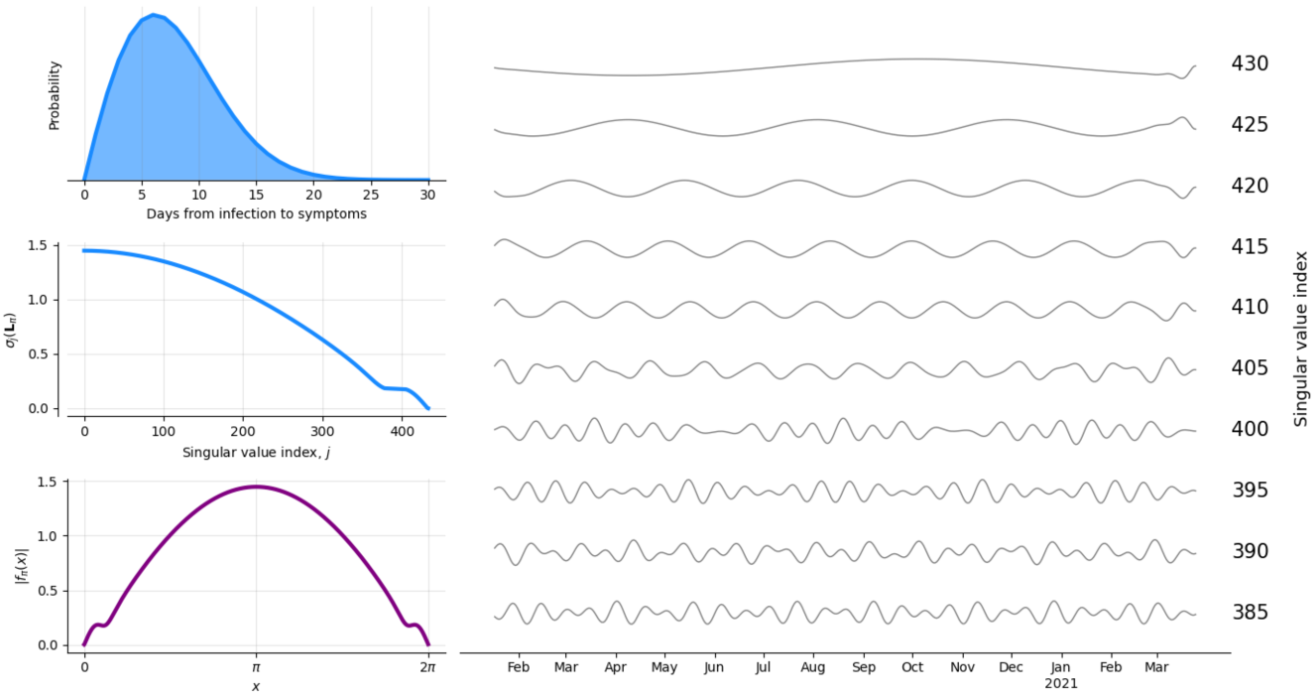

FIGURE 12. The mixture sensitivity pathogenesis distribution $\pi(\tau)$ with $\mu = 1$, $\alpha = 1.97$, and $\lambda = .114$ (top left), the singular value decomposition of $\mathbf{L}_\pi$ (middle left), the generating function $|f_\pi(x)|$ (bottom left), and some modes of $\mathbf{L}_\pi$ (right).

405. There is also now local extrema in the graph of $|f_\pi(x)|$, as well as "noisy" modes at and below singular value index 405.

Similarly, if we instead keep μ and λ as they were originally ($\mu = 1$ and $\lambda = .11$) and slightly decrease α from 1.97 to 1.88, we get the plots shown in Figure 13. This again looks very similar to our original pathogenesis distribution and would suggest a pathogenesis effective rank of around 405.

So, by very slightly modifying the starting pathogenesis distribution (in two different ways), we get two new similar distributions with pathogenesis effective rank ~ 405. It's possible that this alteration/approximation method could be formalized so that the pathogenesis effective rank could be defined even in cases where "elbows" and local extrema are not present. In this case, we would have a pathogenesis effective rank of ~ 405 instead of the Roy-Vetterli effective rank of 371, meaning we could cut out even more of the noise in Figure 2 by projecting onto a smaller effective nullspace.

4.6.2. *Unexplained "smooth" modes with low singular value index.* Another potential drawback to the use of the pathogenesis effective rank is that there are occasionally "smooth" modes with really low singular value index. Since we know that every "reasonable" time series is necessarily in the effective nullspace of $\mathbf{L}_\pi$ (which has a basis of modes with high singular value index), this is not a huge issue. However, based on the intuition we laid out in Subsection 4.1, a "smooth" mode with low singular value index would be indicative of

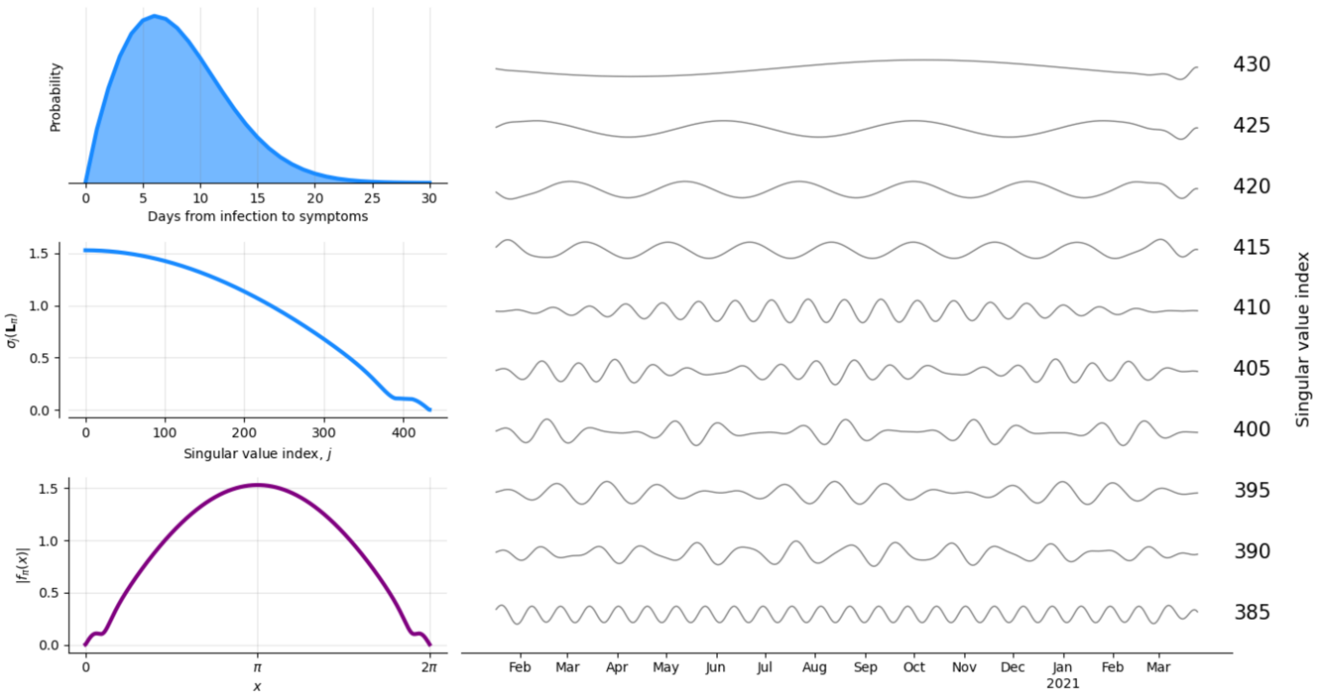

FIGURE 13. The mixture sensitivity pathogenesis distribution $\pi(\tau)$ with $\mu = 1$, $\alpha = 1.88$, and $\lambda = .11$ (top left), the singular value decomposition of $\mathbf{L}_\pi$ (middle left), the generating function $|f_\pi(x)|$ (bottom left), and some modes of $\mathbf{L}_\pi$ (right).

a "signal" that would be cut out by the projection onto the effective nullspace. This goes against the intuition that the effective rank should delineate between "noisy" and "smooth" modes. To better justify this intuition (which motivated our definition of pathogenesis effective rank), we should hope to better understand these unexplained "smooth" modes.

5. Future Work

5.1. **Proving Conjecture 1.** The first priority for future work related to pathogenesis signal processing should be to prove Conjecture 1, and in doing so, verify that the pathogenesis effective rank is the correct notion of effective rank for the purposes of pathogenesis signal processing. Through this process, we would also like to better understand – on an intuitive level – *why* the singular value distribution of $\mathbf{L}_\pi$ typically has an "elbow," and what is different about the "edge cases" in which there is no "elbow."

5.2. **Resolving the potential drawbacks of the pathogenesis effective rank.** With an understanding of *why* the "elbows" in the singular value distributions of the $\mathbf{L}_\pi$ operators are (or are not) occurring, we would like to understand how to best define effective rank in the event that there is no "elbow." One approach could be to formalize the estimation-by-slight-modification process outlined in Subsection 4.6. Otherwise, we could default to the Roy-Vetterli effective rank or come up with some other definition for these cases.

As mentioned in Subsection 4.6, we would also like to better understand the unexplained "smooth" modes occurring occasionally with low singular value index.

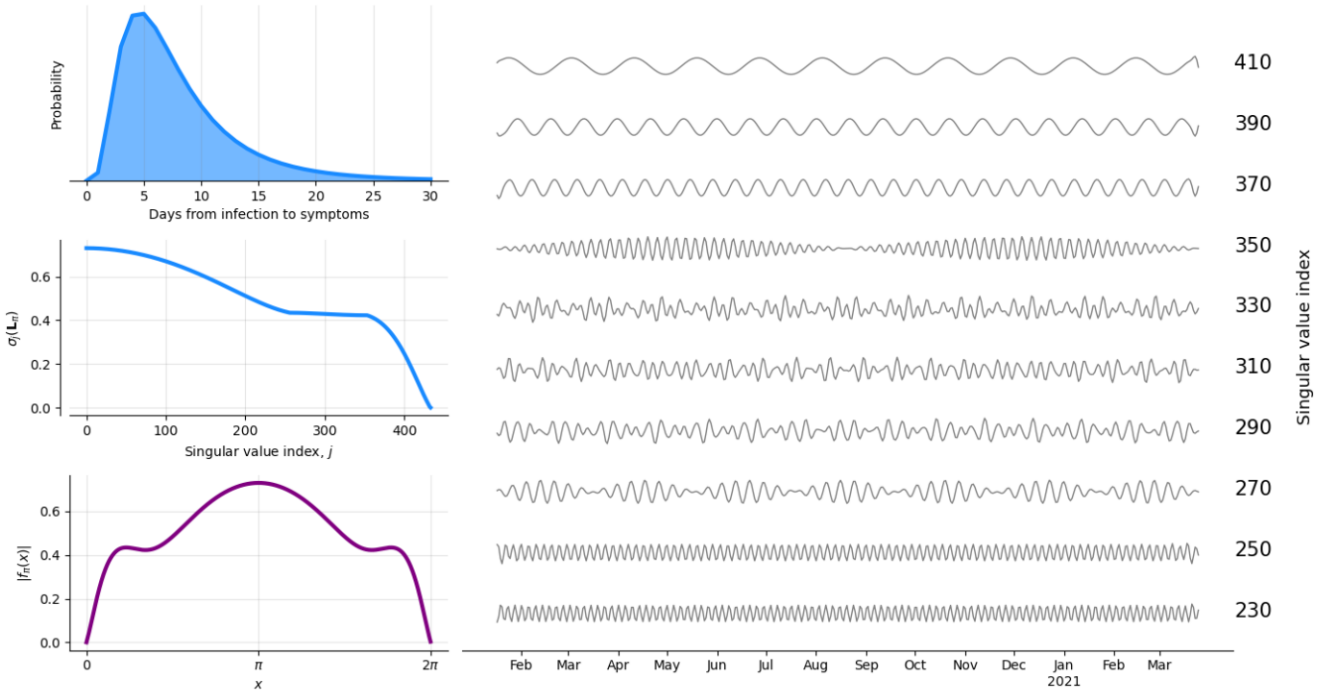

FIGURE 14. The log-normal pathogenesis distribution $\pi(\tau)$ with mean $\mu = 8.06$ and variance $\sigma^2 = 30$ (top left), the singular value decomposition of $\mathbf{L}_\pi$ (middle left), the generating function $|f_\pi(x)|$ (botton left), and some modes of $\mathbf{L}_\pi$ (right).

5.3. **Generalizing Conjecture 1.** Recall from Figures 8 and 9 that there can sometimes be an additional sharp point in the graph of the singular value distribution of $\mathbf{L}_\pi$ (besides the "elbow"). These additional sharp points seem to correspond to the local extrema of $|f_\pi(x)|$ in the same way that the "elbows" do. Additionally, there often seems to be a change in behavior in the modes of $\mathbf{L}_\pi$ at the singular value index corresponding to the sharp point. For example, in Figure 14, there is a sharp point around singular value index 250, at which point the modes switch from "smooth" (albeit of high frequency) to "noisy"; there also appears to be an "elbow" around singular value index 370, at which point the modes become "smooth" again.

To frame this all more precisely, we have the following conjecture which generalizes Conjecture 1:

Conjecture 2. *Let π be a pathogenesis distribution with corresponding Toeplitz matrix $\mathbf{L}_\pi$. Let j be a singular value index at which a "sharp" point in the graph of the singular value distribution of $\mathbf{L}_\pi$ occurs. Then we have the following:*

(i) $\sigma_j(\mathbf{L}_\pi)$ equals the y-value of some local extremum of $|f_\pi|$;
(ii) at singular value index j, there is a change in behavior in modes of $\mathbf{L}_\pi$.

In other words, it appears that "elbows" are just specific instances of sharp points in the graph of the singular value distribution of $\mathbf{L}_\pi$. Perhaps there is more we can say about the relationship between the singular value distribution of $\mathbf{L}_\pi$, the local extrema of $|f_\pi(x)|$, and the behavior of the modes of $\mathbf{L}_\pi$ (beyond the relevance to questions about effective rank).

5.4. **Apply pathogenesis signal processing to multipathogen analysis.** Finally, in future work we would like to apply pathogenesis signal processing to situations where we have data for multiple pathogens collected from the same locations, times, and/or people. In these situations, we would expect there to be commonality in the noise observed across the data for different pathogens. Supposing we had pathogenesis distributions $\pi_1, \ldots, \pi_r$ for the different pathogens along with raw epi-curves $\varphi_1, \ldots, \varphi_r$, this means we would expect the projections of $\varphi_1, \ldots, \varphi_r$ onto the effective ranges of $\mathbf{L}_{\pi_1}, \ldots, \mathbf{L}_{\pi_r}$ to be similarly structured. This could be a fascinating way to study the common causes of noise, and would allow us to identify and remove noise from our models for these pathogens with extremely high confidence.

References

[1] Elbow method (clustering). https://en.wikipedia.org/wiki/Elbow_method_(clustering). Accessed: 2024-03-20.

[2] Scree plot. https://en.wikipedia.org/wiki/Scree_plot. Accessed: 2024-03-20.

[3] G. Barbarino, S.-E. Ekström, S. Serra-Capizzano, and P. Vassalos. Theoretical results for eigenvalues, singular values, and eigenvectors of (flipped) Toeplitz matrices and related computational proposals. https://arxiv.org/pdf/2203.06992.pdf, 2022.

[4] J. Qin, C. You, Q. Lin, T. Hu, S. Yu, and X.-H. Zhou. Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. *Science Advances*, 6(33):eabc1202, 2020.

[5] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. *2007 15th European signal processing conference (IEEE, 2007)*, pages 606–610.

[6] N. Thakkar, R. Burstein, and M. Famulare. Towards robust, real-time, high-resolution COVID-19 prevalence and incidence estimation. *Institute for Disease Modeling*, 2020.

[7] N. Thakkar and M. Famulare. COVID-19 epidemiology as emergent behavior on a dynamic transmission forest. https://arxiv.org/pdf/2205.02150.pdf, 2022.