Characterizing Covid Waves via Spatio-Temporal Decomposition

Kevin Quinn
quinnk@bu.edu
Boston University
Boston, Massachussetts, US

Evimaria Terzi
evimaria@bu.edu
Boston University
Boston, Massachussetts, US

Mark Crovella
crovella@bu.edu
Boston University
Boston, Massachussetts, US

ABSTRACT

In this paper we develop a framework for analyzing patterns of a disease or pandemic such as Covid. Given a dataset which records information about the spread of a disease over a set of locations, we consider the problem of identifying both the disease's intrinsic *waves* (temporal patterns) and their respective spatial *epicenters*. To do so we introduce a new method of spatio-temporal decomposition which we call *diffusion NMF* (D-NMF). Building upon classic matrix factorization methods, D-NMF takes into consideration a spatial structuring of locations (features) in the data and supports the idea that locations which are spatially close are more likely to experience the same set of waves. To illustrate the use of D-NMF, we analyze Covid case data at various spatial granularities. Our results demonstrate that D-NMF is very useful in separating the waves of an epidemic and identifying a few centers for each wave.

CCS CONCEPTS

• **Computing methodologies** → **Non-negative matrix factorization**; • **Applied computing** → **Bioinformatics**.

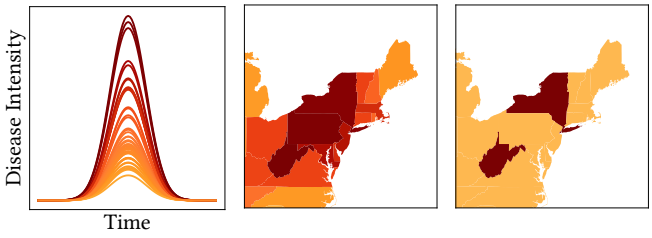
KEYWORDS

COVID-19, NMF, Non-Negative Matrix Factorization, Spatiotemporal Decomposition, Diffusion, D-NMF

ACM Reference Format:
Kevin Quinn, Evimaria Terzi, and Mark Crovella. 2022. Characterizing Covid Waves via Spatio-Temporal Decomposition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539136

1 INTRODUCTION

This paper is concerned with the analysis of data which is observed over time and across spatially related locations (i.e. spatiotemporal data). The questions that our methods address are the following: *(1) What are the waves (temporal patterns) that are intrinsic to the data?* and *(2) What are the spatial epicenters of each wave? Where did the waves have begin or have the most impact?*



(a) Waves of Spread **(b) NMF** **(c) D-NMF**

Figure 1: .
(1a) Depicts a single wave of disease with varying levels of intensity across locations; (1b) Shows a diffused pattern of spread recovered by NMF; (1c) Highlights two distinct points of origin recovered by D-NMF

Our study is motivated by Covid case data collected from the past (roughly) two year period over a variety of spatial scales. Our goal is to develop and deploy techniques that are able to identify waves of a pandemic (as temporal patterns) along with their respective spatial epicenter(s).

One may think to attack the two questions raised above via the techniques of classical Non-Negative Matrix Factorization (NMF) [12] or one of its variants. This approach may find key temporal patterns, but does not take into consideration their *spatial structure*: one wave which is present in location j is more likely to also occur in locations neighboring j than in locations that are further away from j. There may be locations in the output which appear to be strong sources of a particular wave, but are in fact only reflecting a pattern that comes from an epicenter that is spatially close to them.

The challenge of teasing apart the locational aspect of Covid waves is illustrated in Figure 1. The figure shows a hypothetical scenario of a wave which occurs in the northeastern US region. Figure 1a shows the temporal evolution of the wave in different states; some states show a stronger intensity overall than do other states. If we look at the heights of the peaks plotted as a geographical heatmap, we get Figure 1b. From this figure, it is not clear what the epicenter of the wave might be: it could be any combination of states from the set containing New York, Pennsylvania, and West Virginia. However, in this synthetic example, the epicenters of this wave were pre-defined as New York and West Virginia. The methods we develop in this paper are designed to identify those two states as epicenters. In fact, Figure 1c shows the result of applying our methods to the data in Figure 1a, demonstrating its ability to extract the correct epicenters in this example.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539136

To address this challenge, we define and develop *diffusion NMF* (D-NMF). D-NMF decomposes our data, D, into *three* matrices X, V, and K. Similar to an output from NMF, the matrices X and V are low-rank, non-negative matrices which respectively describe the waves (temporal patterns) and their spatial coefficients. The matrix K is the *diffusion kernel*: a square matrix provided as part of the input which encodes our assumptions about the the similarity in intensity between locations. Its purpose is to factor this similarity out of the matrix of spatial coefficients, V. Objectively, our goal is to find X and V such that $\|D - XVK\|_F$ is minimized. With the inclusion of K the columns of X will still encode wave patterns, but the columns of V will now sparsely encode the location-wise intensity of a wave so as to highlight its *epicenter(s)*.

To the best of our knowledge D-NMF is an idea that has not yet been proposed for analyzing disease or pandemic data. We show how one can construct an iterative algorithm for computing the D-NMF decomposition and include extensive experiments with Covid data collected from different locations over an observation period of 630 days. Our experiments show that D-NMF is able to identify interesting waves of the epidemic and, more importantly, distinct epicenters for them.

2 RELATED WORK

To the best of our knowledge, we are the first to propose *diffusion NMF*, the problem of capturing the spatiotemporal characteristics of a phenomenon such as the Covid pandemic. However, from a methodological point of view, our work is related to other works that focus on identification of sources of diffusion, signal deconvolution, and matrix decomposition. Since our application of focus is to study the spread of Covid, our work is also related to other's that seek to identify the pandemic's patterns. Below, we discuss the connections between our work and existing work in these areas.

Identifying patterns in the spread of Covid: The Covid-19 pandemic that started in Dec. 2019 has given rise to many studies which attempt to identify patterns of viral spread using a variety of techniques [1, 2, 7, 8, 10, 11, 16, 20, 26]. Pinpointing key factors that have affected the evolution of the Covid has become an important problem that is pushed forward by works such as these. Some of them focus on predicting the presence of Covid using NMF, timeseries analysis, or other machine learning methods [1, 8, 10, 11, 16, 20, 26], while others focus on understanding spread patterns through the use of mobility data [2, 7, 10]. We do not focus on predicting where or when the Covid will spread, but rather on providing an a-posteriori insight of the spatio-temporal characteristics seen in the data which is available to us. Related to ours is the work of Chen and Zhang [8] who show how NMF can be used to both uncover waves of spread and cluster locations into disjoint groups that have seen similar patterns. Their methods, however, do not make use of spatial similarity within the factorization process itself and are not focused on identifying small sets of epicenters.

Tracing the history of a spread: On a very high level, our work is related to social network problems where the goal is to identify both the backbone of a network and the most probable set initiators for some observed spread [13, 17, 21, 22]. The connection between this work and ours, however, is mostly high-level: in both cases observations are used to decouple the initiators and the observed spread. But, the approaches these works consider are tailored towards particular information propagation models on graphs. They do not consider the same variability in wave patterns which is picked out by matrix factorization with the use of a diffusion kernel.

Signal deconvolution and matrix decompositions: Similarly, our work is also abstractly related to work on image deconvolution [23], where the goal is to recover an image from a set of blurred images in the presence of a known or unknown point spread function. In a way, our diffusion kernel corresponds to the point spread function, which we assume is known. Our problem of recovering waves of spread and their corresponding epicenters is a very specific way of decomposing the input data which fits our purposes, but is not necessarily appropriate for image data. Therefore, our specific methods of signal separation are novel.

The D-NMF problem separates signal by decomposing the input data matrix D into a product of three matrices XVK, where X and V are low-rank non-negative matrices and K is an a-priori known matrix (the diffusion kernel). Since our goal is to find X and V given D and K, our problem is very much inspired by classical non-negative matrix factorization methods [12, 18, 19]. However, the distinguishing factor of our work is that the kernel K forces the non-negative factors X and V to take a particular form, and allows us to translate the columns of X as the waves and the rows of V as information related to their epicenters. This is unique to our problem, although similar approaches have been considered in other application domains [5, 6, 14, 24, 27]. These approaches all focus on finding a non-negative matrix factorization of the input data subject to some application-specific constraints. Our constraints are distinct in that they are specifically designed to to analyze the spread of an endemic or a pandemic such as Covid. Therefore, the similarity between our work and existing work in this area can only be seen at a high level.

3 SPATIO-TEMPORAL DATA DECOMPOSITION

Throughout the paper, we will assume that the input data consists of a matrix $D \in \mathbb{R}^{n \times m}$ where the rows correspond to timestamps and columns correspond to locations. Each entry $D(i, j)$ corresponds to some measure of Covid intensity at time i for location j, for the n timestamps and m locations.

Our analyses in this paper focus on patterns within cumulative Covid case data, but our methods are general and could be applied to any phenomenon that occurs in spatially distinct waves over time. Hence, in describing our methods, we will generally refer to *intensity* as the cumulative incidence of Covid cases. In other words, as a measure of a location's cumulative case count relative to its population – for example, each entry $D(i, j)$ corresponds to the cumulative number of new cases seen by location j from time 0 to time i, divided by j's population. In practice intensity could also mean infection rates, hospitalization rates, or other measures of spread relative to population.

Our goal is to discover *waves* of disease intensity, i.e., general patterns of rising and falling intensity over time that are common throughout locations in the data. At the same time, we want to

identify locations that act as *centers* of the different waves, i.e., places where a particular wave might have begun or where it had the most dramatic impact.

Although classical Non-negative Matrix Factorization (NMF) might be used to identify waves (basis vectors), NMF does not impose any geographical structure on the detected waves. In particular, it does not enable us to identify the epicenters of the different waves. For that, we introduce a new decomposition that enables us to do so, which we call *diffusion NMF* (D-NMF).

3.1 Modeling disease patterns using NMF

The idea behind NMF [19] is to describe the input data as a weighted sum of a small number of additive latent patterns. In other words, NMF assumes that the input data D can be approximated by two low-rank factors W and H such that

$$\|D - WH\|_F \tag{1}$$

is minimized. In the above equation, matrices W and H are constrained to be non-negative (i.e., all their entries are greater or equal to zero). Furthermore, the decomposition is generally low rank — W is of size $n \times r$ and H is of size $r \times m$, with $r \ll \min(n, m)$. We refer to the problem of finding W and H given r and D as the NMF problem.

In our setting, the columns of W correspond to the latent patterns that can be used to describe the data; these are the *waves*. Each of the r columns describes how its wave changes over n units of time. Similarly, each of the r rows of H is a length m vector which stores non-negative coefficients that encode the strength of the appearance of the rth wave at each of the m locations. In other words, NMF provides a low-rank temporal representation of intensity.

3.2 Modeling disease patterns using D-NMF

Although using NMF enables us to extract patterns of the temporal evolution, the NMF decomposition as described above does not provide insight about *spatial* patterns of intensity. In particular, because NMF determines the weights of H independently for each location, it does not assist in determining where each wave was initiated, i.e., what was the epicenter of the wave, and how the waves diffused across the different locations.

In order to enforce geographical constraints upon our model, we introduce a *kernel function* that captures similarity between different geographic locations. This kernel function can be any measure that captures how similar one location is expected to be to another location in terms of disease intensity. In what follows, we assume that the kernel function is generated by a diffusion process (described below), but our algorithms work with whatever kernel function is provided.

Hence, we introduce the $m \times m$ matrix K, with $K(j, j')$ giving the value of the kernel function relating location j to location j'. By incorporating this matrix in our decomposition, we define *diffusion NMF* (D-NMF). In D-NMF, the initial observation matrix D is decomposed into three matrices X, V and K such that:

$$\|D - XVK\|_F \tag{2}$$

is minimized.

In the above equation, X and V are analogous to W and H from NMF, i.e., are rank r non-negative matrices of shape $n \times r$ and $r \times m$ respectively. The additional non-negative $m \times m$ matrix K is a (diffusion) kernel that represents the expected similarity in intensity from one location to another. We call the problem of finding X and V, given the number of waves r, the diffusion matrix K and the data matrix D the D-NMF problem.

The diffusion kernel: The above problem definition is general and can, in principle, work with any non-negative matrix K. However, to effectively identify the epicenters of waves, it is necessary that the chosen kernel captures the some spatial aspect of the disease.

In this paper, we choose the simplest possible model of disease spread to demonstrate our results, without attempting to precisely model the epidemic properties of Covid. For that purpose, we treat disease spread as a diffusion or random-walk process on the adjacency graph of the studied locations. For this, we use a diffusion kernel [15]. More specifically, we adopt the Regularized Laplacian kernel [3, 28] defined as :

$$K = (I + \beta L)^{-1}, \tag{3}$$

where L is the Laplacian of the unweighted, undirected graph defined over the m studied locations. For our graphs, we connect two locations by an edge if they are geographically adjacent; all other locations are disconnected. Note that with this kernel any Laplacian matrix corresponding to an appropriate graph may be used, but we have chosen to restrict ourselves to the graph defined by geographical adjacency for simplicity.

The Regularized Laplacian (RL) as defined in Equation (3) is a symmetric, positive definite, and nonnegative matrix. Hence it meets the requirements for use in D-NMF. Furthermore, it is appropriate for modeling an infectious disease because it has the natural interpretation of a length-bounded random walk, as described in [3]. The random walk associated with the RL is defined as follows: consider a continuous-time random walk on a graph in which node v has degree d_v, and $a_{jv} = 1$ if j and v are adjacent, and is 0 otherwise. At time 0 the walker starts in a given node (say v) and remains in v for an exponentially distributed time with expected duration $1/d_v$. It then moves to a new node ℓ with probability $a_{v\ell}/d_v$, and repeats the process. The kernel function $K(v, q)$ is then proportional to the probability that a random walk that starts in node v is found in node q after an exponentially distributed time with expected duration β.

This parameter β found in the RL describes the 'spread' of the resulting diffusion. We call this parameter the *diffusion parameter*. If β is large, the walker moves far from the source before it is observed; hence, a large β value describes a widely-spread similarity function, i.e., a diffusion that has spread far from its source, covering many locations along the way. It is clear that more highly parameterized models, such as those that posit a different β for each wave, are possible and could be used in place of the RL, but we leave this for future work.

4 ALGORITHMS FOR D-NMF

We start this section by describing an iterative algorithm for the NMF problem as described by Lee and Seung [18]. Then, we discuss the relationship between the D-NMF problem and the original NMF

problem and show how to modify the original algorithm for our purposes.

An iterative algorithm for NMF As a starting point for solving the D-NMF problem, we will first discuss a popular iterative algorithm for the NMF problem which was first proposed by Lee and Seung [18]. Given the number of waves r, this algorithm finds non-negative low-rank matrices W and H that minimize Equation (1). In each iteration the algorithm updates the estimates it has for W and H following the update rules:

$$W \leftarrow W \cdot \frac{DH^T}{WHH^T}, \tag{4}$$

and

$$H \leftarrow H \cdot \frac{W^T D}{W^T WH}. \tag{5}$$

Note that in the equations above $\cdot$ represents element-wise multiplication. This iterative algorithm, which we call `iNMF`, iteratively updates the matrices W, H until convergence. In their paper, Lee and Seung show that this algorithm converges to a local minimum and also demonstrate that the algorithm works very well in practice, i.e., the number of iterations it requires is polynomial in the input size.

An iterative algorithm for D-NMF: Recall that in the D-NMF problem the input consists of the data matrix D, the number of waves r, and the diffusion matrix K (computed given the graph of locations and the diffusion parameter β). The goal is to find rank-r matrices X and V such that Equation (2) is minimized. To solve this problem, we will modify the `iNMF` algorithm and call this modification `diffusion-iNMF`. Still being an iterative algorithm, every iteration updates the current estimates of matrices X and V. Following the ideas of Lee and Seung, we modify the update rules for X and V so that they include the diffusion kernel matrix K. The new update rules are as follows:

$$X \leftarrow X \cdot \frac{DK^T V^T}{XVKK^T V^T}, \tag{6}$$

and

$$V \leftarrow V \cdot \frac{X^T DK^T}{X^T XVKK^T}. \tag{7}$$

The derivation of these rules follow the ones done by Lee and Seung which are described nicely in [4]. They will be provided in an extended version of the paper.

Discussion: Since we assume that the diffusion matrix K is known, it's tempting to think the D-NMF problem is equivalent to the problem of approximating the matrix DK^{-1} with rank-r non-negative matrices X and V i.e. solving an NMF problem for data matrix DK^{-1} (instead of D). However, such an approach would be problematic. While K has positive values for each entry, the same is not necessarily true for K^{-1} and therefore also DK^{-1}. This poses a problem for NMF which always assumes non-negative input. Therefore, the optimal solution to the problem of minimizing $\|D - XVK\|_F$ is different from the optimal solution to the problem of minimizing $\|DK^{-1} - XV\|_F$ (subject to X and V being non-negative in each case).

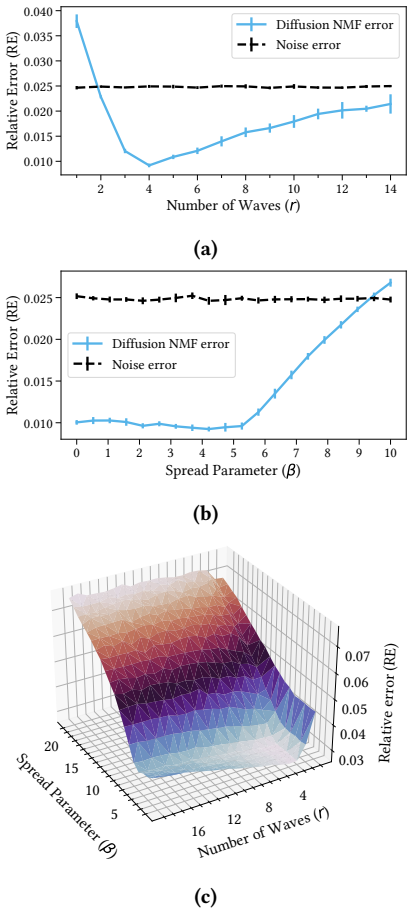


Figure 2: Experiments on synthetic data with true values for $r = 4$ and $\beta = 5$; (2a) Grid-search to determine number of waves (r) used as input; (2b) Grid-search to determine extent of diffusion parameter (β) for RL kernel K; (2c) Grid search jointly on (r, β).

5 EXPERIMENTS

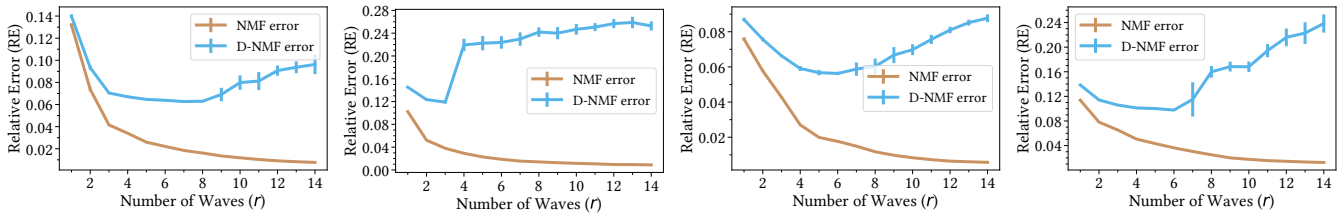
In this section, we experimentally evaluate our approach using both synthetic and real datasets. The code, data, and examples with more detail are all available on github[1].

5.1 Experiments with synthetic data

Before demonstrating the efficacy of our methodology on real datasets we first evaluate it on synthetic data. In doing so, we demonstrate that our grid-search methodology allows us to adequately recover accurate values for r (i.e., number of waves or rank) and β for the RL Kernel (see Equation (3)).

Synthetic datasets: To generate the observation matrix D of the synthetic data we proceed as follows: first, we generate an $n \times r$ matrix for X where each column follows the pattern of a sine wave with a fixed frequency and amplitude. Then we generate a random

[1]https://github.com/KevinQ152/diffusion_nmf



(a) New England Counties (b) California Counties (c) US States (d) European Countries

Figure 3: Relative Error (RE) of NMF and D-NMF for $D_{\mathbf{NE}}$, $D_{\mathbf{CA}}$, $D_{\mathbf{US}}$ and $D_{\mathbf{EU}}$.

sparse $r \times m$ matrix for V in which only 7% of the entries are non-zero. The nonzero entries are positive values that represent points of origin in the diffusion process. Our chosen level of sparseness simulates the case in which a few locations are responsible for the majority of the spread. Finally, we construct a random geometric graph [25] – using m nodes and a distance threshold of 0.3 – and compute its Laplacian, L. Given this L and a chosen value of β, we then form matrix K as per Equation (3). We construct the data, D, per the formula $D = XVK$ and, to simulate random noise, we add a sample from a Gaussian distribution with zero mean and a standard deviation of 0.001 to every element $D_{i,j}$. Note that if the addition of noise resulted in a negative value we set this value equal to 0.

In this report we used $n = 250$, $m = 50$, but our experiments suggest that the size of the dataset does not have much impact on the ability of grid search to pick the right parameters r and β.

Grid search: Given a synthetically-generated matrix D, and for fixed hyperparameters (r, β), we use `iNMF` to estimate $\hat{X}$ and $\hat{V}$. Then, for these values of (r, β) we compute the *relative error* of our solution as follows:

$$\mathrm{RE}(\widehat{X}, \widehat{V} \mid D, r, \beta) = \frac{\|D - \widehat{X}\widehat{V}K\|}{\|D\|}. \tag{8}$$

Note that in order to avoid overfitting we randomly select 80% of the entries in D, use them to estimate $\hat{X}$ and $\hat{V}$, and compute the relative error on the unseen 20% of the data.

Our *grid-search* methodology for finding hyperparameters considers all combinations within a range of values for r and β, and chooses values that are minimal in error while also preferring smaller r and larger β. More details for hyperparameter selection and our train-test procedures are provided within our example notebooks[1].

Results: Figure 2 demonstrates the ability of the grid-search method described above to accurately recover the values of r and β that were used for generating the synthetic data. In this experiment we generated data using $r = 4$ and $\beta = 5$. The results in Figure 2a correspond to keeping $\beta = 5$ and running `diffusion-iNMF` for different values of r ranging from 1 to 14. The y-axis shows the relative error and the x-axis the different values of r tested. The results show that the relative error drops sharply when $r = 4$ and then continues to slightly increase as the rank rises. We consider this increase as evidence of overfitting.

In the plot we also show the relative error which is due to the added noise; it is important to observe that unless the parameters r or β we provide as part of the input are very different from the ones used for generating the data, the error of the solution output by `diffusion-iNMF` is smaller than the error due to noise. Given that the error caused by noise was very small on its own, this implies that our algorithm finds solutions that are very close to the optimal ones.

For the same dataset, Figure 2b shows the relative error (y-axis) achieved by running the `diffusion-iNMF` algorithm for fixed $r = 4$ and different values of $\beta \in (0, 10]$ (x-axis). The results show that for values of β in the interval $(0, 5]$ the error is very low; however as β increases past the true value of 5 the error immediately increases almost linearly. Also note that, as before, the relative error of our algorithm is less than the error that came from adding noise onto our input data. This suggests that our algorithm is cutting through the noise and picking up on the important trends in the data.

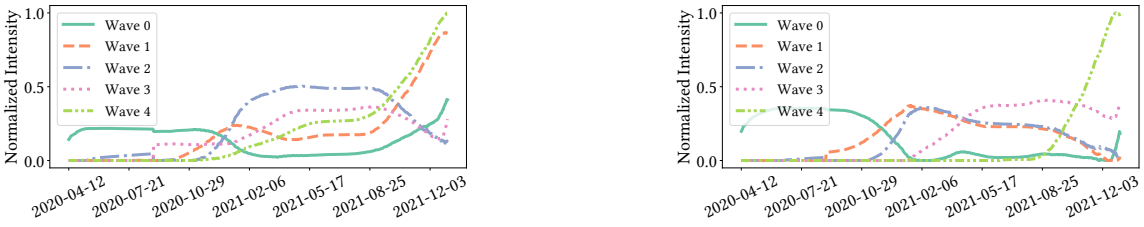
These results demonstrate that by using grid search for one of the hyperparameters (i.e., r or β) we can recover its true value. We also explore the ability of the grid-search methodology to recover the true values of r and β by searching jointly over (r, β) pairs. The result of this experiment is shown in Figure 2c. For this plot we use the same dataset with true values of (r, β) being $(4, 5)$. The relative error of `diffusion-iNMF` achieved for the different values of (r, β) is shown on the z-axis of this 3-d plot. Again, our results are consistent: we can recover the true values of (r, β) using our procedure.

5.2 Experiments with real Covid data

Having shown that `diffusion-iNMF` can accurately recover the wave and diffusion structure in synthetic data, we now use it to characterize wave and diffusion structures of Covid. Using cumulative covid case counts (i.e., infections) reported daily by Johns Hopkins [9], we demonstrate the ability of D-NMF to analyze data at a range of spatial scales: at the US county, US state, and country levels (focusing on European countries). At each level, our experiments demonstrate the ability of D-NMF to discover both distinct waves in the spread of Covid and their spatial epicenters. Whenever we use NMF and D-NMF for our analyses we recover the corresponding matrix decompositions using the `iNMF` and `diffusion-iNMF` algorithms respectively.

5.2.1 Datasets. For all of the following datasets, we collected daily data for the period starting April 12, 2020 and ending January 1, 2022. Therefore, all of our datasets consider 630 timestamps.

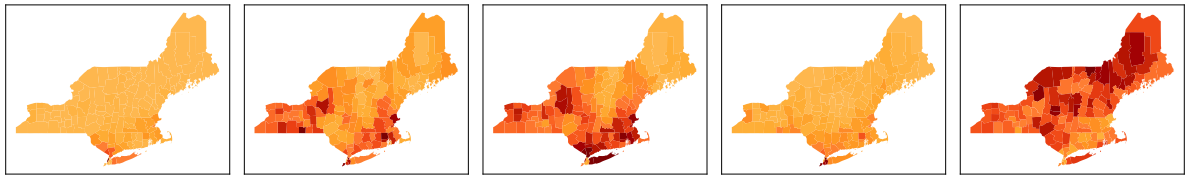
County-level datasets: Our county-level datasets are D_{NE} and D_{CA}. The first (D_{NE}) contains as locations 127 counties in the New England Region (including New York, Massachusetts, Connecticut,



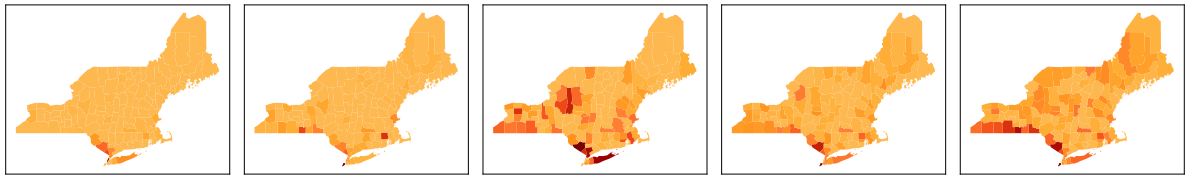
(a) Wave matrix W recovered with NMF.

(b) Wave matrix X recovered with D-NMF.

Figure 4: County-level Covid waves for New England recovered with NMF and D-NMF.



(a) Each row (0-4) of recovered H matrix of NMF corresponding to waves (0-4) in W.



(b) Each row (0-4) of recovered V matrix of D-NMF corresponding to waves (0-4) in X.

Figure 5: County-level coefficient matrices H, V for New England as recovered with NMF and D-NMF; large coefficients (i.e., darker colors) correspond to locations with high intensity for a given wave.

Maine, Vermont, New Hampshire, and Rhode Island). The second dataset (D_{CA}) covers all 58 counties in California. In both, two counties are considered adjacent if they have a shared border, as determined by the US Census.[2] Every entry corresponds to a cumulative case count (since the beginning of the observation period) and the case counts are normalized by the corresponding 2010 census population of the county.[3]

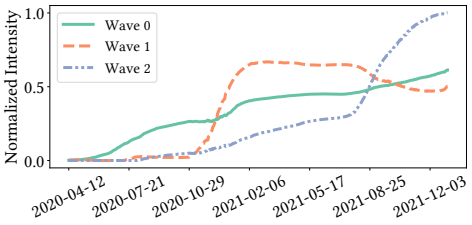
State-level dataset: Our state level dataset, D_{US}, considers the 48 contiguous US states that form a connected adjacency graph (excluding other locations not connected by a land border). Each entry corresponds to a normalized cumulative case count, where normalization is done with the state's 2010 US census population.[4] In this dataset an edge is added between two states if they share a border.[5]

European-countries dataset: Our European-country dataset D_{EU} has as locations 49 countries in and around the region of Europe. Cumulative case counts are normalized using the corresponding country's population as reported in by the World Bank.[6] In this dataset, there is an edge between two countries if they share a land border (roughly).[7]

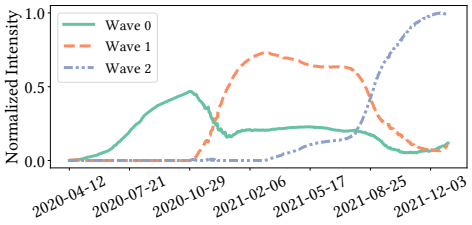
5.2.2 Error analyses for Covid data: We start our analysis of real data by comparing the relative error of the solutions to D-NMF and NMF (as produced by `diffusion-iNMF` and `iNMF` respectively) as a function of the number of waves r. The results for all four Covid datasets are shown in Figure 3. The error bars shown correspond to different repetitions of the algorithm when hiding different portions of the data for training and testing; as before, we use 80% for training and 20% for testing.

We observe that, in all cases, the error for D-NMF is larger than that of NMF. This is expected as D-NMF is a more constrained version of NMF. For all datasets we see that the error for NMF decreases steadily as r increases. This is because NMF is not restricted by a diffusion process and so can increase r freely. D-NMF, however, does tend to increase in error after finding a minimum point. Unlike NMF it is constructed around a diffusion process in which waves are spread and shared amongst locations in a pre-defined way. For this reason, it is not always beneficial to add new waves. The introduction of a new wave may decrease the error for one location but increase the error for others.

[2] https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html
[3] https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html
[4] https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html
[5] Adjacencies obtained from https://data.world/bryon/state-adjacency
[6] https://data.worldbank.org/indicator/SP.POP.TOTL
[7] Adjacencies obtained from https://github.com/geodatasource/country-borders.

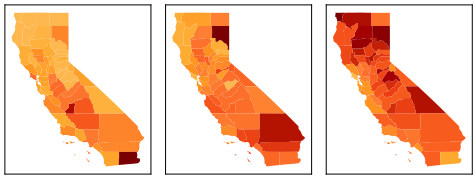


(a) Wave matrix W recovered with NMF.

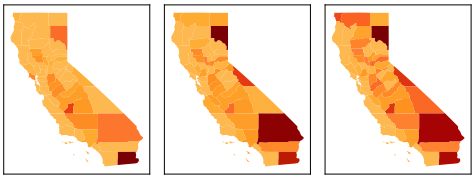


(b) Wave matrix X recovered with D-NMF.

Figure 6: County-level Covid waves for CA recovered with NMF and D-NMF.



(a) Each row (0-2) of recovered H matrix of NMF corresponding to waves (0-2) in W.



(b) Each row (0-2) of recovered V matrix of D-NMF corresponding to waves (0-2) in X.

Figure 7: County-level coefficient matrices H, V for CA as recovered with NMF and D-NMF; large coefficients (i.e., darker colors) correspond to locations with high intensity for a given wave.

5.2.3 Qualitative analyses for county-level Covid data. Figures 4 and 5 summarize the results of our analysis for D_{NE} using D-NMF; the results for NMF are also shown for comparison. For these plots we use $(r, \beta) = (5, 1)$, which are found using the grid-search procedure described in Section 5.1. Figure 4a shows the waves as recovered by NMF and Figure 4b shows the corresponding waves as recovered by D-NMF (both re-scaled to fit exactly in the range of [0,1]). The spatial maps in Figure 5 show the corresponding coefficients of every NMF wave (Figure 5a) and every D-NMF wave (Figure 5b). In the map, darker colors correspond to larger coefficients and therefore are locations with high intensity for a given wave. Note that there are five waves (labeled 0 to 4) and the labeling is consistent across figures.

In Figure 4 we observe that there are clear waves corresponding to spring/summer of 2020 (wave 0), fall/winter of 2020-21 (waves 1-2), mid 2021 (wave 3), and late 2021 (wave 4). The influence of the Omicron variant is clear in the late 2021 wave. Moreover, the New England example shows that D-NMF isolates waves more distinctly than NMF – in NMF wave "peaks" are harder to distinctly determine.

The spatial maps of NMF (Figure 5a) show how the NMF-defined waves are spread very broadly across the region. However the spatial maps of D-NMF (Figure 5b) allow us to see that there are clearer epicenters for each wave. For example, wave 0 (spring/summer 2020) is centred in Manhattan, wave 1 had epicenters in western Massachusetts and the southern tier of New York State, and wave 2 had epicenters in Long Island and central New York. In the corresponding NMF waves, however, the spatial maps show much more dispersion and do not clearly identify a epicenter for each wave.

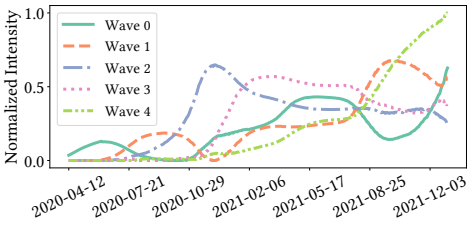
Figures 6 and 7 summarize the results of our analysis for D_{CA} and they have the same semantics as the figures for New England. The only difference being that our grid search identified $(r, \beta) = (3, 1)$ meaning that only three waves were identified in California during this time period.

The results for CA are similar to those of New England. Figure 6 shows that D-NMF separates the three waves in the data more clearly than NMF. Furthermore the spatial maps for NMF (Figure 7a) show that the NMF waves are more widespread across the state compared to the corresponding D-NMF waves (Figure 7b); the latter have significantly less dispersion and more clearly identify the epicenters associated with every wave. Using D-NMF, we can more clearly identify that the second and third waves each had a epicenter in Northern California as well as a epicenter in Southern California.

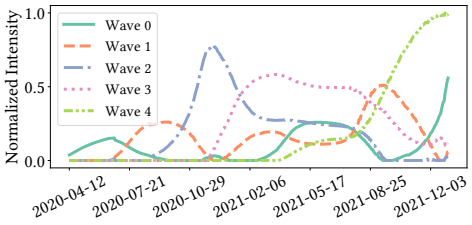
5.2.4 Qualitative analysis of state-level Covid data. Figures 8 and 9 summarize the results of our analysis for D_{US}. Figure 8 shows the waves discovered by NMF and D-NMF (both re-scaled to the range of [0,1]), and the spatial maps shown in Figure 9 show the value of the coefficient associated with each of the US States for every wave. After performing a grid search with input D_{US}, we used the values $(r, \beta) = (5, 1)$ meaning that five waves were identified at the national level in the US.

The main take-away is again that D-NMF isolates specific states as the epicenters of Covid waves more distinctly than NMF does. For example, D-NMF identifies Arizona as a epicenter for wave 0 more clearly than NMF. D-NMF also shows more clearly that Wave 1 is centered on New York, Florida, Louisiana, and Arizona/Nevada. Likewise, wave 2 is very sharply centered in North and South Dakota in Figure 9b, but is much more spread out among states in Figure 9a. Similar observations apply to Waves 3 and 4.

5.2.5 Qualitative analysis of European country-level Covid data. Finally, we conclude by turning to the European country-level dataset,

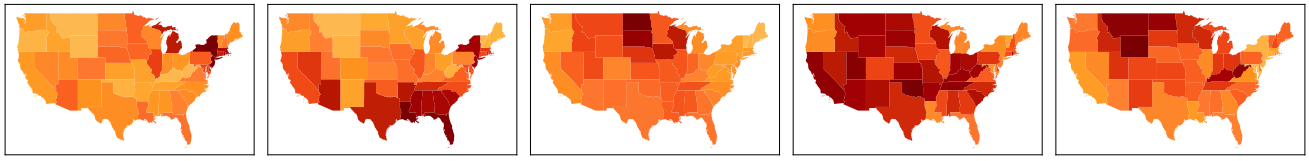


(a) Wave matrix W recovered with NMF.

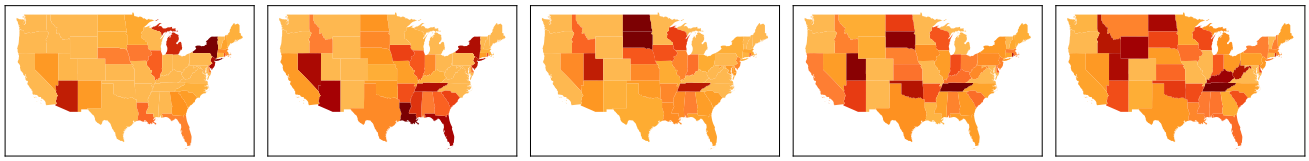


(b) Wave matrix X recovered with D-NMF.

Figure 8: USA state-level Covid waves recovered with NMF and D-NMF.



(a) Each row (0-4) of the recovered H matrix of NMF corresponding to waves (0-4) in W.



(b) Each row (0-4) of the recovered V matrix of D-NMF corresponding to waves (0-4) in X.

Figure 9: USA state-level coefficient matrices H, V recovered with NMF and D-NMF; large coefficients (i.e., darker colors) correspond to locations with high intensity for a given wave.

D_{EU}. Figures 10 and 11 show the waves and their corresponding spatial extent in different European countries. For this dataset, using grid search, we picked $(r, \beta) = (4, 0.5)$. Again, waves were re-scaled to the range of [0,1].

Here, as before, we find that D-NMF isolates specific countries as the epicenters of Covid waves in a sharper way than NMF. For example, D-NMF shows wave 0 as being highly concentrated in Montenegro, The Czech Republic, Sweden, and Spain, but these epicenters of intensity are harder to distinguish in the results from NMF. For other waves, central European countries are also more sharply associated with epicenters in D-NMF as compared to NMF. As before, this effect is also reflected in the appearance of the waves themselves. Wave 0 in NMF has two peaks, but the first of which that occurs in the fall of 2020 is small and unpronounced. In D-NMF, however, this peak becomes more distinguished and recognizable as a key event in the data.

6 CONCLUSIONS

In this paper we have shown how to decompose pandemic or disease related data in a way that simultaneously identifies temporal patterns (waves), as well as spatial locations – epicenters – that describe where each wave is located. To do so, we defined the *diffusion NMF* matrix decomposition problem, D-NMF, and developed an algorithm that attempts to solve it. We then applied D-NMF to the analysis of 630 days of Covid case data at three different spatial scales, and showed its advantages over classical NMF. Particularly, it shows an ability to identify spatial locations as epicenters for Covid waves.

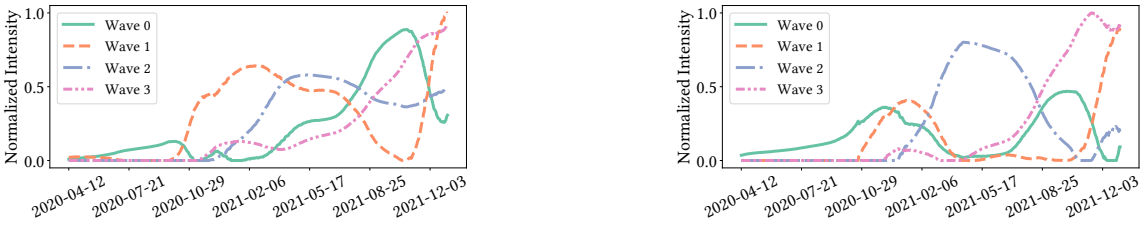
While this first step is encouraging, we believe that it suggests a number of intriguing directions for future work. For example, the form of the diffusion matrix K used affects the results that we find. For our analyses, we chose the simplest possible such matrix: the Regularized Laplacian Kernel computed using the geographical adjacency graph. However, we believe that moving forward it will be worthwhile to consider different diffusion kernels, different graph structures, or even more complex spread parameters.

ACKNOWLEDGMENTS

The authors are thankful for support from the National Science Foundation under grants IIS 1908510 and IIS 1813406, and from the Boston University UROP program.

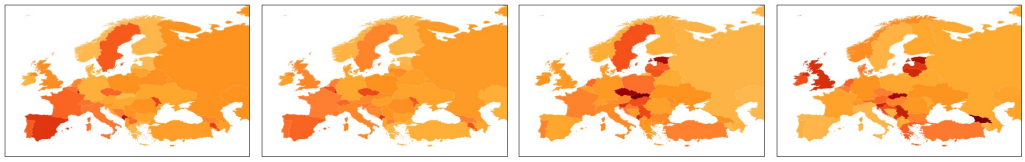
REFERENCES

[1] Madini O. Alassafi, Mutasem Jarrah, and Reem Alotaibi. 2022. Time series predicting of COVID-19 based on deep learning. *Neurocomputing* 468 (2022), 335–344.
[2] Laura Alessandretti. 2022. What human mobility data tell us about COVID-19 spread. *Nature Reviews Physics* 4, 1 (2022), 12–13.
[3] Konstantin Avrachenkov, Pavel Chebotarev, and Alexey Mishenin. 2015. Semi-supervised Learning with Regularized Laplacian. arXiv:1508.04906 [cs.LG]
[4] Juan JosÃĺ Burred. 2014. Detailed derivation of multiplicative update rules for NMF. (2014). https://www.jjburred.com/research/pdf/jjburred_nmf_updates.pdf
[5] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang. 2011. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 8 (2011), 1548–1560.



(a) Wave matrix W recovered with NMF **(b) Wave matrix X recovered with D-NMF.**

Figure 10: European country-level Covid waves recovered with NMF and D-NMF.



(a) Each row (0-3) of the recovered H matrix of NMF corresponding to waves (0-3) in W.

(b) Each row (0-3) of the recovered V matrix of D-NMF corresponding to waves (0-3) in X.

Figure 11: European country-level coefficient matrices H, V recovered with NMF and D-NMF; large coefficients (i.e., darker colors) correspond to locations with high intensity for a given wave.

[6] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. 2008. Non-negative Matrix Factorization on Manifold. In *IEEE International Conference on Data Mining (ICDM)*. 63–72.

[7] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589, 7840 (2021), 82–87.

[8] Jianmin Chen and Panpan Zhang. 2022. Clustering US states by time series of covid-19 new case counts with non-negative matrix factorization. *Journal of Data Science* (2022).

[9] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 20, 5 (2020), 533–534.

[10] Peter Edsberg MÃ¥llgaard, Sune Lehmann, and Laura Alessandretti. 2022. Understanding components of mobility during the COVID-19 pandemic. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380, 2214 (2022).

[11] E. Gecili, A. Ziady, and R.D. Szczesniak. 2021. Forecasting COVID-19 confirmed cases, deaths and recoveries:Revisiting established time series modeling through novel applications for the USA and Italy. *PLoS One* 16, 1 (2021).

[12] N. Gillis. 2014. *Regularization, Optimization, Kernels, and Support Vector Machines*. Chapman & Hall/CRC.

[13] Manuel Gomez-Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Netw. Sci.* 2, 1 (2014), 26–65.

[14] Patrik O. Hoyer. 2004. Non-negative matrix factorization with sparseness constraints. *CoRR* cs.LG/0408058 (2004). http://arxiv.org/abs/cs.LG/0408058

[15] Risi Kondor and John Lafferty. 2002. Diffusion Kernels on Graphs and Other Discrete Input Spaces. *ICML* Vol. 2 (05 2002).

[16] Naresh Kumar and Seba Susan. 2020. COVID-19 Pandemic Prediction using Time Series Forecasting Models. In *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 1–7.

[17] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. 2010. Finding effectors in social networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1059–1068.

[18] Daniel Lee and H. Sebastian Seung. 2001. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*. MIT Press.

[19] David Lee and Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[20] Hanbaek Lyu, Christopher Strohmeier, Georg Menz, and Deanna Needell. 2020. COVID-19 Time-series Prediction by Joint Dictionary Learning and Online NMF. *arXiv e-prints*, Article arXiv:2004.09112 (April 2020), arXiv:2004.09112 pages. arXiv:2004.09112 [cs.LG]

[21] Heikki Mannila and Evimaria Terzi. 2009. Finding Links and Initiators: A Graph-Reconstruction Problem. In *SIAM International Conference on Data Mining, SDM*. SIAM, 1209–1219.

[22] Michael Mathioudakis, Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Antti Ukkonen. 2011. Sparsification of influence networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chid Apté, Joydeep Ghosh, and Padhraic Smyth (Eds.). ACM, 529–537.

[23] Sophocles J. Orfanidis. 2010. *Introduction to Signal Processing*.

[24] Pentti Paatero, Unto Tapper, Pasi Aalto, and Markku Kumal. 1991. Matrix Factorization Methods for Analysing Diffusion Battery Data. *Journal of Aerosol Science* 22 (1991), S273–S276.

[25] Mathew Penrose. 2003. Random Geometric Graphs. *Oxford Studies in Probability* (2003).

[26] Chen Tang, Tiandong Wang, and Panpan Zhang. 2020. Functional data analysis: An application to COVID-19 data in the United States. *arXiv preprint arXiv:2009.08363* (2020).

[27] Wei Xu and Yihong Gong. 2004. Document Clustering by Concept Factorization. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, United Kingdom) *(SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 202âĂŞ209. https://doi.org/10.1145/1008992.1009029

[28] Denny Zhou and Bernhard Schölkopf. 2004. A Regularization Framework for Learning from Graph Data. In *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields* (icml workshop on statistical relational learning and its connections to other fields ed.). https://www.microsoft.com/en-us/research/publication/regularization-framework-learning-graph-data/