EJTAS
ISSN 2786-7447
EUROPEAN JOURNAL OF THEORETICAL AND APPLIED SCIENCES
OPEN ACCESS
PEER REVIEWED

Application of Weibull Distribution to Hidden Markov Model for Non-Negative Factorization Matrix

Edesiri Bridget Nkemnole
Department of Statistics, University of Lagos, Nigeria

Joshua Olaniyi Bamigbode
Department of Statistics, University of Lagos, Nigeria

Suggested Citation
Nkemnole, E.B. & Bamigbode, J.O. (2024). Application of Weibull Distribution to Hidden Markov Model for Non-Negative Factorization Matrix. *European Journal of Theoretical and Applied Sciences, 2*(1), 607-622.
DOI: 10.59324/ejtas.2024.2(1).53

Abstract:

Probabilistic Nonnegative Matrix Factorizations (NMFs) are very useful in statistics when dealing with stochastic signals such as wave fronts, share prices, and volatility as a Nonnegative Matrix Factorization (NMF) approach. Little attention has been made in the literature to developing NMF algorithms that use moving average to exploit data's temporal dependencies. A hidden Markov model (HMM) using a Weibull distribution as the output density function was created in this study. The Weibull HMM was then reformulated as a probabilistic NMF. This demonstrates the connection between the proposed HMM and NMF, and will lead to a novel probabilistic NMF approach in which the model captures temporal dependencies inherently utilizing moving average. Furthermore, the model parameters were estimated using maximum likelihood estimation (MLE). The model's adaptability was compared to the existing probabilistic NMFs models of gamma and lognormal. Our trials with US COVID-19 data revealed that the proposed technique achieves a superior balance of sparsity, the goodness of fit, and temporal modeling than gamma and lognormal models.

Keywords: *Hidden Markov Model (HMM), Maximum Likelihood Estimation (MLE), Moving Average, Nonnegative Matrix Factorization (NMF), Weibull Model.*

Introduction

Markov models are frequently used in probability theory and statistics to simulate the probabilities of distinct states and the rates of transitions between them. In general, the approach is used to model systems. Markov models can also be used to identify patterns, predict outcomes, and learn the statistics of sequential data. The Hidden Markov Model (HMM) was first introduced in 1957 (MacDonald and Zucchini 1997; Cappe et al. 2005; Nkemnole et al. 2013), with several applications in signal processing, medicine, engineering, and management. The HMM is a doubly stochastic process that has an underlying stochastic process that is not directly observable but can be observed via another process that generates a succession of independent random observations. HMM can be defined in the same way via a functional representation known as a state space model (Nkemnole et al. 2013).

HMM have been used to analyze random processes with temporal or spatial structure for over four decades and have been successfully applied to a wide range of applications including but not limited to speech recognition, musical scores, handwriting, and bio-informatics

(Lakshminarayanan and Raich 2013). In text document processing, for example, words display reliance on previous material, and the usage of HMM may result in more effective modeling and inference as compared to an independent word model.

In classification, an HMM can be extracted for each class during the training phase (e.g., using the maximum likelihood (ML) principle), and the Bayes risk minimizing maximum aposteriori classifier can be used to optimally determine which of several HMMs is most likely to generate a sequence of test observations. The well-known Baum-Welch technique (Rabiner 1989) is commonly utilized in performing HMM ML parameter estimation. The Baum-Welch algorithm is an expectation-maximization (EM) generalized implementation of the ML estimator. Baum-Welch inherits the property that generalized EM produces a succession of non-decreasing likelihood values converging to a local maximum of the likelihood. Fast convergence, on the other hand, is not assured. The computational complexity of each iteration of the Baum-Welch method is proportional to the length of the observation series. When dealing with extended sequences, the Baum-Welch technique may be prohibitively expensive if many iterations are necessary. We are interested in developing resolution to this challenge.

Non-Negative Matrix Factorization (NNMF) (Lee and Seung 2001) is a computationally efficient method for factoring a matrix into Non-Negative Factors that has been successfully applied in a variety of applications including text mining, spectral data analysis (Berry et al. 2007), face recognition (Lee and Seung 1999), and so on. There has been minimal previous research into the use of NNMF algorithms for HMM. Hsu et al. (2009) use a creative reparametrization of the conventional HMM problem (Rabiner 1989) to compute a global solution for their suggested parametrization based on singular value decomposition (SVD). The authors of Cybenko and Crespi (2008) use NNMF on higher-order transitions to compute HMM parameters. However, unlike the least squares criterion employed in their research, their NNMF formulation is based on divergence minimization. Furthermore, both papers assume that the observations are discrete. To the best of our knowledge, no previous work on the application of NNMF to HMMS with continuous observations has been published.

Probabilistic Non-negative Matrix Factorization (NMF) techniques are particularly useful when dealing with stochastic signals such as voice. Many scholars have become interested in non-negative matrix factorization (NMF) in recent years. As a result, various ways to obtaining NMF utilizing various criteria have been established (Cichocki 2009). In its most basic form, NMF finds a locally optimum and deterministic approximation of a non-negative matrix x as a product of two non-negative matrices.

To deal with non-negative data, we developed an HMM in which the output density functions are considered to be Weibull distributions. The Weibull distribution was chosen because of its shape characteristics, which allow for more flexible modelling than the exponential, normal, student-t, gamma, and generalized error distributions explored previously. Data scaling is not required when utilizing the Weibull distribution. The Weibull HMM would therefore be rewritten as a probabilistic NMF. Each HMM state is intended to result in a basis vector in the NMF representation, and the data's temporal correlation is enforced by the transition probability between the states. To account for the data's time-varying level, an explicit Weibull prior distribution is evaluated over the gain variable. Following that, the study presents an efficient EM approach for estimating the time-varying and stationary model parameters.

Literature review

Cichocki et al. (2009) establishes a locally optimum and deterministic approximation of a non-negative matrix in the form of a product of two non-negative matrices. Cemgil and Fevotte (2009), Hoffman et al (2010), and Mysore et al (2010) create a numerical technique to establish a probabilistic framework for HMM for a non-

negative matrix factorization. Hoffman et al. (2010) presented a method for performing non-negative matrix factorization (NMF), in which data is supposed to follow an exponential distribution with rate parameters factored using an NMF. Their model is built in such a way that the optimal number of NMF basis vectors can be determined automatically. Mysore et al. (2010) integrated NMF and HMM, assuming that each sample of data was a sum of certain complex-valued Gaussian components with covariance matrices factorized using NMF. More specifically, each component's movement between different states was governed individually by a separate Markov chain. Mysore et al. (2010) also suggested a non-negative HMM with a Markov chain with many chains in each state, and the observed data is supposed to follow a multinomial mixture model. Nonetheless, due to the multinomial distribution assumption, the observed data must be scaled to be integer. The issue here is that the integer scaling levels will directly assume noise level in the model, which may cause side effects. To deal with non-negative data, Mohammadiha et al. (2017) developed an HMM in which the output density functions are considered to be gamma distributions. The gamma distribution source allows for more flexible modeling than the exponential distribution examined by Hoffman. Furthermore, unlike Cemgil and Mysore, the approach does not require data scaling. However, if some of the data points are zeros, this method may fail.

This research work, on the other hand, proposes an HMM in which the output density function is assumed to be Weibull distribution, using moving average data. This method uses moving average rather than actual data points to smooth the curves and prevent zero data points. Furthermore, this method does not require that the data is normally distributed and can be used in measuring volatility in survival and reliability models.

Methods

Parameter Estimation

The HMM is an important statistical model in many fields including bioinformatics (Durbin et al., 1998), econometrics (Kim et al., 1998) and population genetics (Felsenstein & Churchill, 1996); see also Cappe et al. (2005) for a recent overview. Often, one has a range of HMMs parameterized by a parameter vector taking values in some compact subset , of the Euclidian space. The HMM is comprised of a latent state process $\{Xk\}_{k\geq 0}$ which is only indirectly observed through the observation process. Given a sequence of such observations $\hat{Y}_1 \ldots \hat{Y}_n$, in R^m, the objective is to find the parameter vector $\theta * \varepsilon \Theta$, that corresponds to the particular HMM from which the data were generated.

A common approach to Θ^* estimate is maximum likelihood estimation (MLE). The parameter estimate, denoted $\hat{\theta}_n^*$, is obtained via maximizing the log-likelihood.

$$\hat{\theta}_n = \arg max\, \theta^* \epsilon \Theta l_n\,(\theta) \qquad (1)$$

where

$$l_n(\theta) = logP_\theta\left(\hat{Y}_1 \ldots . \hat{Y}_n\right) = \sum_{i=1}^{n} logP_\theta\,(\hat{Y}_i|\hat{Y}_1 \ldots \ldots \hat{Y}_{i-1}) \qquad (2)$$

Unless the model is simple, for example, linear Gaussian or when the range χ of the latent variables $\{X_k\}_{k\geq 0}$ is a finite set, one can seldom evaluate the likelihood analytically. There are a variety of techniques, for example sequential Monte Carlo (SMC), for numerically estimating the likelihood. However, in a wide range of applications these methods cannot be used, for example when the conditional density of the observation $\hat{Y}_k$ given hidden state X_k is intractable, by which we mean that this -density cannot be evaluated analytically and has no unbiased Monte Carlo estimator. Despite this, one is often still able to generate samples from the corresponding processes for different values

of the parameter θ (e.g. Jasra et al., 2012). This has led to the development of methods in which θ * is estimated by taking the value of that maximizes some principled approximation of the likelihood.

One such approach is indirect inference (Gourieroux et al., 1993; Heggland and Frigessi, 2004. However, in the context of HMMs, when one does not adopt a linear Gaussian approximation of the filtering density, which can be very inaccurate as in extended Kalman filter approximations, this method could become more expensive. Estimating θ*for HMMs with an intractable conditional observation density can be performed in a Bayesian context (Campillo & Rossi, 2009). Campillo and Rossi (2009) utilizes the so-called convolution particle filter, which uses ideas from kernel density estimation, to replace the intractable observation density needed for the weight evaluation in the particle filter with their kernel estimates, to sequentially estimate the posterior distribution of θ *.

The convolution particle filter is closely related to another Bayesian approach called approximate Bayesian computation (ABC), which has recently received a great deal of attention. A non-exhaustive list of references includes McKinley et al. (2009), Peters et al. (2010), and Ratmann et al. (2009) In the standard ABC approach, given the data set $Y_1 \ldots Y_n$, the likelihood

$p_\theta(\hat{Y}_1 \ldots \hat{Y}_n)$ is replaced by the approximation

$$P_\theta(d(Y_1 \ldots \ldots Y_n : \hat{Y}_1 \ldots \ldots \hat{Y}_n) \leq \varepsilon \quad (3)$$

where $\{Y_k\}_{k\geq 1}$ denotes the observed state of the HMM with parameter θ , d(.;.) is some suitable metric on the n-fold product space $R^m \ldots R^m$ and $\varepsilon > 0$ is a constant that reflects the accuracy of the approximation. The probability in (1) is computed with respect to the law of $Y_1 \ldots Y_n$ while $\hat{Y}_1 \ldots \hat{Y}_n$ are regarded fixed. We call (1) the ABC likelihood and in practice, it can be estimated using Monte Carlo techniques.

The intuitive justification for the ABC likelihood is that for sufficiently small ε

$$P_\theta(d(Y_1 \ldots \ldots Y_n : \hat{Y}_1 \ldots \ldots \hat{Y}_n) \leq \varepsilon \approx P_\theta(\hat{Y}_1 \ldots \ldots \hat{Y}_n) V^\varepsilon \hat{Y}_1 \ldots \ldots \hat{Y}_n \quad (4)$$

where $V^{\varepsilon}_{\hat{Y}_1 \ldots \hat{Y}_n}$ denotes the volume of the d-ball of radius around the points $\hat{Y}_1 \ldots \hat{Y}_n$. Thus, the probabilities (1) will provide an approximation to the true likelihood, up to the value of some renormalizing factor, which is independent of θ and hence can be ignored. However, in general, it is not at all clear in what sense an approximation to the likelihood must be 'good' in order for the resulting inference procedures to be well behaved. In this study one of the things to do is to resolve this issue by directly investigating the effect of the parameter , not on the quality of the approximations (1), but on the behaviour of the estimates of θ* , which are obtained by maximizing the ABC likelihood as the length of the data set increases.

We shall use the following specialization of the standard ABC likelihood (1) for the HMM (Jasra et al., 2012): we approximate the likelihood associated to a given sequence of observations $\hat{Y}_1 \ldots \hat{Y}_n$ from an HMM with the probability

$$P_\theta(Y_1 \varepsilon B^{\varepsilon}_{\hat{Y}1} \ldots \ldots Y_n \varepsilon B^{\varepsilon}_{\hat{Y}n}) \quad (5)$$

where B^{ε}_{Y} denotes the ball of radius centred around point y. The benefits of the ABC likelihood in (5) is that it retains the Markovian structure of the model. This facilitates both simpler Markov chain Monte Carlo (McKinley et al., 2009) and SMC (Jasra et al., 2012) implementation of the ABC. Furthermore, we are able to compute an unbiased estimate of this ABC likelihood using SMC even when the conditional densities of the observations given the hidden state are intractable. This is a practically useful extension as such HMMs are used in practice.

Hidden Markov Model

A Hidden Markov model (HMM) is a stochastic process with an underlying, unseen Markov chain in which each state generates just one of K potential observations. These state-dependent observable output measurements are apparent to us. So we may say that an HMM is a Markov process with two components, one observable and one unobservable (hidden). The Markov property is fulfilled by the process S_t, which represents the underlying unobserved process of the HMM:

$$P(S_t = j|S_1 = i_1 \ldots . S_{t-1} = i_{t-1}) = P(S_t = j|S_{t-1} = i_{t-1}) \quad (6)$$

meaning that the probability of the process S of being in a state j depends only on the previous state i_{t-1}.

Let

$\pi_k = P(S_1 = k)$ be the initial probability of state k, $k = 1,\ldots,K$.

Let

$$P_{jk} = P(S_t = K|S_{t-1} = j) \quad (7)$$

denote the transition probability, that is, the probability of being in state k at time t given that previous state was j at time $t - 1$, We must also have that $\sum_{k=1}^{K} P_{jk} = 1$ and $P_{jk} > 0$. The initial probabilities $\pi_k = (\pi_1,\pi_2,\ldots,\pi_k)$ together with the transition probability matrix P, where P_{ij} is the elements of the matrix, govern the state switching behaviour of the chain. The number of time spent in each state before jumping to the next state is called the sojourn time. The probability of spending u consecutive time steps in state i under this model is

$$di\,(u) = P(S_{t+u+1} \neq i, S_{t+u} = i, S_{t+u-1} = i, \ldots\ldots, S_{t+2=i}|S_{t+1=i,} S_t \neq i) \quad (8)$$

$$= P_{ii}^{u-1}(1 - P_i)$$

We call $d_i(u)$ the sojourn density. Hence the sojourn time is geometrically distributed for any Markov chain, and the most likely sojourn time for any state is equal to 1.

The probability of an output observation, X_t, depends only on the state that produced the observation, S_t, and not on any other states or any other observations. Let $S = (S_t, t = 0,\ldots,T)$ denote the sequence of unobserved random variables, each with a finite state space $\{1,\ldots, J\}$, and let $X = (X_t, t = 1,\ldots,T)$ denote a corresponding set of observed random vectors. The process $\{X_t\}$ represents the state-dependent process of the HMM and fulfils the conditional (on the hidden states) independence property

$$P(X_t = x_t|X_{1=}x_{1,..}X_{t-1} = x_{t-1,}S_{1=}s_{1..}S_t = s_t) \quad (9)$$

$$P(X_t = x_t|P(S_t = s_t)$$

This is called the emission probability. The probability that you will see x_t given that at the same time you are in state s_t.

The HMM has the functional form

$$P(X,S) = P(X|S)P(S)\prod_{t=1}^{T} P(X_t|S_t)\prod_{t=1}^{T} P\,(S_t|S_{t-1}) \quad (10)$$

Given a specific observation sequence we want to calculate, we must first compute the joint probability of being in a particular hidden state sequence S_t and generating a particular sequence X_t of observable events. Then we must compute the total probability of the observations just by summing over all possible hidden state sequences.

$$P(X) = \sum_S P(X,S) = \sum_S P(X|S)P(S) \quad (11)$$

For a HMM with an observation sequence of T observations and N hidden states, there are N^T possible hidden sequences. In real life situations, even if the length of the sequences N and T are moderate, N^T becomes a very large number. This makes it hard for us to compute the total observation likelihood. Thus, there exist a couple of algorithms we can use that makes it easier for us to compute the probability. The Forward Algorithm, The Backward Algorithm and The Viterbi Algorithm are three of the most used algorithms to compute the probability of a given observation sequence.

Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF or NNMF), also non-negative matrix approximation (Dhillon and Sra, 2005; Rashish and Sra, 2010) is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect.

Consider a data matrix $V \in R_{m\times n}$ with m dimensions and n data points which has only non-negative elements. If we define two matrices, also with only non-negative elements:

$W \in R_{m\times r}$ and $H \in R_{r\times n}$, then NMF can reduce the dimensionality of V through the approximation:

$$V \approx WH \qquad (12)$$

where r < min(m, n) is the number of directions in the subspace of m that we are projecting onto. NMF looks like:

$$[V] = [W] * [H]$$

where the **W** and **H** matrices are both smaller than the original starting matrix, **V**.

The columns of **W** make up the new dimensions we are projecting onto. Each column of **H** represents the coefficients s of each data point in this new subspace.

An explanatory way of looking at NMF, as pointed out by Lee and Seung (2001), is to study each column of **V** at a time. Equation (12) then reduces to

$$\mathbf{v}_j \approx \mathbf{W}\mathbf{h}_j$$

where $\mathbf{v}_j$ and $\mathbf{h}_j$ are column vectors of **V** and **H** respectively.

The original m-dimensional data point, vj, is now represented in this new space by the r-dimensional hj. If we then consider each row of V independently, then the ith row of vj is $V_{i,j}$ and the approximation is:

$$V_{ij} = \sum_{k=1}^{r} W_{ik}H_{kj}$$

i.e. $V_{i,j}$, is approximated as a linear combination of all the elements of h_j.

The Weibull Distribution

The Weibull distribution, named after Swedish mathematician Waloddi Weibull, is a continuous probability distribution. He first developed the distribution as a model for material breaking strength, but in his 1951 publication A Statistical Distribution Function with Wide Applicability, he saw its potential. It is now widely used to evaluate product reliability, examine life data, and model failure times. The Weibull model can also fit data from many other domains, such as biology, economics, engineering sciences, and hydrology (Rinne, 2008).

Probability Density Function (PDF) of Weibull Distribution

The probability density function of a Weibull random variable is

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, x \geq 0 \qquad (13)$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution. Its complementary cumulative distribution function is a stretched exponential function. The Weibull distribution is related to a number of other probability distributions; in particular, it interpolates between the exponential distribution ($k = 1$) and the Rayleigh distribution ($k = 2$ and $\lambda = \sqrt{2}\sigma$. An important aspect of the Weibull distribution is how the values of the shape parameter and the scale parameter affect such distribution characteristics as the shape of the *pdf* curve, the reliability and the failure rate.

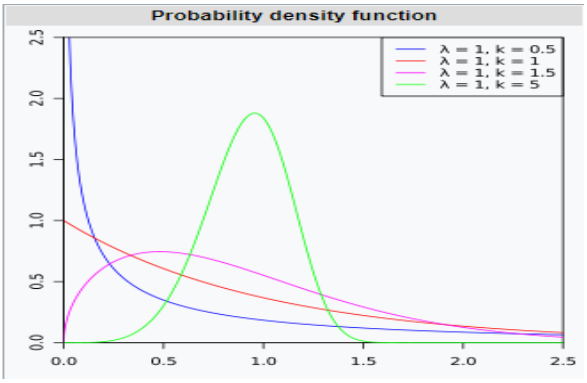

Figure 1. Probability Density Function of Weibull Distribution

The form of the density function of the Weibull distribution changes drastically with the value of k as shown in Figure 1. For $0 < k < 1$, $f(x) \to \infty$ as $x \to 0$ from above and is strictly decreasing. For $k = 1$, $f(x) \to 1/\lambda$ as $x \to 0$ from above and is strictly decreasing. For $k > 1$, $f(x) \to 0$ as $x \to 0$ from above, increases until its mode and decreases after it. The density function has infinite negative slope at $x = 0$ if $0 < k < 1$, infinite positive slope at $x = 0$ if $1 < k < 2$ and null slope at $x = 0$ if $k > 2$. For $k = 1$ the density has a finite negative slope at $x = 0$. For $k = 2$ the density has a finite positive slope at $x = 0$. As k goes to infinity, the Weibull distribution converges to a Dirac delta distribution centered at $x = \lambda$. Moreover, the skewness and coefficient of variation depend only on the shape parameter. A generalization of the Weibull distribution is the hyperbolastic distribution of type III.

Cumulative Distribution Function (CDF) of Weibull Distribution

The cumulative distribution function for the Weibull distribution is given by

$$F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k} \qquad (14)$$

for $x \geq 0$, and $F(x; k; \lambda) = 0$ for $x < 0$.

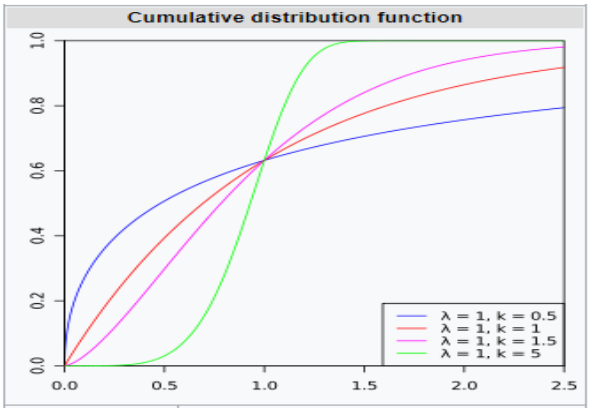

Figure 2. Cumulative Distribution Functions of Weibull

From figure 2, if $x = \lambda$ then $F(x; k; \lambda) = 1 - e^{-1} \approx 0.632$ for all values of k. Vice versa: at $F(x; k; \lambda) = 0.632$ the value of $x \approx \lambda$.

Quantile Function of Weibull Distribution

The quantile (inverse cumulative distribution) function for the Weibull distribution is

$$Q(p) = \lambda[-\ln(1-p)]^{1/k} \qquad (15)$$

for $0 \leq p < 1$. The Weibull variates would be generated from the quantile function in Equation (3).

Survival Function of Weibull Distribution

By definition, the survival function is given by

$$S(x) = 1 - F(x)$$

The survival function of Weibull distribution is given by

$$S(x) = e^{-\left(\frac{x}{\lambda}\right)^k} \quad (16)$$

Hazard Function of Weibull Distribution

By definition, the hazard function is given by

$$h(x) = \frac{f(x)}{1 - F(x)}$$

The hazard rate also known as failure rate h is given by

$$h(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \quad (17)$$

for $0 \leq p < 1$.

Cumulative Hazard of Weibull Distribution

By definition, the cumulative hazard function is given by

$$H(x) = -\ln[1 - F(x)]$$

The hazard rate also known as failure rate h is given by

$$H(x) = \left(\frac{x}{\lambda}\right)^k \quad (18)$$

Reverse Hazard of Weibull Distribution

By definition, the reverse hazard function is given by

$$h(x) = \frac{f(x)}{F(x)}$$

The reverse hazard rate also known as failure rate h is given by

$$h(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}\left[1 - e^{-\left(\frac{x}{\lambda}\right)^k}\right]^{-1} \quad (19)$$

HMM with Weibull Distributions

The proposed HMM models the multidimensional nonnegative signal X with a limited number of hidden states. In the field of speech processing, where the short-time magnitude/power spectral vectors are commonly used as the input matrix for NMF, these hidden states can be identified, for example, with different speech sounds (phones). Let us assume that the hidden random variable S_t (at time t) of the HMM can take one of I available discrete values $i = 1, ..., I$. Because of the nonnegative nature of the signal, the output probability density functions of the HMM are modelled as gamma distributions. Hence, the conditional distribution of each element of X is given as:

$$f(x_{rt}|S_t = i, G_t = g_t) = \frac{k_{ri}}{g_t\lambda_{ri}}\left(\frac{x_{rt}}{g_t\lambda_{ri}}\right)^{k_{ri}-1} e^{-\left(\frac{x_{rt}}{g_t\lambda_{ri}}\right)^{k_{ri}}}, x_{rt} \geq 0 \quad (20)$$

where G_t is the short-term stochastic gain parameter, r is the dimension index, and k_{ri} and λ_{ri} are state-dependent shape and scale parameters. Thus, the expected value and variance are obtained as $E(X_{rt}|S_t = i, G_t = g_t) = k_{ri}g_t\lambda_{ri}$ and $Var(X_{rt}|S_t = i, G_t = g_t) = k_{ri}(g_t\lambda_{ri})^2$. In modeling the speech spectra, the choice of the gamma distribution is motivated by the super-Gaussianity of the speech DFT coefficients.

We assume that, given the hidden state S_t, different elements of the vector X_t are independent. Hence, the HMM output density function is given as:

$$f(X_t|S_t = i, G_t = g_t) = \prod_{r=1}^{R} f(x_{rt}|S_t = i, G_t = g_t). \quad (21)$$

G_t is allowed to take only nonnegative values, and is assumed to have a Weibull distribution for the sake of tractability:

$$f(g_t) = \frac{\phi}{\theta_t}\left(\frac{g_t}{\theta_t}\right)^{\phi-1} e^{-\left(\frac{g_t}{\theta_t}\right)^{\phi}}, \quad g_t \geq 0 \quad (22)$$

with ϕ and θ_t being the shape and scale parameters, respectively. In practice, the scale parameters λ_{ri} are estimated to describe the signal statistics for different states while θ_t is meant to only model the long-term level changes of the signal. The sequence of hidden states is characterized by a first order Markov chain, with initial state probability mass vector p, with elements $p_i = f[S_{t=0} = i]$, and a transition probability matrix q, with elements $q_{ij} = f[S_{t=j} \mid S_{t-1} = i]$ as illustrated in Figure 1.

Weibull HMM as a Probabilistic NMF

The HMM described above can alternatively be formulated as a probabilistic NMF. A usual approach in probabilistic NMFs is to approximate an input matrix with an expected value that is obtained under the model assumptions. Then, the expected value is decomposed into a product of two nonnegative matrices. By doing so, a condensed representation of data can be achieved. Following this approach, we use an expected value of X_t to derive an NMF representation of the given vector x_t. That is, $x_t \approx \hat{\hat{x}}_t = vw_t$ where $\hat{\hat{x}}_t$ refers to the expected value. We calculate $\hat{\hat{x}}_t$ by considering the posterior distribution of the state and gain variables conditioned on the entire sequence of the observed signal over time, x:

$$\hat{x}_t = \sum_{i=1}^{I} \int E(X_t|S_t = i, G_t = g_t) f(S_t = i, g_t|x)\, dg_t . \quad (23)$$

Let us denote the hidden random states by a one-of-I indicator vector S_t i.e. $S_{it} = 1$ if the Markov chain at time t is in state i and $S_{jt} = 0$ for $j \neq i$. Accordingly, the realizations are referred to as s_t. We define the basis matrix $v \times r = [v_{ri}]$ as $v_{ri} = k_{ri}\lambda_{ri}$. Using this notation, we can write $v_t = vs_t$. Therefore, we have $E[X_t \mid s_t; g_t] = g_t v s_t$. Equation (21) can now be written as:

$$\hat{x}_t = v \sum_{s_t} s_t f(s_t|x) \int g_t f(g_t|s_t, x)\, dg_t \quad (24)$$

Denoting $w_t = v \sum_{s_t} s_t f(s_t|x)\, E(g_t|s_t, x)$, we can write: $\hat{x}_t = vw_t$. Thus, we have $x \approx vw$, i.e., x is factorized into two nonnegative factors, a basis matrix v and an NMF coefficients matrix w. In an extremely sparse case where $f(s_t^i \mid x) = 1$ only for one state s_t^i, depending on time t, and all the other states have zero probability, we will have $w_t = s_t^i E(gt \mid s_t^i, x)$.

The conditional state probabilities $f(s_t \mid x)$ can be calculated using the forward-backward algorithm. To finish our NMF derivation, we need to evaluate $E(g_t \mid st; x)$. Noting that g_t depends only on the observation at time t, the posterior distribution of the gain variable can be obtained by using the Bayes rule as:

$$f(g_t|s_t, x) = \frac{f(x_t|g_t, s_t) f(g_t)}{f(x_t|s_t)} \quad (25)$$

Since the denominator of equations (23) is constant, using (19) and (20) gives:

$$lnf(g_t|s_t = 1_i, x) \propto \frac{f(x_t|g_t, s_t)\frac{\phi}{\theta_t}\left(\frac{g_t}{\theta_t}\right)^{\phi-1} e^{-\left(\frac{g_t}{\theta_t}\right)^{\phi}}}{\frac{k_{ri}}{g_t\lambda_{ri}}\left(\frac{x_t}{g_t\lambda_{ri}}\right)^{k_{ri}-1} e^{-\left(\frac{x_{kt}}{g_t\lambda_{ri}}\right)^{k_{ri}}}} \quad (26)$$

$$f(X_t|S_t = i, G_t = g_t) = \prod_{r=1}^{R} f(x_{rt}|S_t = i, G_t = g_t) \quad (27)$$

$$f(g_t) = \frac{\phi}{\theta_t}\left(\frac{g_t}{\theta_t}\right)^{\phi-1} e^{-\left(\frac{g_t}{\theta_t}\right)^{\phi}}, \quad g_t \geq 0 \tag{28}$$

$$f(x_{rt}|S_t = i, G_t = g_t) = \frac{k_{ri}}{g_t\lambda_{ri}}\left(\frac{x_{kt}}{g_t\lambda_{ri}}\right)^{k_{ri}-1} e^{-\left(\frac{x_{kt}}{g_t\lambda_{ri}}\right)^{k_{ri}}}, x_{kt} \geq 0 \tag{29}$$

Equation (27) corresponds to a Generalized Weibull (GW) distribution with parameters $\phi = k_{ri}$ and $\theta_t = g_t\lambda_{ri}$.

Parameter Estimation

We propose an EM algorithm to estimate the model parameters, denoted by $\lambda = \{p; q; a; b; \phi; \theta\}$. Alternatively, the time-invariant parameters p; q; a; b; and ϕ can be obtained given a training data set. To obtain the NMF representation of a new vector in this case, the only unknown parameter is θ, which should be estimated.

In the E step of the EM, a lower bound is obtained on the log-likelihood of data, and in the M step, this lower bound is maximized. Let Z represent the hidden variables in the model. The EM lower bound takes the form

$$L\left(f(z|x,\lambda),\hat{\lambda}\right) = Q\left(\hat{\lambda},\lambda\right) + constant \tag{30}$$

Where

$$Q\left(\hat{\lambda},\lambda\right) = \int f(z|x,\lambda) ln\left(f\left(z,x|\hat{\lambda}\right)\right) dz \tag{31}$$

where λ includes the estimated parameters from the previous iteration of the EM, and λ contains the new estimates to be obtained.

For this problem, Z = {S;G} in which S = {S_1; …S_T} , and G = {G_1; …G_T} g where T is the number of data samples, i.e., the number of columns of x. Now, $Q\left(\hat{\lambda},\lambda\right)$ can be written as

$$Q(\hat{\lambda},\lambda) = \hat{Q}(\hat{\lambda},\lambda) + \sum_{t,i} w_t(i) \int f(g_t|x_t, S_t = i, \lambda)\left[lnf\left(g_t|\hat{\lambda}\right) + lnf\left(x_t|g_t, S_t = i, \hat{\lambda}\right)\right] dg_t$$

Here, $\hat{Q}\left(\hat{\lambda},\lambda\right)$ includes the terms for optimizing the Markov chain parameters p and q, which is done similarly to. The state probabilities ωt (i) = f(St = i | x,λ) are obtained by the forward-backward algorithm in which:

$$f(\mathbf{x}_t \mid S_t = i) = \int_0^{\infty} f(\mathbf{x}_t \mid S_t = i, G_t = g_t)\, f(g_t)\, dg_t. \tag{32}$$

Inserting (30) and (31) in (32), and using the definition of the GW distribution we get:

$$f(\mathbf{x}_t \mid S_t = i) = \frac{2\tau^{\vartheta/2}\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)}{\rho^{\vartheta/2}\theta_t^{\phi}\Gamma(\phi)} \prod_{k=1}^{K} \frac{x_{kt}^{a_{ki}-1}}{b_{ki}^{a_{ki}}\Gamma(a_{ki})}, \tag{33}$$

With

$$\rho = 1/\theta_t,\ \vartheta = \phi - \sum_{k=1}^{K} a_{ki},\ \tau = \sum_{k=1}^{K} x_{kt} b_{ki}^{-1}$$

To obtain the new estimate of the rest of the parameters, (33) is differentiated w.r.t. parameters of interest, and the result is set to zero. Obtaining the gradient w.r.t. $\hat{b}_{ki}$ and setting it to zero yields the following estimate:

$$\hat{b}_{ki} = \frac{\sum_t \omega_t(i)\, x_{kt} E\left(G_t^{-1} \mid \mathbf{x}_t, S_t, \lambda\right)}{\hat{a}_{ki} \sum_t \omega_t(i)} \stackrel{\text{def}}{=} \frac{\mu_{ki}}{\hat{a}_{ki}}. \tag{34}$$

Inserting (33) into (34), and setting the gradient of the objective function w.r.t. $\hat{a}_{ki}$ to zero yields:

$$\varphi\left(\hat{a}_{ki}\right)-\ln\left(\hat{a}_{ki}\right)=\frac{\sum_t \omega_t\left(i\right)\left(\ln x_{kt}-E\left(\ln G_t \mid \mathbf{x}_t, S_t, \lambda\right)-\ln \mu_{ki}\right)}{\sum_t \omega_t\left(i\right)}, \quad (35)$$

Where $\varphi(u) = \frac{d}{du}\ln\Gamma(u)$ is the digamma function. Therefore, **a** is first estimated using (35), then (33) is solved to obtain bˆ. Similarly, φˆ and θˆ are obtained by first estimating the shape parameter φ as:

$$\varphi\left(\hat{\phi}\right)-\ln\left(\hat{\phi}\right)=\frac{\sum_{t,i}\omega_t\left(i\right)\left(E\left(\ln G_t \mid \mathbf{x}_t, S_t, \lambda\right)-\ln \xi_t\right)}{\sum_{t,i}\omega_t\left(i\right)} \quad (36)$$

with ξt defined as:

$$\xi_t=\frac{\sum_i \omega_t\left(i\right)E\left(G_t \mid \mathbf{x}_t, S_t, \lambda\right)}{\sum_i \omega_t\left(i\right)}, \quad (37)$$

and then using φˆ to estimate θt:

$$\hat{\theta}\mathrm{t} = \xi\mathrm{t}/\hat{\varphi}. \quad (38)$$

Results

Data

The data used in this study is a USA COVID-19 daily data collected from 23rd January 2020 to 22nd July, covering 547 days. Another variable collected is the number of COVID-19 hospitalized cases on daily basis covering 368 days, because hospitalized cases were not reported until 15th July 2020. The daily data fluctuates but are not normally distributed.The COVID-19 data increased, decreased or remain the same from day to day. For the hospitalized data, it either increased or not increased. An increase in a previous day, does not guarantee and increase or decrease in the subsequent days. The data are discrete and there are some days that there are no new cases of COVID-19 or hospitalization. Thus, a zero data point might occur. Both gamma and Weibull will not be good fit for discrete data. The moving average is used to make sure zero does not exist in the data point, and produces a smoother curve that are continuous.

Table 1. Summary of COVID-19 Cases and Moving Average

	Min	1st Q.	Median	Mean	3rd Q.	Max	Skewness	Kurtosis
New Cases	0	22,732	44,251	62,673	70,107	300,462	1.5686	4.7379
Moving Ave. Cases	0.2	22,841.8	43,149.9	62,960.3	66,910.6	245,409.1	1.4730	4.1660
Hospitalized Cases	12,218	28,516	37,986	50,620	65,430	133,214	1.0678	2.8924
Moving Ave. Hospitalized	9,810	27,809	37,845	50,195	66,280	128,245	1.0394	2.8511

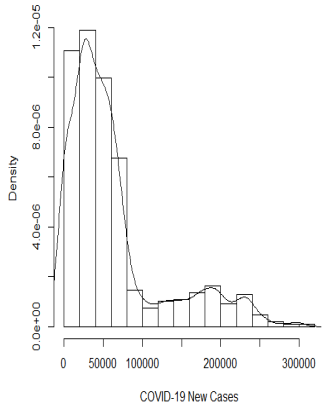

Figure 3. COVID-19 Observed Daily Cases

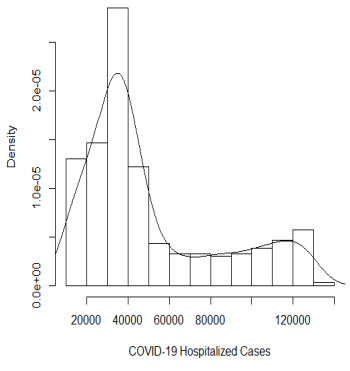

Figure 4. COVID-19 Observed Hospitalised Cases

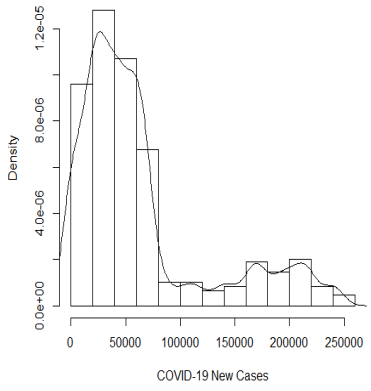

Figure 5. Covid-19 Moving Average

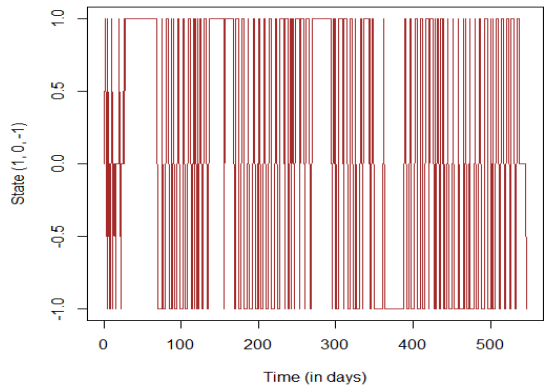

Figure 6. State per Time

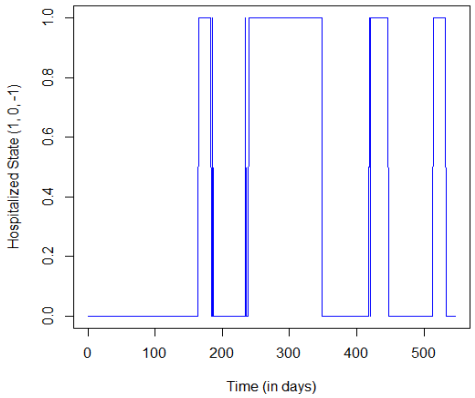

Figure 7. Hospitalized State per Time in Days

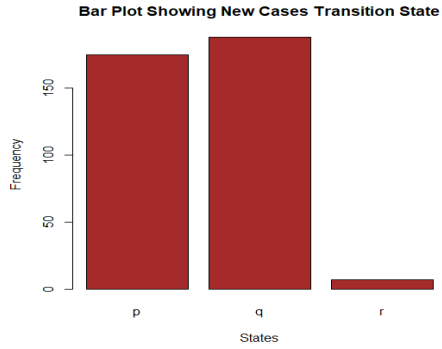

Figure 8. Bar Plot Showing New Cases

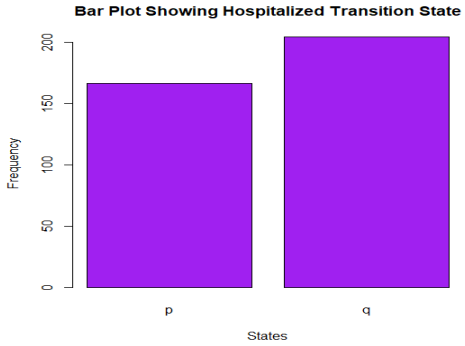

Figure 9. Bar Plot Showing Hospitalized Transition State

The EDA of the data collected is used to show some hidden features in the dataset that might not be easily seen by mere observation of the data. The summary of the data collected on US COVID-19 new and hospitalised cases are shown on Table 1.

Data Transmutation to State

If the COVID-19 new cases from day to day increased, it is positive (1), if it decreased, it is negative (-1), but if it remained the same, it is neither positive nor negative (0).

Let y_t represent the number of new cases reported at day t. Let y_{t+1} be number of new cases reported on day $t + 1$.

So, if $y_{t+1} > y_t$, assign 1; if $y_{t+1} = y_t$, assign 0 and if $y_{t+1} < y_t$, assign -1.

These three states 1, 0 and -1, implies that COVID-19 reported cases can increase, remain the same or decrease. The probability that it will increase is p, the probability that it will remain the same is r and the probability that it will decrease is q.

Thus

$$p + q + r = 1$$

Also, let x_t represent the number of new hospitalised reported cases at day t. Let x_{t+1} be number of hospitalised cases reported on day $t + 1$.

So, if $x_{t+1} > x_t$, assign 1; else assign 0.

These two states 1 and 0 implies that COVID-19 reported hospitalised cases can increase or not. The probability that it will increase is p, the probability that it will not increase is q.

Thus

$$p + q = 1$$

The continuous data for the two variables are achieved with moving average to convert the discrete data to continuous and remove every point with zero entries. Zero entry implies no new case was recorded for that day and this will affect the gamma and Weibull models.

Transition State

Table 2. The table shows the Covid-19 New Cases Observed

		COVID-19 New Cases (Observed)			
		P	**q**	**R**	**Pr**
Hospitalized Cases (Hidden)	**P**	0.3027	0.1459	0.0000	0.4486
	Q	0.1703	0.3622	0.0189	0.5514
	Pr	0.5191	0.4351	0.0457	1

Table 3. Basic Statistics Moving Average (MA) for COVID-19 New and Hospitalised Cases

	New Cases	MA New Cases	Hospitalised	MA Hospitalised
N	547	547	368	368
NAs	0	0	0	0
Minimum	0	0.2	12,218	9,810.0
Maximum	300,462	245,409.1	133,214	128,245.4
1st Quartile	22,732	22,841.85	28,516.25	27,808.72
3rd Quartile	70,107	66,910.55	65,429.50	66,279.68
Mean	62,672.51	62,960.27	50,620.47	50,195.09
Median	44,251.00	43,149.90	37,986.00	37,844.55
Sum	34,281,860	34,439,270	18,628,330	18,471,790
SE Mean	2,695.849	2,622.013	1,730.090	1,741.916
LCL Mean	57,377.00	57,809.80	47,218.34	46,769.70
UCL Mean	67,968.02	68,110.74	54,022.61	53,620.48
Variance	3,975,379,000	3,760,599,000	1,101,502,000	1,116,612,000
Stdev	63,050.60	61,323.73	33,188.88	33,415.75
Skewness	1.5643	1.4690	1.0635	1.034511
Kurtosis	1.7206	1.1508	-0.1233	-0.1735

Transition Matrix for New Cases

$$\begin{array}{c} \\ q \\ r \\ p \end{array} \begin{array}{c} \begin{array}{ccc} q & r & p \end{array} \\ \begin{pmatrix} 0.6878 & 0.0169 & 0.2954 \\ 0.2800 & 0.5600 & 0.1600 \\ 0.2394 & 0.0211 & 0.7394 \end{pmatrix} \end{array}$$

Transition Matrix for Hospitalised Cases

$$\begin{array}{c} \\ q \\ p \end{array} \begin{array}{c} \begin{array}{cc} q & p \end{array} \\ \begin{pmatrix} 0.9811 & 0.0189 \\ 0.0398 & 0.9602 \end{pmatrix} \end{array}$$

trans2<-trans^2
Transition Matrix for New Cases

$$\begin{array}{c} \\ q \\ r \\ p \end{array} \begin{array}{c} \begin{array}{ccc} q & r & p \end{array} \\ \begin{pmatrix} 0.5485 & 0.0273 & 0.2954 \\ 0.3877 & 0.3217 & 0.1600 \\ 0.3476 & 0.0315 & 0.7394 \end{pmatrix} \end{array}$$

trans5<-trans^5
Transition Matrix for New Cases

$$\begin{array}{c} \\ q \\ r \\ p \end{array} \begin{array}{c} \begin{array}{ccc} q & r & p \end{array} \\ \begin{pmatrix} 0.4467 & 0.0392 & 0.5142 \\ 0.4410 & 0.0857 & 0.4734 \\ 0.4287 & 0.0405 & 0.5308 \end{pmatrix} \end{array}$$

trans10<-trans^10
Transition Matrix for New Cases

$$\begin{array}{c} \\ q \\ r \\ p \end{array} \begin{array}{c} \begin{array}{ccc} q & r & p \end{array} \\ \begin{pmatrix} 0.4372 & 0.0417 & 0.5211 \\ 0.4377 & 0.0438 & 0.5185 \\ 0.4369 & 0.0417 & 0.5213 \end{pmatrix} \end{array}$$

trans100<-trans^100
Transition Matrix for New Cases

$$\begin{array}{c} \\ q \\ r \\ p \end{array} \begin{array}{c} \begin{array}{ccc} q & r & p \end{array} \\ \begin{pmatrix} 0.4371 & 0.0418 & 0.5211 \\ 0.4371 & 0.0418 & 0.5211 \\ 0.4371 & 0.0418 & 0.5211 \end{pmatrix} \end{array}$$

Table 4. Evaluation Statistics of the Weibull, Gamma and Log Normal Distributions

	Weibull	Gamma	Log Normal
MLE Parameters	Shape: 3.2143 (0.1269) Scale 0.8116 (0.0110)	Shape: 3.3936 (0.1960) Rate : 4.5648 (0.2841)	Meanlog: -0.4510 (0.0377) Sdlog:0.8812 (0.0266)
-LogL	87.55773	222.0167	460.2831
AIC	179.1155	448.0333	924.5662
BIC	187.7244	456.6422	933.175
K-S Stat.	0.2661	0.3239	0.3348
Cramer-von Mises stat.	9.4202	14.3585554	17.9709
Anderson-Darling stat.	51.6876	70.48894	87.7225

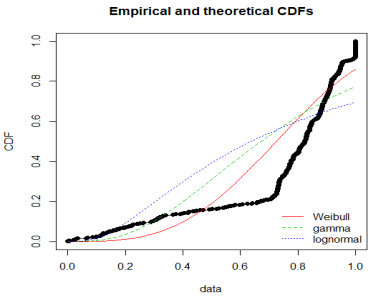

Figure 10. Empirical and Theoretical CDFs

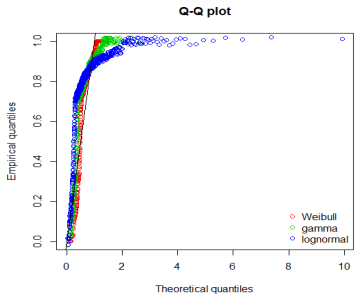

Figure 11. Empirical Quantiles

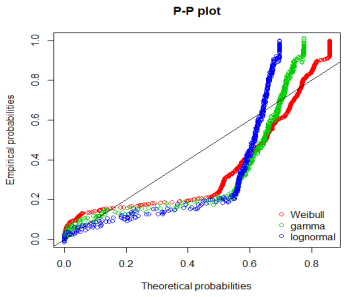

Figure 12. Empirical Probabilities

Discussion

Weibull HMM was proposed in the work to develop an NMF approximation of the COVID-19 laboratory confirmed new cases in a predictive way. The COVID-19 data from the United States are utilized here for demonstrative purposes only. The time-varying scale parameter t was calculated with the value gained from the training data being fixed.

Figure 1 depicts the data's empirical and theoretical cumulative distribution functions, as well as the competing models' NMF representations. The plot illustrates that only Weibull NMF reflected the COVID-19 data since it follows the data in the predicted direction, unlike gamma and lognormal.

The plot of empirical quantiles against theoretical quantiles of the data and the three competing models is shown in Figures 10–12. The Weibull model is a smooth approximation of the power spectra of various data sets. This demonstrates the suggested scheme's potential for coping with noise and unwanted fluctuations when applied to a more realistic application such as voice recognition or enhancement.

To have a better understanding of the obtained factorization, the derived NMF representation

was compared to two state-of-the-art approaches using model selection statistic and criteria. We considered two variants of NHMM: the gamma model and the lognormal model with same number of observations each. Aimed to compare the goodness of fit as a function of the sparsity level, the NMF coefficient vectors, $\mathbf{w}_t$, were constrained to have a specified l_0 norm (number of non-zero elements). For each value of l_0 in KL-NMF, the training was repeated to have the best possible basis matrix under the sparsity constraint.

The number of the consecutive vectors, $\mathbf{x}_t$, for which the same basis vector had the largest coefficient was computed, and it was normalized by the total number of columns in $\mathbf{x}$. In other words, if the system is in the state i in the current time frame, this measure (denoted as P_{rep}) gives the probability of staying in the same state in the next time frame.

The findings were based on an average of 368 days of COVID-19 reported cases. The experiment demonstrates that the proposed approach results in a sparse representation with a tradeoff between accurate fitting and temporal modeling. Weibull HMM approximation accuracy is comparable to gamma HMM and lognormal HMM utilizing negative log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) selection criteria. K-S, Cramer-von Mises, and Anderson-Darling statistics are also employed.

However, the Weibull HMM gives a higher probability to continuously stay in the same state. The Weibull HMM approach provides the best fit to the observed data, and the probability of staying in the same state is the highest using this method. By comparing actual and moving average variants, we see that the moving average provides both better temporal modelling and better fit as each state of the moving average has a greater flexibility to model the observation.

In terms of the computational complexity, Weibull HMM is substantially faster than gamma and competed well with log-normal. When applied for the factorization, the computational requirements of the Weibull HMM is comparable with that of the gamma and lognormal, but for some computation of some criteria it is a little more demanding as gamma and lognormal have more existing known algorithm than Weibull.

Conclusion

The purpose of this theoretical investigation was to develop a probabilistic NMF technique. To use the signal's temporal correlations, we created an HMM using Weibull output density functions (Weibull HMM). Furthermore, another Weibull distribution was explored to govern the signal's long-term level. We demonstrated the analogy between the Weibull HMM and the NMF, and so generated a novel probabilistic NMF. This work serves as the foundation for a variety of applications, including stock price swings, speech and image processing, and daily COVID-19 instances. When the suggested Weibull NMF is compared to other approaches, the Weibull model outperforms the gamma and lognormal models for the US COVID-19 data. COVID-19 instances in the United States peaked and then declined, only to climb again. The likelihood of an increase in new cases in the future days is smaller than that of fewer cases, but this chance is quite close. As a result, it is recommended that the government implement severe steps to prevent the spread and the strain on the hospital.

References

Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., & Plemmons, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal., 52*, 155-173. https://doi.org/10.1016/J.CSDA.2006.11.006

Cappe, O., Moulines, E., & Ryden, T. (2005). *Inference in hidden Markov models*. New York, NY: Springer.

Cemgil A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, *2009*, 785152. https://doi.org/10.1155/2009/785152

Cichocki, A. Zdunek, R. Phan, A. H. & Amari, S. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. New York: John Wiley & Sons.

Cybenko, G. & Crespi, V. (2008). Learning Hidden Markov Models using Non-Negative Matrix Factorization. *Arxiv preprint arXiv:0809.408*.

Dhillon, I.S. & Sra, S. (2005). *Generalized Nonnegative Matrix Approximations with Bregman Divergences (PDF)*. NIPS.

Felsenstein, J., & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, *13*(1), 93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575

Fevotte, C. & Cemgil, A. T. (2009). Non-negative matrix factorisations as probabilistic inference in composite models. *In 17 European Signal Processing Conference (EUSIPCO), 47,* 1913–1917.

Hoffman, M. D., Blei, D. M. & Cook P. R. (2010). Bayesian non-parametric matrix factorization for recorded music. *In Processing International Conference of Machine Learning* (pp. 439–446).

Hsu, D., Kakade, S.M. & Zhang, T. (2009). A spectral algorithm for learning hidden markov models. *Proceedings of the Conference on Learning Theory* (*COLT*).

Kim, S., Shephard N., & Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies, 65,* 361-393. https://doi.org/10.1111/1467-937X.00050

Lakshminarayanan, B. & Raich R. (2013). *Non-negative matrix factorization for parameter estimation in Hidden Markov models*. School of EECS, Oregon State University, Corvallis, OR 97331-5501.

Lee, D.D. & Seung, H.S. (2001). Algorithms for non-negative matrix factorization, *Advances in neural information processing systems, 13,* 556-562.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discretevalued time series, vol 70: Monographs on statistics and applied probability*. London: Chapman & Hall.

Mohammadiha, N., Taghia, J. & Leijon, A. (2017). Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions. *In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)* 4561–4564. http://dx.doi.org/10.1016/j.apacoust.2016.04.024

Mysore, G. J., Smaragdis, P. & Raj, B. (2010). Non-negative hidden Markov modelling of audio with application to source separation. *In Int. Conf. on Latent Variable Analysis and Signal Separation,* 140–148.

Nkemnole, E. B., Abass, O. and Kasumu, R. A. (2013). Parameter Estimation of a Class of Hidden Markov Model with Diagnostics. *Journal of Modern Applied Statistical Methods, 12*(1), 181–197. https://doi.org/10.56801/10.56801/v12.i.652

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the *IEEE,* 77(2), 257–286. https://doi.org/10.1109/5.18626

Rashish, T. & Sra, S. (2010). *Sparse nonnegative matrix approximation: new formulations and algorithms* (PDF). TR.