MIXTURES MATTER: INTERPRETABLE FORECASTING FOR EPIDEMICS

Arnab Sarker
MIT
arnabs@mit.edu

Ali Jadbabaie
MIT
jadbabai@mit.edu

Devavrat Shah
MIT
devavrat@mit.edu

ABSTRACT

We introduce a simple mixture-based model for predicting cases and deaths in an epidemic. The model represents time series of cases and fatalities as a mixture of Gaussians. Empirically, we find it to have low prediction error on COVID-19 case data,[1] with best results when our model selection procedure identifies an appropriate number of Gaussian components. We provide a simple learning algorithm to identify model parameters from data and establish its efficacy theoretically. Furthermore, we show that such a model is the natural outcome of a stochastic process on a graph based on a mechanistic SIR framework. This allows the learned parameters to take on a meaningful interpretation that encodes behaviors, which can enable policy makers to better understand the progress of the pandemic.

1 INTRODUCTION

In the early stages of the pandemic, the public turned to experts in order to understand the potential magnitude of SARS-CoV-2. With very few data points available, many models prognosticated a long period of exponential growth in cases consistent with an SIR model in a large population. Yet, since human behavior has historically ensured that prolonged exponential growth is rarely the case, non-mechanistic models such as that from the Institute of Health Metric and Evaluation (IHME) gained mass attention (Murray, 2020). Rather than assume an underlying process of infection, the non-mechanistic approach makes the implicit assumption that human behavior will somehow result in sub-exponential growth.

However, a common critique of such methods is that they lack interpretability. Because the assumptions on behavior are not explicit, it is difficult to understand the effect of endogeneity. While non-mechanistic methods tend to work well empirically, they lack the same meaningful parameters of traditional SIR models (Kermack & McKendrick, 1927; Holmdahl & Buckee, 2020).

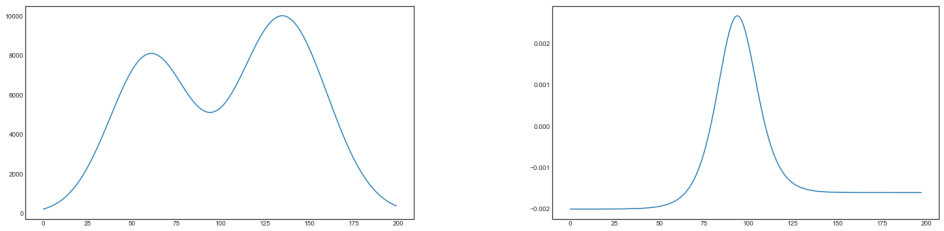
In this work, we focus on a non-mechanistic approach which, upon further examination, can be interpreted as a mechanistic SIR model. That is, the model can be seen as a bridge between interpretable mechanistic models and data-efficient non-mechanistic models. Specifically, we assume that the observed time series of cases has the form

$$N(t) = \sum_{k=1}^{r} M_k e^{-a_k(t-C_k)^2} \qquad t = 0, 1, \ldots, T\,. \tag{1}$$

The idea of modeling cases as a mixture stems from the reality that the disease is spreading to a population which has diverse regional divisions and includes many jurisdictions (Chandrasekhar et al., 2020). Since each region has its own features and policies, we expect the observed case counts to take an additive form. The specific form of the Gaussian time series is chosen in part due to historical prevalence (Farr, 1840; Santillana et al., 2018), and the parameterization is rigorously justified in Section 3. While a much wider variety of function classes can explain sub-exponential growth (Dandekar & Barbastathis, 2020), we restrict to the parsimonious class in equation (1) since the restrictive assumption better justifies applications to out-of-sample prediction.

Formally, equation (1) admits a simple algorithm for learning time series as a mixture of Gaussian curves. Further, such a time-series can be seen as arising from a simple stochastic process on a

[1] This model is available live at http://covidpredictions.mit.edu.



(a) Idealized Number of Cases $N(t)$ (b) Idealized $S(t)$

Figure 1: Idealized Time Series for Learning a Mixture.

graph, allowing for meaningful interpretation of parameters. This interpretation can be validated with mobility data, and shows that policy makers can use this approach to better understand the progress of the pandemic. From an empirical perspective, we have found that our prediction error is relatively small compared to other models, with best results when our model selection procedure selects the correct number of components r.

Ultimately, prediction based on the function class in equation (1) appears to strike a desirable balance between complex, SIR-based models with time-varying parameters, which are interpretable but may overfit to existing data or depend too much on historical assumptions, and other commonly used non-mechanistic models which make overly restrictive parametric assumptions on the curve shape and lack a meaningful interpretation. Moreover, while policy makers in practice may still prefer to use deep-learning based methods to prioritize accuracy Shahid et al. (2020), or agent-based models to allow for refined interpretability Rockett et al. (2020), equation (1) provides a foundation for models which provide both accurate and interpretable forecasting.

2 An Algorithm for Learning Mixtures

In showing that the Gaussian components can be learned from data, we first note that approaches based on expectation-maximization do not fit the problem setting, as such algorithms require samples from a distribution as opposed to observations of a 1-D time series. Instead, to identify the different Gaussian components of the mixture in equation (1), we use a simple algorithm based on taking the second derivative of the log of the time series. First, given observations $N(t)$, we compute the time series

$$S(t) = \log \frac{N(t+1)}{N(t)} - \log \frac{N(t)}{N(t-1)} .$$

The computation of $S(t)$ in the ideal case is shown in Figure 1, and the following two observations can be made about $S(t)$. First, we see that if a single Gaussian component is "dominant," then $S(t)$ is flat, and is approximately equal to $-2a_k$ where a_k is the quadratic coefficient of the dominant cluster. This approximation arises from the similarities $S(t)$ shares with the `log-sum-exp` function. Second, we note that in regions where the two clusters have similar counts, $S(t)$ increases and reaches a local maximum.

In our algorithm, we exploit this local maximum in order to identify the midpoint between Gaussian components, creating disjoint intervals of time in which each interval corresponds to a single dominant Gaussian curve. Once these midpoints are defined, the problem is reduced to identifying the parameters of the dominant Gaussian components in each interval. The task of identifying such parameters is simple as long as the Gaussian curves are well-separated.

To formalize the above claim, we will consider the case $r = 2$, and assume that the Gaussian curves in equation (1) satisfy $M_k \leq M$ and $a_k \geq a$ for for $k = 1, 2$ for some constants M and a. These two assumptions are reasonable as they imply that the observed time series is bounded, and that the shape of the Gaussian curves is not too flat. If such conditions hold, we may say that two Gaussian curves are ϵ-separated for some $\epsilon > 0$ if

$$|C_1 - C_2| \geq 2\sqrt{\frac{1}{a} \log \frac{M}{\epsilon}} .$$

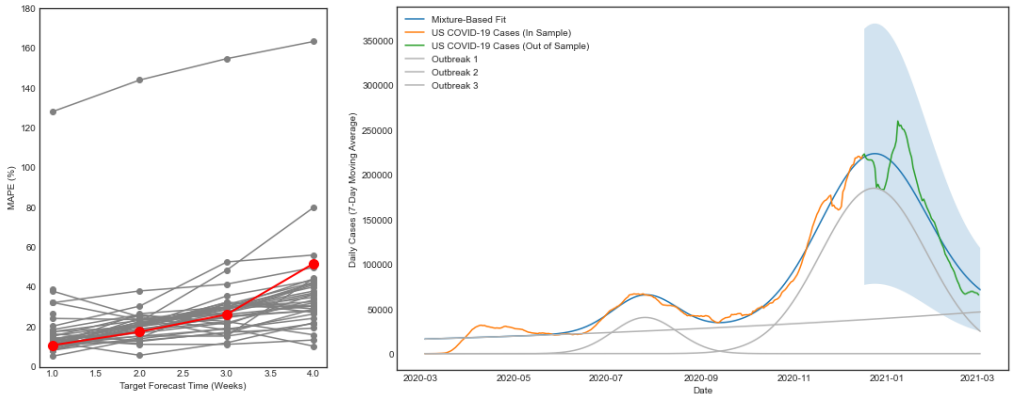


Figure 2: Accuracy of the Mixture Model in Predicting COVID-19 Cases for the United States. (Left) In terms of median accuracy of one week ahead forecasts, the mixture model (red) places 14th out of 34 available models, and 4th among models that do not use data apart from case counts.[2] (Right) An example of a remarkable forecast based on equation (1), when the number of components r is selected appropriately.

The above condition implies that, if $N_1(t) \geq N_2(t)$, then $N_2(t) \leq \epsilon$, and vice versa. Stated differently, if one component is "dominant," the other component will have size at most ϵ.

Proposition 1 (Algorithmic Guarantee)**.** *Assume the Gaussian components are well separated, let* $t^* = \arg\max S(t)$*, and define the intervals* $T_1 = \{0, \ldots, t^* - 1\}$ *and* $T_2 \in \{t^*, \ldots, T\}$*. Then, for* $k = 1, 2$*, the estimates* $\widehat{M}_k = \max_{t \in T_k} N(t)$ *and* $\widehat{C}_k = \arg\max_{t \in T_k} N(t)$ *satisfy*

$$M_k \leq \widehat{M}_k \leq M_k + \epsilon\,, \qquad \textit{and} \qquad |\widehat{C}_k - C_k| \leq \sqrt{\frac{1}{a_k} \log\left(\frac{M_k}{M_k - \epsilon}\right)} \approx \sqrt{\frac{\epsilon}{a_k M_k}}\,. \quad (2)$$

The proof of the claim follows from the definition of ϵ-separated Gaussian components, and the approximation for the bound on $|\widehat{C}_k - C_k|$ follows from a Taylor expansion when $M_k \gg \epsilon$. We note that this Proposition may be easily generalized to the case $r > 2$, as the assumption of well-separated components will make no more than 2 components non-negligible at any particular time t. In practice, the selection of r is performed using the BIC criterion to appropriately trade off in-sample accuracy and model complexity.

We also note a variant of Proposition 1 can also be stated for the case in which the observed series is subject to independent bounded noise. Such a noise condition may be further utilized in order to provide confidence intervals on the predictions, as depicted in Figure 2.

3 Connections to the SIR Model

While equation (1) provides an intuitive mixture model, and performs well empirically as shown in Figure 2, it does not immediately provide interpretation in the same way as traditional mechanistic models (Kermack & McKendrick, 1927). To that end, we introduce a simple stochastic model that captures population heterogeneity, incorporates traditional epidemiological dynamics, and provides a mechanistic justification for the time series in equation (1).

Our model begins with a graph $G = (V, E)$, where each node $v \in V$ represents an individual and edges represent connections between individuals by which the disease may spread. To impose community structure, we will assume that G takes the form of a stochastic block model, in which $V = V_1 \cup V_2$ can be decomposed into two disjoint communities. Edges form with probability p for

[2]Predictions are collected from `https://github.com/reichlab/covid19-forecast-hub` and true case data is provided by Johns Hopkins.

each pair of nodes within the same community, and with a probability $q \ll p$ for individuals from different communities.[3]

The spreading model on the graph G follows from the standard SIR model on networks (Easley et al., 2010). At time $t = 0$, we assume that one individual in V_1 will be infected. Then, for each time step t, each node will independently attempt to infect its neighbors and succeed with probability β. After a node has been infected for one time epoch, it will move to the recovered state, where it can not become infected again.

The final aspect of our model, which encodes human behavior, is the notion of "degree pruning." Specifically, at each time step t, each edge will be removed from the graph with probability $1 - \gamma$, for some $0 < \gamma < 1$. This process represents how individuals may respond to the pandemic, using measures such as social distancing, masking, and vaccination in order to stop the spread of the virus.

It is worth noting that the above model provides but one formalization by which case counts of the form in equation (1) can arise, and we prefer this model for its tractability in analysis. In particular, the model may also arise from an SIR-model with time-varying reproductive rate, or in other network-based models with a degree distribution which emulates the degree pruning parameter above. Even situations in which the spread has spatial heterogeneity can be captured, so long as the simultaneous outbreaks have similar features resulting in global observations of the Gaussian curve, and the temporal separation of Gaussian curves may reflect different waves of the pandemic (Epstein et al., 2008).

With the network-based SIR model defined, we can show that within each community, the number of infected individuals will follow a Gaussian shape in its time series. To be precise, we let $n_i = |V_i|$ and hold the average degree within and outside of each community constant as each n_i goes to infinity. That is, we assume $\lim_{n_i \to \infty} pn_i = c_i$ for some constants c_i, which allows for the following proposition.

Proposition 2 (Gaussian Components)**.** *Let $I_1(t)$ represent the random variable indicating the number of infected individuals in V_1. Further, assume that $q \ll p$, so that infections resulting from V_2 are negligible. Then, in the model above,*

$$\lim_{n_1 \to \infty} \mathbf{E}[I_1(t)] = e^{-\frac{1}{2}\log\gamma t^2 + \log\left(\beta c_1/\sqrt{\gamma}\right)t}\,. \tag{3}$$

The claim above roughly follows from modeling the cases within the community as a branching process, and appealing to the definition of γ. Beyond the spread within a community, we are also able to characterize the spread of the disease between communities, as noted in the following proposition, assuming that the degree between communities is also held constant, i.e. there exists a parameter c_{12} such that $c_{12} = \lim_{n_2 \to \infty} qn_2$.

Proposition 3 (Timing Between Communities)**.** *Define $X(t)$ to be the random variable representing the number of infections in V_2 which result from infected individuals in V_1. Further, define*

$$t_I = \min\left\{ t \mid \lim_{n_2 \to \infty} \lim_{n_1 \to \infty} \mathbf{E}[X(1) + \cdots + X(t)] \geq 1 \right\}.$$

Then,

$$t_I \geq \Phi^{-1}\left(\frac{1}{c_{12}\beta} \times \frac{1}{I_{1,tot}}\right) \times \frac{1}{\sqrt{-\log\gamma}} + t_{1,max} + 1\,,$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal Gaussian, $I_{1,tot} = \int_{-\infty}^{\infty} e^{\frac{1}{2}\log\gamma t^2 + \log\left(\beta c_1\sqrt{\gamma}\right)t} dt$, and $t_{1,max} = \text{argmax}_t\, I(t)\gamma^t$.

Thus, we see that if q is sufficiently small, V_2 will have to wait a nontrivial period of time for its first infection, and it is likely that the Gaussian components in the time series will be well-separated. Ultimately Propositions 2 and 3 allow us to generate the two key components of equation (1), namely its Gaussian components and additive form, in a mechanistic fashion.

[3]For simplicity, we consider the case of two communities, although the results can hold properly for $k > 2$ communities as well, depending on the structure of the block model.

	RETAIL	GROCERY	PARKS	TRANSIT	WORKPLACES	RESIDENTIAL
Correlation	-0.209	-0.286	-0.281	-0.248	0.770	0.109
p-value	0.039	0.004	0.005	0.014	0.454	0.285

Table 1: Comparison of learned γ parameters in each US state to its corresponding Google mobility data. When using mobility in order to predict degree pruning rates, the six regressors above can predict degree pruning with an r^2 value of 0.092; when restricting to outbreaks that occur before July 31st, the r^2 value increases to 0.616. This suggests the importance of mobility in degree pruning earlier in the pandemic, and that other factors may be more important in later stages of the pandemic.

4 DISCUSSION

The ability to learn the components of Gaussians empirically and interpret parameters using the theoretical model provides an opportunity to validate the interpretations of the γ parameters learned from data. Proposition 2 indicates that empirical γ values can be computed from the quadratic parameter of each Gaussian curve. The validation of parameters is shown in Table 1, in which we compare learned γ parameters from case counts in 50 states to corresponding Google Mobility data. The table validates the interpretation of γ as being associated with individual's behavior, with four out of six mobility metrics showing statistical significance in their relationship with γ.

Overall, a key goal of bringing interpretation to non-mechanistic methods is to enable policy makers to make data-driven decisions. In future work, our hope is to better understand the implications of interventions on γ, so that practitioners will have the foresight to effectively mitigate the spread of an epidemic in a closed-loop fashion, ensuring that cases never peak above a specified maximum.

REFERENCES

Arun G Chandrasekhar, Paul S Goldsmith-Pinkham, Matthew O Jackson, and Samuel Thau. Interacting regional policies in containing a disease. *Available at SSRN*, 2020.

Raj Dandekar and George Barbastathis. Quantifying the effect of quarantine control in covid-19 infectious spread using machine learning. *medRxiv*, 2020.

David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

Joshua M Epstein, Jon Parker, Derek Cummings, and Ross A Hammond. Coupled contagion dynamics of fear and disease: mathematical and computational explorations. *PLoS One*, 3(12): e3955, 2008.

William Farr. Progress of epidemics. *Second report of the Registrar General of England and Wales*, pp. 16–20, 1840.

Inga Holmdahl and Caroline Buckee. Wrong but useful—what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine*, 2020.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Christopher JL Murray. Forecasting covid-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv*, 2020. doi: 10.1101/2020.03.27.20043752.

Rebecca J Rockett, Alicia Arnott, Connie Lam, Rosemarie Sadsad, Verlaine Timms, Karen-Ann Gray, John-Sebastian Eden, Sheryl Chang, Mailie Gall, Jenny Draper, et al. Revealing covid-19 transmission in australia by sars-cov-2 genome sequencing and agent-based modeling. *Nature medicine*, 26(9):1398–1404, 2020.

Mauricio Santillana, Ashleigh Tuite, Tahmina Nasserie, Paul Fine, David Champredon, Leonid Chindelevitch, Jonathan Dushoff, and David Fisman. Relatedness of the incidence decay with exponential adjustment (idea) model,"farr's law" and sir compartmental difference equation models. *Infectious disease modelling*, 3:1–12, 2018.

Farah Shahid, Aneela Zameer, and Muhammad Muneeb. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos, Solitons & Fractals*, 140:110212, 2020.