Estimating Undetected COVID-19 Infections—The Case of North Carolina

Gregory Brown*, Eric Ghysels†, Lu Yi‡

August 10, 2020

Abstract

We specify and estimate a time-varying Markov model of COVID-19 cases in North Carolina. We find that the estimated level of undetected infections spiked in early March and remained elevated through May. However, since late April estimated undetected infections have generally declined as a fraction of all cases though it was not until mid-May that detected cases exceeded the estimated number of undetected cases. Our results suggest that the substantial increase in testing capacity over the last few months in North Carolina has identified a higher percentage of infections. However, these findings also indicate that much of the increase in the number of positive tests in June and July represents a true increase in new cases as opposed to an increase resulting from more testing. One shortcoming of our analysis is that we are not able to condition our estimates on the age of people infected or hospitalized which may cause us to underestimate the current number of undetected cases.

*Prof. Gegory W. Brown, Sarah Graham Kenan Distinguished Professor of Finance and Executive Director, Frank H. Kenan Institute of Private Enterprise, The University of North Carolina at Chapel Hill. Corresponding Author: gregwbrown@unc.edu

†Prof. Eric Ghysels, Edward M. Bernstein Distinguished Professor of Economics,Professor of Finance, Kenan-Flager Business School and Faculty Research Director, Rethinc.Labs - Frank H. Kenan Institute of Private Enterprise, The University of North Carolina at Chapel Hill.

‡Lu Yi, Ph.D. Student in Economics, The University of North Carolina at Chapel Hill.

1 Introduction and Summary of Findings

One of the challenges facing policymakers, business leaders, and the general public in understanding the spread of COVID-19 is the fact that many cases go undetected because of testing shortages or infected individuals not seeking a test (e.g., asymptomatic individuals may not even consider the need for a test). Having an accurate estimate of undetected infections can help planners make decisions about testing policy and economic openness, let business leaders better understand risks to their workers and customers, and inform economic projections.

The state of North Carolina (NC) provides an interesting case study in infection modeling because the number of positive tests has grown steadily faster than the number of hospitalizations. Likewise, hospitalizations in NC have grown more quickly than deaths attributed to COVID-19. A very simple way to understand the disconnect between deaths and reported new cases is to estimate the total number of cases statewide using lagged data on the number of deaths and recent estimates for infection fatality rates (see Meyerowitz-Katz and Merone (2020)). Figure 1 shows that these "death-implied" estimates suggest that the number of cases in NC rose rapidly in March and April and then levelled off. This is obviously at odds with the number of new positive tests which was quite low in March and April and has only recently approached the number of death-implied cases.

Of course, much better models have been proposed for estimating the number of undetected infections. In this analysis, we examine a standard 5-state time-varying Markov model based on Gourieroux and Jasiak (2020) (and cites therein) and apply it to data in North Carolina. In our model the population is either susceptible (S), infected and undetected (IU), infected and detected (ID), hospitalized (H), or deceased (D). Recovered cases re-enter the susceptible pool. States are mutually exclusive so we track hospitalized separately from infected and detected. As conditioning variables in our analysis we include both the testing positivity rate and the intensity of testing (i.e., tests conducted per 100,000) and find that these are important factors in the estimation with intuitive relations to infection probabilities. The model provides estimates of undetected infections that are plausible and the model fits observed levels of positive cases, hospitalizations and deaths very well. We examine two different versions of the model and obtain very similar results from each.

We find that the estimated level of IU grew rapidly in early March. Estimated

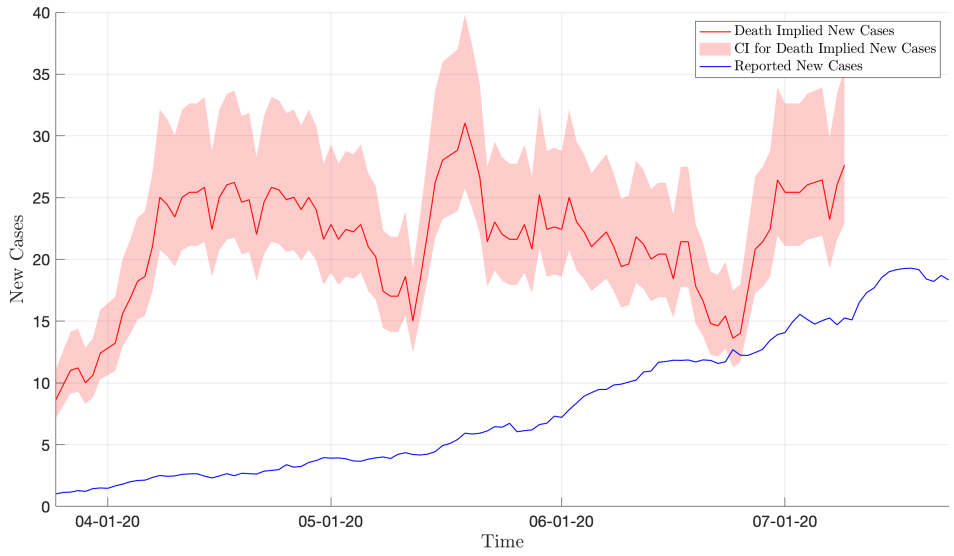

Figure 1: **New Cases in NC (per 100,000, 7-day moving average).**
The red line in the figure shows the death implied new cases, which are calculated using the 7-day average of new reported deaths in NC lagged by 14 days (to reflect the average time between contracting COVID-19 and death) divided by the infection fatality rate of 0.68% estimated by Meyerowitz-Katz and Merone (2020). The confidence band is calculated using the 7-day average of new reported deaths in NC lagged by 14 days divided by the 95% confidence interval of the estimated infection fatality rate. The blue line in the figure shows the reported new cases.

IU cases tapered off in mid-March and then grew in early April before peaking again in mid-April. Since late April, estimated *IU* has declined substantially but according to our estimates it was not until mid-May that the number of detected cases exceeded the number of undetected cases. Our results suggest that the substantial increase in testing capacity over the last few months in North Carolina has been successful in identifying a higher percentage of infections. However, it also suggests that much of the recent increase in the number of positive tests represents a true increase in new cases as opposed to an increase resulting from more testing. One concern about our analysis is that we are not able to condition on the age of those with detected cases or who are hospitalized, and consequently, we may underestimate undetected cases if the age of those infected is declining on average. Our estimates could also underestimate cases if the quality of care has improved over time and reduced hospitalization and death rates in a way the model does not capture.

2 Model

The latent individual history variable $Y_{i,t}$, for individual $i = 1, ..., N$ at time $t = 1, ..., T$, is qualitative polytomous with J alternatives denoted by $j = 1, ..., J$. As in Gourieroux and Jasiak (2020), we assume that $Y_{i,t}$ have the same marginal distribution for all individuals $i = 1, ..., N$ at t fixed, which can be summarized by the J-dimensional vector $p(t)$. The j-th component of the marginal distribution is

$$p_j(t) = P(Y_{i,t} = j).$$

In addition, the individual history variable follows a Markov process with time-varying transition matrix $P\left[p(t-1); \theta\right]$, which gives

$$p(t) = P[p(t-1); \theta]' p(t-1), t = 2, ..., T,$$

with θ being a vector of parameters.

The data of individual histories may not be available in practice. With the assumptions of independent individual histories and homogeneous population of risks, the J-dimensional cross-sectional frequency vector $f(t)$, where $f_j(t)$ is the state j frequency of the population, can be seen as the sample counterpart of $p(t)$. However, the cross-sectional frequencies are only partially observed. A state aggregation matrix A is used to account for the unobserved states and the observations are $\hat{A}_t = Af(t)$ for $t = 1, ..., T$, where A is a $K \times J$ matrix of full rank K. The parameters of interest, θ and the sequence of the unobserved component of $p(t)$, can then be estimated by solving the following optimization problem,

$$\begin{aligned} (\hat{p}(1), ..., \hat{p}(T), \hat{\theta}) = \text{argmin} \sum_{t=2}^{T} \left\| p(t) - P\left[p(t-1), \theta\right]' p(t-1) \right\|_2^2 \qquad (1) \\ \text{s.t. } Ap(t) = Af(t) = \hat{A}_t, t = 1, ..., T, \end{aligned}$$

where $\|.\|_2$ denotes the Euclidean norm.

To model the COVID-19 propagation, we consider a Markov process with 5 states: $1 = S$, for susceptible, $2 = IU$, for Infected and Undetected, $3 = ID$, for Infected and Detected, $4 = H$ for Hospitalized, and $5 = D$ for Deceased. The sum of the

frequencies across all the five states equals to the size of the population. For simplicity, we assume no immunity, hence the recovered cases re-enter the susceptible pool. This assumption lets us avoid having an unobservable recovered state but will have little impact on estimation for low levels of overall infection.

The transition matrix $P[p(t-1);\theta]$ of the Markov process is defined as

$$
\begin{array}{c} \\ \text{S} \\ \text{IU} \\ \text{ID} \\ \text{H} \\ \text{D} \end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{S} & \text{IU} & \text{ID} & \text{H} & \text{D} \end{array} \\
\begin{bmatrix}
1-p_i & p_i(1-p_d) & p_i p_d & 0 & 0 \\
p_{21} & (1-p_{21}-p_{24})(1-p_d) & (1-p_{21}-p_{24})p_d & p_{24} & 0 \\
p_{31} & 0 & 1-p_{31}-p_{34} & p_{34} & 0 \\
p_{41} & 0 & 0 & 1-p_{41}-p_{45} & p_{45} \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
$$

with

$$
\begin{aligned}
p_i &= \text{logist}\left(a_1 + a_2\left(p_2(t-1) + p_3(t-1)\right)\right) + a_3 x_t), \\
p_d &= \text{logist}\left(b_1 + b_2 y_t\right),
\end{aligned}
$$

where $\text{logist}(x) = 1/[1+\exp(-x)]$ is the logistic function, i.e. the inverse of the logit function. The probability of infected p_i follows a multinomial logit model for the competing propagation driven by lagged IU and lagged ID, and it also depends on the testing positivity rate x_t. Conditioning on being infected, the probability of being detected p_d is a function of testing intensity y_t. Each row of the transition matrix sums to one by construction. The structure of zeros indicates that one cannot go backward from ID to IU, patients who died are hospitalized before death, the hospitalized patients will stay in hospital until they recover or die, and death is considered an absorbing state.

In addition, we consider two model specifications for the transition probabilities from state IU and ID to state H. The basic specification assumes constant transition probabilities p_{24} and p_{34}. In this model, there are 11 parameters in $\theta = [a_1, a_2, a_3, b_1, b_2, p_{21}, p_{24}, p_{31}, p_{34}, p_{41}, p_{45}]'$. The full specification assumes time-varying transition probabilities driven by the lagged frequency of the corresponding state with

$$
\begin{aligned}
p_{24} &= \text{logist}\left(c_1 + c_2 p_2(t-1)\right), \\
p_{34} &= \text{logist}\left(d_1 + d_2 p_3(t-1)\right),
\end{aligned}
$$

in which $\theta = [a_1, a_2, a_3, b_1, b_2, c_1, c_2, d_1, d_2, p_{21}, p_{31}, p_{41}, p_{45}]'$ has 13 parameters.

Empirically, $IU(t)$ and $ID(t)$ represent the state of currently infected excluding those hospitalized. The frequency of $ID(t)$ is observable by assumption, while $IU(t)$ is the unobserved state of unidentified infections and will be considered as additional quantities of interest to be estimated jointly. Also, the frequencies of $H(t)$ and $D(t)$ are both observable. Therefore, we have the state aggregation matrix A expressed as

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

3 Data, Estimation and Results

We estimate the two specifications of the time-varying Markov model on Covid-19 propagation data from North Carolina over the period of 133 days between March 4 to August 7, 2020. We use the daily data reported by The Covid Tracking Project. The frequency of $ID(t)$ is measured by the rolling 2-week sum of the new positive tests in NC, which assumes that a person with positive test will either be hospitalized or recover within 14 days. The frequency of $H(t)$ is the actual number of hospitalized in NC on any given date and the frequency of the absorbing state $D(t)$ is measured by the cumulative deaths caused by COVID-19 in NC. In constructing the cross-sectional frequency vector $f(t)$, we do everything in per $100,000$ population to facilitate interpretation as well as comparison to estimated infection rates from other geographies. The daily evolutions of the observed components of $f(t)$ are displayed in Figure 2.

For the two conditioning variables, the test positivity rate x_t is measured by the weekly moving average of the testing positivity rate (i.e., out of all tests) and the test intensity y_t is measured by the rolling 7-day average of tests per day per $100,000$ population as of date t. Figure 3 shows the plots of these two conditioning variables.

The initial frequency is set equal to $100,000$ for state $S(0)$ and 0 for all other states. The model parameters θ and the series of frequencies of the unobserved state $IU(t)$ are then estimated by solving the optimization problem in Equation (1) numerically using the $fminsearch$ function in Matlab. The estimates of the parameters for the basic model are provided in Table 1. Table 2 shows the estimates of the full model.

The mean fitted values are within 7.40% of observed values for the basic model and within 7.21% of observed values for the full model. The comparisons of fitted and observed frequencies for state ID, state H and state D are shown in Figure 5 and 6 in an appendix. We see that the estimated frequencies track the observations closely for both the basic model and the full model.

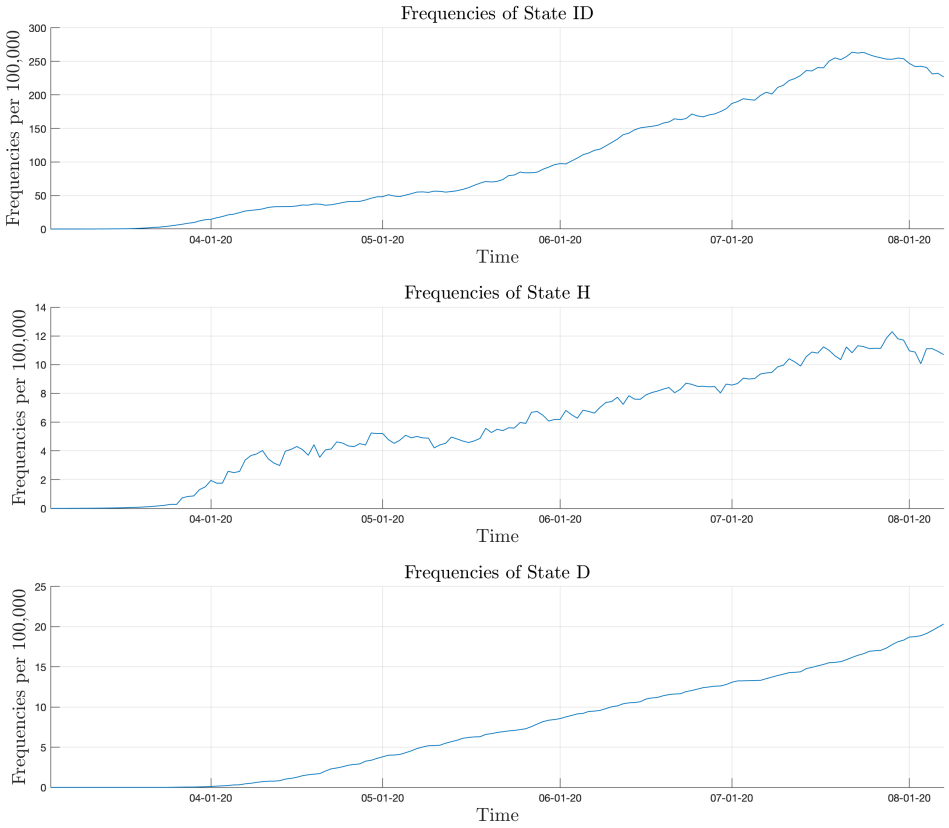

Figure 2: **The Frequencies of the Observed States in NC (per 100,000).** The top panel shows the time series data for the frequency of state $ID(t)$, which are measured by the rolling 2-week sum of the new positive tests in NC. The middle panel shows the time series data for the frequencies of state $H(t)$, which is measured by the actual number of hospitalized in NC at date t. The bottom panel shows the time series data for the frequencies of state $D(t)$, which is measured by the sum of deaths caused by COVID-19 in NC up until date t.

In Table 1, $p_{21} = 0.1516$, which corresponds to an average recovery time of 1 week for state IU, and $p_{31} = 0.0672$, which represents an average recovery time around 2 weeks for state ID. Thus the model estimates that it takes longer for a patient in the detected state to recover which is reasonable considering it is more likely that patients

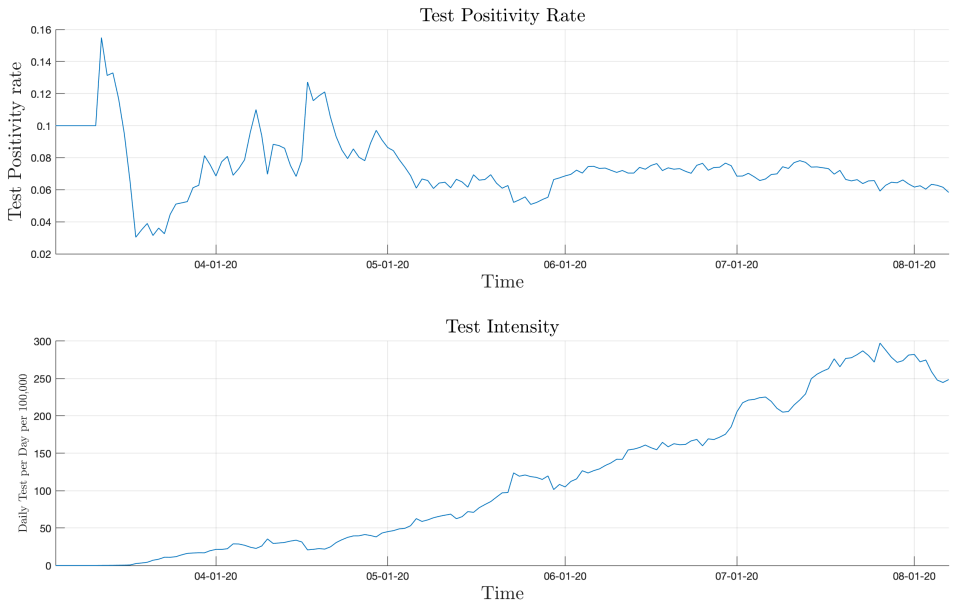

Figure 3: **The Conditioning Variables (x_t and y_t).**
The top panel shows the time series data of the test positivity rates x_t, which is measured by the weekly moving average of the rates of positivity in testing (i.e., out of all tests). The bottom panel shows the time series data of test intensity y_t, which is measured by the rolling 7-day average of tests per day per $100,000$ population as of date t.

with severe cases will get tested (and be detected) thus the overall health condition of state ID is worse than state IU. This is also consistent with the estimated transition probabilities to state H. The probability of transition to state H is 0.0035 from state ID, which is higher than the probability of 0.0002 from state IU. The estimates of p_{33} is 0.9293, which means that people stay in the state ID for an average around 15 days and are then either hospitalized or recover. This is roughly consistent with how we construct the variable representing state ID (i.e. rolling 2-week sum of the positive tests). The mortality rate conditional on being hospitalized is 2.02%, which is higher than the estimated value of 0.68% for the overall infection-fatality rate of COVID-19 in Meyerowitz-Katz and Merone (2020). This is not surprising considering that the severity of the illness is higher for the hospitalized patients than the average severity of all cases. We have positive estimates for both a_2 and a_3 meaning that the probability of being infected, p_i, is increasing with a higher (lagged) frequency of infections and a higher positivity rate of tests. The estimate of b_2 is also positive, which means that if a person is infected, the probability of being detected, p_d, is

increasing with the intensity of testing.

a_1	a_2	a_3	b_1	b_2
-10.1303	0.0023	14.1249	-4.2599	0.0239

	$1 = S$	$2 = IU$	$3 = ID$	$4 = H$	$5 = D$
$2 = IU$	0.1516	Time-varying	Time-varying	0.0002	0
$3 = ID$	0.0672	0	0.9293	0.0035	0
$4 = H$	0.0448	0	0	0.9350	0.0202
$5 = D$	0	0	0	0	1

Table 1: Parameter Estimates of the Basic Model

The estimates of the constant transition probabilities for the full model in Table 2 are very similar to the estimated parameters of the basic model except for a slightly lower rate of recovery given hospitalization. The constant components of p_{24} is very close to zero and it is increasing with the lagged frequency of state IU. The coefficients of the time-varying component of p_{34} is very close to zero, while the constant components are equal to 0.0034, which are similar to the estimates of the p_{34} in the basic model.The estimate of d_2 is negative, meaning that the rate of hospitalization given state ID is decreasing with the lagged frequency of state ID. Considering the increasing frequency of state ID likely contributes to the increasing test intensity and the higher rate of detection, it is reasonable that more cases with light symptoms are detected, thus the hospitalization rate of state ID decreases because of the decreasing proportion of severe cases among all detected cases.

a_1	a_2	a_3	b_1	b_2
-10.1097	0.0018	14.5785	-4.3008	0.0236
c_1	c_2	d_1	d_2	
-35.2790	0.0027	-5.6830	-0.0007	

	$1 = S$	$2 = IU$	$3 = ID$	$4 = H$	$5 = D$
$2 = IU$	0.1526	Time-varying	Time-varying	Time-varying	0
$3 = ID$	0.0625	0	$1 - p_{31} - p_{34}$	Time-varying	0
$4 = H$	0.0328	0	0	0.9474	0.0198
$5 = D$	0	0	0	0	1

Table 2: Parameter Estimates of the Full Model

The time series of the frequencies of state $IU(t)$ are the quantities of primary interest. Figure 4 shows the estimated frequencies of the state using both the basic

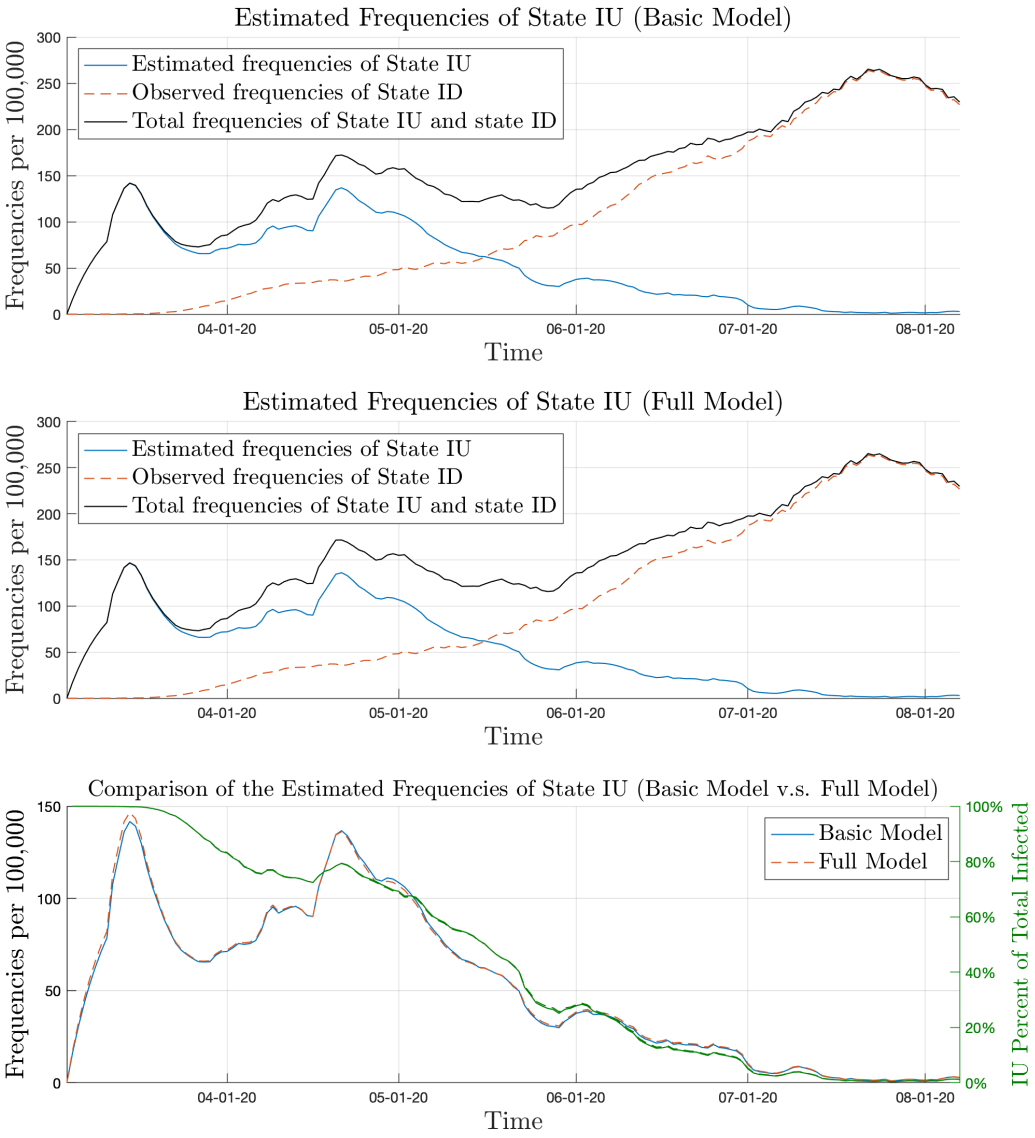

Figure 4: **The Estimated Frequencies of State IU.**
The figure shows the estimated frequencies of state IU for both the basic model (top panel) and the full model (middle panel). In the top two panels, the dashed red line is the observed frequencies of state ID and the solid blue line is the estimated frequencies of state IU. The bottom panel shows the comparison of the estimated frequencies of state IU for the basic model vs. the full model.

model (top panel) and the full model (middle panel). The figure also includes the comparison between the estimates from the two models (bottom panel). We can see that the two lines of estimates are close to each other with the estimates from the full model being slightly higher. From Figure 4, we find that the estimated IU grew rapidly in early March. Estimated IU dipped some in mid-March and then grew slowly until peaking again in mid-April. Since late April, estimated IU has declined substantially but according to our estimates it was not until early-May that the number of detected cases exceeded the number of undetected cases. Our results suggest that the substantial increase in testing capacity over the last few months has been successful in identifying a much higher percentage of infections. However, it also suggests that much of the recent increase in the number of positive tests is in fact an increase in new cases as opposed to an increase related to a higher number of tests.

4 Conclusion

We estimate a model of COVID-19 infections, hospitalizations, recoveries, and deaths. The results of the estimation are intuitive and indicate a high percentage of undetected cases early in our sample period followed by a decline to a much lower percentage of undetected cases by July. Taken at face value, our results suggest that reported cases in North Carolina increasingly reflect the true number of infections in the state. Nonetheless, our model is fairly simple and estimated on aggregate data for just one state. Future work may estimate the model using data from other states (or countries) to obtain better estimates of model parameters. Likewise, it may be possible to estimate this model by region in North Carolina. Given anecdotal evidence that age of detected cases is changing through time, the estimation is also likely to benefit by conditioning estimates on other variables such as the average age of hospitalized patients or the average age of those testing positive.

References

Gourieroux, C. and J. Jasiak (2020). Time varying markov process with partially observed aggregate data; an application to coronavirus.

Meyerowitz-Katz, G. and L. Merone (2020). A systematic review and meta-analysis of published research data on covid-19 infection-fatality rates. *medRxiv*.

Appendix

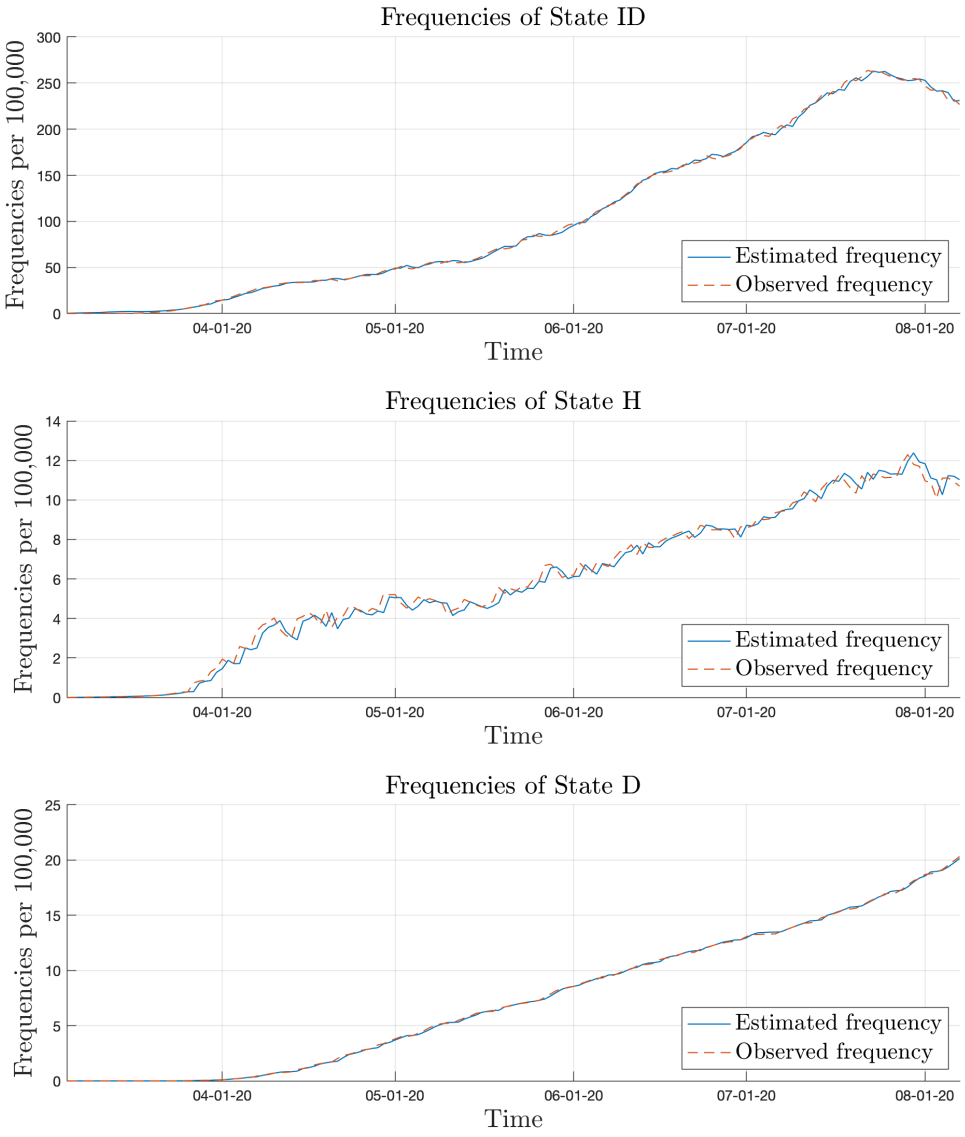

Figure 5: **The Observed and Estimated Frequencies (Basic Model).**
The figure compares the observed frequencies with the estimated frequencies of the basic model for state ID (top panel), state H (middle panel) and state D (bottom panel). The dashed red line is the observed frequencies and the solid blue line is the estimated frequencies.

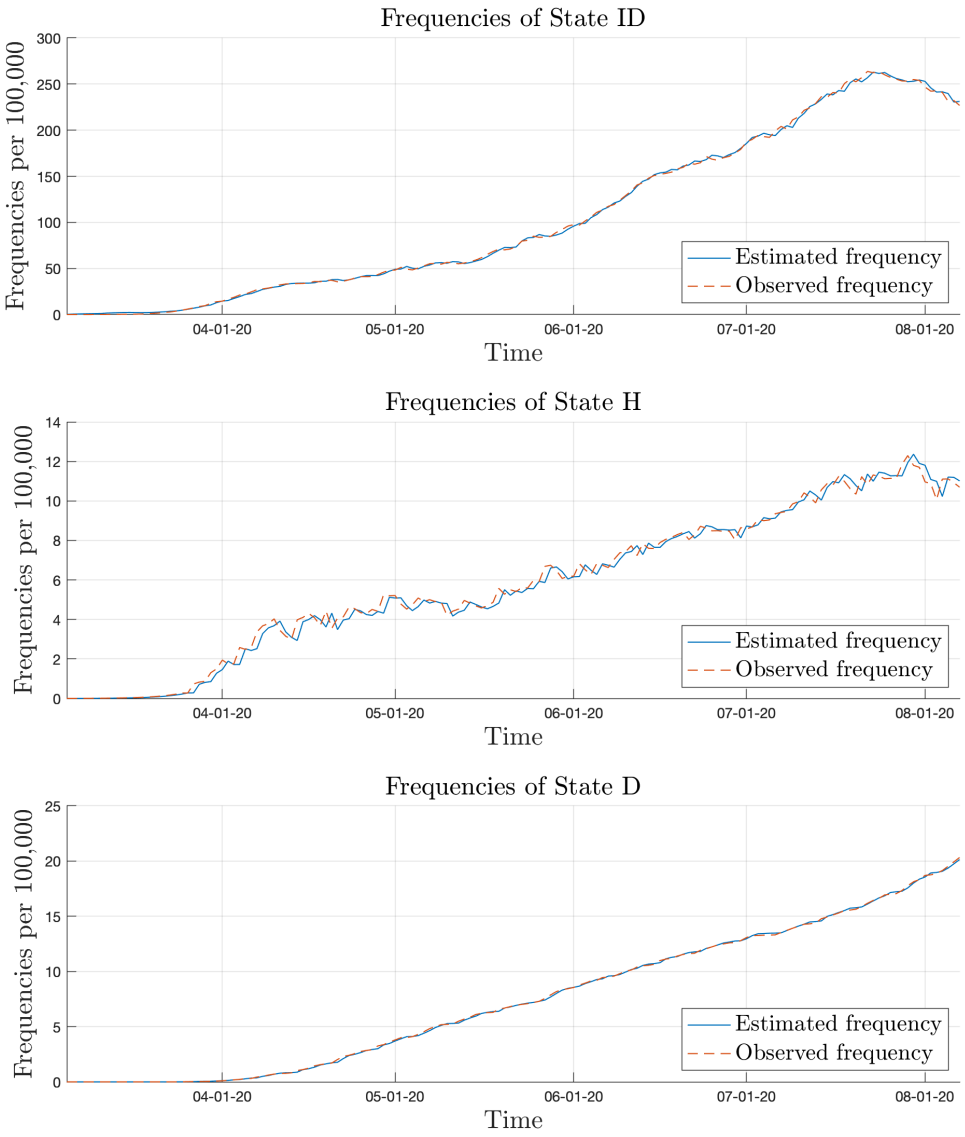

Figure 6: **The Observed and Estimated Frequencies (Full Model).**
The figure compares the observed frequencies with the estimated frequencies of the full model for state ID (top panel), state H (middle panel) and state D (bottom panel). The dashed red line is the observed frequencies and the solid blue line is the estimated frequencies.