Relative Contagiousness of Emerging Virus Variants:

An Analysis of SARS-CoV-2 Alpha and Delta Variants*

Peter Reinhard Hansen[a,b]

[a] *University of North Carolina*[†]

[b] *Copenhagen Business School*

October 4, 2021

Abstract

We propose a simple dynamic model for estimating the relative contagiousness of two virus variants. Maximum likelihood estimation and inference is conveniently invariant to variation in the total number of cases over the sample period and can be expressed as a logistic regression. Using weekly Danish data we estimate the Alpha variant of SARS-CoV-2 to increase the reproduction number by a factor of 1.51 [CI 95%: 1.50, 1.53] relative to the ancestral variant. The Delta variant increases the reproduction number by a factor of 2.17 [CI 95%: 1.99, 2.36] relative to the Alpha variant and a factor of 3.28 [CI 95%: 3.01, 3.58] relative to the ancestral variant. Forecasting the proportion of an emerging virus variant is straight forward and we proceed to show how the effective reproduction number for the new variant can be estimated without contemporary sequencing results. This is useful for assessing the state of the pandemic in real time as we illustrate empirically with the inferred effective reproduction number for the Alpha variant.

Keywords: Covid-19, SARS-CoV-2, Reproduction number, Alpha variant, Delta variant, B.1.1.7, B.1.617.2, Maximum Likelihood, Logistic Regression.

*Thanks to Peter Dalgaard, Emily Dyckman, Claus Ekstrøm, Mogens Fosgerau, Ulrik Gerdes, Martin Vinæs Larsen, Uffe Poulsen, Tom Wenseleers for valuable comments and suggestions. I also thank David Sanders and Michael Krabbe Borregaard for making tutorials and information about the Julia language available and the Danish Patient Safety Authority for sharing data on COVID-19 cases in relation to Euro 2020 games.

[†]Address: University of North Carolina, Department of Economics, 107 Gardner Hall Chapel Hill, NC 27599-3305

1 Introduction

During the fall of 2020, confirmed cases of COVID-19 grew rapidly in the UK with the emergence of the Alpha variant of SARS-CoV-2 (B.1.1.7) formerly known as the British variant, see Rambaut et al. (2020). The Alpha variant was shown to be more contagious than earlier lineages, see Volz et al. (2021) and Washington et al. (2021). Moreover, infection with the Alpha variant was found to increase the risk of hospitalization by about 42%, see Bager et al. (2021). India experienced a similar explosive growth in COVID-19 cases in April 2021 following the emergence of the Delta variant (B.1.617.2), formerly known as the Indian variant. This variant is estimated to increase the risk of hospitalization by about 85%, see Sheikh et al. (2021).

In this paper, we formulate a simple model for two virus variants of an infections disease, where the object of interest is the relative contagiousness, denoted by γ. The time series of new-variant cases to total cases can be modeled as binomially distributed variables with a time-varying parameter, λ_t. The dynamic properties of λ_t are given by the relative contagiousness that can be estimated solely from changes in the relative proportion of the two variant. The analysis is therefore invariant to the reproduction number for the existing lineage and time variation therein. This is convenient because the reproductive number varies substantially over time due to changes in behavior and preventive measures along with other factors. The analysis is also invariant to testing intensity and time variation therein, so long as the variation in sampling does not "favor" one variant over another. Our starting point is maximum likelihood analysis of the sequenced tests, which leads to a simple logistic regression after some straight forward algebra. This greatly simplifies the estimation, inference, and prediction.

The *basic reproduction number*, R_0, is an important characteristic of an infections disease and during a pandemic the *effective reproduction number,* R_t, provides a measure of the direction for cases numbers during the pandemic. The time it takes to determine the genome in positive cases present an obstacle for assessing the reproduction number for a new virus variant. Such data are typically only available with some delay. We show that the highly predictable nature of the proportion of a new variant can be used to infer its reproduction number from the aggregate reproduction number and the most recently available estimate of the proportion, λ. The effective reproduction number for all cases is simpler to compute and the proportion of a new variant can be projected forward, typically with high accuracy. This makes it possible to compute the effective reproduction number of a new variant with a simple formula before contemporaneous sequencing results are available.

We apply the methodology to weekly Danish data using two sample periods. The first sample period, November 9, 2020 to March 14, 2021 (18 weeks), is the period where the Alpha variant made

its inroad in Denmark and the second sample period, May 17 to July 25, 2021 (10 weeks), is the period where the Delta variant grew to dominance in Denmark. The Danish data are excellent for studying the progression of a new variant of SARS-COV-2 because the vast majority of confirmed COVID-19 cases are being sequenced. Moreover, testing is extensive in Denmark. The number of weekly PCR tests varied between nearly 500 thousands and over 1 million tests per week during the two sample periods for a population of about 5.8 million individuals. The proportion of the new-variant cases increased from <0.5% to over 90% for both variants during their respective sample periods, see Figure 1. The progression of the Alpha variant is shown in the left panel Figure 1 and the progression of the Delta variant in the right panel, along with 95% confidence intervals for each week. It can be seen that the Delta variant progressed substantially faster than the Alpha variant. For instance, it took the Alpha variant about eight weeks to increase from <10% to >90%, while this same growth was achieved by the Delta variant in just 4 weeks.

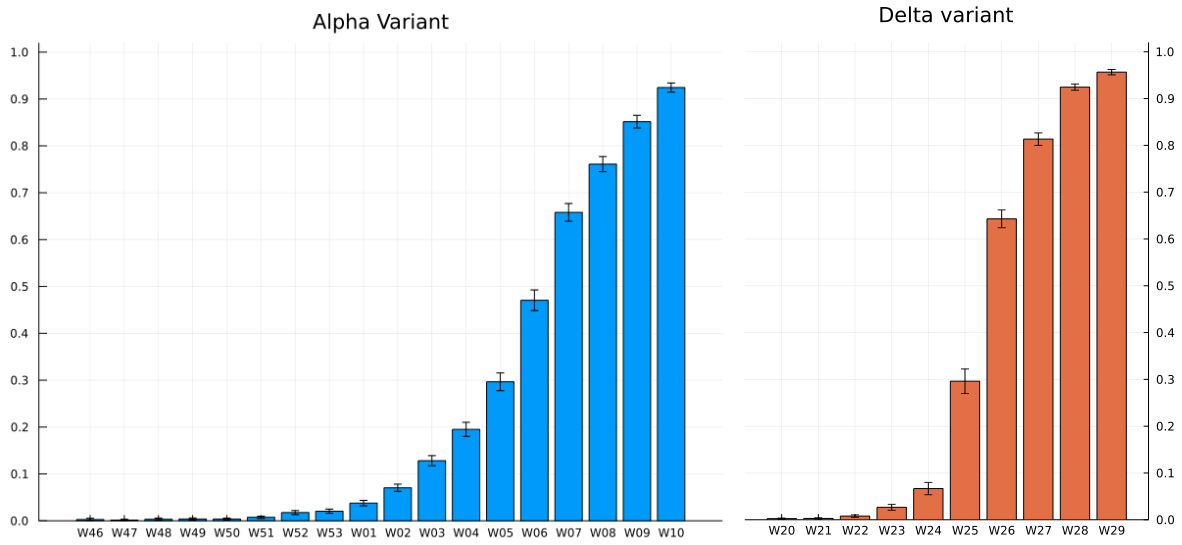

Figure 1: Weekly proportions of the Alpha variant and Delta variant relative to all cases.

The paper is organized as follows: We present the statistical model in Section 2, where the time series of binomially distributed variables is expressed as a logistic regression. We present the data and the empirical results on relative contagiousness in Section 3. Section 4 presents the two auxiliary results on prediction and inferring the latent reproductive number for a new emerging variant. We illustrate how the proportion of new-variant cases can be predicted and derive the associated confidence bands. We apply this to the Alpha variant data and review how accurate the estimated model was at predicting the realized proportion of the Alpha variant out-of-sample. Then we derive the method for estimating the latent reproductive number for a new emerging variant and apply the method to the Alpha variant data.

Section 5 has some concluding remarks, and we present some details about the maximum likelihood estimation and robust standard errors in the Appendix.

2 The Statistical Model

In this section we present the simple dynamic structure for the case numbers of two competing virus variants. The structure is not specific to competing virus variants, but could be used to analyze other competing objects.

2.1 Two Competing Virus Variants

Consider a virus with two variants, A and B, where the number of new cases in period t are denoted A_t and B_t, respectively. We use B to represent a new, emerging variant whereas A represents the older variant. The rate of growth in case numbers for the old variant is denoted, $a_t = A_t/A_{t-1}$, which depends on its contagiousness and the number of "opportunities" the virus has to jump from an infected individual to another person. The latter is heavily influenced by preventive measures and restrictions imposed by health authorities, seasonality, percentage of susceptible people in the population, individual behavior, along with many other things. The new virus variant is subject to the same level of "opportunities", but differs in terms of its contagiousness. Thus, its rate of growth is proportional to a_t, i.e.

$$b_t = B_t/B_{t-1} = \gamma a_t,$$

where γ is the parameter of interest that captures the relative contagiousness of variant B to variant A.

In period t, the genome is determined in N_t of the new cases, where X_t are variant B and $N_t - X_t$ are variant A. We assume N_t is a representative random sample from the population of new cases such that $X_t \sim \text{bin}(N_t, \lambda_t)$, with $\lambda_t = \frac{B_t}{A_t+B_t}$. It follows that

$$\begin{aligned} \lambda_{t+1} = \frac{B_{t+1}}{A_{t+1}+B_{t+1}} &= \frac{\gamma a_{t+1} B_t}{a_{t+1}A_t + \gamma a_{t+1} B_t} \\ &= \frac{\gamma B_t/(A_t+B_t)}{A_t/(A_t+B_t) + \gamma B_t(A_t+B_t)} = \frac{\gamma \lambda_t}{(1-\lambda_t)+\gamma\lambda_t}. \end{aligned} \tag{1}$$

Note that the dynamic equation for λ_t depends on the ratio $\gamma = b_t/a_t$ but not the actual values of a_t and b_t. This greatly simplifies the analysis and estimation and inference about γ becomes invariant to a range of changes during the sample period that influence the rate of change in case numbers. Equation

(1) defines the function $f(x) = \gamma x/[1 + (\gamma - 1)x]$, which is strictly increasing for $\gamma > 1$. Figure 2 presents two examples of $f(x)$. The case $\gamma = 1.86$ is shown in the left panel and $\gamma = 3.16$ in the right panel along with the progression of the weekly observed proportions of the two variants.

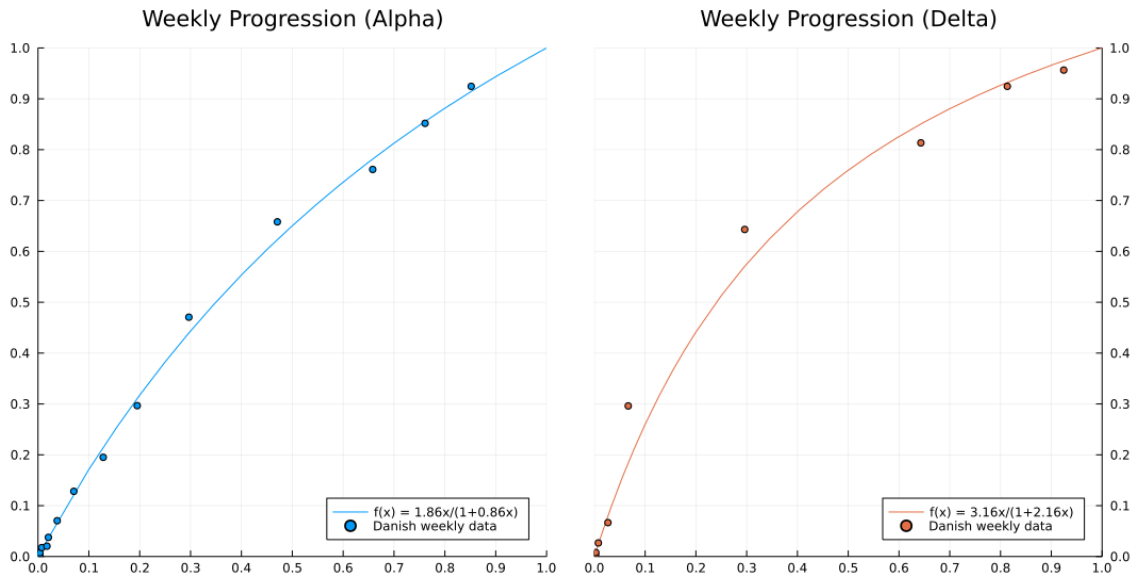

Figure 2: Scatterplot of the empirical proportion of the New Variant this week against that of the previous week. The solid line is $f(x) = \gamma x/[1 + (\gamma - 1)x]$ with $\gamma = 1.86$ for the Alpha variant in the left panel and $\gamma = 3.16$ for the Delta variant in the right panel.

The progression of a far more contagious variant can seem deceptively slow in the early phase. If we take the case $\gamma = 1.86$, which is our estimate for the Alpha variant, then it takes four weeks for the new variant to increase from 1 in 1,000 cases to 1 in a 100 cases. Then another four weeks to increase to 1 in 10 cases. After that, it picks up the pace and the new variant becomes the dominant variant ($\lambda_t > 50\%$) four weeks later and reaches +90% of all cases after another four weeks. So, it can take several months from the moment the first case of a new and more contagious variant is observed to the time when the new variant begins to have a noticeable impact on the total number of cases. Then, in a matter of weeks, it can cause the total number of cases to exhibit explosive growth unless preventive measures are taken. This scenario played out for both the Alpha and Delta variants in many places, such as the August 2021 surge in total cases in Florida, Texas, along with other states in the US.

2.2 The Likelihood Analysis and Logistic Regression

Let N_t denote the total number of new cases in period t for which the genome is identified, and let X_t be the number of cases that are identified as the new emerging variant. The log-likelihood function for

a sample (N_t, X_t), $t = 1, \ldots, T$, is proportional to

$$\ell(\gamma, \lambda_1) \propto \sum_{t=1}^{T} X_t \log \lambda_t + (N_t - X_t) \log(1 - \lambda_t),$$

where λ_t evolves according to (1). The two unknown parameters, the initial value $\lambda_0 \equiv \lambda$ and γ, can be estimated by maximum likelihood, $(\hat{\lambda}, \hat{\gamma}) = \arg\max_{\gamma,\lambda} \ell(\lambda, \gamma)$, and confidence intervals for λ and γ can be obtained with conventional methods. The likelihood can conveniently be expressed as a logistic regression model. For this purpose, we introduce the odds ratio, $\rho_t = \lambda_t/(1 - \lambda_t)$, and it is simple to show that (1) is equivalent to the simple dynamic equation, $\rho_t = \gamma \rho_{t-1}$. This implies that

$$\rho_t = \gamma^t \rho_0 = \exp(\log \rho_0 + \log \gamma \times t) = \exp(\alpha + \beta t),$$

where $\alpha = \log \rho_0$ and $\beta = \log \gamma$. Since $\lambda_t = \rho_t/(1 + \rho_t)$, the structure of the logistic regression model emerges such that

$$\lambda_t = \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)} = \frac{1}{1 + e^{-\alpha - \beta t}}. \tag{2}$$

This model is straight forward to estimate and analyze using standard software implementations, including the generalized linear model package, `glm`, that is implemented in R and Julia. In the empirical analysis we estimate the model by maximum likelihood and compute robust standard errors from the score and hessian of the log-likelihood function, see White (1980). The details are presented in the Appendix.[1]

3 Estimates of Relative Contagiousness for Delta Variants

Weekly data for the sequenced COVID-19 tests were obtained from the Statens Serum Institute, Denmark, and the vast majority of positive COVID-19 tests have their genome identified in Denmark. The weekly numbers positive PCR COVID-19 tests, the number of tests with the genome identified, N_t, and the number of tests for which the new emerging variant was found, X_t, are presented in Table 1 along with the percentages of positive tests for which the genome was determined and the percentage of these tests that were the new variant.

[1] Identical estimates were obtained with the `glm` packaged in Julia, see Besançon et al. (2019) and Lin et al. (2021). The proper command for the `glm` package in Julia is: `glm(@formula(x / n ~ time_trend), [data], wts = n, Binomial())` and in `R` it is: `glm(x/n ~ tt, weights=n, [data], family = binomial)`, see R Core Team (2018) for details. The latter was kindly provided by Peter Dalgaard. The `glm` package computes the non-robust standard errors based on the Fisher information. These were smaller than the robust standard errors, in particular in our analysis of the Delta variant. Robust and non-robust confidence intervals are reported in the Appendix.

Week	Tested (PCR)	Cases C_t	Sequenced N_t (N_t/C_t)	Alpha cases X_t	Alpha proportion X_t/N_t
46	490,543	7,533	1,486 (19.7%)	4	0.27%
47	502,852	8,456	1,941 (23.0%)	3	0.15%
48	502,851	8,774	2,127 (24.2%)	7	0.33%
49	544,578	12,816	2,868 (22.4%)	11	0.38%
50	694,989	21,925	4,226 (19.3%)	16	0.38%
51	883,253	24,579	4,943 (20.1%)	37	0.75%
52	650,374	17,043	3,633 (21.3%)	64	1.76%
53	536,958	14,560	3,916 (26.9%)	80	2.04%
1	563,348	11,311	4,161 (36.8%)	157	3.77%
2	596,048	7,008	4,230 (60.4%)	298	7.04%
3	739,922	5,321	3,688 (69.3%)	473	12.83%
4	768,925	3,616	2,660 (73.6%)	519	19.51%
5	794,917	3,096	2,235 (72.2%)	663	29.66%
6	809,028	2,716	1,974 (72.7%)	929	47.06%
7	833,795	3,335	2,416 (72.4%)	1,590	65.81%
8	956,070	3,688	2,683 (72.7%)	2,042	76.11%
9	1,033,111	3,616	2,699 (74.6%)	2,299	85.18%
10	1,056,404	3,809	2,874 (75.5%)	2,657	92.45%

Week	Tested (PCR)	Cases C_t	Sequenced N_t (N_t/C_t)	Delta cases X_t	Delta proportion X_t/N_t
20	1,167,981	6,867	5,366 (78.1%)	13	0.24%
21	1,013,403	6,698	5,213 (77.8%)	15	0.29%
22	911,764	5,662	4,565 (80.6%)	36	0.79%
23	720,274	2,811	2,467 (87.8%)	66	2.68%
24	575,207	1,649	1,364 (82.7%)	91	6.67%
25	524,837	1,315	1,165 (88.6%)	345	29.61%
26	608,540	2,674	2,418 (90.4%)	1,555	64.31%
27	624,414	4,614	3,322 (72.0%)	2,702	81.34%
28	583,932	6,818	6,253 (91.7%)	5,781	92.45%
29	473,843	5,289	4,800 (90.8%)	4,591	95.65%

Table 1: Weekly Danish data for positive SARS-CoV-2 tests: Cases, sequenced, and Alpha cases. Source: *Status for udvikling af B.1.1.7 og andre mere smitsomme varianter i Danmark*, SSI, April 7, 2021 and *Status for udvikling af SARS-CoV-2 Varianter der overvÃ¥ges i Danmark* SSI, August 27, 2021. Data available at: https://files.ssi.dk/covid19/virusvarianter/status/status-virusvarianter-07042021-dg45 and https://files.ssi.dk/covid19/virusvarianter/status/virusvarianter-covid-19-280721-gd14

A preliminary probing of the data can be done by considering the empirical odds ratios of new-variant cases to old-variant cases. This ratio should be approximately proportional to B_t/A_t, such that the ratio of consecutive odds ratios,

$$\frac{X_t}{N_t - X_t} \Big/ \frac{X_{t-1}}{N_{t-1} - X_{t-1}} = \frac{X_t/X_{t-1}}{(N_t - X_t)/(N_{t-1} - X_{t-1})} \approx b_t/a_t = \gamma.$$

Thus we can use the ratio of consecutive odds ratios as a measurements of γ in week t. These empirical ratios and the corresponding confidence intervals are shown in Figure 3. The crude measures tend to have large confidence intervals early in the sample because the number of new-variant cases is small. The width of the confidence intervals are also influenced by the number of tests that are being sequenced, N_t. For instance, in week 25, this number was relatively small for the simple reason that there were few positive COVID-19 cases in Denmark that week – just 1,315 positive cases of which 1,165 were successfully sequenced. The crude measures for the Alpha variant in the left panel of Figure 3 stabilizes about their average value, 1.73. For the Delta variant, the crude measures are substantially larger and more disperse. The progression of the Delta variant was particularly rapid in weeks 25 and 26.

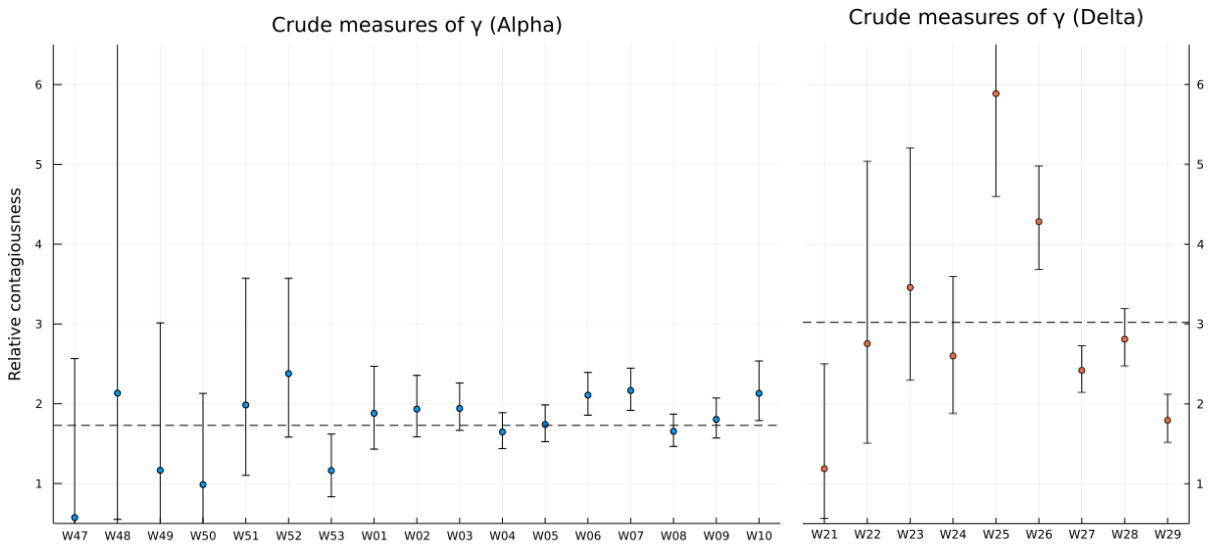

Figure 3: The weekly growth in new-variant cases divided by the weekly growth in old-variant cases, $\frac{X_t}{X_{t-1}} \Big/ \frac{N_t - X_t}{N_{t-1} - X_{t-1}}$, is plotted with 95% confidence intervals. There is more uncertainty associated with the first observations for both the Alpha variant (left panel) and the Delta variant (right panel) because the number of cases of the new variant is quite small in the early phase. The horizontal dotted lines represent the average crude measures: 1.73 in for Alpha and 3.19 for Delta.

The crude measurements of γ in Figure 3 do not fully exploit the information in the data, and the simple sample averages (the dotted lines in Figure 3) do not account for heteroskedasticity and autocorrelation in the measurements errors. To exploit the information in full, we turn to maximum likelihood estimation using the parametrization of the logistic regression. We compute robust standard errors using the Parzen kernel (with bandwidth parameter $K = 4$), see the Appendix. The results are not very sensitive to the choice of bandwidth, but the heteroskedasticity robust standard errors are somewhat larger than the non-robust standard errors, especially for the Delta variant, see Table A.1.

	Alpha vs Ancestral	Delta vs Alpha	Delta vs Ancestral
Per Week			
α	-7.8 [−9.00,−8.50]	-7.81 [−8.75,−6.87]	
β	0.619 [0.601,0.636]	1.152 [1.026,1.278]	
$\gamma_{\text{week}} = \exp(\beta)$	1.86 [1.82,1.89]	3.16 [2.79,3.59]	5.87 [5.17,6.67]
Per Generation (4.7 days)			
$\gamma_{4.7\text{days}} = \exp(\frac{4.7}{7}\beta)$	1.51 [1.50,1.53]	2.17 [1.99,2.36]	3.28 [3.01,3.58]

Table 2: Empirical estimates with 95% confidence intervals computed with robust standard errors.

The maximum likelihood estimates along with 95% confidence intervals are presented in Table 2. The Alpha variant is estimated to be about 86% more contagious per week than the preceding variant, which we refer to as the *ancestral variant.* The Delta variant, which emerged after then Alpha variant had become completely dominant, is estimated to be 216% more contagious than the Alpha variant on a weekly basis. The reproduction number for SARS-CoV-2 is defined for a generation period (the typical time from a person gets infected to the same person infects the next person). For SARS-CoV-2 this period is shorter than a week. The Statens Serum Institut in Denmark use 4.7 days per generation which we adopt in our calculations. We can convert $\hat{\gamma}$ to a period of x days using $\gamma_{x\text{days}} = \exp(\frac{x}{7}\log\gamma_{\text{week}})$, and the estimates for $x = 4.7$ days are presented in the last row of Table 2. The estimates suggest that the Alpha variant has a reproductive number that is about 1.5 times larger than the ancestral variant. The Delta variant is estimated to increase the reproduction number by an additional factor of 2.17, which implies more than a threefold increase relative to the ancestral variant. This is in line with other estimates, which include those for the Alpha variant based on British data by Volz et al. (2021) and those for the Delta variant by Wenseleers (2021). The implication is that it requires a larger proportion $(1 - 1/R_0)$ to be immune to reach *herd immunity.* Suppose that 70% immunity was needed for the ancestral variant. Our estimates of $\gamma_{4.7\text{days}}$ suggest this number increased to about 80% for the Alpha variant and about 90% for the Delta variant.

The estimated model and the observed odds ratios are shown in Figure 4. Overall the model fit looks good, especially for the analysis of the Alpha variant. There are some discrepancies between the data and the linear specification for log odds ratios with the Delta variant. A possible explanation is that many of the COVID-19 cases that were detected in Denmark during the second sample period were contracted abroad. According to the Danish Patient Safety Authority, about 25% of Covid-19

cases were imported cases, primarily by people who had been vacationing in Spain in July.[2] This could potentially influence the progression of the Delta variant because imported cases could be acquired in areas with a higher or a lower Delta proportion than that in Denmark.

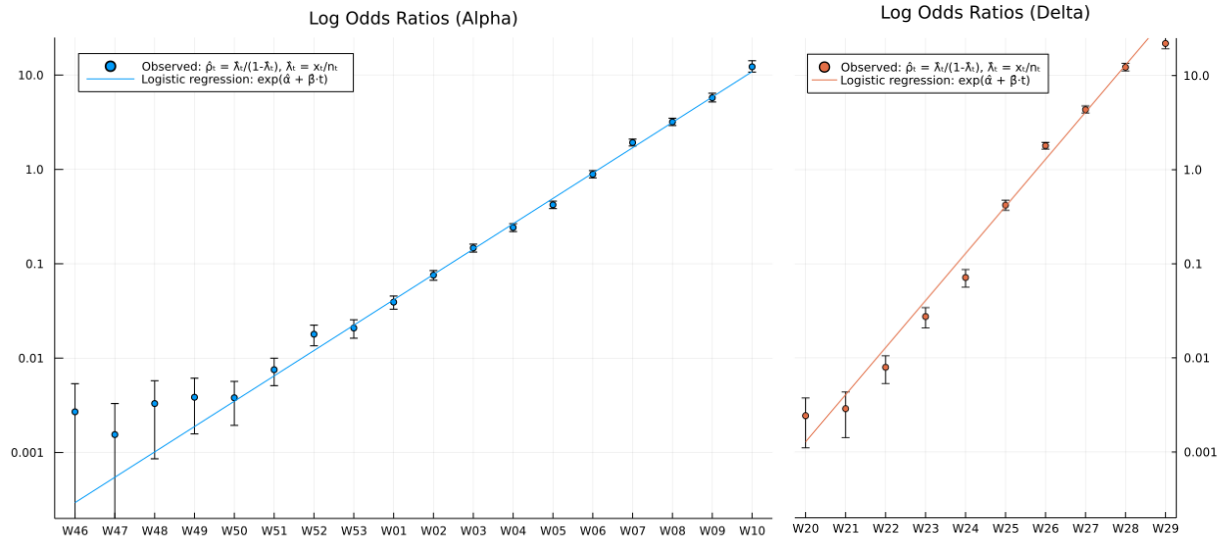

Figure 4: Observed logarithmically transformed odds ratios and the corresponding estimated model, $\hat{\alpha} + \hat{\beta}t$, for the Alpha variant in the left panel and the Delta variant in the right panel.

Another explanation is that the second sample period, where most restrictions were largely abolished, caused the data for the delta variant to be more noisy. During the Alpha sample period restrictions were quite restrictive. In contrast, during the Delta sample period most restrictions were abolished in Denmark, especially in relation to large gatherings. The relaxed restrictions may explain the larger degree of randomness in the progression of the Delta variant. For instance, the Euro 2020 games in Copenhagen may have contributed to the accelerated growth in the Delta variant in Week 25 (see right panel of Figure 4) because spectators at two games accounted for a large fraction of the Delta variant cases. Following the Denmark-Belgium Euro 2020 game in Copenhagen on June 17, 2021, 41 attending spectators tested positive for COVID-19 of which 25 cases (61.0%) were the Delta variant. The following week, on Monday June 21, 2021, Denmark played Russia in Copenhagen at another Euro 2020 game, where 62 cases were subsequently detected among spectators of which 28 (45.2%) were Delta variant cases. These are large numbers and percentages, because the total number of Delta variant cases in Week 24 and Week 25 were 91 and 345, respectively, and Delta variant only accounted for 6.7% in Week 24 and 29.6% in Week 25.

[2]https://www.ssi.dk/aktuelt/nyheder/2021/en-stor-del-af-covid-19-smitten-i-danmark-kommer-fra-de-rejsende

4 Confidence Intervals, Predictions, and Inferring reproduction Number

In this section, we detail detail two ancillary results. First, in Section 4.1, we show the estimated model can be used to predict the proportion of an emerging virus variant and develop methods for quantifying the associated uncertainty. We illustrate these methods with the data for the Alpha variant. Then, in Section 4.2, we develop a simple formula for the reproduction number of the new variant, which does not require concurrent genome data. Instead it projects the most recent estimate of the proportion forward and infer the effective reproduction number from the recent growth in total cases.

4.1 Confidence Sets and Out-of-Sample Analysis

At times T we can estimate α and β, as well as their variance-covariance matrix, $\Sigma_T = \text{var}((\hat{\alpha}_T, \hat{\beta}_T)')$, where $\hat{\alpha}_T$, $\hat{\beta}_T$, and $\hat{\Sigma}_T$ denote the resulting estimates. Point forecasts for the proportion of the new virus variant, λ_t, is given from (2). The h period ahead point forecast, made at time T, is simply

$$\hat{\lambda}_{T+h,T} = 1 \Big/ \Big[1 + \exp\{-\hat{\alpha}_T - \hat{\beta}_T(T+h)\Big] ,$$

and the corresponding confidence bands can be deduced from the asymptotic distribution of $(\hat{\alpha}_T, \hat{\beta}_T)$. The confidence band based on c units of standard deviations is given by

$$1 \Big/ \left[1 + \exp\{-\hat{\alpha}_T - \hat{\beta}_T(T+h) \pm c\sqrt{v(T+h, \hat{\Sigma}_T)}\}\right] , \tag{3}$$

where $c = 1.96$ would correspond to a 95% confidence bands and

$$v(T+h, \hat{\Sigma}_T) \equiv \hat{\text{var}}(\hat{\alpha}_T + \hat{\beta}_T(T+h)) = (1, T+h)\hat{\Sigma}_T(1, T+h)'.$$

The estimated and predicted progression of λ_t for the Alpha variant along with confidence bands (using $c = 2$ and $c = 4$ standard deviations) are presented in Figure 5. The saltires (x-crosses) in Figure 5 are the observed weekly empirical proportion of the Alpha variant.

In the upper left panel of Figure 5, we have estimated the model by maximum likelihood using 4 weeks of data (Week 50-53) which leaves 10 weeks for out-of-sample forecasting. The point forecasts are reasonably close to the realized proportions, but with just four weeks of data for estimation, there is a great deal of uncertainty about the estimated parameters, causing $v(T+h, \hat{\Sigma}_T)$ to be large. With

two additional weeks for estimation (six weeks total), the parameters are more precisely estimated, resulting in tighter confidence bands, as shown in the upper-right panel of Figure 5. With eight or ten weeks for estimation, the parameter estimates become even more accurate, resulting in the even tighter confidence intervals in the two lower panels.

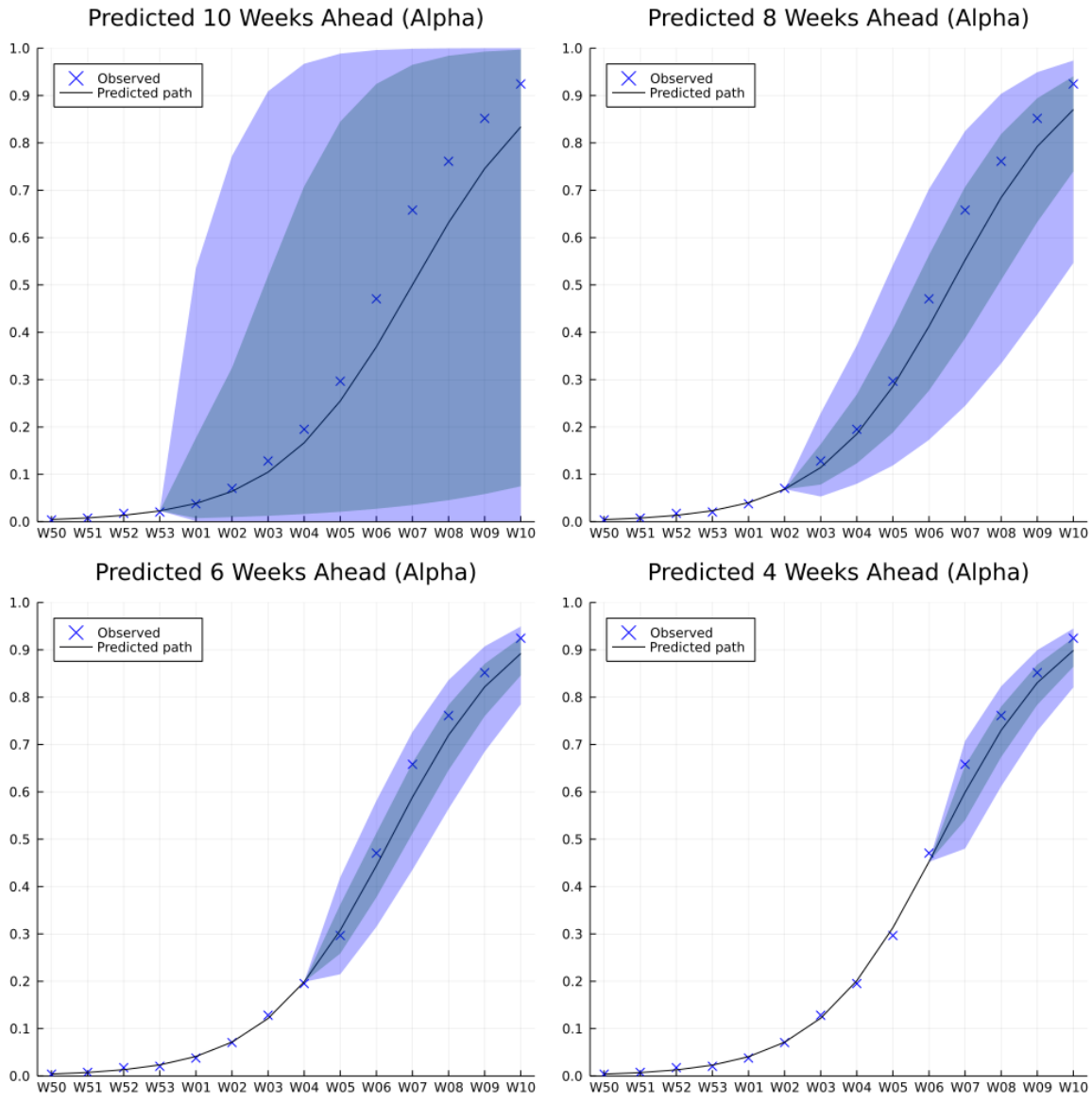

Figure 5: The predicted path for λ_t (solid black line) is shown for when the model is estimated with, 4, 6, 8, and 10 weeks of data, which translates to an out-of-sample period of 10, 8, 6 and 4 weeks, respectively. The shaded areas are the confidence bands using $2\times$ and $4\times$ the standard deviation as defined in (3). The observed proportion of Alpha are indicated with the blue crosses.

The point forecasts are reasonably accurate at horizons up to four weeks, but tend to be below the realized values, especially at longer horizons. This is because the four in-sample estimates of γ ($\hat{\gamma}_{W53} = 1.71$, $\hat{\gamma}_{W02} = 1.76$, $\hat{\gamma}_{W04} = 1.79$, and $\hat{\gamma}_{W06} = 1.81$) are all smaller than the full sample

estimate: $\hat{\gamma} = 1.86$. This highlights that we should expect the out-of-sample forecasting errors to be positively autocorrelated and likely have the same sign as $\gamma - \hat{\gamma}_T$. It should be noted that the confidence bands reflect the uncertainty about λ_{T+h}, while the realized empirical proportions, X_{T+h}/N_{T+h}, are themselves noisy estimates of λ_{T+h}, see the confidence bands in Figure 1.

4.2 Inferring the Reproduction Number for the New Variant in Real Time

We can infer the effective reproduction number for an emerging variant from the effective reproduction number of all cases when combined with knowledge about λ and γ. Let C be the number of all cases in this period, of which $B = \lambda C$ are the new-variant cases and $A = (1-\lambda)C$ are the old-variant cases. If the current reproduction number for all cases is R, then there were C/R cases one generation ago. Similarly, there were $B/R_B = \lambda C/R_B$ new-variant cases and $A/R_A = (1-\lambda)C/R_A$ old-variant cases one generation earlier, where R_A and R_B denote the current reproduction numbers for the old and new variant, respectively. The number of cases for the previous generation have to add up to the total number of cases. Hence, $C/R = \lambda C/R_B + (1-\lambda)C/R_A$, and since $R_A = R_B/\gamma$ it follows that

$$R_B = R_B(\lambda, R, \gamma) = R[\lambda + \gamma(1-\lambda)]. \tag{4}$$

The value of λ to be used in this expression should be the that for the current period, which is typically predicted from earlier periods, and the value of γ to be used in (4), should be the one that corresponds to the same generation period as used to compute R. We estimated $\gamma_{4.7\text{days}} \approx 1.5$ for the Alpha variant and $\gamma_{4.7\text{days}} \approx 2$ for Delta. Thus, based on the Danish data we approximately have,

$$R_{\text{Alpha},T} \approx R_T \times (1.5 - 0.5\lambda_T) \qquad \text{and} \qquad R_{\text{Delta},T} \approx R_T \times (2 - \lambda_T).$$

This formula makes it possible to assess the reproduction number for an emerging variant before concurrent sequencing data are available. The reproduction number, R_T, for all cases can be inferred from the progression in the total number of COVID-19 cases and the proportion of the new variant, λ_T, can be obtained from the estimated model, by projecting forward the most recent knowledge about the proportion, see Figure 4.

4.2.1 Empirical Illustration for the Alpha Variant

We can use (4) to characterize the combinations of (λ, R) that correspond to a particular reproduction number for the Alpha variant. A contour plot for $R_B(\lambda, R)$ based on the point estimate of γ that

corresponds to a generation period, $\hat{\gamma}_{4.7\text{days}} = \exp(\frac{4.7}{7}\log\hat{\gamma})$, is presented in Figure 6. The region above the solid line, $\{(\lambda, c) : R_B(\lambda, R, \hat{\gamma}_{4.7\text{days}}) > 1\}$, are the combinations of λ and R where case numbers for Alpha are increasing, and the region below the solid line is the region where Alpha cases are decreasing. The shaded region about the solid line represent the uncertainty about the threshold, due to uncertainty about γ. The shaded area is given by

$$\{(\lambda, R) : R_B(\lambda, R, \gamma) = 1, \text{ for some } \gamma \in \text{CI}_{95\%}\},$$

where $\text{CI}_{95\%}$ is the 95% confidence interval for $\hat{\gamma}_{4.7\text{days}}$ we obtained in Section 3. Note that the uncertainty interval shrinks to zero as $\lambda \to 1$. The reason is that the limited case, $\lambda = 1$, represents the situation where Alpha cases make up all cases, and its rate of increase can therefore be inferred from the rate of increase in all cases. More formally, the result follows from the fact that $R_B/R = \lambda + \gamma(1-\lambda) \to 1$ as $\lambda \to 1$.

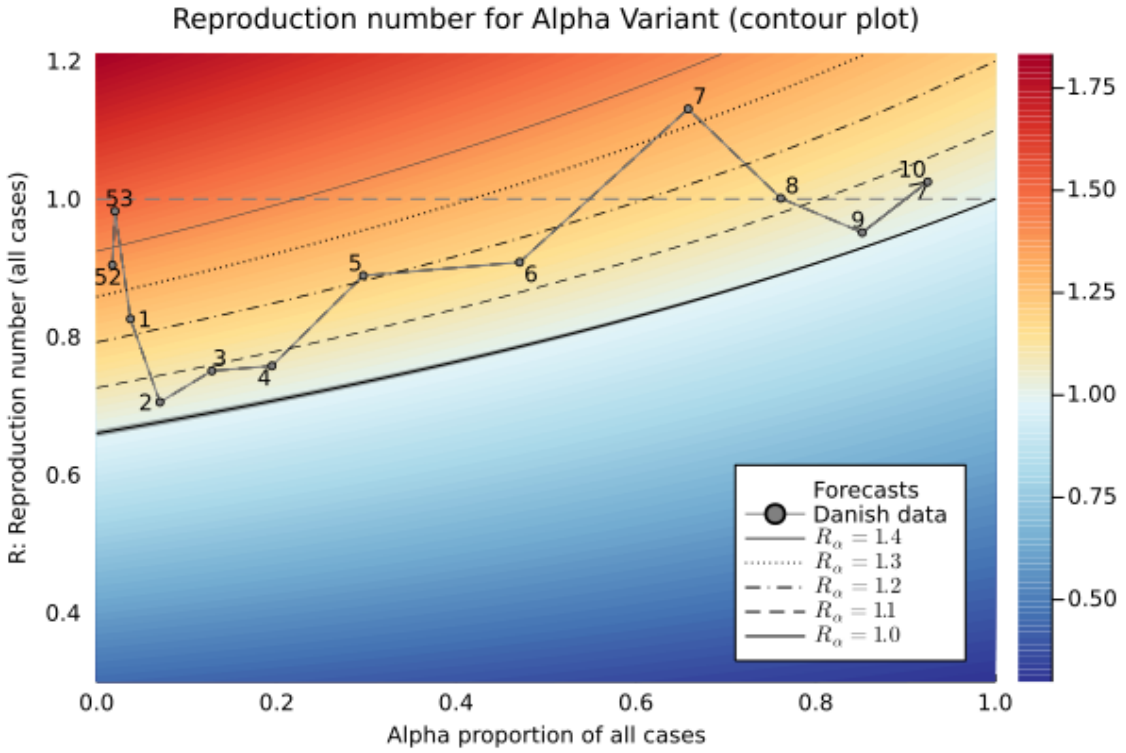

Figure 6: Contour plot for the Alpha variant reproduction number as a function of the proportion of Alpha cases, λ, and the reproduction number for all cases. Alpha cases are increasing above the solid line and decreasing below the solid line. The solid line is based on the estimate of γ and shaded area reflects the statistical uncertainty therein. Danish weekly statistics for (λ_t, R_t) are shown and labelled with the corresponding week number.

A model-free proxy for λ_t is X_t/N_t and a crude estimate of R in week t, is given by $\hat{R}_t \equiv$

$\exp(\frac{4.7}{7}\log\frac{\text{Cases}_t^{Adj}}{\text{Cases}_{t-1}^{Adj}})$, where Cases_t^{Adj} is the number of all cases in week t after adjusting for the testing intensity. The adjustment is given by $\text{Cases}_t^{Adj} = \text{Cases}_t \times \left(\frac{\text{Tested}_t}{M}\right)^{-0.7}$, where M is a baseline number of tests. Statens Serum Institute (2020) The baseline number, M, which does not influence the ratio

$$\frac{\text{Cases}_t^{Adj}}{\text{Cases}_{t-1}^{Adj}} = \frac{\text{Cases}_t}{\text{Cases}_{t-1}} \times \left(\frac{\text{Tested}_t}{\text{Tested}_{t-1}}\right)^{-0.7},$$

and we use this ratio to compute $\hat{R}_t$.

The estimated reproduction number, $\hat{R}_t$, is plotted against the observed proportion of the Alpha variant in Figure 6, labelled with the corresponding week number. All pairs fall above the solid line, where the effective reproduction number for the Alpha variant is greater than one. This indicates that the number of Alpha cases (detected and undetected) was growing throughout the sample period even though the total number of cases was declining most weeks.

5 Discussion

We have shown how the relative contagiousness of a new virus variant can be estimated by maximum likelihood and how robust standard errors can be computed. The underlying structure is that of a logistic regression model. We applied the methodology to weekly Danish data from the periods where the Alpha and Delta variant emerge to become the dominant variants. The methodology can also be applied to data at different frequencies, such as daily data, and to time series with missing data. The analysis can also be extended to situations with more than two competing virus variants and is not specific to the analysis of competing virus variant, but could be applied in a context with other competing objects. We found the Alpha variant increased the contagiousness by about 50% and the Delta variant increased the contagiousness further by more than 100% per generation. To reach herd immunity, it was originally estimated that about 70% of the population needed to be immune, which corresponds to a basic reproductive number equal to $R_0 = 1/(1-0.7) \simeq 3.3$. So, if the Delta variant increases R_0 by a factor of 3, it reduces the fraction of the population that can be without immunity to a third. In this case from 30% to 10%, so that 90% population immunity is needed for herd immunity.

Two new variants of the SARS-CoV-2 have emerged to become dominant in short succession, which suggests that even more contagious variants may emerge in the time to come. Both variants were not only more contagious but were also determined to substantially increase the risk of hospitalization. It is unclear when a more contagious variant will emerge, if at all. It is, however, discomforting that there

were just 18 weeks between the time the Alpha variant made up 90% of all cases to the time the Delta variant surpassed that same threshold. Fortunately, vaccinations have shown to abate transmission and greatly reduce the risk of severe disease. So, vaccination appears to be the most effective measure for slowing the emergence of new variants and preventing new variants from having harmful effects.

References

A. Rambaut, N. Loman, O. Pybus, W. Barclay, J. Barrett, A. Carabelli, T. Connor, T. Peacock, D. L. Robertson, E. Volz, Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations, 2020. URL: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-563.

E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson, Transmission of SARS-CoV-2 lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data, medRxiv (2021). URL: https://www.medrxiv.org/content/early/2021/01/04/2020.12.30.20249034.1. doi:10.1101/2020.12.30.20249034. arXiv:https://www.medrxiv.org/content/early/2021/01/04/2020.12.30.20249034.1.full.pdf.

N. L. Washington, K. Gangavarapu, M. Zeller, A. Bolze, E. T. Cirulli, K. M. Schiabor Barrett, B. B. Larsen, C. Anderson, S. White, T. Cassens, S. Jacobs, G. Levan, J. Nguyen, J. M. Ramirez, C. Rivera-Garcia, E. Sandoval, X. Wang, D. Wong, E. Spencer, R. Robles-Sikisaka, E. Kurzban, L. D. Hughes, X. Deng, C. Wang, V. Servellita, H. Valentine, P. De Hoff, P. Seaver, S. Sathe, K. Gietzen, B. Sickler, J. Antico, K. Hoon, J. Liu, A. Harding, O. Bakhtar, T. Basler, B. Austin, M. Isaksson, P. Febbo, D. Becker, M. Laurent, E. McDonald, G. W. Yeo, R. Knight, L. C. Laurent, E. de Feo, M. Worobey, C. Chiu, M. A. Suchard, J. T. Lu, W. Lee, K. G. Andersen, Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States, medRxiv (2021). URL: https://www.medrxiv.org/content/early/2021/02/07/2021.02.06.21251159. doi:10.1101/2021.02.06.21251159. arXiv:https://www.medrxiv.org/content/early/2021/02/07/2021.02.06.21251159.full.pdf.

P. Bager, J. Wohlfahrt, J. Fonager, M. Albertsen, T. Yssing Michaelsen, C. Holten Mœller, S. Ethelberg, R. Legarth, M. S. Fischer Button, S. M. Gubbels, M. Voldstedlund, K. Mølbak, R. L. Skov, A. Fomsgaard, T. Grove Krause, The Danish Covid-19 Genome Consortium, Increased risk of hospitalisation associated with infection with SARS-CoV-2 lineage B.1.1.7 in Denmark, Lancet (2021). doi:https://doi.org/10.1016/S1473-3099(21)00290-5.

A. Sheikh, J. McMenamin, B. Taylor, C. Robertson, Public Health Scotland, the EAVE II Collaborators, SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness, Lancet (2021) 2461–2462. doi:https://doi.org/10.1016/S0140-6736(21)01358-1.

H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica 48 (1980) 817–838.

M. Besançon, D. Anthoff, A. Arslan, S. Byrne, D. Lin, T. Papamarkou, J. Pearson, Distributions.jl: Definition and modeling of probability distributions in the JuliaStats ecosystem, arXiv e-prints (2019) arXiv:1907.08611. arXiv:1907.08611.

D. Lin, S. Byrne, J. M. White, A. Noack, M. Besançon, D. Bates, D. Widmann, J. Pearson, J. Zito, A. Arslan, K. Squire, M. Schauer, D. Anthoff, T. Papamarkou, J. Drugowitsch, B. Deonovic, A. Sengupta, G. Ragusa, G. Moynihan, B. J. Smith, M. O'Leary, Michael, M. J. Innes, C. Dann, G. Lacerda, I. Dunning, J. Chen, M. Tarek, T. K. Papp, Juliastats/distributions.jl: v0.25.11, 2021. URL: https://doi.org/10.5281/zenodo.5105997. doi:10.5281/zenodo.5105997.

R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. URL: https://www.R-project.org.

T. Wenseleers, Analysis of the growth rate advantage and increase in infectiousness of the SARS-Cov2 variant of concern B.1.617.2, also know as the Delta variant, in India and selected other countris, Github.com (2021). URL: github.com/tomwenseleers/covid-delta.

Statens Serum Institute, Ekspertrapport af d. 23. oktober 2020: Incidens og fremskrivning af COVID-19 tilfælde, 2020. URL: https://files.ssi.dk/ekspertrapport-af-den-23-oktober-2020-incidens-og-fremskrivning-af-covid19-tilflde.

J. Revels, M. Lubin, T. Papamarkou, Forward-mode automatic differentiation in Julia, arXiv:1607.07892 [cs.MS] (2016). URL: https://arxiv.org/abs/1607.07892.

Appendix: Robust Standard Errors of Estimators

While the log-likelihood estimates are identical to those obtained with logistic regression packages, the standard errors provided by most packages are based on the Fisher Information matrix, $\hat{\mathcal{I}}$ (defined below), and for these to be reliable, the model must be correctly specified. We will compute standard errors using the sandwich form of variance-covariance matrix for the estimated parameters, which is detailed next.

We parameterize the log-likelihood with the standard parameterization of the logistic regression, $\theta' = (\alpha, \beta) = (\log \rho_0, \log \gamma)$. The Maximum likelihood estimates are obtained by maximizing $\sum_{t=1}^{T} \ell_t(\theta)$, where $\ell_t(\theta) = X_t \log \lambda_t + (N_t - X_t) \log(1 - \lambda_t)$. To compute robust standard errors we derive the score , $s_t(\theta) = \frac{\partial \ell_t(\theta)}{\partial \theta}$, and hessian, $h_t(\theta) = \frac{\partial^2 \ell_t(\theta)}{\partial \theta \partial \theta'}$. To this end we observe that the derivatives of $\lambda_t(\theta) = \lambda_t(\alpha, \beta) = [1 + e^{-\alpha - \beta t}]^{-1}$ are simply

$$\frac{\partial \lambda_t(\theta)}{\partial \alpha} = -e^{-\alpha - \beta t} \frac{1}{(1 + e^{-\alpha - \beta t})^2} = -\lambda_t(1 - \lambda_t),$$

and similarly of $\partial \lambda_t / \partial \beta = -\lambda_t(\theta)[1 - \lambda_t(\theta)] \times t$, such that the score for the observations in the t-th

week is given

$$s_t(\theta) = [-X_t(1-\lambda_t) + (N_t - X_t)\lambda_t] \begin{bmatrix} 1 \\ t \end{bmatrix} = (N_t\lambda_t - X_t) \begin{bmatrix} 1 \\ t \end{bmatrix}. \tag{A.1}$$

Next, by combining the expression for $\partial\lambda_t(\theta)/\partial\theta$ with (A.1) we obtain,

$$h_t(\theta) = -N_t\lambda_t(1-\lambda_t) \begin{bmatrix} 1 & t \\ t & t^2 \end{bmatrix}.$$

It is now straightforward to compute the information matrices $\hat{\mathcal{J}}_0 = \sum_{t=1}^{T} s_t(\hat{\theta})s_t(\hat{\theta})'$ and $\hat{\mathcal{I}} = -\sum_{t=1}^{T} h_t(\hat{\theta})$, which yields the heteroskedasticity robust variance covariance matrix for $\hat{\theta}$, $\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}\hat{\mathcal{I}}^{-1}$.[3] For heteroskedasticity and autocorrelation (HAC) robust standard errors we compute:

$$\hat{\mathcal{J}}_K = \hat{\mathcal{J}} + \sum_{j=1}^{K+1} k(\tfrac{j}{K}) \sum_{t=1}^{T-j} \left(s_t(\hat{\theta})s_{t+j}(\hat{\theta})' + s_{t+j}(\hat{\theta})s_t(\hat{\theta})' \right),$$

where $k(x)$ is a kernel function with $k(0) = 1$, $k(1) = 0$.

Variance-Estimator	Alpha Variant	Delta Variant
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.5037, 1.5262]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 2.1319, 2.2033]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_0\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4994, 1.5306]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 2.0215, 2.3236]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_1\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4990, 1.5310]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 2.0119, 2.3347]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_2\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4986, 1.5314]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 2.0009, 2.3476]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_3\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4980, 1.5320]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.9949, 2.3546]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_4\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4971, 1.5329]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.9909, 2.3593]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_5\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4962, 1.5339]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.9888, 2.3618]
$\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_6\hat{\mathcal{I}}^{-1}$	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.4952, 1.5349]	$\gamma_{4.7\text{days}} \in$ [CI 95%: 1.9888, 2.3618]

Table A.1: Sensitivity of confidence intervals for relative contagiousness, $\gamma_{4.7\text{days}}$, to the choice of variance-covariance estimator (computed with non-robust and various robust standard errors).

Our empirical results are based on HAC robust estimator, $\hat{\Sigma} = \hat{\mathcal{I}}^{-1}\hat{\mathcal{J}}_K\hat{\mathcal{I}}^{-1}$, with $K = 4$ and the Parzen kernel function for $k(\cdot)$. Standard errors and confidence intervals for α and β are given from the diagonal elements of $\hat{\Sigma}$, which we denote by $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\beta^2$, respectively. The reported 95% confidence intervals for α and β are based on the point estimates ± 1.96 times the corresponding standard error. Those for $\gamma_{4.7\text{days}} = \exp(\frac{4.7}{7}\beta)$ are given by $\exp\{\frac{4.7}{7}(\hat{\beta} \pm 1.96\hat{\sigma}_\beta)\}$.

[3] The numerical derivatives computed by the Julia package, `ForwardDiff`, see Revels et al. (2016), are identical to the analytical expressions for both the score, $s_t(\hat{\theta})$, $t = 1, \ldots, T$, and the hessian, $\sum_{t=1}^{T} h_t(\hat{\theta})$.