Statistical Research on COVID-19 Response

by

Xiaolin Huang
A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Statistical Research on COVID-19 Response

by

Xiaolin Huang

Supervisory Committee

Dr. Xuekui Zhang, Supervisor
(Department of Mathematics and Statistics)

Dr. Li Xing, Departmental Member
(Department of Mathematics and Statistics)

ABSTRACT

COVID-19 has affected the lives of millions of people worldwide. This thesis includes two statistical studies on the response to COVID-19. The first study explores the impact of lockdown timing on COVID-19 transmission across US counties. We used functional principal component analysis to extract COVID-19 transmission patterns from county-wise case counts, and used supervised machine learning to identify risk factors, with the timing of lockdowns being the most significant. In particular, we found a critical time point for lockdowns, as lockdowns implemented after this time point were associated with significantly more cases and faster spread. The second study proposes an adaptive sample pooling strategy for efficient COVID-19 diagnostic testing. When testing a cohort, our strategy dynamically updates the prevalence estimate after each test if possible, and uses the updated information to choose the optimal pool size for the subsequent test. Simulation studies show that compared to traditional pooling strategies, our strategy reduces the number of tests required to test a cohort and is more resilient to inaccurate prevalence inputs. We have developed a dashboard application to guide the clinicians through the test procedure when using our strategy.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My family, for their unwavering support throughout my studies.

My supervisor Dr. Xuekui Zhang, for mentoring, guidance, encouragement, and patience.

Dr. Li Xing, Dr. Youlian Pan, and Dr. Xiaojian Shao, for their invaluable knowledge and expertise.

Dr. You Liang, for her constructive, insightful suggestions.

All the staff in the department, for their immense help and support.

Part I

The Impact of Lockdown Timing on COVID-19 Transmission across US Counties

Chapter 1

Introduction

This work was published by EClinicalMedicine [12] and is included in Part I and Appendix A of this thesis verbatim with minor changes. Xiaolin Huang contributed to the acquisition of the datasets from online resources, data processing, data analysis, preparing the first draft of the manuscript, and participating in revisions. Permission from the publisher is not required to include the publication in this thesis.

Coronavirus disease 2019 (COVID-19) is a global pandemic that has affected over 181 million individuals and killed 3.9 million people across the world as of June 27, 2021 [26]. SARS-CoV-2, the virus responsible for this pandemic, is transmitted through a respiratory route with an average basic reproductive number (commonly denoted as R_0) of $2-3$ [25]. At this R_0, there is an exponential growth in the case counts of COVID-19 in the community, leading to large increases in COVID-19 related morbidity and mortality, which may overwhelm the local health care systems. To reduce COVID-19 transmission, governments around the world have imposed 'lockdowns' of their communities [11]. By limiting resident mobility and inter-personal contact, lockdowns along with other non-pharmacological interventions (NPIs) reduce the spread of COVID-19 in communities [27, 32, 8]. However, the timing of these lockdowns has been extremely variable with no clear consensus on when they should be implemented in communities. Here, we used data from over $3,000$ counties in the United States (US) to determine the relationship between the timing of lockdowns relative to the first appearance of COVID-19 and the trajectory of COVID-19 spread in these communities.

Chapter 2

Methods

2.1 Data sources

2.1.1 COVID-19 case counts during the pandemic across the United States (US) counties

We extracted COVID-19 data from the Johns Hopkins Coronavirus Resource Center [6] and analyzed the daily records of cumulative COVID-19 case counts across 3340 counties in the US from 2020-01–22 to 2021–01–31. We excluded counties that were not included in the US American Community Survey (ACS) [31] 5-year estimates, leaving 3140 counties in the dataset. We further excluded counties that did not report at least five total cases of COVID-19. The final data contained case counts from 3112 counties.

2.1.2 Demographic factors and lockdown across US counties

We extracted demographic, socioeconomic, and health insurance data for each county from the 2015–2019 US Census (using R package tidycensus [33]). Specifically, we fetched the following parameters (which are detailed in Table A.1) from the ACS five-year data profile for each county: socioeconomics (comprising median family income and the Gini Index), demographics (comprising total population, population density, and proportionality of males), health insurance status (private and public coverage of health insurance), household composition (median age), ethnicity, and geographical mobility and mode of transportation. In Figs. A.2–A.4, we display the relationship of these parameters with the COVID-19 count trajectories and have

overlaid these values on a US map in Figs. A.5–A.9. In addition, we determined "lockdown timing", which was calculated as the difference in days between the date on which the county experienced at least five cumulative cases of COVID-19 and the date on which the county first initiated a lockdown [36]. Here, we defined "lockdown" as the date on which "stay-at-home" orders were issued in a county. If a county instituted multiple lockdowns during the follow-up period, we only used the first lockdown in our downstream analysis.

2.1.3 Non-pharmaceutical interventions

We also included data on non-pharmaceutical interventions (NPI), which were defined using terms from the Oxford Covid-19 Government Response Tracker (OxCGRT) [11]. We formated the data to enable calculation of the time interval (in days) from the reporting date of a county of 5 or more cumulative cases of COVID-19 to the initiation date of the NPI in question. The NPIs included 'debt/contract relief' (government preventing termination of services from missing payments), 'public information campaigns' (on COVID-19), 'testing policy' (accessibility to COVID diagnostics), 'contact tracing' (of identified cases), use of 'facial coverings', 'vaccination policy' (availability of vaccines), and 'protection of elderly people'. We excluded NPIs which had more than 30% of missing data. Detailed definitions of NPIs can be found in Table A.1.

2.2 Statistical methods

2.2.1 Modeling the spread of COVID-19 over time in the US counties using unsupervised machine learning

We considered the daily cumulative case count of a county as its trajectory over time, and extracted the patterns using a functional principal component (FPC) analysis [34]. First, we realigned the trajectories to ensure that there were at least five cumulative cases at the start of each trajectory. We then investigated the hidden patterns in these trajectories with FPC analysis. The FPC model is given using the following formula:

$$log(Q_{ij}) = f_i(t_j) = \mu(t_j) + \sum_{k=1}^{m} \xi_{ik}\phi_k(t_j) \tag{2.1}$$

where Q_{ij} is the cumulative case count of the i^{th} county on the j^{th} day.

The FPC model mapped these trajectories onto an m-dimensional functional space spanned by m orthogonal eigen-functions $\phi_k()$. The eigen-functions are ordered by the proportion of variance in the dataset that can be explained by these functions. Each eigen-function describes how individual trajectory differs from $\mu()$, which denotes the average trajectory across all the counties. The coefficient ξ_{ik} is the functional principal component (FPC) score, or the coordinate of the i^{th} county in the k^{th} dimension of the functional space. Practically, ξ_{ik} describes the strength of the k^{th} pattern in the i^{th} county's cumulative case count trajectories. Therefore, the log daily case count trajectory of each county can be modeled as the national average trajectory plus the sum of eigenfunctions (weighted by corresponding FPC scores), as in (2.1). Müller et al. [34, 20] introduced the theoretical details that outline the method by which estimated functions $\mu()$ and $\phi_k()$ as well as coefficients ξ_{ik} are generated. In this work, we estimated these parameters using the R package fdaPACE [3].

2.2.2 Exploring the marginal effect of each risk factor

Using a simple linear regression model, we investigated the unadjusted marginal effects of lockdown timing, characteristics of the counties, and NPIs on COVID-19 transmission across the US. A summary statistics from these linear regression models is provided in Table 3.2.

Fig. 3.3 shows that the observed relationship between the first FPC scores and the timing of lockdowns was non-linear: its appearance was that of a "hockey stick" with an inflection point indicating a significant change in its slope. Thus, we derived three new variables from the timing of lockdown: a binary indicator of lockdown implementation, a slope before inflection point (denotes the effect of the lockdown timing when implemented before the inflection point), and a slope after inflection point (denotes the effect of the lockdown timing when implemented after the inflection point), and used segmented regression to model this relationship. The inflection point for the lockdown variable was ascertained via the significant change in the slopes before and after the inflection point. The statistical technical details of the segmented model are provided in supplementary document under section "*Modeling lockdown effect using segmented regression.*"

2.2.3 Modeling joint effects of all risk factors simultaneously using supervised machine learning

Finally, to explore the joint effects of all risk factors on the first FPC scores, we fitted an elastic net model [37] to these data. Elastic net is a popular machine learning method, which is based on a regularized linear model

$$\xi_{i1} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \alpha_4 X_{i4} + \cdots + \alpha_p X_{ip} \tag{2.2}$$

where (X_{i1}, X_{i2}, X_{i3}) are derived variables from lockdown information as defined in model (A.1) which together represent the effect of lockdown timing, and $(X_{i4}, ..., X_{ip})$ are $(p-3)$ demographic and NPI characteristics of the i^{th} county.

Compared with multiple linear regression, elastic net incorporates various penalties on coefficients and provides better prediction models. First, elastic net can automatically select important predictors in a linear model (2.2) by automatically assigning a zero coefficient to unimportant predictors via a penalty on absolute values of coefficients. Second, the elastic net penalty addresses the issue of multi-collinearity among predictors, which makes models more reliable than multiple regressions. However, elastic net does not provide confidence intervals for coefficients. To capture the uncertainty of the risk estimates, we generated 95% confidence intervals for each coefficient using a re-sample (bootstrap) approach. Specifically, we sampled the counties for replication 1000 times. Next, we applied an elastic net model to each of these random subsets to generate 1000 sets of estimated coefficients, and then built a 95% confidence interval using these coefficients. Here, we fitted all elastic net models using R Package 'glmnet' [9]. Statistical significance was defined by p-value < 0.05. All data analyses were performed using R Statistical Software [24]. The source codes are available to the public by accessing `https://github.com/ubcxzhang/COVID.FPCA/`. The mathematical details and interpretation of this modeling process are provided in section "*Interpretation of fitted Elastic net models*" of the supplementary document.

2.3 Role of the funding

The data analysis is conducted using computing resource offered by
Compute Canada/West Grid. The sponsor had no role in the design of the study, the collection and analysis of the data, or the preparation of the manuscript.

Chapter 3

Results

3.1 Functional principal component analysis of COVID-19 case counts

We performed a Functional Principal Component (FPC) Analysis on the trajectories of COVID-19 spread across 3112 US counties. Strikingly, the first FPC explained a vast majority of the total variance (about 92.86%). The first FPC score represents the weighted average of COVID-19 case counts and the weighted changes in the rate of COVID-19 case counts over time (on an exponential scale), with weights based on the first eigenfunction. Thus, we can use the first FPC score to describe the overall severity of the pandemic for the i^{th} county. In section "*Interpretation of the first FPC scores*" of the supplementary document, we provide the mathematical details to support this interpretation.

Fig. 3.1 shows the average trajectory of COVID-19 daily cumulative counts across all US counties, which is denoted by the function $\mu()$. In the lower panel, the blue/red curve represents the average COVID-19 case count trajectories of counties that implemented a lockdown before/after the inflection point. The shaded area represents the confidence intervals constructed using the interquartile range (i.e., 25–75% quartiles). An early lockdown (before the inflection point) was associated with a lower case count than the national average across the entirety of the follow-up period; whereas the opposite was true for late lockdowns (defined as occurring after the inflection point). Furthermore, an early lockdown was associated with a slower increase in the rate of COVID-19 counts for the first 50 days of the pandemic. The upper panel shows the percentages of counties which implemented lockdown at each day, grouped by

early (blue) or late (red) lockdown. Since we normalized the trajectories by defining day-0 as the day on which a county reported first 5 cumulative cases and "early" versus "late" lockdown was dichotomized based on implementation of a lockdown approximately 7 days prior to day 0, all "early-lockdown" counties were by definition locked-down at day 0. In contrast, the late-lockdown counties did not achieve full lockdown until approximately day 25. The lower panel shows the differences in slopes between red and blue average trajectories over the first 50 days. However, after this period, the two trajectories gradually became approximately parallel, indicating that late-lockdown counties have more cumulative cases across the time range. Fig. 3.2 shows a heat map of the US according to the FPC score for each county. Since we used the first FPC score as a surrogate variable for the overall severity of the pandemic, the darker colored regions represent a more severe outbreak of COVID-19. Thus, counties in the western and eastern coastal states in general demonstrated significantly higher case counts compared with those in the central states. The most severely affected counties were found in New York, Arizona, Florida, and California.

3.2 The marginal effects of risk factors

We employed a simple linear regression to explore the unadjusted relationships between the cumulative case count trajectories and each potential risk factors. Table 3.2 summarizes the results, which include regression coefficients, p-values, and the R^2 statistic. Among the 21 factors we investigated, all of them demonstrated a significant coefficient (p-value < 0.05) for the first FPC score. The marginal R^2 was moderate for these factors, up to 0.34. The variable 'Total Population' ($R^2 = 0.339$) displayed the strongest association, which was followed by the variable, 'Contact Tracing' ($R^2 = 0.212$). The two most negatively correlated factors were 'Median Age' ($R^2 = 0.194$) and 'Proportion of Whites' ($R^2 = 0.101$).

3.3 The impact of implementing a lockdown

Fig. 3.3 shows the relationship between the first FPC scores and the timing of the lockdown, which displays a strong non-linear relationship. To better characterize this relationship, we used a segmented regression model. Compared with a linear regression model, segmented regression improved the fit of the model (i.e., $R^2 = 0.45$ for segmented regression vs. $R^2 = 0.20$ for linear regression). The red lines in Fig. 3.3

show the segments of a fitted line, whose appearance was a "hockey stick" containing an inflection point. Using time zero as the date on which a county reported at least 5 cumulative cases of COVID-19, we identified day -7.76 (i.e., approximately a week before a county reported at least 5 cumulative cases of COVID-19) as the average "inflection" point (the green vertical line in Fig. 3.3) in the segmented regression model. We divided the counties into two groups based on whether or not a lockdown was implemented before this inflection point, and compared the underlying demographic and lockdown features of these two groups (Table 3.1). Note that certain NPIs were negative values because these policies were implemented at an early stage in the pandemic (i.e., before the counties reported 5 or more cumulative cases). The detailed results of the segmented regression model are shown in Table 3.2. Specifically, the two slopes corresponding to the two segments, '*Lockdown Slope before the Inflection Point*' and '*Lockdown Slope after the Inflection Point*' were all positive, corresponding to 0.05069, and 1.95820, respectively.

3.4 Joint modeling for all risk factors for COVID-19

We found that certain risk factors were highly correlated with each other as shown in Fig. A.1. As seen in regression models of marginal effects, most of the risk factors were significantly associated with COVID-19 infection. To investigate their joint effects after adjusting for other variables, we used an elastic net model to determine the relationship of the first FPC scores with these predictors. The confidence intervals, obtained from 1000 bootstraps, are shown in Fig. 3.4 and the mean value and the 95% confidence interval of the model's coefficients are shown in supplementary Table A.2. We note that the elastic net models achieved a much better fit with an R^2 of 0.62, compared with the marginal regression results in which the maximal R^2 was 0.34 for the first FPC score.

We observed that 6 of 24 risk factors demonstrated statistical significance (i.e., their coefficients did not cover zero). For example, '*Lockdown Slope after Inflection point*' and '*Total Population*' were positively associated with the first FPC scores, while '*Median Age*' was negatively associated with the first FPC scores. Other positive risk factors included '*Median Family Income*', '*Gini Index*', and '*Proportion who Moved within the Same County*'. Many other factors became statistically insignificant

in the joint models.

In the elastic net models, the mean of '*Lockdown Slope after Inflection Point*' was 1.048 from 1000 bootstraps. This indicates that, after adjusting for other factors, if a lockdown was implemented after the inflection point, there was an exponential increase in the cumulative COVID-19 case counts in the community over the follow-up time. Specifically, model (A.5) demonstrates that the changes in the daily cumulative case count are a function of the first FPC scores and the first eigenfunction. For each day of delay in implementing a lockdown after the inflection point, the daily cumulative case count increased on average by 5.80% (range 2.36 to 7.03%). For each week of delay in implementing a lockdown after the inflection point, the daily cumulative case count increased on average by 48.36% (range 17.77 to 60.92%). The timing of the lockdown at the county level explained 45% of the total variance (R^2 of segmented regression model) in the cumulative case counts of COVID-19 across the communities.

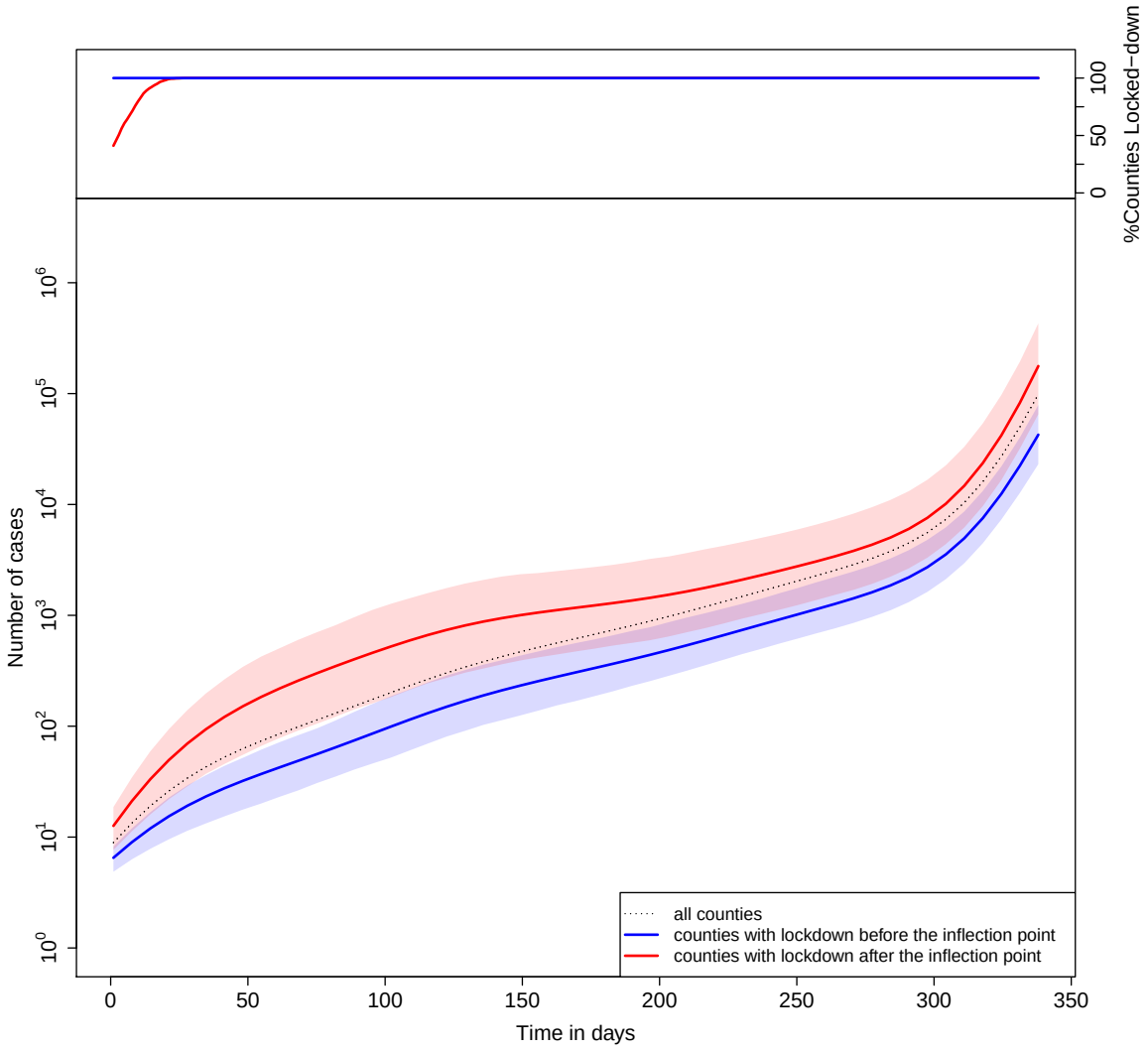

Figure 3.1: The mean curve of COVID-19 cumulative case trajectories. In the upper panel, the red curve shows the (cumulative) percentages of "late-lockdown" counties which locked down during the follow-up period. Day zero is defined as the date on which a county reported at least 5 cumulative COVID-19 cases. Late-lockdown was defined as implementing lockdown after the inflection point (which occurred approximately 7 days prior to day 0). Blue line denotes "early-lockdown" counties. In the lower panel, dotted curve $\mu()$ represents the national average of COVID-19 cases over time. The blue curve represents the average COVID-19 count trajectories of counties that implemented a lockdown before the inflection point, while the red curve represents the average trajectories of counties with lockdown after the inflection point. The shaded area represents confidence bound constructed using interquartile range (i.e., $25-75\%$ quantiles).

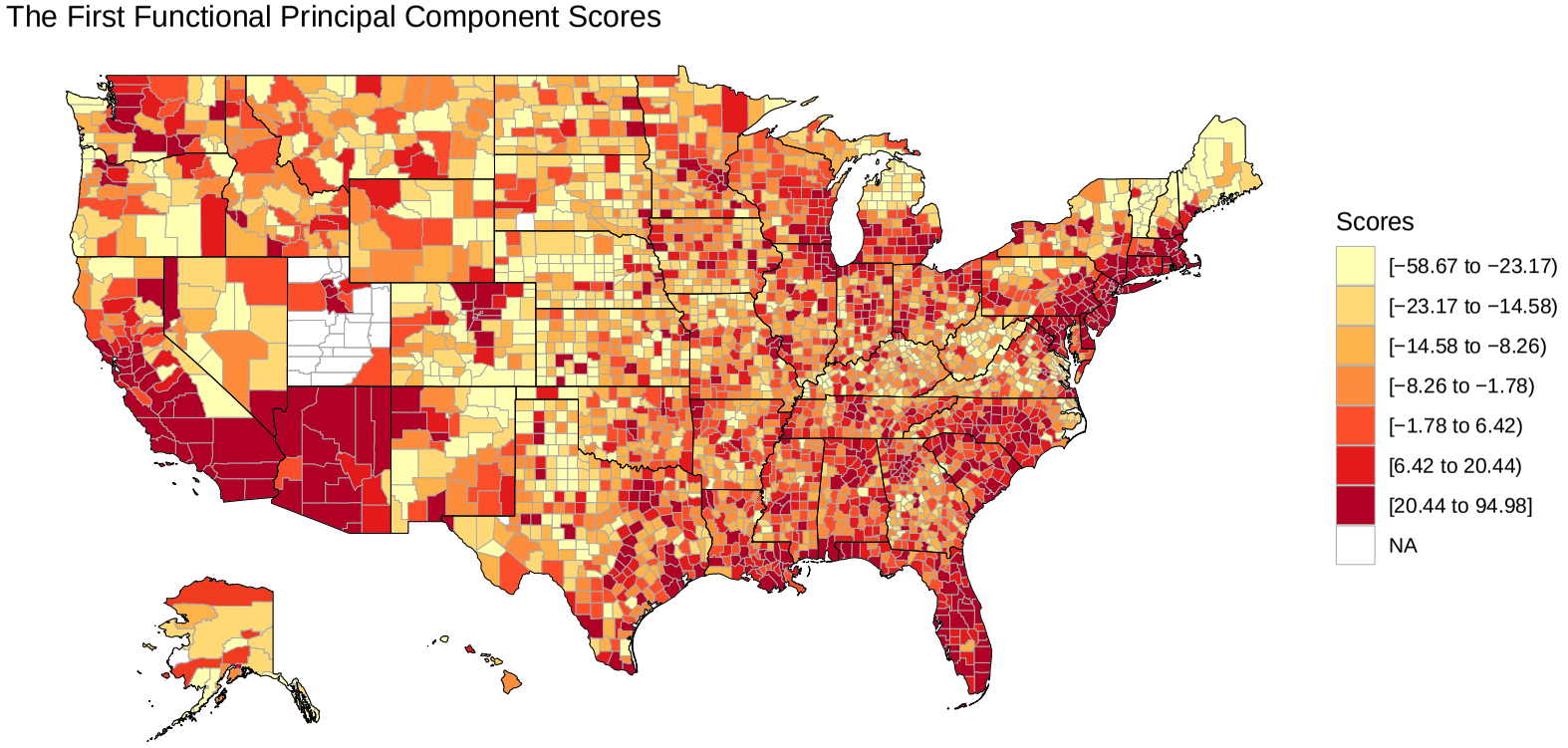

Figure 3.2: Heat map of the United States (US) according to the first FPC scores of counties.

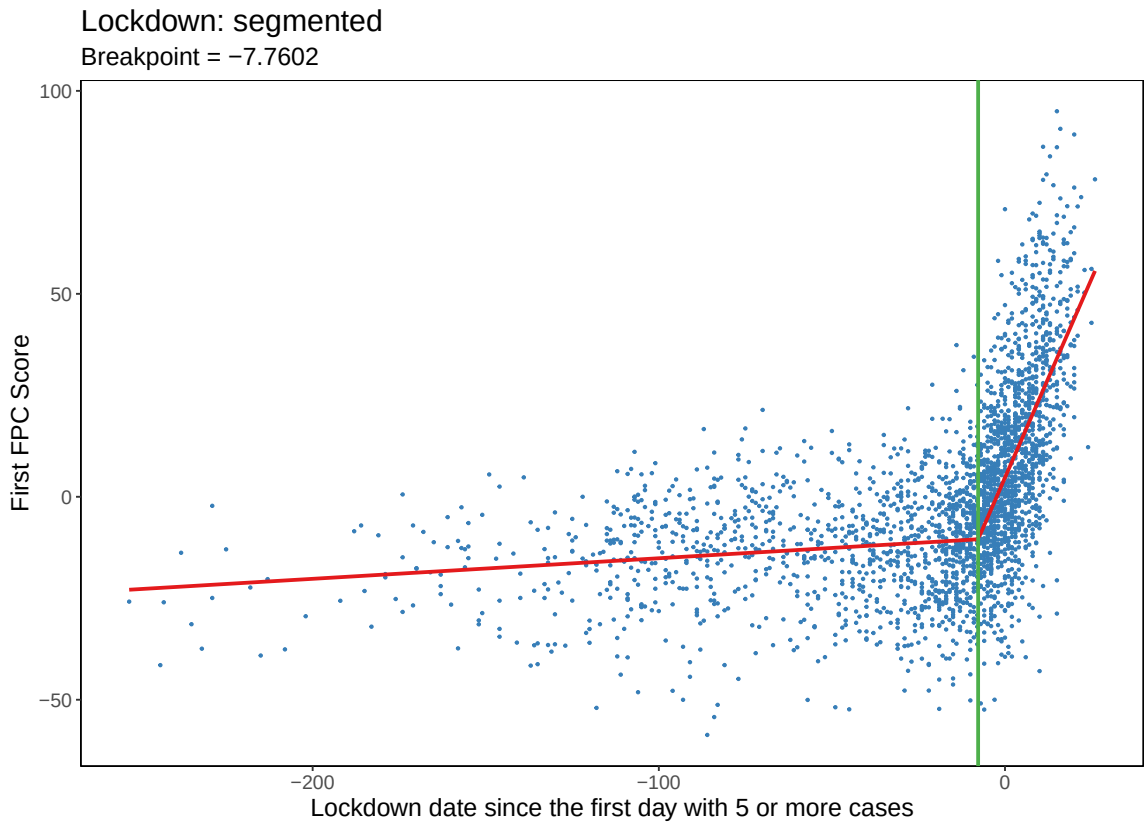

Figure 3.3: Relationship between the first FPC score and the first lockdown date. The x-axis represents the number of days between the lockdown date and the date on which the county reported at least 5 COVID-19 cases. Positive values denote counties that instituted a lockdown after they reported at least 5 cumulative COVID-19 cases, while negative values denote counties that instituted a lockdown before they reported at least 5 cumulative COVID-19 cases. Each blue point represents data of a US county. The red hockey-stick shape line represents two fitted slopes of a segmented regression model. The vertical green line (at −7.8 days) indicates the inflection point on which the slope of the first FPC score significantly changes.

Table 3.1: Baseline characteristics of counties according to implementation of early or late lockdown (defined as whether or not implementation date was before the inflection point, i.e., 7 days before 5 total cases were reported in a county) in the course of the pandemic.

County Characteristics	**Early Lockdown (n=1349)**	**Late Lockdown (n=1378)**
Total Population ($\times 10^3$)	20.6±21	208±479
Population Density (number of people per sq mile)	64.2±322	543±2660
Median Age (years)	42.9±5.51	39.9±4.71
Median Family Income ($\$ \times 10^3$)	61.5±12.3	70.7±19.2
Gini Index	0.441±0.0378	0.453±0.0344
Proportion of Male (%)	50.7±2.76	49.5±1.83
Proportion of Whites (%)	87.4±14.4	77.1±17.4
Proportion of African Americans (%)	5.17±11	14.5±16.8
Proportion of Natives (%)	2.22±7.69	1.1±4.42
Proportion of Asians (%)	0.741±1.72	2.19±3.63
Proportion of Individuals who Used Public Transport (%)	0.438±1.06	1.54±4.44
Proportion who Moved within the Same County (%)	5.62±2.44	6.9±2.6
Proportion with Private Health Insurance (%)	63±10	66.4±9.82
Proportion with Public Health Insurance (%)	42.3±9.09	37.5±8.09
Debt/Contract Relief (days)*	-60.3±45.4	-7.46±8.64
Public Information Campaigns (days)*	-91.4±48.8	-33.3±22.1
Testing Policy (days)*	-117±45.5	-65.2±7.42
Contact Tracing (days)*	-117±45.5	-65.2±7.41
Facial Coverings (days)*	123±143	170±138
Vaccination Policy (days)*	211±45.5	264±9.55
Protection of Elderly People (days)*	-4.44±135	43.6±112

P-values for all variables are smaller than 0.05 based on a Wilcoxon test for differences between early lockdown and late lockdown. Data are shown as mean ± SD.
* days are calculated relative to day 0 (i.e. the date on which counties reported 5 or more cumulative cases of COVID-19). A negative value would indicate that counties implemented these non-pharmacologic intervention (NPI) several days prior to day 0; a positive value would indicate that NPIs were implemented after day 0.

Table 3.2: The unadjusted relationship of the baseline characteristics of the counties with the COVID-19 spread in these communities.

County Characteristics	**Association with the First FPC Score**		
	Coefficient	***P*-value**	R^2
Lockdown Slope Before the Inflection Point	0.05069*	2.43E-08*	0.448*
Lockdown Slope After the Inflection Point	1.95820*	<2E-16*	
Total Population	3.87E-05	5.70E-282	0.33883
Contact Tracing	0.23167	2.53E-163	0.21197
Testing Policy	0.23163	2.60E-163	0.21195
Vaccination Policy	0.22893	3.65E-161	0.20944
Debt/Contract Relief	0.22428	6.07E-155	0.20214
Median Age	-1.8156	2.29E-148	0.19434
Proportion of Asians	343.83	4.19E-143	0.18805
Public Information Campaigns	0.18513	4.24E-134	0.17716
Proportion who Moved within the Same County	328.89	9.24E-116	0.15455
Proportion of Individuals who Used Public Transport	253.39	7.88E-93	0.12541
Proportion of Whites	-41.881	4.94E-74	0.10078
Median Family Income	4.17E-04	8.16E-70	0.095166
Population Density	0.0035699	1.27E-61	0.084159
Proportion with Public Health Insurance	-64.053	2.38E-49	0.067428
Proportion of African Americans	35.071	6.27E-39	0.053
Gini Index	117.79	6.10E-28	0.037569
Proportion of Male	-164.56	1.02E-22	0.030165
Protection of Elderly People	0.027229	9.03E-17	0.021685
Proportion with Private Health Insurance	22.156	4.40E-09	0.010696
Facial Coverings	0.012778	1.94E-06	0.0069405
Proportion of Natives	-16.217	0.0023103	0.002661

Variables are sorted by R^2. The first FPC score is used as a surrogate for COVID-19 spread across the counties. (*Results from segmented regression model; the rest are from linear regression models).

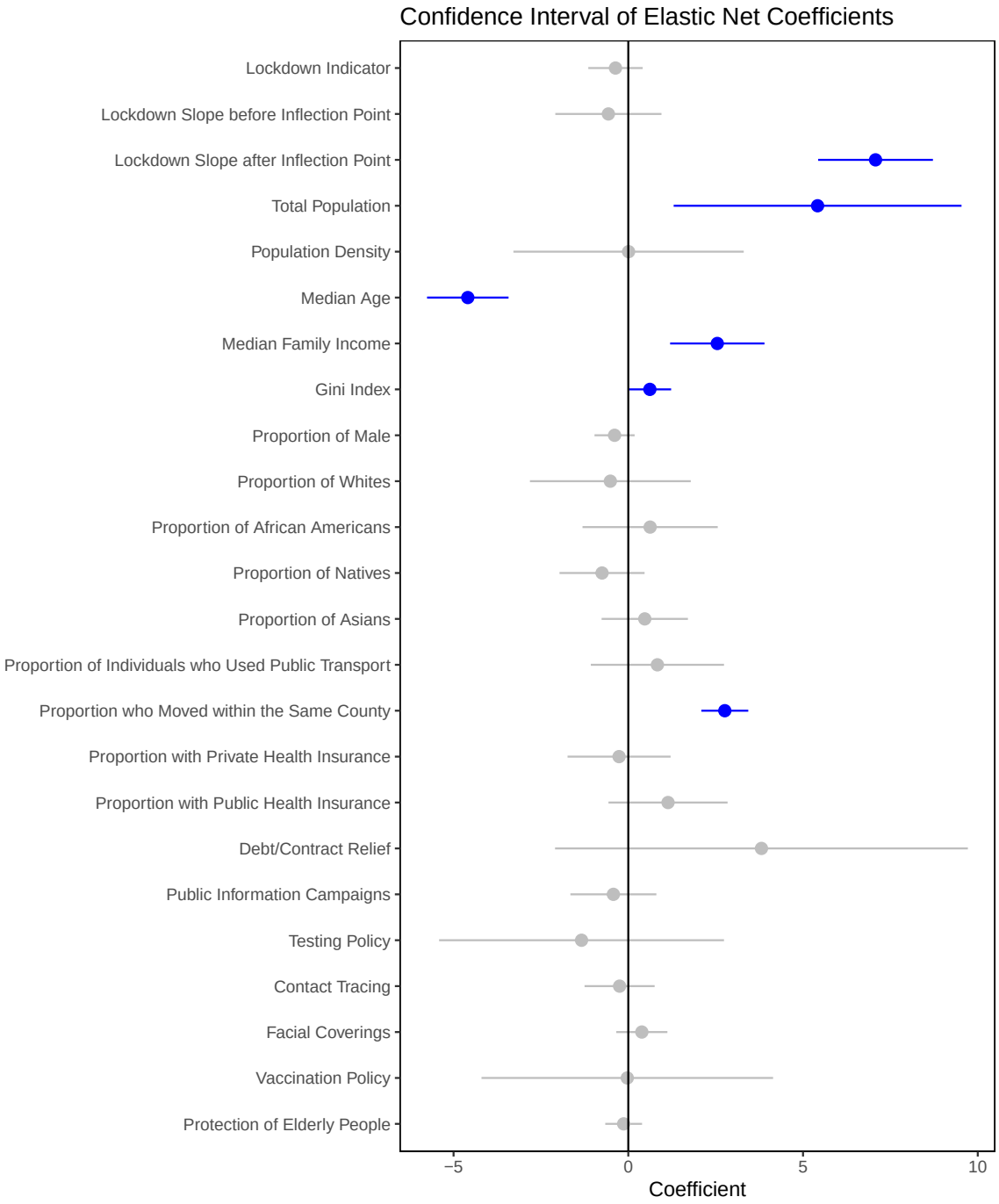

Figure 3.4: The adjusted relationship between standardized characteristics of counties and the first FPC scores, based on results of elastic net models. The effect of every variable is adjusted to other factors listed in the figure. A positive coefficient denotes variables that are positively related to the number of COVID cases. The dot indicates the mean coefficients, and the bar represents the 95% confidence interval. Blue color indicates the significant factors whose 95% confidence interval does not cover 0.

Chapter 4

Discussion

Lockdowns are an effective way of reducing the reproduction number of COVID-19 and controlling the spread of disease in local communities. However, there is no consensus on when governments should take this action. Here, we found that communities, which implemented the lockdown at or prior to the inflection point (defined as 7 days before the date on which at least 5 cumulative cases were first reported in the community) experienced a slower rise in COVID-19 rates over the first 50 days and a lower cumulative count consistently across all time points during follow-up compared with counties that implemented lockdowns after the inflection point (Fig. 3.1). In our models, the timing of the lockdown at the county level explained nearly 50% of the total in COVID-19 case counts across US counties, highlighting the importance of early lockdown implementation in controlling the pandemic at the county level.

Our findings extend data from recent cross-sectional studies that have investigated the relationship of COVID-19 spread in communities with their population characteristics and lockdown measures. By examining the temporal patterns of COVID-19 transmission within and across the US, we demonstrated the relationship between the timing of lockdown implementation and the trajectory of COVID-19, independent of other characteristics, within and across US counties using FPC analysis for the first time. We were able to convert the trajectory of COVID-19 spread for each county into a (first) FPC score, which accounted for 93% of the total variance in the COVID-19 infection trajectories across the US counties. This enabled us to use the first FPC score as a surrogate for infection case counts in these counties and model the relationship of the longitudinal COVID-19 infection pattern with the timing of lockdowns, and other risk factors including the use of NPIs, and demographic characteristics of

US counties.

Based on an elastic net model of risk effects, we found that the most important factors associated with a rapid spread of COVID-19 at a county level were the timing of the lockdown, and certain characteristics of the counties. For example, counties with a larger population experienced a more rapid rate of COVID-19 transmission compared with smaller counties. The heat map (Fig. 3.2) reveals that the most populous states, such as New York, California, and Florida, were most impacted by COVID-19. At a city level, Los Angeles had the highest first FPC score, followed by Chicago, Miami, and New York, which all have large populations. Interestingly, counties with a higher median family income and a higher Gini Index (representing greater spread of income inequality) experienced a more rapid COVID-19 surge, which aligns with the findings by Tan et al. [29] Although COVID-19 becomes more severe among older adults, counties with a lower median age experienced more case counts than older counties. These data are consistent with the observation that case counts generally decrease with increasing age in adulthood [4]. Finally, we found that increased mobility within counties is also associated with increased COVID-19 case counts.

There are many definitions of lockdowns [10]. Here, we defined lockdown as the day on which the local government issued a "stay-at-home" order. To evaluate the robustness of our results, we performed several additional analyses using alternate definitions of lockdown (e.g. the date of school closing, workplace closing, cancelation of public events, restrictions on gatherings, etc.). However, the use of alternate definitions did not materially change the primary results. In every case, the analysis showed a non-linear "hockey-stick" relationship between the date of "lockdown" and the cumulative rise in case counts across the US communities, as shown in Fig. A.11–A.15. Importantly, we found that the definition based on the date of stay-at-home order produced the fewest number of outliers amongst all definitions that were evaluated. Thus, we believe that our *a priori* decision to use the date of issuance of a stay-at-home order was a reasonable choice for our primary analysis, yielding the most robust data.

Note, Principal Component Analysis (PCA) is a popular dimension reduction method. In this work, however, we used FPC analysis instead of PCA for the following three reasons: (1) The model had to account for differential follow-up time across the counties. This occurred because the date on which 5 cumulative cases were reported for each county significantly differed. Differential follow-up time, however, led to an

uneven matrix, preventing the use of PCA. (2) Because FPC analysis considers each trajectory to be a smooth curve, this allows observations to borrow information from their nearby points on the trajectory to improve the quality of results. (3) PCA is not sensitive to the time-order of observations and, thus, not suitable for a "trajectory over time", which again made it unsuitable for our dataset. In contrast, FPC analysis retains all the information of a time-order dataset, making it a preferred choice over PCA.

In this work, we defined "day 0" as the "first instance of detecting more than 5 cases". Our choice of 5 cumulative case counts was based on the fact that with a lower threshold, the uncertainty (or the noise) of the measurement would be significantly increased. On the other hand, we were concerned that a higher threshold cutoff (e.g. 100) may artificially bias the inflection point towards a higher number. For example, if we had used 100-cases to define day 0, we would have discarded all the information collected before 100 cases were reached. Although the choice of 5 was arbitrary, in the literature, we found many incidences where statisticians have chosen 5 as their "magic threshold". To check for the robustness of the case definition, we repeated our analysis using 3-case and 4-case definitions, and found similar results (available upon request).

There were limitations to the study. First, as this was not a randomized controlled trial, unmeasured confounders could have distorted the overall findings. To minimize this possibility, we evaluated only counties in the US and adjusted for the most important characteristics of these counties using well-curated databases. Second, these data were generated in the US and may not apply to other countries around the world, which may have different characteristics and attitudes and adherence to public health policies such as masking and social distancing. Third, we could not fully quantify the stringency of the stay-at-home orders, or the adherence rate of the residents to the lockdown order across the counties. Fourth, in our analysis, we considered the effects of the first lockdown order for each of the counties. It should be noted, however, that some counties experienced multiple lockdowns during the follow-up period, leading to an "on-and-off" effect. Future studies will be needed to evaluate the effects of multiple lockdowns on communities. Finally, we could not address problems related to the quality of data source such as unexplained bias and unobserved errors in the raw data.

Notwithstanding these limitations, our findings have important public health implications. Local state and municipal governments should issue an immediate lock-

down order even when there are a few cases of COVID-19 in their communities (less than 5); any significant delays in lockdown beyond this point are associated with a rapid growth of COVID counts and a higher overall cumulative count trajectory, which will make COVID-19 containment difficult for that community.

Part II

An Adaptive Sample Pooling Strategy for COVID-19 Diagnostic Testing

Chapter 5

Introduction

A version of this work was submitted to the Journal of the American Statistical Association in collaboration with Xuekui Zhang and Li Xing. Xiaolin Huang contributed to data simulation, data processing, data analysis, application development, preparing the first draft of the manuscript, and participating in revisions.

As of May 2022, COVID-19 has affected over 500 million people worldwide. Since the end of 2019, the COVID-19 pandemic has seen multiple peaks in many countries [26]. During each outbreak, more tests were needed to identify the rising number of positive cases and prevent the further spread of the disease. However, as testing resources are limited by factors such as workforce and logistics, testing facilities and laboratories can only process a certain number of tests within a certain period of time. Therefore, it is highly desirable to increase the efficiency of COVID-19 testing by testing more samples using limited resources, which can provide timely results and guidance for actions from the treatment of individual patients to nonpharmaceutical interventions such as quarantines and lockdowns at a community level.

With the emergence of the Omicron variant, testing remains a critical component in the COVID-19 pandemic control at the current stage. Omicron has a reproduction number much higher than its predecessor Delta [17] and is responsible for the latest and highest peak of confirmed cases in the United States [13]. In January 2022, the Omicron outbreak overwhelmed Canada's testing capacity [21]. Omicron also caused the largest outbreak in China since the beginning of the pandemic, and the government had to lockdown several areas of the country [35]. Although Omicron is less likely to cause severe clinical symptoms in patients than the previous variants [18], we still face the challenges of conducting a mass volume of tests. Such scale is especially critical to monitoring the spread of the disease as we phase out various

COVID-related restrictions. Furthermore, testing can provide early warning of new outbreaks caused by new variants of COVID-19 that requires early intervention. In another study, we found that the timing of intervention in early outbreaks is a crucial factor affecting COVID-19 transmission patterns [12].

Pooled sample testing is a popular strategy to improve testing efficiency with a long history dating back to the twentieth century. In 1943, Dorfman proposed the method of combining multiple individual samples into a pool for testing. If the pool is positive, every single sample in the pool will be retested [7]. This method, called Dorfman Pooling, is popular in clinical settings [5]. In addition to Dorfman Pooling, many other types of pooling strategies have been developed, such as binary splitting, where a positive pool is recursively split into two equal-sized pools and tested respectively [16]. As for pooling in viral testing, serum samples were pooled for HIV antibody testing to reduce costs in the 1990s [28]. Another study proposed a method that utilizes the geometry of a hypercube to reduce the number the tests. An initial pooled test was used to estimate the prevalence, and the positive pools were distributed into pools modeled by the slices of a hypercube to identify the positive cases. The strategy was tested with oropharyngeal swab specimens in Rwanda and showed that positive COVID cases could still be detected at a 100-fold dilution [19].

Pooling has been widely used during the COVID-19 pandemic. In Israel, over 130,000 samples were tested in 5 months using pooled testing, and a 76% reduction in tests was achieved [2]. In Vietnam, more than 96,000 people were tested using pooling in 14 days, saving 77% of the resources, totaling about 1.5 million USD [30]. In both studies, there was no significant compromise in sensitivity.

The main benefit of pooling is to reduce the total number of tests when a group of samples is tested negative in one test. Pooling too many samples may include positive cases and lead to positive diagnoses, which still require follow-up testing; pooling too few samples does not fully utilize the benefit of reducing the number of tests. Therefore, the pool size is a crucial factor in a pooling strategy. Many different pool sizes have been used in recent studies. A six-sample pooled strategy improved the testing capacity by around 100% in Kenya [1]. In Seattle, four-sample pooling allowed a greater testing throughput as well as saving testing material [22]. In practice, pooling ten samples per test is the most popular choice in large-scale testing. It was used for COVID screening at an airport in China with the potential to reduce the need for contact tracing [14]. A Malaysian study demonstrated the robustness of pooling ten samples when testing demand is high [15]. A study in Ohio

also showed the feasibility of 10:1 pooling in a low-prevalence setting [23].

Although pooing ten samples per test is the most popular choice in practice for large-cohort testing, the optimal pool size varies in different situations. Hence, pooling with the optimal pool size instead of an empirically chosen one can help reduce the expected number of tests. Dorfman established the concept of expected relative cost, defined as the ratio between the expected number of tests required by Dorfman Pooling to the number of tests required by individual testing. Dorfman calculated the optimal pool sizes at different prevalences by minimizing the expected relative cost [7]. In the Israeli study, the pool size was chosen between 5 (Dorfman's optimal pool size for prevalence at 5%) and 8 (Dorfman's optimal pool size for prevalence at 2%) weekly depending on the previous week's test results. The pool size was also switched to five when the samples came from a subpopulation suspected to have a high prevalence [2].

Intuitively, prevalence plays an important role in determining the optimal pool size. With a higher population prevalence, it is more likely to include a positive sample in the pool, and the pool size should be relatively small. The opposite can also be said for low prevalence scenarios. However, it is difficult to determine the prevalence at the beginning of the testing procedure as the value varies among subpopulations and between different epidemic phases. Therefore, it is necessary to proactively update the prevalence estimate when testing large numbers of samples. One of the limitations of traditional Dorfman Pooling is that it uses a constant pool size for the whole cohort, where the pool size might not be the most suitable as testing progresses. Another limitation is that when a pool tests positive, Dorfman Pooling directly tests each sample individually without further pooling to reduce the number of tests. Also, Dorfman Pooling does not consider the cohort size when determining the pool size, the relationship of which will be discussed in Chapter 7.

To address these limitations, we propose an adaptive sample pooling strategy (ADSP) for conducting diagnostic testing. Specifically, to efficiently conduct COVID-19 diagnostic testing for a large cohort, our novel algorithm dynamically updates the prevalence estimate based on the previous test result (positive or negative) and adaptively determines the optimal pool size of the subsequent test accordingly. Our ADSP also recursively pools samples to reduce the number of tests further. With simulation studies, we illustrate that our ADSP requires fewer tests than Dorfman Pooling in many scenarios.

Chapter 6

Methods

We propose an adaptive sample pooling strategy for large-cohort testing that dynamically chooses the optimal pool size, given the current prevalence estimate and cohort size, to reduce the total number of tests needed to test the cohort. We denote the size of the cohort by n, and the prevalence of infections in this cohort by $p \in (0, 1)$. At any stage of testing this cohort, our ADSP only solves the optimal pool size for the current test since the outcome of the current test is needed to update the information and learn the optimal pool size for the next test.

6.1 Statistical models for learning the optimal adaptive pool size for the current test

We let $V(k, n, p)$ denote the minimum expected number of tests required to test a cohort of n samples with prevalence p if we pool k samples for the current test. We model our dynamic pooling strategy as an optimization problem to solve the optimal pool size k^* for the current test, defined as

$$k^* \equiv \underset{1 \le k \le n}{\operatorname{argmin}} \; V(k, n, p). \tag{6.1}$$

Hence, the minimum expected number of tests required to test this cohort is

$$Y(n, p) \equiv V(k^*, n, p). \tag{6.2}$$

Following the solution of the optimization problem (6.1), we split the n samples

into two groups. One group has k^* samples that are pooled and tested with a single test. The other group has the rest $n-k^*$ samples, which is considered a new cohort for testing. The optimal pool size for the new cohort is solved from the same optimization problem (6.1), except replacing n by $n-k^*$. Following the notation above, $Y(n-k^*, p)$ represents the minimum expected number of tests required by the new cohort.

Testing the pool of k^* samples leads to two possible scenarios. When the pool tests negative, we label each of the k^* samples as negative. We used one test to diagnose these k^* samples, so the minimum expected number of tests required by the original cohort is

$$Y(n, p) = 1 + Y(n - k^*, p) \tag{6.3}$$

when the pool tests negative.

When the pool tests positive, we need to conduct follow-up tests on these k^* samples with potentially further pooling. This cohort of k^* samples is different from the original cohort since we now have the extra information that positive samples exist in this cohort. Therefore, we introduce new notations U and X as the counterparts of V and Y defined above. For a cohort of n samples with prevalence p and the information that positive samples exist, let $U(k, n, p)$ denote the minimum expected number of tests required to test the cohort if we pool k samples for the current test. We define a similar optimization problem to solve for the optimal pool size k^{**} in this situation as

$$k^{**} \equiv \underset{1 \le k < n}{\operatorname{argmin}} \; U(k, n, p). \tag{6.4}$$

and the minimum expected number of tests required for testing this cohort of n samples with extra information is

$$X(n, p) \equiv U(k^{**}, n, p). \tag{6.5}$$

Back to the scenario where the pool from the original cohort tests positive, we used one test to learn that the k^*-sample pool contains positive samples. Thus the minimum expected number of tests for the original cohort is

$$Y(n, p) = 1 + X(k^*, p) + Y(n - k^*, p). \tag{6.6}$$

when the pool tests positive.

Equations (6.3) and (6.6) show that the optimization problem (6.1) can be solved

using a recursive algorithm with regard to n. To combine these two scenarios and derive the recursive formula, we need to calculate the probability of each scenario. Let $Q(p, k^*)$ denote the probability that a pool of k^* samples with the prevalence p is negative, i.e., all k^* samples in the pool are negative. Since the probability that a sample is negative equals $1 - p$, we derive

$$Q(p, k^*) = (1 - p)^{k^*}. \tag{6.7}$$

Combining the two outcomes from the test on the k^*-sample pool, we calculate the expected minimum number of tests as

$$Y(n, p) = Q(p, k^*)[1 + Y(n - k^*, p)] + [1 - Q(p, k^*)][1 + X(k^*, p) + Y(n - k^*, p)]. \tag{6.8}$$

To complete the recursive relationship, it remains to calculate $X(k^*, p)$ in equation (6.8). Similarly, let $Q_{\text{TRUNC}}(n, p, k^{**})$ denote the probability that a pool of k^{**} samples is negative, i.e., all k^{**} samples in the pool are negative, when pooling from a cohort of size n with prevalence p, conditioned on the fact that positive samples exist in the cohort. $Q_{\text{TRUNC}}(n, p, k^{**})$ is given by

$$Q_{\text{TRUNC}}(n, p, k^{**}) = \frac{(1 - p)^{k^{**}}[1 - (1 - p)^{n - k^{**}}]}{1 - (1 - p)^n}. \tag{6.9}$$

where $(1 - p)^{k^{**}}$ equals the probability that the pool is negative, and $1 - (1 - p)^{n - k^{**}}$ equals the probability that at least one of the unpooled samples is positive. Therefore, the probability that the pool is negative and the unpooled samples have at least one positive sample equals $(1 - p)^{k^{**}}[1 - (1 - p)^{n - k^{**}}]$, which is then conditioned on the probability that at least one sample in the cohort is positive which equals $1 - (1 - p)^n$.

When the pool tests positive, we do not know if the unpooled samples in the cohort contain positive samples. So the minimum expected number of tests required for the cohort is

$$X(n, p) = 1 + X(k^{**}, p) + Y(n - k^{**}, p). \tag{6.10}$$

When the pool is negative, we know for sure that the unpooled samples contain positive samples. So the minimum expected number of tests required for the cohort

is

$$X(n,p) = 1 + X(n - k^{**}, p). \tag{6.11}$$

Expected on these two scenarios, we can derive

$$\begin{aligned} X(n,p) = {} & Q_{\text{TRUNC}}(n,p,k^{**})[1 + X(n - k^{**}, p)] \\ & + [1 - Q_{\text{TRUNC}}(n,p,k^{**})][1 + X(k^{**}, p) + Y(n - k^{**}, p)]. \end{aligned} \tag{6.12}$$

In summary, equations (6.8) and (6.12) are used to recursively solve the optimization problem which outputs the optimal pool size for the current test. And the base cases of the recursive relationships are

$$X(0,p) = X(1,p) = Y(0,p) = 0; Y(1,p) = 1. \tag{6.13}$$

6.2 Updating the prevalence estimate

The optimization problem (6.1) discussed above depends on the prevalence parameter p, whose value needs to be updated based on the outcome of the previous test. Next, we discuss how the strategy updates the prevalence estimate based on the outcome of each pooled test.

We model the number of positive samples in a pool of size k as a random variable M following a binomial distribution with a prevalence parameter p. p follows a beta distribution with hyperparameters α and β. In particular, this is a beta-binomial hierarchical model

$$M|p \sim Binom(k,p), \tag{6.14}$$

$$p \sim Beta(\alpha, \beta), \tag{6.15}$$

where the posterior distribution of p has a closed-form solution

$$p|M \sim Beta(\alpha + M, \beta + k - M), \tag{6.16}$$

which can be used to update the information about prevalence based on the observed number of positive cases in the pool.

However, not every outcome from a pooled test provides enough information to

determine the value of M. When we test a pool of size k, there are only two scenarios where we can ascertain the value of M. First, if a pool tests negative, we know that $M = 0$ and thus can update $\beta \leftarrow \beta + k$. Second, since we can only identify a positive sample when it is tested individually, we only update $\alpha \leftarrow \alpha + 1$ when an individual tests positive, i.e., $M = k = 1$. These rules can be summarized by

$$\alpha \leftarrow \alpha + 1, \text{ if } M = k = 1, \tag{6.17}$$

$$\beta \leftarrow \beta + k, \text{ if } M = 0. \tag{6.18}$$

Before testing a cohort, we initiate α and β depending on the prior knowledge of the prevalence. In particular, we consider two cases. First, when there is weak evidence that $p = p^*$, we want to have small α and β that assure $E[p] = p^*$ in order to retain the prior information and allow fast updates of the parameters when testing begins. Therefore, we fix $\alpha = 1$ and calculate β such that $E[p] = p^*$, i.e.,

$$\alpha = 1, \beta = \frac{1 - p^*}{p^*}. \tag{6.19}$$

Second, when there is strong evidence that $p = p^{**}$ learned from N diagnostic samples, one can initialize α and β by

$$\alpha = Np^{**}, \beta = N(1 - p^{**}), \tag{6.20}$$

or just simply assign the number of identified positive samples from the N samples to α and the number of identified negative samples to β, since α and β are the pseudocounts of the positive and negative samples, respectively.

Finally, since our models require a specific value of p, we use the expected value as the value of p throughout the strategy, which is calculated as

$$E[Beta(\alpha, \beta)] = \frac{\alpha}{\alpha + \beta}. \tag{6.21}$$

6.3 Test procedure of the adaptive sample pooling strategy

Next, we introduce how our ADSP tests the whole cohort based on the statistical models discussed above. The entire test procedure consists of three processes that

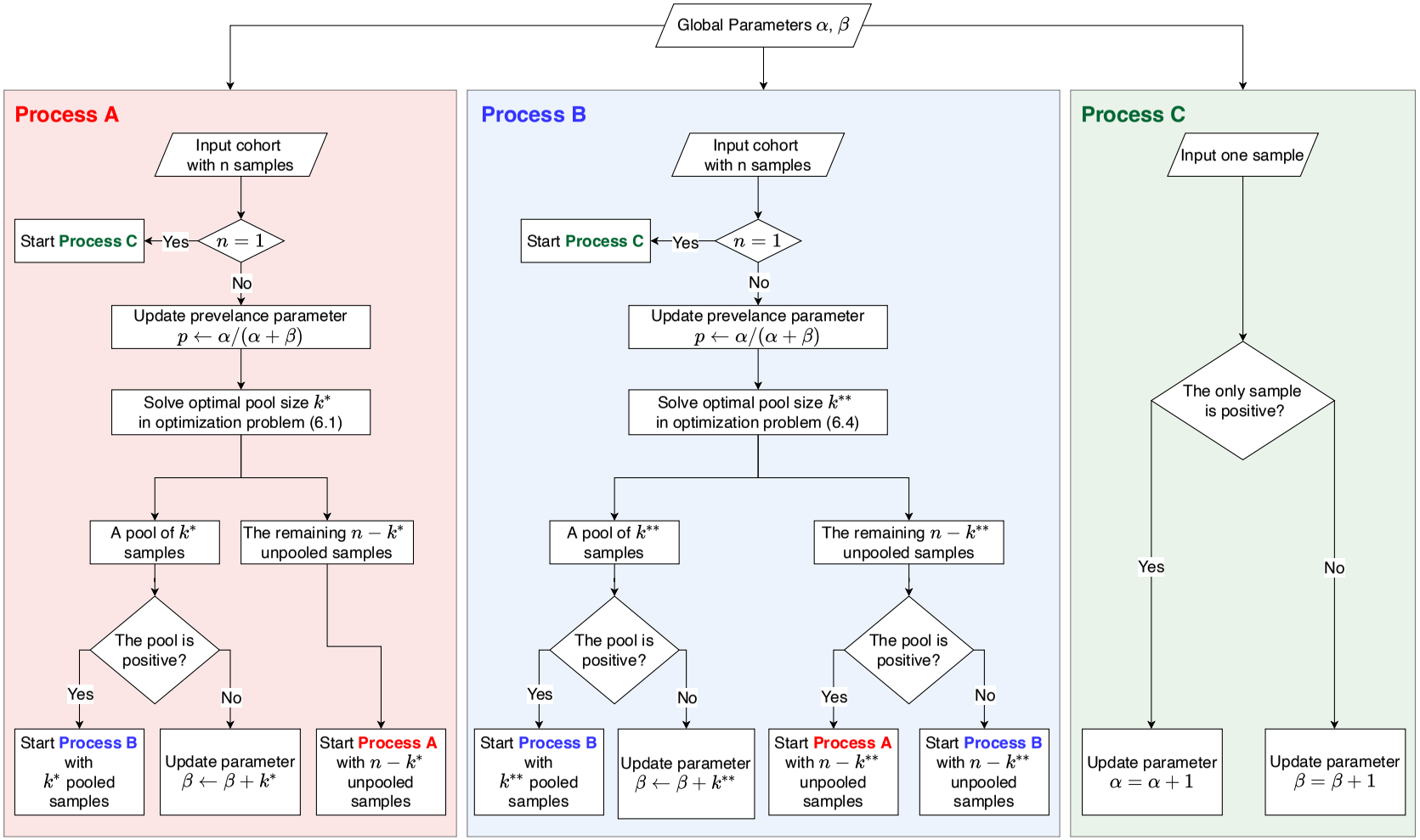

Figure 6.1: Flowchart of the adaptive sample pooling strategy. A cohort is first input into Process A. Depending on the test result of the pool, the samples are input into Process B or another Process A. The processes call each other recursively until an endpoint is reached.

recursively call each other, which are illustrated in Fig. 6.1.

Given a test cohort of n samples, we first initiate α and β using the abovementioned rules. We start the test procedure with Process A, splitting the cohort into two groups. One group is the pool to be tested, whose optimal size is calculated by the optimization problem (6.1). If the test outcome is positive, this pooled group will be further tested using Process B. The other group consists of the remaining samples not included in the pool, which will be used as the input for a new Process A. Process B is similar to Process A except for three major differences. First, the input test cohort of Process B is known to contain at least one positive sample. Second, Process B solves the optimization problem (6.4) to find the optimal pool size. Third, the unpooled samples of Process B can be further tested using either Process A or Process B, depending on the test outcome of the pooled group. Finally, when the cohort size is one, it is tested using Process C.

Note that at the end of each process, we update the value of the hyperparameters α and β if possible, which leads to the update of the prevalence parameter p to determine the optimal pool size for the next test.

6.4 Implementation of the adaptive sample pooling strategy

Before each test, our strategy needs to retrieve the optimal pool size by solving the optimization problem (6.1) or (6.4). With dynamic programming, we can reduce the time complexity of these problems to $O(n^2)$ where n is the cohort size, which can still take several minutes when n is large. When developing our software implementation of this strategy, we included two features to shorten the time for the user to obtain the optimal pool size. First, we precalculated the optimal pool sizes for a set of predefined values of n and p and saved the results in a table. Our software queries the table to instantly retrieve the optimal pool size for the values of n and p desired. Second, in practice, the optimal pool size for a specific p becomes stable when n is large. Therefore, we only calculated the optimal pool size with increasing n until the optimal pool size became stable. In general, smaller p requires larger n for the optimal pool size to become stable. For n past stability, we used the stable pool size to approximate the optimal pool size.

In particular, we precalculated the optimal pool sizes for prevalence from 0.0001

to 0.5 with a resolution of 0.0001. For each p, we calculated the optimal pool size iteratively with increasing n. We considered stability reached when a pool size was optimal for 900 of the last 1000 values of n. For the same combination of n and p, the optimization problem (6.1) can have multiple k that minimize $V(k, n, p)$. The same can be said for the optimization problem (6.4). Before stability, if multiple pool sizes were equally optimal, we chose the largest one. If multiple pool sizes fulfilled our stability criteria, we also chose the largest one as the stable pool size. If the pool size was still not stable when $n = 2000$, we stopped the process and used the pool size that appeared most frequently in the last 1000 values of n. We chose the largest size if there was a tie. With these implementations, our software provides instant queries for the optimal pool size without latency.

Chapter 7

Results

We illustrate the performance of our ADSP with simulation studies. In particular, we show that our ADSP outperforms two other pooling strategies, (1) Dorfman Pooling using the optimal pool size derived by Dorfman (Dorfman-Optimal) [7], and (2) Dorfman Pooling with an empirically decided fixed group size of 10 (Dorfman-10). We also illustrate the change in the prevalence estimate throughout the testing of a cohort using ADSP, as well as the fact that the optimal pool size learned from ADSP becomes stable when the cohort size is large.

In the simulation studies, we used cohorts of size 1000 since large cohort sizes have little impact on the optimal adaptive pool sizes. We generated cohorts with various prevalence ranging from 0.01 to 0.5 with steps of 0.01, which leads to different numbers of positive samples between cohorts. For each prevalence, we randomly generated 100 test cohorts. The number of positive cases was the expectation of a binomial random variable based on each particular prevalence. Each cohort was represented by a vector consisting of 1's and 0's, corresponding to the positive and negative samples, respectively. Samples were tested in order from the beginning to the end of each vector, either in pools or individually. Before the test procedure, we shuffled each vector so that the positions of the positive samples in each cohort were random.

Fig. 7.1 shows the number of tests required by the three pooling strategies when the initial prevalence estimate equals the true value. The dotted lines show the mean number of tests from the 100 cohorts for each prevalence. The shades represent the confidence intervals of two standard deviations from the mean. Note that Dorfman-

Optimal uses the pool size which minimizes the expected relative cost defined by

$$\frac{k+1}{k} - (1-p)^k, \tag{7.1}$$

where k is the pool size. (7.1) is greater than 1, $\forall k \in \mathbb{Z}^+$ when $p \geq 0.31$, suggesting that individual testing is better than Dorfman Pooling [7]. Therefore, when $p \geq 0.31$, we performed individual testing for Dorfman-Optimal to align with Dorfman's goal to minimize the expected relative cost.

The confidence intervals show that ADSP (the red shade) uses fewer tests than Dorfman-10 (the green shade) across all prevalence simulated, and uses fewer tests than Dorfman-Optimal (the blue shade) when prevalence is low. T-tests confirmed our observation that ADSP uses significantly fewer tests than Dorfman-10 across all prevalence tested, and also significantly fewer tests than Dorfman-Optimal when $p \leq 0.37$. When $p > 0.37$, our ADSP is worse than Dorfman-Optimal, which has shifted to individual testing. Therefore, 0.37 can be an upper limit on the prevalence for ADSP to be suitable.

Except for Dorfman-10, all pooling strategies require prevalence as input. However, it is impossible to have a perfectly accurate prevalence estimate before testing in real life. Therefore, we conducted sensitivity analyses on the initial prevalence estimate. To evaluate our strategy in both scenarios where the prevalence is correctly and incorrectly estimated, we considered four different misspecifications on the initial prevalence estimate. In particular, we offset the true prevalence by 50%, 20%, -20%, and -50% and used the value as the initial prevalence estimate. With each initial prevalence estimate, our ADSP initialized the hyperparameters α and β using the equation (6.19). We always compared ADSP with Dorfman-Optimal on the same initial prevalence estimate to ensure fairness.

Both ADSP and Dorfman-Optimal use pool sizes designed to minimize the number of tests required for a particular prevalence. Therefore, we expect both methods to use more tests when the initial prevalence estimate is misspecified. We let δ_{ADSP} and δ_{DO} denote the increase in the number of tests from using a misspecified instead of an accurate initial prevalence estimate in ADSP and Dorfman-Optimal, respectively. δ_{ADSP} and δ_{DO} indicate how heavily the performance of these methods is negatively affected by the misspecification in prevalence. We are interested in $\Delta \equiv \delta_{ADSP} - \delta_{DO}$, the difference between the effect of misspecification on these two methods. A negative Δ suggests that ADSP is more resilient to misspecification.

Fig. 7.2 shows the boxplots of Δ across different true prevalence and misspecification. The initial prevalence input is different for different misspecifications under the same true prevalence. Under each misspecification, Dorfman-Optimal shifts to individual testing either too early or too late, causing a huge gap in each panel except the bottom one, where Dorfman-Optimal never shifts to individual testing within the range of the prevalence simulated. We tested if Δ is different from zero using t-tests at the significance level of 0.05. Cyan boxes indicate Δ is significantly less than zero, and vice versa for the red boxes. There are many more cyan boxes than red boxes, which indicates that ADSP is more resilient to misspecification than Dorfman-Optimal in more scenarios. Also, with either sign, the misspecification with a magnitude of 50% has more cyan boxes than the misspecification with a magnitude of 20%, which indicates that the difference in resilience is even larger with larger misspecifications.

Fig. 7.3 illustrates the changes in the prevalence estimate in the process of testing a cohort using ADSP with a 20% misspecification on the initial prevalence estimate from the true prevalence. The black lines show the mean prevalence of the 100 simulations of each true prevalence, and the shades are the pointwise 95% confidence intervals of the prevalence estimate. In all scenarios tested, the prevalence estimate converges to the true prevalence as the test procedure progresses. The mean estimate becomes relatively stable after around 50 tests. The confidence intervals converge slower for higher prevalences, which is expected because our strategy produces smaller pools for higher prevalences. Since β increases each time by the pool size, it increases more slowly with small pools.

In Chapter 6, we discussed that the optimal pool size becomes stable when the cohort size increases. Fig. 7.4 shows the optimal pool sizes solved from the optimization problem (6.1) across different cohort sizes when the prevalence equals 0.1. A violin plot is shown for each size that appears at least twice. 8 is the most frequent optimal pool size and the stable pool size. Note that when a pool size other than 8 is optimal, there are usually multiple equally optimal pool sizes, with 8 being one of them. The fluctuation in the optimal pool size when n is large can be explained by the tail behavior caused by the optimal pool sizes for small cohorts. First, when n is smaller than the stable pool size, ADSP cannot pool with the stable pool size. Second, the optimization problems that ADSP solves can output solutions different from the stable pool size for small cohorts. For example, when $p = 0.1$ and $n = 10$, ADSP suggests pooling with size 5 instead of the stable pool size 8 because 5 minimizes the expected number of tests in this particular case. The variance in the optimal pool

size caused by this kind of behavior will be propagated into the optimal pool size for large cohorts because we calculated the optimal pool sizes with increasing n, causing the optimal pool size to fluctuate even when n is large.

Fig. 7.5 shows a screenshot of the dashboard application that we developed to guide users through ADSP. To start with, one inputs the cohort size and the prior information on the prevalence on the left panel. One can either input an initial prevalence estimate or the numbers of identified positive and negative cases. The main panel displays the pooling instruction, the progress bar, the groups that remained to be tested, and the testing history. One can see the result of each sample on the right panel. When testing is finished, one can also download the testing result by switching to the table view on the right panel.

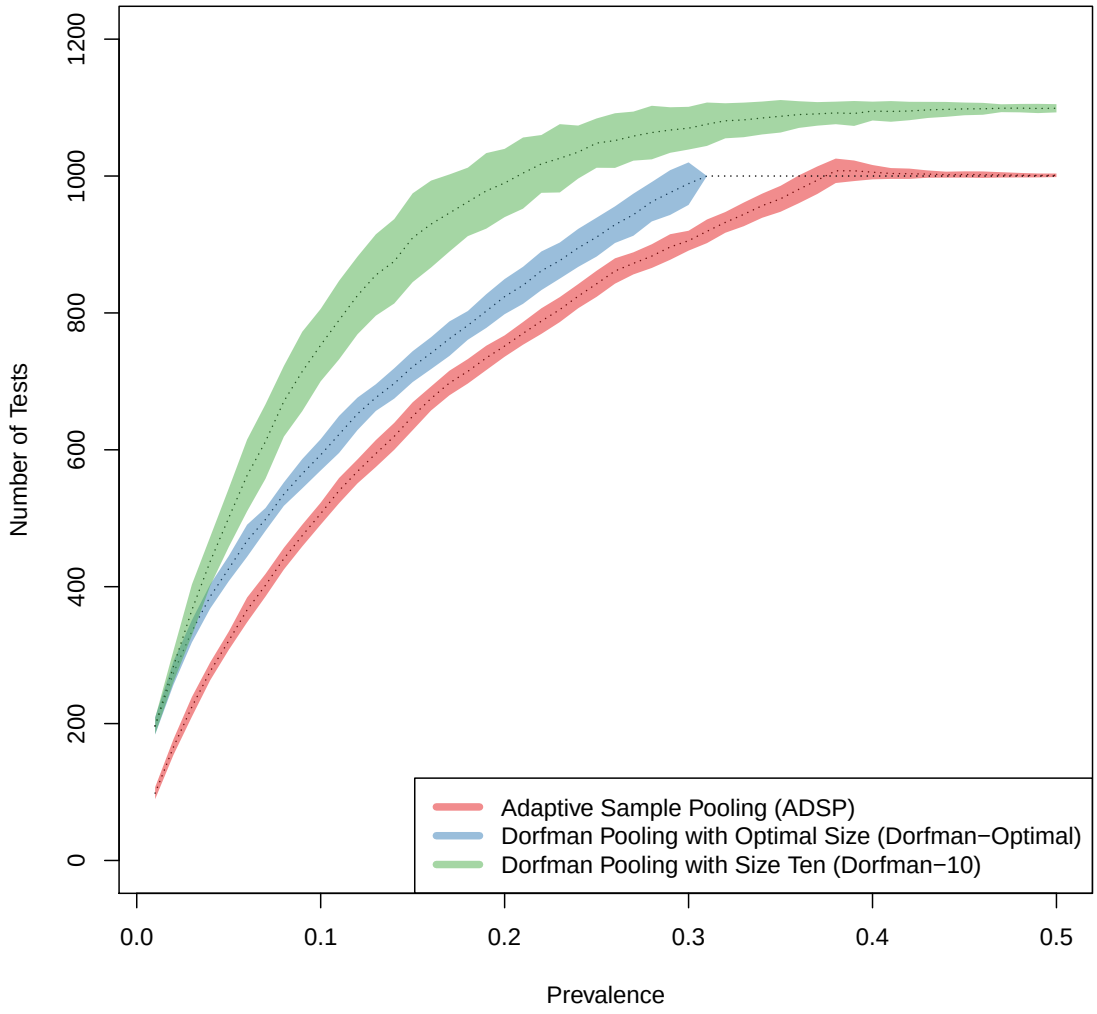

Figure 7.1: Number of tests required for different methods under different prevalence. The dotted lines represent the mean number of tests for each method based on the simulation. The shades represent the estimated pointwise 95% confidence intervals. Note that when prevalence is over 0.3, Dorfman Pooling with the optimal size is equivalent to individual testing.

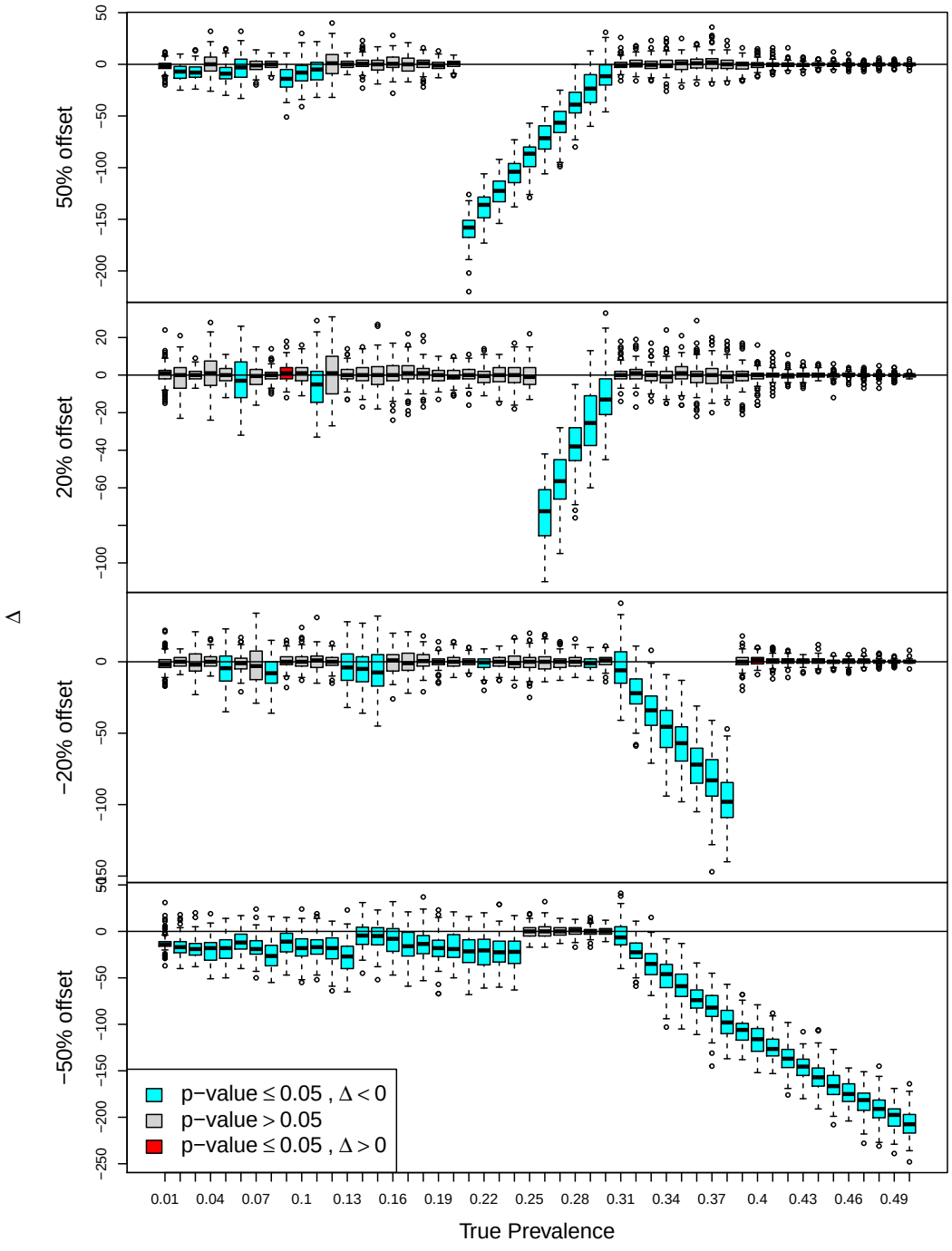

Figure 7.2: Difference in resilience to misspecified prevalence estimates between ADSP and Dorfman-Optimal. The cyan boxes indicate that Δ is significantly less than 0 (i.e., ADSP is more resilient to misspecification). The red boxes indicate that Δ is significantly greater than 0 (i.e., Dorfman-Optimal is more resilient to misspecification). The gray boxes indicate that Δ is not significantly different from 0.

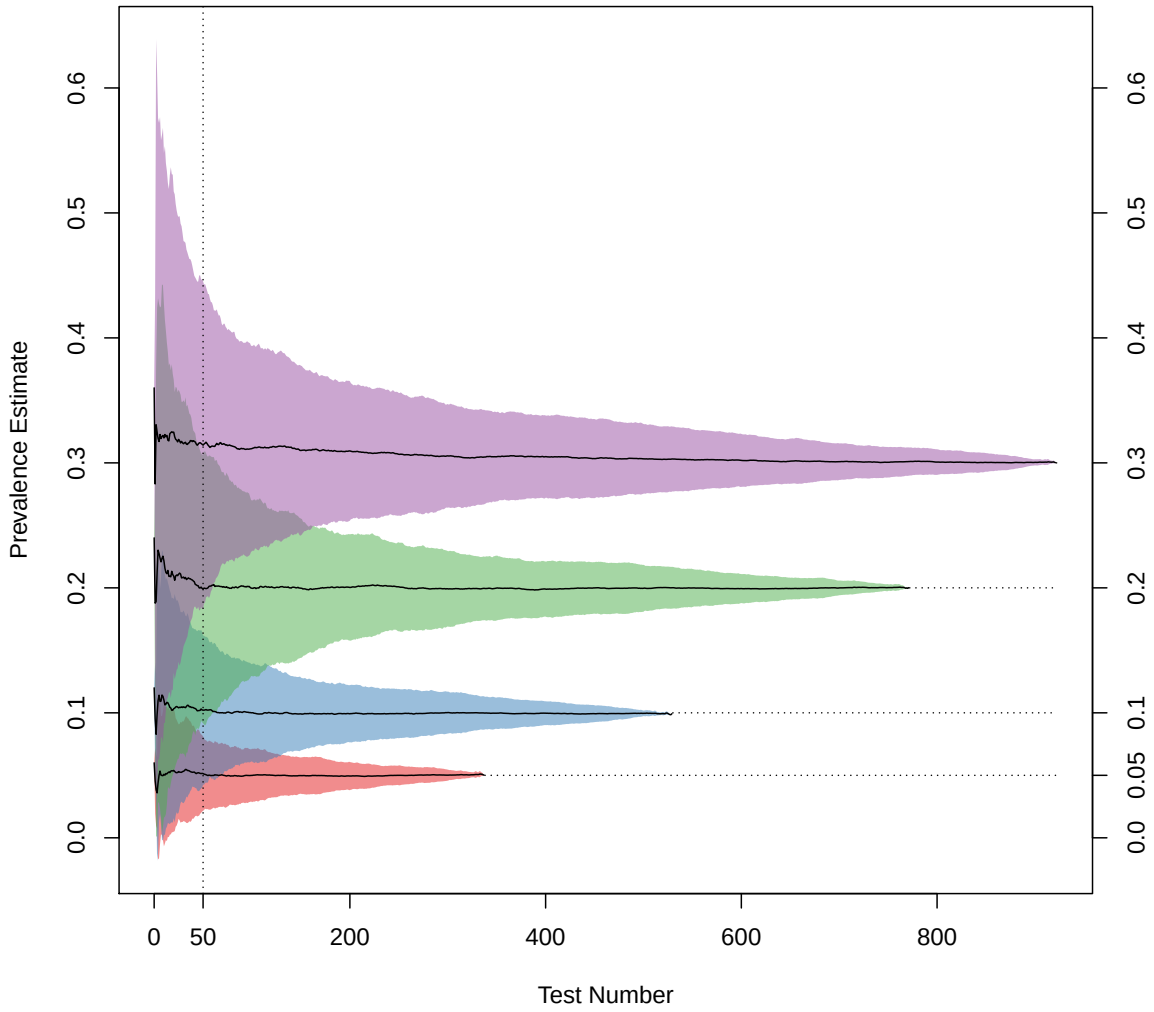

Figure 7.3: Changes in the prevalence estimate during ADSP with a 20% misspecification on the initial value. The black lines are the means of the prevalence estimate from the simulation. The shades represent the estimated pointwise 95% confidence intervals.

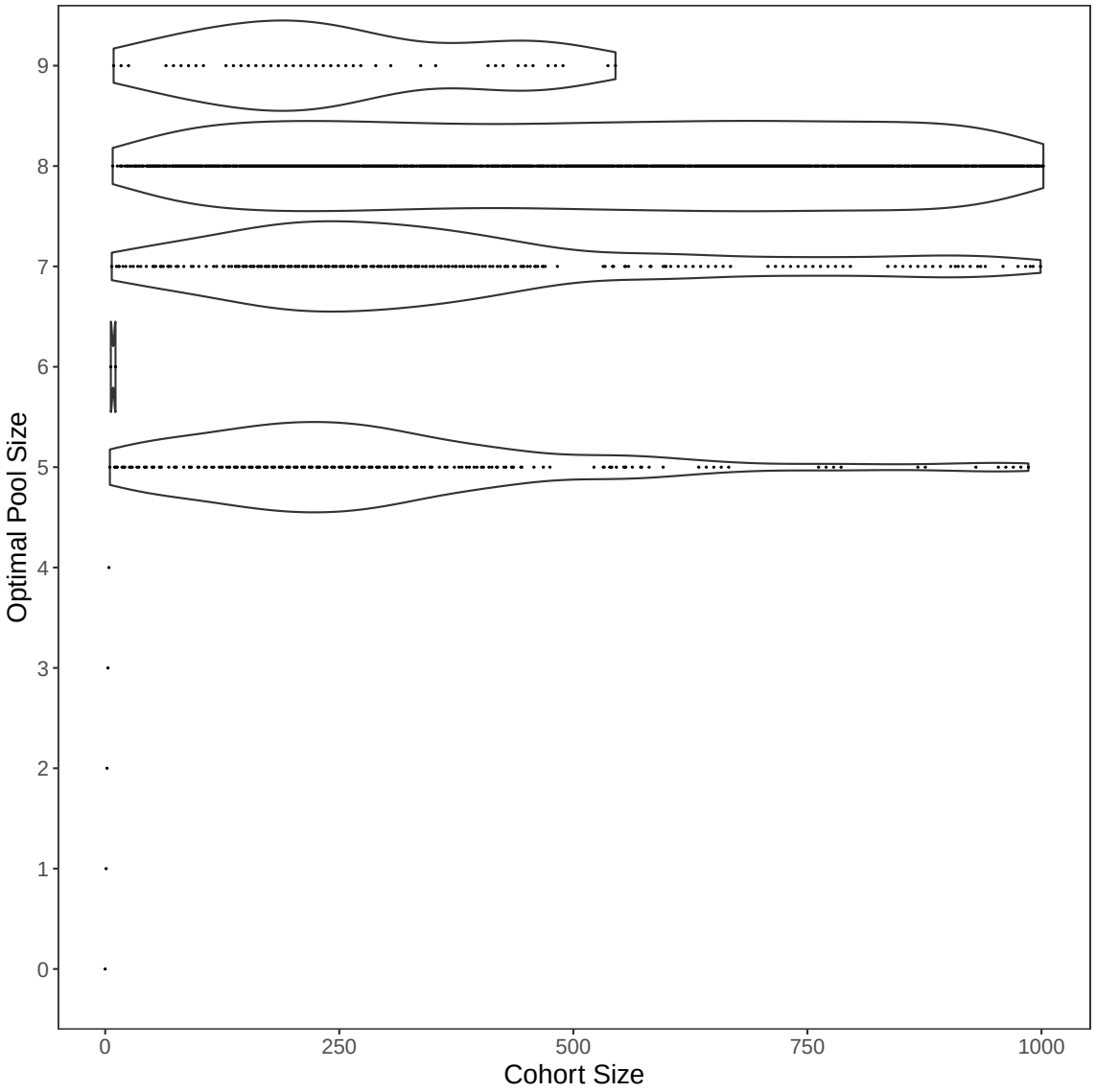

Figure 7.4: The optimal pool size for ADSP across different cohort sizes when the prevalence equals 0.1. Violin plots are shown for each pool size that appears at least twice.

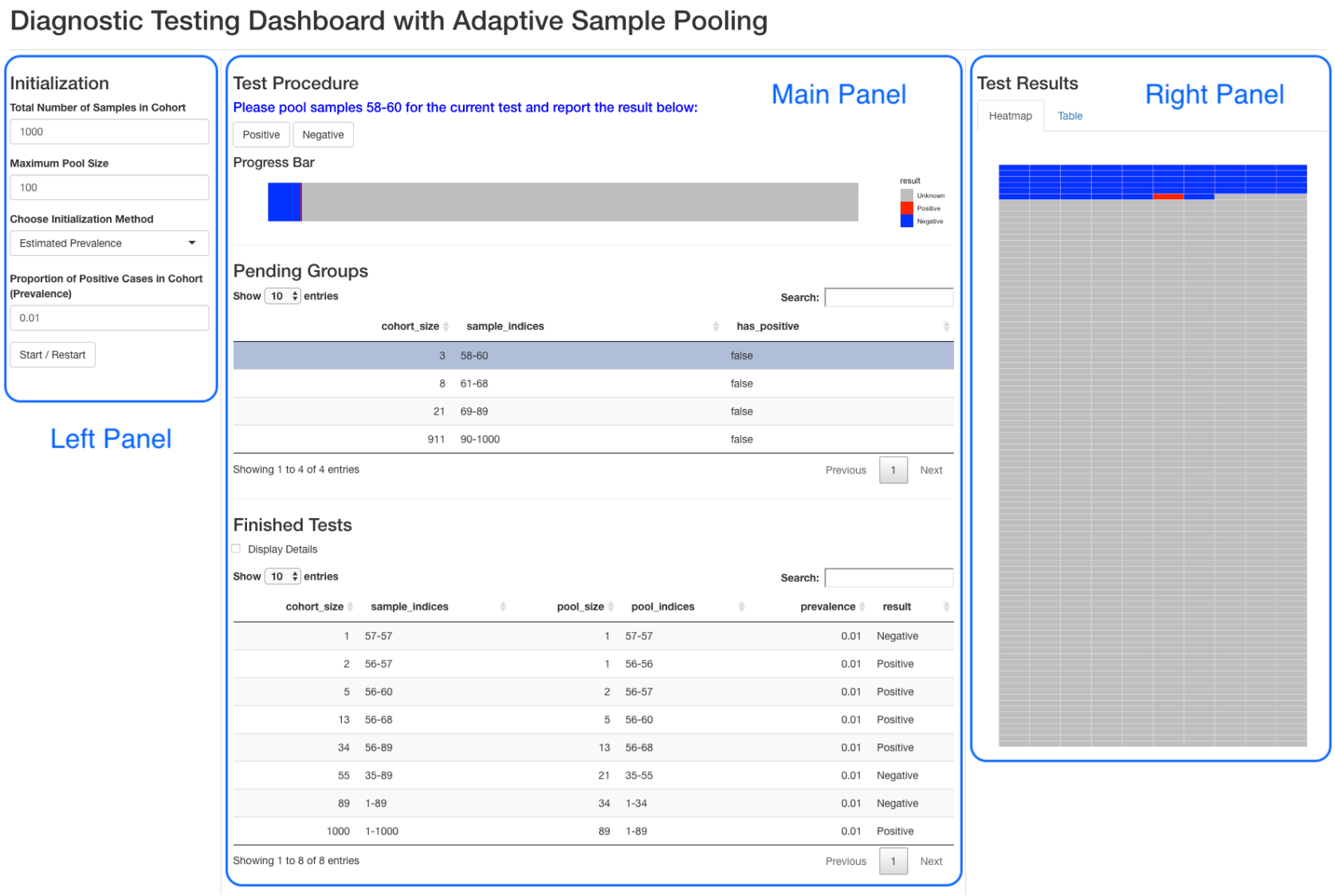

Figure 7.5: Illustration of the dashboard application for ADSP. The left panel initializes the parameters. The main panel displays the test procedure. The right panel shows the test results.

Chapter 8

Discussion

We have proposed an adaptive sample pooling strategy that updates the prevalence estimate after each test if possible, and uses the updated information to decide the pool size for the subsequent test. The simulation studies show that our ADSP uses fewer tests than Dorfman Pooling in many scenarios. We recommend using ADSP when the prevalence is lower than 0.38, and individual testing otherwise.

While other methods use a "single" prevalence value to determine the pooling scheme of the entire cohort, one crucial component of our strategy is the mechanism to dynamically update the prevalence estimate between tests, which allows inaccurate initial prevalence inputs without heavily affecting the method's performance. This mechanism has the potential to be incorporated into other pooling methods to improve their performance. For example, if the hypercube method tests a cohort in batches, our mechanism can be used to update the prevalence estimate between batches [19].

There are limitations to our strategy. For example, when the pool size is too big, positive samples can be diluted to the extent that their viral load cannot be detected. One can address this issue by setting a maximum pool size depending on the sensitivity of the particular test method. This feature is available in our dashboard.

Although this work is motivated by COVID-19 testing, our ADSP can also be applied in other diagnostic testing tasks. Our strategy has the potential to save testing resources (e.g., medical devices, workforce, and time) as fewer tests are needed to diagnose a cohort. With our dashboard application and R package, clinicians can implement ADSP in their diagnostic testing with ease.

Appendix A

Supplementary Document: The Impact of Lockdown Timing on COVID-19 Transmission across US Counties

A.1 Modeling lockdown effect using segmented regression

Figure 3.3 shows that the observed relationship between the first FPC scores and the timing of lockdowns was non-linear: its appearance was that of a "hockey stick" with an inflection point indicating a significant change in its slope. Thus, we used segmented regression to model this relationship and to ascertain the inflection point for the lockdown variable. We defined T_i as the difference in days between the date on which the county experienced at least five cumulative cases of COVID-19 and the date on which the county first initiated a lockdown. We denote β as the inflection point, which is an unknown model parameter requiring estimation. We derived the following three new variables from the lockdown information:

$$X_{i1} = \mathbb{1}[\text{Lockdown}]$$

$$X_{i2} = X_{i1} \cdot T_i$$

$$X_{i3} = X_{i1} \cdot (T_i - \beta) \cdot \mathbb{1}[T_i \geq \beta]$$

and defined the segmented model as:

$$\xi_{i1} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} \tag{A.1}$$

where X_{i1} (*Lockdown Indicator*) indicates whether a lockdown was implemented for the i^{th} county, which enabled the model to incorporate data from counties without a lockdown. X_{i2} (*Lockdown Time*) was the same as the lockdown time T_i if lockdown was implemented by the i^{th} county; otherwise, it was assigned a value of 0. X_{i3} (*Lockdown Time after Inflection Point*) indicates the date of the lockdown of the i^{th} county relative to the inflection point. The model (A.1) is equivalent to three models under different conditions as below:

1. When a lockdown was not implemented in a county, i.e., $X_{i1} = 0$, model (A.1) simplified into a constant $\xi_{i1} = \alpha_0$ where the first FPC score is modeled as a constant that is not related to the lockdown timing.

2. When a lockdown was implemented in a county before the inflection point, time β, i.e., $X_{i1} = 1$ and $T_i < \beta$, then model (A.1) simplified into a linear model of T_i
$$\xi_{i1} = (\alpha_0 + \alpha_1) + \alpha_2 T_i \tag{A.2}$$

3. When a lockdown was implemented in a county after the inflection point, time β, i.e., $X_{i1} = 1$ and $T_i > \beta$, then the model (A.1) simplified into another linear model of T_i with a different slope and intercept

$$\xi_{i1} = (\alpha_0 + \alpha_1 - \alpha_3\beta) + (\alpha_2 + \alpha_3) \cdot T_i \tag{A.3}$$

Note that when a lockdown was implemented exactly at the inflection point in a county, i.e., $X_{i1} = 1$ and $T_i = \beta$, then models (A.2) and (A.3) become the same model, which ensures that the different slopes in the segmented regression models are well connected at the inflection point. Based on these models, the coefficients can be explained by the following: a_2 (lockdown slope before inflection point) denotes the effect of the lockdown timing when implemented before the inflection point, and a_2+a_3 (lockdown slope after the inflection point) represents the effect of the lockdown timing when implemented after the inflection point. The values of model parameters $(a_0, a_1, a_2, a_3, \beta)$ can be estimated using a profile likelihood approach.

A.2 Interpretation of the first FPC scores

The result of a Functional Principal Component (FPC) Analysis on the trajectories of COVID-19 spread across 3,112 US counties shows that the first FPC explained a vast majority of the total variance (about 92.86%). Hence, we approximated the original case trajectory of the i^{th} county using the first FPC, and model (2.1) becomes

$$log(Q_i(t)) \approx \mu(t) + \xi_{i1}\phi_1(t) \tag{A.4}$$

Note that the functions $\mu(t)$ and $\phi_1(t)$ do not depend on the county index i. So, we represented the COVID-19 case trajectory in the i^{th} county, $Q_i(t)$, with a single score ξ_{i1}, i.e., the first FPC score is a surrogate variable for the trajectory of each county. Based on model (2.1) and the shape of the first eigenfunction, $\phi_1(t)$, we interpret the first FPC score as two concepts more friendly to the our readers: the weighted average of COVID-19 case counts and the weighted changes in the rate of COVID counts over time (on an exponential scale), with weights based on the first eigenfunction. Or more generally, we can use the first FPC score to describe the overall severity of the pandemic for the i^{th} county.

When we approximate the original case trajectory of the i^{th} county using the first FPC, model (2.1) can be rewritten as $\xi_{i1} \approx log(Q_i(t))/\phi_1(t) - \mu(t)/\phi_1(t)$. Since this score does not vary with time, we interpret the first FPC score, ξ_{i1}, as the weighted average of all log case counts on the i^{th} trajectory, $log(Q_i(t))$, with time-related weights $1/\phi_1(t)$. Since the day-1 of all of these trajectories are consistently defined as the first day with at least 5 cumulative cases, all trajectories have roughly the same starting points. So, weighted average case counts are tantamount to weighted changes in the rate of COVID counts over time (on an exponential scale).

Specifically, Figure A.10 shows that the first eigenfunction is always positive for the time range that was evaluated. Thus, a higher first FPC score ξ_{i1} associates with a higher case count across the time range. Note that the values of $\phi_1()$ changes across time, which means that the contribution of $\phi_1()$ to the case counts varies with time. Hence, we call it as a "weighted" average.

In short, since the first FPC score, ξ_{i1}, represents the strength of this pattern in the i^{th} county, a county with a higher first FPC score will have more cumulative cases on any given day, compared to another county with a lower first FPC score.

A.3 Interpretation of fitted Elastic net models

We used two-step modelling, which consisted of 1) unsupervised machine learning (FPC analysis) for dimensional reduction and derivation of surrogate variables (FPC scores), and 2) supervised machine learning (elastic net) to determine the association between risk factors and the FPC scores. When $m = 1$, and by combining these two models, we obtain:

$$log(Q_{ij}) = \mu(t_j) + (\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 X_{i3} + \alpha_4 X_{i4} + \cdots + \alpha_p X_{ip}) \cdot \phi_1(t_j) \quad \text{(A.5)}$$

A two-step relationship can be built from this:

1. For the n^{th} predictor X_{in} in model (A.5), if its value increases by S, after adjustment of all other variables, the first FPC score increases by $\alpha_n S$.

2. An increase of the first FPC score by $\alpha_n S$ will then lead to COVID-19 cumulative case count, Q, in model (2.1), multiplied by $exp(\alpha_n S \phi_1(t_j))$ on the j^{th} day after the county reports at least 5 cumulative cases.

In brief, when a risk factor X_{in} increases by S, the daily cumulative case count of the j^{th} day is multiplied by $exp(\alpha_n S \phi_1(t_j))$.

Table A.1: Definition of variables from the American Community Survey and the Oxford Covid-19 Government Response Tracker

Variable	**Definition**
lockdown	Days between the date on which the county experienced at least five cumulative cases of COVID-19 and the date on which the county first initiated a lockdown
Variables below are from the American Community Survey 5-Year Data (2019)	
Socioeconomics	
med_income	Median family income in the past 12 months
gini	Gini index of income inequality
Demographics	
pop	Total population
pop_density	Total population divided by land area
male	Male population divided by the total population
Health Insurance	
private_ins	Proportion of the population with private health insurance
public_ins	Proportion of the population with public health insurance
Household Composition	
median_age	Median age of the population
Ethnicity	
whites	Proportion of Whites in the total population
african_americans	Proportion of African Americans in the total population
natives	Proportion of American Indians and Alaska natives in the total population
asians	Proportion of Asians in the total population
Geographical Mobility and Mode of Transportation	
public_trans	Proportion of workers going to work by public transports
same_county	Proportion of people who moved within same county in the past year
Variables below are from the Oxford Covid-19 Government Response Tracker Calculated as the days from the date on which the county experienced at least five cumulative cases of COVID-19 to the date of the initiation of the policy	
Debt/Contract Relief	Relief for certain type and/or all types of contract
Public Information Campaigns	Public information or coordinated public campaign on COVID-19
Testing Policy	Accessibility of testing for people with symptom or everyone
Contact Tracing	Contact tracing after certain or all confirmed cases

Facial Coverings	Facial coverings are required in specific spaces or all the time outside home
Vaccination Policy	Availability of vaccine of particular/all groups of people
Protection of Elderly People	Isolation and hygiene requirements for the elders in long term care facilities and/or at home

Table A.2: Mean and 95% Confidence Interval of Elastic Net Coefficients (original scale)

Variable	**Low**	**Mean**	**High**
Lockdown Indicator	-3.48	-1.11	1.25
Lockdown Slope before Inflection Point	-0.0523	-0.0143	0.0238
Lockdown Slope after Inflection Point	0.865	1.05	1.23
Total Population	3.90E-06	1.63E-05	2.87E-05
Population Density	-0.00183	4.50E-06	0.00184
Median Age	-1.07	-0.855	-0.638
Median Family Income	7.30E-05	1.55E-04	2.38E-04
Gini Index	0.233	16.9	33.6
Proportion of Male	-41.2	-16.7	7.75
Proportion of Whites	-16.8	-3.06	10.6
Proportion of African Americans	-9	4.29	17.6
Proportion of Natives	-26.4	-10.1	6.26
Proportion of Asians	-27.4	16.9	61.1
Proportion of Individuals who Used Public Transport	-34.7	26.9	88.5
Proportion who Moved within the Same County	79	104	130
Proportion with Private Health Insurance	-16.6	-2.52	11.6
Proportion with Public Health Insurance	-6.31	12.7	31.6
Debt/Contract Relief	-0.0473	0.0859	0.219
Public Information Campaigns	-0.0329	-0.00848	0.016
Testing Policy	-0.123	-0.0304	0.0623
Contact Tracing	-0.0285	-0.00566	0.0172
Facial Coverings	-0.00234	0.00263	0.00759
Vaccination Policy	-0.095	-6.52E-04	0.0937
Protection of Elderly People	-0.00548	-0.0011	0.00327

Table A.3: 10 Counties with the Highest First FPC Scores

Location	**First FPC Score**
Los Angeles, California, US	94.979
Cook, Illinois, US	90.646
Miami-Dade, Florida, US	89.251
Queens, New York, US	86.226
Maricopa, Arizona, US	86.111
Kings, New York, US	83.867
Bronx, New York, US	79.431
Harris, Texas, US	78.206
Suffolk, New York, US	78.074
Nassau, New York, US	76.759

Table A.4: 10 Counties with the Lowest First FPC Scores

Location	**First FPC Score**
Haines, Alaska, US	-58.666
Sioux, Nebraska, US	-56.959
Wrangell, Alaska, US	-54.266
McPherson, Nebraska, US	-53.248
Forest, Pennsylvania, US	-52.422
Hamilton, New York, US	-52.364
Woodson, Kansas, US	-52.252
Kenedy, Texas, US	-52.008
Cheyenne, Colorado, US	-51.853
Petersburg, Alaska, US	-51.259

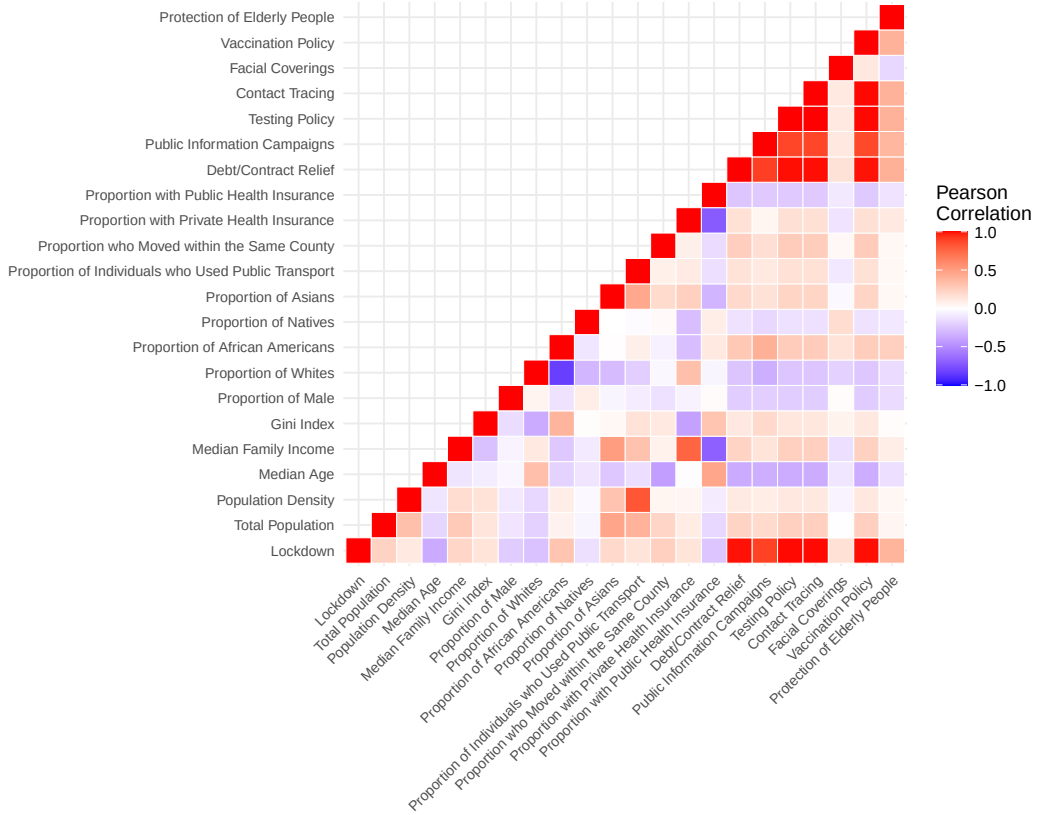

Figure A.1: Heatmap of Pearson Correlation between variables

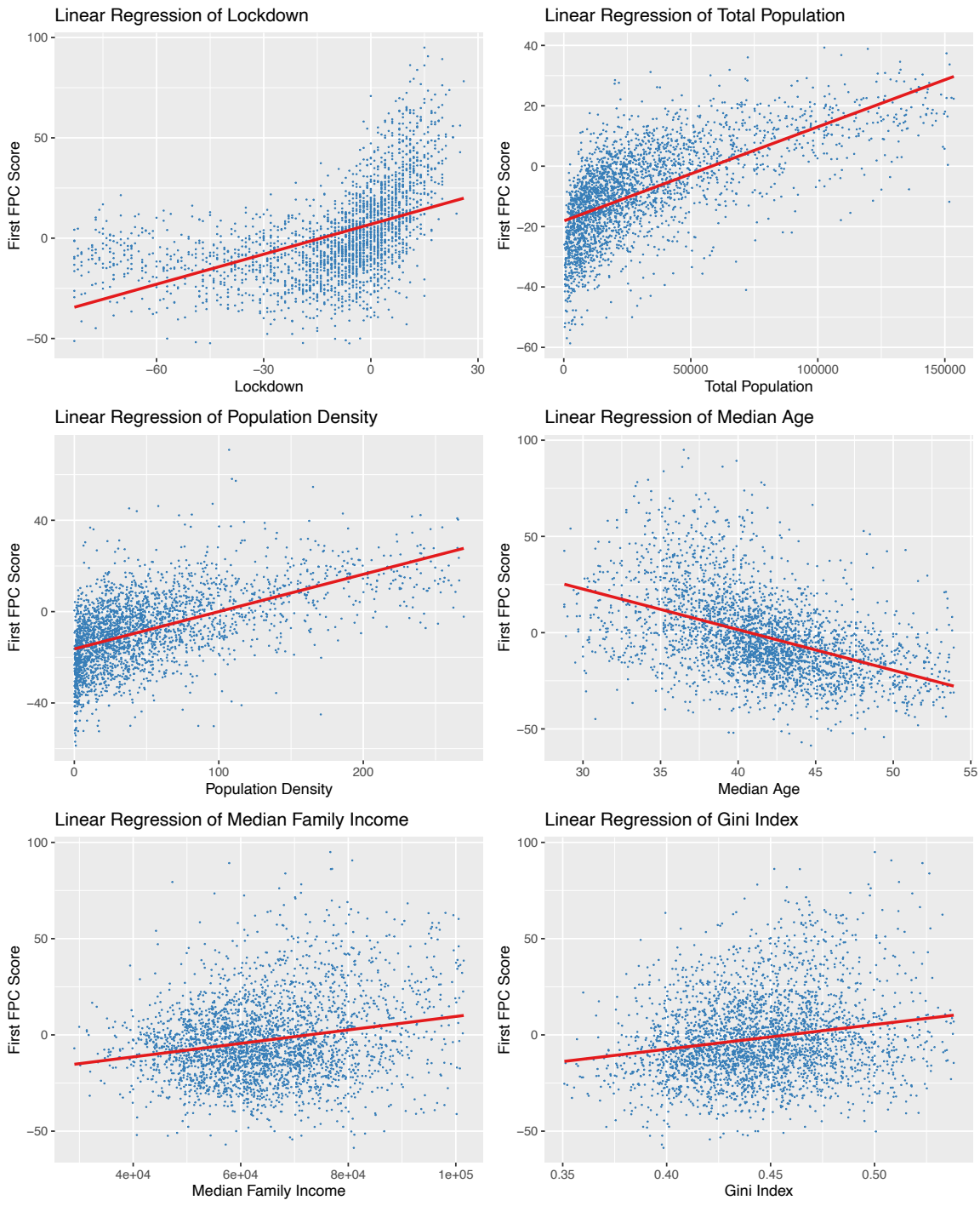

Figure A.2: Plots of the first FPC scores versus lockdown timing and the factors from ACS: total population, population density, median age, median family income, and Gini index. The blue dots show the values of all the US counties studied. The red lines display the fitted linear regression lines.

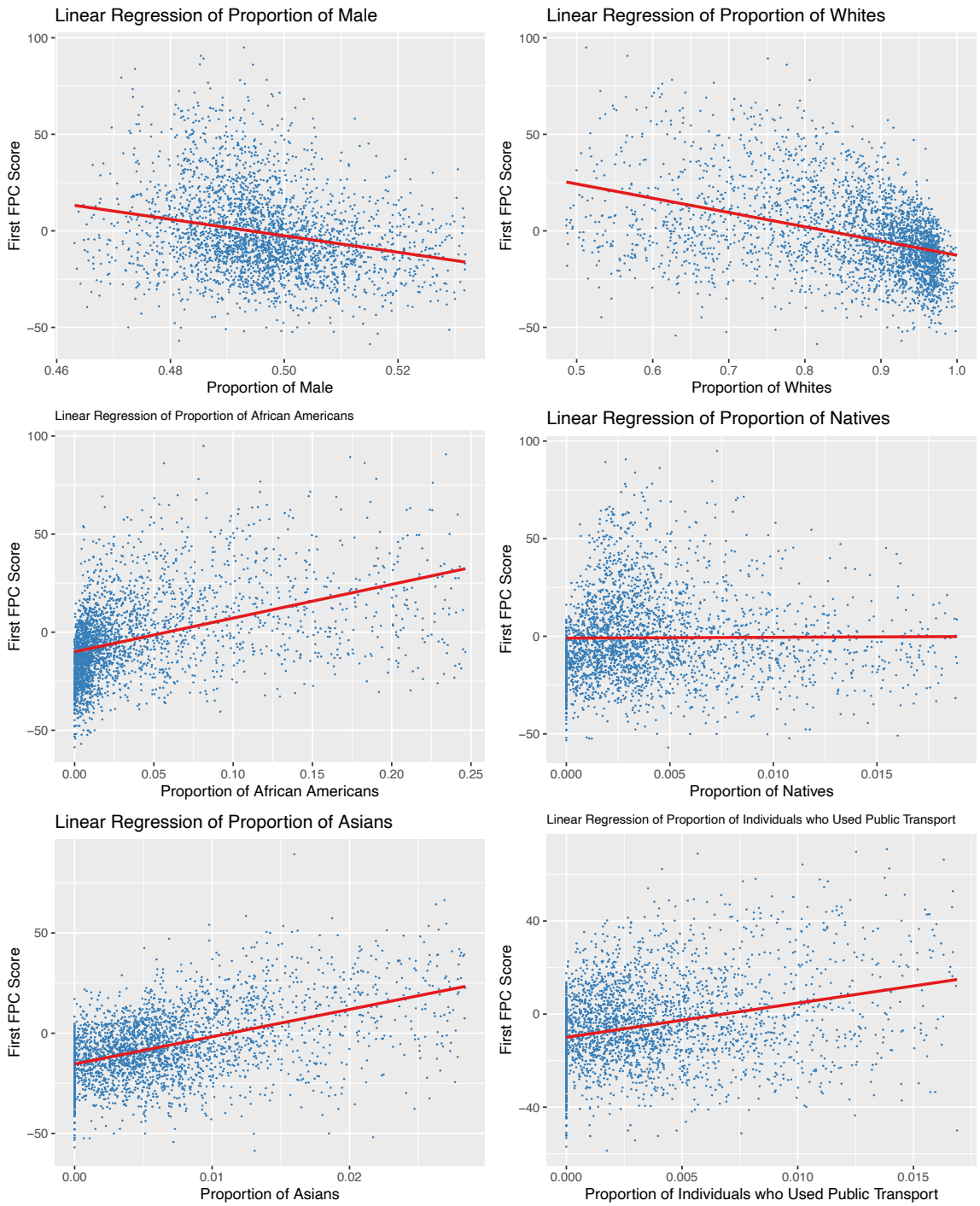

Figure A.3: Plots of the first FPC scores versus the factors from ACS: proportion of male, proportion of Whites, proportion of African Americans, proportion of Natives, proportion of Asians, and proportion of individuals who used public transport. The blue dots show the values of all the US counties studied. The red lines display the fitted linear regression lines.

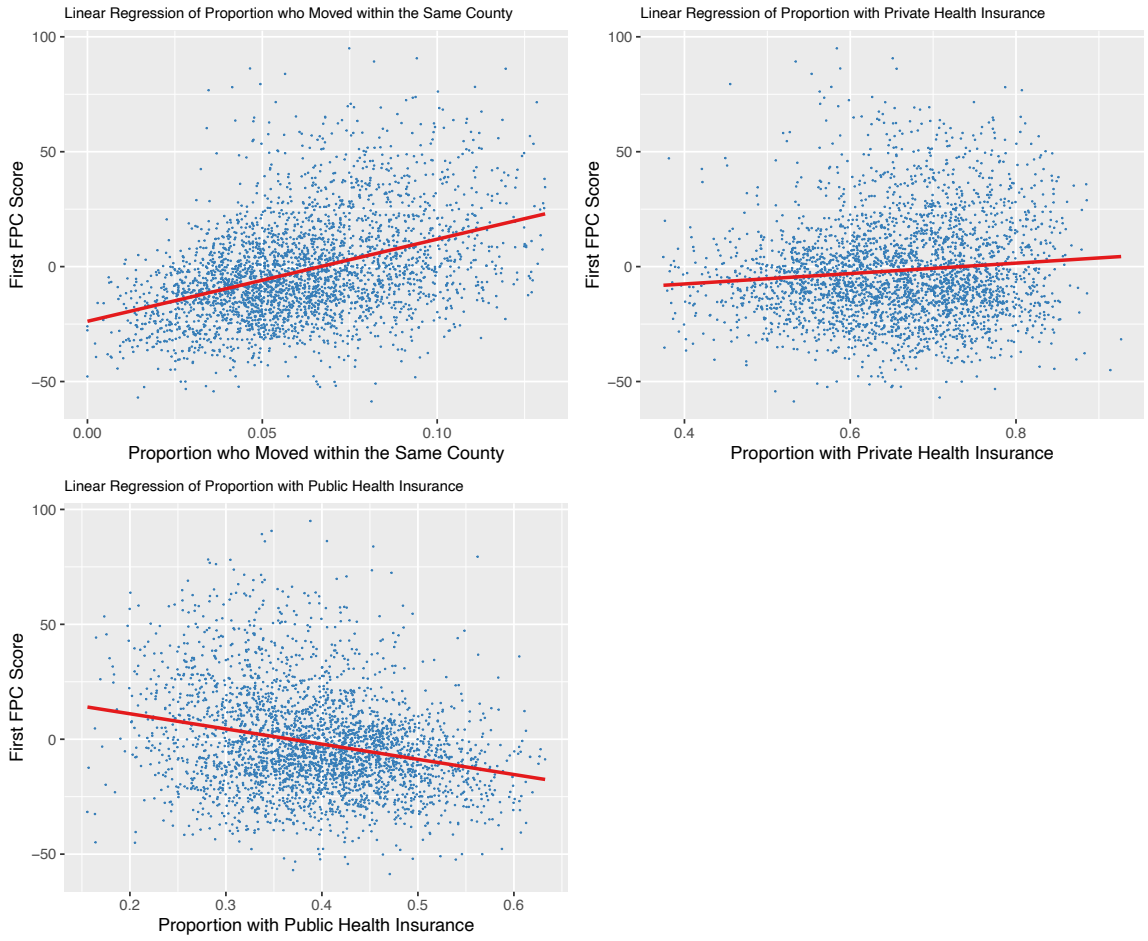

Figure A.4: Plots of the first FPC scores versus the factors from ACS: proportion who moved within same county, proportion with private health insurance, and proportion with public health insurance. The blue dots show the values of all the US counties studied. The red lines display the fitted linear regression lines.

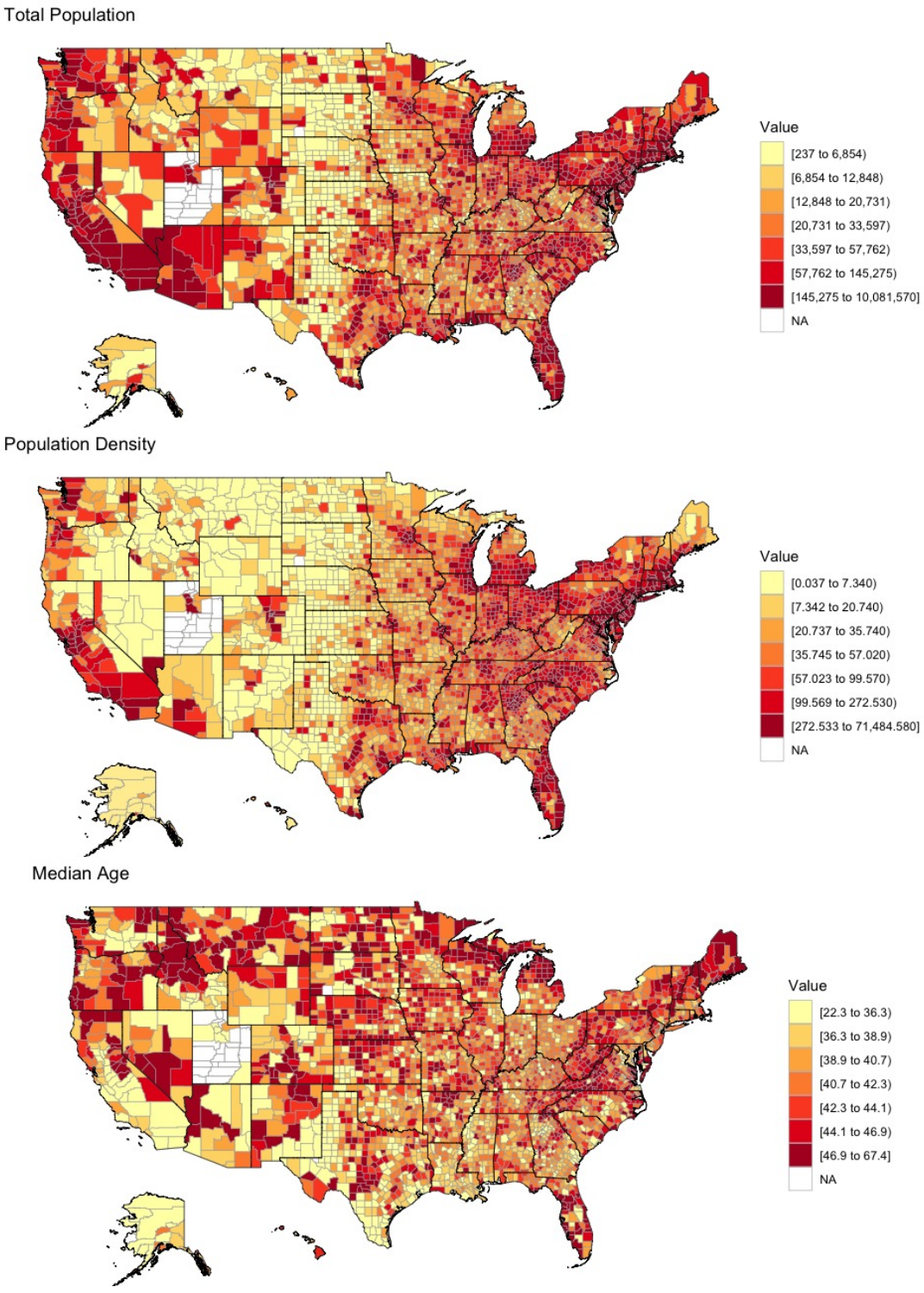

Figure A.5: Heatmap of Total Population, Population Density, and Median Age

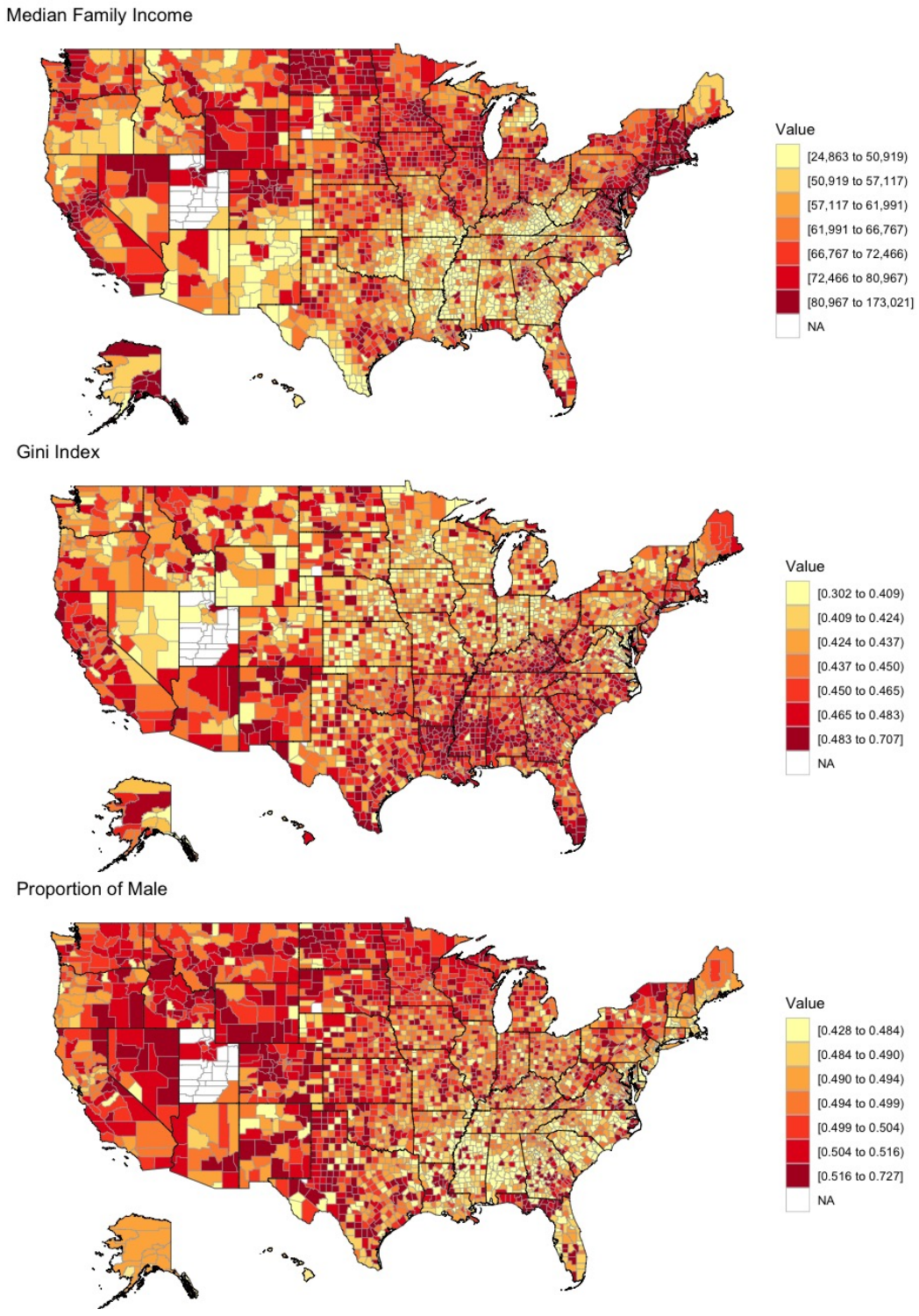

Figure A.6: Heatmap of Median Family Income, Gini Index, and Proportion of Male

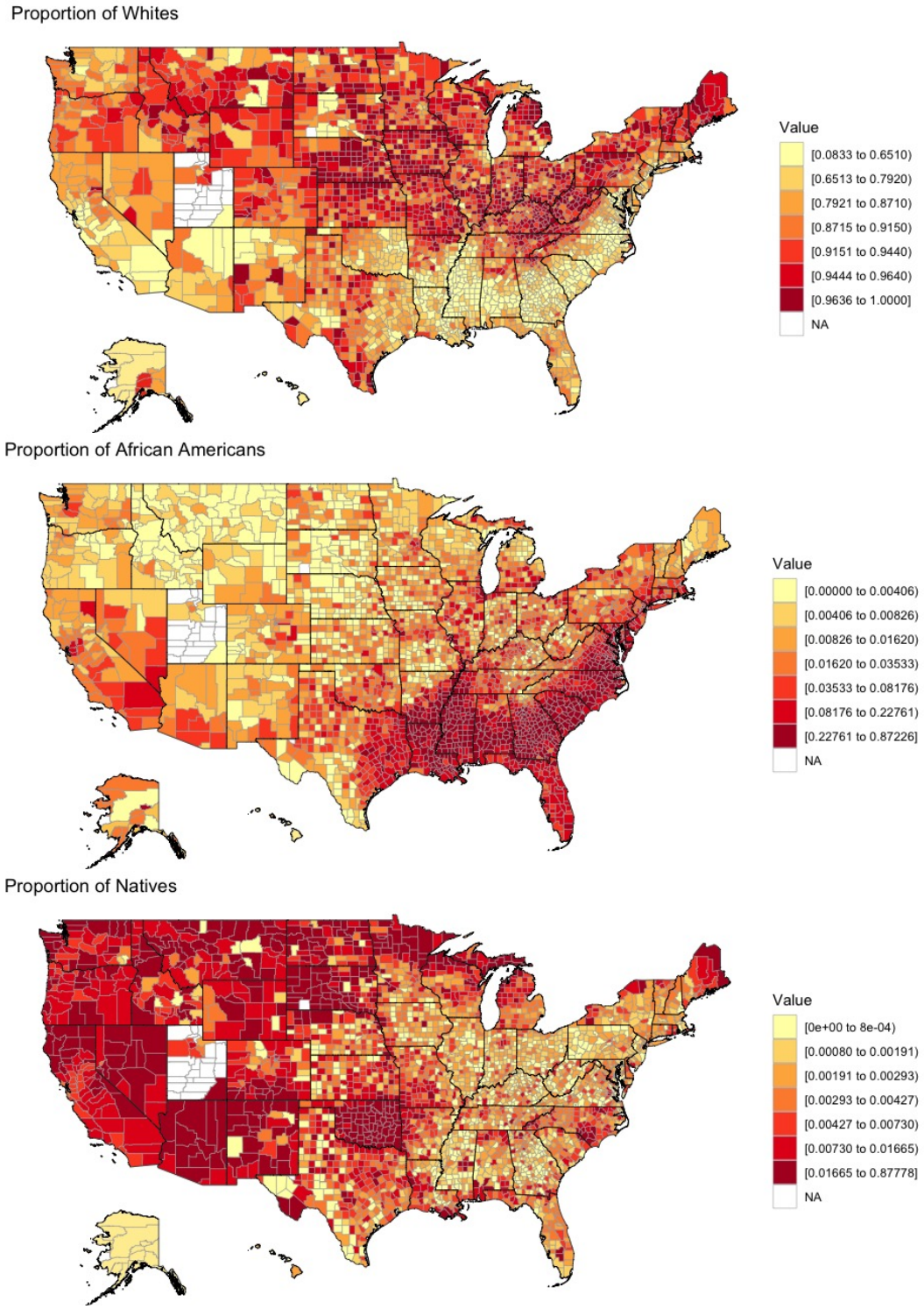

Figure A.7: Heatmap of Proportion of Whites, Proportion of African Americans, and Proportion of Natives

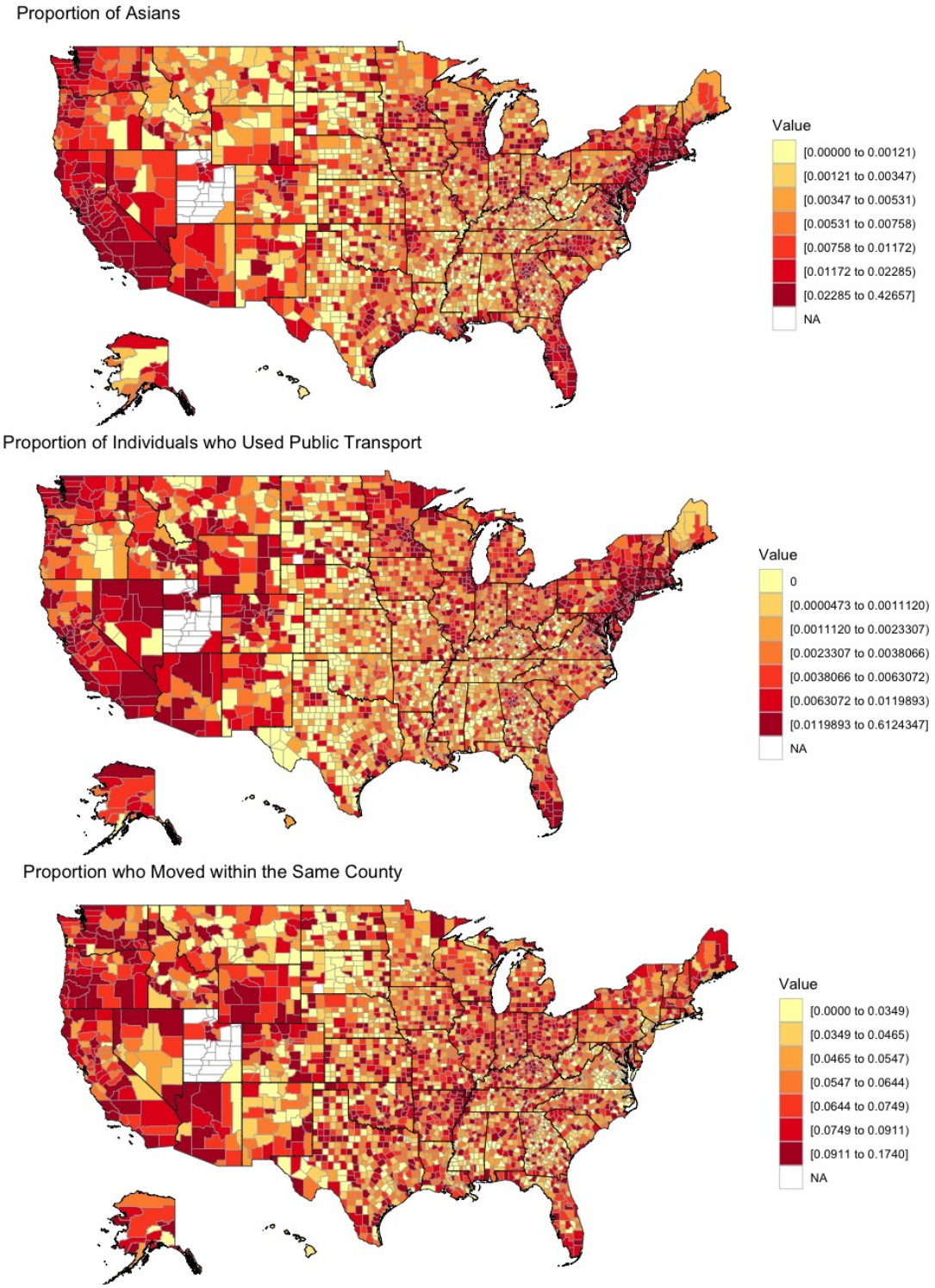

Figure A.8: Heatmap of Proportion of Asians, Proportion of Individuals who Used Public Transport, and Proportion who Moved within the Same County

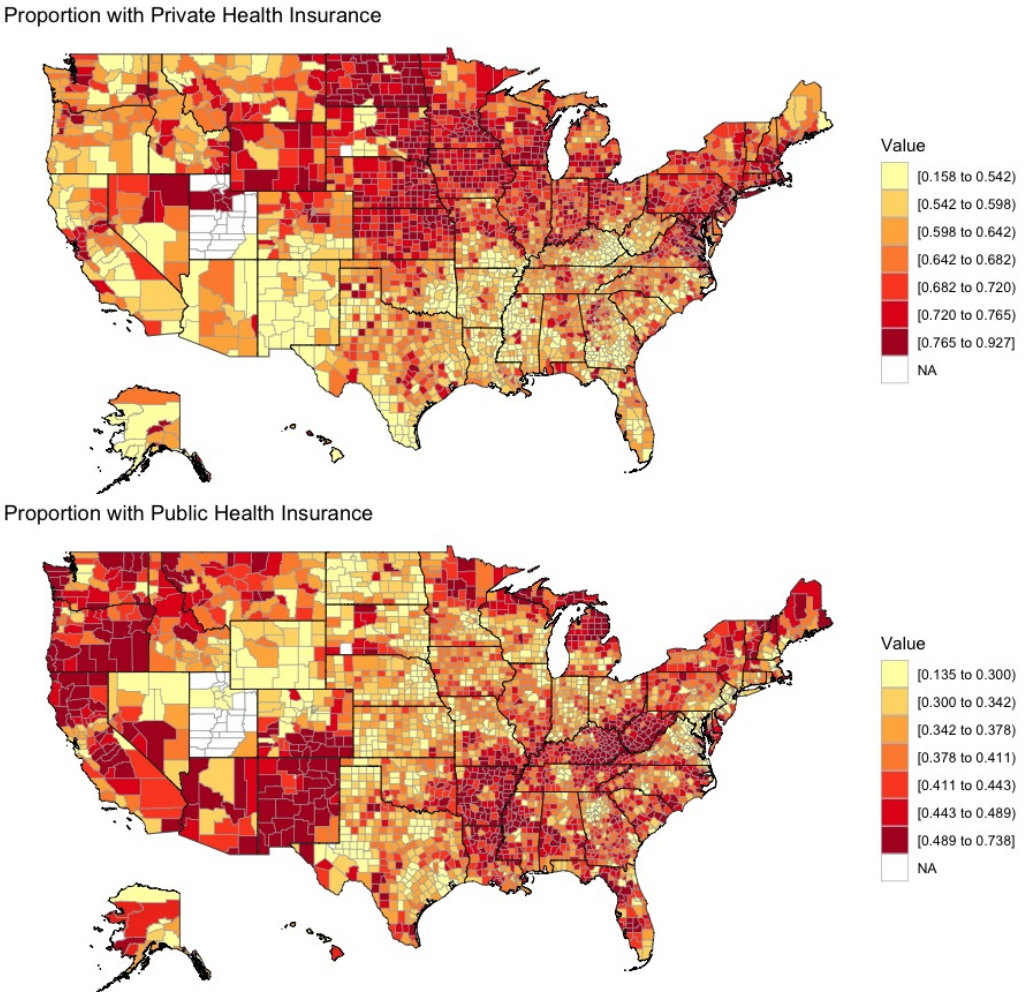

Figure A.9: Heatmap of Proportion with Private Health Insurance and Proportion with Public Health Insurance

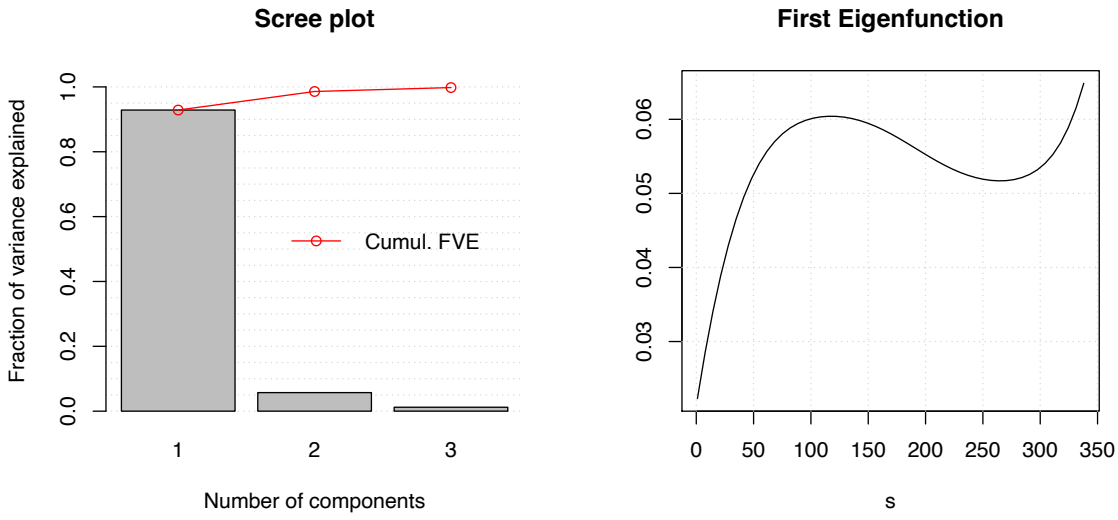

Figure A.10: Scree Plot on the left shows the cumulative proportions of variance explained by the first 1, 2, and 3 FPCs; the plot on the right shows the shape of the First Eigenfunction, i.e., the first FPC $\phi_1(t)$.

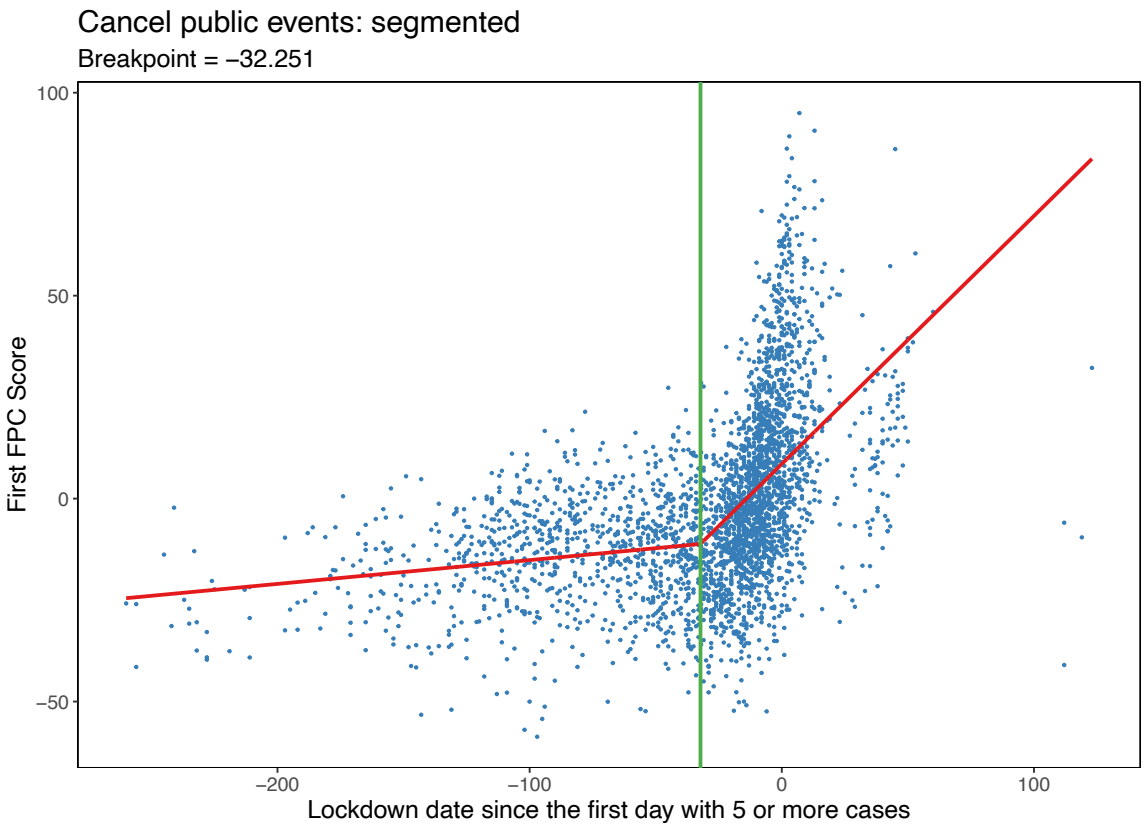

Figure A.11: Segmented Regression of the first FPC score vs the timing of Cancel Public Events

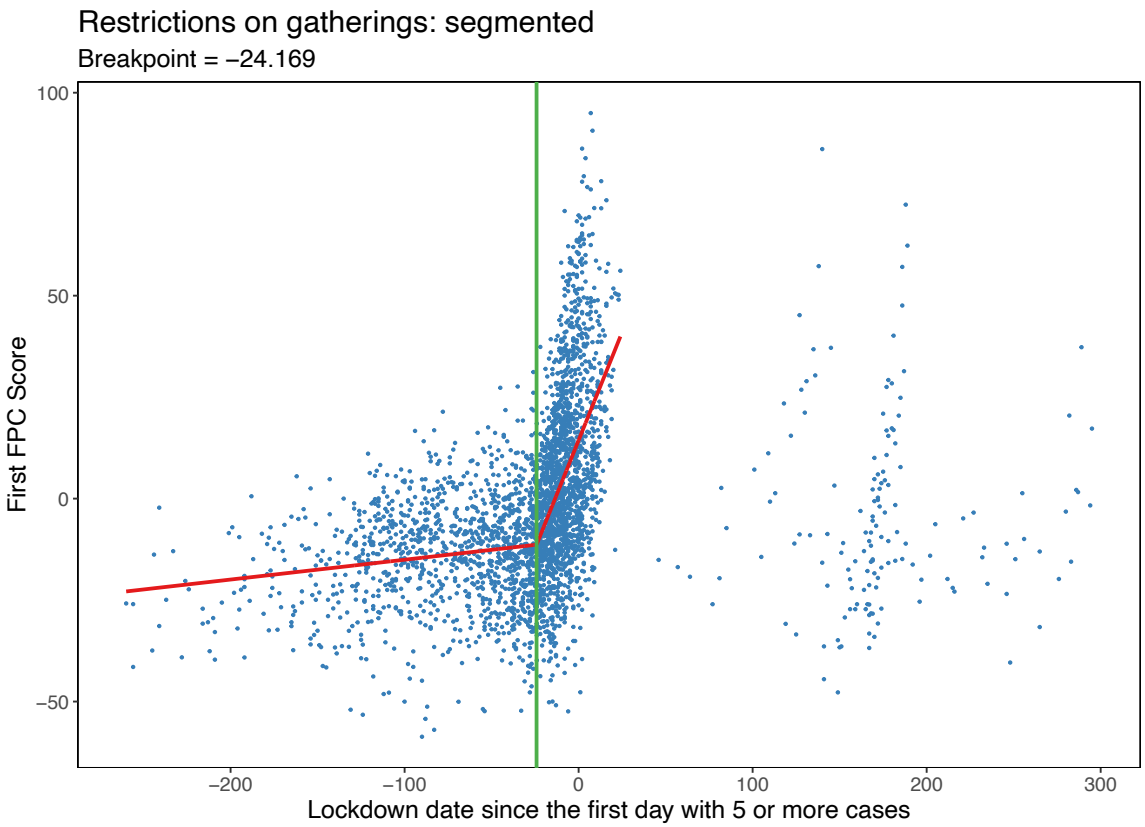

Figure A.12: Segmented Regression of the first FPC score vs the timing of Restrictions on Gatherings

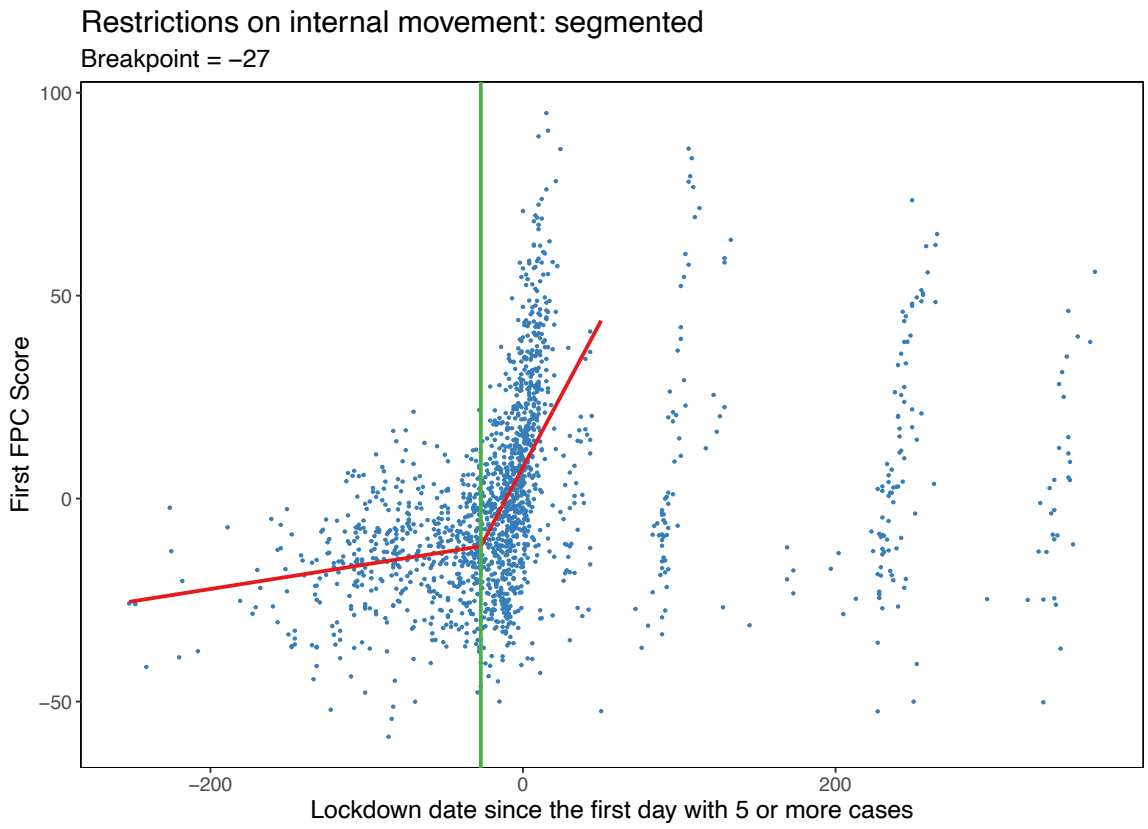

Figure A.13: Segmented Regression of the first FPC score vs the timing of Restrictions on Internal Movement

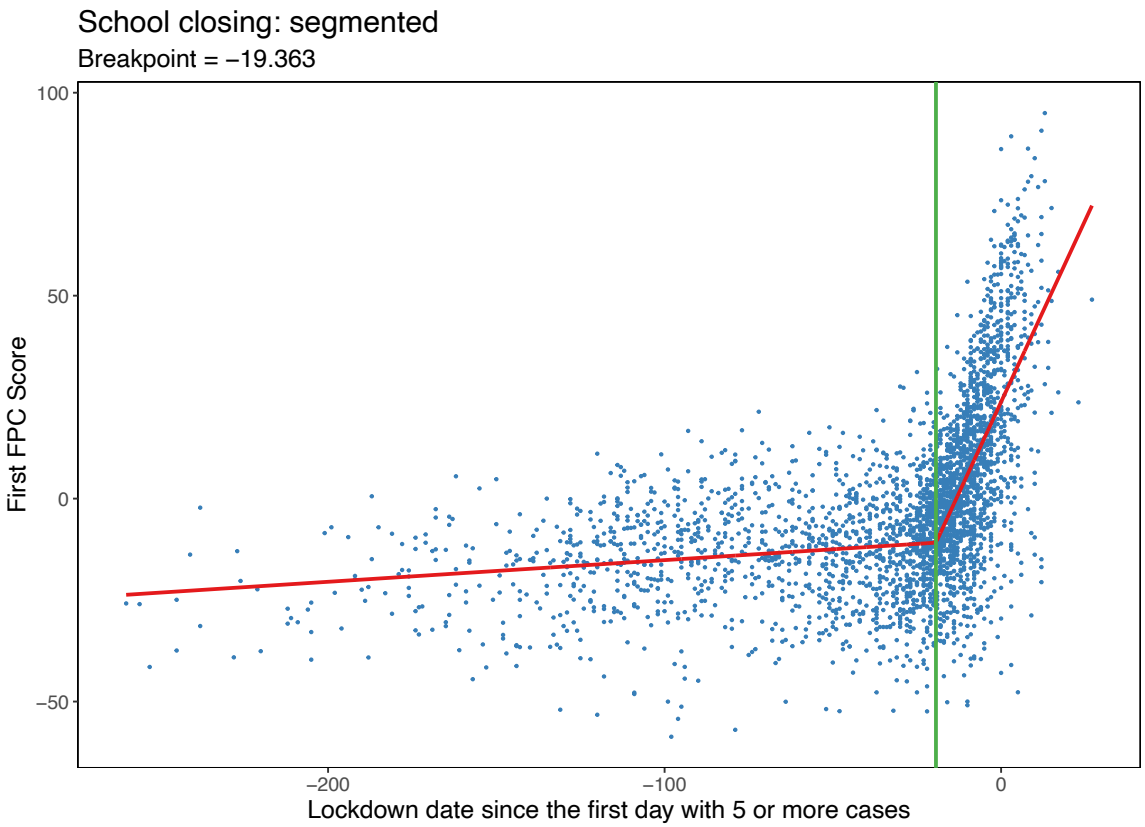

Figure A.14: Segmented Regression of the first FPC score vs the timing of School Closing

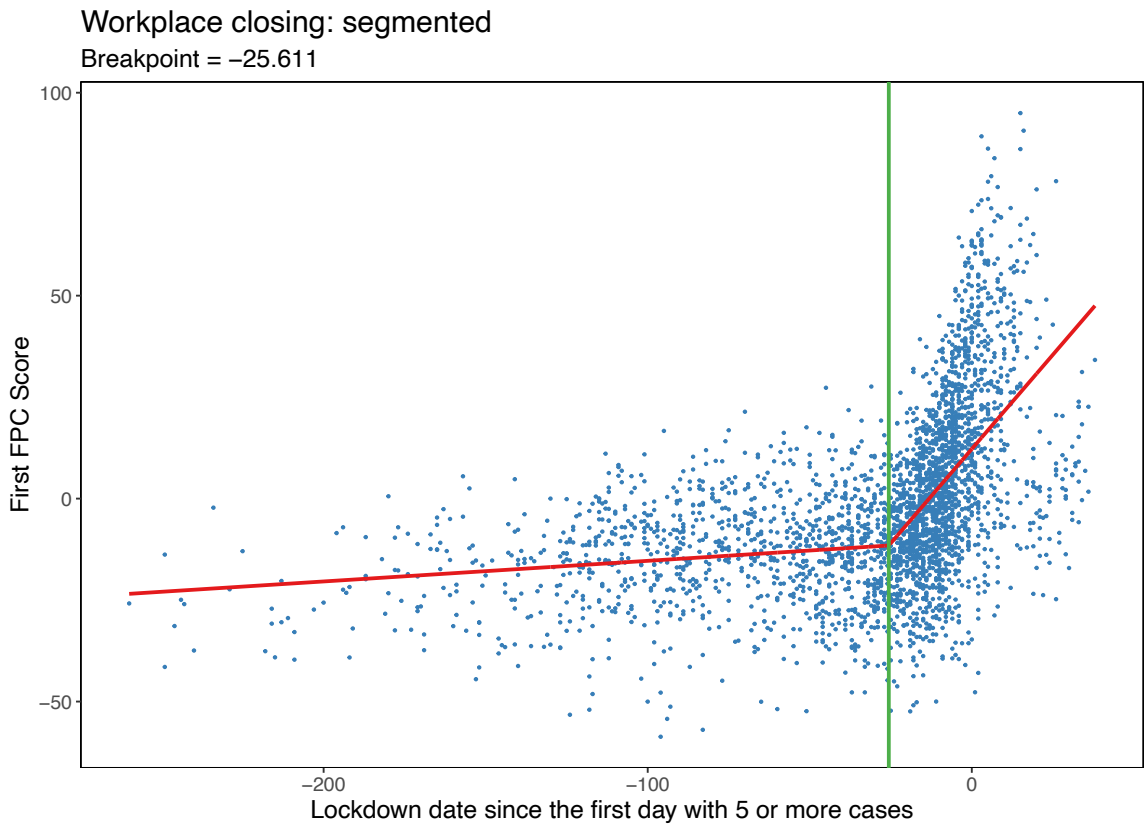

Figure A.15: Segmented Regression of the first FPC score vs the timing of Workplace Closing

Bibliography

[1] Charles N. Agoti, Martin Mutunga, Arnold W. Lambisia, Domtila Kimani, Robinson Cheruiyot, Patience Kiyuka, Clement Lewa, Elijah Gicheru, Metrine Tendwa, Khadija Said Mohammed, Victor Osoti, Johnstone Makale, Brian Tawa, Calleb Odundo, Wesley Cheruiyot, Wilfred Nyamu, Wilson Gumbi, Jedidah Mwacharo, Lydia Nyamako, Edward Otieno, David Amadi, Janet Thoya, Angela Karani, Daisy Mugo, Jennifer Musyoki, Horace Gumba, Salim Mwarumba, Bonface M. Gichuki, Susan Njuguna, Debra Riako, Shadrack Mutua, John N. Gitonga, Yiakon Sein, Brian Bartilol, Shaban J. Mwangi, Donwilliams O. Omuoyo, John M. Morobe, Zaydah R. de Laurent, Philip Bejon, Lynette Isabella Ochola-Oyier, and Benjamin Tsofa. Pooled testing conserves SARS-CoV-2 laboratory resources and improves test turn-around time: experience on the Kenyan Coast. *Wellcome Open Research*, 5:186, February 2021.

[2] Netta Barak, Roni Ben-Ami, Tal Sido, Amir Perri, Aviad Shtoyer, Mila Rivkin, Tamar Licht, Ayelet Peretz, Judith Magenheim, Irit Fogel, Ayalah Livneh, Yutti Daitch, Esther Oiknine-Djian, Gil Benedek, Yuval Dor, Dana G. Wolf, Moran Yassour, and The Hebrew University-Hadassah COVID-19 Diagnosis Team. Lessons from applied large-scale pooling of 133,816 SARS-CoV-2 RT-PCR tests. *Science Translational Medicine*, 13(589):eabf2823, April 2021.

[3] Cody Carroll, Alvaro Gajardo, Yaqing Chen, Xiongtao Dai, Jianing Fan, Pantelis Z. Hadjipantelis, Kyunghee Han, Hao Ji, Hans-Georg Mueller, and Jane-Ling Wang. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2021. `https://CRAN.R-project.org/package=fdapace`.

[4] CDC. COVID Data Tracker. `https://covid.cdc.gov/covid-data-tracker`, 2020. Accessed: 2021-05-23.

[5] Evangeline Ann Daniel, Bennett Henzeler Esakialraj L, Anbalagan S, Kannan Muthuramalingam, Ramesh Karunaianantham, Lucia Precilla Karunakaran, Manohar Nesakumar, Murugesan Selvachithiram, Sathyamurthi Pattabiraman, Sudhakar Natarajan, Srikanth Prasad Tripathy, and Luke Elizabeth Hanna. Pooled Testing Strategies for SARS-CoV-2 diagnosis: A comprehensive review. *Diagnostic Microbiology and Infectious Disease*, 101(2):115432, October 2021.

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

[7] Robert Dorfman. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, 14(4):436–440, December 1943.

[8] Seth Flaxman, Swapnil Mishra, Axel Gandy, H. Juliette T. Unwin, Thomas A. Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W. Eaton, Mélodie Monod, Pablo N. Perez-Guzman, Nora Schmit, Lucia Cilloni, Kylie E. C. Ainslie, Marc Baguelin, Adhiratha Boonyasiri, Olivia Boyd, Lorenzo Cattarino, Laura V. Cooper, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Sabine L. van Elsland, Richard G. FitzJohn, Katy A. M. Gaythorpe, Lily Geidelberg, Nicholas C. Grassly, William D. Green, Timothy Hallett, Arran Hamlet, Wes Hinsley, Ben Jeffrey, Edward Knock, Daniel J. Laydon, Gemma Nedjati-Gilani, Pierre Nouvellet, Kris V. Parag, Igor Siveroni, Hayley A. Thompson, Robert Verity, Erik Volz, Caroline E. Walters, Haowei Wang, Yuanrong Wang, Oliver J. Watson, Peter Winskill, Xiaoyue Xi, Patrick G. T. Walker, Azra C. Ghani, Christl A. Donnelly, Steven Riley, Michaela A. C. Vollmer, Neil M. Ferguson, Lucy C. Okell, Samir Bhatt, and Imperial College COVID-19 Response Team. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, August 2020.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[10] Najmul Haider, Abdinasir Yusuf Osman, Audrey Gadzekpo, George O. Akipede, Danny Asogun, Rashid Ansumana, Richard John Lessells, Palwasha Khan,

Muzamil Mahdi Abdel Hamid, Dorothy Yeboah-Manu, Leonard Mboera, Elizabeth Henry Shayo, Blandina T. Mmbaga, Mark Urassa, David Musoke, Nathan Kapata, Rashida Abbas Ferrand, Pascalina-Chanda Kapata, Florian Stigler, Thomas Czypionka, Alimuddin Zumla, Richard Kock, and David McCoy. Lockdown measures in response to COVID-19 in nine sub-Saharan African countries. *BMJ Global Health*, 5(10):e003319, October 2020.

[11] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 5(4):529–538, April 2021.

[12] Xiaolin Huang, Xiaojian Shao, Li Xing, Yushan Hu, Don D. Sin, and Xuekui Zhang. The impact of lockdown timing on COVID-19 transmission across US counties. *EClinicalMedicine*, 38:101035, August 2021.

[13] Dan Keating, Madison Dong, and Tim Meko. Visualizing the omicron wave striking and rolling across the country. *The Washington Post*, January 2022.

[14] Huiling Li, Kai Sun, David H. Persing, Yi-Wei Tang, and Dingxia Shen. Real-time Screening of Specimen Pools for Coronavirus Disease 2019 (COVID-19) Infection at Sanya Airport, Hainan Island, China. *Clinical Infectious Diseases*, 73(2):318–320, July 2021.

[15] Khai Lone Lim, Nur Alia Johari, Siew Tung Wong, Loke Tim Khaw, Boon Keat Tan, Kok Keong Chan, Shew Fung Wong, Wan Ling Elaine Chan, Nurul Hanis Ramzi, Patricia Kim Chooi Lim, Sulaiman Lokman Hakim, and Kenny Voon. A novel strategy for community screening of SARS-CoV-2 (COVID-19): Sample pooling method. *PLOS ONE*, 15(8):e0238417, August 2020.

[16] Eugene Litvak, Xin M. Tu, and Marcello Pagano. Screening for the Presence of a Disease by Pooling Sera Samples. *Journal of the American Statistical Association*, 89(426):424–434, June 1994.

[17] Ying Liu and Joacim Rocklöv. The effective reproductive number of the Omicron variant of SARS-CoV-2 is several times relative to Delta. *Journal of Travel Medicine*, 29(3):taac037, April 2022.

[18] Christopher J. L. Murray. COVID-19 will continue but the end of the pandemic is near. *The Lancet*, 399(10323):417–419, January 2022.

[19] Leon Mutesa, Pacifique Ndishimye, Yvan Butera, Jacob Souopgui, Annette Uwineza, Robert Rutayisire, Ella Larissa Ndoricimpaye, Emile Musoni, Nadine Rujeni, Thierry Nyatanyi, Edouard Ntagwabira, Muhammed Semakula, Clarisse Musanabaganwa, Daniel Nyamwasa, Maurice Ndashimye, Eva Ujeneza, Ivan Emile Mwikarago, Claude Mambo Muvunyi, Jean Baptiste Mazarati, Sabin Nsanzimana, Neil Turok, and Wilfred Ndifon. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature*, 589:276–280, Jan 2021.

[20] Hans-Georg Müller. *Nonparametric Regression Analysis of Longitudinal Data*. Springer, New York, NY, USA, 1988.

[21] Laura Osman and Stephanie Taylor. Provinces clamour for rapid tests while feds struggle to deliver millions promised. *CTV News*, January 2022.

[22] Garrett A. Perchetti, Ka-Wing Sullivan, Greg Pepper, Meei-Li Huang, Nathan Breit, Patrick Mathias, Keith R. Jerome, and Alexander L. Greninger. Pooling of SARS-CoV-2 samples to increase molecular testing throughput. *Journal of Clinical Virology*, 131:104570, October 2020.

[23] Gary W Procop, Marion Tuohy, Christine Ramsey, Daniel D Rhoads, Brian P Rubin, and Richard Figler. Asymptomatic Patient Testing After 10:1 Pooling Using the Xpert Xpress SARS-CoV-2 Assay. *American Journal of Clinical Pathology*, 155(4):522–526, April 2021.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[25] Bootan Rahman, Evar Sadraddin, and Annamaria Porreca. The basic reproduction number of SARS-CoV-2 in Wuhan is about to die out, how about the rest of the World? *Reviews in Medical Virology*, 30(4):e2111, July 2020.

[26] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. Coronavirus Pandemic (COVID-19). *Our World in Data*, 2020. `https://ourworldindata.org/coronavirus`.

[27] Carlos Santamaria, Francesco Sermi, Spyridon Spyratos, Stefano Maria Iacus, Alessandro Annunziato, Dario Tarchi, and Michele Vespe. Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data. A European regional analysis. *Safety Science*, 132:104925, December 2020.

[28] H. Tamashiro, A. Fauquex, D. Heymann, J. Emmanuel, P. Sato, and W. Maskill. Reducing the cost of HIV antibody testing. *The Lancet*, 342(8863):87–90, July 1993.

[29] Annabel X. Tan, Jessica A. Hinman, Hoda S. Abdel Magid, Lorene M. Nelson, and Michelle C. Odden. Association Between Income Inequality and County-Level COVID-19 Cases and Deaths in the US. *JAMA Network Open*, 4(5):e218799, May 2021.

[30] Ton That Thanh, Nguyen Thi Thanh Nhan, Huynh Kim Mai, Nguyen Bao Trieu, Le Xuan Huy, Ho Thi Thanh Thuy, Le Thanh Chung, Nguyen Ngoc Anh, Nguyen Thi Thu Hong, Bui Thuc Thang, Nguyen Thi Hoai Thu, Le Thi Kim Chi, Nguyen Thi Hanh, Nguyen Huy Hoang, Nguyen Van Vinh Chau, Guy Thwaites, Do Thai Hung, Le Van Tan, and Ngo Thi Kim Yen. The Application of Sample Pooling for Mass Screening of SARS-CoV-2 in an Outbreak of COVID-19 in Vietnam. *The American Journal of Tropical Medicine and Hygiene*, 104(4):1531–1534, April 2021.

[31] US Census Bureau. American Community Survey (ACS). `https://www.census.gov/programs-surveys/acs`, 2021. Accessed: 2021-05-14.

[32] Marco Vinceti, Tommaso Filippini, Kenneth J. Rothman, Fabrizio Ferrari, Alessia Goffi, Giuseppe Maffeis, and Nicola Orsini. Lockdown timing and efficacy in controlling COVID-19 using mobile phone tracking. *EClinicalMedicine*, 25:100457, August 2020.

[33] Kyle Walker and Matt Herman. *tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames*, 2021. `https://CRAN.R-project.org/package=tidycensus`.

[34] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, June 2016.

[35] Huizhong Wu. China battles multiple COVID-19 outbreaks, driven by 'stealth omicron' variant. *Global News*, March 2022.

[36] Jiachuan Wu, Savannah Smith, Mansee Khurana, Corky Siemaszko, and Nigel Chiwaya. Coronavirus lockdowns and stay-at-home orders across the U.S. *NBC News*, March 2020.

[37] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005.