SMALL-SAMPLE ESTIMATION OF THE MUTATIONAL SUPPORT AND THE DISTRIBUTION OF MUTATIONS IN THE SARS-COV-2 GENOME

BY

VISHAL RANA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Adviser:

Professor Olgica Milenkovic

ABSTRACT

The problem of accurately estimating and characterizing different mutations in the viral genomes present within a population is of great importance in tracking and mitigating the spread of the virus and is made difficult by the lack of a sufficient number of sequenced genomes especially during the early stages of an outbreak. We consider the problem of determining the mutational support and distribution of mutations in the SARS-Cov-2 genome and its open reading frames (ORFs). The mutational support refers to the unknown number of sites that are mutated among all the viral strains present in a population. The support and distribution of mutations can be used to guide primer selection for RT-PCR test kits, study the virulence of the virus, discover adaptation mechanisms deployed by the virus to evade the host immune system, as well as to identify new strains that might be circulating in the population early on. We propose new state-of-the-art polynomial estimation techniques using weighted and regularized Chebyshev approximations for small-sample mutational support estimation and we use a modified Good-Turing estimator for distribution estimation. Our differential analysis of mutations in various population subgroups (based on data retrieved from GISAID repository) revealed several important differences including those in the ORF6 and ORF7a regions for older versus younger patients, ORF1b and ORF10 regions for females versus males and in several ORFs for Asia versus Europe and North America. We also found no significant mutations in the primer regions from ORF N chosen by CDC for RT-PCR test kits in any of the subpopulations, which is important for reliability of the test results.

To my parents, for their love and support.

ACKNOWLEDGMENTS

I would like to thank my adviser, Professor Olgica Milenkovic, for her guidance and support. I would also like to thank Eli Chien and Jianhao Peng for their collaboration and mentorship throughout this project.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Viruses undergo frequent mutations introduced mostly during the replication of their genetic material, a process that is prone to errors. Some of these mutations provide survival advantage to the virus by helping it evade the immune system of the host, thus becoming more widespread in the population. The rates of mutations are different for various different viruses and have been extensively explored in virology literature in the past [1, 2]. RNA viruses are known to mutate faster than DNA viruses, as are single stranded viruses compared to double stranded ones. This is due to inaccurate RNA duplication compared to DNA and general structural instability of single stranded genomes [3]. An inverse correlation between the length of the viral genome and the rate of mutation has also been documented. Viruses with shorter genomes mutate much more rapidly than their longer counterparts [4].

The host immune system has memory of viruses it has already encountered and if the host is exposed to such a virus, the immune system proceeds swiftly to eliminate it. However, the virus can mutate and if the rate if mutation is sufficiently high, it can make it harder for the host immune system to identify it, thus slowing down the immune response. This phenomenon is known as *antigenic drift*. This gives fast mutating viruses more time to replicate and spread, and by evading the host immune system such viruses pose a great health risk [5, 6]. On the other hand, some recent studies have shown that high rates of mutation can also be detrimental to survival of the virus on short time scales by triggering a rapid innate immune response by the host [7]. Thus, the mutational landscape of a virus is closely related to its potential to spread among a population and a virus may need to explore a significant number of mutations in its attempt to successfully infect a large number of hosts [8, 9, 10]. However, our understanding of the causes of elevated mutation rates and their correlation with clinical outcomes is still limited. Accurately determining the mutation rates and the distribution of mutations is an important first step in the direction of addressing these questions.

A number of different definitions of viral "mutation rate" can be found in the literature [1, 11]. *Genomic mutation rate* represents the average number of positions at which each viral genome differs from its ancestral genome and is calculated as a product of the per-nucleotide mutation rate and the length of the genome. The per-nucleotide mutation rate of RNA viruses lies in the range $10^{-6} - 10^{-4}$ [11]. Even though it is known that replication errors are not the only source of mutations in viral genomes, replication error rates and mutation rates of viruses are often used interchangeably. Some studies estimate the counts of mutations in sequenced genomes using the genomes of the first infected individual (Patient 0) or, more frequently, the first individual that was sequenced (Patient 1) as a reference. The genomic mutation rate of SARS-Cov-2 is estimated to be 2-3 mutations a month [12]. Since a large carrier population can harbor viruses with widely different mutation rates, it is more challenging to define the genomic mutation rate for such a population.

To define the mutational support of a virus we use the viral genome of Patient 1 as a reference and index all locations along the genome. The mutational support of a single viral genomic sequence equals the set of locations where it disagrees with the reference. The size of the mutational support hence equals the Hamming distance between the reference and the sequence under consideration. The *mutational support of a population (henceforth, mutational support)* of viral genomes equals the size of the union of the individual mutational supports. Only a subset of the infected patient's viral genomes is sequenced at any given time, therefore we do not observe the mutational support of a population directly and estimation methods need to be employed. We can count the number of mutations present in at least one of the sequenced samples, however a simple count based estimator (maximum likelihood estimator) gives good performance only in cases where the number of samples sequenced is significantly larger than the length of the viral genome. In the absence of sufficient number of sequenced samples, the maximum likelihood estimator may return highly inaccurate estimates due to unobserved mutations. This phenomenon, known as the small-sample effect, has been extensively researched in the machine learning community [13, 14].

Nevertheless, to the best of our knowledge, the problems of mutational support and mutational distribution estimation in the small-sample regime have not been addressed in the virology literature. We argue that this problem is of significant relevance as its successful solution may be used to assess the virulence of the virus, guide primer selection for real-time RT-PCR tests during the early stages of an outbreak and correlate mutational rates with elevated risks of heavy symptoms.

Our contributions are two-fold. First, we present new machine learning methods for determining the unknown support of mutations and their distributions given sequencing data from a limited number of Covid-19 patients. The methods use efficient polynomial class estimators and exhibit state-of-the art performance on synthetic datasets. The actual genomic datasets are retrieved from the Global Influenza Surveillance Aid (GISAID) repository during the early stages of the Covid-19 outbreak. In our initial analysis, we only use $< 9,000$ samples, which is a significantly smaller number than the length of the SARS-Cov-2 genome which roughly equals $30,000$. The approach is based on weighted Chebyshev polynomial estimators and adapted Good-Turing distribution estimators, and its accuracy is evaluated based on larger sample set sizes retrieved on later dates. Second, the mutational supports are estimated for three different population types, namely according to geographic region (Asia, Europe, North America (NA)), gender (female/male) and age (< 55, > 55). For European samples retrieved at a later time stage, estimates for females of age < 55 versus males of age > 55 were analyzed as well. The estimates are used to predict mutational hotspots and compare the genomic loci with highest mutation frequency in different subpopulations. For the latter task, we further process the results by using the Jaccard distance as well as the symmetric Kullback-Leibler divergence. Furthermore, to determine if the mutation rates are appropriately low in genomic regions harboring primers used for real-time reverse-transcriptase polymerase chain reaction (RT-PCR) testing [15], we separately scrutinize the N ORF of SARS-Cov-2 samples.

Our analysis reveals several important biological findings. The predicted mutational supports exhibit significant differences in the ORF6 and ORF7a regions in older versus younger patients, and in the ORF1b and ORF10 regions in females versus males. The mutational support of the ORF1b region for young females is almost twice that of old males, while old males have a significantly larger mutational supports for genes S and ORF10. Given that young females are much less likely to develop severe symptoms than old males, the identified potential high-mutation regions may be further examined to identify their potential role in the spread and severity/potency of the virus. Furthermore, it is important to observe that the variance of the support is extremely high in the ORF8 region, close to 200 times higher for patients above 55 years of age compared to patients below 55 years of age. Less surprisingly, there also exist statistically significant differences in the ORFs of Asian versus European and NA samples in the ORF1a,b and other ORFs. Second, despite the fact that we predict that the N region of SARS-Cov-2 will have a very large mutational support, almost all high-probability mutations fall outside of the two regions of paired primers recommended by the CDC for RT-PCR testing.

The remainder of the thesis is organized as follows. In Chapter 2, we describe the data acquisition process, the pre-processing tasks as well as our new small-sample support and distribution estimation algorithms. Chapter 3 contains the most relevant results and the discussion of their biological relevance, and Chapter 4 concludes the work.

CHAPTER 2

MATERIALS AND METHODS

We first describe the organization of the SARS-Cov-2 genome, followed by the data acquisition from GISAID, data pre-processing as well as the entire work-flow pipeline we developed for our analysis. Finally, we describe our polynomial based small-sample support estimator and a modified Good-Turing estimator for distribution estimation.

2.1 Organization of the SARS-Cov-2 genome

A breakdown of the genomic structure of SARS-Cov-2 is depicted in Figure 2.1, and described in detail in [16] and [17]. Understanding the roles played by various ORFs of the viral genome is of importance as it allows one to put the results of the mutational support analysis into proper context: Mutational variability in certain ORFs of different host subpopulations may be indicative of different innate immune responses and evading mechanisms employed by the virus.

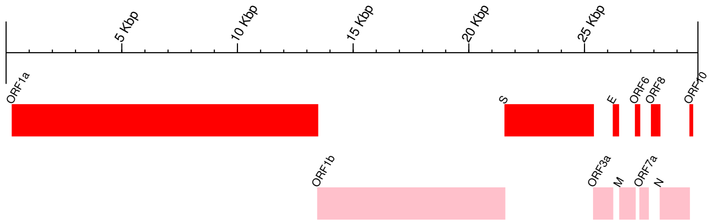


Figure 2.1: Organization of the SARS-Cov-2 genome. (Wuhan-Hu-1, GenBank MN908947)

Typically, coronaviruses have genomes including at least six open reading frames (ORFs). ORF1a and ORF1b constitute the longest component of the genomes and are responsible for encoding two polypeptides, pp1a and pp1ab, which are jointly used to create a family of NSP proteins. This family of polypeptides includes replicase-transcriptase proteins, responsible for promoting cellular mRNA degradation and blocking the translation process in host cells, thereby impairing the operation of the immune response and proofreading. The pp1a/b polypeptides are functionally combined using proteases, such as the native chymotrypsin-like protease. Viral structural proteins are encoded by the sgRNA region, and include the ORF2 or spike (S), ORF5 or membrane (M), ORF4 or envelope (E), and nucleocapsid (N) proteins, as well proteins encoded by the ORF10 sequence. ORF3a encodes a membrane protein that interacts with proteins encoded by ORFs M, S and E and is believed to play an important role in viral release and the generation of cytokine storm; on the other hand, ORF3b encodes proteins that block the induction of interferons with antiviral activity. The ORF6 products are important virulence factors that enable the virus to escape detection by the immune system of the host.

For real time RT-PCR testing and detection of Covid-19, the oligonucleotide primers and probes are selected from the nucleocapsid (N) gene region (per CDC guidelines for the United States [18]), and as provided in panels produced by Integrated DNA Technologies (IDT), including two primer pairs/probe sets. As a control, additional primer/probe sets are used such as the human RNase P gene (RP) which is also included in the panel. Countries like Germany and China have adopted primers from other genomic regions, as outlined in [15]. For individual testing for Covid-19 in the United States, it is of special interest to predict mutation rates in the N region of the genome [15]. High-rate mutations in this region may cause highly undesirable false negatives in the test outcomes. ORF7a encodes for a membrane protein while ORF7b is believed to act as a viral attenuation factor and contributor in human infectivity, similarly to the protein encoded by ORF8. The ORF9b has the role to impede mitochondrial morphology and function and disable the interferon response of the host, while ORF9c appears to block important signaling pathways of the host [17].

2.2 Data acquisition

For the proposed analyses, we used data from the GISAID EpiCoV database [19] which contains sequenced viral strains collected from patients across the world. We downloaded the data at three time points, starting from 04-03-2020, continuing on 04-10-2020 and finishing on 04-14-2020. We then revisited the repository on 10-20-2020 to further evaluate the quality of our predictions regarding the mutational supports. At that point of time, $9,271$ samples from Asia and more than $30,000$ samples from NA and $85,000$ samples from Europe were available.

For samples made available in April as well as in September 2020, we filtered the datasets only to include nearly-complete samples, i.e., those of length $> 29,000$ nts, resulting in a number of samples summarized in Table 2.1. We also downloaded the associated metadata used for patient subtyping. Note that we used results obtained early in the monitoring process in order to evaluate our small-sample estimation schemes. Table 2.1 provides the number of samples available within different categories for each of the three time points.

Table 2.1: Number of samples available for different phenotype classes and data retrieved on three different dates, 04-03-2020, 04-10-2020 and 04-14-2020.

Date	Age		Gender		Region			Total # of samples
	> 55	< 55	Male	Female	Asia	Europe	NA	
04-03	909	1,477	1,349	1,061	510	1,695	818	3,511
04-10	2,373	1,850	2,315	1,956	615	3,194	1,147	5,650
04-14	3,047	3,231	3,526	2,817	636	5,890	1,774	8,893

As the first step in our analysis, we used the sequence alignment software MUSCLE [20] to perform pairwise alignment of all the samples with the SARS-Cov-2 sequence of Patient 1, published under the name Wuhan-Hu-1, admitted to the Central Hospital of Wuhan on December 26, 2019 (GenBank accession number MN909847). Furthermore, we also performed alignment with respect to Patient 1 of two additional continents, Europe and NA. The latter alignment was performed to better determine how the mutational support and mutational distribution depends on a particular geographic context.

For each aligned pair of samples, we generated a "mutation profile," a list containing the positions in the reference genome in which the patient aligned to the reference has a substitution mutation. We did not perform multiple sequence alignment in order to assess the mutation landscape as we need to analyze each patient data separately (each patient and her/his mutations are treated as one sample in the estimation procedure). The mutational profile lists are subsequently aggregated over all the patient samples, resulting in a histogram of mutations across all positions in the viral reference genome. The aggregate profiles are further partitioned according to the 11 genes they are located in on the viral genome depicted in Figure 2.1. The total count of mutations for each location in each gene is used as a sufficient statistic for estimating the mutational support and the distribution of the mutations in each of the 11 genes. The analytic pipeline used is depicted in Figure 2.2.

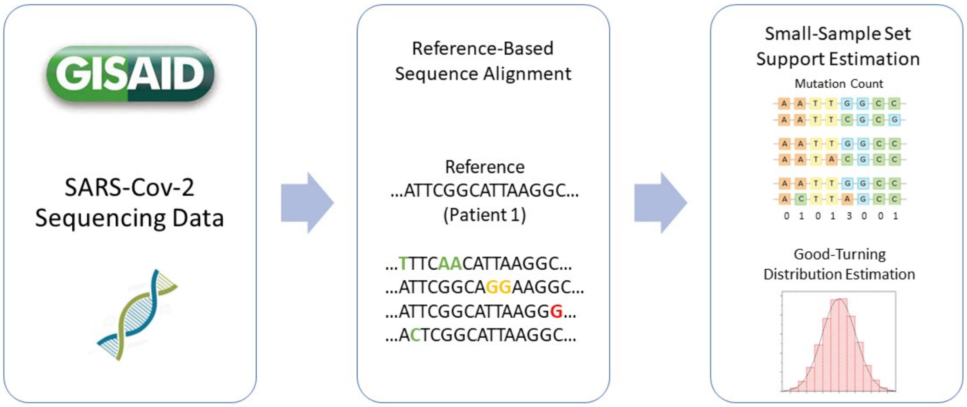


Figure 2.2: The data analysis flowchart: Viral sequencing data is retrieved from the GISAID repository and then aligned against the genome of Patient 1 or regional Patient 1 in a pairwise fashion. The substitutions at different genomic locations for all analyzed pairs of samples are counted and used as sufficient statistics for the estimation procedures.

To adjust for alignment artifacts introduced by sequencing errors, dropouts and alignment gaps, we removed all gaps encountered in the prefixes and suffixes and sufficiently long gaps (> 10 nts) within the actual alignments. Most gaps are encountered at the 5’UTR and 3’UTR regions of the genome, as may be expected from outputs of global alignment algorithms.

As there exists a large body of evidence of stratified susceptibility and severity of symptoms across different racial, age and gender groups [21, 22], we performed four different types of mutational support and distribution analyses. In the first set of tests, we split the patient mutation histograms based on gender (male/female), based on age (under 55/over 55) and based on the geographic location (Asia/ NA/ Europe). The age threshold was set by taking into consideration available sample sizes needed for the analysis and the age profile of patients available on GISAID; the threshold also reflects different risk groups for the development of severe symptoms. In addition, we performed the same analysis for a combination of patient features for settings with sufficiently many samples available early in the pandemic, such as males above 55 years of age/females below 55 years of age, from Europe. Note that in all the described cases, "geographic location" refers to the region of infection of the patient and not the region where he/she was tested and the sample was sequenced.

Since the number of samples per population type may vary significantly, we performed two tests. In one test we used all samples available, while in another we adjusted for difference in sizes of the sets by subsampling the larger of the two classes to make the sample sets of equal sizes. The number of samples available for various patient subgroups is listed in Table 2.1. For data obtained on 04-03-2020, we used all the samples available for all the classes, without balancing the class sizes. For data retrieved on 04-10-2020 and 04-14-2020, we balanced the classes by subsampling from the larger of the two classes for both age- and gender-based subtypes. For different geographical regions, on 04-10-2020, we used all 615 samples from Asia and subsampled Europe and NA to 1000 samples each. Similarly, we used all 636 samples from Asia and subsampled Europe and NA to $1,774$ samples each, for data retrieved on 04-14-2020. It is important to point out that by performing the experiments with different sample set sizes one can compare the quality of the estimates obtained using samples from the early stages of epidemics and those obtained from later stages when more information about individual viral sequences is available. Furthermore, the new machine learning methods outlined in Sections 2.3 and 2.4 apply to any other viral or bacterial dataset collection with the obviously required modifications to account for the genetic profile of the microorganisms.

2.3 New small-sample support estimators

We focus on the *polynomial approximation approach* put forward in [23], and significantly improve on it in practice by introducing new weighted Chebyshev polynomial optimization techniques largely unknown in the machine learning and computational biology community [24]. The weighted approximation method can be seamlessly combined with regularization techniques that use the variance of the estimator in a way that complements features used in maximum likelihood (ML) estimation [25]; and with semi-infinite programming (SIP) solvers that produce the parameters of the estimator. The SIP methods can be solved consistently and highly efficiently through discretization resulting in a small linear program (LP) of size *decreasing with the number of samples.* Interestingly, despite the fact that our estimators are constructed using an LP as is the case for the best performing ML-based approach [26], the ML-LP formulation has a number of variables and constraints that actually *increase* with the number of samples; this difference makes our estimator significantly more efficient as is needed for large scale estimation processes like the ones described in this work, in addition to improving their performance.

Next, we provide a detailed description of our polynomial estimation method. Recall that the support of a discrete probability distribution is defined as the number of symbols with positive probability of occurrence. We define the mutational support of a virus as the total number of genomic sites mutated in any viral genome in any individual (observed or unobserved due to limited testing), compared to a reference genome. As already pointed out, in our case the reference is the genome of Patient 1, the first sequenced SARS-Cov-2 genome or the genome of regional Patient 1.

The most commonly used techniques for support and distribution estimation are ML methods which directly use the empirical counts of the symbols to determine the support or probabilities of interest. It is well known that ML approaches perform poorly for large alphabet sizes (supports) when only a small number of samples from the distribution is available. In this case, they fail to account for samples that have never been observed due to limited sampling. To see why this is the case, assume that we observe 10 samples from a distribution supported on $\{1, \ldots, 100\}$. Clearly, with only 10 samples available, our best possible guess for the support size will be the number of distinct symbols observed which is a number ≤ 10 and far from the correct value 100.

The problem of estimating the support of an unknown probability distribution or estimating the distribution itself in the context of small-sample sets has a long history. The first line of work in this area is attributed to Laplace, as described in [27], who introduced a class of smoothed distribution estimators termed add 1 (or more generally, add constant c estimators). These estimators adjust the counts of observed symbols in order to account for the unseen symbols.

Let $P = (p_1, p_2, \ldots)$ be a discrete distribution over some finite alphabet and let $\mathbf{x}^n$ be a vector of i.i.d. samples drawn according to the distribution P. The problem of interest is to estimate the support size, defined as $S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}$. We use S instead of $S(P)$ to avoid notational clutter. An important assumption used in our estimation methods is that the minimum non-zero probability of the distribution P is greater than $\frac{1}{k}$, for some $k \in \mathbb{R}^+$, i.e., $\inf\{p \in P \,|\, p > 0\} > \frac{1}{k}$. We let D_k denote the space of all probability distributions satisfying $\inf\{p \in P \,|\, p > 0\} > \frac{1}{k}$. A sufficient statistic for $\mathbf{x}^n$ is the empirical distribution (i.e., histogram) $n = (n_1, n_2, \ldots)$, where $n_i = \sum_{j=1}^{n} \mathbf{1}_{\{\mathbf{x}_j = i\}}$ and $\mathbf{1}_A$ stands for the indicator function of the event A.

To determine the quality of an estimator, we use the most frequently studied risk model, the minmax risk under normalized squared loss, defined as

$$R^*(k,n) = \inf_{\hat{S}} \sup_{P \in D_k} \mathbb{E}\left[\left(\frac{\hat{S}(N) - S}{k}\right)^2\right]. \tag{2.1}$$

We seek a support estimator $\hat{S}$ that minimizes

$$\sup_{P \in D_k} \mathbb{E}\left[\left(\frac{\hat{S}(N) - S}{k}\right)^2\right] = \sup_{P \in D_k} \left[\mathbb{E}^2\left(\frac{\hat{S}(N) - S}{k}\right) + var\left(\frac{\hat{S}(N) - S}{k}\right)\right]. \tag{2.2}$$

The first term within the supremum captures the expected bias of the estimator $\hat{S}$. The second term represents the variance of the estimator $\hat{S}$. A“good” estimator should jointly balance out the worst-case contributions of the bias and variance (note that for the case that only the bias is considered directly, and the variance accommodated for by modifying the bias-optimized solution, the underlying estimator was analyzed in [23]).

To introduce our method, we first describe the class of *polynomial estimators*. Given a positive integer parameter L, we say that an estimator $\hat{S}$ is a polynomial class estimator with a threshold parameter L (i.e., a $Poly(L)$ estimator) if it takes the form $\hat{S} = \sum_i g_L(n_i)$, where g_L is defined as

$$g_L(j) = \begin{cases} a_j j! + 1, & \text{if } j < L \\ 1, & \text{otherwise.} \end{cases} \tag{2.3}$$

The coefficients a satisfy $a_j \in \mathbb{R}$ and $a_0 = -1$, (since this choice ensures that $g_L(0) = 0$) and have to be optimized in order to minimize the risk. One can associate an estimator $\hat{S}$ with its corresponding coefficients $\mathbf{a}$, i.e.,

$$Poly(L) = \Big\{ \mathbf{a} \in \mathbb{R}^{L+1} | a_0 = -1 \Big\}.$$

The authors of [23] proposed using a special form of polynomial estimators in which the coefficients a_j correspond to scaled evaluations of a Chebyshev polynomial of order L. The Chebyshev polynomial of the first kind of degree L is defined as

$$T_L(x) = \cos(L \arccos(x)) = \frac{z^L + z^{-L}}{2},$$

where z is the solution of the quadratic equation $z + z^{-1} = 2x$. The polynomial T_L is bounded in the interval $[-1, 1]$ and may be scaled and shifted to lie in an arbitrary interval $[l, r]$ based on

$$R_L(x) = -\frac{T_L(\frac{2x-r-l}{r-l})}{T_L(\frac{-r-l}{r-l})} \triangleq \sum_{j=0}^{L} \tilde{a}_j x^j.$$

Clearly, $R_L(0) = -1$ and $\tilde{a}_0 = -1$.

The Chebyshev polynomial estimator is an estimator for which

$$\tilde{a}_j = \frac{R_L^{(j)}(0)}{j!}, \tag{2.4}$$

and it takes the form $\tilde{S} = \sum_i \tilde{g}_L(n_i)$, where

$$\tilde{g}_L(j) = \begin{cases} \tilde{a}_j j! + 1, & \text{if } j < L, \\ 1, & \text{otherwise,} \end{cases} \tag{2.5}$$

$$\text{with } \ L \triangleq \lfloor c_0 \log k \rfloor, \ [l, r] \triangleq \left[\frac{n}{k}, c_1 \log k\right]. \tag{2.6}$$

The choice values of the constants c_0 and c_1 are $c_0 = 0.558$ and $c_1 = 0.5$ and they are obtained based on an analysis of the bias and variance of the estimator.

The estimator $\tilde{S}$ above is order-optimal *in the exponent* under the unbiased risk. Thus, the estimator can be improved by selecting coefficients of $Poly(L)$ that jointly optimize the bias and variance term in the risk. We show how to accomplish this task by rewriting the original minmax problem as a regularized exponentially weighted Chebyshev approximation problem [24].

In order to jointly optimize the bias and variance term in the squared loss, we start by directly analyzing $\sup_{P \in D_k} \mathbb{E}\left(\frac{S-\hat{S}}{k}\right)^2$. Classical Poissonization arguments lead to

$$\mathbb{E}\left(\frac{S-\hat{S}}{k}\right)^2 = \frac{1}{k^2}\Bigg\{ \sum_{i:\lambda_i>0} \left(\sum_{l=0}^{L} e^{-\lambda_i} a_l^2 \lambda_i^l l!\right) + \sum_{i\neq j:\lambda_i\lambda_j>0} \Big(e^{-\lambda_i} P_L(\lambda_i, \mathbf{a})\Big)\Big(e^{-\lambda_j} P_L(\lambda_j, \mathbf{a})\Big)\Bigg\},$$

where $P_L(\lambda, \mathbf{a}) \triangleq \sum_{l=0}^{L} a_l \lambda^l$. Taking the supremum over D_k we can bound the risk as

$$\begin{aligned} &\leq \sup_{\lambda:\lambda_i \in [\frac{n}{k}, n]} \frac{1}{k^2}\Bigg\{ \sum_{i:\lambda_i>0} \left(\sum_{l=0}^{L} e^{-\lambda_i} a_l^2 \lambda_i^l l!\right) \\ &\quad + \sum_{i\neq j:\lambda_i\lambda_j>0} \Big(e^{-\lambda_i} P_L(\lambda_i, \mathbf{a})\Big)\Big(e^{-\lambda_j} P_L(\lambda_j, \mathbf{a})\Big)\Bigg\} \\ &\leq \sup_{\lambda \in [\frac{n}{k}, n]} \left\{\frac{1}{k}\left(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l!\right) + \Big(e^{-\lambda} P_L(\lambda, \mathbf{a})\Big)^2\right\}. \end{aligned}$$

In the above inequality, we used the Cauchy-Bunyakovsky-Schwarz inequality, the fact that $S \leq k$ and $\left(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l!\right) > 0$, for all $\lambda > 0$. Hence, the optimization problem for the coefficients of the polynomial estimator at hand reads as

$$\inf_{\mathbf{a}\in Poly(L)} \sup_{\lambda\in[\frac{n}{k},n]} \left\{ \frac{1}{k}\bigg(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l! \bigg) + \bigg(e^{-\lambda} P_L(\lambda,\mathbf{a}) \bigg)^2 \right\}. \tag{2.7}$$

Problem (2.7) represents an instance of a *regularized weighted Chebyshev approximation problem.* If we ignore the first term in (2.7), the optimization problem becomes

$$\inf_{\mathbf{a}\in Poly(L)} \sup_{\lambda\in[\frac{n}{k},n]} \bigg(e^{-\lambda} P_L(\lambda,\mathbf{a}) \bigg)^2 .$$

The term $e^{-\lambda}P_L(\lambda,\mathbf{a})$ corresponds to the bias of the estimator. It is straightforward to see that the optimal choice of $\mathbf{a}$ for the above problem is a solution to

$$\inf_{\mathbf{a}\in Poly(L)} \sup_{\lambda\in[\frac{n}{k},n]} \bigg| e^{-\lambda} P_L(\lambda,\mathbf{a}) \bigg| . \tag{2.8}$$

The first term $\frac{1}{k}\bigg(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l! \bigg)$, which corresponds to the variance, may be written as

$$\begin{aligned} \frac{1}{k}\bigg(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l! \bigg) &= \mathbf{a}^T \mathbf{M}(\lambda) \mathbf{a} \\ &\triangleq ||\mathbf{a}||^2_{\mathbf{M}(\lambda)}, \mathbf{M}(\lambda) \triangleq \frac{e^{-\lambda}}{k} Diag(\lambda^0 0!, \lambda^1 1!, ..., \lambda^L L!). \end{aligned}$$

Clearly, $||.||_{\mathbf{M}(\lambda)}$ is a valid norm, and consequently, the first term in (2.7) can be viewed as a regularizer.

Simple algebra reveals that

$$\begin{aligned} \sup_{P\in D_k} \frac{1}{k} |\mathbb{E}(S-\hat{S}(N))| &\leq \sup_{\lambda\in[\frac{n}{k},n]} |e^{-\lambda} P_L(\lambda,\mathbf{a})| \qquad (2.9) \\ &\leq e^{-\frac{n}{k}} \sup_{\lambda\in[\frac{n}{k},n]} |P_L(\lambda,\mathbf{a})| = e^{-\frac{n}{k}} \sup_{\lambda\in[\frac{n}{k},n]} |\sum_{l=0}^{L} a_l \lambda^l|, \qquad (2.10) \end{aligned}$$

where (2.9) is equivalent to (2.8), while (2.10) resembles the problem studied in [23], except for a different optimization interval used within the supremum (the authors of [23] choose a shorter interval in order to decrease the contribution of the variance to the loss). Hence, optimizing (2.9) should produce an estimator with smaller bias as the exponential weight is inherent to the formulation. The modified bound in (2.10) is minimized with respect to the coefficients $\mathbf{a}$, using the minmax property of Chebyshev polynomials [28, 29], resulting in $\tilde{\mathbf{a}}$.

To solve (2.7), we more closely examine some results known about weighted Chebyshev approximations [29] and semi-infinite programs. Solving for the problem directly is difficult, so we instead resort to numerically solving the epigraph formulation of problem (2.7) and proving that the numerical solution is asymptotically consistent.

The epigraph formulation of (2.7) is of the form ([30], Section 6.1)

$$\min_{t,a_1,\ldots,a_L} t \quad \text{subject to}$$
$$\left\{\frac{1}{k}\left(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l!\right) + \left(e^{-\lambda} P_L(\lambda, \mathbf{a})\right)^2\right\} \leq t, \forall \lambda \in [\frac{n}{k}, n], \text{with } a_0 = -1. \tag{2.11}$$

Note that (2.11) is a semi-infinite programming problem. There are many algorithms that can be used to numerically solve (2.11), such as the discretization method, and the central cutting plane, KKT reduction and SQP reduction methods [31, 32]. For simplicity, we focus on the discretization method. For this purpose, we first form a grid of the interval $[\frac{n}{k}, n]$ involving s points, denoted by $\text{Grid}([\frac{n}{k}, n], s)$. Problem (2.11) may consequently be viewed as an LP with infinitely many quadratic constraints, which is not solvable. Hence, instead of addressing (2.11), we focus on solving the relaxed problem

$$\min_{t,a_1,\ldots,a_L} t \quad \text{subject to} \quad \left\{\frac{1}{k}\left(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l!\right) + \left(e^{-\lambda} P_L(\lambda, \mathbf{a})\right)^2\right\} \leq t, \tag{2.12}$$
$$\forall \lambda \in \text{Grid}([\frac{n}{k}, n], s), \text{with } a_0 = -1.$$

The solution of the relaxed problem is asymptotically consistent with the solution of the original problem (i.e., as s goes to infinity, the optimal values of the objectives of the original and relaxed problem are equal). Problem (2.12) is an LP with a finite number of quadratic constraints that may be solved using standard optimization tools. Unfortunately, the number of constraints scales with the length of the grid interval, which in the case of interest is linear in n. This is an undesired feature of the approach, but it may be mitigated through the following theorem which demonstrates that an optimal solution of the problem may be found over an interval of length proportional to the significantly smaller value $\log k$, where $\frac{k}{\log k} \lesssim n$ is the fundamental bound for support estimation. We relegate the proof to Appendix A.

Theorem. For any $\mathbf{a} \in Poly(L)$ and $L = \lfloor c_0 \log k \rfloor$, and $c_0 = 0.558$, let

$$g(\mathbf{a}, \lambda) = \frac{1}{k}\bigg(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l!\bigg) + \Big(e^{-\lambda} P_L(\lambda, \mathbf{a})\Big)^2.$$

Then, we have

$$\sup_{\lambda \in [\frac{n}{k}, n]} g(\mathbf{a}, \lambda) = \begin{cases} \sup_{\lambda \in [\frac{n}{k}, 6.5L]} g(\mathbf{a}, \lambda) & \text{if } \frac{n}{k} \leq 6.5L \\ g(\mathbf{a}, \frac{n}{k}) & \text{if } \frac{n}{k} > 6.5L. \end{cases}$$

Remark. In weighted approximation theory [24], the problem of bounding the interval over which the supremum is achieved is a topic of significant interest, with many important available results. For example, if we ignore the regularization term, we can directly use the Mhaskar-Saff theorem to reduce the length of the interval in the supremum to $\frac{\pi}{2}L$. Our Theorem shows that even when a regularization term is present, we can still restrict the length of the interval to $6.5L$. Our proof differs from that of the more general Mhaskar-Saff theorem, since we exploit the specific structure of the problem.

The optimization problem we need to solve to determine our estimator therefore reads as

$$\begin{aligned} &\min_{t, a_1, \ldots, a_L} t \quad \text{subject to} \\ &\bigg\{\frac{1}{k}\bigg(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l!\bigg) + \Big(e^{-\lambda} P_L(\lambda, \mathbf{a})\Big)^2\bigg\} \leq t, \\ &\forall \lambda \in \text{Grid}([\frac{n}{k}, 6.5L], s), \text{with } a_0 = -1. \end{aligned} \tag{2.13}$$

Since $L = \lfloor c_0 \log k \rfloor$, the length of the optimization interval in (2.13) is proportional to $\log k$.

It seems intuitive to assume that as s grows, the solution of the relaxed semi-infinite program approaches the optimal solution of the original problem (2.11). This intuition can be rigorously justified for the case of objective functions and constraints that are "well-behaved," as defined in [33] and [34]. The first line of work describes the conditions needed for convergence, while the second establishes the convergence rate given that the discretized solver converges. We use these results in conjunction with a number of properties of our objective SIP to establish the claim in the following theorem. The proof is relegated to Appendix A.

Theorem. Let s be the number of uniformly placed grid points on the interval (2.13), and let $d \triangleq \frac{6.5L-\frac{n}{k}}{s-1}$ be the length of the discretization interval. As $d \to 0$, the optimal objective value t_d of the discretized SIP (2.13) converges to the optimal objective value of the original SIP $t^\star$. Moreover, the optimal solution is unique $\mathbf{a}^\star$. The convergence rate of t_d to $t^\star$ equals $O(d^2)$. If the optimal solution of the SIP is a strict minimum of order one (i.e., if $t-t^\star \geq C||\mathbf{a}-\mathbf{a}^\star||$ for some constant $C > 0$ and for all feasible neighborhoods of $\mathbf{a}^\star$), then the solution of the discretized SIP also converges to an optimal solution with rate $O(d^2)$.

In summary, for given parameters k and n, and sample count histograms N, we obtain the optimal coefficients of our polynomial estimators by solving the small LP program described above. An example of our polynomial estimator (henceforth termed Regularized Weighted Chebyshev (RWC) estimator) and its scaled coefficients g_L is shown in Figure 2.3, along with a corresponding example of a Chebyshev estimator (termed the Wu-Yang (WY) estimator). It is easily observed that the coefficients of the two estimators exhibit very different behaviors: Unlike the Chebyshev case, for which the coefficients have to alternate in sign, our estimators are not constrained to obey this pattern.

Remark. It is important to point out that the RWC estimators are "additive": They operate on each symbol separately and the contributions of symbols are linearly combined to obtain the overall support estimate.

We conclude by observing that our RWC estimator can be further (heuristically) improved in practice by optimizing it with respect to a minmax risk that involves a different scaling factor in the denominator. This estimator, termed the RWC-S estimator (to indicate that the scaling is performed using the result of a naive support estimator) is described in more detail in

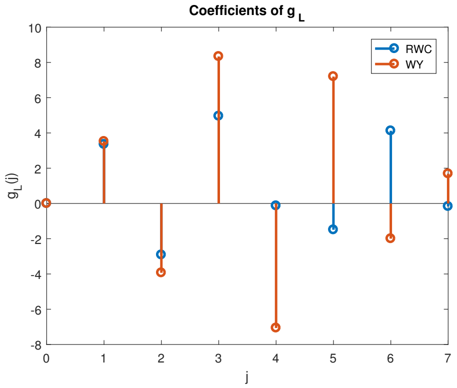


Figure 2.3: The function g_L for RWC and WY estimator. The parameter setting used for the illustration is $n = k = 10^6$ and $c_0 = 0.558$.

Appendix A.

2.4 Small-sample distribution estimation

By far the most frequently used method for distribution estimation in the small-sample regime is the Good-Turing estimator [13], which tries to account for the unseen by adjusting the counts (histograms) of the actually observed symbols. In a slightly modified form the method may be described as follows. For a sequence $\mathbf{x}^n$ of length n over an unknown finite alphabet, we once again let n_i denote the number of times a symbol i appears in $\mathbf{x}^n$. Furthermore, we let φ_t stand for the count of counts, i.e., the number of symbols that appear t times in $\mathbf{x}^n$. The estimator proposed in [14] combines the Good-Turing and ML estimators, the latter being used for the frequently observed symbols. For symbols that appear t times, if $\varphi_{t+1} > \Omega(t)$, then the Good-Turing estimate is used to determine the underlying total probability mass, otherwise the ML estimator is used instead. More precisely, for a symbol appearing t times, if $\varphi_{t+1} > t$ we use the Good-Turing estimator, otherwise we use the ML estimator. If $n_i = t$, the estimated probability of the symbol i is computed according to

$$p_i = \begin{cases} \frac{t}{\eta}, & \text{if } t > \varphi_{t+1}, \\ \frac{\varphi_{t+1}+1}{\varphi_t} \frac{t+1}{\eta}, & \text{otherwise}, \end{cases}$$

where η is a normalization term that ensures that the obtained values are probability masses. The term φ_{t+1} used in the Good-Turing estimator is replaced by $\varphi_{t+1}+1$ so that every symbol has a nonzero probability.

The modified Good-Turing estimator is used instead of the classical Good-Turing estimator as the latter is known to poorly estimate the probabilities of high frequency symbols. Modifications of the Good-Turing estimator that take sampling artifacts/errors into account are also available, and are implemented as described in [35, 36].

2.5 The performance of RWC estimators on synthetic data

Consider a finite alphabet $\mathcal{S} = \{1, \ldots, S\}$. Assume that the probability of symbol $i \in \mathcal{S}$ equals p_i and that you can randomly sample symbols from the alphabet with replacement and record the distribution histogram of N observed symbols. The question of interest is how accurately one can estimate S based on N samples and the parameter k dictating the smallest nonzero probability of the distribution.

For simplicity, assume that the alphabet is $\mathcal{S} = \{1, 2, \cdots, 10\}$ and that

$$p_i = \frac{i}{\sum_{j=1}^{10} j} = \frac{i}{55},\ i \in \{1, 2, \cdots, 10\}.$$

Clearly, $S = 10$ and $k = 55$. For the RWC and RWC-S estimators, we choose $L = \lfloor 0.558 \log k \rfloor = 2$. Now assume we draw $n = 6$ samples from the alphabet according to the specified distribution. In this case, the values of g_L for the RWC-S estimator are given in Table 2.2.

Table 2.2: The g_L values corresponding to the RWC-S estimator for different distinct symbol counts: Note that $\hat{S}_c$ denotes the naive estimator (i.e., the estimator equal to the count of different symbols).

	$\hat{S}_c = 1$	$\hat{S}_c = 2$	$\hat{S}_c = 3$	$\hat{S}_c = 4$	$\hat{S}_c = 5$	$\hat{S}_c = 6$
$g_L(0)$	0	0	0	0	0	0
$g_L(1)$	1.8128	2.4819	2.9427	3.2422	3.4367	3.5758
$g_L(2)$	1.8128	2.3967	1.7205	1.0699	0.5556	0.1663
$g_L(j), \forall j \geq 3$	1	1	1	1	1	1

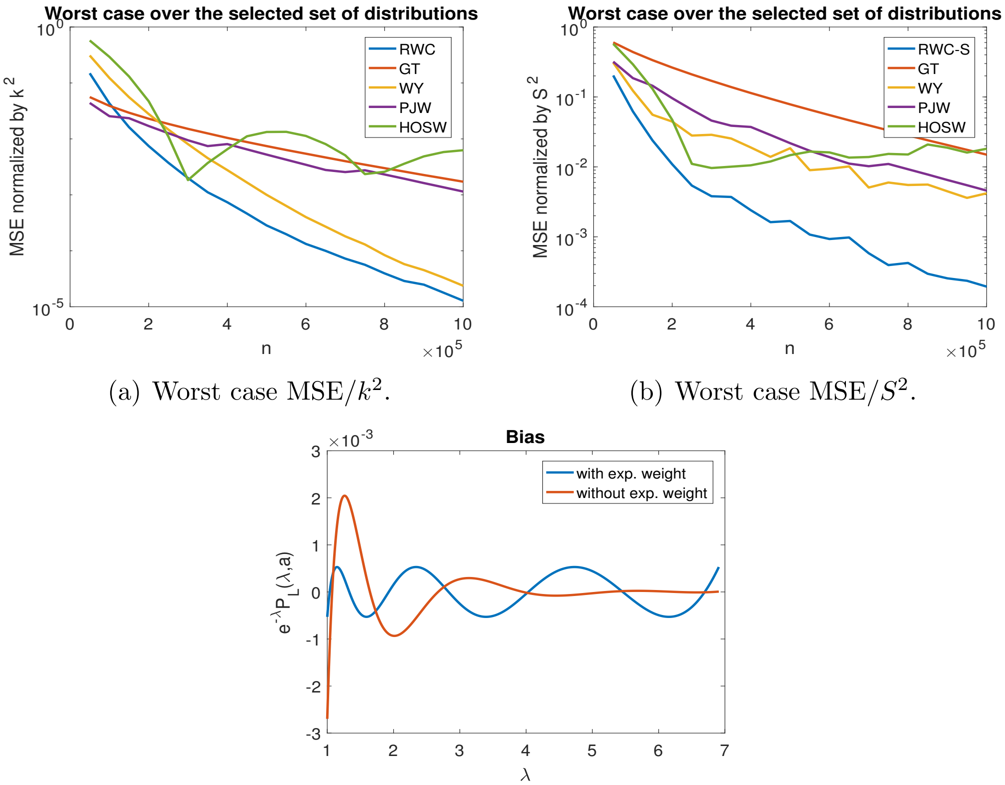
Consider all possible histograms of $n = 6$ symbols in this setting, summarized in Table 2.3. We can clearly see that except for the case $N = [1, 1, 1, 1, 1, 1]$, our estimator provides a significantly better support estimation result. Note that the histogram $N = [1, 1, 1, 1, 1, 1]$ arises only with very small probability (9%), and this probability significantly decreases as n, S, k increase. Nevertheless, even in this case, the risk (mean-square error normalized by S^2) of our RWC-S estimator equals 0.2186 while that of the naive estimator equals 0.319.

Table 2.3: The estimated supports produced by the RWC-S and naive estimators for all possible histogram inputs. The probability of each histogram is computed via a Monte Carlo method with 10^6 independent trials. Bold numbers indicate the best estimation result compared to the ground truth.

Histogram N	[1,1,2,2]	[1,1,1,1,2]	[1,1,1,3]	[1,2,3]	[1,5]	[1,1,4]
RWC-S	**8.6242**	**14.3034**	**10.7266**	**5.6632**	**3.4819**	**6.8854**
Naive	4	5	4	3	2	3
Probability	0.2633	0.3792	0.1374	0.0801	0.0021	0.0253
Histogram N	[2,2,2]	[3,3]	[2,4]	[6]	[1,1,1,1,1,1]	
RWC-S	**5.1615**	**2**	**2.7205**	**1**	21.4548	
Naive	3	**2**	2	**1**	**6**	
Probability	0.0171	0.0025	0.0043	0.0001	0.0886	

We also tested the performance of the estimators on significantly larger sets of synthetic data for which the ground truth distributions and their supports are known. In particular, we compared the RWC method with the Good-Turing (GT) estimator, the WY estimator of [23], the PJW estimator described in [37] and the HOSW estimator of [38]. We did not compare our method with the estimators introduced in [26, 39] due to their high computational complexity [38].

We considered six different distributions: the uniform distribution with $p_i = \frac{1}{k}$, the Zipf distributions with $p_i \propto i^{-\alpha}$, and α equal to 1.5, 1, 0.5 or 0.25, and the Benford distribution with $p_i \propto \log(i + 1) - \log(i)$. We choose the support sizes for the Zipf and Benford distribution so that the minimum non-zero probability mass is roughly 10^{-6}. We run the estimator 100 times to calculate the risk. For both approximation-based estimators, we fix c_0 to be 0.558. With our proposed method, we solve (2.13) on a grid with $s = 1000$ points on the proposed interval $[\frac{n}{k}, 6.5L]$. For the estimator described in [23], we set $c_1 = 0.5$ according to the recommendation made in the cited paper.



(a) Worst case MSE/k^2.

(b) Worst case MSE/S^2.

(c) RWC (weighted) versus classical Chebyshev approximation.

Figure 2.4: Comparison of worst case risks for different ground truth distributions and estimators. In our simulations we set $n = k = 10^6$ and $c_0 = 0.558$.

The GT method used for comparison first estimates the total probability of seen symbols (e.g., sample coverage) according to $\hat{C} = 1 - \frac{h_1}{n}$, and then estimates the support size according to $\hat{S}_{GT} = \frac{\hat{S}_c}{\hat{C}}$; here, $\hat{S}_c$ stands for the (naive) counting estimator. Note that h_1 equals the number of different alphabet symbols observed only once in the n samples.

Figure 2.4(a) shows that the RWC estimator has a significantly better worst-case performance than all other methods when tested on the above described collection of distributions, provided that $n \geq 0.2k$. Also, both RWC and WY estimators have significantly better error exponents compared to the GT, PJW and HOSW estimators. The GT and PJW estimators perform better than RWC if $n \lesssim \frac{k}{\log k}$, which confirms the results of our theoretical analysis as well.

In the second set of experiments, we change the normalization from $(1/k)^2$

to $(1/S)^2$ as was also done in [38]. The RWC-S estimator minimizes an upper bound on the worst-case risk $\mathbb{E}\left(\frac{\hat{S}-S}{S}\right)^2$. As already pointed out, a detailed description of this algorithm and an intuitive explanation of why it outperforms the RWC method is provided in Appendix A. Figures 2.4(b) illustrate that our RWC-S estimator significantly outperforms all other estimators with respect to the worst-case risk normalized by S^2. Moreover, the RWC-S estimator outperforms all known estimators on almost all tested distributions. As illustrated in Figure 2.4(c) we see that a classical Chebyshev approximation introduces a larger bias than our RWC method whenever the underlying distribution is close to uniform (i.e., when $\lambda \sim \frac{n}{k} = 1$). This phenomenon persists even when regularizations is taken into account.

Another common approach to testing support estimators on real data is to estimate the number of distinct words in selected books [26, 23]. Books are chosen as ground truth test cases as the words in a text are not independent and identically distributed (i.i.d.) and hence provide a means to test the performance of estimators optimized for i.i.d. settings. The performance of our approach and those of prior works on Hamlet and Macbeth can be found in Appendix B. In the experiments, we randomly sampled words in the text with replacement and used the obtained counts to estimate the number of distinct words. For simplicity, we set k to the total number of words. For example, as the total number of words in Hamlet equals $30,364$, we set $k = 30,364$. Once again, our method significantly outperforms all other competitive techniques both in terms of convergence rate and the accuracy of the estimated support for all experiments.

The details about data acquisition pipeline, alignment software and implementations of the RWC and RWC-S algorithms may be found at the following GitHub repository: `https://github.com/rana95vishal/Mutational-landscape-SARS-Cov-2`

CHAPTER 3

RESULTS AND DISCUSSION

We proceed to apply our small-sample support and distribution estimation methods on GISAID SARS-Cov-2 genomic datasets. The underlying assumption is that there exists a “ground truth” distribution of mutations, and that most of the mutations cannot be observed due to limited testing. Our studies of the mutational support and mutation distribution are conducted for different patient subpopulations and all ORFs separately in order to determine potential subpopulation differences. As already pointed out, the estimators to be used are additive implying that estimates for individual genes may be summed to obtain the estimate for the whole genome.

First, we observe that by the last small-sample data collection date reported in this thesis, 04-14-2020, the average number of mutations with respect to the reference was 7.93 (for male patients) and 7.96 (for female patients). This difference is statistically insignificant. For patients older than 55 years, this number was 7.33 while for those younger than 55 the recorded values were significantly higher, amounting to 8.377. For three different continents, Asia, Europe and NA, the average number of mutations recorded equaled 13.51, 6.67, and 6.68, respectively. The average number of mutations per patient in Asia is almost twice as large as the corresponding numbers in Europe and NA, which is indicative of the fact that the outbreak started in Asia and that the virus may have been present in the population significantly longer than in Europe and NA. In all cases, the total number of recorded mutations across all patients is too small to allow for accurate prediction of the actual mutational support using frequency methods.

3.1 Mutational support estimation

The first set of results pertains to data collected at a very early stage of the pandemic (04-03-2020) that did not include sufficiently many samples to allow for sample set sizes to be evened out through subsampling. Therefore, for this analysis, all available samples are included, which may create biases due to sample set size differences. The results are listed in Tables 3.1, 3.2 and 3.3. They illustrate the difference in the support estimates for two different age groups, genders and three geographic regions. The nonuniform sample size artifacts do not obscure the most important findings regarding mutation rates in different genes across different age groups, gender and geographic region - the same trends persist even when significantly more samples are used in the analysis, as described next.

Table 3.1: Support sizes of different age groups based on 909 samples for individuals over 55 years of age and $1,477$ samples below 55 years of age. The data was obtained from GISAID by 04-03-2020 and includes all the samples for the two categories available at the given date. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	> 55	< 55	> 55	< 55	> 55	< 55	All Ages
ORF1a	**625**	**764**	**1,280**	**1,544**	**1,209**	**1,454**	**13,203**
ORF1b	**276**	**616**	**570**	**1,301**	**514**	**1,223**	**8,087**
S	160	218	291	420	277	375	3,822
ORF3a	55	73	103	132	92	121	828
E	14	13	23	23	23	22	228
M	34	35	58	63	54	55	669
ORF6	11	25	19	42	19	42	186
ORF7a	24	27	41	45	39	44	366
ORF8	340	340	87	344	235	343	366
N	**66**	**110**	**108**	**197**	**97**	**172**	**1,260**
ORF10	26	29	29	53	33	53	117

Tables 3.4, 3.5 and 3.6 list the results analogous to those reported for 04-03-2020, obtained using datasets retrieved on 04-10-2020. The datasets were sufficiently large to allow for random subsampling to obtain equal sample set sizes for all subpopulations considered (excluding Asia).

Based on the results of Table 3.4, we see that the mutational supports in populations of different age (cutoff at 55 years) differ substantially for the

Table 3.2: Support sizes based on $1,349$ male and $1,061$ female samples. The data was obtained from GISAID by 04-03-2020 and includes all the samples for the two categories available. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Male	Female	Male	Female	Male	Female	Both Genders
ORF1a	**854**	**702**	**1,807**	**1,468**	**1,702**	**1,388**	**13,203**
ORF1b	**348**	**594**	**690**	**1,307**	**640**	**1,234**	**8,087**
S	225	186	447	359	405	329	3,822
ORF3a	68	61	132	111	115	99	828
E	18	10	30	18	29	18	228
M	37	36	62	68	57	60	669
ORF6	13	27	22	49	21	50	186
ORF7a	32	21	55	38	53	38	366
ORF8	340	341	344	592	343	458	366
N	**96**	**85**	**165**	**143**	**146**	**129**	**1,260**
ORF10	26	10	30	17	29	17	117

Table 3.3: Support sizes for different geographical regions based on 510 samples from Asia, $1,695$ from Europe and 818 from NA. The data was obtained from GISAID by 04-03-2020 and includes all the samples for the three categories available at the given date. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively. Maximum support for all the genes is the same as shown in previous tables.

Gene	ML			RWC			RWC-S		
Symbol	Asia	Europe	NA	Asia	Europe	NA	Asia	Europe	NA
ORF1a	**770**	**757**	**397**	**1,645**	**1,558**	**776**	**1,603**	**1,455**	**720**
ORF1b	**279**	**590**	**205**	**566**	**1,251**	**372**	**553**	**1,159**	**345**
S	168	181	131	321	345	254	313	309	230
ORF3a	84	62	38	158	113	71	154	100	63
E	37	11	6	66	19	9	65	19	9
M	30	29	15	53	49	25	50	44	24
ORF6	2	28	5	2	46	8	2	45	7
ORF7a	108	38	49	215	66	90	214	65	89
ORF8	340	27	19	341	46	26	342	43	28
N	**53**	**90**	**68**	**93**	**152**	**122**	**85**	**137**	**114**
ORF10	10	25	9	18	28	15	17	27	14

ORF3a, ORF6 and ORF7a regions (note that ORF1ab and N are shown in **bold** in every table due to their large length and relevance in testing, re-

Table 3.4: Support sizes of different age groups based on $1,850$ samples from each group. The data was retrieved from GISAID on 04-10-2020. The mutational supports between the two groups differ substantially for the genes shown in *italics*. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively.

Gene	ML		RWC		RWC-S		Maximum Support
Name	> 55	< 55	> 55	< 55	> 55	< 55	All Ages
ORF1a	**996**	**934**	**2,039**	**1,857**	**1,896**	**1,743**	**13, 203**
ORF1b	**499**	**484**	**991**	**965**	**924**	**896**	**8,087**
S	265	279	490	547	458	501	3,822
ORF3a	*104*	*79*	*188*	*138*	*171*	*124*	*828*
E	23	19	36	33	36	32	228
M	55	47	98	86	92	77	669
ORF6	*38*	*26*	*65*	*43*	*64*	*41*	*186*
ORF7a	*60*	*31*	*108*	*50*	*103*	*49*	*366*
ORF8	340	341	93	342	236	343	366
N	**140**	**163**	**248**	**294**	**223**	**265**	**1,260**
ORF10	31	28	35	49	39	50	117

Table 3.5: Support sizes for different genders based on $1,956$ samples for each group. The data was retrieved from GISAID on 04-10-2020. The mutational supports between the two groups differ substantially for the genes shown in *italics*. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively.

Gene	ML		RWC		RWC-S		Maximum Support
Name	Male	Female	Male	Female	Male	Female	Both Genders
ORF1a	**1,071**	**1,115**	**2,176**	**2,313**	**2,055**	**2,175**	**13,203**
ORF1b	**500**	**804**	**1,013**	**1,721**	**941**	**1,621**	**8,087**
S	283	293	551	562	509	519	3,822
ORF3a	114	99	216	175	190	158	828
E	*24*	*14*	*37*	*23*	*36*	*22*	*228*
M	52	56	87	101	82	94	669
ORF6	*42*	*30*	*75*	*51*	*74*	*50*	*186*
ORF7a	42	51	74	87	71	84	366
ORF8	341	342	344	345	344	345	366
N	**143**	**162**	**251**	**282**	**226**	**259**	**1,260**
ORF10	*29*	*12*	*33*	*20*	*32*	*19*	*117*

spectively). For ORF7a, the older population exhibits almost twice as many mutations as the younger population, while for ORF6 and ORF3a the corresponding numbers are 1.5 and 1.4, respectively; the estimated mutational

Table 3.6: Support size for three different geographic regions based on 615 samples from Asia and $1,000$ samples from Europe and NA each. The data was retrieved from GISAID on 04-10-2020. The mutational supports between the three groups differ substantially for the genes shown in *italics*. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively. Maximum support for all the genes is the same as shown in previous tables.

Gene	ML			RWC			RWC-S		
Name	Asia	Europe	NA	Asia	Europe	NA	Asia	Europe	NA
ORF1a	**827**	**504**	**470**	**1,768**	**975**	**948**	**1,725**	**919**	**874**
ORF1b	**308**	**271**	**244**	**631**	**531**	**478**	**611**	**491**	**432**
S	*182*	*163*	*142*	*352*	*336*	*269*	*340*	*293*	*243*
ORF3a	*91*	*56*	*39*	*174*	*96*	*74*	*168*	*85*	*63*
E	*37*	*12*	*14*	*66*	*21*	*24*	*65*	*21*	*24*
M	*31*	*23*	*17*	*55*	*38*	*28*	*52*	*35*	*27*
ORF6	*3*	*48*	*15*	*3*	*87*	*26*	*3*	*86*	*25*
ORF7a	*109*	*63*	*51*	*216*	*118*	*94*	*214*	*116*	*93*
ORF8	*340*	*19*	*21*	*335*	*29*	*31*	*339*	*29*	*31*
N	**58**	**72**	**77**	**96**	**121**	**137**	**91**	**108**	**129**
ORF10	*10*	*26*	*7*	*18*	*48*	*10*	*17*	*48*	*10*

supports of the ORF6 and ORF7a regions are close to 1/3 of the whole gene length for individuals older than 55 years. The mutational differences in the ORF6 and ORF7a region persist with an increase in the number of samples (see the Additional Table C.1), with an estimated mutational support for the former region equal to almost 1/2 of the gene length. Furthermore, additional differences are observed in the M region which were not apparent when using smaller sample set sizes. The protein encoded by ORF6 was studied in depth during the SARS epidemics [40] and it has been established that the ORF6 protein impairs the nuclear import complex formation (controlling the transport of innate immune regulatory cargo to the nucleus of cells capable of increasing antiviral defenses). The protein encoded by ORF7a has been implicated in inhibiting bone marrow stromal antigen 2 virion tethering [41]. Bone marrow stromal antigen 2, also known as tetherin, is an interferon-induced protein which, when expressed, reduces the release of enveloped viral particles. The significant number of predicted mutations in the ORF7a region of older patients suggests a similar observation as that made for the ORF3a region - a possible effort by the virus to disable or strongly impair the function of the tetherin antigen.

The results pertaining to female/male patients differ significantly from those pertaining to different age groups. The results are listed in Table 3.5, and imply strong differences in the mutation rates of the ORF1b and ORF10 regions. The mutational support of ORF1b in the female population is $1,621$ compared to 941 in the male population, which amounts to a 8.4% difference with respect to the length of the ORF. A similar result is true for the ORF10 region, for which no well-understood functions are known. Some recent results suggest, based on different evidence, that ORF10 encodes a functional protein in SARS-CoV-2 and that positive selection is driving the evolution of this region [42].

The above described differences persist with increased sample set sizes. The estimated mutational support for ORF1b is 24% and 16% of the length of the region, and for ORF10 18% and 32% of the length of the region, for females and males, respectively (see the Additional Table C.2). Smaller yet possibly relevant differences are also observed for the ORF3a and M regions, but these do not persist with increased sample set sizes.

For samples obtained from Asia, Europe and NA the results show that despite the number of samples for Asia being significantly smaller than that of Europe and NA, the predicted mutational support in all regions is significantly higher except for the N and ORF6 genes (with only 3 mutations observed in the ORF6 gene). This is particularly the case for ORF3a and ORF8, where the mutation rates are more than 2 and 10-fold higher in Asian patients, respectively. It is reasonable to assume that these regions are mutated early on in an epidemic and tend to "accumulate" the number of mutations. Also, the significant differences suggest that the epidemic started *significantly* earlier in Asia than Europe and NA. The ORF3a region is known to encode for a protein that activates the NLRP3 inflammasome [43]. ORF3a proteins are activators of pro-IL-1β gene transcription and protein maturation that trigger activation of the NLRP3 inflammasome. The inflammasome has a dual role of boosting the host defense and driving pathologic inflammation. Based on our findings, one possible explanation for the high mutation rate in this region in older populations is that the virus trying to disable the host's immune system and increase its virulence. Recent results show that the ORF8 protein may be acquired from SARS-related coronaviruses present in bats [44], which could explain the large difference in the mutational support through "adaptation" in a human host (for patients in Asia). The increase in the number of samples available for analysis shows that significant differences in the mutation support of the E, M, ORF6, ORF7a and ORF10 regions exist as well.

Additional Tables C.1, C.2 and C.3 show the trends of increase for the mutational support with increased sample sizes. For data collected by 04-14-2020, this includes roughly $9,000$ samples. All sample set sizes used are equal (except for Asia, for which the sample set sizes available are significantly smaller), therefore allowing for fair comparisons. Additional Table C.1 illustrates that when the sample set sizes are equal, no significant differences are observed in the mutational supports of disparate age groups except in the E, ORF6, ORF7a and ORF8 regions. Given that the difference in the number of mutations in the ORF7a regions persists for several data acquisition dates, the finding appears to be sample-size independent. On the other hand, the significant differences in the number of mutations in the E region is only evident when sufficiently many samples are available. The E region contains the code for the encapsulation protein of viral RNA, in addition to some spike proteins. In older subjects, this region is subjected to a significantly larger number of mutations than in other groups. This may imply that immunity in elder patients may be dependent on generating antibodies for the encapsulation proteins. Clearly, no conclusive explanation is possible based on limited data sets but the results suggest performing further sampling and analysis for this particular ORF in older patients. Although it has been observed that the immune responses of individuals vary significantly due to the initial viral load, physical health, and the hosts microbiome, no definite link between these features and the mutation rates in the above region can be established due to lack of supporting clinical data at GISAID and other Covid-19 data repositories.

Additional Table C.2 illustrates surprisingly few differences in the mutational supports of male and female patients once a sufficiently large number of samples is available: Exceptions are the ORF1b and ORF10 regions. For different geographic regions, the most significant difference observed pertains to the ORF8 region, where samples from Asia exhibit a roughly one order of magnitude larger number of mutations compared to those for samples sequenced in Europe and NA. There also exists a marked difference in the mutational support of ORF7a between patients from Europe on one side and patients from Asia and NA on the other (i.e., a roughly two-fold difference for Europe and NA).

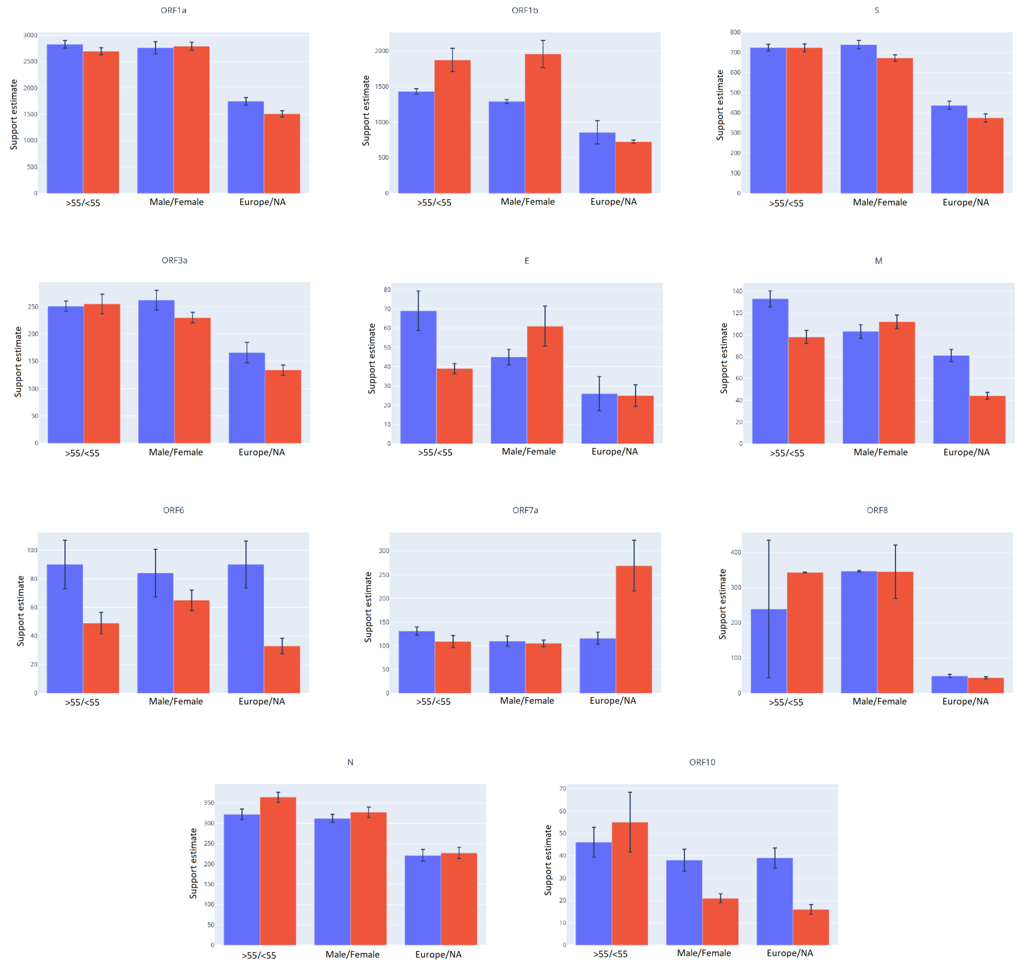


Figure 3.1: Support sizes for all genes and for different groups along with their standard deviation estimates. The estimates are based on data collected by 04-14-2020 and should provide the most accurate assessment of the mutation rate in the small-sample regimes investigated. Estimates are based on 3, 047 samples each for patients above 55 and below 55 years of age, 2, 817 samples each for male and female patients and 1, 774 each for patients exposed in Europe and North America.

Ten additional data collection days (starting on 04-03-2020, ending on 04-14-2020) lead to more than twice the samples, and the results for the latter date are shown in Figure 3.1 along with the standard deviations of the estimators. Note that in order to estimate the variance of an estimator, one needs to subsample the data which requires more samples to start with; hence, the standard deviation is only evaluated for all samples available by 04-

14-2020. The additional data samples show that the N region of the SARC-Cov-2 genome exhibits a much more significant increase in mutations than could have been predicted from early small-set sample sizes, amounting to roughly an average of 23% of the genome, across populations. This finding is significant as it suggests that genomic regions used as identifiers for the virus may mutate much faster than predicted based on small preliminary sample set information. Nevertheless, the N1 and N2 regions used as primer targets for RT-PCR testing (the use of region N3 as a primer has been discontinued) appear to be largely unmutated. This is illustrated in the Additional Table C.11 which lists a total of only 8 mutations observed in these regions in the SARS-Cov-2 genomes of US patients. Similar results for mutations in viral genomes of patients from China are presented in Additional Table C.12.

Table 3.7 provides results for a finer partition of test samples into two categories, one including males over 55 years of age and another females below 55 years of age, with both populations sampled from Europe. The first category has been empirically observed to be at higher risk of infection and for exhibiting more severe symptoms [22]. Substantial mutational differences are observed in the ORF1b, S and ORF10 regions. The differences in the ORF1b and ORF10 genes appear to be mostly gender specific, while the age factor may contribute to the differences in the mutation rates of the S region. Another important finding is that the mutational support of ORF1b is almost twice as large in the low risk population compared to the high risk population. This result may imply that the large mutational support is a result of a highly competitive virus-host interaction which forces the virus to mutate the proteins encoded by ORF1b in order to gain an advantage over the host's immune system.

Figure 3.1 shows the mutational support sizes, along with the standard deviations of the estimates for six different patient categories. Since the true distribution of mutations of various groups of patients is not known, one cannot directly calculate the standard deviation of the support sizes produced by our estimator. To compute the standard deviation, we therefore subsample 85% of the available samples and compute the support size for the returned aggregate mutation profile. Our samples are chosen randomly and uniformly over the whole period of data collection, and for each month of data collection samples are retrieved separately and in proportion to the total number of samples available for that month. Since the number of samples collected and made available during the months of December and January is small, we group these two months together in the subsampling process. Subsampling is repeated 100 times resulting in 100 aggregate mutation profiles and corresponding support size estimates.

The mutational supports generated by our procedure have variances that demonstrate good concentration of the estimates; some exceptions exist, though, and are most likely not a consequence of the estimation procedure itself but rather an indicator of disparately collected datasets or some unknown governing biological process. The latter is supported in part by previously observed high rates of mutations in certain SARS-Cov-2 genes [45, 46]. The results for ORF8 are particularly interesting because the corresponding standard deviations of the mutational support vary significantly across different categories of patients: The standard deviation of the support size is close to 200 times higher for patients above 55 years of age than patients below 55 years of age.

We also performed a collection of tests in which alignments and mutational counts were performed with respect to the first sample from the same geographical region. Hence, for patients from Asia, the alignments and mutation counts are still performed with respect to the genome of the Wuhan-Hu-1 patient. For NA, we used the sample USA/WA1/2020 with ID EPI_ISL_404895, while for Europe we used the sample France/IDF0372/2020 with ID EPI_ISL_406596, both being the chronologically first samples from NA and Europe available at GISAID. For this study, we only used samples retrieved by 04-14-2020. The results are available in Additional Tables C.4-C.7. As expected, the mutational support estimates are lower for both the NA and European sample sets. However, one important and interesting exception pertains to the estimates for the gene N regions and samples from Europe, as well as samples for males above 55 from Europe, which are higher for the alignment and mutation counts performed with respect to Patient 1 in Europe. The same is true for mutational support estimates for gene N and under gender stratification. Additional differences were observed in the mutational support of the ORF6a and ORF7 regions in younger females versus older males when focusing on patients from Europe only and when using Patient 1 from Europe as the alignment reference. These results suggest different mutational patterns for viruses hosted by high-risk populations in Europe versus those in NA and Asia.

Table 3.7: Support size differences between males above 55 years of age and females below 55 years of age from Europe based on $1,078$ samples in each group. The data was retrieved from GISAID by 04-14-2020. The mutational supports between the two groups differ substantially for genes with values shown in *italics*. ORF1ab and N are shown in **bold** due to their large length and relevance in testing, respectively. Maximum support for all the genes is the same as shown in previous tables.

Gene	ML		RWC		RWC-S	
Symbol	M, >55	F, <55	M, >55	F, <55	M, >55	F, <55
ORF1a	**588**	**670**	**1,159**	**1,374**	**1,078**	**1,294**
ORF1b	**349**	**553**	**686**	**1,189**	**638**	**1,117**
S	*209*	*166*	*420*	*329*	*387*	*296*
ORF3a	76	61	138	104	124	96
E	10	9	17	15	16	14
M	27	33	45	58	40	52
ORF6	*15*	*28*	*25*	*47*	*24*	*48*
ORF7a	*31*	*23*	*54*	*36*	*52*	*36*
ORF8	27	28	45	48	43	46
N	**110**	**108**	**197**	**199**	**178**	**183**
ORF10	*27*	*5*	*28*	*7*	*33*	*7*

It is important to note that for some genes and patient categories it appears the RWC estimates roughly double those of the ML estimator, but this is **not a general trend** of the analysis. For example, the mutational support estimates for ORF8 for male and female are approximately equal to ML estimates (Table 3.5) and more pronounced differences exist across the whole subpopulation spectrum. Similar trends are observed for ORF6 in Asian subjects, and ORF10 across different subpopulations. Furthermore, although **the naive ML estimates may lead to similar conclusions regarding the trends of mutations in some ORFs, the degree of the trend and the scale of the mutation rates within different regions cannot be fully understood through the use of ML estimates only.** As an illustrative example, the ML estimator implies that there is no difference in the mutational supports of the ORF8 region in young versus old patients (Table 3.4), as the values equal 340 and 341, respectively. On the other hand, the RWC-S estimator predicts mutational supports of 236 and 343, respectively, which show a very different stratification.

We conclude by pointing out that one way to validate the results for our support estimation methods is to compare the results of the ML mutation counts at a later date with the computed estimates. We compare the mutational supports using the small-sample techniques and the data collected by 04-10-2020 with the actual count (ML estimates) generated from data retrieved by 04-14-2020. In this time period, the number of samples increased by roughly $3,000$, as may be seen from Table 2.1. The results are listed in Table 3.8. As may be seen, the estimates obtained based on data acquired by 04-10-2020 for Europe and NA and all open reading frames are excellent matches for the actual counts obtained by 04-14-2020, indicating that the number of samples was sufficient to predict the growth trend. Much more significant differences are observed for Asia, which can clearly be attributed to the very small sample sizes available from that continent on both 04-10-2020 and 04-14-2020 or potential strong correlations between the mutations in the three aforementioned regions. Other categories that are of interest involve male/female patients for which the actual counts from 04-14-2020 are significantly smaller than the estimates. This is indicative of a large number of potentially unseen mutations harbored by these populations.

Finally, Table 3.9 shows the support estimates for samples from patients from Asia for a more recent date of data collection, 10-20-2020. In this case, almost $10,000$ samples from Asia were readily available which allows one to get significantly improved results for ML estimators. As may be seen, the differences between ML and RWC-S values are significantly smaller, and for some reason even close to equal when a very different trend was true for data collected in April. In particular, the ratio of the number of estimated mutations in the ORF E region for the RWC-S and ML method was close to 1.76 in April, and only 1.24 in October. Similar findings are apparent for other ORFs.

3.2 Distribution estimation

Next, we report on the distribution of mutations in the ORF1a,b and N regions of the SARS-Cov-2 virus obtained using the Good-Turing estimator and once again focus on the traits of different subpopulations. We focus on these regions as the first two regions are the longest genes while the N region is of importance for Covid-19 testing in NA. As may be seen from

Table 3.8: Comparison of small-sample estimates of RWC-S based on data retrieved by 04-10-2020 and the ML estimates based on data retrieved by 04-14-2020.

Gene	Method-Date	Asia	Europe	NA	> 55	< 55	Male	Female
ORF1a	RWC-S (04-10)	1,725	919	874	1,896	1,743	2,055	2,175
	ML (04-14)	835	911	804	1,488	1,439	1,478	1,456
ORF1b	RWC-S (04-10)	611	491	432	924	896	941	1,621
	ML (04-14)	316	477	403	787	953	705	991
S	RWC-S (04-10)	340	293	243	458	501	509	519
	ML (04-14)	188	246	209	431	400	405	389
ORF3a	RWC-S (04-10)	168	85	63	171	124	190	158
	ML (04-14)	93	99	81	156	165	169	140
E	RWC-S (04-10)	65	21	24	36	32	36	22
	ML (04-14)	36	15	15	43	26	30	36
M	RWC-S (04-10)	52	35	27	92	77	82	94
	ML (04-14)	31	51	28	79	62	67	69
ORF6	RWC-S (04-10)	3	86	25	64	41	74	50
	ML (04-14)	3	52	21	53	32	50	40
ORF7a	RWC-S (04-10)	214	116	93	103	49	71	84
	ML (04-14)	109	66	135	86	66	68	72
ORF8	RWC-S (04-10)	339	29	31	236	343	344	345
	ML (04-14)	340	32	29	341	343	343	342
N	RWC-S (04-10)	91	108	129	223	265	226	259
	ML (04-14)	60	139	138	201	219	195	204
ORF10	RWC-S (04-10)	17	48	10	39	50	32	19
	ML (04-14)	11	30	10	35	33	31	13

Figures 3.2 and 3.3 there is a surprisingly small difference in the distribution of the top-20 mutated sites across different age and gender groups, except for a marked difference in the largest probability (in particular, in the N region for populations partitioned according to age and populations partitioned according to gender when including larger sample sets from 04-14-2020 as seen in Figure B.2). This is especially the case for samples partitioned according to gender, despite the fact that the number of mutations in female subjects in the ORF1b region was close to twice as large as that in male subjects. In addition, the probability of having a mutation at the highest probability sites is significantly larger in “younger” than “older” populations. The trend remains the same for data collected by 04-14-2020 as supported by the results in Figure B.2. Figure B.3 gives similar results for alignment per-

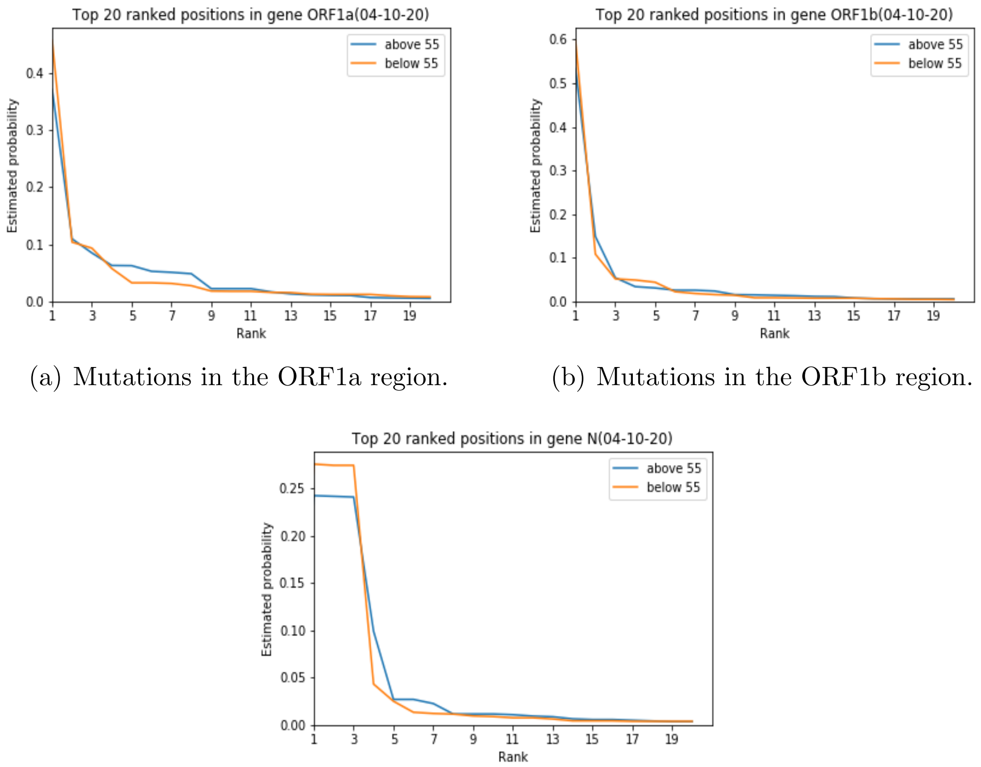
Table 3.9: ML and RWC-S estimates for mutational support in ORFs of patients from Asia based on data collected by 10-20-2020. At that date, substantially more samples $(9,271)$ were available for analysis. The standard deviation values are given in parentheses.

	ORF1a	ORF1b	S	ORF3a	E	M
ML	5,020 (77)	2,691 (42)	1,418 (27)	464 (19)	115 (3)	188 (4)
RWC-S	8,481 (152)	4,435 (80)	2,227 (61)	674 (34)	143 (5)	262 (8)
	ORF6	ORF7a	ORF8	N	ORF10	
ML	90 (3)	304 (9)	363 (1)	510 (6)	45 (2)	
RWC-S	112 (7)	333 (15)	361 (7)	711 (14)	61 (3)	

formed against the first sequenced patient from each region. The situation is completely different when comparing the distributions of mutations across different geographic regions (Figure 3.4), where there are significant differences in the distributions as one would expect. To compactly summarize the differences in the distributions, we also computed all three pairwise symmetric Kulback-Leibler (KL) divergences for the normalized top-20 mutation probabilities as described below. We also list the Jaccard distances between the sets of 20 most frequently mutated sites.

The distributions of mutations only reveal the statistical landscape of the mutation sites but not their exact locations in the genome. The actual mutated sites in the SARS-Cov-2 genomes are depicted in Figures 3.5 and 3.6, in addition to a more detailed set of results presented in the Figures B.4 and B.5. We selected the latest retrieval data for this analysis as it most accurately reflects the positions undergoing most frequent mutations; we also focused on two cohorts of patients for which the mutational landscapes differ the most. The positional stratification of mutations is significant for patients from different continents, especially in the N region of the SARS-Cov-2 genome. The largest spread of probability mass is (as expected) observed for patients from Asia which is indicative of the larger exploration rate for mutations in the region where the outbreak originated.

Additional Table C.8 lists the 10 most frequently mutated sites in the ORF1a region of all previously analyzed patients categories when alignment



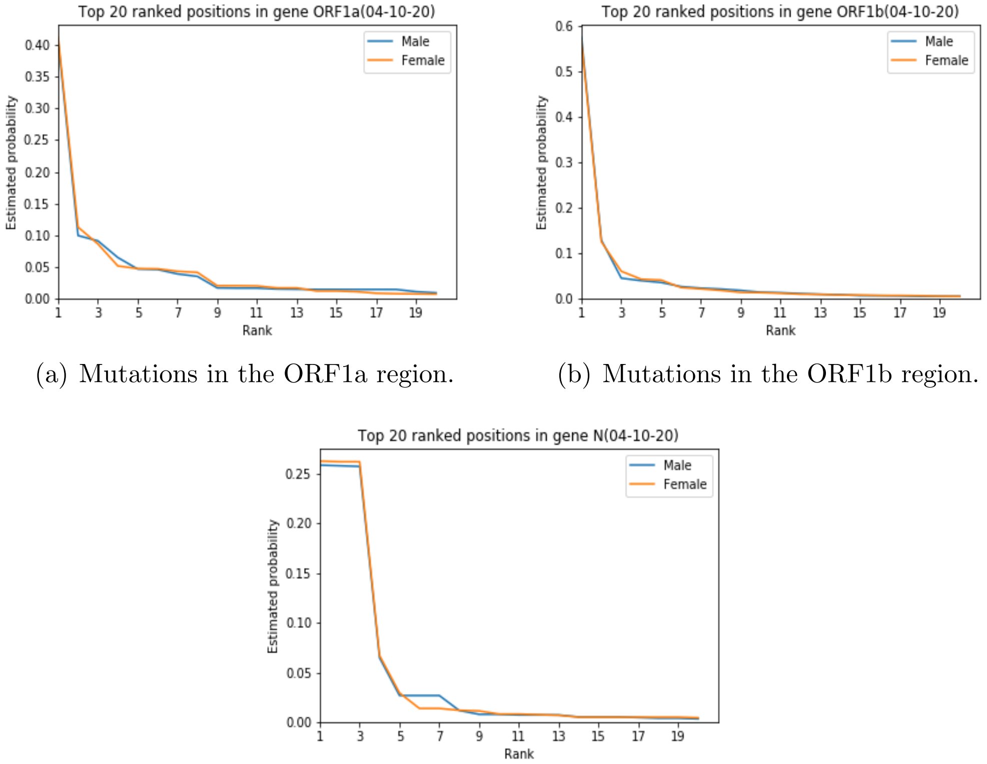
(a) Mutations in the ORF1a region.

(b) Mutations in the ORF1b region.

(c) Mutations in the N region.

Figure 3.2: Comparison of the estimated distributions of mutations in adults <55 of age and adults >55 of age tested by 04-10-2020. Almost all the probability mass is concentrated on five sites. The biggest observed difference occurs in the N region.

is performed with respect to the first patient sequenced in the geographic region. For the age and gender groupings, as expected, the top-ten sites are the same except for one difference encountered in both cases (shown in **bold**). A mutation in position $8,781$ of Asian and NA viral samples appears with high frequency but is surprisingly not present in the list of top mutated sites in the European population. Similarly, Additional Table C.9 lists the 10 most frequently mutated sites in the ORF1b region of all previously analyzed patients categories. As one may expect from the differences in the mutational support, the frequent sites of mutations differ significantly more in this region for different age groups, gender and continents when compared to the ORF1a region. This is especially the case when viewing the results for different geographic regions as except for the top-ranked site and one more site (i.e., sites $14,407$ and $14,804$); all other locations are different. This suggests



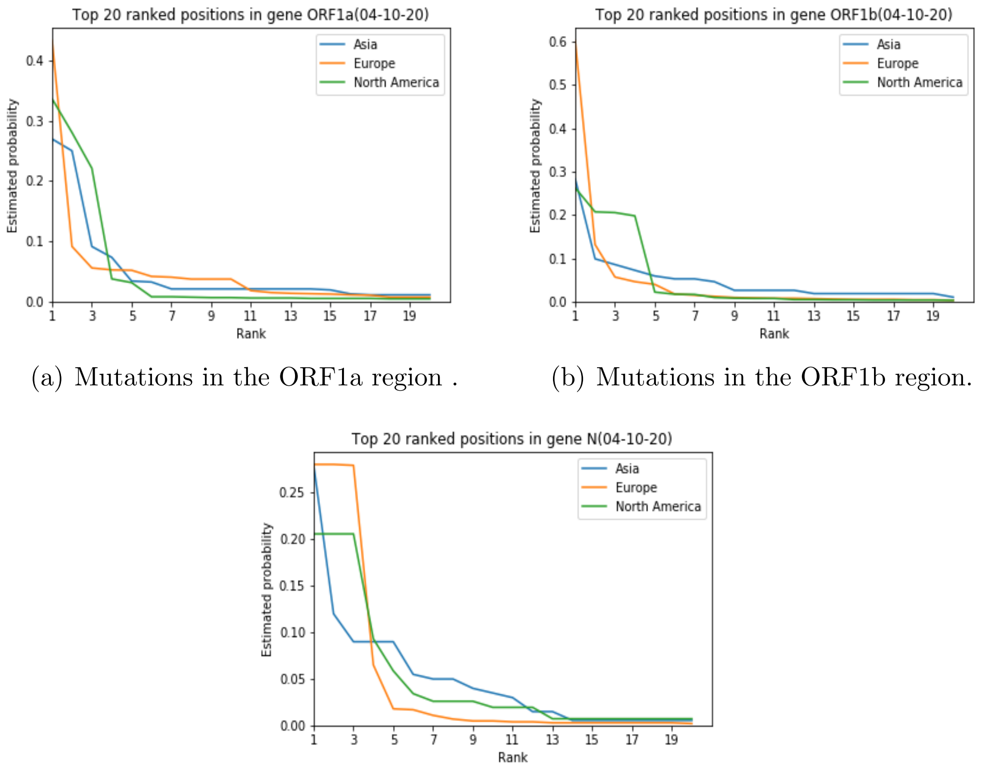
(a) Mutations in the ORF1a region.

(b) Mutations in the ORF1b region.

(c) Mutations in the N region.

Figure 3.3: Comparison of the estimated distributions of mutations in male and female test subjects tested by 04-10-2020. The distributions exhibit no difference except on two sites in the N region.

very different evolution patterns of the virus in the ORF1b genomic region at different regional sites, and more similar mutational patterns for different gender and age categories. Additional Table C.10 suggests significantly fewer stratifications in the mutations of different patient groups in the N region. Gender and age do not appear to play a major role, but marked differences are observed in patients from Asia, Europe and NA (the sites mutated in two regions but not in the third are shown in *italics*). Given the large differences in the mutational sites of patients across different continents in the N region it comes as no surprise that different primers for RT-PCR testing were selected for Asia, Europe and NA. The sites selected for forward and reverse primers by the CDC, i.e., the N1 and N2 region, do not contain a significant number of mutations, as may be easily seen from Additional Table C.11. Similar observations are true for the primers selected in China (Additional Table C.12).



(a) Mutations in the ORF1a region .

(b) Mutations in the ORF1b region.

(c) Mutations in the N region.

Figure 3.4: Differences in the estimated distributions of mutations for different geographic regions based on subjects tested by 04-10-2020. The distributions differ significantly.

Summarizing the Differences in the Distributions Using the Symmetric KL Divergence and the Jaccard Distance

The symmetric KL divergence between two discrete probability distributions p and q is defined as

$$D_s(p,q) = D(p||q) + D(q||p), \quad D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

For the mutation distributions pertaining to Europe-NA, Europe-Asia and Asia-NA, the KL divergences equal 0.672, 0.316 and 0.376 (ORF1a), 0.491, 0.435 and 0.646 (ORF1b), 0.293, 1.021 and 0.303 (N), respectively, for data collected by 04-14-2020. These results indicate that the largest differences in the distributions in the ORF1a region exist between Europe and NA, while the largest differences in the ORF1b region exist between Asia and NA. For

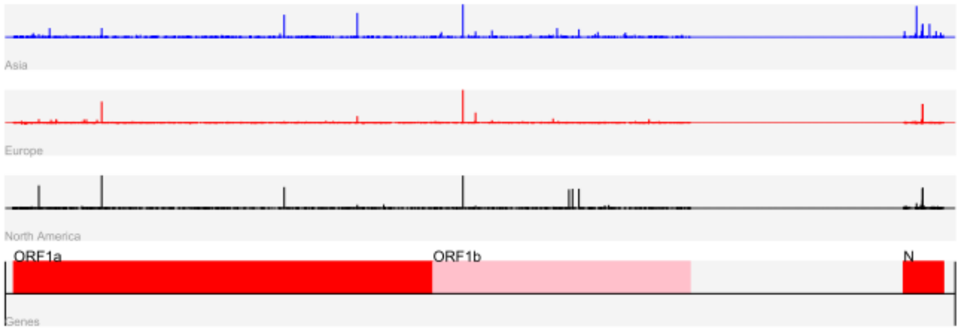


Figure 3.5: Positions of mutations in the SARS-Cov-2 genome for patients across three different continents, for data collected by 04-14-2020. The height of the bar is proportional to the estimated probability of mutation.

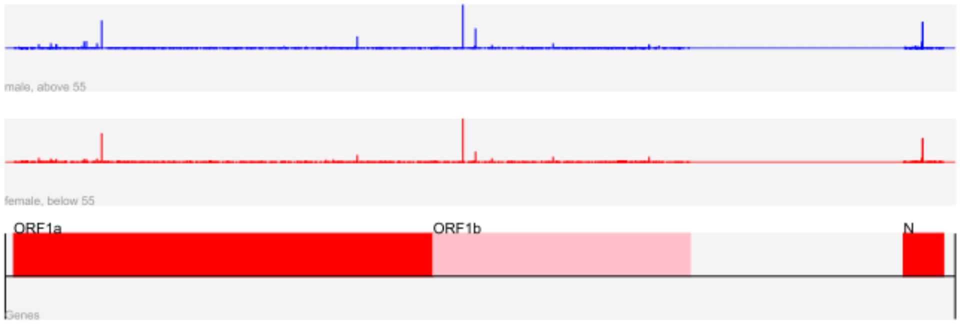


Figure 3.6: Positions of mutations in the SARS-Cov-2 genome for European females of age < 55 and males of age > 55 collected by 04-14-2020. The height of the bar is proportional to the estimated probability.

the N region, a significant difference between the distributions of mutations is observed between Europe and Asia, and at this point, no simple explanation for this finding is possible. Similarly, the corresponding KL divergences based on the samples collected by 04-10-2020 equal 0.788 (which is significantly larger than the one predicted based on data collected on 04-14-2020), 0.328 and 0.371 (ORF1a), 0.743 (which is significantly larger than the one predicted based on data collected on 04-14-2020), 0.615 and 0.0.755 (ORF1b), 0.315, 0.893 and 0.248 (N), respectively. The results for the KL divergences for the N regions suggest relatively small changes in the distribution of mutations in the N region, and larger changes in the ORF1a and ORF1b regions, which is expected.

Since the previously described distribution estimates do not convey the information about the locations of the highest mutated sites but only their frequency of mutations, we also list the Jaccard distances of the sets of mutations specific to each tested subpopulation. For two sets Σ_1 and Σ_2 over the same ground set Σ, the Jaccard distance $J(\Sigma_1, \Sigma_2)$ is defined as:

$$J(\Sigma_1, \Sigma_2) = 1 - \frac{\Sigma_1 \cap \Sigma_2}{\Sigma_1 \cup \Sigma_2}.$$

As may be seen from Table 3.10, the largest distances are observed in the E and ORF10 regions, in the first case when comparing patients from Asia and Europe and in the second case when comparing younger female and older males in Europe. The distances in the N region seem to be significantly smaller, especially between the two categories of patients from Europe. The results for the ORF10 region are rather surprising as they indicate the highest possible difference is observed between males and females on the same continent despite these differences being uniformly small for all other open reading frames. As already pointed out, the function of the ORF10 reading frame is currently unknown but given the marked mutational profiles in high-risk and low-risk profiles it is highly likely that this gene plays an important role in guiding disease symptoms. The exact same trends are observed when using alignments with respect to Patient 1 from the underlying geographic region, as listed in Additional Table C.13.

Table 3.10: The Jaccard distance between sets of mutations from different pairs of geographic regions, based on alignments with respect to Patient 1 from Wuhan. Values in *italics* are the smallest in the category, while values in **bold** are the largest.

	ORF1a	ORF1b	S	ORF3a	E	M	ORF6	ORF7a	ORF8	N	ORF10
Asia - Europe	**0.91**	**0.95**	0.91	0.92	**0.98**	**0.92**	**0.96**	*0.74*	0.91	0.86	*0.89*
Europe-NA	0.88	0.89	0.88	0.84	0.89	0.89	*0.86*	**0.91**	0.89	0.82	0.95
Asia - NA	**0.91**	**0.95**	**0.92**	**0.95**	0.91	0.89	**0.96**	0.89	**0.92**	**0.87**	*0.89*
Male >55 - Female < 55 (Europe)	*0.85*	*0.87*	*0.85*	*0.73*	*0.88*	*0.75*	0.87	*0.83*	*0.83*	*0.77*	**0.97**

CHAPTER 4

CONCLUSION

The problem of determining mutational support and distribution of a virus is crucial for accessing its virulence and for primer selection for real time RT-PCR kits, especially during early stages of a pandemic when insufficient information is available about the virus. An accurate estimate of possible mutations in the viral genome in the absence of a sufficiently large database is important for an early understanding of the adaptation mechanisms employed by the virus as well as the potential differences in its impact on diverse subpopulations. In this thesis, we presented a novel, state-of-the-art estimator for support estimation for the small-sample regime and benchmarked it against existing estimators. We also adapted the Good-Turing estimator for distribution estimation.

We used our estimators for a differential analysis on mutations in the SARS-Cov-2 genome among various population groups including male/female, older/younger and different geographic locations. We observed significant differences in the mutational support of ORF6 and ORF7a between older and younger patients as well as differences in ORF1b and ORF10 between males and females. We also noted that these differences persist with increase in number of samples available. Given that these ORFs play important biological roles in the spread and evolution of the virus, these differences can provide significant insights into why different population groups are impacted differently by the virus. Furthermore, we discovered differences in mutational support among all ORFs while comparing between different geographical locations. Our analysis showed that patients from Asia had comparatively higher mutational support than those from Europe and North America, which can potentially indicate that the virus was in circulation much earlier than expected. We validated our results by comparing the support estimate returned by our estimators on 04-10-2020 with ML estimates from 04-14-2020 as well as comparing the two estimators on a much larger sample set obtained on 10-20-2020.

We observed that even though the N region of the SARS-CoV-2 genome has a high number of mutations, only a few mutations lay in the primer regions for real time RT-PCR kits recommended by CDC for testing in the USA. This is important because frequent mutations in the primer regions can potentially lead to high rates of false negative results. Finally, we compared the distributions of mutations among various population groups and computed pairwise symmetric Kulback-Leibler divergences for normalized top-20 mutated positions as well as Jaccard distance for the sets of all mutations for each population. We emphasize that our estimators are general enough to be adapted for the genomes of any microorganism, making it extremely useful in the early stages for any future outbreak as well.

APPENDIX A

PROOFS

A.1 Proof of the theorem establishing the length of the optimization interval

To prove the result, we need to show that $\forall \lambda \geq 6.5L$, $\frac{\partial}{\partial \lambda} g(\mathbf{a}, \lambda) < 0$. The derivative of the first term in g equals

$$\frac{\partial}{\partial \lambda} \frac{1}{k} \bigg(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l! \bigg) = \frac{1}{k} \bigg(\sum_{l=0}^{L} (\frac{l}{\lambda} - 1) e^{-\lambda} a_l^2 \lambda^l l! \bigg).$$

Clearly, the right-hand side in the above expression is negative for all $\lambda > L$. The second term of the derivative equals

$$\begin{aligned} & \frac{\partial}{\partial \lambda} \bigg(e^{-\lambda} \sum_{l=0}^{L} a_l \lambda^l \bigg)^2 \\ & = 2 \bigg(e^{-\lambda} \sum_{l=0}^{L} a_l \lambda^l \bigg) \bigg(- e^{-\lambda} \sum_{l=0}^{L} a_l \lambda^l + e^{-\lambda} \sum_{l=0}^{L} \frac{l}{\lambda} a_l \lambda^l \bigg) \\ & = 2 e^{-2\lambda} \bigg(\sum_{l=0}^{L} a_l \lambda^l \bigg) \bigg(\sum_{l=0}^{L} (\frac{l}{\lambda} - 1) a_l \lambda^l \bigg). \end{aligned}$$

To analyze the two terms of the derivative, we introduce the vectors $\mathbf{y}, \mathbf{z}, \mathbf{1}$ and the diagonal matrix $\mathbf{D}$ according to

$$\begin{aligned} & \mathbf{y} = (a_0 \lambda^0, a_1 \lambda^1, ..., a_L \lambda^L)^T, \\ & \mathbf{z} = ((\frac{0}{\lambda} - 1), (\frac{1}{\lambda} - 1), ..., (\frac{L}{\lambda} - 1))^T, \\ & \mathbf{1} = (1, 1, ..., 1)^T, \\ & D_{ii} = (-1 + \frac{i-1}{\lambda}) \frac{(i-1)!}{\lambda^{(i-1)}}. \end{aligned}$$

Consequently, we have

$$\frac{\partial}{\partial\lambda}\frac{1}{k}\left(\sum_{l=0}^{L} e^{-\lambda}a_l^2\lambda^l l!\right) = \frac{e^{-\lambda}}{k}\mathbf{y}^T\mathbf{D}\mathbf{y},$$

$$\frac{\partial}{\partial\lambda}\left(e^{-\lambda}\sum_{l=0}^{L} a_l\lambda^l\right)^2 = 2e^{-2\lambda}\mathbf{y}^T\mathbf{1}\mathbf{z}^T\mathbf{y} = e^{-2\lambda}\mathbf{y}^T(\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)\mathbf{y}.$$

Therefore,

$$\frac{\partial}{\partial\lambda}g(\mathbf{a},\lambda) = e^{-2\lambda}\mathbf{y}^T\left(\frac{e^{\lambda}}{k}\mathbf{D} + (\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)\right)\mathbf{y}.$$

To show that $\frac{\partial}{\partial\lambda}g(\mathbf{a},\lambda) < 0$ for all polynomials of degree L whenever $\lambda > CL$, we show that the matrix $\left(\frac{e^{\lambda}}{k}\mathbf{D} + (\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)\right)$ is negative-definite whenever $\lambda > CL$, for some constant $C > 0$. It suffices to show that the sum of the maximum eigenvalues of $\frac{e^{\lambda}}{k}\mathbf{D}$ and $(\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)$ is negative, since $\frac{e^{\lambda}}{k}\mathbf{D}$ is a diagonal matrix. Thus, we turn our attention to determining the maximum eigenvalues of these two matrices. For $\frac{e^{\lambda}}{k}\mathbf{D}$, the maximum eigenvalue satisfies

$$\frac{e^{\lambda}}{k}\max_{i\in\{0,1,\ldots,L\}}\left(-1+\frac{i}{\lambda}\right)\frac{i!}{\lambda^i} \le -\frac{e^{\lambda}}{2k}\min_{i\in\{0,1,\ldots,L\}}\frac{i!}{\lambda^i},$$

since for $\lambda > 2L$, one has $(-1+\frac{i}{\lambda}) \le -\frac{1}{2}$. When $\lambda > L$, it is clear that $\frac{i!}{\lambda^i}$ is decreasing in i, for $i \in \{0,1,\ldots,L\}$, so that

$$\min_{i\in\{0,1,\ldots,L\}}\frac{i!}{\lambda^i} = \frac{L!}{\lambda^L} \ge \left(\frac{L}{e\lambda}\right)^L.$$

The last inequality is a consequence of Stirling's formula, which asserts that $n! \ge (\frac{n}{e})^n$. Combining the above expressions, we obtain

$$\frac{e^{\lambda}}{k}\max_{i\in\{0,1,\ldots,L\}}\left(-1+\frac{i}{\lambda}\right)\frac{i!}{\lambda^i} \le -\frac{e^{\lambda}}{2k}\left(\frac{L}{e\lambda}\right)^L.$$

Next, we derive an upper bound on maximum eigenvalue of the second matrix. The i,j entry of the matrix $(\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)$ equals $\frac{i+j-2}{\lambda} - 2$, and all these values are negative when $\lambda > L$. Moreover, it is clear that the matrix of interest has rank equal to 2. Therefore, the matrix has exactly two nonzero eigenvalues.

Let $\mathbf{A} = -(\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)$. All entries of $\mathbf{A}$ are positive whenever $\lambda > L$. By Gershgorin's theorem, we can upper bound the maximum eigenvalues of the matrix $\mathbf{A}$ by its maximum row sum. It is obvious that the maximum row sum equals

$$2(L+1) - \frac{L(L+1)}{2\lambda}.$$

Moreover, the trace of $\mathbf{A}$ equals

$$2(L+1) - \frac{L(L+1)}{\lambda}.$$

This implies that the minimum eigenvalue of $\mathbf{A}$ is lower bounded by $-\frac{L(L+1)}{2\lambda}$, which directly implies that the maximum eigenvalue of $(\mathbf{1}\mathbf{z}^T + \mathbf{z}\mathbf{1}^T)$ is upper bounded by $\frac{L(L+1)}{2\lambda}$.

Summing up the two previously derived upper bounds gives

$$h(\lambda) \triangleq -\frac{e^\lambda}{2k}\left(\frac{L}{e\lambda}\right)^L + \frac{L(L+1)}{2\lambda},$$

whenever $\lambda > 2L$. Note that $h(\lambda) < 0$ is equivalent to

$$\begin{aligned} &\frac{L(L+1)}{2\lambda} < \frac{e^\lambda}{2k}\left(\frac{L}{e\lambda}\right)^L \\ &\Leftrightarrow \log(L) + \log(L+1) + \log(k) - L\log(L) + L < \lambda + \log(\lambda) - L\log(\lambda). \end{aligned} \tag{A.1}$$

The function $\lambda + \log(\lambda) - L\log(\lambda)$ is nondecreasing in λ whenever $\lambda > L$ since

$$\frac{d}{d\lambda}(\lambda + \log(\lambda) - L\log(\lambda)) = 1 - \frac{L-1}{\lambda}.$$

By the definition of $L = \lfloor c_0 \log(k) \rfloor$, we also have $\log(k) \leq \frac{L+1}{c_0}$. Using $\log(x+1) \leq x$, which holds $\forall x \geq 1$. Hence $\forall \lambda > CL$ where $C > 2$, the sufficient condition for (A.1) to hold is

$$\log(L) + L + \frac{L+1}{c_0} - L\log(L) + L < CL + \log(CL) - L\log(CL).$$

Rearranging terms leads to

$$\left(C - \log(C) - 2 - \frac{1}{c_0}\right) L + \log(C) > \frac{1}{c_0}.$$

Sufficient conditions that ensure that the above inequality holds are $\log(C) \geq \frac{1}{c_0}$ and $(C - \log(C) - 2 - \frac{1}{c_0}) > 0$. The first condition implies $C \geq e^{\frac{1}{c_0}} = 6.0021$, while the second condition holds with $C = 6.5$, for which the first condition is also satisfied. This completes the proof.

A.2 Proof of the convergence rate of the discretized SIP

The proof consists of two parts. In the first part, we establish the conditions for convergence, while in the second part, we determine the convergence rate. For simplicity, we present the proofs for the case without Poisson repeats. We then outline how the analysis can be modified to account for the repeats.

A.2.1 Proof of convergence

We start by introducing the relevant terminology. Let $\Pi \subset \mathbb{R}^{L+1}$ be a closed set of parameters, and let f be a continuous functional on Π. Assume that $B \subset \mathbb{R}$ is compact and that $g : \Pi \mapsto \mathcal{C}(B)$ is a continuous mapping from Π into $\mathcal{C}(B)$, where $\mathcal{C}(B)$ is the space of continuous functions over B equipped with the supremum norm $||\cdot||_\infty$. For each $D \subset B$ let

$$M(D) = \{\mathbf{c} \in \Pi |\, g(\mathbf{c}, x) \leq 0, x \in D\}$$

denote the set of feasible points of the optimization problem

$$\min\ f(\mathbf{c}) \text{ over } \mathbf{c} \in M(D).$$

Assuming that $M(D) \neq \emptyset$, let

$$\mu(D) = \inf\{f(\mathbf{c}) | \mathbf{c} \in M(D)\},$$

and define the level set

$$\text{Level}(\mathbf{c}_0, D) = \{\mathbf{c} \in \Pi |\, f(\mathbf{c}) \leq f(\mathbf{c}_0)\} \cap M(D).$$

We also make the following two assumptions:

- Assumption 1: Fine grid Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. There exists a sequence $\{B_i\}$ of compact subsets of B with $B_i \subset B_{i+1}$, $i \in \mathbb{N}_0$, for which $\lim_{i\to\infty} h(B_i, B) = 0$, such that

$$h(B_i, B) = \sup_{x\in B} \inf_{y\in B_i} ||x - y||.$$

- Assumption 2: Bounded level set $M(B)$ is nonempty, and there exists a $\mathbf{c}_0 \in M(B)$ such that the level set $\text{Level}(\mathbf{c}_0, B_0)$ is bounded and hence compact in $\mathbb{R}^{L+1}$.

Convergence of the discretized method, Theorem 2.1 from [33]:

Under Assumptions 1 and 2, the solution of the discretized problem converges to the optimal solution. More formally, we have

$$\begin{aligned} &\mu(B_i) \leq \mu(B_{i+1}) \leq \mu(B), \forall t \in \mathbb{N}_0 \\ &\lim_{i\to\infty} \mu(B_i) = \mu(B). \end{aligned}$$

If $\mathbf{c}^*$ is the unique optimal solution of the original problem, and $\mathbf{c}_i^*$ is the optimal solution of the discretized relaxation with grid B_i, then

$$\lim_{i\to\infty} ||\mathbf{c}^* - \mathbf{c}_i^*||_2 = 0.$$

It is straightforward to see that our chosen grid is arbitrary fine. Hence, we only need to prove that there exists a $\mathbf{c}_0$ such that the level set $\text{Level}(\mathbf{c}_0, D)$ is bounded.

Let $\mathbf{c} = (\mathbf{a}; t)$ and note that in our setting, $f(\mathbf{c}) = t$. Rewrite $g(\mathbf{c}, \lambda)$ in matrix form as

$$g(\mathbf{c}, \lambda) = \mathbf{a}^T \mathbf{M}(\lambda)\mathbf{a} + \mathbf{a}^T \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \mathbf{a} - t,$$

where

$$\mathbf{\Lambda} \triangleq e^{-\lambda}(\lambda^0, \lambda^1, ..., \lambda^L)^T.$$

Note that only $a_1, ...a_L$ are allowed to vary since we fixed $a_0 = -1$. Obviously, $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is positive semi-definite and the previously introduced $\mathbf{M}(\lambda)$ is positive definite for all $\lambda > 0$. Since the constraints on g are positive definite with respect to $a_1, ...a_L$, g is coercive in $a_1, ...a_L$. Furthermore, for any given t, the set of feasible coefficients $a_1, ...a_L$ is bounded. Therefore, given a t_0, the level set Level$(\mathbf{c}_0, B_0)$ is bounded. This ensures that Assumption 1 holds for our optimization problem.

Next, we prove the uniqueness of the optimal solution $\mathbf{c}^\star$. Note that proving this result is equivalent to proving the uniqueness of $\mathbf{a}^\star$. Hence, we once again refer to the original minmax formulation of our problem,

$$\inf_{\mathbf{a}:a_0=-1} \sup_{\lambda\in[\frac{n}{k}, 6.5L]} \mathbf{a}^T(\mathbf{M}(\lambda) + \mathbf{\Lambda}\mathbf{\Lambda}^T)\mathbf{a} \triangleq \inf_{\mathbf{a}:a_0=-1} \sup_{\lambda\in[\frac{n}{k}, 6.5L]} h_\lambda(\mathbf{a}). \tag{A.2}$$

Clearly, $\forall \lambda \in [\frac{n}{k}, 6.5L]$, the function $h_\lambda(\mathbf{a})$ is strictly convex since $(\mathbf{M}(\lambda) + \mathbf{\Lambda}\mathbf{\Lambda}^T) \succ 0$, $\forall \lambda \in [\frac{n}{k}, 6.5L]$. Taking the supremum over λ preserves strict convexity since $\forall \theta \in (0, 1)$, one has

$$\begin{aligned}
&\sup_{\lambda\in[\frac{n}{k}, 6.5L]} h_\lambda(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \\
&< \sup_{\lambda\in[\frac{n}{k}, 6.5L]} \theta h_\lambda(\mathbf{x}) + (1-\theta)h_\lambda(\mathbf{y}) \\
&\leq \sup_{\lambda\in[\frac{n}{k}, 6.5L]} \theta h_\lambda(\mathbf{x}) + \sup_{\lambda'\in[\frac{n}{k}, 6.5L]} (1-\theta)h_{\lambda'}(\mathbf{y}).
\end{aligned}$$

Hence $\sup_{\lambda\in[\frac{n}{k}, 6.5L]} h_\lambda(\mathbf{a})$ is strictly convex, which consequently implies the uniqueness of $\mathbf{a}^\star$ and hence $\mathbf{c}^\star$.

For the case of samples passed through a Poisson channel, it is not hard to see that the constraints are again strictly convex in $\mathbf{a}$, where one need only replace $\mathbf{M}(\lambda), \mathbf{\Lambda}$ by

$$\begin{aligned}
&\frac{1}{k} e^{-\lambda(1-e^{-\eta})}\, Diag(0!\eta^0 M_{N^*}^{(0)}(0), 1!\eta^1 M_{N^*}^{(1)}(0), ..., L!\eta^L M_{N^*}^{(L)}(0)) \\
&e^{-\lambda(1-e^{-\eta})}(\eta^0 M_{N^*}^{(0)}(0), \eta^1 M_{N^*}^{(1)}(0), ..., \eta^L M_{N^*}^{(L)}(0))^T
\end{aligned}$$

respectively. Thus, a similar analysis is possible and the details are omitted. The proof above along with the previous observation proves the convergence result.

A.2.2 Proof for the convergence rate

In what follows, and for reasons of simplicity, we omit the constraint $a_0 = -1$ in the SIP formulation. The described proof only requires small modifications to accommodate $a_0 = -1$.

Recall that we used B_d to denote the grid with grid spacing d. In order to use the results in [34], we require the convergence assumptions below.

- Assumption 3: Let $\bar{\mathbf{c}}$ be a local minimizer of an SIP. There exists a local solution $\mathbf{c}_d$ of the discretized SIP with grid B_d such that

$$||\mathbf{c}_d - \bar{\mathbf{c}}|| \to 0.$$

  This assumption is satisfied for the SIP of interest as shown in the first part of the proof.

- Assumption 4: The following hold true:
  - There is a neighborhood $\bar{U}$ of $\bar{\mathbf{c}}$ such that the function $\frac{\partial^2}{\partial \lambda^2} g(\mathbf{c}, \lambda)$ is continuous on $\bar{U} \times B$.
  - The set B is compact, nonempty and explicitly given as the solution set of a set of inequalities, $B = \{\lambda \in \mathbb{R} | v_i(\lambda) \leq 0, i \in I\}$, where I is a finite index set and $v_i \in C^2(B)$.
  - For any $\bar{\lambda} \in B$, the vectors $\frac{\partial}{\partial \lambda} v_i(\bar{\lambda}), i \in \{i \in I | v_i(\bar{\lambda}) = 0\}$ are linearly independent.

  Recall that our objective is of the form

$$g(\mathbf{c}, \lambda) = \mathbf{a}^T \mathbf{M}(\lambda) \mathbf{a} + \mathbf{a}^T \boldsymbol{\Lambda} \boldsymbol{\Lambda}^T \mathbf{a} - t,$$

where

$$\mathbf{\Lambda} \triangleq e^{-\lambda}(\lambda^0, \lambda^1, ..., \lambda^L)^T, \ \mathbf{c} = (\mathbf{a}; t),$$
$$\mathbf{M}(\lambda) \triangleq \frac{e^{-\lambda}}{k} Diag(\lambda^0 0!, \lambda^1 1!, ..., \lambda^L L!).$$

It is straightforward to see that the first condition in Assumption 4 holds. For the second condition, recall that $B = [\frac{n}{k}, 6.5L]$. Hence, the second condition can be satisfied by choosing $I = \{1\}$, $v_1(\lambda) = (\lambda - \frac{n}{k})(\lambda - 6.5L)$. Since we only have one variable v_1, it is also easy to see that the third condition is met.

- Assumption 5: The set B satisfies Assumption 4 and all the sets B_d contain the boundary points $\frac{n}{k}, 6.5L$.

  This assumption also clearly holds for the grid of choice. Note that it is crucial to include the boundary points for the proof in [34] to be applicable.

- Assumption 6: $\nabla_{\mathbf{c}} g(\mathbf{c}, \lambda)$ is continuous on $\bar{U} \times B$, where $\bar{U}$ is a neighborhood of $\bar{\mathbf{c}}$. Moreover, there exists a vector ξ such that

  $$\nabla_{\mathbf{c}} g(\bar{\mathbf{c}}, \lambda)^T \xi \leq -1, \ \forall \lambda \in B.$$

  Note that $\nabla_{\mathbf{c}} g(\mathbf{c}, \lambda) = [\nabla_{\mathbf{a}} g(\mathbf{c}, \lambda); \nabla_t g(\mathbf{c}, \lambda)]$ and

  $$\nabla_{\mathbf{a}} g(\mathbf{c}, \lambda) = 2(\mathbf{M}(\lambda) + \mathbf{\Lambda}\mathbf{\Lambda}^T)\mathbf{a}.$$

  Also note that $\forall \lambda \in B$, $\mathbf{M}(\lambda) + \mathbf{\Lambda}\mathbf{\Lambda}^T$ is positive definite. Hence choosing ξ to be colinear with and of the same direction as $[-\mathbf{a}^T \ 1]^T$, as well as of sufficiently large norm, will allow us to satisfy the inequality

  $$\nabla_{\mathbf{c}} g(\bar{\mathbf{c}}, \lambda)^T \xi \leq -1, \ \forall \lambda \in B.$$

  Hence, Assumption 6 holds as well. The next results follow from the above assumptions and observations, and the results in [34].

(Corollary 1 in [34]) Let t_d be the optimal objective value of the discretized SIP used for support estimation with the grid B_d, and let $t^\star$ be the optimal objective value for the original SIP. Since Assumptions 3-6 hold, then for some $c_3 > 0$ and d sufficiently small, we have

$$0 \leq t^\star - t_d \leq c_3 d^2.$$

Consequently, $t_d \to t^\star$ with a convergence rate of $O(d^2)$.

(Theorem 2 in [34]) Assume that all assumptions in Corollary 1 above hold. If there exists a constant $c_4 > 0$ such that

$$t - \bar{t} \geq c_4 ||\mathbf{c} - \bar{\mathbf{c}}||, \ \forall \mathbf{c} \in M(B) \cap \bar{U},$$

then for sufficiently small d and $\sigma > 0$ we have

$$||\mathbf{c}_d - \bar{\mathbf{c}}|| \leq \sigma d^2.$$

This result implies that if $\bar{\mathbf{c}}$ is also a strict minimum of order one, then the solution of the discretized SIP converges to that of the original SIP with rate $O(d^2)$. For the Poisson repeat channel, the constraints are also strictly convex in $\mathbf{a}$. Therefore, a similar analysis is possible and the details are omitted once again. Combining these results completes the proof.

A.3 Additional theoretical results

The result described in the main text follows from Theorem 6.2 in [24].

(Theorem 6.2 from [24]) Let $W(x) = \exp(-Q(x))$ be a weight function, where $Q : \mathbb{R} \mapsto [0, \infty)$ is even, convex, diverging for $x \to \infty$, and such that

$$0 = Q(0) < Q(x), \forall x \neq 0.$$

Then, for any polynomial $P(x)$ of degree $\leq L$, not identical to zero, one has

$$\begin{aligned} \sup_{x \in \mathbb{R}} |P(x)W(x)| &= \sup_{x \in [-M_L, M_L]} |P(x)W(x)|, \\ \sup_{x \in \mathbb{R} \setminus [-M_L, M_L]} |P(x)W(x)| &< \sup_{x \in [-M_L, M_L]} |P(x)W(x)|. \end{aligned}$$

Here, M_L stands for the *Mhaskar-Rakhmanov-Saff* (MSF) number, which is the smallest positive root of the integral equation

$$L = \frac{2}{\pi}\int_0^1 \frac{M_L t Q'(M_L t)}{\sqrt{1-t^2}}dt. \tag{A.3}$$

In our setting, the weight equals $\exp(-x)$. Solving (A.3) gives us an MSF number equal to $M_L = \frac{\pi}{2}L$. Thus, we can restrict our optimization interval to $[\frac{n}{k}, \frac{\pi}{2}L + \frac{n}{k}]$. If there is no regularization term, the optimal interval reduces to $[\frac{n}{k}, \frac{\pi}{2}L + \frac{n}{k}]$.

A.4 Construction of the RWC-S estimator

We introduce the optimization problem needed for minimizing the risk $E\left(\frac{S-\hat{S}}{S}\right)^2$. Poissonization arguments once again establish that

$$\mathbb{E}\left(\frac{S-\hat{S}}{S}\right)^2 = \frac{1}{S^2}\Big\{\sum_{i\in\mathcal{L}}\Big(\sum_{l=0}^{L} e^{-\lambda_i} a_l^2 \lambda_i^l l!\Big) + \sum_{i\neq j\in\mathcal{L}}\Big(e^{-\lambda_i}\sum_{l=0}^{L} a_l\lambda_i^l\Big)\Big(e^{-\lambda_j}\sum_{l=0}^{L} a_l\lambda_j^l\Big)\Big\}.$$

Taking the supremum over D_k, one can further upper bound the risk as

$$\begin{aligned}
&\leq \sup_{\lambda_\ell\in[\frac{n}{k},n],\ \ell\in\mathcal{L}} \frac{1}{S^2}\Big\{\sum_{i\in\mathcal{L}}\Big(\sum_{l=0}^{L} e^{-\lambda_i} a_l^2 \lambda_i^l l!\Big) + \sum_{i\neq j\in\mathcal{L}}\Big(e^{-\lambda_i}\sum_{l=0}^{L} a_l\lambda_i^l\Big)\Big(e^{-\lambda_j}\sum_{l=0}^{L} a_l\lambda_j^l\Big)\Big\}\\
&\leq \sup_{\lambda\in[\frac{n}{k},n]}\left\{\frac{1}{S}\Big(\sum_{l=0}^{L} e^{-\lambda}a_l^2\lambda^l l!\Big) + \Big(e^{-\lambda}\sum_{l=0}^{L} a_l\lambda^l\Big)^2\right\}\\
&\leq \sup_{\lambda\in[\frac{n}{k},n]}\left\{\frac{1}{\hat{S}_c}\Big(\sum_{l=0}^{L} e^{-\lambda}a_l^2\lambda^l l!\Big) + \Big(e^{-\lambda}\sum_{l=0}^{L} a_l\lambda^l\Big)^2\right\},
\end{aligned} \tag{A.4}$$

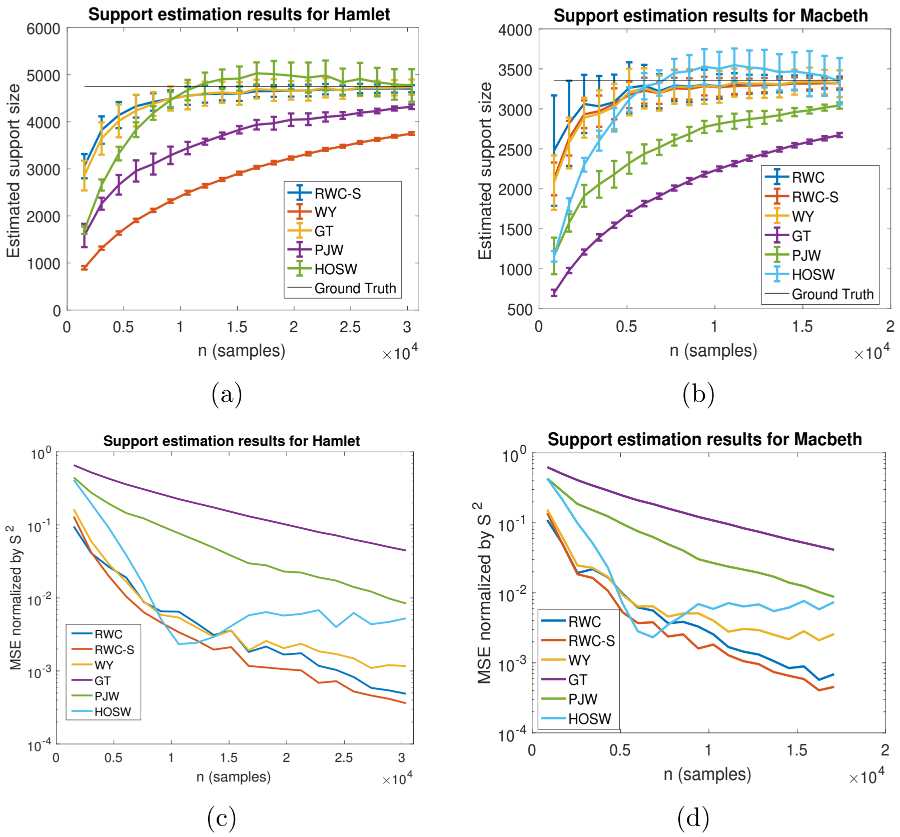
where the last inequality is due to the fact that $\hat{S}_c \leq S$. Note that the only difference between (A.4) and the corresponding optimization problem described in the main text is in terms of changing the normalization from $1/k$ to $1/\hat{S}_c$ in the first term. The expression (A.4) is optimized by the solution of the following problem:

$$\begin{aligned} &\min_{t,\mathbf{a}\in Poly(L)} t \quad \text{s.t.} \\ &\left\{ \frac{1}{\hat{S}_c}\left(\sum_{l=0}^{L} e^{-\lambda} a_l^2 \lambda^l l! \right) + \left(e^{-\lambda} \sum_{l=0}^{L} a_l \lambda^l \right)^2 \right\} \leq t,\ \forall \lambda \in \mathrm{Grid}([\frac{n}{k}, 6.5L], s). \end{aligned} \tag{A.5}$$

APPENDIX B

ADDITIONAL FIGURES

This appendix provides additional figures and results.



(a)

(b)

(c)

(d)

Figure B.1: Comparison of performance of various estimators on non-i.i.d. data with ground truth. The results are obtained over 100 independent trials. (a) and (b) show the mean and standard deviation of the estimators, while (c) and (d) show the MSE normalized by S^2.

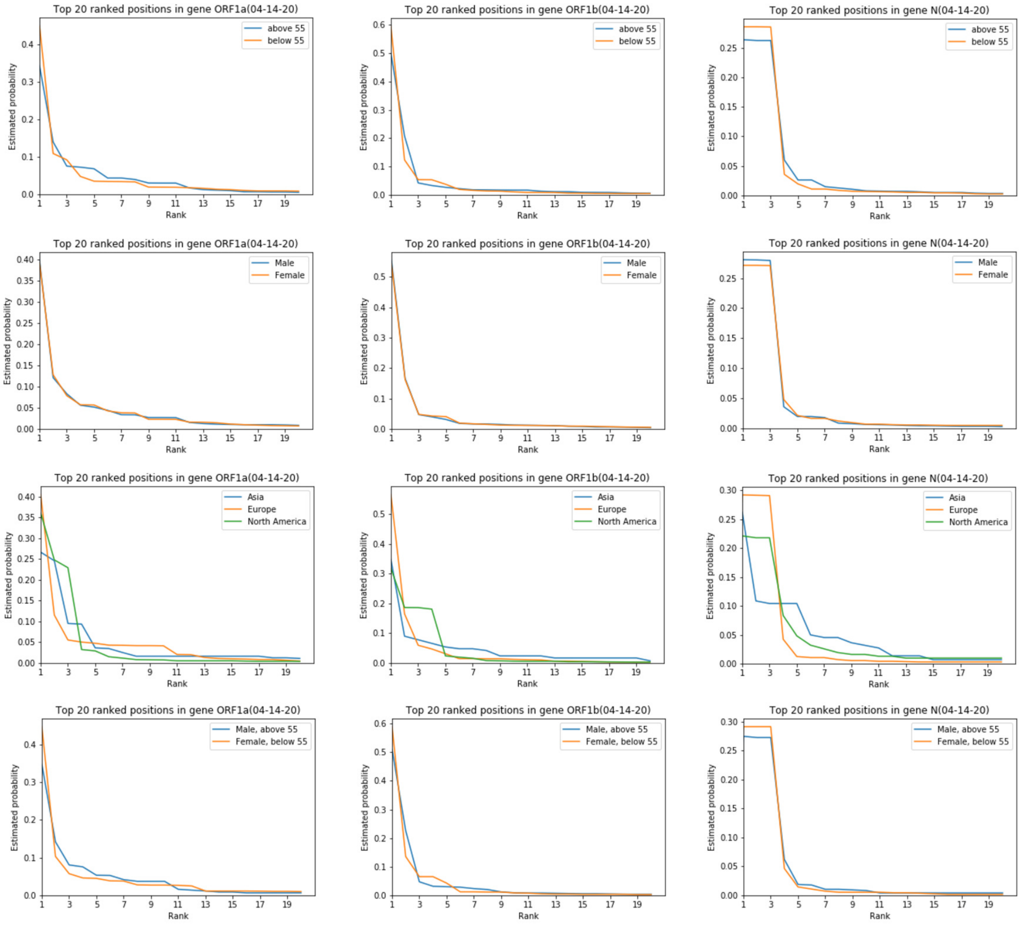


Figure B.2: Comparison of mutations in various groups of patients based on the data collected by 04-14-2020. All the alignments were performed with respect to Patient 1 Wuhan-Hu-1.

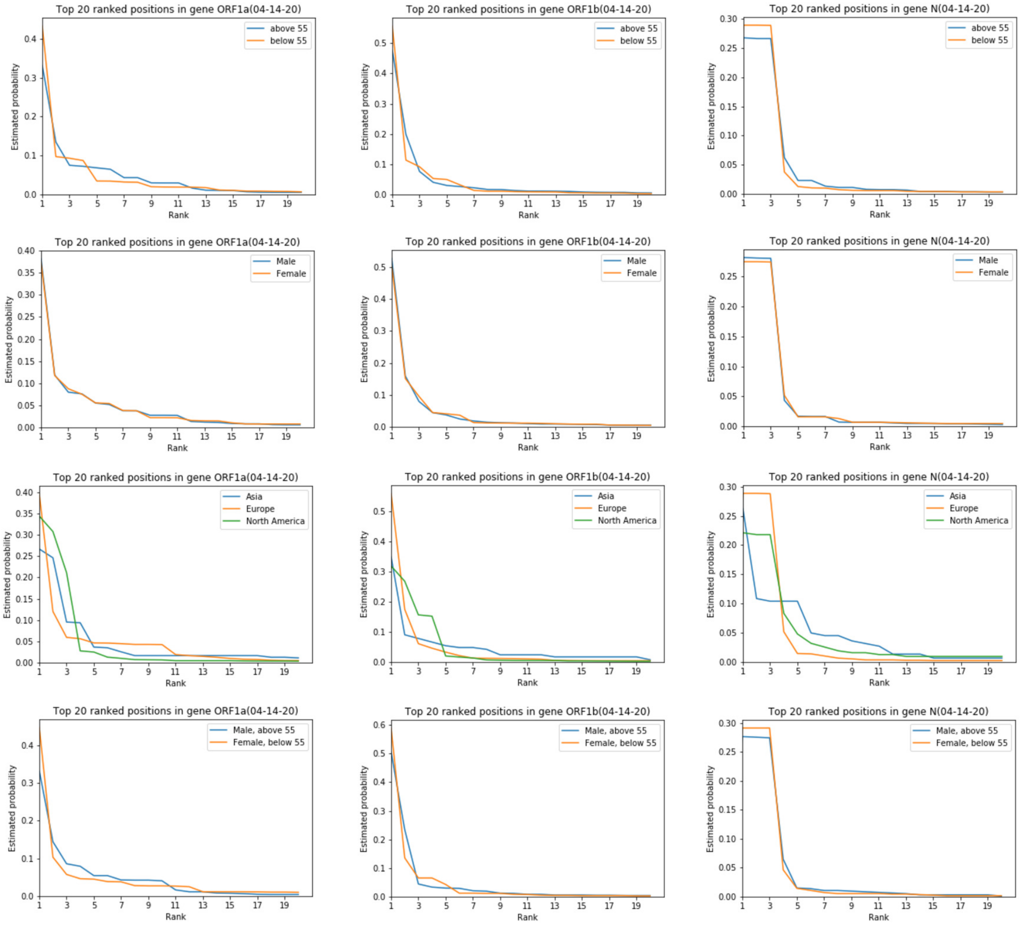


Figure B.3: Comparison of mutations in various groups of patients based on the data collected by 04-14-2020. All the alignments were performed with respect to the first sequenced patient 1 in the corresponding region.

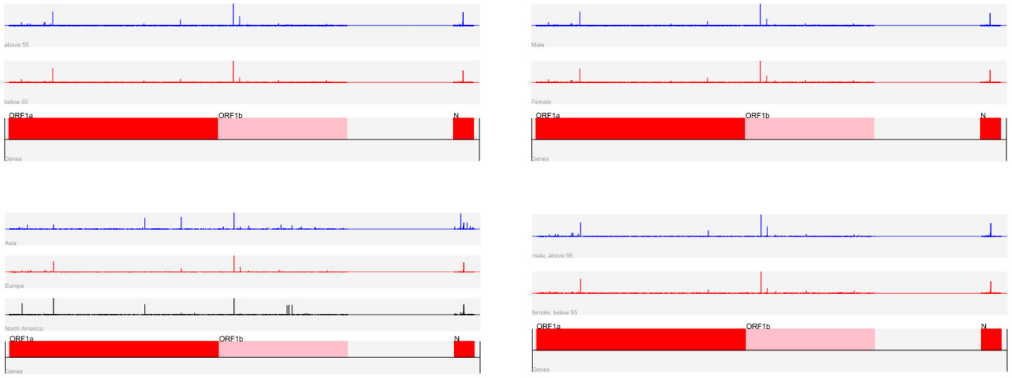


Figure B.4: Positions of mutations in the SARS-Cov-2 genome with high probability of mutations in patients from different categories based on data collected by 04-14-2020. The height of the bar is proportional to the probability of the mutation. All the alignments were performed with respect to Patient 1 Wuhan-Hu-1.

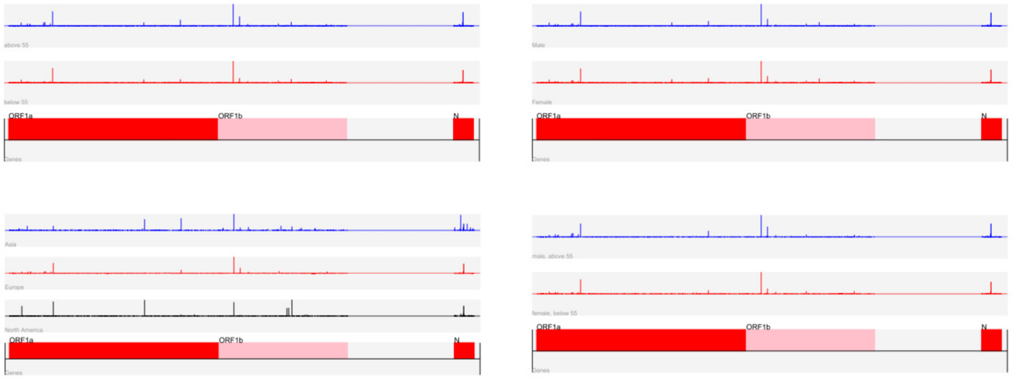


Figure B.5: Positions of mutations in the SARS-Cov-2 genome with high probability of mutations in patients from different categories based on data collected by 04-14-2020. The height of the bar is proportional to the probability of the mutation. All the alignments were performed with respect to the first sequenced patient 1 in the corresponding region.

APPENDIX C

ADDITIONAL TABLES

Additional tables have been provided in a separate Excel file - Supplementary_tables.xlsx.

REFERENCES

[1] R. Sanjuán, M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw, "Viral mutation rates," *Journal of Virology*, vol. 84, no. 19, pp. 9733–9748, 2010.

[2] J. W. Drake and J. J. Holland, "Mutation rates among RNA viruses," *Proceedings of the National Academy of Sciences*, vol. 96, no. 24, pp. 13 910–13 913, 1999.

[3] R. Sanjuán and P. Domingo-Calap, "Mechanisms of viral mutation," *Cellular and Molecular Life Sciences*, vol. 73, no. 23, pp. 4433–4448, 2016.

[4] S. Duffy, L. A. Shackelton, and E. C. Holmes, "Rates of evolutionary change in viruses: Patterns and determinants," *Nature Reviews Genetics*, vol. 9, no. 4, pp. 267–276, 2008.

[5] T. Hoenen, D. Safronetz, A. Groseth, K. Wollenberg, O. Koita, B. Diarra, I. Fall, F. Haidara, F. Diallo, M. Sanogo et al., "Mutation rate and genotype variation of Ebola virus from Mali case sequences," *Science*, vol. 348, no. 6230, pp. 117–119, 2015.

[6] R. M. Ribeiro, H. Li, S. Wang, M. B. Stoddard, G. H. Learn, B. T. Korber, T. Bhattacharya, J. Guedj, E. H. Parrish, B. H. Hahn et al., "Quantifying the diversification of hepatitis c virus (HCV) during primary infection: estimates of the in vivo mutation rate," *PLoS Pathogens*, vol. 8, no. 8, 2012.

[7] J. J. Bull, R. Sanjuan, and C. O. Wilke, "Theory of lethal mutagenesis for viruses," *Journal of Virology*, vol. 81, no. 6, pp. 2930–2939, 2007.

[8] R. Sanjuán, A. Moya, and S. F. Elena, "The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus," *Proceedings of the National Academy of Sciences*, vol. 101, no. 22, pp. 8396–8401, 2004.

[9] A. Acevedo, L. Brodsky, and R. Andino, "Mutational and fitness landscapes of an RNA virus revealed through population sequencing," *Nature*, vol. 505, no. 7485, pp. 686–690, 2014.

[10] C. L. Burch and L. Chao, “Evolvability of an RNA virus is determined by its mutational neighbourhood,” *Nature*, vol. 406, no. 6796, pp. 625–628, 2000.

[11] K. M. Peck and A. S. Lauring, “Complexities of viral mutation rates,” *Journal of Virology*, vol. 92, no. 14, pp. e01 031–17, 2018.

[12] Centre for Health Security, Johns Hopkins University, “SARS-CoV-2 genetics,” https://www.centerforhealthsecurity.org/resources/COVID-19/COVID-19-fact-sheets/200128-nCoV-whitepaper.pdf, 2020.

[13] W. A. Gale and G. Sampson, “Good-Turing frequency estimation without tears,” *Journal of Quantitative Linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[14] A. Orlitsky and A. T. Suresh, “Competitive distribution estimation: Why is Good-Turing good,” in *Advances in Neural Information Processing Systems*, 2015, pp. 2143–2151.

[15] Johns Hopkins University, “Covid-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU),” https://coronavirus.jhu.edu/map.html, 2020.

[16] L. Mousavizadeh and S. Ghasemi, “Genotype and phenotype of COVID-19: Their roles in pathogenesis,” *Journal of Microbiology, Immunology and Infection*, 2020.

[17] GeneTex, “GeneTex review of the function of SARS-Cov-2 ORFs,” https://www.genetex.com/MarketingMaterial/Index/SARS-CoV-2_Genome_and_Proteome, 2020.

[18] Center for Disease Control and Diagnostics, “2019 novel coronavirus (2019-nCoV) real-time RT-PCR diagnostic panel, catalog number 2019-nCoVEUA-01 with 1000 reactions,” https://www.fda.gov/media/134922/download, 2020.

[19] Y. Shu and J. McCauley, “GISAID: Global initiative on sharing all influenza data–from vision to reality,” *Eurosurveillance*, vol. 22, no. 13, 2017.

[20] R. C. Edgar, “MUSCLE: A multiple sequence alignment method with reduced time and space complexity,” *BMC Bioinformatics*, vol. 5, no. 1, p. 113, 2004.

[21] J. Kopel, A. Perisetti, A. Roghani, M. Aziz, M. Gajendran, and H. Goyal, “Racial and gender-based differences in COVID-19,” *Frontiers in Public Health*, vol. 8, p. 418, 2020.

[22] R. C. Rabin, “Why the coronavirus seems to hit men harder than women,” *New York Times*, 2020. [Online]. Available: https://www.nytimes.com/2020/02/20/health/coronavirus-men-women.html

[23] Y. Wu, P. Yang et al., “Chebyshev polynomials, moment matching, and optimal estimation of the unseen,” *The Annals of Statistics*, vol. 47, no. 2, pp. 857–883, 2019.

[24] D. S. Lubinsky, “A survey of weighted polynomial approximation with exponential weights,” *Surveys in Approximation Theory*, vol. 3, pp. 1–105, 2007.

[25] L. Paninski, “Estimation of entropy and mutual information,” *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[26] P. Valiant and G. Valiant, “Estimating the unseen: Improved estimators for entropy and other properties,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2157–2165.

[27] A. Orlitsky, N. P. Santhanam, and J. Zhang, “Always Good-Turing: Asymptotically optimal probability estimation,” *Science*, vol. 302, no. 5644, pp. 427–431, 2003.

[28] A. F. Timan, *Theory of Approximation of Functions of a Real Variable.* Elsevier, 2014, vol. 34.

[29] J. C. Mason and D. C. Handscomb, *Chebyshev Polynomials.* Chapman and Hall/CRC, 2002.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[31] M. López and G. Still, “Semi-infinite programming,” *European Journal of Operational Research*, vol. 180, no. 2, pp. 491–518, 2007.

[32] R. Reemtsen and S. Görner, “Numerical methods for semi-infinite programming: A survey,” in *Semi-Infinite Programming.* Springer, 1998, pp. 195–275.

[33] R. Reemtsen, “Discretization methods for the solution of semi-infinite programming problems,” *Journal of Optimization Theory and Applications*, vol. 71, no. 1, pp. 85–103, 1991.

[34] G. Still, “Discretization in semi-infinite programming: The rate of convergence,” *Mathematical Programming*, vol. 91, no. 1, pp. 53–69, 2001.

[35] F. Farnoud, O. Milenkovic, and N. P. Santhanam, “Small-sample distribution estimation over sticky channels,” in *2009 IEEE International Symposium on Information Theory.* IEEE, 2009, pp. 1125–1129.

[36] F. Farnoud, N. P. Santhanam, and O. Milenkovic, "Alternating Markov chains for distribution estimation in the presence of errors," in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 2017–2021.

[37] D. S. Pavlichin, J. Jiao, and T. Weissman, "Approximate profile maximum likelihood," *arXiv preprint arXiv:1712.07177*, 2017.

[38] H. Yi, A. Orlitsky, A. T. Suresh, and Y. Wu, "Data amplification: A unified and competitive approach to property estimation," in *Advances in Neural Information Processing Systems*, 2018, pp. 8834–8843.

[39] Y. Han, J. Jiao, and T. Weissman, "Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance," *arXiv preprint arXiv:1802.08405*, 2018.

[40] M. Frieman, B. Yount, M. Heise, S. A. Kopecky-Bromberg, P. Palese, and R. S. Baric, "Severe acute respiratory syndrome coronavirus ORF6 antagonizes STAT1 function by sequestering nuclear import factors on the rough endoplasmic reticulum/Golgi membrane," *Journal of Virology*, vol. 81, no. 18, pp. 9812–9824, 2007.

[41] J. K. Taylor, C. M. Coleman, S. Postel, J. M. Sisk, J. G. Bernbaum, T. Venkataraman, E. J. Sundberg, and M. B. Frieman, "Severe acute respiratory syndrome coronavirus ORF7a inhibits bone marrow stromal antigen 2 virion tethering through a novel mechanism of glycosylation interference," *Journal of Virology*, vol. 89, no. 23, pp. 11 820–11 833, 2015.

[42] R. Cagliani, D. Forni, M. Clerici, and M. Sironi, "Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses," *Infection, Genetics and Evolution*, p. 104353, 2020.

[43] K.-L. Siu, K.-S. Yuen, C. Castaño-Rodriguez, Z.-W. Ye, M.-L. Yeung, S.-Y. Fung, S. Yuan, C.-P. Chan, K.-Y. Yuen, L. Enjuanes et al., "Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC," *The FASEB Journal*, vol. 33, no. 8, pp. 8865–8877, 2019.

[44] S. K. Lau, Y. Feng, H. Chen, H. K. Luk, W.-H. Yang, K. S. Li, Y.-Z. Zhang, Y. Huang, Z.-Z. Song, W.-N. Chow et al., "Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination," *Journal of Virology*, vol. 89, no. 20, pp. 10 532–10 547, 2015.

[45] S. Laha, J. Chakraborty, S. Das, S. K. Manna, S. Biswas, and R. Chatterjee, “Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission,” *Infection, Genetics and Evolution*, vol. 85, p. 104445, 2020.

[46] M. R. Islam, M. N. Hoque, M. S. Rahman, A. R. U. Alam, M. Akther, J. A. Puspo, S. Akter, M. Sultana, K. A. Crandall, and M. A. Hossain, “Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity,” *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.