Using Sensitivity Analyses to Approximate Total COVID-19 Infections: State and County level in the United States, March 2021 - March 2022

Quinn White

Submitted to the Department of Statistical and Data Sciences
of Smith College
in partial fulfillment
of the requirements for the degree of
Bachelor of Arts

Ben Baumer, Primary Faculty Advisor
Nicholas Reich, Secondary Faculty Advisor

May 2023

Acknowledgements

Will add

Table of Contents

Abstract

As we have navigated the COVID-19 pandemic, case counts have been a central source of information for understanding transmission dynamics and the effect of public health interventions. However, because the number of cases we observe is limited by the testing effort in a given location, the case counts presented on local or national dashboards are only a fraction of the true infections. Variations in testing rate by time and location impacts the number of cases that go unobserved, which can cloud our understanding of the true COVID-19 incidence at a given time point and can create biases in downstream analyses. Additionally, the number of cases we observe is impacted by the sensitivity and specificity of the diagnostic test. To quantify the number of true infections given incomplete testing and diagnostic test inaccuracy, this work implements probabilistic bias analysis at a biweekly time scale from January 1, 2021 through February 2022. In doing so, we can estimate a range of possible true infections for a given time interval and location. This approach can be applied at the state level across the United States, as well as in some counties where the needed data are available.

Dedication

You can have a dedication here if you wish.

Chapter 1

Introduction

Placeholder

Chapter 2

Background

2.1 Probabalistic Bias Analysis

Often the focus of quantifying error about an effect estimate focuses on random error rather than the systematic error. For example, typical frequentist confidence intervals are frequent in medical and epidemiological literature, although they have faced rising criticism (Greenland et al., 2016). These confidence intervals quantify the fraction of the times we expect the true value to fall in this interval under the assumption that our model is correct. That is, if we ran an experiment 100 times and computed the effect size each time, we would expect the 95% confidence interval to contain the true value to 95 of those times, on average. Neyman stressed this in his original publication formalizing the concept of a confidence interval in 1937 (Neyman, 1937). The nuance that the confidence interval is not the probability that the true value falls within this interval, however, is often lost in the discussion of results, in part because the true meaning of a confidence interval is less intuitive.

The aim of quantitative bias analysis is to estimate systematic error to give a range of possible values for the true quantity of interest. In this sense, it is a type of sensitivity analysis. It can be used to estimate various kinds of biases, from misclassification, as is implemented in this work, as well as selection bias and unmeasured confounding (Petersen, Ranker, Barnard-Mayers, MacLehose, & Fox, 2021). Often, the goal of performing such an analysis is to see how these sources of bias affect our estimates; in particular, under what situations of bias the observed effect would be null.

There are multiple different forms of bias analysis (Lash, Fox, & Fink, 2009). The most simple case, simple bias analysis, is correcting a point estimate for a single source of error. Multidimensional bias analysis extends this to consider sets of bias parameters, but still provides a corrected point estimate rather than a range of plausible estimates. Probabilistic bias analysis, meanwhile, defines probability distributions for bias parameters to generate a distribution of corrected estimates by repeatedly correcting estimates for bias under different combinations of the parameter values. Then, via Monte Carlo we obtain a distribution of corrected estimates that reflect the corrected values under different scenarios of bias, that is,

under different combinations of the bias parameters. This can give us a better idea for the extent of uncertainty about the corrected estimates, although this uncertainty does depend on the specification of the bias parameter distributions. Inherent in bias analysis is the dependence of our results on the specification of bias parameters, which reflect what is known from available data, literature, or theory on the extent of bias that may occur. There is uncertainty about how we define these distributions or values; otherwise, if the precise values of the bias parameters were known, we could simply correct the estimates and probabilistic bias analysis would not be useful.

Although some forms of probabilistic bias analysis can be applied to summarized data, for example, frequencies in a contingency table, the methods are most often implemented with unsummarized data in its original form, as implemented here.

In choosing specific distributions for the bias parameters, different specifications may yield density functions where most of the density is within a similar interval, which means the choice of the specific distribution will not be sensitive to the particular choice of density.

2.2 Background for the Approach

The Bayesian melding approach was proposed by Poole et al. (Poole & Raftery, 2000).

This approach enables us to account for both uncertainty from inputs and outputs of a deterministic model. The initial motivation for the approach was to study the population dynamics of whales in the presence of substantial uncertainty around model inputs for population growth (Poole & Raftery, 2000). However, the framework provided by Poole et al. can applied in any circumstance where we have uncertainty around some quantities θ and ϕ where there is a deterministic function $M : \theta \to \phi$. Due the utility of Bayesian melding in various contexts, since this deterministic model M could take on a wide range of forms, the approach has since been applied in various fields, including urban simulations (Ševčíková, Raftery, & Waddell, 2007), ecology (Robson, 2014), and infectious disease (Powers et al., 2011).

Let $M : \theta \to \phi$ be the deterministic model defined by the function relating a vector of input parameters θ to an output vector ϕ, and suppose we have a prior on θ denoted $f_\theta(\theta)$ and a prior on ϕ denoted $f_\phi^{direct}(\phi)$.

However, note that we actually have two distinct priors on ϕ. There is the prior formed by the distribution induced on ϕ by the prior for θ and the function M, where we denote this induced prior $f_\phi^{induced}(\phi)$. Generally, these priors are based on different sources of information.

If M^{-1} exists, we can write this induced prior $f_\phi^{induced}(\phi) = f_\theta(M^{-1}(\phi))|J(\phi)|$[1]. This result follows from the fact $M(\theta) = \phi$, so we apply a change of variables to obtain the distribution of ϕ from the distribution of $M(\theta)$.

[1] In the continuous case we need to multiply by $|J(\phi)|$, but not in the discrete case (Blitzstein & Hwang, 2019).

In practice, M^{-1} rarely exists exists since θ is often of higher dimensionality then ϕ, in which cases M is not invertible. This means we generally approximate $f_\phi^{induced}$ without acquiring its analytical form.

In addition to this induced prior, we have the prior $f_\phi^{direct}(\phi)$, which does not involve M nor the inputs θ. Since these priors are based on different sources of information and may reflect different uncertainties, often it useful to use both sources of information to inform our estimates. To do so, we need to combine the distributions for $f_\phi^{induced}$ and f_ϕ^{direct} to create a pooled distribution.

Multiple pooling strategies exist for distinct distributions, but one requirement for a Bayesian analysis is that the distribution should be independent of the order in which the prior is updated and the combining of the prior distribution. That is, updating the prior distributions using Bayes' theorem and then combining distributions should yield the same result as combining distributions and then updating this combined distribution; pooling methods that have this property are deemed externally Bayesian. Logarithmic pooling has been shown to be externally Bayesian under some conditions, which are likely to hold in most settings. Furthermore, logarithmic pooling has actually been shown to be the only pooling method where this holds (Genest, McConway, & Schervish, 1986). For this reason, Poole *et al.* recommend proceeding with logarithmic pooling for Bayesian melding.

The logarithmically pooled prior for ϕ by pooling $f_\phi^{induced}$ and f_ϕ^{direct} is

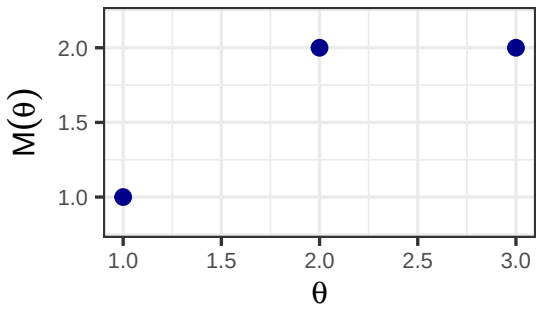
$$f_\phi^{pooled}(\phi) = t(\boldsymbol{\alpha})(f_\phi^{induced}(\phi))^{\alpha}(f_\phi^{direct}(\phi))^{1-\alpha}.$$

The pooling weights are given by $\boldsymbol{\alpha} = (\alpha,\ \ 1-\alpha)$ where $\alpha \in [0, 1]$, and $t(\boldsymbol{\alpha})$ is the normalizing constant. Commonly, a choice of $\alpha = 0.5$ is used to give the priors equal weight. In this case, logarithmic pooling may be referred to as geometric pooling since it is equivalent to taking a geometric mean.

If M is invertible, we can obtain the contrained distributions for the model inputs by simply inverting M. However, M is rarely invertible, so we have to think about how to proceed in the noninvertible case.

2.2.1 Simple Discrete Example

To get intuition for a valid strategy Poole et al. recommend, we consider a mapping $M : \theta \to \phi$ for $\theta \in \mathbb{R}$ and $\phi \in \mathbb{R}$ defined as follows (Figure 2.1). Note the choice of $f_\theta, f_\phi^{direct}$ does not matter here as long as they are valid densities.



θ	$f_\theta(\theta)$	$M(\theta) = \phi$	$f_\phi^{direct}(\phi)$
1	0.3	1	0.4
2	0.2	2	0.6
3	0.5	2	0.6

Figure 2.1: A simple discrete example where M is not invertible.

We see that M is not invertible since $\theta = 1$ and $\theta = 2$ both map to $\phi = 2$, which implies the inverse M^{-1} would not be well defined.

We can generate a sample from the density $f_\phi^{induced}$ by sampling from f_θ and computing $M(\theta)$.

So we have

$$f_\phi^{induced}(1) = f_\theta(1) = 0.3 \qquad \text{(since } \theta = 1 \text{ maps } \phi = 1\text{)}$$
$$f_\phi^{induced}(2) = f_\theta(2) + f_\theta(3) = 0.2 + 0.5 = 0.7 \qquad \text{(since } \theta = 2 \text{ and } \theta = 3 \text{ both map to } \phi = 2\text{)}$$

Then, we can compute the logarithmically pooled pooled prior with $\alpha = 0.5$ by taking $f_\phi^{induced}(\phi)^\alpha f_\phi^{direct}(\phi)^{1-\alpha}$.

This gives us

$$f_\phi^{induced}(\phi)^\alpha f_\phi^{direct}(\phi)^{1-\alpha} = (0.3)^{0.5}(0.4)^{0.5} = 0.3464$$
$$f_\phi^{induced}(\phi)^\alpha f_\phi^{direct}(\phi)^{1-\alpha} = (0.7)^{0.5}(0.6)^{0.5} = 0.6481.$$

To make this a valid density, however, these probabilities must sum to 1, so we renormalize by dividing by (0.3464 + 0.6481). Denoting the pooled prior in phi-space as $f_\phi^{pooled}(\phi)$, this gives us

$$f_\phi^{pooled}(1) = \frac{0.3464}{0.3464 + 0.6481} = 0.3483$$
$$f_\phi^{pooled}(2) = \frac{0.6481}{0.3464 + 0.6481} = 0.6517.$$

We summarize these results and compare $f_\phi^{induced}$, f_ϕ^{direct}, and f_ϕ^{pooled} in Figure 2.2.

ϕ	$f_\phi^{direct}(\phi)$	$f_\phi^{induced}(\phi)$	$f_\phi^{pooled}(\phi)$
1	0.4	0.3	0.3483
2	0.6	0.7	0.6517

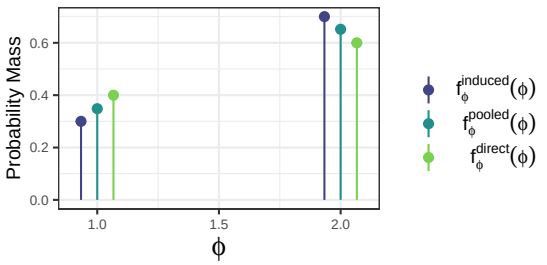


Figure 2.2

However, we also want the pooled prior on the inputs θ, that is, $f_\theta^{pooled}(\theta)$.

Poole et al. reasoned as follows. Since M uniquely maps $\theta = 1$ to $\phi = 1$, the probability that $\theta = 1$ should be equal to the probability $\phi = 1$. That is, we should have $f_\theta^{pooled}(1) = f_\phi^{pooled}(1)$.

However, the relationship for $\theta = 2$ or $\theta = 3$ to ϕ is not one to one. Since $M(2) = 2$ and $M(3) = 2$, the sum of the probabilities for $\theta = 1$ and $\theta = 2$ should be equal to that for $\phi = 2$, that is, $f_\theta^{pooled}(2) + f_\theta^{pooled}(3) = f_\phi^{pooled}(2) = 0.6517$.

The challenge here is how we divide the probability for $f_\phi^{pooled}(2)$, which is defined, among $f_\theta^{pooled}(2)$ and $f_\theta^{pooled}(3)$. The prior for ϕ yields no information to

assist in this choice, because knowing which value ϕ takes on does not give us any information about whether $\theta = 2$ or $\theta = 3$. Thus, the information we have about θ must be taken from $f_\theta(\theta)$.

That is, we can assign a probability for $f_\theta^{pooled}(2)$ by considering the probability that $\theta = 2$ relative to the probability $\theta = 3$, computing

$$f_\theta^{pooled}(2) = f_\phi^{pooled}(2)\Big(\frac{f_\theta(2)}{f_\theta(2) + f_\theta(3)}\Big).$$

That is, if the probability θ takes on the value 2 is lower in this case than the probability $\theta = 3$ which we know from the prior on θ, $f_\theta(\theta)$, then the pooled prior on θ, $f_\theta^{pooled}(2)$, should reflect this.

Using this reasoning, we have

$$f_\theta^{pooled}(2) = (0.7)\frac{0.2}{0.2 + 0.5} = 0.1862$$
$$f_\theta^{pooled}(3) = (0.7)\frac{0.5}{0.2 + 0.5} = 0.4655.$$

The result in this simple example, using $f_\theta(\theta)$ to determine how to distribute the probability for values of ϕ where multiple θ map to ϕ, can be used to derive general formulas to compute $f_\theta^{pooled}(\theta)$ for discrete and continuous distributions (Poole & Raftery, 2000).

2.2.2 General Solution for the Discrete Case

Denote the possible values of θ as $A_1, A_2, \ldots$, the possible values of ϕ as $B_1, B_2, \ldots$, and a mapping $m : \mathbb{N} \to \mathbb{N}$ such that $M(A_i) = B_{m(i)}$ and $C_j = M^{-1}(B_j) = \{A_i : M(A_i) = B_j\}$. Then

$$f_\theta^{pooled}(A_i) = f_\phi^{pooled}(B_{m(i)}) \left(\frac{f_\theta(A_i)}{f_\phi^{induced}(B_{m(i)})}\right).$$

2.2.3 General Solution for the Continuous Case

We denote $B = M(A) = \{M(\theta) : \theta \in A\}$ and $C = M^{-1}(B) = \{\theta : M(\theta) \in B\}$.

Then

$$f_\phi^{pooled}(M(\theta)) = t(\alpha) f_\theta(\theta) \left(\frac{f_\phi^{direct}(M(\theta))}{f_\phi^{induced}(M(\theta))}\right)^{1-\alpha} \tag{2}$$

where $t(\alpha)$ is a renormalizing constant for the choice of α.

2.2.4 Implementation through the Sampling-Importance-Resampling Algorithm

We can obtain the pooled distributions f_θ^{pooled} and f_ϕ^{pooled} by using the Sampling-Importance-Resampling Algorithm.

The steps are as follows.

1. We draw θ from its prior distribution $f_\theta(\theta)$.
2. For every θ_i we compute $\phi_i = M(\theta_i)$ to obtain a sample from the induced distribution.
3. Since the density $f_\phi^{induced}(\phi)$ is unlikely to have an analytical form, we can compute it via a density approximation such as kernel density estimation.
4. Construct weights proportional to the ratio of the prior on ϕ evaluated at $M(\theta_i)$ to the induced prior $f_\phi^{induced}$ evaluated at $M(\theta_i)$. If a likelihood $L_1(\theta)$ for the inputs and $L_2(\phi)$ is available, the weights are

$$w_i = \left(\frac{f_\phi^{direct}(M(\theta_i))}{f_\phi^{induced}(M(\theta_i))} \right)^{1-\alpha} L_1(\theta_i)\ L_2(M(\theta_i)).$$

   However, in this work, no likelihood is available for the variables of interest, so the likelihood is left out of the weights, leaving us with

$$w_i = \left(\frac{f_\phi^{direct}(M(\theta_i))}{f_\phi^{induced}(M(\theta_i))} \right)^{1-\alpha}.$$

5. Sample θ and ϕ from step (1) with probabilities proportional to the weights from (4).

2.3 Bayesian Melding Applied to COVID-19 Misclassification

In this work, we can relate the inputs $\theta = \{P(S_1|\text{untested}), \alpha, \beta\}$ and $\phi = P(S_0|\text{test}_+, \text{untested})$ by the deterministic model $M : \theta \to \phi$ given by $P(S_0|\text{test}_+, \text{untested}) = \frac{\beta(1 - P(S_1|\text{untested}))}{\beta(1 - P(S_1|\text{untested})) + \alpha P(S_1|\text{untested})}$. The derivation of M is in the following section.

Now, we have two distributions on ϕ: the distribution based on data on the asymptomatic rate of infection of COVID-19, and the distribution formed by taking $M(\theta)$ where θ represents the values from the defined distributions of α, β, and $P(S_1|\text{untested}$. With Bayesian melding, we pool these distributions using logarithmic pooling, and then implement the sampling-importance-resampling algorithm to obtain constrained distributions of the inputs θ that are in accordance with information about the asymptomatic rate of the virus.

Due to the uncertainty around our definitions of α and β, it is particularly useful to leverage the information we have about the asymptomatic rate of the virus $P(S_0|\text{test}_+, \text{untested})$ because a large collection of studies has been published in this area. In a meta-analysis pooling data from 95 studies, the pooled estimate among the confirmed population that was asymptomatic was 40.50% [95% CI, 33.50%-47.50%] (Ma et al., 2021). Another meta-analysis including 350 studies estimated the asymptomatic percentage to be 36.9% [95% CI: 31.8 to 42.4%], and, when restricting to screening studies, 47.3% (95% CI: 34.0% -61.0%) (Sah et al., 2021).

This means we have two priors on the asymptomatic rate ϕ, that by taking $M(\theta)$ for sampled values of θ, denoted $f_\phi^{induced}$ in the previous section, and that based on data about the asymptomatic rate, f_ϕ^{direct}.

2.3.1 Distribution of $\theta = \{\alpha, \beta, P(S_1|\textbf{untested})\}$

First, we obtain a sample $\theta_1, \theta_2, \ldots, \theta_k$ from θ (Figure 2.3).

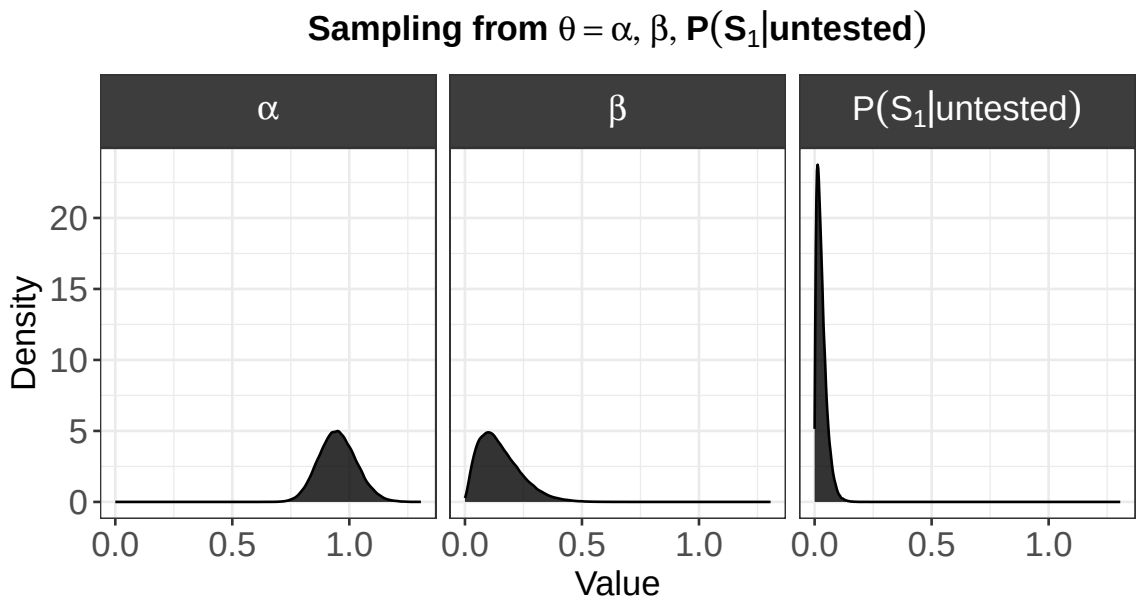


Figure 2.3

(#fig:create theta.png)

2.3.2 Direct Prior and Induced Prior Distributions for $P(S_0|\textbf{test}_+, \textbf{untested})$

Then, taking $M(\theta)$, we can compute the induced distribution $f_\phi^{induced}(M(\theta))$ and compare it to our prior on ϕ from meta-analyses on the asymptomatic rate, $f_\phi^{direct}(\phi)$ (2.4).

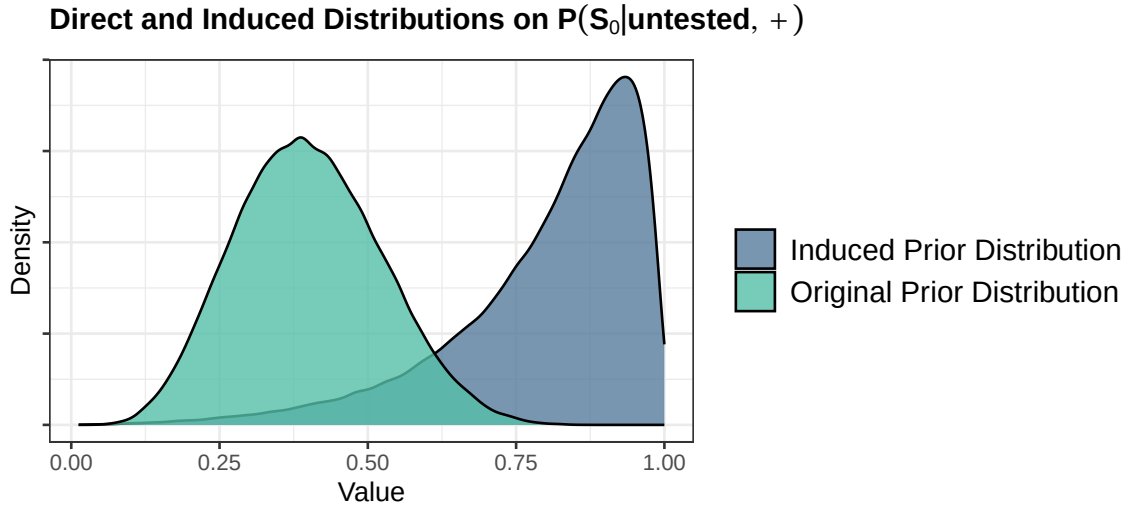


Figure 2.4

2.3.3 Pooling

At this point, we want to obtain the logarithmically pooled distribution of the two priors we have on ϕ, denoted f_ϕ^{pooled}.

Now, as described in greater detail in the section on the Sampling-Importance-Resampling algorithm, the weights are $w_i = \left(\frac{f_\phi^{direct}(M(\theta_i))}{f_\phi^{induced}(M(\theta_i))}\right)^{1-\alpha}$.

We perform a kernel density estimation to approximate the density of $f_\phi^{induced}(\phi)$ at the coordinates $\phi_1, \ldots, \phi_M$. To compute $f_\phi^{direct}(\phi)$, we can use the density function f_ϕ^{direct}.

Once we have these weights, we resample the $\phi_1, \ldots, \phi_M$ to obtain a sample from the target distribution $t(\alpha)\Big(f^{induced}(M(\theta))\Big)^{0.5}\Big(f^{direct}(M(\theta))\Big)^{0.5}$, where $t(\alpha)$ is the normalizing constant needed to make the pooled density valid. We resample $\theta_1, \ldots, \theta_k$ with the same weights to obtain the constrained distributions for the inputs.

We see the melded distributions and pre-melding distributions in Figure 2.5.

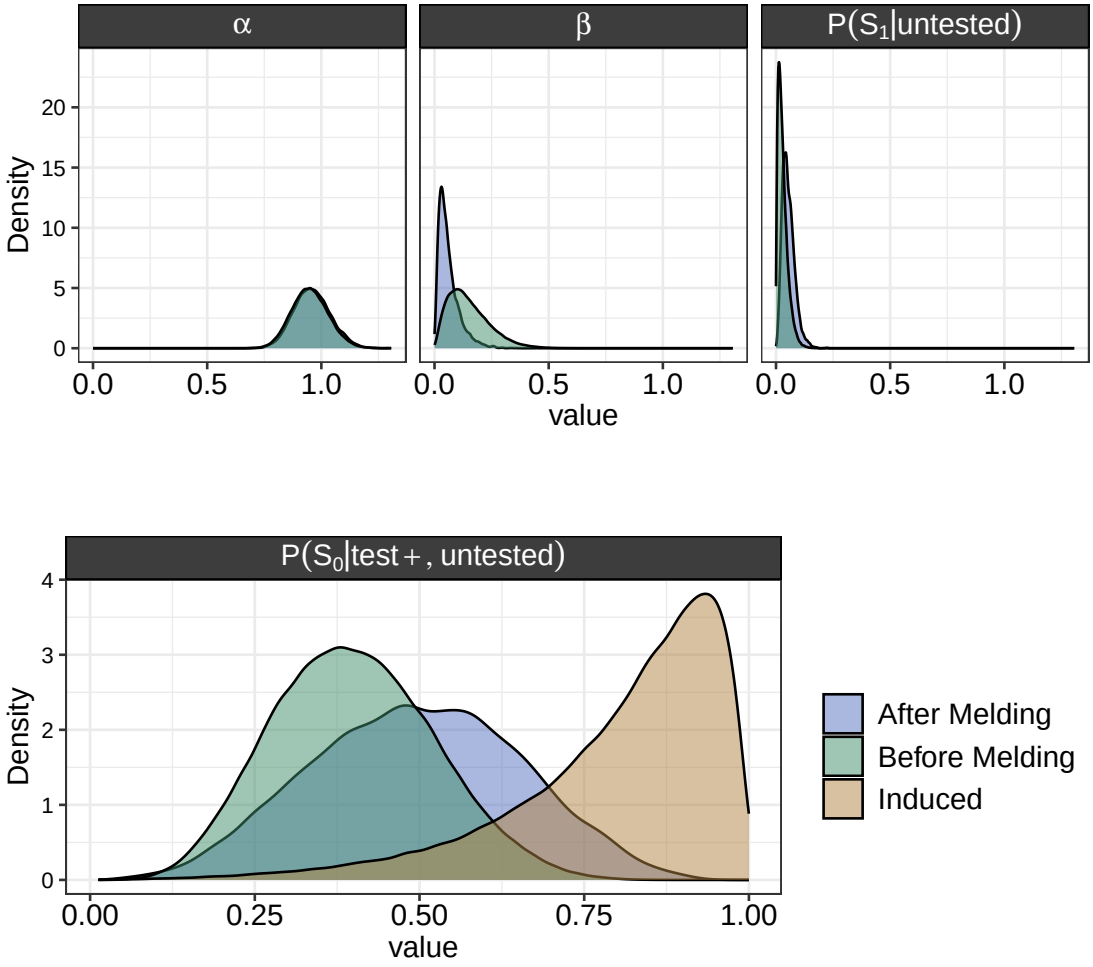
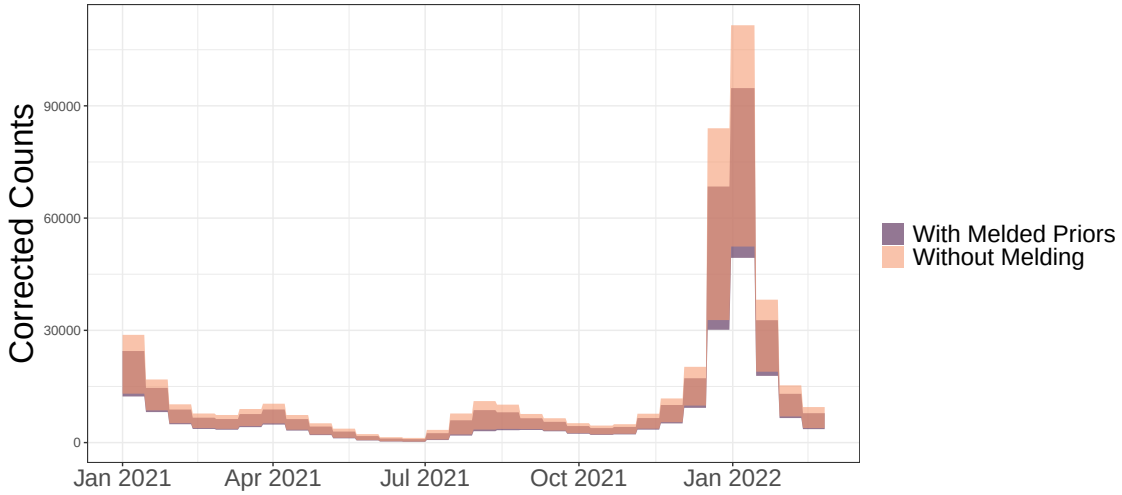


Figure 2.5

Comparing the induced and direct priors on $P(S_0|\text{test}_+, \text{untested})$ above, we see that although they have shared support, some values from the induced distribution we acquire by using M to generate values of ϕ from sampled values of θ are very unlikely to be in accordance with the information we know about the prevalence of SARS-CoV-2 asymptomatic infection. This is where Bayesian melding comes into play. Pooling these distributions enable us to take both the prior on f^{direct} from published analyses on asymptomatic infection, and the induced prior, $f^{induced}$, into account to constrain the distributions of both the model inputs $\theta = \{\alpha, \beta, P(S_1|\text{untested})\}$ and model output $\phi = P(S_0|\text{test}_+, \text{untested})$ to be in accordance with both prior distributions. We then use these constrained distributions as inputs in the probabilistic bias analysis.

2.3.4 Motivating Example

We can see the impact of using melded priors in Suffolk county in} Massachusetts in Figure **??**. Since using the priors without melding allows for asymptomatic rates $P(S_0|\text{test}_+, \text{untested})$ that are extremely high, the upper bound of the estimates will be substantially higher than predicted when using the melded priors, which do not include values where the inputs lead to values of asymptomatic rate that are unsupported by available data (e.g., meta-analyses) on the asymptomatic rate.



2.3.5 Derivation of M

We define θ as the set of bias parameters $\{P(S_1|\text{untested}), \alpha, \beta\}$. The parameters α and β relate the observed overall test positivity rate to the test positivity rate we would obtain if we tested the asymptomatic and symptomatic partitions of the untested population. We define:

- $\alpha = \dfrac{P(\text{test}_+|S_1, \text{untested})}{P(\text{test}_+|\text{tested})}$
- $\beta = \dfrac{P(\text{test}_+|S_0, \text{untested})}{P(\text{test}_+|\text{tested})}$.

The parameter $P(S_1|\text{untested})$ reflects the probability someone among the untested population has moderate to severe COVID-like symptoms.

We relate this set of parameters to the asymptomatic infection rate $\phi = P(S_0|\text{test}_+, \text{untested})$ by the function $M : \theta \to \phi$:

$$M(\theta) = \frac{\beta(1 - P(S_1|\text{untested}))}{\beta(1 - P(S_1|\text{untested})) + \alpha(P(S_1|\text{untested})} = P(S_0|\text{test}_+, \text{untested}).$$

In what follows, we show this equality holds.

Since we have $\alpha = \frac{P(\text{test}_+|S_1,\text{untested})}{P(\text{test}_+|tested)}$ and $\beta = \frac{P(\text{test}_+|S_0, \text{untested})}{P(\text{test}_+|tested)}$, we can write

$$= \frac{\frac{P(\text{test}_+|S_0, \text{untested})}{P(\text{test}_+|tested)}(1 - P(S_1|\text{untested}))}{\frac{P(\text{test}_+|S_0, \text{untested})}{P(\text{test}_+|tested)}(1 - P(S_1|\text{untested})) + \frac{P(\text{test}_+|S_1, \text{untested})}{P(\text{test}_+|tested)}P(S_1|\text{untested})}$$

and cancelling out the term $P(\text{test}_+|tested)$ we have

$$= \frac{P(\text{test}_+|S_0, \text{untested})(1 - P(S_1|\text{untested}))}{P(\text{test}_+|S_0, \text{untested})(1 - P(S_1|\text{untested})) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}.$$

Since $P(S_0|\text{untested}) = 1 - P(S_1|\text{untested})$,

$$= \frac{P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested})}{P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested}) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}.$$

Applying the definition of conditional probability to the term $P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested})$ in the numerator,

$$= \frac{\left(\frac{P(\text{test}_+, S_0, \text{untested})}{P(S_0, \text{untested})}\right)\left(\frac{P(S_0, \text{untested})}{P(\text{untested})}\right)}{P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested}) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}$$

$$= \frac{\left(\frac{P(\text{test}_+, S_0, \text{untested})}{P(S_0, \text{untested})}\right)\left(\frac{P(S_0, \text{untested})}{P(\text{untested})}\right)}{P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested}) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}$$

$$= \frac{\frac{P(\text{test}_+, S_0, \text{untested})}{P(\text{untested})}}{P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested}) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}$$

$$= \frac{P(\text{test}_+, S_0|\text{untested})}{P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested}) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}.$$

We can substitute this result in for the $P(\text{test}_+|S_0, \text{untested})P(S_0|\text{untested})$ term in the denominator to yield

$$= \frac{P(\text{test}_+, S_0|\text{untested})}{P(\text{test}_+, S_0|\text{untested}) + P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested})}$$

With same reasoning, we can simplify

$$P(\text{test}_+|S_1, \text{untested})P(S_1|\text{untested}) = P(S_1, \text{test}_+|\text{untested}),$$

giving us

$$\begin{aligned}
&= \frac{P(\text{test}_+, S_0|\text{untested})}{P(\text{test}_+, S_0|\text{untested}) + P(S_1, \text{test}_+|\text{untested})} \\
&= \frac{P(\text{test}_+, S_0|\text{untested})}{P(\text{test}_+|\text{untested})} \\
&= \frac{\frac{P(S_0, \text{test}_+, \text{untested})}{P(\text{untested})}}{\frac{P(\text{test}_+, \text{untested})}{P(\text{untested})}} \\
&= \frac{P(S_0, \text{test}_+, \text{untested})}{P(\text{test}_+, \text{untested})} \\
&= P(S_0|\text{test}_+, \text{untested}).
\end{aligned}$$

Hence, we have

$$P(S_0|\text{test}_+, \text{untested}) = \frac{\beta(1 - P(S_1|\text{untested}))}{\beta(1 - P(S_1|\text{untested})) + \alpha(P(S_1|\text{untested})}$$

as desired. □

2.4 Sampling-Importance-Resampling Algorithm

2.4.1 Overview

The Sampling-Importance-Resampling Algorithm, introduced in Rubin (1987), is a non-iterative method for approximating a sample from a target probability density function f. This algorithm is fundamental to the implementation of Bayesian melding.

The two main steps are the sampling step and importance resampling step. We have two (generally distinct) sample sizes, where m is the initial sample size and r is the resample size.

In the sampling step, we draw an independent and identically distributed sample of size m from g, $Y_1, Y_2, \ldots, Y_m$. Then, we compute weights $h(Y)$ such that $g \cdot h \propto f$. That is, we set the weights

$$w_i = h(Y_i) = \frac{\frac{f(Y_i)}{g(Y_i)}}{\sum_{i=1}^{m} \frac{f(Y_i)}{g(Y_i)}}.$$

We resample with these defined weights to obtain a sample of size r from $Y_1, Y_2, \ldots, Y_m$. We denote this resample $Z_1, \ldots, Z_r$. With these weights, $Z_1, \ldots, Z_r$ is approximately a sample from f.

The method is most efficient when g is a good approximation of f. The relationship between the sample size m and resample size r also has implications for the quality of the approximation. The algorithm generates independent and identically distributed samples as $m/r \to \infty$, but in most applications m/r between 10 and 20 is appropriate (Rubin, Gelman, & Meng, 2004). The practical implications of this choice are discussed later in this section.

To better understand the use of this algorithm, we provide a proof that formally relates the choice of g, weights h, and the target distribution f. We then follow up with a couple concrete examples where there is a closed formed solution to visualize how the algorithm works in practice.

2.4.2 Proof that Algorithm Obtains Approximate Sample from Target Distribution

To gain further insight into how sampling with weights $w_i = \left(\frac{f_\phi^{direct}(M(\theta_i))}{f_\phi^{induced}(M(\theta_i))}\right)^{0.5}$ approximates a sample from the target distribution the logarithmically pooled distribution f^{pooled}, we first prove a more general result.

Function M Relating Testing Positivity Parameters to Asymptomatic Rate

Suppose we sample $Y_1, Y_2, \ldots, Y_m$ independently and identically distributed with probability density function g and compute the weights

$$w_i = \frac{h(Y_i)}{\sum_{i=1}^m h(Y_i)}$$

for some nonnegative function h defined on the support of Y.
If we sample $Z_1, \ldots, Z_r$ from the discrete distribution $Y_1, \ldots, Y_m$ such that

$$P(Z = Y_i) = \frac{h(Y_i)}{\sum_{i=1}^m h(Y_i)} = w_i,$$

then $Z_1, \ldots, Z_r$ is approximately a sample with density proportional to $h \cdot g$.

Since Z is sampled from Y, we have

$$P(Z \leq x) = \sum_{z_i \leq x} P(Z = z_i) = \sum_{Y_i \leq x} P(Z = Y_i).$$

We can take this sum to be over all possible values of Y by including the indicator function $\mathbb{I}(Y_i \leq x)$, yielding

$$= \sum_{i=1}^m P(Z = y_i) \;\; \mathbb{I}(Y_i \leq x).$$

and since $P(Z = Y_i) = \dfrac{h(Y_i)}{\sum_{i=1}^m h(Y_i)}$ by definition we have

$$\begin{aligned}
&= \sum_{i=1}^m \frac{h(Y_i)}{\sum_{i=1}^m h(Y_i)} \;\; \mathbb{I}(Y_i \leq x) \\
&= \left(\frac{1}{\sum_{i=1}^m h(Y_i)} \right) \sum_{i=1}^m h(Y_i) \;\; \mathbb{I}(Y_i \leq x) \\
&= \frac{\sum_{i=1}^m h(Y_i) \;\; \mathbb{I}(Y_i \leq x)}{\sum_{i=1}^m h(Y_i)} \\
&= \frac{\frac{1}{m}\sum_{i=1}^m h(Y_i) \;\; \mathbb{I}(Y_i \leq x)}{\frac{1}{m}\sum_{i=1}^m h(Y_i)}.
\end{aligned}$$

Now, we need the Weak Law of Large Numbers. That is, if we have a sequence of random variables $X_1, X_2, \ldots$ with finite variance, then,

$$\lim_{n \to \infty} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = E(X_i).$$

Applying this law to both the numerator and denominator, we obtain

$$\begin{aligned}\lim_{m\to\infty}\left(\frac{\frac{1}{m}\sum_{i=1}^{m} h(Y_i)\ \ \mathbb{I}(Y_i \le x)}{\frac{1}{m}\sum_{i=1}^{m} h(Y_i)}\right) &= \frac{E_g[h(Y)\ \ \mathbb{I}(Y \le x)]}{E_g[h(Y)]} \\ &= \frac{\int_{-\infty}^{\infty} h(y)\ \ \mathbb{I}(y \le x)\ g(y)\ dy}{\int_{-\infty}^{\infty} h(y)\, g(y)\ dy} \\ &= \frac{\int_{-\infty}^{x} h(y)\, g(y) dy}{\int_{-\infty}^{\infty} h(y)\, g(y)\ dy} \\ &\propto \int_{-\infty}^{x} h(y)\, g(y) dy.\end{aligned}$$

It follows that the probability density function of Z is proportional to $h \cdot g$.

□

It is easiest to understand the Sampling-Importance-Resampling Algorithm when the resampled distribution has a closed form, which we can see in the following two examples.

Example 1:

Suppose $Y \sim Exp(\lambda)$, so we have the PDF $f_Y(y) = \lambda e^{-\lambda y}$, and we sample $Z_1, \ldots, Z_r$ from $Y_1, \ldots, Y_m$ with weights direction proportional to X, that is, $h(Y) = Y$.

Then $Z_1, \ldots Z_r$ is approximately a sample from $h(x)\ f_Y(y) = y\ \lambda e^{-\lambda y}$.

From the PDF of the gamma distribution, $\frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$ we can recognize that $y \cdot e^{-\lambda y}$ corresponds to the gamma distribution with $\alpha = 2$ and $\beta = \lambda$.

We can see this result by considering Y before and after resampling below (Figure 2.6).

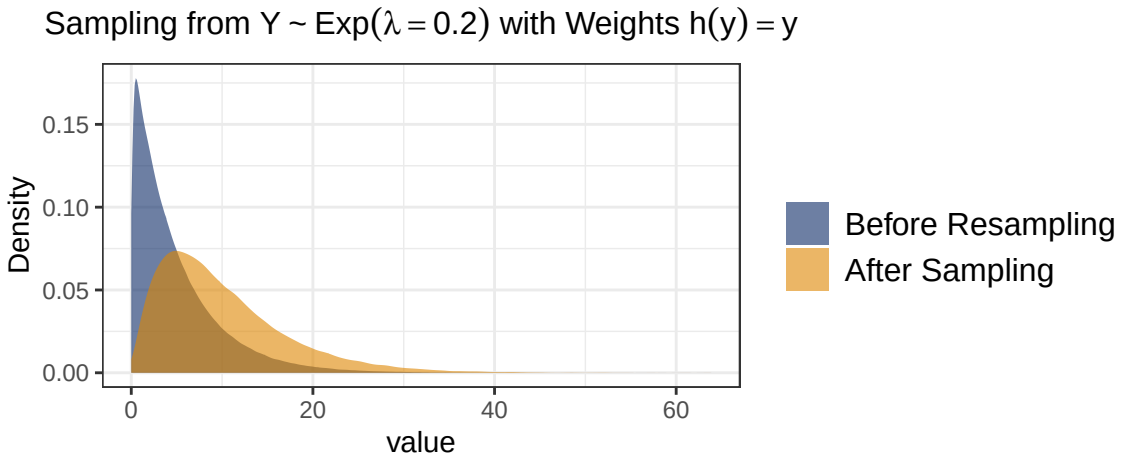


Figure 2.6

Then, we can see that the PDF of the the gamma distribution with $\alpha = 2$ and $\beta = \lambda$ corresponds to the post-sampling distribution as expected (Figure 2.7).

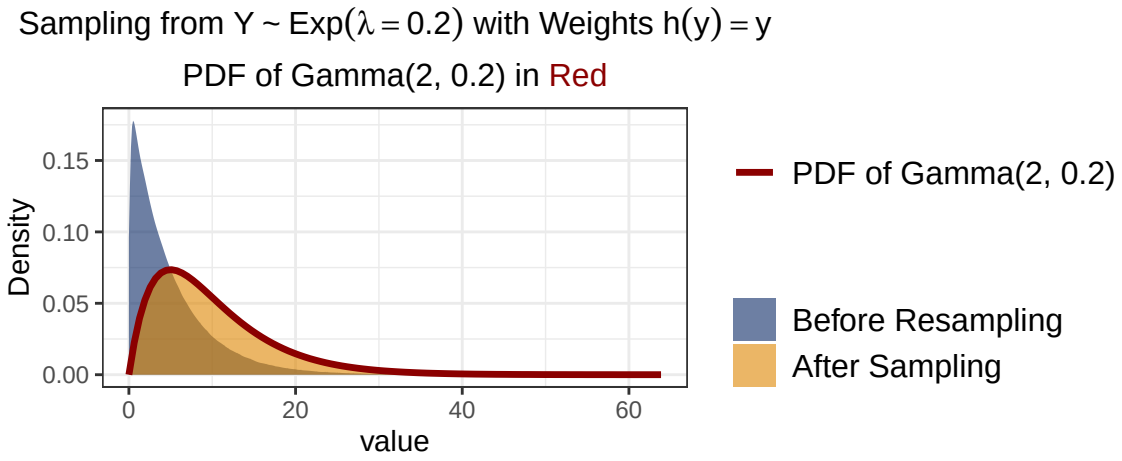


Figure 2.7

Example 2:

Similarly, again suppose $Y \sim Exp(\lambda)$, so $f_Y(y) = \lambda e^{-\lambda y}$. However, now we sample with weights defined by $h(y) = e^{-\lambda y}$. Then our sample $Z_1, \ldots, Z_r$ is approximately a sample from

$$\begin{aligned} h(y)\ f_Y(y) &= e^{-\lambda y} \cdot \lambda e^{-\lambda y} \\ &= e^{-2\lambda y} \end{aligned}$$

which is proportional to the exponential distribution with parameter 2λ.

The distributions before and after resampling are shown in Figure 2.8.

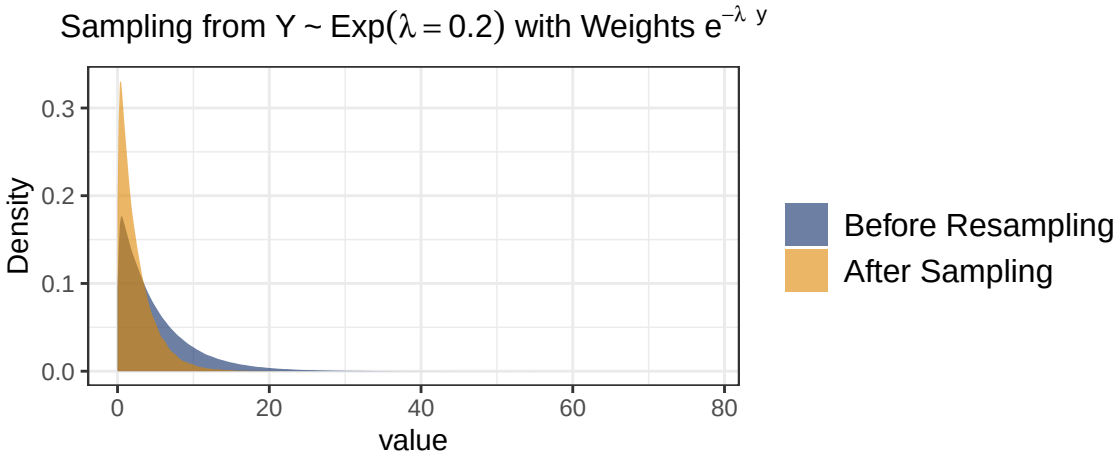


Figure 2.8

and then plotting the PDF of the exponential distribution with parameter 2λ we can see the correspondence to the post-sampling distribution (Figure 2.9).

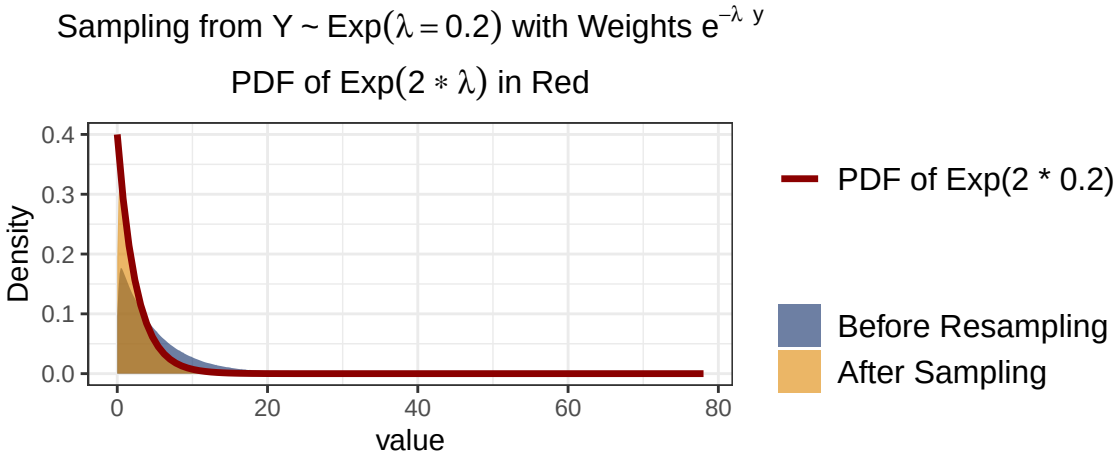


Figure 2.9

2.4.3 Obtaining Logarithmic Pooled Distribution with the Sampling-Importance-Resampling Algorithm

As outlined in Carvalho, Villela, Coelho, & Bastos (2023), we can formally define logarithmic pooling as follows.

If we have a set of densities $\{f_1(\phi), f_2(\phi), \ldots, f_n(\phi)\}$ and corresponding pooling weights $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, then the pooled density is

$$f^{\text{pooled}}(\phi) = t(\boldsymbol{\alpha}) \prod_{i=0}^{n} f_i(\phi)^{\alpha_i},$$

where $t(\boldsymbol{\alpha})$ is the normalizing constant $t(\boldsymbol{\alpha}) = \dfrac{1}{\int_\Phi \prod_{i=0}^{n} f_i(\phi)^{\alpha_i} d\phi}$ to ensure the pooled density is a valid probability density.

The case for this work is more simple: we only have two densities we wish to pool, $f_\phi^{induced}$ and f_ϕ^{direct}, and we assign them equal weights by letting $\boldsymbol{\alpha} = \{.5, .5\}$. This yields

$$f^{pooled}(\phi) = t(\boldsymbol{\alpha}) \left(f^{induced}(\phi)\right)^{0.5} \left(f^{direct}(\phi)\right)^{0.5}.$$

Since our target distribution is $t(\boldsymbol{\alpha}) \left(f^{induced}(\phi)\right)^{0.5} \left(f^{direct}(\phi)\right)^{0.5}$, and we have a sample from $f^{induced}$, we compute the weights such that

$$\begin{aligned} w_i &\propto \frac{\left(f^{induced}(\phi_i)\right)^{0.5} \left(f^{direct}(\phi_i)\right)^{0.5}}{f^{induced}(\phi_i)} \\ &= \frac{\left(f^{direct}(\phi_i)\right)^{0.5}}{\left(f^{induced}(\phi_i)\right)^{0.5}} \\ &= \left(\frac{f^{direct}(\phi_i)}{f^{induced}(\phi_i)}\right)^{0.5}. \end{aligned}$$

Sampling from $f^{induced}$ with these weights will yield a sample with approximately the target density $t(\alpha) \left(f^{induced}(\phi)\right)^{0.5} \left(f^{direct}(\phi)\right)^{0.5}$ from the result in the previous section.

2.4.4 Implications of the Sample Size and Resample Size

When we have an initial sample of size m from g, denoted $Y_1, \ldots, Y_m$, and take a weighted sample of size r, $Z_1, \ldots, Z_r$, the choices of m and r can have notable effects on the estimated distribution of the resample. In particular, when the sample size and resample size do not differ substantially, it becomes more likely that we will sample some element of $Y_1, \ldots, Y_m$ more than once. This can result in irregularities in the estimated distribution of $Z_1, \ldots, Z_r$. We see this in 2.10 when the ratio of m/r

is closer to 1, while the problem reduces as we increase the sample size m compared to the posterior (resample) size r.

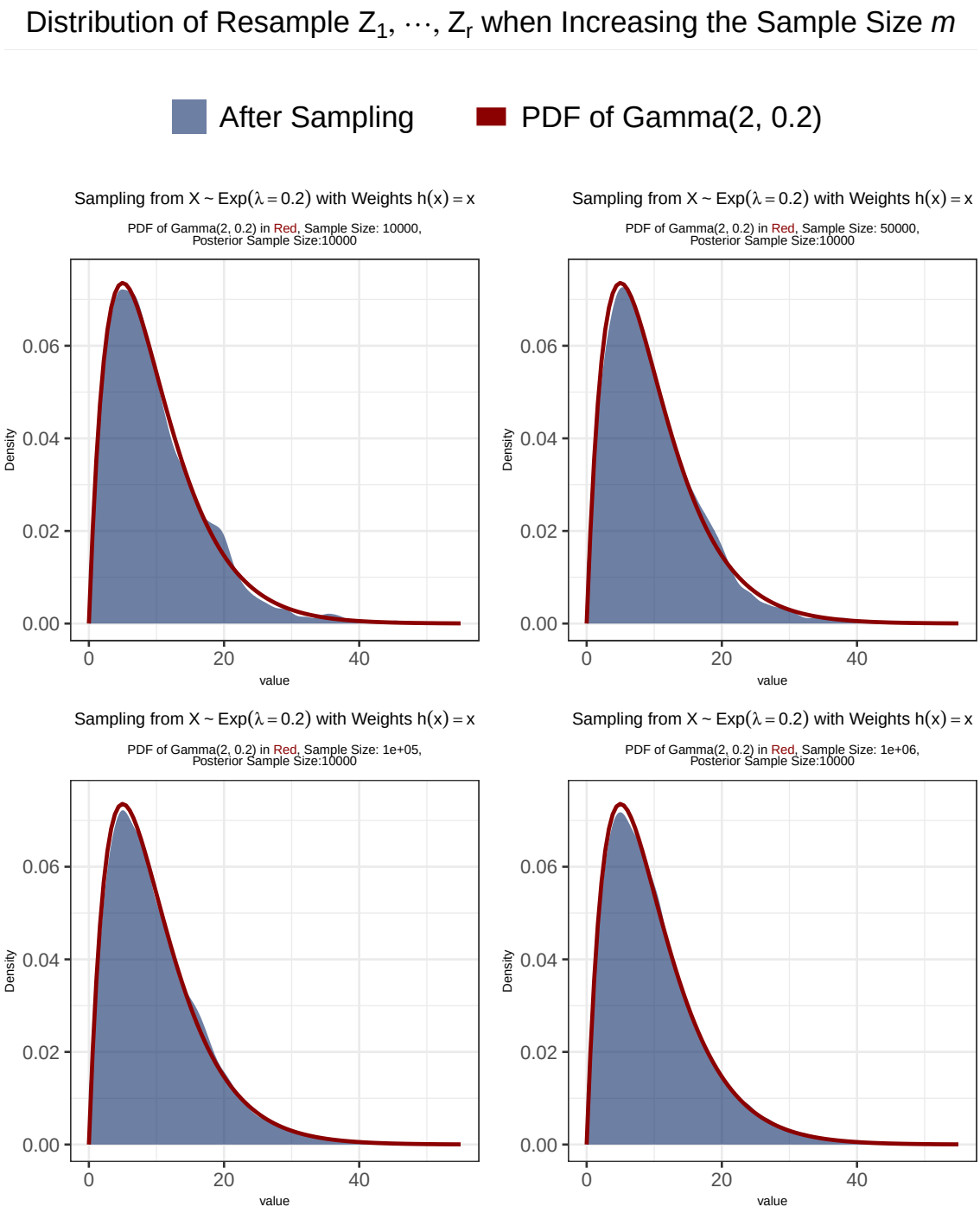


Figure 2.10

When using the Sampling-Importance-Resampling algorithm to obtain the logarithmically pooled distribution, see the effect of this choice has a major impact when we are melding truncated distributions. The pooled distribution is only defined

on the intersection of the supports of the distributions being pooled. Truncation, then, can limit the choices of $Y_1, \ldots, Y_m$ we take when resampling, which can lead to substantial irregularities in the resulting estimated pooled distribution (Figure 2.11).

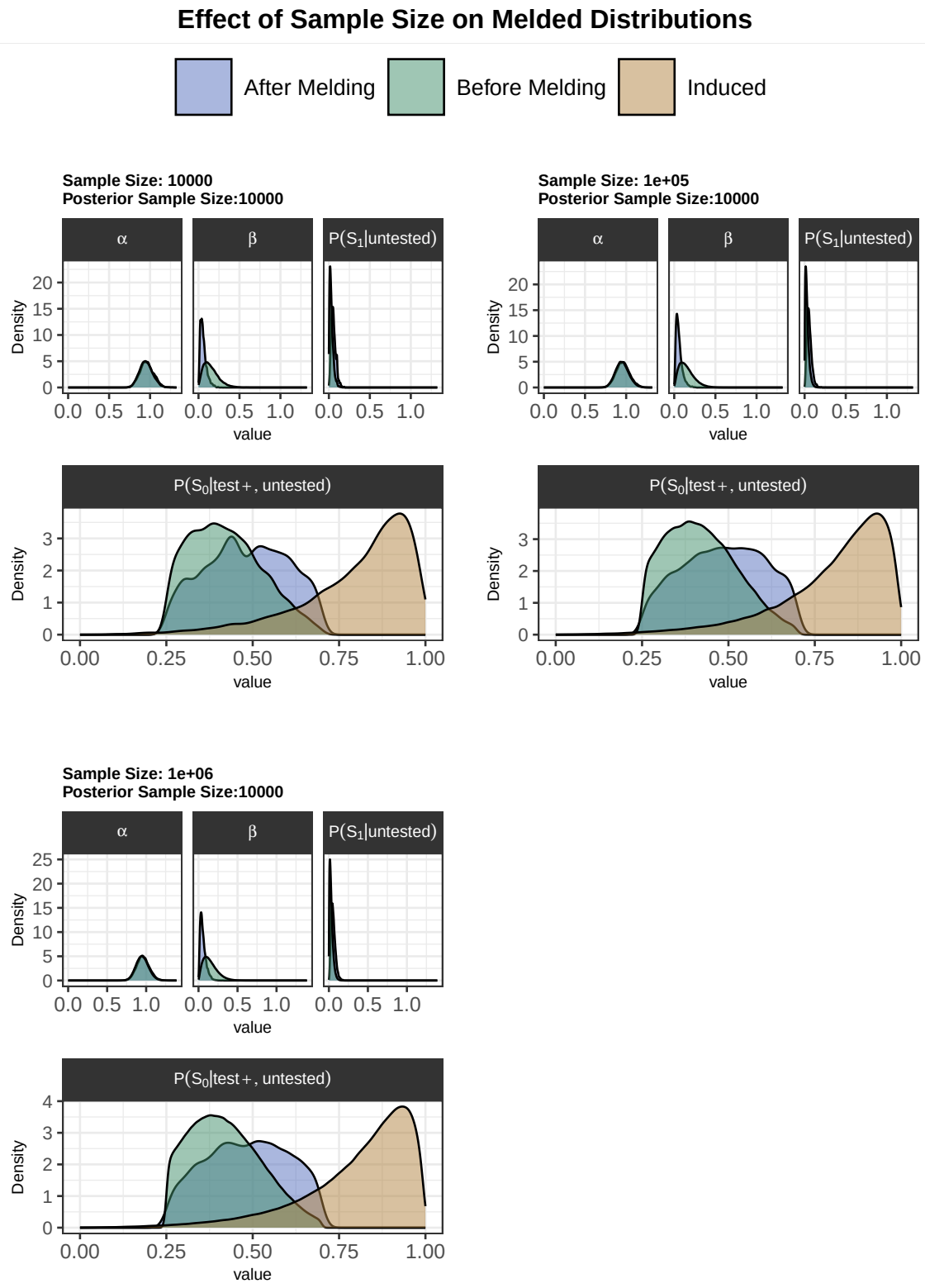

Effect of Sample Size on Melded Distributions
After Melding
Before Melding
Induced
Sample Size: 10000
Posterior Sample Size:10000
Sample Size: 1e+05
Posterior Sample Size:10000
Sample Size: 1e+06
Posterior Sample Size:10000
α
β
P(S1|untested)
P(S0|test+, untested)
Density
value

Figure 2.11

2.5 LOESS Smoothing

2.5.1 Introduction

Locally estimated scatterplot smoothing (LOESS) fits a collection of local regression models to obtain a smooth curve through the observed data (Figure 2.12). It is highly flexible in the sense that we do not have to specify the functional relationship between the predictor and response variable for the entire range of the predictor, which may be impossible in various settings. It is particularly useful when working with time series data with substantial noise.

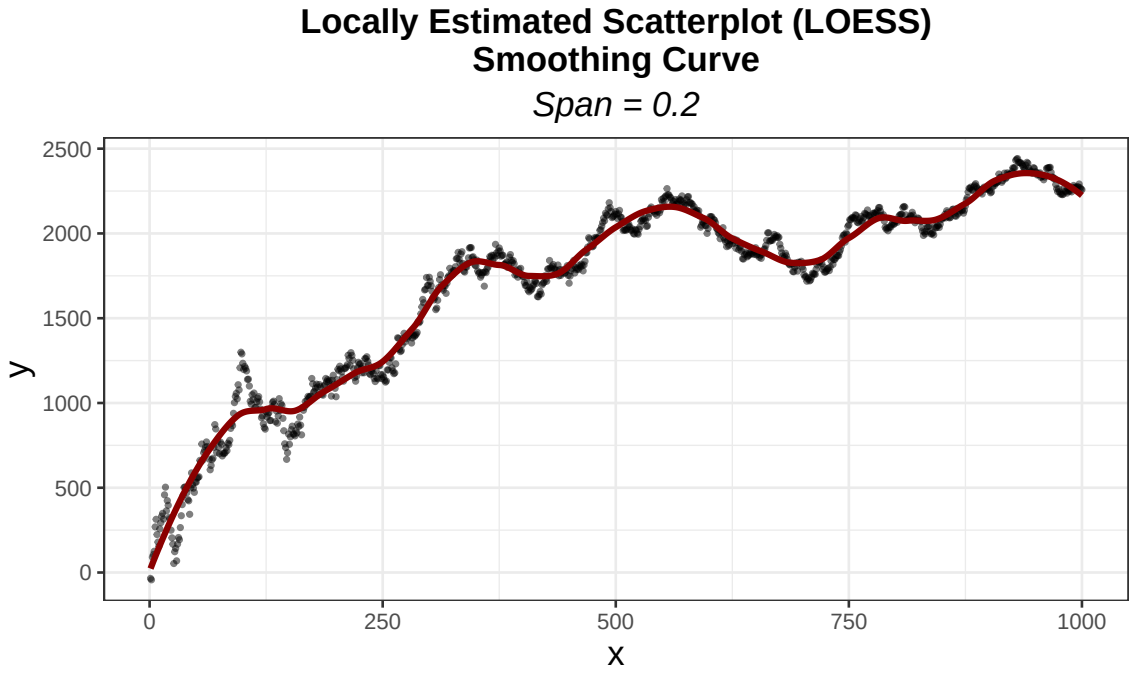


Figure 2.12: LOESS curve fitted with a span of 0.2.

To perform LOESS smoothing, we estimate a set of local regressions (Chambers, 1997). To do this, we must specify the span; this smoothing parameter is the fraction of the data that is used for the local polynomial fit. With a smaller span, the resulting curve will fit the trends more closely, while a larger span reflects broader trends (Figure 2.13).

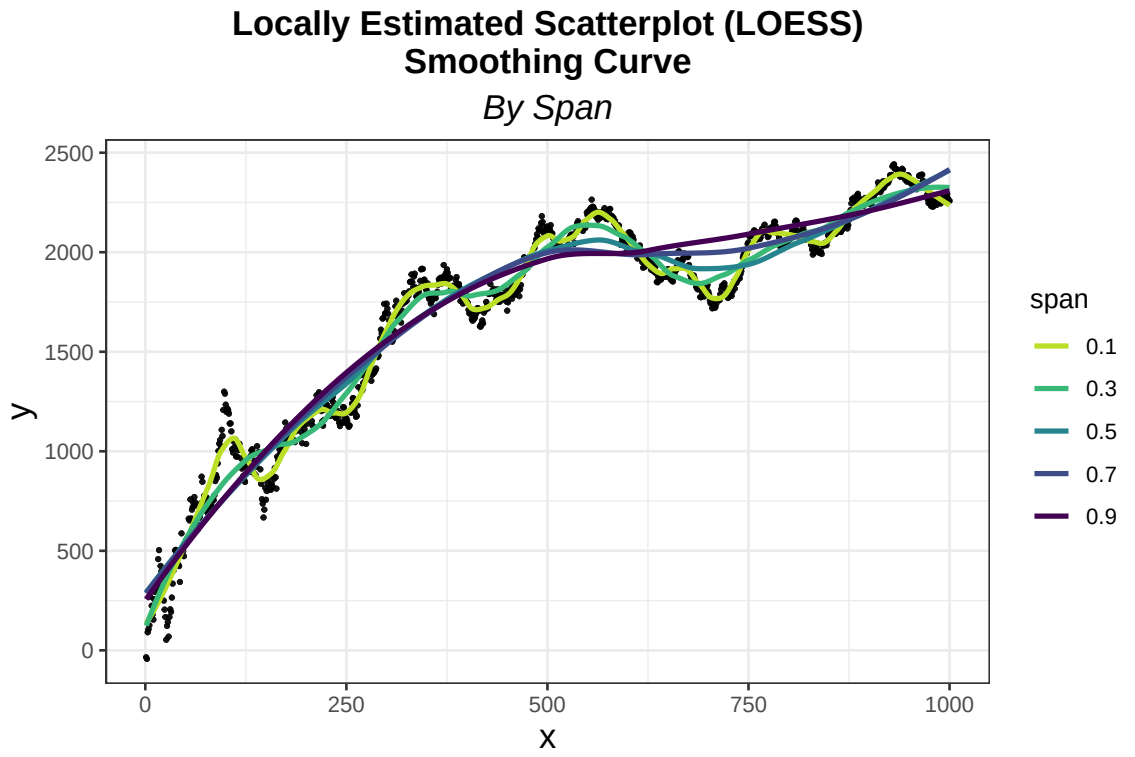


Figure 2.13

2.5.2 Fitting the LOESS Curve

To introduce some notation for the model at hand, we have a dependent variable $\mathbf{y}$ and independent variable $\mathbf{x}$, where $\mathbf{y}$ and $\mathbf{x}$ are related by some unknown function g, that is, $y = g(x) + \boldsymbol{\epsilon}$[2]. When we want to use LOESS smoothing to estimate g, often this function is complex, so we break up the problem into estimating a set of local regressions.

To obtain a predicted value $\hat{g}(x^*)$ for a particular value of the independent variable x^*, we fit a polynomial with greatest weight placed on points in the neighborhood of x^*, where the width of this neighborhood is defined by the choice of smoothing span. Let $\alpha \in (0, 1]$ denote the chosen smoothing span.

For a particular value of x^*, we estimate the predicted value $\hat{g}(x^*)$ by fitting a local regression. We first compute the weights by computing the vector of distances from this point x^*, that is,

$$\Delta(x^*) = |\mathbf{x} - x^*|$$

We define $q = \text{floor}(\alpha n)$, and take $\Delta_q(x^*) \in \mathbb{R}$ to be the q^th smallest distance of $\Delta(x^*)$.

The vector of weights is then

$$T(\Delta(x^*), \Delta_q(x^*))$$

[2]Recall we use bold type for vectors, e.g., $\mathbf{x} \in \mathbb{R}^n$ is a vector with observations $x_i \in \mathbb{R}$.

where T is the tricube weight function given by

$$T(x) = \begin{cases} (1-(x)^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases}.$$

Essentially, this process gives weight to points in the neighborhood of x^*. Consider $x^* = 500$ and smoothing span $= \alpha = .2$.

Then the weights we obtain are given in Figure 2.14.

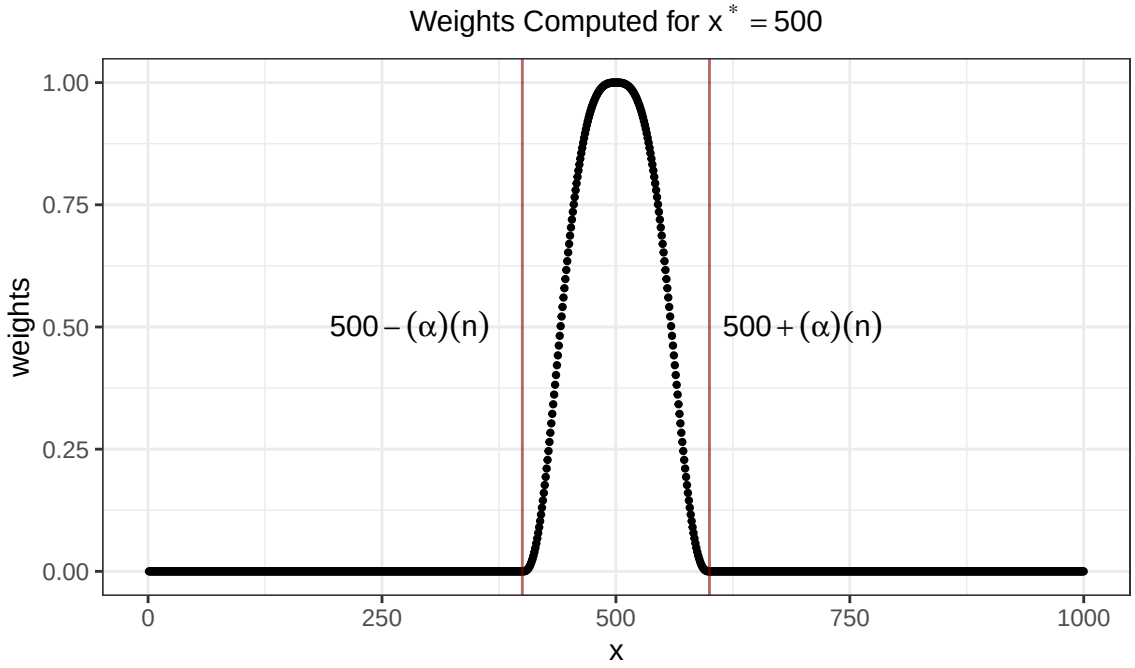


Figure 2.14: The only values with nonzero weights are those within the interval $(500-\alpha(n), 500-\alpha(n))$. That is, the proportion α of the data points closest to x^* will have nonzero weights.

We fit a linear regression with polynomial terms, typically with degree up to 2, with these weights. For example, fitting the model for this same $x^* = 500$, we obtain the polynomial in Figure 2.15.

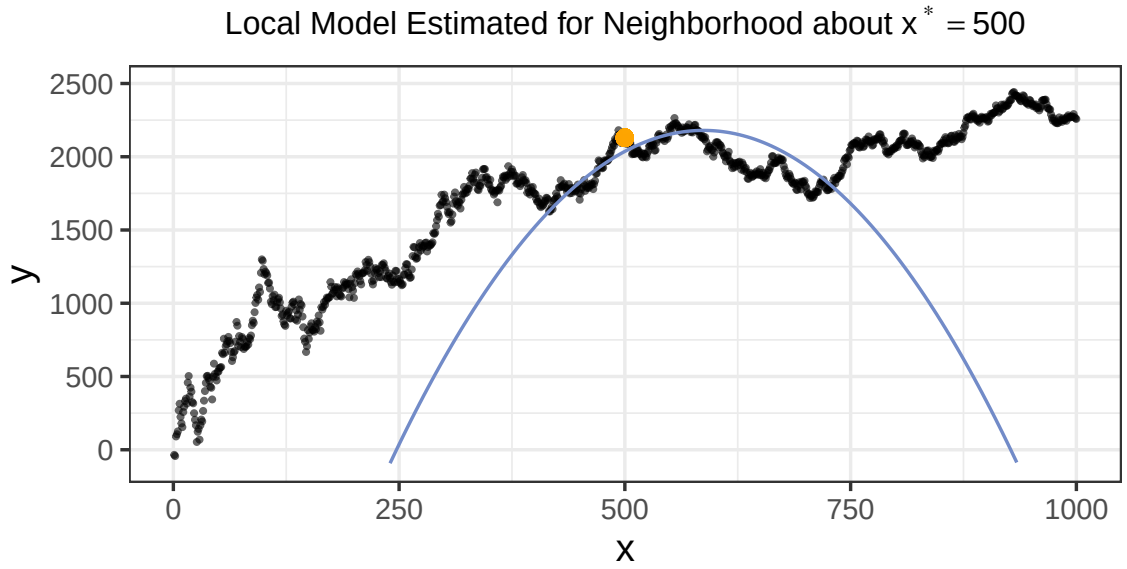


Figure 2.15

By fitting the model for every point in x, we obtain the smoothed line shown in red in Figure 2.16.

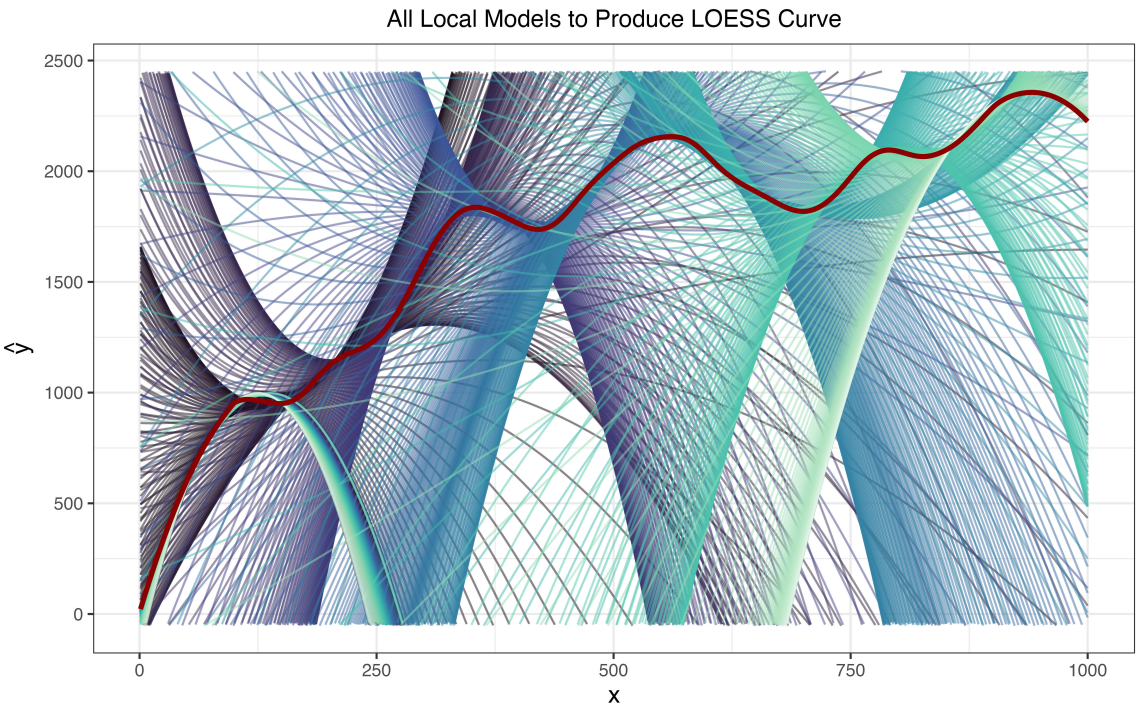


Figure 2.16

Smoothing methods are sensitive to the choice of smoothing parameter h, which represents the fraction of the data that is used for the local polynomial fit.

Methods exist for picking the smoothing parameter h that minimizes the mean squared error between the predicted values from the estimated line and observed values of the dependent variable, for example, leave-one-out cross-validation or generalized cross-validation.

However, for this work, we used LOESS smoothing to smooth survey data from the COVID-19 Trends and Impact Survey (Reinhart et al., 2021). We choose the smoothing parameter for each variable based on domain knowledge regarding the level of noise present for each variable of interest. For example, there is substantial noise in the screening test positivity data that reflect trends that do not represent meaningful differences in the screening test positivity. Some trends in the screening sensitivity may be due to scheduled workplace screenings happening at regular time intervals, and some of the variation may be due to the frequency of screening testing due to other variables, such as the access and cost of testing.

Since the ratio $\frac{\text{screening test positivity}}{\text{overall test positivity}}$ is used to estimate $\beta = \frac{P(\text{test}_+|S_0,\text{untested})}{P(\text{test}_+|tested)}$, the variability in the screening positivity creates substantial variability in our estimates of β.

In light of this variability and the presence of other trends regarding the screening test positivity, we set the span to $\frac{4}{12} = 0.33$ to fit the local regressions for 4-month intervals with the aim to capture the broader trends over time.

There was less variabiity in the smoothing span for the weighted percentage of COVID-like Illness, the estimate of $P(S_1|\text{untested})$. Hence, we set the smoothing

parameter to 0.2 detect trends at a finer time scale.

Sensitivity analyses with modified versions of the smoothing span of β are included in the appendix in the section INCLUDE SECTION.

2.6 Kernel Density Estimation

2.6.1 Overview

When we have a random sample $X_1, \dots X_n$ drawn from the density f_X and we want to estimate f_X at some set of points, we can use kernel density estimation. This is relevant in this work for estimating $f^{induced}$.

We define a kernel function as follows (Wasserman, 2006).

Definition: Kernel Function

A kernel function K is a smooth nonnegative function such that

$$\int K(x)\, dx = 1, \int xK(x)dx = 0, \sigma_k^2 \equiv \int x^2 K(x)dx > 0.$$

The Gaussian kernel $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is commonly used in practice; the tricube kernel, as discussed in the LOESS smoothing section, is another valid kernel function.

The kernel density estimator is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

where h is the smoothing parameter or bandwidth. In Figure 2.17, we see the effect of increasing the bandwidth h: larger values result in smoother curves, while smaller values result in curves that follow the histogram more closely.

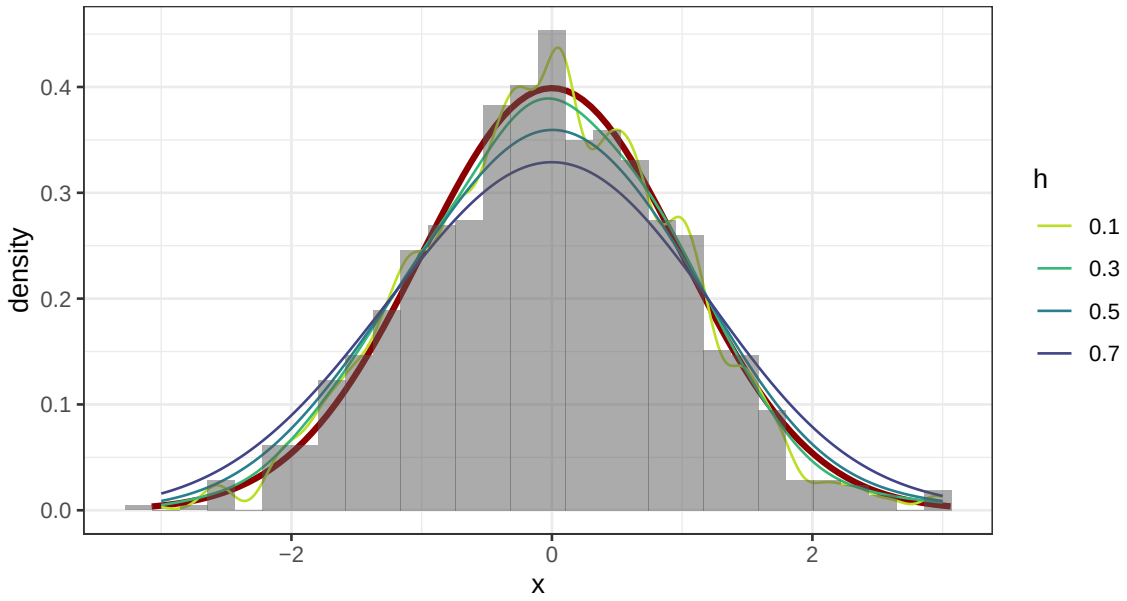


Figure 2.17

2.6.2 Bounded Density Estimation

A question warranting investigation is the choice of kernel given we are working with a bounded variable – the density we seek to estimate, $f^{induced}$ is the density of $P(S_0|\text{untested}, \text{test}_+)$ and hence is bounded between 0 and 1.

One way to handle density estimation for a bounded variable X is by performing a transformation $X = g(Y)$ and then using the change of variables for a probability density to obtain $f_X(x)$ (Aurelien Pelissier, 2022).

Since $X \in [0, 1]$ and we want to transform it to the range $(-\infty, \infty)$, we can let $Y = \text{logit}(X) = \log\left(\frac{X}{1-X}\right)$.

We know if we have $X = g(Y)$, then we can acquire the distribution of X from that of Y by considering the change of variables of the probability density functions f_X and f_y given by

$$f_X(x) = f_Y(g^{-1}(X)) \left|\frac{d}{dx} g^{-1}(X)\right|. \tag{1}$$

Thus, in this case, we have $Y = \text{logit}(X)$, so g^{-1} is the logit function. By definition of the change of variables formula (1), we have

$$f_X(x) = f_Y(\text{logit}(X)) \left|\frac{d}{dx}\text{logit}(X)\right|.$$

Computing the derivative and simplifying, we have

$$\begin{aligned}
&= f_Y(logit(X)) \left|\frac{d}{dx} log(\frac{x}{1-x})\right| \\
&= f_Y(logit(X)) \left|\left(\frac{1-x}{x}\right)(x(1-x)^{-1})'\right| \\
&= f_Y(logit(X)) \left|\left(\frac{1-x}{x}\right)((1-x)^{-1} + x(1-x)^{-2})\right| \\
&= f_Y(logit(X)) \left|\left(\frac{1-x}{x}\right)\left(\frac{(1-x)+x}{(1-x)^2}\right)\right| \\
&= f_Y(logit(X)) \left|\left(\frac{1-x}{x}\right)\left(\frac{1}{(1-x)^2}\right)\right| \\
&= f_Y(logit(X)) \left|\frac{1}{x(1-x)}\right|.
\end{aligned}$$

This means that we compute $Y = logit(X)$ and then estimate the density of the unbounded variable Y, and then we can recover the density f_X by multiplying by $\frac{1}{x(1-x)}$.

In some cases, this approach works well. In Figure 2.18, we simulate a variable $X \sim Beta(3, 2)$ and estimate the density with the transformation approach.

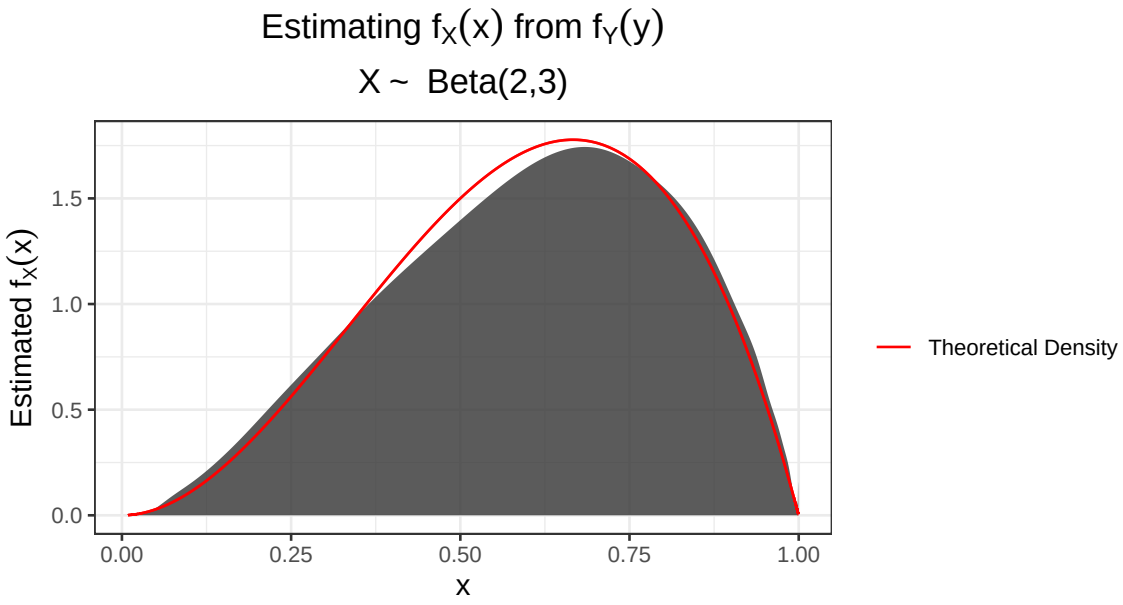


Figure 2.18

We see the difference between using the transformation approach versus estimating the density of X without first transforming it to be unbounded in Figure 2.19.

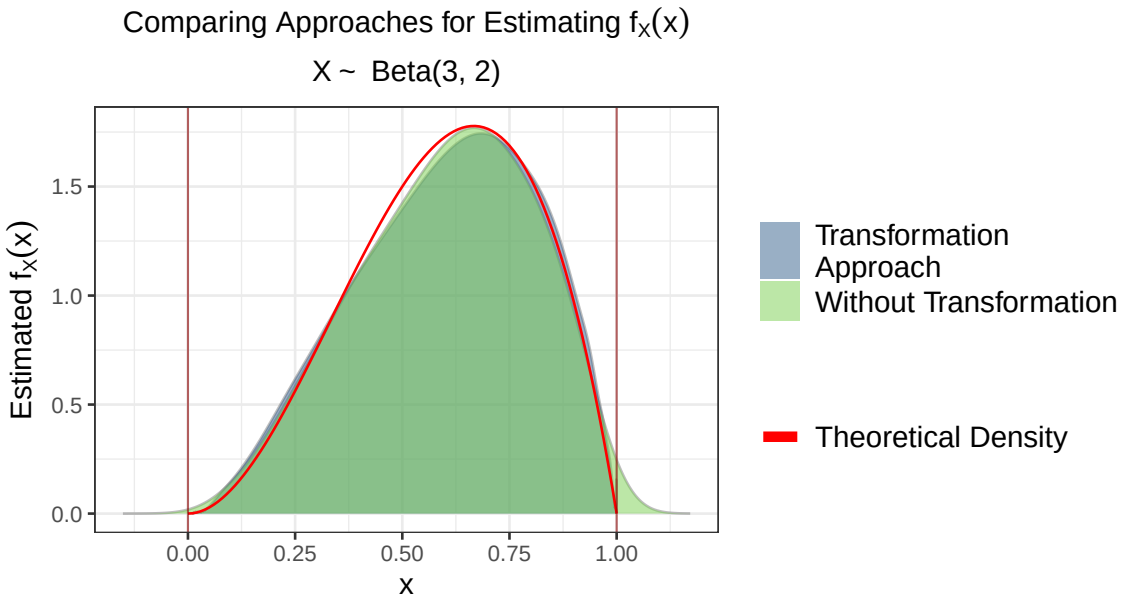


Figure 2.19

However, when we simulate densities that have greater mass toward the boundaries 0 or 1, we see that boundary bias becomes problematic (Figure 2.20). This is evident in panels B, C, D, and G of Figure 2.20, where the estimated density near the boundaries is a poor estimate of the true density.

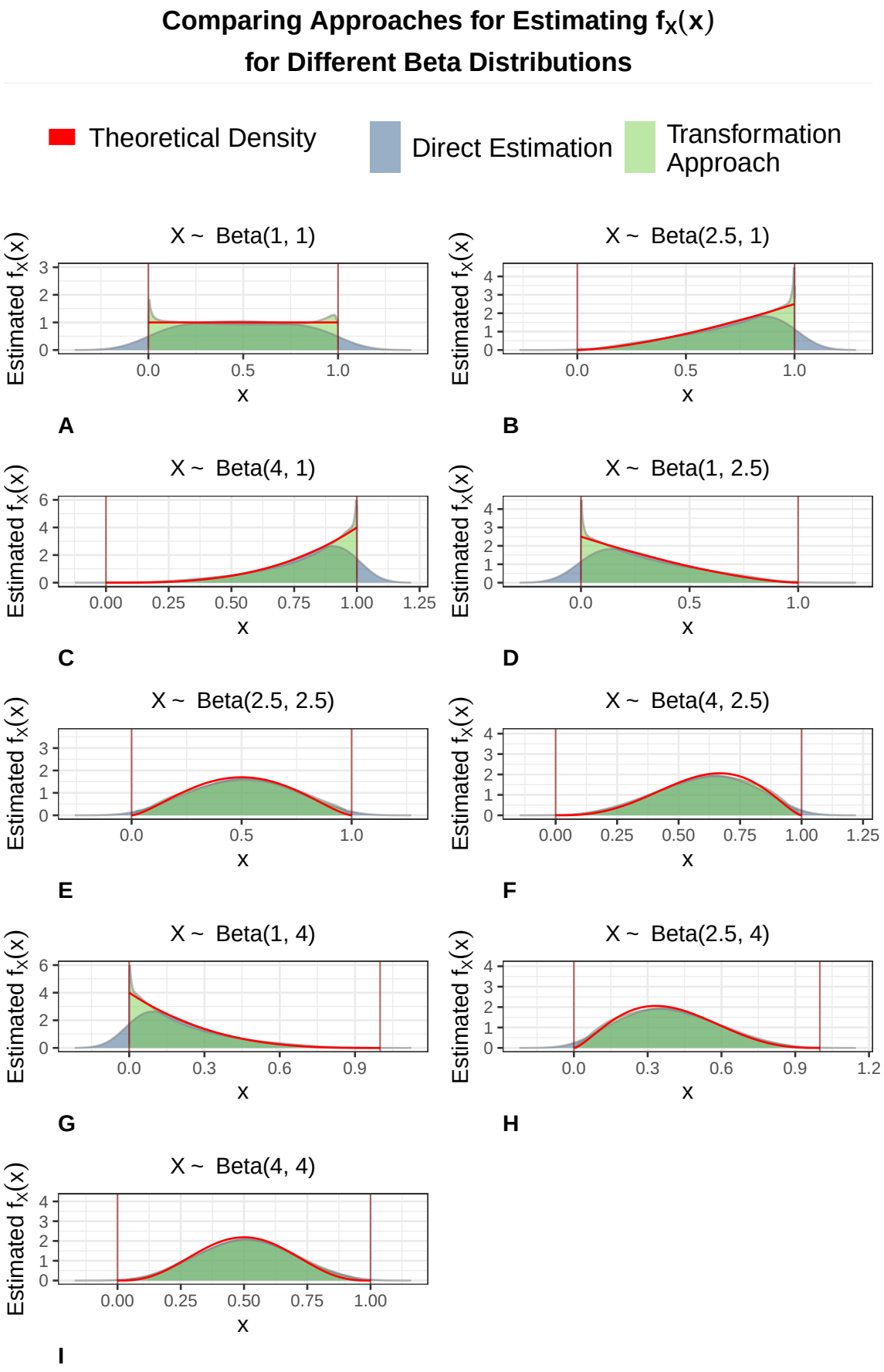

Comparing Approaches for Estimating $f_X(x)$
for Different Beta Distributions
Theoretical Density
Direct Estimation
Transformation Approach
X ~ Beta(1, 1)
X ~ Beta(2.5, 1)
X ~ Beta(4, 1)
X ~ Beta(1, 2.5)
X ~ Beta(2.5, 2.5)
X ~ Beta(4, 2.5)
X ~ Beta(1, 4)
X ~ Beta(2.5, 4)
X ~ Beta(4, 4)
Estimated $f_X(x)$
x
A
B
C
D
E
F
G
H
I

Figure 2.20

An alternative to the transformation approach for density estimation of bounded variables by using beta kernel estimators, which resolves the issue of boundary bias.

As defined in Chen (1999), the most simple beta kernel estimator would be

$$\hat{f}_1(x) = \frac{\sum_{i=1}^{n} K_{x/b+1,\,(1-x)/b+1}(X_i)}{n}$$

where $K_{\text{shape1,shape2}}$ is the density function $Beta(shape1,\ shape2)$.

However, Chen (1999) show that the modified beta kernel estimator $\hat{f}_2(x)$ has lower variance and bias than $\hat{f}_1$, where we define $\hat{f}_2$ as follows:

$$\hat{f}_2(x) = \frac{\sum_{i=1}^{n} K^*_{x,b}(X_i)}{n},$$

$$K^*_{x,b} = \begin{cases} K_{x/b,\,(1-x)/b}(t) & \text{if } x \in [2b, 1-2b] \\ K_{\rho(x),\,(1-x)/b}(t) & \text{if } x \in [0, 2b) \\ K_{x/b,\,\rho(1-x)}(t) & \text{if } x \in (1-2b, 1] \end{cases},$$

$$\rho(x,b) = 2b^2 + 2.5 - \sqrt{b^2 + 6b^2 + 2.25 - x^2 - x/b}.$$

Notably, for beta kernel estimators, the shape of the kernel depends on x (Figure 2.21).

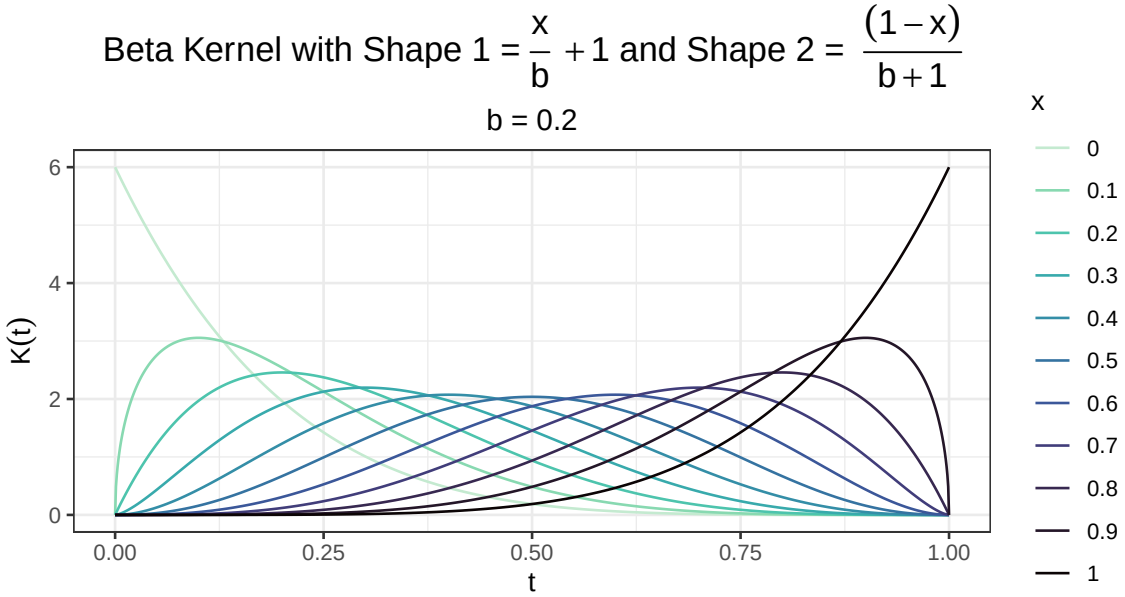


Figure 2.21

As we did in Figure 2.20, we can compare the performance of the beta kernel $\hat{f}_2$ for estimating the density of samples from different beta distributions (Figure 2.22).

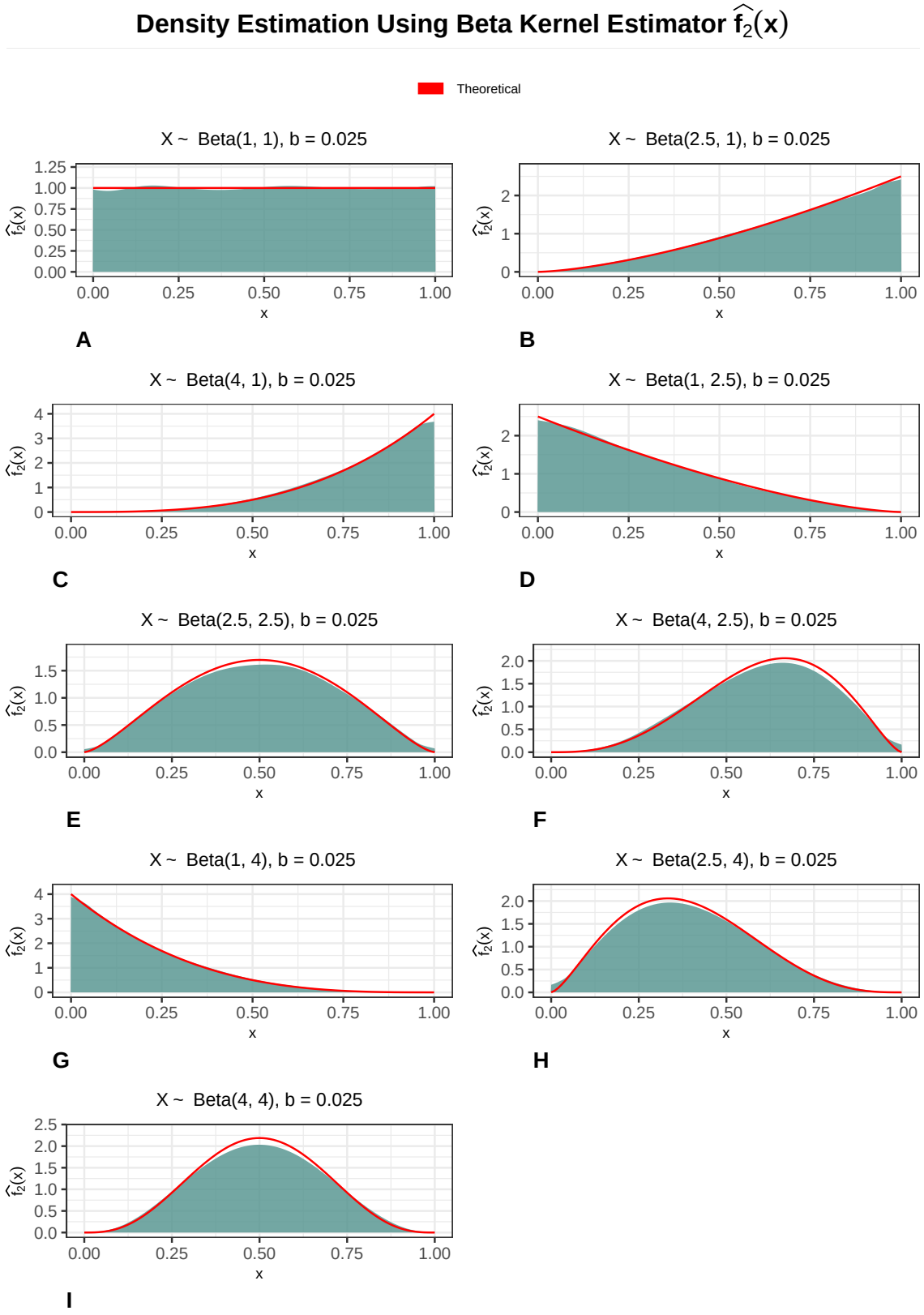

Density Estimation Using Beta Kernel Estimator $\widehat{f}_2(x)$
Theoretical
X ~ Beta(1, 1), b = 0.025
A
X ~ Beta(2.5, 1), b = 0.025
B
X ~ Beta(4, 1), b = 0.025
C
X ~ Beta(1, 2.5), b = 0.025
D
X ~ Beta(2.5, 2.5), b = 0.025
E
X ~ Beta(4, 2.5), b = 0.025
F
X ~ Beta(1, 4), b = 0.025
G
X ~ Beta(2.5, 4), b = 0.025
H
X ~ Beta(4, 4), b = 0.025
I
x
$\widehat{f}_2(x)$

Figure 2.22

Although we can see the advantages of using the beta kernel estimator, in practice, available implementations (e.g., the `bde` package implementation) are much more computationally expensive than using Gaussian kernel density estimation, which make the beta kernel estimator unfeasible for this work. The exact density of the induced distribution is not going to change the estimated counts dramatically relative to the other sources of variability (e.g. the specification of the prior distributions or sources of data to inform these priors), and, as we see in Figure 2.23, Gaussian density estimation does perform reasonably well for a variety of bounded distribution shapes.

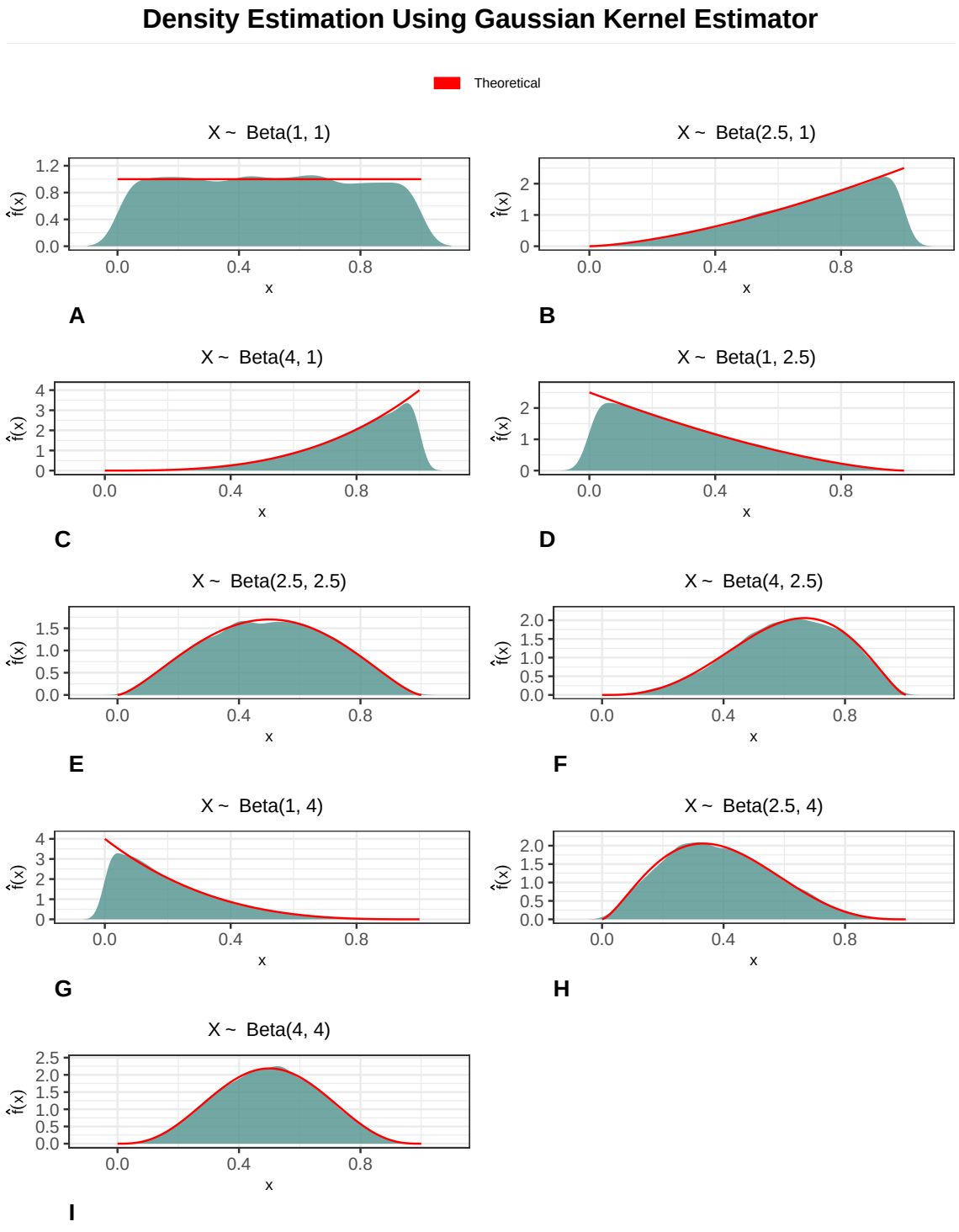

Density Estimation Using Gaussian Kernel Estimator
Theoretical
X ~ Beta(1, 1)
X ~ Beta(2.5, 1)
X ~ Beta(4, 1)
X ~ Beta(1, 2.5)
X ~ Beta(2.5, 2.5)
X ~ Beta(4, 2.5)
X ~ Beta(1, 4)
X ~ Beta(2.5, 4)
X ~ Beta(4, 4)
f̂(x)
x
A
B
C
D
E
F
G
H
I

Figure 2.23

Chapter 3

Definition of Prior Distributions for Bias Parameters

Placeholder

3.1 Background on the Beta Distribution

3.2 Background on the Gamma Distribution

3.3 Definition of Prior Distributions for Incomplete Testing Correction

3.3.1 Defining $P(S_1|\textbf{untested})$

3.3.2 Defining α

3.3.3 Defining β

3.3.4 Defining $P(S_0|\textbf{test}_+, \textbf{untested})$

3.4 Definition of Priors for Test Inaccuracy Correction

3.4.1 Defining Test Sensitivity (S_e)

3.5 Defining Test Specificity (S_p)

3.6 Exploration of the Implications of Changes in the Bias Parameters

3.7 Correction for Incomplete Testing

3.8 Correction for Diagnostic Test Inaccuracy

3.8.1 Derivation of Formula for Correction for Diagnostic Test Inaccuracy

Chapter 4

Details of Implementation

Placeholder

4.1 Version 1: Priors do Not Vary by State or Date

4.2 Version 2-4: Allowing Some Prior Parameters to Vary

Chapter 5

Results

Placeholder

5.1 Comparison to the Covidestim Model

5.1.1 Overview

5.1.2 The Covidestim Model

5.1.3 Assumptions

5.1.4 Comparison to Serological Data

5.1.5 Limitations of this Comparison

5.2 State-level Results

5.3 Relationship Between the Ratio of Estimated to Observed Infections Compared to Testing Rate

5.4 County-level Results

5.4.1 Massachusetts

5.4.2 Michigan

5.5 Cross Correlation Comparison

5.5.1 Background

5.5.2 Cross Correlation Results Comparing Bias Corrected Counts, Covidestim Estimates, and Wastewater Concentrations

Comparison Between Implementations of Probabilistic Bias Analysis

Comparison Between Covidestim, Observed Cases, and Bias Corrected Counts

Takeaways

Chapter 6

Conclusion

The aim of this work is to consider possible scenarios for the extent of unobserved infections over an extended time during the COVID-19 pandemic, and to explore how we can present the uncertainty in the number of incident infections. Throughout the pandemic we often see line charts of observed cases or the test positivity rate. Advice has changed as has testing behavior, with warnings to not consider case counts in isolation, but rather to also look at trends in the test positivity rate (among other indicators). However, presentation of infections as intervals, their widths defined as a direct consequence of assumptions we make about the bias parameters, reflects genuine uncertainty about the number of true infections that may exist in a given area over time. Various models exist to try to get at this quantity of the number of true infections, incorporating a range of sources of data, including COVID-19 deaths, hospitalizations, seroprevalence data, and viral concentrations in wastewater, as well as estimates such as the infection fatality ratio. The strength of applying probabilistic bias analysis to consider possible values of true COVID-19 infections lays in its relative simplicity and transparency of assumptions, in addition to the ease of exploration of possible testing scenarios of the extent to which infections are going undetected. Although there is no ground truth to rigorously assess the accuracy of an estimate of the true number of COVID-19 infections, comparing approaches and understanding where and when they are non concordant provides useful insight into quantifying the range of true infections.

Chapter 7

Appendix

Placeholder

7.1 Smoothing Span

7.1.1 Changing SPAN for LOESS Smoothing of β

7.2 Changing Mean and Variance for Prior Distribution Specifications

7.3 Relationship Between $(X+Y)_\alpha$ and X_α +Y_α for Dependent Variables X, Y

7.3.1 Simulation: Bivariate Normal

7.3.2 Derivation of the Distribution of X+Y for Bivariate Normal

References

Placeholder

Aurelien Pelissier. (2022, February 4). Density Estimation for Bounded Variables. Retrieved March 19, 2023, from https://medium.com/mlearning-ai/density-estimation-for-bounded-variables-7d68f633e772

Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability* (Second edition). Boca Raton: CRC Press.

Carvalho, L. M., Villela, D. A. M., Coelho, F. C., & Bastos, L. S. (2023). Bayesian Inference for the Weights in Logarithmic Pooling. *Bayesian Analysis*, *18*(1). http://doi.org/10.1214/22-BA1311

Chambers, J. M. (Ed.). (1997). *Statistical models in S* (Reprint). London: Chapman & Hall.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, *31*(2), 131–145. http://doi.org/10.1016/S0167-9473(99)00010-9

Genest, C., McConway, K. J., & Schervish, M. J. (1986). Characterization of Externally Bayesian Pooling Operators. *The Annals of Statistics*, *14*(2), 487–501. Retrieved from https://www.jstor.org/stable/2241231

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. http://doi.org/10.1007/s10654-016-0149-3

Lash, T. L., Fox, M. P., & Fink, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York, NY: Springer New York. http://doi.org/10.1007/978-0-387-87959-8

Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., … Liu, M. (2021). Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis. *JAMA Network Open*, *4*(12), e2137257. http://doi.org/10.1001/jamanetworkopen.2021.37257

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *236*(767), 333–380. http://doi.org/10.1098/rsta.1937.0005

Petersen, J. M., Ranker, L. R., Barnard-Mayers, R., MacLehose, R. F., & Fox, M. P. (2021). A systematic review of quantitative bias analysis applied to epi-

demiological research. *International Journal of Epidemiology*, *50*(5), 1708–1730. http://doi.org/10.1093/ije/dyab061

Poole, D., & Raftery, A. E. (2000). Inference for Deterministic Simulation Models: The Bayesian Melding Approach. *Journal of the American Statistical Association*, *95*(452), 1244–1255. http://doi.org/10.1080/01621459.2000.10474324

Powers, K. A., Ghani, A. C., Miller, W. C., Hoffman, I. F., Pettifor, A. E., Kamanga, G., … Cohen, M. S. (2011). The role of acute and early HIV infection in the spread of HIV and implications for transmission prevention strategies in Lilongwe, Malawi: A modelling study. *The Lancet*, *378*(9787), 256–268. http://doi.org/10.1016/S0140-6736(11)60842-8

Reinhart, A., Brooks, L., Jahja, M., Rumack, A., Tang, J., Agrawal, S., … Tibshirani, R. J. (2021). An open repository of real-time COVID-19 indicators. *Proceedings of the National Academy of Sciences*, *118*(51), e2111452118. http://doi.org/10.1073/pnas.2111452118

Robson, B. J. (2014). When do aquatic systems models provide useful predictions, what is changing, and what is next? *Environmental Modelling & Software*, *61*, 287–296. http://doi.org/10.1016/j.envsoft.2014.01.009

Rubin, D. B. (1987). The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, *82*(398), 543. http://doi.org/10.2307/2289460

Rubin, D. B., Gelman, A., & Meng, X.-L. (Eds.). (2004). *Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin's statistical family*. Chichester, West Sussex, England ; Hoboken, NJ: Wiley.

Sah, P., Fitzpatrick, M. C., Zimmer, C. F., Abdollahi, E., Juden-Kelly, L., Moghadas, S. M., … Galvani, A. P. (2021). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, *118*(34), e2109229118. http://doi.org/10.1073/pnas.2109229118

Ševčíková, H., Raftery, A. E., & Waddell, P. A. (2007). Assessing uncertainty in urban simulations using Bayesian melding. *Transportation Research Part B: Methodological*, *41*(6), 652–669. http://doi.org/10.1016/j.trb.2006.11.001

Wasserman, L. (2006). *All of nonparametric statistics*. New York: Springer.