Statistical methods for analyzing deep sequencing data in HIV-1 prevention trial sieve analyses

James Peng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:
Peter B. Gilbert, Chair
Pamela A. Shaw
Linbo Wang

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical methods for analyzing deep sequencing data in HIV-1 prevention trial sieve analyses

James Peng

Chair of the Supervisory Committee:
Peter B. Gilbert
Department of Biostatistics

Understanding how vaccines perform against different pathogen genotypes is crucial for developing effective prevention strategies, particularly for highly genetically diverse pathogens like human immunodeficiency virus-1 (HIV-1). Sieve analysis is a statistical framework used to determine whether a vaccine selectively prevents infection by certain genotypes while allowing breakthrough of other genotypes that evade immune responses. Traditionally, these analyses are conducted with a single sequence available per individual acquiring the pathogen. However, modern deep sequencing technology can provide detailed characterization of within-host viral diversity by capturing up to hundreds of pathogen sequences per person.

In this dissertation, we develop statistical methods for conducting sieve analyses using deep sequencing data. Our objective is to define and study estimands that summarize within-host viral diversity and enable comparisons across treatment groups. When viral variants are characterized by a feature of interest, the within-host viral population induces a distribution over the possible values of that feature. Because only a finite number of sequences is observed for each person, the data provide a sample from this unobserved individual-level distribution. One key issue is that the number of observed sequences, called sequencing depth, can vary across individuals and may be low for some. Throughout, we consider how low and variable sequencing depth affects estimation and inference and propose methods to

address these effects. We show that incorporating deep sequencing data into sieve analyses has the potential to reveal patterns missed by single-sequence analyses by providing a more complete characterization of each individual's viral population.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Peter Gilbert, for sharing with me his broad statistical knowledge and content knowledge, for his time and energy in providing guidance and mentoring me, and for giving me countless opportunities to collaborate with the broader team and share my work at talks and conferences. I am very lucky to have him as an advisor and mentor. It's obvious from his scholarly record that Peter is a brilliant statistician and scientist, but what is more important to me is that Peter is a wonderfully good, compassionate, and generous person; those values are apparent in the work that he does and the way he treats others, and that inspires me to be the same. Thank you, Peter.

I would also like to express immense gratitude to my committee: Pam Shaw, who was always so open to offering her time and energy to help me understand measurement error; Linbo Wang, whose previous papers were the inspiration for many of the ideas in these projects; Amy Willis, who has been so thoughtful in checking in on me and is always encouraging; Thomas Richardson, whose broad and deep knowledge is so valued; and Lillian Cohn, whose expertise in deep sequencing helps bridge the statistics with the application. I am very grateful for my mentors outside of UW – Maya Petersen and Laura Balzer – who inspired me to pursue grad school.

Next, I'm so grateful to my collaborators. In particular, I'd like to thank my collaborator and friend Florian Stijven, a fellow PhD student at KU Leuven who was a visiting student at the Hutch. The collaboration with Florian was the most fruitful academic collaboration that I had in grad school; I admire his independence and desire to understand things fully, and I've learned so much from him. I'm very grateful to Michal Juraska for being generous with his time and energy, for sharing his deep expertise from his years of work in sieve analysis,

and for being an empathetic, patient, and supportive mentor when I was struggling with the research process. I'm also very grateful to my collaborators Craig Magaret, Elena Giorgi, and Allan deCamp, and to the sieve analysis team, for their support and collaboration.

I am grateful for the UW faculty who have dedicated countless hours to sharing their knowledge. I want to thank Minh Vo and the rest of the staff for supporting us through this journey. I also want to thank my cohort – Ethan Ashby, Ameer Dharamshi, Ellen Graham, Jaewon Lim, Katie Paulson, Albert Osom, Taek Son, Rui Wang, and Yinxiang Wu – for helping me get through classes and for always being willing to offer help and support. It's been a pleasure going through this journey with you all.

I'm so grateful for my friends in and outside of Seattle for their support and our connection. Finally, I'm forever grateful and indebted to my parents, Qiyuan and Xuefei; my brother, Ray; my grandma, 外婆; and my late grandpa, 外公, for their love and unwavering support.

Chapter 1
INTRODUCTION

1.1 Background

Since emerging in the early 1980s, human immunodeficiency virus type 1 (HIV-1) has infected over 90 million individuals and claimed 44 million lives globally (World Health Organization, 2025). Although a cure has not been developed, the advent of antiretroviral therapy has changed HIV from a death sentence to a survivable chronic disease and has drastically reduced the risk of transmission by lowering individuals' viral loads (Deeks et al., 2013). In recent years, pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP) have been shown to be extremely effective tools for preventing HIV infection, especially among high-risk populations (Bekker et al., 2022; Ayieko et al., 2022). Both of these tools, however, require regular access to medication and adherence to the prescribed regimen. As of 2025, approximately 1.3 million individuals acquire HIV and 630,000 individuals die from HIV-related causes each year (World Health Organization, 2025).

Given the ongoing substantial public health burden of HIV infection, the development of new, long-lasting, and effective biomedical prevention strategies remains a public health priority. Numerous randomized trials have evaluated interventions to prevent HIV infection, including vaccines and passively administered monoclonal antibodies (Gray et al., 2021; Corey et al., 2021; Gray et al., 2024). In these trials, participants are assigned to an intervention or placebo arm and followed longitudinally with regular HIV testing to identify incident infections. Prevention efficacy is then determined by comparing the rate of HIV infection between the intervention and placebo arms.

One major challenge in developing effective HIV prevention strategies is the substantial genetic diversity of the virus, both across infected individuals and within a single infection.

In fact, within a single individual, the diversity of HIV can match the diversity of influenza viruses circulating across the globe in a single year (Korber et al., 2001). Transmission typically involves a limited number of viral particles that successfully establish infection in a new host, creating a genetic bottleneck (Joseph et al., 2015; Keele et al., 2008; Shaw and Hunter, 2012). Following transmission, rapid viral replication and mutation generate an evolving population of related viral variants within the host. Immune pressures further shape this process, leading to diversification of the viral population over time. The resulting within-host viral population forms a dynamic collection of variants known as a *quasispecies*.

Despite decades of research, no HIV vaccine efficacy trial has demonstrated robust protection against infection. Given the extraordinary genetic diversity of HIV, an intervention that protects against some viral genotypes may be less effective against others, motivating efforts to evaluate how prevention efficacy varies across viral characteristics. To investigate this in prevention trials, viral sequences are obtained from samples drawn at and/or soon after HIV diagnosis from participants who acquire HIV during follow-up. These sequences provide information about the infecting virus and allow investigators to assess whether protection by the intervention differs across viral types. Methods for evaluating genotype-specific prevention efficacy are collectively referred to as *sieve analysis*, drawing an analogy between the intervention and a sieve (Gilbert et al., 1998, 2001). In this analogy, the intervention acts as a barrier that blocks some virus types while allowing others to pass through and establish infection. We see evidence of a *sieve effect* if the intervention efficacy differs across viral genotypes, suggesting that the intervention preferentially prevents infection by certain variants.

To study such effects, researchers must first define features of the virus that may modify efficacy. These features are often summarized by measures of amino acid divergence between the infecting virus and a reference viral sequence (such as a sequence represented inside the vaccine construct in the case of a vaccine intervention). This divergence is typically assessed in biologically relevant regions such as antibody-binding epitopes or other functional sites targeted by treatment-elicited immune responses (Juraska et al., 2024). It may be represented

as a binary or categorical indicator of match or mismatch at specific amino acid residues, or as a continuous measure such as the Hamming distance between the infecting viral sequence and the reference sequence within a defined epitope or viral protein (Figure 1.1a). In sieve analyses, these pathogen features are referred to as *marks*, reflecting that they are observed only among participants who become infected.

Two broad classes of approaches have been used to evaluate sieve effects:

1. *Approach 1 (competing risks / mark-specific efficacy).* After infections are classified according to their observed mark value, these mark-defined infection types are treated as competing risks. Comparing mark-specific infection rates between the intervention and placebo arms yields mark-specific prevention efficacy estimates. Different sieve analysis approaches have been developed for various settings: analyzing categorical marks using competing risks Cox regression (Gilbert et al., 2001), studying continuous marks in proportional hazards models (Sun et al., 2008), analyzing multivariate continuous marks subject to missing values (Juraska and Gilbert, 2016), and cumulative-incidence based estimation (Benkeser et al., 2019), among others (Sun and Gilbert, 2012; Gilbert and Sun, 2015; Follmann and Huang, 2018; Yang et al., 2022; Heng et al., 2025). Such methods have been applied to sieve analyses of HIV (Rolland et al., 2012; Edlefsen et al., 2015; Juraska et al., 2024), malaria (Neafsey et al., 2015), dengue (Juraska et al., 2018), and COVID prevention trials (Magaret et al., 2024).
2. *Approach 2 (case-only comparisons).* Sieve effects are assessed by comparing the distribution of marks among infected participants in the two trial arms. If the intervention preferentially prevents infection by viruses with certain mark values, then the distribution of marks among breakthrough infections in the intervention arm will differ from that in the placebo arm. Although this analysis approach is simpler, it only indirectly addresses the question of interest and is subject to post-randomization selection bias, since viral sequences are observed only among participants who become infected. In addition, this analysis is less interpretable, as it does not quantify the difference in the

vaccine's effectiveness against different pathogen variants. Despite these caveats, these analyses are often presented in complement to Approach 1 (Rolland et al., 2011, 2012; Edlefsen et al., 2015; Neafsey et al., 2015; Hertz et al., 2016; Juraska et al., 2018, 2024; Magaret et al., 2024).

Historically, sieve analyses have relied on Sanger sequencing with single genome amplification, which typically captures only a limited fraction of the viral variants present within an infection (Gregori et al., 2016). Consequently, traditional sieve analyses usually represent each infected participant by a single viral sequence or a single derived mark. Recent advances in *deep viral sequencing* now permit much richer characterization of within-host viral populations by generating up to hundreds of sequences from a single sample, revealing minority variants that would previously have gone undetected.

This shift in sequencing technology fundamentally changes the structure of the data available for sieve analysis. Rather than observing a single sequence mark for each participant who acquires the virus, investigators now observe a sample of sequences drawn from the within-host viral population. Under this framework, binary marks become proportions, categorical marks become compositions, and continuous marks become distributions (Figure 1.1b). Past analyses have typically reduced the sequence set to a single sequence, or the set of mark values to a single mark. For example, one such reduction of a sequence set is a *consensus sequence*, which is obtained by first constructing a sequence with the most frequent residue at each position in the sequence alignment and then finding the closest observed sequence to that constructed sequence. Another example is a reduction of the mark set by taking the modal mark or the mean mark value.

These reductions represent a lost opportunity to examine the data more deeply and may obscure characteristics that are not represented by the dominant viral population. Indeed, the richer data are well-suited to the study of differences between treatment arms that may arise through minority variants. For binary marks, data from a toy example illustrating such a difference is shown in Table 1.1. In this example, while the modal marks are the same in

the two treatment arms, the full sequence sets reveal a small number of vaccine-mismatched viruses in the vaccine arm and none in the placebo arm. Therefore, the intra-individual mark distributions differ between arms even when the modal marks do not. For continuous marks, this could look like the presence of variants with a small proportion of extremely low or extremely high mark values that are observed in individuals in one arm but not the other.

(a)

Subject with infection	AA at position 43 matched to vaccine	AA at position 43	Hamming distance from virus to vaccine
1	1	A	45
2	0	N	79
3	0	G	100

(b)

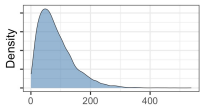
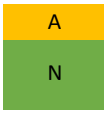
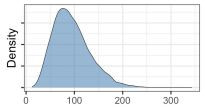
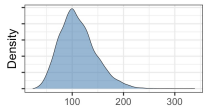
Subject with infection	AA at position 43 matched to vaccine (proportion)	AA at position 43 (composition)	Hamming distance from virus to vaccine (distribution)
1	0.99		
2	0.2		
3	0.05		

Figure 1.1: Sequencing data structure for subjects who are infected in (a) single sequence sieve analyses and (b) multi-sequence sieve analyses.

A further complication is that deep sequencing produces different numbers of sequences across samples, called *sequencing depth*. Sequencing depth determines the resolution with which the within-host viral population can be characterized. It can vary because of the sequencing technology and processing pipeline used, and may also reflect the amount of viral genetic material available in a sample (Raymond et al., 2024). Thus, if an intervention affects post-acquisition viral load, it may induce systematic differences in sequencing depth between study arms. Therefore, ignoring heterogeneity in sequencing depth could lead to bias: samples sequenced more deeply are more likely to reveal rare variants, potentially

Subject ID	Arm	Sanger sequencing	Deep sequencing	
		Sequence match (0) or mismatch (1)	Sequence set match (0) or mismatch (1)	Modal sequence
1	Placebo	0	0, 0, 0, 0, 0	0
2	Placebo	0	0, 0, 0, 0, 0	0
3	Placebo	0	0, 0, 0, 0, 0	0
4	Placebo	0	0, 0, 0, 0, 0	0
5	Placebo	0	0, 0, 0, 0, 0	0
6	Vaccine	0	0, 0, 0, 0, 1	0
7	Vaccine	0	0, 0, 0, 0, 1	0
8	Vaccine	0	0, 0, 0, 0, 1	0
9	Vaccine	0	0, 0, 0, 0, 1	0
10	Vaccine	0	0, 0, 0, 0, 1	0

Table 1.1: Toy dataset illustrating HIV infections in a hypothetical vaccine trial with five infections in each arm. For each individual, Sanger sequencing yields one (or occasionally a small handful of) sequence mark(s), coded as a match (0) or mismatch (1) to the vaccine strain; for simplicity, the table displays a single Sanger-derived mark per individual. In contrast, deep sequencing yields a set of sequence marks per individual. The Sanger-derived sequence marks are identical for infections in the vaccine and placebo arms. However, deep sequencing reveals a small minority of vaccine-mismatched marks that appear only in the vaccine-arm infections. This tail sieve effect is not detectable when only the modal sequence is used.

leading to spurious differences between study arms (Garner, 2011).

Taken together, deep sequencing introduces two fundamental statistical challenges for sieve analysis: (i) representing the distribution of viral variants within participants infected with the virus rather than a single sequence, and (ii) accounting for heterogeneous sequencing depth that affects the probability of detecting rare variants. The goal of this dissertation is to develop statistical methodology for assessing sieve effects using deep sequencing data, with an emphasis on addressing these two challenges.

1.2 Data application

In this dissertation, we will apply our methods to five randomized placebo-controlled HIV-1 prevention efficacy trials: Step/Phambili (HVTN 502/503) (Gray et al., 2010), AMP (HVTN 703/704) (Corey et al., 2021), and Imbokodo (HVTN 705) (Gray et al., 2024). For each of these trials, samples were collected from participants in both treatment arms at the time of virologic HIV-1 diagnosis, and the HIV-1 envelope gene was sequenced. Additionally, for the AMP trials, a second sample was collected 2-4 weeks after diagnosis to examine the change in the virus over time. Table 1.2 displays details on each of the studies.

These trials demonstrate the wide variability in sequencing depth across studies (Figure 1.2). The Step/Phambili studies, which used older Sanger sequencing technology, have very low sequencing depths and distributed similarly across arms (Vaccine: median 5 [range 2–12]; Placebo: 5 [1–12]). In contrast, the AMP studies, which used PacBio deep sequencing, exhibit a wide range of sequencing depths with an overall median of 141.5 sequences per person (range: 1–988), balanced across arms (Treatment: 134 [1–672]; Placebo: 157 [4–988]). Finally, the Imbokodo trial, which also used deep sequencing, demonstrates a scenario with systematic differences between groups. While the overall median is 121 sequences (range: 1–589), vaccine recipients tended to have lower sequencing depths than placebo recipients (Vaccine: 96 [2–493]; Placebo: 146 [1–589]).

Table 1.2: Selected HIV-1 prevention efficacy trials studied in the data applications of this dissertation.

Study (Years)	**Region**	**Intervention**	**Prevention Efficacy (95% CI)**	**N**	**HIV+**	**HIV+ with sequencing**
Step HVTN 502 (2004–2009)	North America, Caribbean, South America, and Australia	MRKAd5 HIV-1 gag/pol/nef subtype B vaccine	−20.0% (−120.0%, 40.0%)	3,000	82	65
Phambili HVTN 503 (2007–2012)	South Africa	MRKAd5 HIV-1 gag/pol/nef subtype B vaccine	−25.0% (−105.0%, 24.0%)	801	62	43
AMP HVTN 703/HPTN 081 (2016–2021)	Sub-Saharan Africa	VRC01 broadly neutralizing monoclonal antibody	8.8% (−45.1%, 42.6%)	1,924	76	74
AMP HVTN 704/HPTN 085 (2016–2021)	Americas and Switzerland	VRC01 broadly neutralizing monoclonal antibody	26.6% (−11.7%, 51.8%)	2,699	98	98
Imbokodo HVTN 705 (2017–2022)	Sub-Saharan Africa	Ad26.Mos4.HIV plus clade C gp140 mosaic vaccine regimen	14.1% (−22.0%, 39.5%)	2,637	119	116

1.3 Data structure and notation

We briefly describe the notation common to the projects in this dissertation. The data come from randomized HIV-1 prevention efficacy trials. Let $i = 1, \ldots, n$ index trial participants. Participants are randomized to either the intervention or placebo arm, with treatment assignment denoted by Z_i. Baseline covariates are denoted by X_i, and a baseline stratum of interest, when applicable, is denoted by S_i.

Participants are followed for a prespecified period and undergo regular HIV-1 testing. Let T_i denote the time of HIV-1 diagnosis and let C_i denote the censoring time, which may

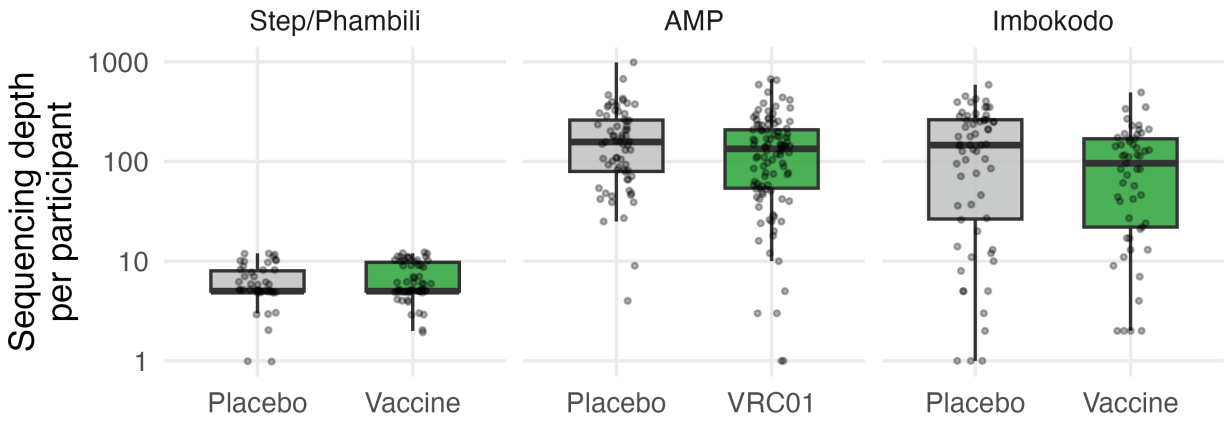


Figure 1.2: HIV-1 Envelope sequencing depth of participants acquiring HIV-1 infection for Step/Phambili (HVTN 502/503), AMP (HVTN 703/704), and Imbokodo (HVTN 705), stratified by study arm. The y-axis is shown on the log10 scale.

occur because of participant dropout or administrative censoring. The observed follow-up time is $\widetilde{T}_i = \min(T_i, C_i)$, and $E_i = \mathbb{1}(T_i \leq C_i)$ indicates whether an HIV-1 diagnosis was observed during follow-up.

For participants with $E_i = 1$, a biological sample is collected and sequenced. Let m_i denote the number of sequences obtained from participant i, which we refer to as the sequencing depth.

With a binary mark, each viral sequence is classified as a 0 or 1. Let Q_i denote the true latent proportion of viruses in the within-host viral population with mark value 1. Let V_{ij} denote the observed mark value for sequence $j = 1, \ldots, m_i$, and let $K_i = \sum_{j=1}^{m_i} V_{ij}$ denote the number of observed sequences with mark value 1. Assuming that the observed sequences are independent draws from the within-host viral population, $K_i \mid m_i, Q_i \sim \text{Binomial}(m_i, Q_i)$. Thus, Q_i represents the underlying proportion of viruses with the mark, whereas K_i/m_i is the corresponding proportion observed.

With a continuous mark, each viral sequence is mapped to a numeric feature. Let Y_i denote the latent distribution of mark values in the within-host viral population of participant

i. For participant i, the observed sequence-level mark values are denoted by Y_{ij} for $j = 1, \ldots, m_i$, where $Y_{ij} \sim Y_i$. Thus, rather than observing Y_i directly, we observe a finite sample of m_i mark values from this distribution.

In the final chapter, two samples are available for each participant. The continuous mark distributions at the two timepoints are denoted by $Y_{1,i}$ and $Y_{2,i}$, the corresponding sequencing depths are denoted by $m_{1,i}$ and $m_{2,i}$, and the time elapsed between the two sample collections is denoted by ℓ_i.

A schematic of notation used in each chapter is presented in Figure 1.3.

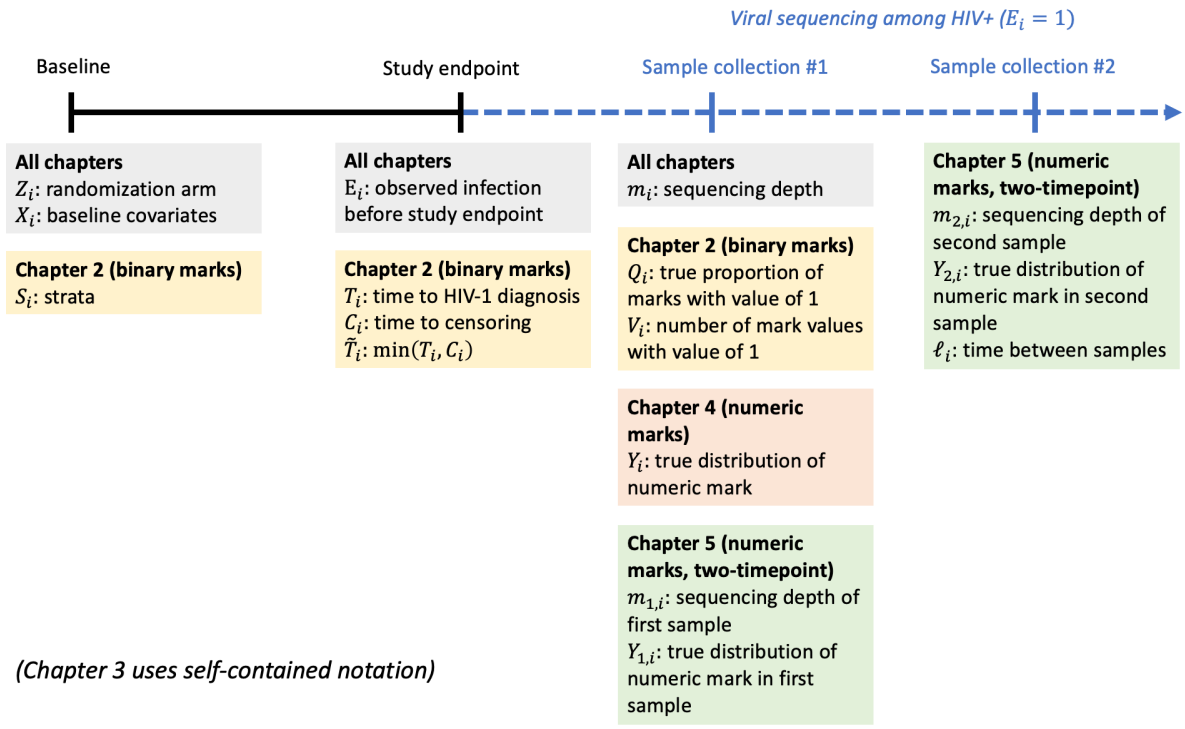


Figure 1.3: Common notation used in all dissertation chapters, except for Chapter 3, which is a self-contained chapter.

1.4 Dissertation organization

This dissertation is organized as follows. In Chapter 2, we discuss a method for estimating vaccine efficacy with respect to binary marks using competing risks Cox regression that corrects for measurement error due to differential sequencing depth. Chapter 3 is a self-contained statistical methods paper that addresses sparse sampling (i.e., low sequencing depth in our application) when estimating Wasserstein barycenters; the method is used in the framework presented in Chapter 4. In that chapter, we present a framework for two (independent) sample comparisons of continuous viral feature distributions arising from deep sequencing data, which can be used to analyze continuous marks in sieve analyses. Chapter 5 describes a method for summarizing and analyzing how distributional change differs between two treatment arms, motivated by data from the AMP trials. Chapter 6 concludes the dissertation, and discusses limitations and directions for future work. In Appendix A, we include a vignette describing the code for `mcbarycenter`, a general-use R package for analyzing Wasserstein barycenter data. Chapter-specific supplementary materials appear in Appendices B–E.

Chapter 2

BINARY FEATURE ANALYSIS USING COMPETING RISKS COX REGRESSION WITH FAILURE TYPE SUBJECT TO MISCLASSIFICATION

James Peng, Michal Juraska, Pamela A. Shaw, Peter B. Gilbert

Chapter Summary

In this chapter, we propose a sieve analysis methodology for *binary marks* to account for intra-individual viral diversity captured through deep sequencing. Our approach estimates vaccine efficacy against viral populations with at least a small threshold of vaccine-mismatched mutations. To account for differential resolution of information from differing sequence counts per person, we use competing risks Cox regression with modeled causes of failure and propose an empirical Bayes approach for the classification model. Simulation studies demonstrate that our approach reduces bias, provides nominal confidence interval coverage, and improves statistical power compared to conventional methods. We apply our method to the HVTN 705 Imbokodo trial, which assessed the efficacy of a heterologous vaccine regimen in preventing HIV-1 infection.

2.1 Introduction

In this chapter, we propose a new approach for estimating sieve effects of binary marks using deep sequencing data. We define an estimand for multi-sequence data that classifies failures according to the presence or absence of a feature within the sequence set while accounting for differences in resolution due to sequencing depth. Estimation is based on a competing risks Cox model with classified failure causes, where an empirical Bayes model provides

participant-specific probabilities of the latent failure type. This formulation treats variable sequencing depth as a measurement error problem involving the failure cause.

Our approach provides an alternative to the competing risks Cox regression method of Van Rompaye et al. (2012) for misclassified failure causes. Whereas their method assumes known and fixed misclassification probabilities, ours estimates participant-specific classification probabilities from sequencing and covariate information. Although empirical Bayes methods are widely used to address measurement error in econometrics (Walters, 2024), they have been less commonly applied in biostatistics and rarely in competing risks settings; see Whittemore (1989) for a related example. We provide theoretical justification and demonstrate good performance in simulations. Related multi-sequence sieve methods include Follmann and Huang (2018), who studied estimands based on pathogen presence and count, and DeCamp (2013), who compared generalized estimating equations with multiple outputation. Unlike these approaches, our method explicitly addresses measurement error caused by variable sequencing depth.

To ground the methodology, we illustrate our approach using deep sequencing data from the Imbokodo/HVTN 705 HIV-1 vaccine efficacy trial (NCT03060629). The trial enrolled females, aged 18–35 years, across five southern African countries and randomized participants 1:1 to receive either placebo or a vaccine regimen comprised of an Ad26-vector mosaic HIV-insert prime followed by a clade C gp140 protein. Although the study did not demonstrate significant efficacy against HIV-1 infection (Gray et al., 2024), viral samples from participants who acquired HIV-1 were deep-sequenced using PacBio technology to characterize within-host diversity in the *env* gene. A full description of the motivating data application is provided in Section 2.5.

2.2 Data structure and estimand

For individual i, denote treatment assignment as $Z_i \in \{0, 1\}$, stratum $S_i \in \{1, ..., L\}$, and other covariates as a p-dimensional vector X_i. Define $W_i := (Z_i, X_i^\top)$. Let T_i be time from randomization until the study endpoint (first HIV-positive sample collection) and C_i be time

to right-censoring. Denote right-censored failure time $\widetilde{T}_i = \min(T_i, C_i)$ and failure indicator $E_i = \mathbb{1}(T_i \leq C_i)$. For participants who experience the study endpoint before censoring, i.e., $E_i = 1$, multiple sequences of the virus are obtained with deep sequencing technology. For each sequence, we consider a binary feature taking values 0 or 1. We observe multiple instances of this feature per individual, which represents the feature's distribution within the viral quasispecies. One example of such a feature, used throughout the following sections, is whether a sequence matches (0) or mismatches (1) a specific residue in a given vaccine-insert virus.

We denote random variable Q_i as the true proportion of sequences circulating in the blood that are mismatched to the vaccine. Note that this variable is not observed. Instead, we observe the proportion among m_i sequences, where m_i is an individual's sequence depth. We denote the binary match/mismatch mark for intra-individual sequences as $V_{i1}, V_{i2}, \ldots, V_{im_i}$, where each $V_{ij} \in \{0, 1\}$ is observed only if $E_i = 1$.

Assumption 2.2.1 (Conditional simple random sampling of sequences)**.** *For each individual i with $E_i = 1$, conditional on $(Q_i, m_i, W_i, S_i, \widetilde{T}_i)$, the observed sequence marks $V_{i1}, V_{i2}, \ldots, V_{im_i}$ are independent and identically distributed as* Bernoulli(Q_i)*. Thus, once the underlying mismatch proportion and sequencing depth are fixed, the remaining observed variables provide no additional information about the sampled sequence marks.*

Let $K_i = \sum_{j=1}^{m_i} V_{ij}$ denote the total number of mismatched sequences observed for individual i. Under Assumption 2.2.1, the conditional distribution of $K_i \mid m_i, Q_i \sim$ Binomial(m_i, Q_i). There are n observations of the data, denoted as $\{X_i, Z_i, S_i, \widetilde{T}_i, E_i, E_iK_i, E_im_i\}_{i=1}^n$.

Similar to the VE_{IF} estimand studied in Follmann and Huang (2018), we wish to measure vaccine efficacy against viral quasispecies with or without some presence of vaccine-mismatched viruses. Formally, we define the mark of interest as $J_{i,q_0} = \mathbb{1}(Q_i \geq q_0)$, indicating that the mismatch proportion is at least a small, fixed threshold q_0. There is no universal standard value for q_0; its choice should be guided by the scientific question, because different thresholds correspond to different definitions of meaningful feature presence. Although

one scientific goal may be to distinguish viral quasispecies with any nonzero presence of the feature, finite sequencing depth prevents reliable classification using an exact zero threshold in practice. The implications of this resolution constraint and considerations for selecting q_0 are discussed in Appendix B.3.

We let $\lambda_{js}(t; z, x)$ denote the covariate-adjusted conditional cause-specific hazard of disease for a viral quasispecies with mark $J_{q_0} = j$ for $j \in \{0, 1\}$:

$$\lambda_{js}(t; z, x) = \lim_{\delta \to 0} \frac{P(T \in [t, t+\delta), J_{q_0} = j \mid T \geq t, Z = z, X = x, S = s)}{\delta} \tag{2.1}$$

That is, $\lambda_{0s}(t; z, x)$ represents the hazard of disease at time t caused by a viral quasispecies with less than a threshold q_0 of mismatched viruses for the $Z = z$ treatment arm with covariates $X = x$ in strata $S = s$, and $\lambda_{1s}(t; z, x)$ represents this cause-specific hazard for a quasispecies with mismatched viruses at least that threshold. Our estimand of interest is the vaccine efficacy against a viral quasispecies with mark $J_{q_0} = j$ for $j \in \{0, 1\}$, denoted as $VE_j(t; x, s)$, which is defined as one minus the mark-specific covariate-adjusted hazard ratio comparing the vaccine and placebo arms:

$$VE_j(t; x, s) = 1 - \frac{\lambda_{js}(t; 1, x)}{\lambda_{js}(t; 0, x)} \tag{2.2}$$

If the value J_{i,q_0} were known for individuals with the study endpoint, we could use a competing risks Cox model to estimate (2.2), where infections by viral quasispecies with and without the presence of vaccine-mismatched viruses ($J_{q_0} = 1$ and $J_{q_0} = 0$) are considered competing failure types (Prentice et al., 1978; Gilbert, 2000). However, for each individual i, J_{i,q_0} is unknown because the true mismatch proportion Q_i is unknown. We could use the observed empirical proportions $\tilde{Q}_i := \frac{K_i}{m_i}$ as a proxy for Q_i and the empirical indicator $\tilde{J}_{i,q_0} := \mathbb{1}(\frac{K_i}{m_i} \geq q_0)$ as a proxy for J_{i,q_0}. However, using this naïve proportion without any correction for measurement error can lead to highly biased results with loss of power in detecting a sieve effect, as shown in our simulation study in Section 2.4.

2.3 Methodology

2.3.1 Competing risks Cox model with modeled failure cause

To account for the fact that we do not observe J_{i,q_0}, we propose a new methodology for competing risks Cox regression with modeled failure cause. Our method will rely on creating a *classification model* of J_{i,q_0} based on its proxies, which include sequencing depth m_i and the number of observed mismatches K_i. We then incorporate these probabilities into our partial likelihood. In order to allow estimation and inference with a Cox model, we make a proportional hazards assumption and a non-informative right censoring assumption:

Assumption 2.3.1 (Proportional hazards)**.** *Treatment assignment Z and covariate vector X have a proportional effect on the hazard for each viral quasispecies type in each stratum. Therefore, for virus type $J_{q_0} = j$ in stratum s, we assume*

$$\lambda_{js}(t; z, x) = \exp\left(\beta_j z + \alpha_j^\top x\right) \lambda_{0,js}(t), \tag{2.3}$$

where $\lambda_{0,js}(t)$ is the stratum- and type-specific baseline hazard, and $\{\beta_0, \beta_1, \alpha_0, \alpha_1\}$ is the vector of regression parameters.

To simplify notation, define $W := (Z, X^\top)^\top$ and $\theta_j := (\beta_j, \alpha_j^\top)^\top$. Then

$$\lambda_{js}(t; w) = \exp\left(\theta_j^\top w\right) \lambda_{0,js}(t).$$

Assumption 2.3.2 (Non-informative right censoring)**.** *Censoring time is independent of failure time and the underlying mismatch proportion conditional on treatment assignment, covariates, and stratum, i.e., $C \perp (T, Q) \mid (W, S)$.*

Under Assumption 2.3.1, our estimand of interest can be written as

$$VE_j(t; x, s) = 1 - e^{\beta_j}, \quad j = 0, 1. \tag{2.4}$$

Equation (2.4) does not depend on time t, covariate vector x, or strata s, so we drop these from the notation for VE from this point forward (e.g. VE_0 and VE_1). If J_{i,q_0} were observed for each individual i, then we can use standard competing risks Cox methodology and construct separate log partial likelihoods, denoted as $\ell_j(\theta_j)$ for $j \in \{0, 1\}$, from the conditional probabilities of an observed event of each type, given one such event was observed in the strata-specific risk set at that time:

$$\ell_j(\theta_j) = \sum_{i=1}^{n} \int_0^\tau \left[\theta_j^\top W_i - \log \left\{ \sum_{l:\, S_\ell = S_i} \mathcal{Y}_l(t)\, \exp\big(\theta_j^\top W_l\big) \right\} \right] dN_{ij}(t) \tag{2.5}$$

where $\mathcal{Y}_i(t) := \mathbb{1}(\widetilde{T}_i \geq t)$ is the at-risk indicator for person i, $N_{ij}(t) := \mathbb{1}(T_i \leq t, E_i = 1, J_{i,q_0} = j)$ is the cause-specific counting process, and τ denote the end of the observation period (any value greater than or equal to the largest observed event time). We can differentiate to obtain the estimating function $U_j(\theta_j)$:

$$U_j(\theta_j) \;=\; \sum_{i=1}^{n} \int_0^\tau \Big\{ W_i - \overline{W}_{j,S_i}(t;\theta_j) \Big\}\, dN_{ij}(t), \tag{2.6}$$

where

$$\overline{W}_{j,s}(t;\theta_j) = \frac{\sum_{l:S_l=s} \mathcal{Y}_l(t)\, W_l\, e^{\theta_j^\top W_l}}{\sum_{l:S_l=s} \mathcal{Y}_l(t)\, e^{\theta_j^\top W_l}}.$$

However, we need to adjust this estimating function to account for the fact that J_{i,q_0} is unobserved for each study endpoint. We define a modified estimating function $U_j^{adj}(\theta_j)$ using the mean score approach, replacing the unknown score term with its expected value given observed variables W_i and $\widetilde{T}_i$ along with auxiliary variables K_i and m_i (Pepe et al., 1994):

$$U_j^{adj}(\theta_j) \;=\; \sum_{i=1}^{n} \int_0^\tau \Big\{ W_i - \overline{W}_{j,S_i}(t;\theta_j) \Big\}\, \nu_{q_0}(j; m_i, K_i, W_i, S_i, \widetilde{T}_i)\, dN_i(t), \tag{2.7}$$

where $\nu_{q_0}(j; m, K, W, S, \widetilde{T}) := P(J_{q_0} = j \mid m, K, W, S, \widetilde{T}, E = 1)$ denotes the classification probabilities that $J_{q_0} = 1$ (i.e., $Q \geq q_0$) or $J_{q_0} = 0$ (i.e., $Q < q_0$) given the observed variables

and $N_i(t) := \mathbb{1}(T_i \leq t, E_i = 1)$ is the counting process for any failure type. Note that, for $E_i = 1$, $\mathbb{E}\Big[dN_{ij}(t) \,\Big|\, m_i, K_i, W_i, S_i, \widetilde{T}_i\Big] = \nu_{q_0}\Big(j; m_i, K_i, W_i, S_i, \widetilde{T}_i\Big)\, dN_i(t)$, so $U_j^{adj}(\theta_j)$ replaces the unobserved cause-specific counting process $dN_{ij}(t)$ with its conditional expectation given the observed variables.

$U_j^{adj}(\theta_j)$ can be seen as a weighted estimating equation, where each event contributes to the equation for each failure type weighted by the probability that the event was that failure type. If ν_{q_0} is known, then we can use standard Cox model theory to show consistency and derive asymptotic variance estimates under the usual regularity conditions (Andersen and Gill, 1982). However, ν_{q_0} will need to be estimated, and we will need to account for the uncertainty in its estimation in the downstream variance estimates.

2.3.2 Classification model ν_{q_0}

Since the classification probabilities $\nu_{q_0}(j; m_i, K_i, W_i, S_i, \widetilde{T}_i)$ are not observed, we will need to estimate them using a model. In this section, we only need to consider estimation of $\nu_{q_0}(1; m_i, K_i, W_i, S_i, \widetilde{T}_i)$, since $\nu_{q_0}(0; m_i, K_i, W_i, S_i, \widetilde{T}_i) = 1 - \nu_{q_0}(1; m_i, K_i, W_i, S_i, \widetilde{T}_i)$. First, by definition, we note that

$$\nu_{q_0}(1; m, K, W, S, \widetilde{T}) = P(Q \geq q_0 \mid m, K, W, S, \widetilde{T}, E = 1) \tag{2.8}$$

$$= \int_{q_0}^{1} f_{Q|m,K}(q \mid m, K, W, S, \widetilde{T}, E = 1)\, dq \tag{2.9}$$

where, with slight abuse of notation, $f_{Q|m,K}(q \mid m, K, W, S, \widetilde{T}, E = 1)$ denotes the conditional density of mismatch proportion Q, given observed variables m, K, W, S, and $\widetilde{T}$ among individuals with $E = 1$. Using Bayes' rule, we can express this density as

$$f_{Q|m,K}(q \mid m, K, W, S, \widetilde{T}, E = 1) = \frac{f_{K|m,Q}(K \mid q, m, W, S, \widetilde{T}, E = 1) \cdot f_{Q|m}(q \mid m, W, S, \widetilde{T}, E = 1)}{f_{K|m}(K \mid m, W, S, \widetilde{T}, E = 1)}. \tag{2.10}$$

Following from Assumption 2.2.1, we have that $f_{K|m,Q}(K \mid q, m, W, S, \widetilde{T}, E = 1)$ corresponds to the probability mass function of a binomial distribution with parameters m and q, which we denote as $f_{\text{binom}}(K; m, q) := \binom{m}{K} q^K (1-q)^{m-K}$. In order to simplify the second term in the numerator $f_{Q|m}(q \mid m, W, S, \widetilde{T}, E = 1)$, we rely on one additional assumption.

Assumption 2.3.3 (Sequence depth conditional independence)**.** *Denote $B := (W, S, \widetilde{T})$. The true mismatch proportion Q is independent of m conditional on B among observed failures, i.e. $Q \perp m | B, E = 1$.*

From Assumption 2.3.3, we have that $f_{Q|m}(q|m, B, E = 1) = f_Q(q|B, E = 1)$. Finally, we can rewrite the denominator as a normalizing constant:

$$f_{Q|m,K}(q|m, K, B, E = 1) = \frac{f_{\text{binom}}(K; m, q) \cdot f_Q(q|B, E = 1)}{\int_0^1 f_{\text{binom}}(K; m, q) \cdot f_Q(q|B, E = 1) dq}. \tag{2.11}$$

The only part of the right hand side that is unknown is $f_Q(q|B, E = 1)$, which represents the *prior (or mixing) distribution* of Q given B for those with $E = 1$. We use parametric binomial mixture model techniques (i.e., parametric deconvolution) to estimate this prior. Indeed, the information we have on K_i mismatches out of the m_i sequences for each individual i for whom $E_i = 1$ can provide information on the true distribution of $Q|B, E = 1$. Under equation (2.11), this corresponds to an empirical Bayes approach, where posterior distributions are calculated using a prior distribution estimated from the data itself (Robbins, 1956).

To enable estimation and inference, we assume that the density of $Q|B = b, E = 1$ arises from a known, correctly specified parametric model:

Assumption 2.3.4 (Parametric model)**.** *The conditional distribution of $Q \mid B = b, E = 1$ is correctly specified by a parametric family $\mathcal{F} = \{f_Q(\cdot; \gamma) : \gamma \in \Gamma\}$. That is, for every b with $\Pr(E = 1 \mid B = b) > 0$, there exists a bin-specific parameter value $\gamma_b \in \Gamma$ such that $Q \mid (B = b, E = 1) \sim f_Q(\cdot; \gamma_b)$.*

While Assumption 2.3.4 may appear restrictive, the spline-based modeling approach with regularization described in the following section allows considerable flexibility in estimating

these distributions (Efron, 2016). If B contains only discrete categorical variables, the density of Q may be estimated separately across levels of B. If B contains continuous variables, the distribution of Q may instead be modeled as a function of B. For simplicity, in the following sections we assume that B contains only discrete categorical variables and estimate the conditional density within each stratum.

In practice, some levels of B may contain too few observations to estimate the distribution of Q reliably. In such settings, stronger conditional independence assumptions may be used to reduce the conditioning set. For example, assuming $Q \perp (m, \widetilde{T}) \mid W, S, E = 1$ yields $f_{Q|m}(q \mid m, W, S, \widetilde{T}, E = 1) = f_Q(q \mid W, S, E = 1)$, so that the conditional distribution of Q need only be estimated across levels of (W, S) rather than $(W, S, \widetilde{T})$. This and other alternatives to Assumption 2.3.3 are discussed in Appendix B.2.

Binomial mixture models

We begin by describing the binomial mixture model setup in the context of our problem. Without loss of generality, fix $B = b$. Our goal is to estimate the conditional density of $Q \mid B = b, E = 1$, where Q has support on $[0, 1]$. Among the total of n observations, suppose the first n' correspond to this subgroup. In the binomial mixture formulation, an unknown distribution of $Q \mid B = b, E = 1$, denoted by G_b with density $g_b(q)$, generates unobserved realizations $\{Q_1, \ldots, Q_{n'}\}$. For each observation $i = 1, \ldots, n'$, we observe pairs (K_i, m_i) satisfying

$$K_i \mid m_i, Q_i \sim \text{Binomial}(m_i, Q_i).$$

The marginal distribution of the observed data $K \mid m, B = b, E = 1$ is thus a binomial mixture with probability mass function

$$f(k \mid m) = \int_0^1 f_{\text{binom}}(k; m, q)\, g_b(q)\, dq. \tag{2.12}$$

Our objective is to estimate mixing density g_b from the observed data. Under Assump-

tion 2.3.4, we assume the mixing density $g_b(q)$ belongs to a parametric family indexed by the parameter vector denoted as γ_b. Thus, from equation (2.12), the marginal likelihood of a single observation $K_i \mid m_i, B_i = b, E_i = 1$ can be written as $f(k_i \mid m_i; \gamma_b) = \int_0^1 f_{\text{binom}}(k_i; m_i, q)\, g(q; \gamma_b)\, dq$. A maximum likelihood estimator (MLE) of γ_b can be calculated as

$$\widehat{\gamma}_b = \arg\max_{\gamma_b} \prod_{i=1}^{n'} f(K_i \mid m_i; \gamma_b) = \arg\max_{\gamma_b} \sum_{i=1}^{n'} \log f(K_i \mid m_i; \gamma_b). \tag{2.13}$$

Although parametric choices such as a Beta mixing distribution provide analytical simplicity, their restricted shapes can be inadequate in practice. We therefore utilize the approach of Efron (2016), which models the mixing density $g_b(q)$ using a low-dimensional exponential family representation based on spline basis functions. Specifically, $g(q; \gamma_b)$ is written as

$$g(q; \gamma_b) = \exp\{\boldsymbol{h}(q)^\top \gamma_b - \phi(\gamma_b)\},$$

where $\boldsymbol{h}(q)$ is a vector of spline basis functions (e.g., natural splines on $[0, 1]$) and $\phi(\gamma_b)$ is the log-normalizing constant ensuring integration to one. The number of spline basis functions controls the smoothness and flexibility of g_b. Point masses at $Q = 0$ and $Q = 1$ may also be included in the mixing model; in this case, the notation $g(q)\, dq$ is understood to include both the continuous component on $(0, 1)$ and the discrete probability masses at the boundaries.

The estimator $\widehat{\gamma}_b$ is obtained via maximum likelihood following the general form in equation (2.13), substituting the spline-based model for g_b. To reduce variability of the estimator, a penalized approach can be utilized:

$$\widehat{\gamma}_b = \arg\max_{\gamma_b} \sum_{i=1}^{n'} \log f(K_i \mid m_i; \gamma_b) - c_0 \|\gamma_b\|^2 \tag{2.14}$$

where $c_0 > 0$ controls the amount of regularization. This penalty shrinks the spline coefficients toward zero, effectively encouraging smoother estimates of $g_b(q)$. This regularization yields lower variance in $\widehat{\gamma}_b$ at the cost of bias but improves numerical stability of the resulting estimates. As discussed in Efron (2016), this framework offers gains in stability

compared to fully nonparametric deconvolution methods while offering more flexibility than rigid parametric priors such as the Beta distribution.

Steps for obtaining classification probabilities

Following from the previous sections, we propose the following procedure for the classification probabilities:

1. Estimate the prior distribution of Q for each level b of B using observations with $E = 1$. For each stratum b, obtain the penalized marginal-likelihood MLE $\hat{\gamma}_b$, and denote the resulting estimated prior density by $g(q; \hat{\gamma}_b)$.
2. For each individual i with $E_i = 1$:
   (a) Following equation (2.11), estimate the individual's posterior density as:

$$\hat{f}_{Q|m,K}(q \mid m_i, K_i, B_i, E_i = 1) = \frac{f_{\text{binom}}(K_i; m_i, q)\, g(q; \hat{\gamma}_{B_i})}{\int_0^1 f_{\text{binom}}(K_i; m_i, q)\, g(q; \hat{\gamma}_{B_i})\, dq}. \tag{2.15}$$

   (b) Following equation (2.9), estimate $\hat{\nu}_{q_0}(1; m_i, K_i, B_i)$ as the probability of $Q \geq q_0$ based on the individual's posterior density:

$$\hat{\nu}_{q_0}(1; m_i, K_i, B_i) = \int_{q_0}^1 \hat{f}_{Q|m,K}(q \mid m_i, K_i, B_i, E_i = 1)\, dq. \tag{2.16}$$

   We then estimate $\hat{\nu}_{q_0}(0; m_i, K_i, B_i) = 1 - \hat{\nu}_{q_0}(1; m_i, K_i, B_i)$.

The estimator for the classification probabilities can be considered as a shrinkage estimator, combining the information from each observation with information from the entire sample. If the information for a given individual's sample is limited (i.e., low sequencing depth), then we rely more heavily on the other individuals in the same stratum of B when estimating the individual's classification probabilities. In contrast, if a given individual has high sequencing depth, then we rely more on the individual's own data when estimating their classification probabilities. We provide more intuition on the connection to shrinkage

estimation through an example in Appendix B.5.

2.3.3 Variance estimation

In summary, Section 2.3.1 presents modifications to the competing risks Cox model that incorporate classified failure types, and Section 2.3.2 describes a procedure to estimate the corresponding classification probability nuisance parameters. Consistency and asymptotic normality of the resulting parameter estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ follow under Assumptions 2.2.1–2.3.4 and standard Cox regularity conditions through the following argument:

1. **Consistency and asymptotic normality of the nuisance parameters.** Under Assumption 2.3.4, the model for the mixing distribution $Q \mid B, E = 1$ parameterized by γ is correctly specified. Standard likelihood theory therefore guarantees that the MLE $\hat{\gamma}$ is consistent and asymptotically normal.
2. **Consistency and asymptotic normality of the Cox parameter estimates.** Under Assumptions 2.2.1 and 2.3.3, the classification probabilities used in the modified Cox estimating equations (equation (2.7)) are deterministic, known functions of γ and the data via equation (2.16). This therefore fits in a standard two–step M-estimation setup, where we first estimate the nuisance parameter γ and then solve the estimating equations that utilize the nuisance parameter. Under Assumption 2.3.1 (proportional hazards) and 2.3.2 (independent right censoring), along with the usual Cox regularity conditions, the resulting estimators $(\hat{\theta}_0, \hat{\theta}_1)$ are consistent and asymptotically normal. The structure parallels the argument in Gao and Tsiatis (2005), which also analyzes estimating equations for a competing risks model that incorporates estimated nuisance parameters.

Because closed-form deconvolution estimators are only available under restrictive parametric assumptions, closed-form analytic variance estimates may be unfeasible in practice. Therefore, we recommend estimating variance through bootstrapping (Efron, 1992). The

resampling procedure can be implemented as follows. For each bootstrap sample: (i) re-estimate the mixing-distribution parameters γ and compute the associated classification probabilities as in Section 2.3.2, and (ii) solve the modified score equations U_j' in Equation (2.7) with these bootstrap-specific classification probabilities substituted, obtaining $\hat{\theta}_0^{*(k)}$ and $\hat{\theta}_1^{*(k)}$. The empirical variance–covariance matrix of the bootstrap replicates

$$\left\{ (\hat{\theta}_0^{*(k)}, \hat{\theta}_1^{*(k)}) : k = 1, \ldots, B \right\}$$

provides a consistent estimator of the sampling covariance of $(\hat{\theta}_0, \hat{\theta}_1)$. These covariance estimates can be used to construct Wald-type or percentile-based confidence intervals.

2.3.4 Hypothesis testing

In addition to the standard Cox regression hypothesis test for each risk type j that type-specific vaccine efficacy equals zero (i.e., $\beta_j = 0$), we consider two additional hypothesis tests. The first test examines whether the vaccine confers any effect against either of the two viral types $J_{q_0} = \{0, 1\}$. Specifically, we test the null hypothesis that vaccine efficacy is zero for *both* risk types:

$$\begin{aligned} H_{A0} &: VE_j = 0 \text{ for } j \in \{0, 1\} \\ H_{A1} &: VE_j \neq 0 \text{ for } j = 0 \text{ or } j = 1. \end{aligned} \tag{2.17}$$

Equivalently, these null hypotheses can be expressed as $H_{A0} : \beta_0 = \beta_1 = 0$ versus $H_{A1} : \beta_0 \neq 0$ or $\beta_1 \neq 0$.

The second test evaluates whether vaccine efficacy varies across the two viral types, thereby testing for a *sieve effect* with respect to the binary mark. The null and alternative hypotheses are:

$$\begin{aligned} H_{B0} &: VE_0 = VE_1 \\ H_{B1} &: VE_0 \neq VE_1 \end{aligned} \tag{2.18}$$

or equivalently, $H_{B0} : \beta_1 - \beta_0 = 0$ versus $H_{B1} : \beta_1 - \beta_0 \neq 0$. This test is analogous to the Lunn–McNeil test for equality of covariate effects in a competing-risks Cox analysis (Lunn

and McNeil, 1995) and serves as the primary test for detecting differential vaccine protection by viral type.

To conduct each test, we first estimate the variance–covariance matrix $\widehat{\Sigma} = \widehat{\text{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)$, obtained empirically from bootstrap replicates of the fitted Cox model. For the first test, we can apply a joint Wald test on the joint null hypothesis that $\beta_0 = 0$ and $\beta_1 = 0$. To do this, we compute the following test statistic:

$$W_{A0} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}^\top \widehat{\Sigma}^{-1} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}. \tag{2.19}$$

We then obtain a p-value of $p = 1 - F_{\chi^2_2}(W_{A0})$, where $F_{\chi^2_2}$ is the cdf of a χ^2_2 distribution with 2 degrees of freedom.

For the second test, we can obtain the following Wald z-statistic:

$$W_{B0} = \frac{\widehat{\beta}_1 - \widehat{\beta}_0}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_1 - \widehat{\beta}_0)}} \tag{2.20}$$

where $\sqrt{\widehat{Var}(\widehat{\beta}_1 - \widehat{\beta}_0)}$ is obtained by applying the delta method on the estimated variance-covariance matrix $\widehat{\Sigma}$. We then obtain a two-sided p-value of $p = 2\Phi(-|W_{B0}|)$ where Φ is the cdf of a standard Normal distribution.

2.4 Simulation study

We study the performance of our proposed method with a numerical study. First, we fix threshold $q_0 = 0.01$. We consider two equal-sized treatment arms: placebo ($Z = 0$) and vaccine ($Z = 1$), with sample sizes in each arm of 1,000, 2,000, and 3,000. We assume a single stratum for all individuals and simulate a binary adjustment covariate X for each individual. We set the probability of the event by $t = 5$ in the placebo arm at roughly 15%. We consider three vaccine efficacy scenarios:

- Setting (a) (no VE): 0% VE for both viral types.

- Setting (b) (positive VE, no sieve effect): 50% VE against both viral types.
- Setting (c) (positive VE with sieve effect): 50% VE against viral type $J_{q_0} = 0$ and 5% VE against viral type $J_{q_0} = 1$.

Details of these exact simulation conditions are shown in Appendix B.6.1. We perform three simulation studies, each with settings (a)–(c), exploring different conditions for simulating sequencing depth m, with K generated as $K|m, Q \sim \text{Binomial}(m, Q)$:

- **Simulation Study #1:** Large, fixed sequencing depth per case with $m = 2000$.
- **Simulation Study #2:** Varying sequencing depth but with equal variation per arm, with m ranging uniformly from 1 to 15 in 40% of endpoint cases, and uniformly from 16 to 1,000 in the remaining endpoint cases.
- **Simulation Study #3:** Unequal sequencing depth across arms. In the placebo arm, m ranges uniformly from 1 to 15 in 20% of endpoint cases and 16 to 1,000 in the remaining cases. In the vaccine arm, m ranges uniformly from 1 to 15 in 40% of endpoint cases and 16 to 1,000 in the remaining cases.

Across all conditions, we compared bias, standard error, and confidence interval coverage for the proposed and uncorrected estimators over 1,000 simulations. The uncorrected estimator used the empirical failure indicator $\widetilde{J}_{i,q_0} = \mathbb{1}(K_i/m_i \geq q_0)$. For the proposed estimator, we assumed the stronger condition $Q \perp (m, \widetilde{T}) \mid Z, X, E = 1$, which holds under the data-generating mechanism. Within each level of (Z, X), we estimated the mixing distribution of Q using penalized splines with $df = 5$ and $c_0 = 0.01$ (Narasimhan and Efron, 2020), augmented with point masses at 0 and 1. Although the spline models are not correctly specified, they provide a flexible and more realistic evaluation than correctly specified parametric models.

Standard errors were estimated using 1,000 bootstrap samples, and Wald confidence intervals were constructed accordingly. We tested for a sieve effect using H_{B0} in Equation (2.18). For the uncorrected estimator, we fit a standard competing-risks Cox model and used the

Lunn–McNeil test for equal covariate effects (Lunn and McNeil, 1995). In settings (a) and (b), where $VE_0 = VE_1$, we compared type I error rates; in setting (c), where $VE_0 \neq VE_1$, we compared power. Appendix B.6 presents additional simulations with correctly specified mixing distributions and varying choices of q_0.

2.4.1 Simulation Study #1

Results for Simulation Study #1 are presented in Appendix B.6.2. Both the proposed and uncorrected estimators perform well, as expected, because misclassification error is rare with large sequencing depths.

2.4.2 Simulation Study #2

Results for Simulation Study #2 are shown in Figure 2.1. When sequencing depth varied similarly between treatment arms, both estimators performed well in settings (a) and (b), where vaccine efficacy was equal across failure types. Although misclassification occurred primarily from $j = 1$ to $j = 0$, it was approximately balanced between arms, yielding minimal bias, near-nominal coverage, and controlled type I error. In setting (c), however, the uncorrected estimator was substantially biased for the $j = 0$ failure type and exhibited confidence interval undercoverage, whereas the proposed estimator remained approximately unbiased, maintained near-nominal coverage, and had greater power to detect the sieve effect.

2.4.3 Simulation Study #3

Results for Simulation Study #3 are shown in Figure 2.2. When sequencing depths differed between treatment arms, differential misclassification led to biased vaccine efficacy estimates and substantial type I error inflation for the uncorrected estimator in settings (a) and (b), with inflation increasing with sample size. The proposed estimator provided much better type I error control, although very mild type 1 error inflation remained because the spline models did not perfectly represent the true mixing distributions. In setting (c), the un-

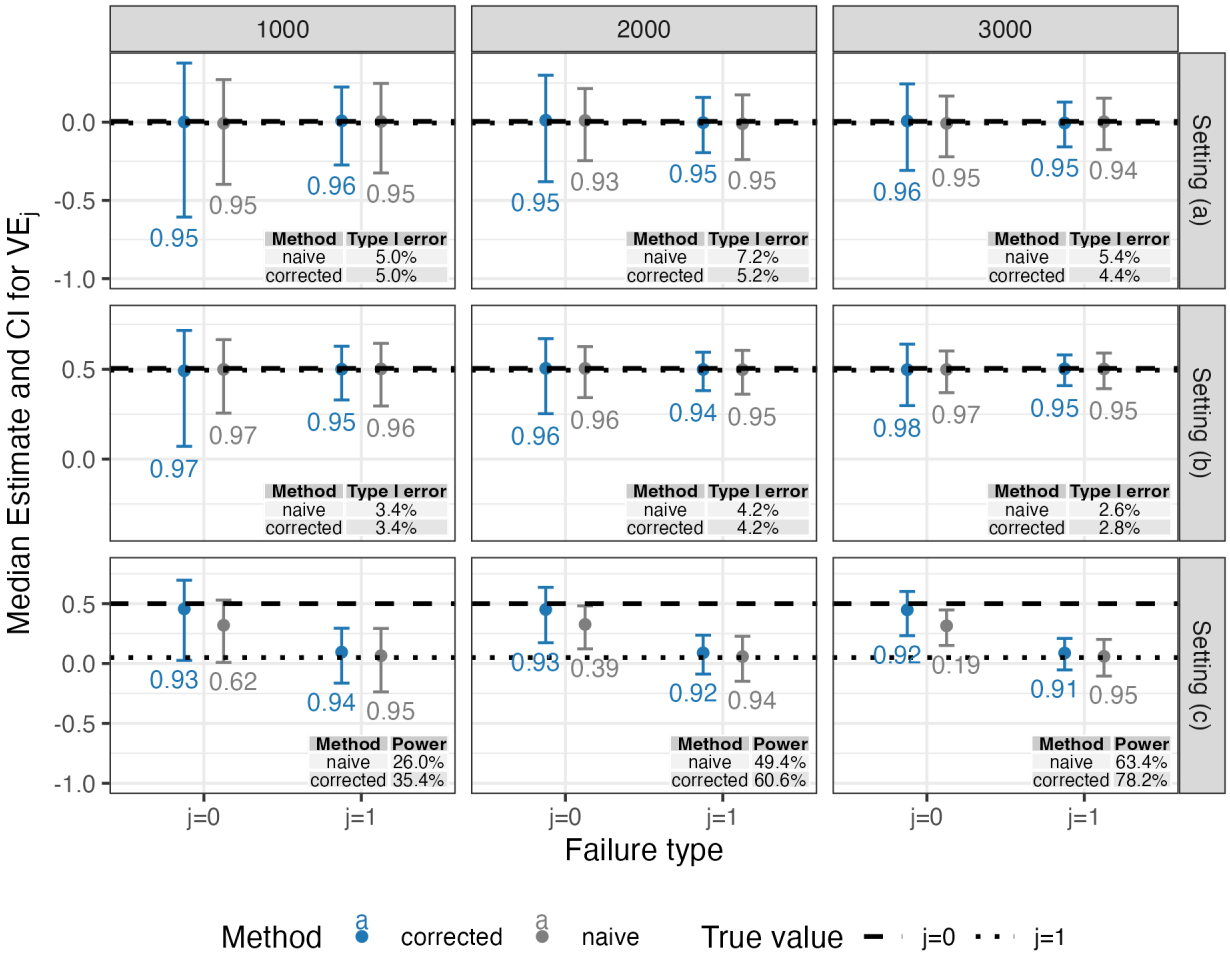


Figure 2.1: Results for Simulation Study #2. The top, middle, and bottom panels display results for settings (a), (b), and (c) respectively. The left, middle, and right panels display results for differing sample sizes per treatment arm. Each graph displays points for median VE estimates for each failure type across all simulations, along with median lower and upper 95% confidence interval bounds. The dashed and dotted horizontal lines are placed at the true values for the $j = 0$ and $j = 1$ failure types, respectively. The numbers below each error bar display the confidence interval coverage. We compare Type I error between the two estimators in settings (a) and (b) and power to detect a sieve effect in setting (c).

corrected estimator had power close to 5% across all sample sizes, whereas power for the proposed estimator increased from 38.4% to 87.2% as the sample size increased from 1,000 to 3,000 participants per arm.

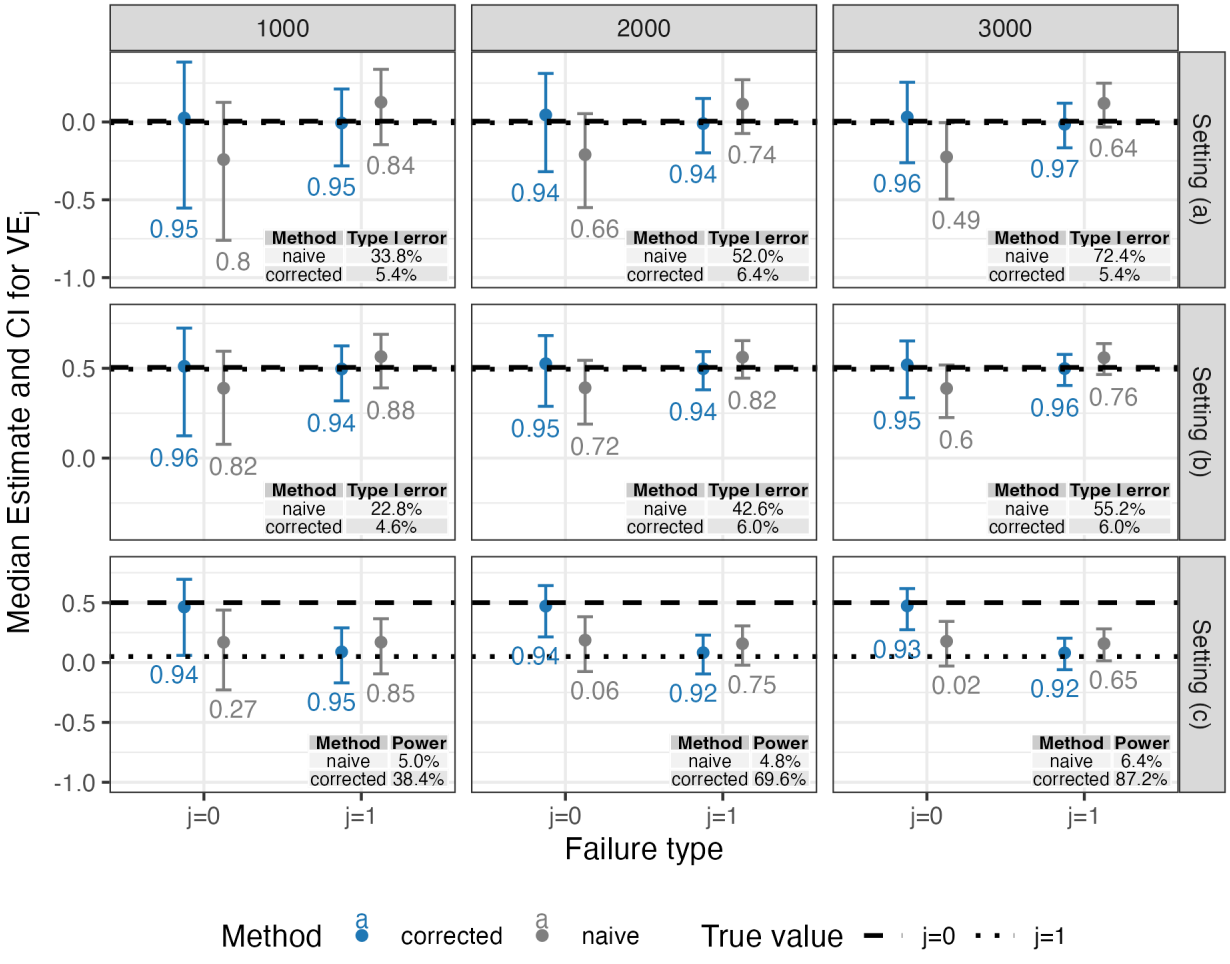


Figure 2.2: Results for Simulation Study #3. The top, middle, and bottom panels display results for settings (a), (b), and (c) respectively. The left, middle, and right panels display results for differing sample sizes per treatment arm. Each graph displays points for median VE estimates for each failure type across all simulations, along with median lower and upper 95% confidence interval bounds. The dashed and dotted horizontal lines are placed at the true values for the $j = 0$ and $j = 1$ failure types, respectively. The numbers below each error bar display the confidence interval coverage. We compare Type I error between the two estimators in settings (a) and (b) and power to detect a sieve effect in setting (c).

2.5 Data application

We applied our proposed methodology to the HVTN 705 HIV-1 vaccine efficacy trial. Two classes of binary sequence features were analyzed: (1) reference-dependent marks indicating whether the amino acid residue in a given sequence position matches or mismatches that in each of three HIV-1 Env vaccine insert sequences (Mos1 and Mos2S in the Ad26-vector prime, and C97ZA in the Clade C boost), and (2) reference-independent amino acid indicators denoting the presence of a specific amino acid at a given position (using the HXB2 numbering system). The median sequencing depth across all samples was 121 (IQR: 23.5–223.0), with lower depths in the vaccine arm (median [IQR]: 96 [22.0–169.0]) than in the placebo arm (146 [26.5–262.5]). Based on these depths, following the guidance in Appendix B.3, we set thresholds of 1% and 99%. Therefore, for each feature, the resulting analyses estimate vaccine efficacy against quasispecies with $< 1\%$ (or $< 99\%$) presence of the feature compared with those with $\geq 1\%$ (or $\geq 99\%$) presence. Age, BMI, and HIV risk score were included as adjustment covariates X, with stratification (S) by the indicator of South Africa. The mixing distributions in the classification model (conditioning on vaccination status only) were estimated using spline-based mixing distributions ($df = 10, c_0 = 1$). For comparison with conventional sieve analyses, we also report results based on the modal sequence mark. Although this benchmark targets a different estimand, it illustrates the conclusions that would be obtained from a single-sequence analysis.

Because the data had thousands of potential sequence features, we perform a pre-screening procedure to reduce the starting sequence set, described in Appendix B.4. Among the 8,458 candidate binary features, 114 feature-threshold combinations passed screening: 71 at the 1% threshold and 43 at the 99% threshold. We identified evidence of differential VE for amino acid leucine at HXB2 position 832 in the envelope protein under the 99% threshold (unadjusted $p < 0.001$; FWER-adjusted $p = 0.023$). Quasispecies with $\geq 99\%$ leucine at this position exhibited high vaccine efficacy (VE estimate: 91.7%; 95% CI: 67.4, 97.9), whereas those with $< 99\%$ leucine showed no evidence of protection (VE estimate: -7.0%; 95% CI: -

55.5, 26.4) (Figure 2.3a). Results from a single-sequence sieve analysis using the naïve $\geq 99\%$ classification (unadjusted $p = 0.027$) and the modal sequence mark (unadjusted $p = 0.086$) showed a similar trend in differential VE but had higher p-values. Figure 2.3b shows the raw data for this viral feature. As shown, the vaccine arm had a lower raw proportion of viral quasispecies exceeding the 99% leucine threshold at this position, and cases in the vaccine arm meeting this threshold tended to have lower sequencing depths. Our method corrects for the possible misclassification into the $\geq 99\%$ bin due to low sequencing depth.

This result suggests that even a presence of a minority of variants with a non-leucine residue at that position may be capable of undermining vaccine protection. One possibility is that substitutions at this position alter the amount or conformation of surface-expressed Env, thereby reducing humoral recognition of infected cells. Alternatively, the lower prevalence of L832 among vaccine recipients may reflect early post-infection adaptation within the LLP-1 motif to evade vaccine-induced immune pressure, consistent with prior reports that adjacent LLP-1 residues participate in mutational escape networks and affect replication, infectivity, and virion maturation.

2.6 Discussion

Our proposed methodology provides a principled framework for evaluating sieve effects in prevention trials that have deep viral sequencing data. The approach is particularly well suited for rapidly evolving viruses that generate substantial within-host diversity, such as HIV, where analyses based solely on a modal or mindist sequence may miss important intra-individual variation. We believe our work has two main innovations in the context of sieve analysis literature: (1) the definition of an estimand in the context of deep sequencing data, and (2) an estimation procedure that explicitly addresses heterogeneous sequencing depth, which is a key feature of the data. In contrast with naïve approaches that ignore measurement error or exclude participants with low sequencing depth, our method retains all available information and explicitly corrects for sequence depth-related measurement error.

Our framework addresses settings in which the true failure type is not directly observed

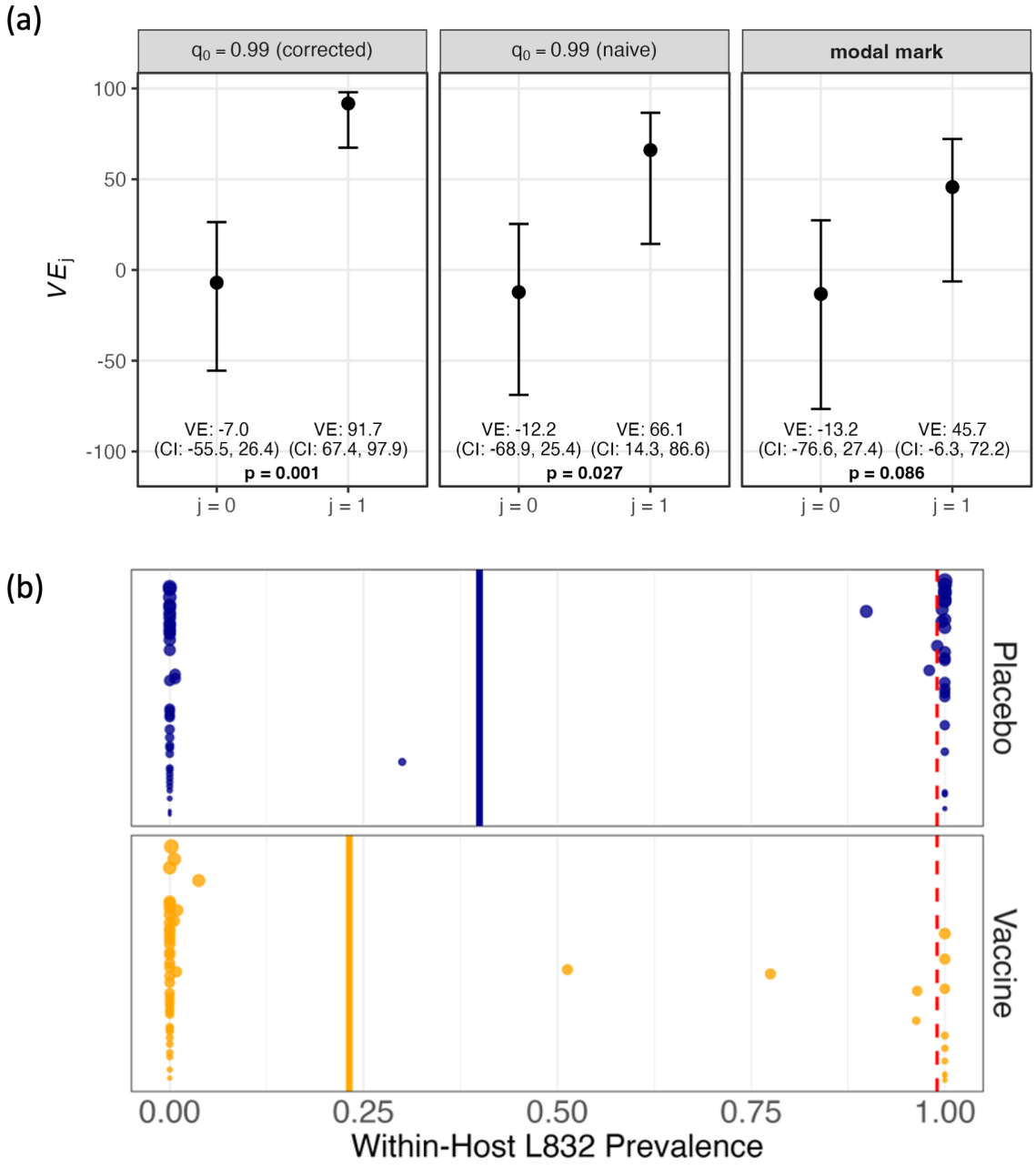


Figure 2.3: (a) Results for the sieve analysis of leucine at HXB2 position 832 comparing the proposed estimator, uncorrected estimator using a raw $\geq 99\%$ classification and uncorrected estimator using the modal mark (i.e., raw $\geq 50\%$ classification). (b) For each individual, the plot shows the raw proportion of the individual's sequences with leucine at HXB2 position 832. The size of the dot is proportional to the sequencing depth for that individual. The dots are ordered vertically by sequencing depth. The solid lines show the average proportion for each arm.

and must instead be inferred from error-prone sequencing data. Unlike prior methods that assume known and fixed misclassification probabilities, our approach allows these probabilities to vary across cases and estimates them from the sequencing data. The empirical Bayes posterior combines individual-level information from the number of mismatches K_i and sequencing depth m_i with information from the estimated mixing distribution. At high sequencing depth, classification relies primarily on the individual's observed sequences; at low sequencing depth, it shrinks toward the population distribution. More broadly, the method illustrates how empirical Bayes approaches can combine subject-specific and population-level information to address measurement error in nontraditional data settings.

One limitation of our proposed approach concerns the modeling assumptions required to construct the classification model. The method requires estimation of the binomial mixing distribution of Q conditional on some combination of $(W, S, \widetilde{T})$. When the covariate vector is high-dimensional, estimation of this conditional distribution may become difficult and may limit the scalability of the approach. One possible simplification is to reduce the conditioning set used in the classification model. However, smaller conditioning sets require stronger conditional independence assumptions to ensure that the resulting posterior classification probabilities remain valid, and these assumptions require appropriate scientific justification; see Appendix B.2 for further discussion. Finally, valid inference depends on correct specification of the mixing distributions of Q. Penalized splines provide a flexible modeling strategy and performed well in our simulation studies, but misspecification may nevertheless lead to biased estimates and invalid confidence intervals or p-values.

Another limitation is that our method uses threshold-based dichotomization of the underlying viral features, reflecting the empirical reality that the observed proportions are often concentrated near 0 or 1. An alternative approach would treat these proportions as continuous features and incorporate a continuous measurement error model to account for variable sequencing depth.

Acknowledgements

We thank the reviewer and Associate Editor of Biometrics for their helpful comments; Craig Magaret for constructing the data set; Elena Giorgi, Paul Edlefsen, Raabya Rossenkhan, Allan deCamp, and James Ludwig for helpful discussions; and the participants and investigators of HVTN 705. Research reported in this publication was supported by the National Institutes of Health under award number R37AI054165, and the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health (NIH) under the U.S. Public Health Service Grant AI068635.

Chapter 3

ESTIMATING THE WASSERSTEIN BARYCENTER OF ONE-DIMENSIONAL DISTRIBUTIONS UNDER SPARSE SAMPLING

James Peng, Florian Stijven, Linbo Wang, Peter B. Gilbert

This chapter presents a method that will be applied in the framework in Chapter 4. The setup and notation in this chapter are self-contained.

Chapter Summary

We study distributional data under sparse sampling where each unit is represented by a probability distribution on the real line observed only through a small i.i.d. sample. A natural notion of central tendency for one-dimensional distributional data is the Wasserstein barycenter, whose quantile function is the pointwise average of the unit-level quantile functions. We focus on pointwise estimation of the Wasserstein barycenter quantile function: at a given quantile level, the target is the population mean of the corresponding unit-level quantiles. A naïve plug-in estimator is the empirical Wasserstein barycenter, which treats observed unit-level empirical distributions as the true latent unit-level distributions. Under sparse sampling, however, this estimator can be severely biased. We propose an approach that avoids directly estimating either the unit-level distributions or the full population law of distributions. We start with the more ambitious goal of characterizing the distribution of latent unit-level quantiles at a given quantile level. We show that this distribution can be written in terms of the marginal distributions of the unit-level CDF values, which can be estimated using binomial mixture methods. This motivates our estimator, the marginal-constructed barycenter (MCB) estimator, obtained by taking the mean of this estimated distribution

of latent unit-level quantiles. We establish conditions under which the MCB estimator is pointwise consistent and asymptotically normal, and show through simulations that it can substantially outperform the empirical Wasserstein barycenter under sparse sampling. We illustrate the method in an analysis of HIV-1 sequence data from the HVTN 502/503 vaccine efficacy trials, using the barycenter to summarize and compare within-participant distributions of viral sequence features when only a small number of sequences are available per participant.

3.1 Introduction

We study the statistical analysis of distributional data, where the units of analysis are distributions on the real line. Our primary motivation comes from HIV-1 studies, where the biological object of interest is not a single homogeneous sample but a heterogeneous population of viral variants within an individual with HIV-1. These studies often summarize each sample by a single consensus sequence or a small number of dominant variants (Edlefsen et al., 2015; deCamp et al., 2017; Juraska et al., 2024). With modern sequencing technologies, however, it is possible to obtain multiple viral sequences from the same individual, giving a more detailed view of the within-host viral population (Gregori et al., 2016). HIV-1 exhibits extensive genetic diversity within an individual, forming a quasispecies of continually evolving viral variants. When a one-dimensional numerical feature, such as a genetic distance or physicochemical property, is computed for each sequence, the resulting collection of values can be viewed as a sample from an individual-specific distribution. This distribution captures the heterogeneity of the within-host viral population.

A common difficulty in this setting is that, despite these technological advances, the number of sequences observed for each individual can be small. In some HIV-1 sequencing studies, there may be fewer than 10 sequences per individual (Mullins et al., 2025). Thus, the scientific target is distributional, but each individual distribution might be observed only through a *sparse sample*. This differs from better-studied examples of distributional data, such as wearable device studies, where accelerometers measure physical activity intensity at

high frequency over multiple days. In that setting, the dense stream of measurements can often be regarded as repeated draws from an underlying subject-specific distribution, where the distribution characterizes the relative amount of time the individual spends at each intensity level (Augustin et al., 2017; Ghosal et al., 2023). The number of accelerometer readings per individual typically is large, so each empirical distribution may be a reliable representation of the true distribution. In contrast, in the HIV-1 setting, the empirical distribution for individuals with low numbers of sequences may be a poor representation of their true individual distribution. More broadly, the challenge of small within-unit sample sizes is not unique to sequencing studies: it can arise from resource constraints or from applications in which each unit naturally contains only a potentially small number of subunits. Additional data examples are discussed in Appendix C.3. Therefore, in this chapter, we focus on this sparse sampling setting, where each unit-level distribution is observed through only a small number of samples.

A key task in the analysis of distributional data is to summarize a collection of distributions by a single representative distribution. In the HIV-1 application, such a summary could describe the average pattern of within-host viral heterogeneity across a population or within a clinically relevant subgroup. This notion of an average depends on the geometry of the space in which the objects are compared. In this paper, we focus on the Wasserstein barycenter, defined as the Fréchet mean of a collection of distributions under the Wasserstein metric (Agueh and Carlier, 2011). Compared with pointwise Euclidean averaging of the distributions, the Wasserstein barycenter better preserves salient features of the component distributions (Bigot et al., 2018; Panaretos and Zemel, 2020). For example, the barycenter of bimodal distributions of the same shape is itself bimodal, whereas the pointwise mean of the densities does not preserve that shape (Figure 3.1). Moreover, for one-dimensional distributions, the barycenter has a simple and interpretable form: its quantile function is the mean of the unit-level quantile functions (Bigot et al., 2017). Since the quantile function fully determines the distribution, we focus on estimating the quantile function of the barycenter.

A simple approach to estimating the barycenter is to treat the observed empirical distri-

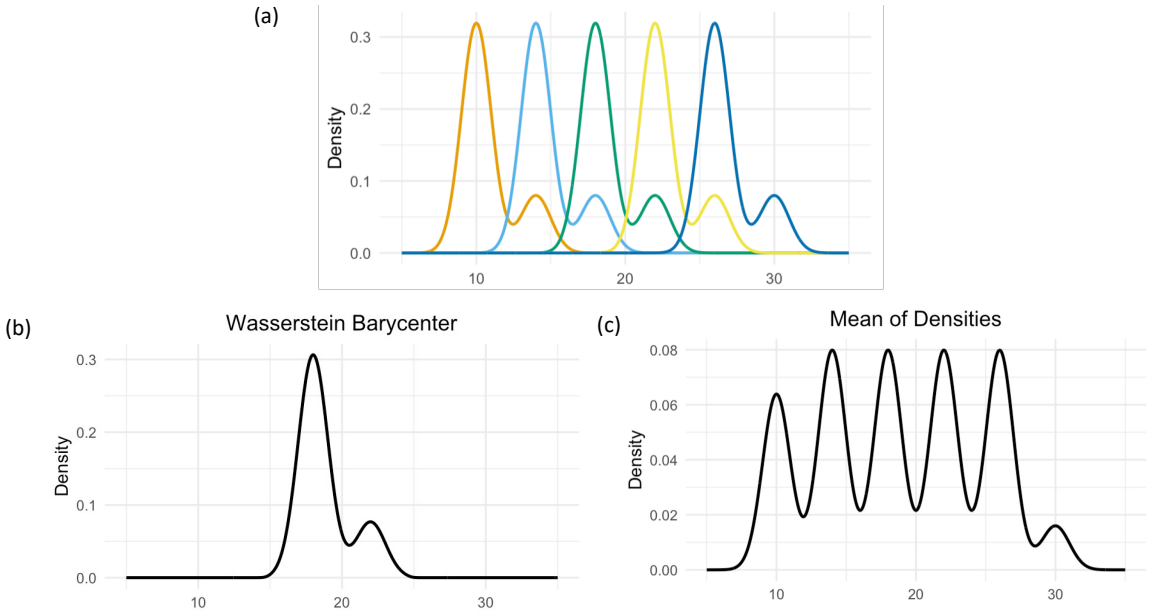


Figure 3.1: (a) The density of five bimodal distributions, (b) the Wasserstein barycenter of the five distributions, (c) the distribution obtained by taking the pointwise mean of the five density functions.

butions as if they were the true unit-level distributions and then compute the corresponding average quantile function. This yields the *empirical Wasserstein barycenter*. Under sparse sampling, however, this estimator can be substantially biased. This is because the barycenter is calculated by averaging unit-level quantile functions, and empirical quantiles, unlike the empirical CDF, are generally biased when observed through only a small sample (Bobkov and Ledoux, 2019). In particular, they are well known to generally be downward biased at upper tail quantiles, and upward biased at lower tail quantiles. Figure 3.2 illustrates the problem in a simulation where 1,000 distributions are drawn from the 5 distributions in Figure 3.1a and only 7 i.i.d. samples are observed from each. The empirical barycenter deviates substantially from the truth at most quantiles, and particularly in the tail quantile levels. Nevertheless, there is hope in that even though the unit-level quantile functions are poorly measured, the observed data still contain information across many independent units. This cross-unit information could be leveraged to recover the barycenter.

Motivated by this problem, we study barycenter estimation under sparse sampling. We make five main contributions. First, we show that when the number of observations per unit is bounded, the barycenter is generally not identifiable from the observed data distribution without additional structure. Second, we derive a representation result, which we call the *marginal construction*. The key idea is to begin with the seemingly more ambitious task of characterizing the distribution of the unobserved unit-level quantiles, from which the barycenter quantile is recovered as the mean. We show that the distribution of unit-level quantiles can be expressed in terms of the marginal distributions of unit-level CDF values. The latter distributions arise as mixing distributions in binomial mixture models. Because the binomial mixture problem is well studied, this representation suggests how structural assumptions on these mixing distributions can restore identifiability of the barycenter. Third, motivated by this representation, we propose the *marginal-constructed barycenter* (MCB) estimator, which estimates the barycenter's quantile function by first estimating the relevant binomial mixing distributions and then plugging those estimates into the representation formula. Fourth, under suitable structural conditions, we show that the MCB estimator's

quantile function is pointwise consistent and asymptotically normal. We then complement these theoretical results with simulations showing that the MCB estimator can outperform the empirical barycenter even when these conditions are not satisfied; Figure 3.2 previews its performance in the simulation described above. Fifth, we show that our framework extends to estimation of several other related distributional estimands, such as weighted barycenters and Fréchet variance.

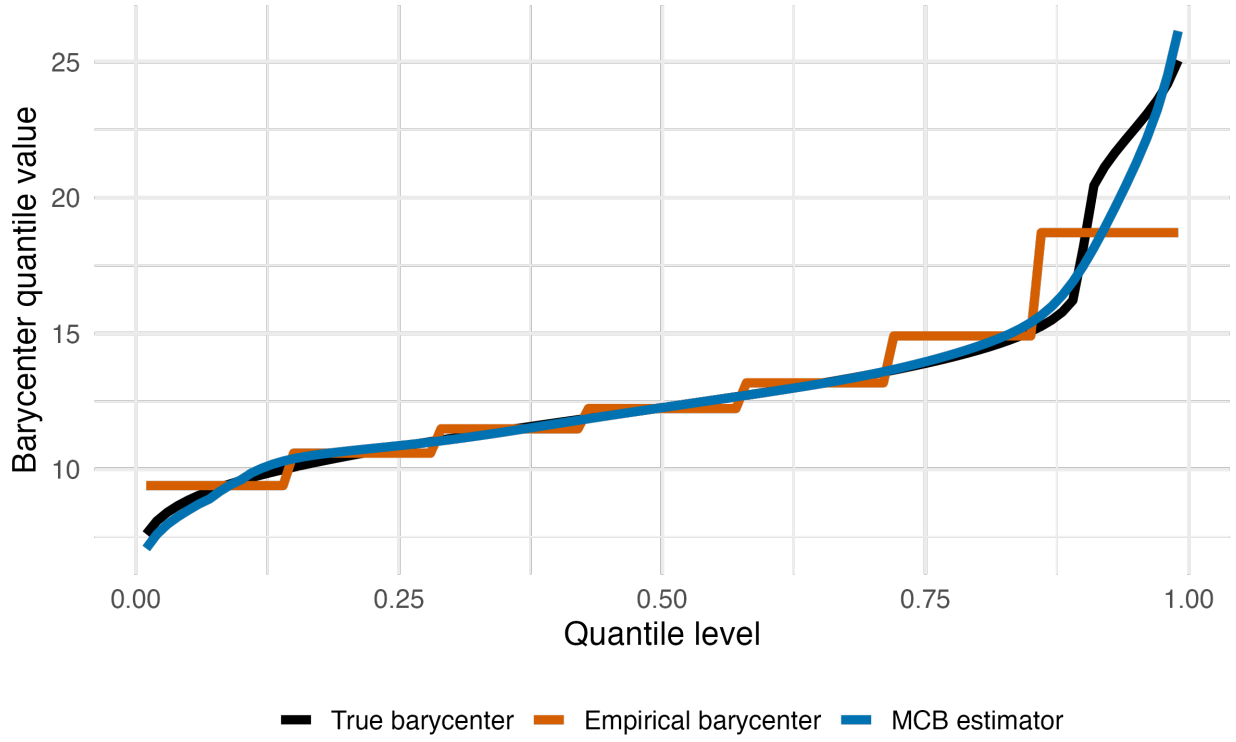


Figure 3.2: Quantile functions of the empirical Wasserstein barycenter (red) and the proposed MCB estimator (blue) for a single simulated data set where 1,000 distributions are drawn from the five distributions in Figure 3.1a and observed through only 7 i.i.d. samples. The true barycenter is shown in black.

To our knowledge, no existing method explicitly corrects for sparse sampling bias in estimating the Wasserstein barycenter. Previous work typically avoids this challenge by assuming that each latent distribution can be accurately recovered from its unit-level obser-

vations, either by the empirical measure or by a smoothed version of it (Bigot et al., 2018; Petersen and Müller, 2019; Panaretos and Zemel, 2020; Lin et al., 2023). These assumptions are natural when the number of observations per distribution is large, but they do not address the sparse setting considered here, where each distribution is observed through only a small, bounded number of samples. The closest related work is Zhou and Müller (2024), who explicitly study sparse sampling in Wasserstein regression. Their results characterize conditions under which the bias from using empirical measures is asymptotically negligible. However, they do not propose a correction to this bias, and consistency requires the per-distribution sample sizes to grow sufficiently quickly relative to the number of distributions. While such increasing-sample-size regimes provide important theoretical guarantees, they offer limited guidance when analyzing a fixed sparse dataset, where the number of observations per distribution is small and cannot be increased asymptotically.

The remainder of the chapter is organized as follows. Section 3.2 formalizes the data structure and defines the estimand of interest. Section 3.3 discusses estimation challenges under sparse sampling and the failure of the empirical barycenter. Section 3.4 presents our main representation result and the resulting MCB estimation procedure. Section 3.5 reviews estimation methods for the binomial mixing problem on which our estimator relies, and Section 3.6 develops the MCB estimator's asymptotic properties. Section 3.7 discusses extensions, including weighting and Fréchet variance estimation. Section 3.8 reports simulation results, Section 3.9 illustrates the proposed method using data from the HVTN 502 and HVTN 503 HIV-1 vaccine efficacy trials, and Section 3.10 concludes.

3.2 Preliminaries

3.2.1 Data structure

We consider a two-stage sampling framework. Each observational unit $i = 1, \ldots, n$ is associated with an unobserved probability distribution ν_i on $\mathcal{I} \subseteq \mathbb{R}$, where $\nu_i \in \mathcal{P}_2(\mathcal{I})$ and $\mathcal{P}_2(\mathcal{I})$ denotes the set of probability measures on $\mathcal{I}$ with finite second moment. We assume

that $\nu_1, \ldots, \nu_n$ are i.i.d. draws from a population distribution Π on $\mathcal{P}_2(\mathcal{I})$. Thus, Π is a distribution on distributions that captures distributional heterogeneity across units. Rather than observing ν_i directly, we observe m_i i.i.d. draws from ν_i, denoted by $(X_{i,j})_{j=1}^{m_i}$, where m_i is drawn from a distribution η on the positive integers and is independent of ν_i. The sampling scheme is summarized in Figure 3.3 and can be written as

$$\begin{aligned} \nu_i &\overset{\text{iid}}{\sim} \Pi, \quad i = 1, \ldots, n, \\ m_i &\overset{\text{iid}}{\sim} \eta, \quad m_i \perp\!\!\!\perp \nu_i \\ X_{i,1}, \ldots, X_{i,m_i} \mid \nu_i, m_i &\overset{\text{iid}}{\sim} \nu_i. \end{aligned} \tag{3.1}$$

We denote the empirical measure associated with the ith unit as $\hat{\nu}_i := \frac{1}{m_i} \sum_{j=1}^{m_i} \delta_{X_{i,j}}$, where δ_a denotes a Dirac mass at the point a. Under this two-stage sampling framework, the full data consist of $\{(\nu_i, m_i, (X_{i,j})_{j=1}^{m_i})\}_{i=1}^n$. Since the latent distributions ν_i are unobserved, the observed data consist of $\{O_i\}_{i=1}^n$, where $O_i := (m_i, (X_{i,j})_{j=1}^{m_i})$. Throughout, $\mathbb{P}$ and $\mathbb{E}$ denote probability and expectation under the full two-stage sampling law. When probability or expectation is taken only over one component of this law, we indicate this with a subscript. For example, $\mathbb{P}_{\nu \sim \Pi}$ and $\mathbb{E}_{\nu \sim \Pi}$ denote probability and expectation over a latent distribution drawn from Π, while $\mathbb{P}_{m \sim \eta}$ and $\mathbb{E}_{m \sim \eta}$ denote probability and expectation over a sample size drawn from η.

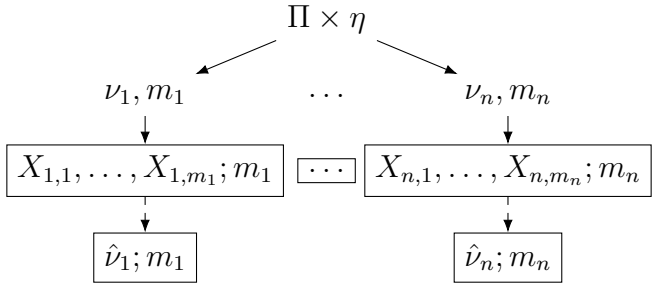


Figure 3.3: Schematic of the two-stage sampling scheme. Full arrows denote sampling, and dashed arrows denote the construction of the empirical measures. Observable elements are surrounded by a rectangle.

For any $\nu \in \mathcal{P}_2(\mathcal{I})$, let F_ν denote its cumulative distribution function, and define its quantile function by

$$F_\nu^-(t) := \inf\{x \in \mathbb{R} : F_\nu(x) \geq t\}, \qquad t \in (0,1).$$

We equip $\mathcal{P}_2(\mathcal{I})$ with the 2-Wasserstein metric, defined for $\nu_1, \nu_2 \in \mathcal{P}_2(\mathcal{I})$ by

$$d_{W_2}^2(\nu_1, \nu_2) := \int_0^1 \left(F_{\nu_1}^-(t) - F_{\nu_2}^-(t)\right)^2 dt.$$

Under this metric, $\mathcal{P}_2(\mathcal{I})$ forms a metric space (Panaretos and Zemel, 2020).

Remark 3.2.1. *The assumption $m_i \perp\!\!\!\perp \nu_i$ is often referred to as non-informative cluster size, meaning that the number of observations for unit i is independent of its latent distribution. Informative cluster size, where m_i depends on ν_i, is a standard concern in the clustered data literature and can invalidate inference; see, e.g., Williamson et al. (2003). Although ν_i is not observed directly, the non-informative cluster size assumption has testable implications (Nevalainen et al., 2017).*

3.2.2 Estimand

Our primary target is the population Wasserstein barycenter of Π, defined as:

$$\bar{\nu}_\Pi \in \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathcal{I})} \mathbb{E}_{\nu \sim \Pi}\left[d_{W_2}^2(\mu, \nu)\right].$$

When the dependence on Π is clear from context, we write $\bar{\nu}$ in place of $\bar{\nu}_\Pi$. For one-dimensional distributions, the Wasserstein barycenter is unique, and its quantile function equals the mean of the quantile functions of measures drawn from Π (Proposition 4.1 in Bigot et al. (2017)):

$$t \mapsto F_{\bar{\nu}}^-(t) = \mathbb{E}_{\nu \sim \Pi}\left[F_\nu^-(t)\right], \qquad t \in (0,1).$$

Thus, inference on the barycenter can be reframed as inference on the pointwise average quantile function of the latent distributions.

For a fixed quantile level $\alpha \in (0,1)$, we define the scalar target

$$q_\alpha^\star := F_{\bar{\nu}}^-(\alpha) = \mathbb{E}_{\nu\sim\Pi}\big[F_\nu^-(\alpha)\big].$$

This is the α-quantile of the barycenter. Our main theoretical development focuses on pointwise inference for $q_\alpha^\star$ at fixed α, since this is the quantity directly characterized by our representation result. The full barycenter is determined by considering the collection $\{q_t^\star : t \in (0,1)\}$, which is exactly the quantile function of $\bar{\nu}$. Although our main focus is the barycenter and its pointwise quantiles, the same ideas extend to other related estimands; these are discussed in Section 3.7.

3.3 Challenges for estimating the barycenter

In this section, we highlight obstacles to estimating the barycenter under sparse sampling. First, the empirical barycenter, the natural plug-in estimator formed by averaging unit-level empirical quantiles, is generally inconsistent. The second obstacle is more fundamental. When the per-unit sample sizes are bounded, q_α^* is not identifiable from the observed data distribution without additional restrictions on Π.

3.3.1 The empirical Wasserstein barycenter

First, we consider an intuitive plug-in estimator for q_α^*. If the latent distributions ν_i were directly observed, a natural estimator of $q_\alpha^\star$ would be the oracle plug-in estimator

$$\hat{q}_\alpha^{\,\mathrm{or}} := \frac{1}{n}\sum_{i=1}^n F_{\nu_i}^-(\alpha).$$

However, since the ν_i are unobserved, this estimator is infeasible. A substitute is to replace each latent distribution ν_i by its empirical measure $\hat{\nu}_i$, yielding

$$\hat{q}_\alpha^{\text{emp}} := \frac{1}{n}\sum_{i=1}^{n} F^-_{\hat{\nu}_i}(\alpha).$$

As α varies over $(0,1)$, this defines the quantile function of the *empirical Wasserstein barycenter* (Bigot et al., 2018).

The next proposition shows that the expectation of $\hat{q}_\alpha^{\text{emp}}$ is governed by order statistics from the population barycenter, averaged over the sample size distribution η. This makes clear why sparse sampling induces bias: for fixed m, the empirical α-quantile behaves like the $\lceil m\alpha \rceil$th order statistic of a sample of size m, whose expectation generally differs from the population α-quantile. This bias is especially pronounced for extreme quantiles, as was seen in Figure 3.2.

Proposition 3.3.1. *For each $m \geq 1$, let $X^{*(m)}_{(1)}, \ldots, X^{*(m)}_{(m)}$ denote the order statistics of an i.i.d. sample of size m drawn from the barycenter $\bar{\nu}$. Then, under the two-stage sampling scheme in* (3.1)*,*

$$\begin{aligned}\mathbb{E}[\hat{q}_\alpha^{\text{emp}}] &= \mathbb{E}_{m\sim\eta}\left[\mathbb{E}_{\bar{\nu}}\left[X^{*(m)}_{(\lceil m\alpha \rceil)}\right]\right] \\ &= \sum_{m'=1}^{\infty} \eta(m')\,\mathbb{E}_{\bar{\nu}}\left[X^{*(m')}_{(\lceil m'\alpha \rceil)}\right], \end{aligned} \tag{3.2}$$

where $\eta(m') = \mathbb{P}_{m\sim\eta}(m = m')$ and, with slight abuse of notation, $\mathbb{E}_{\bar{\nu}}$ denotes expectation with respect to the i.i.d. sample from $\bar{\nu}$ whose order statistic appears inside the expectation.

Proposition 3.3.1 shows that $\hat{q}_\alpha^{\text{emp}}$ is unbiased for $q^\star_\alpha$ if and only if

$$\sum_{m'=1}^{\infty} \eta(m')\,\mathbb{E}_{\bar{\nu}}\left[X^{*(m')}_{(\lceil m'\alpha \rceil)}\right] = q^\star_\alpha, \tag{3.3}$$

which only holds in exceptional cases. Since the left-hand side does not depend on n, the em-

pirical barycenter is generally inconsistent for $q^\star_\alpha$ unless (3.3) holds. Example 3.3.1 illustrates this point in a simple setting.

Example 3.3.1. *Suppose the latent law* Π *is degenerate at* $\mathcal{N}(0,1)$ *(i.e.,* $\nu_i = \mathcal{N}(0,1)$ *for all* i*) and hence* $\bar{\nu} = \mathcal{N}(0,1)$*. Also suppose that all units have the same sample size* m*. Let* $X^{*(m)}_{(1)}, \ldots, X^{*(m)}_{(m)}$ *denote the order statistics of an i.i.d. sample of size* m *drawn from* $\mathcal{N}(0,1)$*. Then, as* $n \to \infty$*,*

$$\hat{q}^{\text{emp}}_\alpha \xrightarrow{P} \mathbb{E}\left[X^{*(m)}_{(\lceil m\alpha \rceil)}\right].$$

For fixed m*, the right-hand side is piecewise constant in* α*, whereas* $q^\star_\alpha = F^{-1}_{\bar{\nu}}(\alpha)$ *is strictly increasing and continuous in* α*, so the equality* $\mathbb{E}\left[X^{*(m)}_{(\lceil m\alpha \rceil)}\right] = q^\star_\alpha$ *can only hold at isolated values of* α*. In particular, the empirical barycenter quantile estimator is inconsistent for almost all* α*. By symmetry, when* m *is odd,* $\alpha = 0.5$ *is one value where it is consistent.*

3.3.2 Non-identifiability

The previous section shows that the empirical Wasserstein barycenter's quantile function is generally pointwise inconsistent. In this section we discuss a more fundamental problem with estimation. Different laws Π on latent distributions may generate the same observed data distribution while having different barycenters. This is formalized in the following proposition.

Proposition 3.3.2. *Assume that the unit-specific sample sizes are bounded, in the sense that* $\mathbb{P}_{m\sim\eta}(m \le C) = 1$ *for some* $C < \infty$*. Then, for every* $\alpha \in (0,1)$*, there exist two distributions on distributions, denoted by* Π_1 *and* Π_2*, which, together with the same* η*, induce the same distribution for the observed data* $O_i = (m_i, (X_{i,j})_{j=1}^{m_i})$*, but satisfy* $q^\star_{\alpha,\Pi_1} \neq q^\star_{\alpha,\Pi_2}$*.*

This non-identifiability is a fundamental obstacle for inference: without additional assumptions on Π, the observed-data distribution does not in general determine $q^\star_{\alpha,\Pi}$. Therefore, consistent estimation of $q^\star_{\alpha,\Pi}$ is generally impossible. The lack of identifiability can be addressed in two ways. One approach, as discussed in the introduction, is to allow the

sampling distribution η to vary with n so that unit-specific sample sizes increase and the barycenter becomes identifiable in the limit (Zhou and Müller, 2024). A second approach is to assume a parametric model for the law Π that is identifiable even when unit-specific sample sizes are bounded. Asymptotic results then follow from standard maximum likelihood theory; see Example 3.3.2.

Example 3.3.2. *A simple fully parametric example is a Gaussian random-intercept model:*

$$\nu_i = \mathcal{N}(\mu_i, \sigma^2), \qquad \mu_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \tau^2).$$

Then the population barycenter is $\bar{\nu} = \mathcal{N}(\mu_0, \sigma^2)$, which is identifiable if some units have at least two observations. Estimation and inference for $\bar{\nu}$ can then be performed by fitting the random-intercept model using maximum likelihood or restricted maximum likelihood (Molenberghs and Verbeke, 2000).

3.4 MCB construction and estimator

3.4.1 Alternative representation via marginal mixing distributions

The empirical barycenter is a natural estimator because it replaces each latent (unobserved) distribution by its unit-level empirical distribution. However, as shown above, this plug-in estimator is generally inconsistent under sparse sampling. The key difficulty is that subject-specific quantiles $F_\nu^-(\alpha)$ can be poorly estimated when only a small number of observations are available for each distribution.

Our strategy is based on a reformulation of the problem. Although our target, $q_\alpha^\star = \mathbb{E}_{\nu\sim\Pi}\{F_\nu^-(\alpha)\}$, is the expectation of the latent α-quantiles, we instead begin with the seemingly more ambitious goal of estimating the *distribution* of the latent α-quantiles across units. The rationale is that this distribution is directly linked to the distribution of the unit-level CDF values. Indeed, for any $x \in \mathcal{I}$,

$$\{F_\nu^-(\alpha) \le x\} \iff \{F_\nu(x) \ge \alpha\}. \tag{3.4}$$

Thus, although the expectation of a random quantile is not a simple transformation of the expectation of the corresponding distribution functions, the distribution of the latent quantiles can be translated into the distribution of the latent CDF values. This reformulation is useful because, as we argue below, estimating the distribution of the latent CDF values can be reduced to the well-studied binomial mixture problem.

To make this precise, for each fixed $x \in \mathcal{I}$, let

$$G(x,t) := \mathbb{P}_{\nu\sim\Pi}\left(F_\nu(x) \le t\right), \qquad t \in [0,1], \tag{3.5}$$

so that $G(x,\cdot)$ is the CDF of the random variable $F_\nu(x)$ when $\nu \sim \Pi$. This is illustrated in Figure 3.4. Equivalently, for each fixed x, the law $G(x,\cdot)$ serves as the mixing distribution in a binomial mixture model. Writing $U_i(x) := F_{\nu_i}(x)$ for the unobserved CDF value of unit i at x and

$$Y_i(x) := \sum_{j=1}^{m_i} \mathbf{1}\{X_{i,j} \le x\} \tag{3.6}$$

for the number of "successes" in the ith unit's sample at x, we have the conditional model

$$Y_i(x) \mid m_i, U_i(x) \sim \mathrm{Bin}\left(m_i, U_i(x)\right). \tag{3.7}$$

By construction, the latent values $U_1(x), \ldots, U_n(x)$ are i.i.d. with CDF $G(x,\cdot)$, so marginally $Y_i(x)$ is a binomial mixture with mixing distribution $G(x,\cdot)$.

Theorem 3.4.1 below is a representation result: it represents the target $q^\star_\alpha$ in terms of the collection of marginal mixing distributions $\{G(x,\cdot) : x \in \mathcal{I}\}$ induced by Π. Hence, $q^\star_\alpha$ is identified with the observed data if the mixing distribution $G(x,\cdot)$ is identifiable for each $x \in \mathcal{I}$. We return to those conditions in Section 3.5.

Theorem 3.4.1 (Representation via marginal mixing distributions)**.** *For $x \in \mathcal{I}$, define $G(x,\alpha-) := \lim_{t\uparrow\alpha} G(x,t)$, and let*

$$H_\alpha(x) := 1 - G(x,\alpha-).$$

Then the CDF of the distribution of α-quantiles $F_\nu^-(\alpha)$, where $\nu \sim \Pi$, equals $x \mapsto H_\alpha(x)$ for $x \in \mathcal{I}$. Consequently,

$$q_\alpha^\star = \int_{\mathcal{I}} x \, dH_\alpha(x) = -\int_{\mathcal{I}} x \, dG(x, \alpha-).$$

The key observation behind Theorem 3.4.1 is the equivalence in (3.4). For each $x \in \mathcal{I}$, this gives

$$\mathbb{P}_{\nu\sim\Pi}\{F_\nu^-(\alpha) \le x\} = \mathbb{P}_{\nu\sim\Pi}\{F_\nu(x) \ge \alpha\} = 1 - G(x, \alpha-) = H_\alpha(x).$$

Therefore, $x \mapsto H_\alpha(x)$ is the CDF of the latent α-quantile $F_\nu^-(\alpha)$ under $\nu \sim \Pi$. Thus, once the marginal mixing distributions $\{G(x, \cdot) : x \in \mathcal{I}\}$ are identified from the observed data, the distribution of the latent α-quantiles is identified as well. Its mean, $q_\alpha^\star$, is then identified by integrating with respect to H_α. We refer to this representation as the *marginal construction* of the barycenter.

3.4.2 Estimation procedure

Theorem 3.4.1 suggests a plug-in strategy for estimating $q_\alpha^\star$. For a fixed $\alpha \in (0, 1)$, suppose that we have an estimator $\hat{G}_n(x, \alpha-)$ of $G(x, \alpha-)$ for each $x \in \mathcal{I}$, and define $\hat{H}_{n,\alpha}(x) := 1 - \hat{G}_n(x, \alpha-)$. Then the *marginal-constructed barycenter* (MCB) estimator is given by

$$\hat{q}_\alpha^{\mathrm{MCB}} := \int_{\mathcal{I}} x \, d\hat{H}_{n,\alpha}(x). \tag{3.8}$$

While the true function $x \mapsto G(x, \alpha-)$ may vary over the entire interval $\mathcal{I}$, the data used to estimate $G(x, \cdot)$ change only at values of x where observations occur. That is, for any two points x and x' lying between the same pair of consecutively observed values, the indicators $\mathbf{1}\{X_{i,j} \le x\}$ and $\mathbf{1}\{X_{i,j} \le x'\}$ will coincide for each i, j, so the data used to estimate the corresponding mixing distributions are identical. Therefore, it suffices to estimate $G(x, \cdot)$ only at these locations.

We implement the MCB estimator for $q_\alpha^\star$ in the following steps, illustrated in Figure 3.5:

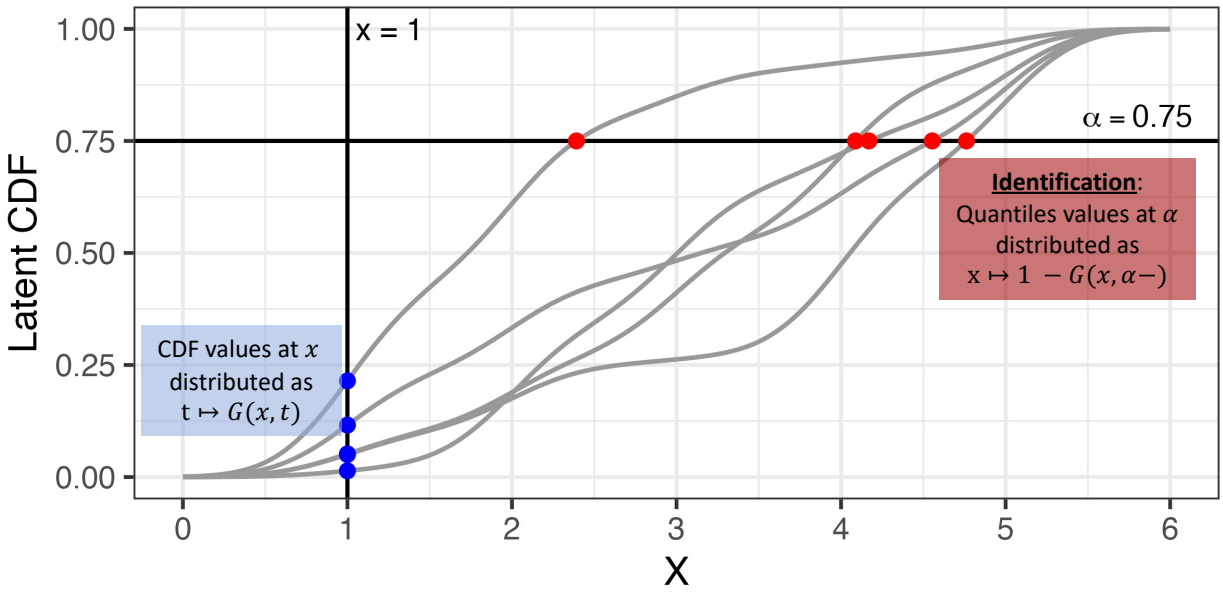


Figure 3.4: The figure shows the CDFs of five latent distributions. The vertical positions of the blue points are the values $F_\nu(1)$ for the five latent distributions; across $\nu \sim \Pi$, these values have CDF $t \mapsto G(1,t)$ by (3.5). The horizontal positions of the red points are the 0.75-quantiles of the latent distributions; by Theorem 3.4.1, these values have CDF $H_{0.75}(x) = 1 - G(x, 0.75-)$.

1. Let
$$\mathcal{X} := \{X_{i,j} : 1 \le i \le n,\ 1 \le j \le m_i\}$$
denote the set of observed points across all units and let $x_1 < \cdots < x_K$ be the sorted values of $\mathcal{X}$, with $K := |\mathcal{X}|$.
2. For each point x_k with $k = 1, \ldots, K-1$, estimate the binomial mixing distribution $t \mapsto G(x_k, t)$ using the Binomial data $\{(Y_i(x_k), m_i)\}_{i=1}^n$. Denote the resulting estimator by $t \mapsto \hat{G}_n(x_k, t)$.
3. With α fixed, define $x \mapsto \hat{H}_{n,\alpha}(x)$ as the stepwise function
$$\hat{H}_{n,\alpha}(x) := \begin{cases} 0, & x < x_1, \\ 1 - \hat{G}_n(x_k, \alpha-), & x \in [x_k, x_{k+1}), \quad k = 1, \ldots, K-1, \\ 1, & x \ge x_K. \end{cases}$$
In practice, we evaluate $\hat{G}_n(x, t)$ at $t = \alpha$ rather than $\alpha-$; this is equivalent under continuity of $t \mapsto G(x, t)$ at $t = \alpha$.
We then compute the MCB estimate for $q_\alpha^\star$ as follows:
$$\begin{aligned} \hat{q}_\alpha^{\mathrm{MCB}} &:= \int_{\mathcal{I}} x \, d\hat{H}_{n,\alpha}(x) \\ &= \sum_{k=1}^{K} x_k \left\{ \hat{G}_n(x_{k-1}, \alpha) - \hat{G}_n(x_k, \alpha) \right\}, \end{aligned}$$
where for notational convenience $\hat{G}_n(x_0, \alpha) := 1$ and $\hat{G}_n(x_K, \alpha) := 0$.

The MCB estimator is modular: its construction is separated from the choice of binomial mixture estimator for the mixing distributions. Specifically, by reformulating barycenter estimation in terms of the marginal mixing distributions $G(x, \cdot)$, the MCB framework allows different estimators of $G(x, \cdot)$ to produce different barycenter estimators. In Section 3.5, we discuss potential binomial mixture estimators. A simple choice is to treat the empirical CDF

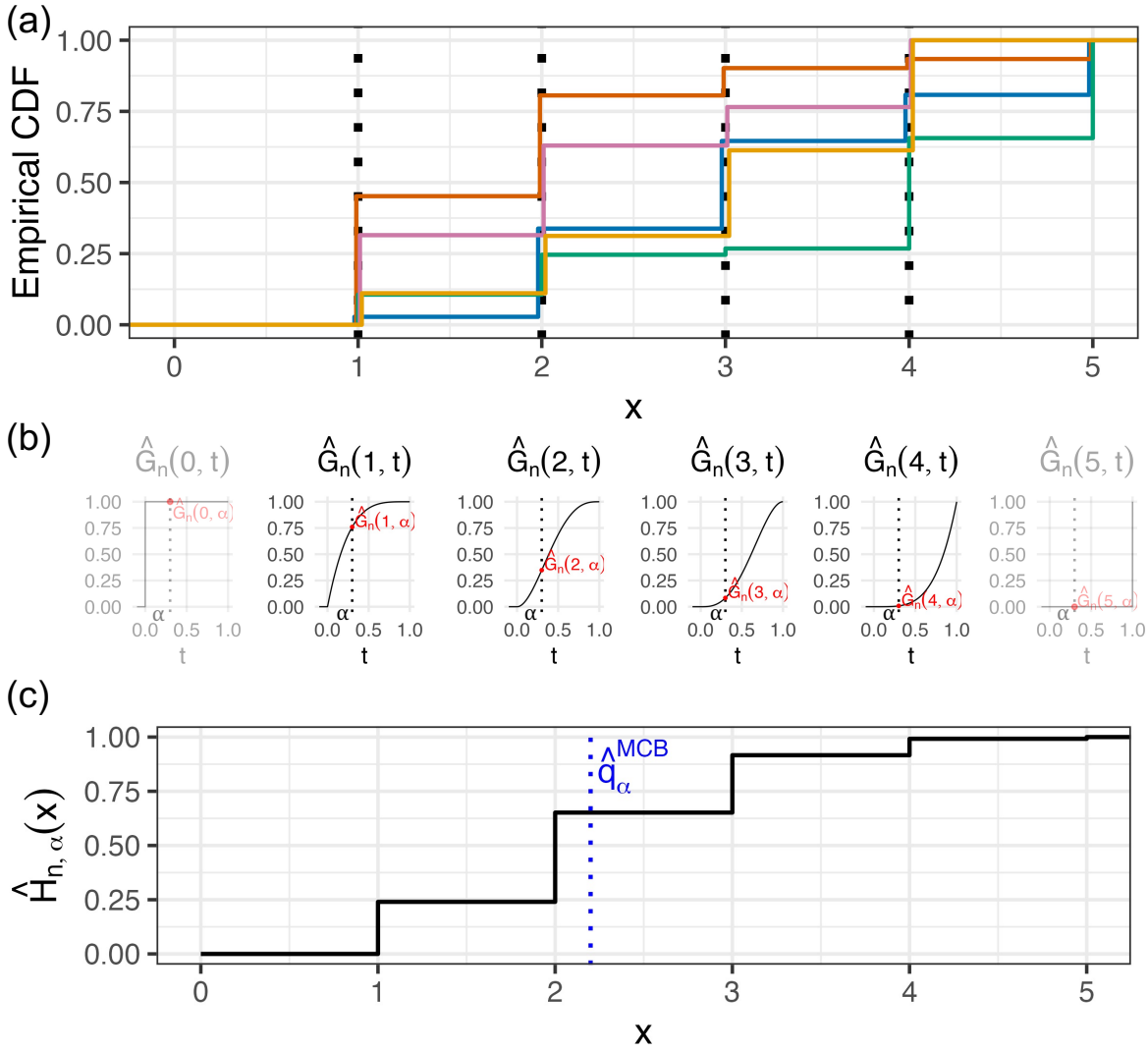


Figure 3.5: (a) Empirical CDFs in a toy example with five unit-level distributions. The dotted vertical lines indicate the grid points $x = 1, 2, 3, 4$ at which the binomial mixing distributions are estimated. (b) Estimated mixing distributions $t \mapsto \hat{G}_n(x, t)$ at each grid point. For points below and above the grid, the boundary distributions are fixed by construction to be degenerate at 0 and 1, respectively, and are shown with reduced opacity. (c) Construction of $\hat{H}_{n,\alpha}(x) = 1 - \hat{G}_n(x, \alpha)$ from the estimated mixing distributions in panel (b). The blue dotted line marks the MCB estimate $\hat{q}_\alpha^{\mathrm{MCB}}$, defined as the mean of the distribution with CDF $\hat{H}_{n,\alpha}$.

values $\{F_{\hat{\nu}_i}(x)\}_{i=1}^n$ as if they were direct draws from the mixing distribution $G(x,\cdot)$, which yields the empirical histogram estimator of the mixing distribution (Ye and Bickel, 2021):

$$\hat{G}_n^{\text{emp}}(x,t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{F_{\hat{\nu}_i}(x) \le t\}, \qquad t \in [0,1].$$

Notably, if we follow the estimation procedure using $\hat{G}_n^{\text{emp}}(x,\cdot)$, we recover the empirical Wasserstein barycenter.

Proposition 3.4.2. *The MCB estimator that uses the empirical mixing distributions $\hat{G}_n^{\text{emp}}(x,\cdot)$ coincides with the empirical Wasserstein barycenter. That is,*

$$\hat{q}_\alpha^{\text{MCB}} = \frac{1}{n}\sum_{i=1}^{n} F_{\hat{\nu}_i}^{-}(\alpha) = \hat{q}_\alpha^{\text{emp}}.$$

Proposition 3.4.2 shows that the empirical barycenter arises as a special case of the MCB estimator. This characterization shows the advantage of our formulation over the empirical barycenter: by replacing the empirical histogram estimator of $G(x,\cdot)$ with alternative estimators within the MCB framework, one may obtain barycenter estimators with improved statistical properties. In particular, if one is willing to impose parametric structure on these mixing distributions, pointwise consistent and asymptotically normal estimators of the barycenter's quantile function can be obtained. More generally, even without parametric structure, the empirical histogram estimator is suboptimal for estimating binomial mixing distributions (Vinayak et al., 2019; Tian et al., 2017), suggesting that other estimators of $G(x,\cdot)$ within the MCB framework may have better performance.

Remark 3.4.1. *For a generic binomial mixture estimator, $x \mapsto \hat{H}_{n,\alpha}(x)$ may fail to be monotone increasing, even though $H_\alpha(x)$ is a distribution function on $\mathcal{I}$. This lack of guaranteed monotonicity arises because the marginal mixing distributions $G(x,\cdot)$ are estimated separately across values of x. In implementation, one may enforce monotonicity by replacing $\hat{H}_{n,\alpha}$ with its projection onto the cone of nondecreasing functions in a post-processing step*

using isotonic regression. General results on isotonic correction of non-monotone estimators imply that, under suitable regularity conditions, such projections are asymptotically equivalent to the unprojected estimator (Westling et al., 2020). Related monotonicity corrections have also been used in Wasserstein estimation problems; see, e.g., Zhou and Müller (2024).

Remark 3.4.2. *In practice, it may be computationally infeasible to estimate mixing distributions at all unique points of $\mathcal{X}$. In that case, we can pre-specify a grid of points $x_0 < x_1 < \cdots < x_L$ in $\mathcal{I}$ and approximate the integral in $\hat{q}_\alpha^{\mathrm{MCB}}$ by a midpoint Riemann–Stieltjes sum. For each grid point x_ℓ, $\ell = 0, \ldots, L$, we estimate the binomial mixing distribution $G(x_\ell, \cdot)$ and denote the estimate by $\hat{G}_n(x_\ell, \cdot)$. We then approximate $\hat{q}_\alpha^{\mathrm{MCB}}$ as*

$$\hat{q}_\alpha^{\mathrm{MCB,approx}} := \sum_{\ell=1}^{L} \bar{x}_\ell \left\{ \hat{G}_n(x_{\ell-1}, \alpha) - \hat{G}_n(x_\ell, \alpha) \right\}, \qquad \bar{x}_\ell := \frac{x_{\ell-1} + x_\ell}{2}.$$

3.5 The fixed-x binomial mixture problem

In this section, we review the estimation of the mixing distribution $G(x, \cdot)$ for a fixed $x \in \mathcal{I}$. This fixed-x binomial mixture problem is the core building block of the MCB framework: once an estimator of $G(x, \cdot)$ is available for each x, it can be plugged into the construction in Section 3.4 to form an estimator of $q_\alpha^\star$. We first clarify when $G(x, \cdot)$ is identifiable under sparse sampling, and then outline potential estimators under structural restrictions. In this section, we draw on the extensive literature on binomial mixture models; see Teicher (1963), Lindsay (1995), Wood (1999), Efron (2016), Tian et al. (2017), and Ye and Bickel (2021), among others.

Recall that, for fixed $x \in \mathcal{I}$, we defined $U_i(x) := F_{\nu_i}(x)$ and $Y_i(x) := \sum_{j=1}^{m_i} \mathbf{1}\{X_{i,j} \le x\}$. Then, conditional on m_i and $U_i(x)$, $Y_i(x) \mid m_i, U_i(x) \sim \mathrm{Bin}\big(m_i, U_i(x)\big)$, and $U_i(x) \sim G(x, \cdot)$. Therefore, inference on $G(x, \cdot)$ reduces to a binomial mixture problem with conditional likelihood

$$L_n\big(G(x, \cdot)\big) = \prod_{i=1}^{n} \int_0^1 \binom{m_i}{Y_i(x)} t^{Y_i(x)} (1-t)^{m_i - Y_i(x)} \, dG(x, t). \tag{3.9}$$

3.5.1 Identifiability

A key difficulty in estimating the mixing distribution $G(x, \cdot)$ is that it is not identifiable from the observed data without additional restrictions (Wood, 1999). This is because the observed data only provide information on the first m_i moments of $G(x, \cdot)$ for each unit i, and if the m_i are bounded, then only finitely many moments of $G(x, \cdot)$ are identified. This is formalized in the following proposition.

Proposition 3.5.1. *Fix $x \in \mathcal{I}$ and suppose that $\mathbb{P}_{m\sim\eta}(m \leq C) = 1$ for some finite constant C. Then the observed distribution of $(Y_i(x), m_i)$ depends on $G(x, \cdot)$ only through its first C moments. Consequently, $G(x, \cdot)$ is not identifiable over the class of all probability distributions on $[0, 1]$.*

Proposition 3.5.1 clarifies where structural assumptions can enter to ensure barycenter identifiability. We showed in Theorem 3.4.1 that the barycenter depends on Π only through the collection of mixing distributions $\{G(x, \cdot) : x \in \mathcal{I}\}$. Proposition 3.5.1 now shows that these mixing distributions are themselves not nonparametrically identifiable from the observed data. Therefore, if structural restrictions on $G(x, \cdot)$ can make the pointwise mixing distributions identifiable from the binomial data, then the barycenter is identifiable through Theorem 3.4.1.

Remark 3.5.1. *Structural assumptions are not the only way to address the non-identifiability of the mixing distributions. For example, if $\mathbb{P}_{m\sim\eta}(m \geq M) > 0$ the first M moments of $G(x, \cdot)$ are identifiable. Hence, given the first M moments, one can obtain bounds for $G(x, \alpha)$ at each x and consequently bounds for $q^\star_\alpha$. We do not pursue this approach in the current paper, but it may be an interesting direction for future work.*

3.5.2 Estimation under structural restrictions

To address the lack of identifiability of the collection of mixing distributions, we restrict attention to estimation under parametric models for each $G(x, \cdot)$. Specifically, we consider

finite-dimensional families for the mixing distribution at every $x \in \mathcal{I}$. Let

$$\mathcal{G}_x = \big\{G(x, \cdot; \beta) : \beta \in \mathcal{B}_x \subseteq \mathbb{R}^d\big\}$$

denote a parametric class of distribution functions on $[0, 1]$. With $m_{\max} := \max\{m : \eta(m) > 0\}$ denoting the largest possible unit-level sample size, Proposition 3.5.1 implies that $d \leq m_{\max}$ is a necessary but not sufficient condition for identification. In the following, we assume that this model is identifiable and correctly specified in the sense that $G(x, \cdot) = G(x, \cdot; \beta_0(x))$ for a unique $\beta_0(x) \in \mathcal{B}_x$, for each $x \in \mathcal{I}$. An estimator of $\beta_0(x)$ may then be obtained by parametric maximum likelihood estimation, and the resulting estimator $\hat{\beta}_n(x)$ induces an estimator $\hat{G}_n(x, \cdot) := G(x, \cdot; \hat{\beta}_n(x))$ of the mixing distribution.

Below, we outline two possible choices for the parametric family $\mathcal{G}_x$.

Beta specification. Suppose that, for every $x \in \mathcal{I}$, $G(x, \cdot)$ is the CDF of a Beta(a_x, b_x) distribution for some $a_x > 0$ and $b_x > 0$. Then the induced model for the observed data at x is a Beta-binomial model with probability mass function:

$$\mathbb{P}\big(Y_i(x) = y \mid m_i, a_x, b_x\big) = \binom{m_i}{y} \frac{B(y + a_x, m_i - y + b_x)}{B(a_x, b_x)}, \qquad y = 0, \ldots, m_i,$$

where $B(\cdot, \cdot)$ denotes the Beta function. For every $x \in \mathcal{I}$, the parameters (a_x, b_x) may be estimated by maximizing the corresponding likelihood over $(0, \infty)^2$, yielding estimators $\hat{a}_x$ and $\hat{b}_x$. The estimated mixing distribution $\hat{G}_n(x, \cdot)$ is then defined as the CDF of the Beta$(\hat{a}_x, \hat{b}_x)$ distribution.

Spline-based classes. A more flexible alternative to the Beta specification models the mixing density $g(x, t) := \partial_t G(x, t)$ via a finite-dimensional basis. Let $Q(t) \in \mathbb{R}^p$ be a spline basis on $[0, 1]$. Note that to satisfy the necessary identification condition above, we will need to restrict the number of free spline coefficients p to satisfy $p \leq m_{\max}$.

The density at each $x \in \mathcal{I}$ can then be modeled as follows (Efron, 2016):

$$g\big(x, t; \beta(x)\big) = \exp\{Q(t)^\top \beta(x) - \phi(\beta(x))\}, \qquad \phi(\beta) = \log \int_0^1 \exp\{Q(u)^\top \beta\}\, du. \tag{3.10}$$

For every $x \in \mathcal{I}$, the parameter $\beta(x)$ may be estimated by maximizing the corresponding likelihood over $\mathbb{R}^p$, yielding an estimator $\hat{\beta}_n(x)$ and inducing an estimated mixing distribution $\hat{G}_n(x, \cdot)$ with density $g(\cdot; \hat{\beta}_n(x))$.

Although we impose a parametric model for each marginal mixing distribution $G(x, \cdot)$, this does not require specifying a fully parametric model for the law Π on unit-level distributions. The restrictions are imposed only on the marginal laws of the random CDF values $F_\nu(x)$, separately for each $x \in \mathcal{I}$. The remaining aspects of the population law Π, including the dependence structure among $\{F_\nu(x) : x \in \mathcal{I}\}$, are left unrestricted as long as they induce these marginal laws. In this sense, the resulting MCB estimator is semiparametric: it uses parametric structure on the features of Π needed to identify the barycenter, while avoiding a full parametric specification of Π.

Remark 3.5.2 (Connection to the Dirichlet process)**.** *The Beta specification includes the one-dimensional marginal distributions induced by a Dirichlet process (DP) as a special case (Ferguson, 1973). If $\nu \sim \mathrm{DP}(\kappa, P_0)$, where P_0 is the DP base measure and $\kappa > 0$ is the concentration parameter, then for each $x \in \mathcal{I}$,*

$$F_\nu(x) \sim \mathrm{Beta}\big(\kappa\, P_0((-\infty, x]),\ \kappa\,(1 - P_0((-\infty, x]))\big).$$

3.5.3 Other methods

Several other estimators for binomial mixing distributions are available, including the nonparametric maximum likelihood estimator (NPMLE) (Kiefer and Wolfowitz, 1956; Wood, 1999), moment-based estimators (Tian et al., 2017), and kernel density-smoothed estimators (Lee et al., 2025). The NPMLE, which maximizes (3.9) over all probability measures on $[0, 1]$, is particularly attractive because it has no tuning parameters. In the bounded m_i

setting, however, the NPMLE may not be unique and its asymptotic analysis is difficult. We reserve the theoretical study of these estimators for future work, although we examine the performance of the NPMLE in our simulation study.

3.6 Asymptotic properties

In this section, we remain agnostic about the particular estimator for the mixing distributions and instead impose high-level conditions on $\hat{G}_n(x, \cdot)$ that are sufficient for consistency and asymptotic normality of the MCB estimator $\hat{q}_\alpha^{\mathrm{MCB}}$ for $q_\alpha^\star$.

3.6.1 Consistency

The MCB estimator is built from pointwise estimates of the marginal mixing distributions $G(x, \cdot)$, but the target $q_\alpha^\star$ depends on these estimates through an integral over x. Specifically, $\hat{q}_\alpha^{\mathrm{MCB}} = \int_{\mathcal{I}} x \, d\hat{H}_{n,\alpha}(x)$, where $\hat{H}_{n,\alpha}(x) = 1 - \hat{G}_n(x, \alpha-)$. Consequently, pointwise consistency of $\hat{G}_n(x, \alpha-)$ for each fixed x does not by itself guarantee consistency of $\hat{q}_\alpha^{\mathrm{MCB}}$; additional conditions are needed to control how the estimation error behaves after integration. Theorem 3.6.1 provides sufficient conditions under which the MCB integral is consistent.

Theorem 3.6.1 (Consistency of the MCB estimator at fixed α)**.** *Assume that:*

(C1) For every continuity point x of $x \mapsto H_\alpha(x)$,

$$\hat{H}_{n,\alpha}(x) = H_\alpha(x) + r_n(x), \qquad r_n(x) = o_P(1).$$

(C2) With probability tending to one, $x \mapsto \hat{H}_{n,\alpha}(x)$ is a cumulative distribution function on $\mathcal{I}$.

(C3) Additionally, at least one of the following holds:

- *(a) the support is bounded, $\mathcal{I} = [a, b] \subset \mathbb{R}$ with $|a|, |b| < \infty$; or*
- *(b) the sequence $(\hat{H}_{n,\alpha})_{n \geq 1}$ is asymptotically uniformly integrable with respect to $f : x \mapsto |x|$, per Definition C.4.7 in Appendix C.4.2.*

Then the MCB estimator is consistent: $\hat{q}_\alpha^{\mathrm{MCB}} = q_\alpha^\star + o_P(1)$.

The conditions in Theorem 3.6.1 are mild. Condition (C1) is simply pointwise consistency of the estimated CDF at each x, condition (C2) can be enforced directly by projecting onto the cone of distribution functions (i.e., isotonic regression), and condition (C3) is automatic when $\mathcal{I}$ is bounded. When $\mathcal{I}$ is unbounded, condition (C3)b prevents mass in the tails from dominating the integral functional but may be more difficult to verify.

While our main interest is in pointwise inference for $q_\alpha^\star$, the same argument also yields convergence in $\mathcal{P}_2(\mathcal{I})$ with respect to the 2-Wasserstein metric under additional regularity conditions.

Corollary 3.6.2. *Assume that, with probability tending to one, $t \mapsto \hat{q}_t^{\mathrm{MCB}}$ is a quantile function, and let $\hat{\nu}_n$ denote the corresponding probability measure. If the assumptions of Theorem 3.6.1 hold for every continuity point of the true barycenter quantile function $t \mapsto F^-_{\bar{\nu}_\Pi}(t)$, then*

$$d_{W_2}(\hat{\nu}_n, \bar{\nu}_\Pi) \xrightarrow{P} 0,$$

if $\mathcal{I}$ is bounded or if $\hat{\nu}_n$ is asymptotically uniformly integrable with respect to $f(x) = x^2$.

3.6.2 Asymptotic linearity

We next study the asymptotic distribution of $\hat{q}_\alpha^{\mathrm{MCB}}$. Theorem 3.6.3 states the conditions under which pointwise asymptotic linearity of $\hat{H}_{n,\alpha}(x)$ for each fixed $x \in \mathcal{I}$ implies asymptotic linearity of the integrated estimator $\hat{q}_\alpha^{\mathrm{MCB}}$.

Theorem 3.6.3 (Asymptotic linearity of the MCB estimator)**.** *Assume that:*

(AL1) The estimator $\hat{H}_{n,\alpha}$ admits the linear expansion

$$\sqrt{n}\big(\hat{H}_{n,\alpha}(x) - H_\alpha(x)\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\alpha(O_i; x) + r_{\alpha,n}(x),$$

where, for each fixed x, the map $O \mapsto \psi_\alpha(O;x)$ has mean zero $\mathbb{E}[\psi_\alpha(O;x)] = 0$ and $\mathbb{E}[\int_{\mathcal{I}} |\psi_\alpha(O;x)|\,dx] < \infty$, and the remainder satisfies $\int_{\mathcal{I}} |r_{\alpha,n}(x)|\,dx = o_P(1)$.

(AL2) With probability tending to one, $x \mapsto \hat{H}_{n,\alpha}(x)$ is a cumulative distribution function on $\mathcal{I}$ satisfying

$$\int_{\mathcal{I}} |x|\, d\hat{H}_{n,\alpha}(x) < \infty.$$

(AL3) Define the integrated influence function

$$\tilde{\psi}_\alpha(O) := -\int_{\mathcal{I}} \psi_\alpha(O;x)\,dx,$$

and assume that $\mathrm{Var}\big(\tilde{\psi}_\alpha(O)\big) =: \sigma_\alpha^2 < \infty$.

Then

$$\sqrt{n}\big(\hat{q}_\alpha^{\mathrm{MCB}} - q_\alpha^\star\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_\alpha(O_i) + o_P(1) \xrightarrow{d} \mathcal{N}(0, \sigma_\alpha^2).$$

The hardest-to-verify assumption is the remainder condition $\int_{\mathcal{I}} |r_{\alpha,n}(x)|\,dx = o_P(1)$, which ensures that the pointwise expansion remains valid after integration over x. Pointwise negligibility of $r_n(x)$ for each fixed x is not enough on its own, since small errors can accumulate when integrated over the whole domain. This condition is automatically satisfied when all distributions are supported on a common finite set of points, since in that case the integral reduces to a finite sum of pointwise remainder terms. Moreover, if the interval $\mathcal{I}$ is bounded, this assumption can be replaced with the uniform asymptotic linearity condition that $\sup_{x\in\mathcal{I}} |r_n(x)| = o_P(1)$.

The asymptotic variance σ_α^2 can be estimated consistently by $\hat{\sigma}_{n,\alpha}^2 := \frac{1}{n}\sum_{i=1}^n \hat{\tilde{\psi}}_\alpha(O_i)^2$ under standard conditions on the estimator $\hat{\tilde{\psi}}_\alpha$ of $\tilde{\psi}_\alpha$. This yields asymptotically valid $(1-\gamma)$ Wald confidence intervals for $q_\alpha^\star$ of the form

$$\left[\hat{q}_\alpha^{\mathrm{MCB}} \;\pm\; z_{1-\gamma/2}\frac{\hat{\sigma}_{n,\alpha}}{\sqrt{n}}\right],$$

where $z_{1-\gamma/2}$ is the $(1-\gamma/2)$-quantile of the standard normal distribution.

In practice, direct estimation of $\tilde{\psi}_\alpha$ requires first estimating the pointwise influence function $\psi_\alpha(\,\cdot\,; x)$ and then integrating:

$$\hat{\tilde{\psi}}_\alpha = -\int_{\mathcal{I}} \hat{\psi}_\alpha(\,\cdot\,; x)\, dx.$$

Even when $\hat{\psi}_\alpha(\,\cdot\,; x)$ has a closed form, this integral is typically evaluated as a large discrete sum, since $\hat{\psi}_\alpha(\,\cdot\,; x)$ changes at the observed points $x \in \mathcal{X}$. The bootstrap avoids this additional calculation and may therefore be more practical, especially when the fixed-x mixture estimator is computationally efficient.

Under stronger uniformity conditions, the pointwise asymptotic linearity result extends to weak convergence of the estimated barycenter quantile process to a Gaussian process.

Corollary 3.6.4 (Weak convergence to a Gaussian process)**.** *Suppose that the conditions of Theorem 3.6.3 hold for all α in a compact subset $\mathcal{A} = [\alpha_{\min}, \alpha_{\max}] \subset (0,1)$ and that the following conditions hold:*

(i) the probability (tending to one) statements from Theorem 3.6.3 hold uniformly over $\alpha \in \mathcal{A}$,

(ii) $\sup_{\alpha\in\mathcal{A}} \int_{\mathcal{I}} |r_{\alpha,n}(x)|\, dx = o_P(1)$, and

(iii) the class $\{\tilde{\psi}_\alpha : \alpha \in \mathcal{A}\}$ is P-Donsker.

Then the stochastic process

$$\sqrt{n}\big(\hat{q}_\alpha^{\mathrm{MCB}} - q_\alpha^\star\big)_{\alpha\in\mathcal{A}}$$

converges weakly in $\ell^\infty(\mathcal{A})$ to a centered Gaussian process with covariance function

$$\Gamma(\alpha, \alpha') := \mathbb{E}\left[\tilde{\psi}_\alpha(O)\, \tilde{\psi}_{\alpha'}(O)\right].$$

Corollary 3.6.4 provides the basis for simultaneous confidence bands over $\mathcal{A}$. In particular, if Z denotes the centered Gaussian process with covariance function Γ from Corollary 3.6.4, and $\sigma_\alpha^2 = \Gamma(\alpha, \alpha)$, then the limiting random variable governing the maximal standardized

error is $T = \sup_{\alpha \in \mathcal{A}} \left| \frac{Z(\alpha)}{\sigma_\alpha} \right|$. Here we assume $\inf_{\alpha \in \mathcal{A}} \sigma_\alpha > 0$ to avoid degeneracy. Let $c_{1-\gamma}$ denote the $(1-\gamma)$ quantile of the distribution of T. Then, if $\sup_{\alpha \in \mathcal{A}} |\sigma_\alpha - \hat{\sigma}_{n,\alpha}| = o_P(1)$, an asymptotic simultaneous $(1-\gamma)$ confidence band is

$$\left[\hat{q}_\alpha^{\mathrm{MCB}} \pm c_{1-\gamma} \frac{\hat{\sigma}_{n,\alpha}}{\sqrt{n}} \right], \qquad \alpha \in \mathcal{A}.$$

In practice, $c_{1-\gamma}$ can be approximated using either a multiplier bootstrap based on the estimated integrated influence functions $\hat{\hat{\psi}}_\alpha(O_i)$, or a nonparametric bootstrap that resamples the observational units O_i with replacement and recomputes $\hat{q}_\alpha^{\mathrm{MCB}}$ over a fine grid of quantile levels.

Example 3.6.1 (Beta specification)**.** *For the Beta specification, in which each mixing distribution $G(x, \cdot)$ is modeled by a Beta distribution, we translate the high-level conditions of Theorem 3.6.3 into lower-level sufficient conditions in Appendix C.5. Under these conditions, the MCB estimator with Beta mixing distributions is asymptotically normal at each fixed α.*

3.7 Extensions

The MCB construction extends beyond the estimation of the barycenter quantile $q_\alpha^\star$. In particular, it can be adapted to weighted barycenters and other functionals of the distribution of α-quantiles $F_\nu^-(\alpha)$. We discuss these extensions briefly and leave a detailed treatment of their asymptotic properties to future work.

3.7.1 Weighted barycenters

Suppose that we observe a covariate $Z_i \in \mathcal{Z}$ for each unit, and that $(\nu_i, Z_i) \overset{\text{iid}}{\sim} \Pi_Z$, where Π_Z is a probability measure on $\mathcal{P}_2(\mathcal{I}) \times \mathcal{Z}$ whose marginal law for ν_i is Π. Further let $w : \mathcal{Z} \to [0, \infty)$ be a measurable weight function satisfying $\mathbb{E}\{w(Z)\} = 1$. We can then

define the weighted barycenter quantile by

$$q_\alpha^{\star,w} := \mathbb{E}\big[w(Z)\, F_\nu^-(\alpha)\big]\,. \tag{3.11}$$

Equivalently, $q_\alpha^{\star,w}$ is the barycenter under the probability measure Π^w, defined as the marginal law for ν under Π_Z^w with $d\Pi_Z^w(\nu,z) := w(z)\, d\Pi_Z(\nu,z)$. Hence, we can also write $q_\alpha^{\star,w} = \mathbb{E}_{\nu\sim\Pi^w}[F_\nu^-(\alpha)]$. Next, we explain how the MCB estimator can be adapted to estimate $q_\alpha^{\star,w}$.

The same intermediate objects as in Section 3.4 can be defined under the probability measure Π^w. In particular, for each $x \in \mathcal{I}$, the random variable $F_\nu(x)$ has the following distribution function under Π^w:

$$G^w(x,t) := \mathbb{P}_{\nu\sim\Pi^w}\big(F_\nu(x) \le t\big), \qquad t \in [0,1]. \tag{3.12}$$

Similarly, letting $H_\alpha^w(x) := 1 - G^w(x, \alpha-)$, the same argument as in Section 3.4 yields the representation

$$q_\alpha^{\star,w} = \int_{\mathcal{I}} x\, dH_\alpha^w(x) = -\int_{\mathcal{I}} x\, dG^w(x,\alpha-). \tag{3.13}$$

The MCB construction extends directly to this setting by replacing the unweighted binomial-mixture problem at each x with a weighted version. Let $w_i := w(Z_i)$ and recall that $Y_i(x) := \sum_{j=1}^{m_i} \mathbb{1}\{X_{i,j} \le x\}$. We can estimate $G^w(x,\cdot)$ using the weighted conditional likelihood

$$L_n\big(G^w(x,\cdot)\big) = \prod_{i=1}^{n} \left\{ \int_0^1 \binom{m_i}{Y_i(x)} t^{Y_i(x)} (1-t)^{m_i - Y_i(x)}\, dG^w(x,t) \right\}^{w_i}. \tag{3.14}$$

Since multiplying all weights by a common constant does not change the maximizer of (3.14), one may equivalently use normalized sample weights. Analogous consistency and asymptotic normality results can be obtained by applying Theorems 3.6.1 and 3.6.3 to the weighted estimators, with G and H_α replaced by G^w and H_α^w. Verifying these conditions will generally require additional assumptions on the weights.

Below, we give two examples of weighted barycenters that may be of interest in applications.

Example 3.7.1 (Conditional barycenter)**.** *Let $Z \in \mathcal{Z} \subset \mathbb{R}$ be a continuous covariate, and let*

$$q^\star_\alpha(z) := \mathbb{E}\big[F^-_\nu(\alpha) \mid Z = z\big] \tag{3.15}$$

denote the α-quantile of the conditional Wasserstein barycenter given $Z = z$. A natural estimator can be obtained by local weighting (Hein, 2009). Let K be a kernel and let $K_h(u) = K(u/h)/h$. For the population target, define $w_{z,h}(Z) := K_h(Z - z)/\mathbb{E}\{K_h(Z - z)\}$. The resulting localized target is $q^\star_{\alpha,h}(z) := \mathbb{E}[w_{z,h}(Z)\, F^-_\nu(\alpha)]$, which is a weighted barycenter quantile of the form (3.11)*. In estimation, the population normalizing constant is replaced by its empirical analogue, giving sample weights $\hat{w}_{z,h}(Z_i) := K_h(Z_i - z)/\{\frac{1}{n}\sum_{j=1}^n K_h(Z_j - z)\}$. The localized target approximates $q^\star_\alpha(z)$ in* (3.15) *as $h \to 0$ under standard smoothing conditions.*

Example 3.7.2 (IPW barycenter)**.** *Suppose $Z = (W, A) \in \mathcal{Z}$, where $A \in \{0, 1\}$ is a treatment indicator and $W \in \mathbb{R}^d$ is a vector of baseline covariates. For $a \in \{0, 1\}$, let $q^{\star,(a)}_\alpha := \mathbb{E}[\mathbb{E}\{F^-_\nu(\alpha) \mid W, A = a\}]$, which is well defined under positivity of treatment assignment. Under standard causal identification assumptions (consistency and exchangeability), $q^{\star,(a)}_\alpha$ identifies the α-quantile of the counterfactual Wasserstein barycenter under assignment of all units to treatment level a, as studied by Lin et al. (2023). The statistical target $q^{\star,(a)}_\alpha$ can also be written as follows under positivity:*

$$q^{\star,(a)}_\alpha = \mathbb{E}\left[\frac{\mathbb{1}(A = a)}{\pi_a(W)}\, F^-_\nu(\alpha)\right], \qquad \pi_a(W) := P(A = a \mid W). \tag{3.16}$$

Thus, $q^{\star,(a)}_\alpha$ is a weighted barycenter quantile of the form (3.11)*, with weight $w_i := \mathbb{1}(A = a)/\pi_a(W_i)$.*

3.7.2 Alternative functionals of the distribution of quantiles

The representation in Theorem 3.4.1 also yields estimators for other functionals of the latent α-quantile distribution. Beyond the mean, a commonly used choice is the variance of the α-quantiles under Π:

$$v_\alpha^\star := \mathrm{Var}_{\nu\sim\Pi}\left(F_\nu^-(\alpha)\right) = \int_{\mathcal{I}} x^2\, dH_\alpha(x) - \left(\int_{\mathcal{I}} x\, dH_\alpha(x)\right)^2 = \int_{\mathcal{I}} x^2\, dH_\alpha(x) - \left(q_\alpha^\star\right)^2. \quad (3.17)$$

A simple plug-in estimator is

$$\hat{v}_{n,\alpha} := \int_{\mathcal{I}} x^2\, d\hat{H}_{n,\alpha}(x) - \left(\hat{q}_\alpha^{\mathrm{MCB}}\right)^2. \quad (3.18)$$

Integrating (3.17) over $\alpha \in (0,1)$ recovers the Fréchet variance under the 2-Wasserstein metric:

$$\mathbb{E}_{\nu\sim\Pi}\left[d_{W_2}^2(\bar{\nu}_\Pi, \nu)\right] = \int_0^1 v_t^\star\, dt. \quad (3.19)$$

Replacing $v_t^\star$ by the plug-in $\hat{v}_{n,t}$ gives an estimator of the Fréchet variance, providing a scalar summary of the heterogeneity of distributions under Π.

3.8 Simulation study

We conduct a simulation study to evaluate the operating characteristics of the MCB estimator and compare its performance with that of the empirical barycenter. We also examine the behavior of the estimator under misspecification of the parametric model for the mixing distributions.

Data are generated using a two-stage sampling procedure. In the first stage, we sample $n = 50, 200, 500$ distributions from Π. We consider two simulation settings for Π:

(A) A Dirichlet process with concentration parameter 2 and base measure given by a modified Kumaraswamy distribution scaled to $(0, 10)$, with shape parameters 0.2 and 0.6 (Sagrillo et al., 2021).

(B) A continuous mixture of two normal distributions. The first component has mean uniformly distributed on $(-3, 0)$ and standard deviation uniformly distributed on $(1, 2)$. The second component has mean uniformly distributed on $(2, 10)$ and standard deviation uniformly distributed on $(1, 2)$, with the mixture weight on the second component uniformly distributed on $(0.1, 0.2)$.

In the second stage, for each distribution sampled in the first stage, we draw m_i uniformly from $\{3, \ldots, 10\}$ and then sample m_i observations from that distribution.

We configure the MCB estimator to estimate the mixing distributions using a Beta model, splines with $df = 7$ (Efron, 2016; Narasimhan and Efron, 2020), or the NPMLE (Koenker and Gu, 2017). The Beta configuration is correctly specified for Simulation Setting A but not for Simulation Setting B, while the spline configuration is misspecified for both settings. To ease computation, we estimate each mixing distribution on 50 equally spaced cutpoints between the minimum and maximum values of the pooled observations in the simulated dataset. We estimate standard errors using the bootstrap with 500 replicates. For the Beta- and spline-configured MCB estimators, we construct Wald confidence intervals based on the bootstrap standard errors. Because theoretical results are not available for the NPMLE-configured estimator, we use the mean of the bootstrap estimates as the point estimate for stability and report percentile-based bootstrap confidence intervals for this estimator.

Across 500 simulation runs, we display the mean estimate, RMSE, mean standard error, and 95% confidence interval coverage for the estimated barycenter quantile function on the grid $0.01, 0.02, \ldots, 0.99$.

3.8.1 Simulation Setting A

Results for Setting A are displayed in Figure 3.6. Overall, the empirical barycenter performs poorly relative to the MCB estimators. It has larger RMSE across most quantile levels, and its confidence interval coverage deteriorates as n increases, indicating persistent bias that is not eliminated by increasing the number of units.

The RMSE results show the clearest advantage for the correctly specified Beta-configured MCB estimator. At $n = 500$, the empirical barycenter had average RMSE of 0.25 across quantile levels, compared with 0.014 for the Beta-configured MCB estimator, 0.066 for the spline-configured MCB estimator, and approximately 0.25 for the NPMLE-configured MCB estimator. Table 3.1 shows that the gains from the MCB estimators are most pronounced for tail quantiles, where bias in the empirical barycenter is especially large.

The coverage results tell a similar story. At $n = 500$, the empirical barycenter's coverage ranged from 0% to 97% and was well below the nominal 95% level for many quantile levels. In contrast, the Beta-configured MCB estimator achieved near-nominal confidence interval coverage across quantile levels and sample sizes, with coverage ranging from 90% to 98% at $n = 500$.

The spline-configured MCB estimator also substantially outperforms the empirical barycenter, despite being misspecified. It has lower RMSE and confidence interval coverage closer to the nominal level, although its coverage decreases with increasing n, as expected under model misspecification. Although we did not study the theoretical properties of the NPMLE-configured MCB estimator, it exhibits close-to-nominal percentile-based confidence interval coverage. Its RMSE, however, exceeds that of the empirical barycenter for central quantile levels. Among the MCB estimators, standard errors are generally largest for the NPMLE configuration and smallest for the Beta configuration.

Overall, these results suggest that, when the parametric submodel is correctly specified or reasonably flexible, the MCB estimator improves RMSE and confidence interval coverage relative to the empirical barycenter under sparse sampling.

3.8.2 Simulation Setting B

Results for Setting B are displayed in Figure 3.7. As in Setting A, the empirical barycenter has substantially larger RMSE than the MCB estimators, with the largest errors occurring in the tails. At $n = 500$, the empirical barycenter had an average RMSE of 0.61, compared with 0.24 for the Beta-configured MCB estimator, 0.16 for the spline-configured MCB estimator,

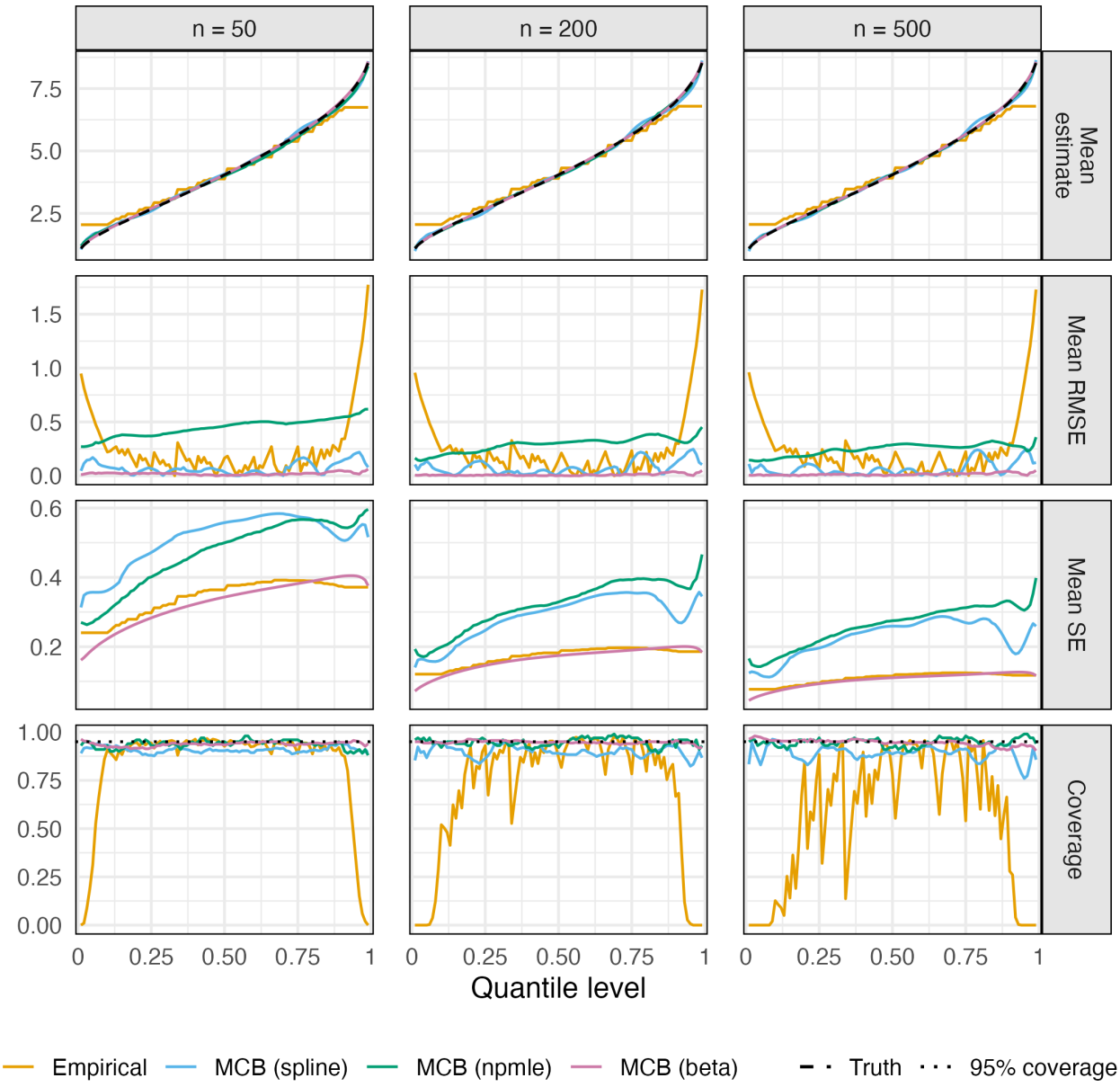


Figure 3.6: Results for Simulation Setting A. Rows display, from top to bottom, the mean estimate, RMSE, mean standard error, and 95% confidence interval coverage for all quantile levels across simulation iterations. Wald confidence intervals are shown for the empirical barycenter, the Beta-configured MCB estimator, and the spline-configured MCB estimator, while percentile confidence intervals are shown for the NPMLE-configured MCB estimator.

Table 3.1: Average, tail, and center RMSE by simulation setting for $n = 500$. Average RMSE is averaged over all quantiles $\alpha \in \{0.01, 0.02, \ldots, 0.99\}$. Lower tail, center, and upper tail RMSE are averaged over quantiles $\alpha \in \{0.01, \ldots, 0.05\}$, $\alpha \in \{0.06, \ldots, 0.94\}$, and $\alpha \in \{0.95, \ldots, 0.99\}$, respectively.

Setting	Estimator	Average RMSE (0.01–0.99)	Lower tail RMSE (0.01–0.05)	Center quantiles RMSE (0.06–0.94)	Upper tail RMSE (0.95–0.99)
A	Empirical barycenter	0.250	0.748	0.166	1.254
	MCB-Beta	0.014	0.007	0.014	0.025
	MCB-Spline	0.066	0.078	0.059	0.184
	MCB-NPMLE	0.245	0.141	0.250	0.273
B	Empirical barycenter	0.605	3.223	0.431	1.077
	MCB-Beta	0.237	0.211	0.241	0.205
	MCB-Spline	0.164	0.692	0.121	0.399
	MCB-NPMLE	0.205	0.264	0.198	0.264

and 0.21 for the NPMLE-configured MCB estimator. As with Simulation Setting A, Table 3.1 shows that this improvement is especially pronounced for tail quantiles, where the empirical barycenter has particularly large RMSE.

Unlike Setting A, both the Beta- and spline-configured MCB estimators are misspecified in Setting B. Despite this misspecification, they still improve point estimation relative to the empirical barycenter, with the spline configuration achieving the lowest average and center-quantile RMSE. Their confidence interval coverage, however, is poor and generally worsens as n increases, as expected when the fitted mixing model is misspecified. The NPMLE-configured MCB estimator achieves close-to-nominal percentile-based confidence interval coverage across nearly all quantile levels, albeit with larger standard error estimates. Overall, these results show that the MCB estimator can greatly improve point estimation even under mixing model misspecification, but that coverage can be poor because the estimator converges to a misspecified limiting target rather than the true barycenter.

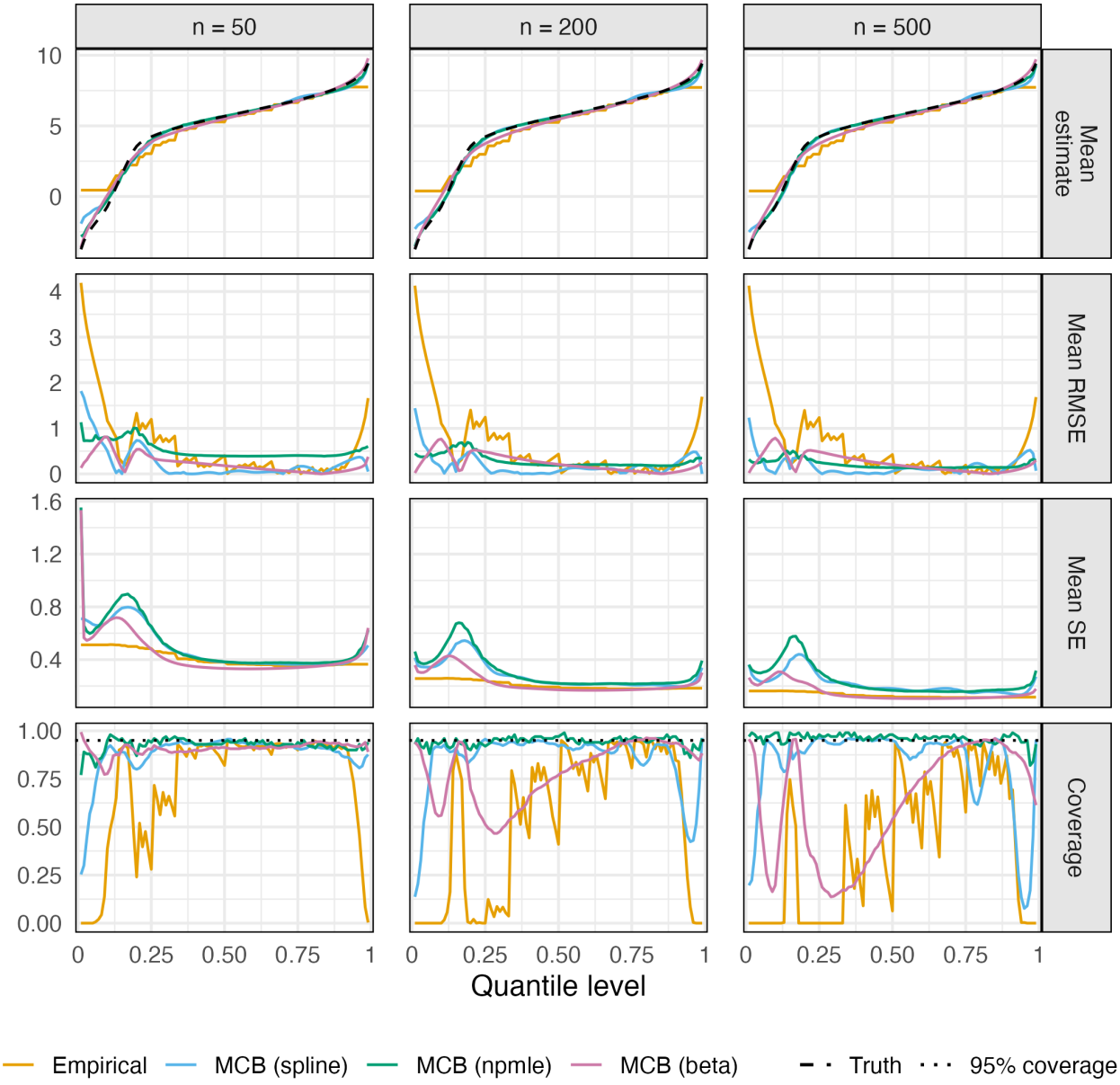


Figure 3.7: Results for Simulation Setting B. Rows display, from top to bottom, the mean estimate, root mean squared error (RMSE), mean standard error, and 95% confidence interval coverage for all quantile levels across simulation iterations. Wald confidence intervals are shown for the empirical barycenter, the Beta-configured MCB estimator, and the spline-configured MCB estimator, while percentile confidence intervals are shown for the NPMLE-configured MCB estimator.

3.9 Data application

We illustrate the proposed method using HIV-1 sequence data from the HVTN 502 (Step) and HVTN 503 (Phambili) vaccine efficacy trials. Additional illustrative data applications are included in Appendix C.3.

Both the Step and Phambili trials were randomized placebo-controlled phase 2b trials of the Merck MRKAd5 HIV-1 subtype B Gag/Pol/Nef vaccine (Gray et al., 2010). The Step trial was conducted in North America, the Caribbean, South America, and Australia and enrolled 3,000 participants; vaccinations were discontinued after an interim analysis showed no efficacy against HIV-1 acquisition (Buchbinder et al., 2008). The Phambili trial evaluated the same vaccine in South Africa, a predominantly subtype C epidemic setting, and was halted early following the Step results (Gray et al., 2011).

Our analysis uses viral sequence data from participants who acquired HIV infection and whose viral sequences were sampled near the time of diagnosis. In the Step trial, sequence data were available from 65 participants (39 vaccine, 26 placebo), with an average of 6.58 sequences per participant (range 1–12). In the Phambili trial, sequence data were available from 43 participants (23 vaccine, 20 placebo), with an average of 6.51 sequences per participant (range 2–12). Within each trial, all viral sequences were aligned using a multiple sequence alignment, and the distances described below were computed from the aligned sequences. We summarize each viral sequence with two numeric features: (1) the physicochemically weighted mismatch distance between the sequence and the vaccine insert across positions 77–85 of the Gag protein, using the HXB2 numbering system, at which SLYNTVATL is the vaccine-strain amino acid sequence and represents a well-known epitope that elicits HIV-1-specific CD8+ T-cell responses; and (2) the same metric computed over the entire Gag protein. For each participant, the subject-specific distribution of a given feature represents the within-host distribution of viral distances from the vaccine insert. Our objective is to compare the barycenters of these subject-specific distance distributions between the vaccine and placebo arms. Because these distances measure how similar the infecting

viral populations are to the vaccine insert, differences between treatment-arm barycenters may suggest that vaccination was associated with infection by viruses with different sequence characteristics.

The MRKAd5 vaccine encoded Gag, Pol, and Nef and was intended to elicit cell-mediated immune responses, particularly CD8+ T-cell responses against epitopes in these proteins. We focus on Gag because it was an important intended target of these responses, with natural history studies associating Gag-specific cellular immune responses with beneficial outcomes. Prior analyses of viral sequences from the Step and Phambili trials also found evidence of vaccine-associated sequence divergence in Gag (Rolland et al., 2011; Hertz et al., 2016). These prior analyses reduced the observed sequences for each participant to a single representative sequence. We revisit the analysis while retaining all observed sequences and treating them as samples from the participant-specific viral population. In our data application, the SLYNTVATL epitope distance and the full-Gag weighted mismatch distance were included to capture divergence at two scales: within a well-known CD8+ T-cell epitope and across the entire Gag protein. Within the resulting distance distributions, the upper tail represents the minority of sequences that are most divergent from the vaccine insert, whereas the lower tail represents those that are most similar. Differences between treatment arms that are concentrated in either tail may therefore reveal vaccine-associated patterns affecting only a subset of the within-host viral population, which could be missed when each participant is represented by a single sequence.

We analyzed the Step and Phambili trials separately. Within each trial, we applied the MCB estimator separately to the vaccine and placebo arms, yielding an estimated barycenter quantile function for each arm. We then compared the vaccine- and placebo-arm barycenters to assess whether the distribution of sequence-derived distances differed between treatment groups. We estimated these barycenter quantile functions on the grid of quantile levels $\alpha = 0.01, 0.02, \ldots, 0.99$ using the Beta-configured MCB estimator, with sensitivity analyses using the spline and NPMLE configurations. The Beta configuration was used because it provides more stable estimates in the presence of small sample sizes. We obtain pointwise

confidence intervals across the quantile grid using the nonparametric bootstrap with $B = 500$ replications and Wald-type intervals based on the bootstrap standard errors.

The estimated barycenter quantile functions are displayed in Figure 3.8 together with 95% pointwise confidence intervals. For the SLYNTVATL feature, the estimated barycenter quantile functions are nearly flat, indicating generally limited within-host heterogeneity in this sequence feature. The MCB estimator and the empirical barycenter closely match across all quantile levels. This agreement is expected when the unit-level distributions are close to point masses: in such settings, even sparsely sampled empirical quantiles can provide accurate estimates of the latent quantiles, leaving little bias for the MCB estimator to correct. There is also a larger difference between the vaccine and placebo barycenters in the Step trial than in the Phambili trial.

For the full Gag protein feature, the estimated quantile functions show greater within-host heterogeneity. In this setting, the MCB estimator and empirical barycenter agree closely in the central quantile range, but differ noticeably in the tails. This pattern is consistent with the simulation results, where sparse-sampling bias was most pronounced for tail quantiles and where the MCB estimator provided the largest correction. For the full Gag protein feature, the vaccine and placebo barycenters appear more separated in the Phambili trial than in the Step trial.

Sensitivity analyses with the spline-configured and NPMLE-configured MCB estimators are shown in Supplementary Figures C.1 and C.2.

3.10 Discussion

To our knowledge, this is the first paper to propose a method that directly corrects for sparse sampling bias in estimating the Wasserstein barycenter of one-dimensional distributions. We approached the problem by first considering the seemingly more ambitious task of estimating the distribution of the latent quantiles across units, from which the barycenter quantile can be recovered as the mean. This distribution can be represented in terms of marginal laws of unit-level CDF values, which arise as mixing distributions in binomial mixture models. This

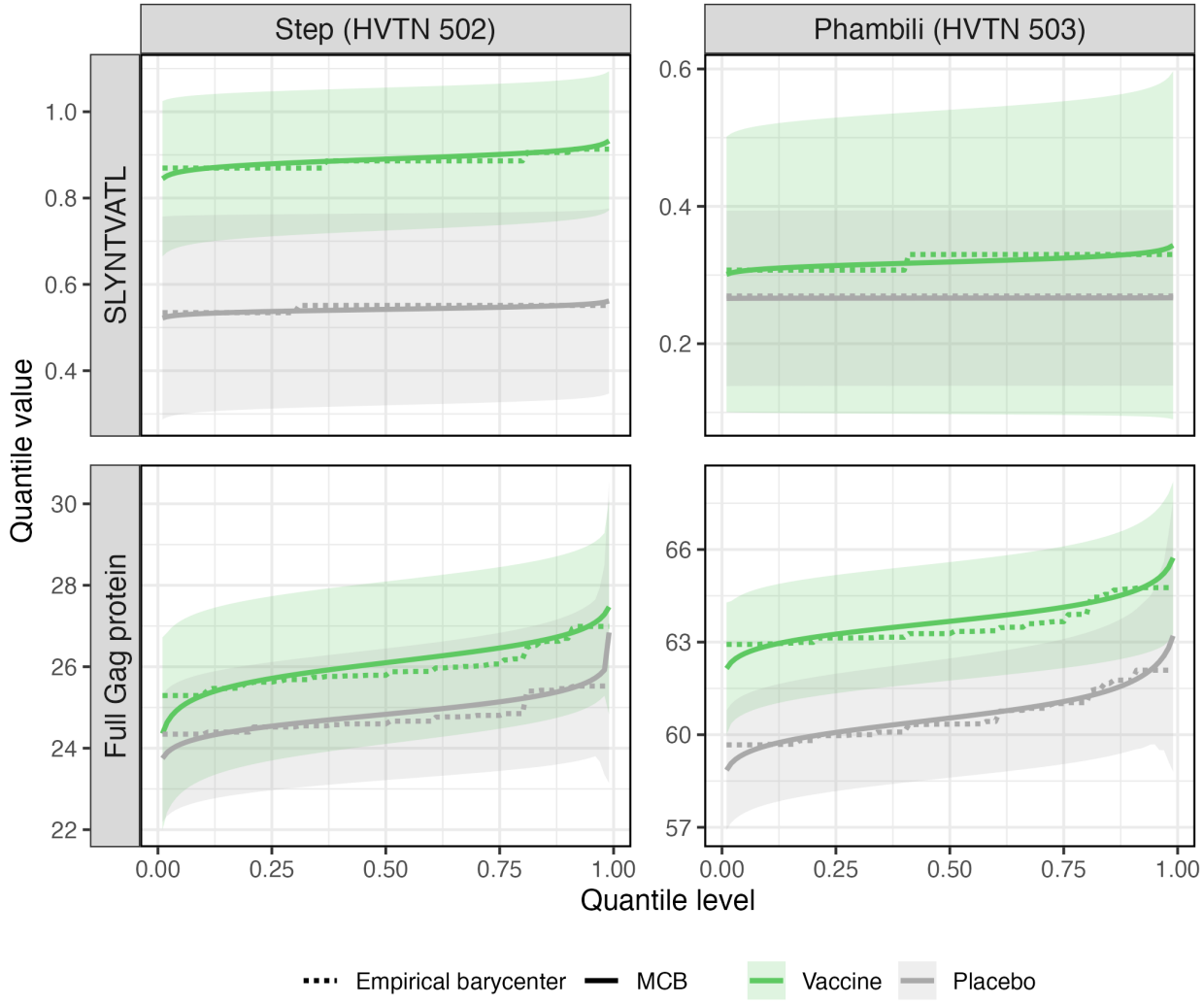


Figure 3.8: Estimated barycenter quantile functions for the SLYNTVATL epitope distance and full Gag protein physicochemical-weighted Hamming distance in the Step and Phambili trials, stratified by treatment arm. The MCB estimator along with its pointwise 95% confidence intervals is displayed with a solid line, while the empirical barycenter is shown with a dotted line.

representation leads to a natural plug-in estimator: we first estimate the relevant binomial mixing distributions and then plug these estimates into the representation formula. Our method can accommodate any suitable estimator of the mixing distributions, allowing it to draw on the extensive existing literature on binomial mixture estimation. In simulations, the proposed approach performed well even when the parametric submodels were misspecified.

Our results suggest that a distributional analysis can also be performed in hierarchical data settings, where the outcomes are typically treated as repeated measurements instead of distributions. Such data are traditionally analyzed using generalized estimating equations or mixed models, which focus on mean effects across groups or covariates (Raudenbush and Bryk, 2002; Gelman, 2007). Distributional methods could in principle be used in these settings because the repeated measurements within each unit can be viewed as i.i.d. draws from a latent unit-level distribution. However, distributional methods are rarely used in these settings because these methods typically assume that each unit-level distribution can be estimated accurately (i.e., requiring a large number of repeated measurements). Our framework relaxes this requirement and allows for inference on the Wasserstein barycenter of a population of unit-level distributions. Comparing barycenters across groups can reveal differences in features, such as tail behavior or spread, that may be missed by analyses focused on means. In Appendix C.3, we illustrate the utility of our method in these settings by reanalyzing three classic data examples taken from hierarchical data analysis textbooks.

One caveat of the proposed approach is that structural assumptions are needed for estimation. These structural assumptions are of a semiparametric nature: we make parametric assumptions on the binomial mixing distributions, which does not fully specify the distribution on distributions Π. Without these structural assumptions, the binomial mixing distributions underlying the marginal construction are not identifiable. Misspecification of these parametric models is a concern, and while the NPMLE-configured MCB estimator is an appealing alternative, its theoretical properties in the MCB procedure are not fully understood. An alternative direction for future work is to use the representation result to construct nonparametric bounds on barycenter quantiles, avoiding the need to impose

structural assumptions on the mixing distributions.

Finally, additional theory is needed for estimation of the marginal mixing distributions across x. The current implementation estimates the mixing distribution pointwise over a grid of x values and then enforces monotonicity through isotonic regression for each quantile level $\alpha \in (0, 1)$ separately. This procedure is convenient because it can use available binomial mixture estimators, but is unlikely to be optimal because it does not impose the monotonicity constraint directly nor globally. We are not aware of any existing theory and methods that could be directly applied to this problem.

3.11 Acknowledgments

We thank Zhenhua Lin, Thomas Richardson, and Michal Juraska for helpful discussions, and Craig Magaret for preparing the data used in the data application. The authors thank the participants and investigators of HVTN 502 and 503. Research reported in this publication was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health, award number R37AI054165 and U.S. Public Health Service Grant AI068635. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Florian Stijven was supported by the Agentschap Innoveren & Ondernemen (VLAIO) and Johnson & Johnson Innovative Medicine through a Baekeland Mandate [grant number HBC.2022.0145].

Chapter 4

TWO-SAMPLE INFERENCE FOR SEQUENCE-DERIVED FEATURE DISTRIBUTIONS FROM DEEP SEQUENCING DATA

James Peng, Florian Stijven, Michal Juraska, Allan C. deCamp, Elena E. Giorgi, Linbo Wang, Peter B. Gilbert

Chapter Summary

In Chapter 2, we discussed a method for analyzing binary marks in sieve analyses with deep sequencing data. In this chapter, we consider continuous numeric marks, such as Hamming distance to a reference sequence. We take the simpler case-only approach described as "Approach 2" in the Introduction, directly comparing the distributions of numeric marks between study arms. When a numeric feature is computed for each sequence, the collection of feature values defines a one-dimensional distribution that characterizes individual-level viral diversity with respect to that feature. Past analyses have typically reduced each individual's distribution to a single summary value, such as the mean or mode. Instead, we propose a framework that retains the full individual-level distribution as the object of analysis. Drawing on existing tools from distributional data analysis, our framework uses Wasserstein barycenters to aggregate individual-level distributions within independent arms, compares the resulting barycenters using differences in their quantile functions, and provides hypothesis tests for equality of the Wasserstein barycenters. Through simulation studies, we examine when low or differential sequencing depth can distort inference and describe practical strategies for addressing these issues. We then illustrate the framework using data from the HIV-1 AMP efficacy trials and a human B cell receptor sequencing study of vaccine-

elicited immune responses. By using the entire collection of sequences for each individual, our approach can reveal arm differences that may be obscured by single-number summaries and can characterize how those differences vary across the feature distribution.

4.1 Introduction

In Chapter 2, we considered binary marks, such as whether a sequence matches or mismatches a reference sequence at a particular amino acid site. In this chapter, we consider continuous numeric marks, such as Hamming distance to a reference sequence or a sequence-specific phenotypic property. We focus on the case-only comparison described as "Approach 2" in the Introduction, in which the distribution of marks is compared directly between study arms. In the HIV-1 prevention trial setting, this comparison is restricted to participants who acquire HIV-1 and therefore does not directly estimate mark-specific prevention efficacy. However, it provides a simpler way to assess whether the viral populations observed among infected participants differ between study arms. Although motivated by sieve analyses, the framework applies more broadly to two independent group comparisons of sequencing data; we include a separate application to a study of vaccine-elicited B cell receptor responses.

With deep sequencing data, each participant contributes multiple sequences, and a numeric feature can be computed for each sequence. The resulting collection of feature values defines an individual-level distribution. Past analyses have often collapsed these data into a single representative sequence or summarized the feature distribution by a single scalar value, such as the mean or mode. This reduction simplifies treatment arm comparisons by allowing for classical two independent sample procedures, such as comparing arm means via a t-test. In contrast, we propose retaining each individual's complete collection of feature values to more fully characterize individual-level molecular diversity. Thus, the primary outcome for each individual is the distribution itself.

Because distributional observations cannot be averaged, differenced, or tested in the same way as real-valued observations, analyzing them requires additional statistical tools. A growing body of literature has developed tools for data in which the observational unit is a

distribution (Agueh and Carlier, 2011; Chen et al., 2023; Lin et al., 2023). A central object in this work is the *Wasserstein barycenter*, which serves as a notion of an average for a set of distributions. Our aim is to translate ideas from this literature into a practical framework for biological sequencing studies. We describe how to compare treatment arms by treating each individual's sequence-derived features as a distribution, summarizing each treatment arm with a Wasserstein barycenter, and evaluating arm differences through pointwise differences in their barycenter quantile functions. Furthermore, we develop hypothesis tests for the equality of arm barycenters, designed to detect different classes of alternatives that may arise in practice. Throughout, we will study the impact of varying sequencing depth on our framework and provide practical guidance for mitigating this potential issue. The proposed methods are implemented in an easy-to-use R package, `mcbarycenter`.

To our knowledge, Wasserstein barycenter-based distributional methods have not previously been applied to the analysis of deep sequencing data. A related paper applies a similar framework in genetics research of single-cell differential expression (Zhang and Guo, 2022). Their method uses Wasserstein barycenter techniques to analyze single-cell sequencing data in a two-sample problem, comparing cases versus controls. While relying on similar ideas, our proposal differs in a few ways: first, we explicitly consider the role of low or differential sequencing depth in producing biased estimates. Second, we consider an estimand—the difference in quantiles—that contrasts the Wasserstein barycenters of the two treatment arms, whereas they consider the Wasserstein distance between the two barycenters as a single summary measure. Third, we approach hypothesis testing differently. Their permutation testing procedure evaluates the null of exchangeability between two groups, which may be automatically violated if the distribution of sequencing depths differs systematically between groups. In contrast, our testing approach is valid even when sequencing depths differ between arms.

The remainder of the chapter is organized as follows. Section 4.2 provides a high-level introduction to the statistical tasks for our proposed analytic approach and presents illustrative toy examples to build intuition. Section 4.3 describes the data structure, estimands, and methodology for estimating the treatment arm-specific Wasserstein barycenters and testing

the hypothesis of equal barycenters between arms. Section 4.4 presents results from simulation studies. Section 4.5 presents the framework applied to the AMP HIV-1 prevention trials and human B cell receptor sequencing study of vaccine-elicited immune responses. Section 4.6 concludes.

4.2 Statistical tasks with distributional data

To build intuition before describing the formal data structure and methodology, we first introduce the framework's statistical tasks and provide illustrative toy examples. These tasks include:

Task 1: Average distribution. First, we define what constitutes a "central" or "average" distribution across individuals. We use the *Wasserstein barycenter*, which provides a notion of averaging distributions under optimal transport geometry (Agueh and Carlier, 2011; Bigot, 2020). Compared to pointwise averages of density functions, the Wasserstein barycenter better preserves the salient features of component distributions. In Figure 4.1(a), we display the densities of five bimodal distributions of the same shape along with their Wasserstein barycenter; as shown, the barycenter has the same bimodal shape as the five input distributions. Additionally, for one-dimensional distributions, the Wasserstein barycenter has a convenient and interpretable representation: its quantile function is equal to the average of the quantile functions of its component distributions (Bigot and Klein, 2018) (Figure 4.1(b)). Given this property, we frame our presentation around quantile functions of distributions rather than their densities. In practice, we summarize the within-host feature distributions for each treatment arm using their mean quantile function—or equivalently, the quantile function of their Wasserstein barycenter.

Task 2: Differences between barycenter distributions. After estimating the barycenter for each treatment arm, we quantify how the barycenter distributions differ between arms. Analyses that reduce a distribution to a single scalar—such as the mean or mode—often miss

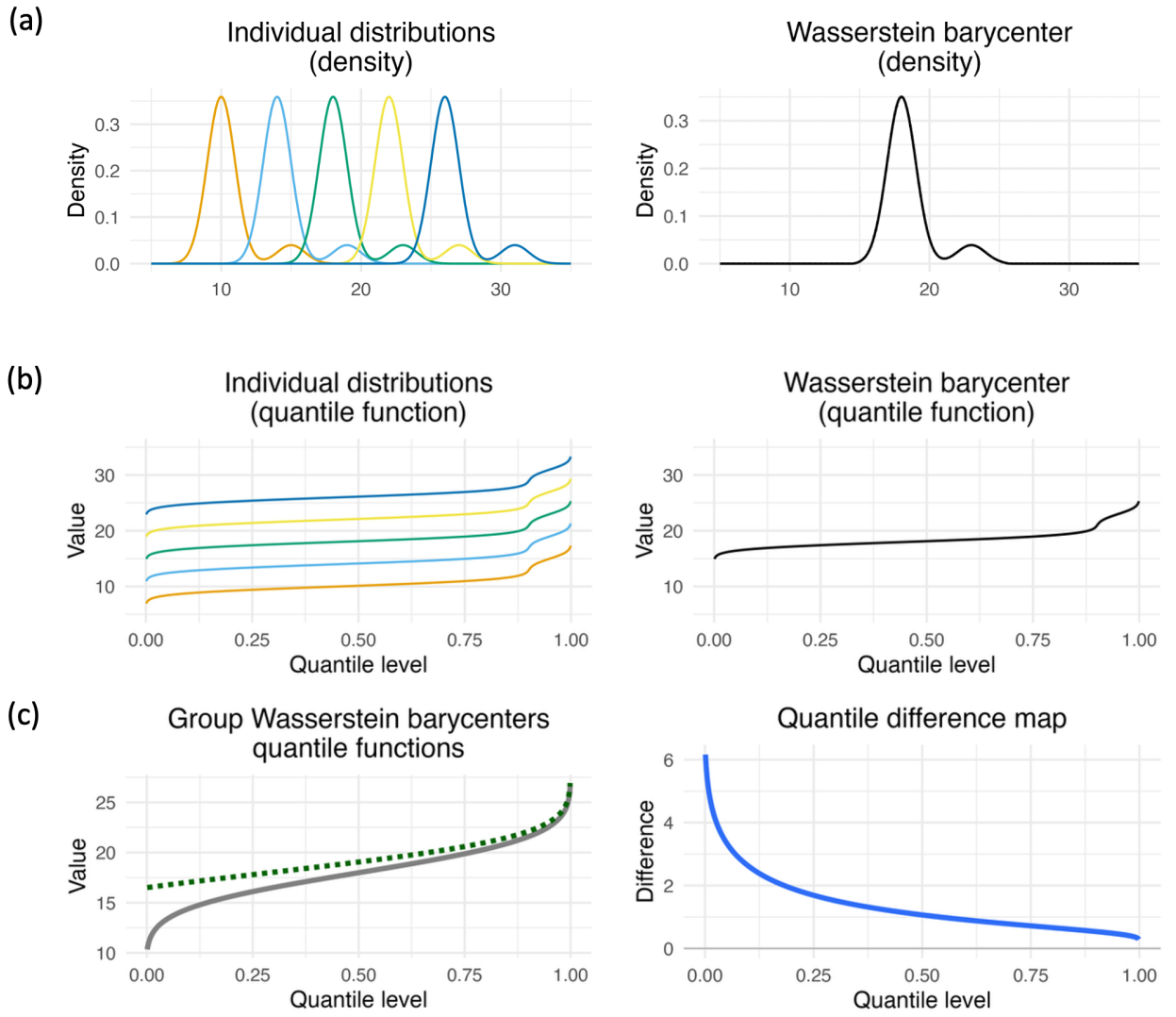


Figure 4.1: (a) Densities of five bimodal distributions and the density of their Wasserstein barycenter. (b) Quantile functions of the same five bimodal distributions and the quantile function of their Wasserstein barycenter. The Wasserstein barycenter quantile function is obtained by taking the pointwise mean of the input quantile functions, (c) Two arm-specific Wasserstein barycenter quantile functions (Arm 0: solid gray; Arm 1: dotted green), and their associated quantile difference map (comparing Arm 1 vs Arm 0).

important differences in shape or tail behavior. Instead, we employ a *quantile difference map*, which compares two distributions by evaluating the differences between them across all quantile levels in $(0, 1)$. Comparing quantiles is a well-established method for contrasting distribution functions (Doksum, 1974), interpreting quantile regression (Koenker and Bassett, 1978), and studying causal treatment effects for both real-valued and distributional outcomes (Firpo, 2007; Lin et al., 2023; Katta et al., 2024).

In Figure 4.1(c), we display a toy example of the Wasserstein barycenter quantile functions for two treatment arms (Arm 0 and Arm 1), together with the corresponding quantile difference map. In this example, the 0.01 quantile of the Arm 1 barycenter, representing a near-minimum sequence feature value, is approximately six units higher than that of Arm 0. In contrast, the median feature value in Arm 1 is only approximately one unit higher than in Arm 0. In the upper tail, the two barycenters converge, with their maximum feature values being approximately equal. Therefore, the comparison across all quantiles allows us to more thoroughly characterize differences across the entire barycenter distributions, including in the tails.

Task 3: Hypothesis testing for equal Wasserstein barycenters. Finally, we require a hypothesis test to determine whether the Wasserstein barycenters differ between treatment arms. Specifically, we test the null hypothesis that the Wasserstein barycenters of the within-host viral feature distributions are identical in the arms. Equivalently, this null asserts that the difference in barycenter quantiles between the two arms is zero across all quantile levels in $(0, 1)$.

By representing the barycenter distribution via its quantile function, this problem can be viewed through the lens of functional data analysis, where mean functional tests are well established (Horváth and Kokoszka, 2012). For distributional data specifically, two-sample tests have been explored by a handful of past works. Dubey and Müller (2019) propose an ANOVA-type test evaluating the equality of barycenters and Fréchet variances; Zhang and Guo (2022) use a permutation test for the strong null of exchangeability between groups;

and Jeong et al. (2026) derive the asymptotic limiting distribution of a scaled, squared Wasserstein distance test statistic. Additionally, Petersen et al. (2021) examine testing in a Wasserstein regression context, which, in its simplest form, can be reduced to a two-sample barycenter problem. Our approach aligns most closely with Jeong et al. (2026). However, we propose more flexible test statistics designed to target various alternative data scenarios that could arise in practice.

4.2.1 Illustrative toy examples

To build intuition, Figure 4.2 illustrates four toy scenarios alongside the corresponding outputs from our framework, which we refer to as a *distributional analysis*. In each scenario, the left panel shows the individual-level feature distributions and their treatment arm-specific Wasserstein barycenters, the middle panel shows the corresponding quantile functions, and the right panel shows the resulting difference in barycenter quantiles between arms. We contrast this approach with a standard *mean-based analysis*, which relies on a two-sample mean comparison after computing the average feature value for each individual's sequence set.

(a) *Similar distributions (Figure 4.2(a)).* Individuals in Treatment Arm 0 exhibit sequence feature distributions similar to those in Arm 1, resulting in similar barycenters. The distributional analysis estimates negligible differences between the two arm barycenters, producing a quantile difference curve nearly flat at 0 across all quantile levels. Similarly, the mean-based analysis estimates a near-zero difference between arm means.

(b) *Uniform differences (Figure 4.2(b)).* Treatment Arm 1 individuals generally exhibit higher sequence feature values than Arm 0, resulting in a uniform difference between the two arm barycenters across quantile levels. Consequently, the distributional analysis yields a consistently positive difference across all quantiles. The mean-based analysis similarly estimates a positive difference in means of roughly the same magnitude.

(c) *Upper-tail difference (Figure 4.2(c)).* For most quantile levels, the Treatment Arm 1

and Arm 0 barycenters are similar, but they diverge at the upper tail. Specifically, the Arm 1 barycenter has elevated upper-tail feature values that are absent from the Arm 0 barycenter. The distributional analysis captures this by producing near-zero estimates for most quantile levels but a sharply positive difference at the upper tail. In contrast, the mean-based analysis yields an attenuated estimate that averages this tail effect across the entire distribution.

(d) *Difference in variability (Figure 4.2(d)).* The Treatment Arm 1 barycenter exhibits less variability than the Arm 0 barycenter, despite sharing the same overall mean. Here, the distributional analysis reveals a positive difference at the lower tail, a negative difference at the upper tail, and an approximately zero difference at the 0.5 quantile. In contrast, the mean-based analysis fails to capture this trend, estimating a null difference between arm means.

In scenarios (a) and (b), both analytical approaches yield similar results and may have comparable power to detect treatment arm differences. In scenario (c), however, the distributional approach can be more sensitive to the upper-tail difference and also characterizes a feature of the barycenter difference that the mean-based analysis obscures. Finally, in scenario (d), the mean-based approach will fail to detect the difference in variability, whereas the distributional approach can capture this difference through the quantile difference map. Together, these examples highlight two advantages of our proposal: it can have power to detect distributional differences that mean-based analyses may miss, and it provides a more resolved view of where in the distribution those differences manifest.

Remark 4.2.1. *In this chapter, we directly compare treatment arms among participants who acquired HIV-1, conditioning on infection. We refer to this comparison as the net sieve contrast. Although conditioning on infection is part of the definition of this estimand, the contrast should not be interpreted as a causal effect within a fixed population because infection is a post-randomization event that may be affected by treatment assignment. In Appendix D.2, we show that, under additional assumptions, the net sieve contrast can be decomposed into*

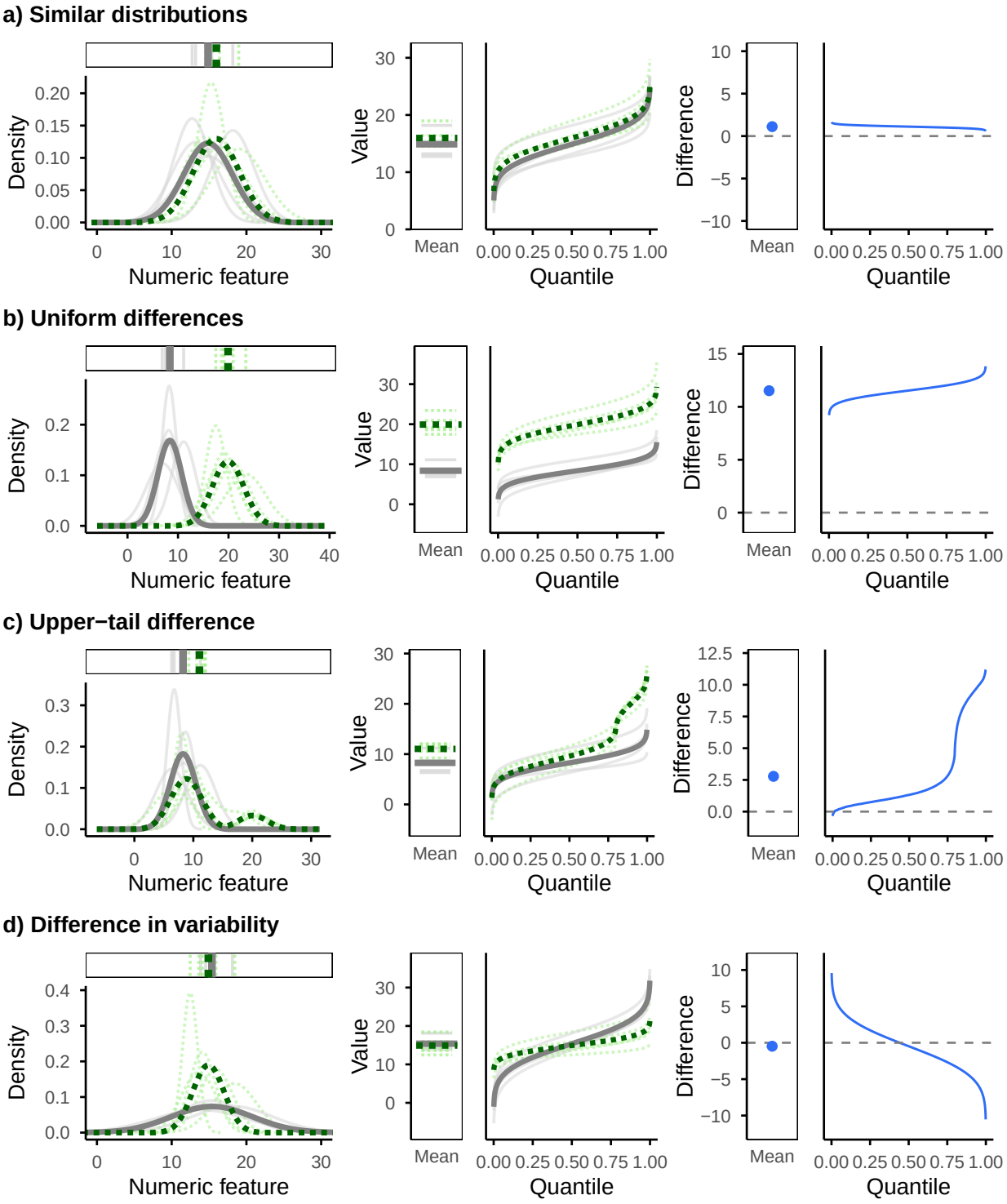


Figure 4.2: Output of a distributional analysis for four hypothetical toy datasets. In each scenario, there are five participants in each arm. In the left panels, light curves show the densities of numeric features for each individual (Arm 0, gray solid; Arm 1, green dotted), while bold curves represent the arm-specific Wasserstein barycenters. The middle panels show the corresponding individual-level and barycenter quantile functions. The right panels display the difference in quantiles between the Arm 1 and Arm 0 barycenters. Panels illustrate the following scenarios: (a) similar distributions, (b) approximately uniform differences, (c) upper-tail difference, and (d) difference in variability.

a breakthrough/post-acquisition component and a prevention sieve component, which have distinct biological interpretations. These individual components are not identified without further assumptions.

4.3 Methodology

4.3.1 Data setup

For each individual i, let $Z_i \in \{0,1\}$ denote the treatment arm indicator and Y_i denote the individual's underlying within-host distribution of a numeric sequence feature. In this chapter, we implicitly condition on HIV-1 infection ($E_i = 1$). Note that Y_i, which is a realization of a *distribution-valued random variable*, is not directly observed; we instead observe m_i feature measurements $Y_{i1}, \ldots, Y_{im_i}$ sampled from this distribution, where m_i denotes the sequencing depth. We assume that, conditional on Y_i, $Y_{i1}, \ldots, Y_{im_i} \overset{\text{i.i.d.}}{\sim} Y_i$. (See Remark 4.3.1 for a brief discussion of this assumption.) We allow the distribution of sequencing depth m_i to differ in each arm, so that $m_i \mid Z_i = z \sim \eta_z$ for $z \in \{0,1\}$, where η_z denotes the distribution of sequencing depths for Arm z. In total, we observe n individuals in our dataset, with n_0 in Arm 0 ($Z = 0$) and n_1 in Arm 1 ($Z = 1$).

From the observed measurements for individual i, we can construct the empirical distribution $\widehat{Y}_i^{\text{emp}} = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta_{Y_{ij}}$. We also consider a smoothed version of this empirical distribution, denoted $\widehat{Y}_i^{\text{sm}}$, obtained by applying a smoothing procedure to $\widehat{Y}_i^{\text{emp}}$, such as kernel density smoothing or smoothing of the quantile function (Bigot et al., 2018). Both $\widehat{Y}_i^{\text{emp}}$ and $\widehat{Y}_i^{\text{sm}}$ can be viewed as error-prone proxies for Y_i, with the amount of error depending on the sequencing depth m_i. Because the sequencing-depth distributions η_0 and η_1 may differ between treatment arms, the amount of measurement error may also differ between arms. As shown in the following sections, this measurement error can bias estimation of the target estimands.

For a generic probability measure λ, let $F_\lambda(y)$ denote its cumulative distribution function for $y \in \mathbb{R}$, and let $F_\lambda^-(u)$ denote its quantile function for $u \in (0,1)$. Because the individual-level outcome is distribution-valued, we measure distances between distributions using the

2-Wasserstein metric. In optimal transport, the Wasserstein distance quantifies the minimum cost of transporting probability mass from one distribution to another; in the case of the 2-Wasserstein distance on Euclidean space, this cost is based on squared Euclidean distance. For one-dimensional distributions, the 2-Wasserstein distance between λ_1 and λ_2 has a convenient representation in terms of their quantile functions:

$$W_2(\lambda_1, \lambda_2) := \left(\int_0^1 \left| F^-_{\lambda_1}(u) - F^-_{\lambda_2}(u) \right|^2 du \right)^{1/2}. \tag{4.1}$$

Using this metric, we define the Wasserstein barycenter of the distribution-valued random variable Y as the probability measure $\mathbb{B}Y$ that minimizes the expected squared Wasserstein distance between a random realization of Y and a candidate probability measure ν:

$$\mathbb{B}Y := \arg\min_{\nu} \mathbb{E}\left[W_2^2(Y, \nu) \right],$$

where the minimum is taken over probability measures ν, and the expectation is taken with respect to the randomness in Y. We denote the corresponding quantile function by $F^-_{\mathbb{B}Y}(u)$. More generally, for a generic random variable X, we define the conditional Wasserstein barycenter of Y given X by

$$\mathbb{B}[Y \mid X] := \arg\min_{\nu} \mathbb{E}\left[W_2^2(Y, \nu) \mid X \right].$$

By the one-dimensional representation of the Wasserstein distance in (4.1), the quantile function of the Wasserstein barycenter is given by the pointwise mean of the quantile functions of the component distributions:

$$F^-_{\mathbb{B}Y}(u) = \mathbb{E}\left[F^-_Y(u) \right], \qquad u \in (0, 1). \tag{4.2}$$

Because of this representation, we frame our methodology around estimating the Wasserstein barycenter's quantile function. We consider two estimands of interest and a correspond-

ing hypothesis test. In Task 1, we wish to estimate the treatment arm-specific Wasserstein barycenter quantile function

$$\mu_z(u) := F^-_{\mathbb{B}[Y|Z=z]}(u), \qquad u \in (0,1), \quad z \in \{0,1\}. \tag{4.3}$$

In Task 2, we estimate the quantile difference map, which compares the quantiles of the treatment arm-specific Wasserstein barycenters between the two arms:

$$\Delta(u) := \mu_1(u) - \mu_0(u), \qquad u \in (0,1). \tag{4.4}$$

Finally, in Task 3, our hypothesis test evaluates the global null hypothesis that the two treatment arm-specific barycenter quantile functions are equal at every quantile level within a compact set $\mathcal{U} \subset (0,1)$:

$$H_0 : \Delta(u) = 0 \text{ for all } u \in \mathcal{U}. \tag{4.5}$$

Remark 4.3.1 (Simple random sample assumption)**.** *The assumption* $Y_{i1}, \ldots, Y_{im_i} \overset{\text{i.i.d.}}{\sim} Y_i$ *may not hold depending on the sequencing technology and processing procedure used. For example, PCR resampling, amplification bias, sequencing error, and bioinformatic filtering can cause the observed sequences to deviate from independent draws from the biological viral population. In such settings,* Y_i *should be interpreted as the distribution of sequence-derived features after the sequencing and processing procedure, rather than the full biological viral population. Molecular barcoding approaches, such as Primer ID and related unique molecular identifier-based methods, can reduce PCR resampling artifacts and sequencing errors, improving the extent to which observed sequences reflect the viral template population present in the biological sample (Zhou et al., 2022).*

4.3.2 Task 1: Estimation and inference for the Wasserstein barycenter

We first aim to estimate the treatment arm-specific barycenter quantile functions $\mu_0(u)$ and $\mu_1(u)$. We first consider two estimators based on the empirical distributions $\widehat{Y}_i^{\text{emp}}$ and $\widehat{Y}_i^{\text{sm}}$

introduced above. Let

$$\widehat{q}_i^{\,\mathrm{emp}}(u) := F^{-}_{\widehat{Y}_i^{\mathrm{emp}}}(u), \qquad \widehat{q}_i^{\,\mathrm{sm}}(u) := F^{-}_{\widehat{Y}_i^{\mathrm{sm}}}(u), \qquad u \in (0,1),$$

denote the empirical and pre-smoothed empirical quantile functions for individual i, respectively. The first estimator is the *empirical* Wasserstein barycenter, which averages the empirical quantile functions within each arm. Specifically, for $z \in \{0,1\}$, we define

$$\widehat{\mu}_z^{\mathrm{emp}}(u) := \frac{1}{n_z} \sum_{i=1}^{n} \mathbb{1}\{Z_i = z\}\, \widehat{q}_i^{\,\mathrm{emp}}(u), \qquad u \in (0,1). \tag{4.6}$$

The second is a *pre-smoothed empirical* Wasserstein barycenter, obtained by replacing each individual's empirical distribution with its smoothed version before averaging the corresponding quantile functions:

$$\widehat{\mu}_z^{\mathrm{sm}}(u) := \frac{1}{n_z} \sum_{i=1}^{n} \mathbb{1}\{Z_i = z\}\, \widehat{q}_i^{\,\mathrm{sm}}(u), \qquad u \in (0,1). \tag{4.7}$$

The empirical and pre-smoothed Wasserstein barycenter quantile function estimators can perform well when sequencing depths are high, because each individual's empirical quantile function is an accurate approximation to their true unobserved quantile function. When sequencing depth is low, however, the empirical quantile function for each individual can be biased as an estimator of the corresponding true quantile function, and this can be especially pronounced in the tails (Bobkov and Ledoux, 2019). As a result, averaging these biased quantile functions across individuals produces a biased estimator of the treatment arm-specific barycenter quantile function (Bigot et al., 2018); see also Chapter 3. Although pre-smoothing can reduce finite-sample variability, it does not in general remove this tail bias.

Motivated by this problem, in Chapter 3, we studied Wasserstein barycenter estimation under sparse sampling, where each individual distribution is observed through only a small

number of measurements. To address the limitations of individual-level empirical quantile estimation in this setting, we developed the marginal-constructed barycenter (MCB) estimator, which uses information pooled across individuals rather than attempting to accurately reconstruct each individual distribution separately. The approach relies on a non-informative sequencing depth assumption, meaning that sequencing depth for an individual is assumed to be independent of that individual's feature distribution. The MCB estimator is based on an alternative representation of the barycenter estimation problem through binomial mixture models, and can be configured with different structural models for these mixtures. While the asymptotic properties require that these models be correctly specified, more flexible implementations, such as spline-based mixture models, can also be used in practice. We denote the corresponding treatment arm-specific barycenter quantile estimator by $\widehat{\mu}_z^{\mathrm{mcb}}(u)$.

The three estimators are therefore suited to datasets with different sequencing depths. The empirical and pre-smoothed empirical barycenters are most appropriate when sequencing depths are large enough to support reliable estimation of individual-level quantiles. The MCB estimator is instead designed for sparse sampling settings, where individual-level distributions may be poorly estimated. In this case, the MCB estimator replaces reliance on accurate individual-level distribution estimation with structural assumptions on the marginal mixture models and on non-informative cluster sizes. We evaluate the performance of each estimator across datasets with different sequencing depths in our simulation studies in Section 4.4.

Each estimator—$\widehat{\mu}_z^{\mathrm{emp}}(u)$, $\widehat{\mu}_z^{\mathrm{sm}}(u)$, and $\widehat{\mu}_z^{\mathrm{mcb}}(u)$—is pointwise asymptotically linear for each $u \in (0,1)$ under estimator-specific regularity conditions; these conditions are stated explicitly in Appendix D.3. Briefly, $\widehat{\mu}_z^{\mathrm{emp}}(u)$ and $\widehat{\mu}_z^{\mathrm{sm}}(u)$ are pointwise asymptotically linear when the estimated quantile function for each individual distribution converges sufficiently quickly to the corresponding true quantile function. The MCB estimator is pointwise asymptotically linear when the binomial mixture models used in the estimation procedure are correctly specified. For notational simplicity, let $\widehat{\mu}_z(u)$ denote any one of the estimators $\widehat{\mu}_z^{\mathrm{emp}}(u)$, $\widehat{\mu}_z^{\mathrm{sm}}(u)$, or $\widehat{\mu}_z^{\mathrm{mcb}}(u)$. Then, under the conditions in Appendix D.3, we have pointwise

asymptotic normality for each fixed $u \in (0,1)$:

$$\sqrt{n_z}\,\{\widehat{\mu}_z(u) - \mu_z(u)\} \rightsquigarrow N\big(0, \sigma_z^2(u)\big)\,, \qquad \sigma_z^2(u) = \mathrm{Var}\{\phi_z(O_i; u) \mid Z_i = z\}\,. \tag{4.8}$$

Accordingly, pointwise asymptotically valid confidence intervals for $\mu_z(u)$ can be constructed as

$$\widehat{\mu}_z(u) \pm z_{1-\alpha/2}\,\frac{\widehat{\sigma}_z(u)}{\sqrt{n_z}}, \tag{4.9}$$

where $\widehat{\sigma}_z(u)$ is a consistent estimator of $\sigma_z(u)$.

Under additional regularity conditions, also stated in Appendix D.3, the estimator $\widehat{\mu}_z$ converges weakly as a stochastic process. Specifically, for any compact set $\mathcal{U} \subset (0,1)$,

$$\sqrt{n_z}\,\{\widehat{\mu}_z - \mu_z\} \rightsquigarrow \mathbb{G}_z \qquad \text{in } \ell^\infty(\mathcal{U}), \tag{4.10}$$

where $\mathbb{G}_z$ is a tight mean-zero Gaussian process with covariance kernel $\Sigma_z(u,v) = \mathrm{Cov}\{\mathbb{G}_z(u), \mathbb{G}_z(v)\}$. This result allows us to perform uniform inference over $u \in \mathcal{U}$. In particular, let $c_{1-\alpha,z}$ denote the $(1-\alpha)$ quantile of $\|\mathbb{G}_z\|_\infty = \sup_{u\in\mathcal{U}} |\mathbb{G}_z(u)|$. Then an asymptotic $(1-\alpha)$ uniform confidence band for μ_z is given by

$$\left\{\widehat{\mu}_z(u) \pm \frac{\widehat{c}_{1-\alpha,z}}{\sqrt{n_z}} : u \in \mathcal{U}\right\}, \tag{4.11}$$

where $\widehat{c}_{1-\alpha,z}$ is a consistent estimator of $c_{1-\alpha,z}$.

4.3.3 Task 2: Estimation and inference for the difference in quantiles

Next, we consider estimation and inference for the quantile difference map $\Delta(u)$. Pointwise asymptotic normality of $\widehat{\Delta}(u)$, constructed using any of $\widehat{\mu}_z^{\mathrm{emp}}(u)$, $\widehat{\mu}_z^{\mathrm{sm}}(u)$, or $\widehat{\mu}_z^{\mathrm{mcb}}(u)$, follows from the asymptotic normality of the treatment arm-specific estimators together with the continuous mapping theorem. Let $n = n_0 + n_1$, and suppose that $n_z/n \to \pi_z \in (0,1)$ for

$z \in \{0,1\}$. Then, for each fixed $u \in (0,1)$,

$$\sqrt{n}\left\{\widehat{\Delta}(u) - \Delta(u)\right\} \rightsquigarrow N\big(0, \sigma^2_\Delta(u)\big)\,, \tag{4.12}$$

where, under independence of the two arms, $\sigma^2_\Delta(u) = \frac{\sigma^2_0(u)}{\pi_0} + \frac{\sigma^2_1(u)}{\pi_1}$.

Weak convergence in $\ell^\infty(\mathcal{U})$ of $\widehat{\Delta}$ follows from weak convergence of the treatment arm-specific estimators together with the continuous mapping theorem and independence of both arms. Specifically,

$$\sqrt{n}\left\{\widehat{\Delta} - \Delta\right\} \rightsquigarrow \mathbb{G}_\Delta \qquad \text{in } \ell^\infty(\mathcal{U}), \tag{4.13}$$

where $\mathbb{G}_\Delta$ is a mean-zero Gaussian process with covariance kernel $\Sigma_\Delta(u,v) = \mathrm{Cov}\{\mathbb{G}_\Delta(u), \mathbb{G}_\Delta(v)\}$, $u, v \in \mathcal{U}$. Under independence of the two arms, $\Sigma_\Delta(u,v) = \frac{\Sigma_0(u,v)}{\pi_0} + \frac{\Sigma_1(u,v)}{\pi_1}$, for $u, v \in \mathcal{U}$. These results can be used to derive both asymptotically valid pointwise confidence intervals and uniform confidence bands, similar to (4.9) and (4.11) in the previous section, respectively.

4.3.4 Task 3: Hypothesis testing

Finally, we describe hypothesis testing for the global null hypothesis $H_0 : \Delta(u) = 0$ for all $u \in \mathcal{U}$. By the weak convergence result in (4.13), the estimated difference process satisfies $\sqrt{n}\left\{\widehat{\Delta} - \Delta\right\} \rightsquigarrow \mathbb{G}_\Delta$ in $\ell^\infty(\mathcal{U})$. Under H_0, $\Delta(u) = 0$ for all $u \in \mathcal{U}$, so the scaled observed process $\sqrt{n}\,\widehat{\Delta}$ converges weakly to $\mathbb{G}_\Delta$.

A global test statistic may be constructed by applying a real-valued functional to the estimated difference process. That is, for a functional $\phi : \ell^\infty(\mathcal{U}) \to \mathbb{R}$, we may consider statistics of the form $T = \phi(\sqrt{n}\widehat{\Delta})$. If ϕ is continuous, or sufficiently regular for the continuous mapping theorem to apply, then under H_0, $T \rightsquigarrow \phi(\mathbb{G}_\Delta)$. Therefore, critical values can be obtained from the limiting Gaussian process, or in practice from a simulation-based procedure that consistently approximates its distribution (Kosorok, 2008).

We consider three global test statistics based on this weak convergence result. The first two are based on the studentized difference process. Let $\widehat{\mathrm{se}}_\Delta(u) := \left\{\widehat{\Sigma}_\Delta(u,u)/n\right\}^{1/2}$ denote

the estimated pointwise standard error of $\widehat{\Delta}(u)$, and define $\widehat{Z}_\Delta(u) = \frac{\widehat{\Delta}(u)}{\widehat{\mathrm{se}}_\Delta(u)}$. Equivalently, $\widehat{Z}_\Delta(u) = \sqrt{n}\widehat{\Delta}(u)/\{\widehat{\Sigma}_\Delta(u,u)\}^{1/2}$ when $\widehat{\Sigma}_\Delta$ estimates the covariance kernel of the limiting process $\mathbb{G}_\Delta$. We define the supremum test statistic T_∞ and the integrated squared test statistic T_{int} as

$$T_\infty := \sup_{u\in\mathcal{U}} \left|\widehat{Z}_\Delta(u)\right| \qquad T_{\text{int}} := \int_{\mathcal{U}} \widehat{Z}_\Delta(u)^2\, du. \tag{4.14}$$

The third statistic T_M is based on the squared functional Mahalanobis semi-distance of the estimated difference process, analogous to Hotelling's T^2 test but for functional data (Joseph et al., 2015):

$$T_M := n\left\langle \widehat{\Delta}, \widehat{\Sigma}_\Delta^{-1}\widehat{\Delta}\right\rangle := n\int_{\mathcal{U}}\int_{\mathcal{U}} \widehat{\Delta}(u)\,\widehat{\Sigma}_\Delta^{-1}(u,v)\,\widehat{\Delta}(v)\,dv\,du. \tag{4.15}$$

In practice, we do not use the inverse covariance operator $\widehat{\Sigma}_\Delta^{-1}$ directly, since covariance inversion is typically unstable or ill-posed in infinite-dimensional function spaces. We therefore compute a regularized version of T_M using a truncated spectral inverse based on the first K empirical functional principal components of $\widehat{\Sigma}_\Delta$. Specifically, if $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_K$ and $\widehat{\psi}_1,\ldots,\widehat{\psi}_K$ denote the first K empirical eigenvalues and eigenfunctions of $\widehat{\Sigma}_\Delta$, then we compute

$$T_M^{(K)} := n\sum_{k=1}^{K} \frac{\left\langle \widehat{\Delta}, \widehat{\psi}_k\right\rangle^2}{\widehat{\lambda}_k}. \tag{4.16}$$

The three test statistics are sensitive to different alternatives. The supremum statistic is most useful for detecting localized differences concentrated at a small number of quantile levels. The integrated squared statistic aggregates evidence across the quantile curve and is useful for detecting accumulated departures from the null. The Hotelling-type statistic accounts for the covariance structure across quantile levels and may be more powerful when the signal lies along low-variance directions of the estimated process (Table 4.1). We will evaluate the performance of each of these test statistics in our simulation study in Section 4.4.

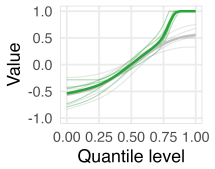
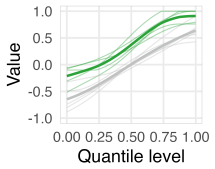
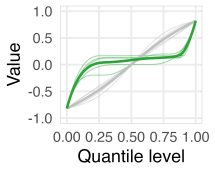
Statistic	Test Statistic	Test Statistic (on grid $\{u_1, ..., u_m\}$)	Sensitive to...	Figure
Supremum (T_∞)	$\sup_{u\in\mathcal{U}} \left\lvert \frac{\widehat{\Delta}(u)}{\widehat{\mathrm{se}}\{\widehat{\Delta}(u)\}} \right\rvert$	$\max_{1\le j\le k} \left\lvert \frac{\widehat{\Delta}(u_j)}{\widehat{\mathrm{se}}\{\widehat{\Delta}(u_j)\}} \right\rvert$	Localized differences at one or a small number of quantile levels	
Integrated squared (T_{int})	$\int_{\mathcal{U}} \left[\frac{\widehat{\Delta}(u)}{\widehat{\mathrm{se}}\{\widehat{\Delta}(u)\}} \right]^2 du$	$\sum_{j=1}^{k} \left[\frac{\widehat{\Delta}(u_j)}{\widehat{\mathrm{se}}\{\widehat{\Delta}(u_j)\}} \right]^2$	Broad differences spread across the quantile curve	
Hotelling-type (T_M)	$\left\langle \widehat{\Delta}, \widehat{\Sigma}_{\Delta}^{-1} \widehat{\Delta} \right\rangle$	$\widehat{\boldsymbol{\Delta}}^\top \widehat{\boldsymbol{\Sigma}}_{\Delta,K}^{-1} \widehat{\boldsymbol{\Delta}}$	Differences after accounting for covariance across quantile levels	

Table 4.1: Candidate test statistics for comparing treatment arm-specific barycenter quantile functions and schematic alternatives to which each statistic is expected to be sensitive. The right-hand panels show toy examples, with light lines representing individual-level quantile functions and dark lines representing the corresponding arm barycenter quantile functions. Arm 0 is shown in gray and Arm 1 in green.

Remark 4.3.2 (Other test statistics)**.** *Other test statistics can be constructed from the estimated difference process depending on the alternatives of interest. For example, one-sided supremum statistics may be useful for directional alternatives, while weighted versions of* T_{int} *can place greater emphasis on specific regions of the quantile function, such as the tails.*

4.3.5 Implementation on a grid

In practice, the procedure can be implemented using a finite-dimensional representation of the estimated quantile functions. In our implementation, the functions are evaluated on a common dense grid in the following steps:

1. Choose a grid $u_1, \ldots, u_m \in (0, 1)$.

2. Estimate the Wasserstein barycenter quantile function for each treatment arm on the grid using $\widehat{\mu}_z^{\mathrm{emp}}(u)$, $\widehat{\mu}_z^{\mathrm{sm}}(u)$, or $\widehat{\mu}_z^{\mathrm{mcb}}(u)$. Denote the resulting generic estimator by $\widehat{\mu}_z(u)$, and write its grid representation as

$$\widehat{\boldsymbol{\mu}}_z = \{\widehat{\mu}_z(u_1), \ldots, \widehat{\mu}_z(u_m)\}^\top.$$

Let $\widehat{\boldsymbol{\Sigma}}_z$ denote the estimated covariance matrix of the limiting process $\sqrt{n_z}(\widehat{\boldsymbol{\mu}}_z - \boldsymbol{\mu}_z)$.

3. Form the estimated difference vector

$$\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0 = \left\{\widehat{\Delta}(u_1), \ldots, \widehat{\Delta}(u_m)\right\}^\top. \tag{4.17}$$

Assuming the two treatment arms are independent, estimate the covariance matrix of the limiting process $\sqrt{n}(\widehat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})$ by

$$\widehat{\boldsymbol{\Sigma}}_\Delta = \frac{\widehat{\boldsymbol{\Sigma}}_0}{\widehat{\pi}_0} + \frac{\widehat{\boldsymbol{\Sigma}}_1}{\widehat{\pi}_1}, \tag{4.18}$$

where $\widehat{\pi}_z = n_z/n$.

4. Compute the observed test statistics (Table 4.1). First define the pointwise standardized differences

$$\widehat{Z}_{\Delta,j} = \frac{\sqrt{n}\widehat{\Delta}(u_j)}{\sqrt{\widehat{\Sigma}_{\Delta,jj}}}, \qquad j = 1, \ldots, m, \tag{4.19}$$

where $\widehat{\Sigma}_{\Delta,jj}$ is the jth diagonal element of $\widehat{\boldsymbol{\Sigma}}_\Delta$.

The supremum statistic is computed as

$$T_\infty = \max_{1 \le j \le m} \left|\widehat{Z}_{\Delta,j}\right|, \tag{4.20}$$

and the integrated squared statistic is approximated by

$$T_{\mathrm{int}} = \sum_{j=1}^{m} w_j \widehat{Z}_{\Delta,j}^2, \tag{4.21}$$

where w_j are numerical integration weights, depending on the chosen grid.
The Hotelling-type statistic is computed using a regularized inverse of $\widehat{\mathbf{\Sigma}}_\Delta$:

$$T_M = n\widehat{\mathbf{\Delta}}^\top \widehat{\mathbf{\Sigma}}_{\Delta,K}^{-1} \widehat{\mathbf{\Delta}}. \tag{4.22}$$

Here $\widehat{\mathbf{\Sigma}}_{\Delta,K}^{-1}$ is obtained by retaining the first K empirical functional principal components of $\widehat{\mathbf{\Sigma}}_\Delta$, chosen to explain a prespecified proportion of the total variance.

5. Approximate the null distributions. For T_∞ and T_{int}, the null distribution is obtained by simulating mean-zero Gaussian difference processes with covariance $\widehat{\mathbf{\Sigma}}_\Delta$:

$$\boldsymbol{G}_b \sim N(0, \widehat{\mathbf{\Sigma}}_\Delta), \qquad b = 1, \ldots, B.$$

Each simulated process is then standardized pointwise:

$$Z_{b,j} = \frac{G_{b,j}}{\sqrt{\widehat{\Sigma}_{\Delta,jj}}}, \qquad j = 1, \ldots, m.$$

The statistics T_∞ and T_{int} are recomputed for each $\boldsymbol{Z}_b := (Z_b(u_1), \ldots, Z_b(u_J))^\top$ to estimate the null distribution for each test statistic. The Monte Carlo p-value is the proportion of simulated null statistics greater than or equal to the observed statistic.
For the Hotelling-type statistic, the implementation uses the chi-square approximation

$$T_M \dot{\sim} \chi^2_K,$$

where K is the number of retained empirical principal components.

These procedures are implemented in an accompanying R package `mcbarycenter`, which provides all three barycenter estimators, the corresponding difference-in-quantiles estimators, and the global testing procedures described above. A vignette illustrating the use of these methods is provided in Appendix A.

Remark 4.3.3 (Multiplier bootstrap)**.** *In step 2 of the procedure above, if estimation of the covariance matrix $\widehat{\mathbf{\Sigma}}_z$ for $z \in \{0, 1\}$ is infeasible due to the density of the grid or for other reasons, one could use the Gaussian multiplier bootstrap to approximate the null distribution of the test statistics T_∞ and T_{int} (Kosorok, 2008, Section 10.1); see also Chernozhukov et al. (2013).*

4.4 Simulation studies

We conduct two simulation studies. The first examines how low sequencing depth, as well as differential sequencing depth between treatment arms, can affect estimation and inference for Wasserstein barycenters. This study is intended to provide practical guidance on the choice of estimator—the empirical Wasserstein barycenter, pre-smoothed empirical barycenter, or MCB estimator—under different sequencing depth settings. The second simulation study evaluates the proposed hypothesis testing procedure, focusing on the validity of each test statistic under the null and its power under a range of alternative scenarios.

In both simulation studies, we estimate each barycenter quantile function on a common grid of quantile levels $\{0.01, 0.02, \ldots, 0.99\}$. We compare the following estimators:

1. Empirical Wasserstein barycenter.
2. Pre-smoothed empirical Wasserstein barycenter. Each individual distribution is pre-smoothed using kernel density estimation with a Gaussian kernel and a bandwidth selected by Silverman's rule of thumb (Silverman, 1986).
3. Correctly specified MCB estimator. A Beta distribution specification is used to fit the binomial mixing distributions, which matches the Dirichlet process specification of the data generating process.
4. Flexible but misspecified MCB estimator. The binomial mixing distributions are fit using splines (Efron, 2016). The spline basis consists of natural splines with 5 degrees of freedom, augmented with point mass basis functions at 0 and 1, and fit using a regularization parameter of 0.01.

The correctly specified MCB estimator is included to evaluate performance when the structural assumptions required by the MCB estimator are satisfied. The spline-based MCB estimator is included to evaluate performance in a more realistic setting in which the mixing distributions may be misspecified.

4.4.1 Simulation Study 1: impact of varying sequencing depth on estimation bias

Setup

In the first simulation study, we examine how sequencing depth affects estimation and pointwise inference for the treatment arm-specific barycenter quantile functions and the quantile difference map. We consider sequencing depth settings chosen to roughly match scenarios observed in the HIV-1 prevention studies described in the introduction (Figure 1.2). The settings considered in the simulation are summarized in Table 4.2.

Table 4.2: Sequencing depth ranges for Simulation Study 1

Sequencing depth scenario	**Sequencing depth (Arm 0)**	**Sequencing depth (Arm 1)**
Low balanced	2 to 15	2 to 15
Low differential	2 to 15	2 to 7
High balanced	2 to 300	2 to 300
High differential	2 to 300	2 to 150

For each sequencing depth setting, we consider two sequence feature scenarios:

1. *Null scenario*: The two treatment arms have equal Wasserstein barycenters. In both arms, each individual-level distribution is generated from a Dirichlet process with base measure Normal$(0, 1)$ and concentration parameter 2.
2. *Alternative scenario*: The two treatment arms have unequal Wasserstein barycenters. In Arm 0, each individual-level distribution is generated from a Dirichlet process with base measure Normal$(0, 1)$ and concentration parameter 2. In Arm 1, each individual-level distribution is generated from a Dirichlet process with concentration parameter

2 and a bimodal Normal mixture base measure, with 80% weight on Normal$(0, 1)$ and 20% weight on Normal$(6, 0.25^2)$.

For each simulation setting, we generate observations for 50 individuals in each treatment arm using a two-stage mechanism. First, for each individual, we sample an individual-level distribution according to the sequence feature scenario described above. Second, we sample that individual's sequencing depth from a discrete uniform distribution over the corresponding range in Table 4.2, and then sample the corresponding number of observations from the individual's distribution. Combining the four sequencing depth settings with the two sequence feature scenarios yields eight total simulation settings. The true arm-specific Wasserstein barycenter quantile functions, and the resulting quantile difference map, are approximated by computing the empirical barycenter of one large dataset with 100,000 individuals per arm and 1,000 sequences per individual. Performance is assessed in terms of bias, estimated standard error, and pointwise confidence interval coverage. Additional simulation studies with 20 and 100 individuals per arm are included in Appendix D.5.

Results

Summary results for the bias, mean standard error, and confidence interval coverage of the estimators of $\Delta(u)$ are shown in Figure 4.3 under the null and Figure 4.4 under the alternative. In each figure, results are averaged separately over the central quantiles (0.06–0.94) and the tail quantiles (0.01–0.05 and 0.95–0.99).

We first consider the low, balanced sequencing depth setting. Under the null scenario, all estimators perform similarly well, with negligible bias and near-nominal confidence interval coverage despite the low sequencing depths. This pattern is expected because, under the null and with the same sequencing depth distribution in the two treatment arms, the bias in the estimated arm-specific barycenters is the same across arms. As a result, when the two estimated barycenter quantile functions are differenced, this bias cancels, producing accurate estimates of $\Delta(u) = 0$ across quantile levels.

Under the alternative scenario, however, differences among the estimators become apparent. Both empirical barycenters exhibit larger bias than the MCB estimators, with the discrepancy most pronounced in the tail quantiles. The pre-smoothed empirical barycenter reduces bias relative to the empirical barycenter, but its coverage remains poor in the tails and it does not perform as well as the MCB estimators. The mean absolute bias for the tail quantiles was 0.454 for the empirical barycenter and 0.345 for the pre-smoothed empirical barycenter, compared with 0.030 and 0.064 for the correctly specified and spline-based MCB estimators, respectively. Confidence interval coverage in the tails was 71% for the empirical barycenter and 65% for the pre-smoothed empirical barycenter, compared with 93% and 91% for the correctly specified and spline-based MCB estimators, respectively.

Next, we consider the low, differential sequencing depth setting. Under the null scenario, the empirical and pre-smoothed empirical barycenters perform reasonably well for the central quantiles, with low bias and near-nominal confidence interval coverage. However, their performance deteriorates in the tails, where both estimators exhibit larger bias and reduced coverage. This pattern is more pronounced under the alternative scenario: the empirical and pre-smoothed empirical barycenters show increased bias even in the central quantiles, with the largest bias and poorest coverage occurring in the tails. These results indicate that when sequencing depth is low overall, systematic differences in sequencing depth between treatment arms can distort estimation and inference on $\Delta(u)$, particularly at the tail quantile levels. In contrast, both MCB estimators reduce bias and improve coverage. The correctly specified MCB estimator has the smallest bias and coverage closest to the nominal level. The flexible spline-based MCB estimator also performs well with respect to both bias and coverage, suggesting that, in this simulation setting, the spline-based specification is sufficiently flexible to approximate the relevant mixing distributions used to estimate the barycenter. Under the alternative scenario, the mean absolute bias in the tail quantiles was 0.902 for the empirical barycenter and 0.462 for the pre-smoothed empirical barycenter, compared with 0.038 and 0.152 for the correctly specified and spline-based MCB estimators, respectively; the corresponding confidence interval coverage was 31%, 71%, 95%, and 89%.

Finally, we consider the high sequencing depth settings. In these settings, the empirical barycenter performs substantially better than in the low sequencing depth settings and is generally similar to the MCB estimators for both central and tail quantiles. This improvement is expected because, with larger sequencing depths, each individual's empirical quantile function provides a better approximation to the corresponding latent quantile function. Although some individuals still have low sequencing depths in these settings, they do not appear to have enough influence to substantially distort the empirical barycenter, particularly in the central part of the distribution. In the alternative scenario, the MCB estimators show slightly lower tail bias and coverage closer to the nominal level compared to the empirical barycenter. The pre-smoothed empirical barycenter has worse performance in the higher sequencing depth settings, especially in the tails. This shows that the smoothing step may introduce additional bias when the empirical distributions are already reasonably well estimated.

More detailed simulation results are provided in Supplementary Figures D.2–D.7. These figures show bias, mean standard error, and confidence interval coverage for both $\Delta(u)$ and the treatment arm-specific barycenter quantile functions, $\mu_z(u)$, across all quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Additional simulations with 20 and 100 individuals per arm showed similar patterns and are presented in Appendix D.5.1. Notably, the MCB estimators continued to perform well even with 20 individuals per arm.

4.4.2 Simulation Study 2: hypothesis testing

Setup

In the second simulation study, we evaluate the proposed hypothesis testing procedure across a range of two sample scenarios. The goal is to assess Type I error under null scenarios with equal Wasserstein barycenters and power under alternatives with different forms of barycenter differences. As in Simulation Study 1, we generate individual-level distributions from Dirichlet processes. We consider four broad classes of scenarios, described in high-level

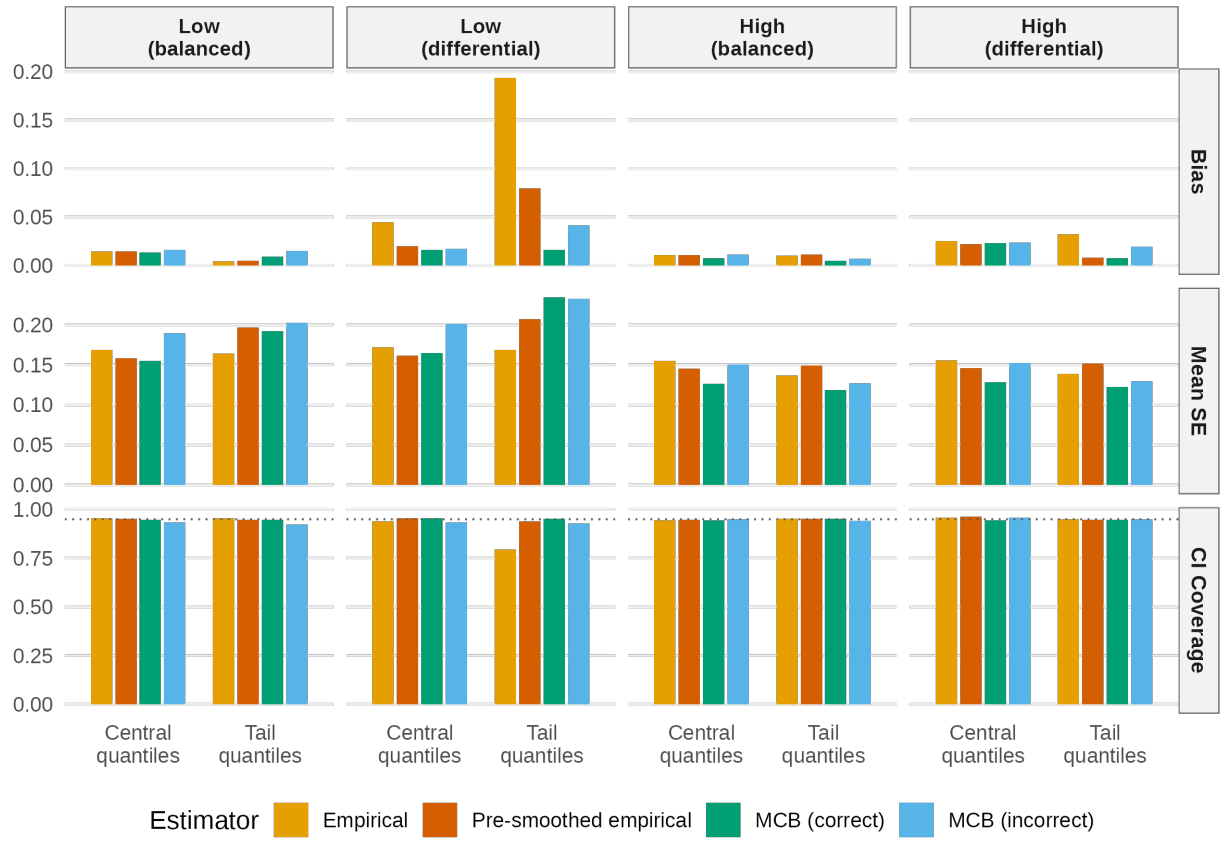


Figure 4.3: Summary of estimator of the difference in quantiles $\Delta(u)$ under the null scenario for Simulation Study 1. Results are shown across four sequencing depth settings: low balanced, low differential, high balanced, and high differential sequencing depth. For each setting, we compare the empirical barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and misspecified spline-based MCB estimator. Performance is summarized by average bias, mean standard error, and pointwise confidence interval coverage for $\Delta(u)$, averaged separately over the central quantiles (0.06–0.94) and tail quantiles (0.01–0.05 and 0.95–0.99). The dotted horizontal line in the coverage panels indicates the nominal 95% coverage level.

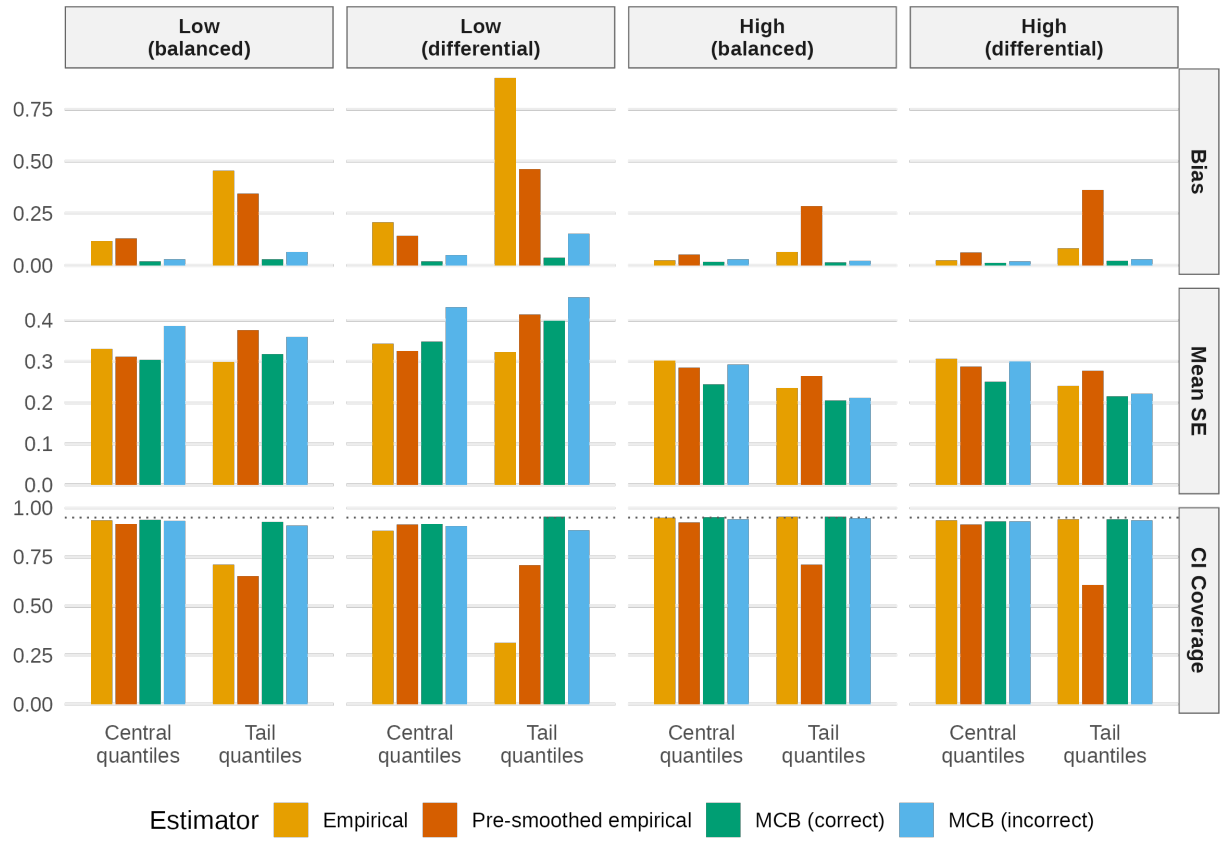


Figure 4.4: Summary of estimator of the difference in quantiles $\Delta(u)$ under the alternative scenario for Simulation Study 1. Results are shown across four sequencing depth settings: low balanced, low differential, high balanced, and high differential sequencing depth. For each setting, we compare the empirical Wasserstein barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and misspecified spline-based MCB estimator. Performance is summarized by average bias, mean standard error, and pointwise confidence interval coverage for $\Delta(u)$, averaged separately over the central quantiles (0.06–0.94) and tail quantiles (0.01–0.05 and 0.95–0.99). The dotted horizontal line in the coverage panels indicates the nominal 95% coverage level.

below with full details provided in Supplementary Table D.1:

1. *Null*: The two treatment arms have identical Wasserstein barycenters. We consider three null settings with different common barycenter shapes: normal, symmetric bimodal, and non-symmetric bimodal.

   For the alternative scenarios (2)–(4), Treatment Arm 0 is generated from a Dirichlet process with a Normal$(0, 1)$ base distribution. The Arm 1 data-generating distribution is chosen so that the resulting barycenter differs from the Arm 0 barycenter in the following ways:

2. *Uniform differences*: The Treatment Arm 1 barycenter differs from the Arm 0 barycenter by an approximately constant shift across quantile levels. We examine three shift magnitudes, labeled "Uniform shift = 0.1, 0.2, 0.3."
3. *Upper-tail differences*: The Treatment Arm 1 barycenter differs from the Arm 0 barycenter primarily in the upper tail. We examine upper-tail shifts of increasing magnitude, labeled "Upper-tail shift = 1, 2, 3."
4. *Shape differences*: The Treatment Arm 1 barycenter differs from the Arm 0 barycenter in overall shape. These alternatives are generated using a bimodal base distribution for Arm 1, with increasing separation between the modes, labeled "Bimodal = 0.2, 0.4, 0.6."

The treatment arm-specific Wasserstein barycenters for all scenarios are shown visually in Supplementary Figure D.1. To focus on the behavior of the test statistics apart from the low sequencing depth bias studied in Simulation Study 1, the main simulation is conducted under the high, balanced sequencing depth setting described in Table 4.2. (Results for the other sequencing depth settings are provided in Appendix D.5.2.) For each scenario, we simulate 100 individuals per arm and apply each proposed test statistic (T_∞, T_{int}, and T_M) to test the null hypothesis of equal Wasserstein barycenters. We repeat this procedure over 500 simulated datasets and estimate the rejection probability as the proportion of datasets

in which the null hypothesis is rejected. This rejection probability estimates the Type I error under the null scenarios and power under the alternative scenarios. Under the alternative scenarios, we compare the power of the proposed distributional tests with a mean-based analysis, in which the sequence features are averaged within each individual and the arms are compared using an unequal-variance two-sample t-test.

Results

Results are shown in Figure 4.5. Under the null settings, all barycenter estimators and all three test statistics had rejection probabilities close to the nominal 5% level.

For the uniform difference alternatives, the three test statistics performed similarly, with rejection probabilities increasing as the size of the mean shift increased. In these scenarios, the correctly specified and spline-based MCB estimators tended to have slightly higher power compared to the empirical barycenters. The distributional tests also generally had similar power to the mean-based analysis. This is expected because the difference between the two treatment arm barycenters is approximately constant across quantile levels, so averaging the sequence features within each individual does not remove the main signal. For example, in the mean shift = 0.2 setting, the rejection probabilities for the supremum test were 70.0%, 66.0%, 75.5%, and 78.0% for the empirical barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and spline-based MCB estimator, respectively, compared with 68.0% for the mean-based analysis.

For the upper-tail alternative scenarios, the supremum test statistic generally had the largest rejection probabilities among the three proposed test statistics, especially as the magnitude of the upper-tail shift increased. The correctly specified and spline-based MCB estimators tended to have the highest power, while the empirical and pre-smoothed empirical barycenters performed similarly. In this setting, the distributional tests had substantially higher power than the mean-based analysis. This is because the treatment arm difference is concentrated in the upper tail, and averaging the sequence features within each individual attenuates this localized signal. For example, in the upper-tail shift = 2 setting, the rejection

probabilities for the supremum test were 45.0%, 43.5%, 61.0%, and 59.0% for the empirical barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and spline-based MCB estimator, respectively, compared with 23.0% for the mean-based analysis.

Finally, for the shape-difference alternatives, the Hotelling-type test and supremum test had the largest rejection probabilities, while the integrated squared test tended to have lower power. The Hotelling-type test was particularly powerful for the MCB estimators, consistent with the fact that this statistic accounts for covariance across quantile levels and can be more sensitive to structured departures from the null that occur across the quantile function. As in the previous settings, the correctly specified MCB estimator generally had the largest rejection probabilities. The mean-based analysis had little power in this setting because the two treatment arms have similar average feature values despite having different barycenter shapes. For example, in the bimodal = 0.4 setting, the rejection probabilities for the supremum test were 72.5%, 83.5%, 85.0%, and 72.0% for the empirical barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and spline-based MCB estimator, respectively, compared with 3.5% for the mean-based analysis.

4.4.3 Recommendations from simulation studies

Simulation Study 1 provides guidance for choosing the barycenter estimator. In low sequencing depth settings, especially when sequencing depths differed between treatment arms or tail inference was of interest, the empirical barycenter performed poorly. In these settings, we recommend using an MCB estimator as the primary approach, provided that there are enough individuals to estimate the binomial mixtures. Although a correctly specified MCB estimator is not available in practice, the spline-configured MCB estimator substantially improved performance relative to the empirical barycenter and continued to perform well even with 20 individuals per arm (Appendix D.5.1). In high sequencing depth settings, both the empirical barycenter and spline-based MCB estimator performed well across balanced and differential sequencing depth scenarios and can be used for analysis. The empirical barycenter has the advantage of being less model-dependent, whereas the spline-based MCB

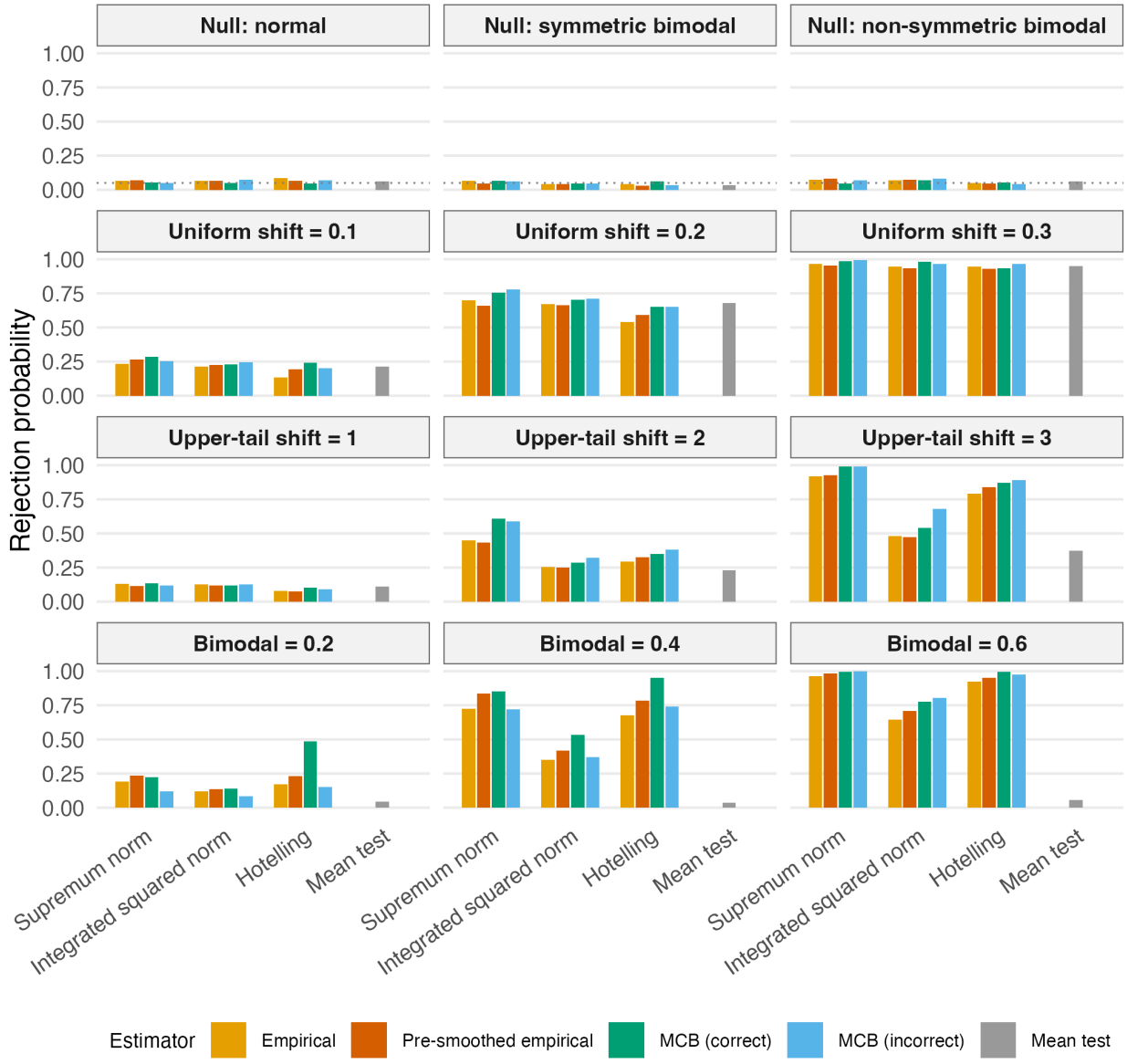


Figure 4.5: Rejection probability across all simulations for each of the three null scenarios and nine alternative scenarios, for each estimator. The mean test shows the result from averaging the sequence features for each person and doing a two-sample unequal variance t-test.

estimator provided modest bias reduction, particularly in the tails, when the mixture model specification closely matched the truth. Thus, the choice between the two estimators may be guided by the goals of the analysis and the plausibility of the modeling assumptions, with either approach representing a suitable option in higher sequencing depth studies. The pre-smoothed empirical barycenter showed mixed performance. Pre-smoothing sometimes reduced bias in low sequencing depth settings, but did not perform as well as the MCB estimators; in higher sequencing depth settings, it introduced additional bias.

These recommendations also depend on the quantile grid. Our simulations used the grid $\{0.01, 0.02, \ldots, 0.99\}$. More resolved grids, such as $\{0.001, 0.002, \ldots, 0.999\}$, would require significantly higher sequencing depths for the empirical barycenter to perform well, and may increase the benefit of the MCB estimator, depending on how it is configured. Conversely, if the scientific question focuses primarily on central quantiles, the empirical barycenter may be adequate at lower sequencing depths.

Simulation Study 2 provides guidance for choosing among the three global test statistics. The integrated squared statistic T_{int} performed similarly to the other statistics under uniform mean shifts, but had lower power for upper-tail and shape difference alternatives. The supremum statistic T_∞ performed best for localized upper-tail differences, while the Hotelling-type statistic T_M performed well for shape differences. When there is no strong reason to prioritize a specific alternative, we recommend the supremum statistic T_∞ as the default as it consistently performed well across the simulation settings.

Finally, Simulation Study 2 shows when the proposed distributional framework is most useful relative to a mean-based analysis. When treatment arm differences were approximately constant across quantile levels, the distributional tests had power similar to the mean-based test. In contrast, when differences were concentrated in the upper tail or reflected differences in barycenter shape, the distributional tests had significantly higher power because they retained information across the full quantile function. Thus, the proposed framework is most useful when meaningful differences occur in specific regions of the feature distribution, which may be obscured by a mean-based analysis.

4.5 Data applications

We apply our approach to the AMP HIV-1 prevention trials (HVTN 703/704) and a study on vaccine-elicited B cell receptor maturation. The goal of these applications is to illustrate how the proposed distributional analysis compares with a standard mean-based analysis and how the quantile difference map can be used to interpret differences between treatment arms in the AMP trials and between vaccination schedule–timepoint groups in the B cell study. Data applications of three additional HIV-1 prevention trials are presented in Appendix D.4.

4.5.1 AMP HIV-1 prevention trials (HVTN 703/704)

The AMP trials tested whether the broadly neutralizing antibody VRC01 could prevent HIV-1 acquisition (Corey et al., 2021). In these trials, HIV-negative participants were randomized to receive a placebo, low-dose VRC01 (10 mg/kg), or high-dose VRC01 (30 mg/kg) every 8 weeks. Overall efficacy was modest, with an estimated efficacy of 29.0% (95% CI: -4.0 to 51.6) for the high-dose treatment arm and 7.2% (95% CI: -33.3 to 35.3) for the low-dose arm, pooled across both AMP trials. Among participants who acquired HIV-1 during follow-up, viral samples were deep-sequenced using PacBio sequencing technology to characterize within-host diversity in the HIV-1 Env protein (Westfall et al., 2024). Previous analyses have found that VRC01 was much more protective against viruses that were highly sensitive to VRC01 neutralization (Corey et al., 2021).

For each viral sequence, we considered three numeric features related to VRC01 susceptibility as previously considered in Juraska et al. (2024): (1) Hamming distance to a VRC01-sensitive reference sequence at pre-selected sites, (2) predicted IC80, and (3) epitope distance. Predicted IC80 is the sequence-predicted concentration of VRC01 required to inhibit viral infection by 80%, with larger values indicating greater predicted resistance to VRC01 neutralization. Epitope distance is a structure-informed measure of how different the VRC01-binding epitope in a given viral sequence is from epitopes in highly VRC01-sensitive reference sequences, with larger values indicating greater divergence from VRC01-sensitive

viruses. Our goal was to compare these feature distributions between the placebo treatment arm and the VRC01 arm (pooling both low-dose and high-dose arms). As the primary analysis, we applied our framework using the empirical barycenter on the grid $0.01, 0.02, \ldots, 0.99$. We conducted a sensitivity analysis using the spline-configured MCB estimator with 5 degrees of freedom augmented with two point mass basis functions at 0 and 1. We tested for equality of barycenters using the supremum test statistic T_∞. For comparison, we also report a mean-based analysis, in which the features are first averaged within each individual and the arms are then compared using an unequal-variance two-sample t-test.

Results for the three features are shown in Figure 4.6, and analogous results using the spline-configured MCB estimator are shown in Supplementary Figure D.8. In each row of Figure 4.6, the left-hand panels show the treatment arm-specific mean summaries and estimated barycenter quantile functions, while the right-hand panels show the corresponding estimated mean differences and quantile difference maps.

The three features show different patterns of separation between the placebo treatment arm and the pooled VRC01 arm. For Hamming distance at the pre-selected sites, the estimated placebo and VRC01 empirical barycenters are nearly indistinguishable. The estimated quantile difference is close to zero across the full range of quantile levels, and there is no evidence of a difference between arms using either the mean-based analysis ($p = 0.95$) or the empirical barycenter-based analysis ($p = 0.93$).

For predicted IC80, the VRC01 treatment arm's estimated barycenter quantile functions are larger than that of the placebo arm across much of the distribution, with the separation most apparent among higher quantiles. Neither the mean-based analysis ($p = 0.13$) nor the distributional analysis ($p = 0.12$) provide evidence of a difference between arms. Nevertheless, the quantile difference map suggests that the estimated difference between the two arms is not well summarized by a uniform shift: the estimated difference is positive over most quantile levels but becomes larger in the upper tail. The sensitivity analysis using the spline-configured MCB estimator shows a similar but more pronounced upper-tail difference ($p = 0.04$).

Finally, for epitope distance, the VRC01 treatment arms have higher values than the placebo arm across nearly all quantile levels. Both the mean-based analysis and the empirical barycenter-based analysis show strong evidence of a difference between arms ($p < 0.001$ for both comparisons). The estimated quantile difference map shows a relatively stable difference across most quantile levels, although it decreases at the extreme upper tail, where estimates are typically less stable. Thus, for the epitope distance feature, the distributional analysis supports the same conclusion as the mean-based analysis, while also showing that the result generally reflects uniform difference between the placebo and VRC01 barycenter distributions rather than a difference restricted to a specific portion of the distribution.

4.5.2 Vaccine-elicited B cell receptor maturation

To illustrate the use of our framework outside the sieve analysis setting, we apply it to a study of vaccine-elicited B cell receptor maturation. Willis et al. (2025) studied a germline-targeting HIV vaccine strategy designed to guide rare B cells toward making VRC01-class broadly neutralizing antibodies. The first vaccine, eOD-GT8 60mer (eOD), served as a germline-targeting prime to activate rare VRC01-class precursor B cells, while the second vaccine, core-z28v2 60mer (core), served as a heterologous boost intended to promote further maturation of those cells.

The study included two phase 1 trials: IAVI G002 in North America, which evaluated both the priming and boosting strategy, and IAVI G003 in Rwanda and South Africa, which evaluated the priming strategy only. In IAVI G002, participants were assigned to different vaccination schedules involving eOD and/or core, while G003 participants received two doses of eOD at weeks 0 and 8, as shown in Figure 1 of Willis et al. (2025). Samples were collected at multiple time points from each participant. Antigen-specific B cells were sorted from PBMC samples and analyzed by B-cell receptor (BCR) sequencing to characterize VRC01-class responses at the antibody-sequence level, allowing the researchers to measure response frequencies, heavy- and light-chain somatic hypermutation, and the accumulation of mutations associated with more mature VRC01-class broadly neutralizing antibodies. The

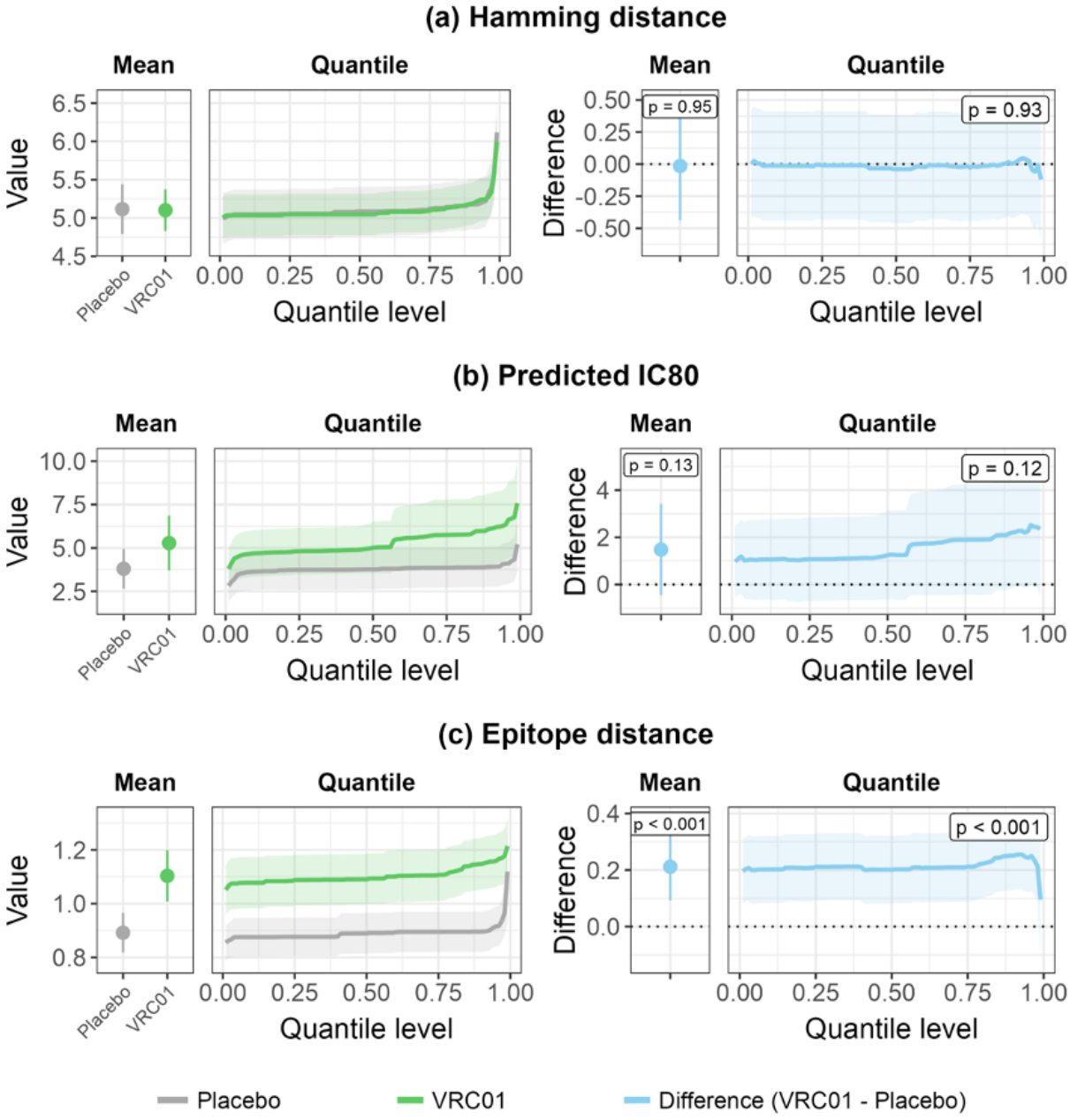


Figure 4.6: Using the empirical Wasserstein barycenter estimator, analysis of three VRC01 susceptibility-associated features among AMP participants who acquired the primary endpoint of HIV-1 infection: (A) Hamming distance to a VRC01-sensitive reference sequence at pre-selected sites, (B) predicted IC80, and (C) epitope distance. For each feature, the left panels show the arm-specific mean summaries and estimated empirical barycenter quantile functions for the placebo arm and the pooled VRC01 arms. The right panels show the corresponding VRC01-minus-placebo estimated mean difference and quantile difference map. Shaded bands indicate pointwise 95% confidence intervals. Reported p-values correspond to the mean-based unequal-variance two-sample t-test for the mean panels and the supremum test for equality of barycenters for the quantile panels.

analysis compared six vaccination schedule–timepoint combinations. Labels give the PBMC sampling week followed by the immunogens administered at week 0, at weeks 0 and 8, or at weeks 0, 8, and 16: wk 8 eOD, wk 16 eOD→eOD, wk 16 eOD→core, wk 24 eOD→eOD, wk 24 eOD→core, and wk 24 eOD→eOD→core. The descriptions of each group and their sample sizes are provided in Supplemental Table D.2.

We apply our framework to these data by analyzing the heavy-chain percent mutation feature for each sequence. For each group, we estimate the Wasserstein barycenter on the grid $0.01, \ldots, 0.99$. Because most groups have small participant sample sizes (< 20) and most participants had high sequencing depth, we use the empirical barycenter. We conduct one representative two-sample comparison: wk24 eOD→core versus wk24 eOD→eOD. We test the equality of barycenters using the supremum test statistic T_∞.

The results are shown in Figure 4.7. Panel (a) displays the distributional and mean-based summaries for all six groups. The wk24 eOD→core group shows the highest estimated heavy-chain mutation levels across nearly all quantiles, whereas the wk8 eOD group shows the lowest mutation levels. The mean-based summaries exhibit a similar ordering across groups. Panel (b) focuses on the comparison between wk24 eOD→core and wk24 eOD→eOD. The estimated barycenter difference, defined as wk24 eOD→core minus wk24 eOD→eOD, is -0.037% at the 0.01 quantile (95% CI: -0.48, 0.41), 0.47% at the median (95% CI: 0.074, 0.87), and 1.3% at the 0.99 quantile (95% CI: 0.44, 2.2). Thus, the estimated group difference is not uniform across the mutation distribution; instead, it becomes more pronounced among sequences with higher mutation rates. The supremum test rejects the null hypothesis of equal barycenters ($p < 0.001$), providing strong evidence that the two groups differ in their mutation barycenter distributions. The corresponding mean-based analysis estimates an average difference of 0.607% (95% CI: 0.20, 1.0; $p = 0.005$).

4.6 Discussion

We proposed a framework for two-sample comparisons of sequence-derived numeric feature distributions from deep sequencing data. Rather than reducing each individual's sequence

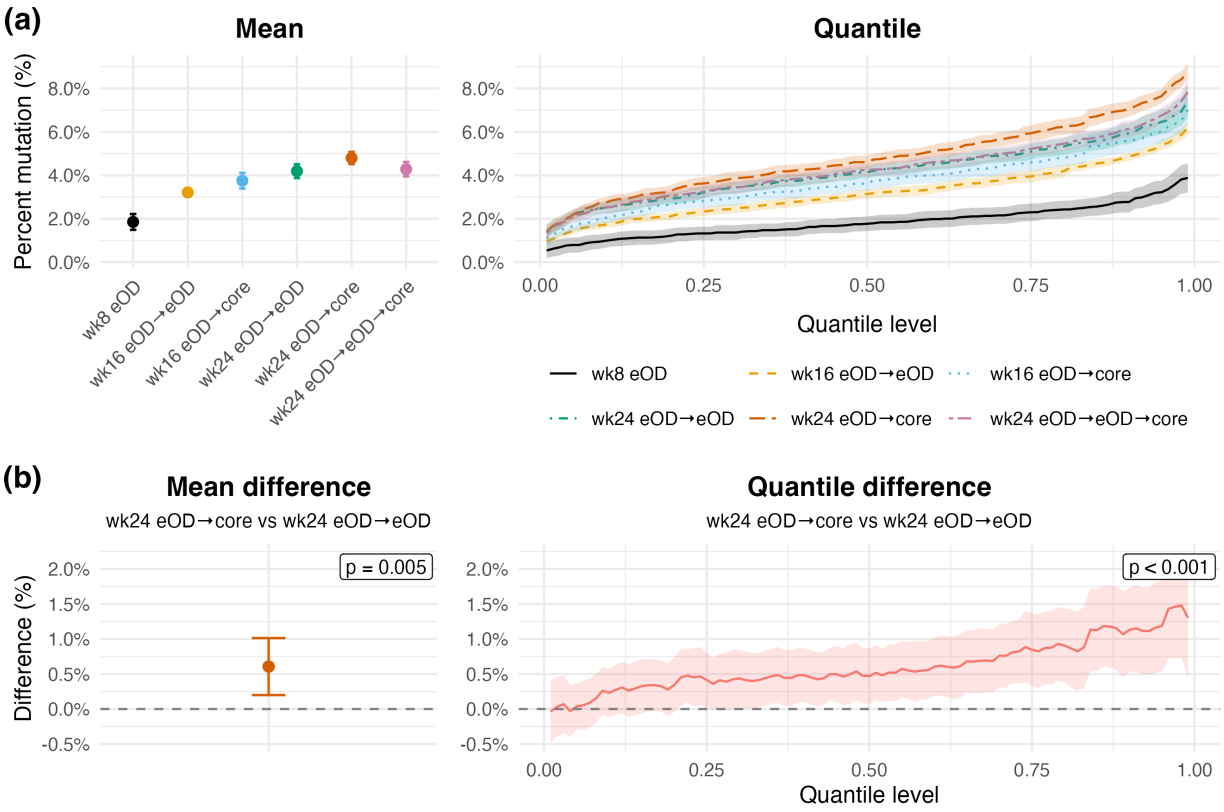


Figure 4.7: BCR sequencing analysis of vaccine-elicited VRC01-class responses using heavy-chain percent mutation as the sequence-level feature. (a) Estimated group summaries for the six vaccination schedule–timepoint groups. The left panel shows the mean percent mutation for each group, and the right panel shows the corresponding empirical Wasserstein barycenter quantile functions. (b) Two-sample comparison of week 24 eOD→core versus week 24 eOD→eOD. The left panel shows the estimated difference in mean percent mutation, and the right panel shows the estimated quantile difference map, defined as the week 24 eOD→core barycenter quantile function minus the week 24 eOD→eOD barycenter quantile function. Vertical bars and shaded bands indicate pointwise 95% confidence intervals, and the dashed horizontal line marks zero difference. Reported p-values correspond to the unequal-variance two-sample t-test for the mean comparison and the supremum test for equality of Wasserstein barycenters for the quantile comparison.

set to a scalar summary, the approach treats the sequence-derived feature values for each individual as a distribution-valued outcome. Treatment arms are summarized using Wasserstein barycenters, and arm differences are described through quantile difference maps. This allows researchers to compare arms across the full within-host feature distribution and to distinguish uniform shifts from localized tail differences or differences in distributional shape. In contrast, mean-based analyses can identify average differences but may obscure where in the distribution those differences arise.

A key practical consideration is sequencing depth. When sequencing depth is low, an individual's empirical distribution may poorly approximate the corresponding within-host feature distribution, which can bias estimation of the treatment arm barycenter, particularly in the tails. In our simulations, the empirical barycenter performed poorly in low sequencing settings, especially when sequencing depths differed between arms. In such settings, the MCB estimator is preferable, provided that there are enough individuals to estimate the binomial mixture components. In higher sequencing depth settings, both the empirical barycenter and spline-configured MCB estimator performed well on the quantile grid considered here, even when some individuals had low sequencing depths. The performance of the empirical barycenter is consistent with the theoretical results of Zhou and Müller (2024), which allow some units to have few observations when others have many. Thus, both estimators can be used in higher sequencing depth applications, with the empirical barycenter offering a less model-dependent approach and the spline-configured MCB estimator providing an alternative that may be particularly useful in sparse settings or when sequencing depth is variable.

The quantile grid should also be chosen with the observed sequencing depths and scientific target of interest. Our simulations considered quantiles from 0.01 to 0.99, but less extreme grids, such as 0.2 to 0.8, may be reliable at lower sequencing depths. Conversely, inference farther into the tails requires greater sequencing depth and is more sensitive to estimator choice. Developing data-adaptive tools for selecting the quantile grid would improve usability of our framework and reduce reliance on manual choices.

Our presentation focused on the two independent sample setting. In randomized studies (not subject to post-randomization selection bias), this contrast can be interpreted as an average treatment effect on the treatment arm-level feature distributions. In observational studies, however, the unadjusted comparison is generally descriptive, since baseline confounders may differ between groups. To estimate an average treatment effect in such settings, the analysis must adjust for baseline confounder variables under standard causal assumptions such as consistency, exchangeability, and positivity. Existing approaches for causal inference with distributional outcomes provide one path forward for estimating covariate-adjusted treatment effects on barycenter quantile functions (Lin et al., 2023). The empirical barycenter and pre-smoothed empirical barycenter can be incorporated into these approaches directly, whereas extending the MCB estimator to covariate-adjusted settings requires additional work. With discrete covariates, one could estimate covariate-specific barycenters and then marginalize over a target covariate distribution to obtain an adjusted average treatment effect. For continuous covariates, inverse probability weighting, as in Example 7.2 of Chapter 3, may provide a useful direction.

Several methodological extensions are also natural. The present work considered one-dimensional sequence features, but some sequence-derived summaries are multivariate and would require alternatives to quantile-function representations. Longitudinal extensions would allow investigators to study how within-host feature distributions evolve over time and whether those trajectories differ between treatment arms. Relatedly, paired or repeated-measures versions of the proposed tests would be useful when the same individuals are observed under multiple conditions or across multiple visits.

4.7 Acknowledgments

The authors thank the participants and investigators of HVTN 703/704 and IAVI G002/G003, and Craig Magaret for constructing the HVTN 703/704 trial datasets. The authors also thank Lillian Cohn, Paul Edlefsen, Raabya Rossenkhan, and James Ludwig for their helpful discussions and insights. This research was supported by the National Institute of Allergy

and Infectious Disease of the National Institutes of Health under grants UM1 AI068635 and R37AI054165. The content is the sole responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases and the National Institutes of Health.

Chapter 5

COMPARING DISTRIBUTIONAL CHANGE OVER TWO TIMEPOINTS: APPLICATION TO THE AMP STUDIES

Chapter Summary

The AMP trials collected viral sequences at two early timepoints from participants who acquired HIV-1. Motivated by these data, we propose an approach for comparing within-host viral evolution between treatment arms. For each participant, the change in the distribution of a sequence-derived feature is summarized by a normalized quantile shift map, which describes how the within-host feature distribution changes across quantile levels. Individual normalized shift maps are averaged to estimate treatment arm-specific mean normalized shift maps, and differences between arms are assessed using a global supremum-based test. For estimation, each participant's normalized shift map is estimated by using their empirical quantile functions from the observed sequences at each timepoint and taking their time-normalized difference; this quantity is then used in a plug-in estimator.

Because sequencing depth was low for some AMP participants, the empirical distributions may not closely approximate the underlying within-host distributions. Therefore, we examined the proposed estimator's performance in simulations using sequencing depths comparable to those in AMP. In these simulations, we found that the empirical estimator performed well across most quantile levels, although slight bias and reduced confidence interval coverage were observed in the tail quantile levels. Applying the proposal to the AMP analysis, the mean normalized quantile shift map did not differ significantly between the low-dose VRC01 and placebo arms ($p = 0.9$). In contrast, the high-dose VRC01 arm exhibited positive shifts in VRC01 epitope distance which were not observed in the placebo arm ($p < 0.001$), a pattern consistent with a shift toward increased VRC01 resistance. In future work, we aim to

extend this framework to explicitly account for low sequencing depth.

5.1 Introduction

After infection is established, the HIV-1 viral quasispecies evolves over time within the host. Rather than existing as a static genome, the viral population comprises an interconnected swarm of related variants (Domingo et al., 2012). These variants may interact and respond collectively as the viral population adapts to changing fitness landscapes (Andino and Domingo, 2015; Singh et al., 2023). Figure 5.1 illustrates these dynamics using a toy example of early HIV-1 infection. Timepoint (a) represents the genetic bottleneck at transmission, when one or a few transmitted/founder variants establish infection. Timepoint (b) shows early diversification of the founder population as mutations accumulate during rapid viral replication. Timepoint (c) illustrates how host immune pressure, including HIV-specific T-cell responses and antibody-mediated responses, can favor escape variants and shift the evolution of the quasispecies in a particular direction.

External interventions can further reshape the evolution of the quasispecies. Antiretroviral therapy (ART), for example, suppresses viral replication and typically limits ongoing HIV-1 evolution if treatment is effective. However, when viral replication continues in the presence of drug pressure, resistant variants may be selected, shifting the quasispecies toward a drug-resistant population that can evade treatment (Kijak et al., 2002; Metzner et al., 2009). Similar concerns arise with broadly neutralizing antibodies (bNAbs), such as VRC01, which are a promising approach for HIV-1 prevention, treatment, and cure (Tebas, 2025). Like ART, bNAbs can impose strong selective pressure on the viral population, potentially enriching for variants with increased antibody resistance. These examples motivate studying how the within-host collection of viral variants shifts under different sources of selective pressure.

The AMP trials (HVTN 703/HVTN 704), described in Section 4.5.1, provide an opportunity to study viral quasispecies evolution under VRC01-mediated selective pressure (Corey et al., 2021). Among participants who acquired HIV-1, viral *env* sequences were generated

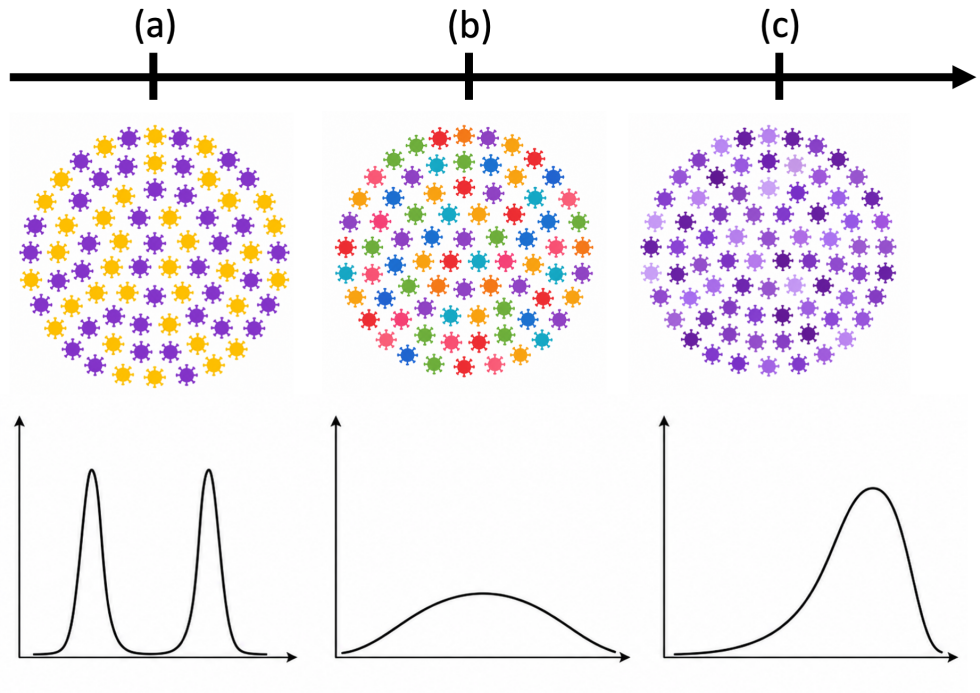


Figure 5.1: Toy example of HIV-1 quasispecies composition at (a) time of infection, (b) shortly post-infection, and (c) selection pressure in response to immune or treatment. The toy schematics at the top represent the diversity of HIV-1 in the quasispecies, with different colors representing different viral types. The distributions shown as density functions on the bottom portion of the figure represent an illustrative distribution of a numeric feature across all viruses in the quasispecies at the particular timepoint.

using PacBio SMRT-UMI deep sequencing from samples collected at diagnosis and again several weeks later. In one analysis of these two-timepoint data, Williamson et al. (2026) found that, among VRC01 recipients co-infected with sensitive and resistant viral lineages, VRC01-sensitive lineages tended to decrease in frequency over 5–20 days, whereas resistant lineages increased, suggesting early post-acquisition selection favoring VRC01-resistant viruses. In a related analysis, Bai et al. (submitted) studied post-infection dynamics with respect to epitope distance, a measure that describes how far a sequence is from a VRC01-sensitive reference sequence at sites relevant to VRC01 binding. They found that the slope of change in VRC01 epitope distance—defined as the difference in epitope distance between sequences at each timepoint divided by the number of days between timepoints—was significantly higher among VRC01 recipients than among placebo recipients. However, this analysis reduced each participant's viral population to a single representative sequence at each timepoint. We extend this analysis by using the entire set of sequences for each participant.

As in the previous chapters, we represent each sequence by a numeric feature, so that the collection of sequence features from a given participant at a given timepoint can be viewed as a one-dimensional distribution. In the AMP application, our primary feature is VRC01 epitope distance. The observed data for an individual therefore resemble the bottom panel of Figure 5.1: a distribution of epitope distances observed at each timepoint. The goal of this chapter is to develop a framework for characterizing how these within-host distributions change over time and for determining whether their patterns of change differ across treatment arms.

We will treat the evolution of the quasispecies in the individual with respect to the feature of interest as functional data. Specifically, we represent each timepoint-specific distribution by its quantile function and define an individual-level quantile shift map as the time-normalized difference between the quantile functions at the two timepoints. This function describes the per-unit-time change in each quantile of the within-host distribution. We then estimate and compare the mean normalized quantile shift maps across treatment groups and use a global testing procedure to assess whether the groups differ at any quantile level.

In the AMP application, this approach allows us to study whether changes in the within-host distribution of VRC01 epitope distance differ between the VRC01 and placebo arms.

5.2 Methodology

5.2.1 Data structure and quantile representation

For each individual i, let $Z_i \in \{1, 2\}$ denote treatment arm. We suppose that each individual is observed at two timepoints, indexed by $t \in \{1, 2\}$, and let $s_{t,i}$ denote the sampling time for individual i at timepoint t, with $s_{1,i} < s_{2,i}$. We define the elapsed time between the two measurements as $\ell_i := s_{2,i} - s_{1,i}$. Let $Y_{t,i}$ denote individual i's underlying within-host distribution of a numeric sequence feature at timepoint t. Thus, $Y_{1,i}$ and $Y_{2,i}$ represent the individual's latent (unobserved) feature distributions at the first and second timepoints, respectively. We assume that, for each individual i, both latent distributions have finite second moments.

The latent distributions $Y_{1,i}$ and $Y_{2,i}$ are not directly observed. Instead, at each timepoint t, we observe $m_{t,i}$ feature measurements $Y_{t,i1}, \ldots, Y_{t,im_{t,i}}$, where $m_{t,i}$ denotes the number of observed feature measurements for individual i at timepoint t. We assume that, conditional on the latent timepoint-specific distribution $Y_{t,i}$ and $m_{t,i}$,

$$Y_{t,i1}, \ldots, Y_{t,im_{t,i}} \overset{\text{i.i.d.}}{\sim} Y_{t,i}, \qquad t \in \{1, 2\}.$$

This assumption is made separately at each timepoint. The latent distributions $Y_{1,i}$ and $Y_{2,i}$ may be statistically dependent within an individual, since they arise from repeated measurements on the same individual over time. We assume that individuals are independent and that the individual-level data-generating process is identically distributed within each group. In total, we observe n individuals, with n_1 in Group 1 ($Z = 1$) and n_2 in Group 2 ($Z = 2$), so $n = n_1 + n_2$.

From the observed measurements for individual i at timepoint t, we construct the empir-

ical distribution

$$\widehat{Y}_{t,i}^{\text{emp}} = \frac{1}{m_{t,i}} \sum_{j=1}^{m_{t,i}} \delta_{Y_{t,ij}}, \qquad t \in \{1, 2\}.$$

For a generic probability measure λ, let $F_{\lambda}^{-}(u)$ denote its quantile function for $u \in (0, 1)$. To simplify notation, for each individual i and timepoint t, we write

$$q_{t,i}(u) := F_{Y_{t,i}}^{-}(u), \qquad u \in (0, 1),$$

for the quantile function of the individual's latent within-host feature distribution at timepoint t. Therefore, $Y_{t,i}$ denotes the latent distribution itself, whereas $q_{t,i}$ denotes its quantile-function representation. The corresponding empirical plug-in estimator of $q_{t,i}$ is defined as

$$\widehat{q}_{t,i}^{\text{emp}}(u) := F_{\widehat{Y}_{t,i}^{\text{emp}}}^{-}(u), \qquad u \in (0, 1).$$

5.2.2 Normalized quantile shifts

For each individual i, we define the *quantile shift map* as

$$D_i(u) := q_{2,i}(u) - q_{1,i}(u), \qquad u \in (0, 1).$$

The function $D_i(u)$ gives the change in the uth quantile of the individual's latent within-host feature distribution between the two timepoints. Positive values of $D_i(u)$ indicate that the uth quantile of the feature distribution is larger at timepoint 2 than at timepoint 1, while negative values indicate that the corresponding quantile value is smaller at timepoint 2.

Because the elapsed time between measurements may vary across individuals, we define the *normalized quantile shift map* as

$$R_i(u) := \frac{D_i(u)}{\ell_i} = \frac{q_{2,i}(u) - q_{1,i}(u)}{\ell_i}, \qquad u \in (0, 1). \tag{5.1}$$

The function $R_i(u)$ gives the average rate of change in the individual's uth quantile over the

interval between the two measurements.

The empirical analogue of the quantile shift map is

$$\widehat{D}_i^{\text{emp}}(u) := \widehat{q}_{2,i}^{\text{emp}}(u) - \widehat{q}_{1,i}^{\text{emp}}(u), \qquad u \in (0,1).$$

Similarly, the empirical analogue of the normalized quantile shift map is

$$\widehat{R}_i^{\text{emp}}(u) := \frac{\widehat{D}_i^{\text{emp}}(u)}{\ell_i} = \frac{\widehat{q}_{2,i}^{\text{emp}}(u) - \widehat{q}_{1,i}^{\text{emp}}(u)}{\ell_i}, \qquad u \in (0,1).$$

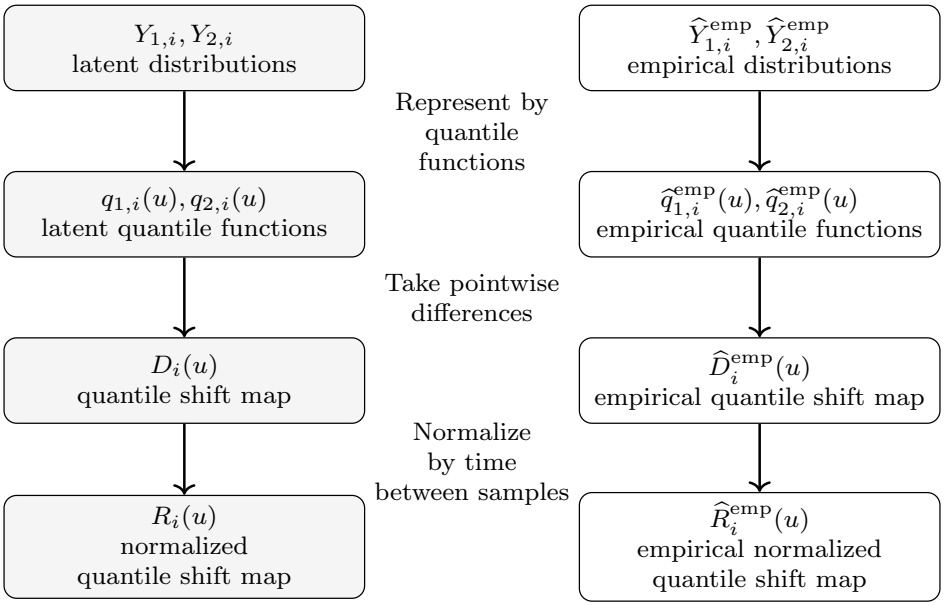


Figure 5.2: Schematic of the latent and empirical quantities used to construct an individual normalized quantile shift map. The left column shows the latent within-host distributions and derived quantities, and the right column shows their empirical counterparts estimated from the observed sequence-level measurements.

The normalized quantile shift map has a natural geometric interpretation in Wasserstein space. Suppose two real-valued observations x_1 and x_2 are collected at times s_1 and s_2, with $s_1 < s_2$. In Euclidean space, the quantity $(x_2 - x_1)/(s_2 - s_1)$ is the velocity of the linear path connecting observations x_1 and x_2 over the elapsed time $s_2 - s_1$ (Figure 5.3(a)). For

distribution-valued outcomes observed at times s_1 and s_2, the analogous path is a constant-speed geodesic in Wasserstein space (Villani, 2009). In the one-dimensional setting, this geodesic has a simple form because the Wasserstein distance can be expressed in terms of quantile functions.

Specifically, let λ_1 and λ_2 be two probability distributions on $\mathbb{R}$ with finite second moments. The squared 2-Wasserstein distance between λ_1 and λ_2 is

$$W_2^2(\lambda_1, \lambda_2) = \int_0^1 \{F_{\lambda_2}^-(u) - F_{\lambda_1}^-(u)\}^2 \, du.$$

Thus, in one dimension, Wasserstein geometry can be represented through quantile functions in the common linear space $L^2(0,1)$.

The constant-speed Wasserstein geodesic from λ_1 to λ_2, observed over times $s_1 < s_2$, is obtained by linear interpolation of their quantile functions (Panaretos and Zemel, 2019). Let $\omega(s) := \frac{s-s_1}{s_2-s_1}$ denote the relative position of time s between s_1 and s_2, so that $\omega(s) = 0$ at the first timepoint and $\omega(s) = 1$ at the second timepoint. Let λ_s denote the distribution along the geodesic at time s, and let $q_s := F_{\lambda_s}^-$ denote its quantile function. Then

$$q_s(u) = \{1 - \omega(s)\}F_{\lambda_1}^-(u) + \omega(s)F_{\lambda_2}^-(u), \qquad u \in (0,1).$$

Differentiating $q_s(u)$ with respect to s yields the constant velocity of the geodesic, expressed as a function of the quantile level:

$$v_{\lambda_1,\lambda_2}(u) = \frac{F_{\lambda_2}^-(u) - F_{\lambda_1}^-(u)}{s_2 - s_1}, \qquad u \in (0,1).$$

This velocity gives the per-unit-time rate of change in each quantile along the geodesic (Figure 5.3(b)).

Applying this construction to individual i, with $\lambda_1 = Y_{1,i}$, $\lambda_2 = Y_{2,i}$, and $s_2 - s_1 = \ell_i$, the velocity of the constant-speed Wasserstein geodesic connecting the two latent distributions

is

$$v_i(u) = \frac{q_{2,i}(u) - q_{1,i}(u)}{\ell_i} = R_i(u).$$

Thus, R_i is the velocity function of the constant-speed Wasserstein geodesic connecting the individual's latent distributions at the two timepoints. At each quantile level u, $R_i(u)$ gives the change per unit time in the corresponding quantile along this geodesic.

5.2.3 Group-level estimands and hypothesis testing

This interpretation motivates treating R_i as an individual-level rate of distributional change. Assuming the latent distributions have finite second moments, each R_i belongs to the common space $L^2(0,1)$, so the normalized quantile shift maps can be averaged across individuals.

Our primary estimand is the group-specific *mean normalized quantile shift map*. For $z \in \{1,2\}$, we define

$$\rho_z(u) := \mathbb{E}\{R_i(u) \mid Z_i = z\}, \qquad u \in (0,1). \tag{5.2}$$

Equivalently,

$$\rho_z(u) = \mathbb{E}\left[\frac{q_{2,i}(u) - q_{1,i}(u)}{\ell_i} \mid Z_i = z\right], \qquad u \in (0,1).$$

Thus, $\rho_z(u)$ gives the mean change per-unit of time in the uth quantile across individuals in group z.

We test whether the two groups have equal mean normalized quantile shift maps over a compact set of quantile levels $\mathcal{U} \subset (0,1)$. The global null hypothesis is

$$H_0 : \rho_1(u) = \rho_2(u) \text{ for all } u \in \mathcal{U}.$$

5.2.4 Estimation and inference procedure

We use a plug-in estimation strategy in which each latent timepoint-specific distribution is replaced by the corresponding empirical distribution. This yields an estimated normal-

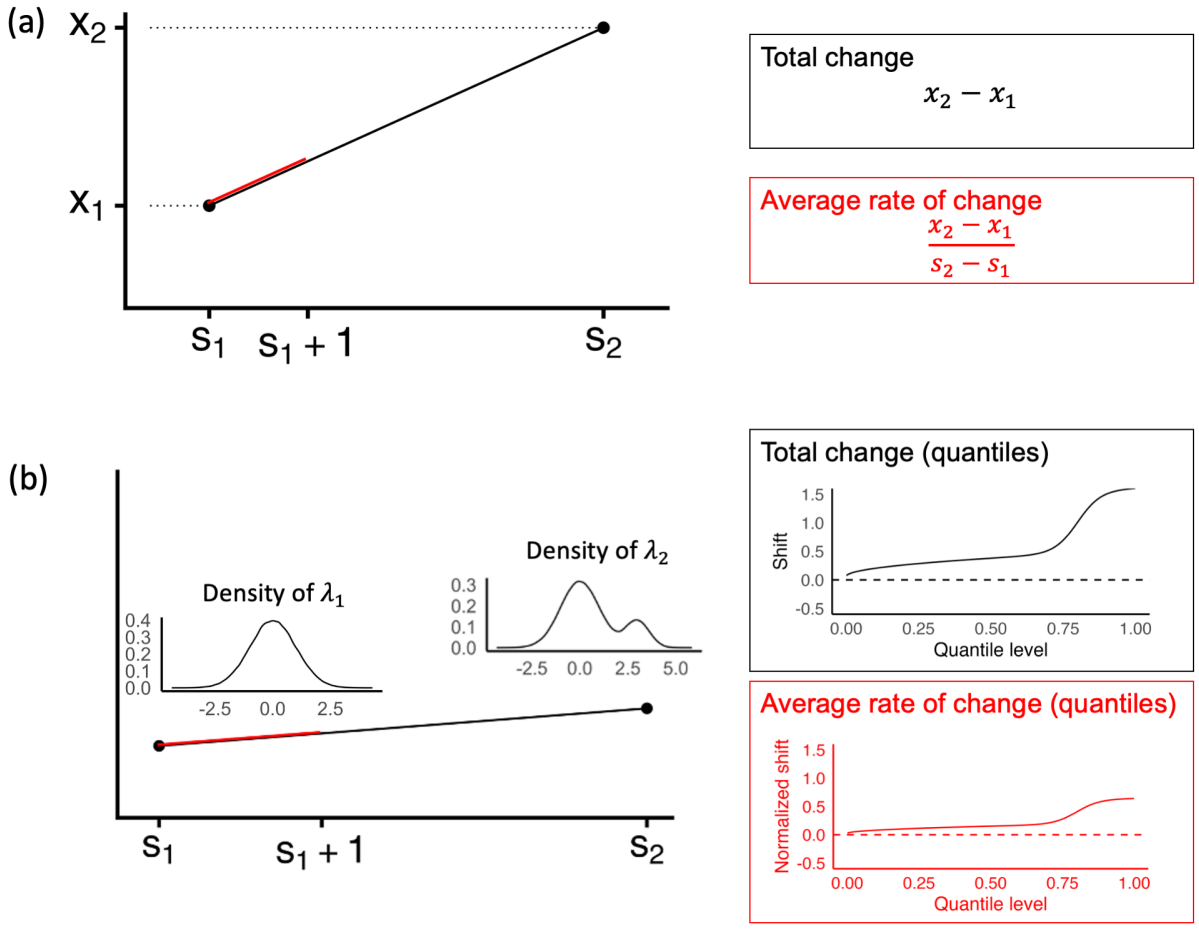


Figure 5.3: Euclidean and Wasserstein interpretations of change between two timepoints. (a) For real-valued observations x_1 and x_2 collected at times $s_1 < s_2$, the total change is $x_2 - x_1$, and the average rate of change is $(x_2 - x_1)/(s_2 - s_1)$. (b) For distributions λ_1 and λ_2, the total change is represented by the quantile shift map $F^-_{\lambda_2}(u) - F^-_{\lambda_1}(u)$, and the average rate of change is represented by the corresponding normalized quantile shift map. In both panels, the red segment represents the change over one unit of time along the linear path in (a) and the constant-speed Wasserstein geodesic in (b).

ized quantile shift map, $\widehat{R}_i^{\text{emp}}$, for each individual. We estimate the group-specific mean normalized quantile shift map by averaging these individual maps:

$$\widehat{\rho}_z^{\text{emp}}(u) := \frac{1}{n_z} \sum_{i:Z_i=z} \widehat{R}_i^{\text{emp}}(u), \qquad u \in (0,1).$$

To establish pointwise asymptotic normality for $\widehat{\rho}_z^{\text{emp}}(u)$ as an estimator of $\rho_z(u)$, we assume that the individual-level estimation error is negligible relative to the between-individual sampling variation. Specifically, for each fixed $u \in (0,1)$, suppose that

$$\frac{1}{\sqrt{n_z}} \sum_{i:Z_i=z} \left\{ \widehat{R}_i^{\text{emp}}(u) - R_i(u) \right\} = o_p(1). \tag{5.3}$$

Under the condition above, the central limit theorem applied to the individual normalized latent quantile shift maps gives

$$\sqrt{n_z} \left\{ \widehat{\rho}_z^{\text{emp}}(u) - \rho_z(u) \right\} \rightsquigarrow N\left(0, \sigma_z^2(u)\right), \tag{5.4}$$

where $\sigma_z^2(u) = \text{Var}\{R_i(u) \mid Z_i = z\}$. We estimate $\sigma_z^2(u)$ using the sample variance of the estimated individual shift maps:

$$\widehat{\sigma}_z^2(u) = \frac{1}{n_z - 1} \sum_{i:Z_i=z} \left\{ \widehat{R}_i^{\text{emp}}(u) - \widehat{\rho}_z^{\text{emp}}(u) \right\}^2.$$

Provided that $\widehat{\sigma}_z^2(u)$ is consistent for $\sigma_z^2(u)$, an approximate $100(1-\alpha)\%$ pointwise confidence interval for $\rho_z(u)$ is

$$\widehat{\rho}_z^{\text{emp}}(u) \pm z_{1-\alpha/2} \frac{\widehat{\sigma}_z(u)}{\sqrt{n_z}}.$$

For simultaneous inference over a set of quantile levels, pointwise convergence is not sufficient. Let $\mathcal{U} \subset (0,1)$ be compact, and suppose that the centered latent normalized

quantile shift map process satisfies

$$\frac{1}{\sqrt{n_z}} \sum_{i:Z_i=z} \{R_i - \rho_z\} \rightsquigarrow \mathbb{G}_z \qquad \text{in } \ell^\infty(\mathcal{U}).$$

This weak convergence may be established under conditions including stochastic equicontinuity. We also assume that the error from estimating the individual normalized quantile shift maps is uniformly negligible over $\mathcal{U}$:

$$\sup_{u \in \mathcal{U}} \left| \frac{1}{\sqrt{n_z}} \sum_{i:Z_i=z} \left\{ \widehat{R}_i^{\text{emp}}(u) - R_i(u) \right\} \right| = o_p(1). \tag{5.5}$$

It follows that

$$\sqrt{n_z}\,\{\widehat{\rho}_z^{\text{emp}} - \rho_z\} \rightsquigarrow \mathbb{G}_z \qquad \text{in } \ell^\infty(\mathcal{U}), \tag{5.6}$$

where $\ell^\infty(\mathcal{U})$ is the space of bounded functions on $\mathcal{U}$ equipped with the supremum norm. The limiting process $\mathbb{G}_z$ is a Gaussian process with mean zero and covariance kernel

$$\Sigma_z(u, v) = \text{Cov}\{R_i(u), R_i(v) \mid Z_i = z\}, \qquad u, v \in \mathcal{U}.$$

The conditions in Equations (5.3) and (5.5) require the aggregate error from estimating the individual normalized quantile shift maps to be asymptotically negligible relative to the sampling variation across individuals. These conditions may be difficult to justify when sequencing depth is low for some individuals, because the empirical quantile functions may not closely approximate the latent quantile functions. Therefore, to stress test our estimator, we will examine the performance of this estimator under sequencing depths similar to those observed in the AMP trials in our simulation studies.

5.2.5 Hypothesis testing

Global hypothesis testing can be based on the Gaussian process convergence result (5.6) from the previous section. We consider the null hypothesis $H_0 : \Delta_\rho(u) = 0$ for all $u \in \mathcal{U}$,

where $\Delta_\rho(u) = \rho_2(u) - \rho_1(u)$. The alternative hypothesis is that $\Delta_\rho(u) \neq 0$ for at least one $u \in \mathcal{U}$.

We focus on a global supremum test statistic designed to detect departures from the null at any quantile level in $\mathcal{U}$. Let $\widehat{\Delta}_\rho(u) = \widehat{\rho}_2^{\text{emp}}(u) - \widehat{\rho}_1^{\text{emp}}(u)$. Assuming that the two groups are independent, we estimate its pointwise standard error by

$$\widehat{\text{se}}\{\widehat{\Delta}_\rho(u)\} = \left\{ \frac{\widehat{\sigma}_2^2(u)}{n_2} + \frac{\widehat{\sigma}_1^2(u)}{n_1} \right\}^{1/2}.$$

We then form the studentized difference process $\widehat{Z}_\Delta(u) = \widehat{\Delta}_\rho(u)/\widehat{\text{se}}\{\widehat{\Delta}_\rho(u)\}$. Studentization places the estimated differences across quantile levels on a comparable scale because the variability of $\widehat{\Delta}_\rho(u)$ may vary with u. The supremum statistic is

$$T_\infty = \sup_{u \in \mathcal{U}} \left| \widehat{Z}_\Delta(u) \right| = \sup_{u \in \mathcal{U}} \left| \frac{\widehat{\Delta}_\rho(u)}{\widehat{\text{se}}\{\widehat{\Delta}_\rho(u)\}} \right|.$$

Under H_0 and the assumptions above, including consistency of the estimated standard errors, the studentized difference process $\widehat{Z}_\Delta$ converges weakly to a Gaussian process Z_Δ with mean zero and unit pointwise variance. By the continuous mapping theorem, $T_\infty \rightsquigarrow \sup_{u \in \mathcal{U}} |Z_\Delta(u)|$. Because the distribution of this supremum depends on the unknown covariance structure of Z_Δ and is generally not available in closed form, we approximate it by simulating Gaussian processes using an estimate of the limiting covariance kernel.

In practice, we approximate the supremum statistic on a grid $u_1, \ldots, u_K \in \mathcal{U}$, so that

$$T_\infty \approx \max_{1 \leq k \leq K} \left| \frac{\widehat{\Delta}_\rho(u_k)}{\widehat{\text{se}}\{\widehat{\Delta}_\rho(u_k)\}} \right|.$$

For each group z, we estimate the covariance between the individual shift maps at u_k and u_ℓ by

$$\widehat{\Sigma}_z(u_k, u_\ell) = \frac{1}{n_z - 1} \sum_{i: Z_i = z} \left\{ \widehat{R}_i^{\text{emp}}(u_k) - \widehat{\rho}_z^{\text{emp}}(u_k) \right\} \left\{ \widehat{R}_i^{\text{emp}}(u_\ell) - \widehat{\rho}_z^{\text{emp}}(u_\ell) \right\}.$$

Because the two groups are independent, the estimated covariance matrix $\widehat{\boldsymbol{\Sigma}}_\Delta$ of the group difference has entries $\widehat{\Sigma}_{\Delta,k\ell} = \widehat{\Sigma}_2(u_k, u_\ell)/n_2 + \widehat{\Sigma}_1(u_k, u_\ell)/n_1$ for $k, \ell = 1, \ldots, K$.

To approximate the null distribution, we simulate $\boldsymbol{G}_b \sim N(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_\Delta)$ for $b = 1, \ldots, B$ and compute

$$T_\infty^{(b)} = \max_{1 \le k \le K} \left| \frac{G_{b,k}}{\sqrt{\widehat{\Sigma}_{\Delta,kk}}} \right|.$$

The Monte Carlo p-value is the proportion of simulated statistics $T_\infty^{(b)}$ that are at least as large as the observed test statistic. We reject H_0 at level α when this p-value is less than or equal to α.

5.3 Simulation study

5.3.1 Simulation Study 1: estimating normalized quantile shift maps

In Simulation Study 1, we simulate individuals with differing baseline distributions and participant-specific normalized quantile shift maps, with the goal of evaluating estimation of the mean normalized quantile shift map. Specifically, we assess the performance of the empirical estimator under sample sizes and sequencing depths similar to those observed in the AMP trials. For each individual, the timepoint 1 distribution is a Normal distribution, with its mean sampled from Uniform$(-1, 1)$ and its standard deviation sampled from Uniform$(0.5, 2)$. Sequencing depths at the two timepoints are sampled independently from the integers 1 through 300, and the elapsed time between samples is generated from Uniform$(0.5, 2)$. We consider sample sizes of $n = 20, 50$, and 100 individuals.

For each individual, we generate a normalized quantile shift map $R_i(u)$ according to one of four settings and construct the timepoint 2 quantile function as

$$q_{2,i}(u) = q_{1,i}(u) + \ell_i R_i(u).$$

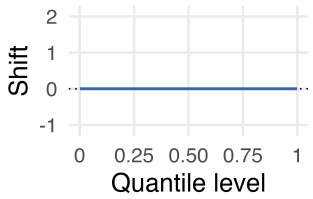
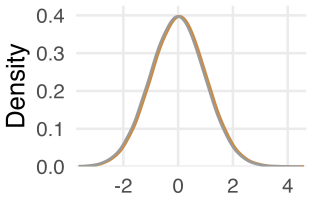
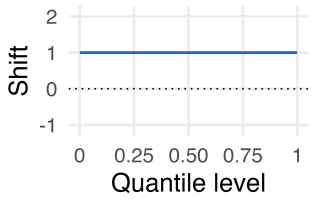
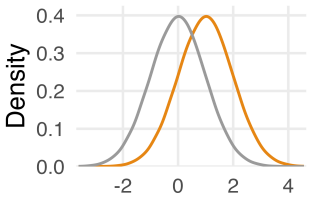
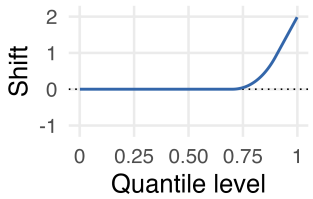
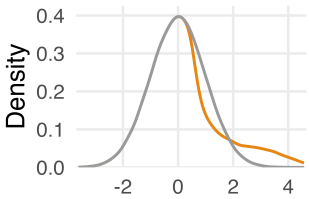
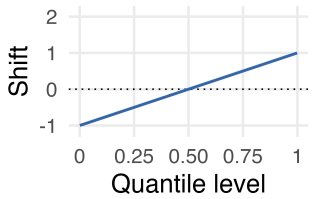
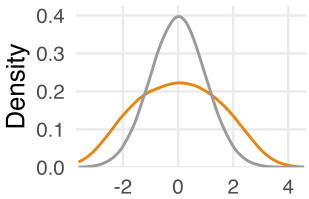
The four settings are:

1. **No mean change:** Each individual's normalized quantile shift map is constant across quantile levels, with the constant sampled from Uniform$(-1, 1)$. The mean normalized quantile shift map is therefore zero across all quantile levels.
2. **Uniform shift:** Each individual's normalized quantile shift map is constant across quantile levels, with the constant sampled from Uniform$(0, 2)$. The mean normalized quantile shift map is constant across quantile levels, corresponding to an average location shift of the entire distribution.
3. **Upper tail shift:** Each individual's normalized quantile shift map is zero below a randomly selected quantile sampled from Uniform$(0.7, 0.9)$. Above this quantile, the map increases linearly from zero to a value sampled from Uniform$(1, 3)$ at the upper boundary. Change therefore occurs primarily at upper quantile levels.
4. **Variability increase:** Each individual's normalized quantile shift map is linear, with its intercept sampled from Uniform$(-2, 0)$ and its slope sampled from Uniform$(1, 3)$. The mean normalized quantile shift map decreases at lower quantiles and increases at upper quantiles, corresponding to an increase in distributional variability over time.

These settings are summarized in Table 5.1.

After generating the latent distributions at both timepoints, we simulate observed sequence feature values from each distribution using the sampled sequencing depths. From the simulated data, we estimate each individual's empirical quantile shift map and divide it by the elapsed time between observations to obtain an estimated normalized quantile shift map. We then average these maps across individuals to estimate the mean normalized quantile shift map. For each setting and sample size, we simulate 500 datasets and compare the estimated mean normalized quantile shift map with its known truth over the quantile grid $u \in 0.01, \ldots, 0.99$. We examine pointwise bias, mean estimated standard error, and coverage of 95% pointwise confidence intervals.

Table 5.1: Quantile shift settings considered in Simulation Study 1.

Setting	**Quantile shift map**	**Description**	**Visual**
No mean change		The mean normalized quantile shift map is zero across all quantile levels, indicating no systematic change in the feature distribution over time.	
Uniform shift		All quantiles shift by the same amount, corresponding to an average location shift of the entire feature distribution.	
Upper-tail shift		Change occurs primarily at upper quantile levels, corresponding to an average shift among sequences with larger feature values.	
Variability increase		Lower quantiles decrease while upper quantiles increase, corresponding to an increase in the variability of the feature distribution over time.	

Results

Results from Simulation Study 1 are shown in Figure 5.4. Across the four data-generating settings, the empirical estimator generally recovered the true mean normalized quantile shift map well. In the no mean change, uniform shift, and variability increase settings, bias was small across most quantile levels for all sample sizes considered ($n = 20, 50, 100$), and confidence interval coverage was close to the nominal level. As expected, the mean standard error decreased as the number of individuals increased, although standard errors were larger near the boundaries of the quantile domain.

Bias and undercoverage were greatest in the upper-tail shift setting. In this scenario, the true shift map changes rapidly near the highest quantile levels, and the estimator exhibited some bias in the upper tail. This bias is likely due to finite sequencing depth, since estimating extreme quantiles is difficult when some individuals contribute relatively few sequences. Coverage also decreased near the highest quantiles, particularly as sample size increased and confidence intervals became narrower. However, even under sequencing depths sampled uniformly from 1 to 300 sequences per person per timepoint, the magnitude of the bias was small relative to the size of the true upper-tail shift. In supplementary simulations with sequencing depth fixed at 1,000 sequences per person per timepoint (Supplementary Figure E.1), the upper-tail bias was reduced and confidence interval coverage was closer to nominal, supporting the interpretation that the observed bias is primarily driven by low sequencing depth rather than by the estimator itself.

5.3.2 Simulation Study 2: hypothesis testing

In Simulation Study 2, we evaluate the operating characteristics of the global hypothesis testing procedure for comparing mean normalized quantile shift maps between two arms. We simulate two treatment arms with equal group sizes of $n = 20, 50$, or 100 individuals per arm. Baseline distributions, elapsed times, and observed sequence derived feature values are generated as described in Simulation Study 1. At each timepoint, sequencing depth is

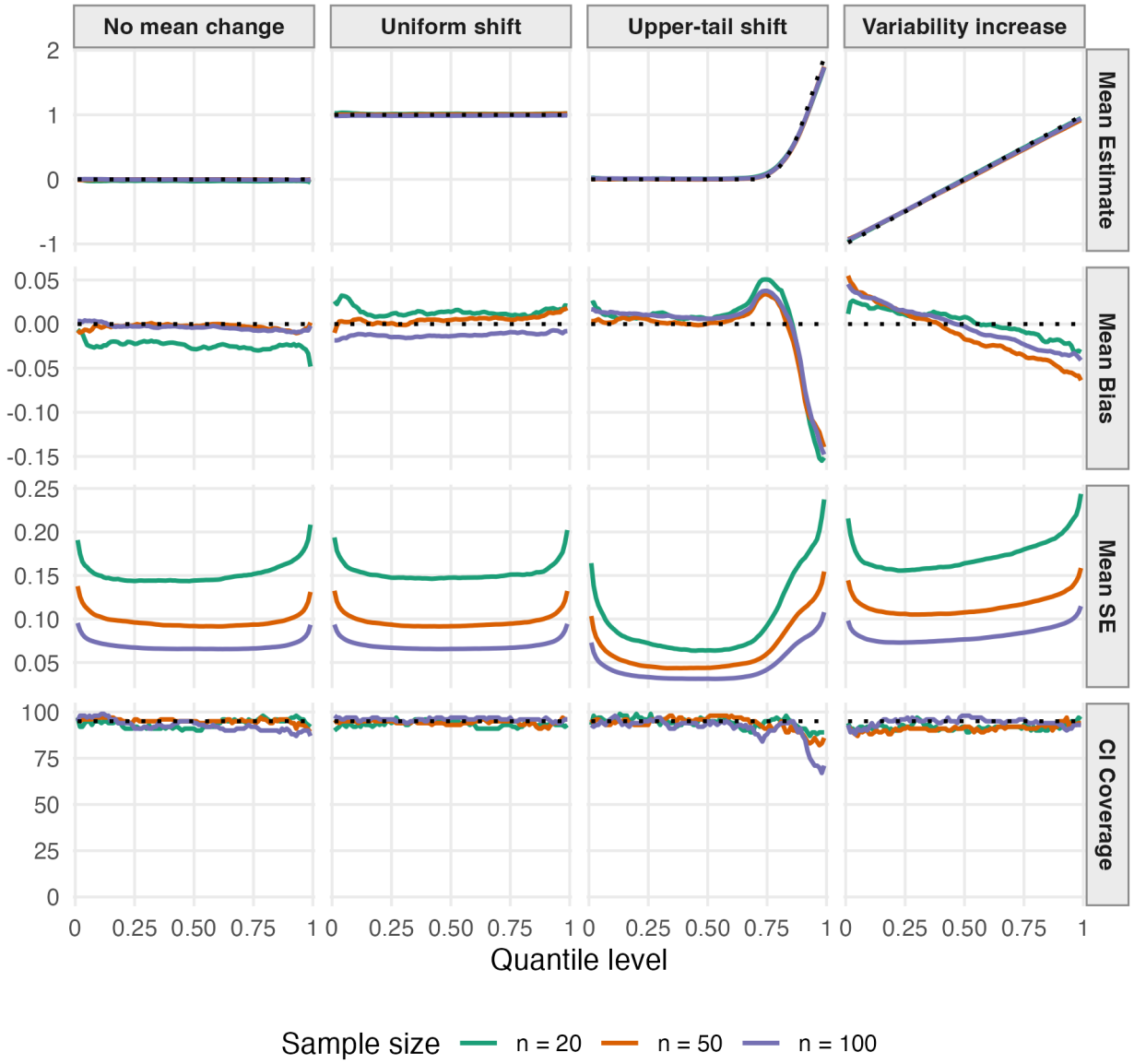


Figure 5.4: Estimation performance for the mean normalized quantile shift map under four data-generating settings in Simulation Study 1. Columns correspond to the true shift pattern: no change, uniform shift, linear variability change, and upper-tail shift. Rows show the estimated mean shift map, pointwise bias, mean standard error, and pointwise confidence interval coverage across quantile levels. Colored curves correspond to sample sizes $n = 20, 50, 100$. Black dotted lines indicate the true mean shift map in the top row, zero bias in the second row, and nominal 95% coverage in the bottom row. Sequencing depths were sampled uniformly from 1 to 300 sequences per individual per timepoint.

sampled independently from the integers 1 through 300.

We consider four normalized quantile shift settings. In the no mean change setting, each individual's normalized quantile shift map is constant across quantile levels, with the constant sampled from Uniform$(-0.1, 0.1)$. The mean normalized quantile shift map is therefore zero across all quantile levels. In the uniform shift setting, each individual's map is constant across quantile levels, with the constant sampled from Uniform$(0.05, 0.35)$. In the upper tail shift setting, the quantile level at which the shift begins is sampled from Uniform$(0.82, 0.92)$, and the map then increases linearly from zero to a value sampled from Uniform$(0.25, 0.70)$ at the upper boundary. In the variability increase setting, each individual's map is linear, with its intercept sampled from Uniform$(-0.5, 0)$ and its slope sampled from Uniform$(0.25, 0.75)$.

For each simulated dataset, we estimate the arm specific mean normalized quantile shift maps and apply the global supremum test described above over the quantile grid $\{0.01, \ldots, 0.99\}$. The null distribution is approximated using 1,000 Gaussian process draws, and the null hypothesis is rejected when the Monte Carlo p-value is less than 0.05.

We consider ten pairwise comparisons among the four settings. These comparisons are shown visually Supplemental Figure E.2. Four comparisons evaluate type I error by assigning the same setting to both arms:

1. No mean change versus no mean change,
2. Uniform shift versus uniform shift,
3. Upper tail shift versus upper tail shift,
4. Variability increase versus variability increase.

The remaining six comparisons evaluate power by assigning different settings to the two arms:

1. No mean change versus uniform shift,
2. No mean change versus upper tail shift,
3. No mean change versus variability increase,

4. Uniform shift versus upper tail shift,
5. Uniform shift versus variability increase,
6. Upper tail shift versus variability increase.

For each comparison and group size, we generate 500 simulated datasets and estimate the rejection probability as the proportion of datasets for which the null hypothesis is rejected.

Results

Results from Simulation Study 2 are shown in Figure 5.5. Under the four null scenarios, in which the two groups were generated from the same normalized quantile shift setting, empirical rejection rates were generally close to the nominal 5% level and remained approximately stable across group sizes. Overall, these results suggest that the test adequately controlled type I error under the data-generating settings considered. Under the alternative scenarios, power increased as the group size increased from $n = 20$ to $n = 100$ per arm. Power was highest for the comparison between the uniform-shift and variability-increase settings and lowest for the comparison between the no-mean-change and upper-tail-shift settings.

5.4 Data application

We apply our framework to the AMP trials, pooling both HVTN 703 and HVTN 704 studies. In total, 146 participants were included in the analysis: 55 in the placebo arm, 52 in the low-dose VRC01 arm, and 39 in the high-dose VRC01 arm (Table 5.2). Sampling times were separated by a median of 15 days, with a range of 5 to 35 days. We compare each VRC01 arm with the placebo arm. As a benchmark, we also considered a mean-based analysis in which we first computed the mean sequence-level feature for each individual at each timepoint, calculated the slope of change between the two timepoints, and then used an unequal-variance t-test to compare mean slopes between treatment arms.

A median of 141.5 viral sequences per individual were analyzed at diagnosis, with a range of 1 to 676 sequences, and a median of 148.5 sequences were analyzed post-diagnosis, with

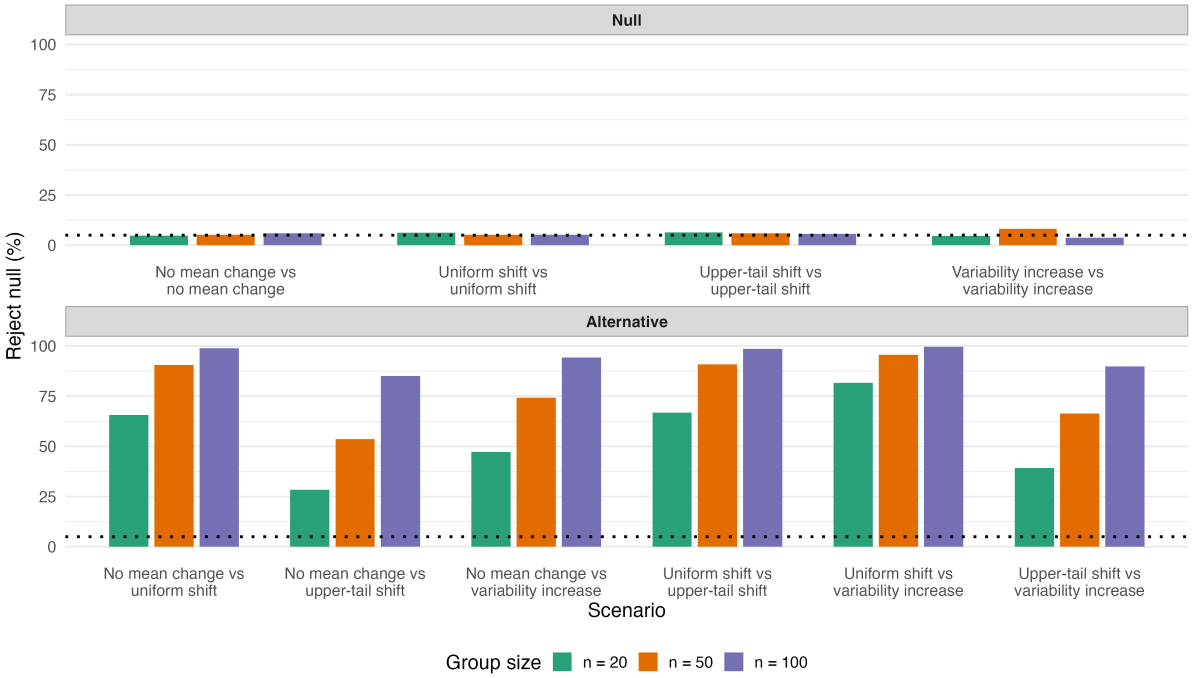


Figure 5.5: Empirical rejection rates for the global test comparing mean normalized quantile shift maps between two arms in Simulation Study 2. The top panel shows same-pattern comparisons, which evaluate rejection under the null hypothesis of equal mean shift maps. The bottom panel shows different-pattern comparisons, which evaluate power under alternatives where the two arms have different mean shift maps. Bars correspond to group sizes $n = 20, 50, 100$ per arm. The black dotted horizontal line indicates the nominal 5% significance level. Sequencing depths were sampled uniformly from 1 to 300 sequences per individual per timepoint.

a range of 2 to 1567 sequences. Results from the distributional analysis are shown in Figure 5.6. In the placebo arm, the estimated mean normalized quantile shift map was close to zero across most quantile levels, suggesting little systematic change over time in the distribution of VRC01 epitope distances. The low-dose VRC01 arm showed a similar overall pattern, although the estimated mean normalized quantile shift map was positive in the upper tail, suggesting that, on average across participants, the upper quantiles of the epitope-distance distribution increased between the two timepoints. However, the confidence intervals in this region overlapped zero, and the mean quantile shift map in the low-dose arm was not significantly different from that in the placebo arm ($p = 0.9$).

In contrast, the estimated mean normalized quantile shift map for the high-dose VRC01 arm was positive across most quantile levels in the high-dose arm, indicating that the distribution of epitope distances on average increased between diagnosis and follow-up. At the 0.01 quantile, the estimated mean normalized shift was 0.00080 (95% CI: -0.0021, 0.0037); at the median, the estimated mean normalized shift was 0.0077 (95% CI: -0.00059, 0.016); and at the 0.99 quantile, the estimated mean normalized shift was 0.012 (95% CI: 0.0061, 0.017). Overall, the mean normalized quantile shift map in the high-dose arm was significantly different from that in the placebo arm ($p < 0.001$). These results suggest that high-dose VRC01 may have exerted post-acquisition selective pressure on the viral population, shifting the within-host quasispecies toward larger VRC01 epitope distances and, therefore, toward increased VRC01 resistance.

Dosage arm	Count	Days between samples, median (range)	Sequencing depth at diagnosis, median (range)	Sequencing depth post-diagnosis, median (range)
Placebo	55	14 (5–30)	159 (1–475)	150 (10–930)
Low-dose	52	18 (5–35)	142 (1–676)	164 (2–1567)
High-dose	39	17 (5–30)	131 (3–593)	133 (9–841)

Table 5.2: Sample timing and sequencing depth by dosage arm. Sequencing depth is summarized separately for the diagnosis and post-diagnosis samples.

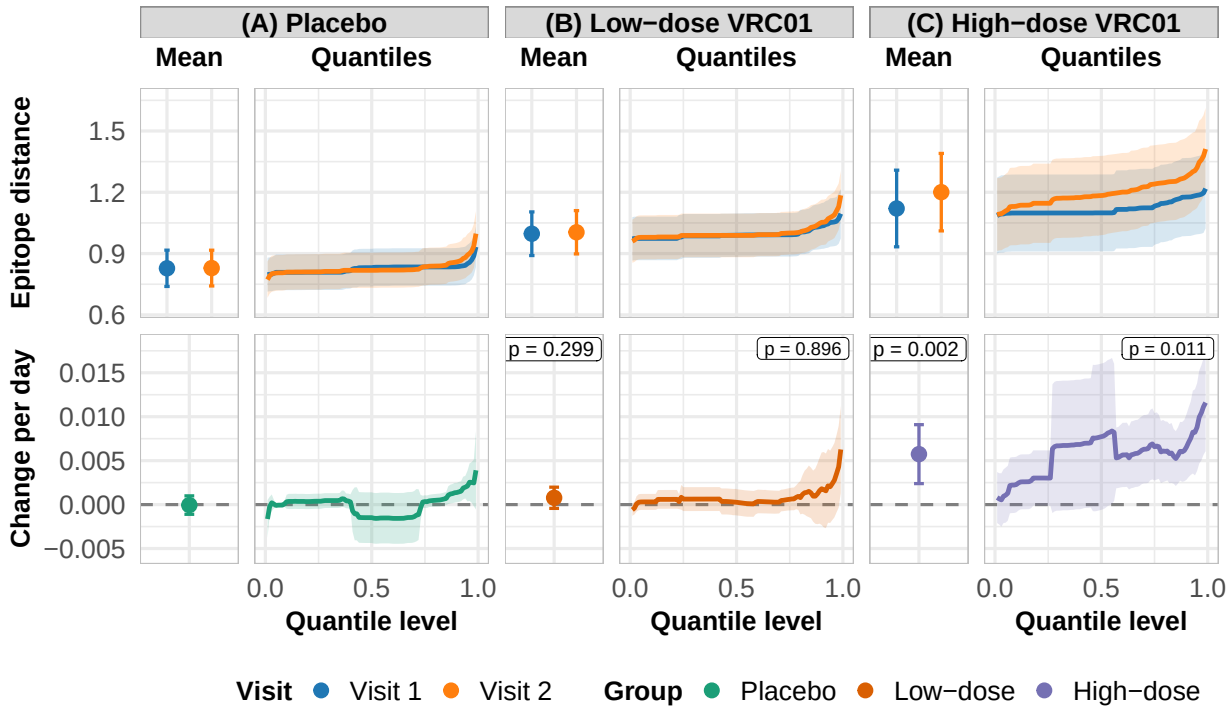


Figure 5.6: Top panels show participant-level means and Wasserstein barycenter quantiles at diagnosis and post-diagnosis; p-values compare each VRC01 arm to placebo at diagnosis only. Bottom panels show the mean per-day change in participant-level mean values and distributional quantiles from diagnosis to post-diagnosis, with p-values comparing each VRC01 arm with placebo.

5.5 Discussion

In this chapter, we propose a framework for comparing changes in within-host viral distributions observed at two timepoints. For each individual, distributional change is summarized by a normalized quantile shift map, which describes the average per-unit-time change at each quantile of the feature distribution. These individual-level maps are treated as functional data and averaged within treatment arms to estimate arm-specific mean functions of distributional change. Comparing these mean maps allows us to assess whether the viral populations evolve differently across treatment groups and to identify the quantile levels at which those differences occur. The normalized quantile shift map also has a geometric interpretation as the velocity of the constant-speed Wasserstein geodesic connecting an individual's two within-host distributions.

One limitation of this approach is that normalizing by time does not completely address differences in when samples were collected relative to infection, particularly if the trajectory of evolution is not constant over time. However, frequent HIV testing and the use of the same sampling protocol across arms should help minimize this concern; we did not observe systematic differences in sampling intervals across arms.

Estimation relies on empirical quantile functions constructed from the observed viral sequences. This plug-in approach is most reliable when the empirical distributions closely approximate the underlying within-host distributions. In simulations using sequencing depths similar to those observed in the AMP trials, the estimator generally performed well across most quantile levels. However, performance worsened slightly near the tails, where quantile estimation is inherently more difficult. Future work should explicitly account for sparse sequencing, for example through the marginal-constructed barycenter estimator described in Chapter 3.

An additional limitation is that the analysis is restricted to participants who acquired HIV-1 after randomization and had sequence data available at both timepoints. Because HIV-1 acquisition and inclusion in the post-infection analysis are post-randomization events

that may be affected by treatment assignment, conditioning on this subset can introduce selection bias (Hudgens and Halloran, 2006). The observed differences between treatment arms should therefore be interpreted as descriptive associations among the included participants rather than as causal effects of VRC01 on post-acquisition viral evolution.

Finally, the proposed method is limited to two timepoints and therefore summarizes distributional change as an average shift over the observed interval. With additional timepoints, the framework could be extended to estimate and compare mean distributional trajectories across treatment arms by using Wasserstein regression methods (Chen et al., 2023).

Chapter 6
CONCLUSION AND FUTURE WORK

In this dissertation, we developed statistical tools for conducting sieve analyses with deep sequencing data, with a particular emphasis on measurement error due to potentially low sequencing depth. In Chapter 2, we studied binary marks using a competing risks Cox model. We classified viral quasispecies according to whether the true proportion of sequences with a feature exceeded a prespecified threshold. Because this classification is unknown and estimated with different levels of precision depending on sequencing depth, we used an empirical Bayes approach to estimate the probability that each quasispecies belonged to each class. In Chapter 3, we developed a method for estimating Wasserstein barycenters under sparse sampling, which was then used in Chapter 4. Although this method's development was motivated by our application, we believe that the question addressed in this Chapter—how one can estimate average tail behavior of a set of distributions with potentially few tail observations—is of broader interest. Chapter 4 provided a practical framework for summarizing and comparing continuous feature distributions between groups, along with an R package implementing the proposed methods. Finally, Chapter 5 provided a proof-of-concept framework for characterizing within-host viral evolution over time and comparing patterns of change between groups.

Our main lesson from these projects is that statistical tools for deep sequencing data should target estimands which relate to intra-individual diversity, since this is the main information gained by observing multiple sequences. Representing this diversity, summarizing it across individuals, and comparing these summaries between groups can reveal effects that would be missed if each sequence set were reduced to a single sequence. At the same time, analyses of within-host diversity must consider sequencing depth, particularly when minority

variants or distributional tails are of interest. In the studies with deep sequencing considered here, low sequencing depth was often less of a practical problem than expected. However, adequate sequencing depth is not guaranteed, and differential sequencing depth between groups remains a potential concern.

There are several limitations of our work. The method in Chapter 2 relies on assumptions for the classification model in addition to the standard assumptions of the Cox model. Violation of these assumptions may lead to biased estimation and incorrect inference. The methods in Chapters 4 and 5 compare distributions among participants who acquired HIV after randomization. Because infection is a post-randomization event, these comparisons are subject to selection bias and cannot generally be interpreted causally without additional assumptions (Hudgens and Halloran, 2006; Shepherd et al., 2006). The resolution of the estimands in each of the chapters also required manual choices, including the threshold for binary marks in Chapter 2 and the quantile grid for continuous marks in Chapter 4 and 5. Although we provided some practical guidance for these choices, a more principled, data-adaptive approach would be preferred.

There are several potential areas for future work. One is the development of meaningful vaccine efficacy estimands for continuous marks. With single-sequence data, previous work has described how vaccine efficacy varies with the value of a continuous mark (Sun and Gilbert, 2012; Juraska and Gilbert, 2016). With multi-sequence data, each individual instead has an associated mark distribution. A direct extension would require describing vaccine efficacy against all possible mark distributions, which is likely too complex for the amount of data available. Principled ways to reduce or summarize these distributions while retaining scientifically relevant information are an important direction for future work.

Another area for future work is the development of causal estimands for sieve analyses with deep sequencing data. Indeed, none of the estimands considered in this dissertation are causally defined. Hazard-based estimands, such as the one considered in Chapter 2, are particularly difficult to interpret causally. One possible direction would be to extend the cumulative-incidence-based causal vaccine efficacy estimands proposed by Benkeser et al.

(2020) and Yang et al. (2022) to accommodate distribution-valued marks. For the case-only analyses in Chapters 4 and 5, causal effects could be defined using principal stratification, for example by conditioning on the always-infected principal stratum (Frangakis and Rubin, 2002; Gilbert et al., 2003). Such an approach would require extending existing principal stratification methods for scalar post-infection outcomes to functional or distribution-valued outcomes.

BIBLIOGRAPHY

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis* **43,** 904–924.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10,** 1100–1120.

Andino, R. and Domingo, E. (2015). Viral quasispecies. *Virology* **479-480,** 46–51.

Augustin, N. H., Mattocks, C., Faraway, J. J., Greven, S., and Ness, A. R. (2017). Modelling a response as a function of high-frequency count data: the association between physical activity and fat mass. *Statistical Methods in Medical Research* **26,** 2210–2226.

Ayieko, J., Petersen, M. L., Kamya, M. R., and Havlir, D. V. (2022). PEP for HIV prevention: are we missing opportunities to reduce new infections? *Journal of the International AIDS Society* **25,** e25942.

Bekker, L.-G., Pike, C., and Hillier, S. L. (2022). HIV prevention: better choice for better coverage. *Journal of the International AIDS Society* **25,** e25872.

Benkeser, D., Gilbert, P., and Carone, M. (2019). Estimating and testing vaccine sieve effects using machine learning. *Journal of the American Statistical Association* **114,** 1038–1049.

Benkeser, D., Juraska, M., and Gilbert, P. B. (2020). Assessing trends in vaccine efficacy by pathogen genetic distance. *Journal De La Societe Francaise De Statistique (2009)* **161,** 164–175.

Bigot, J. (2020). Statistical data analysis in the Wasserstein space. *ESAIM: Proceedings and Surveys* **68,** 1–19.

Bigot, J., Gouet, R., Klein, T., and López, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Annales De L'institut Henri Poincaré, Probabilités Et Statistiques* **53,** 1–26.

Bigot, J., Gouet, R., Klein, T., and López, A. (2018). Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electronic Journal of Statistics* **12,**.

Bigot, J. and Klein, T. (2018). Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics* **22,** 35–57.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons. Google-Books-ID: 6ItqtwaWZZQC.

Bobkov, S. and Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society.

Buchbinder, S., Mehrotra, D. V., Duerr, A., Fitzgerald, D. W., Mogg, R., Li, D., et al. (2008). Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the STEP study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *The Lancet* **372,** 1881–1893.

Chen, Y., Lin, Z., and Müller, H.-G. (2023). Wasserstein regression. *Journal of the American Statistical Association* **118,** 869–882.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* **41,** 2786–2819.

Corey, L., Gilbert, P. B., Juraska, M., Montefiori, D. C., Morris, L., Karuna, S. T., et al. (2021). Two randomized trials of neutralizing antibodies to prevent HIV-1 acquisition. *New England Journal of Medicine* **384,** 1003–1014.

Crauel, H. (2002). *Random probability measures on Polish spaces.* CRC Press, 0 edition.

DeCamp, A. (2013). Assessing vaccine effects in HIV-1 vaccine trials: antigenic maps, antigen selection, and sieve analysis. *University of Washington Libraries* OCLC: 882504213.

deCamp, A. C., Rolland, M., Edlefsen, P. T., Sanders-Buell, E., Hall, B., Magaret, C. A., et al. (2017). Sieve analysis of breakthrough HIV-1 sequences in HVTN 505 identifies vaccine pressure targeting the CD4 binding site of Env-gp120. *PLOS ONE* **12,** e0185959.

Deeks, S. G., Lewin, S. R., and Havlir, D. V. (2013). The end of AIDS: HIV infection as a chronic disease. *The Lancet* **382,** 1525–1533.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics* **2,** 267–277.

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews* **76,** 159–216.

Dubey, P. and Müller, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106,** 803–821.

Edlefsen, P. T., Rolland, M., Hertz, T., Tovanabutra, S., Gartland, A. J., deCamp, A. C., et al. (2015). Comprehensive sieve analysis of breakthrough HIV-1 sequences in the RV144 vaccine efficacy trial. *PLOS Computational Biology* **11,** e1003973.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer, New York, NY.

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103,** 1–20.

Elston, D., Moss, R., Boulinier, T., Arrowsmith, C., and Lambin, X. (2001). Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology* **122,** 563–569.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* pages 209–230.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75,** 259–276.

Follmann, D. and Huang, C.-Y. (2018). Sieve analysis using the number of infecting pathogens. *Biometrics* **74,** 1023–1033.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Gao, G. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika* **92,** 875–891.

Garner, C. (2011). Confounded by sequencing depth in association studies of rare alleles. *Genetic Epidemiology* **35,** 261–268.

Gelman, A. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Ghosal, R., Varma, V. R., Volfson, D., Hillel, I., Urbanek, J., Hausdorff, J. M., et al. (2023). Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's disease. *Biostatistics (Oxford, England)* **24,** 539–561.

Gilbert, P. (2000). Comparison of competing risks failure time methods and time-independent methods for assessing strain variations in vaccine protection. *Statistics in Medicine* **19,** 3065–3086.

Gilbert, P., Self, S., Rao, M., Naficy, A., and Clemens, J. (2001). Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. *Journal of Clinical Epidemiology* **54,** 68–85.

Gilbert, P., Self, S. G., and Ashby, M. A. (1998). Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* **54,** 799–814.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59,** 531–541.

Gilbert, P. B. and Sun, Y. (2015). Inferences on relative failure rates in stratified mark-specific proportional hazards models with missing marks, with application to HIV vaccine efficacy trials. *Journal of the Royal Statistical Society. Series C, Applied Statistics* **64,** 49–73.

Gray, G., Allen, M., Moodie, Z., Churchyard, G., Bekker, L.-G., Nchabeleng, M., et al. (2011). Safety and efficacy of the HVTN 503/Phambili study of a clade-b-based HIV-1 vaccine in South Africa: a double-blind, randomised, placebo-controlled test-of-concept phase 2b study. *The Lancet Infectious Diseases* **11,** 507–515.

Gray, G., Buchbinder, S., and Duerr, A. (2010). Overview of STEP and Phambili trial results: two phase IIb test-of-concept studies investigating the efficacy of MRK adenovirus type 5 gag/pol/nef subtype b HIV vaccine. *Current Opinion in HIV and AIDS* **5,** 357–361.

Gray, G. E., Bekker, L.-G., Laher, F., et al. (2021). Vaccine efficacy of ALVAC-HIV and bivalent subtype c gp120–MF59 in adults. *New England Journal of Medicine* **384,** 1089–1100.

Gray, G. E., Mngadi, K., Lavreys, L., et al. (2024). Mosaic HIV-1 vaccine regimen in southern African women (Imbokodo/HVTN 705/HPX2008): a randomised, double-blind, placebo-controlled, phase 2b trial. *The Lancet Infectious Diseases* **24,** 1201–1212.

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J. I., Quer, J., and Domingo, E. (2016). Viral quasispecies complexity measures. *Virology* **493,** 227–237.

Hein, M. (2009). Robust nonparametric regression with metric-space valued output. *Advances in Neural Information Processing Systems* **22,**.

Heng, F., Sun, Y., Li, L., and B. Gilbert, P. (2025). Estimation and hypothesis testing of strain-specific vaccine efficacy with missing strain types with application to a COVID-19 vaccine trial. *Statistics in Medicine* **44,** e10345.

Hertz, T., Logan, M. G., Rolland, M., Magaret, C. A., Rademeyer, C., Fiore-Gartland, A., et al. (2016). A study of vaccine-induced immune pressure on breakthrough infections in the Phambili phase 2b HIV-1 vaccine efficacy trial. *Vaccine* **34,** 5792–5801.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200 of *Springer Series in Statistics*. Springer, New York, NY.

Hudgens, M. G. and Halloran, M. E. (2006). Causal vaccine effects on binary postinfection outcomes. *Journal of the American Statistical Association* **101,** 51–64.

Jeong, S., Yoon, H., Morris, J. S., and Yang, H. (2026). Wasserstein tests for equality of several groups for distributional data. *Computational Statistics & Data Analysis* **222,** 108385.

Joseph, E., Galeano San Miguel, P., and Lillo Rodríguez, R. E. (2015). Two-sample Hotelling's T^2 statistics based on the functional Mahalanobis semi-distance. *DES - Working Papers. Statistics and Econometrics. WS* Number: ws1503.

Joseph, S. B., Swanstrom, R., Kashuba, A. D. M., and Cohen, M. S. (2015). Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. *Nature Reviews Microbiology* **13,** 414–425.

Juraska, M., Bai, H., deCamp, A. C., Magaret, C. A., Li, L., Gillespie, K., et al. (2024). Prevention efficacy of the broadly neutralizing antibody VRC01 depends on HIV-1 envelope sequence features. *Proceedings of the National Academy of Sciences* **121,** e2308942121.

Juraska, M. and Gilbert, P. B. (2016). Mark-specific hazard ratio model with missing multivariate marks. *Lifetime Data Analysis* **22,** 606–625.

Juraska, M., Magaret, C. A., Shao, J., Carpp, L. N., Fiore-Gartland, A. J., Benkeser, D., et al. (2018). Viral genetic diversity and protective efficacy of a tetravalent dengue vaccine in two phase 3 trials. *Proceedings of the National Academy of Sciences of the United States of America* **115,** E8378–E8387.

Katta, S., Parikh, H., Rudin, C., and Volfovsky, A. (2024). Interpretable causal inference for analyzing wearable, sensor, and distributional data. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR.

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 7552–7557.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* **27,** 887–906.

Kijak, G. H., Simon, V., Balfe, P., Vanderhoeven, J., Pampuro, S. E., Zala, C., et al. (2002). Origin of human immunodeficiency virus type 1 quasispecies emerging after antiretroviral treatment interruption in patients with therapeutic failure. *Journal of Virology* **76,** 7000–7009.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46,** 33–50.

Koenker, R. and Gu, J. (2017). REBayes: an r package for empirical Bayes mixture methods. *Journal of Statistical Software* **82,** 1–26.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin* **58,** 19–42.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

Lee, J., Baćak, V., and Kennedy, E. H. (2025). Learning smooth populations of parameters with trial heterogeneity. arXiv:2507.23140 [math] version: 1.

Lin, Z., Kong, D., and Wang, L. (2023). Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85,** 378–398.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics* **5,** i–163.

Lunn, M. and McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics* **51,** 524–532.

Magaret, C. A., Li, L., deCamp, A. C., Rolland, M., Juraska, M., Williamson, B. D., et al. (2024). Quantifying how single dose Ad26.COV2.s vaccine efficacy depends on spike sequence features. *Nature Communications* **15,** 2175.

Metzner, K. J., Giulieri, S. G., Knoepfel, S. A., Rauch, P., Burgisser, P., Yerly, S., et al. (2009). Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clinical Infectious Diseases: an Official Publication of the Infectious Diseases Society of America* **48,** 239–247.

Molenberghs, G. and Verbeke, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Mullins, J. I., Deng, W., Giorgi, E. E., Magaret, C. A., Rolland, M., Bhattacharya, T., et al. (2025). Long-read deep sequencing reveals high rates of multilineage transmission and rapid viral population changes in acute HIV infection. *Biorxiv* page 2025.10.15.682663.

Narasimhan, B. and Efron, B. (2020). deconvolveR: a g-modeling program for deconvolution and empirical Bayes estimation. *Journal of Statistical Software* **94,** 1–20.

Neafsey, D. E., Juraska, M., Bedford, T., et al. (2015). Genetic diversity and protective efficacy of the RTS,s/AS01 malaria vaccine. *New England Journal of Medicine* **373,** 2025–2037.

Nevalainen, J., Oja, H., and Datta, S. (2017). Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in Medicine* **36,** 2630–2640.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application* **6,** 405–431.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. Springer International Publishing, Cham.

Pepe, M. S., Reilly, M., and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference* **42,** 137–160.

Petersen, A., Liu, X., and Divani, A. A. (2021). Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics* **49,** 590–611.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics* **47,** 691.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34,** 541–554.

Price, P. N., Nero, A. V., and Gelman, A. (1996). Bayesian prediction of mean indoor Radon concentrations for Minnesota counties. *Health Physics* **71,** 922–936.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*, volume 1. SAGE.

Raymond, S., Jeanne, N., Vellas, C., Nicot, F., Saune, K., Ranger, N., et al. (2024). HIV-1 genotypic resistance testing using single molecule real-time sequencing. *Journal of Clinical Virology* **174,** 105717.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3.1, pages 157–164. University of California Press.

Rolland, M., Edlefsen, P. T., Larsen, B. B., Tovanabutra, S., Sanders-Buell, E., Hertz, T., et al. (2012). Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature* **490,** 417–420.

Rolland, M., Tovanabutra, S., deCamp, A. C., Frahm, N., Gilbert, P. B., Sanders-Buell, E., et al. (2011). Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nature Medicine* **17,** 366–371.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Sagrillo, M., Guerra, R. R., and Bayer, F. M. (2021). Modified Kumaraswamy distributions for double bounded hydro-environmental data. *Journal of Hydrology* **603,** 127021.

Schmon, S. M., Deligiannidis, G., Doucet, A., and Pitt, M. K. (2021). Large-sample asymptotics of the pseudo-marginal method. *Biometrika* **108,** 37–51.

Shaw, G. M. and Hunter, E. (2012). HIV-1 transmission. *Cold Spring Harbor Perspectives in Medicine* **2,** a006965.

Shepherd, B. E., Gilbert, P. B., Jemiai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62,** 332–342.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability* .

Singh, K., Mehta, D., Dumka, S., Chauhan, A. S., and Kumar, S. (2023). Quasispecies nature of RNA viruses: lessons from the past. *Vaccines* **11,** 308.

Sun, Y. and Gilbert, P. B. (2012). Estimation of stratified mark-specific proportional hazards models with missing marks. *Scandinavian Journal of Statistics, Theory and Applications* **39,** 34–52.

Sun, Y., Hyun, S., and Gilbert, P. (2008). Testing and estimation of time-varying cause-specific hazard ratios with covariate adjustment. *Biometrics* **64,** 1070–1079.

Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics* **46,** 515–522.

Tebas, P. (2025). Future of bNAbs in HIV treatment. *Current HIV/AIDS Reports* **22,** 34.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics* pages 1265–1269.

Tian, K., Kong, W., and Valiant, G. (2017). Learning populations of parameters. arXiv.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

Van Rompaye, B., Jaffar, S., and Goetghebeur, E. (2012). Estimation with Cox models: cause-specific survival analysis with misclassified cause of failure. *Epidemiology* **23,** 194–202.

Villani, C. (2009). *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg.

Vinayak, R. K., Kong, W., Valiant, G., and Kakade, S. M. (2019). Maximum likelihood estimation for learning populations of parameters. arXiv.

Walters, C. R. (2024). Empirical Bayes methods in labor economics.

Westfall, D. H., Deng, W., Pankow, A., Murrell, H., Chen, L., Zhao, H., et al. (2024). Optimized SMRT-UMI protocol produces highly accurate sequence datasets from diverse populations — application to HIV-1 quasispecies. *Virus Evolution* **10,** veae019.

Westling, T., van der Laan, M. J., and Carone, M. (2020). Correcting an estimator of a multivariate monotone function with isotonic regression. *Electronic Journal of Statistics* **14,** 3032.

Whittemore, A. S. (1989). Errors-in-variables regression using Stein estimates. *The American Statistician* **43,** 226–228.

Williamson, C., Moodley, C., Magaret, C. A., Giorgi, E. E., Rolland, M., Westfall, D. H., et al. (2026). Influence of the broadly neutralizing antibody VRC01 on HIV breakthrough virus populations in antibody-mediated prevention trials. *Nature Communications* **17,** 4780.

Williamson, J. M., Datta, S., and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59,** 36–42.

Willis, J. R., Prabhakaran, M., Muthui, M., et al. (2025). Vaccination with mRNA-encoded nanoparticles drives early maturation of HIV bnAb precursors in humans. *Science* **389,** eadr8382.

Wood, G. R. (1999). Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics* **27,** 1706–1721.

World Health Organization (2025). HIV statistics, globally and by WHO region, 2025.

Yang, G., Balzer, L. B., and Benkeser, D. (2022). Causal inference methods for vaccine sieve analysis with effect modification. *Statistics in Medicine* **41,** 1513–1524.

Ye, Y. and Bickel, P. J. (2021). Binomial mixture model with u-shape constraint. arXiv.

Zhang, M. and Guo, F. R. (2022). BSDE: barycenter single-cell differential expression for case–control studies. *Bioinformatics* **38,** 2765–2772.

Zhou, S., Hill, C. S., Spielvogel, E., Clark, M. U., Hudgens, M. G., and Swanstrom, R. (2022). Unique molecular identifiers and multiplexing amplicons maximize the utility of deep sequencing to critically assess population diversity in RNA viruses. *ACS Infectious Diseases* **8,** 2505–2514.

Zhou, Y. and Müller, H.-G. (2024). Wasserstein regression with empirical measures and density estimation for sparse data. *Biometrics* **80,** ujae127.

Appendix A

R PACKAGE FOR WASSERSTEIN BARYCENTER ANALYSIS

In this vignette, we describe the `mcbarycenter` R package, which can be used to implement the three tasks presented in the main body of this paper. The R package is available here.

```
devtools::install_github("jpspeng/mcbarycenter")
```

We discuss usage of the package in a general two-sample distributional setting. Specifically, suppose we have n independent units, indexed by $i = 1, \ldots, n$, each with an associated underlying distribution Y_i. We do not observe each Y_i directly. Instead, for each unit i, we observe m_i independent samples from Y_i:

$$Y_{ij} \sim Y_i, \qquad j = 1, \ldots, m_i.$$

Additionally, each unit i may have an associated group label $Z_i \in \{1, 2\}$, and we may want to compare units across groups.

In this vignette, we use the sparse dataset example `sparse_dist_data` data frame available in the package. We also use the `dplyr` package for data manipulation.

```
library(mcbarycenter)
library(dplyr)
```

A.0.1 Preliminaries

The dataset should include two columns: an ID column, which contains the individual ID, and a value column, which contains samples from the individual's underlying distribution.

In many functions, we specify the names of these columns using the arguments `id_col` and `val_col`.

The sample data also has a group column, which takes the values 1 or 2 and describes the groups that we wish to compare.

```
head(sparse_dist_data)
```

```
      group id value
1     1  1  0.20
2     1  1  0.20
3     1  1  0.20
4     1  2 -0.35
5     1  2 -0.12
6     1  2  0.08
```

We subset the data into the two groups:

```
sample_data1 <- sparse_dist_data %>% filter(group == 1)
sample_data2 <- sparse_dist_data %>% filter(group == 2)
```

We will also need to specify a grid to do our analysis. The functions default to the grid 0.01, ..., 0.99 but more resolved grid could be of interest if the number of samples per distribution is generally higher.

A.0.2 Exploratory analysis

To begin, we may wish to visualize each individual's empirical quantile function. This allows us to see general patterns in the raw data and identify individuals whose data look different from others.

To visualize the empirical quantile curve for each individual distribution, we can use the `graph_individual_quantiles()` function. This function computes empirical quantiles

separately for each ID and then plots the resulting curves. The code below displays the empirical quantile functions for the first three ids.

```
graph_individual_quantiles(sample_data1 %>% filter(id %in% c(1, 2, 3)),
                           id_col = "id",
                           val_col = "value")
```

When `interactive = TRUE`, which is the default, the function returns an interactive `plotly` widget. The curves are visible by default, and clicking a line highlights the selected individual. Double-clicking resets the plot.

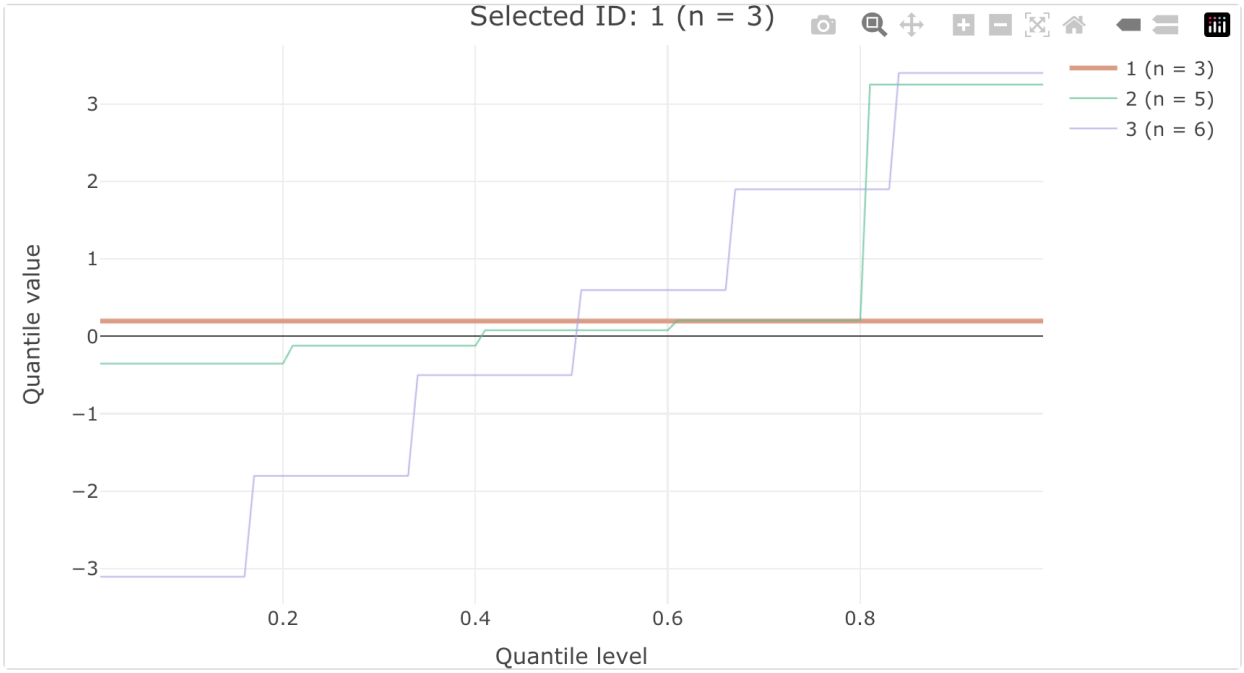


Figure A.1: Output of the `graph_individual_quantiles()` function.

In the figure above, we can see the individual quantile functions for the first three ids. As shown, individual 1 (sample size $m_1 = 3$) has a homogeneous sample, indicated by a flat empirical quantile function. Individual 2 (sample size $m_2 = 5$) has an extended right tail; that is, their empirical distribution has a small proportion of high values. Individual 3 (sample size $m_3 = 6$) has wide variability in their observed empirical distribution.

A.0.3 Task 1: Wasserstein barycenter estimation

We begin with the first analytic task: estimating the Wasserstein barycenter's quantile function, equivalently the mean quantile function, for each group. We denote the Wasserstein barycenter quantile function for Group 1 as $\mu_1(u)$ and that for Group 2 as $\mu_2(u)$, for $u \in (0, 1)$.

Empirical Wasserstein barycenter. To estimate the empirical Wasserstein barycenter's quantile function, we can use the `empbary()` function. This function computes empirical quantiles for each subject and then averages those quantiles pointwise. The main arguments are:

- `df`: the input data frame containing the observed samples.
- `id_col`: the name of the column identifying the individual or unit.
- `val_col`: the name of the numeric column containing observations from each individual's underlying distribution.
- `weight_col`: an optional column containing id-level weights. If `NULL`, the empirical barycenter is unweighted.
- `quantile_grid`: the grid of quantile levels at which the barycenter quantile function is estimated. The default is `seq(0.01, 0.99, 0.01)`.
- `quantile_type`: the interpolation rule passed to `stats::quantile()` through its `type` argument. The default is `1`.

```
emp_res_grp1 <- empbary(
  df = sample_data1,
  id_col = "id",
  val_col = "value",
  quantile_grid = seq(0.01, 0.99, 0.01), # optional, defaulted to 0.01, 0.02, ..., 0.99
  quantile_type = 1 # optional, defaulted to quantile_type = 1
)
```

```
emp_res_grp2 <- empbary(
  df = sample_data2,
  id_col = "id",
  val_col = "value"
)
```

This function call returns an `empbary_result` object containing the estimated quantile curve in `res`. The `res` object contains estimates, standard error estimates, and 95% pointwise confidence intervals for each quantile level. The returned object also includes an estimated covariance matrix in `cov` and the standardized input data in `data`.

```
head(emp_res_grp1$res)
```

```
  quantile  estimate         se     ci_lo     ci_hi
1     0.01 -1.233814 0.05900223 -1.349456 -1.118172
2     0.02 -1.233814 0.05900223 -1.349456 -1.118172
3     0.03 -1.233814 0.05900223 -1.349456 -1.118172
4     0.04 -1.233814 0.05900223 -1.349456 -1.118172
5     0.05 -1.233814 0.05900223 -1.349456 -1.118172
6     0.06 -1.233814 0.05900223 -1.349456 -1.118172
```

Marginal-constructed barycenter. To estimate the marginal-constructed barycenter estimator, we can use the `mcbary()` function, which implements the procedure discussed in Chapter 3. Briefly, the method transforms the problem of estimating the Wasserstein barycenter's quantile function into the problem of estimating the mixing distributions of a sequence of binomial mixture data.

The main arguments include:

- `df`: the input data frame containing the observed samples.
- `id_col`: the name of the column identifying the individual or unit.

- `val_col`: the name of the numeric column containing observations from each individual's underlying distribution.
- `method`: the mixture estimation method. Options include `"spline"`, `"npmle"`, and `"beta"`.
- `x_grid`: a user-specified threshold grid used to construct the binomial mixture problems.
- `cutpoints`: an optional number of evenly spaced thresholds. Either `x_grid` or `cutpoints` must be supplied.
- `bootstrap_samples`: the number of bootstrap resamples used for uncertainty estimation.
- `quantile_grid`: the grid of quantile levels at which the barycenter's quantile function is estimated.
- `use_midpoint`: whether to use interval midpoints between adjacent thresholds when constructing the estimated distribution of the α-quantile. Defaulted to `FALSE`. Defaulted to 0.01, 0.02, ..., 0.99.
- `estimate_endpoints`: whether to estimate endpoint intervals when constructing the estimated distribution of the α-quantile. Defaulted to `FALSE`.
- `use_isotonic_dist`: whether to enforce monotonicity in the cumulative mixture curves across threshold values. Defaulted to `FALSE`.
- `use_isotonic_output`: if `TRUE`, whether to enforce monotonicity in the final quantile function estimate. Defaulted to `FALSE`.

For the mixing distribution specification `method = "spline"`, we can additionally specify the arguments `pDegree`, which sets the degrees of freedom of the natural spline basis used for estimation, and `c0`, which sets the penalty constant in the spline deconvolution procedure. We may also specify `mass_at_endpoints`, which, when set to `TRUE`, includes point masses at 0 and 1 in the spline specification. For the nonparametric maximum likelihood specification `method = "npmle"`, we can additionally specify the argument `backend`, which can be either

"mixsqp" (the default) or "REBayes"; the latter is typically faster, but requires a working installation of Mosek.

In the example below, we use the Beta mixture specification with 100 cutpoints and 20 bootstrap samples. The number of bootstrap samples is kept small for illustration.

```
mcb_res_grp1 <- mcbary(
  df = sample_data1,
  id_col = "id",
  val_col = "value",
  method = "beta",
  cutpoints = 100,
  use_midpoint = TRUE,
  estimate_endpoints = TRUE,
  bootstrap_samples = 20
)

mcb_res_grp2 <- mcbary(
  df = sample_data2,
  id_col = "id",
  val_col = "value",
  method = "beta",
  cutpoints = 100,
  use_midpoint = TRUE,
  estimate_endpoints = TRUE,
  bootstrap_samples = 20
)
```

The function returns an mcbary_result object. As with empbary(), the estimated quantile curve and uncertainty summaries are stored in res. The returned object also includes a bootstrap covariance matrix in cov, the estimated mixing distributions in mixtures, the chosen estimation method in method, and the standardized input data in data.

```
head(mcb_res_grp1$res)
```

```
  quantile estimate estimate_bs    se  ci_lo  ci_hi pct_ci_lo pct_ci_hi
1     0.01   -2.426      -2.548 0.238 -2.891 -1.960    -2.950    -2.189
2     0.02   -2.119      -2.160 0.194 -2.498 -1.739    -2.559    -1.883
3     0.03   -1.960      -1.995 0.160 -2.273 -1.646    -2.326    -1.783
4     0.04   -1.836      -1.867 0.141 -2.113 -1.560    -2.146    -1.644
5     0.05   -1.734      -1.751 0.114 -1.957 -1.511    -2.013    -1.570
6     0.06   -1.646      -1.654 0.093 -1.828 -1.464    -1.902    -1.505
```

Graphing. To visualize these quantile functions, we can use the `graph_quantiles()` function. This function takes as input `empbary_result` objects, `mcbary_result` objects, or any data frame with the columns `quantile` and `estimate`. If confidence intervals are available, they can also be displayed using the columns `ci_lo` and `ci_hi`. We can include multiple objects in the function call, and the function will graph all of the results on the same plot.

The code below graphs the empirical barycenter and MCB estimator for each group.

```
graph_quantiles(
  emp_grp1 = emp_res_grp1,
  emp_grp2 = emp_res_grp2,
  mcb_grp1 = mcb_res_grp1,
  mcb_grp2 = mcb_res_grp2,
  show_ci = TRUE # optional, defaults as TRUE
)
```

We can also visualize the intermediate estimated mixing distributions using `graph_mixtures()`. This can be useful for checking the intermediate objects used to construct the MCB estimator.

```
graph_mixtures(mcb_res_grp1)
```

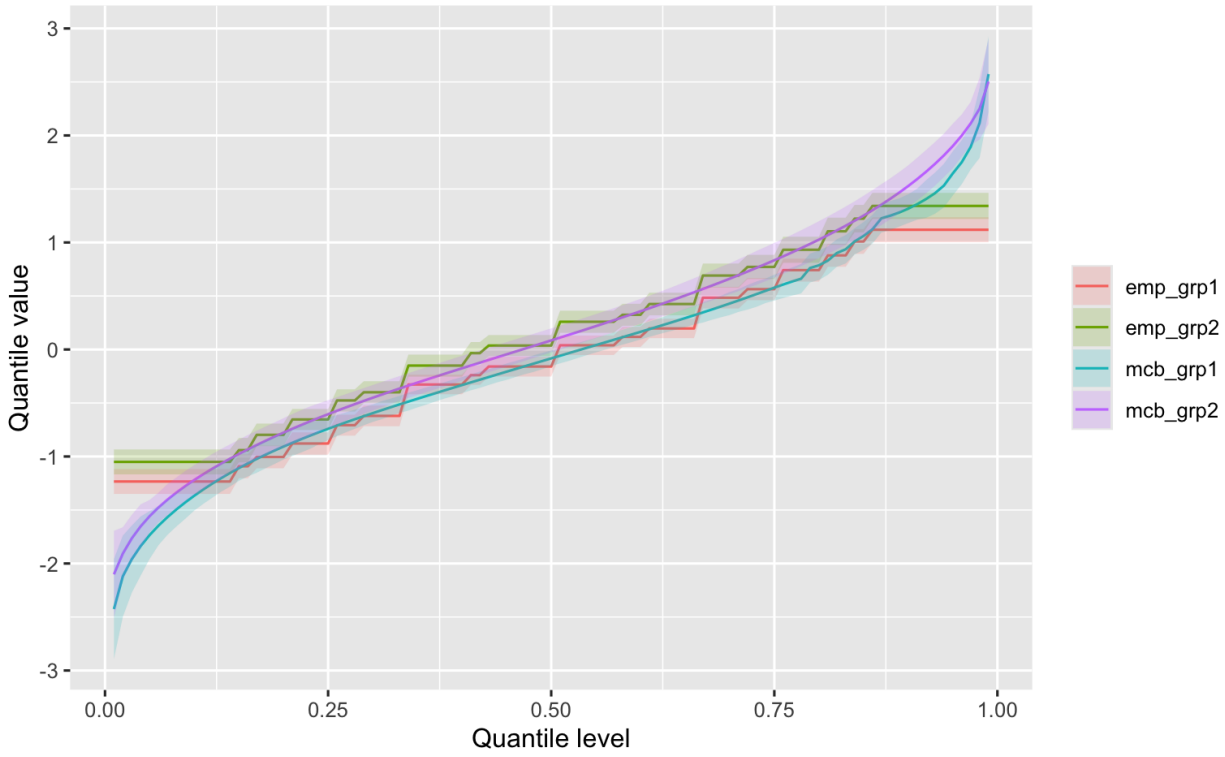


Figure A.2: Empirical and marginal-constructed barycenter estimates for Groups 1 and 2.

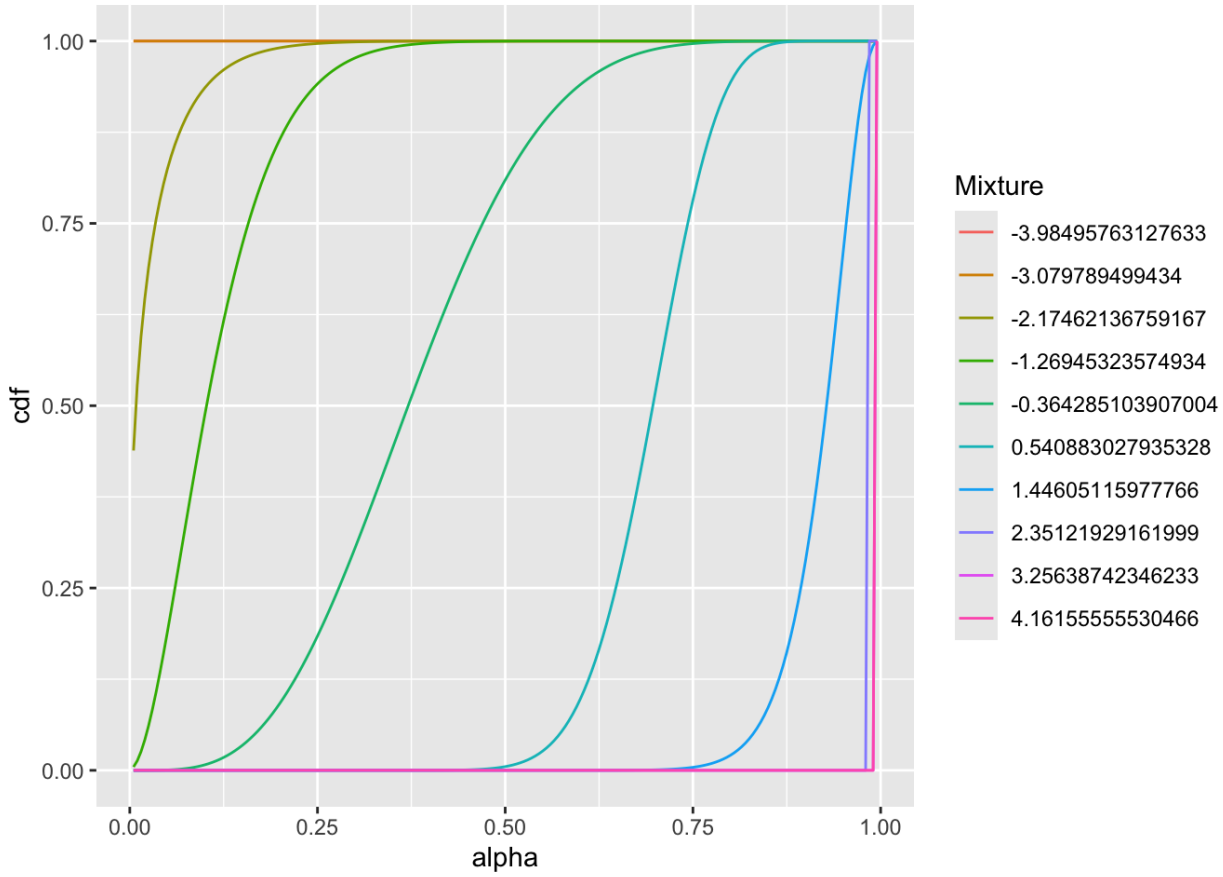


Figure A.3: Estimated mixing distributions from the `mcbary()` procedure for Group 1.

A.0.4 Task 2: Difference in quantiles

The second analytic task is to estimate the difference between the barycenter quantile functions for the two groups. We define the difference in quantiles as

$$\Delta(u) = \mu_1(u) - \mu_2(u), \qquad u \in (0, 1).$$

Our goal is to estimate $\Delta(u)$ over a grid of quantile levels.

To compute this difference, we can use the `quantile_diff()` function. This function takes two estimated barycenter objects and returns the pointwise difference between their estimated quantile functions, along with their estimated standard errors and 95% confidence intervals. In the example below, we compute the difference between the two MCB estimates.

```
quantile_diff_res <- quantile_diff(mcb_res_grp2, mcb_res_grp1)


head(quantile_diff_res)
```

```
  quantile  estimate        se       ci_lo     ci_hi
1     0.01 0.3247018 0.3151857 -0.29305083 0.9424544
2     0.02 0.2118083 0.2306834 -0.24032298 0.6639395
3     0.03 0.1946393 0.1952820 -0.18810627 0.5773850
4     0.04 0.1840011 0.1755934 -0.16015565 0.5281578
5     0.05 0.1762177 0.1376481 -0.09356763 0.4460030
6     0.06 0.1695939 0.1168332 -0.05939498 0.3985828
```

We can graph the estimated difference curve using `graph_quantiles()`.

```
graph_quantiles(mcb_difference = quantile_diff_res)
```

This plot shows how the estimated group difference varies across the distribution. Positive values indicate quantile levels at which the estimated Group 2 barycenter is larger than the estimated Group 1 barycenter, while negative values indicate quantile levels at which the estimated Group 1 barycenter is larger.

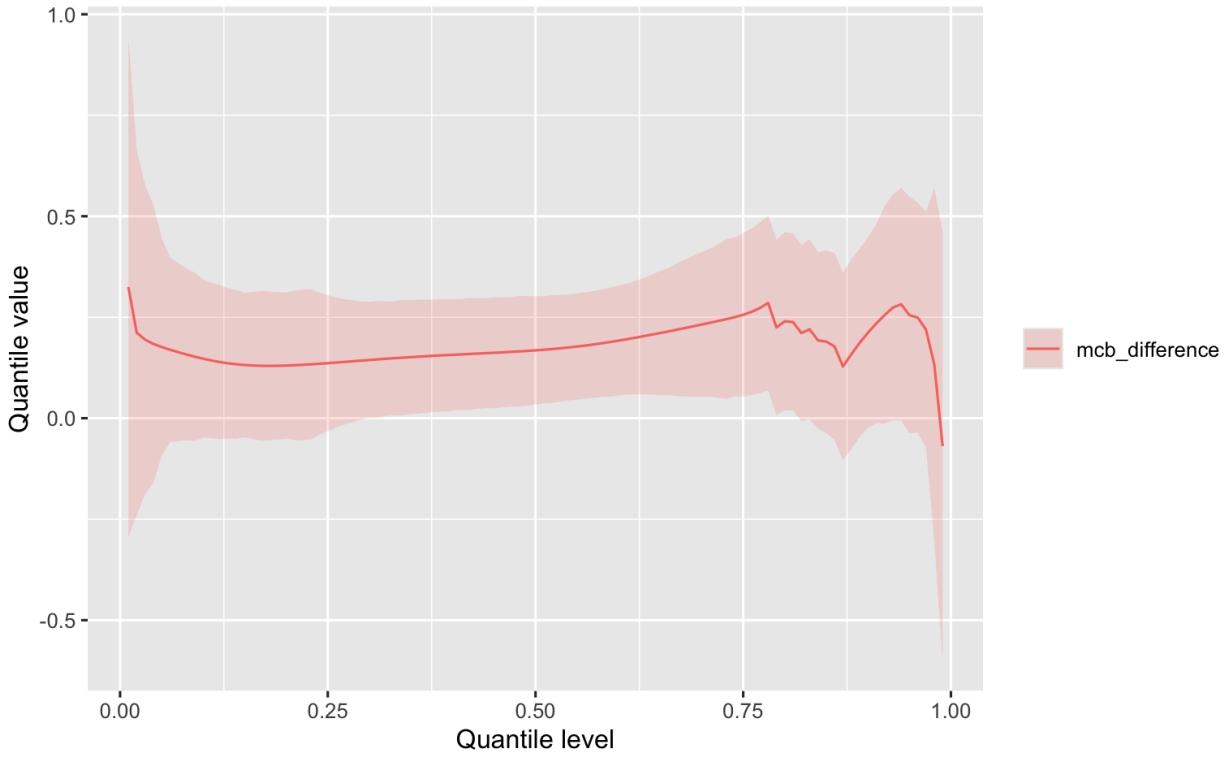


Figure A.4: Estimated difference between the MCB barycenter quantile functions for Group 2 versus Group 1.

A.0.5 Task 3: Hypothesis testing

The third analytic task is to test whether the barycenter quantile functions differ between groups. In this setting, the null hypothesis is that the two groups have the same Wasserstein barycenter over the unit interval:

$$H_0 : \mu_1(u) = \mu_2(u) \quad \text{for all } u \in (0, 1).$$

Equivalently, in terms of the difference in quantiles,

$$H_0 : \Delta(u) = 0 \quad \text{for all } u \in (0, 1).$$

To carry out this test, we can use the `equalbary_test()` function. The function takes two estimated barycenter objects as input. These objects should be estimated on the same quantile grid; the hypothesis test is performed on this grid. The argument `B` specifies the number of resamples used to approximate the null distribution, and `seed` can be specified for reproducibility. For the Hotelling's T^2-type test, we can specify the proportion of total variance to retain through the `variance_explained` argument, and the test uses the first K empirical functional principal components needed to explain at least this proportion of the variance; the default is 0.95.

```
comparison <- equalbary_test(
  mcb_res_grp1,
  mcb_res_grp2,
  B = 1000,
  seed = 1
)
```

The output includes results for three test statistics: the supremum norm (in sublist `$sup_norm`), the integrated squared norm (in sublist `$int_sq_norm`), and the Hotelling's T^2-type test statistic (in sublist `$hotelling`). Each of these test statistics are described in

Table 4.1. Each sublist contains the two-sided p-value for the test (`pval`), the observed test statistic (`T_obs`), and the simulated null (`T_sim`) test statistics. In the code below, we extract the p-values from using each of the three test statistics.

```
comparison$sup_norm$pval
comparison$int_sq_norm$pval
comparison$hotelling$pval
```

```
[1] 0.07
[1] 0.005
[1] 0.1442255
```

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

In Appendix B.1, we present two extensions to our method: allowing categorization into more than two bins and handling missing sequencing data. In Appendix B.2, we discuss alternatives to Assumption 2.3.3 (sequence depth conditional independence). In Appendix B.3, we discuss a strategy to select the threshold q_0. In Appendix B.4, we propose a pre-screening procedure for selecting binary features to test. In Appendix B.5, we provide additional intuition for how the classification model acts as a shrinkage estimator through a visual example. In Appendix B.6, we provide more details on our simulation study and include one additional simulation.

The R code used for the simulations and data analysis is available at `https://github.com/jpspeng/multiseq_cox_sieve`.

B.1 Methodology extensions

B.1.1 Categorization of > 2 bins

Instead of having two bins defined by a single cutoff threshold q_0, suppose that we would like to define our failure type into l bins defined by cut points $q := \{q_0, q_1, ..., q_l\}$, where $q_0 = 0$, $q_l = 1$, and $l > 2$. For individual i who acquires HIV-1, we now define the mark variable $J_{i,q} \in \{1, ..., l\}$ as an l-level categorical variable based on whether the true proportion Q_i is

within each pair of cut points:

$$J_{i,q} = \begin{cases} 1, & \text{if } q_0 \leq Q_i \leq q_1 \\ 2, & \text{if } q_1 < Q_i \leq q_2 \\ \ldots & \\ l & \text{if } q_{l-1} < Q_i \leq q_l \end{cases} \tag{B.1}$$

A potential example of interest would be to set cutpoints $\{q_1, q_2\}$ such that q_1 is close to 0 and q_2 is close to 1.

Our estimands of interest are now vaccine efficacy VE_j against viral quasispecies with mark $J_q = j$ for $j \in \{1, 2, ..., l\}$. To facilitate estimation and inference with a Cox model in this setting, we now assume a proportional hazards assumption for the l failure types:

$$\begin{aligned} \lambda_{1s}(t; z, x) &= \exp\left(\theta_1^\top w\right) \lambda_{0,1s}(t), \\ \lambda_{2s}(t; z, x) &= \exp\left(\theta_2^\top w\right) \lambda_{0,2s}(t), \\ &\vdots \\ \lambda_{ls}(t; z, x) &= \exp\left(\theta_l^\top w\right) \lambda_{0,ls}(t), \end{aligned} \tag{B.2}$$

where $\lambda_{0,js}(t)$ is the strata-specific baseline hazard for each failure type $j = 1, \ldots, l$ and θ_j: $j \in \{1, ..., l\}$ is the vector of parameters to estimate. We use the same modified estimating equation, for $j \in \{1, ..., l\}$, as Equation (2.7):

$$U_j^{adj}(\theta_j) \;=\; \sum_{i=1}^{n} \int_0^\tau \left\{ W_i - \overline{W}_{j,S_i}(t; \theta_j) \right\} \nu_q(j; m_i, K_i, W_i, S_i, \widetilde{T}_i)\, dN_i(t),$$

where the classification probabilities $\nu_q(j; m_i, K_i, W_i, S_i, \widetilde{T}_i)$ are now defined for each $j \in \{1, ..., l\}$. As with before, each event contributes to the estimating equation for each failure type $j \in \{1, ..., l\}$ weighted by the probability that the event was that failure type. To estimate this probability, we follow the same procedure as described in 2.3.2, except that

equation (2.16) is replaced with

$$\widehat{\nu}_q(j; m_i, K_i, B_i) = \int_{q_{j-1}}^{q_j} \widehat{f}_{Q|m,K}(q|m_i, K_i, B_i, E_i = 1)dq \tag{B.3}$$

We note that $\sum_{j=1}^{l} \nu_q(j; \cdot) = 1$.

B.1.2 Missing sequencing data

In some settings, sequencing data may be missing for a subset of individuals who acquire HIV-1 during follow-up. We describe an approach for estimation and inference in the presence of missing sequence information using inverse probability weighting (IPW) and augmented inverse probability weighting (AIPW) (Gao and Tsiatis, 2005). Let R indicate whether an individual's sequence information, (K, m), is obtained, and let A denote a vector of auxiliary covariates predictive of missingness. We observe $(W_i, S_i, \widetilde{T}_i, A_i)$ for individuals with $E_i = 0$ and

$$(W_i, S_i, \widetilde{T}_i, A_i, R_i, R_iK_i, R_im_i)$$

for individuals with $E_i = 1$.

For endpoint cases, let

$$O_i = (W_i, S_i, A_i, \widetilde{T}_i)$$

denote the variables observed regardless of whether sequencing is obtained. We assume that sequencing information is missing at random (MAR) (Rubin, 1976), so that

$$P(R = 1 \mid J_{q_0}, O, E = 1) = P(R = 1 \mid O, E = 1) =: \pi(O). \tag{B.4}$$

We also require the usual positivity condition: for some constant $\epsilon > 0$,

$$\pi(O) \geq \epsilon \quad \text{almost surely among individuals with } E = 1. \tag{B.5}$$

For sequenced endpoint cases, define

$$\nu^R_{q_0}(j; m, K, O) = P(J_{q_0} = j \mid m, K, O, R = 1, E = 1). \tag{B.6}$$

This probability can be estimated using the classification model procedure described in Section 2.3.2.

Define the observed score component

$$H_{ij}(\theta_j) = \int_0^\tau \left\{ W_i - \overline{W}_{j,S_i}(t; \theta_j) \right\} \, dN_i(t), \tag{B.7}$$

where $N_i(t)$ is the counting process for an endpoint of either failure type, as defined in the main text.

An IPW estimating function is

$$U_j^{\mathrm{IPW}}(\theta_j) = \sum_{i:\, E_i=1,\, R_i=1} \frac{H_{ij}(\theta_j)}{\pi(O_i)} \nu^R_{q_0}(j; m_i, K_i, O_i). \tag{B.8}$$

Endpoint cases with missing sequencing information do not contribute directly to this estimating equation. Instead, the inverse probability weights re-balance the sequenced cases to represent the full population of endpoint cases.

To construct an AIPW estimator, define the outcome regression

$$m_j(O_i) = P(J_{i,q_0} = j \mid O_i, E_i = 1). \tag{B.9}$$

Under the MAR assumption,

$$\mathbb{E}\left\{ \nu^R_{q_0}(j; m, K, O) \mid O, R = 1, E = 1 \right\} = m_j(O).$$

Therefore, $m_j(O)$ can be estimated by regressing the estimated classification probabilities $\widehat{\nu}^R_{q_0}(j; m_i, K_i, O_i)$ on O_i among sequenced endpoint cases and predicting for all endpoint cases.

The AIPW estimating function is

$$U_j^{\mathrm{AIPW}}(\theta_j) = \sum_{i:\, E_i=1} H_{ij}(\theta_j) m_j(O_i) + \sum_{i:\, E_i=1,\, R_i=1} \frac{H_{ij}(\theta_j)}{\pi(O_i)} \left\{ \nu_{q_0}^R(j; m_i, K_i, O_i) - m_j(O_i) \right\}. \tag{B.10}$$

Provided that the classification model is correctly specified, the AIPW estimator is doubly robust with respect to the missingness and outcome-regression models. Under the stated MAR and positivity assumptions, it remains consistent if either $\pi(O)$ or $m_j(O)$ is correctly specified.

Estimation and inference can be performed as follows:

1. Estimate the missingness probability $\pi(O) = P(R = 1 \mid O, E = 1)$ among endpoint cases.
2. Estimate $\nu_{q_0}^R(j; m, K, O)$ among sequenced endpoint cases using the classification-model procedure described in Section 2.3.2.
3. Estimate $m_j(O)$ by regressing $\widehat{\nu}_{q_0}^R(j; m, K, O)$ on O among sequenced endpoint cases and predicting for all endpoint cases.
4. Construct either $U_j^{\mathrm{IPW}}(\theta_j)$ using Equation (B.8) or $U_j^{\mathrm{AIPW}}(\theta_j)$ using Equation (B.10), and solve the corresponding estimating equation for $\widehat{\theta}_j$.
5. Estimate the variance using a nonparametric bootstrap that re-estimates the classification, missingness, and outcome-regression models within each bootstrap sample.

Under the identification and modeling assumptions stated in the main text and standard regularity conditions, the IPW estimator is consistent when the classification and missingness models are correctly specified. The AIPW estimator is consistent when the classification model is correctly specified and either the missingness model or the outcome-regression model is correctly specified. The resulting estimators are asymptotically normal under the corresponding regularity conditions.

B.2 Alternatives to Assumption 2.3.3

When estimating the conditional density $f_Q(q \mid B, E = 1)$ across conditioning groups indexed by $B = (W, S, \widetilde{T})$, some groups may contain very few observations, which can lead to unstable or poorly identified estimates of the mixing distribution. To address this, we consider stronger versions of the conditional independence assumption that reduce the size of the conditioning set and simplify the estimation of $f_Q(q \mid B, E = 1)$. In this section, we present a general form of the assumption, examine several choices of the conditioning set, and provide sufficient conditions under which the corresponding conditional independence assumptions hold.

Assumption B.2.1 (General form of sequence depth conditional independence)**.** *Let B be a user-chosen subset of $(W, S, \widetilde{T})$ and let $\widetilde{B}$ denote the remaining components of $(W, S, \widetilde{T})$. Among observed failures,*

$$Q \perp (m, \widetilde{B}) \mid B, E = 1.$$

Under Assumption B.2.1, the density needed in the classification model simplifies from $f_{Q|m}(q \mid m, W, S, \widetilde{T}, E = 1)$ to $f_Q(q \mid B, E = 1)$, reducing the dimension of the conditioning set over which the mixing distribution of Q must be estimated.

Case 1: $B = (W, S, \widetilde{T})$

Setting $B = (W, S, \widetilde{T})$ yields the least restrictive version of the assumption:

$$Q \perp m \mid W, S, \widetilde{T}, E = 1,$$

which is identical to Assumption 2.3.3 in the main text. This assumption allows Q to be associated with failure time and requires only that sequencing depth provide no additional information about Q after conditioning on $(W, S, \widetilde{T})$.

In this case, the conditional distribution $f_Q(q \mid W, S, \widetilde{T}, E = 1)$ must be modeled as a function of all components of $(W, S, \widetilde{T})$. If $\widetilde{T}$ is continuous, this requires an appropriate

smoothing or discretization strategy rather than separate estimation at each observed failure time.

Case 2: $B = (W, S)$

A more restrictive but potentially more practical choice is $B = (W, S)$, which corresponds to the assumption

$$Q \perp (m, \widetilde{T}) \mid W, S, E = 1.$$

Under this assumption, we only need to estimate $f_Q(q \mid W, S, E = 1)$, reducing the complexity of the deconvolution problem. The following conditions are sufficient for this independence assumption:

(i) Depth non-informativeness. Assume that

$$Q \perp m \mid W, S, \widetilde{T}, E = 1.$$

(ii) Proportional Q-specific hazards. Let $\lambda_s(t, q; w)$ denote the continuous mark-specific hazard density for a failure at time t with mismatch proportion $Q = q$, conditional on $W = w$ and $S = s$. Assume that

$$\lambda_s(t, q; w) = a_s(t; w) r_s(q; w),$$

where $a_s(t; w)$ and $r_s(q; w)$ are nonnegative functions. Thus, failure hazards may differ across values of Q, but their relative magnitudes do not vary over time. Because the time-dependent factor $a_s(t; w)$ cancels when the hazard is normalized over q, the conditional distribution of Q among failures does not depend on failure time. Under the censoring condition stated above, this implies

$$Q \perp \widetilde{T} \mid W, S, E = 1.$$

Together, conditions (i)–(ii) imply

$$Q \perp (m, \widetilde{T}) \mid W, S, E = 1.$$

Case 3: $B = Z$

Because $W = (Z, X)$, choosing $B = Z$ corresponds to assuming

$$Q \perp (m, \widetilde{T}, X, S) \mid Z, E = 1.$$

This is the most restrictive version of Assumption B.2.1, but it may be appropriate when (X, S) primarily capture risk factors unrelated to the viral quasispecies composition. The following conditions are sufficient:

(i) Depth non-informativeness. Assume that

$$Q \perp m \mid Z, X, S, \widetilde{T}, E = 1.$$

(ii) Proportional Q-specific hazards. As in Case 2, assume that

$$\lambda_s(t, q; z, x) = a_s(t; z, x) r_s(q; z, x).$$

Under the censoring condition stated above, this implies

$$Q \perp \widetilde{T} \mid Z, X, S, E = 1.$$

(iii) No residual association with adjustment covariates or stratum. Assume that

$$Q \perp (X, S) \mid Z, E = 1.$$

Thus, among observed failures, (X, S) provide no additional information about the mis-

match proportion after conditioning on treatment assignment. Because this assumption is made conditional on $E = 1$, it does not follow from treatment randomization and requires more substantive justification.

Together, conditions (ii)–(iii) imply

$$Q \perp (\widetilde{T}, X, S) \mid Z, E = 1.$$

Combining this result with condition (i) yields

$$Q \perp (m, \widetilde{T}, X, S) \mid Z, E = 1.$$

The three cases reflect a trade-off between the strength of the assumptions and the complexity of estimating the mixing distribution. Case 1 uses the largest conditioning set and makes the weakest independence assumption, allowing the distribution of Q to vary with failure time. Case 2 removes failure time from the conditioning set and is easier to implement, but requires the relative hazards associated with different values of Q to remain stable over follow-up. Case 3 additionally removes (X, S) and therefore requires these variables to provide no information about Q beyond treatment assignment among observed failures. In practice, the largest conditioning set that permits stable estimation should be preferred, with smaller conditioning sets used only when their stronger assumptions are scientifically plausible.

B.3 Threshold q_0 selection

While the goal of our analysis may be to test for the presence of any mismatch (i.e., set threshold $q_0 = 0$), the resolution of our data is limited by the sequencing depth of samples. Therefore, we may wish to specify a lower threshold q_0 with higher sequencing depths and a higher q_0 with lower depths. Specifically, we may wish to set our threshold q_0 to be the smallest mismatch proportion detectable, akin to the limit of detection (LOD) in assays. For a given sequencing depth, we can define the LOD as the minimum true mismatch proportion such that there is at least some pre-specified (e.g. $\geq 80\%$) probability of detection (POD) of the mismatch. This can be calculated as

$$LOD = 1 - (1 - POD)^{1/depth}.$$

Table B.1 presents the LOD for various combinations of sequencing depth and POD.

Sequence Depth	**LOD (60% POD)**	**LOD (80% POD)**	**LOD (95% POD)**
5	0.175	0.275	0.451
10	0.095	0.138	0.259
50	0.019	0.032	0.059
100	0.010	0.016	0.030
500	0.002	0.003	0.006
1000	0.001	0.002	0.003

Table B.1: Limits of detection (LOD) for different sequencing depths and probabilities of detection (POD).

In our data, each sample has its own sequencing depth and, consequently, its own limit of detection. However, our method requires the specification of a single detection threshold q_0 for the entire dataset. Therefore, we define q_0 as the maximum of the treatment arm-specific LODs computed at the median sequencing depth within each treatment arm. For example, if the median sequencing depths are 50 in the vaccine arm and 100 in the placebo arm, the respective LODs (assuming an 80% POD) are approximately 3.2% and 1.6%. In this

case, we set $q_0 = 0.032$ so that the chosen threshold reflects the most conservative detection limit between study arms. For some binary features, we may also wish to consider another threshold q_0 set as 1 minus this threshold (e.g., in our running example, this would be the presence of a non-mismatch).

B.4 Screening for viable marks

In practice, sequence datasets may include hundreds or even thousands of binary features. Testing all features simultaneously in a single multi-sequence analysis would require a substantial multiplicity adjustment, which can severely diminish statistical power. To mitigate this issue, we propose a two-step screening procedure, agnostic to treatment assignment, to reduce the feature set prior to analysis. In the first step, we exclude marks exhibiting insufficient *inter-individual* variability, as these provide little power to detect a sieve effect even under the most ideal conditions (Tarone, 1990). In the second step, we remove marks with insufficient *intra-individual* variability; such features can be adequately analyzed using standard single sequence sieve methods. Both screens depend on the chosen threshold q_0 defining the failure type J_{i,q_0}.

One way to implement the first screen is as follows. We first restrict attention to cases with observed infections ($E = 1$). Then, without examining treatment-arm–specific mark distributions, we assess whether a feature could in principle exhibit sufficient separation between treatment arms to have a detectable sieve effect under the most favorable configuration. Because the time-to-event sieve analysis involves considerably more structure — risk sets, censoring, covariate adjustment, and classification probabilities — using a simple cross-sectional comparison of case counts can ease calculation. To implement this, we identify the minimum number of primary endpoints of one virus type in one treatment arm (with zero of that type in the other arm) required for the corresponding Fisher's exact test p-value to fall below a chosen significance threshold (e.g., 0.05). Although the true virus types J_{i,q_0} are unknown and potentially misclassified, we can use the naive labels $\widetilde{J}_{i,q_0} = \mathbb{1}(K_i/m_i \geq q_0)$ for the screening. We retain only those marks for which both $n_{E=1,\widetilde{J}=1} = \sum_{i=1}^{n} \mathbb{1}(E_i = 1, \widetilde{J}_{i,q_0} = 1)$ and $n_{E=1,\widetilde{J}=0} = \sum_{i=1}^{n} \mathbb{1}(E_i = 1, \widetilde{J}_{i,q_0} = 0)$ exceed this threshold, thereby eliminating features that lack sufficient power to detect a sieve effect.

In the second screen, we exclude marks for which there is insufficient intra-individual diversity relative to the q_0 threshold, such that the resulting multi-sequence analysis would

be expected to yield conclusions similar to a single sequence sieve analysis. An illustration is given in Figure B.1. In hypothetical data (a), the virus classifications obtained from the modal sequence ($q_0 = 0.5$, top panel) and from applying a 5% mismatch threshold ($q_0 = 0.05$, bottom panel) agree for all but one individual. In this setting, the multi-sequence analysis would not provide a meaningfully different result to the modal sequence analysis. In contrast, in hypothetical data (b), a substantial proportion of individuals change failure-type classification when using the 5% threshold, indicating that the multi-sequence analysis at that threshold may capture information not available from the modal sequence alone. As in the first screen, we rely on the naive labels $\widetilde{J}_{i,q_0}$ since the true virus types are unknown. A practical approach is to exclude marks for which fewer than a specified proportion (e.g., 10%) of individuals are reclassified under the chosen threshold.

B.4.1 Screening for real data analysis

In the data application presented in the manuscript, we applied two screening criteria. First, we screened for adequate inter-individual diversity: we required at least four primary endpoints with raw feature proportions $< 1\%$ (or $< 99\%$) and at least four with proportions $\geq 1\%$ (or $\geq 99\%$). Second, we required sufficient intra-individual diversity: the dichotomization induced by the 1% (or 99%) threshold had to differ from the dichotomization based on the mindist-sequence value in at least 10% of primary endpoints.

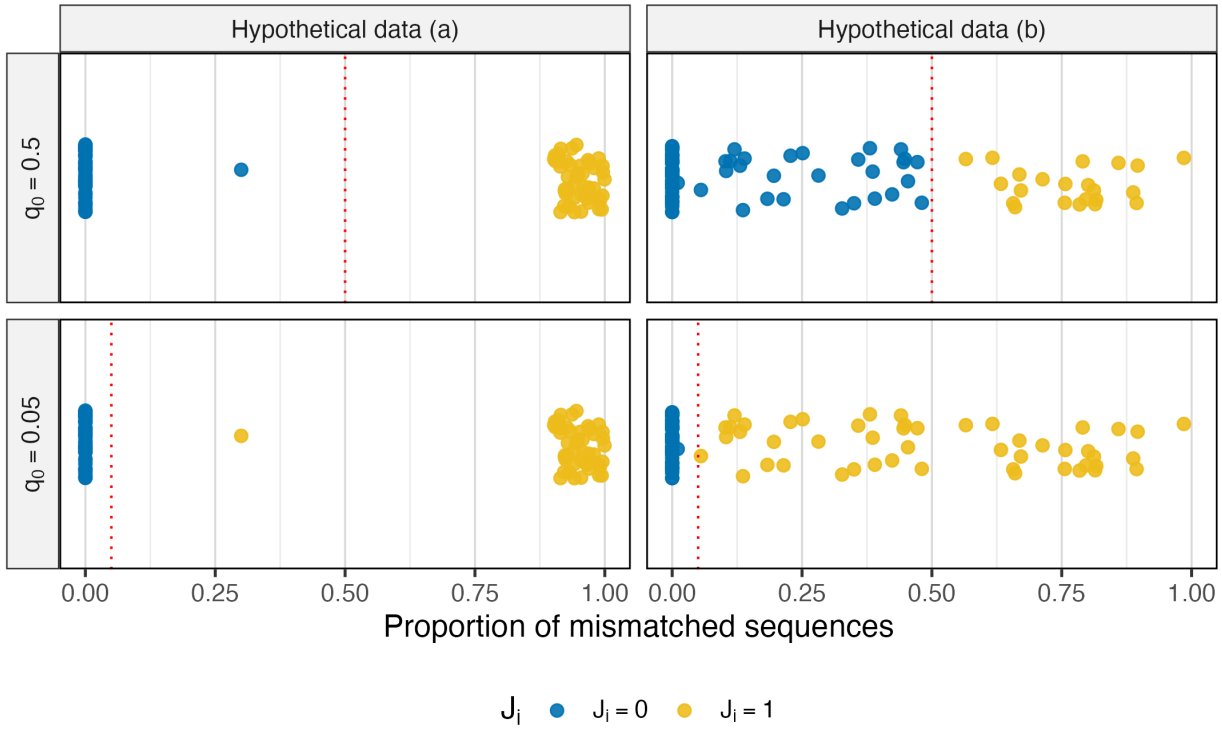


Figure B.1: Illustration of the second screening step based on intra-individual diversity. Panels (a) and (b) show hypothetical datasets in which each point represents the proportion of mismatched sequences for an individual, with colors indicating the resulting virus-type classification. The dashed red line denotes the threshold q_0. In scenario (a), changing q_0 from 0.5 (top) to 0.05 (bottom) produces nearly identical classifications, indicating that a multi-sequence analysis cannot provide additional information beyond that provided from a modal mark ($q_0 = 0.5$) approach. In scenario (b), a substantial proportion of individuals change classification under the lower threshold, indicating that the multi-sequence method may meaningfully differ from the single-sequence analysis for this mark.

B.5 Classification model as a shrinkage estimator

The estimator for the classification probabilities can be considered as a shrinkage estimator, combining the information from each observation with information from the entire sample. Equation (2.16) can be written as follows:

$$\hat{f}_{Q|m,K}(q \mid m_i, K_i, B_i, E_i = 1) = \frac{f_{\text{binom}}(K_i; m_i, q)\, \widehat{f}_Q(q \mid B_i, E_i = 1)}{\int_0^1 f_{\text{binom}}(K_i; m_i, q)\, g(q; \hat{\gamma}_{B_i})\, dq}. \tag{B.11}$$

In this equation, $f_{\text{binom}}(K_i; m_i, q)$ represents the information from the individual's observation, while $\widehat{f}_Q(q \mid B_i, E_i = 1)$ represents the information from all the observations. If an observation's sequencing depth m_i is large, the estimator $\widehat{f}_{Q|m,K}$ relies more on the observation rather than the sample. In contrast, if an observation's sequencing depth is small, the estimator $\widehat{f}_{Q|m,K}$ relies more on information from the entire sample's distribution (Whittemore, 1989).

We can see this depicted in a simple example shown in Figure B.2. In this example, we restrict to a single stratum defined by $B = b$, and assume that the prior for $Q \mid B = b, E = 1$ is estimated to be a Beta$(2, 2)$ distribution. Suppose we have three observations in this stratum, enumerated as $i = 1, 2, 3$, each with $K_i = 0$ mismatches but different sequencing depth: (1) $m_1 = 1$, (2) $m_2 = 10$, and (3) $m_3 = 100$. For each of these observations, we display the components of (2.16), including:

- $f_{\text{binom}}(K_i; m_i, q)$,
- the prior $\widehat{f}_Q(q \mid B = b, E = 1)$, and
- the posterior $\widehat{f}_{Q|m,K}(q \mid m_i, K_i, B_i, E_i = 1)$.

In the first observation with $m_1 = 1$ (red), the posterior $\widehat{f}_{Q|m,K}(q \mid m_i, K_i, B_i, E_i = 1)$ (right panel) is only slightly changed from the prior $\widehat{f}_Q(q \mid B = b, E = 1)$ (middle panel). However, for the third observation with sequencing depth $m_3 = 100$ (blue), the posterior is heavily weighted toward the observation (left panel) and differs greatly from the prior. This

illustrates how the estimator combines individual- and sample-level information, with the relative influence of each determined by the sequencing depth m_i.

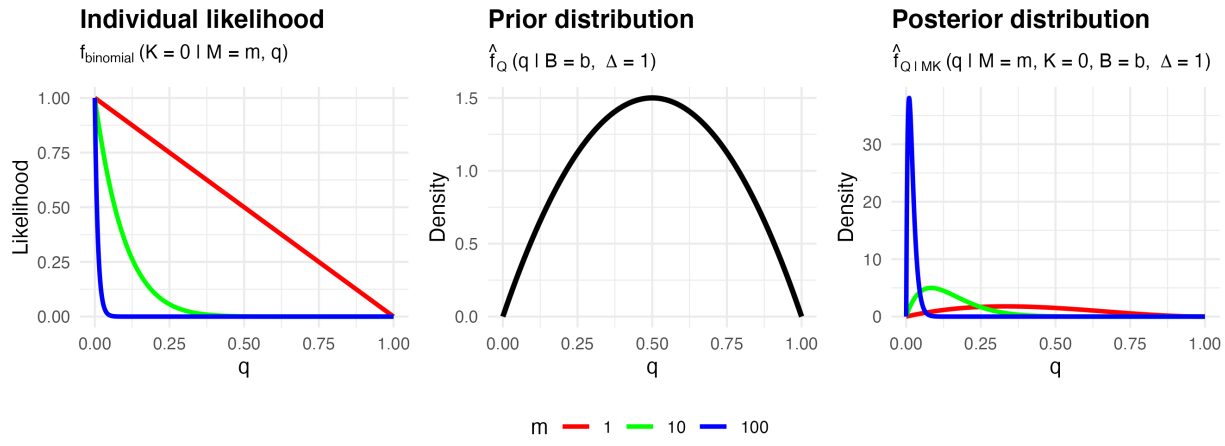


Figure B.2: The components of equation (B.11) in a toy example.

B.6 Simulation details and additional simulations

B.6.1 Details on Simulation Study Setup

Failure times are generated with exponential distributions: $T_0 \sim \text{Exponential}(\gamma_0)$ and $T_1 \sim \text{Exponential}(\gamma_1)$, where the rate parameters depend on treatment assignment: $\gamma_0 \mid (Z, X) = 0.01 \exp(\beta_0 Z - 0.105X)$ and $\gamma_1 \mid (Z, X) = 0.03 \exp(\beta_1 Z - 0.223X)$. The two failure times are simulated independently, but there will be a correlation between T_0 and T_1 for each individual induced by their shared vaccination status and covariate. If $T_0 < T_1$, the failure time is T_0 with failure type $J_{q_0} = 0$; otherwise, the failure time is T_1 with $J_{q_0} = 1$. We draw the true mismatch probabilities Q using truncated Beta distributions as follows:

$$Q \mid J_{q_0} \sim \begin{cases} \text{Beta}(a = 0.5,\, b) \text{ truncated to } [0,\, 0.01), & \text{if } J_{q_0} = 0, \\ \text{Beta}(a = 0.5,\, b) \text{ truncated to } [0.01,\, 1], & \text{if } J_{q_0} = 1. \end{cases}$$

The shape parameter b of the Beta distribution differs for Settings (a)–(c) described below, so that the true marginal distribution of $Q|E = 1$ arises approximately from a Beta distribution. We assume administrative censoring at $t = 5$. With this setup, the probability of the event by $t = 5$ in the placebo arm is roughly 15%.

We consider three vaccine efficacy scenarios:

- Setting (a) (no VE): $\beta_0 = \beta_1 = 0$, corresponding to 0% VE for both viral types. For this setting, we set the Beta distribution's shape parameter $b = 5.7$.
- Setting (b) (positive VE, no sieve effect): $\beta_0 = \log(1 - 0.5)$ and $\beta_1 = \log(1 - 0.5)$, resulting in 50% VE against both viral types. For this setting, we set the Beta distribution's shape parameter $b = 5.7$.
- Setting (c) (positive VE with sieve effect): $\beta_0 = \log(1 - 0.5)$ and $\beta_1 = \log(1 - 0.05)$, giving 50% VE against viral type $J_{q_0} = 0$ and 5% VE against viral type $J_{q_0} = 1$. For this setting, we set the Beta distribution's shape parameter $b = 3.8$.

B.6.2 Simulation Study #1 – Results

Results for Simulation Study #1, in which each endpoint case has a large sequencing depth (m = 2,000), are displayed in Figure B.3. Across all three settings and sample sizes, the uncorrected estimator performs well: its median estimates are close to the true vaccine efficacies, and its nominal 95% confidence intervals achieve near-nominal coverage. This is expected in this high sequencing depth scenario because misclassification of the empirical indicator $\mathbb{1}(K_i/m_i >= q_0)$ is extremely rare. The corrected estimator also performs well under these conditions, with point estimates and confidence intervals closely matching those of the uncorrected estimator. For settings (a) and (b), where the true vaccine efficacies are equal across viral types, both estimators exhibit appropriate type I error rates when testing the null hypothesis H_{B0}. In setting (c), which includes a true sieve effect, both estimators show high power to reject H_{B0}, with power of 52.8% (uncorrected) and 50.6% (corrected) for 1,000 participants per treatment arm, 86.2% (uncorrected) and 86.4% (corrected) for 2,000 per arm, and 96.6% (uncorrected) and 97.0% (corrected) for 3,000 per arm.

B.6.3 Simulation Study #4

In this additional simulation study, we confirm the properties of the estimator when the mixing distribution is specified correctly. Although this scenario is unrealistic for real data, it allows us to evaluate the estimator's operating behavior under perfectly satisfied model assumptions.

Data generating mechanism

Failure times are generated under a competing risks framework with two latent failure times. For each individual, the latent times for the two failure types are drawn independently as

$$T_0 \sim \text{Exponential}(\gamma_0), \qquad T_1 \sim \text{Exponential}(\gamma_1),$$

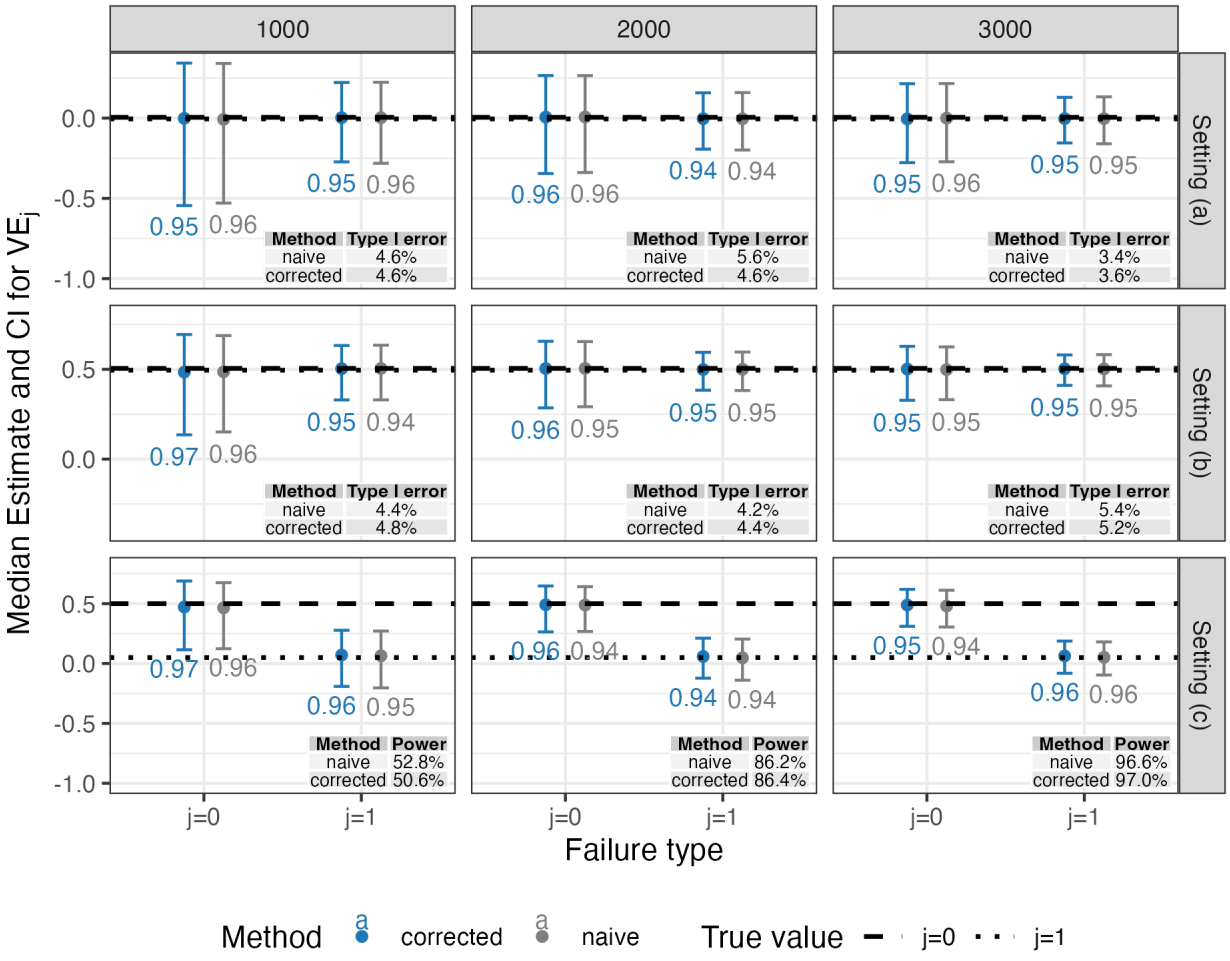


Figure B.3: Results for Simulation Study #1. The top, middle, and bottom panels display results for settings (a), (b), and (c) respectively. The left, middle, and right panels display results for differing sequencing depths per treatment arm. Each graph displays points for median VE estimates for each failure type across all simulations, along with median lower and upper 95% confidence interval bounds. The dashed and dotted horizontal lines are placed at the true values for the $j = 0$ and $j = 1$ failure types, respectively. The numbers below each error bar display the confidence interval coverage. We compare Type I error between the two estimators in settings (a) and (b) and power to detect a sieve effect in setting (c).

where the rate parameters depend on treatment assignment Z. Specifically,

$$\gamma_0 = \kappa_Z(1 - p_Z), \qquad \gamma_1 = \kappa_Z p_Z,$$

where κ_Z controls the overall failure rate in treatment group Z, and p_Z determines the relative frequency of the two failure types. In the simulations, we set $\kappa_0 = 0.04$ and $\kappa_1 = 0.03$, and we vary p_z for $z \in \{0, 1\}$ in Settings (a) and (b), as described below.

The failure type is defined as

$$J_{q_0} = \begin{cases} 0, & \text{if } T_0 < T_1, \\ 1, & \text{if } T_1 < T_0. \end{cases}$$

Administrative censoring is imposed at $t = 5$.

For individuals who experience a failure ($E = 1$), we generate the true mismatch probability Q from truncated Beta distributions so that the failure type corresponds to whether Q exceeds a fixed threshold $q_0 = 0.01$:

$$Q \mid (J_{q_0}, Z) \sim \begin{cases} \text{Beta}(\alpha_Z, \beta_Z) \text{ truncated to } [0, q_0], & J_{q_0} = 0, \\ \text{Beta}(\alpha_Z, \beta_Z) \text{ truncated to } (q_0, 1], & J_{q_0} = 1. \end{cases}$$

The marginal distribution of Q among failures in treatment group $Z = z$ is therefore $\text{Beta}(\alpha_z, \beta_z)$.

We consider two configurations for the distribution of Q:

- **Setting (a):** $Q \mid (E = 1, Z = z) \sim \text{Beta}(1, 50)$ for both treatment groups. This corresponds to a true $VE_j = 0.25$ for both $j \in \{1, 2\}$. To generate this, we input $p_0 = p_1 = 1 - F_{\text{Beta}(1,50)}(0.01) = 0.99^{50} \approx 0.605$.
- **Setting (b):** $Q \mid (E = 1, Z = 0) \sim \text{Beta}(1, 50)$ and $Q \mid (E = 1, Z = 1) \sim \text{Beta}(1, 60)$.

This corresponds to a true $VE_0 = 0.14$ and $VE_1 = 0.32$. To generate this, we input

$$p_0 = 1 - F_{\text{Beta}(1,50)}(0.01) = 0.99^{50} \approx 0.605, \qquad p_1 = 1 - F_{\text{Beta}(1,60)}(0.01) = 0.99^{60} \approx 0.547.$$

We consider the case in which sequencing depth variability differs between treatment arms. In the placebo arm, m ranges uniformly from 1 to 15 in 20% of endpoint cases and from 16 to 1,000 in the remaining cases. In the vaccine arm, m ranges uniformly from 1 to 15 in 40% of endpoint cases and from 16 to 1,000 in the remaining cases. Simulation sample sizes of $n = 1000$, 2000, and 3000 individuals per treatment arm are considered.

Across these sets of conditions, we compare bias, standard error, and confidence interval coverage between the proposed estimator and the uncorrected estimator (which uses the empirical, possibly incorrect indicator $\widetilde{J}_{i,q_0} = \mathbb{1}(K_i/m_i \geq q_0)$ as the failure cause) over 1,000 simulations. We estimate the mixing distributions of Q in each level of Z with Beta distributions. Standard errors are estimated using 1,000 bootstrap samples, and Wald confidence intervals are constructed based on these estimates.

In Setting (a), the null hypothesis $VE_0 = VE_1$ holds. Therefore, we compare the type I error rates when testing the null H_{B0} between the corrected and uncorrected methods. In Setting (b), the null hypothesis of equal vaccine efficacies across virus types does not hold, so we compare the power to detect the sieve effect and reject H_{B0} between the two estimators.

Simulation results

Results are shown in Figure B.4. In Setting (a), where the null hypothesis of equal vaccine efficacies holds, the uncorrected estimator exhibits substantial bias in opposite directions for the two failure types: VE_0 is underestimated while VE_1 is overestimated. This produces confidence interval coverage below the nominal level, with coverage worsening as sample size increases. The resulting type I error rate is highly inflated, increasing from 34.3% with $n = 1000$ individuals per treatment arm to 78.3% with $n = 3000$ individuals per treatment arm. In contrast, the corrected estimator is approximately unbiased, achieves near-nominal

confidence interval coverage, and maintains type I error close to the nominal 5% level across sample sizes.

In Setting (b), where the null hypothesis of equal vaccine efficacies does not hold, the uncorrected estimator again shows substantial bias, underestimating VE_0 and overestimating VE_1. Although this leads to high apparent power to reject H_{B0}, the rejection is driven in part by bias in the estimated vaccine efficacies rather than accurate recovery of the true sieve effect. The corrected estimator remains closer to the true values and achieves near-nominal confidence interval coverage for both failure types. Its power to detect the sieve effect increases with sample size, from 15.3% at $n = 1000$ individuals per treatment arm to 28.7% at $n = 3000$ individuals per arm.

B.6.4 Simulation Study #5

In the main simulation studies, we fixed the threshold at $q_0 = 0.01$. In this additional simulation study, we examine the sensitivity of the proposed estimator to the choice of q_0 by evaluating performance at $q_0 \in \{0.01, 0.02, 0.05\}$. This simulation uses a data-generating mechanism in which, under the sieve setting, vaccine efficacy varies with the underlying mismatch proportion Q, so that viral populations with lower and higher mismatch proportion have different failure rates.

Data generating mechanism

We use the same basic setup as in the main simulation studies, with from two equal-sized treatment arms (2,000 participants each), placebo ($Z = 0$) and vaccine ($Z = 1$), and a binary covariate $X \sim \text{Bernoulli}(0.5)$. Administrative censoring is imposed at $t = 5$.

Instead of generating failure times from two discrete failure types, we generate failure times from a continuous model. Let $f_0(q)$ denote the distribution of Q among placebo-arm failures, and let $h(q)$ denote the vaccine-arm hazard ratio for an infection with mismatch

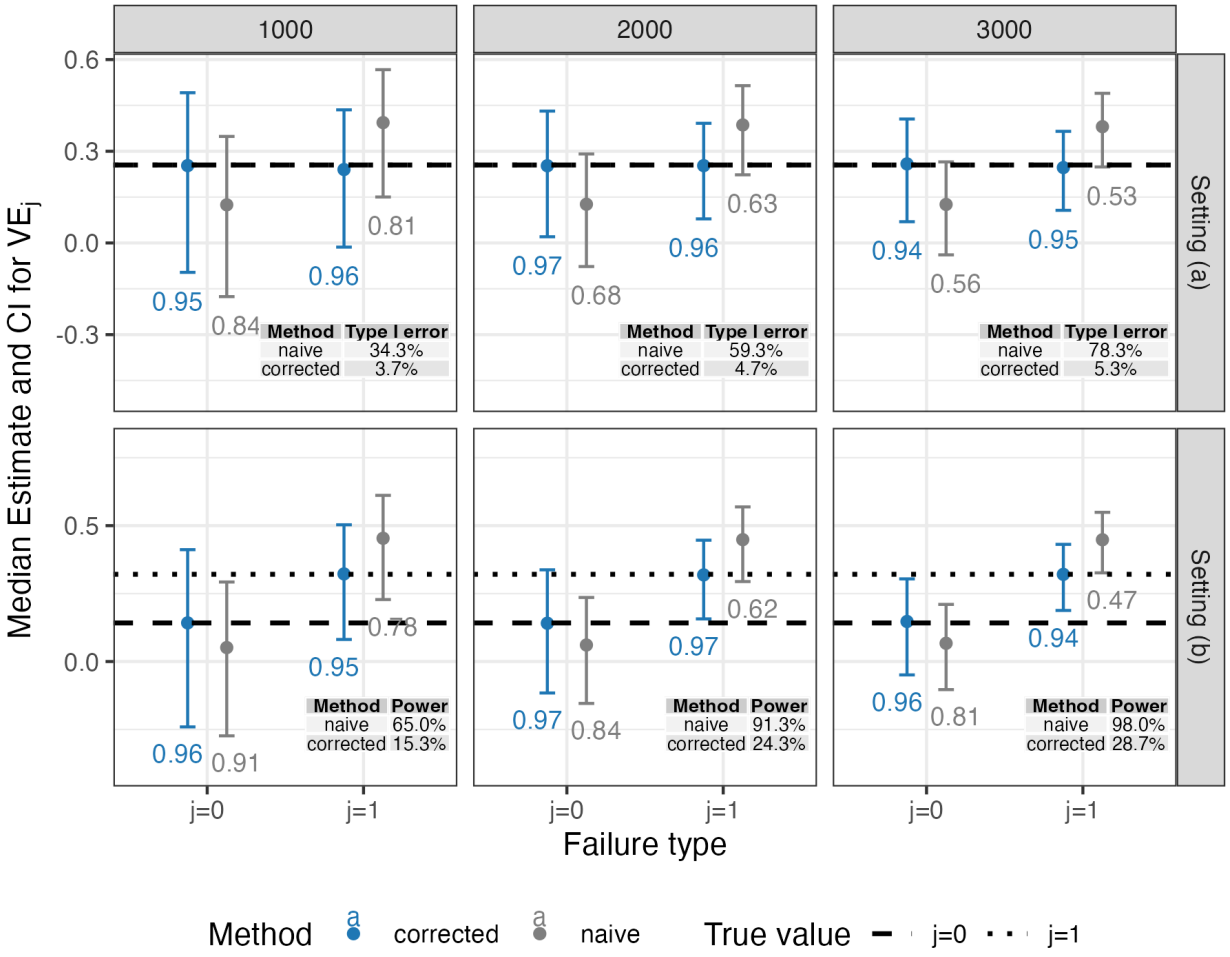


Figure B.4: Results for Simulation Study #4. The top and bottom panels display results for settings (a) and (b) respectively. The left, middle, and right panels display results for differing sample sizes per treatment arm. Each graph displays points for median VE estimates for each failure type across all simulations, along with median lower and upper 95% confidence interval bounds. The dashed and dotted horizontal lines are placed at the true values for the $j = 0$ and $j = 1$ failure types, respectively. The numbers below each error bar display the confidence interval coverage. We compare Type I error between the two estimators in (a) and power to detect a sieve effect in (b).

proportion q. The marginal vaccine-arm hazard ratio is

$$A_1 = \int_0^1 h(q) f_0(q)\, dq,$$

with $A_0 = 1$ for the placebo arm. Failure times are generated as

$$T \sim \text{Exponential}\left\{\lambda_0 \exp(\alpha X) A_Z\right\},$$

where $\lambda_0 = -\log(0.85)/5$, $\alpha = \log(0.9)$, and $A_Z = A_0$ for $Z = 0$ and $A_Z = A_1$ for $Z = 1$.

For individuals who experience a failure ($E = 1$), Q is drawn from $f_0(q)$ in the placebo arm and from

$$f_1(q) = \frac{h(q) f_0(q)}{\int_0^1 h(u) f_0(u)\, du}$$

in the vaccine arm. Therefore, when $h(q)$ is larger for larger values of q, vaccine-arm failures are enriched for viral populations with higher mismatch proportions.

We specify the placebo failure-mark distribution as

$$f_0(q) = 0.45\,\text{Beta}(0.35, 80) + 0.35\,\text{Beta}(2, 35) + 0.20\,\text{Beta}(4, 30).$$

This distribution places substantial mass near zero while retaining non-negligible mass above each of the thresholds 0.01, 0.02, and 0.05.

We consider two vaccine efficacy settings:

- **Setting (a): positive VE, no sieve effect.** We set $h(q) = 0.58$ for all q, corresponding to a constant vaccine efficacy of 42% for all values of the mismatch proportion.
- **Setting (b): positive VE with a sieve effect.** We set

$$h(q) = 0.50 + 0.45\,\text{expit}\left\{\frac{q - 0.02}{0.01}\right\}.$$

This corresponds to stronger vaccine protection against viral populations with low

mismatch burden and weaker vaccine protection against viral populations with higher mismatch burden.

For each simulated dataset, we evaluate the estimators separately at $q_0 = 0.01$, 0.02, and 0.05. The true vaccine efficacy for failure type j at threshold q_0 is

$$VE_j(q_0) = 1 - \frac{\int_{\mathcal{Q}_j(q_0)} h(q) f_0(q)\, dq}{\int_{\mathcal{Q}_j(q_0)} f_0(q)\, dq},$$

where $\mathcal{Q}_0(q_0) = \{q : q < q_0\}$ and $\mathcal{Q}_1(q_0) = \{q : q \geq q_0\}$. In Setting (a), $VE_0(q_0) = VE_1(q_0)$ for all values of q_0. In Setting (b), the two vaccine efficacies differ, with the magnitude of the sieve effect depending on the threshold.

We use the same sequencing-depth mechanism as Simulation Study #3, in which sequencing depth variability differs between treatment arms. In the placebo arm, m ranges uniformly from 1 to 15 in 20% of endpoint cases and from 16 to 1,000 in the remaining cases. In the vaccine arm, m ranges uniformly from 1 to 15 in 40% of endpoint cases and from 16 to 1,000 in the remaining cases. Given m and Q, we generate $K \mid m, Q \sim \text{Binomial}(m, Q)$.

Across these conditions, we compare bias, standard error, and confidence interval coverage between the proposed estimator and the uncorrected estimator at each value of q_0. For the proposed estimator, classification probabilities are estimated separately for each threshold using penalized spline mixtures with df $= 5$ and $c_0 = 0.01$, augmented with point mass basis functions at 0 and 1. We use the same sample sizes, number of simulation replicates, bootstrap procedure, and Wald confidence interval construction as in the main simulation studies.

In Setting (a), the null hypothesis $VE_0(q_0) = VE_1(q_0)$ holds for all values of q_0. Therefore, we compare type I error rates when testing H_{B0} between the corrected and uncorrected methods. In Setting (b), the null hypothesis of equal vaccine efficacies does not hold, so we compare power to detect the sieve effect and reject H_{B0} across the three threshold choices.

Simulation results

Results are shown in Figure B.5. In Setting (a), where there is positive vaccine efficacy but no sieve effect, both estimators produce median estimates close to the true values across the three thresholds. However, the uncorrected estimator shows some inflation of the type I error rate at smaller thresholds, with type I error of 12.3% and 11.0% for $q_0 = 0.01$ and $q_0 = 0.02$, respectively. In contrast, the corrected estimator maintains type I error closer to the nominal 5% level across thresholds.

In Setting (b), where vaccine efficacy varies with the underlying mismatch proportion Q, the uncorrected estimator attenuates the contrast between $VE_0(q_0)$ and $VE_1(q_0)$. This is most evident at smaller thresholds, where the uncorrected estimator tends to underestimate $VE_0(q_0)$ and overestimate $VE_1(q_0)$. The corrected estimator is closer to the true values and achieves near-nominal confidence interval coverage across most settings. The corrected estimator also has higher power to detect the sieve effect than the uncorrected estimator at all thresholds, with power of 53.3%, 59.3%, and 46.3% for $q_0 = 0.01$, 0.02, and 0.05, respectively, compared to 23.0%, 30.0%, and 33.3% for the uncorrected estimator. The larger confidence intervals for $VE_1(q_0)$ at higher thresholds reflect the smaller number of endpoint cases with $Q \geq q_0$.

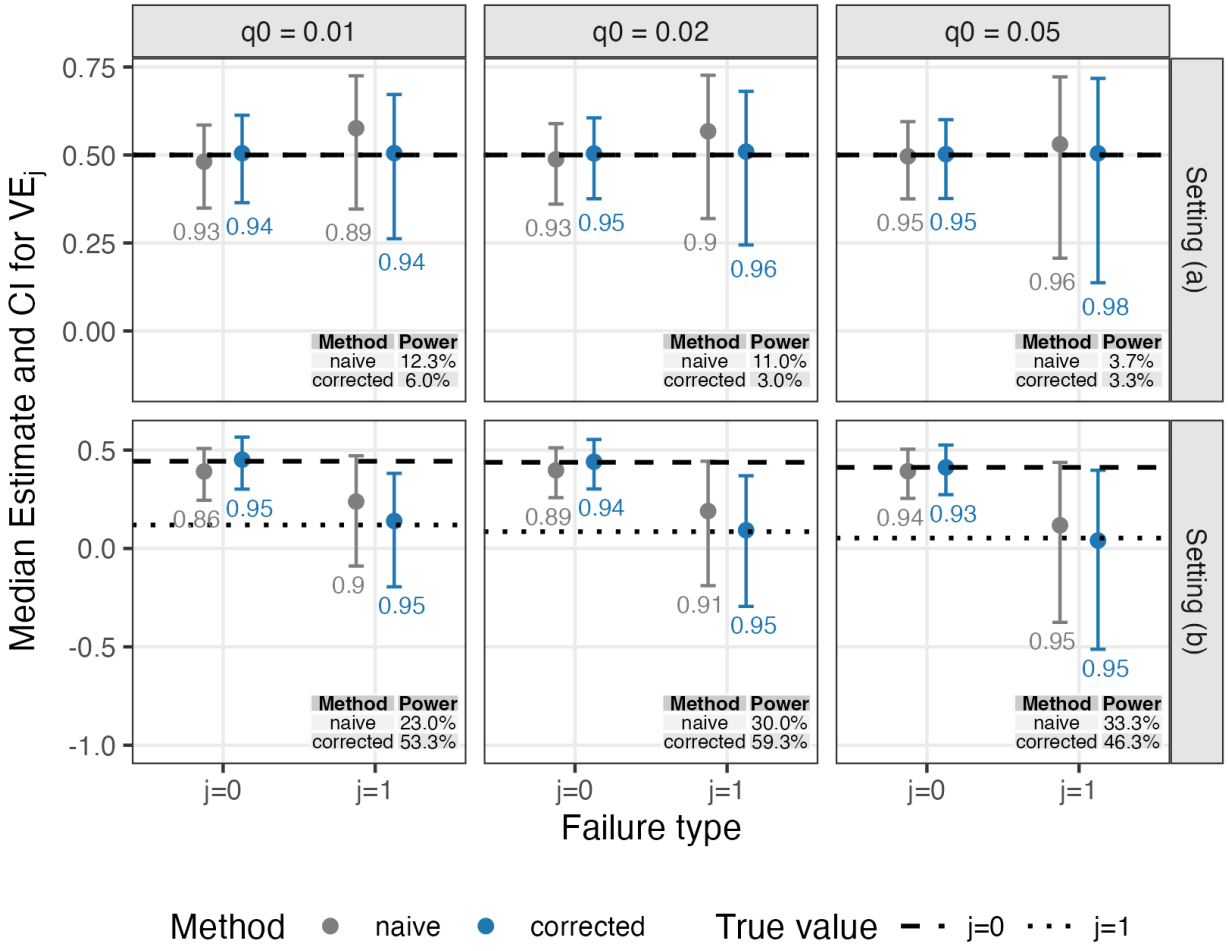


Figure B.5: Results for Simulation Study #5. The top and bottom panels display results for settings (a) and (b) respectively. The left, middle, and right panels display results for differing thresholds q_0. Each graph displays points for median VE estimates for each failure type across all simulations, along with median lower and upper 95% confidence interval bounds. The dashed and dotted horizontal lines are placed at the true values for the $j = 0$ and $j = 1$ failure types, respectively. The numbers below each error bar display the confidence interval coverage. We compare Type I error between the two estimators in (a) and power to detect a sieve effect in (b).

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

The technical results in these supplementary materials were derived in collaboration with Florian Stijven.

In Appendix C.1, we provide proofs of the main results from Chapter 3. In Appendix C.3, we show the method applied to additional data applications. In Appendix C.4, we provide technical results for weak convergence. In Appendix C.5, we provide asymptotic theory for parametric binomial mixture estimators.

C.1 Proofs of main results

C.1.1 Proof of Proposition 3.3.2

We first state and prove two lemmas that are needed for the proof of Proposition 3.3.2. Lemma C.1.1 shows that, for any fixed $\alpha \in (0,1)$ and $C \in \mathbb{N}$, there exist two distinct probability measures on $[0,1]$ that have the same first C moments but different CDF values at $\alpha-$. Lemma C.1.2 is used in the proof of the Lemma C.1.1.

Lemma C.1.1. *Let G be a probability measure on $[0,1]$ with Lebesgue density bounded away from zero. For any $\alpha \in (0,1)$ and $C \in \mathbb{N}$, there exist two distinct probability measures G_1 and G_2 on $[0,1]$ such that $\int_0^1 t^k \, dG_1(t) = \int_0^1 t^k \, dG_2(t)$ for all $k = 0, 1, \ldots, C$, but $\int_0^1 \mathbf{1}(t < \alpha) \, dG_1(t) \neq \int_0^1 \mathbf{1}(t < \alpha) \, dG_2(t)$.*

Proof. We will construct a signed measure $\nu^\star$ with bounded Lebesgue density such that $\int_0^1 t^k \, d\nu^\star(t) = 0$ for all $k = 0, 1, \ldots, C$, but $\int_0^1 \mathbf{1}(t < \alpha) \, d\nu^\star(t) \neq 0$. Then we can take $G_1 := G + \epsilon\nu^\star$ and $G_2 := G - \epsilon\nu^\star$ for some sufficiently small $\epsilon > 0$ (such that G_1 and G_2 are probability measures) to get the desired result.

Write g for the Lebesgue density of G and set $c := \inf_{x\in[0,1]} g(x) > 0$ (this infimum is positive by assumption).

Signed measure $\nu^\star$. Let ν be a signed measure on $[0,1]$ with Lebesgue density a polynomial of degree $M \geq 2C+2$ and coefficients $\boldsymbol{a} := (a_0, a_1, \ldots, a_M)$. The moment conditions $\int_0^1 x^k \, d\nu(x) = 0$ for $k = 0, 1, \ldots, C$ translate to

$$\int_0^1 x^k \sum_{j=0}^M a_j x^j \, dx = \sum_{j=0}^M a_j \int_0^1 x^{k+j} \, dx = \sum_{j=0}^M a_j \frac{1}{k+j+1} = 0 \quad \text{for } k = 0, 1, \ldots, C.$$

Let H be the $(C+1)\times(M+1)$ Hilbert matrix with entries $H_{i,j} = \frac{1}{i+j-1}$ for $i = 1, 2, \ldots, C+1$ and $j = 1, 2, \ldots, M+1$. The previous display can be written as $H\boldsymbol{a} = \boldsymbol{0}$.

We further have that

$$\int_0^\alpha d\nu(x) = \sum_{j=0}^M a_j \int_0^\alpha x^j \, dx = \sum_{j=0}^M a_j \frac{\alpha^{j+1}}{j+1}.$$

By Lemma C.1.2 below, the vector $\boldsymbol{y} := (y_1, \ldots, y_{M+1})$ with $y_j := \alpha^j/j$ does not lie in the row space of H for $\alpha \in (0,1)$. Since $\ker(H)$ is the orthogonal complement of the row space of H, there exists a nonzero vector $\boldsymbol{a}^\star \in \ker(H)$ such that $\boldsymbol{y} \cdot \boldsymbol{a}^\star \neq 0$. Equivalently,

$$\sum_{j=0}^M a_j^\star \frac{\alpha^{j+1}}{j+1} \neq 0.$$

Taking $\nu^\star$ to have Lebesgue density $\sum_{j=0}^M a_j^\star x^j$ yields the desired signed measure.

Alternative probability measure. Let $b := \sup_{x\in[0,1]} |\sum_{j=0}^M a_j^\star x^j| < \infty$, which is finite because a polynomial is bounded on a compact set. Choose $\epsilon > 0$ such that $0 < \epsilon < \frac{c}{b}$. Then $G_1 := G + \epsilon\nu^\star$ and $G_2 := G - \epsilon\nu^\star$ are valid probability measures, and $\int_0^1 t^k \, dG_1(t) = \int_0^1 t^k \, dG_2(t)$ for all $k = 0, 1, \ldots, C$, but $\int_0^1 \mathbf{1}(t < \alpha) \, dG_1(t) \neq \int_0^1 \mathbf{1}(t < \alpha) \, dG_2(t)$. □

Lemma C.1.2. *Choose $C, M \in \mathbb{N}$ such that $2C+2 \leq M+1$ and let $y_j := \frac{\alpha^j}{j}$ for $j = 1, 2, \ldots, M+1$. If $\boldsymbol{y} := (y_1, y_2, \ldots, y_{M+1})$ lies in the row space of the $(C+1)\times(M+1)$*

Hilbert matrix H with entries $H_{i,j} = \frac{1}{i+j-1}$ for $i = 1, 2, \ldots, C+1$ and $j = 1, 2, \ldots, M+1$, then $\alpha = 1$ or $\alpha = 0$.

Proof. Assume that $\boldsymbol{y}$ lies in the row space of H. Then there exists $\boldsymbol{b} := (b_1, b_2, \ldots, b_{C+1})$ such that

$$y_j = \sum_{i=1}^{C+1} b_i H_{i,j} = \sum_{i=1}^{C+1} b_i \frac{1}{i+j-1} \quad \text{for } j = 1, 2, \ldots, M+1.$$

Define

$$D(j) := j(j+1)\cdots(j+C) = \prod_{k=0}^{C}(j+k).$$

For each $i = 1, 2, \ldots, C+1$, the quotient $D(j)/(i+j-1)$ is a polynomial in j of degree C, and hence

$$D(j)\, y_j = \sum_{i=1}^{C+1} b_i \frac{D(j)}{i+j-1} \quad \text{for } j = 1, 2, \ldots, M+1.$$

is a polynomial in j of degree at most C. Consequently, the $(C+1)$-th forward difference $\Delta^{C+1}[D(j)\, y_j]$ is zero for $j = 1, 2, \ldots, C+1$. For $j > C+1$, the forward difference may not be defined as it depends on values of $D(j)\, y_j$ for $j > 2C+2$.

On the other hand, by definition

$$D(j)\, y_j = \frac{D(j)\, \alpha^j}{j} =: P(j)\, \alpha^j \quad \text{for } j = 1, 2, \ldots, M+1,$$

where P is a polynomial of degree C. The $(C+1)$-th forward difference is as follows for $j = 1, 2, \ldots, C+1$:

$$\Delta^{C+1}[P(j)\, \alpha^j] = \sum_{r=0}^{C+1} (-1)^{C+1-r} \binom{C+1}{r} P(j+r)\, \alpha^{j+r} = \alpha^j \sum_{r=0}^{C+1} (-1)^{C+1-r} \binom{C+1}{r} \alpha^r P(j+r).$$

Assume $\alpha \neq 0$, divide both sides by α^j, and denote

$$S(j) := \sum_{r=0}^{C+1} (-1)^{C+1-r} \binom{C+1}{r} \alpha^r P(j+r).$$

From the previous paragraph, we have that $\Delta^{C+1}[P(j)\,\alpha^j] = \Delta^{C+1}[D(j)\,y_j] = 0$ and thus $S(j) = 0$ for $j = 1, 2, \ldots, C+1$. Since $S(j)$ is a polynomial of the same degree as P, which is C, it follows that S is the zero polynomial and thus all coefficients of S are zero. From the definition of P follows that the leading coefficient is $p_C := 1$. From the previous display follows that the leading coefficient of S is

$$p_C \sum_{r=0}^{C+1} (-1)^{C+1-r} \binom{C+1}{r} \alpha^r = p_C(\alpha - 1)^{C+1},$$

where the equality follows from the binomial theorem. This can only be zero for $\alpha = 1$. We have before excluded the case $\alpha = 0$ when we divided by α^j, so the only other possibility is $\alpha = 0$. □

The proof of Proposition 3.3.2 is given below.

Proof. We provide a simple example showing the failure of identification. Let $\mathcal{I} := \{0, 1\}$. Then each latent distribution ν_i is Bernoulli, determined by the parameter $p_i := \nu_i(\{0\}) = F_{\nu_i}(0)$. Its quantile function is

$$F^-_{\nu_i}(\alpha) = \begin{cases} 0 & \text{if } p_i \geq \alpha, \\ 1 & \text{if } p_i < \alpha. \end{cases}$$

Hence, for any fixed $\alpha \in (0, 1)$, we have $q^\star_\alpha = \mathbb{E}_{\nu\sim\Pi}\big[F^-_\nu(\alpha)\big] = \mathbb{P}_{\nu\sim\Pi}(p < \alpha)$, where p relates to ν as defined above. Thus, it suffices to show that the distribution function $t \mapsto \mathbb{P}_{\nu\sim\Pi}(p \leq t)$ is not identifiable.

Under this construction, the observed data reduce to the pair (m_i, Y_i) where $Y_i := \sum_{j=1}^{m_i} \mathbf{1}\{X_{i,j} = 0\}$. Conditional on (m_i, p_i) we have $Y_i \mid m_i, p_i \sim \text{Bin}(m_i, p_i)$, so the marginal law of (Y_i, m_i) is a binomial mixture with mixing distribution $t \mapsto \mathbb{P}_{\nu\sim\Pi}(p \leq t)$. If $\mathbb{P}_{m\sim\eta}(m \leq C) = 1$ for some $C < \infty$, this mixture model identifies at most the first C moments of the mixing distribution (Wood, 1999), but not the full mixing distribution itself.

By Lemma C.1.1, there exist two distinct laws on p_i (and hence two distinct laws Π on distributions) that induce the same observed data distribution while yielding different values of $\mathbb{P}_{\nu\sim\Pi}(p < \alpha)$, and therefore different values of $q^\star_\alpha$. □

Remark C.1.1. *Note that Proposition 3.3.2 does not rule out the existence of observed-data distributions that can only be induced by a single law* Π. *For example, if* $\mathbb{P}(X_{i,1} = 0) = 1$, *then the observed-data law must be induced by the law* Π *that places unit mass on the point-mass distribution concentrated at 0 (i.e., the degenerate distribution at 0). The proposition shows that there exist observed-data distributions that can be induced by multiple distinct laws* Π, *and therefore the barycenter quantiles are not identifiable in general.*

C.1.2 Proof of Proposition 3.3.1

Proof. For each unit i, define $k_i := \lceil m_i\alpha \rceil$. Since $\hat{\nu}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \delta_{X_{i,j}}$, the empirical quantile of $\hat{\nu}_i$ at level α is the k_ith order statistic of the observations from unit i. That is,

$$F^-_{\hat{\nu}_i}(\alpha) = X^{(m_i)}_{i,(k_i)}.$$

Therefore, by the definition of $\hat{q}^{\text{emp}}_\alpha$,

$$\begin{aligned}
\mathbb{E}[\hat{q}^{\text{emp}}_\alpha] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} F^-_{\hat{\nu}_i}(\alpha)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[F^-_{\hat{\nu}_i}(\alpha)\big] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[X^{(m_i)}_{i,(\lceil m_i\alpha\rceil)}\Big].
\end{aligned}$$

Because the units are identically distributed under the sampling scheme (3.1), each term in the sum has the same expectation. Hence

$$\mathbb{E}[\hat{q}^{\text{emp}}_\alpha] = \mathbb{E}\Big[X^{(m_1)}_{1,(\lceil m_1\alpha\rceil)}\Big]$$

$$= \sum_{m'=1}^{\infty} \eta(m')\, \mathbb{E}\Big[X^{(m')}_{1,(\lceil m'\alpha \rceil)} \,\Big|\, m_1 = m'\Big],$$

where $\eta(m') = \mathbb{P}(m_1 = m')$.

Now fix $m' \geq 1$ and set $k := \lceil m'\alpha \rceil$. Conditional on $m_1 = m'$, the two-stage sampling scheme becomes

$$\nu_1 \sim \Pi, \qquad X_{1,1}, \ldots, X_{1,m'} \mid \nu_1 \overset{\text{i.i.d.}}{\sim} \nu_1.$$

Since $m_1 \perp\!\!\!\perp \nu_1$, conditioning on $m_1 = m'$ does not change the law of ν_1. Therefore, for each fixed m', we can use Lemma A.1 in Bigot et al. (2018), which gives

$$\mathbb{E}\Big[X^{(m')}_{1,(k)} \,\Big|\, m_1 = m'\Big] = \mathbb{E}_{\bar{\nu}}\Big[X^{*(m')}_{(k)}\Big],$$

where $X^{*(m')}_{(1)}, \ldots, X^{*(m')}_{(m')}$ are the order statistics of an i.i.d. sample of size m' drawn from the barycenter $\bar{\nu}_\Pi$.

Substituting $k = \lceil m'\alpha \rceil$ into the previous identity and then into the preceding sum yields

$$\begin{aligned}\mathbb{E}[\hat{q}^{\text{emp}}_\alpha] &= \sum_{m'=1}^{\infty} \eta(m')\, \mathbb{E}_{\bar{\nu}}\Big[X^{*(m')}_{(\lceil m'\alpha \rceil)}\Big] \\ &= \mathbb{E}_{m' \sim \eta}\Big[\mathbb{E}_{\bar{\nu}}\Big[X^{*(m')}_{(\lceil m'\alpha \rceil)}\Big]\Big].\end{aligned}$$

This proves the claim. □

C.1.3 Proof of Theorem 3.4.1

Proof. We will show that $G(x, \alpha-) = \mathbb{P}_{\nu \sim \Pi}\{F^-_\nu(\alpha) > x\}$, from which directly follows $\mathbb{P}_{\nu \sim \Pi}\{F^-_\nu(\alpha) \leq x\} = 1 - G(x, \alpha-)$ as required.

Consider the definition of $G(x, \alpha-)$:

$$\begin{aligned}G(x, \alpha-) &= \lim_{t \uparrow \alpha} \mathbb{P}_{\nu \sim \Pi}\{F_\nu(x) \leq t\} \\ &= \mathbb{P}_{\nu \sim \Pi}\{F_\nu(x) < \alpha\},\end{aligned}$$

where the second equality follows because the events $\{F_\nu(x) \leq t\}$ increase to $\{F_\nu(x) < \alpha\}$ as $t \uparrow \alpha$. We now show that $F_\nu(x) < \alpha \iff F_\nu^-(\alpha) > x$, which will imply that $G(x, \alpha-) = \mathbb{P}_{\nu\sim\Pi}\{F_\nu^-(\alpha) > x\}$ and, therefore, will complete the proof.

$F_\nu(x) < \alpha \implies F_\nu^-(\alpha) > x$. Start by assuming that $F_\nu(x) < \alpha$. Because $x \mapsto F_\nu(x)$ is a cumulative distribution function, it is right-continuous, which implies that $\lim_{t\downarrow x} F_\nu(t) = F_\nu(x) < \alpha$ and, hence, there exists a $x^* > x$ such that $F_\nu(x^*) < \alpha$. Because F_ν is a distribution function, it is also non-decreasing, which implies that $x^* \leq \inf\{t : F_\nu(t) \geq \alpha\} = F_\nu^-(\alpha)$. Hence, $x < x^* \leq F_\nu^-(\alpha)$ and thus, as required, $x < F_\nu^-(\alpha)$.

$F_\nu(x) < \alpha \impliedby F_\nu^-(\alpha) > x$. Start by assuming that $F_\nu^-(\alpha) > x$. Now also assume that $F_\nu(x) \geq \alpha$, which implies that $F_\nu^-(\alpha) = \inf\{t : F_\nu(t) \geq \alpha\} \leq x$. This is a contradiction; hence, we must have that $F_\nu(x) < \alpha$. □

C.1.4 Proof of Proposition 3.4.2

Proof. Define $C_i(x) := \mathbf{1}\{F_{\hat{\nu}_i}(x) \geq \alpha\}$. By the same equivalence used in the proof of Theorem 3.4.1, $C_i(x) = \mathbf{1}\{F_{\hat{\nu}_i}^-(\alpha) \leq x\}$. Moreover, $1 - \hat{G}_n^{\text{emp}}(x, \alpha-) = \frac{1}{n}\sum_{i=1}^n C_i(x)$ and therefore

$$\begin{aligned}
\hat{q}_\alpha^{\text{MCB}} &= \int_{\mathcal{I}} x \, d\left(1 - \hat{G}_n^{\text{emp}}(x, \alpha-)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{I}} x \, dC_i(x) \\
&= \frac{1}{n}\sum_{i=1}^n F_{\hat{\nu}_i}^-(\alpha) = \hat{q}_\alpha^{\text{emp}},
\end{aligned}$$

where the third equality follows because C_i is the distribution function of a point mass at $F_{\hat{\nu}_i}^-(\alpha)$. □

C.1.5 Proof of Proposition 3.5.1

Proof. For any $m \in \{1, \dots, C\}$ and $y \in \{0, \dots, m\}$,

$$\mathbb{P}\big(Y_i(x) = y \mid m_i = m\big) = \int_0^1 \binom{m}{y} t^y (1-t)^{m-y} \, dG(x,t).$$

For fixed (m, y), the function $t \mapsto t^y(1-t)^{m-y}$ is a polynomial in t of degree at most m, and hence of degree at most C. Therefore each conditional count probability depends only on the first C moments of $G(x, \cdot)$. Since distinct probability distributions on $[0,1]$ can share the same first C moments, the mixing distribution is not identifiable. □

C.1.6 Proof of Theorem 3.6.1

Proof. Let A_n denote the event that $x \mapsto \hat{H}_{n,\alpha}(x)$ is a cumulative distribution function on $\mathcal{I}$. By Condition (C2), $\mathbb{P}(A_n) \to 1$. On A_n, let $Q_{n,\alpha}$ denote the probability measure with distribution function $\hat{H}_{n,\alpha}$, and let Q_α denote the probability measure with distribution function H_α.

By Condition (C1) and Lemma C.4.5, $Q_{n,\alpha}$ converges weakly in probability to Q_α:

$$Q_{n,\alpha} \overset{P}{\rightsquigarrow} Q_\alpha.$$

If $\mathcal{I}$ is bounded, then $x \mapsto x$ is a bounded continuous function on $\mathcal{I}$, and therefore

$$\int_{\mathcal{I}} x \, d\hat{H}_{n,\alpha}(x) = \int_{\mathcal{I}} x \, dQ_{n,\alpha}(x) \xrightarrow{P} \int_{\mathcal{I}} x \, dQ_\alpha(x) = \int_{\mathcal{I}} x \, dH_\alpha(x).$$

If $\mathcal{I}$ is unbounded but Condition (C3)b holds, then the same conclusion follows from Lemma C.4.8. Hence

$$\hat{q}_\alpha^{\mathrm{MCB}} = \int_{\mathcal{I}} x \, d\hat{H}_{n,\alpha}(x) \xrightarrow{P} \int_{\mathcal{I}} x \, dH_\alpha(x) = q_\alpha^\star.$$

This proves the claim. □

C.1.7 Proof of Corollary 3.6.2

Proof. By the assumptions of the corollary, with probability tending to one, $t \mapsto \hat{q}_t^{\mathrm{MCB}}$ is a quantile function. Let $\hat{\nu}_n$ denote the corresponding probability measure. By Theorem 3.6.1, for every continuity point t of the true barycenter quantile function $t \mapsto F^-_{\bar{\nu}_\Pi}(t)$,

$$F^-_{\hat{\nu}_n}(t) = \hat{q}_t^{\mathrm{MCB}} = q_t^\star + o_P(1) = F^-_{\bar{\nu}_\Pi}(t) + o_P(1).$$

By Lemma C.4.6, this implies that $\hat{\nu}_n$ converges weakly in probability to $\bar{\nu}_\Pi$:

$$\hat{\nu}_n \overset{P}{\rightsquigarrow} \bar{\nu}_\Pi.$$

As explained in the text before Lemma C.4.5 and in the proof of Lemma C.4.5, weak convergence in probability (as defined in Definition C.4.4) is equivalent to the following: for every subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_l)_{l\geq 1}$ such that

$$P\left(\omega : \hat{\nu}^\omega_{m_l} \rightsquigarrow \bar{\nu}_\Pi\right) = 1.$$

By Villani (2009, Theorem 6.9), convergence in the d_{W_2} topology is equivalent to weak convergence together with convergence of second moments. We next argue along subsequences. Hence, it remains to show that, for every subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_l)_{l\geq 1}$ such that

$$\int_{\mathcal{I}} x^2 \, d\hat{\nu}_{m_l}(x) \to \int_{\mathcal{I}} x^2 \, d\bar{\nu}_\Pi(x).$$

If $\mathcal{I}$ is bounded, then $x \mapsto x^2$ is a bounded continuous function on $\mathcal{I}$. Therefore, weak convergence in probability implies

$$\int_{\mathcal{I}} x^2 \, d\hat{\nu}_n(x) = \int_{\mathcal{I}} x^2 \, d\bar{\nu}_\Pi(x) + o_P(1),$$

which implies the desired conclusion along subsequences. If $\mathcal{I}$ is unbounded, the assumed asymptotic uniform integrability with respect to $f : x \mapsto x^2$, together with Lemma C.4.8, gives the same conclusion.

The results above imply that, for every subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_l)_{l\geq 1}$ such that

$$P\left(\omega : \hat{\nu}^{\omega}_{m_l} \rightsquigarrow \bar{\nu}_{\Pi} \text{ and } \int_{\mathcal{I}} x^2 \, d\hat{\nu}^{\omega}_{m_l}(x) \to \int_{\mathcal{I}} x^2 \, d\bar{\nu}_{\Pi}(x)\right) = 1.$$

Villani (2009, Theorem 6.9) now implies that, for every subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_l)_{l\geq 1}$ such that

$$d_{W_2}(\hat{\nu}_{m_l}, \bar{\nu}_{\Pi}) \to 0,$$

which immediately implies that $d_{W_2}(\hat{\nu}_n, \bar{\nu}_{\Pi}) \xrightarrow{P} 0$. □

C.1.8 Proof of Theorem 3.6.3

Proof. Without loss of generality, assume that $\mathcal{I} = \mathbb{R}$. For any distribution function F for which $\int_{\mathcal{I}} |x| \, dF < \infty$, we have that

$$\int_{\mathcal{I}} x \, dF = \int_0^{+\infty} 1 - F(t) \, dt - \int_0^{+\infty} F(-t) \, dt.$$

From the above display and Condition AL2 it follows that, with probability tending to one,

$$\int_{\mathcal{I}} x \, d\hat{H}_{n,\alpha} - \int_{\mathcal{I}} x \, dH_{\alpha} = -\int_0^{+\infty} \{\hat{H}_{n,\alpha}(x) - H_{\alpha}(x)\} \, dx - \int_0^{+\infty} \{\hat{H}_{n,\alpha}(-x) - H_{\alpha}(-x)\} \, dx.$$

Multiplying both sides by $\sqrt{n}$ and using Condition AL1, we obtain

$$\sqrt{n}\left(\int x \, d\hat{H}_{n,\alpha} - \int x \, dH_{\alpha}\right) = -\int_0^{+\infty} \left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\alpha}(O_i; x) + r_{\alpha,n}(x)\right\} dx$$

$$
\begin{aligned}
&- \int_0^{+\infty} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\alpha(O_i; -x) + r_{\alpha,n}(-x) \right\} dx \\
&= -\frac{1}{\sqrt{n}} \int_{-\infty}^{+\infty} \sum_{i=1}^{n} \psi_\alpha(O_i; x)\, dx - \int_{-\infty}^{+\infty} r_{\alpha,n}(x)\, dx \\
&= -\frac{1}{\sqrt{n}} \int_{-\infty}^{+\infty} \sum_{i=1}^{n} \psi_\alpha(O_i; x)\, dx + o_P(1).
\end{aligned}
$$

The second equality follows from the integrability conditions on $x \mapsto r_{\alpha,n}(x)$ and $x \mapsto \psi_\alpha(O; x)$ where $x \mapsto \psi_\alpha(O; x)$ is integrable with probability one because $E\left[\int_{-\infty}^{+\infty} |\psi_\alpha(O; x)|\, dx\right] < \infty$ by Condition AL1. The final equality follows because $\int_{\mathcal{I}} |r_{\alpha,n}(x)|\, dx = o_P(1)$. Moving the sum outside the integral gives

$$
\begin{aligned}
\sqrt{n}\left(\int x\, d\hat{H}_{n,\alpha} - \int x\, dH_\alpha \right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ - \int_{-\infty}^{+\infty} \psi_\alpha(O_i; x)\, dx \right\} + o_P(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_\alpha(O_i) + o_P(1).
\end{aligned}
$$

Condition AL1 also requires that $E\left[\int_{-\infty}^{+\infty} |\psi_\alpha(O; x)|\, dx\right] < \infty$, which allows us to apply Fubini's theorem to obtain $E\left[\tilde{\psi}_\alpha(O)\right] = 0$ from $E[\psi_\alpha(O; x)] = 0$. The asymptotic normality now follows from Condition AL3 and the central limit theorem. □

C.1.9 Proof of Corollary 3.6.4

Proof. This is a standard application of uniform asymptotic linearity and weak convergence of the empirical process indexed by the influence functions.

Uniform asymptotic linearity. By condition (i) of the corollary, Theorem 3.6.3 applies uniformly to all $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. Specifically, the following asymptotic linearity representation holds for all $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ with probability tending to one:

$$
\sqrt{n}\big(\hat{q}_\alpha^{\mathrm{MCB}} - q_\alpha^\star\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_\alpha(O_i) + \tilde{r}_{\alpha,n}
$$

where $\tilde{r}_{\alpha,n} := -\int_{\mathcal{I}} r_{\alpha,n}(x)\,dx$ with $\sup_{\alpha\in[\alpha_{\min},\alpha_{\max}]} |\tilde{r}_{\alpha,n}| = o_P(1)$ by condition (ii).

Weak convergence of empirical process. By condition (iii), the class $\{\tilde{\psi}_\alpha : \alpha \in [\alpha_{\min}, \alpha_{\max}]\}$ is P-Donsker, which implies that

$$\mathbb{G}_n(\tilde{\psi}_\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[\tilde{\psi}_\alpha(O_i) - \mathbb{E}[\tilde{\psi}_\alpha(O)]\Big]$$

converges weakly in $\ell^\infty([\alpha_{\min}, \alpha_{\max}])$ to a centered Gaussian process $\mathbb{G}$ with covariance function

$$\Gamma(\alpha,\alpha') := \mathbb{E}\left[\mathbb{G}(\tilde{\psi}_\alpha)\mathbb{G}(\tilde{\psi}_{\alpha'})\right] = \mathbb{E}\left[\tilde{\psi}_\alpha(O)\,\tilde{\psi}_{\alpha'}(O)\right] - \mathbb{E}\left[\tilde{\psi}_\alpha(O)\right]\mathbb{E}\left[\tilde{\psi}_{\alpha'}(O)\right] = \mathbb{E}\left[\tilde{\psi}_\alpha(O)\,\tilde{\psi}_{\alpha'}(O)\right],$$

where the last equality follows from the mean-zero property of the influence function $\tilde{\psi}_\alpha$.

Conclusion. From the above two steps, we have that,

$$\sqrt{n}\big(\hat{q}_\alpha^{\mathrm{MCB}} - q_\alpha^\star\big) = \mathbb{G}_n(\tilde{\psi}_\alpha) + o_P(1),$$

where (i) $\mathbb{G}_n(\tilde{\psi}_\alpha)_{\alpha\in[\alpha_{\min},\alpha_{\max}]}$ converges weakly in $\ell^\infty([\alpha_{\min}, \alpha_{\max}])$ to the Gaussian process $\mathbb{G}$ with covariance $\Gamma(\alpha,\alpha')$ and (ii) the remainder term is $o_P(1)$ uniformly in $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. Hence, by Slutsky's theorem (van der Vaart, 2000, Theorem 18.10(iv)), we have that $\sqrt{n}\big(\hat{q}_\alpha^{\mathrm{MCB}} - q_\alpha^\star\big)$ converges weakly in $\ell^\infty([\alpha_{\min}, \alpha_{\max}])$ to the same Gaussian process $\mathbb{G}$. □

C.2 Additional figures and tables

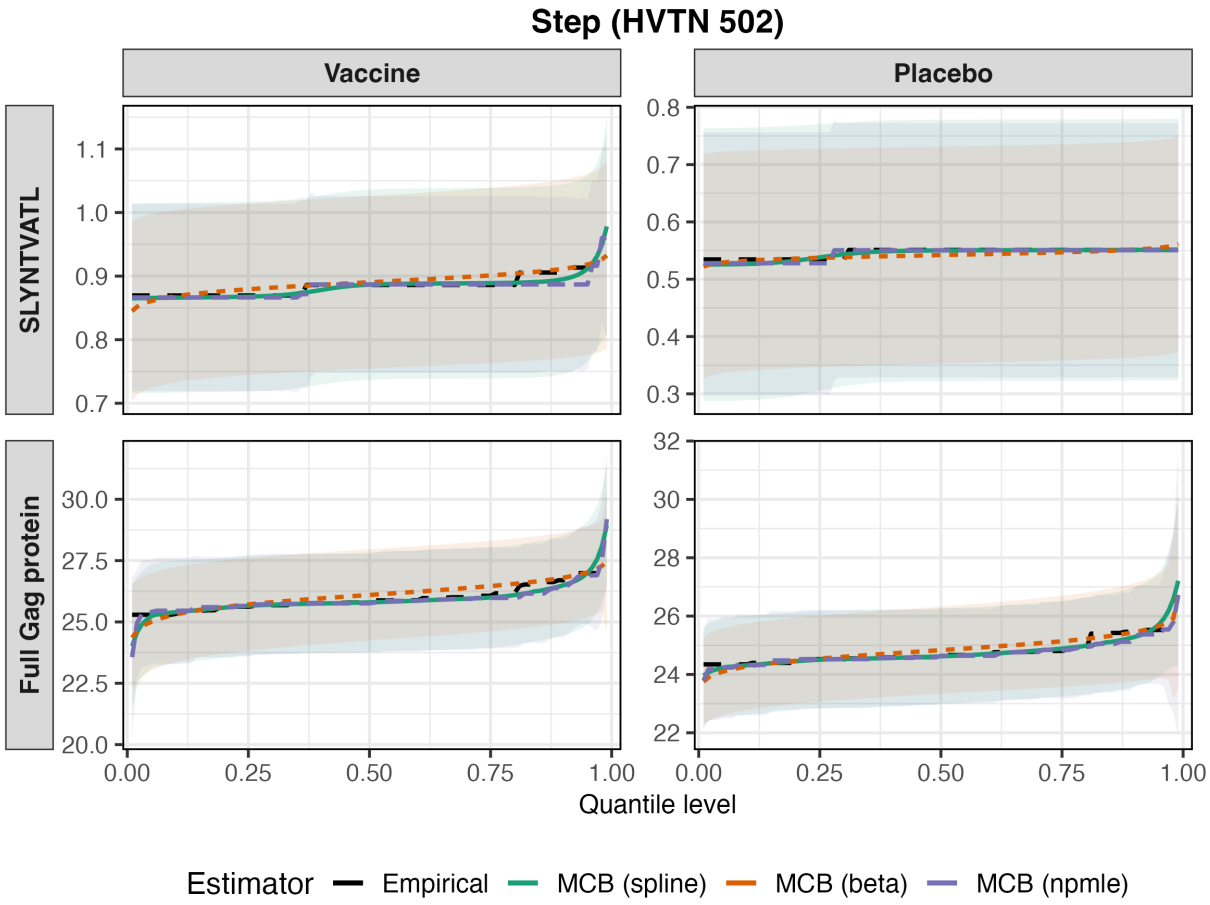


Figure C.1: Estimated barycenter quantile functions, comparing the empirical barycenter, MCB (spline), MCB (beta), and MCB (npmle), for the SLYNTVATL epitope distance and full Gag protein physicochemical-weighted Hamming distance in the Step trial, stratified by treatment arm.

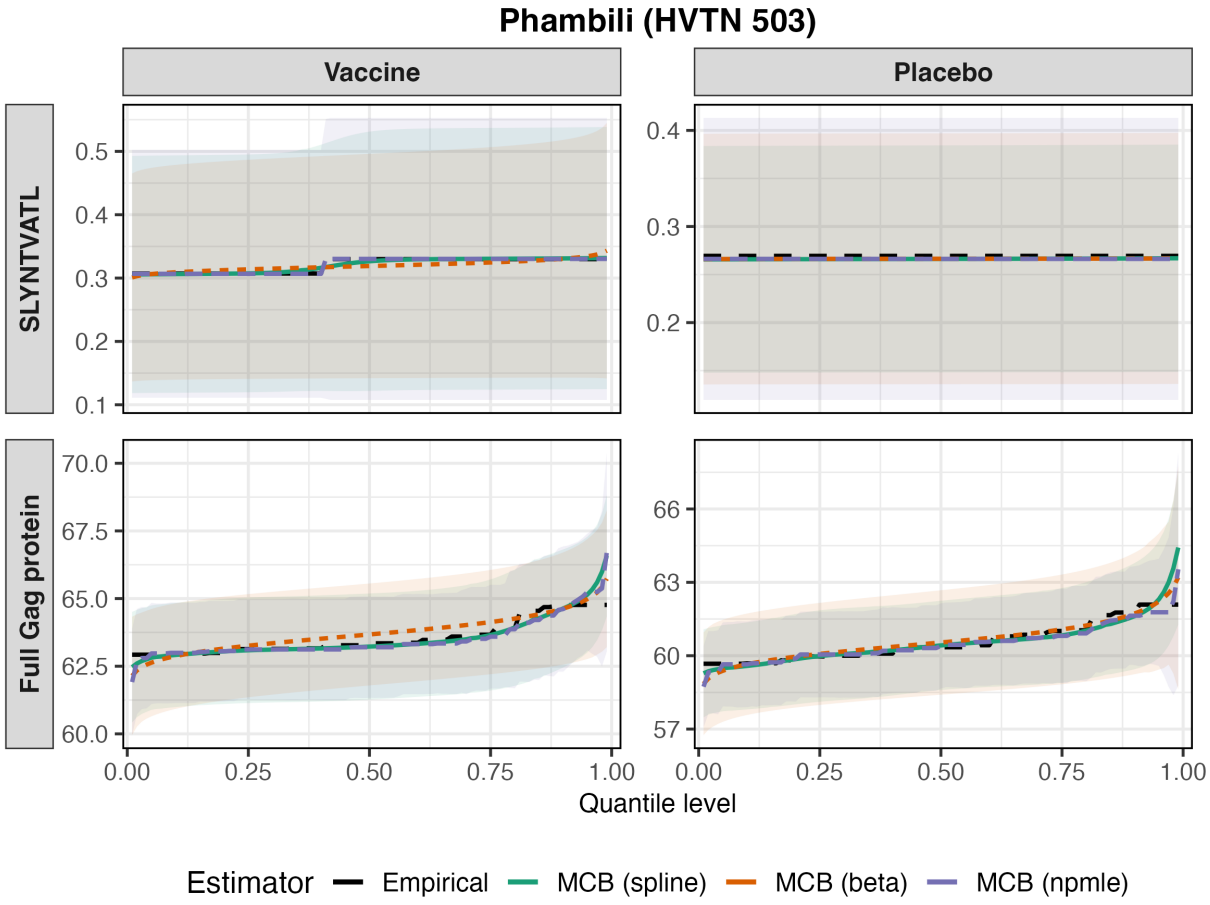


Figure C.2: Estimated barycenter quantile functions, comparing the empirical barycenter, MCB (spline), MCB (beta), and MCB (npmle), for the SLYNTVATL epitope distance and full Gag protein physicochemical-weighted Hamming distance in the Phambili trial, stratified by treatment arm.

C.3 Additional data applications

C.3.1 Public versus Catholic schools

Data: We use the High School and Beyond dataset as presented in the Raudenbush and Bryk (2002) hierarchical data analysis textbook. The data consist of a nationally representative sample of U.S. public and Catholic schools from the 1982 survey. The dataset includes 7,185 students nested within 160 schools, consisting of 90 public schools and 70 Catholic schools. For each school, the data include an indicator of school type (public or Catholic) as well as the school's mean student socioeconomic status. At the student level, the dataset includes math achievement scores and an indicator of whether the student identifies as a minority. Figure C.3 summarizes the school-level sample sizes by school type.

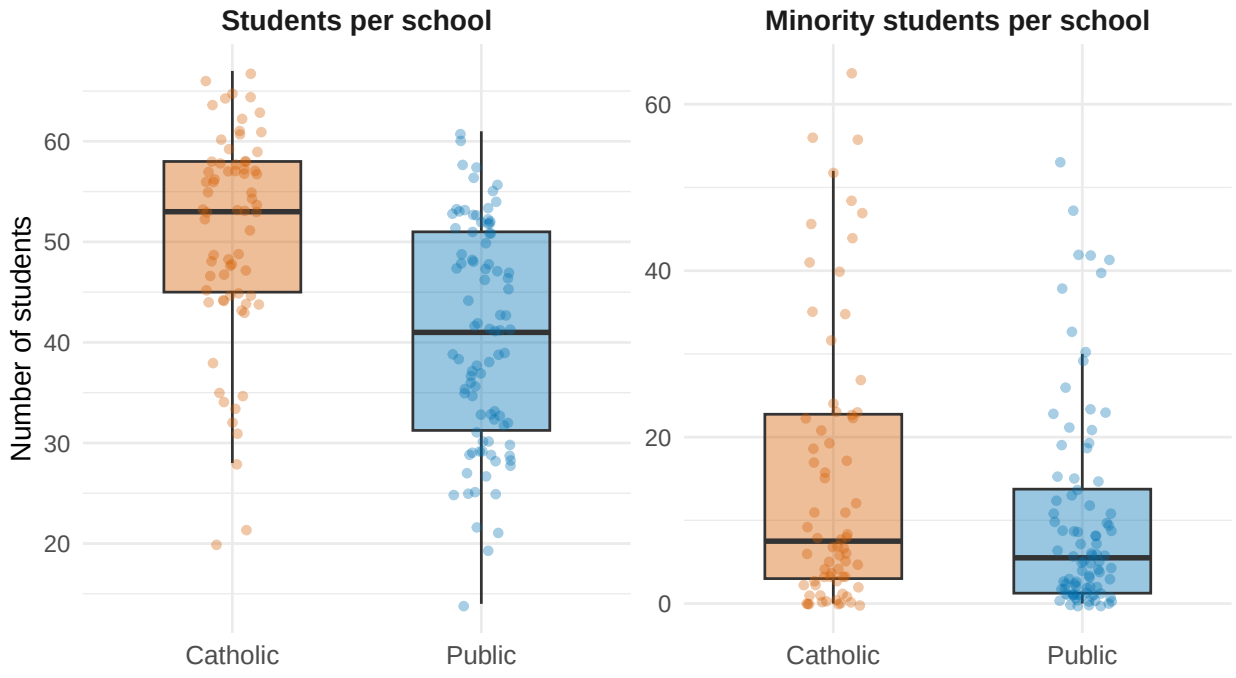


Figure C.3: Distribution of the number of students sampled per school, overall and among minority students, stratified by school type. Points represent individual schools.

Question of interest: How does the barycenter distribution of math scores for public schools compare with that for Catholic schools?

Analysis: Using the MCB estimator, we estimate the Wasserstein barycenter of the distribution of math scores for Catholic schools and compare it with that of public schools. Differences between the two groups are assessed by examining the difference in their Wasserstein barycenter quantile functions. We then repeat this analysis subsetting for minority students. Our primary MCB specification uses splines to estimate the mixing distribution ($df = 5$); for comparison, we also report results from the Beta-configured MCB estimator and the empirical barycenter.

Results: Results are shown in Figure C.4. For the full student sample, the estimated barycenter quantile function for Catholic schools lies above that for public schools for all quantile levels. The quantile difference plot suggests that the Catholic-public gap is concentrated among lower-to-middle achieving students: for example, the 25th percentile of the Catholic-school distribution is roughly 4 points higher than the corresponding percentile for public schools, whereas the difference is minimal among students near the 99th percentile. The MCB spline and Beta estimators produce very similar results, and both align closely with the empirical barycenter.

Among minority students, the Catholic-public difference is larger across nearly all quantile levels. Using the spline-based MCB estimator, the differences are most pronounced in the middle of the distribution, reaching their largest values around the 0.50-0.60 quantiles before tapering off toward the upper tail. The spline and Beta MCB estimators generally closely agree with one another, but both differ from the empirical barycenter, particularly in the upper tail where the MCB estimates for the difference-in-quantiles are lower.

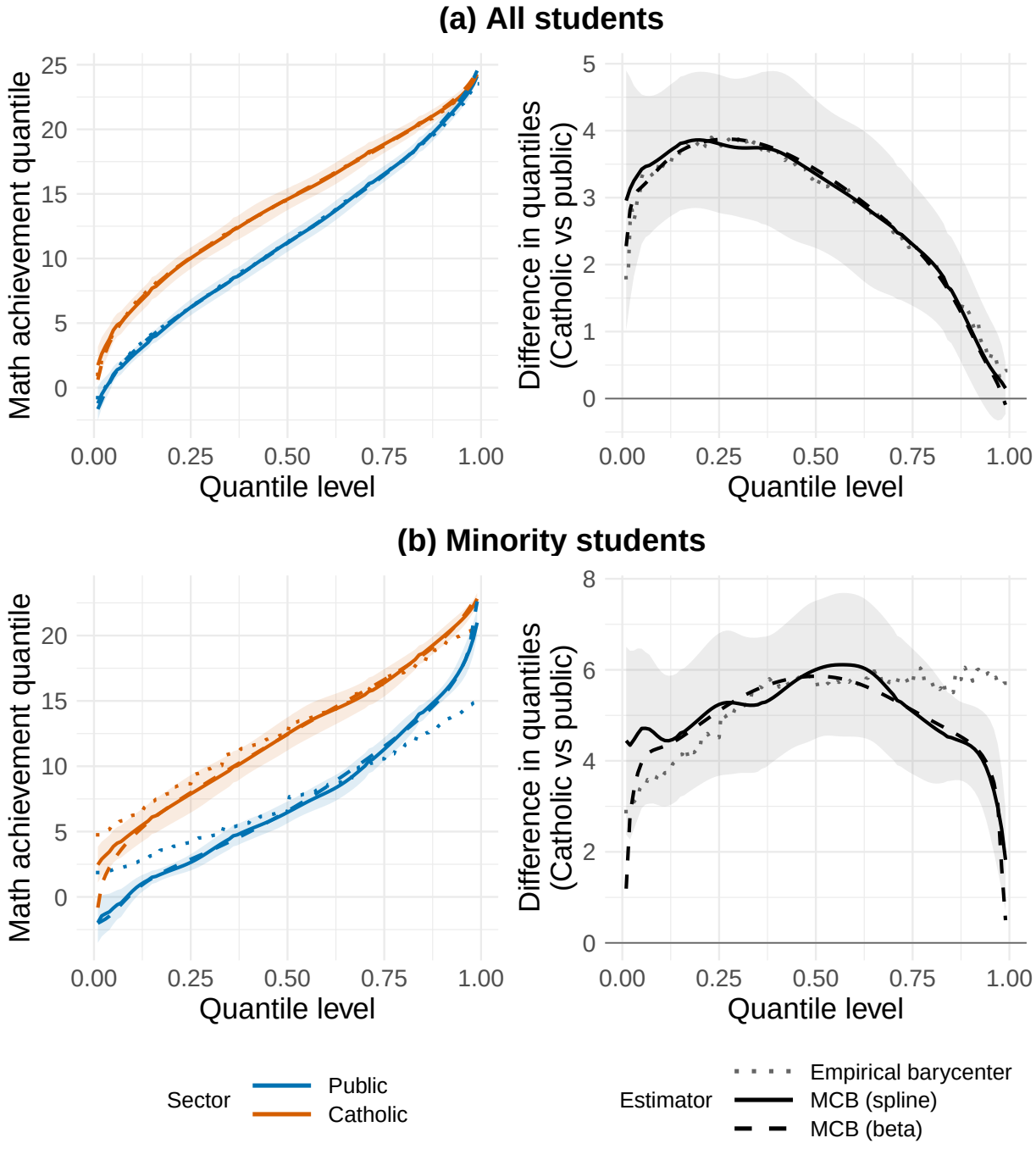


Figure C.4: Estimated Wasserstein barycenters of math achievement score distributions for public and Catholic schools, shown for all students and for minority students. Left panels display the estimated quantile functions by school type, while right panels show the difference in quantile functions comparing Catholic to public schools. Results are shown for the MCB estimators (spline and Beta configurations) and the empirical barycenter. The shaded areas represent pointwise confidence intervals based on the spline-configured MCB estimator.

C.3.2 Ticks on grouse

Data: To study parasite aggregation among hosts, Elston et al. (2001) collected data on sheep tick burdens among red grouse chicks sampled in the Glas Choille study area in Scotland. The sample includes 403 chicks from 118 broods observed across three years: 1995, 1996, and 1997, with a median of 3 chicks per brood (range: 1-10). For each chick, the response variable is the number of sheep ticks counted, with counts taken from the chick's head. Brood-level information includes the year of observation and altitude.

Question of interest: Does the barycenter distribution of tick counts differ between broods captured at lower versus higher altitudes?

Analysis: We first divide the broods into two groups based on altitude, classifying broods as low-altitude or high-altitude depending on whether their capture altitude is below or above the sample mean. Using the MCB estimator, we estimate the barycenter of the distribution of tick counts for broods in the high-altitude group and compare it with that of broods in the low-altitude group. Differences between the two groups are assessed by examining the difference in their estimated quantile functions. Our primary MCB specification uses splines to estimate the mixing distribution ($df = 5$); for comparison, we also report results from the Beta-configured MCB estimator and the empirical barycenter.

Results: Results are shown in Figure C.5. The estimated barycenters suggest substantially higher tick burdens among broods captured at lower altitudes compared with those captured at higher altitudes. The quantile difference plot shows that this difference increases across quantile levels, indicating that the altitude-related difference is most pronounced among chicks with the highest tick counts. The spline and Beta MCB estimators closely agree with one another. Compared with the empirical barycenter, both MCB estimators produce smaller estimated differences in the lower tail and larger estimated differences in the upper tail.

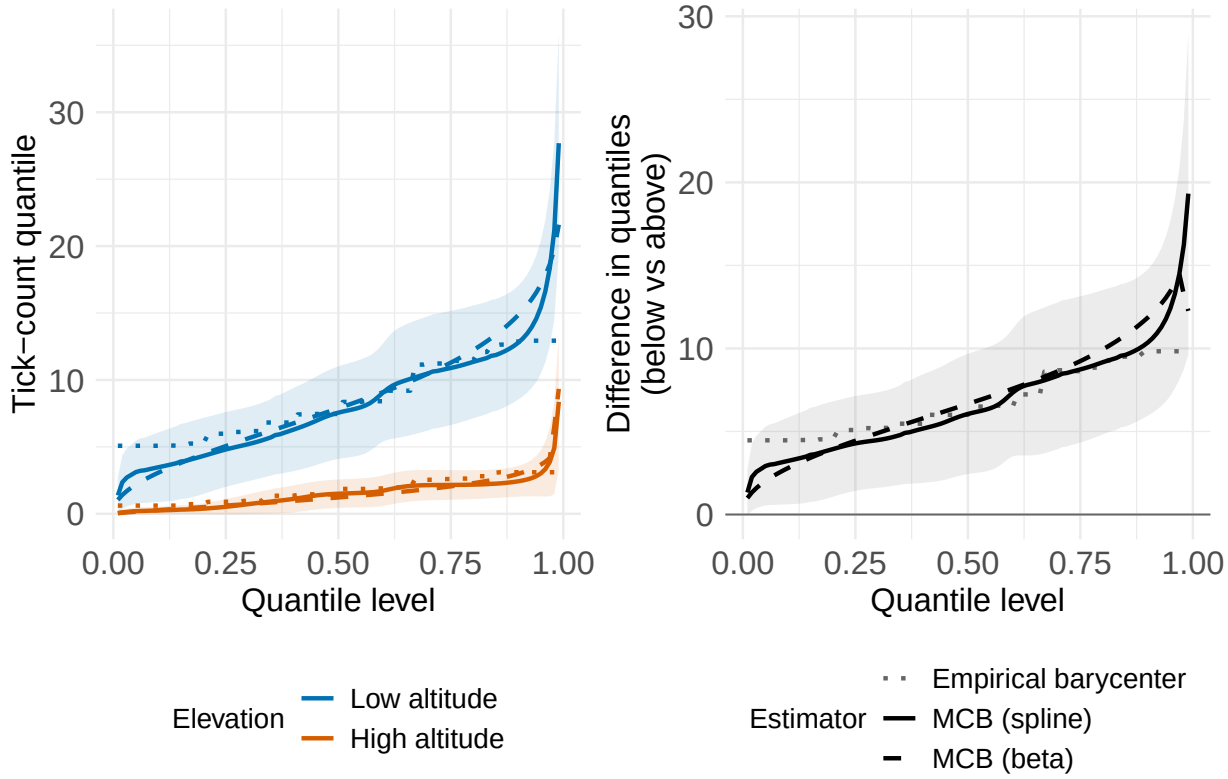


Figure C.5: Estimated Wasserstein barycenters of tick-count distributions for broods captured at low and high altitudes. The left panel displays the estimated quantile functions by altitude group, while the right panel shows the difference in quantile functions comparing low-altitude broods to high-altitude broods. Results are shown for the MCB estimators (spline and Beta configurations) and the empirical barycenter. The shaded areas represent pointwise confidence intervals based on the spline-configured MCB estimator.

C.3.3 Radon exposure in Minnesota

Background To study variation in indoor radon concentrations, Price et al. (1996) analyzed radon measurements from households across Minnesota counties. The radon dataset consists of 919 households from 85 counties, with a median of four household measurements per county (range: 1–100) (Gelman, 2007). For each household, the response variable is the log radon measurement in log picoCuries per liter, and household-level information includes whether the measurement was taken in the basement or on the first floor. County-level information includes average soil uranium content.

Question of interest: Is the barycenter of the household-level radon distribution within a county associated with the county's soil uranium concentration?

Analysis: We estimate the conditional barycenter of radon measurements given a county's mean log-uranium level using the kernel-regression conditional barycenter estimator described in Example 3.7.1. Our primary MCB specification uses splines to estimate the mixing distribution ($df = 5$); for comparison, we also report results from the empirical barycenter.

Results: Figure C.6 displays the estimated conditional barycenter of household log-radon distributions at five county log-uranium levels. The estimated barycenter shifts upward as county log-uranium increases, indicating higher household radon levels in counties with greater soil uranium content.

C.4 Technical tools for weak convergence

This section develops the technical tools used in the proofs of the consistency results in Theorem 3.6.1 and Corollary 3.6.2. Since our estimators are naturally defined through estimated distribution functions or quantile functions, we need a way to translate pointwise convergence statements into convergence of the induced random probability measures. In section C.4.1, we review deterministic results showing that, for monotone functions such as

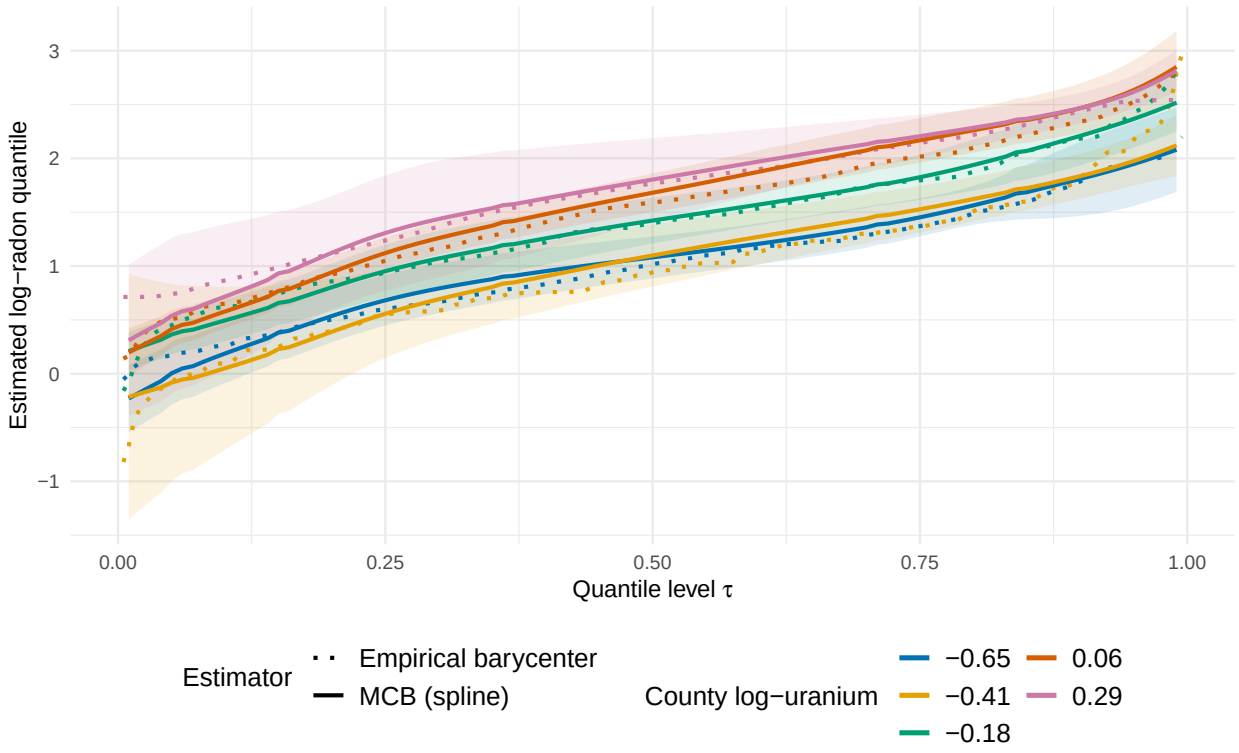


Figure C.6: Estimated conditional barycenters of household log-radon distributions across county log-uranium levels. The curves show estimated log-radon quantile functions at five fixed values of county log-uranium. Results are shown for the empirical barycenter and for the MCB estimator with a spline configuration. Shaded bands denote pointwise confidence intervals for the spline-configured MCB estimator.

CDFs and quantile functions, convergence on a countable dense set of continuity points is enough to imply convergence at all continuity points. Then, in section C.4.2, we extend these ideas to random probability measures and define weak convergence in probability. The main consequence is that pointwise convergence in probability of either the CDFs or the quantile functions at continuity points is equivalent to weak convergence in probability of the associated measures. Finally, we state a uniform integrability condition under which weak convergence in probability also implies convergence of unbounded moments.

C.4.1 Deterministic sequences of distribution functions and quantile functions

Lemma C.4.1. *Let $(F_n)_{n\geq 1}$ be a deterministic sequence of cumulative distribution functions on $\mathbb{R}$ and F a distribution function on $\mathbb{R}$. There exists a countable set of continuity points of $x \mapsto F(x)$ that is dense in $\mathbb{R}$, which we denote by $\mathcal{D} \subset \mathbb{R}$. If $F_n(x) \to F(x)$ for every $x \in \mathcal{D}$, then $F_n(x) \to F(x)$ for every continuity point of $x \mapsto F(x)$.*

Proof. **Part 1: Existence of $\mathcal{D}$.** Note that the set of discontinuity points of $x \mapsto F(x)$, denoted by C, is at most countable because F is non-decreasing càdlàg and bounded. If we remove these points from $\mathbb{R}$, the remaining set of points $\mathbb{R} \setminus C$ is dense in $\mathbb{R}$.[1] A countable subset of $\mathbb{R} \setminus C$ can now be constructed as follows: Let $\mathcal{B}$ be the collection of open intervals with rational endpoints. This set contains a countable number of intervals, denoted by $(B_n)_{n\geq 1}$. Since $\mathbb{R} \setminus C$ is dense, $B_n \cap (\mathbb{R} \setminus C)$ is non-empty for every $n \in \mathbb{N}$. For every $n \in \mathbb{N}$, let d_n be an element of $B_n \cap (\mathbb{R} \setminus C)$. The set $\{d_n : n \geq 1\}$ is countable and dense in $\mathbb{R}$.

Part 2: Pointwise convergence at all continuity points. Let x be a continuity point of $x \mapsto F(x)$. Hence, for every $\epsilon > 0$, there exists a $\delta > 0$ such that $|F(y) - F(x)| < \epsilon$ for every $y \in B_\delta(x)$. Now choose $l < x$ and $u > x$ such that $u, l \in \mathcal{D} \cap B_\delta(x)$ for $\mathcal{D}$ the set mentioned in the first part of the lemma.

[1] Assume for contradiction that $\mathbb{R} \setminus C$ is not dense in $\mathbb{R}$. Then there exists a $x \in C$ and $\delta > 0$ such that $B_\delta(x) \cap (\mathbb{R} \setminus C) = \emptyset$. This implies that $B_\delta(x) \subset C$, which is a contradiction because C is countable and thus cannot contain a non-empty open interval.

Because F_n is non-decreasing for all $n \geq 1$, we have that

$$F_n(l) \leq F_n(x) \leq F_n(u).$$

We consequently have that

$$F(l) = \lim_{n\to\infty} F_n(l) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq \lim_{n\to\infty} F_n(u) = F(u),$$

where the equalities follow from $u, l \in \mathcal{D}$. By definition of u, l, we have that $|F(l) - F(u)| < \epsilon$. Hence, from the above display follows that

$$\left|\limsup_{n\to\infty} F_n(x) - \liminf_{n\to\infty} F_n(x)\right| \leq |F(u) - F(l)| < \epsilon.$$

Because ϵ was arbitrary, this implies that $\liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = \lim_{n\to\infty} F_n(x) = F(x)$. □

Lemma C.4.2. *Let $(Q_n)_{n\geq 1}$ be a deterministic sequence of quantile functions and Q a quantile function. There exists a countable set of continuity points of $t \mapsto Q(t)$ that is dense in $(0,1)$, which we denote by $\mathcal{D} \subset (0,1)$. If $Q_n(t) \to Q(t)$ for every $t \in \mathcal{D}$, then this implies that $Q_n(t) \to Q(t)$ for every continuity point of $t \mapsto Q(t)$.*

Proof. The proof has the same structure as the proof of Lemma C.4.1.

Part 1: Existence of $\mathcal{D}$. The set of discontinuity points of $t \mapsto Q(t)$, denoted by C, is countable because Q is a non-decreasing function on the unit interval. If we remove these points from $(0,1)$, the remaining set of points $(0,1) \setminus C$ is dense in $(0,1)$. A countable subset of $(0,1) \setminus C$ can now be constructed as follows: Let $\mathcal{B}$ be the collection of open intervals with rational endpoints in $(0,1)$. This set contains a countable number of intervals, denoted by $(B_n)_{n\geq 1}$. Since $(0,1) \setminus C$ is dense, $B_n \cap ((0,1) \setminus C)$ is non-empty for every $n \in \mathbb{N}$. For every $n \in \mathbb{N}$, let d_n be an element of $B_n \cap ((0,1) \setminus C)$. The set $\{d_n : n \geq 1\}$ is countable and dense in $(0,1)$.

Part 2: Pointwise convergence at all continuity points Let t be a continuity point of $t \mapsto Q(t)$. Hence, for every $\epsilon > 0$, there exists a $\delta > 0$ such that $|Q(y) - Q(t)| < \epsilon$ for every $y \in B_\delta(t)$. Now choose $l < t$ and $u > t$ such that $u, l \in \mathcal{D} \cap B_\delta(t)$ for $\mathcal{D}$ the set mentioned in the first part of the lemma.

Because Q_n is non-decreasing for all $n \geq 1$, we have that

$$Q_n(l) \leq Q_n(t) \leq Q_n(u).$$

We consequently have that

$$Q(l) = \lim_{n\to\infty} Q_n(l) \leq \liminf_{n\to\infty} Q_n(t) \leq \limsup_{n\to\infty} Q_n(t) \leq \lim_{n\to\infty} Q_n(u) = Q(u),$$

where the equalities follow from $u, l \in \mathcal{D}$. By the definition of u, l, we have that $|Q(l) - Q(u)| < \epsilon$. Hence, from the above display follows that

$$\left|\limsup_{n\to\infty} Q_n(t) - \liminf_{n\to\infty} Q_n(t)\right| \leq |Q(u) - Q(l)| < \epsilon.$$

Because ϵ was arbitrary, this implies that $\liminf_{n\to\infty} Q_n(t) = \limsup_{n\to\infty} Q_n(t) = \lim_{n\to\infty} Q_n(t) = Q(t)$. □

C.4.2 Random probability measures and weak convergence in probability

In what follows, we will focus on weak convergence in probability. We start by summarizing weak convergence of a deterministic sequence of probability measures and then extend this to weak convergence in probability of a sequence of random probability measures. The latter will largely follow Schmon et al. (2021); although, we present some novel technical results[2].

[2]We have not found these statements and proofs in the literature. However, we do not claim that we are the first to present these results, especially since these results can be proven along the same lines as some well-known results.

Weak convergence of deterministic sequences of measures

Let $(\mu_n)_{n\geq 1}$ be a deterministic sequence of probability measures on $\mathcal{I} \subset \mathbb{R}$, and let F_{μ_n} be the corresponding cumulative distribution functions (CDFs). One definition of weak convergence to μ, with CDF F_μ, is as follows:

$$F_{\mu_n}(x) \to F_\mu(x) \quad \text{for all continuity points of } x \mapsto F_\mu(x). \tag{C.1}$$

By the Portmanteau Lemma, this is equivalent to the following definition:

$$\mathbb{E}_{\mu_n} f(X) \to \mathbb{E}_\mu f(X) \quad \forall\ f \in C_b(\mathcal{I}),$$

where $C_b(\mathcal{I})$ is the set of bounded continuous functions on $\mathcal{I}$.

Random probability measures

Before we define weak convergence in probability of a sequence of random probability measures, we more rigorously define random probability measures. We restate the definition of Schmon et al. (2021), but specialized to random probability measures on subsets of the real line, denoted by $\mathcal{I} \subset \mathbb{R}$ for $\mathcal{B}(\mathcal{I})$ the corresponding Borel σ-algebra. Further let $\mathcal{P}(\mathcal{I})$ be the set of probability measures on the measurable space $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$.

Definition C.4.3 (Random probability measure)**.** A random probability measure is a map $\nu : \Omega \times \mathcal{B}(\mathcal{I}) \to [0,1]$ such that for every $B \in \mathcal{B}(\mathcal{I})$ the map $\omega \mapsto \nu(\omega, B) = \nu^\omega(B)$ is measurable while $\nu(\omega, \cdot) \in \mathcal{P}(\mathcal{I})$ for almost every $\omega \in \Omega$.

By Crauel (2002, Proposition 3.3), we have that $\omega \mapsto \mathbb{E}_{\nu^\omega} f := \int_{\mathcal{I}} f \, d\nu^\omega$ for $f \in C_b(\mathcal{I})$ is measurable if ν is a random probability measure.

Weak convergence in probability

Weak convergence in probability is defined below for random probability measures on (a subset of) the real line. This can easily be extended to more general probability measures, but this is not required for our purposes.

Definition C.4.4 (Weak convergence in probability)**.** Let $\mathcal{P}(\mathcal{I})$ be the set of all probability measures on $(\mathcal{I}, \mathcal{B}(\mathcal{I}))$ and let $(\nu_n)_{n\geq 1}$ be a sequence of random probability measures in $\mathcal{P}(\mathcal{I})$. This sequence converges weakly in probability to $\mu \in \mathcal{P}(\mathcal{I})$ if

$$\mathbb{E}_{\nu_n} f(X) = \mathbb{E}_{\mu} f(x) + o_P(1) \quad \forall\ f \in C_b(\mathcal{I}),$$

where $C_b(\mathcal{I})$ is the set of bounded continuous functions on $\mathcal{I}$.

Note that, in the above definition, $\mathbb{E}_{\nu_n} f(X)$ is a random variable because $\omega \mapsto \mathbb{E}_{\nu_n^\omega} f(X)$ is measurable (which follows from ν_n being a random probability measure, see previous subsection).

Definition C.4.4 is different from the definition of weak convergence in probability in Schmon et al. (2021). However, Schmon et al. (2021, Theorem 1 in Supplementary Material) show that their definition is equivalent to ours. The following lemma is a useful extension of the Portmanteau lemma, which complements Schmon et al. (2021, Theorem 1 in Supplementary Material).

Lemma C.4.5. *Let $(\nu_n)_{n\geq 1}$ be a sequence of random probability measures and $(F_{\nu_n})_{n\geq 1}$ the corresponding random cumulative distribution functions. Further let μ and F_μ be a probability measure and corresponding cumulative distribution function on $\mathcal{I}$. Then, $F_{\nu_n}(x) = F_\mu(x) + o_P(1)$ for every x a continuity point of $x \mapsto F_\mu(x)$ if and only if ν_n converges weakly in probability to μ.*

Proof. $\Rightarrow$**.** The difficulty of the proof in this direction is that the $o_P(1)$ term in $F_{\nu_n}(x) = F_\mu(x) + o_P(1)$ only holds pointwise. In other words, $\|F_{\nu_n} - F_\mu\|_\infty$ may not converge in

probability to zero. This is addressed by considering a countable dense subset of continuity points and by arguing along subsequences.

By definition of random probability measures, $\omega \mapsto F_{\nu_n^\omega}(x) := \nu_n^\omega((-\infty, x])$ is measurable for any $x \in \mathcal{I}$. Hence, $F_{\nu_n}(x)$ is a random variable for any $x \in \mathcal{I}$. For ease of notation, we further let $F_n^\omega := F_{\nu_n^\omega}$ and $F_n := F_{\nu_n}$.

By Lemma C.4.1, there exists a countable subset of continuity points of $x \mapsto F_\mu(x)$ that is dense in $\mathbb{R}$, which we denote by $\mathcal{D}$. Order these points in the sequence $(x_n)_{n\geq 1}$. Now consider an arbitrary subsequence $(n_k)_{k\geq 1}$ of $(n)_{n\geq 1}$.

Let x_1 be the first element of $(x_n)_{n\geq 1}$. Because $x_1 \in \mathcal{D}$, it is a continuity point of $x \mapsto F_\mu(x)$ and, consequently, we have that $F_n(x_1) = F(x_1) + o_P(1)$. Hence, the subsequence $(n_k)_{k\geq 1}$ has a further subsequence, denoted by $(n_l^1)_{l\geq 1}$, such that $F_{n_l^1}(x_1) \to F(x_1)$ P-a.s. Or equivalently, $P(A_1) = 1$ for $A_1 := \left\{\omega \in \Omega : F_{n_l^1}^\omega(x_1) \to F_\mu(x_1)\right\}$.

We can repeat the above construction for every element of $(x_n)_{n\geq 1}$, where the subsequence $(n_l^2)_{l\geq 1}$ is a subsequence of $(n_l^1)_{l\geq 1}$, the subsequence $(n_l^3)_{l\geq 1}$ is a subsequence of $(n_l^2)_{l\geq 1}$, and so on. We now define the subsequence $(m_l)_{l\geq 1}$ element-wise as $m_l = n_l^l$. By construction, for any $k \in \mathbb{N}$ and $\omega \in \Omega$, we have that $F_{n_l^k}^\omega(x_k) \to F_\mu(x_k) \implies F_{m_l}^\omega(x_k) \to F_\mu(x_k)$ because $(m_l)_{l\geq 1}$ is eventually a subsequence of $(n_l^k)_{l\geq 1}$. This implies that

$$\left\{\omega \in \Omega : F_{n_l^k}^\omega(x_k) \to F_\mu(x_k)\right\} =: A_k \subset \tilde{A}_k := \left\{\omega \in \Omega : F_{m_l}^\omega(x_k) \to F_\mu(x_k)\right\}.$$

Because $P(A_k) = 1$ for every $k \in \mathbb{N}$, we have that $P\left(\bigcap_{k\geq 1} A_k\right) = 1$ and, because $A_k \subset \tilde{A}_k$ for every k, also that $P\left(\bigcap_{k\geq 1} \tilde{A}_k\right) = 1$.

For any $\omega \in \bigcap_{k\geq 1} \tilde{A}_k$, we have that $F_{m_l}^\omega(x) \to F_\mu(x)$ for all $x \in \mathcal{D}$. By Lemma C.4.1, this implies that $F_{m_l}^\omega(x) \to F_\mu(x)$ for all continuity points. By the Portmanteau lemma, this implies that $F_{m_l}^\omega$ converges weakly to F_μ, or equivalently, that $\nu_{m_l}^\omega$ converges weakly to μ.

Note that the initial subsequence $(n_k)_{k\geq 1}$ was arbitrary. Hence, we have that for every

subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_l)_{l\geq 1}$ such that

$$P\left(\omega : \nu^{\omega}_{m_l} \rightsquigarrow \mu\right) = 1.$$

This is the definition of weak convergence in probability given in Schmon et al. (2021, Definition 3)[3].

$\Leftarrow$**.** By the definition of weak convergence in probability, for every subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(n_l)_{l\geq 1}$ such that

$$P\left(\omega : \nu^{\omega}_{n_l} \rightsquigarrow \mu\right) = 1.$$

For all ω in the above set, we have that $F_{\nu^{\omega}_{n_l}}(x) \to F_{\mu}(x)$ for every x a continuity point. Because the initial subsequence was arbitrary, this implies that $F_{\nu_n}(x) = F_{\mu}(x) + o_P(1)$ for every x a continuity point. □

The following lemma resembles the previous one, but considers pointwise convergence of the quantile functions (instead of the distribution functions) of a sequence of random probability measures. The structure of the proof below also resembles the proof of Lemma C.4.5.

Lemma C.4.6 (Relation between quantile and cdf estimators)**.** *Let $(\nu_n)_{n\geq 1}$ be a sequence of random probability measures and μ a probability measure on $\mathcal{I} \subset \mathbb{R}$. Denote the corresponding quantile functions by $F^{-}_{\nu_n}$ and F^{-}_{μ}, respectively. We then have that $F^{-}_{\nu_n}(t) = F^{-}_{\mu}(t) + o_P(1)$ for every t a continuity point of $t \mapsto F^{-}_{\mu}(t)$ if and only if ν_n converges weakly in probability to μ.*

Proof. $\Rightarrow$**.** For ease of notation, we further let $Q^{\omega}_n := F^{-}_{\nu^{\omega}_n}$, $F_n := F_{\nu_n}$, and $Q = F^{-}_{\mu}$.

By Lemma C.4.2, there exists a countable set of continuity points of $t \mapsto Q(t)$ that is dense in $(0, 1)$, which we denote by $\mathcal{D}$. Order these points in the sequence $(t_n)_{n\geq 1}$. Now

[3]As explained in the text before Lemma C.4.5, this is equivalent to our Definition C.4.4.

consider an arbitrary subsequence $(n_k)_{k\geq 1}$ of $(n)_{n\geq 1}$.

Let t_1 be the first element of $(t_n)_{n\geq 1}$. Because $t_1 \in \mathcal{D}$, it is a continuity point of $t \mapsto Q(t)$ and, consequently, we have that $Q_n(t_1) = Q(t_1) + o_P(1)$. Hence, the subsequence $(n_k)_{k\geq 1}$ has a further subsequence, denoted by $(n_l^1)_{l\geq 1}$, such that $Q_{n_l^1}(t_1) \to Q(t_1)$ P-a.s. Or equivalently, $P(A_1) = 1$ for $A_1 := \left\{\omega \in \Omega : Q^\omega_{n_l}(t_1) \to Q(t_1)\right\}$.

We can repeat the above construction for every element of $(t_n)_{n\geq 1}$, where the subsequence $(n_l^2)_{l\geq 1}$ is a subsequence of $(n_l^1)_{l\geq 1}$, the subsequence $(n_l^3)_{l\geq 1}$ is a subsequence of $(n_l^2)_{l\geq 1}$, and so on. We now define the subsequence $(m_l)_{l\geq 1}$ element-wise as $m_l = n_l^l$. By construction, for any $k \in \mathbb{N}$ and $\omega \in \Omega$, we have that $Q^\omega_{n_l^k}(t_k) \to Q(t_k) \implies Q^\omega_{m_l}(t_k) \to Q(t_k)$ because $(m_l)_{l\geq 1}$ is eventually a subsequence of $(n_l^k)_{l\geq 1}$. This implies that

$$\left\{\omega \in \Omega : Q^\omega_{n_l^k}(t_k) \to Q(t_k)\right\} =: A_k \subset \tilde{A}_k := \left\{\omega \in \Omega : Q^\omega_{m_l}(t_k) \to Q(t_k)\right\}.$$

Because $P(A_k) = 1$ for every $k \in \mathbb{N}$, we have that $P\left(\bigcap_{k\geq 1} A_k\right) = 1$ and, because $A_k \subset \tilde{A}_k$ for every k, also that $P\left(\bigcap_{k\geq 1} \tilde{A}_k\right) = 1$.

For any $\omega \in \bigcap_{k\geq 1} \tilde{A}_k$, we have that $Q^\omega_{m_l} \to Q(t)$ for all $t \in \mathcal{D}$. By Lemma C.4.2, this implies that $Q^\omega_{m_l}(t) \to Q(t)$ for all continuity points. By van der Vaart (2000, Lemma 21.1), this implies that $F^\omega_{m_l}$ converges weakly to F_μ, or equivalently, that $\nu^\omega_{m_l}$ converges weakly to μ.

Note that the initial subsequence $(n_k)_{k\geq 1}$ was arbitrary. Hence, we have that for every subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_l)_{l\geq 1}$ such that

$$P\left(\omega : \nu^\omega_{m_l} \rightsquigarrow \mu\right) = 1.$$

This is the definition of weak convergence in probability given in Schmon et al. (2021, Definition 3)[4].

$\Leftarrow$. By the definition of weak convergence in probability, for every subsequence $(n_k)_{k\geq 1}$,

[4]As explained in the text before Lemma C.4.5, this is equivalent to our Definition C.4.4.

there exists a further subsequence $(n_l)_{l\geq 1}$ such that

$$P\left(\omega : \nu_{n_l}^{\omega} \rightsquigarrow \mu\right) = 1.$$

For all ω in the above set, van der Vaart (2000, Lemma 21.2) implies that $F^{-}_{\nu^{\omega}_{n_l}}(t) \to F^{-}_{\mu}(t)$ for every t a continuity point of $t \mapsto F^{-}_{\mu}(t)$. Because the initial subsequence was arbitrary, this implies that $F^{-}_{\nu_n}(t) = F^{-}_{\mu}(t) + o_P(1)$ for every t a continuity point of $t \mapsto F^{-}_{\mu}(t)$. □

Definition C.4.7 (Asymptotic uniform integrability)**.** The sequence of random probability measures ν_n^{ω} on $\mathbb{R}^k$ is asymptotically uniformly integrable with respect to $f : \mathbb{R}^k \to \mathbb{R}$ if for every subsequence, there exists a further subsequence $\nu^{\omega}_{m_n}$ such that for all $\omega \in A$ with $P(A) = 1$:

$$\lim_{M\to\infty} \lim_{n\to\infty} \mathbb{E}_{\nu^{\omega}_{m_n}} \left[|f(X)| \cdot \mathbf{1}\left\{|f(X)| > M\right\}\right] = 0.$$

Lemma C.4.8 (Convergence of moments)**.** *Let ν_n be a sequence of random probability measures on $\mathbb{R}^k$ converging weakly in probability to μ. Let $f : \mathbb{R}^k \to \mathbb{R}$ be a continuous function where $|\mathbb{E}_{\mu} f(X)| < \infty$. Then $\mathbb{E}_{\nu_n} f(X) \xrightarrow{P} \mathbb{E}_{\mu} f(X)$ if the sequence of random probability measures ν_n is asymptotically uniformly integrable with respect to f.*

Proof. Without loss of generality, assume that $f(X)$ is non-negative; otherwise, one can follow the same arguments below for the negative and positive parts of $f(X)$ separately. Let $(\nu^{\omega}_{l_n})_{n\geq 1}$ be a subsequence of $(\nu^{\omega}_n)_{n\geq 1}$. By the definition of asymptotic uniform integrability, there exists a further subsequence $(\nu^{\omega}_{m_n})_{n\geq 1}$, and set A with $P(A)$ such that for all $\omega \in A$:

$$\lim_{M\to\infty} \lim_{n\to\infty} \mathbb{E}_{\nu^{\omega}_{m_n}} \left[|f(X)| \cdot \mathbf{1}\left\{|f(X)| > M\right\}\right] = 0.$$

By the definition of weak convergence in probability[5], there is a further subsequence $\nu^{\omega}_{m'_n}$ such that $\mathbb{E}_{\nu^{\omega}_{m'_n}} g(X) \to \mathbb{E}_{\mu} g(X)$ for all $g \in C_b(\mathbb{R}^k)$ and $\omega \in A'$ with $P(A') = 1$. To simplify notation, we denote the latter subsequence and set simply by ν_{m_n} and A, respectively.

[5]Our Definition C.4.4 is not formulated in terms of subsequences, but this definition is equivalent to Schmon et al. (2021, Definition 3 in Supplementary Material) which is formulated in terms of subsequences.

The proof is complete if we can show that $\mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\mu} f(X) \to 0$ for all $\omega \in A$ because the initial subsequence $(\nu^{\omega}_{l_n})_{n \geq 1}$ was arbitrary. Indeed, this would imply that for every subsequence of $(\nu_n)_{n \geq 1}$, there exists a further subsequence $(\nu_{m_n})_{n \geq 1}$ where $\mathbb{E}_{\nu_{m_n}} f(X) \overset{a.e.}{\to} \mathbb{E}_{\mu} f(X)$. This in turn implies that $\mathbb{E}_{\nu_n} f(X) \overset{P}{\to} \mathbb{E}_{\mu} f(X)$.

We now show that $\mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\mu} f(X) \to 0$ for all $\omega \in A$. Select $\omega \in A$ and apply the triangle inequality as follows:

$$\begin{aligned}
\left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\mu} f(X) \right| \leq & \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \wedge M \right| \\
& + \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \wedge M - \mathbb{E}_{\mu} f(X) \wedge M \right| \\
& + \left| \mathbb{E}_{\mu} f(X) \wedge M - \mathbb{E}_{\mu} f(X) \right| .
\end{aligned}$$

For the expressions above, we first take the limit with respect to $n \to \infty$ and, second, with respect to $M \to \infty$. The left-hand side then simply is $\lim_{n \to \infty} \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\mu} f(X) \right|$ (as there is no dependence on M). We now consider the three expressions of the right-hand side separately.

We can upper bound the first expression of the right-hand side as follows:

$$\begin{aligned}
\left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \wedge M \right| &= \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \mathbf{1}(f(X) > M) + \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \mathbf{1}(f(X) \leq M) \right. \\
&\quad \left. - \mathbb{E}_{\nu^{\omega}_{m_n}} M \mathbf{1}(f(X) > M) - \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \mathbf{1}(f(X) \leq M) \right| \\
&= \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \mathbf{1}(f(X) > M) - \mathbb{E}_{\nu^{\omega}_{m_n}} M \mathbf{1}(f(X) > M) \right| \\
&\leq \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \mathbf{1}(f(X) > M) \right|
\end{aligned}$$

The inequality follows from the fact that $f(X)$ is non-negative and that $f(X)\mathbf{1}(f(X) > M) > M\mathbf{1}(f(X) > M)$. From asymptotic uniform integrability follows

$$\lim_{M \to \infty} \lim_{n \to \infty} \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \mathbf{1}(f(X) > M) = 0.$$

Hence, we have that $\lim_{M \to \infty} \lim_{n \to \infty} \left| \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) - \mathbb{E}_{\nu^{\omega}_{m_n}} f(X) \wedge M \right| = 0$.

Because $\nu_{m_n}^{\omega}$ converges weakly to μ and $x \mapsto f(x) \wedge M$ is bounded and continuous, we have that $\lim_{n\to\infty} \mathbb{E}_{\nu_{m_n}^{\omega}} f(X) \wedge M = \mathbb{E}_{\mu} f(X) \wedge M$. Hence, it follows that

$$\lim_{M\to\infty} \lim_{n\to\infty} \left| \mathbb{E}_{\nu_{m_n}^{\omega}} f(X) \wedge M - \mathbb{E}_{\mu} f(X) \wedge M \right| = 0.$$

The third expression is upper bounded by $\mathbb{E}_{\mu} f(X) \mathbf{1}(f(X) > M)$ (following the same argument as before). Because $|\mathbb{E}_{\mu} f(X)| < \infty$, we have that $\lim_{M\to\infty} f(X)\mathbf{1}(f(X) > M) = 0$.

The above results now imply that $\mathbb{E}_{\nu_{m_n}^{\omega}} f(X) - \mathbb{E}_{\mu} f(X) \to 0$ for all $\omega \in A$, which completes the proof. □

C.5 Asymptotic theory for parametric binomial mixture estimators

This section develops the asymptotic theory for the parametric binomial mixture estimators used in the main text. For each threshold $x \in \mathcal{I}$, the observable count $Y_i(x) = \sum_{j=1}^{m_i} \mathbf{1}\{X_{ij} \leq x\}$ follows a binomial mixture distribution, where the mixing distribution is the law of $F_{\nu_i}(x)$ for $\nu_i \sim \Pi$. In the parametric setting, we assume that this mixing distribution belongs to a family $G(\cdot; \beta(x))$ with $\beta(x) \in \mathbb{R}^p$, so that the target is the function-valued parameter $x \mapsto \beta_0(x)$.

We first consider a general class of parametric mixture estimators. In this setting, $\hat{\beta}_n(x)$ is obtained by solving estimating equations pointwise in x, and we analyze the resulting estimator as an element of $\ell^{\infty}(K)^p$ for compact subsets $K \subset \mathcal{I}$. We state a general Z-estimation result giving sufficient conditions for uniform asymptotic linearity of $\hat{\beta}_n$, provide sufficient conditions for these assumptions to hold, and then apply a delta method argument to obtain uniform asymptotic linearity of $G(\alpha; \hat{\beta}_n(x))$.

We then specialize the general theory to the Beta–Binomial model. We first define the model and derive the score, Hessian, Fisher information matrix, and gradient of the Beta CDF. We next verify the conditions needed for uniform asymptotic linearity of both $\hat{\beta}_n$ and $G(\alpha; \hat{\beta}_n(x))$ on K. We also verify the conditions used in Theorem 3.6.3 for the integrated estimator $\int_K G(\alpha; \hat{\beta}_n(x))\, dx$, which is the object needed for the asymptotic distribution results

in the main text.

The final part of the section collects the supporting technical results and proofs.

C.5.1 General parametric mixture estimators

We consider a general semiparametric model in which the mixing distribution at $x \in \mathcal{I}$, denoted by $t \mapsto G(x,t)$, is indexed by a parameter $\beta(x) \in \mathbb{R}^p$. To emphasize this dependence, we also write the mixing distribution as $t \mapsto G(t;\beta(x))$, where we assume the same parametric submodel for the mixing distribution at each $x \in \mathcal{I}$. The collection of mixing distributions on $\mathcal{I}$ is indexed by the function-valued parameter $\beta \in \ell^\infty(\mathcal{I})^p$, with true value $\beta_0 \in \ell^\infty(\mathcal{I})^p$. We study estimation of β_0 by pointwise parametric maximum likelihood and use Z-estimation theory to analyze $\hat{\beta}_n$ on compact subsets $K \subset \mathcal{I}$.

Estimation of β_0 and $G(\cdot,\alpha)$

We consider estimation of $\beta_0(x)$ uniformly on a compact interval $K \subset \mathcal{I}$, where the estimator $\hat{\beta}_n(x)$ is obtained as the solution of empirical estimating equations. We assume throughout that the estimating equations at x arise from the binomial mixture problem at x. Hence, the estimator $\hat{\beta}_n \in \ell^\infty(K)$ is defined through a set of binomial mixture problems.

Let $(O_i)_{i=1}^n$ for $O_i := \left\{m_i, (X_{i,j})_{j=1}^{m_i}\right\}$ be the observed data sampled i.i.d. from P as described in Section 3.2.1 of the main text. The observed data used for estimating the mixture distribution at x are $\{Y_i(x), m_i\}_{i=1}^n$ for $Y_i(x) := \sum_{j=1}^{m_i} I(X_{i,j} \leq x)$. Let the estimating function for the binomial mixture at x be denoted by $\psi(O_i;\beta(x),x) \in \mathbb{R}^p$, which only depends on O_i through $(Y_i(x), m_i)$. Define the empirical and population estimating functions pointwise as follows:

$$\Psi_n(\beta(x),x) := \mathbb{P}_n\psi(O_i;\beta(x),x), \quad \text{and} \quad \Psi(\beta(x),x) := P\psi(O_i;\beta(x),x). \tag{C.2}$$

We further view $x \mapsto \beta_0(x)$ as the function-valued estimand and define the corresponding

empirical and true estimating functions as operators from $\ell^\infty(K)^p$ into $\ell^\infty(K)^p$:

$$\Psi_n : \ell^\infty(K)^p \to \ell^\infty(K)^p, \quad \text{where} \quad \Psi_n(\beta)(x) := \Psi_n(\beta(x), x)$$
$$\Psi : \ell^\infty(K)^p \to \ell^\infty(K)^p, \quad \text{where} \quad \Psi(\beta)(x) := \Psi(\beta(x), x).$$

The space $\ell^\infty(K)^p$ is equipped with the norm $\|h\|_\infty := \sup_{x\in K}\|h(x)\|$ for $\|\cdot\|$ the Euclidean norm. The estimator $x \mapsto \hat{\beta}_n(x)$, further denoted by $\hat{\beta}_n$, is defined as an estimator satisfying $\|\Psi_n(\hat{\beta}_n)\|_\infty = 0$. We now reproduce Kosorok (2008, Theorem 13.4) with some modifications to the current setting, which states the conditions under which $\hat{\beta}_n$ is uniformly asymptotically linear on K.

Theorem C.5.1. *Assume that $\Psi(\beta_0) = 0$ and the following hold:*

(A) $\beta \mapsto \Psi(\beta)$ satisfies the following:

$$\|\Psi(\beta_n)\|_\infty \to 0 \quad \textit{implies} \quad \|\beta_n - \beta_0\|_\infty \to 0 \quad \textit{for any} \quad (\beta_n)_{n\geq 1}.$$

(B) The class $\{\psi(\cdot;\beta(x),x) : \beta \in \ell^\infty(K)^p, x \in K\}$ is P-Glivenko-Cantelli

(C) The class $\{\psi(\cdot;\beta(x),x) : \|\beta - \beta_0\|_\infty < \delta, x \in K\}$ is P-Donsker for some $\delta > 0$.

(D) The following holds:

$$\sup_{x\in K} P\|\psi(\cdot;\beta_n(x),x) - \psi(\cdot;\beta_0(x),x)\|^2 \to 0, \quad \textit{as} \quad \beta_n \to \beta_0 \quad \textit{in} \quad \ell^\infty(K)^p.$$

(E) $\beta \mapsto \Psi(\beta)$ is Fréchet-differentiable at β_0 with continuously invertible derivative $\dot{\Psi}_{\beta_0}$.

Then $n^{1/2}\left(\hat{\beta}_n - \beta_0\right) = -n^{1/2}\dot{\Psi}_{\beta_0}^{-1}\Psi_n(\beta_0) + o_P(1)$ where $n^{1/2}\left(\Psi_n - \Psi\right)(\beta_0) \rightsquigarrow Z$ for $Z \in \ell^\infty(K)^p$ a tight, mean zero Gaussian limiting distribution.

Condition (A) is a standard identifiability condition. Conditions (B) and (C) restrict the complexity of the class of estimating functions, and should hold for most parametric models. Condition (D) follows from regularity conditions that depend on the parametric

model. Condition (E) is a smoothness condition that is closely tied to the uniform asymptotic linearity of $\hat{\beta}_n$.

The proposition below provides sufficient conditions for conditions (A), (D), and (E) to hold.

Proposition C.5.2. *Suppose the following conditions hold:*

(1) $\beta(x) \mapsto \psi(O_i; \beta(x), x)$ *is twice continuously differentiable for every* $x \in K$*, with probability one. The first and second-order partial derivatives are denoted by* $\dot{\psi}(O_i; \beta(x), x)$ *and* $\ddot{\psi}(O_i; \beta(x), x)$*.*

(2) For every $x \in K$*, there exists an envelope function* $M_x(O_i)$ *with* $E[M_x(O_i)] < \infty$ *such that, for all* $\beta(x)$ *in a neighborhood of* $\beta_0(x)$*, we have*

$$\|\dot{\psi}(O_i; \beta(x), x)\| \leq M_x(O_i), \quad \|\ddot{\psi}(O_i; \beta(x), x)\| \leq M_x(O_i).$$

(3) There exists a convex open set $U \subset \mathbb{R}^p$ *containing an* ε*-enlargement of* $\{\beta_0(x) : x \in K\} \subset \mathbb{R}^p$ *such that*

$$\sup_{x \in K} \sup_{\beta(x) \in U} P\|\dot{\psi}(O_i; \beta(x), x)\|^2 < \infty, \qquad \sup_{x \in K} \sup_{\beta(x) \in U} P\|\ddot{\psi}(O_i; \beta(x), x)\| < \infty.$$

(4) There exists $c > 0$ *such that*

$$\|P\dot{\psi}(O_i; \beta_0(x), x)v\| \geq c\|v\| \quad \textit{for all } x \in K \textit{ and } v \in \mathbb{R}^p.$$

Then $\beta \mapsto \Psi(\beta)$ *satisfies conditions (A), (D), and (E) of Theorem C.5.1 where the Fréchet derivative and its continuous inverse are defined pointwise as follows:*

$$(\dot{\Psi}h)(x) = P\dot{\psi}(O_i; \beta_0(x), x)h(x), \quad \textit{for } h \in \ell^\infty(K)^p,$$
$$(\dot{\Psi}^{-1}f)(x) = (P\dot{\psi}(O_i; \beta_0(x), x))^{-1}f(x), \quad \textit{for } f \in \ell^\infty(K)^p.$$

In the following section, we will also verify conditions (B) and (C) for the Beta specification; however, general and directly useful sufficient conditions for (B) and (C) seem to be difficult to obtain.

Finally, the uniform asymptotic linearity of $\hat{\beta}_n$ on K, following from Theorem C.5.1, implies uniform asymptotic linearity of $G(\alpha; \hat{\beta}_n(\cdot))$ on K under smoothness conditions on the map $\beta(x) \mapsto G(\alpha; \beta(x))$.

Proposition C.5.3. *Assume the conditions of Theorem C.5.1 hold. Further assume that $\beta(x) \mapsto G(\alpha; \beta(x))$ is continuously differentiable on a convex compact set $U \subset \mathbb{R}^p$ containing an ε-enlargement of $\{\beta_0(x) : x \in K\} \subset \mathbb{R}^p$, with derivative $\dot{G}(\alpha; \beta_0(x))$. Then*

$$n^{1/2}\Big(G(\alpha; \hat{\beta}_n(x)) - G(\alpha; \beta_0(x))\Big) = -n^{1/2}\dot{G}(\alpha; \beta_0(x))\dot{\Psi}_{\beta_0}^{-1}\Psi_n(\beta_0)(x) + r_n(x), \quad \sup_{x \in K}|r_n(x)| = o_P(1).$$

C.5.2 Beta–Binomial

This subsection applies the general Z-estimation theory from the previous section to the Beta specification. The structure is as follows:

1. **Preliminaries**: We define the Beta–Binomial model, derive the score vector ψ, the Hessian $\dot{\psi}$, and the Fisher information matrix $I(x)$.
2. **Asymptotic linearity of $\hat{\beta}_n$ and $G(\alpha; \hat{\beta}_n)$**: We establish uniform asymptotic linearity of $\hat{\beta}_n$ and $G(\alpha; \hat{\beta}_n(x))$ with explicit influence function formulas.
3. **Conditions of Theorem 3.6.3**: We verify the three conditions required for the integrated estimator $\int_K G(\alpha; \hat{\beta}_n(x))\, dx$ to be asymptotically normal, which are used in the main text for $K = \mathcal{I}$.

Preliminaries

For each $x \in K$, let the mixing distribution be Beta with parameter $\beta(x) := (a_x, b_x)^\top \in (0, \infty)^2$. The function-valued parameter is therefore $\beta \in \ell^\infty(K)^2$ with true value $\beta_0 \in \ell^\infty(K)^2$ where $\beta_0(x) = (a_{0,x}, b_{0,x})^\top$. Conditional on m_i, the model is $Y_i(x) \sim$ Beta-Binomial(m_i, a_x, b_x),

with log-likelihood contribution

$$\ell_i(\beta(x), x) = \log B(Y_i(x) + a_x, m_i - Y_i(x) + b_x) - \log B(a_x, b_x).$$

The corresponding estimating function is the score function

$$\psi(O_i; \beta(x), x) = \begin{pmatrix} s_a(O_i; \beta(x), x) \\ s_b(O_i; \beta(x), x) \end{pmatrix},$$

where

$$\begin{aligned} s_a(O_i; \beta(x), x) &= \psi_0(Y_i(x) + a_x) - \psi_0(a_x) + \psi_0(a_x + b_x) - \psi_0(m_i + a_x + b_x), \\ s_b(O_i; \beta(x), x) &= \psi_0(m_i - Y_i(x) + b_x) - \psi_0(b_x) + \psi_0(a_x + b_x) - \psi_0(m_i + a_x + b_x), \end{aligned} \tag{C.3}$$

and ψ_0 denotes the digamma function. The Hessian is

$$\dot{\psi}(O_i; \beta(x), x) = \begin{pmatrix} \partial_a s_a & \partial_b s_a \\ \partial_a s_b & \partial_b s_b \end{pmatrix},$$

where

$$\begin{aligned} \partial_a s_a &= \psi_1(Y_i(x) + a_x) - \psi_1(a_x) + \psi_1(a_x + b_x) - \psi_1(m_i + a_x + b_x), \\ \partial_b s_b &= \psi_1(m_i - Y_i(x) + b_x) - \psi_1(b_x) + \psi_1(a_x + b_x) - \psi_1(m_i + a_x + b_x), \\ \partial_b s_a &= \partial_a s_b = \psi_1(a_x + b_x) - \psi_1(m_i + a_x + b_x), \end{aligned}$$

and ψ_1 denotes the trigamma function. The entries of the Fisher information matrix, defined as $I(x) := -P\dot{\psi}(O_i; \beta_0(x), x)$, are

$$\begin{aligned} I_{aa}(x) &= -P\left[\psi_1(Y_i(x) + a_{0,x})\right] + \psi_1(a_{0,x}) - \psi_1(a_{0,x} + b_{0,x}) + P\left[\psi_1(m_i + a_{0,x} + b_{0,x})\right], \\ I_{bb}(x) &= -P\left[\psi_1(m_i - Y_i(x) + b_{0,x})\right] + \psi_1(b_{0,x}) - \psi_1(a_{0,x} + b_{0,x}) + P\left[\psi_1(m_i + a_{0,x} + b_{0,x})\right], \\ I_{ab}(x) &= -\psi_1(a_{0,x} + b_{0,x}) + P\left[\psi_1(m_i + a_{0,x} + b_{0,x})\right]. \end{aligned} \tag{C.4}$$

The Beta CDF with parameters (a_x, b_x) evaluated in α equals $I_\alpha(a_x, b_x)$ where $I_\alpha(a_x, b_x)$ is the regularized incomplete beta function defined as follows:

$$I_\alpha(a_x, b_x) := \frac{\mathrm{B}_\alpha(a_x, b_x)}{\mathrm{B}(a_x, b_x)}, \quad \text{for} \quad \mathrm{B}_\alpha(a_x, b_x) := \int_0^\alpha t^{a_x - 1}(1-t)^{b_x - 1}\, dt.$$

This function is differentiable with respect to a_x and b_x for $a_x, b_x > 0$ and $\alpha \in (0, 1)$. The partial derivatives are given by

$$\begin{aligned}
\partial_{a_x} I_\alpha(a_x, b_x) &= \frac{1}{\mathrm{B}(a_x, b_x)} \int_0^\alpha t^{a_x - 1}(1-t)^{b_x - 1} \log t \, dt - I_\alpha(a_x, b_x)\big(\psi_0(a_x) - \psi_0(a_x + b_x)\big), \\
\partial_{b_x} I_\alpha(a_x, b_x) &= \frac{1}{\mathrm{B}(a_x, b_x)} \int_0^\alpha t^{a_x - 1}(1-t)^{b_x - 1} \log(1-t) \, dt - I_\alpha(a_x, b_x)\big(\psi_0(b_x) - \psi_0(a_x + b_x)\big).
\end{aligned} \tag{C.5}$$

Uniform Asymptotic Linearity of $\hat{\beta}_n$ and $G(\alpha; \hat{\beta}_n)$ on K

Uniform Asymptotic Linearity of $\hat{\beta}_n$ on K. We verify the conditions of Theorem C.5.1 next. For this, we make the following assumptions on the data generating-process and the true parameter β_0:

(B1) $\{(a_{0,x}, b_{0,x}) : x \in K\} \subset (0, \infty)^2$ is covered by a compact subset of $(0, \infty)^2$,

(B2) $1 \leq m_i < C < \infty$ with probability one and $P(m_i > 1) > 0$.

Conditions (A), (D), and (E) of Theorem C.5.1 follow from Proposition C.5.2 where the conditions of this proposition are verified as follows:

1. The twice continuous differentiability of $\beta(x) \mapsto \psi(O_i; \beta(x), x)$ follows from the polygamma functions being infinitely differentiable on $(0, \infty)$.
2. The existence of an envelope function $M_x(O_i)$ with finite expectation follows from the fact that $1 \leq m_i < C < \infty$ with probability one and that the digamma and trigamma functions are bounded on all compact subsets of $(0, \infty)$.
3. The uniform boundedness of $P\|\dot{\psi}(O_i; \beta(x), x)\|^2$ and $P\|\ddot{\psi}(O_i; \beta(x), x)\|$ follows from the

fact that $1 \leq m_i < C < \infty$ with probability one and that the digamma and trigamma functions are bounded on all compact subsets of $(0, \infty)$.

4. This follows from the uniform nonsingularity of the Fisher information matrix $I(x)$, which itself follows from Lemma C.5.7, which is applicable under Assumptions B1 and B2.

Conditions (B) and (C) of Theorem C.5.1 follow from Proposition C.5.10 under condition (i). Note that condition (B) only holds by Proposition C.5.10 if we restrict the parameter space to a compact subset of $(0, \infty)^2$ that contains $\{(a_{0,x}, b_{0,x}) : x \in K\}$.

It now follows from Theorem C.5.1 that $\hat{\beta}_n$ is uniformly asymptotically linear on K with influence function $I(x)^{-1}\psi(O_i; \beta_0(x), x)$.

Uniform Asymptotic Linearity of $G(\alpha; \hat{\beta}_n(x))$ on K The uniform asymptotic linearity of $\hat{\beta}_n$ on K implies the uniform asymptotic linearity of $G(\alpha; \hat{\beta}_n(x))$ on K under the smoothness conditions of Proposition C.5.3, which are satisfied in the current setting. Indeed, the map $\beta(x) \mapsto G(\alpha; \beta(x))$ is continuously differentiable on $(0, \infty)^2$, which contains a convex compact set that contains a ε-enlargement of $\{\beta_0(x) : x \in K\}$. The influence function of $G(\alpha; \hat{\beta}_n(x))$ is then

$$\mathrm{IF}(O_i; x) := \dot{G}(\alpha, \beta_0(x)) I^{-1}(x) \psi(O_i; \beta_0(x), x). \tag{C.6}$$

Remark C.5.1. *Assumption B1 requires that the distribution of $F_\nu(x)$ for x near the endpoints of $\mathcal{I}$ is not approaching a point mass at 0 or 1, which implies that the subject-specific distributions F_ν have non-zero mass near the endpoints of $\mathcal{I}$. This ensures uniform bounds on the scores and guarantees that the Beta-binomial Fisher information is nonsingular uniformly over $\mathcal{I}$.*[6] *If the mixing distribution of the endpoints of $\mathcal{I}$ are point masses at zero and*

[6]We suspect that this assumption can be relaxed. However, our current strategy of proof requires this assumption. This strategy involves showing uniform asymptotic linearity of $(\hat{a}_{n,x}, \hat{b}_{n,x})$, which in turn implies uniform asymptotic linearity of $I_\alpha(\hat{a}_{n,x}, \hat{b}_{n,x})$. We suspect that uniform asymptotic linearity may hold for $I_\alpha(\hat{a}_{n,x}, \hat{b}_{n,x})$, but not for $(\hat{a}_{n,x}, \hat{b}_{n,x})$, under relaxed conditions.

one, then $a_x \to 0$ *as* $x \to \inf \mathcal{I}$*, and* $b_x \to 0$ *as* $x \to \sup \mathcal{I}$*. Hence,* $\{\beta_0(x) : x \in \mathcal{I}\}$ *is not contained in a compact set. Furthermore, we then have that* $\|I(x)\|^{-1} \to \infty$ *as* x *approaches the endpoints of* $\mathcal{I}$*.*

Remark C.5.2. *Assumption B2 is reasonable in practically any application with sparse distributional data. We expect that this assumption will change to* $P(m_i > p) > 0$ *if a parametric binomial mixture submodel with* $p + 1$ *parameters is used at every* $x \in \mathcal{I}$*.*

Conditions of Theorem 3.6.3

We now verify the conditions of Theorem 3.6.3 for the Beta specification. In addition to Assumptions B1 and B2 above, we make the following additional assumption:

(B3) There exist $c_1, c_2 > 0$ such that $c_1|x_1 - x_2| \leq |H_{0,\alpha}(x_1) - H_{0,\alpha}(x_2)| \leq c_2|x_1 - x_2|$ for all $x_1, x_2 \in K$, where $H_{0,\alpha}(x) := 1 - G(\alpha; \beta_0(x))$.

To ensure that $\hat{H}_{n,\alpha}(x) := 1 - G(\alpha; \hat{\beta}_n(x))$ is a valid CDF, we further apply isotonic regression to the initial estimator $x \mapsto 1 - G(\alpha; \hat{\beta}_n(x))$ to obtain a non-decreasing estimator, which we denote by $\hat{H}^{\text{iso}}_{n,\alpha}(x)$. Westling et al. (2020) show that, if $\hat{H}_{n,\alpha}(x)$ is uniformly asymptotically linear on K and Assumption B3 holds, then $\hat{H}^{\text{iso}}_{n,\alpha}(x)$ is also uniformly asymptotically linear on K with the same (pointwise) influence function as $\hat{H}_{n,\alpha}(x)$. In the arguments below, we use $\hat{H}_{n,\alpha}$ and $H^{\text{iso}}_{n,\alpha}$ interchangeably.

Condition AL1. As shown in the previous subsection, $G(\alpha, \hat{\beta}_n(x))$ is uniformly asymptotically linear on K with influence function given in (C.6). Because K is assumed to be compact, the integrated remainder is also $o_P(1)$.

Below for condition AL3, we verify that $\text{IF}(O_i; x)$ is uniformly bounded on K for all O_i, which implies that $\mathbb{E}\big[\int_K |\psi_\alpha(O; x)|\big] < \infty$.

Condition AL2. By construction, isotonic regression applied to $x \mapsto 1 - G(\alpha; \hat{\beta}_n(x))$ produces a non-decreasing function $\hat{H}^{\text{iso}}_{n,\alpha}$ on K. Without loss of generality, $\hat{H}^{\text{iso}}_{n,\alpha}$ can be

defined to be right-continuous with limits 0 and 1 at the boundaries of K. Since K is compact, the first moment of $\hat{H}^{\text{iso}}_{n,\alpha}$ is finite. Hence, $\hat{H}^{\text{iso}}_{n,\alpha}$ is a valid CDF (i.e., non-decreasing, right-continuous, with limits 0 and 1 at the boundaries) with finite first moment.

Condition AL3. This condition requires that $\text{Var}(\tilde{\psi}_\alpha) \leq E[\int_K |\text{IF}(O_i, x)|^2 \, dx]$ is finite, where $\text{IF}(O_i, x)$ is the influence function given in equation (C.6). This influence function is a product of three components: (i) partial derivatives $\partial_{a_x} I_\alpha$ and $\partial_{b_x} I_\alpha$, (ii) entries of $I(x)^{-1}$, and (iii) score functions s_a and s_b. The partial derivatives do not depend on the data and are continuous functions of $\beta_0(x)$. Since $\{\beta_0(x) : x \in K\}$ is covered by a compact subset of $(0, \infty)^2$, they are bounded on K. The smallest eigenvalue of $I(x)$ is bounded away from zero by Lemma C.5.7, so the entries of $I(x)^{-1}$ are bounded on K. The score functions themselves are bounded on K because they only involve digamma functions evaluated at $a_{0,x}$, $Y_i(x) + a_{0,x}$, $b_{0,x}$, $a_{0,x} + b_{0,x}$, $m_i - Y_i(x) + b_{0,x}$, and $m_i + a_{0,x} + b_{0,x}$, which belong to a compact subset of $(0, \infty)$ due to Assumptions B1 and B2. Hence, $\text{IF}(O_i, x)$ is uniformly bounded on K, and $E[\int_K |\text{IF}(O_i, x)|^2 \, dx]$ is finite.

Remark C.5.3. *Assumption B3 basically requires that the distribution of subject-specific α-quantiles, which has CDF $x \mapsto H_{0,\alpha}(x)$, has a Lebesgue density bounded between $c_1 > 0$ and $c_2 < \infty$. This assumption is required for applying the theoretical results in Westling et al. (2020) which imply that the isotonic regression step does not change uniform asymptotic linearity. This condition can probably be relaxed, especially the lower bound (Westling et al., 2020).*

C.5.3 Technical Results and Proofs

Function-Analytic Results

Lemma C.5.4 (Fréchet differentiability and continuous invertibility of pointwise operators)**.** *Let K be an arbitrary index set and define*

$$\ell^\infty(K)^p = \Big\{\theta : K \to \mathbb{R}^p \text{ such that } \|\theta\|_\infty := \sup_{x\in K} \|\theta(x)\| < \infty\Big\}.$$

Let $g : K \times \mathbb{R}^p \to \mathbb{R}^p$ and define the operator

$$T : \ell^\infty(K)^p \to \ell^\infty(K)^p, \qquad (T\theta)(x) := g(x, \theta(x)).$$

Fix $\theta_0 \in \ell^\infty(K)^p$ and assume:

1. *For every $x \in K$, the map $y \mapsto g(x, y)$ is twice continuously differentiable with first and second-order partial derivatives denoted by $\dot{g}(x, y)$ and $\ddot{g}(x, y)$ respectively.*
2. *Let $R(\theta_0)$ be the range of θ_0 on K. There exists a convex set $U \subset \mathbb{R}^p$ containing an ε-enlargement[7] of $R(\theta_0)$ such that*

$$\sup_{x\in K}\sup_{y\in U} \|\dot{g}(x, y)\| < \infty, \quad \sup_{x\in K}\sup_{y\in U} \|\ddot{g}(x, y)\| < \infty.$$

3. *There exists $c > 0$ such that*

$$\|\dot{g}(x, \theta_0(x))v\| \geq c\|v\| \quad \text{for all } x \in K \text{ and } v \in \mathbb{R}^p.$$

Then:

(i) T is Fréchet differentiable at θ_0 where the derivative $\dot{T}_{\theta_0} : \ell^\infty(K)^p \to \ell^\infty(K)^p$ is the

[7]The ε-enlargement of A, denoted by A^ε, is defined as $A^\varepsilon := \{x \in \mathbb{R}^p : \text{dist}(x, A) < \varepsilon\}$ where $\text{dist}(x, A) := \inf_{y\in A}\|x - y\|$.

following bounded linear operator:

$$(\dot{T}_{\theta_0} h)(x) = \dot{g}(x, \theta_0(x))\, h(x), \qquad h \in \ell^\infty(K)^p.$$

(ii) The derivative $\dot{T}_{\theta_0}$ has a continuous inverse defined pointwise as follows:

$$(\dot{T}_{\theta_0}^{-1} f)(x) := \dot{g}(x, \theta_0(x))^{-1} f(x), \qquad f \in \ell^\infty(K)^p.$$

Proof. Let $\varepsilon > 0$ be the value corresponding to the ε-enlargement in condition 3. For any $h \in \ell^\infty(K)^p$ with $\|h\|_\infty < \varepsilon$, we then have that $\theta_0(x) + h(x) \in U$ for all $x \in K$.

(Fréchet differentiability): Let $j \in \{1, \ldots, p\}$ and use subscript j to denote the j'th element of g such that $g_j : K \times \mathbb{R}^p \to \mathbb{R}$ and $g = (g_1, \ldots, g_p)$. For each $x \in K$, by the standard multivariate Taylor expansion and condition 1,

$$g_j(x, \theta_0(x) + h(x)) = g_j(x, \theta_0(x)) + \dot{g}_j(x, \theta_0(x))h(x) + R_j(h)(x),$$

where

$$R_j(h)(x) = \frac{1}{2} h(x)^\top \ddot{g}_j(x, \tilde{\theta}(x))h(x),$$

for some $\tilde{\theta}(x)$ on the line segment between $\theta_0(x)$ and $\theta_0(x) + h(x)$. Because U is convex and $\theta_0(x), \theta_0(x) + h(x) \in U$, the point $\tilde{\theta}(x)$ also belongs to U for all $x \in K$.

By uniform boundedness of the Hessian (condition 2), we have that

$$\|R_j(h)\|_\infty := \sup_{x \in K} |R_j(h)(x)| \leq \|h\|_\infty^2 \cdot \sup_{x \in K} \sup_{y \in U} \|\ddot{g}_j(x, y)\| < \infty.$$

This holds for any $j \in \{1, \ldots, p\}$ and thus implies that $\|R(h)\|_\infty < \infty$. Letting $\dot{T}_{\theta_0}(h)(x) = \dot{g}(x, \theta_0(x))\, h(x)$, this thus shows that

$$\frac{\|T(\theta_0 + h) - T(\theta_0) - \dot{T}_{\theta_0}(h)\|_\infty}{\|h\|_\infty} = \frac{\|R_{\theta_0}(h)\|_\infty}{\|h\|_\infty} = O(\|h\|_\infty) = o(1) \quad \text{as} \quad \|h\|_\infty \to 0.$$

Linearity and boundedness of $\dot{T}$ follows directly as follows, for any $h_1, h_2 \in \ell^\infty(K)^p$ and any $x \in K$:

$$\dot{T}_{\theta_0}(h_1 + h_2)(x) := \dot{g}(x, \theta_0(x))\ (h_1(x) + h_2(x)) = \dot{T}_{\theta_0}(h_1)(x) + \dot{T}_{\theta_0}(h_2)(x),$$
$$\|\dot{T}_{\theta_0}(h)\|_\infty = \sup_{x \in K} \|\dot{g}(x, \theta_0(x))h(x)\| \leq \sup_{x \in K} \|\dot{g}(x, \theta_0(x))\| \cdot \|h\|_\infty < \infty.$$

This proves Fréchet differentiability.

(Continuous invertibility): The inverse of $\dot{T}_{\theta_0}$ is defined pointwise as follows $\dot{T}^{-1}(f)(x) := \dot{g}(x, \theta_0(x))^{-1} f(x)$ where the matrix inverse exists by condition 3. Also, by condition 3, we have for any $h \in \ell^\infty(K)^p$ and some $c > 0$ that,

$$\|\dot{T}_{\theta_0}(h)\|_\infty := \sup_{x \in K} \|\dot{g}(x, \theta_0(x))h(x)\| \geq c \sup_{x \in K} \|h(x)\| = c\|h\|_\infty,$$

which shows that $\dot{T}_{\theta_0}$ is invertible. □

Corollary C.5.5. *Under the assumptions of Lemma C.5.4 and for $\theta_0 \in \ell^\infty(K)^p$ such that $\|T\theta_0\|_\infty = 0$, T satisfies the following:*

$$\|T\theta_n\|_\infty \to 0 \quad \textit{implies} \quad \|\theta_n - \theta_0\|_\infty \to 0 \quad \textit{for any} \quad (\theta_n)_{n \geq 1}.$$

Proof. Let $(\theta_n)_{n \geq 1}$ be any sequence such that $\|T\theta_n\|_\infty \to 0$. By Lemma C.5.4, T is Fréchet differentiable at θ_0 with derivative $\dot{T}_{\theta_0}$.

$$\begin{aligned}
\|\dot{T}_{\theta_0}(\theta_n - \theta_0)\|_\infty &\leq \|\dot{T}_{\theta_0}(\theta_n - \theta_0) - T(\theta_n)\|_\infty + \|T(\theta_n)\|_\infty \\
&\leq \|T(\theta_n) - T(\theta_0) - \dot{T}_{\theta_0}(\theta_n - \theta_0)\|_\infty + \|T(\theta_n)\|_\infty \\
&\leq o\left(\|\theta_n - \theta_0\|_\infty\right) + o(1)
\end{aligned}$$

By continuous invertibility (see last part of the proof of Lemma C.5.4), we have that $c\|\theta_n - \theta_0\|_\infty \leq \|\dot{T}_{\theta_0}(\theta_n - \theta_0)\|_\infty$ for some $c > 0$. Together with the previous display, this gives the

following:

$$\|\theta_n - \theta_0\|_\infty \leq c^{-1}\|\dot{T}_{\theta_0}(\theta_n - \theta_0)\|_\infty \leq o\left(\|\theta_n - \theta_0\|_\infty\right) + o(1).$$

This implies that $\|\theta_n - \theta_0\|_\infty = o(1)$. □

Lemma C.5.6. *Let $I(\alpha)$ be a positive definite matrix for each $\alpha \in A$, where A is compact, and assume that the entries of $I(\alpha)$ are continuous in α on A. Then there exists a constant $c > 0$ such that $\inf_{\alpha\in A} \lambda_{min}(I(\alpha)) \geq c$, where $\lambda_{min}(I(\alpha))$ is the smallest eigenvalue of $I(\alpha)$.*

Proof. Define the function

$$f : A \to \mathbb{R}, \quad f(\alpha) := \lambda_{\min}(I(\alpha)).$$

Since the entries of $I(\alpha)$ are continuous in α, and the eigenvalues of a matrix depend continuously on the matrix entries, f is continuous on A. A continuous real-valued function on a compact set attains its minimum; therefore, there exists some $\alpha_0 \in A$ such that

$$\inf_{\alpha\in A} \lambda_{\min}(I(\alpha)) = \min_{\alpha\in A} \lambda_{\min}(I(\alpha)) = \lambda_{\min}(I(\alpha_0)).$$

Since $I(\alpha)$ is positive definite for all $\alpha \in A$, we have that $\lambda_{\min}(I(\alpha_0)) > 0$. Set $c := \lambda_{\min}(I(\alpha_0))$. Hence, $\inf_{\alpha\in A} \lambda_{\min}(I(\alpha)) = c > 0$. □

Beta-binomial

Lemma C.5.7. *Let $I(x)$ be the Fisher information of the beta-binomial mixing distribution for $x \in K$ with the entries defined in* (C.4)*. Assume that $\{(a_{0,x}, b_{0,x}) : x \in K\}$ belongs to a compact subset of $(0, \infty)^2$, that $1 \leq m_i < C < \infty$ with probability one, and that $P(m_i > 1) > 0$. Then there exists $c > 0$ such that $\lambda_{\min}(I(x)) \geq c$ for all $x \in K$.*

Proof. Let A be the compact subset containing $\{(a_{0,x}, b_{0,x}) : x \in K\}$. We now use Lemma C.5.6 where we have to show that $I(\alpha, \beta)$, defined as the Fisher information matrix for the beta-binomial mixture with parameters (α, β), is positive definite for all $(\alpha, \beta) \in A$ and its

entries are continuous in (α, β). We further denote the entries of the Fisher information matrix at (α, β) by $I_{\alpha,\alpha}$, $I_{\beta,\beta}$, and $I_{\alpha,\beta}$.

By standard likelihood theory, we have that $I_{\alpha,\alpha} = \mathbb{E}_{(\alpha,\beta)}[s_\alpha^2]$, $I_{\beta,\beta} = \mathbb{E}_{(\alpha,\beta)}\left[s_\beta^2\right]$, and $I_{\alpha,\beta} = \mathbb{E}_{(\alpha,\beta)}[s_\alpha s_\beta]$ where the scores are defined in (C.3) (replacing (a, b) with (α, β)). Since $I(\alpha, \beta)$ is a covariance matrix, it is positive semidefinite. It is positive definite if and only if the determinant is positive, which is equivalent to $I_{\alpha,\alpha} I_{\beta,\beta} - I_{\alpha,\beta}^2 > 0$. Because the digamma function is strictly increasing on $(0, \infty)$, we have that $\mathbb{E}_{(\alpha,\beta)}\left[s_\alpha^2\right] = 0$ only if the distribution of Y is degenerate. The distribution of Y is not degenerate for $(\alpha, \beta) \in A \subset (0, \infty)^2$ and $m_i \geq 1$, and similarly $\mathbb{E}_{(\alpha,\beta)}\left[s_\beta^2\right] > 0$. Moreover, by Cauchy–Schwarz we always have $I_{\alpha,\beta}^2 \leq I_{\alpha,\alpha} I_{\beta,\beta}$, with equality if and only if $s_\beta = c\, s_\alpha$ almost surely for some constant c.

If $P(m_i = 1) = 1$, then $Y \in \{0, 1\}$ and the beta-binomial model reduces to a Bernoulli model with success probability $\alpha/(\alpha + \beta)$, so the score components are colinear and $I(\alpha, \beta)$ is singular.

Now assume $P(m_i > 1) > 0$. Since m_i is integer-valued and $1 \leq m_i < C$ almost surely, there exists $m_* \geq 2$ such that $P(m_i = m_*) > 0$. Suppose, toward a contradiction, that $s_\beta = c\, s_\alpha$ almost surely for some constant c and $(\alpha, \beta) \in (0, \infty)^2$. Then this relation also holds conditional on $m_i = m_*$. Under the beta-binomial model with $(\alpha, \beta) \in (0, \infty)^2$, every value $y \in \{0, \ldots, m_*\}$ has strictly positive conditional probability, so we must have

$$s_\beta(y; m_*, \alpha, \beta) = c\, s_\alpha(y; m_*, \alpha, \beta), \qquad y = 0, \ldots, m_*.$$

Taking first differences in y gives, for $y = 0, \ldots, m_* - 1$,

$$s_\alpha(y+1) - s_\alpha(y) = \psi_0(y + 1 + \alpha) - \psi_0(y + \alpha) = \frac{1}{y + \alpha},$$

$$s_\beta(y+1) - s_\beta(y) = \psi_0(m_* - y - 1 + \beta) - \psi_0(m_* - y + \beta) = -\frac{1}{m_* - y - 1 + \beta}.$$

Hence

$$c = c\frac{s_\alpha(y+1) - s_\alpha(y)}{s_\alpha(y+1) - s_\alpha(y)} = \frac{s_\beta(y+1) - s_\beta(y)}{s_\alpha(y+1) - s_\alpha(y)} = -\frac{y+\alpha}{m_* - y - 1 + \beta} \quad \text{for} \quad y = 0, \ldots, m_* - 1.$$

The right-hand side depends on y when $m_* \geq 2$, which is impossible for a constant c. This contradiction shows that equality in Cauchy–Schwarz cannot occur, so

$$I_{\alpha,\beta}^2 < I_{\alpha,\alpha} I_{\beta,\beta},$$

and therefore $I(\alpha, \beta)$ is positive definite for all $(\alpha, \beta) \in (0, \infty)^2$.

For continuity, note that $m_i \in \{1, \ldots, \lceil C \rceil - 1\}$ almost surely and, conditional on $m_i = m$, Y takes values in the finite set $\{0, \ldots, m\}$. Since expectations $E_{(\alpha,\beta)}[\cdot]$ are finite weighted sums, we have that, for $(\alpha_n, \beta_n) \to (\alpha, \beta) \in (0, \infty)^2$,

$$I_{\alpha_n,\alpha_n} = \mathbb{E}_{(\alpha_n,\beta_n)}\left[s_{\alpha_n}^2\right] \to I_{\alpha,\alpha} = \mathbb{E}_{(\alpha,\beta)}\left[s_\alpha^2\right] \quad \text{as} \quad (\alpha_n, \beta_n) \to (\alpha, \beta),$$

if for each fixed m and y (i) the weights $P_{(\alpha_n,\beta_n)}(Y = y \mid m_i = m)$ converge to $P_{(\alpha,\beta)}(Y = y \mid m_i = m)$ and (ii) s_α is continuous in (α, β) for each fixed m and y. This is indeed the case because (i) the beta-binomial pmf is continuous in (α, β) for each fixed m and y and (ii) s_α is continuous in (α, β) for each fixed m and y because the digamma function is continuous on $(0, \infty)$. The same arguments apply to $I_{\alpha,\beta}$ and $I_{\beta,\beta}$. □

Proposition C.5.8. *Consider the data-generating mechanism described in Section 3.2.1 where the observed data are $O := (m, X_1, \ldots, X_m) \sim P$. Assume that $1 \leq m < C < \infty$ with probability one. Then the function class*

$$\mathcal{F} := \left\{ \sum_{j=1}^{m} \mathbf{1}\left(X_j \leq x\right) : x \in \mathbb{R} \right\}$$

is P-Donsker.

Proof. Since $m \leq C$ almost surely, we may embed observations into the fixed-dimensional space

$$O = (m, X_1, \ldots, X_{\lfloor C \rfloor}),$$

where we define $X_j = 0$ for $j > m$. Then the class $\mathcal{F}$ can be written as

$$\mathcal{F} = \left\{ \sum_{j=1}^{\lfloor C \rfloor} \mathbf{1}(X_j \leq x,\, j \leq m) : x \in \mathbb{R} \right\}.$$

For each fixed $j \in \{1, \ldots, \lfloor C \rfloor\}$, define

$$\mathcal{G}_j := \{\mathbf{1}(X_j \leq x,\, j \leq m) : x \in \mathbb{R}\}.$$

We can write

$$\mathbf{1}(X_j \leq x,\, j \leq m) = \mathbf{1}(X_j \leq x)\, \mathbf{1}(j \leq m).$$

Since $\mathbf{1}(j \leq m)$ is a fixed measurable function and the class

$$\{\mathbf{1}(X_j \leq x) : x \in \mathbb{R}\}$$

is a VC class (with VC index 2), it follows that $\mathcal{G}_j$ is a VC-subgraph class (Kosorok, 2008, Lemma 9.9).

The class $\mathcal{F}$ is the finite sum of $\lfloor C \rfloor$ such classes:

$$\mathcal{F} = \left\{ \sum_{j=1}^{\lfloor C \rfloor} g_j : g_j \in \mathcal{G}_j \right\}.$$

Since the finite sum of P-Donsker classes is P-Donsker (Kosorok, 2008, Corollary 9.32), it suffices to verify that each $\mathcal{G}_j$ is P-Donsker.

By Kosorok (2008, Theorem 9.3), $\mathcal{G}_j$ satisfies the uniform entropy condition of Kosorok (2008, Theorem 8.19). To show that $\mathcal{G}_j$ is P-Donsker, it remains to verify that $\mathcal{G}_j$ admits a

square-integrable envelope and is P-measurable.

For all $g \in \mathcal{G}_j$,

$$0 \leq g(O) \leq 1,$$

so $\mathcal{G}_j$ admits the square-integrable envelope $G_j(O) := 1$.

It remains to verify P-measurability. We show that $\mathcal{G}_j$ is pointwise measurable. Let $\mathcal{G}_{j,0} \subset \mathcal{G}_j$ be the subclass obtained by restricting x to $\mathbb{Q}$. Then $\mathcal{G}_{j,0}$ is countable. For any $x_0 \in \mathbb{R}$ and corresponding $g \in \mathcal{G}_j$, choose a sequence $(q_n)_{n\geq 1} \subset \mathbb{Q}$ such that $q_n \downarrow x_0$. Let $g_n \in \mathcal{G}_{j,0}$ denote the corresponding functions. Then, for every O,

$$g_n(O) = \mathbf{1}(X_j \leq q_n) \cdot \mathbf{1}(j \leq m) \to \mathbf{1}(X_j \leq x_0) \cdot \mathbf{1}(j \leq m) = g(O).$$

Hence, $\mathcal{G}_j$ is pointwise measurable and therefore P-measurable.

We conclude that $\mathcal{G}_j$ satisfies the uniform entropy condition, admits a square-integrable envelope, and is P-measurable. Therefore, $\mathcal{G}_j$ is P-Donsker by Kosorok (2008, Theorem 8.19). □

Corollary C.5.9. *Consider the data-generating mechanism described in Section 3.2.1 where the observed data are $O := (m, X_1, \ldots, X_m) \sim P$. Assume that $1 \leq m < C < \infty$ with probability one. Then the classes of functions $\psi_0 \circ \mathcal{F}_1$, $\psi_0 \circ \mathcal{F}_2$, and $\psi_0 \circ \mathcal{F}_3$ for*

$$\begin{aligned}
\mathcal{F}_1 &:= \left\{ \sum_{j=1}^{m} \mathbf{1}(X_j \leq x) + \theta : x \in \mathbb{R}, \theta \in [l, u] \subset (0, \infty) \right\}, \\
\mathcal{F}_2 &:= \{ m + \theta : \theta \in [l, 2u] \subset (0, \infty) \}, \\
\mathcal{F}_3 &:= \left\{ m - \sum_{j=1}^{m} \mathbf{1}(X_j \leq x) + \theta : x \in \mathbb{R}, \theta \in [l, u] \subset (0, \infty) \right\},
\end{aligned}$$

are P-Donsker.

Proof. Since ψ_0 is continuously differentiable on $(0, \infty)$, it is Lipschitz on $[l, 2u + C]$. Hence, $\psi_0 \circ \mathcal{F}$ is the Lipschitz transformation of $\mathcal{F}$. If $\mathcal{F}$ is a P-Donsker class taking values in

$[l, 2u + C]$, it follows that $\psi_0 \circ \mathcal{F}$ is also P-Donsker by Kosorok (2008, Theorem 9.31).

The class $\mathcal{F}_1$ is the sum of $\left\{\sum_{j=1}^{m} \mathbf{1}(X_j \leq x) : x \in \mathbb{R}\right\}$ and a class of constant functions $\{\theta : \theta \in [l, u]\}$. The first class is P-Donsker by Proposition C.5.8, and the second class is trivially P-Donsker. Hence, $\mathcal{F}_1$ is P-Donsker. The class $\mathcal{F}_2$ is trivially P-Donsker, and $\mathcal{F}_3$ is P-Donsker because it is the sum of the P-Donsker class $\mathcal{F}_1$ and the fixed function m.

The three classes $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ take values in $[l, 2u + C]$. Hence, $\psi_0 \circ \mathcal{F}_1$, $\psi_0 \circ \mathcal{F}_2$, and $\psi_0 \circ \mathcal{F}_3$ are P-Donsker. □

Proposition C.5.10. *Consider the data-generating mechanism described in Section 3.2.1 where the observed data are* $O := (m, X_1, \ldots, X_m) \sim P$. *Assume that* $1 \leq m < C < \infty$ *with probability one. Let* $A \subset (0, \infty)^2$ *be compact and define* s_α *and* s_β *as in* (C.3). *Then, the classes of functions* $\mathcal{F}_\alpha := \{s_\alpha(\cdot; \alpha, \beta) : (\alpha, \beta) \in A\}$ *and* $\mathcal{F}_\beta := \{s_\beta(\cdot; \alpha, \beta) : (\alpha, \beta) \in A\}$ *are* P*-Donsker.*

Proof. Without loss of generality, we assume that $A = [l, u]^2$ where $l > 0$ and $u < \infty$. From (C.3) follows that the classes $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$ are contained in the sum of the following classes of functions:

$$\begin{aligned}
\mathcal{F}_1 &:= \left\{ \psi_0 \left(\sum_{j=1}^{m} \mathbf{1}(X_j \leq x) + \theta \right) : \theta \in [l, u] \right\} \\
\mathcal{F}_2 &:= \{\psi_0(m + \theta) : \theta \in [l, 2u]\} \\
\mathcal{F}_3 &:= \left\{ \psi_0(m - \sum_{j=1}^{m} \mathbf{1}(X_j \leq x) + \theta) : \theta \in [l, u] \right\} \\
\mathcal{F}_4 &:= \{\theta : \theta \in [\psi_0(l), \psi_0(2u)]\},
\end{aligned}$$

where ψ_0 is the digamma function. The classes $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ are P-Donsker by Corollary C.5.9. The class $\mathcal{F}_4$ is trivially P-Donsker. □

Proof of Proposition C.5.2

Proof. Let P denote the distribution of O_i and let Pf denote the expectation of $f(O_i)$ when $O_i \sim P$. We verify that conditions (E), (A), and (D) of Theorem C.5.1 hold.

Conditions (E) and (A). By dominated convergence (which holds by condition (2)), we have that the first and second-order partial derivatives of $\beta(x) \mapsto \Psi(\beta(x), x)$ at $\beta_0(x)$ are given by $\dot{\Psi}(\beta(x), x) = P\dot{\psi}(O_i; \beta(x), x)$ and $\ddot{\Psi}(\beta(x), x) = P\ddot{\psi}(O_i; \beta(x), x)$, respectively. We then have by condition (3) that

$$\sup_{x \in K} \sup_{\beta(x) \in U} \|\dot{\Psi}(\beta(x), x)\| < \infty, \quad \sup_{x \in K} \sup_{\beta(x) \in U} \|\ddot{\Psi}(\beta(x), x)\| < \infty.$$

The above display together with condition (4) is sufficient to apply Lemma C.5.4, which implies that $\beta \mapsto \Psi(\beta)$ is Fréchet differentiable at β_0 with Fréchet derivative $\dot{\Psi}_{\beta_0} : \ell^\infty(K)^p \to \ell^\infty(K)^p$ with a continuous inverse $\dot{\Psi}_{\beta_0}^{-1} : \ell^\infty(K)^p \to \ell^\infty(K)^p$ defined pointwise below, thus satisfying condition (E) of Theorem C.5.1.

$$(\dot{\Psi}_{\beta_0} h)(x) = P\dot{\psi}(O_i; \beta_0(x), x) h(x), \quad \text{for } h \in \ell^\infty(K)^p,$$
$$(\dot{\Psi}_{\beta_0}^{-1} f)(x) = (P\dot{\psi}(O_i; \beta_0(x), x))^{-1} f(x), \quad \text{for } f \in \ell^\infty(K)^p.$$

Condition (A) then follows from Corollary C.5.5.

Condition (D). Let $\beta_n \to \beta_0$ in $\ell^\infty(K)^p$. By condition (3), for all sufficiently large n and all $x \in K$, both $\beta_n(x)$ and $\beta_0(x)$ belong to U. By the mean-value theorem, for each component $j \in \{1, \ldots, p\}$ and each $x \in K$ there exists $\tilde{\beta}_{n,j}(x)$ on the segment between $\beta_n(x)$ and $\beta_0(x)$ such that

$$\psi_j(O_i; \beta_n(x), x) - \psi_j(O_i; \beta_0(x), x) = \dot{\psi}_j(O_i; \tilde{\beta}_{n,j}(x), x)\{\beta_n(x) - \beta_0(x)\}.$$

Because U is convex, each $\tilde{\beta}_{n,j}(x)$ also belongs to U. Hence,

$$\sup_{x\in K} P\left\|\psi(O_i;\beta_n(x),x)-\psi(O_i;\beta_0(x),x)\right\|^2 \leq \left(\sup_{x\in K}\sup_{\beta(x)\in U} P\left\|\dot{\psi}(O_i;\beta(x),x)\right\|^2\right)\|\beta_n-\beta_0\|_\infty^2.$$

The first term of the right-hand side is finite by condition (3), and the second term converges to zero by assumption. Hence, the left-hand side converges to zero, which implies that condition (D) of Theorem C.5.1 holds. □

Proof of Proposition C.5.3

Proof. Because $\hat{\beta}_n$ is uniformly consistent for β_0 on K, we have that $\hat{\beta}_n(x)\in U$ with probability approaching one uniformly on K. We further assume that $\hat{\beta}_n(x)\in U$ for all $x\in K$. Hence, we can apply the mean-value expansion to $G(\alpha;\hat{\beta}_n(x))-G(\alpha;\beta_0(x))$, for $\tilde{\beta}_n(x)$ between $\hat{\beta}_n(x)$ and $\beta_0(x)$.

$$\begin{aligned} n^{1/2}\left(G(\alpha;\hat{\beta}_n(x))-G(\alpha;\beta_0(x))\right) &= n^{1/2}\dot{G}(\alpha;\tilde{\beta}_n(x))\left(\hat{\beta}_n(x)-\beta_0(x)\right) \\ &= n^{1/2}\dot{G}(\alpha;\beta_0(x))\left(\hat{\beta}_n(x)-\beta_0(x)\right) \\ &\quad + n^{1/2}\left(\dot{G}(\alpha;\tilde{\beta}_n(x))-\dot{G}(\alpha;\beta_0(x))\right)\left(\hat{\beta}_n(x)-\beta_0(x)\right) \end{aligned}$$

Note that $n^{1/2}\|\hat{\beta}_n-\beta_0\|_\infty = O_P(1)$ by Theorem C.5.1. We also have that $\|\dot{G}(\alpha;\tilde{\beta}_n(x))-\dot{G}(\alpha;\beta_0(x))\|_\infty = o_P(1)$ because $\dot{G}(\alpha;\beta(x))$ is uniformly continuous in $\beta(x)$ on U by compactness of U and continuity of $\dot{G}(\alpha;\cdot)$ on U. Hence, the second term of the above display is uniformly $o_P(1)$.

Moreover, $\sup_{\beta(x)\in U}\dot{G}(\alpha;\beta(x))<\infty$ since $\dot{G}(\alpha;\beta(x))$ is continuous on the compact set U. The first term of the last display is $-n^{1/2}\dot{G}(\alpha;\beta_0(x))\dot{\Psi}_{\beta_0}^{-1}\Psi_n(\beta_0)(x)+\dot{G}(\alpha;\beta_0(x))r_n(x)$ with $\sup_{x\in K}|r_n(x)|=o_P(1)$ by Theorem C.5.1. Hence, $\sup_{x\in K}\|\dot{G}(\alpha;\beta_0(x))r_n(x)\|=o_P(1)$, which gives the required uniform linearization. □

Appendix D

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

In Appendix D.1, we provide supplemental tables and figures. In Appendix D.2, we discuss post-randomization selection bias and interpretation of the net effect estimand. In Appendix D.3, we provide conditions for pointwise and uniform convergence of the estimated barycenter quantile functions. In Appendix D.4, we show the method applied to additional datasets. In Appendix D.5, we provide additional simulation results.

D.1 Supplemental tables and figures

Table D.1: Base measure configurations for the Dirichlet processes used in Simulation Study 2. All Dirichlet processes were configured with a concentration parameter of 2.

Scenario name	**Group 1 configuration**	**Group 2 configuration**
Null: normal	$N(0,1)$	$N(0,1)$
Null: symmetric bimodal	$50\%\, N(-1.5,0.5) + 50\%\, N(1.5,0.5)$	$50\%\, N(-1.5,0.5) + 50\%\, N(1.5,0.5)$
Null: non-symmetric bimodal	$70\%\, N(-1,0.5) + 30\%\, N(2,0.5)$	$70\%\, N(-1,0.5) + 30\%\, N(2,0.5)$
Mean shift = 0.1	$N(0,1)$	$N(0.1,1)$
Mean shift = 0.2	$N(0,1)$	$N(0.2,1)$
Mean shift = 0.3	$N(0,1)$	$N(0.3,1)$
Upper-tail shift = 1	$N(0,1)$	$95\%\, N(0,1) + 5\%\, N(1,0.25)$
Upper-tail shift = 2	$N(0,1)$	$95\%\, N(0,1) + 5\%\, N(2,0.25)$
Upper-tail shift = 3	$N(0,1)$	$95\%\, N(0,1) + 5\%\, N(3,0.25)$
Bimodal: 0.2	$N(0,1)$	$50\%\, N(-1,0.2) + 50\%\, N(1,0.2)$
Bimodal: 0.4	$N(0,1)$	$50\%\, N(-1,0.4) + 50\%\, N(1,0.4)$
Bimodal: 0.6	$N(0,1)$	$50\%\, N(-1,0.6) + 50\%\, N(1,0.6)$

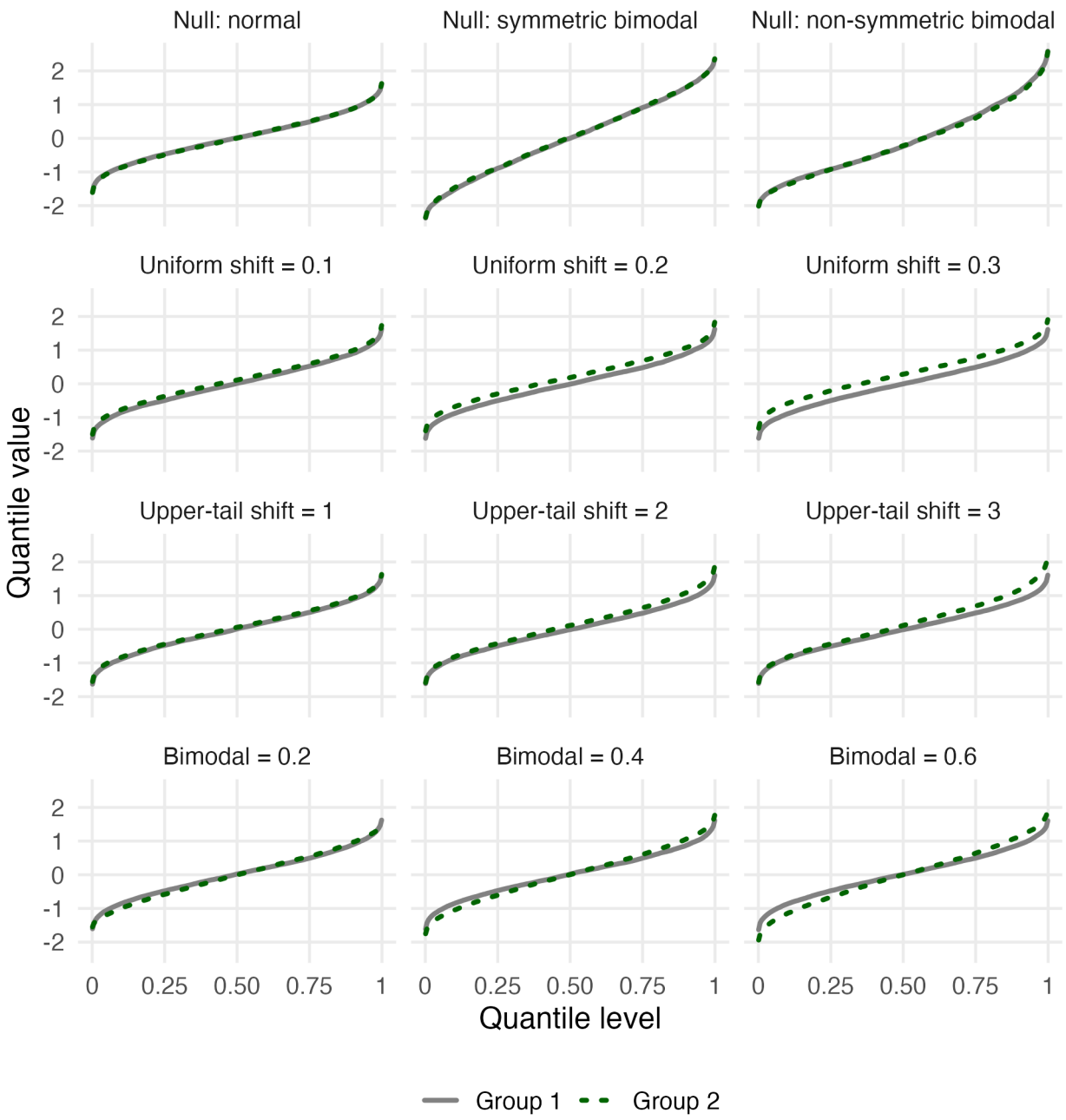


Figure D.1: Wasserstein barycenter quantile functions for Group 1 (solid gray) and Group 2 (dotted green) for the conditions studied in Simulation Study 2.

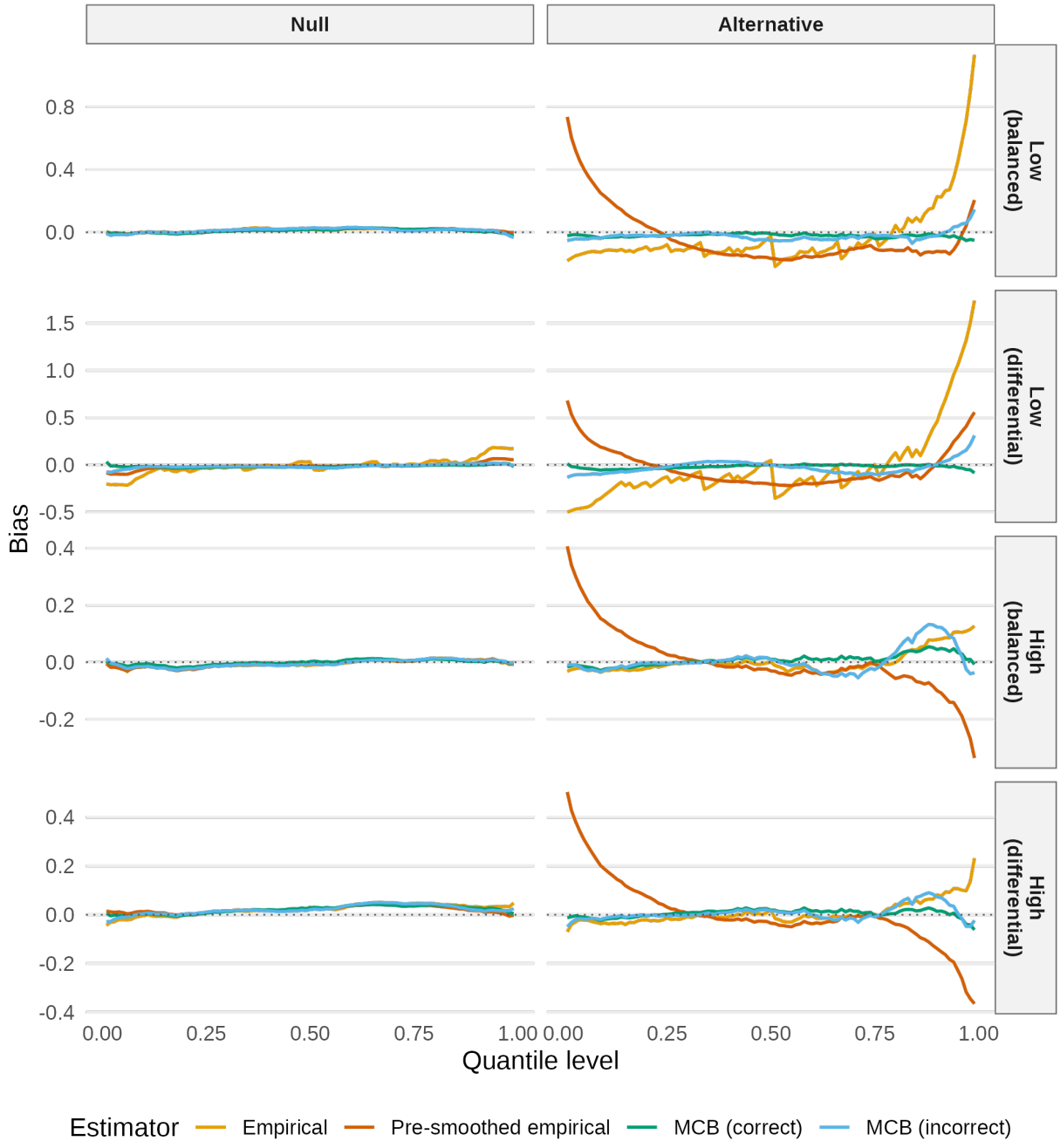


Figure D.2: Bias of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth settings: low balanced, low differential, high balanced, and high differential. Each curve shows the bias of $\widehat{\Delta}(u)$ across quantile levels $u \in 0.01, 0.02, \ldots, 0.99$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

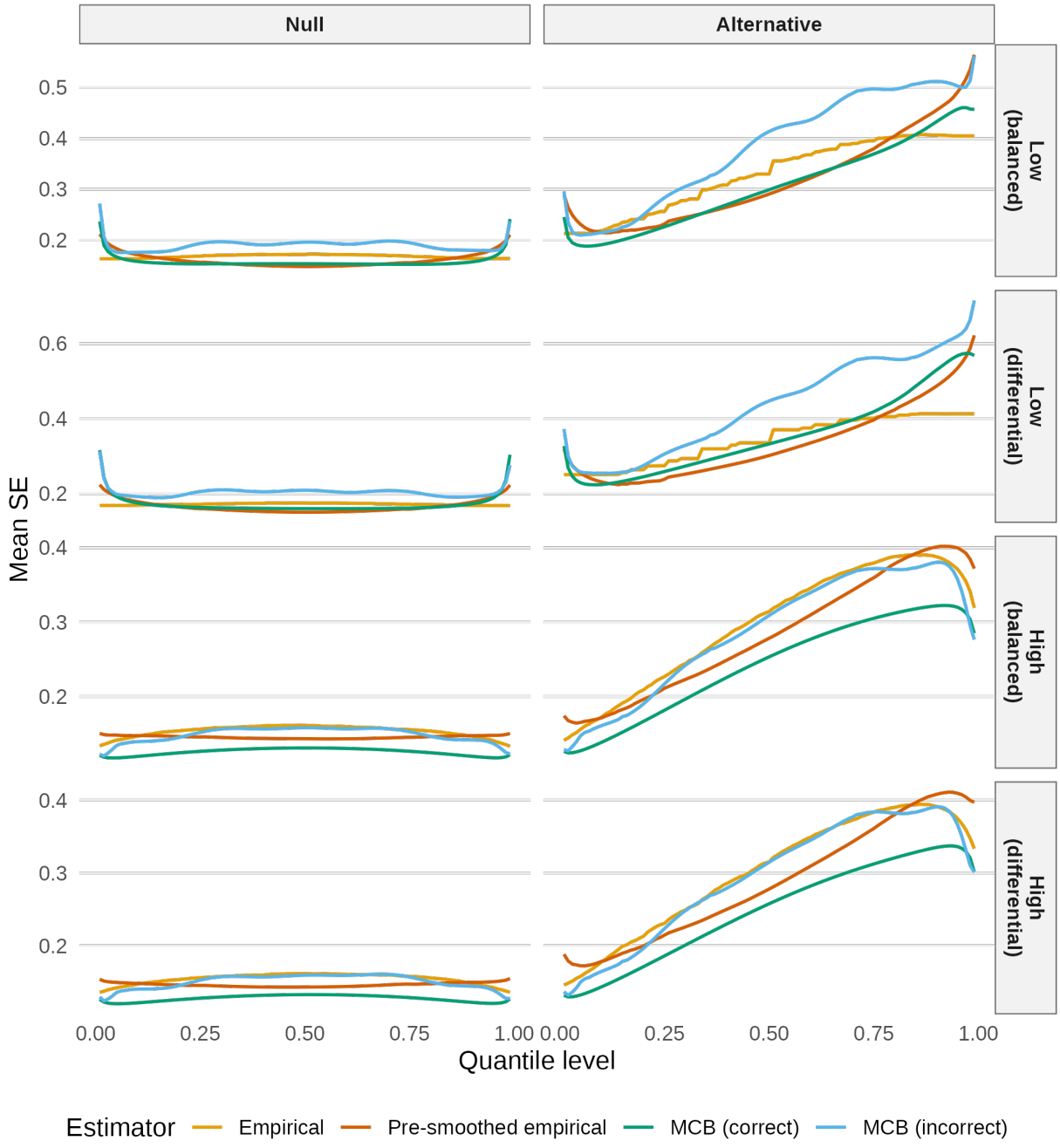


Figure D.3: Mean standard error of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth settings: low balanced, low differential, high balanced, and high differential. Each curve shows the mean standard error of $\widehat{\Delta}(u)$ across quantile levels $u \in 0.01, 0.02, \ldots, 0.99$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

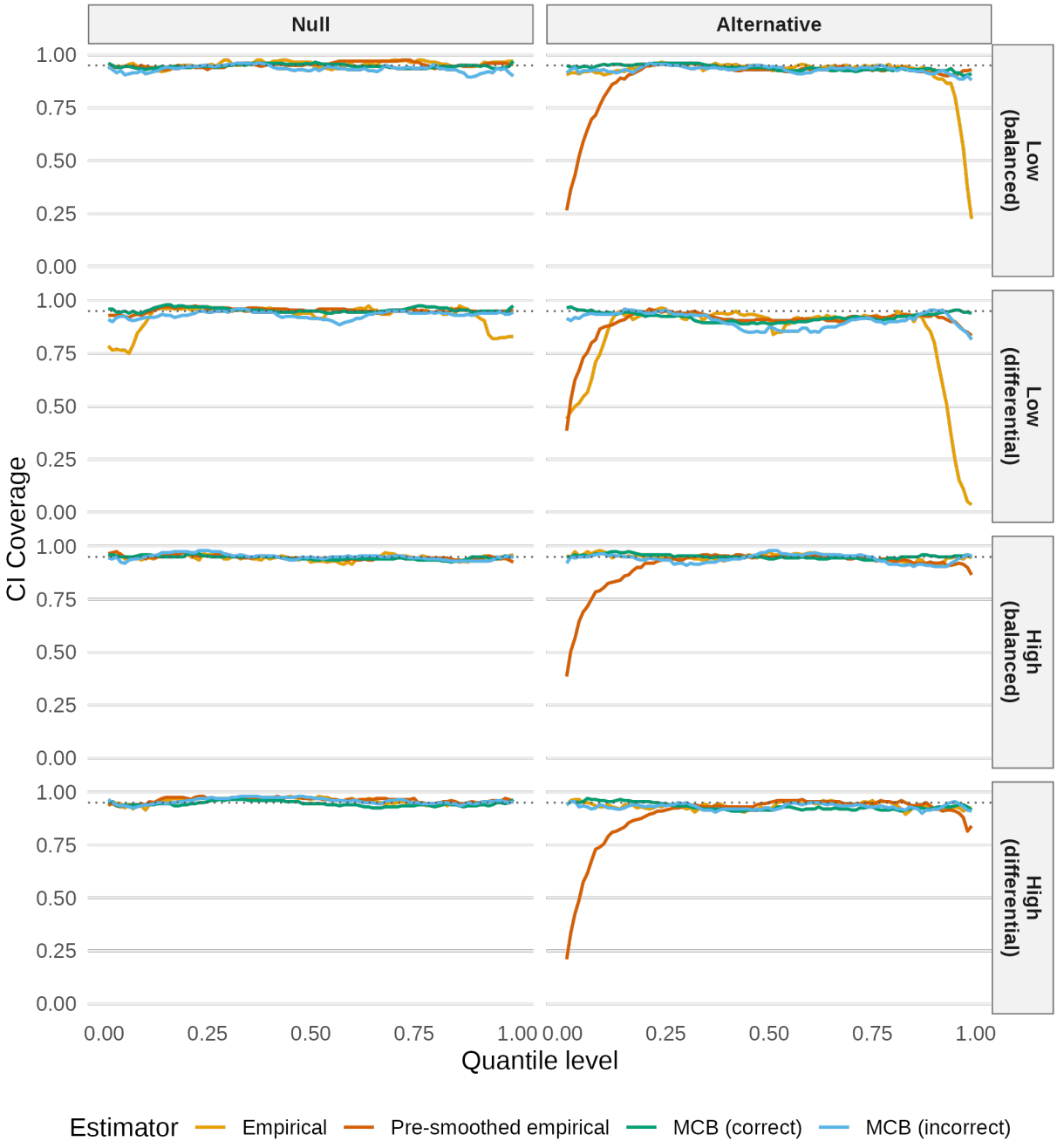


Figure D.4: 95% confidence interval coverage of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth settings: low balanced, low differential, high balanced, and high differential. Each curve shows the 95% confidence interval coverage of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

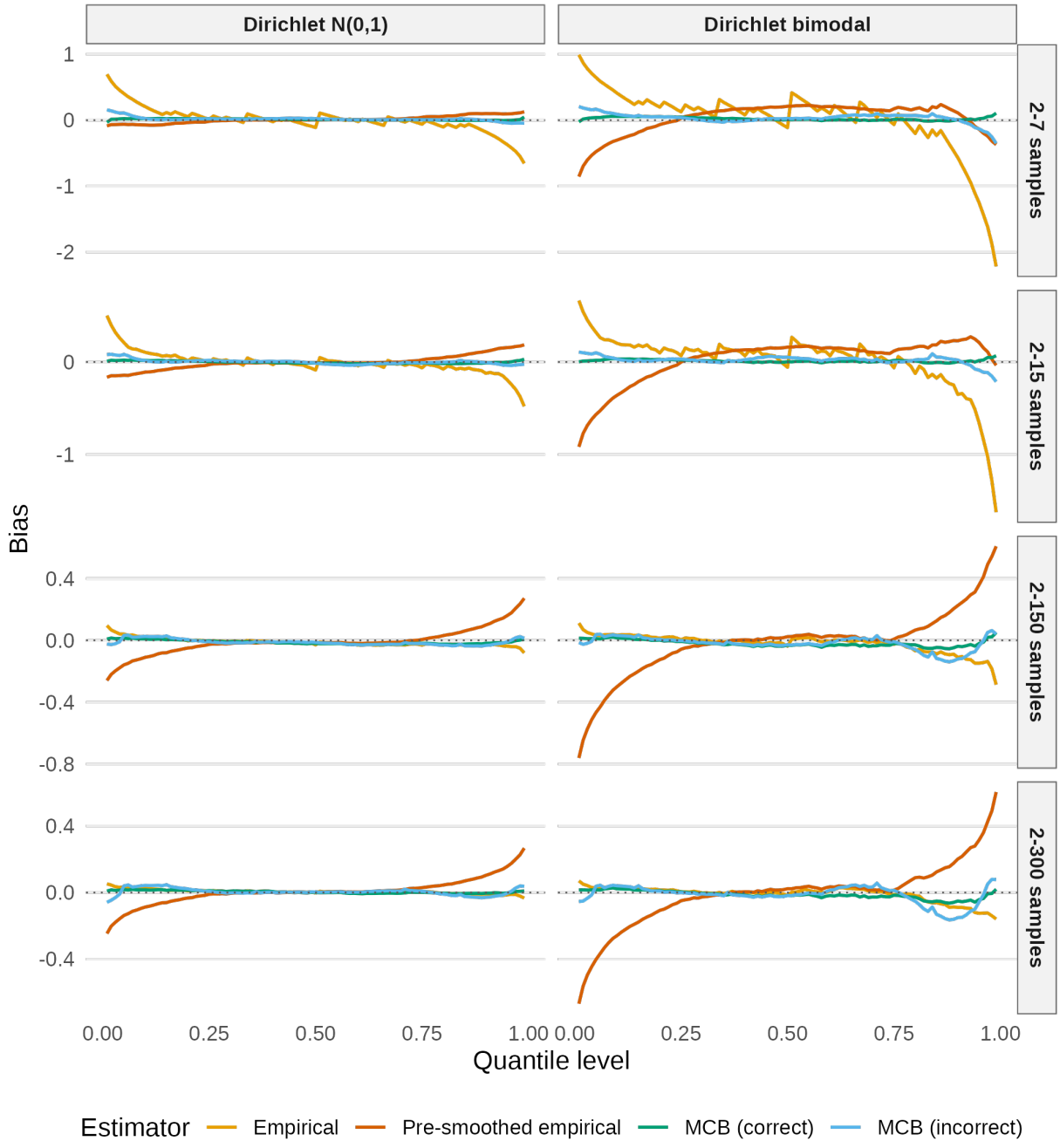


Figure D.5: Bias of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the bias of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

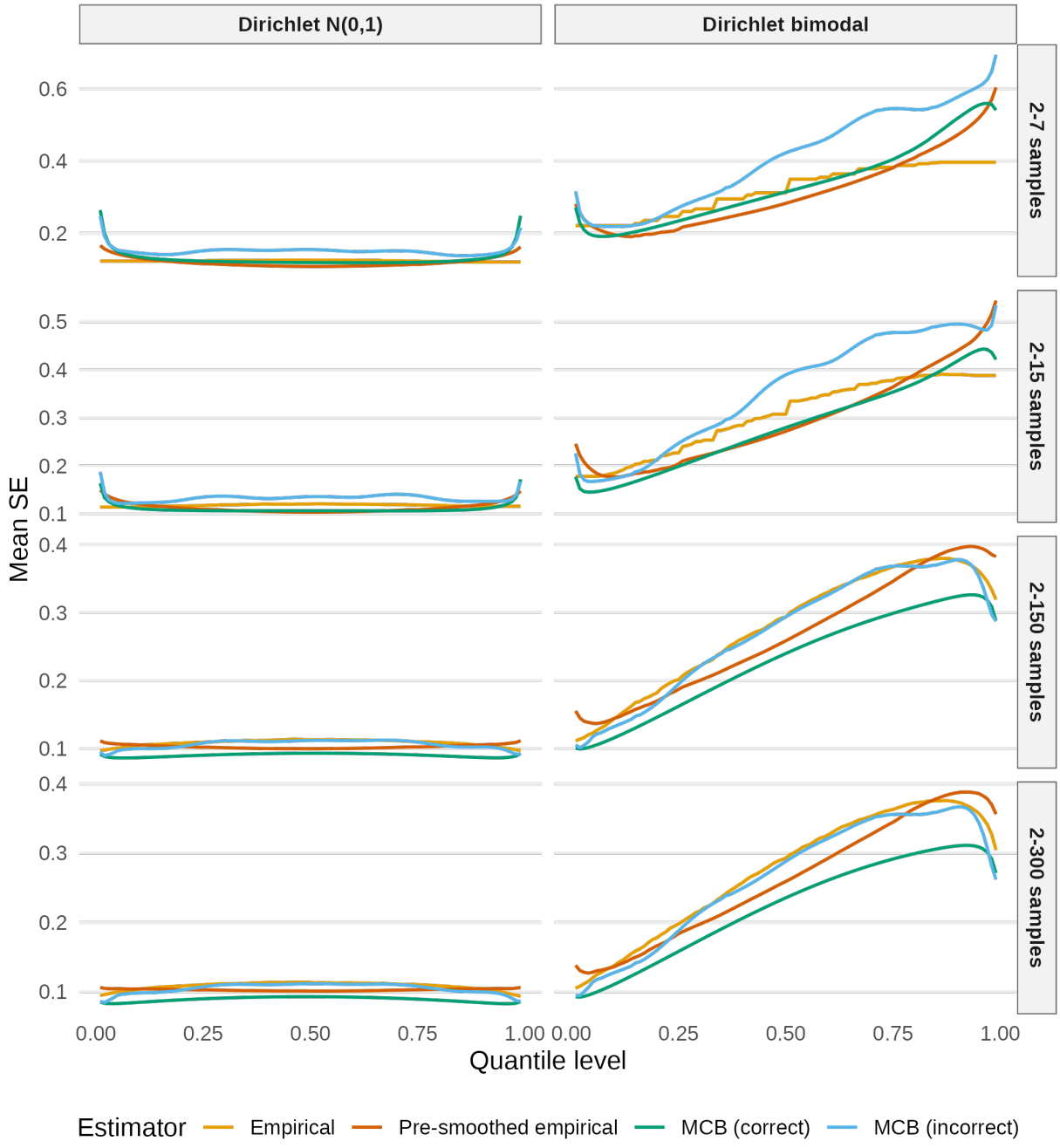


Figure D.6: Mean standard error of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the mean standard error of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

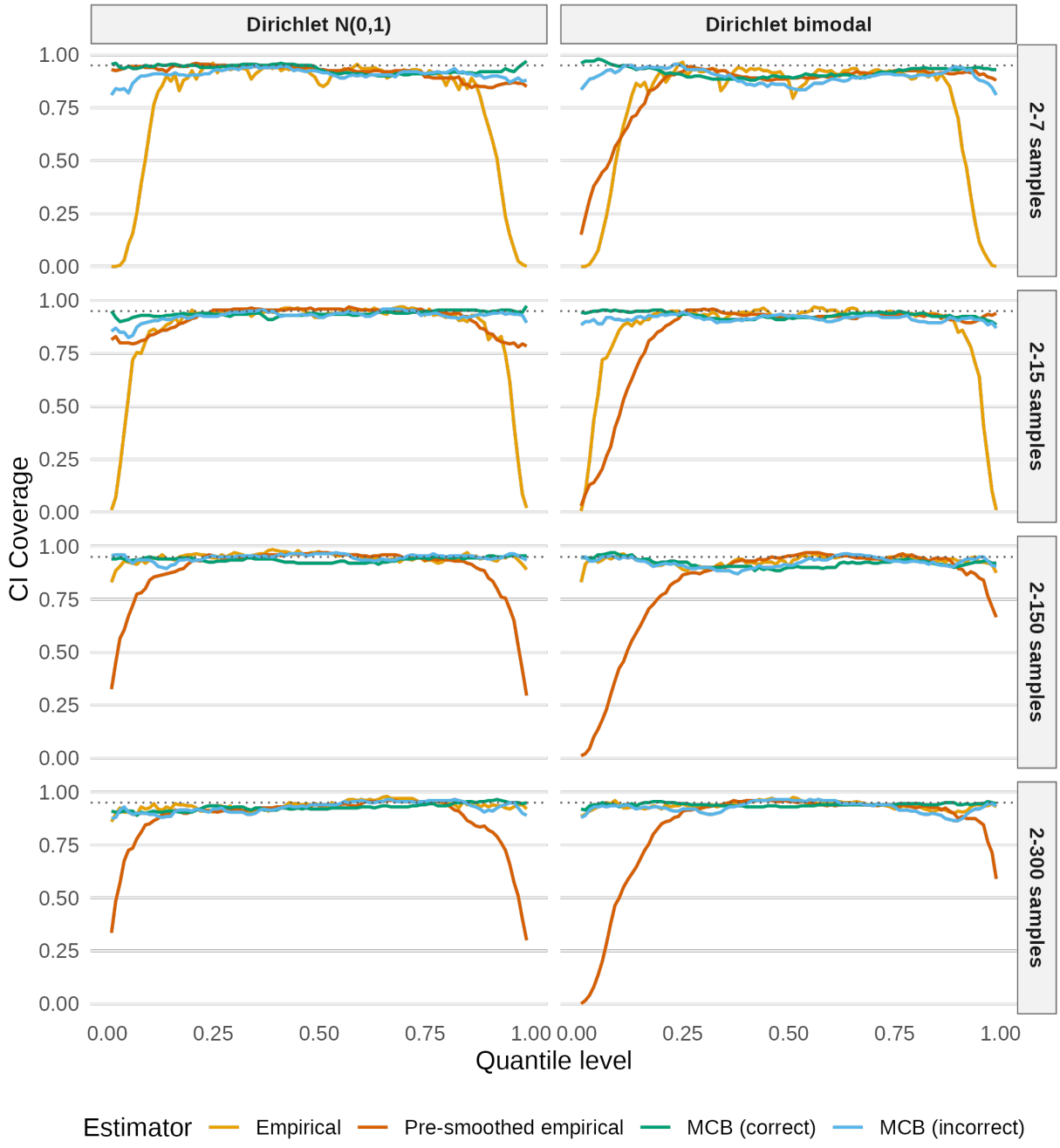


Figure D.7: 95% confidence interval coverage of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the 95% confidence interval coverage of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

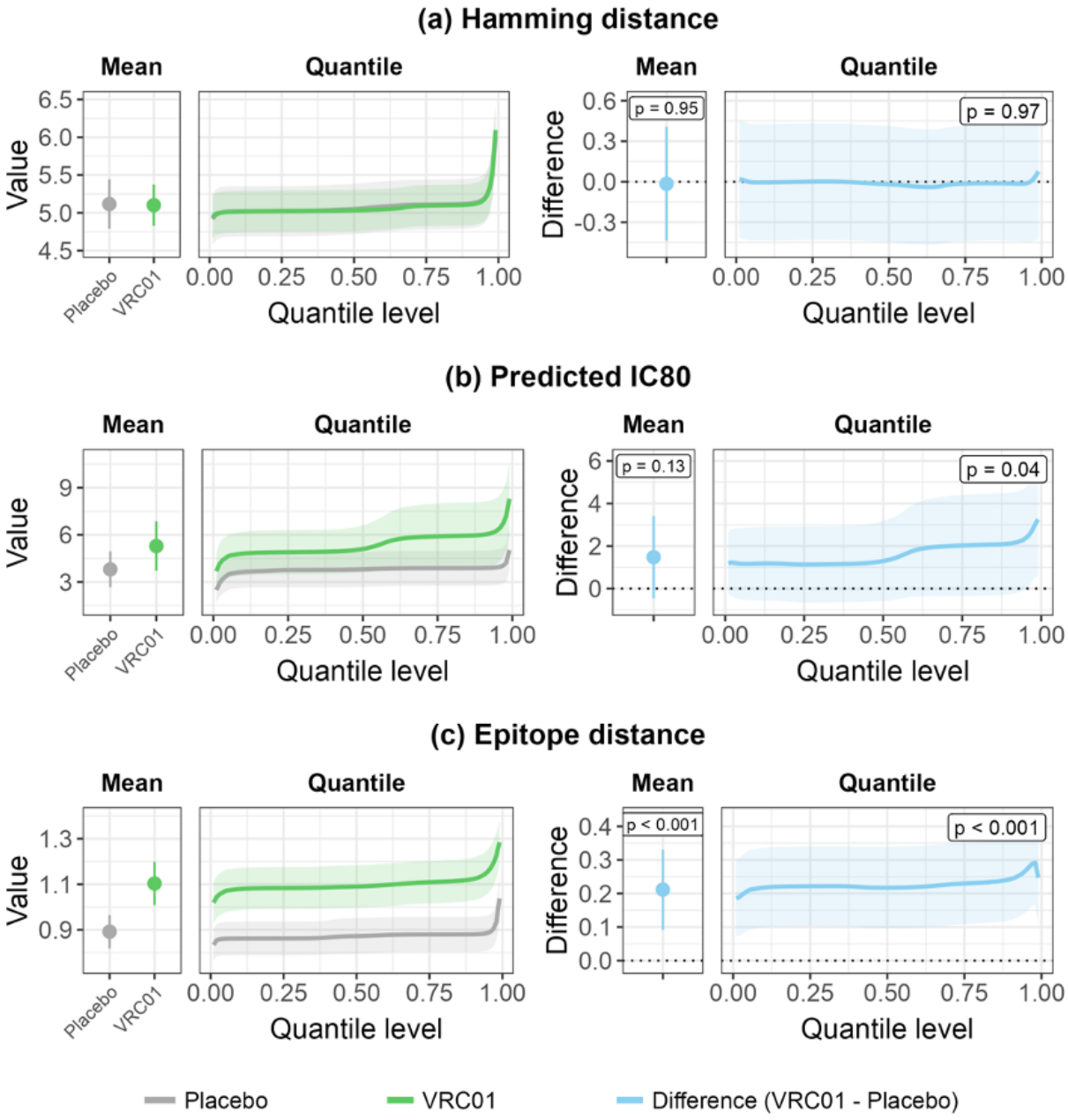


Figure D.8: Using the spline-configured MCB estimator, analysis of three VRC01 susceptibility-associated features among AMP breakthrough infections: (A) Hamming distance to a VRC01-sensitive reference sequence at pre-selected sites, (B) predicted IC80, and (C) epitope distance. For each feature, the left panels show the arm-specific mean summaries and estimated barycenter quantile functions for the placebo arm and the pooled VRC01 arms. The right panels show the corresponding VRC01-minus-placebo estimated mean difference and quantile difference map. Shaded bands indicate pointwise 95% confidence intervals. Reported p-values correspond to the mean-based unequal-variance two-sample t-test for the mean panels and the supremum test for equality of barycenters for the quantile panels.

Table D.2: BCR data application vaccination schedules, PBMC sampling timepoints, number of participants, and median (IQR) number of sequences per participant.

Analysis label	Vaccination schedule	PBMC sample	Unique participants	Median sequences (IQR)
wk8 eOD	eOD at week 0	Week 8	16	34 (21.8, 61.2)
wk16 eOD→eOD	eOD at week 0; eOD at week 8	Week 16	24	101.5 (53, 262.8)
wk16 eOD→core	eOD at week 0; core at week 8	Week 16	16	121 (62.8, 318.5)
wk24 eOD→eOD	eOD at week 0; eOD at week 8	Week 24	13	160 (53, 187)
wk24 eOD→core	eOD at week 0; core at week 8	Week 24	17	265 (36, 494)
wk24 eOD→eOD→core	eOD at week 0; eOD at week 8; core at week 16	Week 24	13	217 (143, 593)

D.2 Interpretation of post-infection estimand

In the main text, we consider the difference between the treatment arm-specific barycenter quantile functions, which we refer to in this Appendix as the *net sieve contrast*. This contrast compares the feature distributions among cases under treatment assignment with those among cases under placebo assignment. In this Appendix, we show that, under standard causal assumptions, the net sieve contrast can be decomposed into a *breakthrough/post-acquisition component* and a *prevention sieve contrast*, which each have a distinct biological interpretation. These individual components cannot generally be identified without additional assumptions (Frangakis and Rubin, 2002; Gilbert et al., 2003; Shepherd et al., 2006).

D.2.1 Preliminaries

Whereas the data setup in Chapter 4 implicitly conditions on infection, in this Appendix we make this conditioning explicit. Let $Z_i \in \{0, 1\}$ denote randomized treatment assignment, where $Z_i = 1$ indicates assignment to treatment and $Z_i = 0$ indicates assignment to placebo. Let $E_i \in \{0, 1\}$ denote the observed infection indicator, where $E_i = 1$ indicates that participant i acquired HIV-1 during follow-up.

Let $E_i(1)$ and $E_i(0)$ denote the potential infection indicators under treatment and placebo assignment, respectively. Let $Y_i(z)$ denote the potential within-host distribution of a numeric sequence feature under assignment z. Because a viral feature distribution is defined only following infection, $Y_i(z)$ is defined only when $E_i(z) = 1$. For each $u \in (0, 1)$, define

$$q_i(z, u) := F^-_{Y_i(z)}(u),$$

where $q_i(z, u)$ is the uth quantile of participant i's potential within-host feature distribution under assignment z.

Assumption D.2.1 (Randomization)**.** *Treatment assignment is independent of the potential*

infection indicators and potential within-host feature distributions:

$$Z_i \perp\!\!\!\perp \{E_i(0), E_i(1), Y_i(0), Y_i(1)\}.$$

In addition, both treatment assignments occur with positive probability:

$$0 < P(Z_i = 1) < 1.$$

Assumption D.2.2 (Consistency)**.** *The observed infection indicator equals the potential infection indicator under the assigned treatment:*

$$E_i = Z_i E_i(1) + (1 - Z_i) E_i(0).$$

Among participants with $E_i = 1$, the observed within-host feature distribution equals the corresponding potential distribution:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

Assumption D.2.3 (Monotonicity)**.** *Treatment cannot cause infection in a participant who would not have been infected under placebo:*

$$E_i(1) \leq E_i(0).$$

Under randomization and consistency, the treatment arm-specific barycenter quantile functions defined in the main text can be written as

$$\mu_z(u) = \mathbb{E}\{q_i(z, u) \mid E_i(z) = 1\}, \qquad z \in \{0, 1\}.$$

Therefore, the net sieve contrast is

$$\begin{aligned}\Delta(u) &= \mu_1(u) - \mu_0(u) \\ &= \mathbb{E}\{q_i(1,u) \mid E_i(1) = 1\} - \mathbb{E}\{q_i(0,u) \mid E_i(0) = 1\}.\end{aligned}$$

This contrast compares the mean uth quantile of the within-host feature distributions among infections that would occur under treatment with the corresponding mean among infections that would occur under placebo.

Define the principal stratum G_i by the joint potential infection indicators $\{E_i(1), E_i(0)\}$. Under monotonicity, participants belong to one of three principal strata. The always-infected stratum, denoted $G_i = II$, contains participants for whom $E_i(1) = E_i(0) = 1$. The protected stratum, denoted $G_i = PI$, contains participants for whom $E_i(1) = 0$ and $E_i(0) = 1$. The never-infected stratum, denoted $G_i = NN$, contains participants for whom $E_i(1) = E_i(0) = 0$. Monotonicity rules out participants who would be infected under treatment but not under placebo.

D.2.2 Decomposition

Under monotonicity, all participants who would be infected under treatment belong to the always-infected stratum. In contrast, participants who would be infected under placebo are a mixture of the always-infected and protected strata. Let

$$\rho := P\{G_i = PI \mid E_i(0) = 1\}$$

denote the proportion of infections that would occur under placebo that would be prevented under treatment. Under randomization, consistency, and monotonicity, this quantity is identified by

$$\rho = 1 - \frac{P(E_i = 1 \mid Z_i = 1)}{P(E_i = 1 \mid Z_i = 0)}.$$

The placebo barycenter quantile function can then be expressed as the mixture

$$\mu_0(u) = (1-\rho)\mathbb{E}\{q_i(0,u) \mid G_i = II\} + \rho\mathbb{E}\{q_i(0,u) \mid G_i = PI\},$$

whereas the treatment barycenter quantile function is

$$\mu_1(u) = \mathbb{E}\{q_i(1,u) \mid G_i = II\}.$$

It follows that the net sieve contrast can be decomposed as

$$\Delta(u) = \underbrace{\mathbb{E}\{q_i(1,u) - q_i(0,u) \mid G_i = II\}}_{\text{breakthrough/post-acquisition component}} + \underbrace{\rho\left[\mathbb{E}\{q_i(0,u) \mid G_i = II\} - \mathbb{E}\{q_i(0,u) \mid G_i = PI\}\right]}_{\text{prevention sieve component}}.$$

The two components have the following biological interpretations:

- *Breakthrough/post-acquisition component.* The first component is a causal contrast within the always-infected principal stratum. It describes how treatment changes the uth quantile of the within-host feature distribution among participants who would become infected under either assignment. When viral samples are collected shortly after infection, this component may reflect founder virus selection at acquisition and/or early post-acquisition viral evolution.
- *Prevention sieve component.* The second component describes the contribution from selective prevention, i.e. a prevention sieve effect. The contrast

  $$\mathbb{E}\{q_i(0,u) \mid G_i = II\} - \mathbb{E}\{q_i(0,u) \mid G_i = PI\}$$

  compares, under placebo assignment, the feature distributions of infections that would occur despite treatment with those of infections that would be prevented. Multiplying

this contrast by ρ accounts for the proportion of placebo infections that treatment would prevent.

The two components of this decomposition are not generally identified separately from the observed trial data because principal stratum membership is not observed among infected placebo recipients. From the case-only contrast alone, we cannot determine how much of the observed difference is due to the selective prevention of infections with particular viral features and how much is due to breakthrough or post-acquisition effects among participants who would be infected under either assignment.

The decomposition also suggests several possible scenarios. If treatment has no prevention efficacy, so that $\rho = 0$, then the prevention sieve component is zero and the net sieve contrast equals the breakthrough/post-acquisition component. Conversely, the net sieve contrast may be nonzero even when the breakthrough/post-acquisition component is zero if treatment selectively prevents infections with particular feature distributions. When both components are nonzero, they reinforce one another if they point in the same direction and partially cancel if they point in opposite directions. In many settings, it is reasonable to expect the two components to point in the same direction. For example, a treatment that preferentially prevents infections by viruses with low Hamming distance may also reduce the fitness of low-Hamming-distance viruses following breakthrough infection. Thus, although the net sieve contrast is not a causal effect within a fixed principal stratum, it remains a useful estimand because it summarizes the overall distributional difference among infections under treatment versus placebo, incorporating both the prevention sieve component and the breakthrough/post-acquisition component.

D.3 Conditions for pointwise and uniform convergence of the estimated barycenter quantile function

In the following, without loss of generality, suppose that we have unobserved realizations of distributional random variable Y denoted as $Y_1, ..., Y_n$. Let $\widehat{Y}_i^{\text{emp}}$ denote the observed empirical distribution for individual i and $\widehat{Y}_i^{\text{sm}}$ denote the pre-smoothed empirical distribution. Suppose that our target is the Wasserstein barycenter $\mathbb{B}[Y]$, as its quantile function $F^-_{\mathbb{B}[Y]}(u)$. In the following, we give conditions for pointwise asymptotic normality and Gaussian process convergence for the quantile function of the empirical barycenter $\widehat{\mu}^{\text{emp}}(u) := \frac{1}{n}\sum_{i=1}^n F^-_{\hat{Y}_i^{emp}}(u)$ and the pre-smoothed empirical barycenter $\widehat{\mu}^{\text{sm}}(u) := \frac{1}{n}\sum_{i=1}^n F^-_{\hat{Y}_i^{sm}}(u)$. The asymptotic study of the marginal-constructed barycenter $\widehat{\mu}^{\text{mcb}}(u)$ is presented in Chapter 3.

Both empirical estimators can be handled using the same argument, because both are obtained by first estimating each individual-level quantile function and then averaging these estimated quantile functions across individuals.

Let

$$q_i(u) := F^-_{Y_i}(u), \qquad \mu(u) := F^-_{\mathbb{B}[Y]}(u) = \mathbb{E}\{q_i(u)\}, \qquad u \in (0,1),$$

where the last equality follows from the one-dimensional representation of the Wasserstein barycenter. For $a \in \{\text{emp}, \text{sm}\}$, define

$$\widehat{q}_i^a(u) := F^-_{\widehat{Y}_i^a}(u), \qquad \widehat{\mu}^a(u) := \frac{1}{n}\sum_{i=1}^n \widehat{q}_i^a(u), \qquad u \in (0,1),$$

where $a = \text{emp}$ corresponds to the empirical distribution $\widehat{Y}_i^{\text{emp}}$, and $a = \text{sm}$ corresponds to the pre-smoothed empirical distribution $\widehat{Y}_i^{\text{sm}}$. Define the average first-stage quantile estimation error

$$R_n^a(u) := \frac{1}{n}\sum_{i=1}^n \left\{\widehat{q}_i^a(u) - q_i(u)\right\}.$$

The role of the following conditions is to require that the first-stage error from estimating the individual-level quantile functions is asymptotically negligible relative to the $n^{-1/2}$

variation from averaging the latent quantile functions q_i. These conditions may hold for either the empirical or pre-smoothed estimator when the within-individual sample sizes are sufficiently large with respect to the number of subjects and the smoothing procedure, if used, has sufficiently small bias.

Proposition D.3.1 (Pointwise asymptotic normality of empirical and pre-smoothed empirical barycenters)**.** *Fix $u \in (0,1)$ and let $a \in \{\text{emp}, \text{sm}\}$. Suppose that the following conditions hold:*

1. *The individual-level quantile value has finite, nonzero variance:*

$$0 < \text{Var}\{q_i(u)\} < \infty.$$

2. *The average first-stage quantile estimation error is negligible at the $n^{-1/2}$ scale:*

$$\sqrt{n}\, R_n^a(u) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\{\widehat{q}_i^a(u) - q_i(u)\}\right] = o_p(1).$$

Then

$$\sqrt{n}\,\{\widehat{\mu}^a(u) - \mu(u)\} \rightsquigarrow N\left(0, \text{Var}\{q_i(u)\}\right).$$

Equivalently, both $\widehat{\mu}^{\text{emp}}(u)$ and $\widehat{\mu}^{\text{sm}}(u)$ are pointwise asymptotically normal whenever their corresponding first-stage estimation errors satisfy Condition 2.

Proof. For either $a = \text{emp}$ or $a = \text{sm}$, add and subtract the average of the latent individual-level quantile functions:

$$\begin{aligned}\widehat{\mu}^a(u) - \mu(u) = &\frac{1}{n}\sum_{i=1}^{n}\{q_i(u) - \mu(u)\} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\{\widehat{q}_i^a(u) - q_i(u)\}.\end{aligned}$$

Since $\mu(u) = \mathbb{E}\{q_i(u)\}$, the first term satisfies

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\{q_i(u) - \mu(u)\}\right] \rightsquigarrow N\left(0, \mathrm{Var}\{q_i(u)\}\right)$$

by the central limit theorem. By Condition 2, the second term is $o_p(1)$ after multiplication by $\sqrt{n}$. The result follows by Slutsky's theorem. □

Proposition D.3.2 (Weak convergence of empirical and pre-smoothed empirical barycenters)**.** *Let $\mathcal{U} = [\epsilon, 1-\epsilon]$ for some $\epsilon \in (0, 1/2)$, and let $a \in \{\mathrm{emp}, \mathrm{sm}\}$. Define the latent empirical process*

$$\mathbb{Z}_n(u) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{q_i(u) - \mathbb{E}q_i(u)\}, \qquad u \in \mathcal{U}.$$

Suppose that the following conditions hold:

1. *The latent quantile functions q_i are i.i.d. random elements of $\ell^\infty(\mathcal{U})$ with finite second moment in sup-norm:*

$$\mathbb{E}\left[\|q_i\|_{\infty,\mathcal{U}}^2\right] < \infty, \qquad \|q_i\|_{\infty,\mathcal{U}} := \sup_{u\in\mathcal{U}}|q_i(u)|.$$

2. *The latent empirical process $\mathbb{Z}_n$ is stochastically equicontinuous over $\mathcal{U}$. That is, for every $\varepsilon > 0$,*

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty} P\left(\sup_{\substack{u,v\in\mathcal{U}\\ |u-v|<\delta}}|\mathbb{Z}_n(u) - \mathbb{Z}_n(v)| > \varepsilon\right) = 0.$$

3. *The average first-stage quantile estimation error is negligible uniformly over $\mathcal{U}$:*

$$\sqrt{n}\,\|R_n^a\|_{\infty,\mathcal{U}} = o_p(1).$$

Then

$$\sqrt{n}\,\{\widehat{\mu}^a - \mu\} \rightsquigarrow \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{U}),$$

where $\mathbb{G}$ is a tight mean-zero Gaussian process with covariance kernel

$$\Sigma(u, v) = \mathrm{Cov}\{q_i(u), q_i(v)\}, \qquad u, v \in \mathcal{U}.$$

Proof. For either $a = \mathrm{emp}$ or $a = \mathrm{sm}$,

$$\begin{aligned}\widehat{\mu}^a - \mu &= \frac{1}{n}\sum_{i=1}^{n}\{q_i - \mu\} + R_n^a \\ &= \frac{1}{n}\sum_{i=1}^{n}\{q_i - \mathbb{E}(q_i)\} + R_n^a.\end{aligned}$$

Multiplying by $\sqrt{n}$ gives

$$\sqrt{n}\,\{\widehat{\mu}^a - \mu\} = \mathbb{Z}_n + \sqrt{n}R_n^a.$$

For any finite collection $u_1, \ldots, u_k \in \mathcal{U}$, the random vectors

$$(q_i(u_1) - \mathbb{E}q_i(u_1), \ldots, q_i(u_k) - \mathbb{E}q_i(u_k))^\top$$

are i.i.d. mean-zero random vectors with finite second moments by Condition 1. Therefore, by the multivariate central limit theorem,

$$(\mathbb{Z}_n(u_1), \ldots, \mathbb{Z}_n(u_k))^\top \rightsquigarrow N_k(0, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma}_{j\ell} = \mathrm{Cov}\{q_i(u_j), q_i(u_\ell)\}, \qquad 1 \le j, \ell \le k.$$

For each fixed collection $u_1, \ldots, u_k \in \mathcal{U}$, $(\mathbb{Z}_n(u_1), \ldots, \mathbb{Z}_n(u_k))^\top$ converges to a mean-zero multivariate Gaussian vector with covariance matrix $\{\Sigma(u_j, u_\ell)\}_{j,\ell=1}^k$.

Condition 2 gives stochastic equicontinuity of $\mathbb{Z}_n$ over $\mathcal{U}$. Under the usual asymptotic measurability conditions for empirical processes, finite-dimensional convergence together

with stochastic equicontinuity yields

$$\mathbb{Z}_n \rightsquigarrow \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{U}),$$

where $\mathbb{G}$ is a tight mean-zero Gaussian process with covariance kernel Σ (Billingsley, 2013).

By Condition 3,

$$\sqrt{n}\,\|R_n^a\|_{\infty,\mathcal{U}} = o_p(1).$$

Therefore,

$$\sqrt{n}\,\{\widehat{\mu}^a - \mu\} = \mathbb{Z}_n + o_p(1)$$

in $\ell^\infty(\mathcal{U})$. The result follows by Slutsky's theorem in $\ell^\infty(\mathcal{U})$. □

Remark D.3.1 (Interpretation of the first-stage negligibility conditions)**.** *The empirical and pre-smoothed empirical barycenters differ only through the first-stage estimate $\widehat{q}_i^a$ of the latent quantile function q_i. The propositions above therefore separate the asymptotic variation into two parts: the between-individual variation in the latent quantile functions, and the within-individual first-stage estimation error. When the first-stage error is $o_p(n^{-1/2})$, either pointwise or uniformly over $\mathcal{U}$, it does not contribute to the first-order limiting distribution. In sparse sequencing settings, this negligibility condition may fail, especially in the tails; in that case the empirical or pre-smoothed empirical barycenter may be biased for the target barycenter quantile function.*

D.4 Analysis of additional datasets

In this section, we present additional analyses of HIV-1 sequence data. First, we analyze additional sequence-derived features from the AMP trial (HVTN 703/704) that were not included in the main text. We then analyze three additional HIV-1 prevention trials: Imbokodo (HVTN 705), Phambili (HVTN 503), and Uhambo (HVTN 702).

In each trial, participants were randomized to active treatment or placebo, and HIV-1 diagnosis was the endpoint. Among participants who reached the endpoint, we analyzed available HIV-1 sequence data and compared sequence-derived features between treatment arms. For each feature, we report unadjusted p-values from the proposed distributional analysis and from a mean-based analysis based on averaging feature values within each individual.

D.4.1 Imbokodo (HVTN 705)

Table D.3: P-values from distributional and mean-based two-sample analyses in the Imbokodo (HVTN 705) data. Results are shown for each sequence-derived feature. The empirical and MCB columns report p-values from the proposed Wasserstein barycenter-based tests, while the mean-based analysis column reports p-values from unequal variance two-sample t-tests applied after averaging the feature values within each individual.

Variable	Empirical	MCB	Mean-based analysis
`hdist.zspace.c97za.hvtn505.cd4bs.kmer`	0.003	0.008	0.006
`hdist.zspace.c97za.c_ab`	0.009	0.007	0.007
`hdist.zspace.mos2.hvtn505.cd4bs.kmer`	0.027	0.037	0.021
`hdist.zspace.mos1.hvtn505.cd4bs.kmer`	0.059	0.044	0.046
`hdist.zspace.c97za.c_ab_no364`	0.090	0.084	0.061
`hdist.zspace.mos2.c_ab_no364`	0.110	0.085	0.086
`hdist.zspace.mos1.v5`	0.168	0.173	0.118
`hdist.zspace.c97za.v5`	0.168	0.140	0.118
`hdist.zspace.mos1.hvtn505.cd4bs.antibody`	0.172	0.148	0.115
`hdist.zspace.mos2.c_ab`	0.192	0.234	0.158
`hdist.zspace.c97za.hvtn505.cd4bs.antibody`	0.196	0.122	0.129
`num.sequons.v1v2`	0.205	0.462	0.334

Continued on next page

Table D.3

Variable	Empirical	MCB	Mean-based analysis
hdist.zspace.mos1.c_ab	0.230	0.222	0.194
nonconserved.cysteine.count	0.297	0.361	0.216
hdist.zspace.c97za.gp41	0.307	0.446	0.299
hdist.zspace.mos1.gp120	0.316	0.262	0.206
hdist.zspace.mos1.gp41	0.333	0.257	0.151
hdist.zspace.mos2.gp120	0.394	0.295	0.231
hdist.zspace.mos2.v2_ab	0.454	0.438	0.469
hdist.zspace.mos2.gp41	0.461	0.485	0.879
hdist.zspace.mos2.hvtn505.cd4bs.antibody	0.481	0.366	0.365
hdist.zspace.c97za.v2	0.531	0.472	0.480
hdist.zspace.c97za.gp120	0.537	0.322	0.363
hdist.zspace.mos1.v2_ab	0.565	0.425	0.622
cysteine.count	0.573	0.805	0.356
hdist.zspace.mos2.v2	0.594	0.670	0.522
num.sequons.v5	0.628	0.603	0.511
charge.v2	0.633	0.702	0.702
hdist.zspace.mos1.v2	0.692	0.578	0.622
hdist.zspace.c97za.adcc	0.737	0.182	0.793
hdist.zspace.1428v1v2	0.753	0.728	0.794
length.v5	0.753	0.829	0.998
hdist.zspace.mos1.adcc	0.784	0.553	0.795
hdist.zspace.mos2.adcc	0.786	0.529	0.795
length.v1v2	0.810	0.496	0.740
hdist.zspace.mos2.v5	0.819	0.607	0.835
hdist.zspace.mos1.c_ab_no364	0.830	0.892	0.719
hdist.zspace.c97za.v2_ab	0.906	0.880	0.906

D.4.2 Phambili (HVTN 503)

Table D.4: P-values from distributional and mean-based two-sample analyses in the Phambili (HVTN 503) data. Results are shown for each sequence-derived feature. The empirical and MCB columns report p-values from the proposed Wasserstein barycenter-based tests, while the mean-based analysis column reports p-values from unequal variance two-sample t-tests applied after averaging the feature values within each individual.

Variable	Empirical	MCB	Mean-based analysis
`hdist.raw.nef`	0.016	0.022	0.019
`hdist.zspace.nef`	0.022	0.045	0.025
`hdist.zspace.gag`	0.038	0.024	0.052
`hdist.raw.gag`	0.049	0.038	0.057
`hdist.zspace.pol`	0.632	0.380	0.339
`hdist.zspace.slyntvatl`	0.687	0.668	0.683
`hdist.raw.pol`	0.695	0.097	0.475
`hdist.raw.slyntvatl`	0.827	0.813	0.823

D.4.3 AMP (HVTN 703/704)

Table D.5: P-values from distributional and mean-based two-sample analyses in the AMP (HVTN 703/704) data. Results are shown for each sequence-derived feature. The empirical and MCB columns report p-values from the proposed Wasserstein barycenter-based tests, while the mean-based analysis column reports p-values from unequal variance two-sample t-tests applied after averaging the feature values within each individual.

Variable	Empirical	MCB	Mean-based analysis
`epitope.dist.b`	< 0.001	< 0.001	< 0.001
`epitope.dist.c`	< 0.001	< 0.001	< 0.001
`epitope.dist.any`	< 0.001	< 0.001	< 0.001
`dist.zspace.sites.vrc01.zhao`	< 0.001	0.002	< 0.001
`epitope.dist.unweighted.any`	< 0.001	< 0.001	< 0.001
`epitope.dist.unweighted.b`	0.001	< 0.001	0.002
`epitope.dist.unweighted.c`	0.003	< 0.001	0.002
`pred.prob.res`	0.032	0.032	0.027
`pred.prob.sens`	0.034	0.032	0.027
`pred.ic80`	0.118	0.041	0.132
`hdist.zspace.sites.binding.all`	0.128	0.061	0.072
`length.gp120`	0.215	0.619	0.245
`dist.zspace.sites.vrc01.cale`	0.218	0.150	0.134
`num.cysteine.gp120`	0.226	0.192	0.199
`num.pngs.v5`	0.329	0.290	0.205
`length.v1v2`	0.351	0.316	0.248
`num.pngs.v1v2`	0.434	0.444	0.387
`num.pngs.gp120`	0.473	0.517	0.305
`length.v5`	0.596	0.597	0.464
`hdist.zspace.sites.preselect.all`	0.928	0.973	0.945

D.4.4 Uhambo (HVTN 702)

Table D.6: P-values from distributional and mean-based two-sample analyses in the Uhambo (HVTN 702) data. Results are shown for each sequence-derived feature. The empirical and MCB columns report p-values from the proposed Wasserstein barycenter-based tests, while the mean-based analysis column reports p-values from unequal variance two-sample t-tests applied after averaging the feature values within each individual.

Variable	Empirical	MCB	Mean-based analysis
`length.hypervariable.V4.tier1`	0.158	0.248	0.406
`cysteine.count.V3.tier1`	0.197	0.419	0.146
`num.sequons.V5.tier1`	0.243	0.130	0.830
`num.sequons.V2.tier1`	0.282	0.397	0.264
`num.sequons.V1.tier1`	0.292	0.475	0.234
`hdist.zspace.96ZM651.V2.linear.epitope.tier2`	0.320	0.365	0.250
`length.hypervariable.V2.tier1`	0.348	0.324	0.375
`num.sequons.V4.tier1`	0.416	0.612	0.644
`hdist.zspace.1086.gp120.tier2`	0.416	0.522	0.391
`cysteine.count.gp160.tier2`	0.430	0.334	0.648
`cysteine.count.V1.tier1`	0.448	0.481	0.425
`num.sequons.V3.tier1`	0.462	0.154	0.435
`hdist.zspace.96ZM651.V3.tier2`	0.557	0.706	0.580
`hdist.zspace.TV1.gp120.tier2`	0.593	0.780	0.735
`hdist.zspace.96ZM651.tier2`	0.594	0.806	0.853
`hdist.zspace.96ZM651.V5.tier2`	0.700	0.776	0.518
`num.sequons.CD4bs.Loop.D.tier1`	0.767	0.246	0.952
`cysteine.count.V4.tier1`	0.772	0.727	0.784
`length.hypervariable.V5.tier1`	0.831	0.894	0.860
`length.hypervariable.V1.tier1`	0.842	0.910	0.773
`hdist.zspace.96ZM651.CD4.binding.site.tier2`	0.901	0.838	0.941
`cysteine.count.V2.tier1`	0.961	0.872	0.963

D.5 Additional simulation results

D.5.1 Simulation Study 1

In addition to 50 individuals per group, we conducted Simulation Study 1 with 20 and 100 individuals per group. Results are shown below.

20 individuals per group

Null scenario summary results

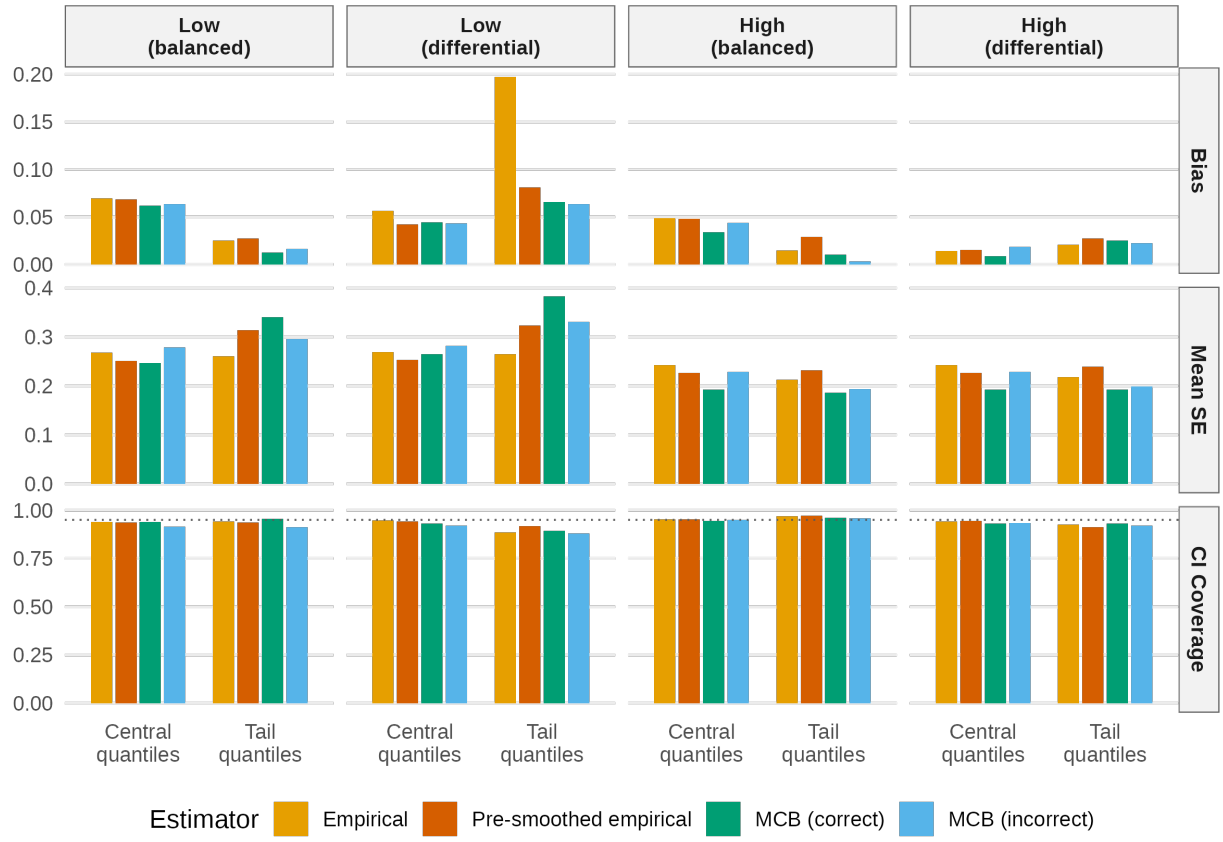


Figure D.9: Summary of estimation of the difference in quantiles $\Delta(u)$ under the null scenario for Simulation Study 1 with 20 individuals per group. Results are shown across four sequencing depth settings: low balanced, low differential, high balanced, and high differential sequencing depth. For each setting, we compare the empirical Wasserstein barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and misspecified spline-based MCB estimator. Performance is summarized by average bias, mean standard error, and pointwise confidence interval coverage for $\Delta(u)$, averaged separately over the central quantiles (0.06–0.94) and tail quantiles (0.01–0.05 and 0.95–0.99). The dotted horizontal line in the coverage panels indicates the nominal 95% coverage level.

Alternative scenario summary results

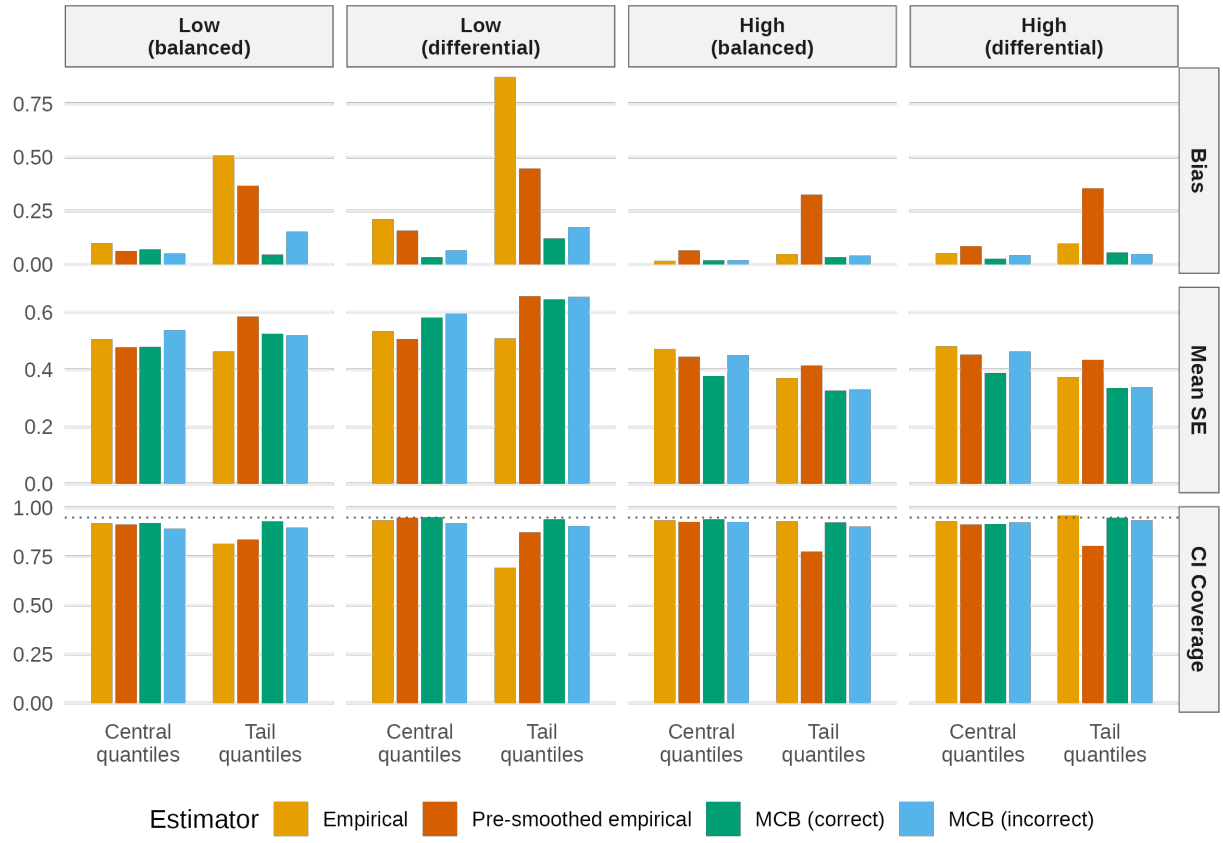


Figure D.10: Summary of estimation of the difference in quantiles $\Delta(u)$ under the alternative scenario for Simulation Study 1 with 20 individuals per group. Results are shown across four sequencing depth settings: low balanced, low differential, high balanced, and high differential sequencing depth. For each setting, we compare the empirical Wasserstein barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and misspecified spline-based MCB estimator. Performance is summarized by average bias, mean standard error, and pointwise confidence interval coverage for $\Delta(u)$, averaged separately over the central quantiles (0.06–0.94) and tail quantiles (0.01–0.05 and 0.95–0.99). The dotted horizontal line in the coverage panels indicates the nominal 95% coverage level.

Quantile difference map $\Delta(u)$: pointwise bias

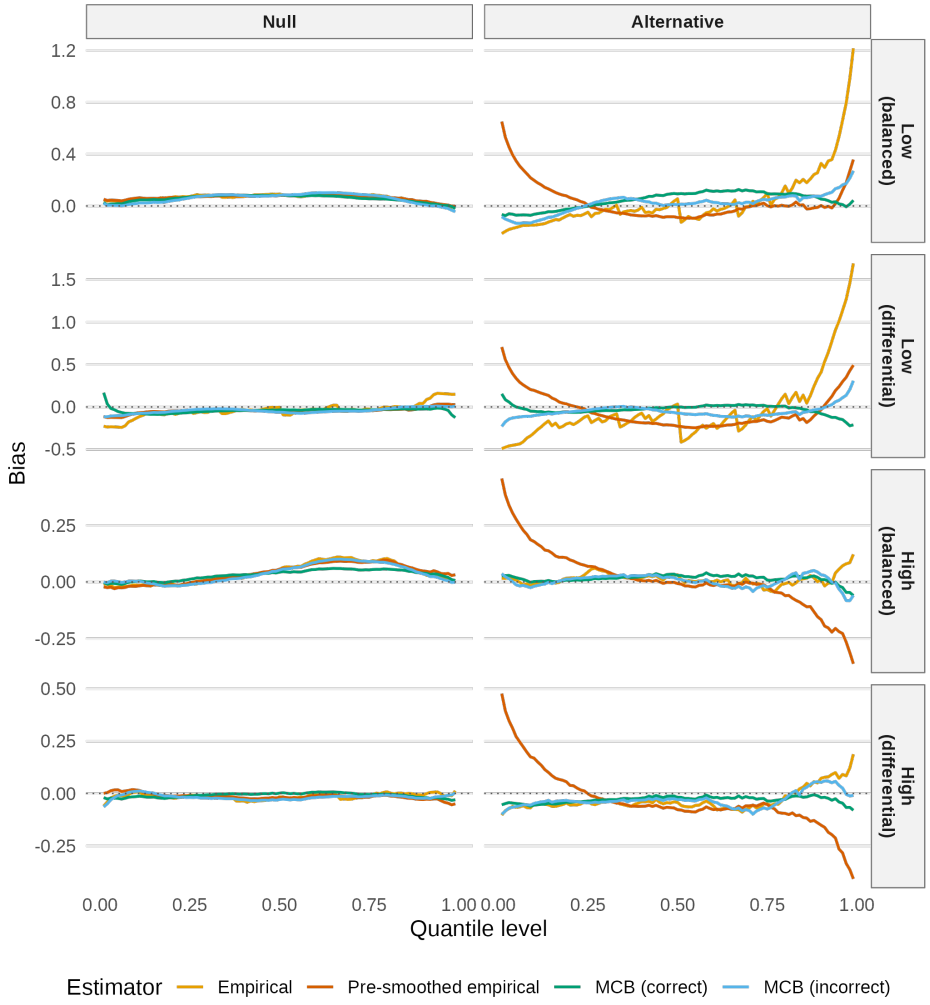


Figure D.11: Bias of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1 with 20 individuals per group. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth scenarios: low balanced, low differential, high balanced, and high differential. Each curve shows the bias of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Quantile difference map $\Delta(u)$: pointwise mean standard error

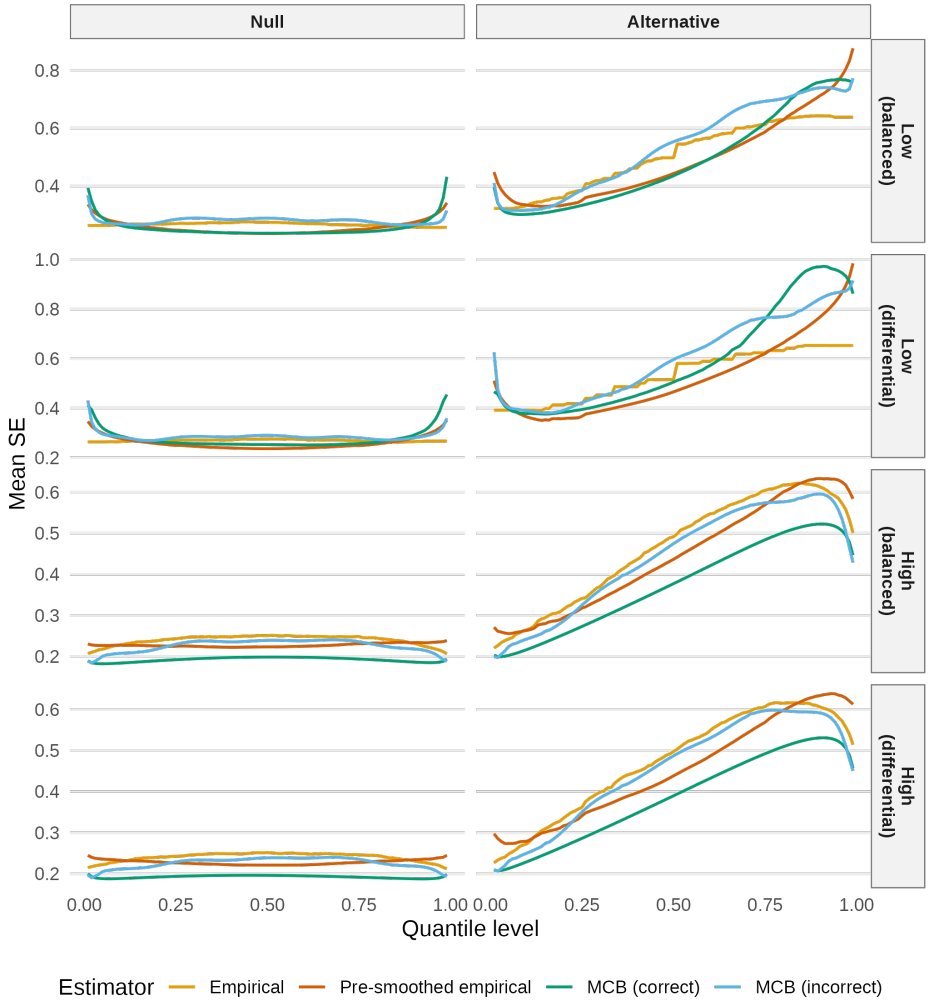


Figure D.12: Mean standard error of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1 with 20 individuals per group. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth scenarios: low balanced, low differential, high balanced, and high differential. Each curve shows the mean standard error of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Quantile difference map $\Delta(u)$: pointwise confidence coverage

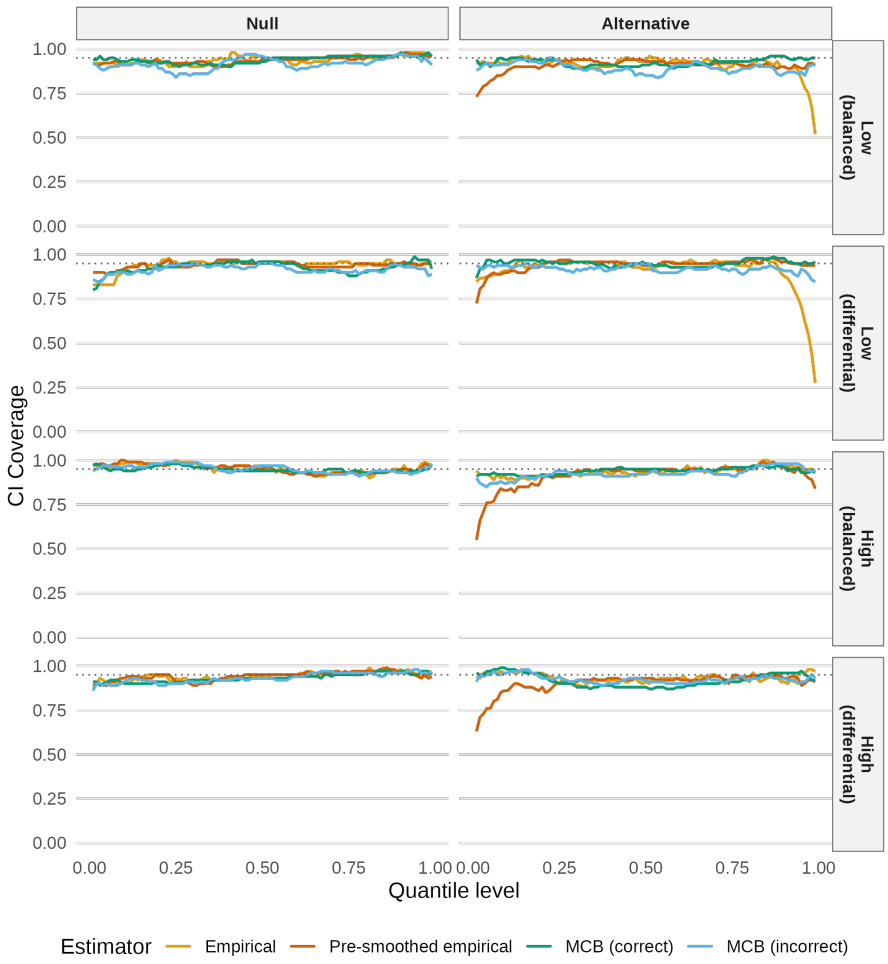


Figure D.13: 95% confidence interval coverage of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1 with 20 individuals per group. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth scenarios: low balanced, low differential, high balanced, and high differential. Each curve shows the 95% confidence interval coverage of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Group-specific barycenter $\mu_z(u)$: pointwise bias

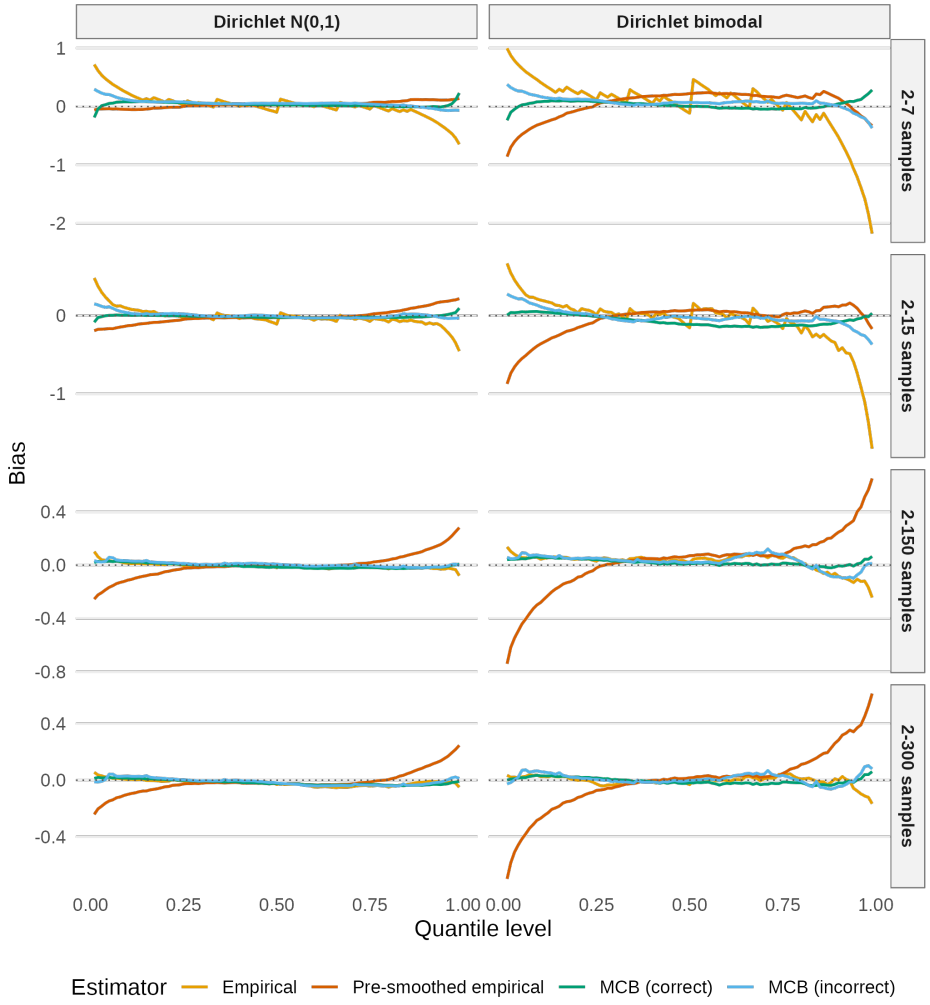


Figure D.14: Bias of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1 with 20 individuals per group. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the bias of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Group-specific barycenter $\mu_z(u)$: pointwise mean standard error

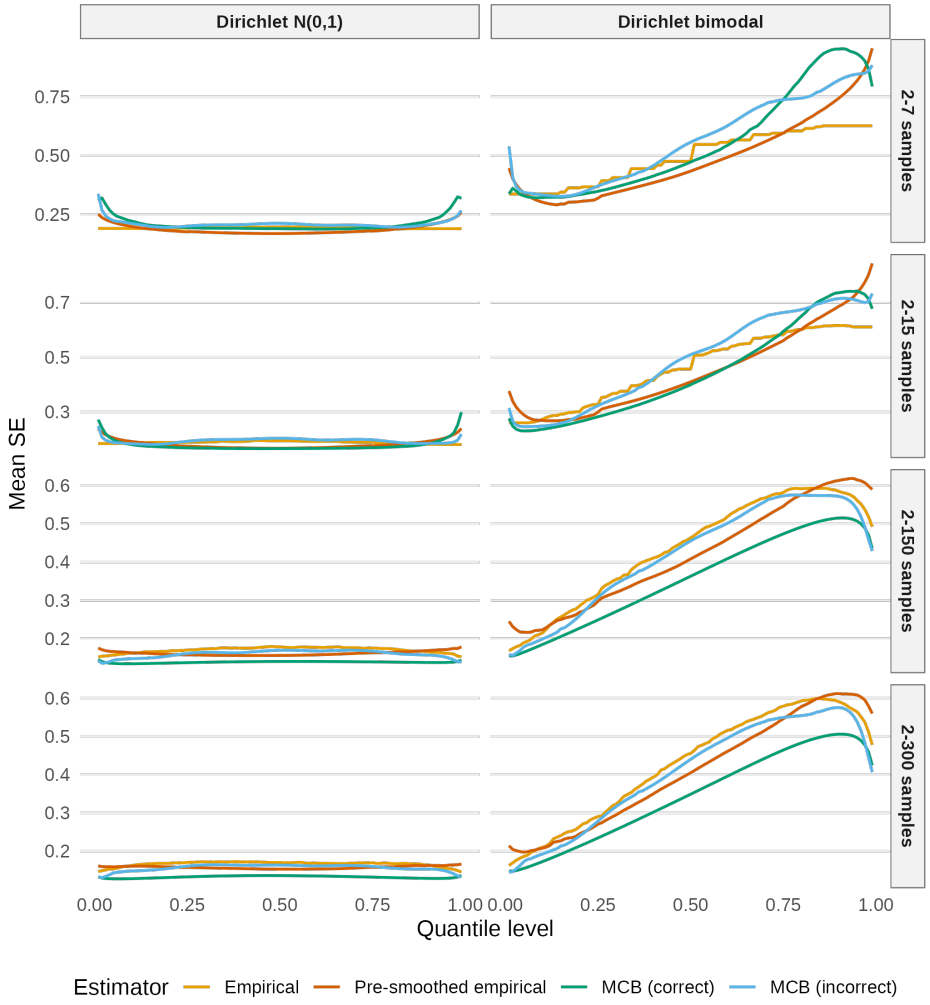


Figure D.15: Mean standard error of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1 with 20 individuals per group. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the mean standard error of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Group-specific barycenter $\mu_z(u)$: pointwise confidence interval coverage

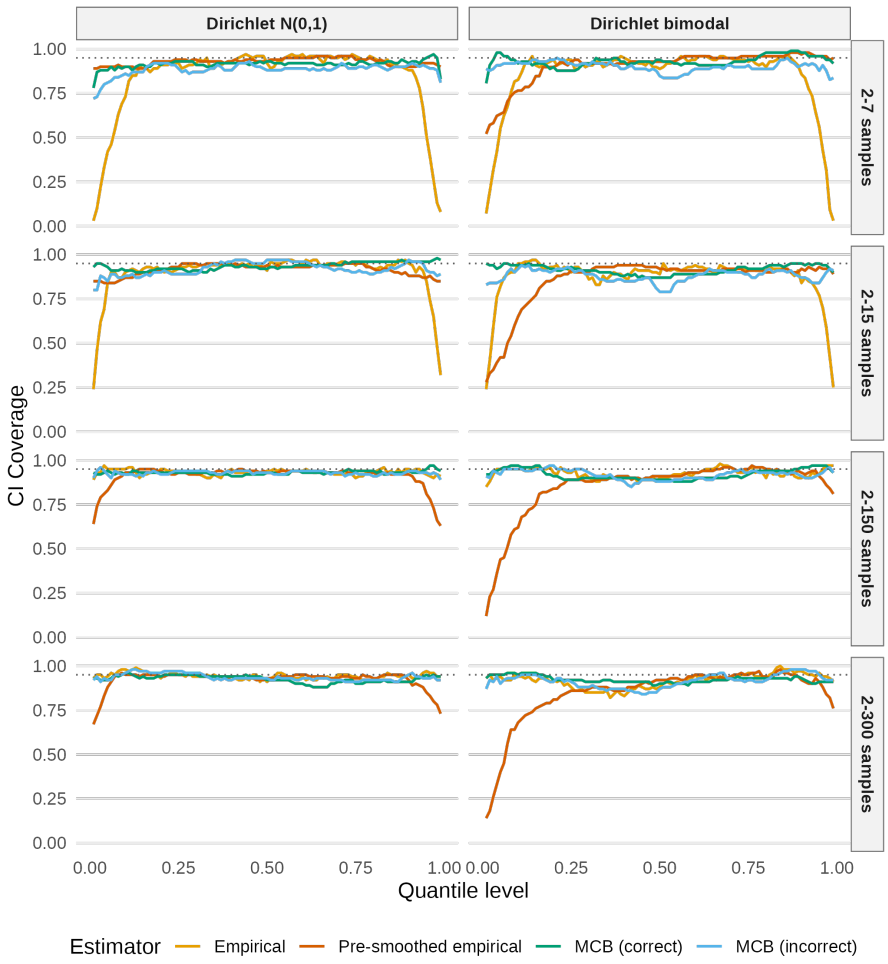


Figure D.16: 95% confidence interval coverage of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1 with 20 individuals per group. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the 95% confidence interval coverage of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

100 individuals per group

Null scenario summary results

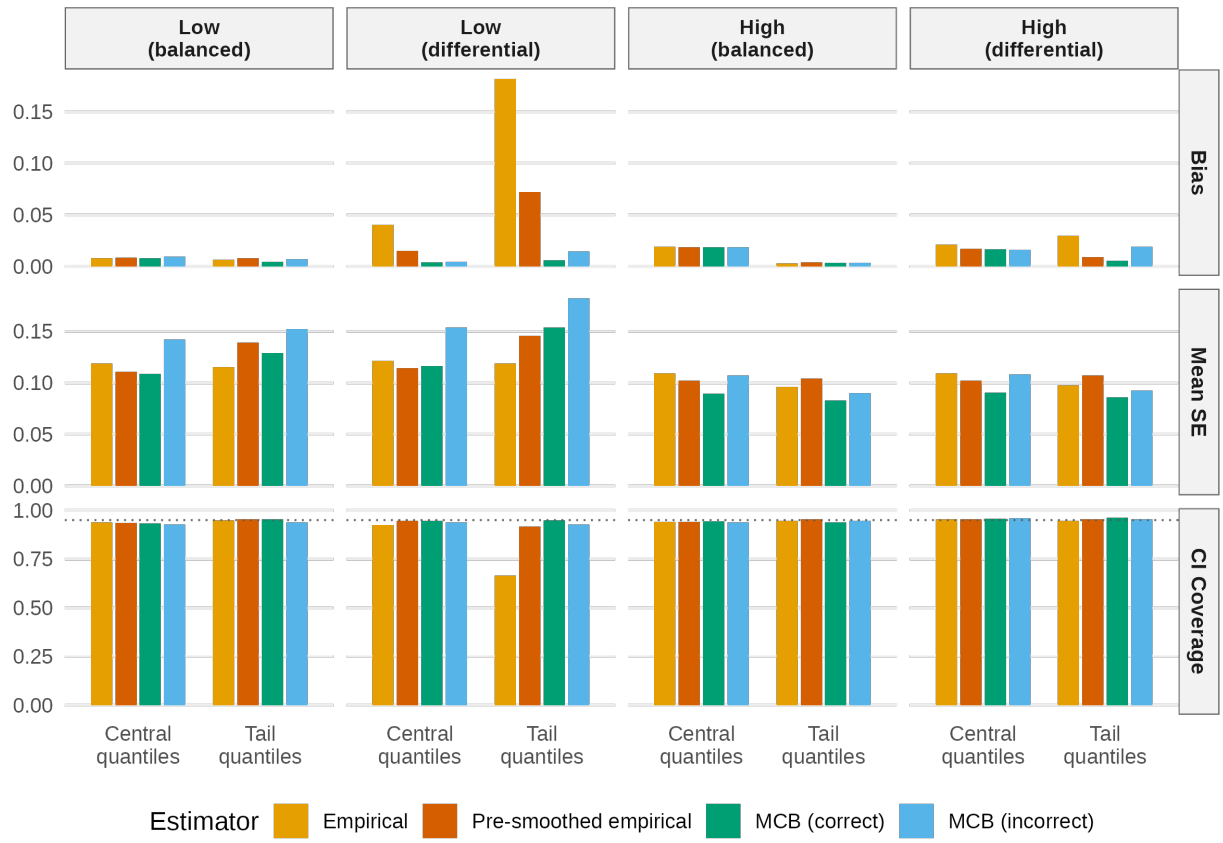


Figure D.17: Summary of estimation of the difference in quantiles $\Delta(u)$ under the null scenario for Simulation Study 1 with 100 individuals per group. Results are shown across four sequencing depth settings: low balanced, low differential, high balanced, and high differential sequencing depth. For each setting, we compare the empirical Wasserstein barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and misspecified spline-based MCB estimator. Performance is summarized by average bias, mean standard error, and pointwise confidence interval coverage for $\Delta(u)$, averaged separately over the central quantiles (0.06–0.94) and tail quantiles (0.01–0.05 and 0.95–0.99). The dotted horizontal line in the coverage panels indicates the nominal 95% coverage level.

Alternative scenario summary results

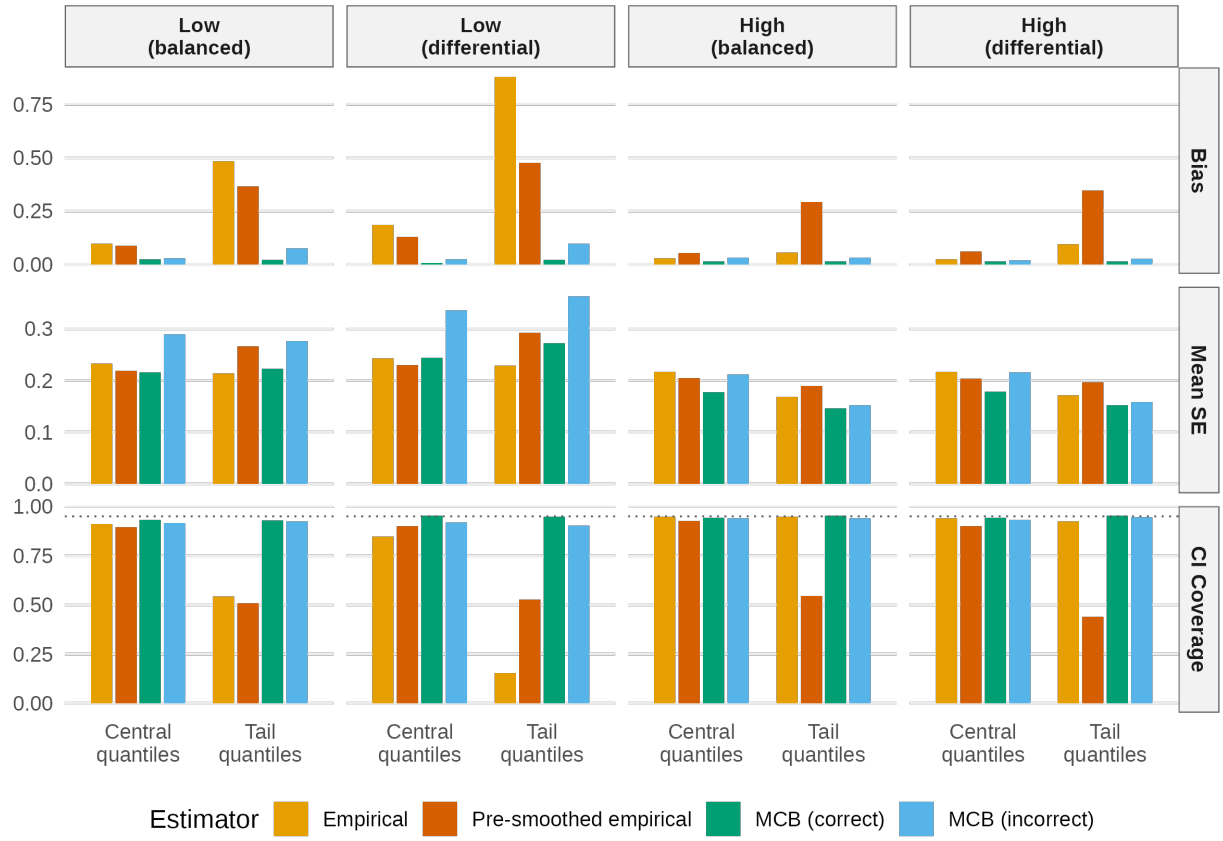


Figure D.18: Summary of estimation of the difference in quantiles $\Delta(u)$ under the alternative scenario for Simulation Study 1 with 100 individuals per group. Results are shown across four sequencing depth settings: low balanced, low differential, high balanced, and high differential sequencing depth. For each setting, we compare the empirical Wasserstein barycenter, pre-smoothed empirical barycenter, correctly specified MCB estimator, and misspecified spline-based MCB estimator. Performance is summarized by average bias, mean standard error, and pointwise confidence interval coverage for $\Delta(u)$, averaged separately over the central quantiles (0.06–0.94) and tail quantiles (0.01–0.05 and 0.95–0.99). The dotted horizontal line in the coverage panels indicates the nominal 95% coverage level.

Quantile difference map $\Delta(u)$: pointwise bias

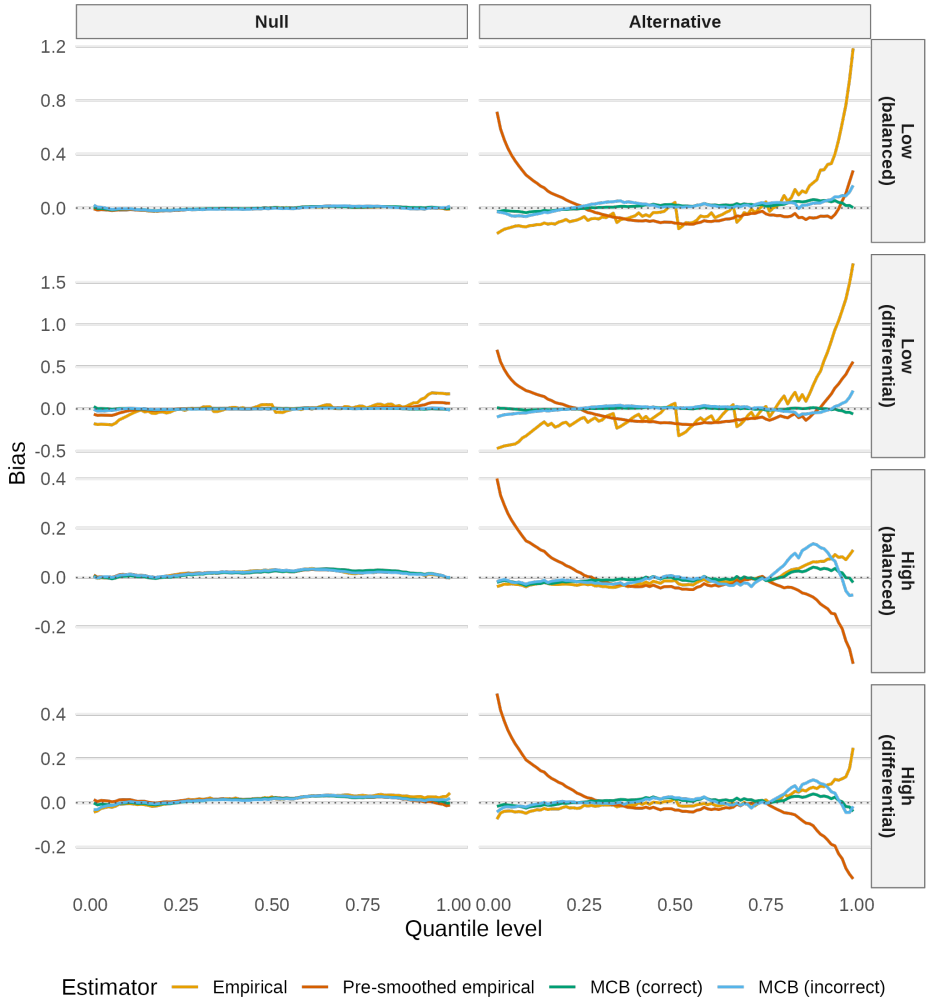


Figure D.19: Bias of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1 with 100 individuals per group. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth scenarios: low balanced, low differential, high balanced, and high differential. Each curve shows the bias of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Quantile difference map $\Delta(u)$: pointwise mean standard error

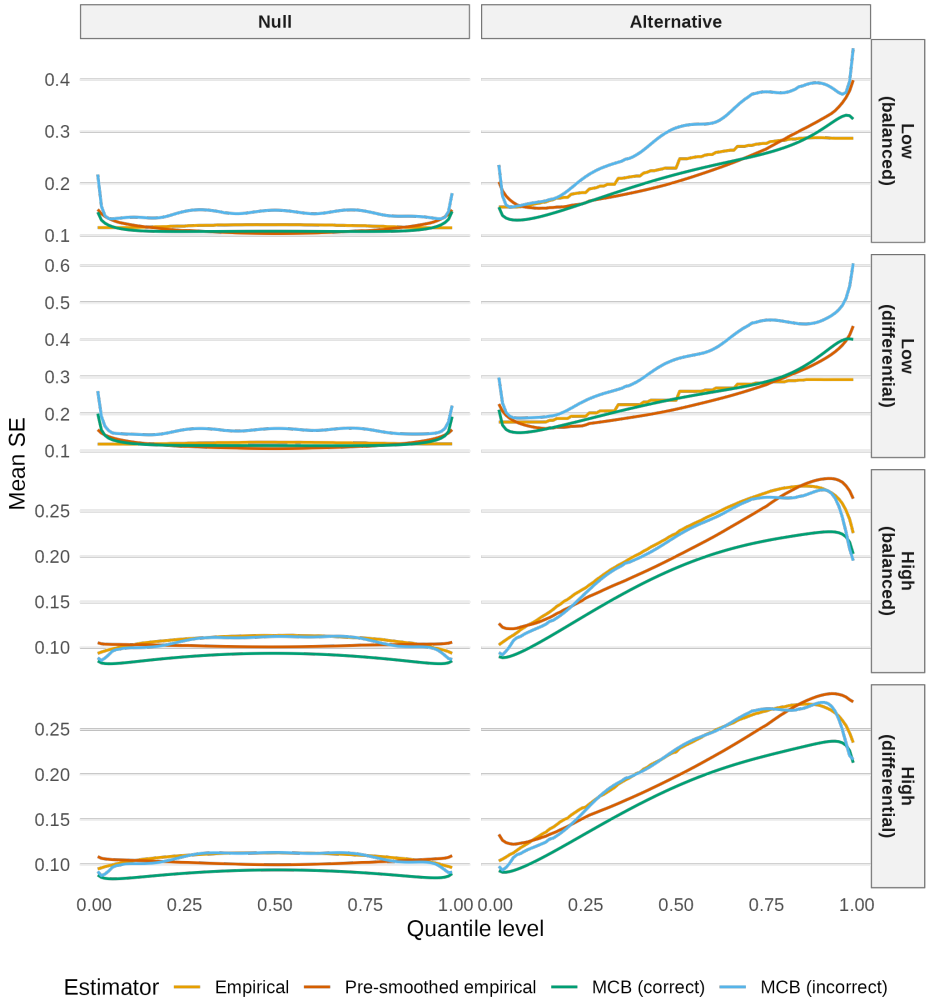


Figure D.20: Mean standard error of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1 with 100 individuals per group. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth scenarios: low balanced, low differential, high balanced, and high differential. Each curve shows the mean standard error of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Quantile difference map $\Delta(u)$: pointwise confidence coverage

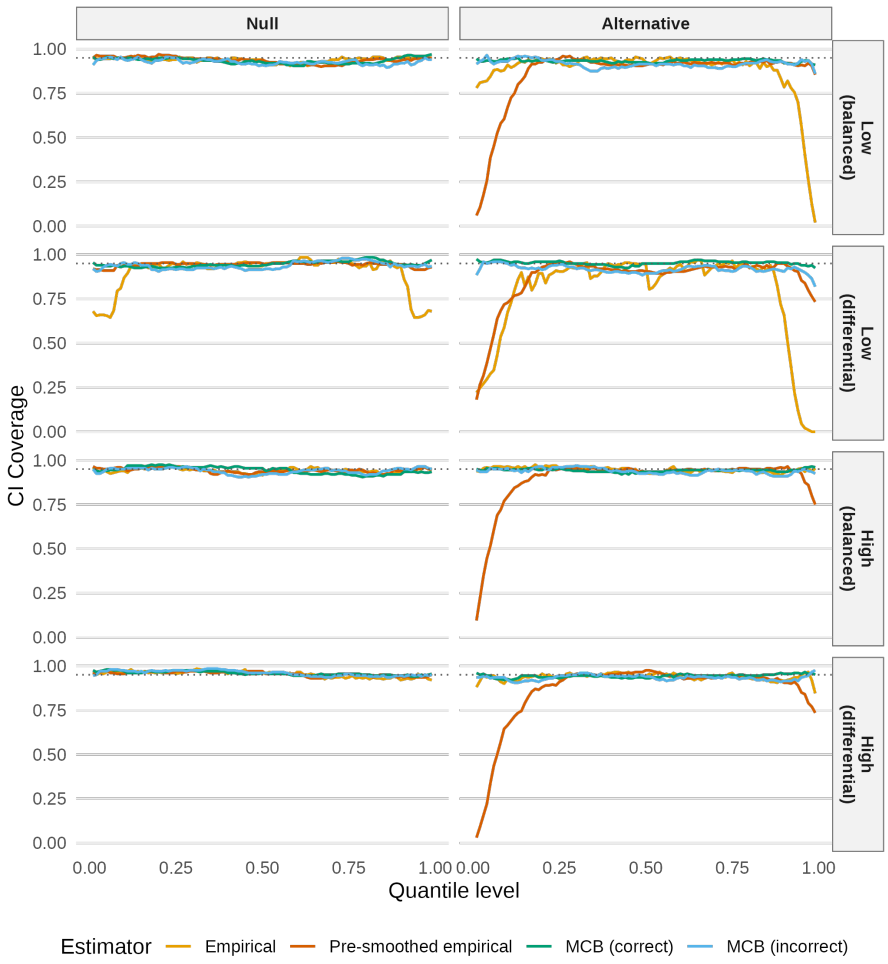


Figure D.21: 95% confidence interval coverage of the estimated quantile difference map $\widehat{\Delta}(u)$ in Simulation Study 1 with 100 individuals per group. Results are shown under the null and alternative sequence feature scenarios, across sequencing depth scenarios: low balanced, low differential, high balanced, and high differential. Each curve shows the 95% confidence interval coverage of $\widehat{\Delta}(u)$ across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Group-specific barycenter $\mu_z(u)$: pointwise bias

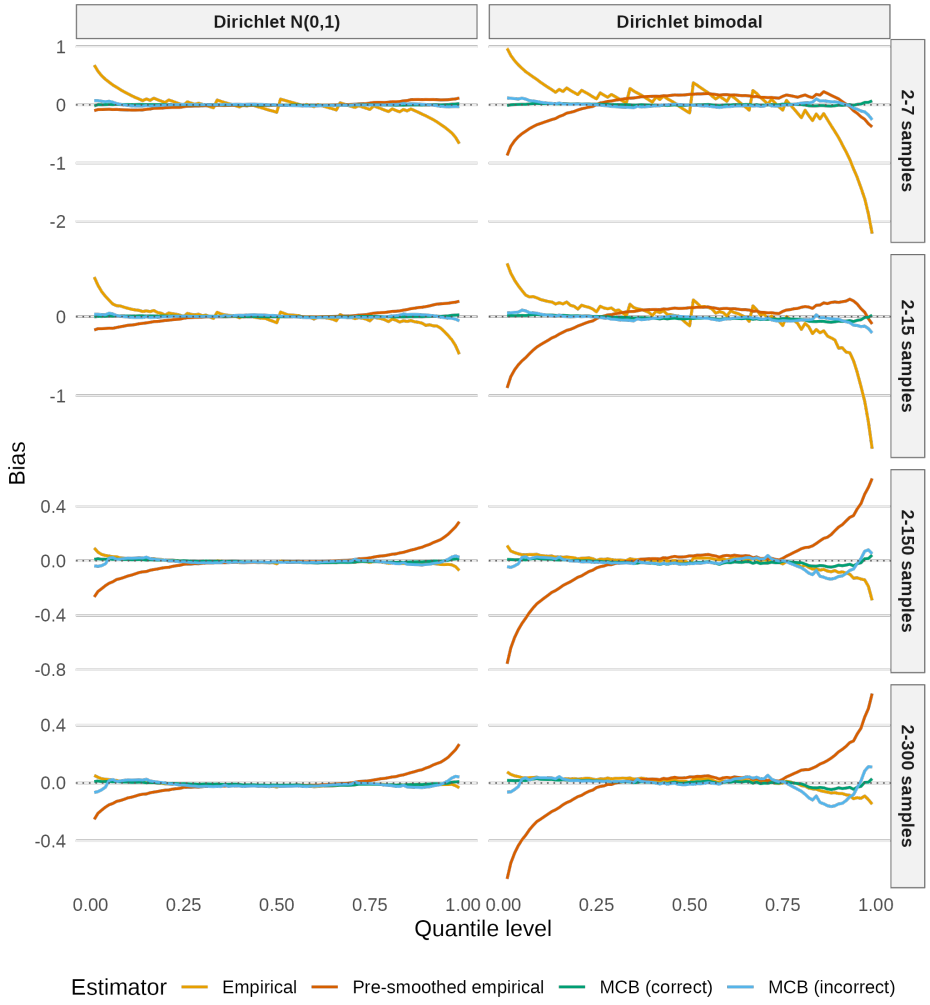


Figure D.22: Bias of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1 with 100 individuals per group. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the bias of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Group-specific barycenter $\mu_z(u)$: pointwise mean standard error

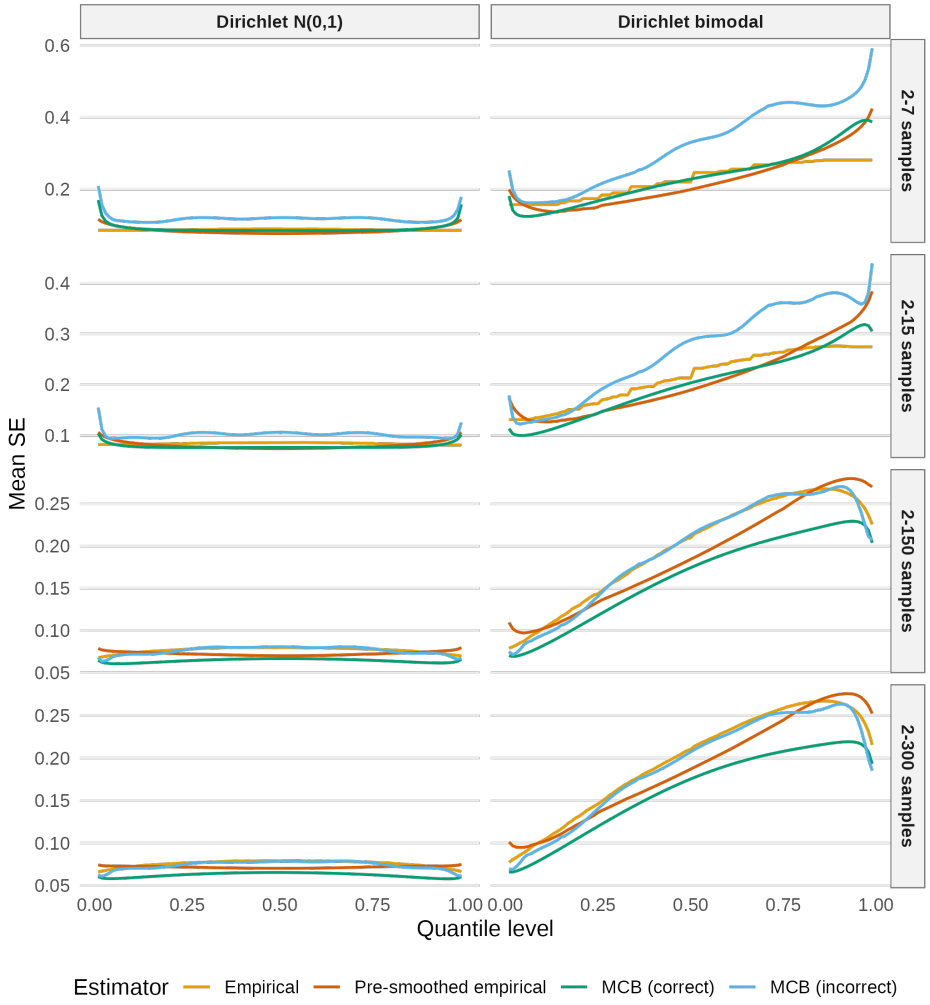


Figure D.23: Mean standard error of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1 with 100 individuals per group. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the mean standard error of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

Group-specific barycenter $\mu_z(u)$: pointwise confidence interval coverage

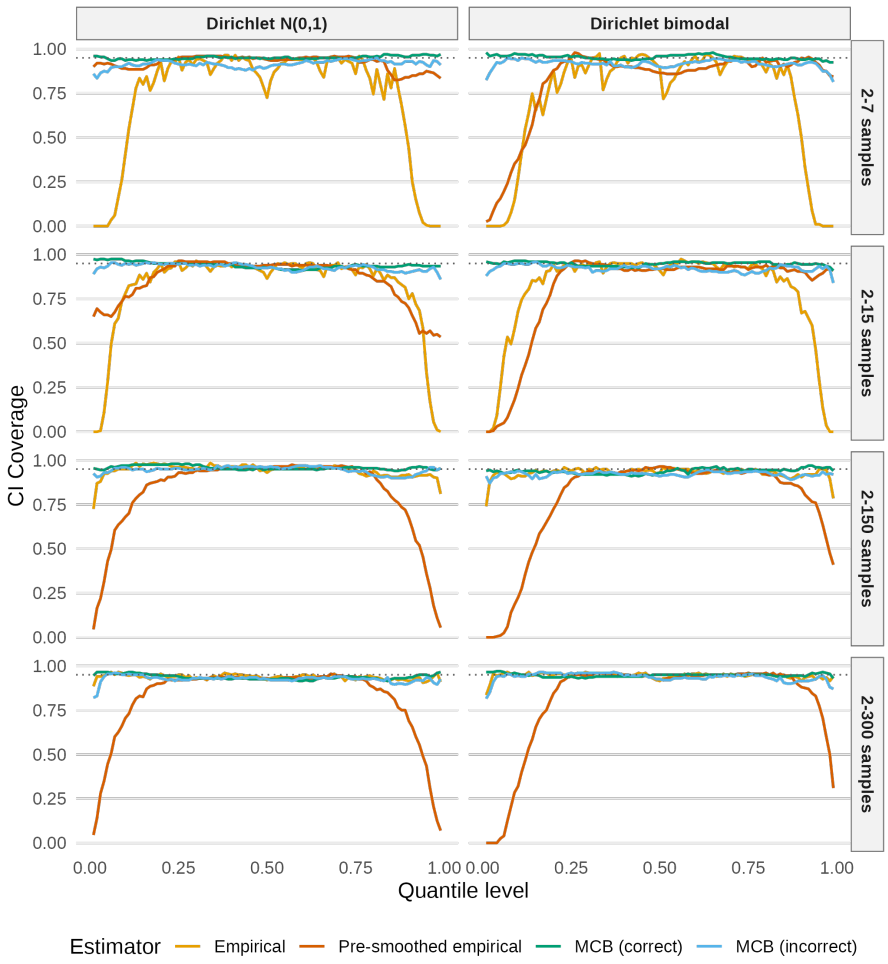


Figure D.24: 95% confidence interval coverage of the estimated group-specific Wasserstein barycenter quantile functions in Simulation Study 1 with 100 individuals per group. Results are shown for two data-generating settings: individual-level distributions generated from a Dirichlet process with a Normal$(0, 1)$ base measure, and individual-level distributions generated from a Dirichlet process with a bimodal Normal mixture base measure. Results are shown across sequencing depths of 2–7, 2–15, 2–150, and 2–300. Each curve shows the 95% confidence interval coverage of the estimated barycenter quantile function across quantile levels $u \in \{0.01, 0.02, \ldots, 0.99\}$. Estimators include the empirical Wasserstein barycenter, the pre-smoothed empirical Wasserstein barycenter, the correctly specified MCB estimator, and the misspecified spline-based MCB estimator.

D.5.2 Simulation Study 2

In the main text, Simulation Study 2 was presented under the high balanced sequencing depth scenario. Here, we provide the corresponding results under the remaining sequencing depth scenarios: low balanced, low differential, and high differential. Results are shown below.

Low differential

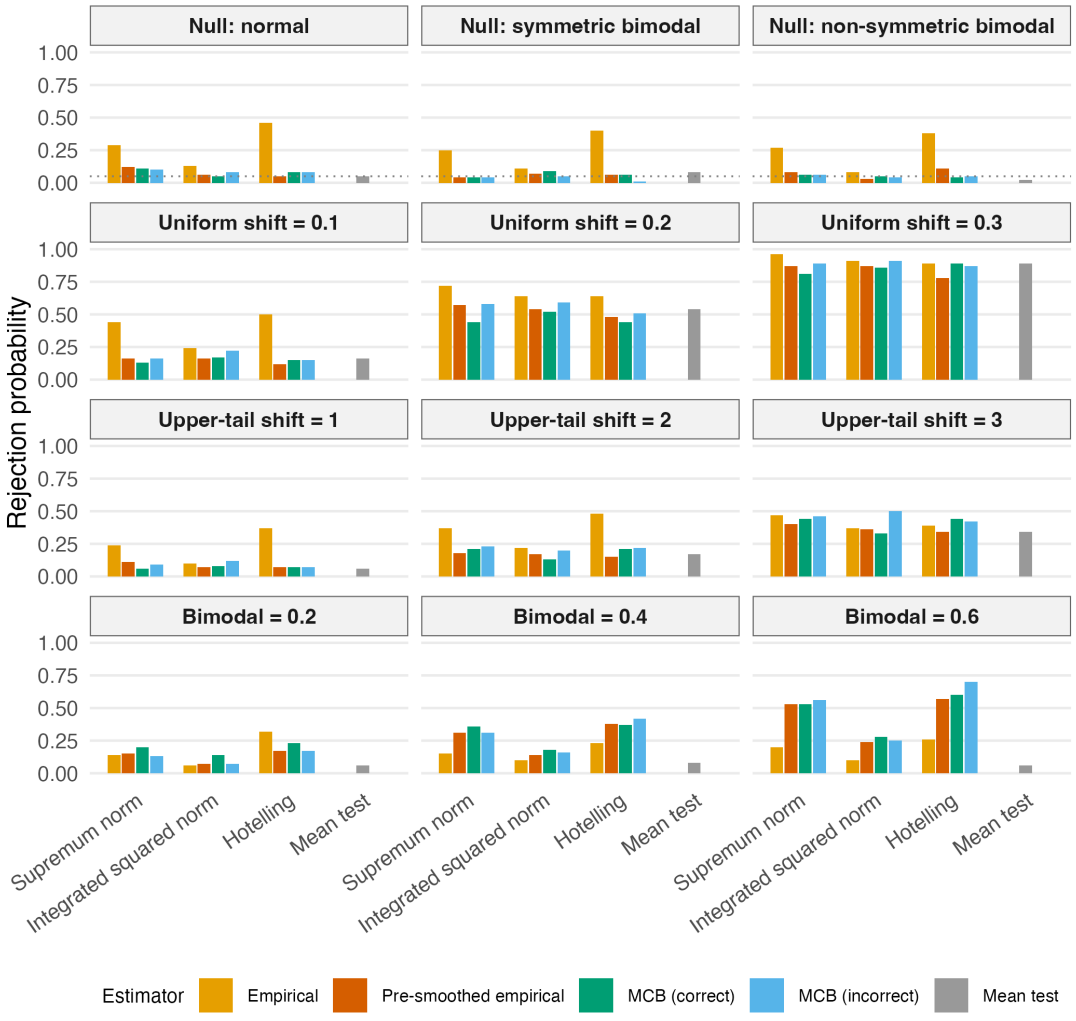


Figure D.25: Rejection probability across all simulations for each of the three null scenarios and nine alternative scenarios, for each estimator, under low differential sequencing depth. The mean test shows the result from averaging the sequence features for each person and doing a two-sample unequal variance t-test.

Low balanced

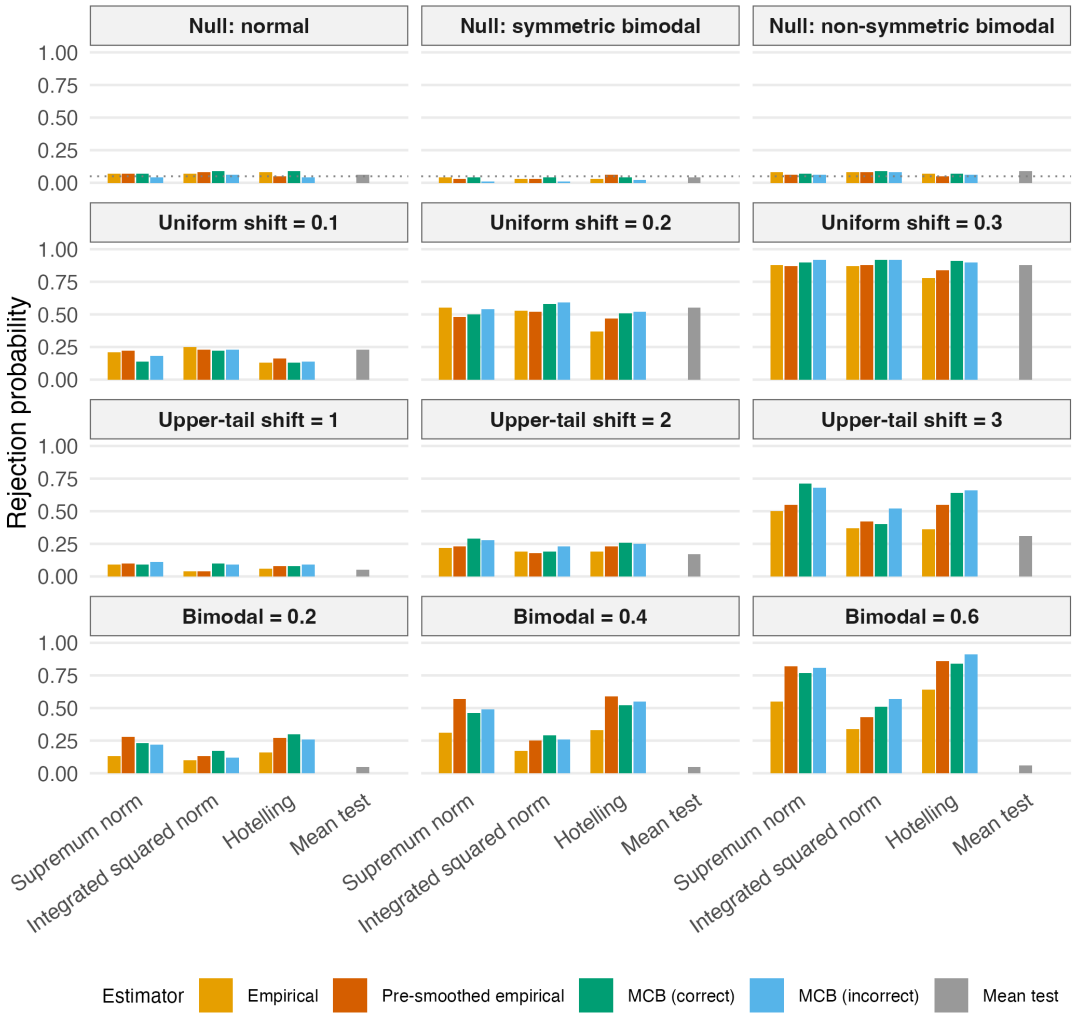


Figure D.26: Rejection probability across all simulations for each of the three null scenarios and nine alternative scenarios, for each estimator, under low balanced sequencing depth. The mean test shows the result from averaging the sequence features for each person and doing a two-sample unequal variance t-test.

High differential

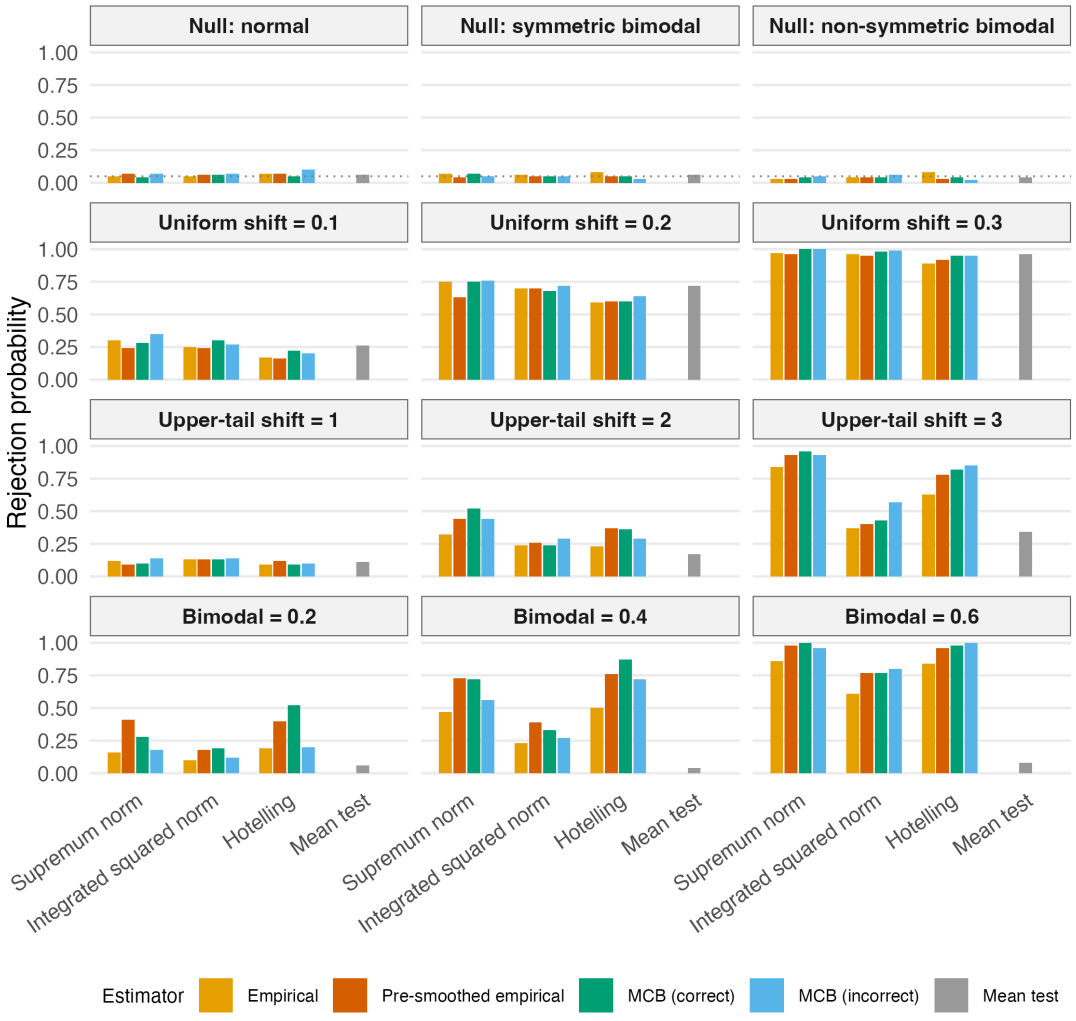


Figure D.27: Rejection probability across all simulations for each of the three null scenarios and nine alternative scenarios, for each estimator, under high differential sequencing depth. The mean test shows the result from averaging the sequence features for each person and doing a two-sample unequal variance t-test.

Appendix E

SUPPLEMENTARY MATERIALS FOR CHAPTER 5

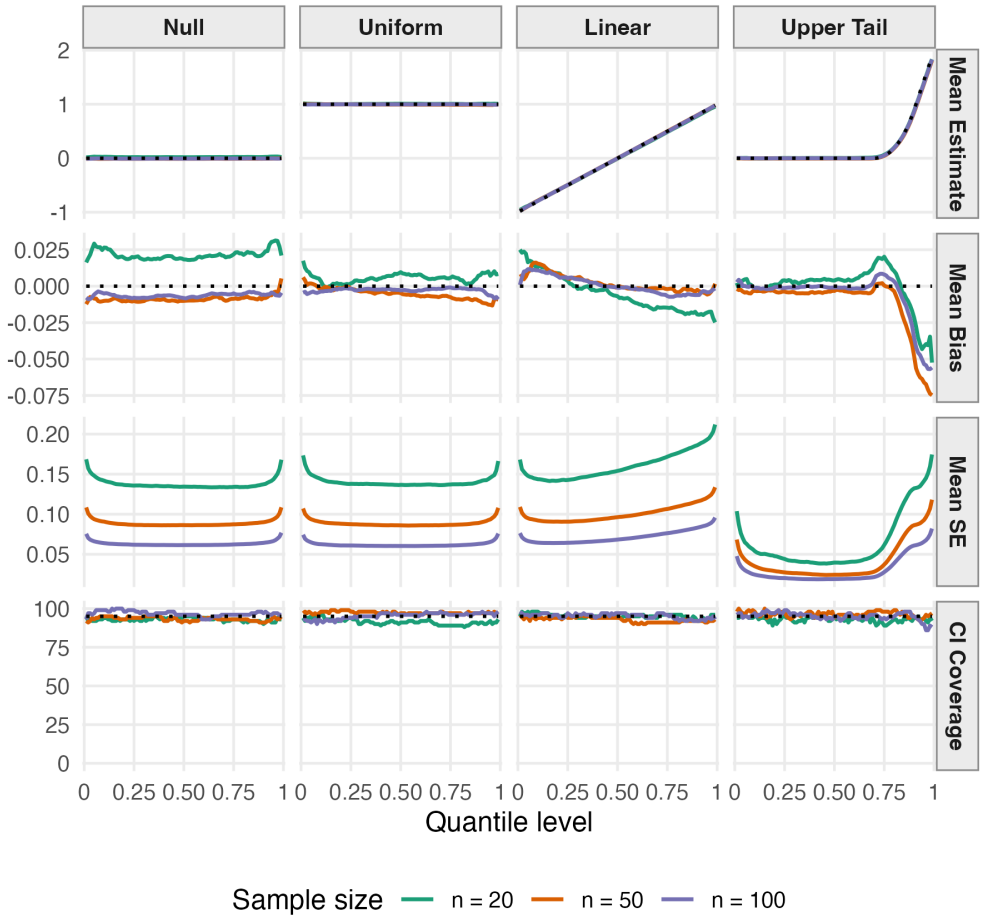


Figure E.1: Estimation performance for the mean normalized quantile shift map under four data-generating settings in Simulation Study 1 with 1,000 sequences per individual per timepoint. Columns correspond to the true shift pattern: no change, uniform shift, linear variability change, and upper-tail shift. Rows show the estimated mean shift map, pointwise bias, mean standard error, and pointwise confidence interval coverage across quantile levels. Colored curves correspond to sample sizes $n = 20, 50, 100$. Black dotted lines indicate the true mean shift map in the top row, zero bias in the second row, and nominal 95% coverage in the bottom row. Sequencing depths were sampled uniformly from 1 to 300 sequences per individual per timepoint.

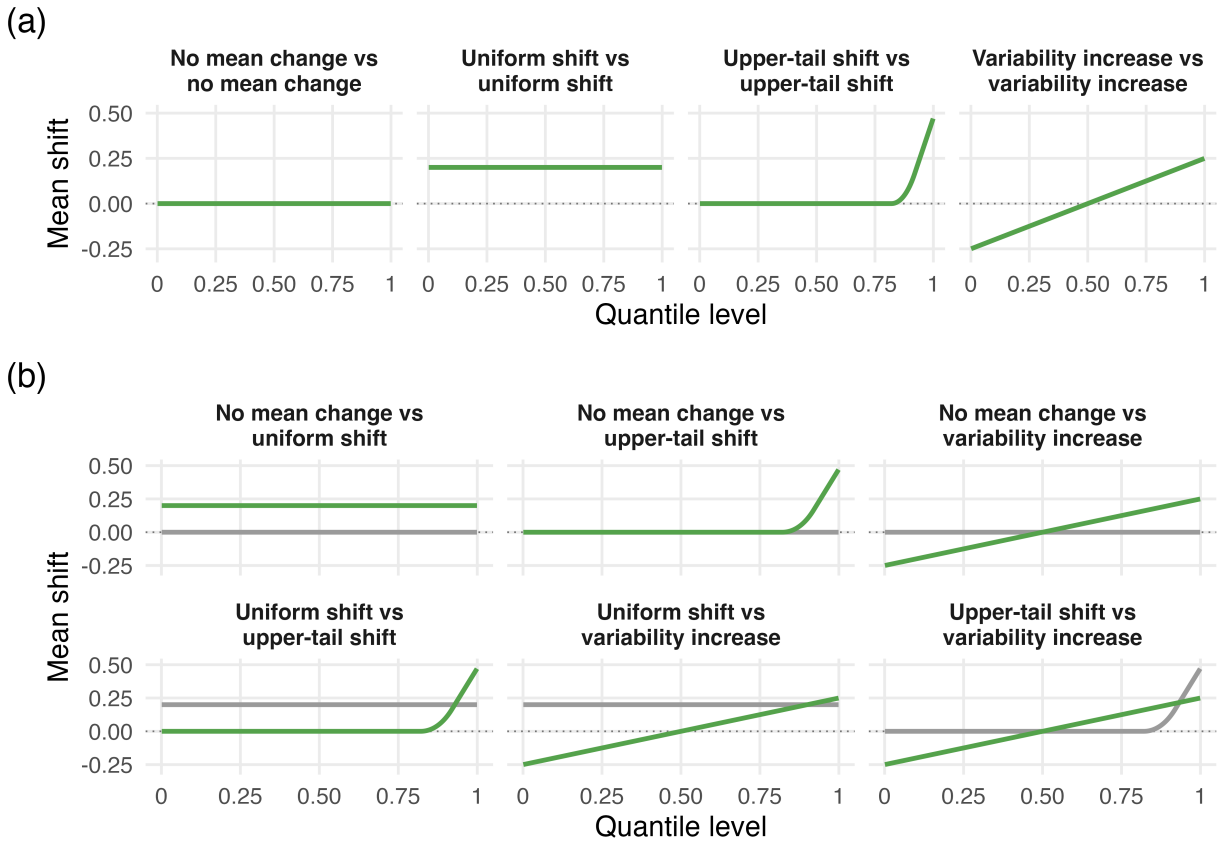


Figure E.2: Mean normalized quantile shift maps for the (a) null scenarios and (b) alternative scenarios in Simulation Study 2. The gray curve represents the mean normalized shift map of the placebo arm, and the green curve represents that of the vaccine arm.