# Novel data-adaptive multivariate testing procedures, with applications to HIV research

Adam Elder

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

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Marco Carone, Chair
Alex Luedtke, Chair
Yen-Chi Chen

Program Authorized to Offer Degree:
Biostatistics

University of Washington

# Abstract

Novel data-adaptive multivariate testing procedures, with applications to HIV research

Adam Elder

Chairs of the Supervisory Committee:

Alex Luedtke
Department of Statistics

Marco Carone
Department of Biostatistics

**Project 1:** Construct a generic data-adaptive framework for multivariate point null testing. Specifically, a data-driven optimality criterion is proposed for selecting among a large collection of candidate test statistics. This framework can be applied in a wide array of problems. It is illustrated on data from HVTN 505, a phase IIB HIV vaccine efficacy trial.

**Project 2:** Extend the framework developed in Aim 1 for testing a functional null hypothesis. The test described in this framework connects to the test in Aim 1 by projecting the function estimator into a finite dimensional vector space using a finite collection of the coefficients from the Fourier transformation. We also provide arguments that the described test can consider more coefficients as sample size grows while still maintaining desirable testing properties.

**Project 3:** Develop novel methodology for estimating open-label effectiveness for trials in which many or all study participants have switched off the placebo arm of the trial. The method developed accounts for changes in the population's distribution of baseline characteristics and adherence behavior. The method developed is used to estimate the open-label effectiveness of a vaginal ring containing Dapivirine using data from the MTN-20 (ASPIRE) and MTN-25 (HOPE) clinical trials.

# Contents

# Figures

# Tables

**Acknowledgements**

This dissertation would not have been possible without the lifetime worth of support from my entire family, especially my parents, May Goldenberg and Edward Elder. Their unwavering trust and love helped me maintain confidence persevere through the many setbacks encountered during my dissertation.

I would also like to thank all my friends both from graduate school and outside of graduate school. Your support throughout these years made me happier and more resilient. In particular I would like to thank all of the roommates I have had the good fortune of living with during these past six year. Thank you Bobby, Walter, Nathan, and James for all of the energy and joy you have brought to my life. I would also like to thank Joe and Rubi for being lifelong friends and helping me stay grounded during my time in graduate school. Last, I would like to thank Subodh and Kendrick for being my statistical confidants and friends during graduate school.

The University of Washington Department of Biostatistics consists of many intelligent and helpful people and I am grateful to have been surrounded by such talented individuals during my time in the program. I would like to give a special thanks to Gitana Garofalo, Minh Vo, Noah Simon, and Barbra Richardson who have gone above and beyond their their normal duties to provide me and other students with the support we need. I also very thankful to Ali Shojaie for his useful feedback and questions as a supervisory committee member and to Yen-Chi Chen for his insightful comments and suggestions as a reading committee member.

Finally, I would like to thank my two committee chairs, Marco Carone and Alex Luedtke for all of their support and guidance over these past six years. I have been inspired by Marco's passion for statistical methodology and the art of scientific writing. The brilliance, persistence, and attention to detail that Alex brings to our work together has taught me how to be a better scientist and statistician. Having the opportunity to work with such immensely talented and kind individuals has been an honor.

**Dedication**

To my advisors, my friends, and my family.

# Chapter 1

# A general adaptive framework for multivariate point null testing

As a common step in refining their scientific inquiry, investigators are often interested in performing some screening of a collection of given statistical hypotheses. For example, they may wish to determine whether any one of several patient characteristics are associated with a health outcome of interest. Existing generic methods for testing a multivariate hypothesis — such as multiplicity corrections applied to individual hypothesis tests — can easily be applied across a variety of problems but can suffer from low power in some settings. Tailor-made procedures can attain higher power by building around problem-specific information but typically cannot be easily adapted to novel settings. In this work, we propose a general framework for testing a multivariate point null hypothesis in which the test statistic is adaptively selected to provide increased power. We present theoretical large-sample guarantees for our test under both fixed and local alternatives. In simulation studies, we find that tests created using our framework perform nearly as well as tailor-made methods when the latter are available, and we illustrate how our procedure can be used to create tests in two settings in which tailor-made methods are not currently available.

## 1.1 Introduction

Addressing a scientific question often involves performing simultaneous inference on components of a vector-valued statistical parameter and, in particular, assessing whether this parameter deviates from a specific null value of scientific interest. Indeed, testing of a multivariate point null hypothesis arises commonly in applications. For example, it may be of interest to determine whether any of several variables are related to a particular health outcome, as often occurs in genetics (Gao et al., 2008), neurology (Flandin and Friston, 2019), and vaccine development (Borthwick et al., 2014), among other fields. General-purpose strategies (e.g., construction of Wald-type test statistics) exist for performing a hypothesis test of a univariate point null with a specified (asymptotic) type I error; in many cases, such strategies can be shown to yield optimal tests. The corresponding

problem for a multivariate point null poses a much greater challenge.

A valid test of a multivariate null hypothesis can be constructed on the basis of multiple tests of univariate null hypotheses in a manner that controls the family-wise type I error rate. For decades, the Bonferroni correction has been used to derive multiple hypothesis testing procedures. Early examples of its use appear in Dunn (1959, 1961). Refinements of the Bonferroni correction have been proposed by various authors, including, for example, Holm (1979), Simes (1986), Hommel (1988), Hochberg (1988) and S. Holland and DiPonzio Copenhaver (1988). Bonferroni-type correction procedures are broadly applicable and easily implemented. However, because they do not leverage knowledge of the dependence between the test statistics involved, they may yield low power in some circumstances. Some authors, including Lehmann and Romano (2005) and Dudoit and van der Laan (2008) have proposed alternative strategies to mitigate this problem by accounting for the joint behavior of the test statistics. These procedures, in particular, allow users to specify the desired trade-off between type I and II errors by controlling, for example, the false discovery rate or family-wise error rate of the test. Nevertheless, despite these improvements, the use of multiple testing techniques to assess a single multivariate hypothesis, while convenient, comes at a price. The ability to determine which null hypothesis (if any) to reject, while potentially valuable, could come at the cost of lower power for detecting deviations from the multivariate point null. Indeed, for any multiple testing procedure that achieves family-wise type I error control, there exists a calibrated test of the multivariate null with at least as much power. In fact, a more powerful test of the multivariate null would be expected to exist since such a test does not need to account for rejections of a univariate null that holds when others do not.

Approaches for multivariate testing have been proposed and typically account for the correlation between individual test statistics. Such methods can be categorized based on how an aggregate test statistic is constructed. In some procedures (e.g., Donoho and Jin, 2004), a summary test statistic is built using estimators of underlying univariate parameters, whereas in others (e.g., Liu and Xie, 2020) $p$-values from multiple univariate tests are directly combined. Unfortunately, these procedures are usually tailored to a specific parameter and statistical model (e.g., Donoho and Jin, 2004) or make assumptions about the data-generating mechanism that can fail in practice (e.g., sparsity conditions or parametric modeling assumptions). Additionally, some procedures do not allow the use of flexible learning strategies in the construction of the involved test statistics (Breiman, 2001). While the use of flexible learners is often critical to obtaining asymptotic guarantees in nonparametric and semiparametric models, it can also cause poor finite-sample performance of testing procedures, especially when the adaptive nature of the test statistic is not taken into account (see, e.g., Leeb and Pötscher, 2005, 2006). While more recent proposals address many of these potential issues (e.g., Pan et al., 2014; McKeague and Qian, 2015; Xu et al., 2016), they

provide techniques for use in specific applications rather than general-purpose templates for use in a variety of problems. Thus, while procedures for multivariate testing with good performance characteristics have been devised for certain settings, in many cases, there is little guidance for investigators beyond crude approaches such as the Bonferroni correction. In this paper, we propose and study a general-purpose procedure for constructing a test of a multivariate point null hypothesis that can be used for a broad range of statistical parameters and models. Our procedure benefits from an explicit accounting of the joint behavior of the test statistic and incorporates data-driven selection of the involved tuning parameters to optimize test performance for the application at hand. As such, it can be expected to provide improved performance compared to existing strategies in many contexts.

This chapter is organized as follows. In Section 1.2, we introduce the testing problem considered and provide working examples with which we will illustrate the implementation and performance of our proposed procedure. We formally describe our procedure in Section 1.3, and provide a theoretical study of its properties in Section 1.4. In Section 1.5, we illustrate through simulation studies that the proposed framework yields novel tests with comparable power to tailor-made procedures in settings in which specialized methods already exist, and has good operating characteristics in settings in which problem-specific methods do not currently exist. In Section 1.6, we use our procedure to test for the existence of a correlate of risk of HIV infection using data from the HVTN 505 HIV vaccine trial. In Section 1.7, we provide concluding remarks. Technical proofs as well as additional simulation results and details on our data analysis are provided in the Appendix.

## 1.2 Problem setup

Suppose that we have at our disposal observations $X_1, X_2, \ldots, X_n$ drawn independently from a common unknown distribution $P_0 \in \mathcal{M}$, where the statistical model $\mathcal{M}$ encodes known restrictions on $P_0$. In the developments below, we are primarily interested in cases in which $\mathcal{M}$ is a nonparametric or semiparametric model, although this is not a requirement for the developments presented. We denote by $\mathcal{X}$ the union of the support of $P$ for each $P \in \mathcal{M}$. Suppose that $\Psi_1, \Psi_2, \ldots, \Psi_d$ form a collection of real-valued statistical parameters defined on $\mathcal{M}$. For each $j \in \{1, 2, \ldots, d\}$, we define $\psi_{j0} := \Psi_j(P_0) \in \mathbb{R}$ to be the evaluation of $\Psi_j$ on $P_0$, and write $\psi_0 := (\psi_{10}, \psi_{20}, \ldots, \psi_{d0})$. In this article, for a given (known) vector $\psi_* := (\psi_{1*}, \psi_{2*}, \ldots, \psi_{d*}) \in \mathbb{R}^d$, we consider testing

$$H_0 : \psi_0 = \psi_* \quad \text{versus} \quad H_1 : \psi_0 \neq \psi_* \ . \tag{1.1}$$

Without loss of generality, we consider the case $\psi_* = (0, 0, \ldots, 0)$ since otherwise we may instead take $\Psi_j$ to be its null-centered counterpart $P \mapsto \Psi_j(P) - \psi_{j*}$.

The setup we consider is sufficiently broad to include a large variety of examples. For concreteness, we present here three particular examples that we will use throughout as an illustration of our general results.

**Example 1: correlation.** In our first and simplest example, we consider the data unit $X = (W, Y)$, where $W := (W_1, W_2, \ldots, W_d)$ represents a vector of real-valued covariates and $Y$ is some outcome of interest, and the parameter of interest $\Psi_j(P) := \text{corr}_P(W_j, Y)$ is the marginal correlation between $W_j$ and $Y$ under $P$. We are interested in testing the multivariate null hypothesis that none of the components of $W$ are marginally correlated with $Y$ in a nonparametric model. For this problem, there exist several competing approaches in the literature, and we will compare a test derived using our proposal to several of these existing approaches.

**Example 2: coefficients of a working log-linear regression model under missingness.** In our second example, we instead consider the data unit $X = (W, U, \Delta)$, where $W := (W_1, W_2, \ldots, W_d)$ again represents a vector of real-valued covariates, $\Delta$ is an indicator that the binary outcome $Y$ is observed, and $U := \Delta Y$ equals $Y$ if $\Delta = 1$ and is set to zero otherwise. In other words, this data unit is similar to that defined in Example 1 but with the outcome value possibly missing. We focus here on coefficients indexing the least-squares projection of the true conditional success probability onto the log-linear regression model $\log \text{pr}(Y = 1 \,|\, W_j = w_j) = \alpha_0 + \alpha_j w_j$. Assuming missingness at random, that is, that $Y$ and $\Delta$ are independent conditionally upon $W$, the parameter

$$\Psi_j(P) := \frac{\text{cov}_P \left[W_j, \log E_P \left\{P(U = 1 \,|\, \Delta = 1, W) \,|\, W_j\right\}\right]}{\text{var}_P(W_j)} \tag{1.2}$$

identifies the coefficient associated with $W_j$ in the projection onto the log-linear working model, and simplifies to $\alpha_j$ when this working model holds true. This parameter represents a measure of association between positive outcome $Y$ and covariate $W_j$ for use when $Y$ is possibly missing at random given $W$. We are interested in testing, within a nonparametric model, the multivariate null hypothesis that all coefficients of this working log-linear model equal zero.

**Example 3: coefficients of a working effect modification model for randomized trials.** In our third example, we consider the data unit $X = (W, A, Y)$, where $W := (W_1, W_2, \ldots, W_d)$ once more represents a vector of real-valued covariates, $A \in \{0, 1\}$ is a binary treatment variable, and $Y$ is a binary outcome of interest, and focus on the interaction coefficient of the least-squares projection of the true conditional success probability onto the logistic model $\text{logit}\, \text{pr}\left(Y = 1 \,|\, W_j = w, A_j = a\right) = \alpha_{0j} + \alpha_{1j} a + \alpha_{2j} w + \delta_j wa$. This coefficient provides a measure of the degree to which $W_j$ modifies the effect of $A$ on $Y$ in a randomized trial. The parameter of interest can be expressed as

$$\Psi_j(P) := \underset{\gamma}{\text{argmin}} \min_{\alpha} E_P \left[\text{logit}\ P(Y = 1 \,|\, A, W_j) - \alpha_0 - \alpha_1 A - \alpha_2 W_j - \gamma W_j A\right]^2 ,$$

which identifies the interaction coefficient in this working model, and simplifies to $\delta_j$ when the working logistic model above holds. Once more, we are interested in testing, within a nonparametric model, the multivariate null hypothesis that each $\Psi_j(P)$ is equal to zero.

## 1.3 Proposed testing procedure

### 1.3.1 Non-adaptive test

While the test we ultimately propose is adaptive, it can be viewed as a refinement of non-adaptive counterparts, which we begin by describing. We define $\mathcal{M}_0 := \{P \in \mathcal{M} : \Psi_j(P) = 0 \text{ for each } j = 1, 2, \ldots, d\}$ to be the collection of all distributions in $\mathcal{M}$ under which the null hypothesis (1.1) is true. Suppose that an estimator $\psi_n := (\psi_{1n}, \psi_{2n}, \ldots, \psi_{dn})$ of $\psi_0$ is available, and that for each $P \in \mathcal{M}$, $n^{1/2}(\psi_n - \psi_0)$ tends in distribution to a random vector $U_0$ following the $d$-dimensional normal distribution $Q_0$ with mean zero and positive definite covariance matrix $\Sigma_0 = \Sigma_0(P_0)$. We define $U_n := n^{1/2}\psi_n$, and note that $U_n$ tends in distribution to $U_0$ provided $P_0 \in \mathcal{M}_0$. In this work, the statistic $U_n$ will be used as a basis for the tests we construct. Our primary focus is on applications in which $\psi_n$ is an asymptotically linear estimator of $\psi_0$, in which case $\Sigma_0$ can be characterized in terms of the (multivariate) influence function of $\psi_n$. Below, we will utilize knowledge of this influence function to determine what values of $\psi_n$ are far enough from the zero vector to warrant rejecting the null hypothesis. Often, this task is accomplished by identifying a multivariate region $\Theta_0 \subseteq \mathbb{R}^d$ such that the test rejecting $H_0$ if and only if $U_n \in \Theta_0$ has type I error that tends to the nominal type I error $\alpha \in (0, 1)$ as $n \to \infty$. Provided $\Theta_0$ is a continuity set of $Q_0$, this property is achieved if $\int I\{u \in \Theta_0\}\, dQ_0(u) = \alpha$ whenever $P_0 \in \mathcal{M}_0$. There are typically infinitely many choices of $\Theta_0$, and it may be unclear which to select in practice. Instead, for a given norm $\varphi$ on $\mathbb{R}^d$, we propose to search for a univariate region $\Theta_0^* \subseteq \mathbb{R}$ such that $\int I\{\varphi(u) \in \Theta_0^*\}\, dQ_0(u) = \alpha$ whenever $P_0 \in \mathcal{M}_0$. Then, an asymptotically calibrated test is defined by rejecting $H_0$ if and only if $\varphi(U_n) \in \Theta_0^*$. Use of the norm $\varphi$ thus allows conversion of the original multivariate problem into a univariate one.

In practice, there are many choices for $\varphi$, and as we will see, the norm used plays an important role in determining the performance of the resulting test. As an example, we consider the $\ell_p$-norm $\varphi_p$ defined as $(z_1, z_2, \ldots, z_p) \mapsto \|z\|_p := (z_1^p + z_2^p + \ldots + z_d^p)^{\frac{1}{p}}$ along with regions of the form $\Theta_0^*(r) = [r, \infty)$. The choice $r_0 := \min\{r : \int I\{\|u\|_p \geq r\}\, dQ_0(u) \leq \alpha\}$ ensures that $\Theta_0^* := \Theta_0^*(r_0)$ provides a calibrated test, in the sense that the test rejecting $H_0$ if and only if $\|U_n\|_p \in \Theta_0^*$ has asymptotic type I error equal to $\alpha$. The corresponding $p$-value is given by $\int I\{\|u\|_p \geq \|U_n\|_p\}\, dQ_0(u)$. Different choices of $p$ may yield tests with a different power profile over various alternatives. To explore this phenomenon, we may consider a simple example comparing tests resulting from the Euclidean

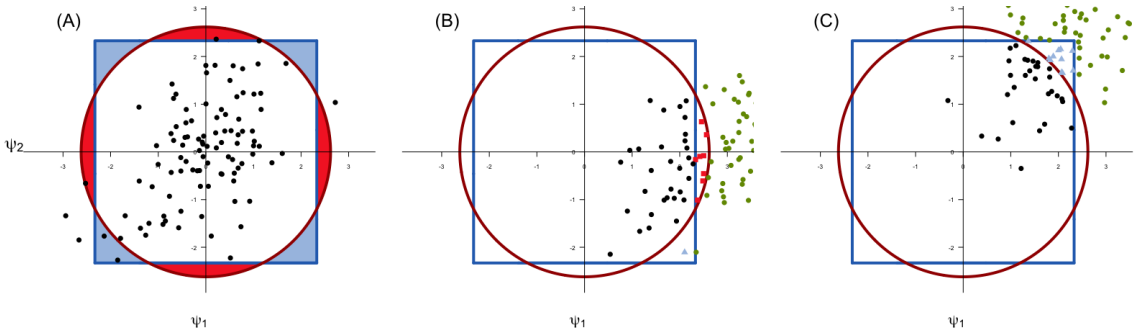


Figure 1.1: Plots of 100 observations from a limiting distribution of a hypothetical vector of parameter estimators in $\mathbb{R}^2$ (A) under the null, (B) under an alternative with $\psi_1 = 0, \psi_2 \neq 0$, and (C) under an alternative with $\psi_1, \psi_2 \neq 0$. The 95% quantiles for the data based on the max (blue) and $\ell_2$ (red) norms under the null are given in all three panels. If a test statistic fell within the blue regions the test would fail to reject $H_0$ if the $\ell_\infty$ norm was used but would reject $H_0$ if the $\ell_2$ norm was used. The converse is true for the red regions. Depending on the alternative, the $\ell_\infty$ norm (B) or the $\ell_2$ norm(C) will achieve higher power.

norm ($p = 2$) versus the maximum norm ($p = \infty$). Figure 1.1 illustrates the behavior of these two tests in the case $d = 2$. In Panel A, 100 draws are taken from a multivariate normal distribution $U_0$ with mean zero and identity covariance matrix. The red circle and blue square represent the boundary of the region $\Theta_0$ of these two tests constructed using empirical estimates of the $95^{\text{th}}$ percentile of the distributions of $\|U_0\|_2$ and $\|U_0\|_\infty$, respectively. All observations in Panel A except the five with the largest $\ell_2$-norm are contained within the red circle. Similarly, the blue square contains all observations in Panel A except the five with the largest $\ell_\infty$-norm. Observations that fall within the blue shaded region result in rejection of the null hypothesis if the $\ell_2$-norm is used to define the test but not if the $\ell_\infty$-norm is instead used. Conversely, observations that fall in the red shaded region result in rejection of the null hypothesis if the $\ell_\infty$-norm is used to define the test but not if the $\ell_2$-norm is instead used. The same square and circle are redrawn in panels B and C to illustrate the behavior of the test under alternatives in which either (B) $\psi_{10} = 0$ and $\psi_{20} \neq 0$, or (C) $\psi_{10} \neq 0$ and $\psi_{20} \neq 0$. While both constructions of a rejection region result in valid asymptotic type I error control, depending on the alternative, one test will outperform the other in power. In Panel B, shifting each observation in only a single direction has a larger impact on the maximum norm of the observations compared to the $\ell_2$-norm since the maximum norm only considers the largest coordinate. This is shown by the numerous observations (given by red squares) outside of the blue box (equivalent to rejecting $H_0$) and inside the red circle (equivalent to failing to reject $H_0$). In contrast, there is only a single observation outside the red circle and inside the blue box (given by blue triangles). The converse trend is shown in panel C, where the $\ell_2$-norm performs better because it takes into account both coordinates of the shift.

### 1.3.2 Adaptive norm selection

We denote by $\mathcal{F}$ the collection of all norms defined on $\mathbb{R}^d$. So far, we have argued that a test can be defined based on any $\varphi \in \mathcal{F}$ and that the choice of $\varphi$ can influence the power of the test. In many scenarios, it may not be clear a priori which of several tests should be preferred in a given setting since the power of each test depends on details of the true (unknown) alternative. In order to compare any of several candidate norms, we must first choose an objective criterion for adjudicating, in the setting at hand, the performance of the test statistic $\varphi(U_n)$ for a given norm $\varphi$.

For this purpose, suppose that $\Gamma_0^d : \mathbb{R}^d \times \mathcal{F} \to [0, \infty)$ provides a local measure of test inefficiency. Specifically, we stipulate that for any $x \in \mathbb{R}^d \backslash \{0\}$ and $\varphi \in \mathcal{F}$, greater values of $\Gamma_0^d(x, \varphi)$ indicate a larger asymptotic type II error — and so, lower power — for the test based on the test statistic $\varphi(U_n)$ under a location shift by $x$ of the null limiting distribution of $U_n$ under sampling from $P_0$. In this work, we focus on two particular measures, although our theoretical results are stated in generality. The first, which we refer to as the *acceptance rate* measure, is defined as

$$\Gamma_{\mathrm{ar},0}^d : (x, \varphi) \mapsto \int \mathbf{1}\{\varphi(u + x) \leq c_0\} \, dQ_0(u) \ , \tag{1.3}$$

where $c_0 := \min\{c \geq 0 : \int \mathbf{1}\{\varphi(u) \leq c\} \, dQ_0(u) \geq 1 - \alpha\}$ is the smallest cutoff value such that the test rejecting $H_0$ if and only if $\varphi(U_n) > c_0$ has asymptotic type I error equal to $\alpha$. This measure can be interpreted as the asymptotic type II error of the test based on $\varphi(U_n)$ in the context of a sequence of local alternatives under which $\psi_0 = \psi_0^{(n)} := xn^{-1/2}$. While it is intuitively simple and straightforward to estimate in practice, this measure can suffer from the fact that its output is constrained in the interval $[0, 1 - \alpha]$, so that it becomes less informative — and thus less useful for discriminating norms — in settings in which the distribution of $\Gamma_0^d(U_n, \varphi)$ is concentrated near zero for each norm $\varphi$ considered. Additionally, in view of the exponential tails of the normal distribution, $\Gamma_0^d(x, \varphi)$ tends to zero rapidly as $x$ tends away from the origin, thereby rendering onerous the task of achieving sufficient relative precision when approximating $\Gamma_0^d(U_n, \varphi)$ using Monte Carlo methods. These difficulties motivate the consideration of an alternative measure defined as

$$\Gamma_{\mathrm{mf},0}^d : (x, \varphi) \mapsto \min\left\{ s \geq 0 : \int \mathbf{1}\{\varphi(u + sx) \leq c_0\} \, dQ_0(u) \leq \tau \right\} \tag{1.4}$$

for some user-specified $\tau \in (0, 1 - \alpha)$. We refer to this as the *multiplicative factor* measure since it provides the smallest factor $\kappa$ such that the asymptotic type II error of the test based on $\varphi(U_n)$ is no greater than $\tau$ in the context of a sequence of local alternatives under which $\psi_0^{(n)} := \kappa x n^{-1/2}$.

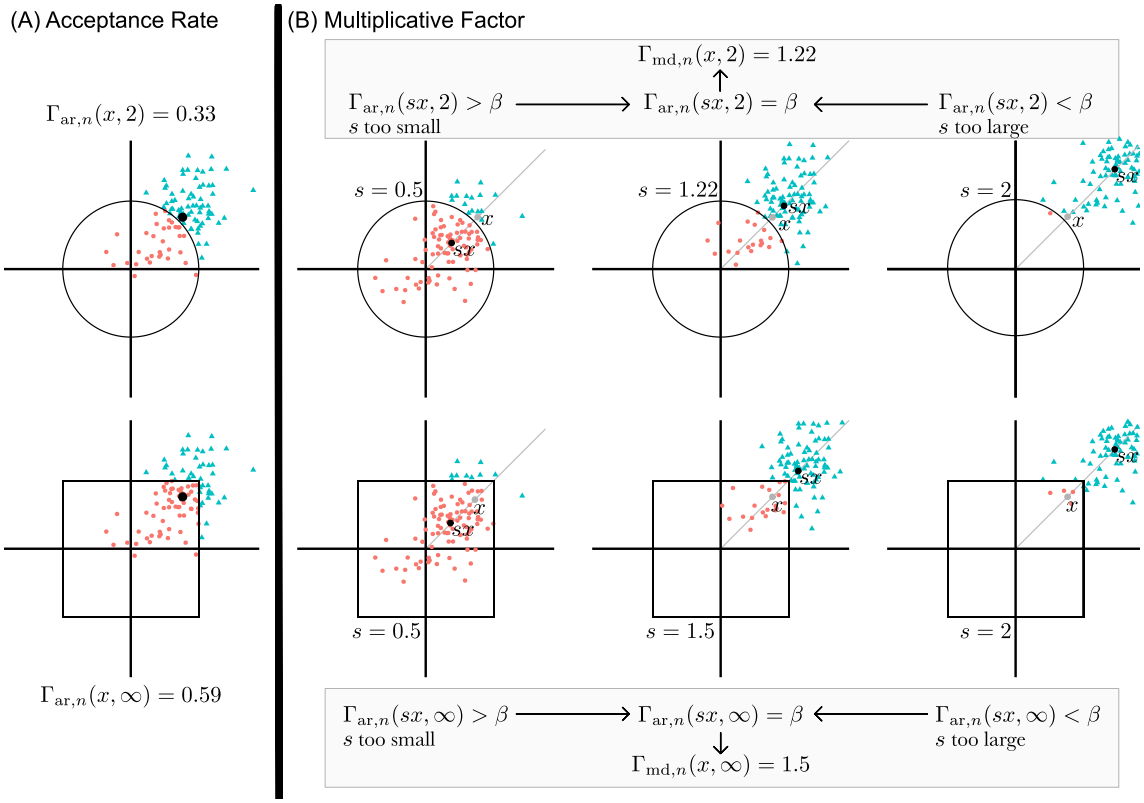


Figure 1.2: This figure illustrates the two proposed performance measures. In each figure, the estimated acceptance region of the norm-based test is shown, with the top row using the $\ell_2$ norm and the bottom row using $\ell_\infty$ norm. In panel $A$, the estimated acceptance rate measure is shown. The measure is estimated by taking draws from the estimated null distribution, shifting each draw by $x$, and calculating the number of observations inside the acceptance region. In panel B, estimation of the multiplicative factor measure is shown. On the left, an initial guess of 0.5 for the multiplicative factor is considered. Because this results in an acceptance rate that is larger than $\beta$ (here $\beta = 0.8$), a larger guess of $s$ is considered until the acceptance rate is equal to $\beta$.

This measure avoids the drawback of the acceptance rate by operating on a multiplicative scale, though it does so at the expense of simplicity of interpretation and computational ease. The computation of these two measures is illustrated in Figure 1.2

To motivate our method, suppose that we consider a finite collection $\mathcal{F}_0 := \{\varphi_1, \varphi_2, \ldots, \varphi_K\} \subset \mathcal{F}$ of norms on $\mathbb{R}^d$, which we wish to discriminate based on a given local measure of test inefficiency $\Gamma_0^d$. Suppose also that an estimator $\Gamma_n^d$ of $\Gamma_0^d$ based on $X_1, X_2, \ldots, X_n$ is available. Then, it is sensible to consider $\Gamma_n^d(U_n, \varphi)$ as an estimated local measure of test inefficiency for a given norm $\varphi$, where local here refers to the consideration of local alternatives defined by $U_n$ itself. As a first attempt at developing a test based on adaptive norm selection, we could consider using the test statistic $\varphi_{k_n(U_n)}(U_n)$ with $k_n(U_n) := \operatorname{argmin}_k \Gamma_n^d(U_n, \varphi_k)$ — this amounts to considering the univariate summary $\varphi(U_n)$ based on the norm $\varphi \in \mathcal{F}_0$ with the smallest estimated local measure of test inefficiency. However, the test statistic $\varphi_{k_n(U_n)}(U_n)$ appears difficult to make valid inference with since its limit distribution is difficult to derive — for example, the lack of continuity of $\varphi_{k_n(U_n)}(U_n)$ as a function of $U_n$ precludes the use of a continuous mapping theorem. More

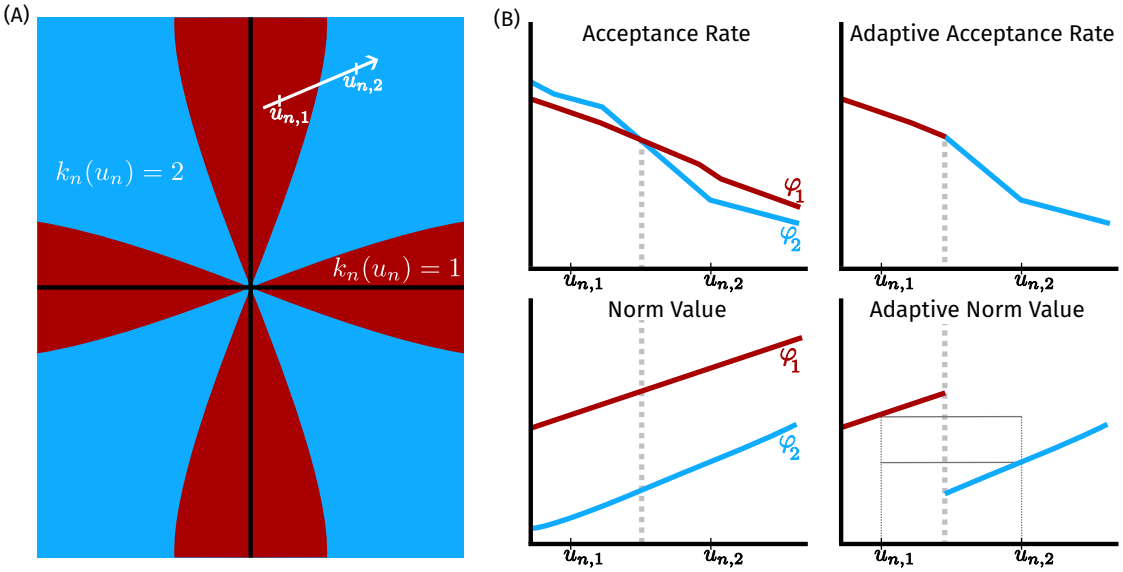


Figure 1.3: This figure illustrates two issues that could arise when using the adaptive norm value as a test statistic. (A) shows regions of $\mathbb{R}^2$ in which $\varphi_1$ (dark red) or $\varphi_2$ (light blue) have better (hypothetical) acceptance rate value. A line segment containing two points $u_{n,1}$ and $u_{n,2}$ is also shown, and the points along this line segment form the $x$-axis of the four figures in (B). The arrow indicates the direction along the line segment in which both $\varphi_1$ and $\varphi_2$ increase. The top left display in (B) shows the hypothetical values of the acceptance rate measure along the $u_n$-values shown on the white arrow in (A), and the top right panel shows the adaptive version of this measure (the pointwise minimum of the individual acceptance rate measures). The bottom left display indicates the norm values, and the bottom right display shows the adaptive norm value wherein the norm with lowest acceptance rate is used. As shown by the two horizontal line segments in this display, the adaptive norm value does not necessarily increase as $u_n$-values are taken further away from the origin. Additionally, the discontinuity of the adaptive norm value is apparent in the bottom right display.

importantly, this test statistic produces an undesirable ordering in the space of alternatives, as illustrated in Figure 1.3 with a simple example in which $\mathcal{F}_0 = \{\varphi_1, \varphi_2\}$ contains only two given norms, and the parameter of interest takes values in $\mathbb{R}^2$. In the figure, for each alternative, the color indicates which of $\varphi_1$ (dark red) and $\varphi_2$ (light blue) is preferred to define a test. However, the norm $\varphi_1$ takes values that are substantially larger than $\varphi_2$ for observations that are similar. As a result, when considering the points $u_{n,1}$ and $u_{n,2}$, even though $\varphi_1(u_{n,1}) < \varphi_1(u_{n,2})$ and $\varphi_2(u_{n,1}) < \varphi_2(u_{n,2})$, it is also true that $\varphi_{k_n(u_{n,2})}(u_{n,2}) < \varphi_{k_n(u_{n,1})}(u_{n,1})$. Thus, even though $u_{n,2}$ is further away from the null as measured by both norms, the relative size of the norms makes it appear as though $u_{n,1}$ is more surprising than $u_{n,2}$ under the null according to the adaptive norm test statistic.

In view of these challenges, we consider another strategy for building an aggregate test statistic. We observe that for any norm $\varphi$, if $\Gamma_0^d$ were known, $\Gamma_0^d(U_n, \varphi)$ could serve as a sensible alternative to the test statistic $\varphi(U_n)$, with smaller values of $\Gamma_0^d(U_n, \varphi)$ supporting rejection of the null hypothesis. The use of $\Gamma_0^d(U_n, \varphi)$ as a test statistic has desirable properties. First, the interpretation of the realizations of $\Gamma_0^d(U_n, \varphi)$ depends neither on the norm used nor on the limiting distribution $Q_0$. As

a result, the value of $\Gamma_0^d(u_n, \varphi)$ can be directly compared across choices of $\varphi$, thereby facilitating the construction of an adaptive test statistic. Second, the ordering induced on the parameter space by $\Gamma_0^d$ is sensible. To illustrate this, suppose first that two realizations $u_{n,1}$ and $u_{n,2}$ of $U_n$ fall on a common ray, that is, $u_{n,1} = \beta_1 v$ and $u_{n,2} = \beta_2 v$ for some direction $v \in \mathbb{R}^d$ and non-negative values $\beta_1, \beta_2 \in \mathbb{R}$. For $\beta_1 > \beta_2$, we expect $u_{n,1}$ to be less likely than $u_{n,2}$ under the null, and indeed, in these settings $\Gamma_0^d(u_{n,1}, \varphi) < \Gamma_0^d(u_{n,2}, \varphi)$ under regularity conditions introduced in the next section. Suppose instead that $u_{n,1}$ and $u_{n,2}$ are such that $\varphi(u_{n,1}) = \varphi(u_{n,2})$. In this case, depending on $Q_0$, either $u_{n,1}$ or $u_{n,2}$ could be more likely. A test statistic based on $\Gamma_0^d$ allows for consideration of $Q_0$ and thus permits differentiation of $u_{n,1}$ and $u_{n,2}$ even when these realizations may be undifferentiated by $\varphi$.

In practice, since $\Gamma_0^d$ is unknown, the test statistic $\Gamma_n^d(U_n, \varphi)$ would be used instead of $\Gamma_0^d(U_n, \varphi)$. We observe that, by definition,

$$Z_n := \Gamma_n^d(U_n, \varphi_{k_n(U_n)}) = \min\{\Gamma_n^d(U_n, \varphi_1), \Gamma_n^d(U_n, \varphi_2), \ldots, \Gamma_n^d(U_n, \varphi_K)\} \; . \tag{1.5}$$

As such, the adaptive test statistic $Z_n$ is a continuous transformation of the single-norm statistics $\Gamma_n^d(U_n, \varphi_1), \Gamma_n^d(U_n, \varphi_2), \ldots, \Gamma_n^d(U_n, \varphi_K)$. Provided $u \mapsto \Gamma_0^d(u, \varphi)$ is continuous for each $\varphi \in \mathcal{F}_0$, this implies that a non-degenerate limit distribution can be derived for the test statistic $Z_n$, thereby facilitating valid inference. Specifically, under regularity conditions, we may expect $Z_n$ to converge in distribution to the random variable $Z_0 := \min\{\Gamma_0^d(U_0, \varphi_1), \Gamma_0^d(U_0, \varphi_2), \ldots, \Gamma_0^d(U_0, \varphi_K)\}$, where $U_0$ is distributed according to $Q_0$. This motivates an adaptive test in which we reject $H_0$ if and only if $Z_n > \chi_n$, where $\chi_n$ is any consistent estimator of the $(1-\alpha)$-quantile $\chi_0$ of the distribution of $Z_0$.

We note here that our proposed procedure was inspired by the proposal of Zhang and Laber (2015), which can be considered a special case of our framework. Their method, which focuses on the specific problem of testing for null correlations in the setting of univariable linear models, is recovered by taking $\psi_n$ to be the vector of sample correlations, $\mathcal{F}_0$ to be a collection of sum-of-squares norms, and $\Gamma_0^d$ to be the observed $p$-value for the test based on $\varphi(U_n)$. The sum-of-squares norm is defined as

$$\jmath_k : (z_1, z_2, \ldots, z_d) \mapsto \left\{\sum_{j=1}^k z_{(d-j+1)}^2\right\}^{1/2}$$

for any fixed $k \in \{1, 2, \ldots, d\}$ and with $z_{(r)}^2$ denoting the $r^{th}$ order statistic based on $z_1^2, z_2^2, \ldots, z_d^2$ for each $r = 1, 2 \ldots, d$. A proof that the sum-of-squares norm is indeed a proper norm is provided in Lemma 6 of the Appendix.

### 1.3.3 Implementation of proposed adaptive test

Suppose that $\Sigma_n$ is a consistent estimator of $\Sigma_0$, and denote by $Q_n$ the distribution function of the normal distribution with mean zero and covariance matrix $\Sigma_n$. An estimator $\Gamma_n^d$ can be derived by replacing $Q_0$ with $Q_n$ in the definition of $\Gamma_0^d$. We set $Z_n^* := \min\{\Gamma_n^d(\bar{U}_n, \varphi_1), \Gamma_n^d(\bar{U}_n, \varphi_2) \ldots, \Gamma_n^d(\bar{U}_n, \varphi_K)\}$, where $\bar{U}_n$ represents a random draw from $Q_n$, and note that $Z_n^*$ serves as a natural proxy for a random draw from the null limit distribution of $Z_n$. Because the distribution of $Z_n$ is difficult to calculate in practice, we instead define our cutoff value $\chi_n^*$ as the $(1-\alpha)$-quantile of $Z_n^*$. Below, we will establish properties of the test in which we

$$\text{reject } H_0 \text{ if and only if } Z_n > \chi_n^* \,. \tag{1.6}$$

While an analytic form $\chi_n^*$ is not currently available, its value can be approximated with an arbitrary level of accuracy using the following steps:

1. for $M$ large, conditionally on $Q_n$, generate independent draws $\bar{U}_{n,1}, \bar{U}_{n,2}, \ldots, \bar{U}_{n,M}$ from $Q_n$;

2. set $\bar{Z}_{n,m} := \min\{\Gamma_n^d(\bar{U}_{n,m}, \varphi_1), \Gamma_n^d(\bar{U}_{n,m}, \varphi_2), \ldots, \Gamma_n^d(\bar{U}_{n,m}, \varphi_K)\}$ for $m = 1, 2, \ldots, M$;

3. compute the sample $(1-\alpha)$-quantile $\chi_{n,m}^*$ based on $\{\bar{Z}_{n,1}, \bar{Z}_{n,2}, \ldots, \bar{Z}_{n,M}\}$.

For concreteness of discussion, suppose that, for each $j = 1, 2, \ldots, K$, $\psi_{jn}$ is an asymptotically linear estimator of $\psi_{j0}$ with influence function $\phi_j : \mathcal{X} \to \mathbb{R}$, in the sense that

$$\psi_{jn} = \psi_{j0} + \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i) + o_P(n^{-1/2})$$

with $E_0\{\phi_j(X)\} = 0$ and $var_0\{\phi_j(X)\} < \infty$. Suppose that the form of each $\phi_j$ is known up to some dependence on the unknown data-generating distribution $P_0$. Asymptotic linearity of $\psi_n$ readily implies that, under the null hypothesis, $U_n$ tends to a random vector following a multivariate normal distribution with mean zero and covariance matrix $\Sigma_0$ with $jk^{th}$ element $\Sigma_{jk0} := \int \phi_j(x)\phi_k(x)\, dP_0(x)$. We will require a consistent estimator $\Sigma_n$ of $\Sigma_0$ in our developments — a natural candidate is the empirical cross-moment estimator, defined entrywise as $\Sigma_{jkn} := \frac{1}{n}\sum_{i=1}^{n} \phi_{jn}(X_i)\phi_{kn}(X_i)$, where $\phi_{jn}$ and $\phi_{kn}$ are estimators of the influence functions $\phi_j$ and $\phi_k$, respectively. While for simplicity this empirical estimator is employed in all simulations and data analyses reported below, more sophisticated procedures for covariance estimation — e.g., as described by Ledoit and Wolf (2004, 2020) — could be used instead. The implementation of our approach also requires the selection of a collection $\mathcal{F}_0$ of norms. In this article, we explicitly consider the $\ell_p$ and sum-of-squares norms.

## 1.4 Large-sample properties of proposed test

In this section, we establish conditions under which the adaptive test outlined in (1.6) is guaranteed to have desirable statistical properties. In addition to type I error control and consistency against fixed alternatives, we will show that our proposed test has nontrivial power against local alternatives. For each theorem in this Section, a proof is provided in the Appendix.

Since $\Gamma_0^d$ depends on $Q_0$ only through $\Sigma_0$, we explicitly denote the local measure of test inefficiency as a fixed mapping $(u, \Sigma, \varphi) \mapsto \Gamma^d(u, \Sigma, \varphi)$ for which we have that $\Gamma_0^d(u, \varphi) = \Gamma^d(u, \Sigma_0, \varphi)$ for each $u \in \mathbb{R}^d$ and $\varphi \in \mathcal{F}_0$. For simplicity, we consider an arbitrary estimator $\Gamma_n^d$ of $\Gamma_0^d$ of the form $\Gamma_n^d(u, \varphi) = \Gamma^d(u, \Sigma_n, \varphi)$ for each $u \in \mathbb{R}^d$ and $\varphi \in \mathcal{F}_0$, where $\Sigma_n$ is any consistent estimator of $\Sigma_0$. By explicitly representing the dependence of $\Gamma_0^d$ and $\Gamma_n^d$ on $\Sigma_0$ and $\Sigma_n$, respectively, via $\Gamma^d$, the consistency of $\Gamma_n^d$ to $\Gamma_0^d$ can be established as a consequence of a simple continuity condition on $\Gamma^d$. We introduce the following conditions on the local measure of test inefficiency relative to a given norm $\varphi$, where we denote by $\mathbb{V}_d$ the space of all positive definite $d \times d$ matrices:

C1) $(u, \Sigma) \mapsto \Gamma^d(u, \Sigma, \varphi)$ is continuous and non-negative on $\mathbb{R}^d \times B_0$ for some neighborhood $B_0 \subset \mathbb{V}_d$ of $\Sigma_0$;

C2) $\int I\{\Gamma_0^d(u, \varphi) = t\}\, dQ_0(u) = 0$ for every $t \geq 0$;

C3) $\Gamma^d(x_s, \Sigma, \varphi) \to 0$ uniformly over $\Sigma \in B_1$ for some neighborhood $B_1 \subset \mathbb{V}_d$ of $\Sigma_0$ for every sequence $x_1, x_2, \ldots$ of elements of $\mathbb{R}^d$ such that $\varphi(x_s) \to \infty$;

C4) $u \mapsto \Gamma_0^d(u, \varphi)$ is quasi-concave, in the sense that $\{u : \Gamma_0^d(u, \varphi) \geq a\}$ is convex for every $a \geq 0$;

C5) $u \mapsto \Gamma_0^d(u, \varphi)$ is centrally symmetric, in the sense that $\Gamma_0^d(u, \varphi) = \Gamma_0^d(-u, \varphi)$ for every $u \in \mathbb{R}^d$.

The result below states that, under mild conditions, the proposed test has valid type I error rate and power tending to one under each fixed alternative as sample size tends to infinity.

**Theorem 1.** *Suppose that conditions C1–C2 hold for each $\varphi \in \mathcal{F}_0$. Under sampling from $P_0$, as $n \to \infty$, the rejection rate of the proposed test* (1.6)*:*

*a) tends to $\alpha$ if $P_0 \in \mathcal{M}_0$;*

*b) tends to 1 if $P_0 \notin \mathcal{M}_0$ provided condition C3 also holds for some $\varphi \in \mathcal{F}_0$.*

Since in practice studies are typically designed to have power substantively below one in view of cost and other logistic constraints, studying the asymptotic behavior of the proposed test for these settings is of interest and motivates consideration of local alternatives. Specifically, a local

alternative to $P_0 \in \mathcal{M}_0$ is a one-dimensional parametric submodel $\{P_t\} \subset \mathcal{M}$ of $\mathcal{M}$ dominated by $P_0$ and such that the Radon-Nikodym derivative of $P_t$ relative to $P_0$ satisfies, for $t$ in a neighborhood of zero,

$$\frac{dP_t}{dP_0}(x) = 1 + tg(x) + tr_t(x) \tag{1.7}$$

for some element $g$ in the tangent space of $\mathcal{M}$ but not in the tangent space of $\mathcal{M}_0$ at $P_0$, and where $r_t$ is a remainder term tending to zero in a uniform sense (Pfanzagl, 1990). The estimator $\psi_n$ is said to be regular at $P_0 \in \mathcal{M}$ if the limit distribution of $n^{1/2}(\psi_n - \psi_0)$ under sampling from $P_0$ and of $n^{1/2}(\psi_n - \psi_0^{(n)})$ under sampling from $P_0^{(n)} := P_t|_{t=n^{-1/2}}$ is the same, where we write $\psi_0^{(n)} := \Psi(P_0^{(n)})$ and $\{P_t\}$ is any local alternative to $P_0$. We note that, for any such sequence $P_0^{(n)}$, it holds that $\psi_0^{(n)} = an^{-1/2} + o(1)$ for some $a \in \mathbb{R}^d \setminus \{0\}$. The following theorem states that, under certain regularity conditions, if the estimator $\psi_n$ is regular, then the proposed test is locally unbiased in the sense that it has non-trivial power under local alternatives.

**Theorem 2.** *Suppose that $P_0 \in \mathcal{M}_0$, and let $P_0^{(n)}$ be a sequence of local alternatives converging to $P_0$. Suppose also that conditions C1, C2, C4 and C5 hold for each $\varphi \in \mathcal{F}_0$, that condition C3 holds for some $\varphi \in \mathcal{F}_0$, and that $\psi_n$ is a regular estimator of $\psi_0$ at $P_0$. Then, the rejection rate $\pi_n$ of the proposed test under sampling from $P_0^{(n)}$ satisfies that* $\liminf_n \pi_n > \alpha$.

This theorem guarantees that the asymptotic rejection rate is greater under local alternatives than it is under the null. Theorems 1 and 2 indicate that the proposed test has desirable properties provided several conditions on the local measure of test inefficiency used hold. The next result establishes that the two measures presented in Section 3, namely the acceptance rate and multiplicative factor measures, indeed satisfy all required conditions, and therefore, can be used in our procedure.

**Theorem 3.** *Both the acceptance rate measure* (1.3) *and the multiplicative factor measure* (1.4) *satisfy conditions C1–C5 for each norm $\varphi$.*

## 1.5 Numerical examples

In this section, we discuss the implementation and evaluate the performance of our proposed test in the context of the three working examples introduced in Section 1.2.

In each example, we consider all combinations of sample size $n \in \{100, 200, 500\}$ and covariate vector dimension $d \in \{10, 50, 100\}$. The multiplicative factor measure (1.4) is used throughout. We compare a variety of competing procedures, including adaptive and non-adaptive versions of our test. The non-adaptive tests use the $\ell_2$ and maximum absolute value norms and are referred

to as the $\ell_2$ and $\ell_\infty$ tests, respectively. The first adaptive version of our test selects over the $\ell_1$, $\ell_2$, $\ell_4$, $\ell_6$ and $\ell_\infty$ norms, and is referred to as the adaptive $\ell_p$ test. The second adaptive test selects over various versions of the $\jmath_k$ norm — specifically, over $k \in \{1, 3, 5, 6, 8, 10\}$ when $d = 10$, $k \in \{1, 11, 21, 30, 40, 50\}$ when $d = 50$, and $k \in \{1, 21, 41, 60, 80, 100\}$ when $d = 100$ — and is referred to as the sum-of-squares test. We note that $\jmath_1 = \ell_\infty$ and $\jmath_d = \ell_2$. We contrast the performance of these adaptive procedures with two existing all-purpose methods for multiple testing. Each all-purpose method (including ours) uses the same covariance matrix estimator $\Sigma_n$ and parameter estimator $\psi_n$. The first is a test based on the Bonferroni-corrected $p$-value $d \times \min(p_{1n}, p_{2n}, \ldots, p_{dn})$ computed from individual $p$-values $p_{jn} := 2\,[1 - \Phi(z_j)]$, where $z_j := n^{1/2}|\psi_{nj}|/\sigma_{jn}$, $\sigma_{jn}$ is a consistent estimator of the asymptotic standard deviation $\sigma_{0j}$ of $n^{1/2}\psi_{jn}$ under the null hypothesis, and $\Phi$ represents the standard normal distribution function. In our simulations, we take $\sigma_{jn}$ to be the root of the empirical second moment of $\phi_{nj}(X_1), \phi_{nj}(X_2), \ldots, \phi_{nj}(X_n)$, where $\phi_{nj}$ is a consistent plug-in estimator of the influence function $\phi_{0j}$ of $\psi_{nj}$. The second is the more recent Cauchy combination test (referred to here as the Cauchy test) described by Liu and Xie (2020) based on the test statistic $d^{-1} \sum_{j=1}^{d} \tan\{(2p_{jn} - 3/2)\pi\}$. Under certain conditions, including mutual independence of $\psi_{n1}, \psi_{n2}, \ldots, \psi_{nd}$, this test statistic has a limiting Cauchy distribution under the null hypothesis. However, Liu and Xie (2020) show that even when independence fails to hold, $p$-values computed using the Cauchy distribution are approximately valid for large realizations of the test statistic. For this reason, and in view of its simplicity, we include this test as a comparator in our simulation studies.

### 1.5.1 Example 1: correlation

In this example, we consider the settings described in the first example of McKeague and Qian (2015) and Zhang and Laber (2015). The vector $W = (W_1, W_2, \ldots, W_d)$ of covariates is generated from a normal distribution with mean zero and covariance matrix with diagonal and off-diagonal terms equal to 1 and $\rho$, respectively. Three distinct conditional outcome distributions are considered. In each setting, we generate $\varepsilon$ as a standard normal variable independent of $W$. Conditionally on $W$ and $\epsilon$, we separately consider

(Setting 1) $Y = \varepsilon$;

(Setting 2) $Y = 0.25W_1 + \varepsilon$;

(Setting 3) $Y = 0.15\,(W_1 + \ldots + W_5) - 0.1\,(W_6 + \ldots + W_{10}) + \varepsilon$.

In this example, the sampling distribution of each test statistic — and thus cutoffs upon which to construct valid tests — can also be determined using two different methods. The standard approach, discussed above and referred to as the parametric bootstrap test, estimates the limiting

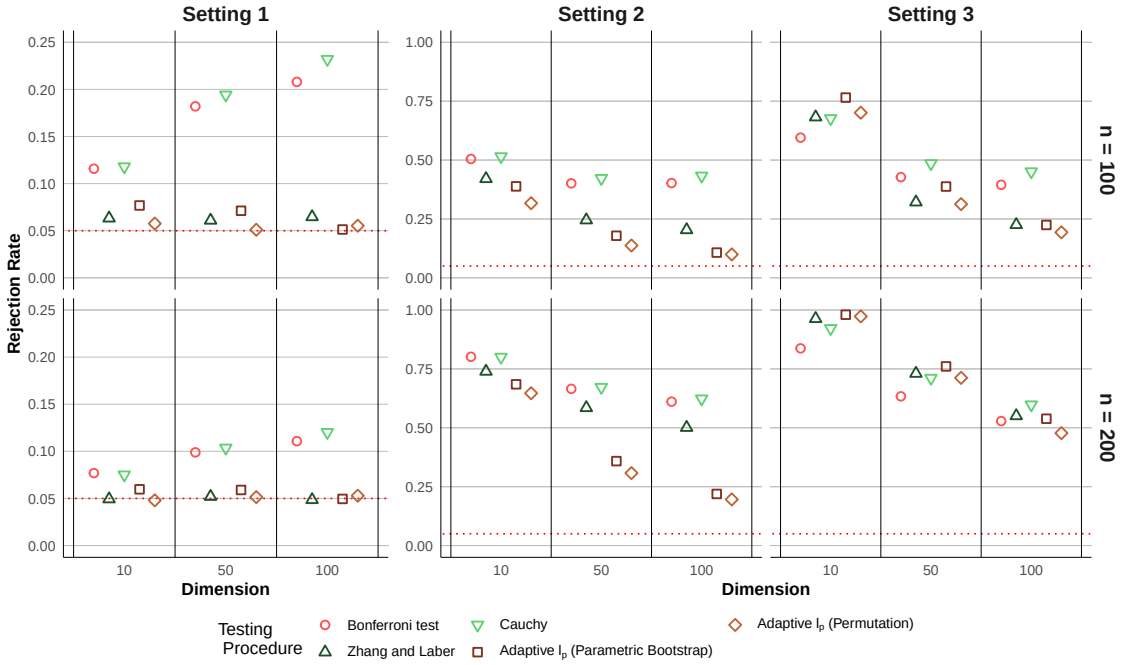


Figure 1.4: Empirical rejection rate of various tests applicable in Example 1 under different data-generating mechanisms, at different sample sizes, and for covariate vectors with no correlation across components and of different lengths.

distribution of $n^{1/2}(\psi_n - \psi_0)$ using a mean-zero normal distribution with an estimated covariance matrix. A permutation approach, which typically provides better calibration than the parametric bootstrap in smaller samples, can also be used in this particular example. A permutation-based approximation of the $p$-value of the test based on $Z_n$ can be obtained by independently generating modifications of the original dataset in which the outcome vector has been randomly permuted across observations, re-computing $Z_n$ for each such permuted dataset, and computing the fraction of permuted datasets for which the re-computed $Z_n$ value is larger than the original $Z_n$ value. In this simulation, permutation-based and parametric bootstrap versions of our adaptive test were compared to three competing tests, namely the test of Zhang and Laber (2015), the Bonferroni test, and the Cauchy test. We note here that the test of Zhang and Laber (2015) leverages knowledge about the data-generating mechanism, whereas other procedures considered instead make use of nonparametric parameter and covariance estimators.

The empirical rejection rates of the different tests considered are shown for the three described settings in Figures 1.4 and 1.5 with $\rho = 0$ and $\rho = 0.8$, respectively. Results for the intermediate setting $\rho = 0.5$ are provided in Figure A.1 in the Appendix. Results for $n = 500$ are not shown because power is very close to one for Settings 2 and 3. Because $H_0$ holds in Setting 1, we expect the rejection rates for this setting to be close to the nominal level 0.05. Figures 1.4 and 1.5 illustrate that this is achieved by every testing procedure evaluated except for the Bonferroni test and the Cauchy test. In Settings 2 and 3, $H_0$ does not hold and the plots convey the empirical power of

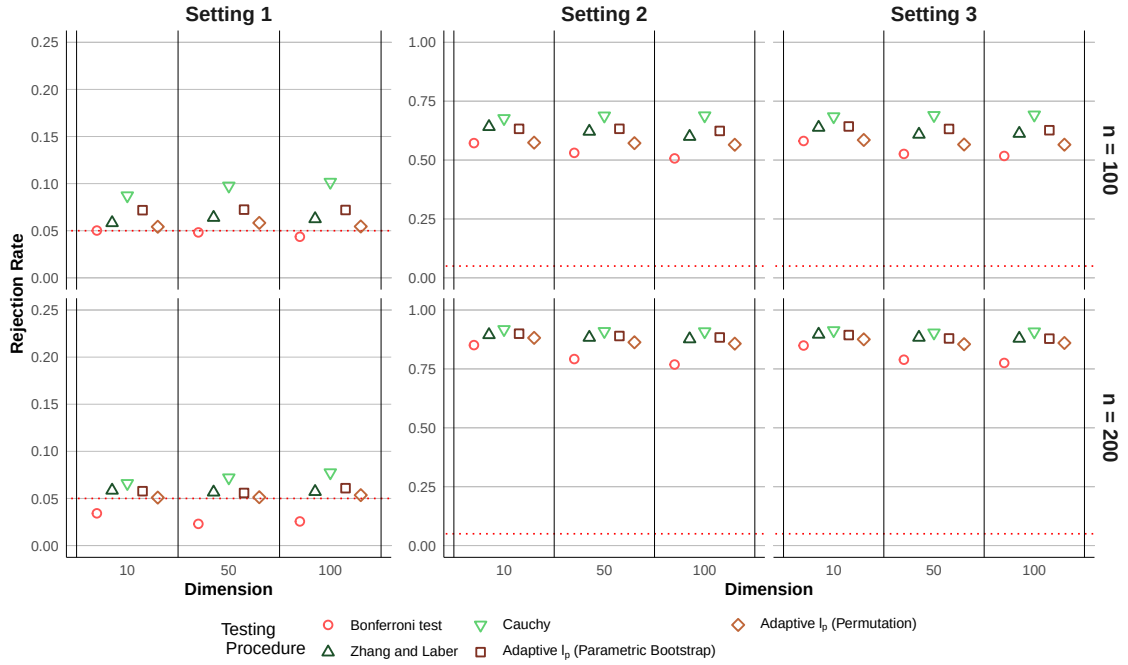


Figure 1.5: Empirical rejection rate of various tests applicable in Example 1 under different data-generating mechanisms, at different sample sizes, and for covariate vectors with high correlation (80%) across components and of different lengths.

the tests considered.

In most scenarios in which $H_0$ does not hold all tests have similar empirical power. In most settings, the test proposed by Zhang and Laber (2015) slightly outperforms all other tests, and in settings in which it is not the best, it only performs slightly worse than the best test. The most noticeable differences in performance are found in Setting 2 at sample size $n = 200$ under mutual independence of covariate component in which the Zhang and Laber (2015) test outperforms all others by a substantial margin. This superior performance is expected since the data-generating mechanism matches the model assumed in this method, whereas the other tests considered are nonparametric and therefore valid under weaker conditions.

Both of our adaptive tests perform similarly, with the permutation-based test having lower power but achieving better type I error control than the parametric bootstrap-based test. The relatively higher empirical power of the parametric bootstrap test compared to the permutation-based test likely stems from the fact that the null hypothesis tested is weaker for the latter (null marginal associations) than for the former (joint independence), and that the parametric bootstrap test is imperfectly calibrated, as evidenced by its slightly inflated type I error.

In this simulation, the Bonferroni test and the Cauchy test are anti-conservative, especially in settings in which there is no correlation between covariates. The failure of these tests to achieve nominal type I error control can mostly be attributed to difficulty estimating the variance of $\psi_n$. Figure A.2 in the Appendix shows the distribution of $p_{1n}$ in the setting in which there is no

correlation between covariate components. This distribution has a large spike near zero, although the spike is less pronounced at sample size $n = 200$ and is expected to dissipate as sample size further increases. We observe that when the $p$-value $p_{1n}$ is computed using the true standard error of $\psi_n$, this spike vanishes, suggesting that the poor small-sample calibration stems from estimation of the standard error. The observed over-representation of small $p$-values (relative to the uniform distribution) causes a large inflation in type I error for the Bonferroni test. The $p$-value of the Cauchy test is also sensitive to small values of $p_{jn}$ due to the vertical asymptote of the tangent function used to define the test statistic.

### 1.5.2 Example 2: coefficients of a working log-linear regression model under missingness

In the second example, conditionally on $W = w$, the binary outcome $Y$ is simulated from the logistic regression model $\text{pr}(Y = 1 \mid W = w) = \text{expit}(\beta_1 w_1 + \beta_2 w_2 + \ldots + \beta_d w_d)$. In all scenarios, the conditional missingness probability is given by $\text{pr}(\Delta = 1 \mid W = w) = \text{expit}(0.5 + 0.15 w_{d-1} - 0.275 w_d)$, and the vector $W = (W_1, W_2, \ldots, W_d)$ of covariates is drawn from a multivariate normal with mean zero and covariance matrix with diagonal and off-diagonal entries equal to 1 and 0.5, respectively. We separately consider the following settings defined by different values for the regression coefficient vector:

(Setting 1) $\beta_1 = \ldots = \beta_d = 0$;

(Setting 2) $\beta_1 = 0.6$, $\beta_2 = \ldots = \beta_d = 0$;

(Setting 3) $\beta_1 = \ldots = \beta_5 = 0.32$, $\beta_6 = \ldots = \beta_{10} = -0.32$, $\beta_{11} = \ldots = \beta_d = 0$;

(Setting 4) $\beta_1 = \ldots = \beta_5 = 0.23375$, $\beta_6 = \ldots = \beta_{10} = 0.4675$, $\beta_{11} = \ldots = \beta_d = 0$.

Thus, the null hypothesis holds in Setting 1 but not in any of Settings 2, 3, and 4. In this example, an influence function-based estimator of the covariance matrix $\Sigma_0$ was used, and conditional mean functions involved were estimated using either an elastic net (Simon et al., 2013; Tibshirani et al., 2012; Friedman et al., 2010; Tibshirani et al., 2012) or loess smoother.

In the null setting (Setting 1), we find that the type one error of all tests is near (though still slightly above) the 0.05 type one error rate. In general, the type one error is higher in settings with smaller sample size and larger dimension (as expected). In Setting 2, all tests have similar power with the Cauchy test slightly outperforming and the Bonferroni test slightly under-performing all other tests. The differences in performance are larger for settings with higher dimension. In Setting 3, the $\ell_\infty$ based test outperforms all others, especially in the sample size 500 setting. In Setting 4, all tests except the $\ell_\infty$ and Bonferroni test perform nearly identically well for each sample size and dimension. In Setting 3, ten covariates are associated with the outcome, which

would suggest that norms accounting for the many non-null associations would perform (relatively) better, as seen in Figure 1.1. Unexpectedly, the $\ell_\infty$ test had the largest power. This finding may be driven by the fact that other norms place larger importance on smaller component values. In this setting, while only ten covariates are directly associated with the outcome, all other covariates are still marginally associated with the outcome through their correlation with other covariates. While the $\ell_\infty$ norm considers only the covariate most strongly associated with the outcome, other norms consider all covariates. Covariates that are indirectly associated with the outcome thus have a small (though still non-zero) association with the outcome. If the additional variability introduced by including these covariates is too large, it may be detrimental to test performance. This explanation is supported by results presented in Figure 1.6, wherein we find that when all covariates are truly associated with the outcome ($d = 10$), all considered tests have comparable power, but that differences emerge in larger dimensions. It may also be that the low power of the adaptive and $\ell_2$ tests are a consequence of various linear effects on the outcome canceling each other out. Because covariate vector components are highly correlated and there are an equal number of positive and negative $\beta$ values of the same magnitude, the combined effect from all covariates could be small. This would make it more difficult to discern the marginal effect of any single covariate. In Setting 4, ten covariates are directly associated with the outcome, just like in Setting 3, though unlike Setting 3, all non-null regression coefficients are positive. These differences result in a reversal of which tests are optimal, with the $\ell_\infty$ test having the lowest power of the non-adaptive tests, and the adaptive tests and the $\ell_2$-based test all perform nearly equally, with the latter narrowly outperforming the former.

Overall, we see that depending on the scenario, the norm on which a test is based could be unimportant (Setting 2) or a source of substantial differences between tests (Setting 3). In settings in which the choice of the norm is consequential, the adaptive test does not outperform all fixed norm tests but does provide consistent performance across all settings. This example also suggests that common guidelines from the high-dimensional statistics literature on which norm should perform best in a given scenario may not be reliable, even when the data-generating mechanism is known a priori.

### 1.5.3 Example 3: coefficients of a working effect modification model for randomized trials

In this example, the covariate vector $W$ is drawn from a multivariate normal distribution with mean zero and a covariance matrix with diagonal and off-diagonal entries equal to 1 and 0.5, respectively. Given $W = w$, the binary exposure $A \in \{0, 1\}$ is drawn from a binomial distribution with a success probability of 0.5, as in a standard randomized trial. Finally, given $(A, W) = (a, w)$,

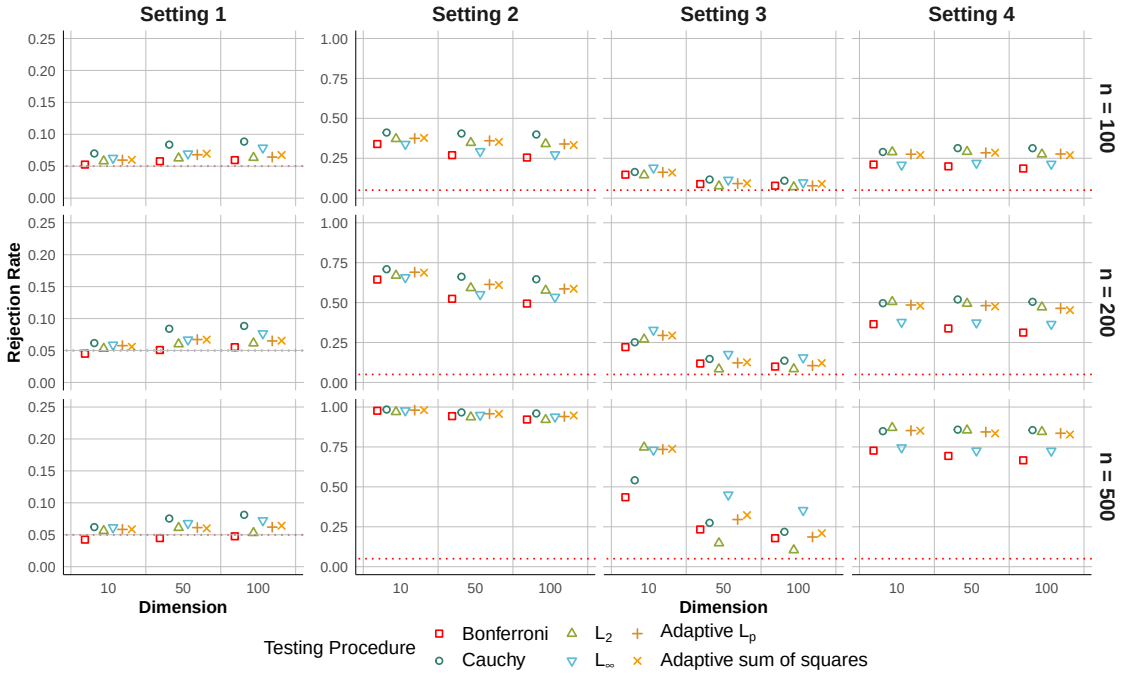


Figure 1.6: Empirical rejection rate of various tests applicable in Example 2 under different data-generating mechanisms, at different sample sizes, and for covariate vectors with moderate correlation (50%) across components and of different lengths.

the binary outcome $Y \in \{0, 1\}$ is drawn from a Bernoulli distribution with success probability given by

$$\text{logit pr}\left(Y = 1 \mid W = w, A = a\right) = \beta_0 a + \sum_{j=1}^{d} (\beta_j + \gamma_j a) w_j \ .$$

We set $\beta_0 = 0.2$, $\beta_1 = \ldots = \beta_{d/2} = 0.7/d$ and $\beta_{d/2+1} = \ldots = \beta_d = 0$, where $d$ is even, and consider the following settings:

(Setting 1) $\gamma_1 = \ldots = \gamma_d = 0$;

(Setting 2) $\gamma_1 = 1.2$ and $\gamma_2 = \ldots = \gamma_d = 0$;

(Setting 3) $\gamma_1 = \ldots = \gamma_5 = -0.8$, $\gamma_6 = \ldots = \gamma_{10} = 0.8$ and $\gamma_{11} = \ldots = \gamma_d = 0$;

(Setting 4) $\gamma_1 = \ldots = \gamma_5 = 0.07$, $\gamma_6 = \ldots = \gamma_{10} = 0.14$ and $\gamma_{11} = \ldots = \gamma_d = 0$.

Thus, the null hypothesis holds in the first setting, and the alternative holds in the three other settings. Calculation of parameter estimates and estimated influence functions required for inference was implemented using code adapted from the `ltmle` package in `R` (Lendle et al., 2017).

In the null setting (Setting 1), we find that the type one error of all norm-based tests is somewhat above the nominal level, and larger for small sample sizes and high dimensions. The Bonferroni test is slightly conservative and the Cauchy test is slightly anti-conservative in all settings. In Setting 2, all tests have similar power with the adaptive tests slightly outperforming others at lower sample sizes and the Bonferroni test under-performing in all settings. The differences in performance are larger when the dimension is higher. In Setting 3, the $\ell_\infty$ based test almost always outperforms


Figure 1.7: Empirical rejection rates for the Bonferroni test, the Cauchy test, the non-adaptive $\ell_p$, and $\ell_\infty$ tests, and the adaptive $\ell_p$ and $\ell_\infty$ tests in Example 3 across Settings 1–4, different sample sizes and covariate vector dimensions.

all other tests, with the largest differences in the sample size 500 setting. In Setting 4, the $\ell_\infty$ and Bonferroni test under-perform all other tests, of which all have nearly identical power.

## 1.6 Assessing correlates of risk of HIV infection in HVTN 505

Between 2008 and 2013, a cohort of 2,504 circumcised men and transgender persons who have sex with men were recruited in the United States to participate in HVTN 505, a phase IIB preventative efficacy trial of a DNA and recombinant adenovirus serotype 5 HIV vaccine (Neidich et al., 2019). While the vaccine under study was not found to be efficacious in preventing HIV infection, secondary analyses were conducted to study the association between the immune response to the vaccine and the risk of infection. This response was measured using a large number of biomarkers, including levels of various antibodies, T cells, and Fc-gamma receptors. These analyses indicated the possibility of a qualitative interaction, whereby the vaccine may lower or raise the rate of HIV-1 acquisition for different subgroups, depending on the immune response (Fong et al., 2018; Gilbert et al., 2020). Estimates of how well each biomarker group can predict future HIV-1 infection are suggestive of which groups protect against HIV-1.

In our analysis, we consider the same groupings of biomarkers as in Neidich et al. (2019). For

| Biomarker Group | $\ell_2$ | $\ell_\infty$ | adaptive $\ell_p$ | adaptive ssq |
|---|---|---|---|---|
| IgG + IgA | 0.127 | 0.149 | 0.147 | 0.153 |
| IgG3 (Immuno Globulin G3 Group) | 0.000 | 0.003 | 0.000 | 0.000 |
| T Cells | 0.000 | 0.000 | 0.000 | 0.000 |
| Fx Ab | 0.062 | 0.116 | 0.052 | 0.049 |
| IgG + IgA + IgG3 | 0.002 | 0.006 | 0.002 | 0.002 |
| IgG + IgA + T Cells | 0.003 | 0.000 | 0.000 | 0.001 |
| IgG + IgA + IgG3 + T Cells | 0.000 | 0.000 | 0.000 | 0.000 |
| IgG + IgA + IgG3 + Fx Ab | 0.004 | 0.004 | 0.002 | 0.002 |
| T Cells + Fx Ab | 0.000 | 0.000 | 0.000 | 0.000 |
| All markers | 0.000 | 0.001 | 0.000 | 0.000 |

Table 1.1: $p$-values for each combination of biomarker group and test type. This analysis is based on data from the HVTN 505 clinical trial, and the null hypothesis tested is that the biomarkers from the given group are not associated with the risk of HIV infection.

each set of biomarkers, we test the null hypothesis that no biomarker is associated with the risk of infection using four tests derived from our framework. Two of these tests are adaptive (selecting across $\ell_p$ and sum-of-squares norms, respectively), whereas the other two are non-adaptive (based on the $\ell_2$ and $\ell_\infty$ norms, respectively). The association parameter we focus on is the biomarker-specific regression coefficient from a weighted univariable working logistic regression model. Weighting accounts for the informative biomarker missingness induced by the two-phase study design. Additional details on the HVTN 505 trial and our analysis strategy are provided in the Appendix.

The results of these tests are summarised in Table 1.1. Each column (except the first) corresponds to a test type and each row to a group of biomarkers considered by Neidich et al. (2019). With the exception of the Fx Ab and IgG+IgA groups, each test of association between a biomarker group and risk of infection has a $p$-value less than 0.01 for all considered tests. For the IgG+IgA group, the tests yield $p$-values that are all similar, though the $\ell_2$ test gives a slightly smaller $p$-value. The tests for the functional antibody (Fx Ab) biomarker group give similar $p$-values to one another except for the $\ell_\infty$ test, which yields a $p$-value roughly twice as large as the others. Thus, in all but one setting, the choice of testing procedure has little impact on results. For the test of the Fx Ab biomarker group, the adaptive tests provide similar $p$-values, whereas $p$-values for the non-adaptive tests differ more.

In Figure 1.8, we focus on the testing results for the Fx Ab group. The gray histogram in each panel shows an approximation of the estimated null limiting distribution of $Z_n$ for each considered test. The dashed red and solid black vertical lines intersect the $x$-axis at $Z_n^*$ and the $5^{th}$ percentile of the estimated limiting distribution, respectively. Both adaptive tests have distributions that are centered and more concentrated around a smaller value. Because the adaptive tests select the pointwise minimum among all norms considered, this phenomenon is expected. Figure A.3 of the Appendix shows this summary for every biomarker group from Table 1.1.

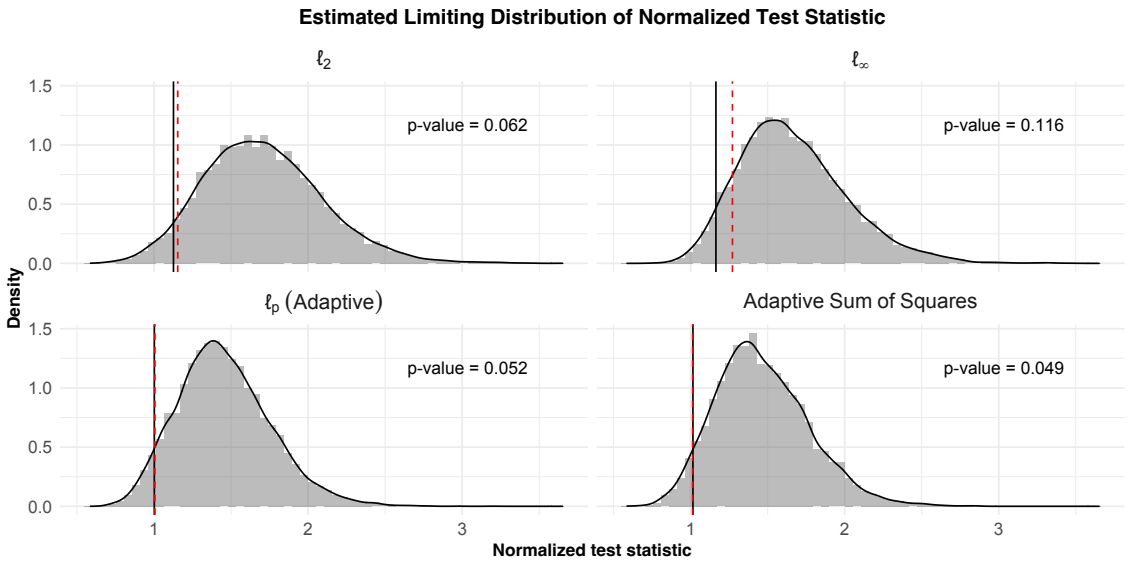


Figure 1.8: Estimated limiting distributions of the multiplicative factor measure for both non-adaptive ($\ell_2$ and maximum absolute deviation) and adaptive (adaptive $\ell_p$ and adaptive sum-of-squares) testing procedures. The black vertical line in each plot represents the 0.05 quantile of the limiting distribution, and the dashed red vertical line represents the value of the test statistic. This analysis is based on data from the HVTN 505 clinical trial, and the null hypothesis tested is that the biomarkers from the Fx Ab group are not associated with the risk of HIV infection.

## 1.7 Concluding remarks

We have described a general framework for constructing tests of a multivariate point null hypothesis in settings in which an asymptotically linear estimator of the underlying target parameter is available. Tests created using this framework leverage knowledge of the parameter estimator and its estimated sampling distribution to adaptively build a test statistic that provides good power under alternatives suggested by the data. Tests constructed using our framework have desirable asymptotic guarantees under the null, fixed alternatives, and local alternatives. We studied the performance of tests constructed using our framework in simulation studies and find these tests have comparable performance to tailor-made methods in settings in which specialized methods currently exist and have favorable properties in settings in which they do not.

The framework we described is quite general, allowing users to specify the parameter of interest and to utilize as much or as little information of the data-generating mechanism as is known. However, it does require an estimator of the covariance matrix of the parameter estimator. For most common parameters, such (non-parametric) estimators already exist and in novel settings constructing these estimators can be facilitated using influence functions. Such analytic derivations could pose a challenge for the implementation of this method in novel settings, though work has been done to allow for such computations to be carried out numerically, which could remove this hurdle (Carone et al., 2019).

Finally, while we have focused on point null hypotheses in this paper, our proposed framework can also be used to test certain composite null hypotheses. For example, the composite null hypothesis $H_0 : \psi_{10} = \psi_{20} = \ldots = \psi_{d0}$ can be equivalently stated as $H_0 : \psi^*_{10} = \psi^*_{20} = \ldots = \psi^*_{(d-1)0} = 0$, where we define $\Psi^*_j(P) := \Psi_j(P) - \Psi_d(P)$ and write $\psi^*_{j0} := \Psi^*_j(P_0)$. Indeed, whenever the composite null hypothesis of interest can be restated as a point null hypothesis (of lower dimension) based on a different parametrization, the methods we have proposed can be used directly.

The code used to run all examples and the data analysis is maintained on GitHub: `https://github.com/adam-s-elder/amvpnt_code`. This code requires the amp package `https://cran.r-project.org/package=amp`.

## Acknowledgements

The authors would like to thank Brian Williamson for his generous help in providing data and guidance on the data analysis. This work was supported by NIH grants DP2-LM013340, R01-HL137808, R01-AI029168 and UM1-AI068635. The opinions expressed in this article are those of the authors and do not necessarily represent the official views of the NIH.

# Chapter 2

# A general adaptive framework for testing a functional null hypothesis

## 2.1 Introduction

It is often of interest to provide inference for function-valued parameters. Because these parameters are infinite-dimensional, and in some cases, non-pathwise differentiable, estimation and inference for these parameters can be more difficult than in the finite-dimensional setting. Currently, multiple methods exist for addressing these issues, though none is a panacea.

A simple way to construct a functional null test is to assume that the data-generating mechanism falls within some restricted space of functions (for example, a linear model). Such techniques are easy to apply and can provide asymptotic guarantees for standard estimators, such as those derived from maximum likelihood, when the model is correctly specified. Even issues arising from model misspecification can be addressed in some settings by instead specifying a working model whose parameter is still well-defined in settings when the working model does not hold (Buja et al., 2019a,b; Whitney et al., 2019). However, even in settings where correct inference is achieved for the parameter itself, a functional null tests based on such estimators may still be inconsistent. This can happen whenever the working model parameter has a preimage that contains elements both inside and outside of the functional null model space. As an example, consider testing the null hypothesis that a function is constant across its support. If a linear working model for the function-valued parameter is used when the true function is a parabola, it is possible the slope parameter from this working model is zero. As a result, a test based on this working model could be inconsistent.

To avoid these issues, non-parametric or semi-parametric models for the function valued parameter may be employed. When the parameter is pathwise differentiable then constructing a

consistent test is possible using a multitude of methods, e.g., those described in Westling and Carone (2020) and Hudson et al. (2021). While the two cited articles utilize different test statistics, both rely on a univariate summary of the original function-valued parameter that is indexed by a norm. As was seen in the previous chapter, the selection of the norm used can influence the performance of the test and suggests that an adaptive selection of the norm could result in the favorable finite-sample performance of the test.

While the previous chapter introduces an adaptive test, this was done in the context of a finite-dimensional vector of parameters. By defining a vector of parameters evaluated at various points on the function's support, it is possible to construct an adaptive test of a functional null using the framework outlined in the first chapter. However, the selection of the number and location of the values at which to evaluate the function introduces multiple complications. Additionally, to attain consistency for all alternatives, it would likely be necessary for the number of locations at which the function is evaluated to grow with sample size. Providing theoretical results for such a test could also prove challenging, especially since the sequence of vectors for such a test has a covariance matrix that is changing dimension and approaching degeneracy as the number of grid points grows.

Fourier transforms provide an alternative means to derive a low-dimensional approximation of the function of interest. While the Fourier transformation still maps to an infinite-dimensional space, lower dimensional approximations are easy to implement and frequently used. In practice, such approximations consider some number of the earliest elements of the transformation. This approximation has been shown to be effective in many areas of statistics including time-series analysis (Bloomfield, 2010), spatial statistics (Subba Rao, 2018) and Gaussian Process models (Hensman et al., 2018). There is also theoretical support for using such finite-dimensional approximations. Indeed for any square-integrable function, the Fourier series approximation of that function will converge to the function with respect to the $\ell_2$ norm (Zygmund, 2002), suggesting that a (large enough) finite-dimensional Fourier approximation will contain most of the information about the given function.

The test described in this chapter uses the Fourier transform to define a test statistic, the form of which builds on work from the previous chapter. We show that if the function estimator satisfies certain regularity conditions, the finite-dimensional test achieves the same desirable properties as the tests described in Chapter 1 in most settings. We also provide arguments that our test approaches that of the (nearly non-parametric) infinite-dimensional test as $d$ grows. We show via simulations that the performance of the test is maintained as the number of coefficients used to define the test statistic grows.

This chapter is organized as follows. In Section 2.2, we introduce the testing problem considered

and provide context for the use of the Fourier transform to help define a test. We formally describe the initial test in Section 2.3, and described an alternative test in Section 2.4. In Section 2.5 we study the performance of the two considered tests in a variety of settings. In Section 2.6 we provide concluding remarks.

## 2.2 Problem setup

In contrast to the first chapter, we now let the parameter $\Psi$ map into a function space. We suppose that $\Psi$ maps from $\mathcal{M}$ to $L^2([0,1])$, the set of all square-integrable functions on $[0,1]$. For each $P \in \mathcal{M}$, denote by $f_P := \Psi(P) : [0,1] \to \mathbb{R}$ the function-valued parameter of interest. In this chapter, we restrict ourselves to parameters $P \mapsto f_P(t), t \in \mathbb{R}$, that are pathwise differentiable with some gradient $D_t(P)$.

For a given (known) function $\psi_* \in L^2([0,1])$, we consider testing

$$H_0 : f_{P_0} = \psi_* \quad \text{versus} \quad H_1 : f_{P_0} \neq \psi_* \ . \tag{2.1}$$

Without loss of generality, we consider the case $\psi_*(t) = 0$ for each $t$ since otherwise we may instead take $\Psi$ to be its null-centered counterpart $P \mapsto f_P - \psi_*$.

As in the previous chapter, we suppose $X_1, X_2, \ldots, X_n$ are observations drawn independently from a common unknown distribution $P_0 \in \mathcal{M}$, with the statistical model $\mathcal{M}$ encoding known restrictions on $P_0$. If the outcome space of $\Psi$ can be indexed by a finite-dimensional vector, then estimation and inference for $\Psi(P_0)$ can be simplified to estimation and inference for the indexing vector. However, here we focus on cases in which $\mathcal{M}$ is a non-parametric or semi-parametric model space. Denote by $\mathcal{X}$ the union of the support of each $P \in \mathcal{M}$. Also, let $\mathcal{M}_0$ denote the set of $P$ for which $\Psi(P) = 0$

Unlike in the finite-dimensional case when element-wise estimation was applicable, the functional null test requires consideration of an entire function. Both estimation of and inference of such quantities can be more difficult than for finite-dimensional quantities. A common method to construct a test in this setting is to instead consider a finite collection of summaries of the function. Commonly, these summaries $S$ map from the infinite-dimensional parameter space into $\mathbb{R}^d$. Such summaries are often selected so that, if $P \in \mathcal{M}_0$, it follows that $S(\Psi(P)) = 0$ and, for $P \notin \mathcal{M}_0$, it holds that $S(\Psi(P)) \neq 0$.

A straightforward version of this approach takes $S$ to be function evaluations at a set of points in the support of $f_P$. This method would simplify the functional null test to a multivariate point null test, a setting for which there are many existing testing procedures. However, it is possible that, for a fixed set of grid points, such a test would be inconsistent if the true value of the

function at each grid point was zero and the function took a non-zero value at another point. This shortcoming can be addressed by adding more grid points, though such additions also come with a computational cost. Because making the grid finer reduces the risk of false negatives, it is natural to consider a test that increases the fineness of the grid as sample size increases.

Two inconveniences arise when considering such a test. The first is the question of how to choose the number and location of the grid points. These decisions will need to be considered alongside considerations of the function estimation procedure. Second, as the number of grid points increases, the covariance matrix of the vector of function estimator evaluations becomes nearly degenerate.

### 2.2.1 Defining a test statistic using a Fourier transformation

Because of the issues raised for the grid-based approach, we consider an alternative vector-valued parameter to define the test. This alternative is the vector of Fourier coefficients arising from the Fourier transformation. This transformation is well studied and frequently used to reduce the dimensionality of estimation problems in a variety of settings. While this parameter is also infinite-dimensional, it provides some key advantages over the grid-based approach. We start by defining the transformation. Let $\mathcal{F}$ denote the map from $L^2([0,1])$ to $\mathbb{R}^\infty$ defined by:

$$\mathcal{F} : f \mapsto (\mathcal{F}_1(f), \mathcal{F}_2(f), \ldots) \text{ where } \mathcal{F}_l(f) = c_l := \begin{cases} 2\int_0^1 f(t)dt, \text{ if } l = 1, \\ 2\int_0^1 f(t)\cos(\pi l t)dt, \text{ if } l \text{ is even}, \\ 2\int_0^1 f(t)\sin(\pi(l-1)t)dt, \text{ otherwise.} \end{cases}$$

One could consider a test that uses the coefficients from this transformation as a test statistic. While there are an infinite number of coefficients arising from the Fourier transform, finite-dimensional approximations of the vector can be used and have favorable properties compared to the finite-dimensional grid-based method. The most important of these properties is that for smooth functions the earlier Fourier coefficients capture the majority of information about the function and take larger values while the later coefficients capture more fine-grained information and tend to take smaller values. This suggests two potential beneficial properties of a test that uses the Fourier coefficients to define the test statistic. First, a finite-dimensional version of the test could perform nearly as well as an infinite-dimensional version of the test since most of the information about the function estimator is captured by the finite-dimensional test statistic. Second, it suggests that a sequence of finite-dimensional tests with an increasing number of coefficients could have behavior that converges to some desirable limit. This could occur if the later Fourier coefficients become smaller as $d$ grows such that the error induced by not including them shrinks

to zero.

While vectors residing in $\mathbb{R}^d$ were used in the previous chapter and we provided a method for constructing a test using these vectors, it is now of interest to consider vectors in $\mathbb{R}^\infty$. To this end, new notation is introduced to allow for the construction of an infinite-dimensional test statistic that is comparable to finite-dimensional versions of the test statistic. Denote by $v^{d,0} \in \mathbb{R}^\infty$ the vector defined by $v_i^{d,0} = v_i$ for $i \in \{1, 2, \ldots, d\}$ and $v_i = 0$ for $i \in \{d+1, d+2, \ldots\}$. Similarly, for any matrix $A$ in $\mathbb{R}^\infty \otimes \mathbb{R}^\infty$, we denote by $A^{d,0}$ a matrix defined by $A_{ij}^{d,0} = A_{ij}$ for $i, j \leq d$ and zero otherwise. Lastly, we let $v^{-d} = v - v^{d,0}$, which is the error induced by considering the approximate vector rather than the vector itself. Such vectors and matrices are easily compared since each is infinite-dimensional for every $d$. In addition, these vectors also allow for consideration of the finite-dimensional test, so long as a $\Gamma$ can be defined such that $\Gamma(v^{d,0}, \Sigma^{d,0}, \varphi) = \Gamma^d(v^d, \Sigma^d, \varphi)$.

Next, to further motivate the use of the Fourier transform in this setting, two useful results are reviewed. To precisely state these results, we introduce the inverse Fourier transformation and the notion of a finite-dimensional Fourier approximation. The inverse Fourier transform, denoted by $\mathcal{F}^{-1}$ maps from $\mathbb{R}^\infty$ into $L^2([0,1])$, defined by:

$$c \mapsto f_c \text{ where } f_c(t) := c_1/2 + \sum_{i=2,4,6,\ldots} c_i \times \cos(\pi i t) + \sum_{i=3,5,7\ldots} c_i \times \sin(\pi(i-1)t).$$

The $d$-dimensional Fourier approximation of a function $f$ is denoted by $f^d$ and is equal to $\mathcal{F}^{-1}[\mathcal{F}^{d,0}(f)]$. The first result that we review is the Riesz–Fischer theorem which states that, for any function $f \in L^2([0,1])$, the sequence $\{f^d\}_{d=1}^\infty$ will converge to $f$. Specifically,

$$\lim_{d\to\infty} \left( \int_0^1 |f^d(t) - f(t)|^2 dt \right)^{1/2} = 0.$$

The second result considered is the Hausdorff-Young inequality (Zygmund, 2002) which states that, for any function $f$ in $L^1([0,1])$, each Fourier coefficient is well-defined and, for any $p \in (1, 2]$,

$$\left( \sum_{n=0}^\infty |\mathcal{F}(f)_n|^{p'} \right)^{1/p'} \leq \left( \int_0^1 |f(t)|^p dt \right)^{1/p},$$

where $p' \in [2, \infty)$ is the Hölder conjugate of $p$.

Due to the Riesz-Fischer theorem mentioned, we expect that if the function of interest is in $L^2([0,1])$, the approximation error from using a $d-$dimensional Fourier approximation of the function, $f^d$, will shrink to zero in as $d$ increase. That is, $\left( \int_0^1 |f^d(t) - f(t)|^2 dt \right)^{1/2}$ will converge to zero. Next, the Hausdorff-Young inequality implies that, given this convergence, the vector $\mathcal{F}(f - f^d)$ will converge to zero in the $\ell_2$ sense as $d$ becomes large. Lastly, note that, due to the linearity of the Fourier transform, we have that $\mathcal{F}(f - f^d) = \mathcal{F}(f) - \mathcal{F}(f^d)$. Taking this all together

suggests that the finite-dimensional approximation of the vector of Fourier coefficients converges to that of the actual vector with respect to the $\ell_2$ norm. This result suggests that it could be possible to show that, as $d$ increases, the $d$-dimensional test based on $u_n^d := \sqrt{n}\mathcal{F}^{d,0}(f_n)$ will approach some infinite-dimensional test. Because an infinite-dimensional test includes all the Fourier coefficients, it would not run the same risk of inconsistency that exists for the $d$-dimensional test.

For any single function in $L^2([0,1])$, we can expect that using a finite Fourier approximation can allow for an arbitrarily accurate approximation of the function. However, when defining the test statistic using the Fourier transform of the function estimator, it is necessary to account for the randomness of this transformed estimator. To better understand the expected asymptotic behavior of the test statistic, note that we anticipate that the normalized estimator $\sqrt{n}[f_n - f]$ converges in distribution to a Gaussian process $\mathbb{G}_k$ with mean zero and some covariance function $k$. Let $\mathcal{K}$ denote the set of bivariate, exchangeable functions and denote by $\Sigma$ the function mapping from $\mathcal{K}$ to $\mathbb{R}^\infty \otimes \mathbb{R}^\infty$ defined by

$$k \mapsto \begin{bmatrix} \mathcal{F}\mathcal{F}_{1,1}(k) & \mathcal{F}\mathcal{F}_{1,2}(k) & \dots \\ \mathcal{F}\mathcal{F}_{2,1}(k) & \mathcal{F}\mathcal{F}_{2,2}(k) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix},$$

where $k_l(x) = \mathcal{F}_l(k(x,\cdot))$ and $\mathcal{F}\mathcal{F}_{l,i}(k) = \mathcal{F}_i(k_l)$. Due the linearity of the Fourier transform, for any Gaussian process $\mathbb{G}_{\phi,k}$ with mean function $\phi$ and covariance function $k$, it follows that $\mathcal{F}(\mathbb{G}_{\phi,k})$ is a Gaussian process with mean $\mathcal{F}(\phi)$ and covariance matrix $\Sigma(k)$ (Papoulis and Pillai, 2002).

Similar to the vector of Fourier coefficients, if the covariance matrix of the $\mathcal{F}(\mathbb{G}_k)$ is almost surely bounded with respect to the $\ell_2$ norm, we expect the $d$-dimensional approximation of the covariance matrix, $\Sigma^{d,0}$, can approximate $\Sigma$ to an arbitrary level of accuracy (given a large enough $d$). We expect this can be shown using the Riesz–Fischer theorem and the Hausdorff-Young inequality.

Thus, building a test around the parameters $\mathcal{F}(f)_1, \dots, \mathcal{F}(f)_d$ rather than $f(x_1), \dots, f(x_d)$ provides practical benefits, making both computation and derivation of theoretical results easier. The described properties of the Fourier transform should prove useful when defining a test statistic that takes values of the coefficients from the transform. In the finite-dimensional setting, we hope that the initial coefficients contain most of the signal from the function estimator. In the infinite-dimensional setting, for sufficiently smooth functions, the Fourier transform coefficients can be accurately approximated using a finite-dimensional approximation. This fact may be used to show the sequence of tests in which the number of coefficients increases with sample size will approach some stable limit and have favorable properties.

## 2.3 Proposed testing procedure

As described above, using the Fourier transform coefficients as a test statistic provides a simple but powerful summary of the function estimator that can be used to construct a finite-dimensional test of the functional null. In what follows we describe a straightforward method of constructing a test using the Fourier transform and describe the properties of the function estimator that we expect to be sufficient to guarantee favorable test performance. To simplify notation, let

$$\Sigma_n := \mathcal{F}\mathcal{F}(\kappa_n),\ \Sigma := \mathcal{F}\mathcal{F}(\kappa),\ \mathbb{S}_{\Sigma,n} := \mathcal{F}(\mathbb{G}_{k_n}),\ \mathbb{S}_{\Sigma} := \mathcal{F}(\mathbb{G}_k).$$

For a test based on $u_n^d$ and $\Sigma_n^d$ to satisfy the conditions of Theorem 1 (for a valid $\Gamma^d$), it is sufficient to show $u_n^d$ is an asymptotically linear estimator of $\mathcal{F}^d(f_{P_0})$ and $\Sigma_n^d$ is a consistent estimator of $\Sigma^d$. To see that this holds in some settings, consider for now a one-step estimator $f_n$ of $f$ of the form

$$f_n : t \mapsto f_{\widehat{P}}(t) + P_n\Phi_t(\widehat{P}),$$

where $P_n$ is the empirical distribution and $\widehat{P}$ is some estimate of $P$. For now, suppose that for each $t \in [0,1]$, $P \mapsto f_P(t)$ is pathwise differentiable with a gradient $\Phi_t(P)$. It follows that, for each $t$,

$$f_n(t) - f_{P_0}(t) = (P_n - P_0)\Phi_t(P_0) + (P_n - P_0)[\Phi_t(\widehat{P}) - \Phi_t(P_0)] + \mathrm{Rem}_t(\widehat{P}, P_0). \tag{2.2}$$

Suppose also that the following two conditions hold:

$$\sup_{t\in[0,1]} |(P_n - P_0)[\Phi_t(\widehat{P}) - \Phi_t(P_0)]| = o_p(n^{-1/2}),$$

$$\sup_{t\in[0,1]} |\mathrm{Rem}_t(\widehat{P}, P_0)| = o_p(n^{-1/2}).$$

In this case, it is possible to show that the Fourier coefficients of the functions $\epsilon_{1,n} : t \mapsto \sqrt{n}(P_n - P_0)[\Phi_t(\widehat{P}) - \Phi_t(P_0)]$ and $\epsilon_{2,n} : t \mapsto \sqrt{n}\mathrm{Rem}_t(\widehat{P}, P_0)$ shrink to zero as $n$ increases. Note by the Hausdorff-Young inequality that

$$\begin{aligned}\left(\sum_{i=0}^{\infty} |\mathcal{F}(\epsilon_{1,n} + \epsilon_{2,n})_i|^{p'}\right)^{1/p'} &\leq \left(\int_0^1 |(P_n - P_0)[\Phi_t(\widehat{P}) - \Phi_t(P_0)] + \mathrm{Rem}_t(\widehat{P}, P_0)|^p dt\right)^{1/p} \\ &\leq o_p(1/\sqrt{n}).\end{aligned}$$

Next, suppose that $\{z \mapsto \Phi_t(P_0)(z) : t \in [0,1]\}$ is a Donsker class. If each of the above conditions holds, then

$$\begin{aligned}
\sqrt{n}\left[\mathcal{F}\{f_n - f_{P_0}\}\right] &= \sqrt{n}\mathcal{F}\left\{[(P_n - P_0)\Phi(P_0)] + (P_n - P_0)[\Phi(\widehat{P}) - \Phi(P_0)] + \mathrm{Rem}(\widehat{P}, P_0)\right\} \\
\sqrt{n}\left[\mathcal{F}\{f_n\} - \mathcal{F}\{f_{P_0}\}\right] &= \sqrt{n}\mathcal{F}\left\{[(P_n - P_0)\Phi(P_0)]\right\} + \sqrt{n}\mathcal{F}\left\{(P_n - P_0)[\Phi(\widehat{P}) - \Phi(P_0)] + \mathrm{Rem}(\widehat{P}, P_0)\right\} \\
&= \sqrt{n}\mathcal{F}\{[(P_n - P_0)\Phi(P_0)]\} + o_p(1).
\end{aligned}$$

If it is the case that

$$\{\sqrt{n}(P_n - P)\Phi_t(P_0) : t \in [0,1]\} \rightsquigarrow \{\mathbb{G}\Phi_t(P_0) : t \in [0,1]\}$$

in $\ell^\infty([0,1])$, then we expect that by the continuous mapping theorem that

$$\sqrt{n}\left[\mathcal{F}\{f_n - f_{P_0}\}\right] \rightsquigarrow \mathcal{F}\{\mathbb{G}\Phi_t(P_0)\}.$$

Remembering that the Fourier transformation of a Gaussian process is also a Gaussian process (Papoulis and Pillai, 2002), the above finding, would imply that under the null, $\sqrt{n}\left[\mathcal{F}^d\{f_n\} - \mathcal{F}^d\{f_{P_0}\}\right]$ converges in distribution to a multivariate normal distribution with mean zero and covariance matrix $\Sigma^d(\Phi\Phi)$, where $\Phi\Phi$ is the map $(x, y) \mapsto \Phi_x(P_0)\Phi_y(P_0)$ with $x, y \in [0,1]$. Lastly, note that if a consistent estimator $\Phi_n$ of $\Phi(P_0)$ exists, it can be used to define a consistent estimator of $\Sigma^d(\Phi\Phi)$. In conclusion, we expect that $u_n^d$ is an asymptotically linear estimator of $\mathcal{F}^d(f)$ and $\Sigma^d(\Phi_n\Phi_n)$ is a consistent estimator of $\Sigma^d(\Phi\Phi)$. If this holds, it would then follow that a test based on these estimators would satisfy the properties described in Theorem 1.

One caveat of the described test is that it may be inconsistent in the sense that the first $d$ Fourier coefficients may be zero while the true function is non-zero at some point. To be more precise, denote by $\mathcal{M}_0^d$ the set of all probability distributions $P$ for which $\mathcal{F}^d\Psi(P) = 0$. Note that, for each $d$, $\mathcal{M}_0^{d+1} \subseteq \mathcal{M}_0^d$ and $\mathcal{M}_0^\infty$ contains only functions that are almost everywhere equal to 0. If it were the case that $f \in \mathcal{M}_0^d$, but $f \notin \mathcal{M}_0^\infty$ the limiting distribution of $u_n^d$ (and every $u_n^{d^*}$ for $d^* < d$) would have mean zero and as a result, the test based on $u_n^d$ would likely have a low probability of rejecting the null hypothesis, even for large samples.

### 2.3.1 Allowing dimension to grow with sample size

As has been hinted at before, it is natural to consider the validity of the test that considers all Fourier coefficients. In what follows an extension of the testing framework outlined in Chapter 1 will be introduced that uses an infinite-dimensional vector estimator as its test statistic. To show that such a test has desirable properties, findings from Chapter 1 will be built upon. In particular,

we seek to show that if a sequence of $d$-dimensional tests is considered, the behavior of these tests approaches some limit as $d$ and $n$ become large. If this can be shown, it is possible to obtain some of the same desirable properties for the infinite-dimensional version of the test as were found for the finite-dimensional test.

It is difficult to make comparisons between tests that are defined using $u_n^d$ if $d$ is different between the tests. To make such comparisons more straightforward, recall that $u_n^{d,0} = (u_n^d, 0, 0, \ldots)$. For now, assume we may construct a performance measure $\Gamma$ such that for each $d$, we have

$$\Gamma(v^{d,0}, \Sigma^{d,0}, \varphi) = \Gamma^d(v^d, \Sigma^d, \varphi) \text{ for each } v,\ \Sigma,\ \text{and } \varphi.$$

Note that for $\Gamma_{ar}$ and $\Gamma_{md}$, the definitions may be used in the infinite-dimensional setting by replacing the finite-dimensional probability density function with an infinite-dimensional version. Next, to ease notional complexity, let $Q_{n,d}$ denote the random variable $\Gamma^d\left(u_n^d, \Sigma_n^d\right) = \Gamma\left(u_n^{d,0}, \Sigma_n^{d,0}\right)$, $\bar{Q}_{n,d}$ denote the random variable $\Gamma^d\left(\mathbb{S}_{\Sigma_n^d}, \Sigma_n^d\right) = \Gamma\left(\mathbb{S}_{\Sigma,n}^{d,0}, \Sigma_n^{d,0}\right)$, and $Q_d$ denote the random variable $\Gamma^d\left(\mathbb{S}_{\Sigma}^{d}, \Sigma^{d}\right) = \Gamma\left(\mathbb{S}_{\Sigma}^{d,0}, \Sigma^{d,0}\right)$. Notation is also introduced for the infinite-dimensional version of the test. Let $Q_n$ denote the random variable $\Gamma\left(u_n, \Sigma_n\right)$, $\bar{Q}_n$ denote the random variable $\Gamma\left(\mathbb{S}_{\Sigma,n}, \Sigma_n\right)$, and $Q$ denote the random variable $\Gamma\left(\mathbb{S}_{\Sigma}, \Sigma\right)$. Let $F_{n,d}, \bar{F}_{n,d}, F_d, F_n, \bar{F}_n$ and $F$ denote the cumulative distribution functions of $Q_{n,d}, \bar{Q}_{n,d}, Q_d, Q_n, \bar{Q}_n$ and $Q$, respectively, and denote the corresponding quantile functions of each random variable with a "$-1$" superscript. Next, we seek to quantify the relationship between the finite and infinite-dimensional versions of the test. For context, suppose that $f_n$ is some asymptotically linear estimator of $f$, where $\sqrt{n}(f_n - f)$ converges to a Gaussian process $\mathbb{G}$ with covariance function $\kappa$ and let $\kappa_n$ be some consistent estimator of $\kappa$ (in the $\ell_\infty$ sense). Next, suppose that $\Gamma^d$ satisfies C1-C3 from Chapter 1. With these assumptions in place, consider the error introduced by using a finite-dimensional test to approximate the infinite-dimensional test, namely

$$\left|\text{pr}\left\{Q_{n,d} < \bar{F}_{n,d}^{-1}(\alpha)\right\} - \text{pr}\left\{Q_n < \bar{F}_n^{-1}(\alpha)\right\}\right|.$$

Being able to show that this error term becomes small as $n$ and $d$ increase should make it possible to show that the infinite-dimensional version of the described test has the desirable properties outlined in Theorem 1.

To see this, suppose it has been established that the above quantity converges to zero. This convergence can be used to show the infinite-dimensional tests has a rejection rate that tends to $\alpha$ under the null. Let $T_n^d$ denote the random variable that is equal to one if the $d$-dimensional test rejects the null and zero otherwise. Let $T_n^\infty$ denote the random variable that is equal to one if the infinite-dimensional test rejects the null and is zero otherwise. Letting $\varepsilon_{d,n} := |\text{pr}(T_n^\infty =$

$1) - \text{pr}(T_n^d = 1)|$, suppose it has been shown that $\varepsilon_{d,n} \to 0$ as $d$ and $n$ become large. Additionally, for each fixed $d$ it is known that $\text{pr}(T_n^d = 1)$ converges to $\alpha$. Thus, the sum $\text{pr}(T_n^d = 1) + \varepsilon_{d,n}$ can be made arbitrarily close to $\alpha$. Let $\varepsilon$ be given. First select a $D_\varepsilon$ and $N_{\varepsilon,1}$ such that $\varepsilon_{d,n} < \varepsilon/2$ for all $d > D_\varepsilon$ and $n > N_{\varepsilon,1}$ and select an $N_{\varepsilon,2}$ such that $|\text{pr}(T_n^{D_\varepsilon+1} = 1) - \alpha| < \varepsilon/2$ for all $n > N_{\varepsilon,2}$. Thus, for all $n$ sufficiently large, it follows that

$$\begin{aligned}|\text{pr}(T_n^\infty = 1) - \alpha| &= |\text{pr}(T_n^\infty = 1) - \text{pr}(T_n^{D_\varepsilon+1} = 1) + \text{pr}(T_n^{D_\varepsilon+1} = 1) - \alpha| \\ &\leq |\text{pr}(T_n^\infty = 1) - \text{pr}(T_n^{D_\varepsilon+1} = 1)| + |\text{pr}(T_n^{D_\varepsilon+1} = 1) - \alpha| < \varepsilon/2 + \varepsilon/2.\end{aligned}$$

Because the $\varepsilon$ chosen was arbitrary, we may then conclude that $\lim_{n\to\infty} \text{pr}(T_n^\infty = 1) = \alpha$. A key piece of the above logic is that $\varepsilon_{d,n}$ can be bounded uniformly with respect to both $d$ and $n$ at the same time. That is, while $\varepsilon_{D_\varepsilon+1,N_{\varepsilon,1}} < \varepsilon/2$, it must also be the case that $\varepsilon_{D_\varepsilon+1,n} < \varepsilon/2$ for each $n > N_{\varepsilon,1}$. A similar argument could be made that would show that when $P \notin M_0^\infty$, the infinite-dimensional test would approach a rejection rate of one. The only additional argument would be finding a second $D$ such that $M_0^D$ does not contain the true function and choosing $D$ to be at least this large. While such an infinite-dimensional test can never be carried out in practice, these results suggest that the sequence of tests for which dimension increases with sample size will converge to a test with the properties outlined in Theorem 1. Additionally, the infinite-dimensional test would avoid most of the issues of inconsistency that exist for the finite-dimensional test.

## 2.4 An alternative testing procedure

Considering the finite-dimensional test described earlier, note that the vector of Fourier coefficients is non-standardized in the sense that the entries of $u_n^d$ may have standard errors that are quite different from one another. In previous settings, the parameters being estimated were often themselves standardized. As an example, the correlation between the outcome of interest and each covariate was considered in Chapter 1, rather than the covariance. However, this lack of standardization could cause more problems in this setting when the parameters are the coefficients of a Fourier transform. While many counter-examples exit, it is generally expected that for smooth functions the Fourier coefficients tend to rapidly shrink towards zero as the index of the coefficient increases. In other words, it is expected that the latter entries of the vector of Fourier coefficient are much smaller than the earlier entries of the vector. Even for non-smooth functions, this behavior can be observed. As an example, the Fourier representation of the Brownian bridge was studied

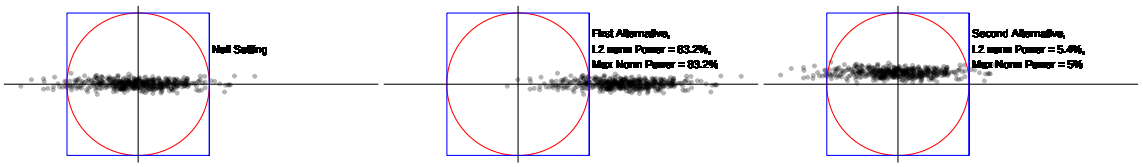


Figure 2.1: Empirical estimates of the performance of a simplified test under the null and two alternatives

by Loève (1978) and has been shown to be given by:

$$\sum_{i=1}^{\infty} \frac{\sqrt{2}W_i}{\pi i} \sin(\pi i t) \text{ for each } t \in (0,1)$$

where each $W_j$ is an independent standard normal random variable. The above representation indicates that, for the non-smooth Brownian bridge, the $j^{\text{th}}$ Fourier coefficients will have a standard error $1/j$ the size of the first coefficient.

These differences in coefficient size could potentially lead to poor test performance for settings in which the latter Fourier coefficients are non-zero. The reason for this can be seen by studying how the rejection region for a simple test is constructed. Consider a test that uses $\ell_2(u_n^d)$ as its test statistic. As described in Chapter 1, this test may use a circular acceptance region, where the radius of the circle, $c_\alpha$, is selected such that $\text{pr}(\ell_2(u_n^d) < c_\alpha)$ converges to $1-\alpha$ when the null holds. Figure 2.1 shows an example of such an estimated acceptance region and a region based on the max norm. In this example, the vector of parameter estimators resides in $\mathbb{R}^2$ and the first element of the vector has a far larger estimated variance than the second. As a result, the size of the acceptance region for both tests is determined nearly entirely by the estimated variance of the first element of the vector. Next, consider two separate alternatives, in which $u_n^2$ has the same covariance as in the null setting. For the first alternative, $u_n^2$ has a mean of $(3 \times \sqrt{\text{var}(u_{n,1})}, 0)$ and for the second alternative $u_n^2$ has a mean of $(0, 3 \times \sqrt{\text{var}(u_{n,2})})$, as is shown in the second and third panels of Figure 2.1, respectively. Because both alternatives are three standard deviations away from zero, both alternatives could be considered equally far away from the null. However, because of the symmetry of the acceptance region, the described tests have vastly different rejection rates for the two alternatives. While tests based on $\Gamma_{ar}$ and $\Gamma_{md}$ construct their acceptance regions differently than the simple tests described above, these performance measures are defined in part by the simple tests that were just described. Thus, the rejection regions of tests defined using these measures are still impacted by the lack of standardization.

One method to mitigate the trend observed above is to standardize the test statistic. To describe standardization, we first recall the definition of a matrix square root. For a matrix, $\Sigma$, we define the square root of $\Sigma$ as the matrix $A$ such that $A^t A = \Sigma$. The inverse square root

of the matrix $\Sigma$ is defined as $A^{-1}$ and is denoted by $\Sigma^{-1/2}$. The standardized test statistic is the original test statistic, but pre-multiplied by $\Sigma^{-1/2}$. This multiplication will result in the test statistics having a limiting distribution with an identity covariance matrix. Such a test statistic is frequently used for defining tests, including the foundational Wald test. If tests using such a standardization were used in the previous example, similar rejection rates for both alternatives would have been observed.

Standardization provides a simple strategy to ensure that each element of the estimator vector takes values relative to the estimated standard error of the estimator. While this can be beneficial in some settings, it may not always improve the performance of the test. In particular, when the Fourier transform defines the vector of parameters, standardization makes the test sensitive to alternatives in which the latter coefficients are non-zero. However, when these latter coefficients are equal to (or close to) zero, each standardized coefficient can add sizable amounts of noise to the test statistic.

Additionally, the technique used to show that the non-standardized test would approach some limit as the dimension of the test grows will not be usable for the standardized test. To show the limit of the finite-dimensional non-standardized tests converged to that of the infinite-dimensional test, it was shown that the size of the latter entries of the vector of Fourier coefficients would shrink toward zero. Intuitively, for these tests, the information in the part of the vector not included in the finite-dimensional approximation shrinks towards zero as the dimension grows. In contrast, each entry of the standardized test is constructed to have an equivalent amount of information. Thus ignoring the latter terms of the standardized vector could introduce large errors, even as dimension increases.

For now, we introduce the finite-dimensional version of the standardized test described above. Rather than using the estimator $u_n^d$ as the test statistic and $\Sigma_n^d$ as the covariance estimator, the standardized testing procedure uses $\left(\Sigma_n^d\right)^{-1/2} u_n^d$ as the estimator and $I_d$ as the covariance estimator where $I_d$ is the $d$ by $d$ identity matrix. By the continuous mapping theorem and the assumption that matrix inverses and matrix square roots are continuous functions near $(\Sigma^d)^{-1}$ and $\Sigma^d$, respectively, we expect that $\left(\Sigma_n^d\right)^{-1/2}$ is a consistent estimator of $\left(\Sigma^d\right)^{-1/2}$ (Bhatia, 2013). This and another use of the continuous mapping theorem implies that $\left(\Sigma_n^d\right)^{-1/2} u_n^d$ will converge in distribution to a multivariate normal distribution with mean zero and covariance matrix $I_d$. If this is the case, then $\left(\Sigma_n^d\right)^{-1/2} u_n^d$ and $I_d$ satisfy the requirements of the parameter and covariance estimators outlined in Theorem 1, respectively.

## 2.5 Numerical examples

The performance of our testing procedures is now studied in a variety of settings. In each setting, the baseline covariate $W$ is drawn from a uniform $[0, 1]$ distribution. Conditional on $W = w$, treatment level is defined by $A := \varepsilon_a + w/2$, where $\varepsilon_a$ is a uniform $[0, 0.5]$ random variable independent of $W$. Conditional on $A = a$, $Y = f(a) + \varepsilon_y$, where, independently of $A$ and $W$, $\varepsilon_y$ is drawn from a normal distribution with mean zero and standard deviation of 0.4. The mean function $f$ is one of six functions:

- Null setting: $f(x) = 0$;
- Linear setting: $f(x) = 0.25x$;
- Parabolic setting: $f(x) = 1.5[0.25 - (x - 0.5)^2]$;
- Discontinuous setting: $f(x) = \begin{cases} 4 - 1.5x \text{ if } 0 \leq x \leq 1/3 \\ 1 + 2x \text{ if } 0 < x \leq 2/3 \\ 3.7\sin(2\pi x) + 1.1 \text{ if } 0 < x \leq 1; \end{cases}$
- Multiple Fourier setting: $\mathcal{F}^{-1}(c^*)$, where $c_5^{*,1} = -0.4, c_6^{*,1} = 0.5, c_7^{*,1} = -0.5, c_8^{*,1} = 0.4$ and each other $c_j^{*,1} = 0$; or
- Single Fourier setting: $f = \mathcal{F}^{-1}(c^{*,2})$, where $c_4^{*,2} = 0.8$ and each other $c_j^{*,2} = 0$.

For any function $f$ with support on $[0, 1]$, we define the standardized primitive of $f$ as the function $t \mapsto \int_0^t f(u)du - t\left[\int_0^1 f(u)du\right]$. For each test considered, the parameter of interest is the standardized primitive of the conditional mean function. Figure 2.2 shows each of the functions described above, along with their corresponding standardized primitive functions. In the bottom row of the figure, the first 11 Fourier transform coefficients of each standardized primitive function are shown. As was done in Chapter 1, we consider both adaptive and non-adaptive versions of the testing procedure and compare them to the method described by (Westling, 2020), which is referred to here as the Westling test. This method requires users to specify the norm used to define the test. For simplicity, only the performance of the Westling test based on the Euclidean distance is presented here. We consider the non-adaptive $\ell_2$, and $\ell_\infty$ versions of our test as well as two adaptive versions of our test. The first adaptive test is the $\ell_p$ test and selects over the $\ell_1$, $\ell_2$, $\ell_4$, $\ell_6$, and max norms. The second adaptive test selects over various versions of the $\jmath_k$ norm — specifically, over $k \in \{1, 2, 3\}$ when $d = 3$, $k \in \{1, 2, 3, 4, 5\}$ when $d = 5$, and $k \in \{1, 4, 6, 8, 11\}$ when $d = 11$, $k \in \{1, 6, 11, 16, 21\}$ when $d = 21$— and is referred to as the sum-of-squares test. We note that $\jmath_1 = \ell_\infty$ and $\jmath_d = \ell_2$.

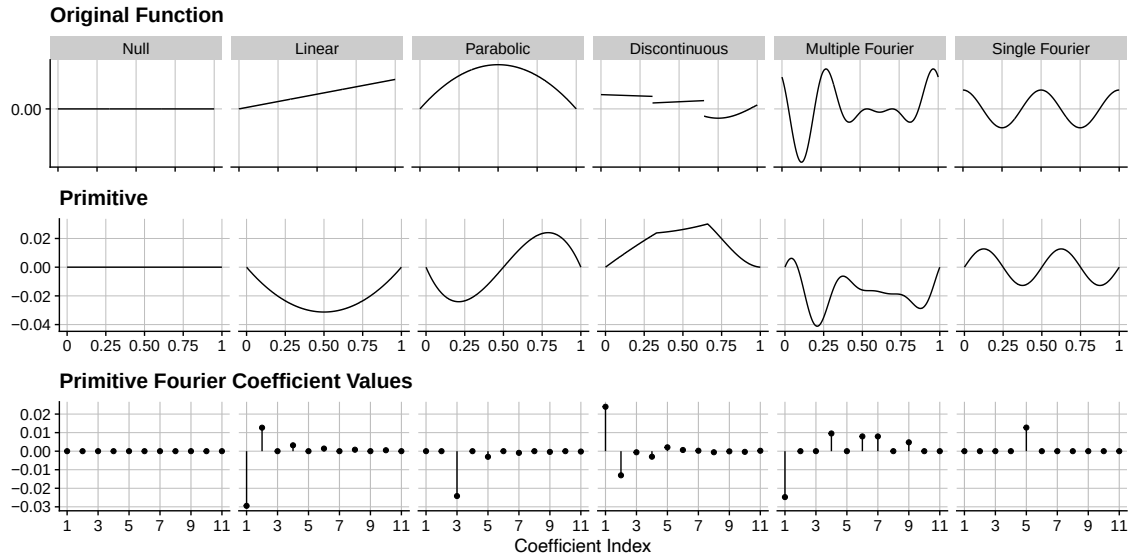


Figure 2.2: The six considered simulation settings.

In each example, we consider all combinations of sample size $n \in \{200, 400\}$ and number of basis functions $d \in \{3, 5, 11, 21\}$. The multiplicative factor measure (1.4) is used throughout. In this example, all tests (including that of Westling, 2020) use the same function estimator and corresponding covariance estimator. In each setting these estimators are calculated using the *ctsCausl* R package (Westling, 2020). The Fourier coefficients of the standardized primitive function estimator and the corresponding IC estimator are approximated using numerical integration.

In Figures 2.3, 2.4, and 2.5, the null hypothesis holds in the first (far left) setting but not in any of the remaining settings. In each panel of each figure, the horizontal dotted red line represents the 0.05 $\alpha$ level of all tests. The thicker horizontal blue line shows the rejection rate of the Westling test. Each of the remaining tests is defined by the number of Fourier coefficients used to define the test statistic (shown along the x-axis,) with points indicating the approximate rejection rate of the test and the line going through each point indicating the exact 95% confidence interval of the approximation.

### 2.5.1 Non-Standardized test

We start by discussing the performance of the finite-dimensional, non-standardized test, with results summarized in Figure 2.3. In the null setting, we find that the type one error of all tests is above the 0.05 type one error rate. In the linear setting for both sample sizes, all tests have similar power across both the number of Fourier coefficients and across the different testing procedures. In the parabolic setting, the adaptive $\ell_p$- and $\ell_\infty$-based tests have lower rejections rates than the adaptive sum of squares and $\ell_2$-based tests, with little to no difference in the performance of the tests across the number of coefficients used. The Westling test rejects the null at a rate slightly

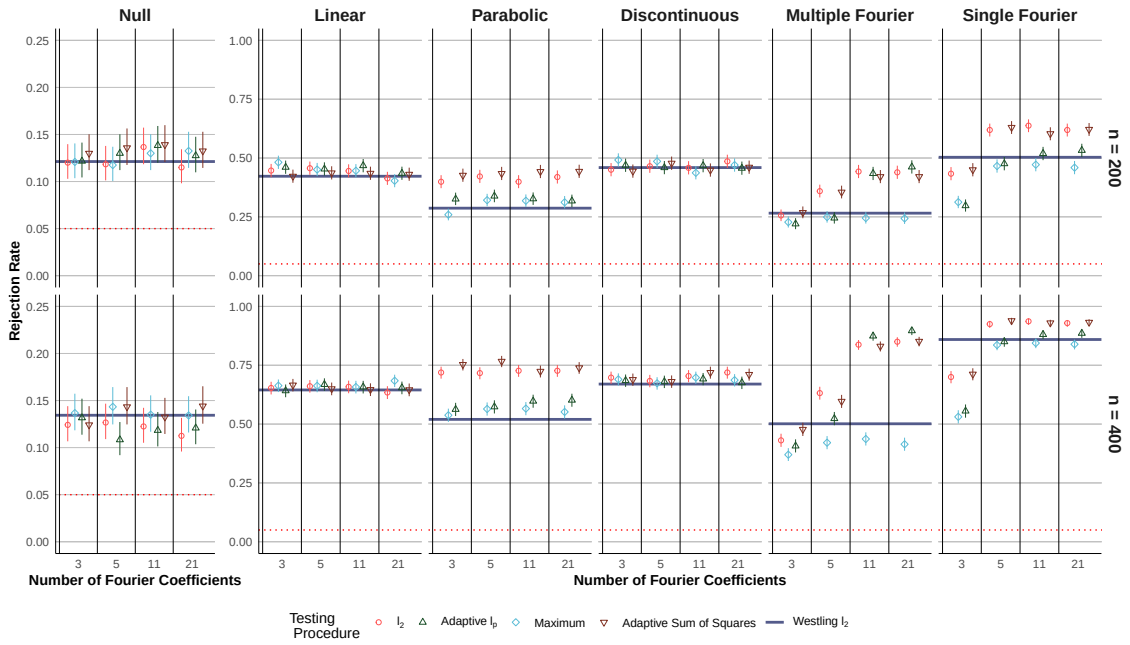


Figure 2.3: Empirical rejection rate of various tests under different data-generating mechanisms, at different sample sizes, and for different numbers of Fourier coefficients used to define the test statistic. The test statistic used to define the test is not standardized in any of the settings shown.

lower than the max norm-based tests. In the discontinuous setting, each test has a comparable performance for each combination of sample size and number of Fourier coefficients. In the multiple Fourier setting, the rejection rate for each norm-based test is larger for tests that use more Fourier coefficients. The exception to this is the max norm-based test, which maintains a low power across all settings compared to the other tests. The Westling test has a slightly higher rejection rate than each of the other tests when only 3 Fourier coefficients are considered, but has a lower rejection rate than most tests that are defined using 11 or 21 Fourier coefficients. These differences are more pronounced in the sample size 400 setting. In the single Fourier setting, while the differences in test performance are less stark than they were in the multiple Fourier setting, the higher dimensional versions of the tests again have higher rejection rates and the maximum norm has the lowest rejection rate of the tests considered.

Considering Figure 2.2, in the parabolic and single Fourier settings, the fact that there is one Fourier coefficient much larger than the others gives reason to think that the max norm would perform well. Thus, the relatively low power observed in these settings is unexpected. Two possible explanations for this observation are the finite sample bias of the function estimator and the lack of standardization of the Fourier coefficient estimator. Both of these phenomena may be observed in Figure B.1 in the appendix of this chapter. In this Figure, for each mean function and sample size considered, a violin plot is shown of the sampling distribution of each of the first six Fourier coefficients, with the black horizontal line showing the median of each sampling distribution. The

blue point shows the true value of each Fourier coefficient in each setting. Note first that in the two settings mentioned, the large Fourier coefficients appear late in the vector, and the corresponding estimators are biased towards zero. Also note that in the single Fourier setting, the estimator for the third coefficient has a non-zero mean whereas the true value is zero. Because Figure B.1 only shows the sampling distribution of the vector of Fourier coefficients rather than the estimated covariance matrix of the Fourier coefficient estimator which determines the rejection region of the test, the construction of the rejection region cannot be fully understood from the figure alone. However, it is still worth noting the large variance of the earlier coefficients. If the estimated variance of these coefficients is also large, this may lead to a stringent cutoff value for the test. As discussed earlier, because the norm-based acceptance regions are symmetric, smaller coefficient estimates may not be large enough to reject the null. This can be true even if, when considering the standard error of the estimator, the smaller coefficient estimate would be quite unlikely to be observed under the null.

In the first chapter, the choice between the adaptive $\ell_p$ and the adaptive sum of squares norm appeared to have little effect on the performance of the test. However, in the parabolic setting, the choice between the two tests appears to be somewhat important. Because the sum of square norm considers only a subset of the vector entries, it could potentially reduce the amount of variance of the test statistic, and as a result, the corresponding test may be able to detect smaller deviations from the null. In contrast, the $\ell_p$ norm-based test will consider all vector values and place a non-trivial amount of weight on each coefficient, potentially resulting in a more stringent cutoff value for the test.

Throughout the settings, using more Fourier coefficients to define the test statistic resulted in either no change in the rejection rate or an increase in the rejection rate. This observation is consistent with the earlier note that the later Fourier coefficients of the function estimator can be expected to have low variance. Thus, including such coefficients could potentially increase power in settings in which the added coefficient is non-zero, but is unlikely to substantially increase the size of the acceptance region. Depending on the setting, the choice of norm can be consequential. From the considered settings, the $\ell_2$ based test or adaptive sum of squares tests had either the highest or nearly the highest power and all tests had a similar type one error rate to all other tests.

### 2.5.2 Standardized test

Next, the performance of the standardized testing procedure is studied across the same settings considered for the non-standardized test. The results of these simulations are summarized in Figure 2.4. In the null setting, the type one error of all tests is above the 0.05 type one error rate when 3 coefficients are considered. Type one error becomes smaller as the number of coefficients used

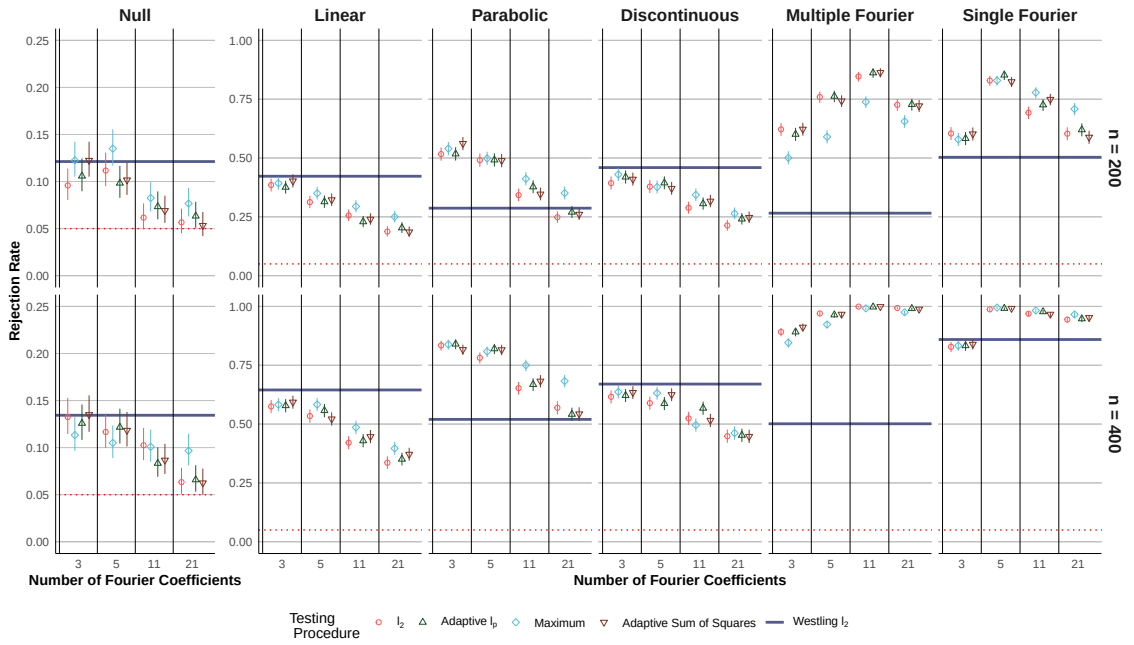


Figure 2.4: Empirical rejection rates of various tests under different data-generating mechanisms for different sample sizes and different numbers of Fourier coefficients used to define the test. For all but the Westling test, the vector of Fourier coefficients used to define the test is standardized in each of the settings shown.

to define the test statistic increases, and is only slightly above the 0.05 rate when 21 coefficients are considered. In nearly all settings, the rejection rate of each test besides the Westling test is similar. In the linear setting for both sample sizes, the Westling test has the highest rejection rate, and the other tests have similar power that decreases as the number of Fourier coefficients used to define the test increases. In the parabolic setting, the Westling test has the lowest rejection rate, and again, the power of the other tests diminishes as the number of coefficients used increases. In the discontinuous setting, the powers of all tests are roughly the same when 3 Fourier coefficients are used for the test statistic. Once again, the rejection rate of the test decreases as the number of Fourier coefficients used to define the test statistic increases. In the multiple Fourier setting, the Westling test has a rejection rate that is much lower than all other tests. The rejection rate of the other tests is higher when 11 Fourier coefficients are used to define the test statistic. The maximum norm-based test has a slightly lower rejection rate across the number of Fourier coefficients, though these differences become smaller as the number of coefficients grows. In the single Fourier setting, the Westling test has a lower rejection rate than the other tests in all but one setting. In this setting, the rejection rate of the other tests is highest when five Fourier coefficients are used.

In contrast to the non-standardized test, the performance of the standardized test was strongly influenced by the number of Fourier coefficients used. Additionally, there was no single number of coefficients that performed the best across all simulation settings. For settings where the earlier Fourier coefficients of the true mean function were much larger than the later coefficients (linear,

parabolic, and discontinuous), only considering the first few coefficients led to better performance. In these settings, the rejection rate of the test decreased rapidly with the number of coefficients. This trend can be better understood by considering the sampling distribution of each entry of the standardized vector of Fourier coefficient estimators, as is summarized in Figures B.2 and B.3.

In contrast to the sampling distribution of the Fourier coefficient estimators from the non-standardized test statistic (see Figure B.1), the variance of the standardized Fourier coefficient estimator sampling distribution is similar across the different coefficients. As a result including additional coefficients that contain only noise can be more detrimental in the standardized setting because each additional coefficient adds a meaningfully large amount of noise. Additionally, while the estimated covariance matrix used to standardize the test statistic does not converge to a degenerate matrix for any fixed dimension, error from estimating this matrix could also inflate the variance of the standardized estimator. As discussed earlier, the variances of the later Fourier coefficients are quite small and, as a result, a relatively large amount of error could be introduced when estimating the $\widehat{\Sigma}^{-1/2}$ that defines the standardized vector of Fourier coefficients. Lastly, note that the standardized estimator is expected to have a limiting distribution with an identity covariance matrix. In Chapter 1, it was observed in settings in which the entries of the vector of parameter estimators were independent from one another that the power of the test diminished more rapidly as dimension grew compared to settings in which the entries of the vector were highly correlated (see Figure 1.4). Thus, attempting to enforce independence of the entries of the test statistic may result in the test becoming sensitive to the number of coefficients used to define it.

However, the addition of Fourier coefficients does not decrease power in all alternative settings. As observed in the single Fourier and multiple Fourier settings, it is possible to include too few coefficients and have a lower rejection rate. It is worth noting that, even in these settings, it is still possible to include too many coefficients. While performance can improve from including the signal of a non-zero coefficient, there can also be a loss of performance from including too many coefficients that only add noise to the vector of Fourier coefficient estimators.

Overall, the norm used to define the standardized test appears to have little, if any, effect on the performance of the test. However, the number of Fourier coefficients does play an important role and appears to be related to the relative size of the Fourier coefficients of the conditional mean function that defines the data generating mechanism. Unfortunately, because this function is usually not known a priori, it could be difficult to provide guidance on how many coefficients to use when applying the test in practice.

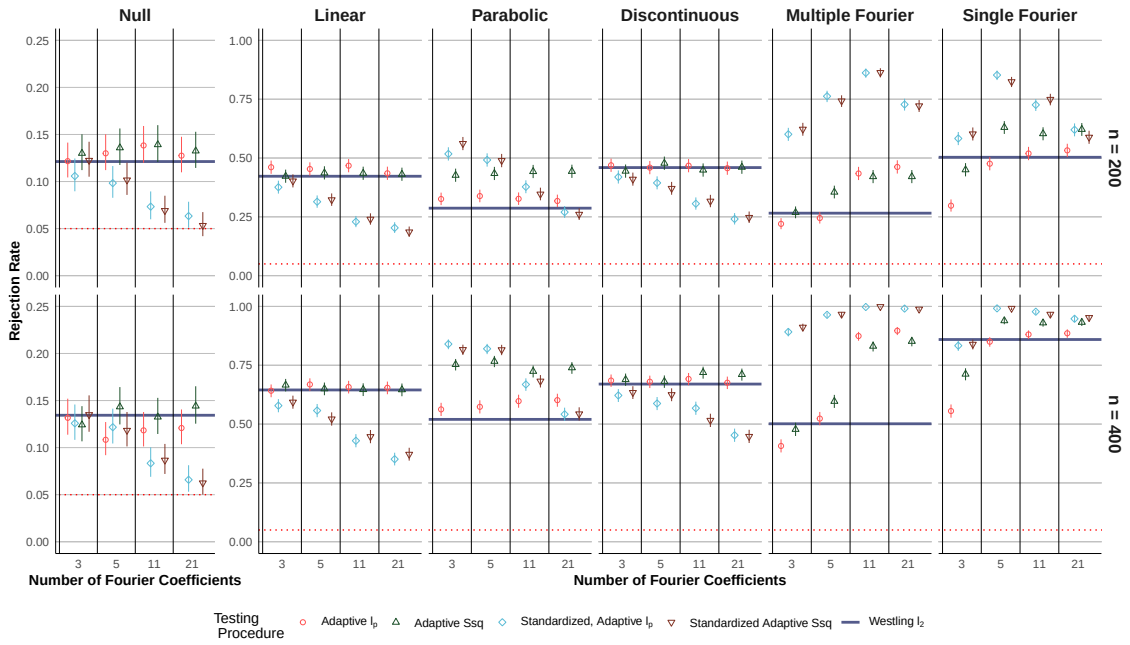


Figure 2.5: Empirical rejection rate of various tests under different data-generating mechanisms, at different sample sizes, and for different numbers of Fourier coefficients used to define the test statistic.

### 2.5.3 Comparing the standardized and non-standardized tests

We conclude by comparing the performance of the standardized and non-standardized tests, with results summarized in Figure 2.5. For simplicity, here we only make comparisons between the adaptive versions of the standardized and non-standardized tests. Tests that use the standardized vector of Fourier coefficients to define the test are referred to as standardized tests and tests that use the non-standardized vector of Fourier coefficients are referred to as non-standardized tests.

In the null setting, all testing methods have inflated type one error. However, the standardized test tends to have a lower type one error. As the number of coefficients used to define the test increases, the type-one error decreases for the standardized tests, but not for the non-standardized tests. In the linear setting, all the tests have similar performance when the first 3 Fourier coefficients define the test statistic. However, the rejection rate of the standardized test diminishes as the number of coefficients used increases and is lower in these settings compared to the non-standardized test which maintains a similar rejection rate across the different number of coefficients. In the parabolic setting, the standardized test has a higher rejection rate relative to the non-standardized test when fewer coefficients are considered and a lower rejection rate when more coefficients are considered. In the discontinuous setting, the performance of the tests is similar when fewer coefficients are used, but differences appear as the dimension of the test statistic increases. In the multiple Fourier setting, the standardized tests have a higher power across all settings compared to the non-standardized tests, though in the $n = 400$ setting these differences shrink as the number

of coefficients considered increases. In the single Fourier setting, the standardized tests tend to outperform the non-standardized tests, though the differences in performance are smaller than they were in the multiple Fourier setting and rejection rates are nearly identical when considering 21 coefficients.

The two testing schemes, standardized and non-standardized, each had settings in which they outperformed the other. For settings in which the first few Fourier coefficients of the mean function are large relative to the latter coefficients, the non-standardized tests have higher power. Conversely, in settings where the opposite is true, the standardized estimator allows the smaller signals that also have a low variance to have a large impact on the test, and thus larger power. While comparisons were originally made holding the number of Fourier coefficients constant, it is also worth considering comparisons across both the testing scheme and the number of coefficients. Because the non-standardized testing method would only improve or stay the same as the number of coefficients increases it would be natural to only consider the non-standardized tests that use 21 coefficients. When comparing the high dimensional non-standardized test to all standardized tests, the non-standardized test tends to have higher power. The only setting in which this is not the case is the multiple Fourier setting. However, it is also worth noting that the non-standardized test in the null setting has a higher type-one error, especially in the 21 coefficient setting.

## 2.6 Conclusion

In this chapter, we introduced a test of a functional null hypothesis. This test leverages the Fourier transform to project the function-valued parameter into $\mathbb{R}^\infty$. The dimension of this projection can be further reduced by considering only the first $d$ of the Fourier coefficients. Both the finite and infinite-dimensional versions of the test are expected to share the consistency and unbiasedness of the test described in the first chapter for most alternatives. An alternative test was also proposed to mitigate some of the issues of the original test arising from differences in the magnitude of different Fourier coefficients. The performance of each of these tests was studied in a variety of settings. While both the standardized and non-standardized tests had strengths and weaknesses, it is possible that some alternative version of the test could better balance the trade-offs made by these two tests. One potential option is a partially standardized test in which only the first, say, 11 elements of the vector of Fourier coefficients are standardized, even as the dimension of the test increases.

It was observed in the numerical study that the non-standardized test achieved similar rejection rates so long as the number of Fourier coefficients was 11 or greater. Because of the lower computational cost, lower dimensional tests are preferable if there are no losses in performance.

However, such tests also risk being inconsistent for some alternatives. A potential solution to this is to add the $\ell_2$ norm (or any norm) of the function as an additional element of the parameter vector, in addition to the Fourier coefficients that are currently considered. Such a test could achieve consistency while maintaining the advantages introduced by using the Fourier transformation.

# Chapter 3

# Estimating open-label effectiveness in trials with arm switching

## 3.1 Introduction

Of the 1.5 million new HIV infections in the last year, about half were among women. In Sub-Saharan Africa, home to roughly two-thirds of people living with HIV, over 60% of new infections were among women and girls (AIDS, 2021). One of the more promising preventions strategies is the use of Pre-Exposure Prophylaxis (PrEP) to prevent new infections (Baeten et al., 2012; Van Damme et al., 2012; Grant et al., 2010; Choopanya et al., 2013). While PrEP has been found to be efficacious in clinical trials, challenges remain for obtaining high levels of adherence across the entire population (Choopanya et al., 2013; Rees et al., 2015). Thus, to effectively combat HIV, it is critical to find interventions that are not only efficacious but also easily adopted by the at-risk population.

One such prevention modality is a vaginal ring containing Dapivirine. The double-blind, placebo-control, ASPIRE clinical trial found one such ring to be efficacious in preventing HIV (Baeten et al., 2012), and a separate trial (The Ring Study) found a similar ring to be efficacious (Nel et al., 2016). Results from each trial led to a corresponding open-label extension of each trial: HOPE, the open-label extension of ASPIRE, and DREAM, the open-label extension of The Ring Study. Both open-label extension studies assessed the uptake and use of the ring in a real-world setting and found elevated levels of adherence relative to the original studies despite less frequent follow-up visits (Baeten et al., 2021; Nel et al., 2021). Based in part on these trial results, in 2021 the WHO updated their guidelines to recommend the use of a ring (alongside other preventive measures) for women at heightened risk of HIV (Phillips, 2021).

While each of the four studies provided useful information, no single study would be sufficient to

establish open-label effectiveness by itself. While both initial trials had active comparator groups, the studies were carried out under settings that may not reflect the environment in which the ring would be used. Conversely, while the extension trials studied the ring in a more pragmatic setting, neither includes a placebo comparison group. Differences between the initial trial settings and the open-label studies include uncertainty of the efficaciousness of the prevention modality, uncertainty of arm assignment, and visit frequency. Existing analyses of the open-label extension trials make comparisons between the open-label cohort and a counterfactual placebo group, but avoid making direct claims about open-label effectiveness because the groups may not be directly comparable (Baeten et al., 2021; Nel et al., 2021). We attempt to make such comparisons using information from both the ASPIRE and HOPE studies. Under certain assumptions, the procedure described in this chapter consistently estimates the effectiveness of the ring in a counterfactual trial that is equivalent to HOPE except that it also includes a placebo arm. These assumptions include standard causal assumptions and a bridging assumption which asserts that holding individuals' characteristics and adherence fixed, the (multiplicative) risk reduction provided by the ring is the same between the two trials.

## 3.2 Background

The ASPIRE study was a placebo-control, double-blind, clinical trial with fixed-sized block randomization, stratified according to site in a 1:1 ratio. All trial participants received a ring, either containing 25mg of Dapivirine if a participant was in the active arm or a placebo ring for those in the placebo arm. Participants had monthly follow-up visits in which they returned the ring used in the previous month and received a new ring. During these visits, they were also tested for HIV. Those who became HIV positive during the trial were no longer given new rings. Four weeks after discontinuation of the ring, trial participants had a follow-up visit to determine if infection occurred at any point during the use of the ring. The follow-up period varied between individuals, with a median follow-up time of 1.6 years (ICR of 1.1 to 2.3).

Of the 1456 participants who eventually enrolled in HOPE, the median time that passed between exiting ASPIRE and enrolling in HOPE was 2 years (ICR of 1.8 to 2.1 years). Trial participants did not have access to the Dapivirine rings in between the ASPIRE and HOPE. Participants who were interested in enrolling in HOPE after ASPIRE were screened before entering the study. To enroll in HOPE an individual must have participated in ASPIRE, had a negative serological test for HIV at HOPE enrollment, have been using an effective form of contraception, not have been pregnant or breastfeeding, and have been otherwise healthy and have had no contraindications to use of the ring.

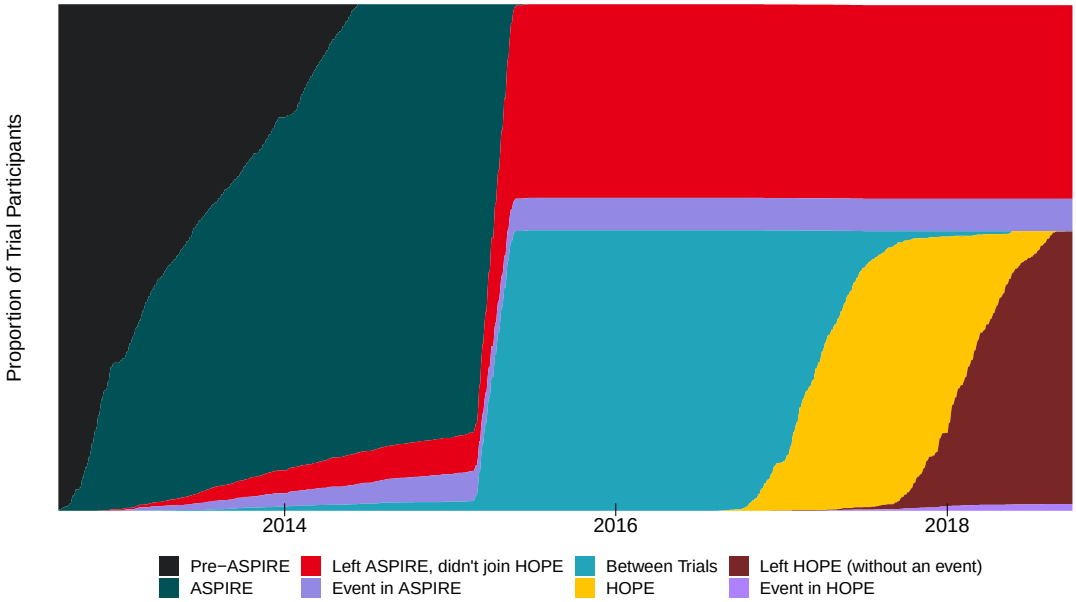


Figure 3.1: Plot showing the distribution of study participants, before, in-between, and after studies from the day before ASPIRE started to the day after HOPE ended. Each vertical strip shows the proportion of ASPIRE trial participants in each of the possible phases of trial participation. In this figure, participants are counted as having completed a trial if they are lost to follow-up.

The HOPE study was a phase 3B open-label extension of the ASPIRE trial. During the HOPE study, participants' visits occurred once a month for the first three months after entering the study. After this period, visits occurred once every three months. During each visit, participants were provided enough rings (one per month) to last until the next visit and were tested for HIV. Participants visited the clinic 4 weeks after discontinuation of the ring to determine if infection occurred at any point during the use of the ring. To assess adherence, the amount of Dapivirine released from each ring was measured. If the measured amount was above the chosen threshold the individual was considered to be adherent during the period that ring was being used.

## 3.3 Methods

### 3.3.1 Model

As the primary purpose of the HOPE trial was to assess the uptake and use of the vaginal ring, including a placebo group was unnecessary. Additionally, doing so would have been unethical given that the ASPIRE trial found the ring to be effective. This lack of a control group presents challenges for estimating the open-label effectiveness of the ring in the HOPE trial population which will be addressed in this chapter. Intuitively, the placebo group from the ASPIRE trial may be a useful comparison group for this estimate, but multiple factors must be accounted for to achieve a valid estimate. First, it is unlikely that the population level risk of infection was the same in HOPE and

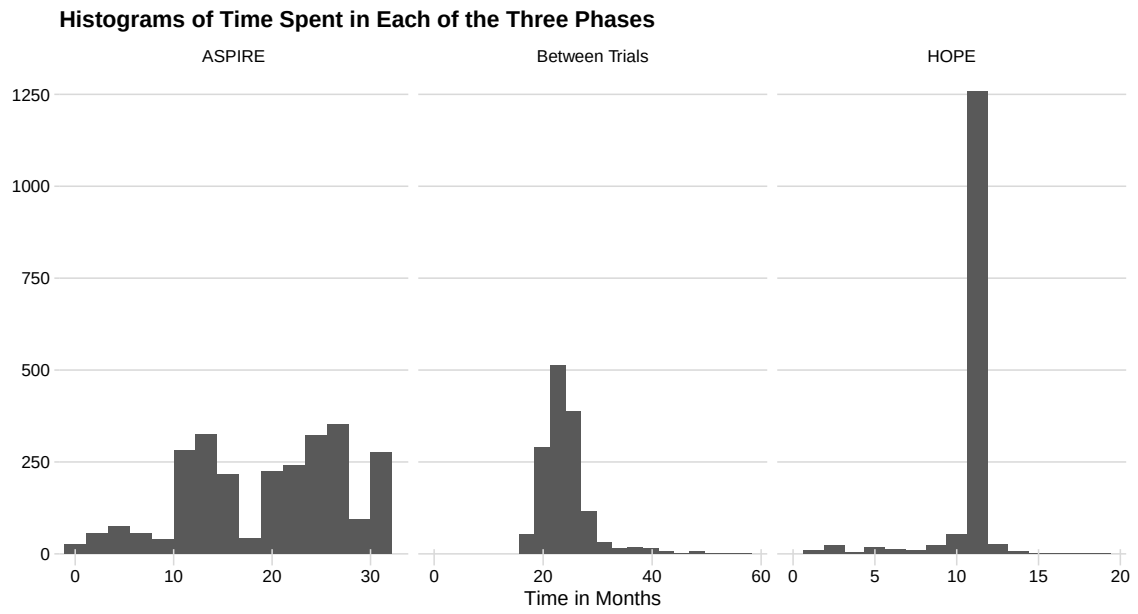


Figure 3.2: Three plots showing, from left to right, the distributions of time (in days) spent in ASPIRE, time waiting between the two trials and time spent in HOPE.

ASPIRE. To be eligible for HOPE, participants must have completed the ASPIRE trial without becoming infected so the population of HOPE could have had a lower baseline risk of infection than ASPIRE, and these differences could lead to an incorrect estimate of the parameter of interest. Second, the studies were carried out over different time periods. Thus, even if the populations of the two trials were the same, one population may be less likely to become infected due to exogenous factors such as the baseline incidence during each trial. Lastly, adherence between the two studies could be different, which could in turn result in differences in effectiveness between the two trials. It would not be unreasonable to observe differences in adherence between a trial conducted to estimate effectiveness (in which patients are followed closely but the prevention modality has not been proven effective) and an open-label extension trial (in which patients are followed less closely but effectiveness has been established). To account for these factors, we use causal inference methods and subject area expertise to identify and estimate open-label effectiveness in the HOPE trial while making the most reasonable assumptions possible.

To precisely define the parameter of interest and the assumptions required to identify it, we now introduce notation for the data from the two trials. Throughout, let upper case letters denote random variables and their lower case counterparts denote realizations of these random variables. In ASPIRE, let $D$ be an indicator for being in the active arm. Throughout, subscripts $j \in \{h, s\}$ indicate whether observations were taken during ASPIRE ($s$) or HOPE ($h$). Let $L_j$ be the set of baseline variables measured at enrollment, and note that each of these variables is measured at baseline for both trials. Let $T_j$ denote the time in months from enrollment in the given trial to

| | Trial | |
|---|---|---|
| Arm | ASPIRE | HOPE |
| Placebo | 10% | ? |
| Active | 5% | 4% |
| Risk Ratio | 0.5 | ? |

Table 3.1: Hypothetical observed 1-Year incidences in each observed arm of each study.

when an individual becomes HIV positive and let $C_j$ denote the time in months from enrollment to when an individual becomes right censored in the given trial. At times $t \in \{0, 1, \ldots, \tau\}$ months after enrollment in clinical trial $j \in \{s, h\}$, for uncensored individuals (that is, those with $C_j \geq t$), an indicator $A_{j,t}$ of adherence is measured. For all $t_j > T_j$, we let $A_{j,t_j} = \star$. We denote the history of adherence in trial $j$ through time $t$ by $\bar{A}_{j,t} := (A_{j,1}, A_{j,2}, \ldots, A_{j,t})$. For individuals who have an event, or are lost to follow-up during ASPIRE, the values of all HOPE variables are set to $\star$. Let $R$ be an indicator that an individual enrolled in HOPE. A description of what the baseline and adherence variables are for the ASPIRE and HOPE studies is provided in the Appendix of this chapter.

Let $T_h^*$ denote the counterfactual time from enrollment to becoming HIV positive if, contrary to fact, individuals in HOPE had been given a placebo ring. The parameter of interest is

$$\frac{\text{pr}(T_h > \tau \mid R = 1)}{\text{pr}(T_h^* > \tau \mid R = 1)}.$$

Because $T_h^*$ is not observed, it is not possible to identify this parameter without further assumptions. In what follows, identification will be achieved using standard causal assumptions and an additional bridging assumption that relates the data generating mechanisms of the two trials.

## 3.4 Parameter identification

There are many possible ways to identify the incidence in a counterfactual placebo arm in HOPE, each of which uses some combination of assumptions and the observed data. To motivate the final identifying assumption used for the analysis, several candidate identification assumptions are considered here. While each of the considered identification strategies is different, they all aim to tackle the same problem. In particular, when identifying the parameter, there are four populations of special interest: the ASPIRE trial placebo and active arms, the HOPE active arm, and the counterfactual HOPE placebo arm. Table 3.1 considers a hypothetical set of data, but like the actual data, three of the four subgroups have a measure for 1-year HIV infection incidence so the main challenge is in identifying the counterfactual HOPE placebo arm incidence.

The first and simplest identifying assumption considered here is the constant risk ratio as-

sumption. This assumption states that the risk ratio comparing the one-year incidence between the active and placebo arms in HOPE is the same as it was in ASPIRE. In particular, the assumption is that

$$\frac{\text{pr}(T_s > 12 \mid D = 1)}{\text{pr}(T_s > 12 \mid D = 0)} = \frac{\text{pr}(T_h > 12 \mid R = 1)}{\text{pr}(T_h^* > 12 \mid R = 1)}.$$

The implications of this assumption and how it could be used to identify the parameter are shown in Table 3.2. While this assumption has the advantage of being simple, it is unlikely this assumption would hold. To see why, first note that, because the HOPE population only includes individuals who remained HIV-1 negative throughout ASPIRE, we expect the population of HOPE to be different from that of ASPIRE. Hence, if the prevention effectiveness of the ring is heterogeneous throughout the population, the population level relative risks will be different between the two trials. Moreover, adherence was higher in HOPE than in ASPIRE, so if adherence reduces risk, we would expect larger effects in HOPE. Considering this assumption from a different angle, the assumption states that the risk ratio in the ASPIRE trial is transportable to the HOPE study and, as a result, there would be no need to even use data from HOPE to estimate the parameter of interest.

To account for some of the above deficiencies, we next consider the constant stratified risk ratio assumption. This assumption is similar to the previously described assumption but aims to properly account for differences in the distribution of baseline covariates between the two trials. This new assumption states that the risk ratio from ring use within subgroups of the population (as defined by the baseline covariates) is the same between the two trials. In particular, the assumption is that the following holds almost surely according to draws from the distribution that generated the baseline covariates $L_h$ of the HOPE participants:

$$\frac{\text{pr}(T_s > 12 \mid D = 1, L_s = l)}{\text{pr}(T_s > 12 \mid D = 0, L_s = l)}\bigg|_{l=L_h} = \frac{\text{pr}(T_h > 12 \mid R = 1, L_h)}{\text{pr}(T_h^* > 12 \mid R = 1, L_h)}.$$

Identification is still straightforward in this setting but requires more work than simply calculating the risk ratio observed in ASPIRE. In particular, the parameter of interest can be identified with a summary of the observed data distribution as follows:

$$\begin{aligned}\frac{\text{pr}\left(T_h \leq 12 | R = 1\right)}{\text{pr}\left(T_h^* \leq 12 | R = 1\right)} &= \frac{\text{pr}\left(T_h \leq 12 | R = 1\right)}{E\left[\text{pr}\left(T_h^* \leq 12 | L_h, R = 1\right)\right]} \\ &= \frac{\text{pr}\left(T_h \leq 12, R = 1\right)}{E\left[\frac{\text{pr}_P(T_h \leq 12 | L_h, R=1)\text{pr}_P(T_s \leq 12 | D=0, L_s=l)|_{l=L_h}}{\text{pr}_P(T_s \leq 12 | D=1, L_s=l)|_{l=L_h}}\right]}.\end{aligned}$$

Table 3.3 provides an example of what such an assumption would imply in a hypothetical setting

| | Trial | |
|---|---|---|
| Arm | ASPIRE | HOPE |
| Placebo | 10% | $4\% \times 10/5 = \mathbf{8\%}$ |
| Active | 5% | 4% |
| Risk Ratio | 0.5 | **0.5** |

Table 3.2: Hypothetical observed 1-Year incidences in each arm of each study under the constant risk ratio assumption. Bolded percentages indicate non-observed quantities that are implied by the constant risk ratio assumption.

| | Age $\geq 25$ | | Age $< 25$ | |
|---|---|---|---|---|
| Arm | ASPIRE | HOPE | ASPIRE | HOPE |
| Placebo | 20% | $12\% \times 20/14 \approx \mathbf{17.1\%}$ | 8% | $2\% \times 8/3 \approx \mathbf{5.3\ \%}$ |
| Active | 14% | 12% | 3% | 2% |
| Risk Ratio | 0.7 | **0.7** | 0.38 | **0.38** |

Table 3.3: Hypothetical observed 1-Year incidences in each arm of each study under the constant stratified risk ratio assumption. Bolded percentages indicate non-observed quantities that are implied by the constant stratified risk ratio assumption.

where there is a single baseline covariate that is 1 if an individual is at least 25 years of age and 0 otherwise. In this example, the estimate of $\text{pr}(T_h^* > 12 \mid R = 1)$ would be a weighted average of 17.1 and 5.3 with weights equal to the proportion of individuals in the HOPE trial who are at least 25 years old and less than 25 years old, respectively. Unlike the previous identifying assumption this assumption seeks to account for differences in the risk of infection or differences in the ring's effectiveness between trials that result from differences in the study populations' baseline covariates. However, the assumption still may not hold due to differences in the risk of infection that result from differences in adherence between the two trials.

We refer to the final assumption considered (and the one adopted in this chapter) as the constant stratified fixed adherence risk ratio assumption. This assumption aims to account for differences between the trial populations, not only with respect to the differences in baseline variables between trials but also with respect to the adherence patterns observed in each trial. While this assumption appears to be more plausible than the previous assumptions, it also requires identification of an additional counterfactual outcome, which the others did not. In particular, this outcome is the counterfactual event time in ASPIRE that would have been observed if, possibly contrary to fact, individuals displayed adherence patterns that were the same as what was observed in HOPE.

To identify the counterfactual event time in ASPIRE under an alternative adherence regime, it is necessary to account for the interplay between the event time and the time-varying adherence. As described by Hernán MA and Robins JM (2020), simply conditioning on the observed adherence pattern and assuming consistency and exchangeability is unlikely to lead to valid results. Methods exist for identifying such counterfactual outcomes in similar settings that include time-varying

covariates, and such an approach is described by Bang and Robins (2005). However, because our data generating mechanism does not include any time-varying variables other than adherence, loss to follow-up status, and HIV infection status, using such an approach may introduce unnecessary complications for identification, as well as estimation, and variance quantification. To this end, alternative adherence and censoring variables are introduced that allow for identification while avoiding the common issues arising when working with time-varying adherence variables. With these alternative variables, it is possible to write our parameter in a form similar to other commonly studied parameters of interest in causal inference and survival analysis. This will prove convenient, as it will greatly simplify both the derivation and implementation of our estimation strategy.

### 3.4.1 Defining the data generating mechanism and our bridging assumption

To define the alternative adherence random variable, note that, once an individual has an event, there are no further measurements available that inform on the adherence that would have been observed for them had they not yet been infected. Consequently, all future adherence values are set to be equal to $\star$ — put another way, if $T = j$, then $A_k = \star$ for each $k > j$. Thinking of the reverse implication, if $A_k$ is not equal to $\star$, then $T > k$. As a result, conditioning on the entire vector $\bar{A} = \bar{a}$ (with the entries of $\bar{a} \neq \star$) will exclude all individuals who had an event before the last time point (otherwise $A_\tau$ would equal $\star$). To avoid this issue, we define a counterfactual adherence variable, $\bar{A}^\dagger$. This variable, which we refer to as immune adherence, is defined as the adherence pattern that would have been observed if, possibly contrary to fact, the individual did not experience an event or become censored. We conceptualize an individual's immune adherence as the adherence an individual would have had if, unbeknownst to them, they were not susceptible to HIV infection. Since they have no knowledge of their immunity, we assume that their immune adherence is the same as their observed adherence through the time of infection — that is, we assume that $\bar{A}^\dagger_Y = \bar{A}_Y$, where $Y := \min(T, C)$. If this assumption is correct, then, given $(Y, \bar{A}^\dagger)$, $\bar{A}$ satisfies the following:

$$A_j = \begin{cases} A_j^\dagger \text{ if } Y \geq j \\ \star \text{ otherwise .} \end{cases}$$

With immune adherence now defined, it is possible to describe the assumed casual data generating mechanism more explicitly via a structural causal model (Pearl, 2009; Anderson, 1955). Let $U_1, U_2, \ldots$ denote uniform random variables that are independent from all observed random

variables and each other. The assumed model is as follows:

$$
\begin{aligned}
& & R &= f_{h,0}(T_s, C_s, L_s, U_5), \\
L_s &= f_{s,1}(U_1), & L_h &= f_{h,1}(L_s, R, U_6), \qquad (3.1) \\
\bar{A}_s^\dagger &= f_{s,2}(L_s, U_2), & \bar{A}_h^\dagger &= f_{h,2}(L_h, R, U_7), \\
T_s &= f_{s,3}(L_s, \bar{A}_s^\dagger, D, U_3), & T_h &= f_{h,3}(L_h, \bar{A}_h^\dagger, R, U_8), \qquad (3.2) \\
C_s &= f_{s,4}(L_s, \bar{A}_s^\dagger, D, U_4), & C_h &= f_{h,4}(L_h, \bar{A}_h^\dagger, R, U_9).
\end{aligned}
$$

Using this structural causal model, it is now straightforward to consider counterfactual outcomes by setting immune adherence to be equal to some fixed $\bar{a}$. We define $T_s^{\bar{a}} := f_{s,3}(L_s, \bar{a}_s, D, U_3)$ and $T_h^{\bar{a}} := f_{h,3}(L_h, \bar{a}_h, R, U_7)$. However, the bridging assumption used involves survival times under stochastic adherence regimes in which adherence is random, but not determined (as above) by $f_{s,2}$ or $f_{h,2}$. To describe this random assignment precisely, let $\pi_t(\cdot \mid \bar{a}_{t-t}, l_0)$ to denote a generic conditional distribution of immune adherence given baseline covariates $l$ and immune adherence before time $t$. Letting $\bar{\pi}(\bar{a}|l) := \pi_1(l) \prod_{i=2}^{\tau} \pi_t(a_i \mid \bar{a}_{i-1}, l)$, we will let $\bar{A}_{h,t}^{\dagger,\bar{\pi}}$ denote a random variable whose distribution is equal to that of the adherence pattern that would have been seen if, possibly contrary to fact, immune adherence had been randomly assigned based on baseline covariates. More concretely, letting $\bar{U}_j^* := (U_{j,t}^*)_{t=1}^{\tau}$ be a vector of independent exogenous uniform random variables that are independent of all other variables under consideration, we recursively define, from $t = 0, 1, \ldots, \tau$, by

$$
A_{j,t}^{\dagger,\bar{\pi}} = I\left\{U_{j,t}^* \leq \pi_t\left(1 \mid \bar{A}_{j,t-1}^{\dagger,\bar{\pi}}, L_j\right)\right\} =: f_{j,t}^{\bar{\pi}}(U_{j,t}^*, \bar{A}_{j,t-1}^{\dagger,\bar{\pi}}, L_j). \qquad (3.3)
$$

We let $f_j^{\bar{\pi}}(\bar{u}_j^*, l) = (f_{j,t}^{\bar{\pi}}(u_{j,t}^*, \bar{a}_{j,t-1}^{\dagger,\bar{\pi}}, l_j))_{t=0}^{\tau}$ and note that $\bar{A}_j^{\dagger,\bar{\pi}} := f_j^{\bar{\pi}}(\bar{U}_j^*, L)$. The stochastic adherence regime $A_{j,t}^{\dagger,\bar{\pi}}$ may be used in the structural causal model to define the counterfactual outcome of interest. This can be done for any $\bar{\pi}$ and either trial:

$$
\begin{aligned}
& & R &= f_{h,0}(T_s, C_s, L_s, U_5), \\
L_s &= f_{s,1}(U_1), & L_h &= f_{h,1}(U_5, L_s, R), \\
\bar{A}_s^{\dagger,\bar{\pi}} &= f_s^{\bar{\pi}}(L_s, \bar{U}_s^*), & \bar{A}_h^{\dagger,\bar{\pi}} &= f_h^{\bar{\pi}}(L_h, R, \bar{U}_h^*), \\
T_s^{\bar{\pi}} &= f_{s,3}(L_s, \bar{A}_s^{\dagger,\bar{\pi}}, D, U_3), & T_h^{\bar{\pi}} &= f_{h,3}(L_h, \bar{A}_h^{\dagger,\bar{\pi}}, R, U_7), \\
C_s^{\bar{\pi}} &= f_{s,4}(L_s, \bar{A}_s^{\dagger,\bar{\pi}}, D, U_4), & C_h^{\bar{\pi}} &= f_{h,4}(L_h, \bar{A}_h^{\dagger,\bar{\pi}}, R, U_8).
\end{aligned}
$$

While the definition of our parameter, the bridging assumption, and the identification results that follow would be well defined for any stochastic adherence regime $\bar{\pi}$, we now define the particular

stochastic adherence regime considered in our analysis. In particular, we consider the stochastic intervention

$$\bar{\pi}_{P,h}(\bar{a}_\tau|l) := \text{pr}(\bar{A}_h^\dagger = \bar{a}_\tau|L_h = l, R = 1),$$

where we have made the dependence of this intervention on the underlying probability distribution $P$ that generated the data explicit in the notation. With the stochastic adherence regime now defined, we can precisely state the bridging assumption that we work with:

$$\left.\frac{\text{pr}\left(T_s^{\bar{\pi}_{P,h}} \leq \tau|D = 1, L_s = l\right)}{\text{pr}\left(T_s \leq \tau|D = 0, L_s = l\right)}\right|_{l=L_h} = \frac{\text{pr}\left(T_h \leq \tau|L_h, R = 1\right)}{\text{pr}\left(T_h^\star \leq \tau|L_h, R = 1\right)} \; P - \text{a.s.}$$

When this assumption holds, we have that

$$\begin{aligned}\frac{\text{pr}\left(T_h \leq \tau|R = 1\right)}{\text{pr}\left(T_h^\star \leq \tau|R = 1\right)} &= \frac{\text{pr}\left(T_h \leq \tau|R = 1\right)}{E\left[\text{pr}\left(T_h^\star \leq \tau|L_h, R = 1\right)\right]} \\ &= \frac{\text{pr}\left(T_h \leq \tau, R = 1\right)}{E\left[\frac{\text{pr}(T_h \leq \tau|L_h, R=1)\text{pr}(T_s \leq \tau|D=0, L_s=l)|_{l=L_h}}{\text{pr}(T_s^{\bar{\pi}} \leq \tau|D=1, L_s=l)|_{l=L_h}}\right]}.\end{aligned} \tag{3.4}$$

### 3.4.2 Identifying counterfactual means

Considering (3.4), note there are still multiple challenges to identifying the parameter of interest. First, the outcome of interest is right censored. Second, by the law of total expectation, $\text{pr}_P\left(T_s^{\bar{\pi}} \leq \tau|D = 1, L\right)$ corresponds to an expectation taken over conditional probabilities for each possible fixed adherence pattern that could be realized under the stochastic adherence regime $\bar{\pi}$. Depending on the observed event or censoring time, the observed adherence pattern may also be partially or completely censored. To address both issues simultaneously, an alternative censoring variable is introduced that accounts for both the first time an individual fails to follow a specified adherence pattern and their right censoring time $C_j$. This variable, which we refer to as $C_j^*$, is defined as follows:

$$C_j^*(\bar{a}_\tau) = \min(C_j, c_j(\bar{a}_\tau, \bar{A}_j^\dagger)), \text{ where } c_j(\bar{a}_\tau, \bar{A}_j^\dagger) = \sum_{i=0}^{\tau} I\{\bar{A}_i^\dagger = \bar{a}_i\}.$$

Using this censoring variable, let $Y_j^*(\bar{a}_\tau) := \min(T_j, C_j^*(\bar{a}_\tau))$ and $\Delta_j^*(\bar{a}_\tau) = I\{T_j \leq C_j^*(\bar{a}_\tau)\}$.

Now considering the counterfactual event times in ASPIRE under the given adherence regime:

$$\begin{aligned}\text{pr}(T_s^{\bar{\pi}} \geq \tau|L_s = l, D = 1) &= E\left[\text{pr}(T_s^{\bar{\pi}} \geq \tau|\bar{A}_s^{\dagger,\bar{\pi}}, L_s = l, D = 1)|L_s = l, D = 1\right] \\ &= \sum_{\bar{a}_\tau} \text{pr}(T_s^{\bar{\pi}} \geq \tau|\bar{A}_s^{\dagger,\bar{\pi}} = \bar{a}_\tau, L_s = l, D = 1)\bar{\pi}^\dagger(\bar{a}_\tau|l).\end{aligned}$$

Since $\bar{A}^{\dagger}_{j,T} = \bar{A}_{j,T}$, it follows that, conditional on $\bar{A}^{\dagger}_{j} = \bar{a}_j$, $T^{\bar{\pi}}_s = T_s$. Thus the above display continues:

$$= \sum_{\bar{a}_\tau} \text{pr}(T^{\bar{a}_\tau}_s \geq \tau | \bar{A}^{\dagger,\bar{\pi}}_s = \bar{a}_\tau, L_s = l, D = 1) \bar{\pi}^{\dagger}(\bar{a}_\tau | l).$$

Note that by the structural causal model (see equation 3.3 and the definition of $T^{\bar{a}_\tau}_j$), $\bar{A}^{\bar{\pi},\dagger}_s$ is independent of $T^{\bar{a}_\tau}_s$ conditional on $L_s$ and $D$. Thus, the display continues as

$$= \sum_{\bar{a}_\tau} \text{pr}(T^{\bar{a}_\tau}_s \geq \tau | L_s = l, D = 1) \bar{\pi}^{\dagger}(\bar{a}_\tau | l).$$

To identify $\text{pr}(T^{\bar{\pi}}_s \geq \tau | L_s = l, D = 1)$, we will show that, for each $\bar{a}_\tau$ and almost all $l$ as drawn from the HOPE distribution, it is possible to identify $\text{pr}(T^{\bar{a}_\tau}_s \geq \tau | L_s = l, D = 1)$ using the observed data. We start by noting that, by definition (as provided by the SCM), $T^{\bar{a}_\tau}_s$ is independent of $C^*_s(\bar{a}_\tau)$ given $L$ and $D$. This fact can be seen by recalling that $T^{\bar{a}}_s := f_{s,3}(L_s, \bar{a}_s, D, U_3)$, conditional on $L$ and $D$, is only random through the exogenous random variable $U_3$. As with $T_s$ and $C_s$, we observe neither $T^{\bar{a}_\tau}_s$ or $C^*_s(\bar{a}_\tau)$ fully. However, as will be shown, we do observe a censored version of these two variables given by

$$Y^{\bar{a}_\tau}_s(\bar{a}_\tau) = \min\{T^{\bar{a}_\tau}_s, C^*_s(\bar{a}_\tau)\}, \ \Delta^{\bar{a}_\tau}(\bar{a}_\tau) = I\{T^{\bar{a}_\tau}_s < C^*_s(\bar{a}_\tau)\},$$

With these observed variables, it is possible to identify $\text{pr}(T^{\bar{\pi}}_s \geq \tau | L_s = l, D = 1)$ using identification methods from survival analysis. We start by noting that both $\Delta^*_s(\bar{a}_\tau) := I\{T_s \leq C^*_s(\bar{a}_\tau)\}$ and $Y^*_s(\bar{a}_\tau) := \min\{T_s, C^*_s(\bar{a}_\tau)\}$ are observed. Considering first the missingness indicator:

$$\begin{aligned}
\Delta^*_s(\bar{a}_\tau) &= I\{T_s \leq C^*_s(\bar{a}_\tau)\} \\
&= I\{T_s \leq C_j, T_s \leq c_j(\bar{a}_\tau, \bar{A}^{\dagger}_s)\},
\end{aligned}$$

for the second statement in the indicator function to be true, it must be the case that $\bar{A}^{\dagger}_{j,T_j} = \bar{a}_{T_j}$ and consequently $T_s = T^{\bar{a}_\tau}_s$. Thus the previous display continues as:

$$\begin{aligned}
&= I\{T_s \leq C_j, T_s \leq c_j(\bar{a}_\tau, \bar{A}^{\dagger}_s), \bar{A}^{\dagger}_{T_s} = \bar{a}_T\} \\
&= I\{T^{\bar{a}_\tau}_s \leq C_j, T^{\bar{a}_\tau}_s \leq c_j(\bar{a}_\tau, \bar{A}^{\dagger}_s), \bar{A}^{\dagger}_T = \bar{a}_T\} && \text{by consistency} \\
&= I\{T^{\bar{a}_\tau}_s \leq C^*_s(\bar{a}_\tau)\}.
\end{aligned}$$

Thus, it follows that $\Delta_s^{\bar{a}_\tau}(\bar{a}_\tau)$ is equal to the observed $\Delta_s^*(\bar{a}_\tau)$. Also, note that

$$\begin{aligned} Y_s^*(\bar{a}_\tau) &= \min\{T_s, C_s^*(\bar{a}_\tau)\} \\ &= T_s\Delta_s^*(\bar{a}_\tau) + C_s^*(\bar{a}_\tau)[1 - \Delta_s^*(\bar{a}_\tau)]. \end{aligned}$$

If $\Delta_s^*(\bar{a}_\tau) = 1$, it follows by consistency that $T_s = T_s^{\bar{a}_\tau}$. Thus, the above display continues as:

$$\begin{aligned} &= T_s^{\bar{a}_\tau}\Delta_s^*(\bar{a}_\tau) + C_s^*(\bar{a}_\tau)[1 - \Delta_s^*(\bar{a}_\tau)] \\ &= \min\{T_s^{\bar{a}_\tau}, C_s^*(\bar{a}_\tau)\}. \end{aligned}$$

Thus, the observed $Y_s^*(\bar{a}_\tau)$ is equal to $Y_s^{\bar{a}_\tau}(\bar{a}_\tau)$. We may then conclude that using the observed data and survival estimation techniques it is possible to estimate $\text{pr}(T_s^{\bar{a}_\tau} \geq \tau | L_s = l, D = 1)$.

One of the remaining hurdles is that this identified probability is conditional on baseline covariates. Fortunately, identification and estimation of such conditional survival outcomes has been studied previously (van der Laan and Dudoit, 2003; Hothorn et al., 2006; van der Laan and Gruber, 2011). Our chosen estimation strategy and its implementation will rely most heavily on the approach described in Westling et al. (2021). We begin by defining the identifiable parameters used for the estimation of the survival function. We attempt to closely follow the notation used by Westling et al. (2021) but note that adherence is treated differently here than it is in the cited article. In particular, in our derivations, adherence defines the censoring and time to event random variables, rather than being considered as a random variable itself, as it is in Westling et al., 2021. Let

$$\begin{aligned} F_{j,1}(P, u, \bar{a}_\tau | l) &:= \text{pr}_P(Y_j^*(\bar{a}_\tau) \leq u, \Delta_j^*(\bar{a}_\tau) = 1 | L_j = l), & F_{j,1}(P, u | l) &:= \text{pr}_P(Y \leq u, \Delta_j^* = 1 | L_j = l), \\ R_j(P, u, \bar{a}_\tau | l) &:= \text{pr}_P(Y_j^*(\bar{a}_\tau) \geq u | L_j = l), & R_j(P, u | l) &:= \text{pr}_P(Y_j^* \geq u, L_j = l), \\ \Lambda_j(P, t, \bar{a}_\tau | l) &:= \int_0^t \frac{F_{j,1}(P, du, \bar{a}_\tau | l)}{R_j(P, u, \bar{a}_\tau | l)}, & \Lambda_j(P, t | l) &:= \int_0^t \frac{F_{j,1}(P, du | l)}{R_j(P, u | l)}, \\ S_j(P, t, \bar{a}_\tau | l) &:= \prod_{(0,t]} \{1 - \Lambda_j(P, du, \bar{a}_\tau | l)\}, & S_j(P, t | l) &:= \prod_{(0,t]} \{1 - \Lambda_j(P, du | l)\}, \\ \theta_j(P, t, \bar{a}_\tau) &:= E_P\left[S_j(P, t, \bar{a}_\tau | L_j)\right], & \theta_j(P, t) &:= E_P\left[S_j(P, t | L_j)\right], \end{aligned}$$

where $\prod$ denotes the Riemann-Stieltjes product integral. Let $S_s^\dagger(P, l) := S_s(P_{D=1}, \tau | l)$, where $P_{D=1}$ is the distribution of $P$ conditional on $D = 1$. Under the specified SCM, $S_s^\dagger(P, l)$ identifies $\text{pr}(T^{\bar{a}_\tau} \geq \tau | L_s = l, D = 1)$. Thus,

$$Q(\bar{\pi}_{h,P}, P, l) := \sum_{\bar{a}_\tau} S_s^\dagger(P, \bar{a}_\tau | l)\bar{\pi}_{h,P}(\bar{a}_\tau | l)$$

identifies $\text{pr}(T_s^{\bar{\pi}} \geq \tau | L_s = l, D = 1)$. Similarly, the above definitions can be used to identify the other probabilities in (3.4). Define $S_h(P, l) := S_h(P_{R=1}, \tau | l)$ where $P_{R=1}$ is the distribution of $P$ conditional on $R = 1$, and note that $S_h(P, l)$ is equal to $\text{pr}_P(T_h \leq \tau | L_h = l, R = 1)$ so long as $T_h \perp C_h \mid L_h, R = 1$. Also, let $S_s(P, l) := S_s(P_{D=0}, \tau | l)$ where $P_{D=0}$ is the distribution of $P$ conditional on $D = 0$, which identifies $\text{pr}_P(T_s \leq \tau | L_s = l, D = 0)$ so long as $T_s \perp C_s \mid L_s, D = 0$. Now that each of the probabilities in (3.4) has been identified, we have identified the parameter:

$$\frac{\text{pr}(T_h \leq \tau | R = 1)}{E\left[\frac{\text{pr}_P(T_h \leq \tau | L_h, R=1)\text{pr}_P\left(T_s^{\bar{\pi}^\star} \leq \tau | D=0, L_s=l\right)|_{l=L_h}}{\text{pr}_P(T_s^{\bar{\pi}} \leq \tau | D=1, L_s=l)|_{l=L_h}}\right]} = \frac{E[1 - S_h(P, L_h)]}{E\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{1-Q(\bar{\pi}_{h,P}, P, L_h)}\right]}. \tag{3.5}$$

### 3.4.3 Influence curve derivation

Now that the parameter of interest has been identified, constructing an estimator is a natural next step. We use a one-step estimator for this purpose (Bickel et al., 1998). An important step in deriving a one-step estimator and its corresponding standard error is determining the influence curve of the parameter. As we did for parameter identification, both the derivation and computation of the influence curve will build on results from Westling et al. (2021). In the cited article, the influence curve of $\theta_j(\cdot, t, \bar{a}_\tau)$ was provided for any $(t, \bar{a}_\tau)$. While our parameter of interest is not written entirely in terms of $\theta_s$ and $\theta_h$, knowledge of their influence curves can be informative. First note that each of the functions $S_h, S_s$ and $S_s^\dagger$ are functions of $P$ and baseline covariates. Further, they appear in our parameter of interest within an expectation over the baseline covariates in HOPE. Because $\theta_j$ is defined as the expected value of $S_j$ with respect to baseline covariates, knowledge of its influence curve is informative. To understand how the influence curve of $\theta_j$ can be useful, let $D$ denote the gradient of $\theta_j$, and let $Z_j$ denote all observed variables in the trial specified by the subscript, except baseline covariates. The gradient of $\theta_j$ can be broken down into $D = D_{Z|L} + D_L$, where $D_{Z|L}$ is the portion of the gradient that arises from perturbing the conditional distribution of $P_{Z|L}$ while keeping $P_L$ fixed. Similarly, $D_L$ is the portion of the gradient arising from perturbing $P_L$ while keeping the conditional distribution $P_{Z|L}$ fixed. From the definition of $\theta_j$, it follows that $D_L = S(P, t, l) - \theta(P, t)$, and thus it follows that

$$\begin{aligned} D_{Z|L}(z, l) &= D(z, l) - D_L(l) \\ &= D(z, l) - [S(P, t, l) - \theta(P, t)]. \end{aligned}$$

Next, note that $E[D_{Z|L}(Z, L) \mid L] = 0$ almost surely. Last, letting $\{P_\varepsilon : \varepsilon\}$ be a smooth univariate submodel through $P$ such that $P_0 = P$, if $P_{\varepsilon,L} = P_L$ for each $\varepsilon$ and the submodel has a score

$u_{Z|L} \in L_0^2(P_{Y|Z})$ at $\varepsilon = 0$, then, for $P$-almost all $l$,

$$\frac{d}{d\varepsilon} S(P_\varepsilon, t, l) = \int D_{Z|L}(z, l) u_{Z|L}(z, l) P_{Z|L}(dz|l).$$

This can be shown to follow from the fact that the mapping $l \mapsto S(P_\varepsilon, t, l)$ is invariant to changes in the distribution of baseline covariates. In the derivation of the influence curve of our parameter of interest, it will be important to calculate $\frac{d}{d\varepsilon} S(P_\varepsilon, t, l)$, and the above findings provide a method of doing so provided that the form of the influence curve of $\theta_j$ is known. Fortunately, this has been calculated by Westling et al. (2021). To express these influence curves, it is necessary to introduce new notation. Let

$$F_{j,0}(P, u, \bar{a}_\tau|l) := \text{pr}_P(Y_j^*(\bar{a}_\tau) \leq u, \Delta_j^*(\bar{a}_\tau) = 0 | L_j = l), \quad F_{j,0}(P, u|l) := \text{pr}_P(Y \leq u, \Delta_j^* = 0 | L_j = l),$$
$$H_j(P, u, \bar{a}_\tau|l) := \int_0^u \frac{S_j(P, -s, \bar{a}_\tau|l)}{S_j(P, s, \bar{a}_\tau|l)} \frac{F_{j,0}(P, ds, \bar{a}_\tau|l)}{R_j(P, s, \bar{a}_\tau|l)}, \qquad H_j(P, u|l) := \int_0^u \frac{F_{j,0}(P, ds|l)}{R_j(P, s|l)} \frac{S_j(P, -s|l)}{S_j(P, s|l)},$$
$$G_j(P, t, \bar{a}_\tau|l) := \prod_{(0,t]} \{1 - H_j(P, du, \bar{a}_\tau|l)\}, \qquad G_j(P, t|l) := \prod_{(0,t]} \{1 - H_j(P, du|l)\}.$$

Theorem 2 of Westling et al. (2021) implies that for a given $\bar{a}_\tau$ and $t$, the influence curve of $\theta_j(\cdot, t, \bar{a}_\tau)$ is $\phi_{j,t}(y, \delta, \bar{a}_\tau, l) - \theta_j(\cdot, t, \bar{a}_\tau)$, where

$$\phi_{j,t}(y, \delta, \bar{a}_\tau, l) := S_j(P, \tau, \bar{a}_\tau, l) \left[ 1 - \left\{ \frac{I(y \leq t, \delta = 1)}{S_j(P, y, \bar{a}_\tau|l) G_j(P, y, \bar{a}_\tau|l)} - \int_0^{\tau \wedge y} \frac{\Lambda_j(P, du, \bar{a}_\tau|l)}{S_j(P, u, \bar{a}_\tau|l) G_j(P, u, \bar{a}_\tau|l)} \right\} \right].$$

Similarly, for a given $t$, the influence curve of $\theta_j(\cdot, t)$ is given by $\phi_{j,t}(y, \delta, l) - \theta_j(P, t)$, where

$$\phi_{j,t}(y, \delta, l) := S_j(P, \tau, l) \left[ 1 - \left\{ \frac{I(y \leq \tau, \delta = 1)}{S_j(P, y|l) G_j(P, y|l)} - \int_0^{\tau \wedge y} \frac{\Lambda_j(P, du|l)}{S_j(P, u|l) G_j(P, u|l)} \right\} \right].$$

It follows from the previous findings and influence curves derived by Westling et al. (2021) that:

$$\frac{d}{d\varepsilon} S_j(P_\varepsilon, t, \bar{a}_\tau, l) = S_j(P, t, \bar{a}_\tau, l) \left[ \int_0^{t \wedge y} \frac{\Lambda_j(P, du, \bar{a}_\tau|l)}{S_j(P, u, \bar{a}_\tau|l) G_j(P, u, \bar{a}_\tau|l)} - \frac{I(y \leq t, \delta = 1)}{S_j(P, y, \bar{a}_\tau|l) G_j(P, y, \bar{a}_\tau|l)} \right].$$

Define the function $\phi_s^\dagger$ such that $\phi_s^\dagger(y, \delta, \bar{a}_\tau|l) = \frac{d}{d\varepsilon} S_s(P_{D=1,\varepsilon}, \tau, \bar{a}_\tau, l)$. Similarly to the above,

$$\frac{d}{d\varepsilon} S_j(P_\varepsilon, t, l) = S_j(P, t, l) \left[ \int_0^{t \wedge y} \frac{\Lambda_j(P, du|l)}{S_j(P, u|l) G_j(P, u|l)} - \frac{I(y \leq t, \delta = 1)}{S_j(P, y|l) G_j(P, y|l)} \right].$$

Let $\phi_h(y, \delta|l, 1) := \frac{d}{d\varepsilon} S_h(P_{R=1,\varepsilon}, \tau, l)$, and let $\phi_s(y, \delta|l, 0) := \frac{d}{d\varepsilon} S_s(P_{D=0,\varepsilon}, \tau, l)$. Each of these functions is referred to later in this chapter as a pseudo gradient. The above building blocks will assist in the derivation of the influence of the parameter of interest that is now presented. To

simplify the influence curve derivation, define

$$\Psi_{den}(P_1, P_2, P_3, P_4, P_5) := E_{P_5}\left[\frac{\{1 - S_h(P_1, L_h)\}\{1 - S_s(P_2, L_h)\}}{1 - Q(\bar{\pi}_{h,P_3}, P_4, L_h)}\right].$$

Using the above definition, it is possible to perturb each of the five probability distributions separately to determine the pathwise derivative of $\Psi^*_{den}$ defined by $\Psi^*_{den}(P) := \Psi_{den}(P, P, P, P, P)$. Throughout these derivations, we denote probability densities with the lowercase p. Subscripts indicate if the probabilities are taken with respect to ASPIRE (s) or HOPE (h). Conditional probability density functions condition on the random variables indicated to the right of | and for observations from the trial indicated by the subscript on $p$. As an example, $p_s(y, \delta|l, d)$ denotes the joint probability density of $Y_s$ and $\Delta_s$ conditional on the event $D = d$ and $L_s = l$. Throughout, $l$ denotes a generic realization of baseline covariates (that could come from either trial).

Now, considering $\Psi_{den}$, the first pathwise derivative considered perturbs the distribution that indexes the HOPE survival probability (conditional on HOPE baseline covariates $L$). Letting $u^1_h \in L^2_0(P_{Y_h,\Delta_h|L_h,R=1})$, note that:

$$\begin{aligned}
&\frac{d}{d\varepsilon}\Psi_{den}(P_\varepsilon, P, P, P, P)|_{\varepsilon=0} \\
&= \frac{d}{d\varepsilon}E_P\left[\frac{\{1 - S_h(P_\varepsilon, L_h)\}\{1 - S_s(P, L_h)\}}{1 - Q(\bar{\pi}_{h,P}, P, L_h)}\right]|_{\varepsilon=0} \\
&= \iint -u_{h,1}(y, \delta, l, 1)\phi_h(y, \delta|l, 1)p_h(y, \delta|l, R = 1)dyd\delta\frac{1 - S_s(P, l)}{1 - Q(\bar{\pi}_{h,P}, P, l)}p_h(l|R = 1)dl \\
&= \sum_{r=0}^{1}\iint -u_{h,1}(y, \delta, l, 1)\frac{I\{r = 1\}}{p_h(r)}\phi_h(y, \delta|l, 1)p_h(y, \delta|l, 1)dyd\delta\frac{1 - S_s(P, l)}{1 - Q(\bar{\pi}_{h,P}, P, l)}p_h(l|r)p_h(r)dl \\
&= \sum_{r=0}^{1}\iint -u_{h,1}(y, \delta, l, 1)\frac{I\{r = 1\}}{p_h(r)}\phi_h(y, \delta|l, 1)\frac{1 - S_s(P, l)}{1 - Q(\bar{\pi}_{h,P}, P, l)}p_h(y, \delta, l, r)dyd\delta dl.
\end{aligned}$$

Next, let the function $\phi^*_1$ be defined by $\phi^*_1(y, \delta, l, r) = -\frac{I\{r=1\}}{p_h(r)}\phi_h(y, \delta|l)\frac{1-S_s(P,l)}{1-Q(\bar{\pi}_{h,P},P,l)}$ and note that $E[u_{h,1}(Y_h, \Delta_h, L_h, 1)|L_h = l, R = 1]$ is equal to zero. Defining $\phi_1$ by $\phi_1(y, \delta, l, r) = \phi^*_1(y, \delta, l, r) - E[\phi^*_1(Y_h, \Delta_h, L_h, R)|L_h = l, R = r]$, the previous display continues as

$$= \sum_{r=0}^{1}\iint u_{h,1}(y, \delta, l, 1)\phi_1(y, \delta, l, r)p_h(y, \delta, l, r)dyd\delta dl = \int u_{h,1}(y, \delta, l, 1)\phi_1(y, \delta, l, r)dP$$

Next, considering perturbations of $P_2$, the derivation begins similarly. Letting $u_{s,2} \in L^2_0(P_{Y_s,\Delta_s|L_s,D=0})$, note that:

$$\begin{aligned}
&\frac{d}{d\varepsilon}\Psi_{den}(P, P_\varepsilon, P, P, P)|_{\varepsilon=0} \\
&= \frac{d}{d\varepsilon}E_P\left[\frac{\{1 - S_{h,\tau}(P, L_h)\}\{1 - S_h(P_\varepsilon, L_h)\}}{1 - Q(\bar{\pi}_{h,P}, P, L_h)}\right]|_{\varepsilon=0}
\end{aligned}$$

$$= \iint -u_{s,2}(y,\delta,l,0)\phi_s(y,\delta|l,0)p_s(y,\delta|l,D=0)dyd\delta\frac{1-S_h(P,l)}{1-Q(\bar{\pi}_{h,P},P,l)}p_h(l|R=1)dl$$
$$= \sum_{d=0}^{1}\iint -u_{s,2}(y,\delta,l,0)\frac{I\{d=0\}}{p_s(d|l)}\phi_s(y,\delta|l,0)p_s(y,\delta|l,d)p_s(d|l)dyd\delta\frac{1-S_h(P,l)}{1-Q(\bar{\pi}_{h,P},P,l)}p_h(l|R=0)dl$$
$$= \sum_{d=0}^{1}\iint -u_{s,2}(y,\delta,l,0)\frac{I\{d=0\}}{p_s(d|l)}\phi_s(y,\delta|l,0)\frac{1-S_h(P,l)}{1-Q(\bar{\pi}_{h,P},P,l)}p_s(y,\delta,d|l)dyd\delta p_h(l|R=1)dl$$
$$= \sum_{d=0}^{1}\iint -u_{s,2}(y,\delta,l,0)\left[\frac{I\{d=0\}}{p_s(d|l)}\phi_s(y,\delta|l,0)\frac{1-S_h(P,l)}{1-Q(\bar{\pi}_{h,P},P,l)}\frac{p_h(l|R=1)}{p_s(l)}\right]p_s(y,\delta,d|l)p_s(l)dlyd\delta.$$

Next, let the function $\phi_2^*$ be defined by

$$\phi_2^*(y,\delta,l,d) = -\frac{I\{d=0\}}{p_s(d|l)}\phi_s(y,\delta|l)\frac{1-S_h(P,l)}{1-Q(\bar{\pi}_{h,P},P,l)}\frac{p_h(l|R=1)}{p_s(l)},$$

and note that $E[u_{s,2}(Y_s,\Delta_s,L_s,0)|L_s=l,D=0]$ is equal to zero. Defining $\phi_2$ by $\phi_2(y,\delta,l,d) = \phi_2^*(y,\delta,l,d) - E[\phi_2^*(Y_s,\Delta_s,L_s,D)|L_s=l,D=0]$, the previous display continues as

$$= \sum_{d=0}^{1}\iint u_{s,2}(y,\delta,l,0)\phi_2(y,\delta,l,d)p_s(y,\delta,l,d)dyd\delta dl = \int u_{s,2}(y,\delta,l,0)\phi_2(y,\delta,l,d)dP.$$

Unlike for the time-varying covariate case, here $\bar{\pi}$ is separate from any expectation over what was observed during ASPIRE which will simplify the calculation of the IC. Let $u_{h,3} \in L_0^2(P_{\bar{A}_{h,\tau}|L_h,R=1})$ and note that:

$$\frac{d}{d\varepsilon}Q(\bar{\pi}_{h,P_\varepsilon},P,L)|_{\varepsilon=0} = \sum_{\bar{a}_\tau} S_s^\dagger(P,\bar{a}_\tau,L)\bar{\pi}_{h,P}(\bar{a}_\tau|l,R=1)u_{h,3}(\bar{a}_\tau,l,1).$$

Now, considering the entirety of $\frac{d}{d\varepsilon}\Psi_{den}(P,P,P_\varepsilon,P,P)$, note that:

$$\frac{d}{d\varepsilon}\Psi_{den}(P,P,P_\varepsilon,P,P) = E_P\left[\frac{\{1-S_{h,\tau}(P,L_h)\}\{1-S_{s,\tau}(P,L_h)\}}{(1-Q(\bar{\pi}_P,P,L))^2}\frac{d}{d\varepsilon}Q(\bar{\pi}_{P_\varepsilon},P,L)\right].$$

Plugging in what was found as the value of $\frac{d}{d\varepsilon}Q(\bar{\pi}_{P_\varepsilon},P,L)$:

$$E_P\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{(1-Q(\bar{\pi}_P,P,L))^2}\frac{d}{d\varepsilon}Q(\bar{\pi}_{P_\varepsilon},P,L)\right].$$
$$= \int\left[\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\sum_{\bar{a}_\tau}S_s^\dagger(P,\bar{a}_\tau,l)\bar{\pi}_{h,P}(\bar{a}_\tau|l)u_{h,3}(\bar{a}_\tau,l,1)\right]p_h(l|R=1)dl$$
$$= \sum_{r=0}^{1}\int\left[\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\frac{I\{r=1\}}{p_h(r)}\sum_{\bar{a}_\tau}S_s^\dagger(P,\bar{a}_\tau,l)\bar{\pi}_{h,P}(\bar{a}_\tau|l)u_{h,3}(\bar{a}_\tau,l,1)\right]p_h(l|r)p_h(r)dl$$
$$= \int\sum_{\bar{a}_\tau}u_{h,3}(\bar{a}_\tau,l,r)\left[\frac{I\{r=1\}}{p_h(r)}\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\})}{(1-Q(\bar{\pi}_P,P,l))^2}S_s^\dagger(P,\bar{a}_\tau,l)\right]p_h(\bar{a}_\tau,l,r)dl$$
$$= \int\sum_{\bar{a}_\tau}u_{h,3}(\bar{a}_\tau,l,1)\left[\frac{I\{r=1\}}{p_h(r)}\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\left\{S_s^\dagger(P,\bar{a}_\tau,l)-Q(\bar{\pi}_P,P,l)\right\}\right]p_h(\bar{a}_\tau,l,r)dl.$$

Next, let the function $\phi_3^*$ be defined by

$$\phi_3^*(\bar{a}_\tau, l, r) = \frac{I\{r=1\}}{p_h(r)} \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2} \left\{S_s^\dagger(P,\bar{a}_\tau,l) - Q(\bar{\pi}_P,P,l)\right\},$$

and note that $E[u_{h,3}(\bar{A}_{h,\tau}, L_h, 1)|L_h = l, R=1]$ is equal to zero. Defining $\phi_3$ by $\phi_3(\bar{a}_\tau, l, r) = \phi_3^*(\bar{a}_\tau, l, r) - E[\phi_3^*(\bar{A}_{h,\tau}, L_h, R)|L_h = l, R=1]$, the previous display continues as

$$= \int \sum_{\bar{a}_\tau} u_{h,3}(\bar{a}_\tau, l, 1)\phi_3(\bar{a}_\tau, l, r) p_h(\bar{a}_\tau, l, r) dl = \int u_{h,3}(\bar{a}_\tau, l, 1)\phi_3(\bar{a}_\tau, l, r) dP.$$

Letting $u_{s,4} \in L_0^2(P_{Y_s,\Delta_s|L,D=1})$, now consider perturbations of the ASPIRE outcome:

$$\begin{aligned}\frac{d}{d\varepsilon} Q(\bar{\pi}_P, P_\varepsilon, l)|_{\varepsilon=0} &= \frac{d}{d\varepsilon} \sum_{\bar{a}_\tau} S_s^\dagger(P_\varepsilon, \bar{a}_\tau, L)\bar{\pi}_{h,P}(\bar{a}_\tau|l)|_{\varepsilon=0} \\ &= \sum_{\bar{a}_\tau} \int \phi_s^\dagger(y,\delta,\bar{a}_\tau|l) u_{s,4}(y,\delta,l,1) p_s(y_s,\delta_s|l,D=1)\bar{\pi}_{h,P}(\bar{a}_\tau|l).\end{aligned}$$

Now, considering the entirety of $\frac{d}{d\varepsilon}\Psi_{den}(P,P,P,P_\varepsilon,P)$, note that:

$$\begin{aligned}\frac{d}{d\varepsilon}\Psi_{den}(P,P,P,P_\varepsilon,P)|_{\varepsilon=0} &= \frac{d}{d\varepsilon} E_P\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{(1-Q(\bar{\pi}_P,P_\varepsilon,L_h))}\right]|_{\varepsilon=0} \\ &= E_P\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{(1-Q(\bar{\pi}_P,P,L))^2}\frac{d}{d\varepsilon}Q(\bar{\pi}_P,P_\varepsilon,L)\right]|_{\varepsilon=0}.\end{aligned}$$

Plugging in what was found as the value of $\frac{d}{d\varepsilon}Q(\bar{\pi},P_\varepsilon,L)$:

$$\begin{aligned}&= E_P\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{(1-Q(\bar{\pi}_P,P,L))^2}\frac{d}{d\varepsilon}Q(\bar{\pi}_{P_\varepsilon},P,L)\right] \\ &= E_P\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{(1-Q(\bar{\pi}_P,P,L))^2}\sum_{\bar{a}_\tau}\int \phi_s^\dagger(y,\delta,\bar{a}_\tau|L)u_{s,4}(y,\delta,L,1)p_s(y,\delta|L,D=1)\bar{\pi}_{h,P}(\bar{a}_\tau|L)dyd\delta\right] \\ &= \iint \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\sum_{\bar{a}_\tau}\phi_s^\dagger(y,\delta,\bar{a}_\tau|l)u_{s,4}(y,\delta,l,1)p_s(y,\delta|l,D=1)\bar{\pi}_{h,P}(\bar{a}_\tau|l)p_h(l)dyd\delta dl \\ &= \iint u_{s,4}(y,\delta,l,1)\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\sum_{\bar{a}_\tau}\phi_s^\dagger(y,\delta,\bar{a}_\tau|l)\bar{\pi}_{h,P}(\bar{a}_\tau|l)p_s(y,\delta|l,D=1)p_h(l)dyd\delta dl \\ &= \iint u_{s,4}(y,\delta,l,1)\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\frac{p_h(l)}{p_s(l)}\sum_{\bar{a}_\tau}\phi_s^\dagger(y,\delta,\bar{a}_\tau|l)\bar{\pi}_{h,P}(\bar{a}_\tau|l)p_s(y,\delta|l,D=1)p_s(l)dyd\delta dl.\end{aligned}$$

Next, to ease notation, let $\phi_s^{\dagger,\bar{\pi}}(y,\delta|l) := \sum_{\bar{a}_\tau}\phi_s^\dagger(y,\delta,\bar{a}_\tau|l)\bar{\pi}_{h,P}(\bar{a}_\tau|l)$. Thus the above display continues as

$$\begin{aligned}&= \iint u_{s,4}(y,\delta,l,1)\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\frac{p_h(l)}{p_s(l)}\phi_s^{\dagger,\bar{\pi}}(y,\delta|l)p_s(y,\delta|l,D=1)p_s(l)dyd\delta dl \\ &= \sum_{\bar{a}_\tau,d=0}^{1}\iint u_{s,4}(y,\delta,l,1)\frac{I\{d=1\}}{p(d|l)}\frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2}\frac{p_h(l)}{p_s(l)}\phi_s^{\dagger,\bar{\pi}}(y,\delta|l)p_s(y,\delta|l,d)p_s(d|l)p_s(l)dyd\delta dl\end{aligned}$$

$$= \sum_{\bar{a}_\tau, d=0}^{1} \iint u_{s,4}(y,\delta,l,1) \left[ \frac{I\{d=1\}}{p(d|l)} \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2} \frac{p_h(l)}{p_s(l)} \phi_s^{\dagger,\bar{\pi}}(y,\delta|l) \right] p_s(y,\delta,l,d) dy d\delta dl.$$

Next, let the function $\phi_4^*$ be defined by

$$\phi_4^*(y,\delta,l,d) = \frac{I\{d=1\}}{p(d|l)} \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{(1-Q(\bar{\pi}_P,P,l))^2} \frac{p_h(l)}{p_s(l)} \phi_s^{\dagger,\bar{\pi}}(y,\delta|l),$$

and note that $E[u_{s,4}(Y_s,\Delta_s,L_s,1)|L_s=l,D=1]$ is equal to zero. Defining $\phi_4$ by $\phi_4(y,\delta,l,d) = \phi_4^*(y,\delta,l,d) - E[\phi_4^*(Y_s,\Delta_s,L_s,D)|L_s=l,D=1]$, the previous display continues as

$$= \sum_{d=0}^{1} \iint u_{s,4}(y,\delta,l,1)\phi_4(y,\delta,l,d) p_s(y,\delta,l,d) dy d\delta dl = \int u_{s,4}(y,\delta,l,1)\phi_4(y,\delta,l,d) dP.$$

Lastly, let $u_{h,5} \in L_0^2(P_{L_h,R=1})$. Thus, perturbing HOPE baseline characteristics, note that:

$$\begin{aligned}
\frac{d}{d\varepsilon}\Psi_{den}(P,P,P,P,P_\varepsilon)|_{\varepsilon=0} &= \int u_{h,5}(l,1) \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{1-Q(\bar{\pi}_P,P,l)} p_h(l|R=1) dl \\
&= \sum_{r=0}^{1} \int u_{h,5}(l,1) \frac{I\{r=1\}}{p_h(r)} \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{1-Q(\bar{\pi}_P,P,l)} p_h(l|r) p_h(r) dl.
\end{aligned}$$

Next, let $\phi_5^*(l,r) := \frac{I\{r=1\}}{p_h(r)} \frac{\{1-S_h(P,l)\}\{1-S_s(P,l)\}}{1-Q(\bar{\pi}_P,P,l)}$ and note that $E[u_{h,5}(L_h,1)|R=1] = 0$. Letting $\phi_5$ be defined by $\phi_5(l,r) = \phi_5(l,r) - \Psi_{den}(P)$, the previous display continues as:

$$= \int u_{h,5}(l,1)\phi_5(l,r) dP.$$

With each of the element-wise derivatives it is now possible to calculate:

$$\begin{aligned}
\frac{d}{d\varepsilon}\Psi_{den}^*(P_\varepsilon) = \frac{d}{d\varepsilon}\Psi_{den}(P_\varepsilon,P,P,P,P) + \frac{d}{d\varepsilon}\Psi_{den}(P,P_\varepsilon,P,P,P) + \frac{d}{d\varepsilon}\Psi_{den}(P,P,P_\varepsilon,P,P) \dots \\
+ \frac{d}{d\varepsilon}\Psi_{den}(P,P,P,P_\varepsilon,P) + \frac{d}{d\varepsilon}\Psi_{den}(P,P,P,P,P_\varepsilon) \\
= \int u_{h,1}(y,\delta,l,1)\phi_1(y,\delta,l,r) dP + \int u_{s,2}(y,\delta,l,0)\phi_2(y,\delta,l,d) dP + \int u_{h,3}(\bar{a}_\tau,l,1)\phi_3(\bar{a}_\tau,l,r) dP \dots \\
+ \int u_{s,4}(y,\delta,l,1)\phi_4(y,\delta,l,d) dP + \int u_{h,5}(l,1)\phi_5(l,r) dP.
\end{aligned}$$

Next, note that because each $\phi_i$ has a conditional mean of zero (when conditioning on the correct variable) the unique score function within each integral can be replaced by some $u \in L_0^2(P)$ without changing the value of the integral. Thus the above display continues as

$$\begin{aligned}
= \int u(y_h,\delta_h,\bar{a}_{\tau,h},l_h,1,y_s,\delta_s,l_s,d)[\phi_1(y_h,\delta_h,l_h,r) + \phi_2(y_s,\delta_s,l_s,d) + \phi_3(\bar{a}_{\tau,h},l_h,r) \\
+ \phi_4(y_s,\delta_s,l_s,d) + \phi_5(l_r,r)] dP.
\end{aligned}$$

Lastly, we use the delta method to derive the influence curve of the identified parameter. Note that the parameter can be written as:

$$\Psi : P \mapsto \frac{E\left[1 - S_h(P, L_h)\right]}{E_P\left[\frac{\{1-S_h(P,L_h)\}\{1-S_s(P,L_h)\}}{1-Q(\bar{\pi}_{h,P},P,L_h)}\right]} = \frac{E\left[1 - S_h(P, L_h)\right]}{\Psi^*_{den}(P)} = \frac{1 - \theta_h(P_{R=1}, \tau)}{\Psi^*_{den}(P)}.$$

Thus, the pathwise derivative is given by:

$$\frac{d}{d\varepsilon}\Psi(P_\varepsilon) = \frac{-\frac{d}{d\varepsilon}\theta_h(P_{R=1,\varepsilon}, \tau)\Psi^*_{den}(P) - [1 - \theta_h(P_{R=1}, \tau)]\frac{d}{d\varepsilon}\Psi^*_{den}(P_\varepsilon)}{\Psi^*_{den}(P_\varepsilon)^2}.$$

Letting $\phi_{den} = \phi_1 + \phi_2 + \phi_3 + \phi_4 + \phi_5$, it follows from the above formulation that the influence curve of the parameter is:

$$\frac{-\left[\phi_{h,\tau} - \theta_h(P, \tau)\right]\Psi^*_{den}(P) - [1 - \theta_j(P_{R=1}, \tau)]\phi_{den}}{\Psi^*_{den}(P_\varepsilon)^2}.$$

### 3.4.4 Computation

To compute the one-step estimator, it is necessary to calculate the influence curve of the parameter as well as the estimated distribution used for the plug-in. While much of the code for doing this is written from scratch, other software packages are also used at various intermediate steps. In particular, all survival-based parameters and influence functions are calculated using code from the R package CFsurvival which is the implementation of the work in Westling et al. (2021). This package was copied and modified for use in this setting to provide estimates of the conditional survival functions and their corresponding pseudo gradients. While estimates of the pseudo gradients of $S_h$ and $S_s$ are not usually accessible when using the functions provided by the CFsurvival package, they are calculated as an intermediate step to providing the influence function of $\theta_j$ and thus slight modifications of the CFsurvival code allow for calculation of these influence functions.

Apart from the survival functions and their corresponding influence functions, the influence curve is defined by two additional high-dimensional quantities. First is the ratio of the probability densities of the baseline characteristics in HOPE and ASPIRE: $p_h(l)/p_s(l)$. Here, the Bayes rule is used to simplify the estimation. Letting $H$ denote the random variable that is equal to 1 for individuals enrolled in HOPE and 0 for individuals during ASPIRE:

$$\begin{aligned}\frac{p_h(l)}{p_s(l)} &= \frac{f(l|H=1)}{f(l|H=0)} = \frac{\frac{\text{pr}(H=1|L=l)f(l)}{\text{pr}(H=1)}}{\frac{\text{pr}(H=0|L=l)f(l)}{\text{pr}(H=0)}} \\ &= \frac{\text{pr}(H=1|L=l)\text{pr}(H=0)}{\text{pr}(H=0|L=l)\text{pr}(H=1)} = \frac{\text{pr}(H=1|L=l)[1-\text{pr}(H=1)]}{[1-\text{pr}(H=1|L=l)]\text{pr}(H=1)}.\end{aligned}$$

Here, the ratio of the probability density ratio of the high dimensional outcome $L$ is estimated using

the conditional probability of trial given covariates done here using logistic regression. Second, estimation of $\pi_j$ is carried out using a sequence of logistic regression models in which adherence at each time point is predicted using adherence from previous time points and baseline covariates.

To estimate the parameter using the plugin estimator, each of the quantities in the identified parameter is estimated. Each of the survival quantities is estimated using the CFsurvival package, the adherence propensity function that defines $Q$ is estimated as described above, and the empirical is used for the distribution of the baseline covariates in HOPE. When estimating the parameter a current intermediate step is to estimate the counterfactual outcome under each fixed treatment regime. While the small number of adherence measurements in this chapter's setting (4) made computation of all combinations of adherence possible, as the number of adherence measurements grows, brute force calculation of the parameter estimate becomes infeasible. However, in settings with large numbers of adherence measurements, taking an average over a random sample of the possible adherence combinations allows for a good approximation of the estimate.

### 3.4.5 Numerical validation

To validate the described estimation strategy, a simulation study was conducted in a setting where the true data generating mechanism is known.

To mimic the observed data, each simulated dataset consists of two trials, with participants of the second trial consisting of a subset of those of the first. The baseline covariate in this study is a categorical variable with five different levels. Individuals' underlying propensities to adhere to treatment, have an event, enroll in the second trial after the first, or become censored vary across levels of the baseline covariate. Adherence at each time point is random, based on an individual's baseline covariate, the previous adherence, and exogenous noise. The data-generating mechanism is such that individuals are far more likely to adhere to treatment during the second trial. Data are simulated across 12 different settings, consisting of each combination of four sample sizes and three effect sizes. The number of participants in the first trial is either $500, 1000, 2000$, or 4000 (the number of participants in the second trial varies based on the number of events in the first trial and the number of eligible participants who choose to enroll in the second trial). The full details of the data generating mechanism can be found at `https://github.com/adam-s-elder/HA_da_simple/blob/master/sim_dat/simple_data_sim.R`.

The results of the simulation study suggest that the developed estimator and software package are working as expected. Confidence intervals across all settings are either very slightly (less than 1%) anti-conservative or conservative and the estimated standard error, on average, is roughly equal to the standard deviation of the estimator. While not perfectly monotone with sample size in each setting, bias tends to decrease with increased sample size.

| Sample Size | Estimate Average | Truth | Percent Bias | Coverage | Estimator SD | SE Estimator Average |
|---|---|---|---|---|---|---|
| **Null Model** | | | | | | |
| 500 | 1.07 | 1.00 | 8.11 | 94.91 | 0.23 | 0.22 |
| 1000 | 1.05 | 1.00 | 7.76 | 94.13 | 0.16 | 0.15 |
| 2000 | 1.03 | 1.00 | 5.26 | 95.19 | 0.11 | 0.11 |
| 4000 | 1.02 | 1.00 | 10.09 | 96.17 | 0.07 | 0.08 |
| **Smaller Effect** | | | | | | |
| 500 | 0.82 | 0.75 | 6.13 | 94.53 | 0.25 | 0.24 |
| 1000 | 0.78 | 0.75 | 2.47 | 97.09 | 0.16 | 0.17 |
| 2000 | 0.76 | 0.75 | 0.88 | 97.60 | 0.11 | 0.12 |
| 4000 | 0.76 | 0.75 | 0.10 | 97.40 | 0.08 | 0.08 |
| **Larger Effect** | | | | | | |
| 500 | 0.66 | 0.62 | 2.90 | 94.92 | 0.27 | 0.26 |
| 1000 | 0.61 | 0.62 | 0.36 | 97.70 | 0.16 | 0.18 |
| 2000 | 0.59 | 0.62 | 3.71 | 98.00 | 0.11 | 0.13 |
| 4000 | 0.59 | 0.62 | 14.10 | 98.47 | 0.07 | 0.09 |

Table 3.4: Summary of results from a simulation study conducted across a variety of effect sizes and sample sizes. The percent bias is $\text{bias}^2/\text{MSE}$. Coverage is the percentage of 95% confidence intervals that covered the true value. The estimator SD is the standard deviation of the sampling distribution of the parameter estimator. SE Estimator average is the mean of the sampling distribution of the standard error estimator.

## 3.5 Results

Based on the methodology described earlier, the estimated relative risk of HIV-1 infection within a year of starting HOPE is 0.83 (95% CI:0.36-1.30) of the HOPE study participants who received Dapivirine rings compared to if they had instead received placebo rings. While this estimate is based on a prespecified analysis plan, a subsequent sensitivity analysis revealed that two decisions played an important role in determining the value of the estimate. The first is the cutoff value for the level of Dapvirine that was required to have left the ring for an individual to be considered adherent. For the primary analysis in HOPE, and in our analysis as well, the cutoff value for adherence was set to be 0.9 mg. The second is the 12-month period in ASPIRE that was used to define our bridging assumption. Because it is necessary to compare the two trials on a single time scale (so that adherence measures are comparable) and because some ASPIRE trial participants stayed in the trial for longer than one year, it is necessary to select which one-year period to compare to. Choosing a different one-year period to calculate incidence results in a different bridging assumption and a different estimand. For our prespecified analysis, the starting month for ASPIRE was set to be month 0 (baseline).

While the communicated results from this study (Baeten et al., 2021) did not directly estimate open-label effectiveness, it did estimate an expected HIV-1 infection incidence of "4.4 per 100 person-years (3.2-5.8) among a population matched on age, site, and presence of a sexually transmitted infection from the placebo group of ASPIRE." The estimation strategy employed in this chapter provides an estimate of the counterfactual HOPE placebo arm HIV-1 infection inci-

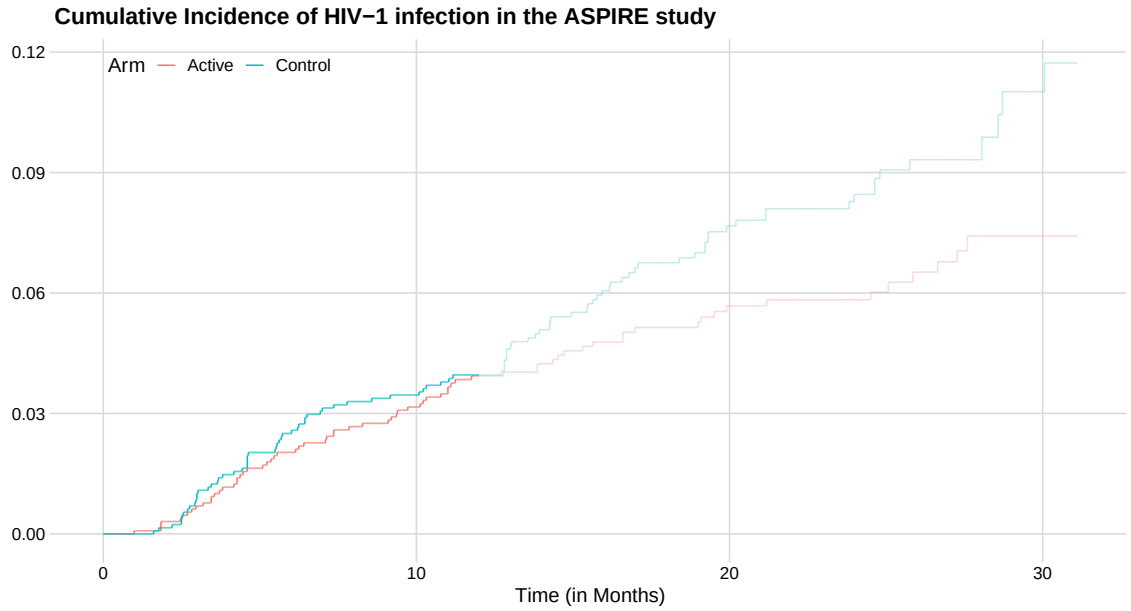


Figure 3.3: Cumulative incidence across during the entirety of the ASPIRE trial. When calculating the primary endpoint (HOPE open-label effectiveness), only the first 12 months are considered, and the remainder of the trial (the dimmed-out section) is ignored

dence of 5.0 per 100 person-years (1.2-8.7). This estimate is similar to that found in (Baeten et al., 2021), but has a noticeably wider confidence interval (which is not unsurprising considering this confidence interval accounts for differences in adherence between trials).

Because ASPIRE found the ring intervention to be effective and the ASPIRE trial saw an increase in adherence, it is somewhat surprising that the open-label effectiveness was not found to be significantly far away from a null effect. To understand one potential explanation for this observation, consider the denominator in (3.4). Note that if $\text{pr}\left(T_s^{\bar{\pi}^\star} \leq \tau | D = 0, L\right)$ is equal to $\text{pr}\left(T_s^{\bar{\pi}} \leq \tau | D = 1, L\right)$, then the parameter estimate will always be equal to one, regardless of what was observed in HOPE. Additionally, note that at 12 months after ASPIRE enrollment, there is little difference in estimated incidence between the two arms of the trial — this is in contrast to the differences between estimated incidences seen at nearly all other times post-enrollment (see Figure 3.3). This suggests that our estimate of effectiveness would be null for the HOPE trial, but there are at least two potential mechanisms by which the HOPE open-label effectiveness estimate could nevertheless have been found to be protective. First, it is possible that those who had the highest level of adherence during ASPIRE received a large benefit, and that during HOPE, adherence was higher across the board. Second, it is possible different subgroups of individuals in ASPIRE received different protective benefits from the ring and that the individuals who received large protective benefits from the ring were over-represented in the HOPE trial (relative to the ASPIRE trial).

As previously discussed two choices made in the prespecified analysis could have potentially been made differently. The first of these choices is the level of Dapivirine in a ring required for an individual to be considered adherent. The second is the period of time in the ASPIRE trial used to define the comparison group used in the bridging assumption. In Figure C.1 estimates and confidence intervals are given for different combinations of these alternative choices. Discussion of these results is provided in the appendix of this chapter.

## 3.6 Discussion

In this chapter, a method for estimating the open-label effectiveness in clinical trials with arm switching was developed. This technique accounts for differences in the population of a study shifting over time as well as population-level changes to the adherence observed during different parts of the study.

This method builds on techniques in causal inference and survival analysis and allows for non-parametric estimation of nuisance parameters that could otherwise result in bias of the parameter estimate. Subject area experts were consulted when deciding on the bridging assumption used to identify the parameter to make the least restrictive assumptions possible such that identification of the parameter of interest was still possible. While care was taken in the development of these assumptions, the validity of the analysis naturally rests on their validity.

Calculation of the estimate (and its confidence interval) was carried out using a combination of already-existing software and code written specifically for use in this setting. The validity of this software was supported by a small-scale simulation study. The calculated estimate and confidence interval find a beneficial, but non-significant effectiveness for the ring. In sensitivity analyses, it is seen that considering different 12-month periods of ASPIRE as the comparator group in estimation results in a large difference in the estimated effectiveness.

While we applied our method to a single dataset in this chapter, both the bridging assumption and the methodology applied could prove useful in other settings. As an example, a nearly identical approach to the one used here could be applied to the data from The Ring Study and its corresponding open-label extension, the DREAM study. The method outlined here could also be used to estimate open-label effectiveness for many open-label extension trials that were preceded by a placebo-control trial with a comparable population. While multiple approaches allow for combining data across studies, the method outlined in this chapter can account for differences in the study populations and differences in the adherence patterns between the two trials so long as measures of adherence are recorded during both studies.

# Bibliography

AIDS, U. (2021). Global HIV & AIDS statistics — Fact sheet. https://www.unaids.org/en/resources/fact-sheet.

Anderson, T. W. (1955). The Integral of a Symmetric Unimodal Function over a Symmetric Convex Set and Some Probability Inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176.

Baeten, J. M., Donnell, D., Ndase, P., Mugo, N. R., Campbell, J. D., Wangisi, J., Tappero, J. W., Bukusi, E. A., Cohen, C. R., Katabira, E., Ronald, A., Tumwesigye, E., Were, E., Fife, K. H., Kiarie, J., Farquhar, C., John-Stewart, G., Kakia, A., Odoyo, J., Mucunguzi, A., Nakku-Joloba, E., Twesigye, R., Ngure, K., Apaka, C., Tamooh, H., Gabona, F., Mujugira, A., Panteleeff, D., Thomas, K. K., Kidoguchi, L., Krows, M., Revall, J., Morrison, S., Haugen, H., Emmanuel-Ogier, M., Ondrejcek, L., Coombs, R. W., Frenkel, L., Hendrix, C., Bumpus, N. N., Bangsberg, D., Haberer, J. E., Stevens, W. S., Lingappa, J. R., and Celum, C. (2012). Antiretroviral Prophylaxis for HIV Prevention in Heterosexual Men and Women. *New England Journal of Medicine*, 367(5):399–410.

Baeten, J. M., Palanee-Phillips, T., Mgodi, N. M., Mayo, A. J., Szydlo, D. W., Ramjee, G., Gati Mirembe, B., Mhlanga, F., Hunidzarira, P., Mansoor, L. E., Siva, S., Govender, V., Makanani, B., Naidoo, L., Singh, N., Nair, G., Chinula, L., Parikh, U. M., Mellors, J. W., Balán, I. C., Ngure, K., van der Straten, A., Scheckter, R., Garcia, M., Peda, M., Patterson, K., Livant, E., Bunge, K., Singh, D., Jacobson, C., Jiao, Y., Hendrix, C. W., Chirenje, Z. M., Nakabiito, C., Taha, T. E., Jones, J., Torjesen, K., Nel, A., Rosenberg, Z., Soto-Torres, L. E., Hillier, S. L., Brown, E. R., and MTN-025/HOPE Study Team (2021). Safety, uptake, and use of a dapivirine vaginal ring for HIV-1 prevention in African women (HOPE): An open-label, extension study. *The lancet. HIV*, 8(2):e87–e95.

Bang, H. and Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973.

Bhatia, R. (2013). *Matrix Analysis*. Springer Science & Business Media. Google-Books-ID: lh4BCAAAQBAJ.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer New York. Google-Books-ID: lSnTm6SC_SMC.

Bloomfield, P. (2010). *Fourier Analysis of Time Series: An Introduction*. Wiley, New York.

Blumenson, L. E. (1960). A Derivation of n-Dimensional Spherical Coordinates. *The American Mathematical Monthly*, 67(1):63–66.

Borthwick, N., Ahmed, T., Ondondo, B., Hayes, P., Rose, A., Ebrahimsa, U., Hayton, E.-J., Black, A., Bridgeman, A., Rosario, M., Hill, A. V., Berrie, E., Moyle, S., Frahm, N., Cox, J., Colloca, S., Nicosia, A., Gilmour, J., McMichael, A. J., Dorrell, L., and Hanke, T. (2014). Vaccine-elicited Human T Cells Recognizing Conserved Protein Regions Inhibit HIV-1. *Molecular Therapy*, 22(2):464–475.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019a). Models as Approximations I: Consequences Illustrated with Linear Regression. *Statistical Science*, 34(4):523–544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., and Zhao, L. (2019b). Models as Approximations II: A Model-Free Theory of Parametric Regression. *Statistical Science*, 34(4):545–565.

Carone, M., Luedtke, A. R., and van der Laan, M. J. (2019). Toward Computerized Efficient Estimation in Infinite-Dimensional Models. *Journal of the American Statistical Association*, 114(527):1174–1190.

Choopanya, K., Martin, M., Suntharasamai, P., Sangkum, U., Mock, P. A., Leethochawalit, M., Chiamwongpaet, S., Kitisin, P., Natrujirote, P., Kittimunkong, S., Chuachoowong, R., Gvetadze, R. J., McNicholl, J. M., Paxton, L. A., Curlin, M. E., Hendrix, C. W., and Vanichseni, S. (2013). Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): A randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, 381(9883):2083–2090.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994.

Dudoit, S. and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. Springer-Verlag, New York.

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, 30(1):192–197.

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64.

Flandin, G. and Friston, K. J. (2019). Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, 40(7):2052–2054.

Fong, Y., Shen, X., Ashley, V. C., Deal, A., Seaton, K. E., Yu, C., Grant, S. P., Ferrari, G., deCamp, A. C., Bailer, R. T., Koup, R. A., Montefiori, D., Haynes, B. F., Sarzotti-Kelsoe, M., Graham, B. S., Carpp, L. N., Hammer, S. M., Sobieszczyk, M., Karuna, S., Swann, E., DeJesus, E., Mulligan, M., Frank, I., Buchbinder, S., Novak, R. M., McElrath, M. J., Kalams, S., Keefer, M., Frahm, N. A., Janes, H. E., Gilbert, P. B., and Tomaras, G. D. (2018). Modification of the Association Between T-Cell Immune Responses and Human Immunodeficiency Virus Type 1 Infection Risk by Vaccine-Induced Antibody Responses in the HVTN 505 Trial. *The Journal of Infectious Diseases*, 217(8):1280–1288.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22.

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32(4):361–369.

Gilbert, P. B., Blette, B. S., Shepherd, B. E., and Hudgens, M. G. (2020). Post-randomization Biomarker Effect Modification Analysis in an HIV Vaccine Clinical Trial. *Journal of Causal Inference*, 8(1):54–69.

Grant, R. M., Lama, J. R., Anderson, P. L., McMahan, V., Liu, A. Y., Vargas, L., Goicochea, P., Casapía, M., Guanira-Carranza, J. V., Ramirez-Cardich, M. E., Montoya-Herrera, O., Fernández, T., Veloso, V. G., Buchbinder, S. P., Chariyalertsak, S., Schechter, M., Bekker, L.-G., Mayer, K. H., Kallás, E. G., Amico, K. R., Mulligan, K., Bushman, L. R., Hance, R. J., Ganoza, C., Defechereux, P., Postle, B., Wang, F., McConnell, J. J., Zheng, J.-H., Lee, J., Rooney, J. F.,

Jaffe, H. S., Martinez, A. I., Burns, D. N., and Glidden, D. V. (2010). Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men. *New England Journal of Medicine*, 363(27):2587–2599.

Hensman, J., Durrande, N., and Solin, A. (2018). Variational Fourier Features for Gaussian Processes. *arXiv*, page 52.

Hernán MA and Robins JM (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Oxford University Press*, 75(2):383–386.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, 7(3):355–373.

Hudson, A., Carone, M., and Shojaie, A. (2021). Inference on function-valued parameters using a restricted score test. *arXiv:2105.06646 [stat]*.

Janes, H. E., Cohen, K. W., Frahm, N., De Rosa, S. C., Sanchez, B., Hural, J., Magaret, C. A., Karuna, S., Bentley, C., Gottardo, R., Finak, G., Grove, D., Shen, M., Graham, B. S., Koup, R. A., Mulligan, M. J., Koblin, B., Buchbinder, S. P., Keefer, M. C., Adams, E., Anude, C., Corey, L., Sobieszczyk, M., Hammer, S. M., Gilbert, P. B., and McElrath, M. J. (2017). Higher T-Cell Responses Induced by DNA/rAd5 HIV-1 Preventive Vaccine Are Associated With Lower HIV-1 Infection Risk in an Efficacy Trial. *The Journal of Infectious Diseases*, 215(9):1376–1385.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2020). The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*, 20(1):187–218.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer-Verlag, New York, third edition.

Lendle, S. D., Schwab, J., Petersen, M. L., and van der Laan, M. (2017). Ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software*, 81(1):1–21.

Liu, Y. and Xie, J. (2020). Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. *Journal of the American Statistical Association*, 115(529):393–402.

Loève, M. (1978). *Probability Theory II*. Number 46 in Graduate Texts in Mathematics. Springer.

McKeague, I. W. and Qian, M. (2015). An Adaptive Resampling Test for Detecting the Presence of Significant Predictors. *Journal of the American Statistical Association*, 110(512):1422–1433.

Neidich, S. D., Fong, Y., Li, S. S., Geraghty, D. E., Williamson, B. D., and et al. (2019). Antibody Fc effector functions and IgG3 associate with decreased HIV-1 risk. *The Journal of Clinical Investigation*, 129(11):4838–4849.

Nel, A., van Niekerk, N., Baelen, B. V., Malherbe, M., Mans, W., Carter, A., Steytler, J., van der Ryst, E., Craig, C., Louw, C., Gwetu, T., Mabude, Z., Kotze, P., Moraba, R., Tempelman, H., Gill, K., Kusemererwa, S., Bekker, L.-G., Devlin, B., and Rosenberg, Z. (2021). Safety, adherence, and HIV-1 seroconversion among women using the dapivirine vaginal ring (DREAM): An open-label, extension study. *The Lancet HIV*, 8(2):e77–e86.

Nel, A., van Niekerk, N., Kapiga, S., Bekker, L.-G., Gama, C., Gill, K., Kamali, A., Kotze, P., Louw, C., Mabude, Z., Miti, N., Kusemererwa, S., Tempelman, H., Carstens, H., Devlin, B., Isaacs, M., Malherbe, M., Mans, W., Nuttall, J., Russell, M., Ntshele, S., Smit, M., Solai, L., Spence, P., Steytler, J., Windle, K., Borremans, M., Resseler, S., Van Roey, J., Parys, W., Vangeneugden, T., Van Baelen, B., and Rosenberg, Z. (2016). Safety and Efficacy of a Dapivirine Vaginal Ring for HIV Prevention in Women. *New England Journal of Medicine*, 375(22):2133–2143.

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A Powerful and Adaptive Association Test for Rare Variants. *Genetics*, 197(4):1081–1095.

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Boston, 4th ed edition.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, second edition.

Pfanzagl, J. (1990). *Estimation in Semiparametric Models: Some Recent Developments*. Lecture Notes in Statistics. Springer-Verlag, New York.

Phillips, H. (2021). PrEP vaginal ring set for rollout after WHO endorsement.

Rees, H., Delany-Moretlwe, S., Lombard, C., Baron, D., Panchia, R., Myer, L., Schwartz, J., Doncel, G., and Gray, G. (2015). FACTS 001 Phase III Trial of Pericoital Tenofovir 1% Gel for HIV Prevention in Women.

Rinott, Y. (1976). On Convexity of Measures. *The Annals of Probability*, 4(6):1020–1026.

S. Holland, B. and DiPonzio Copenhaver, M. (1988). Improved Bonferroni-Type Multiple Testing Procedures. *Psychological Bulletin*, 104:145–149.

Shorack, G. R. (2017). *Probability for Statisticians*. Springer, 2nd edition edition.

Simes, J. R. (1986). An improved Bonferroni procedure for multiple tests of significance. *Oxford University Press*, 73(3):751–754.

Simon, N., Friedman, J., and Hastie, T. (2013). A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression. http://arxiv.org/abs/1311.6529.

Subba Rao, S. (2018). Statistical inference for spatial statistics defined in the Fourier domain. *The Annals of Statistics*, 46(2).

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 74(2):245–266.

Tong, Y. L. (2012). *The Multivariate Normal Distribution*. Springer Science & Business Media.

Van Damme, L., Corneli, A., Ahmed, K., Agot, K., Lombaard, J., Kapiga, S., Malahleha, M., Owino, F., Manongi, R., Onyango, J., Temu, L., Monedi, M. C., Mak'Oketch, P., Makanda, M., Reblin, I., Makatu, S. E., Saylor, L., Kiernan, H., Kirkendale, S., Wong, C., Grant, R., Kashuba, A., Nanda, K., Mandala, J., Fransen, K., Deese, J., Crucitti, T., Mastro, T. D., and Taylor, D. (2012). Preexposure Prophylaxis for HIV Infection among African Women. *New England Journal of Medicine*, 367(5):411–422.

van der Laan, M. and Dudoit, S. (2003). Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

van der Laan, M. and Gruber, S. (2011). Targeted Minimum Loss Based Estimation of an Intervention Specific Mean Outcome. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Westling, T. (2020). Nonparametric Tests of the Causal Null With Nondiscrete Exposures. *Journal of the American Statistical Association*, 0(0):1–12.

Westling, T. and Carone, M. (2020). A unified study of nonparametric inference for monotone functions. *The Annals of Statistics*, 48(2):1001–1024.

Westling, T., Luedtke, A., Gilbert, P., and Carone, M. (2021). Inference for treatment-specific survival curves using machine learning.

Whitney, D., Shojaie, A., and Carone, M. (2019). Comment: Models as (deliberate) approximations. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 34(4):591–598.

Xu, G., Lin, L., Wei, P., and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika*, 103(3):609–624.

Zhang, Y. and Laber, E. B. (2015). Comment. *Journal of the American Statistical Association*, 110(512):1451–1454.

Zygmund, A. (2002). *Trigonometric Series*. Cambridge University Press.

# Appendix A

# Chapter one appendix

## A.1 Technical lemmas

We first state and prove technical lemmas that will be used in the proof of Theorems 1, 2 and 3. The first lemma below indicates when regularity conditions on the individual local measures of test inefficiency in $\mathcal{F}_0$ imply corresponding conditions for the adaptive local measure of test inefficiency upon which our test is derived. This lemma serves as a fundamental building block in the proof of Theorems 1 and 2.

Throughout the Supplement, we define $\Gamma^{d,*} : (x, \Sigma) \mapsto \min_{\varphi \in \mathcal{F}_0} \Gamma(x, \Sigma, \varphi)$ and $\Gamma_0^{d,*} : x \mapsto \Gamma^*(x, \Sigma_0)$. Also, for any event $\mathcal{E}$ involving only the random variable $U_0$, we denote by $\mathrm{pr}(\mathcal{E})$ the evaluation of the probability of $\mathcal{E}$ under $U_0 \sim Q_0$. Finally, we refer to the following conditions:

C1*) $(u, \Sigma) \mapsto \Gamma^{d,*}(u, \Sigma)$ is continuous and non-negative on $\mathbb{R}^d \times B_0^*$ for some neighborhood $B_0^* \subset \mathbb{V}_d$ of $\Sigma_0$;

C2*) $\int I\{\Gamma_0^{d,*}(u) = t\}\, dQ_0(u) = 0$ for every $t \geq 0$;

C3*) for at least one $\varphi \in \mathcal{F}_0$, $\Gamma^{d,*}(x_s, \Sigma) \to 0$ uniformly over $\Sigma \in B_1^*$ for some neighborhood $B_1^* \subset \mathbb{V}_d$ of $\Sigma_0$ for every sequence $x_1, x_2, \ldots$ of elements of $\mathbb{R}^d$ such that $\varphi(x_s) \to \infty$;

C4*) $u \mapsto \Gamma_0^{d,*}(u)$ is quasi-concave, in the sense that $\{u : \Gamma_0^{d,*}(u) \geq a\}$ is convex for every $a \geq 0$;

C5*) $u \mapsto \Gamma_0^{d,*}(u)$ is centrally symmetric, in the sense that $\Gamma_0^{d,*}(u) = \Gamma_0^{d,*}(-u)$ for every $u \in \mathbb{R}^d$.

**Lemma 1.** *If any combination of C1, C2, C4 and C5 hold for every element of $\mathcal{F}_0$, then the respective combination of C1*, C2*, C4* and C5* hold as well. Additionally, if C3 holds for at least one element of $\mathcal{F}_0$, then C3* holds as well.*

*Proof of Lemma 1.* Suppose that C1 holds for each $\varphi \in \mathcal{F}_0$. Denote by $B_0(\varphi) \subset \mathbb{V}_d$ the neighborhood of $\Sigma_0$ over which C1 holds for $\varphi \in \mathcal{F}_0$. Because the minimum function is continuous and

the composition of continuous functions is also continuous, it follows that $\Gamma^{d,*}$ is continuous on $\mathbb{R}^d \times B_0^*$, where $B_0^* := \cap_{\varphi \in \mathcal{F}_0} B_0(\varphi) \subset \mathbb{V}_d$ is itself a neighborhood of $\Sigma_0$. Additionally, the minimum of non-negative values is necessarily non-negative. Thus, C1* holds. Suppose that C2 holds for each $\varphi \in \mathcal{F}_0$. Because $\text{pr}\{\Gamma_0^d(U_0, \varphi) = c\} = 0$ for each $c \in \mathbb{R}$ and $\varphi \in \mathcal{F}_0$, it follows that

$$\text{pr}\left\{\Gamma_0^{d,*}(U_0) = c\right\} \;\leq\; \text{pr}\left\{\bigcup_{\varphi \in \mathcal{F}_0} \left\{\Gamma_0^d(U_0, \varphi) = c\right\}\right\} \;\leq\; \sum_{\varphi \in \mathcal{F}_0} \text{pr}\left\{\Gamma_0^d(U_0^d, \varphi) = c\right\} \;=\; 0\ ,$$

establishing C2*. Suppose that C3 holds for $\varphi_0 \in \mathcal{F}_0$, and denote by $B_1^* \subset \mathbb{V}_d$ the neighborhood of $\Sigma_0$ over which C3 holds. By definition, we have that $0 \leq \Gamma^{d,*}(x_s, \Sigma) \leq \Gamma^d(x_s, \Sigma, \varphi_0)$ for each $\Sigma$, and so, $0 \leq \sup_{\Sigma \in B_1^*} \Gamma^{d,*}(x_s, \Sigma) \leq \sup_{\Sigma \in B_1^*} \Gamma^d(x_s, \Sigma, \varphi_0)$. This establishes C3* since $\sup_{\Sigma \in B_1^*} \Gamma^d(x_s, \Sigma, \varphi_0) \to 0$ by assumption. Suppose that C4 holds for each $\varphi \in \mathcal{F}_0$. Defining $M_\varphi(a) := \{u \in \mathbb{R}^d : \Gamma_0^d(u, \varphi) \geq a\}$ and $M^*(a) := \{u \in \mathbb{R}^d : \Gamma_0^{d,*}(u) \geq a\}$, we note that $M^*(a) = \cap_{\varphi \in \mathcal{F}_0} M_\varphi(a)$. By assumption, $M_\varphi(a)$ is convex for each $a \geq 0$ and $\varphi \in \mathcal{F}_0$. Since the intersection of a finite number of convex sets is convex, $M^*(a)$ is convex for each $a \geq 0$, thus proving C4*. Suppose that C5 holds for each $\varphi \in \mathcal{F}_0$. Because $u \mapsto \Gamma_0^d(u, \varphi)$ is centrally symmetric for each $\varphi \in \mathcal{F}_0$, we have that $\Gamma_0^{d,*}(-u) = \min_{\varphi \in \mathcal{F}_0} \Gamma_0^d(-u, \varphi) = \min_{\varphi \in \mathcal{F}_0} \Gamma_0^d(u, \varphi) = \Gamma_0^{d,*}(u)$ for each $u \in \mathbb{R}^d$, and so, C5* holds. □

The following lemmas establish technical properties for certain sets, functions and probability statements considered in the proof of Theorems 1 and 2.

**Lemma 2.** *The density function $f$ of the $d$-variate normal distribution with mean zero is quasi-concave, that is, the set $\{x \in \mathbb{R}^d : f(x) \geq \kappa\}$ is convex for each $\kappa \in \mathbb{R}$.*

*Proof of Lemma 2.* By Tong (2012), the $d$-variate normal probability density function is log-concave. All log-concave functions are quasi-concave in view of Section 3.5.1 of Boyd et al. (2004). □

**Lemma 3.** *Let $C$ be a convex subset of $\mathbb{R}^d$, and define $C_\mu := \{c + \mu : c \in C\}$ for $\mu \in \mathbb{R}^d$. For any $\mu_1, \mu_2 \in \mathbb{R}^d$ and $t \in [0, 1]$, the set $tC_{\mu_1} + (1-t)C_{\mu_2} = \{tc_1 + (1-t)c_2 : c_1 \in C_{\mu_1}, c_2 \in C_{\mu_2}\}$ is equal to $C_{t\mu_1 + (1-t)\mu_2}$.*

*Proof of Lemma 3.* Let $x \in tC_{\mu_1} + (1-t)C_{\mu_2}$, so that there exist $c_1, c_2 \in C$ such that $x = (c_1 + \mu_1)t + (c_2 + \mu_2)(1-t)$. Since we can rewrite $x = c_1 t + c_2(1-t) + \mu_1 t + \mu_2(1-t)$ with $c_1 t + c_2(1-t) \in C$ by the convexity of $C$, we have that $x \in C_{\mu_1 t + \mu_2(1-t)}$. Hence, we find that $tC_{\mu_1} + (1-t)C_{\mu_2} \subseteq C_{\mu_1 t + \mu_2(1-t)}$. To show the reverse inclusion, let $y \in C_{\mu_1 t + \mu_2(1-t)}$, so that there exists $c \in C$ such that $y = c + \mu_1 t + \mu_2(1-t) = (c + \mu_1)t + (c + \mu_2)(1-t)$. This implies that $y \in tC_{\mu_1} + (1-t)C_{\mu_2}$. Hence, we also find that $C_{\mu_1 t + \mu_2(1-t)} \subseteq tC_{\mu_1} + (1-t)C_{\mu_2}$. □

**Lemma 4.** *Suppose that $B$ is a closed, bounded and centrally symmetric subset of $\mathbb{R}^d$, and let $f$ denote the density function of the $d$-dimensional normal distribution with mean zero and positive definite covariance matrix. For any non-zero $h \in \mathbb{R}^d$, the function $g_h : \beta \mapsto \int_B f(t - \beta h)\, dt$ is strictly decreasing.*

*Proof of Lemma 4.* A minimizer $x_0 \in \operatorname{argmax}\{f(x) : x \in B\}$ exists because $B$ is closed and bounded and $f$ is continuous. We also have that $-x_0 \in \operatorname{argmax}\{f(x) : x \in B\}$ because $B$ and $f$ are both centrally symmetric. Let $\Sigma$ be the covariance matrix indexing $f$, and define $x_0^* := \operatorname{sign}(h^\top \Sigma^{-1} x_0) \cdot x_0 \in B$. In particular, we note that $x_0^* + h \in \{x + h : x \in B\}$. We also note that

$$\begin{aligned} 2[\log f(x_0^*) - \log f(x_0^* + h)] &= (x_0^* + h)^\top \Sigma^{-1}(x_0^* + h) - (x_0^*)^\top \Sigma^{-1} x_0^* = 2h^\top \Sigma^{-1} x_0^* + h^\top \Sigma^{-1} h \\ &= 2|h^\top \Sigma^{-1} x_0| + h^\top \Sigma^{-1} h \;>\; 0\,, \end{aligned}$$

and so, $f(x_0^*) > f(x_0^* + h)$ for $h \in \mathbb{R}^d \neq 0$. Set $u := [f(x_0^*) + f(x_0^* + h)]/2$ and note that $f(x_0^*) > u > f(x_0^* + h)$. This implies that $x_0^* + h$ is an element of $\{x + h : x \in B, f(x) > u\}$ but not of $\{x + h : x \in B\} \cap \{x : f(x) > u\}$. Thus, in view of Corollary 1 of Anderson (1955), $g_h$ is strictly decreasing. □

**Lemma 5.** *Let constants $b, c \in \mathbb{R}$ and strictly ray monotone function $g : \mathbb{R}^d \to \mathbb{R}$ be given. If $Z$ is a non-degenerate $d$-variate normal random variable with distribution $P$ and density function $f$, it holds that $P\{g(Z) = c\} = \int I\{g(z) = c\} f(z)\, dz = 0$.*

*Proof of Lemma 5.* We focus on $d \geq 3$ as the case $d \in \{1, 2\}$ is straightforward. To evaluate the integral $\int I\{g(z) = c\} f(z)\, dz$, we reparametrize $\mathbb{R}^d$ into $\{(r, \gamma_1, \gamma_2, \ldots, \gamma_{d-1}) : r \in [0, \infty), \gamma_1, \gamma_2, \ldots, \gamma_{d-2} \in [0, \pi), \gamma_{d-1} \in [0, 2\pi)\}$ as in Blumenson (1960), setting $z = rt(\gamma)$ with $t_1(\gamma) := \cos(\gamma_1)$, $t_j(\gamma) := \sin(\gamma_1)\sin(\gamma_2)\ldots\sin(\gamma_{j-1})\cos(\gamma_j)$ for $j = 2, 3, \ldots, d-1$, and $t_d(\gamma) := \sin(\gamma_1)\sin(\gamma_2)\ldots\sin(\gamma_{d-1})$. Here, we write $\gamma := (\gamma_1, \gamma_2, \ldots, \gamma_{d-1})$ and $t(\gamma) := (t_1(\gamma), t_2(\gamma), \ldots, t_d(\gamma))$. We also define $|J_d| := r^{d-1}\sin(\gamma_1)^{d-2}\sin(\gamma_2)^{d-3}\ldots\sin(\gamma_{d-2})$, and note that $a_\gamma := |J_d| r^{1-d}$ depends on $\gamma$ but not $r$. This change of variable allows us to write

$$\begin{aligned} \int I\{g(z) = c\} f(z)\, dz &= \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} \int_0^{\infty} I\left\{g\left(rt(\gamma)\right) = c\right\} f\left(rt(\gamma)\right) |J_d|\, dr\, d\gamma_1 \ldots d\gamma_{d-1} \\ &= \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} a_\gamma \left[\int_0^{\infty} r^{d-1} I\left\{g\left(rt(\gamma)\right) = c\right\} f\left(rt(\gamma)\right)\, dr\right] d\gamma_1 \ldots d\gamma_{d-1}\,, \end{aligned}$$

where we note that in the innermost integral $t(\gamma)$ is fixed when integrating over $r$. Thus, the latter integral has the form $\int_0^\infty r^{d-1} I\{g(rv) = c\} f(rv)\, dr$ for some $v \in \mathbb{R}^d$. Because $g$ is strictly ray monotone, the function $g_v : r \mapsto g(rv)$ is strictly monotone. Thus, the indicator function $I\{g(rv) = c\}$ can only equal one for a single value of $r$, and so, the innermost integral and thus

the probability of interest equal zero. □

## A.2 Proof of Theorems 1 and 2

We now prove Theorems 1 and 2. Below, we denote convergence in distribution by $\rightsquigarrow$. We refer to $\bar{U}_n$ as a random draw from the normal distribution with mean zero and covariance matrix $\Sigma_n$ independent of $X_1, X_2, \ldots, X_n$ given $\Sigma_n$. We define $\bar{A}_n^* := \Gamma^{d,*}(\bar{U}_n, \Sigma_n)$, $A_n^* := \Gamma^{d,*}(U_n, \Sigma_n)$ and $A_0^* := \Gamma^{d,*}(U_0, \Sigma_0)$, and denote by $q_n$ and $q_0$ the $\alpha$-quantile of $\bar{A}_n^*$ and $A_0^*$, respectively.

*Proof of Theorem 1.* In view of Lemma 1, if conditions C1*–C2* hold for each $\varphi \in \mathcal{F}_0$, then C1*–C2* hold, and if in addition C3 holds for some $\varphi \in \mathcal{F}_0$, then C3* holds as well. Since $\Sigma_n$ converges in probability to $\Sigma_0$, we have that $(\bar{U}_n, \Sigma_n)$ converges in distribution to $(U_0, \Sigma_0)$. In view of C1*, this implies that $\bar{A}_n^* \rightsquigarrow A_0^*$ by the continuous mapping theorem. By Lemma 21.2 of van der Vaart (2000), we have that $q_n \to q_0$ in probability since the distribution function of $A_0^*$ is continuous at $q_0$ by C2*. Suppose that $P_0 \in \mathcal{M}_0$. The asymptotic linearity of $\psi_n$ and consistency of $\Sigma_n$ imply that $(U_n, \Sigma_n) \rightsquigarrow (U_0, \Sigma_0)$, and so, $A_n^* \rightsquigarrow A_0^*$ by the continuous mapping theorem in view of C1*. By C2*, this implies that the type I error $P_0(A_n^* \leq q_n)$ tends to $P_0(A_0^* \leq q_0) = \alpha$.

Now, suppose instead that $P_0 \notin \mathcal{M}_0$. For any $\varepsilon > 0$, the rejection rate $P_0(A_n^* \leq q_n) = 1 - P_0(A_n^* > q_n)$ is bounded below by $1 - P_0(A_n^* \geq q_0 - \varepsilon) - P_0(|q_n - q_0| > \varepsilon)$. The term $P_0(|q_n - q_0| > \varepsilon)$ tends to zero in probability by the consistency of $q_n$ for $q_0$. Since $q_0 > 0$ by C2*, we can choose $\varepsilon \in (0, q_0)$ above, and for any such choice, we have that

$$\liminf_{n\to\infty} P_0(A_n^* \leq q_n) \geq 1 - \limsup_{n\to\infty} P_0(A_n^* \geq q_0 - \varepsilon) \ .$$

It then directly follows that the power $P_0(A_n^* \leq q_n)$ of the test tends to one provided $A_n^*$ tends to zero in probability. We thus show that the latter statement holds. First note that a sequence of random variables converges in probability if and only if each subsequence of this sequence contains a further subsequence that converges almost surely to the same limit (Shorack, 2017). Let $\varphi_0 \in \mathcal{F}_0$ be such that $\Gamma^{d,*}(x_s, \Sigma) \to 0$ uniformly over $\Sigma \in B^*$ for some neighborhood $B^*$ of $\Sigma_0$ for every sequence $x_1, x_2, \ldots$ of elements of $\mathbb{R}^d$ such that $\varphi_0(x_s) \to \infty$; the existence of $\varphi_0$ is guaranteed by C3*. By the reverse triangle inequality, we have that $\varphi_0(n^{1/2}\psi_n) \geq \varphi_0(n^{1/2}\psi_0) - \varphi_0(n^{1/2}(\psi_n - \psi_0)) = n^{1/2}\varphi_0(\psi_0) - O_P(1)$ in view of the fact that $n^{1/2}(\psi_n - \psi_0)$ is bounded in probability. As $n^{1/2}\varphi_0(\psi_0) \to +\infty$, this shows that $V_n := 1/\varphi_0(n^{1/2}\psi_n)$ tends to zero in probability. Let $V_{n_1}, V_{n_2}, \ldots$ with $1 \leq n_1 < n_2 < \ldots$ be an arbitrary subsequence of $V_1, V_2, \ldots$, and note that $V_{n_k}$ tends to zero in probability as $k \to \infty$. There must then exist a further subsequence $V_{n_{k_1}}, V_{n_{k_2}}, \ldots$ with $1 \leq k_1 < k_2 < \ldots$ that converges to zero almost surely, and so, defining

$U'_j := U_{n_{k_j}} = n_{k_j}^{1/2}\psi_{n_{k_j}}$, we have that $\varphi_0(U'_j)$ diverges almost surely as $j \to \infty$. Thus, it follows that $\Gamma^{d,*}(U'_j, \Sigma)$ converges to zero almost surely uniformly over $\Sigma \in B^*$. Since we have argued that every subsequence $\Gamma^{d,*}(U_{n_k}, \Sigma)$ has a further subsequence that converges to zero almost surely, we have shown that $\Gamma^{d,*}(U_n, \Sigma)$ tends to zero in probability uniformly over $\Sigma \in B^*$. For each $\varepsilon > 0$, we then have that

$$\begin{aligned} P_0(A_n^* > \varepsilon) &= P_0(A_n^* > \varepsilon, \Sigma_n \in B^*) + P_0(A_n^* > \varepsilon, \Sigma_n \notin B^*) \\ &\leq \sup_{\Sigma \in B^*} P_0\{\Gamma^{d,*}(U_n, \Sigma) > \varepsilon\} + P_0(\Sigma_n \notin B^*) \ , \end{aligned}$$

which implies the claim since the first and second summands tend to zero in view of C3* and the consistency of $\Sigma_n$, respectively. □

*Proof of Theorem 2.* In view of Lemma 1, the fact that conditions C1, C2, C4 and C5 hold for each $\varphi \in \mathcal{F}_0$ and that condition C3 holds for some $\varphi \in \mathcal{F}_0$ implies that C1*–C5* hold. Since $\Sigma_n$ is consistent and $\psi_n$ is regular and asymptotically linear, under any sequence $P_n^{(0)}$ of local alternatives, it holds that $U_n \rightsquigarrow U_0 + h$ for some $h \in \mathbb{R}^d$, and in view of C1*, $A_n^* = \Gamma^{d,*}(U_n, \Sigma_n) \rightsquigarrow \Gamma_0^{d,*}(U_0 + h)$. Also, in view of C1* and C2*, it follows that $\bar{A}_n^* \rightsquigarrow A_0^*$ and the distribution function of $A_0^*$ is continuous. Therefore, $q_n$ tends to $q_0$ in probability under this sequence of local alternatives. Lastly, $q_0$ is a continuity point of $\Gamma_0^{d,*}(U_0 + h)$ using a change of variables argument and C2*. Thus, it follows that $P_n^{(0)}(A_n^* \leq q_n) \to \mathrm{pr}\left\{\Gamma_0^{d,*}(U_0 + h) \leq q_0\right\}$. We define the function $g_h : \mathbb{R} \to \mathbb{R}$ pointwise as

$$g_h(\beta) := \mathrm{pr}\left\{\Gamma_0^{d,*}(U_0 + \beta h) \geq q_0\right\} = \int_{B_0} f_0(x - \beta h)\, dx \ , \tag{A.1}$$

where we define $B_0 := \{x \in \mathbb{R}^d : \Gamma_0^{d,*}(x) \geq q_0\}$ and denote by $f_0$ the density of the multivariate normal distribution with mean zero and covariance $\Sigma_0$. By Lemma 2, the probability density function of the multivariate normal distribution with mean zero is centrally symmetric and quasi-concave, and so, $g_h$ is non-increasing in view of Theorem 1 of Anderson (1955). Corollary 1 of Anderson (1955) states that $g_h$ is in fact strictly decreasing provided $\{\omega + h : \omega \in B_0, f_0(\omega) > u\} \neq \{\omega + h : \omega \in B_0\} \cap \{\omega + h : f_0(\omega) > u\}$. Lemma 4 indicates that this condition is satisfied if $B_0$ is closed, bounded and centrally symmetric. By C5*, $\Gamma_0^{d,*}$ is centrally symmetric, and so, $B_0$ is also centrally symmetric. Also, since $u \mapsto \Gamma_0^{d,*}(u)$ is continuous, it is also upper semicontinuous, and therefore, $B_0$ is closed. It remains to show that $B_0$ is bounded. Let $\varphi_0 \in \mathcal{F}_0$ be such that $\Gamma^{d,*}(x_s, \Sigma) \to 0$ uniformly over $\Sigma \in B^*$ for some neighborhood $B^*$ of $\Sigma_0$ for every sequence $x_1, x_2, \ldots$ of elements of $\mathbb{R}^d$ such that $\varphi_0(x_s) \to \infty$; the existence of $\varphi_0$ is guaranteed by C3*. Suppose that $B_0$ is not bounded, that is, for each $r = 1, 2, \ldots$, there exists some $v_r \in B_0$ for

which $\varphi_0(v_r) > r$. Because $\varphi_0(v_r) \to \infty$, it follows that $\sup_{\Sigma \in B^*} \Gamma^d(v_r, \Sigma, \varphi_0) \to 0$, and since $0 \leq \Gamma^{d,*}(v_r, \Sigma) \leq \Gamma^d(v_r, \Sigma, \varphi_0)$ for each $\Sigma \in B^*$, this also implies that $\sup_{\Sigma \in B^*} \Gamma^{d,*}(v_r, \Sigma) \to 0$. However, this is a contradiction since by definition $\sup_{\Sigma \in B^*} \Gamma^{d,*}(v_r, \Sigma) \geq \Gamma_0^{d,*}(v_r^d) \geq q_0$ for every $r$. Thus, no such sequence exists, and instead there exists some $r_0 > 0$ such that $\varphi_0(v) < r_0$ for every $v \in B_0$. Thus, $B_0$ must be bounded. It follows finally that $g_h$ is strictly decreasing, and so, the power of the proposed test under local alternatives tends to $1 - g_h(1) > 1 - g_h(0) = \mathrm{pr}\{\Gamma_0^{d,*}(U_0) < q_0\} = \alpha$. □

## A.3 Proof of Theorem 3

We now show that both local measures of test inefficiency discussed in this paper satisfy regularity conditions C1–C5 irrespective of the norm $\varphi$ used.

*Proof of Theorem 3.* **Part 1:** acceptance rate measure.

**C1.** Non-negativity is clear. To establish the continuity of $(x, \Sigma) \mapsto \Gamma^d_{\mathrm{ar}}(x, \Sigma, \varphi)$, we first show that $\zeta : (x, \Sigma, c) \mapsto \int I\{\varphi(t) < c\} f_\Sigma(t - x)\, dt$ is continuous, where $f_\Sigma$ is the density function of the $d$-dimensional normal distribution with mean zero and covariance matrix $\Sigma$. Fix $x_0 \in \mathbb{R}^d$, $\Sigma_0' \in \mathbb{V}_d$ sufficiently close to $\Sigma_0$ to ensure that it is invertible, and $c_0 \in \mathbb{R}$. Consider an arbitrary sequence $(x_1, \Sigma_1, c_1), (x_2, \Sigma_2, c_2), \ldots$ in $\mathbb{R}^d \times \mathbb{V}_d \times \mathbb{R}$ tending to $(x_0, \Sigma_0', c_0)$. Since the smallest eigenvalue of $\Sigma_j$ converges to that of $\Sigma_0'$, $\Sigma_j$ is invertible for all $j$ sufficiently large. Hence, without loss of generality, we suppose that $\Sigma_j$ is invertible for all $j$. By the triangle inequality, for any $j$, we have that

$$|\zeta(x_j, \Sigma_j, c_j) - \zeta(x_0, \Sigma_0', c_0)| \;\leq\; |\zeta(x_j, \Sigma_j, c_j) - \zeta(x_0, \Sigma_0', c_j)| + |\zeta(x_0, \Sigma_0', c_j) - \zeta(x_0, \Sigma_0', c_0)| \,. \tag{A.2}$$

We first show that $\sup_{c \in \mathbb{R}} |\zeta(x_j, \Sigma_j, c) - \zeta(x_0, \Sigma_0, c)| \to 0$, which implies that the first summand tends to zero. Let $T_1, T_2, \ldots$ be independent random vectors with $T_j$ following the $d$-dimensional normal distribution with mean $x_j$ and covariance matrix $\Sigma_j$, and let $T_0$ be an independent random vector following the $d$-dimensional normal distribution with mean $x_0$ and covariance matrix $\Sigma_0'$. Because the moment generating function of $T_j$ converges pointwise to that of $T_0$, we have that $T_j \rightsquigarrow T_0$, and by the continuity of norms, it follows that $\varphi(T_j) \rightsquigarrow \varphi(T_0)$. Hence, the distribution function $F_j$ of $\varphi(T_j)$ tends to the distribution function $F_0$ of $\varphi(T_0)$ at all continuity points of $F_0$. Because $T_j$ is a non-degenerate normal random vector and all norms are strictly ray increasing, $F_j$ is everywhere continuous for each $j$ in view of Lemma 5, so that $F_j(c) \to F_0(c)$ for each $c \in \mathbb{R}$. Moreover, by Lemma 2.11 in van der Vaart (2000), this convergence is uniform, that is, $\sup_{c \in \mathbb{R}} |F_j(c) - F_0(c)| \to 0$. Since the continuity of $F_j$ everywhere implies that $\zeta(x_j, \Sigma_j, c_0) = F_j(c_0)$ for each $j$, we find that

$\sup_{c\in\mathbb{R}} |\zeta(x_j, \Sigma_j, c) - \zeta(x_0, \Sigma_0, c)| \to 0$, as claimed. That the second summand in (A.2) also tends to zero follows from the fact that $\zeta(x_0, \Sigma_0, c_j) = F_0(c_j) \to F_0(c_0) = \zeta(x_0, \Sigma_0, c_0)$ since $c_0$ is necessarily a continuity point of $F_0$.

For each $\Sigma \in \mathbb{V}_d$, we define $c_\alpha(\Sigma) := \min\{c > 0 : \text{pr}\{\varphi(U) < c\} \geq 1 - \alpha\}$ with $U$ a multivariate normal random vector with mean zero and covariance $\Sigma$. We wish to show that $c_\alpha$ is continuous at $\Sigma_0$. We first note that $c_\alpha(\Sigma_j)$ is the $(1-\alpha)$-quantile of $\varphi(T_j)$ in the setting in which $x_1 = x_2 = \ldots = 0$. Since we have already shown that the distribution function of $\varphi(T_j)$ converges uniformly to that of $\varphi(T_0)$, it follows from Lemma 21.2 of van der Vaart (2000) that the quantile function $F_j^{-1}$ of $\varphi(T_j)$ converges to the quantile function $F_0^{-1}$ of $\varphi(T_0)$. Thus, we have that $c_\alpha(\Sigma_j) = F_j^{-1}(1-\alpha) \to F_0^{-1}(1-\alpha) = c_\alpha(\Sigma_0')$, thereby establishing that $c_\alpha$ is a continuous function in a neighborhood of $\Sigma_0$. Since $\Gamma_{\text{ar}}^d(x, \Sigma, \varphi) = \zeta(x, \Sigma, c_\alpha(\Sigma))$ for each $(x, \Sigma)$, $\Gamma_{\text{ar}}^d$ is a composition of continuous functions, thereby implying C1.

**C2.** Fix $x \in \mathbb{R}^d$ and define $g_x : \beta \mapsto \int_W f_{\Sigma_0}(t - \beta x)\, dt$ with $W := \{t \in \mathbb{R}^d : \varphi(t) \leq c_\alpha(\Sigma_0)\}$, so that $g_x(\beta) = \Gamma_{\text{ar}}^d(\beta x, \Sigma_0, \varphi)$. In view of from Lemma 4, $g_x$ is strictly decreasing provided $W$ is closed, bounded and centrally symmetric. Because $\varphi$ is a norm, it is centrally symmetric, and thus, so is $W$. Moreover, the hypograph $\{(t, c) : \varphi(t) \leq c\}$ is closed and therefore upper semicontinuous by the continuity of $\varphi$. This, in turn, implies that $W$ is closed. Finally, we can show that $W$ is bounded similarly as was done for the set $B_0$ in the proof of Theorem 2. Since this establishes that $x \mapsto \Gamma^d(x, \Sigma_0, \varphi)$ is ray-decreasing, we find that $\text{pr}\left\{\Gamma_{\text{ar}}^d(U, \Sigma_0, \varphi) = c\right\} = 0$ for every $c \in \mathbb{R}$ by Lemma 5.

**C3.** For any sequence $x_1, x_2, \ldots$ of elements in $\mathbb{R}^d$ with $\varphi(x_s) \to \infty$, we have that

$$\begin{aligned}\Gamma_{\text{ar}}^d(x_s, \Sigma_0, \varphi) &= \text{pr}\left\{\varphi(U_0 + x_s) < c_\alpha(\Sigma_0)\right\} \\ &\leq \text{pr}\left\{\varphi(U_0) + \varphi(x_s) < c_\alpha(\Sigma_0)\right\} = 1 - \text{pr}\left\{c_\alpha(\Sigma_0) - \varphi(U_0) \leq \varphi(x_s)\right\}\end{aligned}$$

by the triangle inequality. Because the random variable $c_\alpha(\Sigma_0) - \varphi(U_0)$ is bounded in probability, it follows from the above inequality that $\Gamma_{\text{ar}}^d(x_s^d, \Sigma_0, \varphi)$ tends to zero since $\varphi(x_s) \to \infty$.

Now, suppose that there is no $\varepsilon > 0$ over which, for every sequence $x_s$ for which $\varphi(x_s) \to \infty$, $\Gamma_{\text{ar}}^d(x_s, \Sigma, \varphi) \to 0$ uniformly over all $\Sigma$ in a neighborhood of $\Sigma_0$. There must then exist some $\delta > 0$ and sequences $\Sigma_1, \Sigma_2, \ldots$ and $x_1, x_2, \ldots$ such that $\varphi(x_s) \to \infty$ and $\Sigma_s \to \Sigma$ but $\Gamma_{\text{ar}}^d(x_s, \Sigma_s, \varphi) > \delta$ for every $s$. By the continuity of $\varphi$ and $c_\alpha$, we have that $c_\alpha(\Sigma_s) - \varphi(U_s) \rightsquigarrow c_\alpha(\Sigma_0) - \varphi(U_0)$, where $U_0, U_1, U_2, \ldots$ is a sequence of independent random $d$-vectors with $U_s$ following the multivariate normal distribution with mean zero and covariance $\Sigma_s$. By Lemma 2.11 of van der Vaart (2000), this implies the uniform convergence of the corresponding distribution functions, and so, it follows

that

$$\begin{aligned}&\left|\text{pr}\left\{c_\alpha(\Sigma_0) - \varphi(U_0) \leq \varphi(x_s)\right\} - \text{pr}\left\{c_\alpha(\Sigma_s) - \varphi(U_s) \leq \varphi(x_s)\right\}\right| \\ &\qquad \leq \sup_x \left|\text{pr}\left\{c_\alpha(\Sigma_0) - \varphi(U_0) \leq x\right\} - \text{pr}\left\{c_\alpha(\Sigma_s) - \varphi(U_s) \leq x\right\}\right| \to 0\end{aligned}$$

Since we have already established above that $\text{pr}\left\{c_\alpha(\Sigma_0) - \varphi(U_0) \leq \varphi(x_s)\right\} \to 1$, it must then also be that $\text{pr}\left\{c_\alpha(\Sigma_s) - \varphi(U_s) \leq \varphi(x_s)\right\} \to 1$, and so, $\Gamma^d_{\text{ar}}(x_s, \Sigma_s, \varphi) \leq 1 - \text{pr}\left\{c_\alpha(\Sigma_s) - \varphi(U_s) \leq \varphi(x_s)\right\} \to 0$. This is a contradiction. As such, there must exist some neighborhood of $\Sigma_0$ such that the convergence of $\Gamma^d_{\text{ar}}(x^d_s, \Sigma, \varphi)$ to zero is uniform over $\Sigma$ in this neighborhood.

**C4.** Let $x \in \mathbb{R}^d$ be given. Defining $A_x := \{\omega \in \mathbb{R}^d : \varphi(\omega + x) < c_\alpha(\Sigma_0)\}$, we note that

$$\begin{aligned}\Gamma^d_{\text{ar}}(x, \Sigma_0, \varphi) &= \int I\left\{\varphi(t) < c_\alpha(\Sigma_0)\right\} f_{\Sigma_0}(t - x)\, dt \\ &= \int I\left\{\varphi(u + x) < c_\alpha(\Sigma_0)\right\} f_{\Sigma_0}(u)\, du \;=\; \text{pr}(U_0 \in A_x)\ .\end{aligned}$$

Suppose that $x_1, x_2 \in \mathbb{R}^d$ are such that $\Gamma^d_{\text{ar}}(x_1, \Sigma_0, \varphi) \geq c$ and $\Gamma^d_{\text{ar}}(x_2, \Sigma_0, \varphi) \geq c$. Then, we can write that $c = c^t c^{1-t} \leq \Gamma^d_{\text{ar}}(x^d_1, \Sigma_0, \varphi)^t \Gamma^d_{\text{ar}}(x_2, \Sigma_0, \varphi)^{1-t}$. Theorem 1 of Rinott (1976) states that $\nu(A_y)^t \nu(A_z)^{1-t} \leq \nu(tA_y + (1-t)A_z)$ for any distribution $\nu$ with log-concave density function, where $tA_y + (1-t)A_z := \{t\omega_1 + (1-t)\omega_2 : \omega_1 \in A_y, \omega_2 \in A_z, t \in [0,1]\}$. The multivariate normal distribution has a log-concave density, as shown, for example (see, e.g., Theorem 4.2.1 of Tong, 2012), and so, it holds that $\text{pr}(U_0 \in A_{x_1})^t \text{pr}(U_0 \in A_{x_2})^{1-t} \leq \text{pr}\{U_0 \in tA_{x_1} + (1-t)A_{x_2}\}$. It remains to show that $\text{pr}\{U_0 \in tA_{x_1} + (1-t)A_{x_2}\} = \Gamma^d_{\text{ar}}(tx_1 + (1-t)x_2, \Sigma_0, \varphi) = \text{pr}\{U_0 \in A_{tx_1 + (1-t)x_2}\}$. This is implied by Lemma 3 and the fact that each $A_x$ is convex by the convexity of norms, since this lemma shows that $tA_{x_1} + (1-t)A_{x_2} = A_{tx_1 + (1-t)x_2}$. Thus, we obtain that

$$c \;\leq\; \text{pr}\left\{U_0 \in tA_{x_1} + (1-t)A_{x_2}\right\} \;=\; \text{pr}\left\{U_0 \in A_{tx_1 + (1-t)x_2}\right\} \;=\; \Gamma^d_{\text{ar}}(tx_1 + (1-t)x_2, \Sigma_0, \varphi)\ .$$

Thus, we have established that $x \mapsto \Gamma^d_{\text{ar}}(x, \Sigma^d_0, \varphi)$ is quasi-concave.

**C5.** In view of the facts that $U_0$ and $-U_0$ have the same distribution and that $\varphi$ is centrally symmetric, for any $x \in \mathbb{R}^d$, we have that

$$\begin{aligned}\Gamma^d_{\text{ar}}(x, \Sigma_0, \varphi) \;&=\; \text{pr}\left\{\varphi(U_0 + x) < c_\alpha(\Sigma_0)\right\} \;=\; \text{pr}\left\{\varphi(-U_0 + x) < c_\alpha(\Sigma_0)\right\} \\ &= \text{pr}\left\{\varphi(U_0 - x) < c_\alpha(\Sigma_0)\right\} \;=\; \Gamma^d_{\text{ar}}(-x, \Sigma_0, \varphi)\ .\end{aligned}$$

**Part 2:** multiplicative factor measure.

**C1.** Again, non-negativity is clear. For $(x, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}$, define $\Lambda_{x,\Sigma} : \mathbb{R}^+ \to (0, 1-\alpha]$ pointwise as $\Lambda_{x,\Sigma}(s) := \Gamma^d_{\text{ar}}(sx, \Sigma, \varphi)$. Since $x \mapsto \Gamma^d_{\text{ar}}(x, \Sigma, \varphi)$ is continuous and strictly ray-decreasing,

$s \mapsto \Lambda^d_{x,\Sigma}(s)$ is also continuous and strictly decreasing. We note that $\Lambda_{x,\Sigma}(0) = 1-\alpha > \tau$ and $\lim_{s\to\infty} \Lambda_{x,\Sigma}(s) = 0$, and therefore, $\Gamma^d_{\mathrm{mf}}(x,\Sigma,\varphi) = \min\{s \geq 0 : \Lambda_{x,\Sigma}(s) \leq \tau\}$ equals the inverse $\Lambda^{-1}_{x,\Sigma}(\tau)$ of $\Lambda_{x,\Sigma}$ at $\tau$.

Let sequences $x_1, x_2, \ldots \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2, \ldots \in \mathbb{V}^d$ such that $(x_k, \Sigma_k) \to (x, \Sigma)$ be given, and denote $\Lambda_k := \Lambda_{x_k,\Sigma_k}$ for each $k$ and $\Lambda := \Lambda_{x,\Sigma}$. The continuity of $x \mapsto \Gamma^d_{\mathrm{ar}}(x,\Sigma,\varphi)$ implies that $\Lambda_k(s) \to \Lambda(s)$ for each $s$. In view of the continuity and monotonicity of the bounded functions $\Lambda_1, \Lambda_2, \ldots$ and $\Lambda$, an adaptation of arguments used to prove Lemma 2.11 of van der Vaart (2000) can be used to show that $\sup_{s\geq 0} |\Lambda_k(s) - \Lambda(s)| \to 0$. We prove by contradiction that $\Lambda_k^{-1}(\tau) \to \Lambda^{-1}(\tau)$. Suppose this is not so. Then, there exists $\epsilon > 0$ and natural numbers $k_1 < k_2 < \ldots$ such that (i) $\inf_j[\Lambda^{-1}_{k_j}(\tau) - \Lambda^{-1}(\tau)] \geq \epsilon$ or (ii) $\inf_j[\Lambda^{-1}(\tau) - \Lambda^{-1}_{k_j}(\tau)] \geq \epsilon$. Suppose that (i) holds. By the monotonicity of $\Lambda_k$, we have that $\tau < \Lambda_{k_j}(\Lambda^{-1}(\tau) + \epsilon)$ for all $j$, and so,

$$\begin{aligned} \tau &< \Lambda(\Lambda^{-1}(\tau)+\epsilon) + \Lambda_{k_j}(\Lambda^{-1}(\tau)+\epsilon) - \Lambda(\Lambda^{-1}(\tau)+\epsilon) \\ &\leq \Lambda(\Lambda^{-1}(\tau)+\epsilon) + \sup_{s\geq 0} |\Lambda_{k_j}(s) - \Lambda(s)| \ . \end{aligned}$$

As $\Lambda$ is strictly decreasing, $\Lambda(\Lambda^{-1}(\tau)+\epsilon) < \tau$. This yields a contradiction since the latter summand has been shown to tend to zero. A similar argument can be made if (ii) holds instead. We have thus shown that $x \mapsto \Gamma^d_{\mathrm{mf}}(x,\Sigma,\varphi)$ is continuous.

**C2.** Let $v \in \mathbb{R}^d$ be given, and define $g_v : \mathbb{R} \to \mathbb{R}$ pointwise as

$$g_v(\beta) := \Gamma^d_{\mathrm{mf}}(\beta v, \Sigma_0^d, \varphi) = \min\{s \geq 0 : \mathrm{pr}\{\varphi(U_0 + s\beta v) \geq c_\alpha(\Sigma_0)\} \geq 1-\tau\} \ .$$

We note that $g_v(\beta) = g_v(1)/\beta$, and so, $x \mapsto \Gamma^d_{\mathrm{mf}}(x,\Sigma_0,\varphi)$ is strictly ray-decreasing. Hence, the conditions of Lemma 5 are satisfied, and it follows that $\mathrm{pr}\{\Gamma^d_{\mathrm{mf}}(U_0,\Sigma_0,\varphi) = c\} = 0$ for each $c \geq 0$.

**C3.** Let a sequence $x_1, x_2, \cdots \in \mathbb{R}^d$ such that $\varphi(x_x) \to \infty$ be given. Let $\varepsilon > 0$ be given, and set $\tilde{x}_j := \varepsilon x_j$ for each $j$. The sequence $\tilde{x}_1, \tilde{x}_2, \ldots$ also has the property that $\varphi(\tilde{x}_j) = \varphi(\varepsilon x_j) = \varepsilon\varphi(x_j) \to \infty$. Using condition C3 established in Part 1, there exists some $N > 0$ and a neighborhood $B_0$ of $\Sigma_0$ such that $\Gamma^d_{\mathrm{ar}}(\tilde{x}_j,\Sigma,\varphi) \leq \tau$ for each $n > N$ and $\Sigma \in B_0$. As $\Gamma^d_{\mathrm{ar}}(\tilde{x}^d_j,\Sigma,\varphi) \to 0$ and $\Gamma^d_{\mathrm{mf}}(x_j,\Sigma,\varphi)$ is defined as the smallest $s$ such that $\Gamma^d_{\mathrm{ar}}(sx_j,\Sigma,\varphi) \leq \tau$, it follows that $\limsup_j \Gamma^d_{\mathrm{mf}}(x_j,\Sigma,\varphi) \leq \varepsilon$ uniformly over $\Sigma \in B_0$. Since $\varepsilon > 0$ is arbitrary, it must be the case that $\Gamma^d_{\mathrm{mf}}(x^d_j,\Sigma,\varphi) \to 0$ uniformly over $\Sigma \in B_0$.

**C4.** Suppose that $x_1, x_2 \in \mathbb{R}^d$ are such that $\Gamma^d_{\mathrm{mf}}(x_1,\Sigma_0,\varphi) \geq c$ and $\Gamma^d_{\mathrm{mf}}(x_2,\Sigma_0,\varphi) \geq c$. When establishing condition C2 in Part 1, it was shown that $s \mapsto \mathrm{pr}\{\varphi(U_0+sx) \geq c_\alpha(\Sigma_0)\}$ is continuous and strictly increasing. Hence, if $\Gamma^d_{\mathrm{mf}}(x,\Sigma_0,\varphi) \geq x$, then $\mathrm{pr}\{\varphi(U_0+cx) \geq c_\alpha(\Sigma_0)\} \geq 1-\tau$, which implies that $\mathrm{pr}\{\varphi(U_0+cx_1) < c_\alpha(\Sigma_0)\} > \tau$ and $\mathrm{pr}\{\varphi(U_0+cx_2) < c_\alpha(\Sigma_0)\} > \tau$. Using condition C4 established in Part 1, we find that $\mathrm{pr}\{\varphi(U_0+c\{tx_1+(1-t)x_2\}) < c_\alpha(\Sigma_0)\} > \tau$ or, equivalently,

$\text{pr}\{\varphi(U_0+c\{tx_1+(1-t)x_2\}) \geq c_\alpha(\Sigma_0)\} \leq 1-\tau$. Thus, it follows that $\Gamma^d_{\text{mf}}(tx^d_1+(1-t)x_2, \Sigma_0, \varphi) \geq c$, and so, $x \mapsto \Gamma^d_{\text{mf}}(x, \Sigma_0, \varphi)$ is quasi-concave.

**C5.** Using the fact that the density function of a mean-zero multivariate normal distribution is centrally symmetric, we have that

$$\begin{aligned}\Gamma^d_{\text{mf}}(x, \Sigma_0) &= \min\{s \geq 0 : \text{pr}\{\varphi(U_0 + sx) \geq c_\alpha(\Sigma_0)\} \geq 1-\tau\} \\ &= \min\{s \geq 0 : \text{pr}\{\varphi(-U_0 + sx) \geq c_\alpha(\Sigma_0)\} \geq 1-\tau\} \\ &= \min\{s \geq 0 : \text{pr}\{\varphi(U_0 - sx) \geq c_\alpha(\Sigma_0)\} \geq 1-\tau\} \;=\; \Gamma^d_{\text{mf}}(-x, \Sigma_0)\end{aligned}$$

for each $x \in \mathbb{R}^d$, thereby establishing that $x \mapsto \Gamma^d_{\text{mf}}(x, \Sigma^d_0, \varphi)$ is centrally symmetric. □

## A.4 Additional technical lemma

The sum-of-squares function $\jmath_k : \mathbb{R}^d \to \mathbb{R}$ is defined as $x \mapsto \sqrt{\sum_{i=1}^k x^2_{(d-i+1)}}$, where $x^2_{(j)}$ is the $j^{th}$ order statistic of components of $x$.

**Lemma 6.** *The function $\jmath_k$ is a norm for each $k \in \{1, 2, \ldots, d\}$.*

*Proof of Lemma 6.* Fix $k \in \{1, 2, \ldots, d\}$. We must show that $\jmath_k$ is point-separating, absolutely homogeneous and additive, which then implies the claim. First, we note that if $\jmath_k(x) = 0$, then it must be that $0 \leq x^2_{(1)} \leq \ldots \leq x^2_{(d)} \leq \sum_{i=1}^k x^2_{(d-i+1)} = 0$, and so, $x_1 = x_2 = \ldots = x_d = 0$. Second, we note that, for any $a > 0$ and $x \in \mathbb{R}^d$,

$$\jmath_k(ax) = \sqrt{\sum_{i=1}^k \{ax_{(d-i+1)}\}^2} = \sqrt{\sum_{i=1}^k a^2 x^2_{(d-i+1)}} = a\sqrt{\sum_{i=1}^k x^2_{(d-i+1)}} = a \cdot \jmath_k(x) \;.$$

Finally, we let $x = (x_1, x_2, \ldots, x_d)^\top$ and $y = (y_1, y_2, \ldots, y_d)^\top$ be elements of $\mathbb{R}^d$, and define $z := x + y$. Without loss of generality, suppose that $|z_1| \leq |z_2| \leq \ldots \leq |z_d|$. Then, we have that

$$\begin{aligned}\jmath_k(z) &= \sqrt{\sum_{i=1}^k z^2_{(d-i+1)}} \;=\; \sqrt{\sum_{i=1}^k z^2_{d-i+1}} \;=\; \sqrt{\sum_{i=1}^k (x_{d-i+1} + y_{d-i+1})^2} \\ &\leq \sqrt{\sum_{i=1}^k x^2_{d-i+1}} + \sqrt{\sum_{i=1}^k y^2_{d-i+1}} \;\leq\; \sqrt{\sum_{i=1}^k x^2_{(d-i+1)}} + \sqrt{\sum_{i=1}^k y^2_{(d-i+1)}} \;=\; \jmath_k(x) + \jmath_k(y) \;,\end{aligned}$$

where the first inequality follows from the subaddativity of the $\ell_2$ norm on $\mathbb{R}^k$. □

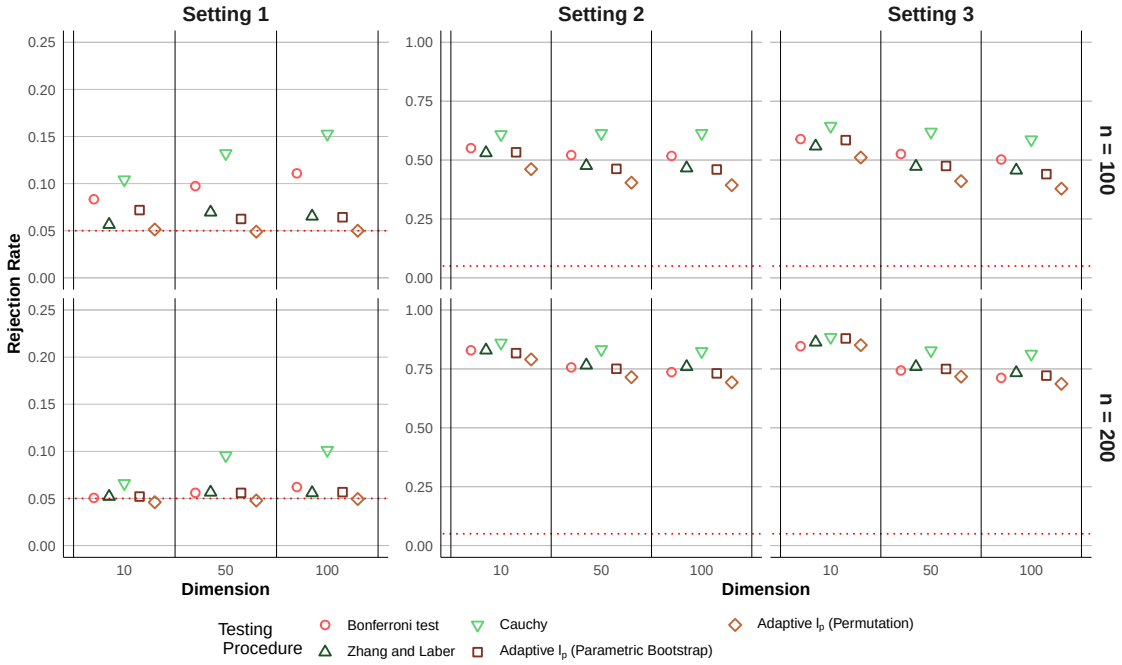


Figure A.1: Simulation study-based empirical rejection rate of various tests applicable in Example 1 under different data- generating mechanisms, at different sample sizes, and for covariate vectors with moderate correlation (50%) across components and of different length.

## A.5 Additional figures

## A.6 Additional information on data analysis

During the HVTN 505 clinical trial, vaccinations were administered at Months 0, 1 and 6. To measure the immune response to vaccination, samples were collected from individuals at Month 7. After trial completion, samples were analyzed for 25 primary endpoint vaccine cases (HIV-1 infected between Month 7 and Month 24) and 125 randomly sampled frequency matched vaccine controls (HIV-1 negative at Month 24) (Janes et al., 2017). Baseline covariates and infection status were recorded for all study participants.

We denote the vector of recorded immune response biomarkers as $S := (S_1, S_2, \ldots, S_d)$, and denote by $W$ and $Y$ the baseline covariate vector and infection status, respectively. The biomarker vector $S$ is only recorded on a subset of participants, and the variable $\Delta$ indicates those patients, with $\Delta = 1$ if $S$ is recorded and 0 otherwise. For each group of biomarkers considered in Neidich et al. (2019), we test the null hypothesis that these biomarkers are not associated with risk of infection. The measure of association used for each biomarker $S_j$ is the $\beta_1$-coefficient value indexing the KL projection of the conditional log-odds of infection $\log(\text{odds}(Y = 1|S_j = s))$ onto a linear working model $\beta_1 + \beta_2 s$.

Denoting by $P_F$ a candidate distribution for the full-data unit $(W, S, Y)$, we first define the

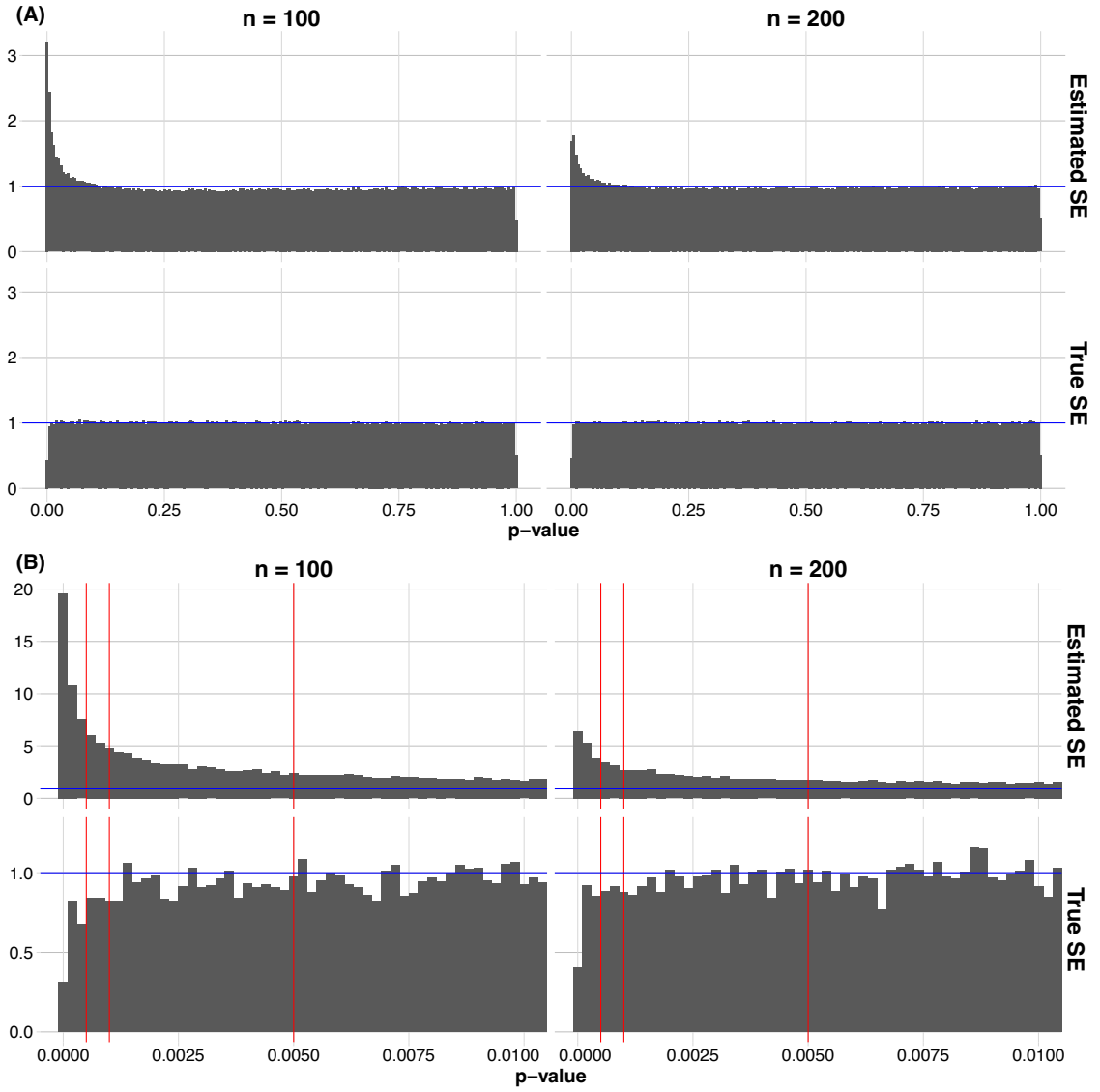


Figure A.2: Simulation-based empirical sampling density of the $p$-value $p_{1n}(\sigma) := 2[1 - \Phi(n^{1/2}|\psi_{n1}|/\sigma)]$ for $\sigma$ equal to either the true asymptotic standard deviation $\sigma_0$ or its influence function-based estimator $\sigma_n$ based on Example 1. Data are generated from the setting in which all covariates are independent of each other and the outcome. Panel (A) shows sampling densities on $[0, 1]$. Panel (B) shows the same densities but is restricted to the interval $[0, 0.01]$. In each panel, displays in the top and bottom rows show, respectively, the sampling density when $\sigma_0$ is estimated or known. Displays in the left and right columns show, respectively, results for $n = 100$ or $n = 200$. The blue horizontal line represents the theoretical standard uniform density of $p$-values under the null, and the red vertical lines (left to right) in Panel (B) are the largest single covariate $p$-value that results in rejection of the Bonferroni test for dimension $d$ equal to 100, 50 and 10.

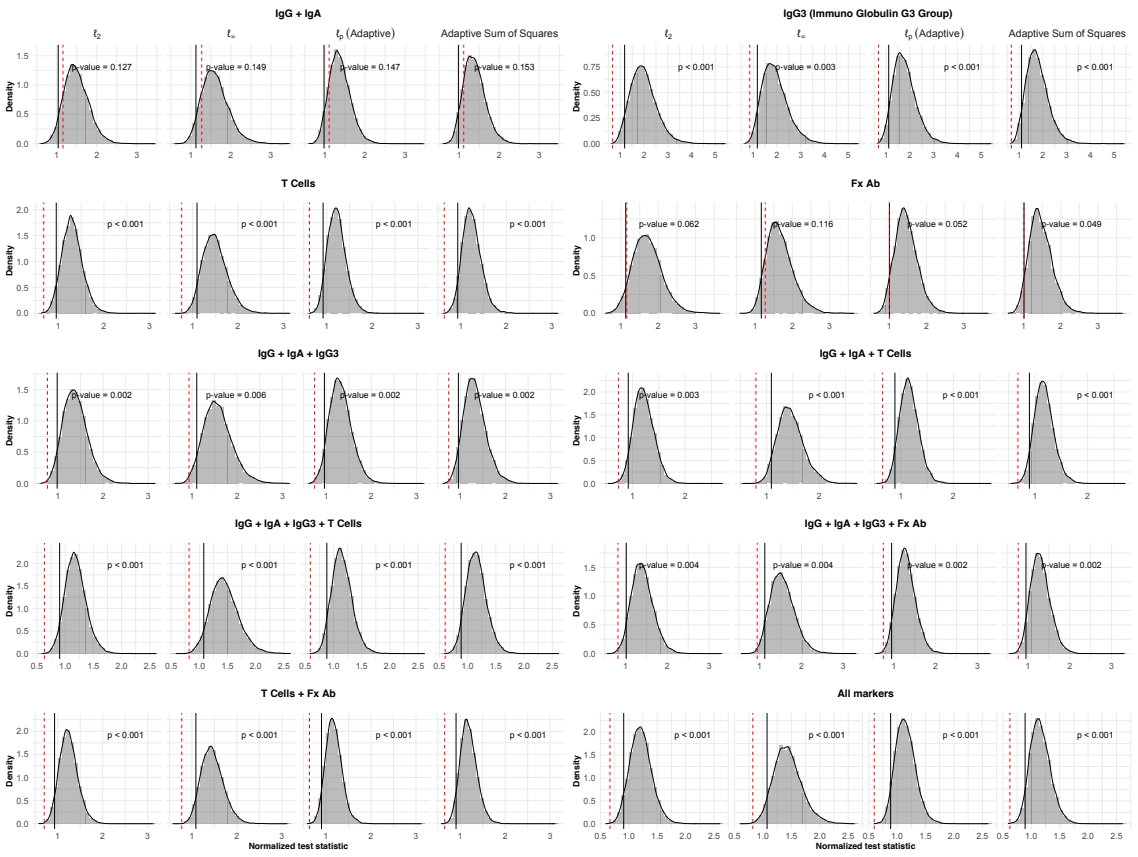


Figure A.3: Estimated limiting distribution of the multiplicative factor measure for both non-adaptive ($\ell_2$ and maximum absolute deviation) and adaptive (adaptive $\ell_p$ and adaptive sum-of-squares) testing procedures. The black and dashed red vertical lines in each plot denote the value of the $5^{th}$ percentile of the limiting distribution and of the test statistic, respectively. This analysis is based on data from the HVTN 505 clinical trial, and the null hypothesis tested is that the biomarkers from the Fx Ab group are not associated with risk of HIV infection.

full-data parameter for biomarker $S_j$ to be

$$\boldsymbol{\beta}_{j,F}(P_F) := \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ E_{P_F}\left[R_j(\boldsymbol{\beta})(S_j, Y)\right] \ ,$$

where we also define $R_j(\boldsymbol{\beta}) : (s, y) \mapsto y \log\{\operatorname{expit}(\beta_1+\beta_2 s)\}+(1-y)\log\{1-\operatorname{expit}(\beta_1+\beta_2 s)\}$. For the two-phase design, the observed-data unit is $X := (W, \tilde{S}, \Delta, Y)$ with $\tilde{S} := \Delta S$. Each participant's probability of being sampled in the second phase depends on their outcome and baseline covariate vector but not on the biomarker vector. In other words, $\Delta$ and $S$ are independent given $(W, Y)$. In this particular study, all cases were sampled but controls were sampled based on BMI, race and ethnicity (Janes et al., 2017). Under this assumption, the full-data parameter can be expressed as the observed-data parameter

$$\boldsymbol{\beta}_j(P) := \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, E_P\left[\frac{\Delta}{P(\Delta = 1 \mid Y, W)} \cdot R_j(\boldsymbol{\beta})(\tilde{S}_j, Y)\right] \ ,$$

where $P$ is a candidate distribution of the observed-data unit.

In the context considered, $(W, Y)$ have a finite support under the true sampling distribution $P_0$. Thus, the parameter value $\boldsymbol{\beta}_{j0} := \boldsymbol{\beta}_j(P_0)$ can be estimated using the plug-in estimator $\boldsymbol{\beta}_{jn} := \boldsymbol{\beta}_j(P_n)$, where $P_n$ is the empirical distribution based on $X_1, X_2, \ldots, X_n$; in practice, this estimator can be obtained using weighted univariable logistic regression with empirically computed weights. The estimator $\boldsymbol{\beta}_{jn}$ is a vector $(\beta_{jn,1}, \beta_{jn,2})$, with components giving estimators of the constant and slope of the best linear model approximation to the true conditional log-odds of risk of HIV infection, respectively.

The influence function of $\boldsymbol{\beta}_{jn}$ is given by

$$x \mapsto -M_{j0}^{-1}\left[\frac{\delta}{\pi_0(w, y)}\nabla R_j(\boldsymbol{\beta}_{j0})(w, \tilde{s}, y) + \left\{1 - \frac{\delta}{\pi_0(w, y)}\right\}\xi_{j0}(w, y)\right] \ ,$$

where we define pointwise the nuisance functions $\pi_0(w, y) := E_0\left(\Delta \mid W = w, Y = y\right)$ and $\xi_{j0}(w, y) := E_0\left[\nabla R_j(\boldsymbol{\beta}_{j0})(W, \tilde{S}, Y) \mid \Delta = 1, W = w, Y = y\right]$ as well as the normalization matrix

$$M_{j0} := E_0\left[\frac{\Delta}{\pi_0(W, Y)}\nabla^2 R_j(\boldsymbol{\beta_{j0}})(W, \tilde{S}, Y)\right] .$$

Here, defining $m_{\boldsymbol{\beta}} : s \mapsto \operatorname{expit}(\beta_1 + \beta_2 s)$, we can compute $\nabla R_j(\boldsymbol{\beta})(w, s, y) = [y - m_{\boldsymbol{\beta}}(s)]\left[\begin{smallmatrix}1\\ s\end{smallmatrix}\right]$ and $\nabla^2 R_j(\boldsymbol{\beta})(w, s, y) = -m_{\boldsymbol{\beta}}(s)[1 - m_{\boldsymbol{\beta}}(s)]\left[\begin{smallmatrix}1 & s\\ s & s^2\end{smallmatrix}\right]$. In particular, the influence function of $\psi_{jn} := \beta_{j2,n}$ is given by

$$\phi_{j0} : x \mapsto \frac{\delta}{\pi_0(w, y)}\left\{a_{j0} + b_{j0}\tilde{s}\right\}\left\{y - m_{\boldsymbol{\beta}_{j0}}(\tilde{s})\right\}$$

$$+\left\{1-\frac{\delta}{\pi_0(w,y)}\right\}E_0\left[\left(a_{j0}+b_{j0}\tilde{S}\right)\left\{Y-m_{\boldsymbol{\beta}_{j0}}(\tilde{S})\right\}\Big|\,\Delta=1,W=w,Y=y\right]\ ,$$

where $a_{j0}$ and $b_{j0}$ are the $[2,1]$ and $[2,2]$ entries of $-M_{j0}^{-1}$. This implies that $n^{1/2}\left(\psi_n-\psi_0\right)$ converges in distribution to a mean-zero multivariate normal distribution with covariance matrix $\Sigma_0$ with entry $[j,k]$ given by $\Sigma_{jk}:=E_0\left[\phi_{j0}(X)\phi_{k0}(X)\right]$. As such, a natural estimator $\Sigma_n$ of $\Sigma_0$ is defined entrywise as $\Sigma_{jk,n}:=\frac{1}{n}\sum_{i=1}^{n}\phi_{jn}(X_i)\phi_{kn}(X_i)$ with

$$\begin{aligned}\phi_{jn}:x\mapsto\ &\frac{\delta}{\pi_n(w,y)}\left\{a_{jn}+b_{jn}\tilde{s}\right\}\left\{y-m_{\boldsymbol{\beta}_{jn}}(\tilde{s})\right\}\\&+\left\{1-\frac{\delta}{\pi_n(w,y)}\right\}E_n\left[\left(a_{jn}+b_{jn}\tilde{S}\right)\left\{Y-m_{\boldsymbol{\beta}_{jn}}(\tilde{S})\right\}\Big|\,\Delta=1,W=w,Y=y\right]\ ,\end{aligned}$$

where $\pi_n$ is an estimator of $\pi_0$, $a_{jn}$ and $b_{jn}$ are the $[2,1]$ and $[2,2]$ entries of $-M_{jn}^{-1}$ with $M_{jn}:=-\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{\pi_n(W_i,Y_i)}m_{\boldsymbol{\beta}_{jn}}(\tilde{S}_i)[1-m_{\boldsymbol{\beta}_{jn}}(\tilde{S}_i)]\begin{bmatrix}1&\tilde{S}_i\\\tilde{S}_i&\tilde{S}_i^2\end{bmatrix}$, and $E_n$ denotes an empirical expectation relative to the distribution of $\tilde{S}$ given $\Delta=1$ and $(W,Y)$.

When solving for efficient influence function, if we have solved everything, but we are still not integrating over everything, we can add an integral over everything else and add the conditional probability of everything else.

## A.7 Parameter of interest

Let $Y$ be the binary outcome of interest, let $\underline{W}=(W_1,\ldots,W_d)$ be a vector of covariates, and let $\Delta$ be an indicator for observing $Y$. Our parameter of interest for each $W_j$ is the $\beta$ coefficient for risk ratio corresponding each $W_j$.

$$\begin{aligned}\Psi_j\left(P^{\text{full}}\right)&=\frac{\text{Cov}\left(\log\left(Pr\left(Y=1|W_j\right)\right),W_j\right)}{\text{Var}(W_j)}\\\Psi_{j,1}\left(P^{\text{full}}\right)&=E\left[\log\left(Pr\left(Y=1|W_j\right)\right)W_j\right]\\\Psi_{j,2}\left(P^{\text{full}}\right)&=E\left[\log\left(Pr\left(Y=1|W_j\right)\right)\right]\\\Psi_{j,3}\left(P^{\text{full}}\right)&=E\left[W_j\right]\\\Psi_{j,4}\left(P^{\text{full}}\right)&=E\left[W_j^2\right]\end{aligned}$$

## A.8 Identifying the parameter

In the data generating mechanism for the second data example from the first chapter, we assume that data is missing at random : $Y\perp\Delta|\underline{W}$.

$$
\begin{aligned}
\Psi_{j,1}\left(P^{\text{full}}\right) &= \frac{\text{Cov}\left(\log\left(Pr\left(Y=1|W_j\right)\right), W_j\right)}{\text{Var}(W_j)} && \text{by law of total expectation} \\
&= \frac{\text{Cov}\left(\log\left(E\left[Pr\left(Y=1|W_j, W_{-j}\right)|W_j\right]\right), W_j\right)}{\text{Var}(W_j)} && \text{by conditional independence} \\
&= \frac{\text{Cov}\left(\log\left(E\left[Pr\left(Y=1|\Delta=1, W\right)|W_j\right]\right), W_j\right)}{\text{Var}(W_j)} \\
\tilde{\Psi}_j\left(P^{\text{obs}}\right) &= \frac{\text{Cov}\left(\log\left(E\left[Pr\left(\Delta Y=1|\Delta=1, W\right)|W_j\right]\right), W_j\right)}{\text{Var}(W_j)}
\end{aligned}
$$

# Appendix B

# Chapter two appendix

## B.1 Additional figures

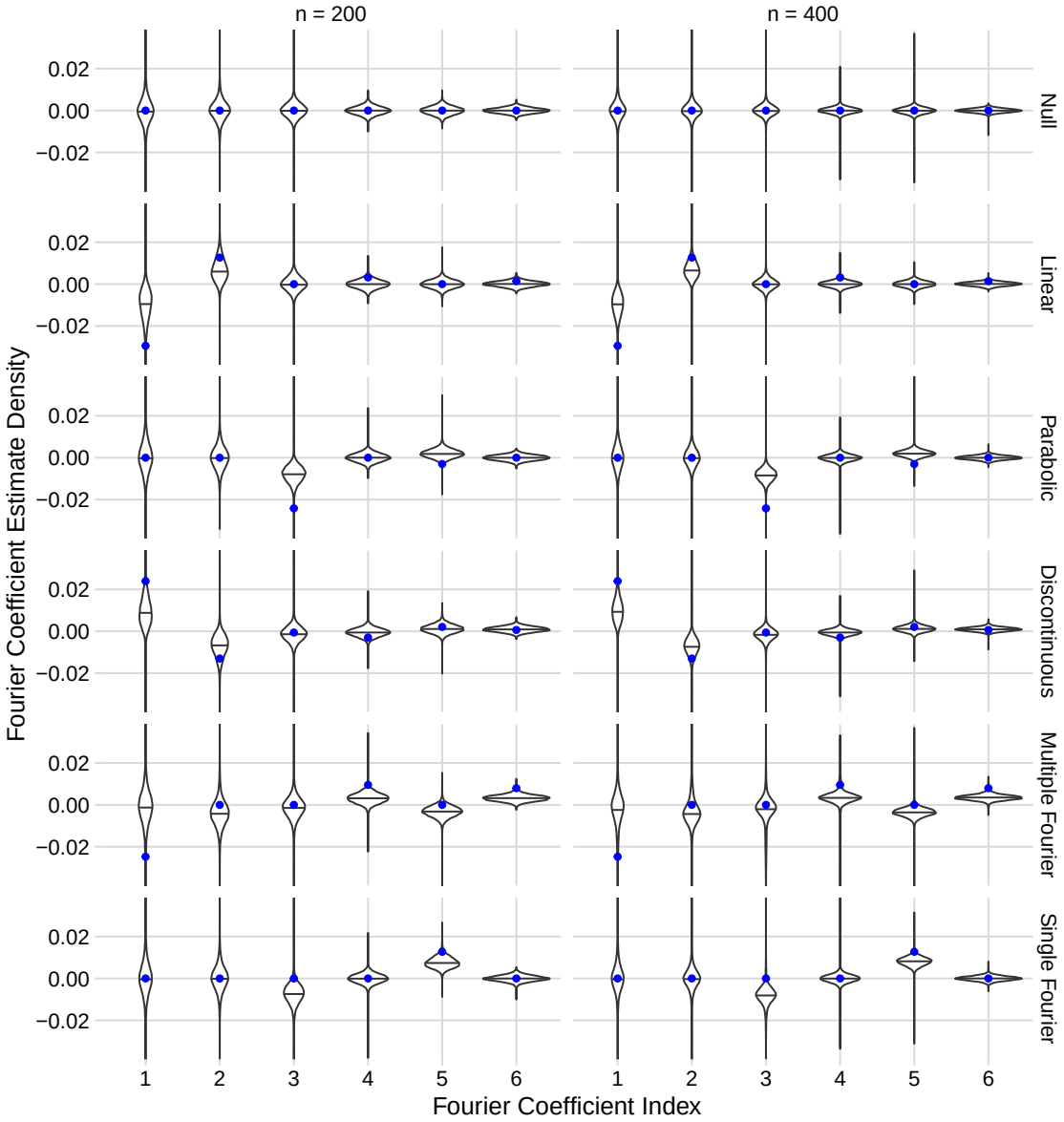


Figure B.1: Empirical sampling distribution of the scaled parameter estimates for the first six elements of the parameter estimator vector for the non-standardized test statistic. The black vertical line shows the median of the given sampling distribution. The blue point shows the true value of each Fourier coefficient in each setting.

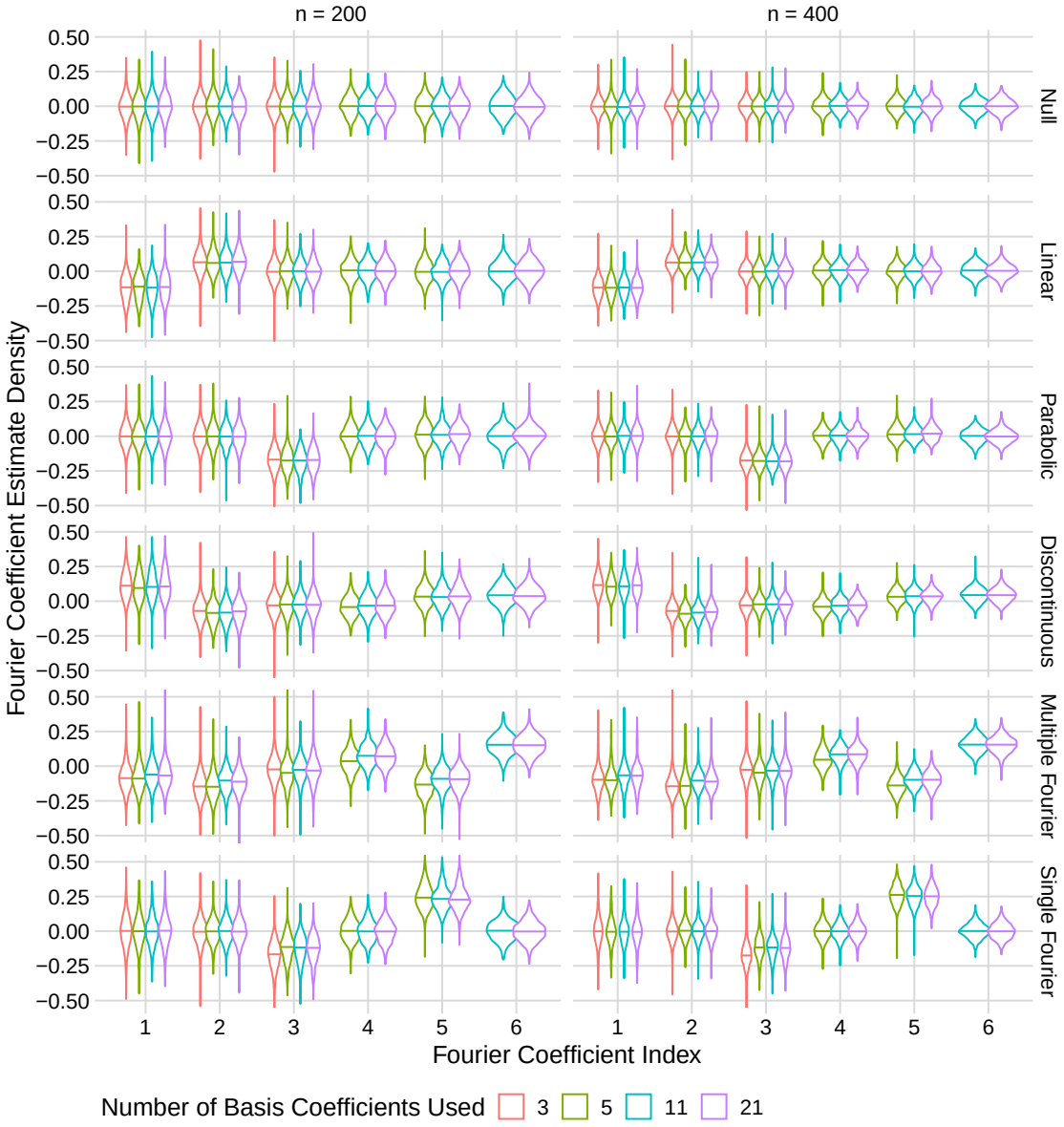


Figure B.2: Empirical sampling distribution of the scaled parameter estimates for the first six elements of the parameter estimator vector for the standardized test statistic across sample size and number of basis coefficients. The color of each violin plot indicates the number of basis functions used to define the test statistic. The vertical line shows the median of the given sampling distribution in each setting.

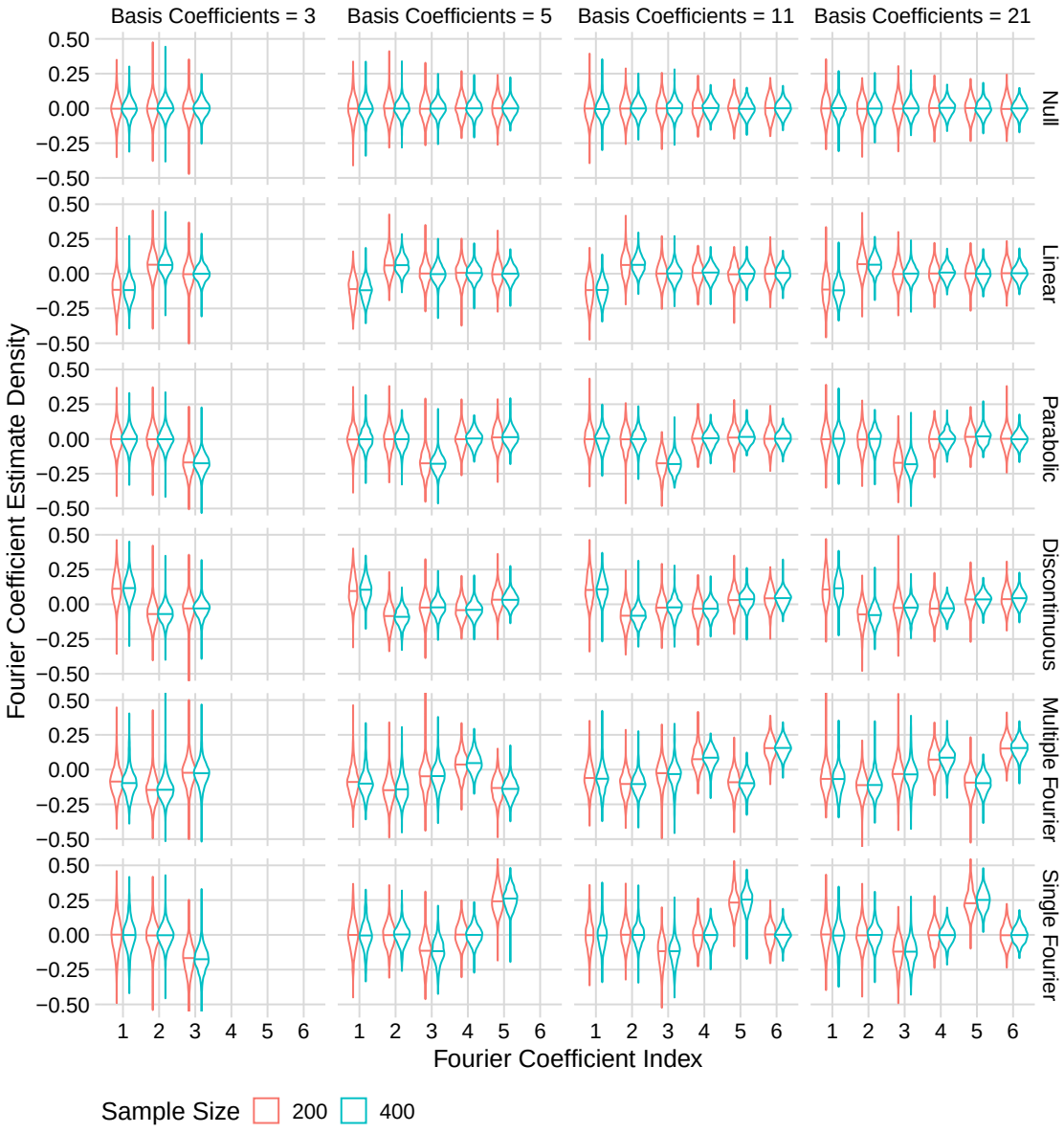


Figure B.3: Empirical sampling distribution of the scaled parameter estimates for the first six elements of the parameter estimator vector for the standardized test statistic across sample size and number of basis coefficients. The color of each violin plot indicates the size of the sample used to estimate the parameter. The vertical line shows the median of the given sampling distribution in each setting.

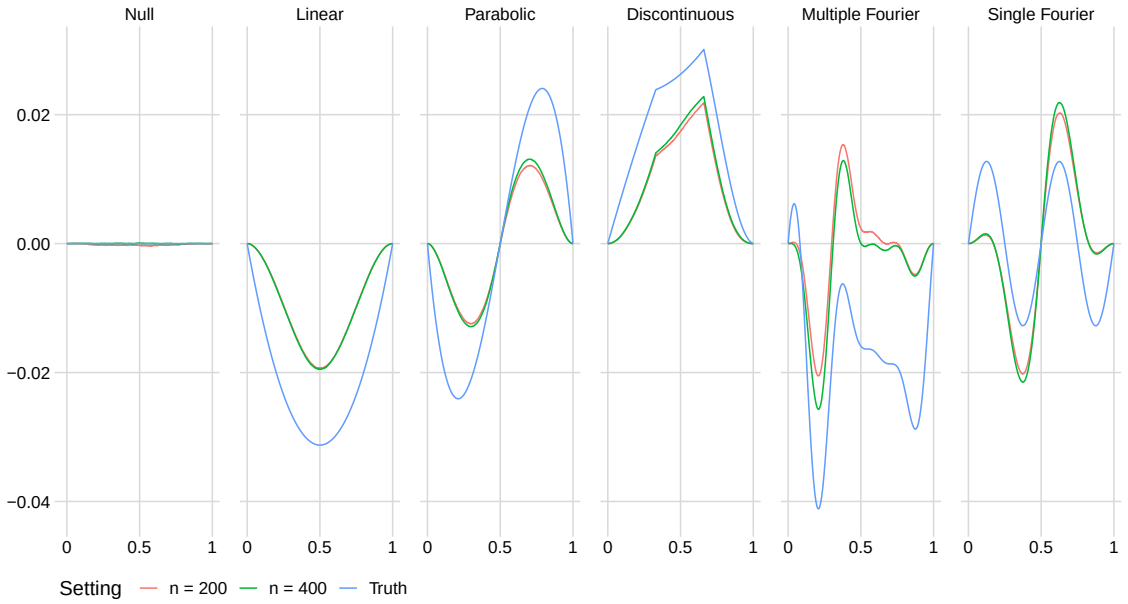


Figure B.4: Average estimated function in each setting for both considered sample sizes.

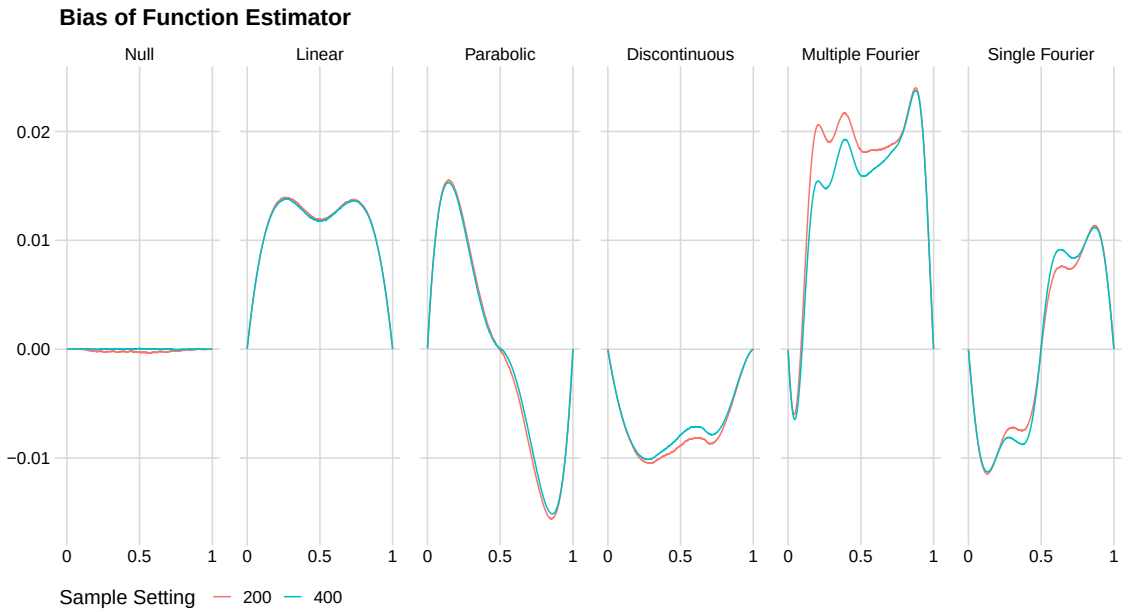


Figure B.5: Bias of the function estimator in each setting for both considered sample sizes.

# Appendix C

# Chapter three appendix

## Appendices

## C.1 Details of data analysis

#### Baseline covariates

Baseline covariates were recorded at the beginning of each trial. The covariates baseline considered in this analysis where age, main partner HIV status, number of partners in the last three months, number of sex acts in the last seven days, number of sex acts in which a condom was used, an indicator for if an individual was worried about using the ring, marital status, and study site.

All but one study site included participants in both trials. The Umkomaas study site, which only enrolled patients during ASPIRE, was collapsed into the Isipingo site. The Isipingo site is the closest geographically to the Umkomaas site, and all participants who were seen at the Umkomaas site during ASPIRE were seen at the Isipingo site if they enrolled in HOPE.

### Adherence measure

Adherence was measured on individuals in the active arm of the ASPIRE trial and all individuals in the HOPE trial. During the first year of the ASPIRE trial, to determine adherence, the levels of Dapivirine using ultra-performance liquid chromatography-tandem mass spectrometry assay (Clinical Pharmacology Analytical Laboratory). However, a plasma Dapivirine higher than 95 pg per milliliter was usually attained within 8 hours of continuous use, making it possible appear adherent with only a single days use of the ring. After the first year in ASPIRE and throughout HOPE, adherence was measured by the level of Dapivirine left in the ring using acetone extraction and high pressure liquid chromatography (Parexel). To account for the missing ring-based adherence measures from the first year of ASPIRE, values were sequentially imputed starting from the last

month in which an individual had Dapivirine concentrations and working backwards.

In the ASPIRE trial, rings were dispensed and collected at each of the monthly visits, making it possible to have month level information on adherence levels of the active arm participants. However, during HOPE, when visits were only once every three months, each participant received and returned three rings at each visit. As a result measurements of adherence can only be determined on three-month intervals. To make adherence levels comparable across the trials, adherence levels were determined at roughly three-month intervals for both trials. The average Dapvirine release for each three-month period was determined and if the average monthly release rate was above 0.9 mg an individual would be classified as adherent, and otherwise they would be deemed non-adherent. In a sensitivity analysis we also consider the other mentioned release monthly Dapvirine release rate cutoff (1.5 mg). As a result, each individual has four adherence measurements throughout the trial, each of which is a binary variable. In an analysis of the ASPIRE trial, the minimum level of Dapivirine released to be deemed adherent trial was 0.9mg. This is in contrast to the minimum level of 1.5 in the analysis of HOPE (Baeten et al., 2021). In our analysis, a single threshold was used to asses adherence during both trials. This threshold was 0.9mg in the main analysis, but in a sensitivity analysis we also consider the 1.5mg threshold.

In the counterfactual outcome framework mentioned earlier, individuals are treated as censored at the point in which they are no longer following the chosen adherence pattern. This time point is chosen to be the date of the last visit in which a ring is returned for a given three month period.

### Event and Censoring time

While nearly all individuals in the HOPE trial participated for close to 12 months, the amount of time spent in the ASPIRE trial was far more variable (see Figure 3.2). To make fair comparisons between the two trials a one year subsection of ASPIRE was considered. In the main analysis, this year was the first year of ASPIRE. However, in sensitivity analyses, other one year periods are considered starting, starting 3, 6, 9, 12, 15, and 18 months after enrollment.

## C.2 Sensitivity analyses

To understand the impact of decisions made regarding the adherence cutoff and ASPIRE comparison group time period, we carried out estimation of the parameter using each combination of potential cutoff values and time periods. A summary of these analyses is given in Figure C.1. Some minor differences in the estimates exist between analyses that use the same ASPIRE starting month but different adherence cutoffs. However, larger differences are observed across ASPIRE starting month. As was mentioned earlier, the estimated ASPIRE effectiveness is an important

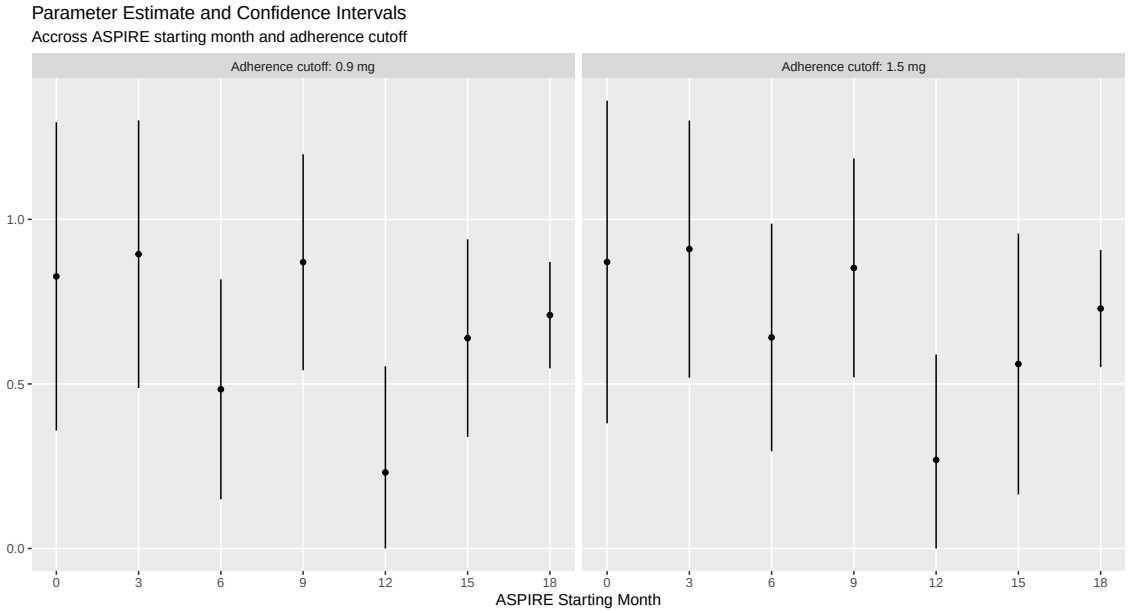


Figure C.1: Plot showing the parameter estimates and confidence intervals for different analysis choices with regards to the adherence cutoff value and ASPIRE starting month used.

factor when estimating the HOPE effectiveness. As seen in Figure 3.3, the incidences in the first 12 months of ASPIRE shows almost no large difference between the two arms, though later in the study, differences do appear. Thus, it is not surprising that using these latter time periods would lead to a noticeable change in the estimated effectiveness compared to when the first 12 months of ASPIRE are used.

To study this connection more closely, multiple simple analyses of the trial were conducted each one based on a data set constructed to reflect the survival and censoring times observed if the ASPIRE trial started at the specified month and lasted for only twelve months. Figure C.2 shows the observed HIV infection incidence curves for each of the different starting times. Above each plot is the estimated hazard ratio from a Cox proportional hazards model fit to the data for the specified starting month. The model includes a single predictor variable that is an indicator for being in the active arm of the ASPIRE trial. In Table C.1, each of these hazard ratios is compared to the estimated effectiveness using a particular starting month. It can be seen that the estimated effectiveness closely tracks the estimated hazard ratio, where rows that have lower hazard ratios also tend to have an estimated effectiveness risk ratio that is lower as well.

When evaluating the validity of each proposed ASPIRE starting month, the most important factor to consider is which time period is most likely to satisfy the bridging assumption. Using the first twelve months of ASPIRE is a natural choice because it allows for comparisons between the first year of both trials. If, for example, the effect of the ring only began after a year of use, then using any other period in ASPIRE would result in the bridging assumption being false. While less

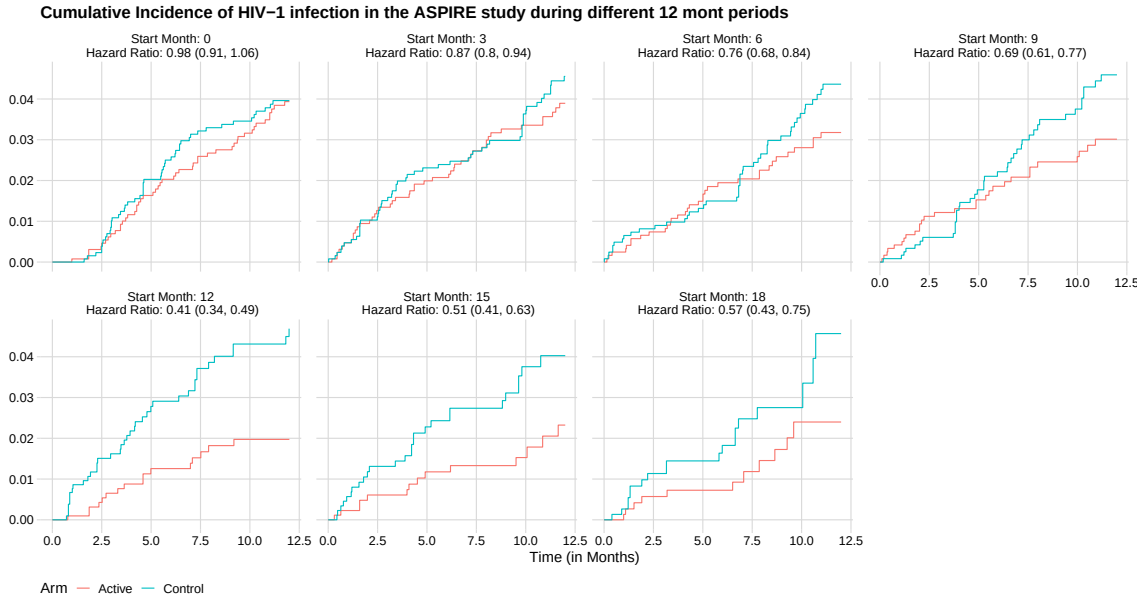


Figure C.2: Plots showing different one-year periods over which the ASPIRE trial data could be used.

| Starting Month | Cox Model Hazard Ratio | Effectiveness Estimate |
|---|---|---|
| 0 | 0.98 (0.91, 1.06) | 0.83 (0.36, 1.30) |
| 3 | 0.87 (0.80, 0.94) | 0.89 (0.49, 1.30) |
| 6 | 0.76 (0.68, 0.84) | 0.48 (0.15, 0.82) |
| 9 | 0.69 (0.61, 0.77) | 0.87 (0.54, 1.20) |
| 12 | 0.41 (0.34, 0.49) | 0.23 (0.00, 0.55) |
| 15 | 0.51 (0.41, 0.63) | 0.64 (0.34, 0.94) |
| 18 | 0.57 (0.43, 0.75) | 0.71 (0.55, 0.87) |

Table C.1: Comparison of a simple Cox analysis to the estimated effectiveness for different ASPIRE starting months. To estimate effectiveness it is necessary to select the 12 month period in ASPIRE used to identify the parameter. The Cox Model Hazard Ratio column shows the estimated hazard ratio from a Cox proportional hazards model with a single predictor that is an indicator for being assigned to the active arm of the ASPIRE trial. Each row corresponds to the model being fit on a different set of data, determined by the starting month. For each row, only data from the twelve month period starting at the starting month is used to fit the Cox model. The effectiveness estimate uses the 0.9 mg adherence cutoff.

intuitive, choosing later ASPIRE starting months could be a good option because doing so would result in the populations that defined the bridging assumption being closer to one another in time (relative to other starting months) and thus more likely to resemble one another. One other factor worth considering is that using an earlier ASPIRE starting month will result in an estimate based on a larger sample size since all individuals at least started the ASPIRE trial.