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

Adam Elder

April 16th 2021

Co-advised by Alex Luedtke and Marco Carone
University of Washington

Project One: An Adaptive Multivariate Point Null Test

Motivation

Let $O = (Y, X_1, \ldots, X_d)$ and let $O_1, O_2, \ldots, O_n$ be drawn independently from a common unknown distribution P_0 in the statistical model $\mathcal{M}$.

We are interested in whether the outcome Y is associated with any of the covariates $X_1, \ldots, X_d$.

- Let the measure of association (for example correlation) between Y and X_i be denoted by ψ_i.
- Let $\hat{\psi}_i$ be the corresponding estimate of ψ_i based on $O_1, O_2, \ldots, O_n$.
- We wish to test the multivariate point null:

$$H_0 : \psi_1 = \psi_2 = \ldots = \psi_d = 0$$

Motivation

We will apply this method in the setting of HIV vaccine research.

- Each covariate is a biomarker, and the outcome is HIV infection.
- The null hypothesis in this setting is that no biomarkers are associated with infection.

Previous Work

Previous work in this area can be broken down into two categories

- Construct a test for each ψ_i, and correct for multiple testing. That is, define

$$H_{0i} : \psi_i = 0 \text{ v.s. } H_{1i} : \psi_i \neq 0,$$

  and generate p-values for each test that correct for multiple hypothesis testing. Then if any of the d hypothesis tests rejects, reject the multivariate null.
- Directly test the multivariate point null.

Previous work in multiple hypothesis testing

Work in multiple hypothesis testing began with Tukey in 1953[1] and was followed by Bonferroni and others[2].

There are many modern improvements on these initial procedures that provide more power, and allow users to specify the trade-offs to be made between type-1 and type-2 errors[3].

[1] Miller 1981.

[2] Hochberg 1988; Holm 1979; S. Holland and DiPonzio Copenhaver 1988.

[3] Lehmann and Romano 2005; Dudoit and van der Laan 2008.

Downsides of multiple hypothesis testing

While it is possible to test a multivariate point null using multiple hypothesis testing procedures, there can be a loss of power when compared with methods that directly test the multivariate point null.

- Multiple hypothesis testing methods allow users to know which hypothesis was rejected.
- This knowledge can come at the cost of power when testing a multivariate point null.

Previous work in multivariate point null tests

More recent work in multivariate point null tests provides more power than previous methods, but are often difficult to generalize to other parameters and data-generating mechanisms[4]. Here we propose a general-purpose testing procedure that

- provides more power than the easily applicable multiple hypothesis correction methods, and
- is applicable for a wide variety of data-generating mechanisms and parameters of interest.

[4] Donoho and Jin 2004; McKeague and Qian 2015; Pan et al. 2014; Xu et al. 2016.

Creating a test

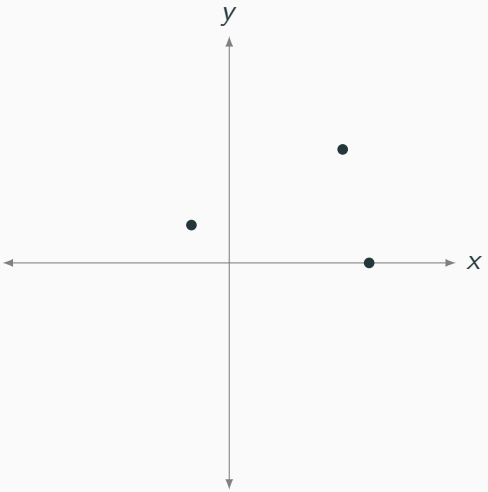
Creating a test can usually be broken down into three steps:

- Picking a test statistic
- Determining the (limiting) distribution of the test statistic under H_0
- Determining which values of the test statistic would be extreme under H_0.

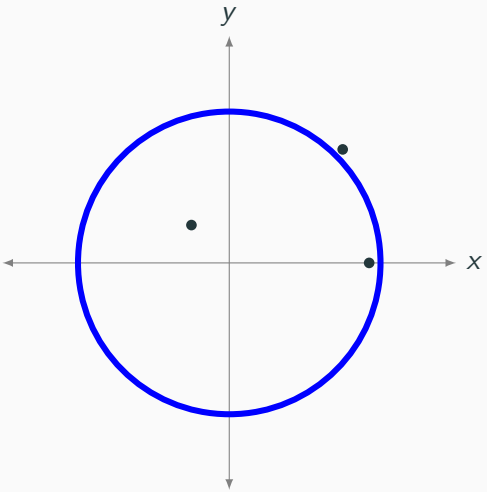
We will consider the last item briefly

Which observation is the most extreme?

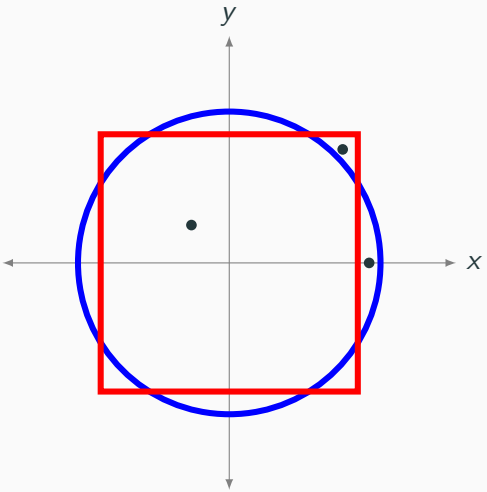
Which observation is the most extreme?



Which observation is the most extreme?



Which observation is the most extreme?



Understanding the vector of parameter estimates

Define $\underline{\psi} = (\psi_1, \ldots, \psi_d)$ and $\underline{\hat{\psi}} = \left(\hat{\psi}_1, \ldots, \hat{\psi}_d\right)$

Suppose that for each $P \in \mathcal{M}$,

$$\sqrt{n}\left(\underline{\hat{\psi}} - \underline{\psi}\right) \xrightarrow{d} Z \sim N\left(0, \Sigma_P\right)$$

Transformation to create a test statistic

Defining which observations are more extreme in higher dimensions is difficult. We will consider a simple example in which the test statistic is the ℓ_2 or Euclidean norm of the vector of parameter estimators.

$$\hat{t} = \sqrt{\sum_{i=1}^{d} \hat{\psi}_i^2}$$

Next we consider the limiting distribution of $\sqrt{n}\hat{t}$ under H_0.

Obtaining the limiting distribution under the null

Under the null:

$$\sqrt{n}\underline{\hat{\psi}} \xrightarrow{d} Z \sim N\left(0, \Sigma_P\right).$$

The continuous mapping theorem tells us that under the null

$$\|\sqrt{n}\underline{\hat{\psi}}\|_2 \xrightarrow{d} \|Z\|_2, \quad Z \sim N\left(0, \Sigma_P\right).$$

While knowing the exact distribution of $\|Z\|_2$ is difficult, we may sample from $\|Z\|_2$ to carry out a test.

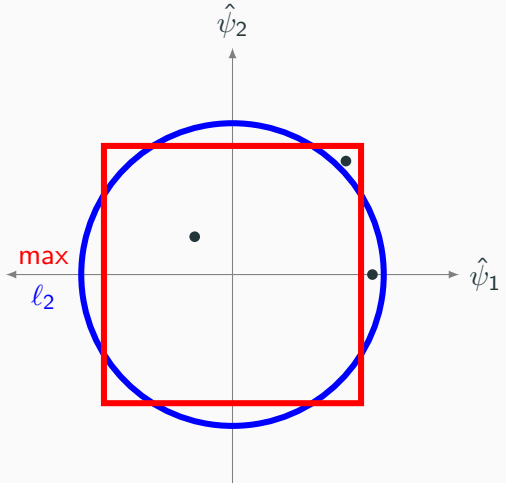
Running a test

First, take many draws $\underline{z}_1, \ldots, \underline{z}_B$ from an estimate of the distribution of Z. Next, $\|\underline{z}_1\|_2, \ldots, \|\underline{z}_B\|_2$ can be used to approximate the limiting distribution of $\|\sqrt{n}\hat{\psi}\|_2$ under H_0. Using these draws, a p-value can be calculated:

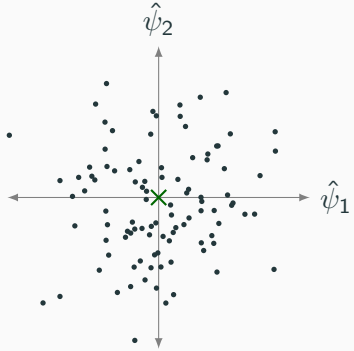
$$\frac{1}{B}\sum_{i=1}^{B} I\{\|\sqrt{n}\hat{\underline{\psi}}\|_2 \leq \|\underline{z}_i\|_2\}$$

Choosing a norm

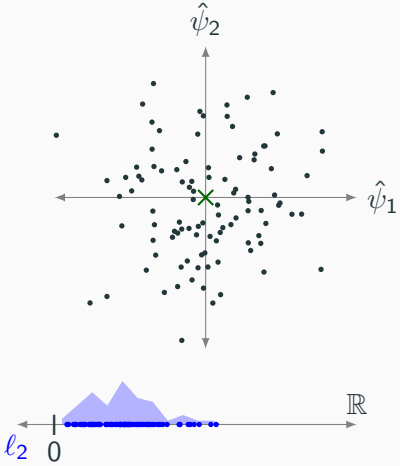
While using the Euclidean norm is a valid way to define our test, a different norm may provide better power:



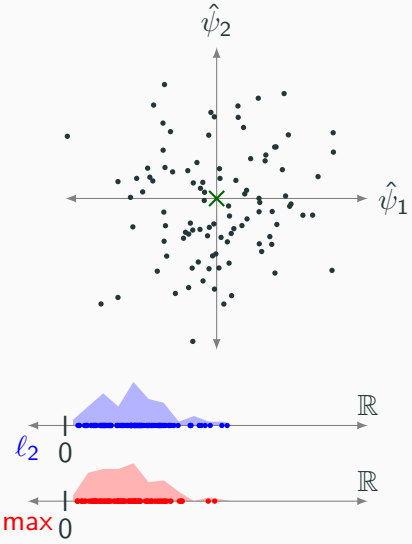
Comparing powers between norms



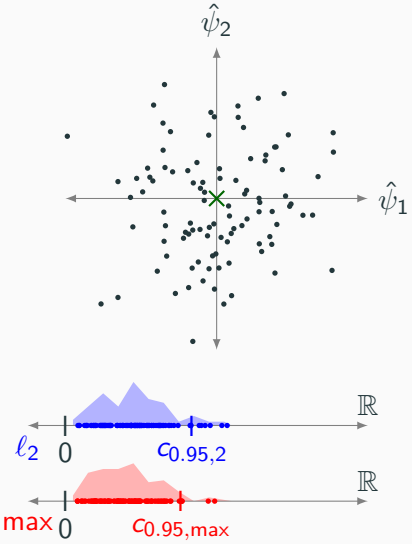
Comparing powers between norms



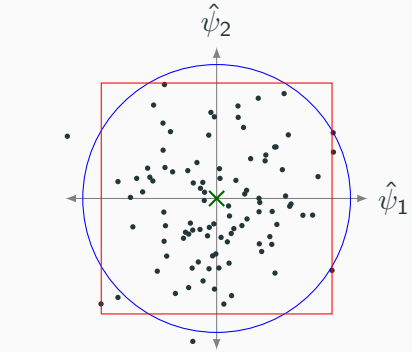
Comparing powers between norms



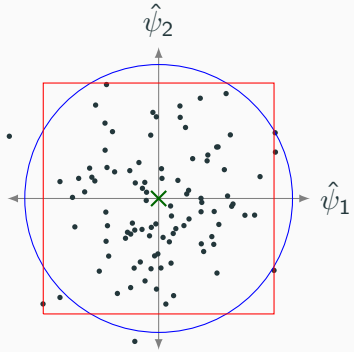
Comparing powers between norms

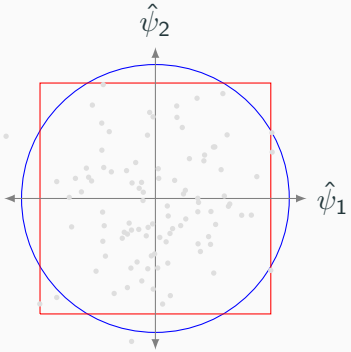


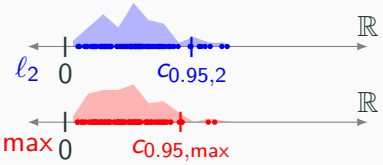
Comparing powers between norms



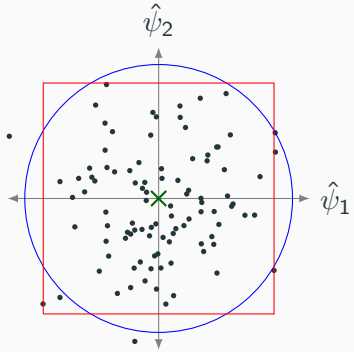
Comparing powers between norms

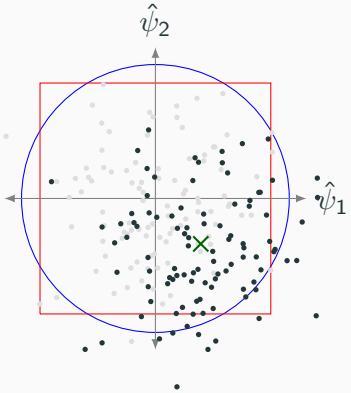




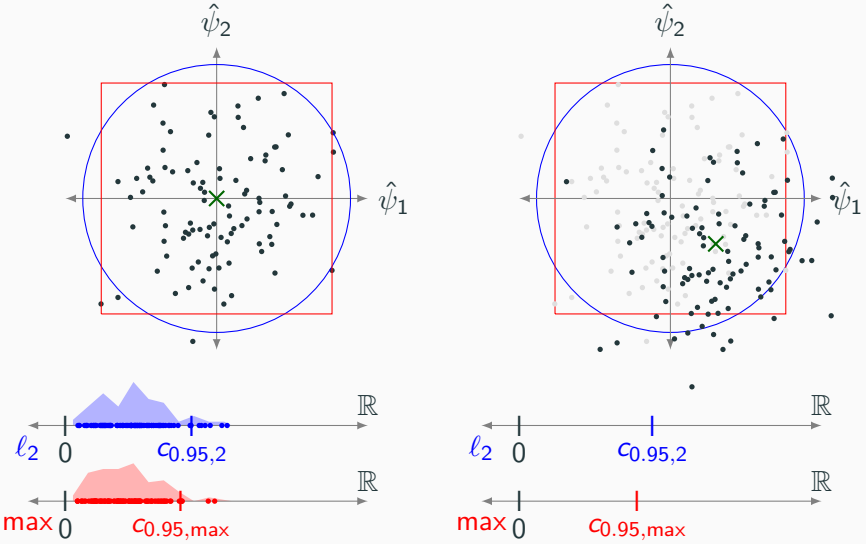


Comparing powers between norms

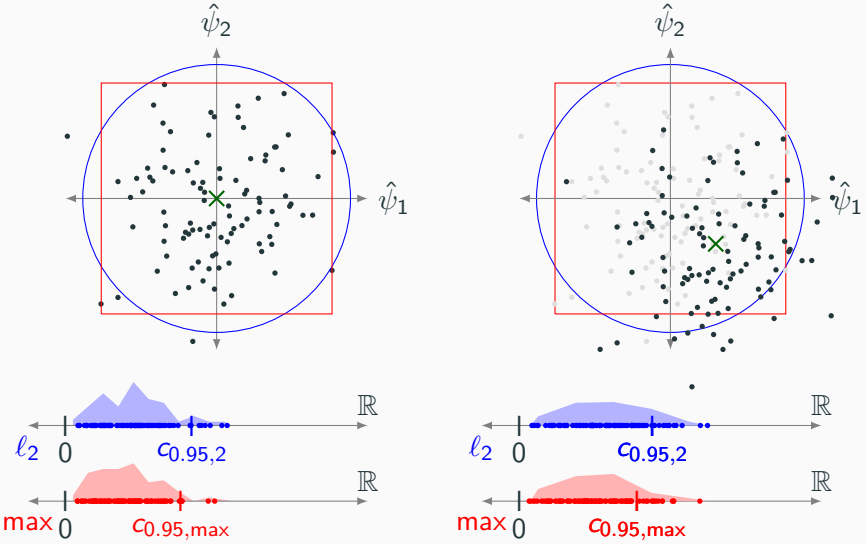




Comparing powers between norms



Comparing powers between norms



Choosing a norm

While it may be possible to know a priori which norm will perform optimally, this will not generally be the case.

We have developed a method that adaptively chooses a norm while controlling type-1 error. This method requires measuring the "performance" of each norm for any potential alternative ψ.

This idea was inspired in part by the work of Ian McKeague and a commentary on this work from Yichi Zhang and Eric Laber[5].

[5] McKeague and Qian 2015; Zhang and Laber 2015.

Defining a performance metric

Here, we denote by Γ a performance metric that maps from $\mathbb{R}^d \times \mathbb{M}_{d\times d}$ into $\mathbb{R}$ where $\mathbb{M}_{d\times d}$ contains all positive definite $d \times d$ matrices.

An example of such a performance metric is the acceptance rate performance metric:

$$\Gamma_{\|\cdot\|}(\omega, \Sigma) = \Pr(\|Z + \omega\| \leq c_\alpha(\Sigma)) \text{ where}$$
$$c_\alpha(\Sigma) \equiv \inf\{c : \Pr(\|Z\| < c) \geq \alpha\} \text{ and } Z \sim N(0, \Sigma).$$

Performance metric

There are many alternative performance metrics we might consider:

- The average power across alternatives near ω.
- The p-value of a test using the specified norm.
- The factor one must multiply ω by to achieve 80% power. This metric has the advantage of taking values in $(0, \infty)$.

Defining a performance metric

A primary benefit of all the previously mentioned performance metrics is that these metrics are comparable across different norms.

- This is not the case for the norms themselves
- For every $\omega \in \mathbb{R}^d$, it holds that $\|\omega\|_1 \geq \|\omega\|_2 \geq \ldots \geq \|\omega\|_\infty$.

For a set of potential norms $\|\cdot\|_1, \|\cdot\|_2, \ldots, \|\cdot\|_k$, the adaptive performance metric is defined as

$$\Gamma^*(\omega, \Sigma) = \min\left(\Gamma_{\|\cdot\|_1}(\omega, \Sigma), \Gamma_{\|\cdot\|_2}(\omega, \Sigma), \ldots, \Gamma_{\|\cdot\|_k}(\omega, \Sigma)\right).$$

What has changed

While both the use of a performance metric and the adaptive selection of a norm add complexity to our method, the underlying testing procedure remains largely unchanged.

Taking draws from the limiting distribution

As the test has been described, p-values are calculated by taking draws $z_1, \ldots, z_B$ from the normal limiting distribution Z, then computing

$$\frac{1}{B}\sum_{i=1}^{B} I\left\{\Gamma^*(\sqrt{n}\hat{\psi}, \Sigma) > \Gamma^*(z_i, \Sigma)\right\}$$

In practice, we will not know Σ, and will instead use a consistent estimator $\hat{\Sigma}$ of Σ and take draws from $\hat{Z} \sim N(0, \hat{\Sigma})$.

Permutation based estimated limiting distribution

For certain data-generating mechanisms and parameters, it is possible to use a permutation of the observed data to take draws $z_1^{\#}, \ldots, z_B^{\#}$ from the distribution of $\sqrt{n}\hat{\psi}$ under the null.

- Doing this requires more computation but can provide better finite sample performance.
- The permutation-based test has a slightly different null hypothesis.

Theoretical results

We now turn to some theoretical results of the described test.

- The theoretical results state that for any Γ satisfying a set of conditions, a test defined using Γ has desirable properties.
- Next we describe these conditions placed on the performance metric Γ.
- Note that the acceptance rate performance metric we have considered here does satisfy all of these conditions (both for fixed norms and for the adaptive version of the metric).
- We will assume throughout that $\hat{\psi}$ is an asymptotically linear estimator of ψ and $\widehat{\Sigma}$ converges in probability to Σ.

Guarentees for tests based on Γ

Theorem (Asymptotic Consistency and type-1 error control)

If Γ satisfies conditions 1 and 2, then the test defined by

$$\text{reject } H_0 \text{ if } \Gamma(\sqrt{n}\hat{\psi}, \hat{\Sigma}) \leq F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}(\alpha) \tag{1}$$

has a probability of rejection that converges to α as $n \to \infty$ under H_0.

If additionally, Γ satisfies condition 3, then the test defined in equation (1) is also consistent.

Guarantees for tests based on Γ

Theorem (Non-trivial local unbiasedness)

Let $P_{n^{-1/2}}$ be a sequence of local alternatives converging to some P_0 under which the null holds. If Γ satisfies conditions 1, 2, 4, and 5, and the estimator $\hat{\psi}$ of ψ is regular and asymptotically linear at P_0, then the test defined in equation (1) has power greater than α under $P_{n^{-1/2}}$ for all n large enough.

Conditions on Γ

1. The performance metric Γ is continuous and non-negative on $\mathbb{R}^d \times \mathbb{M}_{d\times d}$ where $\mathbb{M}_{d\times d}$ contains all positive definite $d \times d$ matrices.

   This means the performance metric is smooth with respect to both the vector and matrix arguments.

2. $\Pr(\Gamma(Z, \Sigma) = t) = 0$ where $Z \sim N(0, \Sigma)$ for every t and positive definite Σ.

   The random variable defined by $\Gamma(Z, \Sigma)$ has a continuous cumulative distribution function.

Conditions on Γ

3. $\Gamma(\sqrt{n}\hat{\psi}, \widehat{\Sigma}) \xrightarrow{p} 0$ under sampling from any fixed alternative.

   Under fixed alternatives, as n gets larger $\sqrt{n}\hat{\psi}$ will move further away from the origin. This condition requires that as this happens, the performance metric converges in probability to zero.

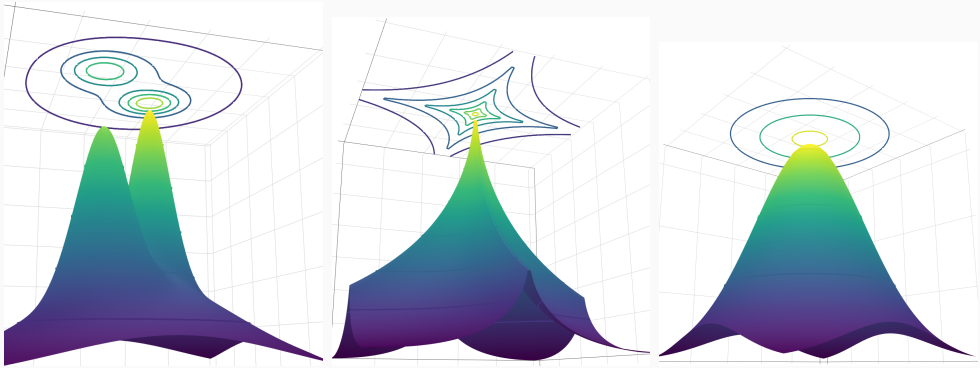
Conditions on Γ

4. $\Gamma(\cdot, \Sigma)$ is quasi-concave for every positive definite Σ (for every k, the set $\{\omega : \Gamma(\omega, \Sigma) \geq k\}$ is convex).

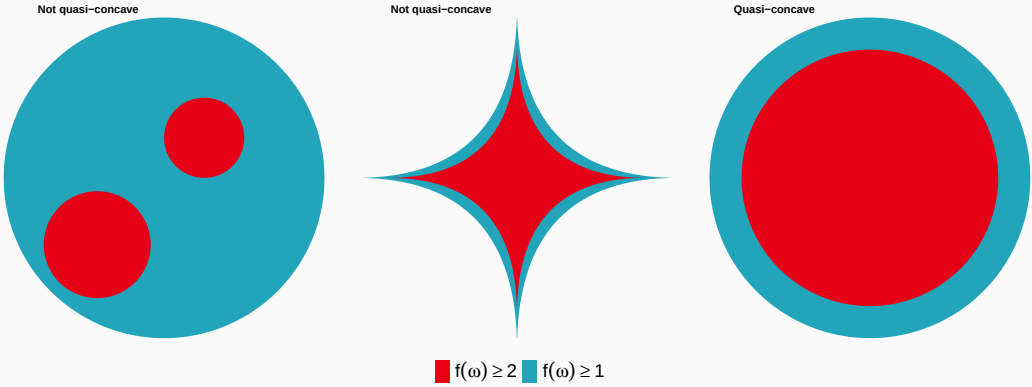
   We will cover this item in the next slides.

5. $\Gamma(\cdot, \Sigma)$ is centrally symmetric for every positive definite Σ, i.e., $\Gamma(-\omega, \Sigma) = \Gamma(\omega, \Sigma)$.

Visualizing quasi-concavity



Visualizing quasi-concavity



Examples

We now turn to the examples in which we studied our test.

For these examples, we consider two different sets of norms. The first set of norms is the ℓ_p norm we have considered up to this point:

$$\ell_p(\omega) = \sqrt[p]{\sum_{i=1}^{d} |\omega_i|^p}.$$

The other norm considered is referred to the sum of squares norm and is defined by:

$$\jmath_k(\omega) = \sqrt{\sum_{i=1}^{k} \omega_{(d-i+1)}^2}$$

where $\omega_{(1)}^2, \ldots, \omega_{(d)}^2$ are the order statistics of $\omega_1^2, \ldots, \omega_d^2$.

Example 1: Correlation

In our first and simplest example, we consider the data unit $X = (W_1, W_2, \ldots, W_d, Y)$, where $W_1, W_2, \ldots, W_d$ represent real-valued covariates and Y is some outcome of interest, and take the parameter of interest

$$\psi_j := corr(W_j, Y)$$

to be the marginal correlation between W_j and Y.

Example 1: Correlation

Data are generated from a linear model:

$$Y = W_1\beta_1 + \ldots + W_d\beta_d + \varepsilon,$$

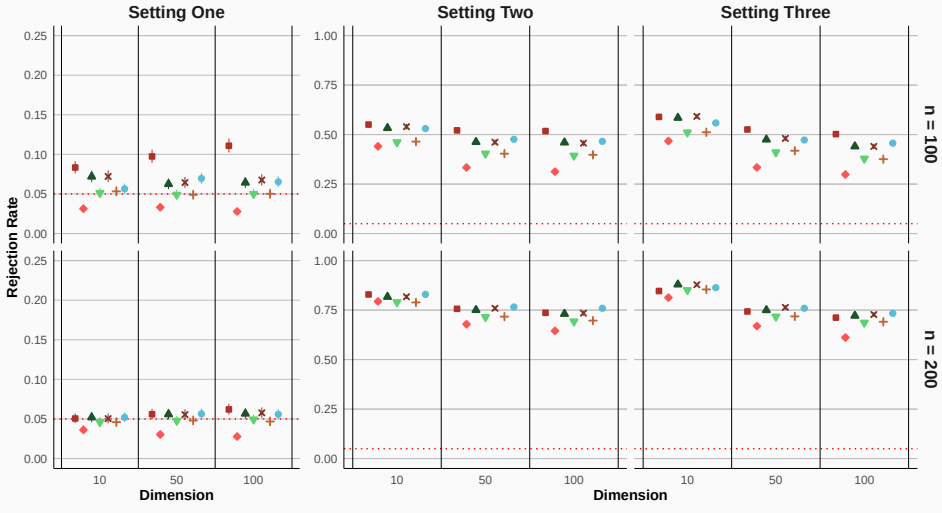
where

- ε is independent of $(W_1, \ldots, W_d)$
- the between W correlation is 0, 0.5, or 0.8 for all W.

We consider three different settings defined by different sets of β values.

- In setting one (the null setting), all $\beta_i = 0$
- In setting two, $\beta_1 = 1/4$ and all other $\beta_i = 0$
- In setting three, $\beta_1 \ldots \beta_5 = 0.15$, $\beta_6 \ldots \beta_{10} = -0.1$, and all other $\beta_i = 0$

Example 1 (Between W correlation is 0.5)



Example 2: Working log-linear regression model under missingness

In our second example, the data unit is $(W_1, W_2, \ldots, W_d, U, \Delta)$, where

- $W_1, W_2, \ldots, W_d$ represent real-valued covariates,
- Δ 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 the first example, but with potential missingness of the outcome value in some observations.

Example 2: Working log-linear regression model under missingness

For the second example, outcome data are simulated using a binomial model defined by

$$\log(\Pr(Y = 1|W)) = \beta_0 + W_1\beta_1 + \ldots + W_d\beta_d.$$

In all settings, the probability of missingness is given by

$$\text{logit}(\Pr(\Delta = 1|W)) = -0.25 + 1W_{d-1} - 1.5W_d,$$

and when $\Delta = 0$, Y is missing.

Example 2: Working log-linear regression model under missingness

We focus here on coefficients of the working log-linear regression model

$$\log\left[\Pr(Y = 1 \mid W_1 = w_1, \ldots, W_d = w_d)\right] = \alpha + \beta_1 w_1 + \ldots + \beta_d w_d \ .$$

Assuming that Y is missing at random given W, that is, that Y and Δ are independent conditionally upon W, the parameter

$$\psi_j := \frac{cov\{W_j, \log E[\Pr(Y = 1 \mid \Delta = 1, W) \mid W_j]\}}{var(W_j)} \tag{2}$$

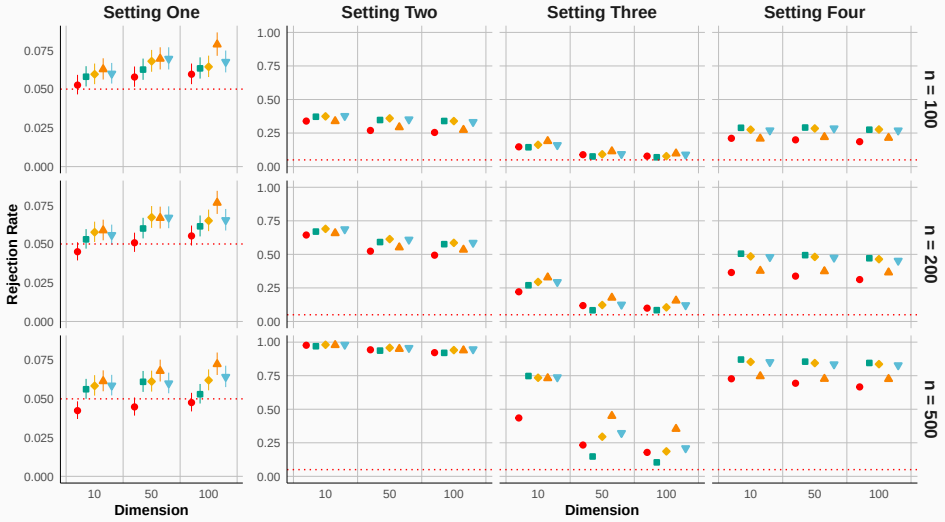
identifies the coefficient associated to W_j in this working model, and simplifies to β_j when the working log-linear model holds.

Example 2: Working log-linear regression model under missingness

In each setting, the vector of covariates $W = (W_1, \ldots W_d)$ is drawn from a multivariate normal with mean zero and covariance matrix Σ, where $\Sigma_{i,j} = 1$ for $i = j$ and $\Sigma_{i,j} = 0.5$ for $i \neq j$ in all four settings.

- In the first (null) setting, $\beta_1, \ldots, \beta_d = 0$.
- In the second setting, $\beta_1 = 0.6$ and $\beta_2, \ldots, \beta_d = 0$.
- In the third setting, $\beta_1, \ldots, \beta_5 = 0.32$, $\beta_6, \ldots, \beta_{10} = -0.32$ and $\beta_{11}, \ldots, \beta_d = 0$.
- In the last setting, $\beta_1, \ldots, \beta_5 = 0.03375$, $\beta_6, \ldots, \beta_{10} = 0.0675$ and $\beta_{11}, \ldots, \beta_d = 0$.

Example 2: Working log-linear regression model under missingness



Application of method

We applied the testing framework to data from the HVTN 505 clinical trial, a phase IIB preventative HIV vaccine efficacy trial[6].

- Secondary analyses studied the effects of the vaccine immune response on the risk of infection.
- This response was measured using a large number of biomarkers, including antibodies, T cells, and $F_{c\gamma}$ receptors.
- One such analysis measured the immune response among 25 cases and 125 randomly sampled frequency-matched vaccine controls.

[6] Neidich et al. 2019.

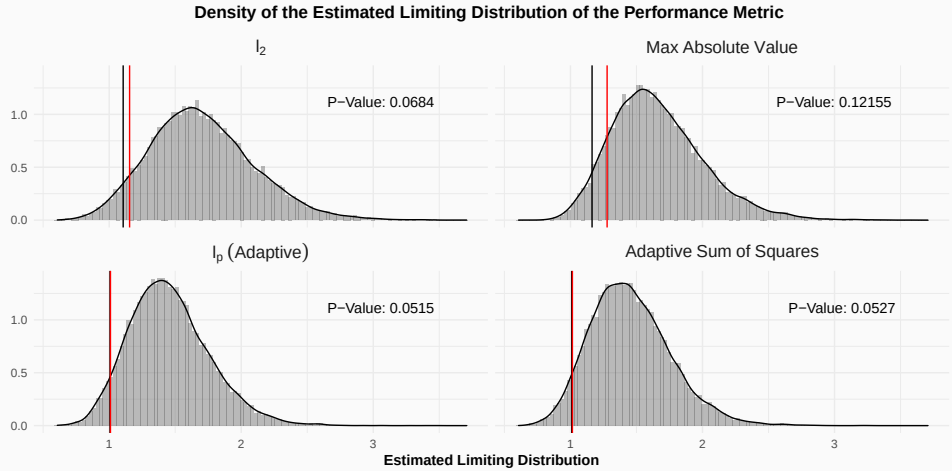
Application of method

Thank you to:

- Brian Williamson, Fred Hutch
- Peter Gilbert, Fred Hutch
- Youyi Fong, Fred Hutch

We conducted a nearly identical analysis, with a different measure of association. Here we show a test for a single biomarker set and test if any biomarkers in this set are associated with HIV infection.

Application of method



Next steps

What we have done so far on this project

- Derivation of testing procedure
- Multiple examples
- Data application
- Derivation of theoretical results
- Completed manuscript

What remains to be done

- Journal submission

Project Two: Extension to functional null hypotheses

Motivation

It is a common for a scientific study to be interested in the effect of some continuous variable (denoted by Λ) on the outcome of interest, Y.

- It is common to discretize Λ when estimating these effects.
- This can allow for interpretable contrasts between groups with different levels of Λ.
- However, this discretization could potentially over-simplify the true relationship. When causal parameters are of interest, this can falsify some causal assumptions.

We will consider studying the relationship between Λ and Y using a function that takes Λ as an argument.

Motivation

For this project, we study the relationship between body mass index (BMI) and the presence of a T cell response in data from 11 phase I/II clinical trials conducted through the HIV Vaccine Trials Network.

Specifically, we seek to test the null hypothesis that the dose-response curve relating BMI to the probability of having a positive T cell response is flat.

To test this hypothesis, we will study a standardized [7] primitive function of the dose response curve. This standardized primitive function will be equal to zero everywhere if and only if the dose response curve is flat.

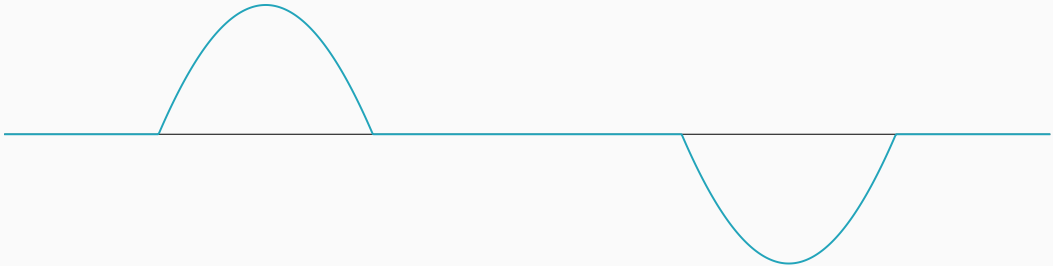
[7] The primitive minus the primitive's closest linear projection.

To formalize a test, let f be the function of interest, defined on $A \subset \mathbb{R}$. We wish to test if f is equal to zero over all of A:

$$H_0 : f(\lambda) = 0 \text{ for all } \lambda \in A \quad \text{versus} \quad H_1 : f(\lambda) \neq 0 \text{ for some } \lambda \in A \, .$$

- Values other than zero can easily be considered.
- Now the parameter value $\Psi(P)$ is the function f rather than the vector ψ.

The new null hypothesis

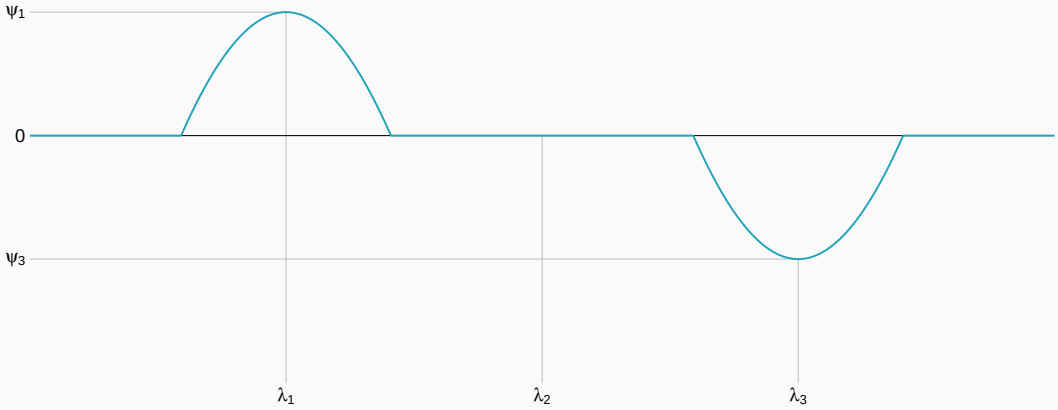


An initial testing method

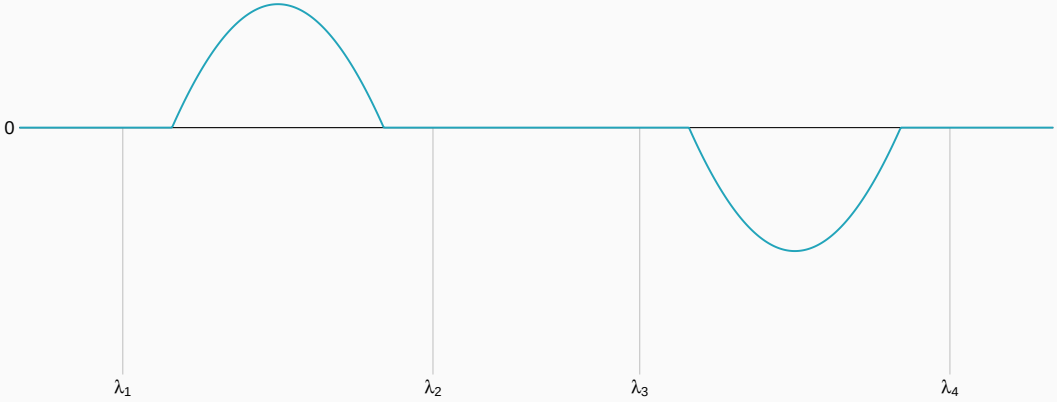
A possible approach for constructing a test of the above hypothesis consists of first defining $\psi := (\psi_1, \psi_2, \ldots, \psi_d)$ for $\psi_j := f(\lambda_j)$ for some collection of grid points $\underline{\lambda}_d := (\lambda_1, \lambda_2, \ldots, \lambda_d)$ and then testing the finite-dimensional hypothesis $\psi = \underline{0}$ as described in our previous work.

- If the null hypothesis $\psi = \underline{0}$ fails to hold, then the functional null hypothesis will also fail to hold.
- However, the converse need not be true.

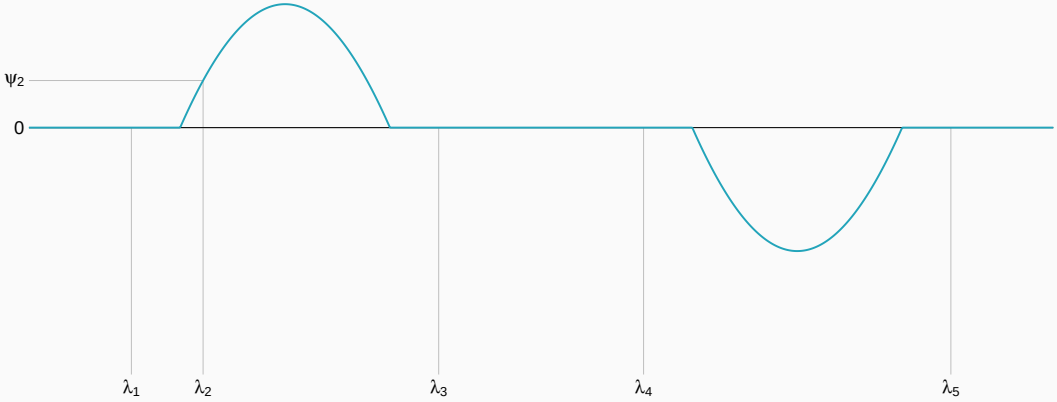
One vector-based approximation



Another vector-based approximation



Yet another vector-based approximation



Increasing the number of grid points

Typically, as the dimension of ψ increases, the set of functions for which H_0 holds for the vector-based null hypothesis but not for the functional null hypothesis becomes small.

- A potential downside to increasing the number of grid points is that the function must be estimated at a larger number of points.
- The number of grid points needed in practice can be expected to depend on the amount of smoothness assumed for the function f, with greater smoothness allowing a coarser grid without too much loss of power.

Project goals

The primary goal of this project is to extend the theory derived in the first project to describe a framework that performs well for testing the functional null hypothesis.

- The theory shown previously does not provide guarantees for when dimension grows with sample size, as we require here.
- It will be important to show that when using a smooth function to define ψ, the dimension can grow with sample size.

Norms of functions

To use the theoretical results found earlier, we will extend the definition of norms beyond vectors to functions. Define

- Q: a probability distribution on A.
- $\|\cdot\|_Q$: a norm function indexed by Q.

Here, we define $\|\cdot\|_Q^p$ by

$$f \mapsto \|f\|_Q^p = \left[\int_A |f(\lambda)|^p \, dQ(\lambda)\right]^{1/p}, p \geq 1.$$

For simplicity, here we will take $A = [0, 1]$ and Q_0 to be a uniform distribution.

Visualizing $\|f\|_{Q_0}^1$

Comparison of norms

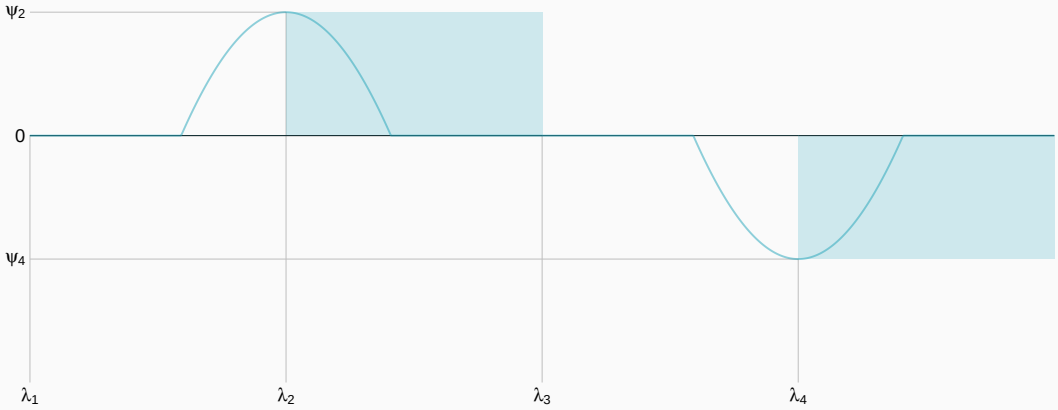
To compare the functional norm to the corresponding vector-based norm, let

- $\underline{\lambda}_d$ be d independent draws from Q_0,
- Q_d be empirical distribution of $\lambda_1, \lambda_2, \ldots, \lambda_d$, and
- $f(\underline{\lambda}_d) := (f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_d))$.

Now, note that:

$$\frac{1}{\sqrt[p]{d}} \|f(\underline{\lambda}_d)\|^p = \sqrt[p]{\frac{1}{d} \sum_{i=1}^{d} |f(\lambda_i)|^p} = \|f\|_{Q_d}^p$$

Visualizing $\|f\|_{Q_4}^1$



Defining a functional test

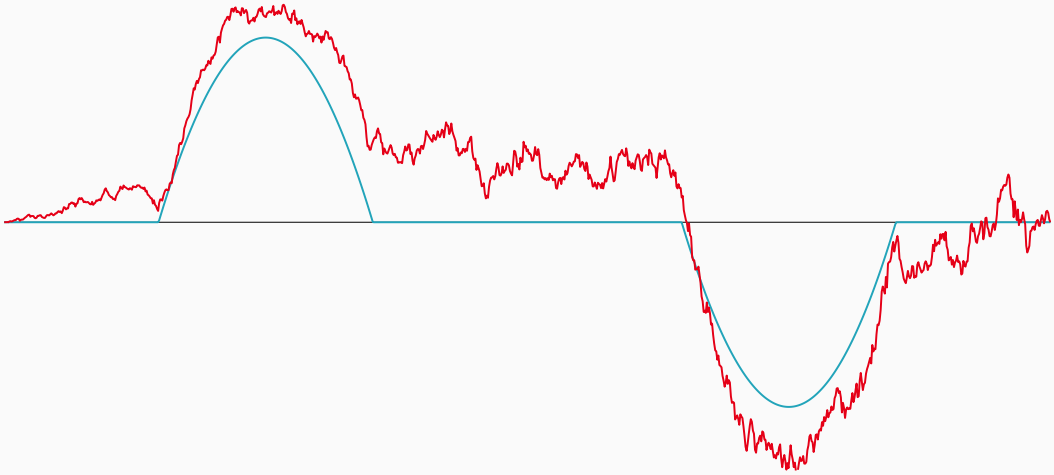
To define the proposed testing procedure, suppose that we have an estimator f_n of f such that

$$\{\sqrt{n}\,[f_n(\lambda) - f(\lambda)] : \lambda \in A\} \rightsquigarrow \{\mathbb{G}_P^*(\lambda) : \lambda \in A\}$$

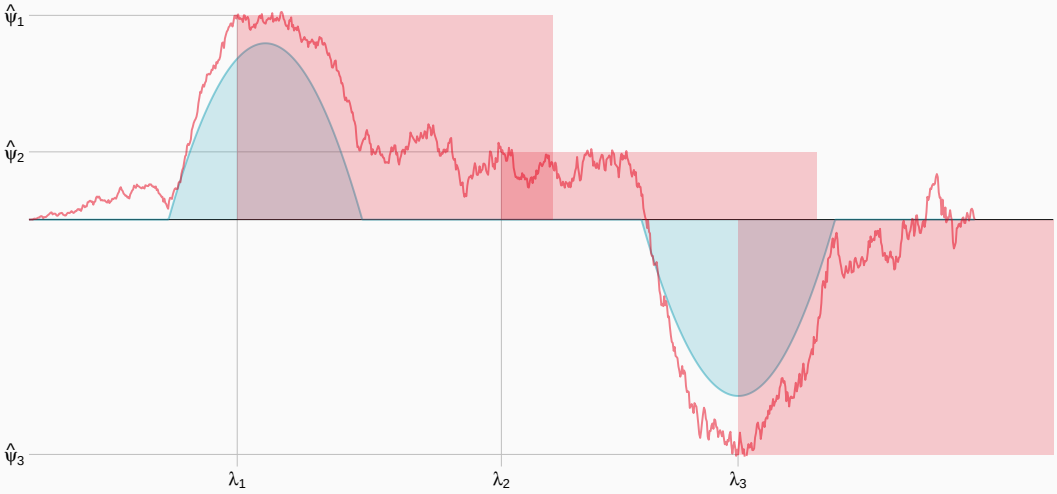
holds, where $\mathbb{G}_P^*$ is some tight mean-zero Gaussian process indexed by P.

For now, consider the random variable $\|f_n\|_{Q_d}$. We expect that as we grow the size of the grid (by increasing d) and sample size (n) the above random variable will converge in probability to $\|f\|_{Q_0}$.

Comparing f and f_n



Comparing $\|f\|_{Q_0}^1$ and $\|f_n\|_{Q_3}^1$



Previous work

Previous work[8] has shown that for some classes of functions and function estimators, that as $n \to \infty$,

$$\|f_n\|_{Q_n}^p \xrightarrow{p} \|f\|_Q^p \ .$$

Additionally, this work has

- found methods of estimating the limiting distribution of $\sqrt{n}f_n$ and $\|\sqrt{n}f_n\|_{Q_n}$ under the null, and
- shown tests defined using $\|\sqrt{n}f_n\|_{Q_n}$ as their test statistic achieve type-1 error control, consistency, and non-trivial local unbiasedness.

[8] Westling 2020; Westling and Carone 2020; Westling, van der Laan, and Carone 2020.

Previous work

While this work does provide general results for constructing a test, it does require users to choose which norm they would like to use.

Our project builds upon this previous work and aims to construct a test that adaptively selects across a variety of norms.

- Again, this will be done with the use of performance metrics.
- Instead of taking a vector as an argument, these **functional performance metrics** will take a function as an argument.

Functional performance metrics

Similar to functional norms, we will define functional performance metrics.

- Because performance metrics are defined using norms, the functional performance metrics will be defined in part by Q (the distribution used to define the functional norm).
- To compare functional performance metrics to their vector-based counterparts, the empirical distribution Q_d can be used.

Next steps

We will want to show that the estimated functional performance metric converges to an estimable limit as both the number of grid points and sample size increase. To do this, we will leverage the theoretical results from our original project and extend them to the problem described here.

Preliminary simulation

Next, we compare the non-adaptive test described earlier[9] with an adaptive version of this test.

- For this comparison, the same algorithm and computations are used to calculate f_n and estimate its corresponding limiting distribution.
- The tests are differentiated by the test statistic used.
  - The three non-adaptive versions of the test use $\|f_n\|_{Q_n}^1$, $\|f_n\|_{Q_n}^2$, and $\|f_n\|_{Q_n}^\infty$ as their test statistic, respectively.
  - The two adaptive tests select over either multiple ℓ_p norms, multiple sum of squares norms.

[9] Westling 2020.

Preliminary simulation

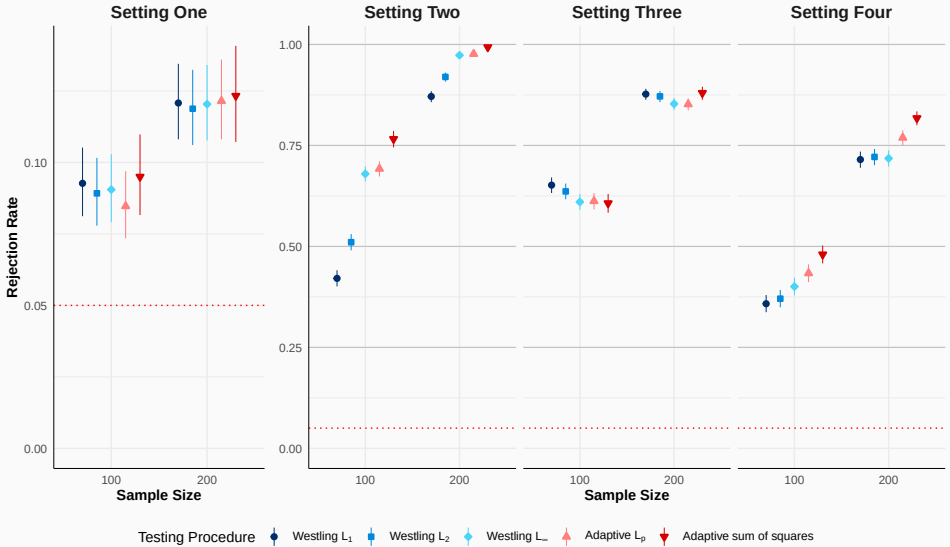
In all settings,

- The random variables W and ε_1 are drawn independently from a uniform distribution on $[0, 1]$.
- The random variable A is defined as $(W + \varepsilon_1)/2$.
- The outcome variable Y is equal to $f(A) + \varepsilon_2$ where ε_2 independent of A and is normally distributed with mean zero and standard deviation 0.4.

Each setting is differentiated by the function used to define Y.

- In the first setting $f(a) = 0$
- In the second setting $f(a) = \sin^{20}(2\pi a)$
- In the third setting $f(a) = a/2$
- In the fourth setting $f(a) = -0.65\sin(\pi(a - 1))$

Preliminary simulation



Next steps

What we have done so far

- Derivation of testing procedure
- Initial example simulations

What remains to be done

- Derivation of theoretical results
- Data application
- Manuscript completion

Project Three: Estimating Open-Label Efficacy in trials with arm switching

Introduction

The work in this project is applied to estimating the efficacy of a dapivirine vaginal ring at preventing HIV-1 infection. Data for this work come from two clinical trials:

- The ASPIRE trial: A phase 3, multi-site, double-blind randomized, clinical trial
- The HOPE trial: A phase 3b, open-label extension of the ASPIRE trial

Acknowledgements

This is joint work with

- Elizabeth Brown, Fred Hutch
- Holly Janes, Fred Hutch

We would like to thank:

- The ASPIRE/HOPE study participants
- The ASPIRE/HOPE study teams
- Daniel Szydlo, SCHARP

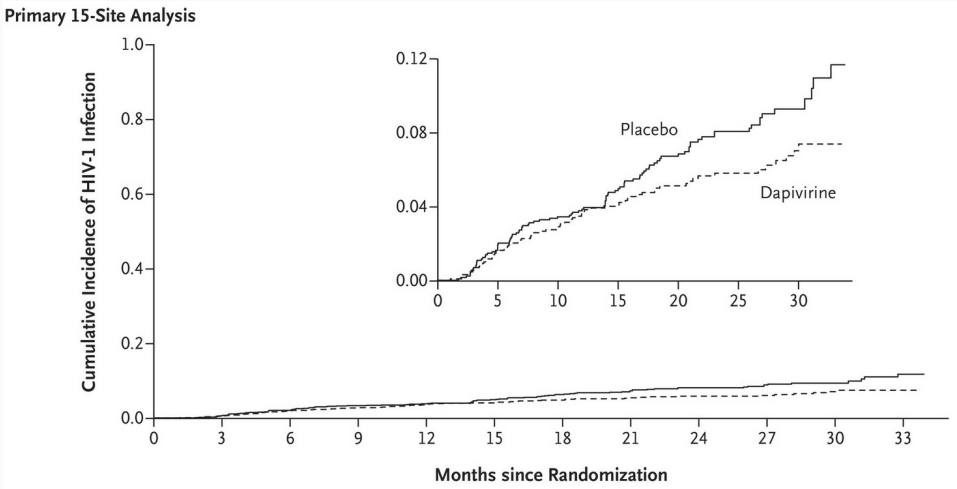
This work was funded by NIH grant 1R56AI143418-01 (PI: Janes).

The ASPIRE trial

- The ASPIRE trial found the vaginal ring containing dapivirine to provide a 27% reduction (95% CI: $1 - 46\%$) in infections.
- When two sites were excluded due to low retention and adherence (approved by the DSMB, and reviewed by regulatory agencies) a 37%(95% CI: $12 - 56\%$) reduction in infections was found[10].

[10] Baeten, Palanee-Phillips, Brown, et al. 2016.

The ASPIRE trial



No. at Risk

Placebo	1306	1280	1241	1203	1106	954	820	702	587	417	256	65
Dapivirine	1308	1285	1234	1204	1100	967	817	708	588	444	253	68

The HOPE trial

- Participants of the ASPIRE trial who were HIV negative after the trial ended were eligible to enroll in the extension trial, HOPE.
- The HOPE trial recorded an HIV-1 incidence of 2.7 per 100 person-years (95% CI 1.9–3.8%).
- In the main analysis of the HOPE trial, this incidence is compared to an expected incidence of 4.4 per 100 person-years (3.2–5.8) among a bootstrapped population from the ASPIRE placebo group, matched to the HOPE population based on age, site, and presence of a curable sexually transmitted infection during enrollment in ASPIRE[11].

[11] Baeten, Palanee-Phillips, Mgodi, et al. 2021.

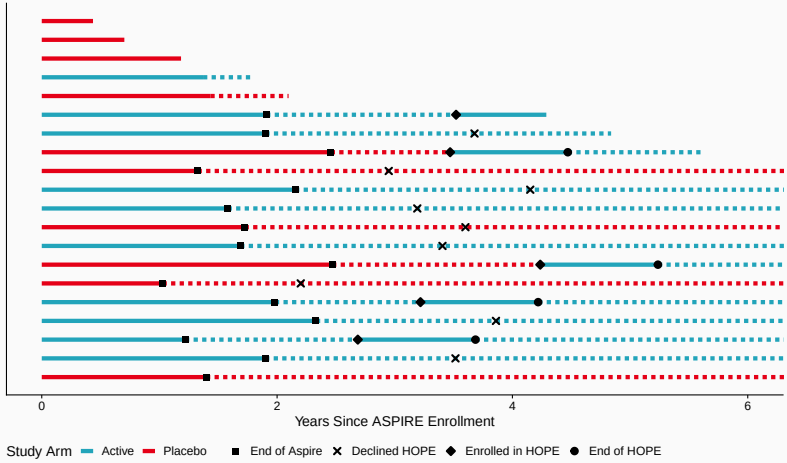
Primary Objective

The primary goal of our work is to estimate the open-label efficacy of the vaginal ring containing dapivirine in the HOPE trial.

- Because there was no placebo arm in the HOPE trial, the parameter of interest cannot be identified without additional assumptions.
- After discusssion with subject area experts, we chose assumptions that we found to be the most believable while still being sufficient to identify the parameter.
- Our estimator uses the data from both the HOPE and ASPIRE trials.

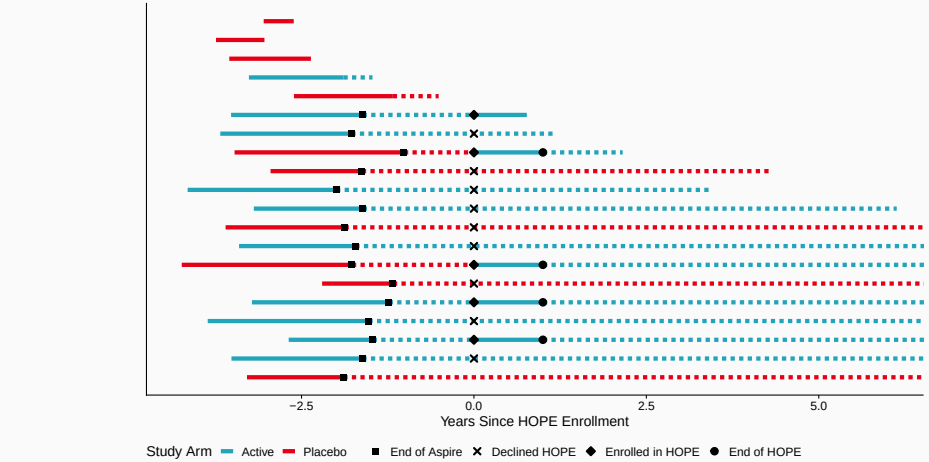
Trial logistics

Note the data shown in this and the following figures are not drawn from the trials.



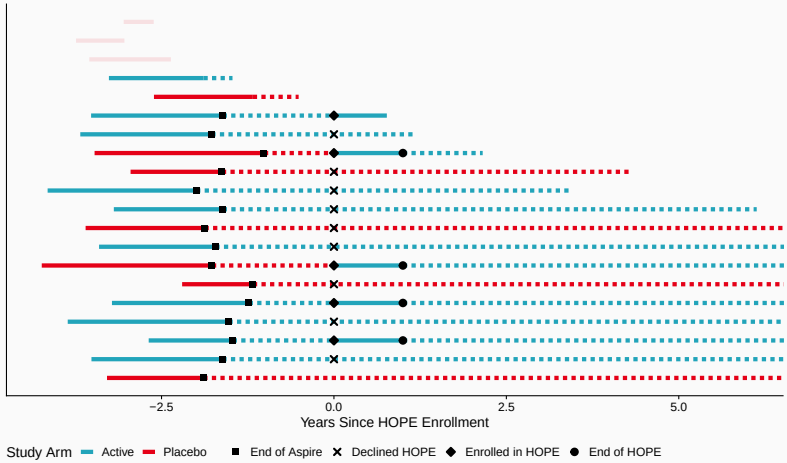
Trial logistics

When estimating open-label efficacy in HOPE, time zero chosen to be enrollment in HOPE:



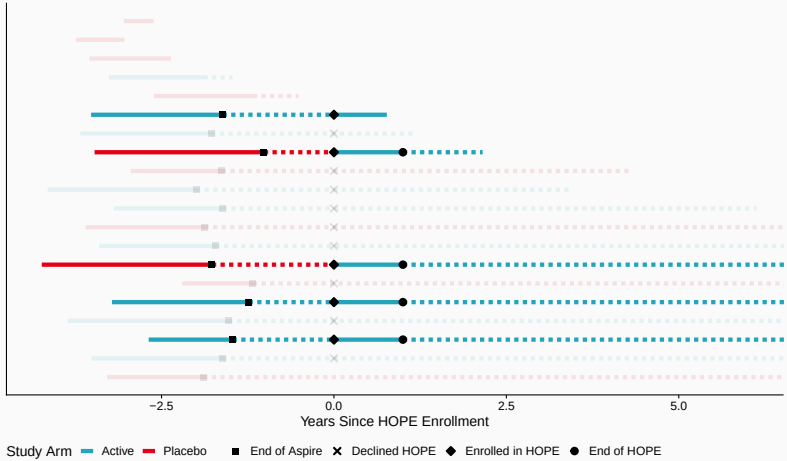
Trial logistics

Some individuals from the ASPIRE trial are ineligible for HOPE because they became HIV-1 positive during ASPIRE.



Trial logistics

Some individuals who were eligible to enroll in HOPE chose not to.



Simplification of our objective

Essentially, the objective of this project is to fill in the last cell of the following 2 by 2 table:

Year 1 HIV-1 incidence		
	Trial	
Arm	ASPIRE	HOPE
Placebo	10%	?
Active	5%	4%
Risk Ratio	0.5	?

Assuming a constant risk ratio

One simple way to estimate the missing cell is to assume a constant risk ratio:

Under this assumption, the relative risk in the ASPIRE and HOPE arms is the same, and a simple calculation provides an estimate of the HOPE placebo arm incidence:

$$4\%/0.5 = 8\%$$

Year 1 HIV-1 incidence		
	Trial	
Arm	ASPIRE	HOPE
Placebo	10%	? = 4% / 0.5
Active	5%	4%
Risk Ratio	0.5	? = 0.5

Weaknesses of the constant risk ratio assumption

We expect this assumption will not hold because the HOPE population only includes individuals who remained HIV negative throughout ASPIRE. Thus, we expect the population of HOPE to be different than that of ASPIRE.

- If the ring has different effects throughout the population, the population level relative risks will be different between the two trials.
- Also, adherence was higher in HOPE than ASPIRE, so if adherence reduces risk, we would expect larger effects in HOPE.
- Put another way, the constant risk ratio assumption assumes the risk ratio in the ASPIRE trial is transportable to the HOPE study (data from HOPE is not used to estimate the relative risk).

The constant stratified risk ratio assumption

Next, consider the assumption of a constant risk ratio within subgroups:

Year 1 HIV-1 incidence Among those age ≥ 25		
	Trial	
Arm	ASPIRE	HOPE
Placebo	20%	? = 12% / 0.7
Active	14%	12%
Risk Ratio	0.7	? = 0.7

Year 1 HIV-1 incidence Among those age < 25		
	Trial	
Arm	ASPIRE	HOPE
Placebo	8%	? = 2% / 0.38
Active	3%	2%
Risk Ratio	0.38	? = 0.38

The constant stratified risk ratio assumption

To understand how this assumption would be used to estimate open-label efficacy, we now introduce some notation. We define

- Y as the indicator of HIV-1 infection within one year of trial initiation,
- D as the indicator of being in the active arm,
- X as the baseline covariates, and
- $Y(0)$ as the counterfactual infection outcome that would have been observed, had the participant received a placebo ring.

The constant stratified risk ratio assumption can be written as:

$$\frac{E_{HOPE}[Y|X]}{E_{HOPE}[Y(0)|X]} = \frac{E_{ASPIRE}[Y|D=1,X]}{E_{ASPIRE}[Y|D=0,X]} \text{ almost surely.}$$

By rearranging the above display, and taking an expectation over X we can estimate the incidence in the counterfactual HOPE placebo arm.

The constant stratified risk ratio assumption

Estimation of

$$E_{HOPE}[E_{HOPE}[Y(0)|X]] = E_{HOPE}\left[E_{HOPE}[Y|X]\frac{E_{ASPIRE}[Y|D=0,X]}{E_{ASPIRE}[Y|D=1,X]}\right].$$

could be carried out using **targeted maximum likelihood estimation**.

This estimation strategy would allow for flexible estimation of the marginal expectations shown above while still allowing for the construction of asymptotically valid confidence intervals that account for the dependence between the ASPIRE and HOPE datasets.

The constant stratified risk ratio assumption

This assumption, while weaker, still may not hold in this setting.

Advantages

- This method will account for differences in the risk of infection between trials.
- This method will account for differences in the rings efficacy between trials resulting from differences in the study populations.

Disadvantages

- This assumption may not hold due to the differences in adherence between the two trials.

Accounting for adherence

Last we consider how to account for adherence. Adherence is not added to the list of other baseline variables because

- adherence varies across time and
- we expect adherence to be independent of infection outcome (conditional on covariates) for individuals in the placebo arm.

The time-varying aspect of adherence can still be accounted for using modern techniques in causal inference[12].

[12] J. Robins 1986; Bang and J. M. Robins 2005.

The fixed adherence, constant stratified risk ratio assumption

To describe a new assumption that accounts for adherence, we define

- A as adherence, and
- $Y(1, a)$ as the counterfactual infection outcome that would have been observed if the individual had adherence level a and was assigned a dapivirine ring.

We now consider the fixed adherence, constant stratified risk ratio assumption:

$$\frac{E_{HOPE}[Y|X]}{E_{HOPE}[Y(0)|X]} = \frac{\sum_a E_{ASPIRE}[Y(1,a)|D=1,X]\Pr_{HOPE}(A=a|X)}{E_{ASPIRE}[Y|D=0,X]} \text{ almost surely.}$$

Accounting for adherence

Similar to the constant stratified risk ratio assumption, this new assumption allows for identification of the counterfactual HOPE placebo arm incidence by solving the equivalence for $E_{HOPE}[Y(0)|X]$

- Using the fixed adherence, constant stratified risk ratio assumption, the estimator of open-label efficacy and its corresponding confidence limits will be constructed using targeted maximum likelihood estimation.
- We will build upon and adapt the current TMLE methodology and code for estimation and inference in this setting[13]

[13] van der Laan and Gruber 2011; Lendle et al. 2017.

Next steps

What we have done so far

- Data acquisition, cleaning and initial descriptives
- Identification of the causal parameter and justification of the identifying assumptions

What remains to be done

- Writing code for estimation.
- Data simulation for estimation validation.
- Sensitivity analyses
- Manuscript completion

Thank you for listening!

References

Baeten, Jared M., Thesla Palanee-Phillips, Elizabeth R. Brown, et al. (Dec. 2016). "Use of a Vaginal Ring Containing Dapivirine for HIV-1 Prevention in Women". In: *New England Journal of Medicine* 375.22. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1506110, pp. 2121–2132. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1506110. URL: https://doi.org/10.1056/NEJMoa1506110 (visited on 03/03/2021).

Baeten, Jared M., Thesla Palanee-Phillips, Nyaradzo M. Mgodi, et al. (Feb. 2021). "Safety, uptake, and use of a dapivirine vaginal ring for HIV-1 prevention in African women (HOPE): an open-label, extension study". eng. In: *The lancet. HIV* 8.2, e87–e95. ISSN: 2352-3018. DOI: 10.1016/S2352-3018(20)30304-0.

Bang, Heejung and James M. Robins (2005). "Doubly Robust Estimation in Missing Data and Causal Inference Models". en. In: *Biometrics* 61.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2005.00377.x, pp. 962–973. ISSN: 1541-0420. DOI: https://doi.org/10.1111/j.1541-0420.2005.00377.x. URL: http://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00377.x (visited on 03/18/2021).

Donoho, David and Jiashun Jin (June 2004). "Higher criticism for detecting sparse heterogeneous mixtures". en. In: *The Annals of Statistics* 32.3, pp. 962–994. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053604000000265. URL: https://projecteuclid.org/euclid.aos/1085408492 (visited on 04/18/2019).

References ii

Dudoit, Sandrine and Mark van der Laan (2008). *Multiple Testing Procedures with Applications to Genomics*. en. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-49316-9. DOI: 10.1007/978-0-387-49317-6. URL: https://www.springer.com/gp/book/9780387493169 (visited on 01/11/2021).

Hochberg, Yosef (Dec. 1988). "A sharper Bonferroni procedure for multiple tests of significance". en. In: *Biometrika* 75.4, pp. 800–802. ISSN: 0006-3444. DOI: 10.1093/biomet/75.4.800. URL: https://academic.oup.com/biomet/article/75/4/800/423177 (visited on 04/18/2019).

Holm, Sture (Jan. 1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6, pp. 65–70. DOI: 10.2307/4615733.

Lehmann, Erich L. and Joseph P. Romano (2005). *Testing Statistical Hypotheses*. en. 3rd ed. Springer Texts in Statistics. New York: Springer-Verlag. ISBN: 978-0-387-98864-1. DOI: 10.1007/0-387-27605-X. URL: https://www.springer.com/us/book/9780387988641 (visited on 01/11/2021).

Lendle, Samuel D. et al. (2017). "ltmle: An r package implementing targeted minimum loss-based estimation for longitudinal data". In: *Journal of Statistical Software, Articles* 81.1, pp. 1–21. ISSN: 1548-7660. DOI: 10.18637/jss.v081.i01. URL: https://www.jstatsoft.org/v081/i01.

McKeague, Ian W. and Min Qian (Oct. 2015). "An Adaptive Resampling Test for Detecting the Presence of Significant Predictors". In: *Journal of the American Statistical Association* 110.512, pp. 1422–1433. ISSN: 0162-1459. DOI: 10.1080/01621459.2015.1095099. URL: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2015.1095099 (visited on 04/18/2019).

Miller, Rupert G. Jr (1981). *Simultaneous Statistical Inference*. en. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag. ISBN: 978-1-4613-8124-2. URL: https://www.springer.com/la/book/9781461381242 (visited on 06/06/2019).

References iii

Neidich, Scott D. et al. (Nov. 2019). "Antibody Fc effector functions and IgG3 associate with decreased HIV-1 risk". en. In: *The Journal of Clinical Investigation* 129.11, pp. 4838–4849. ISSN: 0021-9738. DOI: 10.1172/JCI126391. URL: https://www.jci.org/articles/view/126391 (visited on 08/04/2020).

Pan, Wei et al. (Aug. 2014). "A Powerful and Adaptive Association Test for Rare Variants". en. In: *Genetics* 197.4, pp. 1081–1095. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.114.165035. URL: https://www.genetics.org/content/197/4/1081 (visited on 07/12/2019).

Robins, James (Jan. 1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". en. In: *Mathematical Modelling* 7.9, pp. 1393–1512. ISSN: 0270-0255. DOI: 10.1016/0270-0255(86)90088-6. URL: https://www.sciencedirect.com/science/article/pii/0270025586900886 (visited on 03/29/2021).

S. Holland, Burt and Margaret DiPonzio Copenhaver (July 1988). "Improved Bonferroni-Type Multiple Testing Procedures". In: *Psychological Bulletin* 104, pp. 145–149. DOI: 10.1037/0033-2909.104.1.145.

van der Laan, Mark and Susan Gruber (Aug. 2011). "Targeted Minimum Loss Based Estimation of an Intervention Specific Mean Outcome". In: *U.C. Berkeley Division of Biostatistics Working Paper Series*. URL: https://biostats.bepress.com/ucbbiostat/paper290.

Westling, Ted (Dec. 2020). "Nonparametric Tests of the Causal Null With Nondiscrete Exposures". In: *Journal of the American Statistical Association* 0.0, pp. 1–12. ISSN: 0162-1459. DOI: 10.1080/01621459.2020.1865168. URL: https://doi.org/10.1080/01621459.2020.1865168 (visited on 04/05/2021).

Westling, Ted and Marco Carone (Apr. 2020). "A unified study of nonparametric inference for monotone functions". In: *The Annals of Statistics* 48.2, pp. 1001–1024. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/19-AOS1835. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-2/A-unified-study-of-nonparametric-inference-for-monotone-functions/10.1214/19-AOS1835.full (visited on 04/05/2021).

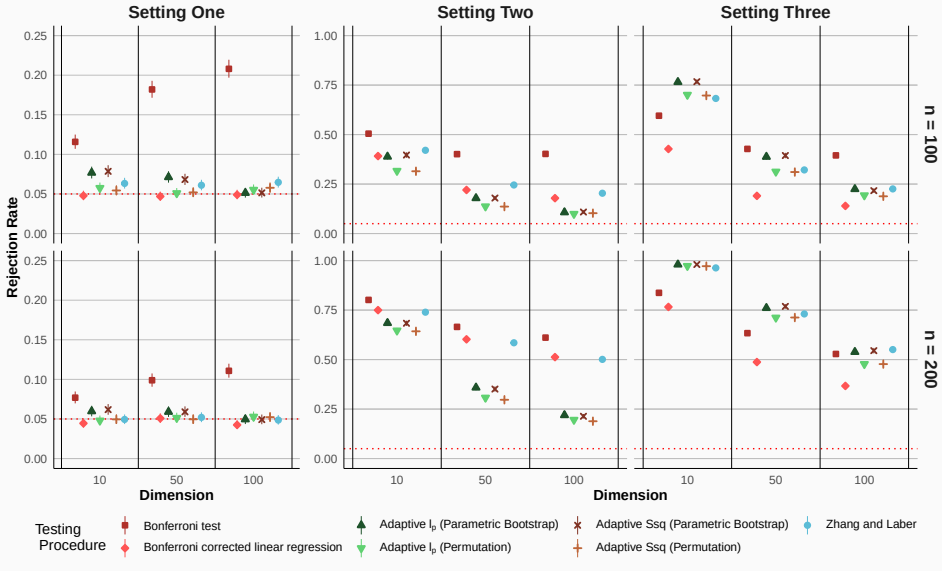
References iv

Westling, Ted, Mark van der Laan, and Marco Carone (Jan. 2020). "Correcting an estimator of a multivariate monotone function with isotonic regression". In: *Electronic Journal of Statistics* 14.2, pp. 3032–3069. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/20-EJS1740. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-14/issue-2/Correcting-an-estimator-of-a-multivariate-monotone-function-with-isotonic/10.1214/20-EJS1740.full (visited on 04/05/2021).

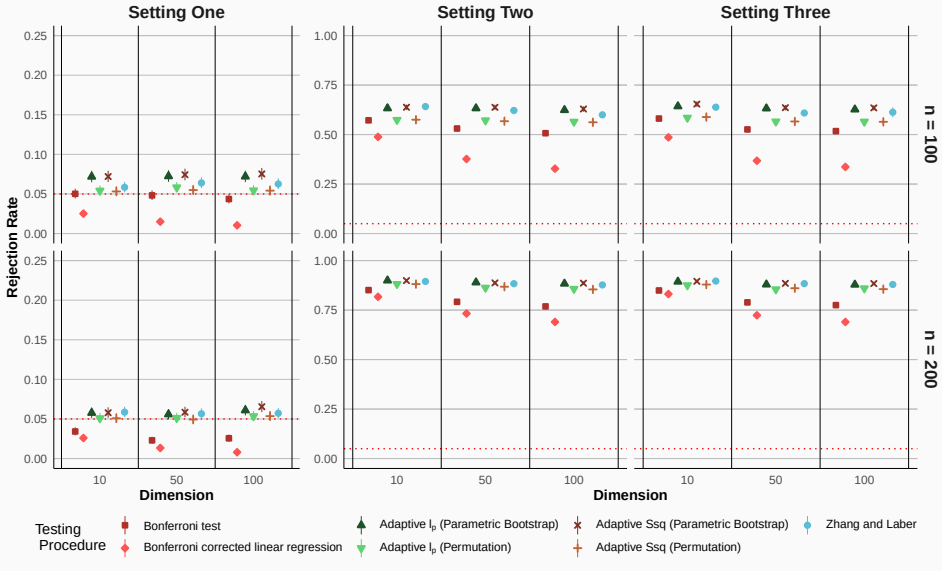
Xu, Gongjun et al. (Sept. 2016). "An adaptive two-sample test for high-dimensional means". en. In: *Biometrika* 103.3, pp. 609–624. ISSN: 0006-3444. DOI: 10.1093/biomet/asw029. URL: https://academic.oup.com/biomet/article/103/3/609/1744173 (visited on 07/12/2019).

Zhang, Yichi and Eric B. Laber (Oct. 2015). "Comment". In: *Journal of the American Statistical Association* 110.512, pp. 1451–1454. ISSN: 0162-1459. DOI: 10.1080/01621459.2015.1106403. URL: https://amstat.tandfonline.com/doi/full/10.1080/01621459.2015.1106403 (visited on 04/18/2019).

Example 1 (Between X correlation is 0)



Example 1 (Between X correlation is 0.8)



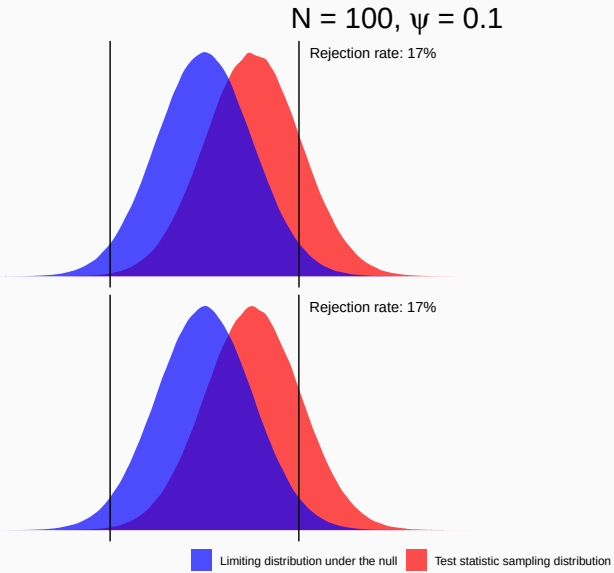
Considering two tests

Consider two different tests. While both tests have an estimate of ψ that is consistent, the second test has a standard error that shrinks towards zero at a rate slower than $n^{1/2}$.

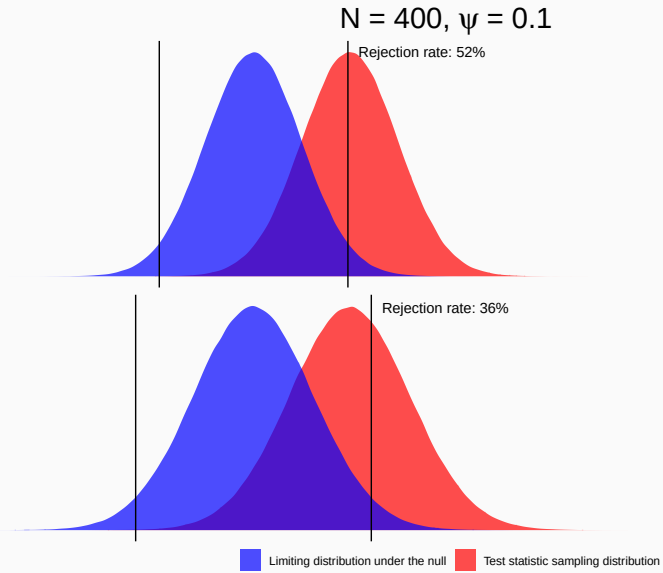
- It is still possible for both tests to achieve type-1 error control and consistency.
- This can happen as long as the standard error of the estimator does not shrink too slowly.

In the following slides, we compare the sampling distribution of test statistic $\sqrt{n}\hat{\psi}$ to the estimated limiting distribution for $\sqrt{n}\hat{\psi}$ under the null.

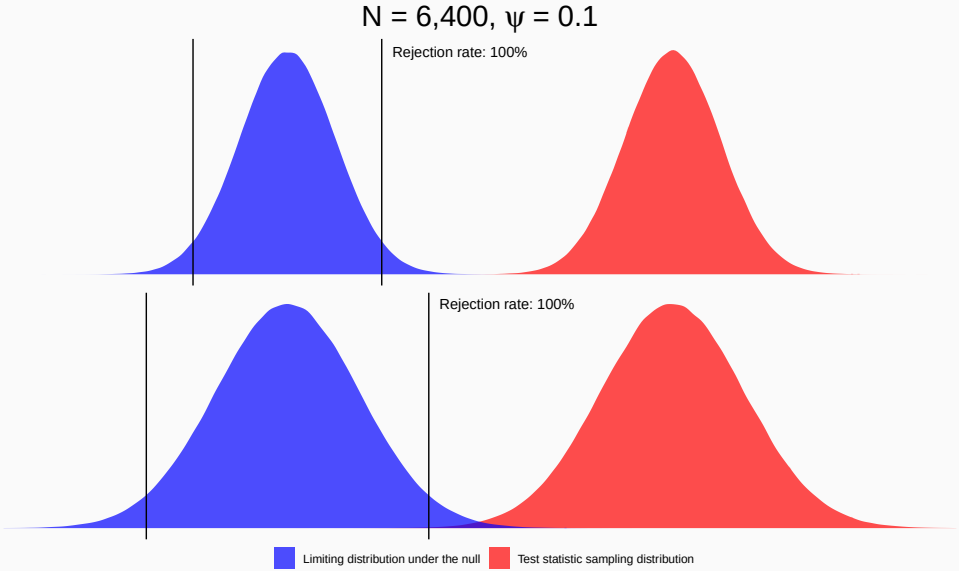
Considering two tests



Considering two tests



Considering two tests

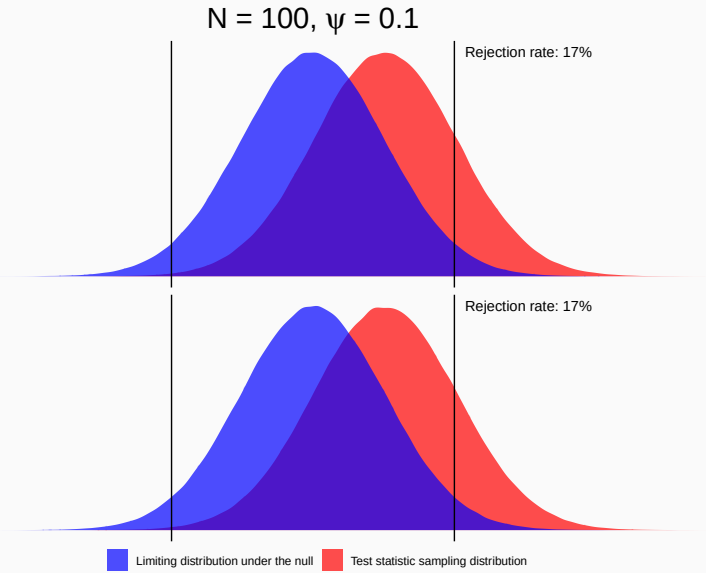


Considering two tests

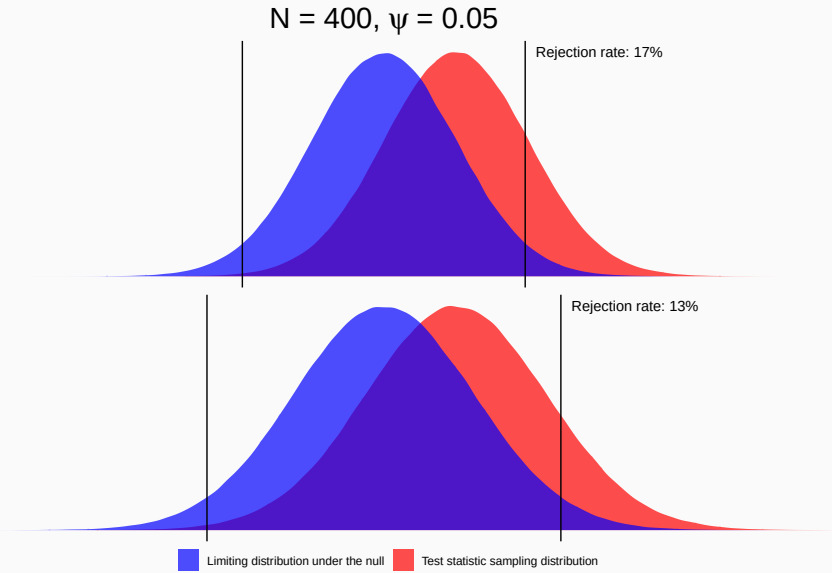
We now consider the above tests under a sequence of local alternatives.

- Under this sequence of local alternatives as sample size grows, the true value of ψ shrinks towards the null ($\psi = 0$ at a rate of $1/sqrtn$).
- Under this sequence of alternatives, a test based on an inefficient estimator will have a rejection rate approaching α.

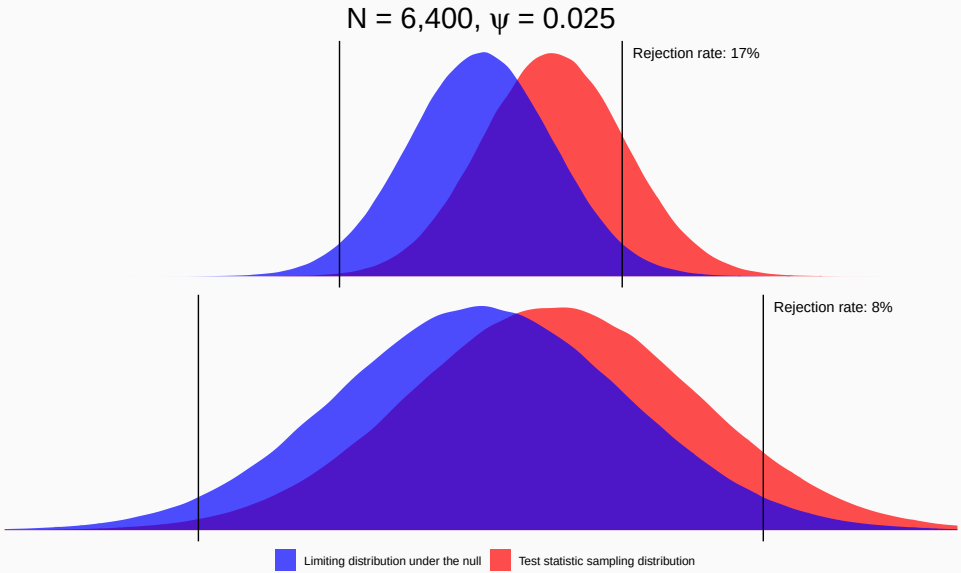
Considering local alternatives



Considering local alternatives



Considering local alternatives



Coefficients of a working effect modification model for randomized trials

In our third example, we consider the data unit $Z = (W_1, W_2, \ldots, W_a, A, Y)$, where $W_1, W_2, \ldots, W_d$ represent real-valued covariates, A is a binary treatment variable, and Y is a binary outcome of interest.

Coefficients of a working effect modification model for randomized trials

We wish to consider the interaction coefficient of the least-squares projection of the true conditional counterfactual success probability onto the logit-linear regression model

$$\text{logit}\,\text{pr}(Y(a) = 1 \mid W_j = w) = \alpha_j + \beta_j a + \gamma_j w + \delta_j wa \ .$$

This coefficient provides a measure of the degree to which W_j modifies the effect of A on Y. For simplicity, we assume that treatment allocation is randomized, so that the counterfactual outcome $Y(a)$ corresponding to treatment level a is independent of A for each $a \in \{0, 1\}$. In this case, the parameter

$$\Psi_j(P) := \underset{\delta}{\text{argmin}} \min_{(\alpha,\beta,\gamma)} E_P \left\{\text{logit}\, E_P[\, P(Y = 1 \mid W, A) \mid W_j\,] - (\alpha + \beta A + \gamma W_j + \delta W_j A)\right\}^2$$

identifies the interaction coefficient in this working model, and again, simplifies to δ_j when the working logit-linear structural model above.

Coefficients of a working effect modification model for randomized trials

In all settings for this example, the working model described in section is used to take draws from Y and the vector of covariates W is draw from a multivariate normal with mean zero and covariance matrix Σ where $\Sigma_{i,j} = 1$ for $i = j$ and $\Sigma_{i,j} = 0.5$ for $i \neq j$.

The random variable A is drawn from a binomial distribution independently of W.

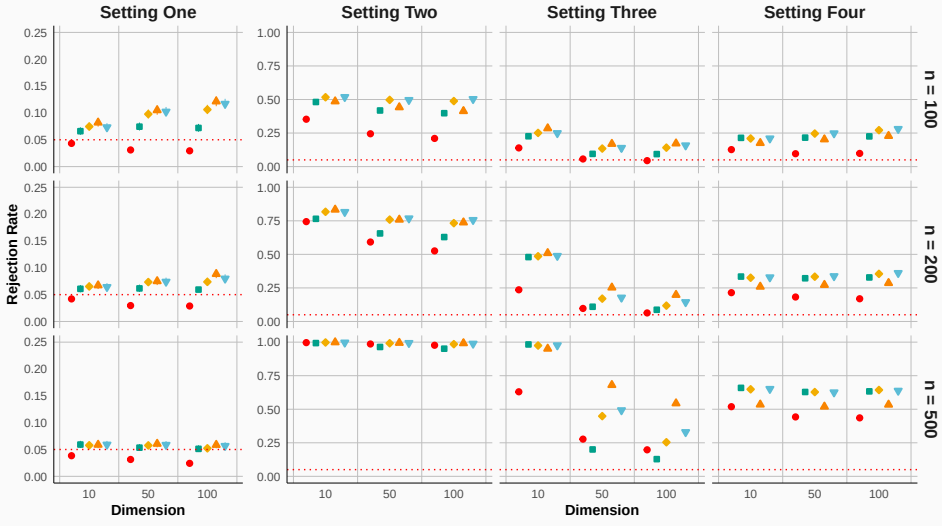
For all three settings,

$$\text{logit}\left(Pr(Y = 1|w, a)\right) = \alpha a + \sum_{i=1}^{d} \beta_i w_i + \sum_{j=1}^{d} \gamma_j w_j a.$$

Additionally, $\alpha = 0.2$, $\beta_1, \ldots, \beta_{d/2} = 2/\sqrt{d}$, and $\beta_{d/2}, \ldots, \beta_d = 0$ in each setting.

- In setting one: $\gamma_1, \ldots, \gamma_d = 0$.
- In setting two: $\gamma_1 = 3$ and $\gamma_2, \ldots, \gamma_d = 0$.
- In setting three: $\gamma_1, \ldots, \gamma_5 = 0.6$, $\gamma_6, \ldots, \gamma_{10} = -0.6$, and $\gamma_{11}, \ldots, \gamma_d = 0$.
- In setting four: $\gamma_1, \ldots, \gamma_5 = 0.09$, $\gamma_6, \ldots, \gamma_{10} = 0.18$, and $\gamma_{11}, \ldots, \gamma_d = 0$.

Example 3


Setting One
Setting Two
Setting Three
Setting Four
n = 100
n = 200
n = 500
Rejection Rate
Dimension
Dimension
Testing Procedure
Bonferroni
l_2
Adaptive l_p
l_∞
Adaptive sum of squares

Functional performance metrics

Let Γ_∞ denote a functional performance metric. For now, consider the functional acceptance rate performance metric:

$$\Gamma_\infty(f, P, Q) := \Pr\{\|f + \mathbb{G}_P^*\|_Q > c_\alpha(P, Q)\} \text{ with}$$
$$c_\alpha(P, Q) := \inf\{c : \Pr\{\|\mathbb{G}_P^*\|_Q > c\} < \alpha\} \ .$$

Comparing the two acceptance rate performance metric

Let $P(\underline{\lambda}_d)$ be the probability distribution of $Z_{P_d} := (\mathbb{G}_P^*(\lambda_1), \mathbb{G}_P^*(\lambda_2), \ldots, \mathbb{G}_P^*(\lambda_d))$. Also, let Σ_{P_d} be the covariance matrix of Z_{P_d}. Recalling the vector-based performance metric,

$$\Gamma(\omega, \Sigma_{P_d}) = Pr(\|Z + \omega\| \leq c_\alpha), \text{ where } Z \sim N(0, \Sigma_{P_d}) \ ,$$

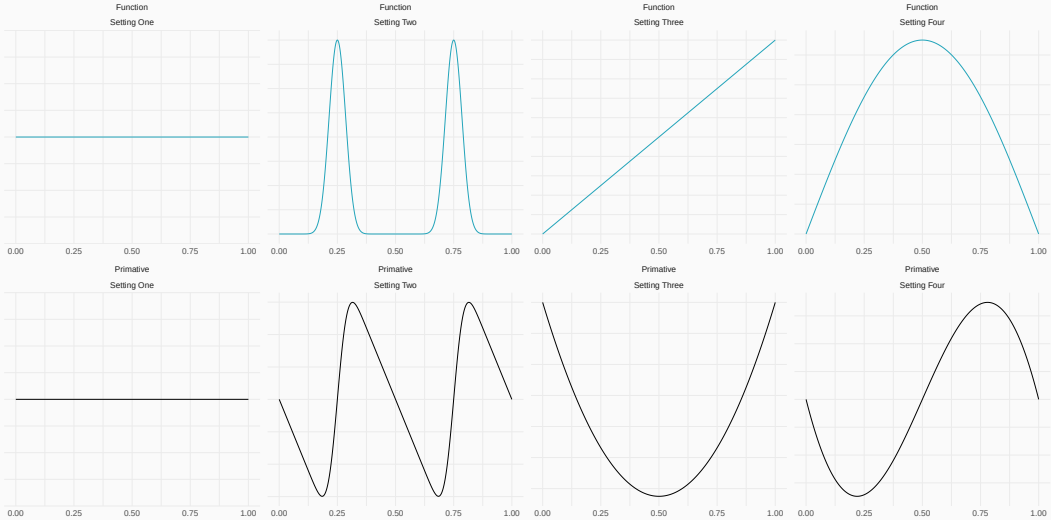
and the functional performance metric,

$$\Gamma_\infty(f, P, Q) := \mathrm{Pr}\{\|f + \mathbb{G}_P^*\|_Q > c_\alpha(P, Q)\} \ ,$$

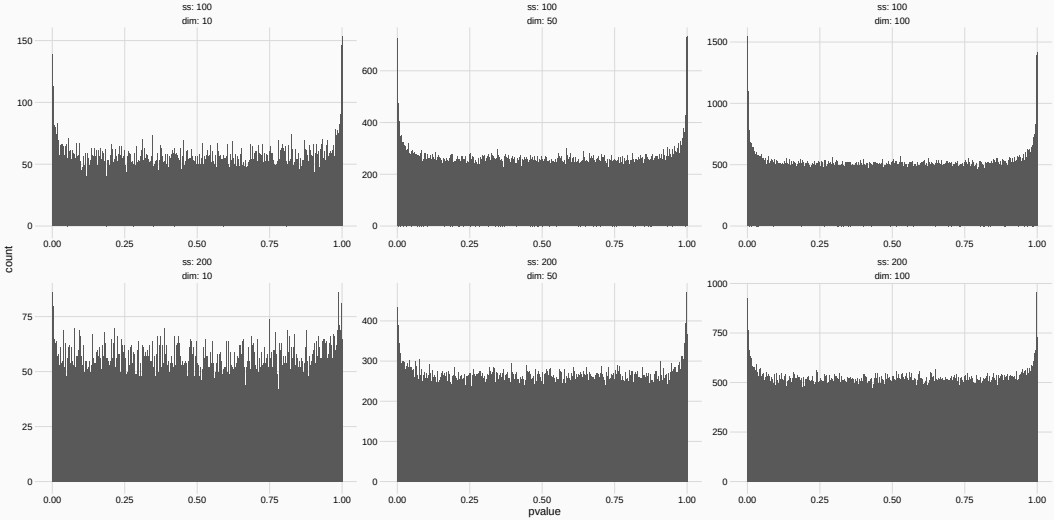
note that

$$\Gamma(f(\underline{\lambda}_d), \Sigma_{P_d}) = \Gamma_\infty(f, P, Q_d) \ .$$

Preliminary simulation



Bonferroni p-values



Bonferroni p-values