Abstract

In many scientific disciplines including genetics, neurology, and vaccine development researchers are often interested in understanding the relationship between an outcome and a large set of covariates. A first step to investigating such relationships is to screen covariates for an association with the outcome. This screening can be achieved using a variety of methods, from general methods such as those based on Bonferroni adjustment to techniques tailored to a single problem. Existing general methods that can be applied across a variety of problems frequently have lower power. Tailor-made procedures tend to attain higher power by building their procedures around problem-specific information that makes adaptation to new settings difficult. We propose a general framework to test for the existence of an association between an outcome and any number of covariates in which the estimator is adaptively selected to optimize an estimate of local power. We present theoretical asymptotic guarantees for our test under fixed and local alternatives. We then show numerically that tests created using our framework can perform nearly as well as tailor-made methods. Finally, we develop tests in two settings in which tailor-made methods are not available.

1 Introduction

To broaden scientific understanding in a field it is typical to seek an understanding of the relationships between various observations. It is common for the set of potentially related observations to be quite large while the number of truly related observations is small. Thus an important first step is to find which variables are truly related to each other. This initial screening step has become increasingly important as the research collect information on more and more potentially related observations.

One method to check for the existence of such relationships or associations is to test the association of each covariate with the outcome. Because of the many tests this procedure could consider, it is desirable for the testing procedure to account for the potentially large number of hypotheses that are simultaneously being tested when using this approach. Work with simultaneous hypothesis testing began with Tukey in 1953 [Miller]. Previous work by Bonferroni was used by [Dunn, a,b] to come up with some of the first multiple hypothesis testing procedures. Further improvements were proposed by [Hochberg, Holm, S. Holland and DiPonzio Copenhaver]. Bonferroni-based correction procedures are easy to apply to already existing tests, while guaranteeing family-wise error control. However because these tests ignore the joint distribution of the test statistics, they suffer from low power, especially in cases where the probability of rejecting each hypothesis is highly correlated.

Newer procedures [Donoho and Jin] and add other papers here using modern statistical theory to provide asymptotic guarantees for their methods and to account for the correlation between parameter estimates.

However, these methods are made for specific parameters and data generating mechanisms and do not account for the irregularity of the estimator on which their test is based. This irregularity can cause poor performance in scenarios in which the probablility of rejecting the null is far away from the α level of the test and one. Subsequent work [McKeague and Qian, Pan et al., Xu et al.] addressed these concerns by accounting for the flexible nature of these adaptive tests. However these methods were still created for a specific measure of association and data generating mechanisms in mind. Thus while much has been done to improve the performance of tests in certain settings investigators may still use bonferroni based methods because advanced methods do not exist for their particular situation As a result, investigators may still choose to use bonferroni correction based tests if they are interested in a different measure of association or are unsure of how deviations from the assumed data generating mechanism could effect the properties of the testing procedure.

In this article we propose a testing procedure that can be used for a wide variety of data generating mechanisms and parameters of interest. This test also achieves comparable power to tailor made procedures in some settings. Section 2 describes the data generating mechanisms and the corresponding parameters that are considered to evaluate the performance of the test and the competing test. Section 3 describes the testing procedure. Section 4 provides the theoretical guarantees of the described procedure. Section 5 shows the performance of our testing procedure in the three considered simulation settings. Section 6 shows the application of our method to real world data. Section 7 summarizes the findings of this article and notes potential weaknesses and extensions of our procedure.

2 Working Examples

Throughout all examples let $X_1, \ldots, X_n$ be independent identically distributed draws from some distribution P, and let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. Let $X_i = (Y_i, W_{1i}, \ldots, W_{di})$, $i \in \{1, \ldots n\}$ where Y is the outcome of interest, and each W is a covariate. Let $\psi_1 = \Psi_1(P), \ldots, \psi_a = \Psi_a(P)$ be measures of association between Y and some combination of the W_js. While the results found in this article are valid for any integer a, for the remainder of this article assume $a = d$, the number of covariates. Also in this article, let ψ_j correspond to a measure of association between Y and W_j. The null hypothesis for our test will be the strong null:

$$H_0 : \psi_1 = \psi_2 = \cdots = \psi_a = 0 \text{ versus } H_1 : \psi_j \neq 0 \text{ for some } j \in \{1, \ldots, a\}.$$

Finally, let $\mathcal{M}$ denote the set of all possible distributions, and let $\mathcal{M}_0 \subset \mathcal{M}$ be the subset of distributions in $\mathcal{M}$ satisfying H_0. We consider three different simulated data settings to study the performance of the test.

2.1 Correlation Parameter

In the first example, our method is compared to a Bonferroni correct marginal testing method and the parametric bootstrap analog of the adaptive resampling test described in [Zhang and Laber] that built on the adaptive resampling test proposed by [McKeague and Qian]. This example uses the same settings as those used in the first example of [McKeague and Qian]. The parameter of interest $\psi_j(P)$ is the correlation between the outcome of interest and the j'th covariate. The vector of covariates in this setting will be generated from a normal distribution with mean zero and a variance covariance of Σ with Σ_{ij} equal to ρ when $i \neq j$ and equal to 1 when $i = j$. Three different models for the outcome of interest (Y) will be considered. For every setting $\varepsilon \sim N(0,1)$ and is independent of all W. In the first setting $Y = \varepsilon$, in the second setting $Y = W_1/4+\varepsilon$, and in the third setting $Y = \sum_{k=1}^{10} \beta_k W_k + \varepsilon$ where $\beta_k = 0.15$ for $k = \{1, \ldots, 5\}$, and $\beta_k = -0.1$ for $k = \{6 \ldots 10\}$. Sample sizes of 100 and 200, dimensions of 10, 50, 100, 150, and 200, and correlations (ρ) of $0, 0.5$, and 0.8 are considered. All combinations of model, sample size, dimension and correlation are considered, and the performance of each test is measured for every settings.

2.2 Missing Data Example

In the second example Y is binary variable and Δ is a missingness indicator. When $\Delta = 0$, Y is not observed. The parameter of interest is the risk ratio under a poission working model for the probability that $Y = 1$.

$$\Psi_j\left(P^{\text{full}}\right) = \frac{\text{Cov}\left(\log\left(Pr\left(Y=1|W_j\right)\right), W_j\right)}{\text{Var}(W_j)}.$$

The identifying assumption is $\Delta \perp\!\!\!\perp Y|W$, and when this assumption holds the corresponding identifiable parameter is:

$$\tilde{\Psi}_j\left(P^{\text{obs}}\right) = \frac{\text{Cov}\left(\log\left(E\left[Pr\left(\Delta Y = 1|\Delta = 1, W = W\right)|W_j\right]\right), W_j\right)}{\text{Var}(W_j)}.$$

The data are drawn from a binomial model:

$$\log(Pr(Y=1|W)) = \beta_0 + W^\top \beta.$$

and in all settings, the probability of missingness is given by

$$\text{logit}(\Pr(A=1|w)) = -0.25 + 1w_{d-1} - 1.5w_d \text{ where d is the number of covariates.}$$

There will be three settings considered which determine the values of the βs in the data generating model. While it is not necessary for the covariates to be normally distributed, in each setting the vector $\boldsymbol{W}$ is draw from a multivariate normal with mean zero and covariance matrix Σ_{DE2} where $\Sigma_{DE2,i,j} = 1$ for $i = j$ and 0.6 for $i \neq j$. A is drawn from a binomial distribution independent from W. For all three settings

$$\text{logit}\left[Pr(Y = 1|w, a)\right] = \sum_{i=1}^{d} \beta_i w_i$$

In the first setting $\beta_1, \ldots, \beta_d = 0$. In the second setting $\beta_1 = 3$ and $\beta_2, \ldots, \beta_d = 0$. In the last setting, $\beta_1, \ldots, \beta_5 = 1$, $\beta_6, \ldots, \beta_{10} = -1$ and $\beta_{11}, \ldots, \beta_d = 0$. Data are generated using all three settings with every possible combination of sample size ($n = 5,000$, or $10,000$), and dimension ($d = 10, 50, 100, 150$, or 200).

2.3 Marginal Structural Model Example

The third data example is a marginal structural model in which we test if the average treatment effect of a binary treatment A is modified by any covariates W_j. The marginal structural model for each W_j is defined by the working model:

$$\text{logit}\left[Pr(Y^{(a)} = 1|w)\right] = \beta_0 + \beta_1 a + \beta_2 w_j + \beta_3 w_j a \tag{1}$$

in which our parameter of interest is:

$$(\beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*) = \text{argmin}_{\beta_0,\beta_1,\beta_2,\beta_3} \int \left\{\text{logit}\left[Pr\left(Y^{(a)} = 1|w\right)\right] - (\beta_0 + \beta_1 a + \beta_2 w_j + \beta_3 w_j a)\right\}^2 dP(w, a)$$

The parameter of interest is β_3^*, but it is worth noting that this parameter is different from a usual logistic regression parameter because we are marginalizing over all other w_i.

In all simulation settings the working model is used to take draws from Y. In each setting, the vector $\boldsymbol{W}$ is draw from a multivariate normal with mean zero and covariance matrix Σ_{DE3} where $\Sigma_{DE3,i,j} = 1$ for $i = j$ and 0.6 for $i \neq j$. A is drawn from a binomial distribution independent from W. For all three settings

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

In every setting, $\alpha = 0.2$, $\beta_1, \ldots, \beta_{d/2} = 2/\sqrt{d}$, and $\beta_{d/2}, \ldots, \beta_d = 0$. In the first (null) setting, $\gamma_1, \ldots, \gamma_d = 0$. In the second setting, $\gamma_1 = 3$ and $\gamma_2, \ldots, \gamma_d = 0$. In the final setting $\gamma_1, \ldots, \gamma_5 = 0.15$,

$\gamma_6, \ldots, \gamma_{10} = -0.15$, and $\gamma_{11}, \ldots, \gamma_d = 0$. Data are generated using all three settings with every possible combination of sample size ($n = 1,000$, or $2,000$), and dimension ($d = 10, 30$, or 75).

3 Proposed Testing Procedure

For some test statistic $\hat{\boldsymbol{t}}$, any test of $H_0 : P \in \mathcal{M}_0$ versus $H_1 : P \notin \mathcal{M}_0$ can be characterized by checking if $\hat{\boldsymbol{t}}$ falls within acceptance region $\Theta_0(P) \subset \mathbb{R}^d$. Letting $\hat{\boldsymbol{t}} \xrightarrow{P_0} Z \sim Q(P)$ for any $P_0 \in \mathcal{M}_0$, the region $\Theta_0(P)$ can be chosen so the probability of rejection under the null is controlled asymptotically:

$$Pr_Q\{Z \notin \Theta_0(P)\} = 1 - \alpha \text{ for every } P \in \mathcal{M}_0. \tag{2}$$

While there are many regions satisfying (2), we focus for now on a particular class of regions defined using ℓ_p norms which will naturally lead to a straightforward testing procedure. For simplicity, first consider regions defined using an ℓ_2 norm:

$$\Theta_0(r) = \{\omega : \|\omega\|_2 \leq r\}. \tag{3}$$

A region satisfying (2) and (3) has a radius defined by:

$$r_\alpha(P) = \min\{r : Pr_Q(\|Z\|_2 \leq r) \geq 1 - \alpha\}. \tag{4}$$

To construct a test using the region defined above, let $\hat{\boldsymbol{\Psi}} : \mathbb{R}^d \to \mathbb{R}$ be an estimator of $\boldsymbol{\psi}$, and let $\hat{\boldsymbol{\psi}} \equiv \hat{\boldsymbol{\Psi}}(x)$ be an estimate of $\boldsymbol{\psi}$. Letting $\hat{\boldsymbol{t}} \equiv \sqrt{n}\hat{\boldsymbol{\psi}}$, suppose for now that $\hat{\boldsymbol{t}}$ converges in law to a normal distribution $Q(P)$ when P is contained in the model space $\mathcal{M}_0$. The test can be defined by

$$\text{reject } H_0 \text{ if } \|\sqrt{n}\hat{\boldsymbol{\psi}}\|_2 \geq r_\alpha(P_0), \tag{5}$$

and the corresponding p-values are defined by $\Pr_Q(\|Z\|_2 \geq \|\sqrt{n}\hat{\boldsymbol{\psi}}\|_2)$.

So far, we have outlined a method to construct a test of H_0, but it is easy to imagine other tests defined using different norms. With so many potential tests to consider, it is natural to wonder if tests based on different norms have different power, and if so how can someone select the norm with optimal power. To explore the first question, consider a simple example comparing the test described in equation (5) with a test that is identical except it uses the maximum absolute deviation or ℓ_∞ norm (maximum absolute value) in place the ℓ_2 norm. Letting $\boldsymbol{x} = (x_1, \ldots, x_d)$ then $\ell_\infty(\boldsymbol{x}) = \max\{|x_1|, \ldots, |x_d|\}$. Figure 1, panel (A)

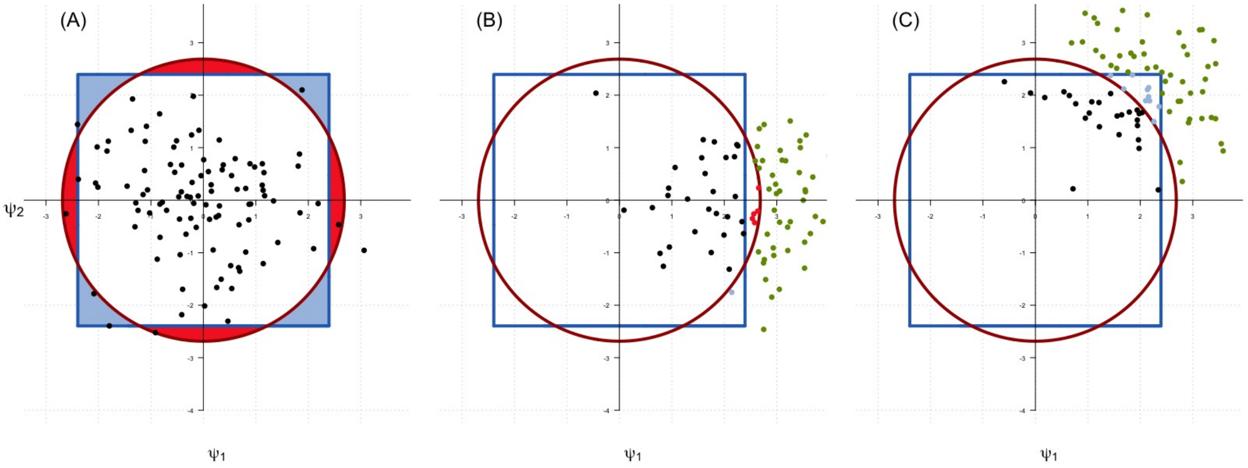

Figure 1: Plots of 100 observations from a limiting distribution of a hypothetical vector of parameter estimators in $\mathbb{R}^2$ (A) under the null, (B) under an alternative with $\psi_1 = 0, \psi_2 \neq 0$, and (C) under an alternative with $\psi_1, \psi_2 \neq 0$. The 95% quantiles for the data based on the max (blue) and ℓ_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 ℓ_∞ norm was used, but would reject H_0 if the ℓ_2 norm was used. The converse is true for the red regions. Depending on the alternative, the ℓ_∞ norm (B) or the ℓ_2 norm(C) will achieve higher power.

illustrates these two tests in $\mathbb{R}^2$. One hundred draws are taken from Y_Σ, a surrogate for the estimated limiting distribution of $\hat{\boldsymbol{t}}$. The random variable Y_Σ is a bivariate normal distribution with mean zero and identity covariance matrix. All observations in panel (A) except the five with the largest ℓ_2 norm are contained within the red circle. The blue square contains all observations in panel (A) except the five with the largest ℓ_∞ norm. The circle and square represent the acceptance regions of tests using empirical estimates of the 95^{th} percentile of $\ell_2(Y_\Sigma)$ and $\ell_\infty(Y_\Sigma)$ respectively corresponding to $r_\alpha(P)$ from (4). The same square and circle are redrawn in panels (B) and (C) to illustrate the performance of the test under two alternatives. Observations that fall within the blue shaded region would result in a rejected null hypothesis if the ℓ_2 norm was used to define the test, but not if the ℓ_∞ norm was used. The converse is true of the red shaded region. Panel B shows draws from an alternative in which $\psi_2 \neq 0$ and $\psi_1 = 0$, and panel C shows draws from an alternative in which ψ_1 and $\psi_2 \neq 0$.

While both acceptance regions are created to achieve asymptotic type one error control, depending on the alternative one test will outperform the other. Panel B shows an alternative in which only ψ_2 is non-zero. Because the max norm only considers the largest coordinate, shifting each observation in only a single direction will have larger impact on the max norm of the observations compared to the ℓ_2 norm. This trend is shown by the numerous red observations 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 that is outside the red circle, and inside the blue box. The converse trend is shown in panel C. Here, the ℓ_2 norm performs better, because it takes into account both coordinates of the shift whereas the ℓ_∞ norm can only take into account one of these coordinate shift.

The efficiency that can be gained from using the correct norm can be quite large, especially in high dimension settings. Work done by [Pinelis] states that even for dimensions as small as 3 the gains in asymptotic efficiency for selecting the correct norm can become arbitrarily large between two potential norms Review this.

3.1 Adaptive selection of a norm

In the previous section we showed that a test can be defined with a summarizing function and norm. However, the choice of norm can influence the power of the test. In many scenarios it will not be clear a priori which test will have maximal power because the power of each test depends on the unknown value of the true alternative. The procedure proposed in this section data adaptively selects a norm to attain some of the power improvements that can be attained from selecting the best norm without needing to specify the norm a priori. It will be shown via simulation that this procedure can achieve greater power than a test with a fixed norm. It will also be shown with both theory and simulation that this adaptive procedure maintains type 1 error control for large sample sizes.

To define norm data adaptive selection rules it will be important to measure the performance of each test. These measures will be referred to as performance metrics and will be denoted by Γ. While these metrics can be quite general, they should provide an ordering in which better performance results in a more extreme measure of performance. These metrics should provide reasonable comparisons across norms. For example if one were to use the ℓ_p of the test statistic as the performance metric, the first criteria would be satisfied because for most definitions of "large" as x grows large, so does $\|x\|_p$. However, the second criateria is not satisfied because $\ell_1(x) \geq \ell_2(x) \geq \cdots \geq \max(x)$, it does not satisfy the second criteria.

In order to achieve both criteria mentioned previously, the performance metrics considered in this paper will be functions of both the test statistic $\hat{\boldsymbol{t}}$ and the limiting distribution of the test statistic Y under the null. The limiting distributions will allow us to understand how extreme $\hat{\boldsymbol{t}}$ is relative to what we would see under the null which allows us to make comparisons across norms.

As an asside, note that the limiting distribution of $\hat{\boldsymbol{t}}$ under the null will always be a multivariate normal with mean zero with some covariance Σ. Thus potential tests can be functions of the finite dimensional Σ matrix rather than the infinite dimension distribution $Q(P)$. To maintain consistency throughout this paper performance metrics will have the added condition that they become smaller as performance improves. This condition is discussed more in section 4.1. If a candidate performance metric becomes large as $\hat{\boldsymbol{t}}$ grows large, the reciprocal of this performance metric can be used. Performance metrics can be thought of analogously to the p-value of a test in which a non-adaptive version of the test was used.

The first performance metric we consider is the acceptance rate of the test defined in (5) if $\psi = \hat{\boldsymbol{t}}$:

$$\Gamma(x, \Sigma) = Pr_Q(\|Y + x\| \leq c_{0.8}) \text{ where } c_{0.8} \equiv \min_c\{c : Pr(\|Y\| < c) \geq 0.8\} \text{ and } Y \sim N(0, \Sigma). \tag{6}$$

We can define a test using Γ:

$$\text{reject } H_0 \text{ if } \Gamma(\hat{\boldsymbol{t}}, \Sigma) \leq c_{1-\alpha} \text{ where } c_{1-\alpha} = \min_c\{c : \Pr(\Gamma(Y, \Sigma) \geq c) < \alpha\} \text{ where } Y \sim N(\mathbf{0}, \Sigma)$$

Considering the Γ for which it is possible to compare across norms, we can now define a new, adaptive Γ^* which is the pointwise maximum of all of the considered Γs. First define $\Gamma_1(x, \Sigma), \ldots, \Gamma_p(x, \Sigma)$ as collection of performance metrics which only differ by the norm used in their definition. Next, define

$$\Gamma^*(x, \Sigma) = \min\left\{\Gamma_1(x, \Sigma), \ldots, \Gamma_p(x, \Sigma)\right\}.$$

While this function is more complicated than before, distribution of $\Gamma^*(Y, \Sigma)$ can still be compared to $\Gamma^*(\hat{\boldsymbol{t}}, \Sigma)$ to obtain a p-value. Also, while it may be difficult to obtain the exact distribution of $\Gamma^*(Y, \Sigma)$, the distribution is a function of Σ, so obtaining good approximations of $\Gamma^*(Y, \Sigma)$ is possible by taking many draws from Y.

3.2 Obtaining the null distribution

The described procedure requires knowledge of the limiting distribution of $\sqrt{n}\hat{\boldsymbol{\psi}}$ when $P \in \mathcal{M}_0$. To obtain an estimate of this limiting distribution, we require that each of the estimators $\sqrt{n}\hat{\boldsymbol{\psi}}_1, \ldots, \sqrt{n}\hat{\boldsymbol{\psi}}_d$ of $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_d$ is asymptotically linear. That is for each $j \in \{1, \ldots, d\}$:

$$\hat{\psi}_j = \psi_j + \frac{1}{n}\sum_{i=1}^{n} D_j(\boldsymbol{x}_i) + o_p(1/\sqrt{n}) \text{ for some function } D_j$$

Where the D_j function will be referred to as an influence function. Most standard estimators of association are asymptotically linear and have known influence functions.There also exists a growing body of literature describing asymptotically linear estimators and their corresponding functions Need citations here.

When there is a fixed number of covariates, the Cramer-Wold device can be used to show that $\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})$ estimates is asymptotically normal with mean zero, and variance covariance matrix given by $\Sigma =$

$E_{P_0}\left[D(X)D(X)^\top\right]$:

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \xrightarrow{d} Y \sim N\left(0, \Sigma\right)$$

Under H_0, $\sqrt{n}\hat{\boldsymbol{\psi}}$ converges to Y, and Σ can be consistently estimated using $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} D(\boldsymbol{x}_i)D(\boldsymbol{x}_i)^\top$. In practice, a consistent estimator of Σ, $\hat{\Sigma}$ will be used in place of Σ for the calculation of $\hat{\boldsymbol{t}}$. Thus, the test statistic will be $\Gamma(\hat{\boldsymbol{t}}, \hat{\Sigma})$ will be compared to $\Gamma(\hat{Y}, \hat{\Sigma})$ where $\hat{Y} \sim N\left(0, \hat{\Sigma}\right)$.

3.3 Using a permutation test for the test statistic

While the above approach works asymptotically, for small sample sizes there can be greater than α type one error. To avoid this, a permutation based test can be used in certain settings.

Letting Π be the set of all $n!$ one-to-one functions mapping from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$, let $\pi_1, \ldots, \pi_B$ be drawn uniformly from Π with replacement. To carry out the permutation test test, B draws are taken from the null distribution where each the j^{th} draw is the test statistic $\sqrt{n}\hat{\boldsymbol{\psi}}_j^{\#}$ for a permuted data set $(y_{\pi_j(1)}, x_{1,1}, \ldots, x_{1,d}), \ldots, (y_{\pi_j(n)}, x_{n,1}, \ldots, x_{n,d})$. After taking these draws, a p-value can be calculated $B^{-1}\sum_{i=1}^{B} I\{\sqrt{n}\hat{\boldsymbol{\psi}} \geq \sqrt{n}\hat{\boldsymbol{\psi}}_i^{\#}\}$ and the corresponding test rejects the null if this $p-$value is less than α.

This approach can only be applied in settings in which the null hypothesis corresponds to all used covariates being jointly independent of the outcome. For example if the parameter of interest is a conditional association measure as is the case in the second and third examples, this procedure will not work.

4 Theoretical Results

While we try to keep theoretical results as general as possible, in order to establish consistency and non-trivial local ubiasedness of the testing procedure we place restrictions on the norms and performance metrics.

Performance metric and norms considered

In the examples considered in this article two performance metrics are considered. The first is the estimated acceptance rate performance metric as described in equation (6). The other performance metric considered is the multiplicative distance $\hat{\boldsymbol{t}}$ performance metric defined by:

$$\Gamma(t, \Sigma) = \min\left\{s : \Pr(\|Z + st\|_p \geq c_\alpha) \geq 0.8\right\} \tag{7}$$

Both performance metrics are considered to provide multiple examples of viable performance metrics and because each has its own advantages and disadvantages. The estimated acceptance rate performance metric has the advantage of being intuitively simple and relatively simple to calculate. However this metric also is bounded between zero and one which has the potential to cause issues when calculating the limiting distribution if most of the mass of $\Gamma(Y, \Sigma)$ is concentrated near 1. The multiplicative distance performance metric does not face this constraint, but is more difficult to understand intuitively. Additionally this metric is more computationally costly, and the 0.8 used is somewhat arbitrary and it is not clear what effect changing this value would have on the power of tests based on this metric (we hope it has little to no effect).

For both metrics, two sets of norms are considered for the adaptive test. Letting $\boldsymbol{x} = (x_1, \ldots, x_d)$ the ℓ_p norm is defined by

$$\ell_p(\boldsymbol{x}) = \sqrt[p]{\sum_{i=1}^{d} |x_i|^p}$$

The second set of norms referred to as the sum of squares norm is a function mapping from $\mathbb{R}^d$ to $\mathbb{R}$ defined by

$$\|\boldsymbol{x}\|_k = \sum_{i=1}^{k} x_{(d-i+1)}$$

where $x_{(1)}, \ldots, x_{(d)}$ are the order statistics of $x_1, \ldots, x_d$. When adaptive versions of either performance metric are used in practice one must specify which norms to select over. In this work the possible indices (p's) for the ℓ_p norm are $1, 2, 4, 6$, and ∞ and the possible indices for the sum of squares norm are the six evenly spaced values between 1 and d (the dimension of x). If any of these values are not whole numbers they are rounded to the nearest whole number.

Lemmas 8 and 9 show that the estimated acceptance rate and multiplicative distance performance metrics satisfy these restrictions respectively.

4.1 Consistency under fixed Alternatives

Consider the following conditions on the performance metric $\Gamma : \mathbb{R}^d \otimes \mathbb{R}^{d \times d} \mapsto \mathbb{R}$:

1. The performance metric Γ is continuous and is non-negative

2. $\Pr\left(\Gamma(Z, \Sigma) = 0\right) = 0$ where $Z \sim N(0, \Sigma)$

3. $E\left[\left|\Gamma^*_{\hat{\Sigma}}(\sqrt{n}\hat{\boldsymbol{\psi}})\right|\right] \to 0$ for all fixed $P \notin \mathcal{M}_0$, all consistent estimators $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ and all consistent estimators $\hat{\Sigma}$ of Σ.

Theorem 1. *If $\hat{\boldsymbol{\psi}}$ is a consistent estimator of ψ, $\hat{\Sigma}$ is a consistent estimator of Σ, and Γ satisfies conditions 1 - 3 then the test defined by*

$$\textit{reject } H_0 \textit{ if } \Gamma^*(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma}) \leq F^{-1}_{\Gamma^*(\hat{Y},\hat{\Sigma})}(\alpha)$$

is consistent.

Local Unbiasedness

Consider a sequence of local alternatives P_n in which the value of $\Psi(P_n) = \psi_n = \underline{h}/\sqrt{n}$ shrinks towards zero as n grows. We wish to show that our test is non-trivially locally unbiased, that is for large n:

$$Pr_P\left(\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma}) \leq F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}(\alpha)\right) > \alpha$$

To achieve this, we fist assume our estimator $\sqrt{n}\hat{\boldsymbol{\psi}}$ is consistent and regular at $\boldsymbol{\psi}$. The estimator $\hat{\boldsymbol{\psi}}$ is regular if and only if for every sequence ψ_n with $\sqrt{n}(\psi_n - \psi) \to \underline{h}$, the sequence $\sqrt{n}(\hat{\psi} - \psi_n) \xrightarrow{P_{\psi_n}} Z$, where Z can be different for different ψ, but not for different t. Two additional conditions on Γ are sufficient to have tests based on Γ be locally unbiasedness.

4. Γ is unimodal for every Σ (for every k, the set $\{x : \Gamma(x, \Sigma) \geq k\}$ is convex).

5. Γ is centrally symmetric for every Σ, $(\Gamma(-x, \Sigma) = \Gamma(x, \Sigma))$.

Theorem 2. *Let P_n be a sequence of local alternatives under which the value of $\Psi(P_n)$ is $\underline{h}/\sqrt{n}$, and let $Y_\Sigma \sim N(\mathbf{0}, \Sigma)$ and $\hat{Y}_\Sigma \sim N\left(\mathbf{0}, \hat{\Sigma}\right)$. If Γ satisfies conditions 1, 2, 4, and 5, the estimator $\hat{\Sigma}$ of Σ is consistent, and the estimator $\hat{\psi}$ of ψ is regular at $\boldsymbol{\psi}$. Then the test defined by*

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

has power greater than α under P_n.

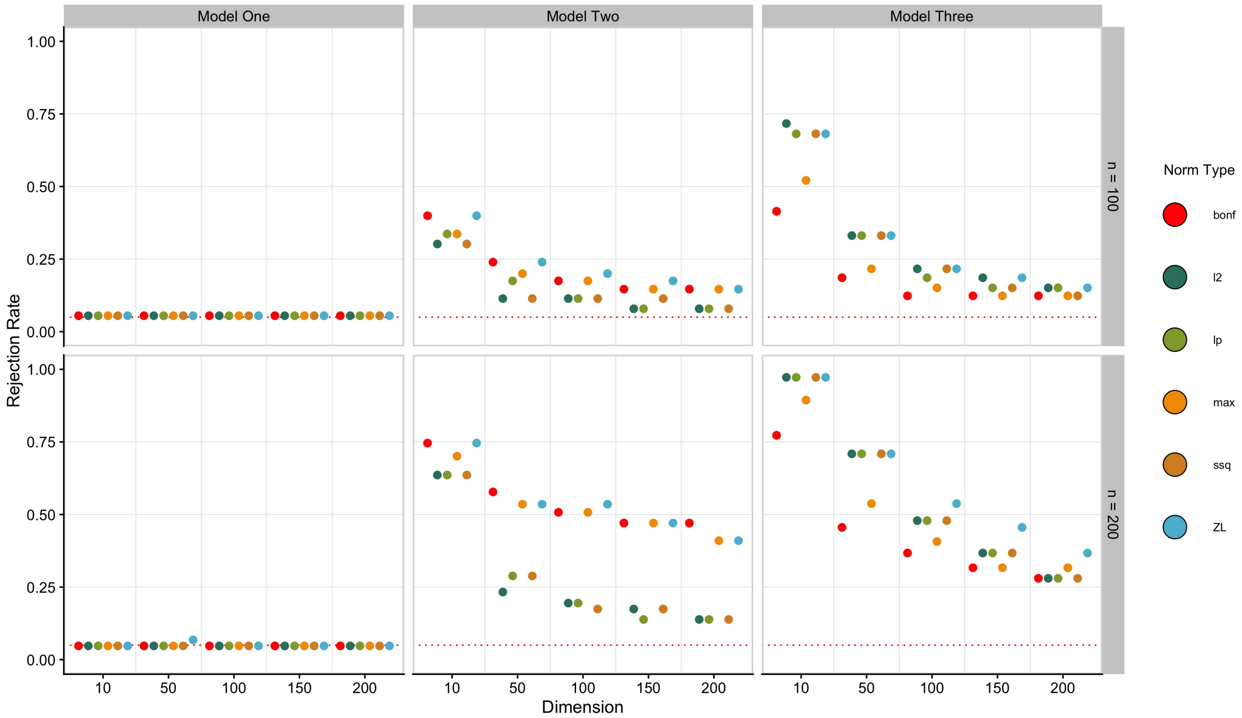

Figure 2: Display of simulations for vector of covariates in this setting will be generated from a normal distribution with mean zero and a variance covariance of Σ with Σ_{ij} equal to 0 when $i \neq j$ and equal to 1 when $i = j$. Three different models for the outcome of interest (Y) will are considered. Letting $\varepsilon \sim N(0,1)$ and be independent of W, in the first model $Y = \varepsilon$, in the second $Y = W_1/4$, and in the third $Y = \sum_{k=1}^{10} \beta_k W_k + \varepsilon$ where $\beta_k = 0.15$ for $k = \{1, \ldots, 5\}$, and $\beta_k = -0.1$ for $k = \{6 \ldots 10\}$. Sample sizes of 100 and 200, dimensions of 10, 50, 100, 150, and 200 are considered.

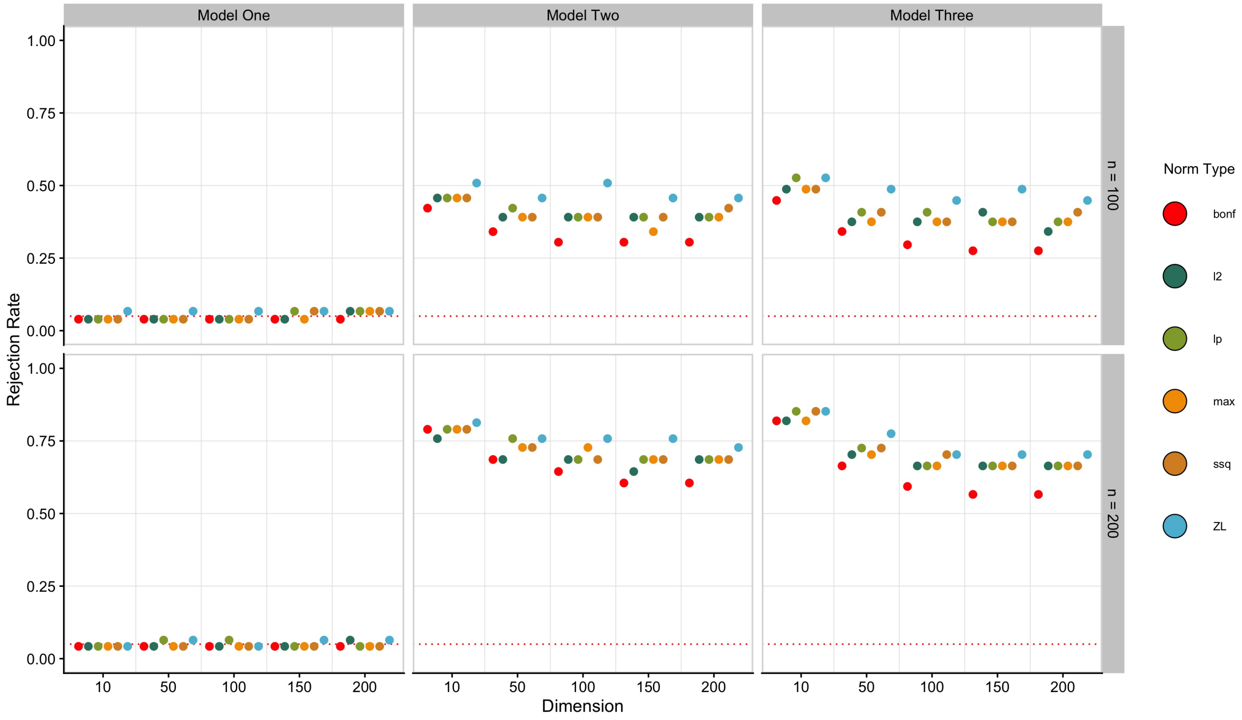

Figure 3: The same simulation settings as those used in Figure 2, but $\Sigma_{ij} = 0.5$ for $i \neq j$.

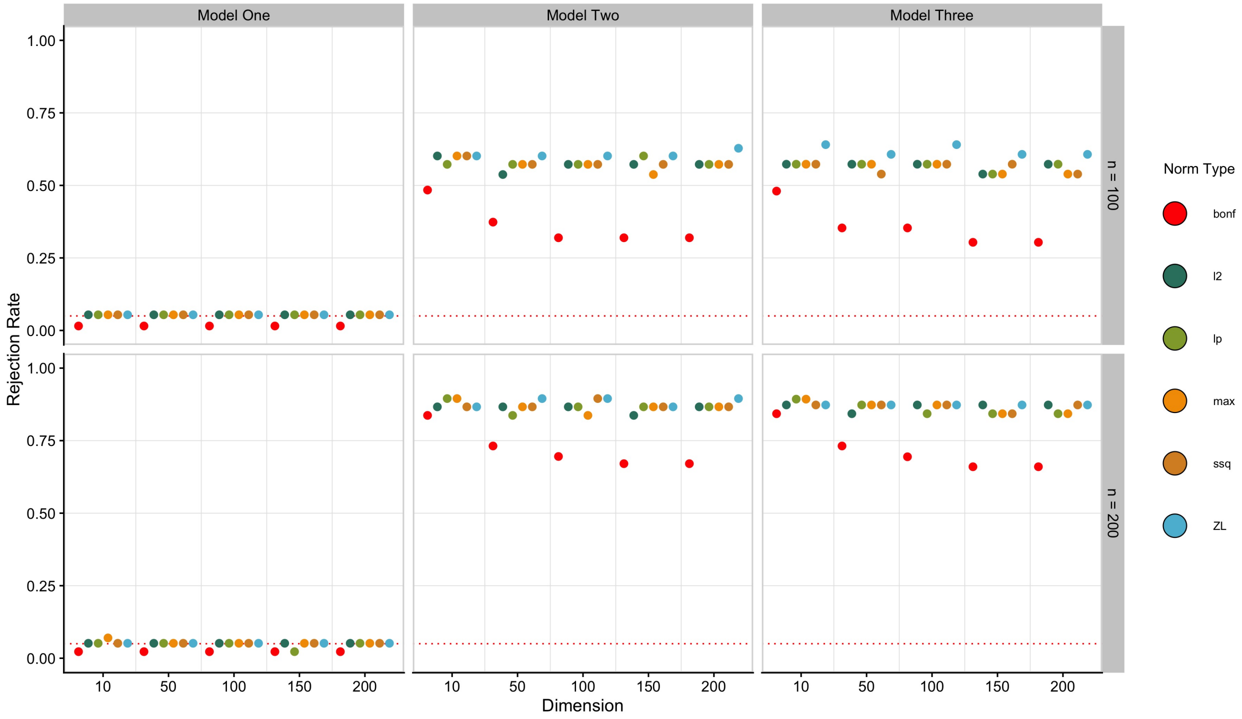

Figure 4: The same simulation settings as those used in Figure 2, but $\Sigma_{ij} = 0.8$ for $i \neq j$.

5 Simulation Study

5.1 Correlation

In figures (2, 3, 4), the rejection rates of six different tests are shown for a wide variety of settings. Each test has the null hypothesis that $\psi_1, \ldots, \psi_d = 0$ where $\psi_i = \Psi_i(P)$ is the correlation between the outcome Y, and the j^{th} covariate W_j. A description of the data generating mechanism can be found in section 2. Red dots indicate a bonferroni adjusted marginal test for the estimated correlation between each covariate and the outcome. Light blue dots indicate the performance of the test proposed by [Zhang and Laber]. The other colors correspond to different variants of our test. Dark green and yellow dots indicate the performance of our test using only the ℓ_2 or maximum absolute value norm respectively. The light green dots indicate the performance of our test when it adaptively select one of the ℓ_p norms. Brown dots indicate our tests performance when the test adaptively selects over various sum of squares norms. The dotted red line in each plot indicates the 0.05 rejection rate that should be observed when H_0 holds.

Because H_0 is true for model one, we expect the rejection rates for this model to be 0.05. Figures 2, 3, and 4 show these rates are achieved by every testing procedure except the bonferroni based test which is somewhat conservative. For the other two models, H_0 does not hold so these plots compare the powers between the different testing procedures. The bonferroni based test has the lowest power in all of the considered settings and for larger numbers of covariates this differences is larger. All other tests have similar power in most settings, with the test proposed by [Zhang and Laber] performing slightly better in most settings where a difference exists. All tests perform better for larger sample sizes, but tend to perform similarly for varying dimension and generating model.

One setting in which the adaptive test performs poorly is in settings in which a single covariate is correlated with the outcome of interest and no covariates are correlated as seen for model 2 in figure 2. This behavior could be due to the parameter estimates inability to be sparse even in settings when sparsity occurs. The many small errors in the parameter estimate across many covariates leads to a preference for the ℓ_2 norm that obtains an overly optimistic estimate of power due to the accumulation of many small effects across all the covariates.

Two Phase Sampling Risk Ratio

The for the second example was carried out after deriving the influence of the parameter given in equation 1. Marginal expectations were estimated using either an elastic net [Simon et al., Tibshirani et al., Friedman et al., Tibshirani et al.] or a superlearner [Polley and Laan].

Figure 5.1 shows the rejection rates of four different tests across the settings described in section 6. Each test has the null hypothesis that $\psi_1, \ldots, \psi_d = 0$ where $\psi_i = \Psi_i(P)$ is the risk ratio that $Y = 1$ comparing two observations which differ by a unit in the covariate W_j. Data are generated using three different settings. In the first setting no covariates are directly associated with the outcome (H_0 holds), in the second setting a single covariate has a strong, direct association with the outcome (H_0 does not hold), and in the third setting ten covariates are directly associated with the outcome (H_0 does not hold). Each color of dots represents a different version of our test. Red and light green dots indicate the performance of our test using only the ℓ_2 or maximum absolute value norm respectively. Dark green dots indicate the performance of our test when it adaptively select one of the ℓ_p norms. Yellow dots indicate our tests performance when the test adaptively selects over various sum of squares norms. The dotted red line in each plot indicates the 0.05 rejection rate that should be observed when H_0 holds.

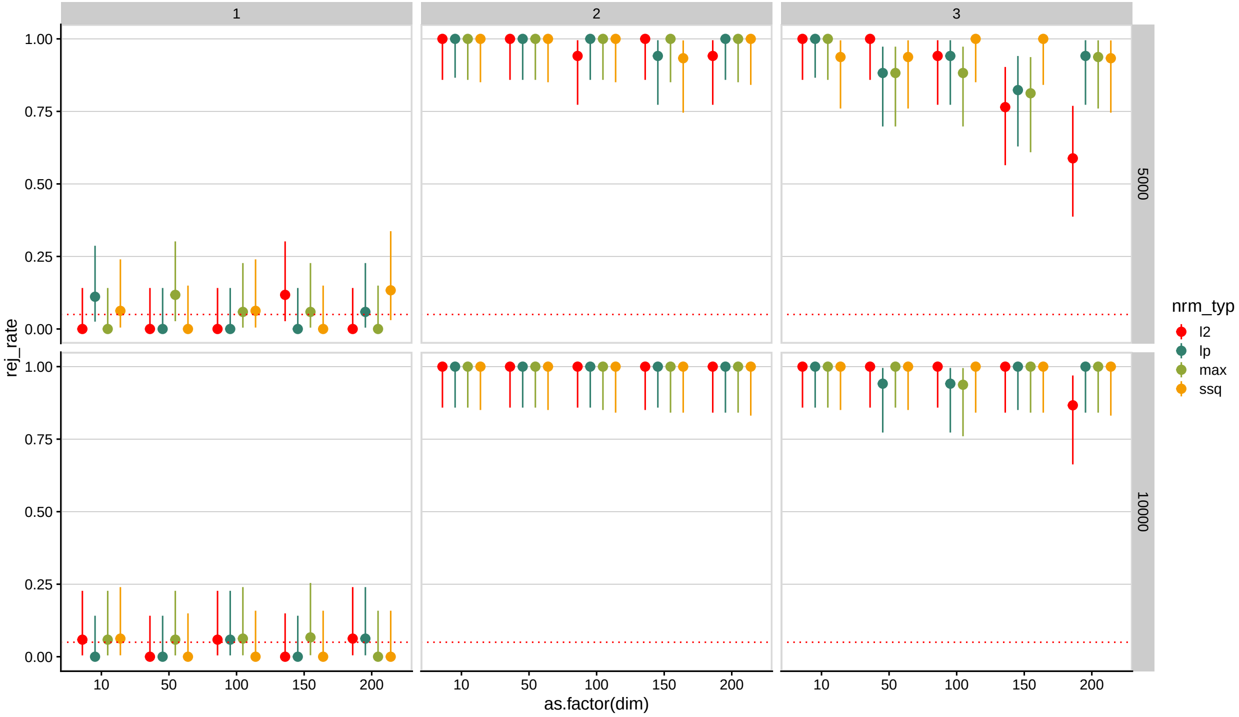

Figure 5: Rejection rates across three different data generating mechanisms. Here we are using the multiplicative distance our test statistic is from obtaining 80% power as our measure of performance.

Marginal Structural Model

Testing for this example was carried out using code from the ltmle package [Lendle et al.] to estimate the influence function of our parameter. Using this estimate of the influence function, a parametric bootstrap was used to estimate the limiting distribution of our test statistic. The multiplicative distance performance metric was used for these simulations. Data are generated using three different settings. In the first setting no covariates modifies the treatment effect (H_0 holds), in the second setting a single covariate modifies the treatment effect (H_0 does not hold), and in the third setting ten covariates modify the treatment(H_0 does not hold). Each color of dots represents a different version of our test. Red and light green dots indicate the performance of our test using only the ℓ_2 or maximum absolute value norm respectively. Dark green dots indicate the performance of our test when it adaptively select one of the ℓ_p norms. Yellow dots indicate our tests performance when the test adaptively selects over various sum of squares norms. The dotted red line in each plot indicates the 0.05 rejection rate that should be observed when H_0 holds.

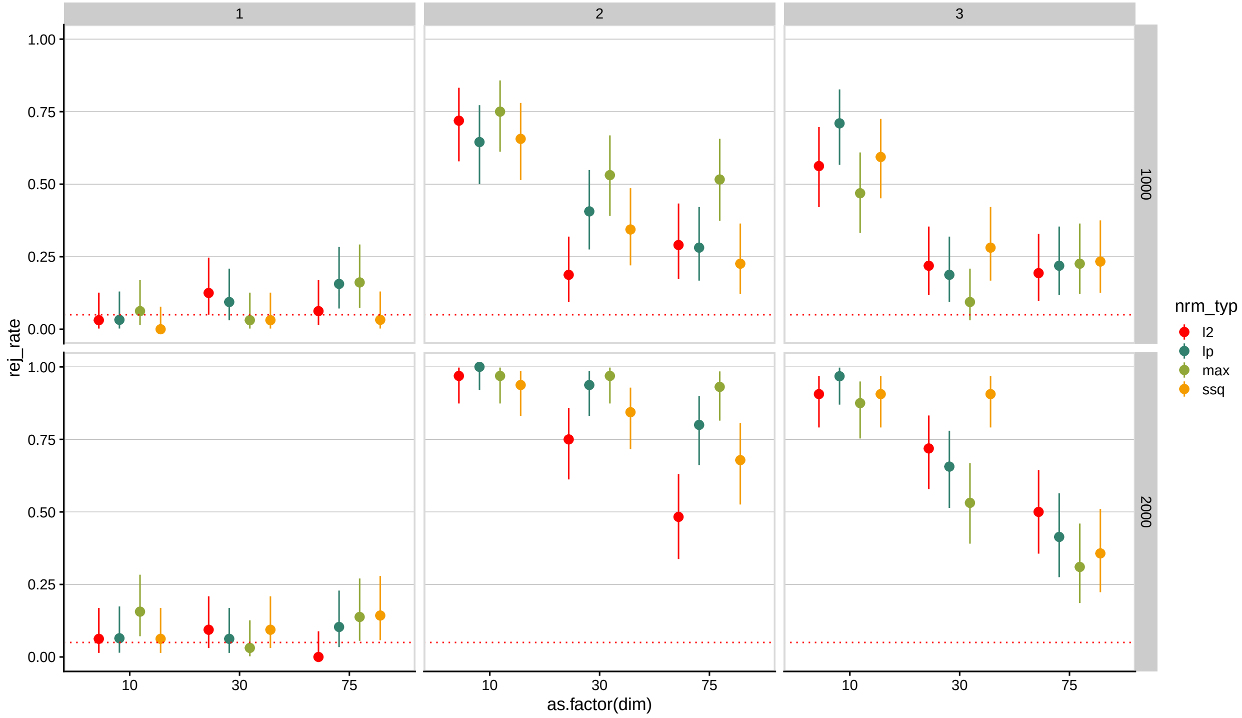

Figure 6: Rejection rates across a range of settings when a marginal structural model was used to define the parameter of interest.

Because H_0 is true for the first setting, we expect the rejection rates in this setting to be 0.05. Figure 6 show results consistent with this being the case. For the other two settings H_0 does not hold so these plots compare the powers between the different testing procedures. While results are not definitive it appears that for the second setting in which a single covariate modifies the treatment effect it appears that the max norm has the best performance followed by the adaptive procedures and then the ℓ_2 norm. In the third setting in which many covariates are associated with the outcome the adaptive norms appear to once again have middling performance now with the ℓ_2 norm performing the best and the max norm performing the worst. It was expected that the non-adaptive tests would excel where they did and that the adaptive versions tend to perform well across multiple settings. It is unclear from these results if, for a given setting, at least one non-adaptive version of the test will outperform the adaptive version of the test.

One setting in which the adaptive test performs poorly is in settings in which a single covariate is correlated with the outcome of interest and no covariates are correlated as seen for model 2 in figure 2. This behavior could be due to the parameter estimates inability to be sparse even in settings when sparsity occurs. The many small errors in the parameter estimate across many covariates leads to a preference for the ℓ_2 norm that obtains an overly optimistic estimate of power due to the accumulation of many small effects across all the covariates.

6 Data Application

Data from Peter Gilbert

7 Discussion

When developing anything there is frequently a trade off to between generalizability and performance. However the reduction in power of bonferroni-based methods compared to taylor made procedures is often larger than is needed. We have developed a generalizable testing procedure that still can outperform bonferroni based methods. We have demonstrated this in section 5 and shown novel settings in which the method can also be performed. These settings show that our test can be applied to parameters in non-standard sampling schemes (see section 2.2) and complex parameters (see section 2.3).

Our procedure does rely on asymptotic results to achieve correct type one error rates which can cause issues when the number of covariates of interest is larger than sample size. To address this issue, in certain cases a our test can be carried out using permutations of the data to obtain better finite sample properties.

While we have outlined some of the ways in which our method can be used, there are other possible applications and improvements.

While we focused on the studying the limiting distribution of $\sqrt{n}\hat{\psi}$ once can also consider the limiting distribution of $\sqrt{n}\hat{\Sigma}^{-1/2}\hat{\psi}$. This estimator will always have a multivariate normal limiting distribution with an identity covariance matrix, which simplifies notation by always having a single limiting distribution. It also simplifies proofs, and potentially could allow for analytical solutions for which norm (or weighted average of norms) would provide the best power. However, this estimator runs into issues when $p > n$ and $\hat{\Sigma}^{-1/2}$ becomes impossible to estimate because of the rank deficiency of the variance estimator. While this problem still technically exist when indexing our Γs by $\hat{\Sigma}$, using a multiplier bootstrap to take draws from $N(0, \hat{\Sigma})$

avoids the computation issues of having a degenerate estimate of the limiting distributions variance matrix.

Mention choice of norms and choice of performance metric

While changing the test statistic as described above has major hurdles to overcome to be implemented in $p > n$ settings, there are other extensions or generalizations of our procedure that are possible without much change. While we considered one set of Γ funcitons in this article, any set of reasonable Γ functions could be used. These Γ functions could even be picked by machine learning methods used for classification. The classifier would provide a probability that any observation was generated from the null distribution. Doing this provides a more rich set of Γ over which to select, and has the potential to provide large increases in test performance. One would expect that a classifier would perform better better when data are generated at an alternative versus at the null. Thus one could reject when the classification performance is especially large.

While our procedure currently only selects a single norm to calculate the test statistic, it is possible to also consider an estimator that is a weighted average of the many different norms. The weights would be estimates of the probability that each norm was optimal. These probabilities could be estimated by taking many draws from a normal with mean $\sqrt{n}\hat{\psi}$ and variance $\hat{\Sigma}$ and finding the optimal norm for each of these draws. This method while being more computationally costly, could potentially be shown theoretically to be optimal in the sense that it would maximize the average power of all tests based on the specified family of norms for each local alternative.

While ℓ_p norms were used throughout this paper, one issue observed with our test is that it selects the ℓ_2 norm more frequently than it should, likely because of the many small estimates for parameters that are not associated with the outcome. One could consider a slightly modified norm that sets to zero all component values of the vector that are less than some value. This would hopefully solve issues of small values close to zero causing incorrect selection of norms that perform well when there are many small effects.

add notes about limitation of our method (need to find influence function). Also discuss difficulty with convergence when using the parametric bootstrap.

In this article chose tests in which each parameter corresponded to a single covariate, and each covariate only had a single corresponding parameter. However it would also be possible to have tests where each parameter corresponded to a set of many covariates or each covariate or group of covariates had multiple measures of association.

8 Conclusion

9 Appendix

This appendix is primarily focused on proving the consistency and local power of our testing procedure.

For multiple proofs it will be necessary to use the dominated convergence theorem, and find a dominating measure for $\phi(x, \mu, \Sigma)$ and $\phi(x, \mu_n, \Sigma_n)$ where ϕ is the pdf for a multivariate normal distribution.

Lemma 3. *The function* $\phi : \mathbb{R}^d \otimes \mathbb{R}^d \otimes \mathbb{R}^{d \times d} \to \mathbb{R}$ *defined by*

$$\phi : (x, \mu, \Sigma) \mapsto det(\Sigma)^{-1/2}(2\pi)^{-k/2} exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)$$

is continuous and $\phi_{\Sigma,\mu} : x \mapsto \phi(x, \Sigma, \mu)$ *and* $\phi_{\mu_n,\Sigma_n} : x \mapsto \phi(x, \Sigma_n, \mu_n)$ *are dominated by an integrable function.*

Proof. Since matrix inverses, matrix determinants, matrix multiplication, and the exponential function are all continuous, it follows that the multivariate normal distribution is also continuous with respect to Σ, μ and x. Because of this, we also know that for large enough n that

$$\left|\frac{1}{2}(x_n-\mu_n)\Sigma_n^{-1}(x_n-\mu_n) - \frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right| < \varepsilon$$

for any ε we choose. When $\varepsilon = \log(2)$ it is the case that

$$\begin{aligned}
\frac{\psi(x, \mu_n, \Sigma_n)}{\psi(x, \mu, \Sigma)} &= \frac{\frac{\det(\Sigma_n)^{-1/2}}{(2\pi)^{k/2}}\exp\left(-\frac{1}{2}(x_n-\mu_n)\Sigma_n^{-1}(x_n-\mu_n)\right)}{\frac{\det(\Sigma)^{-1/2}}{(2\pi)^{k/2}}\exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)} \\
&= \frac{\det(\Sigma_n)^{-1/2}\exp\left(-\frac{1}{2}\left((x_n-\mu_n)\Sigma_n^{-1}(x_n-\mu_n) - (x-\mu)\Sigma^{-1}(x-\mu)\right)\right)}{\det(\Sigma)^{-1/2}} \\
&\leq \frac{\det(\Sigma_n)^{-1/2}\exp\left(\left|-\frac{1}{2}\left((x_n-\mu_n)\Sigma_n^{-1}(x_n-\mu_n) - (x-\mu)\Sigma^{-1}(x-\mu)\right|\right)\right)}{\det(\Sigma)^{-1/2}} \\
&\leq \frac{\det(\Sigma_n)^{-1/2}\exp\left(\log(2)\right)}{\det(\Sigma)^{-1/2}} \\
&\leq \frac{\det(\Sigma_n)^{-1/2}2}{\det(\Sigma)^{-1/2}}
\end{aligned}$$

Next, note that

$$\det(\Sigma_n)^{-1/2} - \det(\Sigma)^{-1/2} \leq \left|\det(\Sigma_n)^{-1/2} - \det(\Sigma)^{-1/2}\right| \leq \det(\Sigma)^{-1/2} \Rightarrow \det(\Sigma_n)^{-1/2} \leq 2\det(\Sigma)^{-1/2}$$

Therefore, we find that

$$\frac{\det(\Sigma_n)^{-1/2}2}{\det(\Sigma)^{-1/2}} \leq \frac{\left[2\det(\Sigma_n)\right]^{-1/2}2}{\det(\Sigma)^{-1/2}} = 2\sqrt{2}$$

Thus, for large enough n, we can bound $\phi(x, \mu_n, \Sigma_n)$ by $2\sqrt{2}\phi(x, \mu, \Sigma)$ which is integrable.

□

Lemma 4. *Let $\hat{Z} \sim N(0, \hat{\Sigma})$ and $Z \sim N(0, \Sigma)$. If $\hat{\Sigma}$ is a consistent estimator of Σ, $\Gamma : \mathbb{R}^d \otimes \mathbb{R}^{d^2}$ is continuous, and $Pr(\Gamma(Z, \Sigma) = t) = 0$ for each t. It follows that the distribution function of $\Gamma(\hat{Z}, \hat{\Sigma})$ denoted by $F_{\Gamma(\hat{Z},\hat{\Sigma})}$ converges pointwise to the distribution function of $\Gamma(Z, \Sigma)$ denoted by $F_{\Gamma(Z,\Sigma)}$ and the quantile function of $\Gamma(\hat{Z}, \hat{\Sigma})$ denoted by $F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}$ converges pointwise to the quantile function of $\Gamma(Z, \Sigma)$ denoted $F^{-1}_{\Gamma(Z,\Sigma)}$ where $F^{-1} : [0, 1] \to \mathbb{R}$ is defined By $F^{-1}(p) = \inf\{x : F(x) \geq p\}$*

Proof. From Lemma 3 we know that we there exists a dominating function for the probability density functions of $\hat{Z}$ and Z. Applying dominated convergence theorem shows that

$$\Pr\left(\Gamma(\hat{Z}, \hat{\Sigma}) \leq t\right) = \int I\{\Gamma(x, \hat{\Sigma}) \leq t\}\phi(x, \hat{\mu}, \hat{\Sigma})dx \to \int I\{\Gamma(x, \Sigma) \leq t\}\phi(x, \mu, \Sigma)dx = \Pr\left(\Gamma(Z, \Sigma)) \leq t\right).$$

The convergence of the distribution functions shown above and lemma 21.1 of [Van der Vaart] imply the pointwise convergence of the corresponding quantile functions. □

9.1 Proof of Theorem 1 (Test Consistency)

Proof. Let $Z \sim N(\mathbf{0}, \Sigma)$ and $\hat{Z} \sim N\left(\mathbf{0}, \hat{\Sigma}\right)$. Lemma 4 and the assumptions of the theorem imply the pointwise convergence of the distribution function of $\Gamma(\hat{Z}, \hat{\Sigma})$ denoted by $F_{\Gamma(\hat{Z},\hat{\Sigma})}$ to the distribution function of $\Gamma(Z, \Sigma)$ denoted by $F_{\Gamma(Z,\Sigma)}$ and the pointwise convergence of the quantile function of $F_{\Gamma(\hat{Z},\hat{\Sigma})}$ to the quantile function of $F_{\Gamma(Z,\Sigma)}$. By the assumption of the theorem $F^{-1}_{\Gamma(Z,\Sigma)}(\alpha) > 0$, so

$$\begin{aligned}\Pr_P\left(\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma}) \leq F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}(\alpha)\right) &= 1 - \Pr_P\left(\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma}) \geq F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}(\alpha)\right) \\ &\geq 1 - \frac{E\left[|\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma})|\right]}{F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}(\alpha)} \xrightarrow{p} 1.\end{aligned}$$

□

While $\Gamma(x, \Sigma)$ may not grow close to zero as x becomes large, one can use transformations of the original Γ to achieve this property. We show that the performance metrics used in this paper satisfy the satisfy conditions 1-3 in lemmas 8 and 9.

9.2 Unbiasedness at local alternatives

To prove non-trivial local unbiasedness, we rely heavily on a result of [Eaton and Perlman]. It states that for two centrally symmetric, unimodal functions, f_1 and f_2, the function $g : \mathbb{R}^d \mapsto \mathbb{R}$ defined by the convolution:

$$g(\underline{h}) = \int f_1(x) f_2(x - \underline{h})$$

is centrally symmetric and ray decreasing. A function $f : \mathbb{R}^p \to \mathbb{R}$ is centrally symmetric if $f(-x) = f(x)$ for every x. A function is unimodal if for every k, the set $\{x : f(x) \geq k\}$ is convex. A function f on $\mathbb{R}^p$ is ray decreasing if for every x in $\mathbb{R}^p$, the function $g(\beta) = f(\beta x), \beta \in \mathbb{R}^+$ is a non-increasing function of β. Additional conditions are required to hold for $g(\beta)$ to be a strictly decreasing function of β. These conditions are shown to hold in lemma 13 for Γ that satisfy conditions 1, 2, 4, and 5.

Our testing procedure can be rewritten in using the above form, where $f_1(x)$ is $I\{\Gamma^*_\Sigma(x) > F^{-1}_{\Gamma^*_\Sigma(Z)}(\alpha)\}$, and f_2 is a multivariate pdf with mean 0. The ray decreasing quality of g can then be used to show that

$$Pr_P\left(\Gamma(Z + \underline{h}, \Sigma) \leq F^{-1}_{\Gamma(Z,\Sigma)}(\alpha)\right) > Pr_P\left(\Gamma(Z, \Sigma) \leq F^{-1}_{\Gamma(Z,\Sigma)}(\alpha)\right) = \alpha$$

Because the multivariate normal pdf is centrally symmetric and unimodal (see lemma 10), all that remains to be shown is that the function $(x, \Sigma) \mapsto I\{\Gamma(x, \Sigma) > F^{-1}_{\Gamma(Z,\Sigma)}(\alpha)\}$ is unimodal and centrally symmetric. Lemma 11 tells us that this indicator function is unimodal if Γ is unimodal and centrally symmetric.

In lemmas 8 and 9, we show that the Γ used in this article are unimodal and centrally symmetric.

9.3 Proof of Theorem 2 (Non-Trivial Local Unbiasedness)

Proof. Lemma 4 and the assumptions of the theorem imply the pointwise convergence of the distribution function of $\Gamma(\hat{Z}, \hat{\Sigma})$ denoted by $F_{\Gamma(\hat{Z},\hat{\Sigma})}$ to the distribution function of $\Gamma(Z, \Sigma)$ denoted by $F_{\Gamma(Z,\Sigma)}$ and the pointwise convergence of the quantile function of $F_{\Gamma(\hat{Z},\hat{\Sigma})}$ to the quantile function of $F_{\Gamma(Z,\Sigma)}$. Thus

$$Pr_P\left(\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}},\hat{\Sigma})\leq F^{-1}_{\Gamma(\hat{Z},\hat{\Sigma})}(\alpha)\right)\xrightarrow{p}Pr_P\left(\Gamma(Z+\underline{h},\Sigma)\leq F^{-1}_{\Gamma(Z,\Sigma)}(\alpha)\right).$$

Breaking down the quantity on the right, the only random part inside of the probability statement is Z. Define the function $g^*:\mathbb{R}^d\to\mathbb{R}$ by the quantity inside the probability statement on the right hand side of the the above display:

$$\begin{aligned}g(\underline{h})=Pr_P\left(\Gamma(Z+\underline{h},\Sigma)\leq F^{-1}_{\Gamma(Z,\Sigma)}(\alpha)\right)&=\int I\left\{\Gamma(x,\Sigma)\leq c_\alpha\right\}\phi(x-\underline{h})dx\\&=\int\left(1-I\left\{\Gamma(x,\Sigma)\geq c_\alpha\right\}\right)\phi(x-\underline{h})dx\\&=1-\int I\left\{\Gamma(x,\Sigma)\geq c_\alpha\right\}\phi(x-\underline{h})dx.\end{aligned}\tag{8}$$

Because Γ is unimodal and centrally symmetric it follows from Lemma 11 that $I\left\{\Gamma(x,\Sigma)\geq c_\alpha\right\}$ is also centrally symmetric and unimodal. Thus, the subtracted quantity in (8) is ray decreasing with respect to $\underline{h}$ by [Anderson] and attains it maximum at $\underline{h}=0$. To show that $g(\underline{h}a)$ is increasing for $a\in[0,1]$ and $g(\underline{h})>g(0)$ where

$$g(0)=Pr_P\left(\Gamma(Z,\Sigma)\leq F^{-1}_{\Gamma(Z,\Sigma)}(\alpha)\right),$$

The asymptotic probability of rejecting H_0 when H_0 holds. □

9.4 Extension of results from Γ to Γ^*

While theorems 1 and 2 place conditions on our adaptive estimator Γ^*, it may be difficult to directly prove conditions 2-6 hold for Γ^*. Here we show that if conditions 2-6 hold for each $\Gamma_1,\ldots,\Gamma_k$ and $\Gamma^*(x)=\max\{\Gamma_1(x),\ldots,\Gamma_k(x)\}$, then conditions 2-6 also hold for Γ^*.

Lemma 5. *Let $\Gamma_1,\ldots,\Gamma_{Sigma,k}$ be such that each Γ_i is continuous with respect to x and Σ is non-negative, and at least one Γ_j has the property that $E\left[\Gamma_j(\sqrt{n}\hat{\boldsymbol{\psi}},\hat{\Sigma})\right]\to0$ for all fixed $P\notin\mathcal{M}_0$. Then if $\Gamma(x)=\max\{\Gamma_1(x),\ldots,\Gamma_k(x)\}$ satisfies conditions 2-4.*

Proof. Because the max function is continuous and the composition of continuous functions is also continuous it follows that $\Gamma^*(x)=\max\{\Gamma_1(x),\ldots,\Gamma_k(x)\}$ is also continuous. Because $\Gamma_i(Z)=0$ for every $i\in\{1,\ldots,k\}$, it follows that

$$\Pr(\Gamma^*(Z,\Sigma)=0) = \Pr\left(Z \in \bigcup_{i=1}^{k} \{\omega : \Gamma_i(\omega,\Sigma)=0\}\right) \leq \sum_{i=1}^{k} \Pr(\Gamma_{\Sigma,1}(Z)=0) = 0$$

□

Lemma 6. *Let $\Gamma_1, \ldots, \Gamma_k$ all be centrally symmetric, unimodal functions. Then $\Gamma^* = \min\{\Gamma_1, \ldots, \Gamma_k\}$ is also centrally symmetric and unimodal.*

Proof. Because each Γ_i is a centrally symmetric and Γ^* is a function of x only through Γ_i it follows that Γ^* is also centrally symmetric:

$$\Gamma^*(x) = \min\{\Gamma_1(x), \ldots, \Gamma_k(x)\} = \min\{\Gamma_1(-x), \ldots, \Gamma_k(-x)\} = \Gamma^*(-x)$$

The set $M^* \equiv \{x : \Gamma^*(x) \geq k\}$ contains all of the x for which $\Gamma \geq k$ for all $i \in \{1, \ldots, d\}$. Thus M^* is the intersection of all sets $M_i \equiv \{x : \Gamma_i(x) \geq k\}$. Each M_i is convex because each Γ_i is unimodal, and the intersection of a countable number of convex sets is convex. Thus M^* is convex and Γ^*_Σ is unimodal. □

Performance Metric Specific proofs

In this section, we show that both performance metrics we consider satisfy conditions 2-6. This requires setting conditions on the norms that index the performance metrics. The requirements we place on the norms are that

- each norm obeys the triangle inequality: $\|x\|_p + \|y\|_p \geq \|x+y\|_p$ for every x and y in $\mathbb{R}^d$
- For each $\omega \in \mathbb{R}$, the set $\{x : \|x\| < \omega\}$ is convex.
- each norm has a gradient which is finite at ψ for all possible values of ψ.

The second requirement mentioned above will allow us to bound the variance of the normed vector of parameter estimators as described in the following lemma. This fact will be used to prove both condition 4 ($E\left[\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma})\right] \to 0$ for all fixed $P \notin \mathcal{M}_0$.) for both performance metrics.

Lemma 7. *Let $\hat{Z} \sim N(0,\hat{\Sigma})$ with $\hat{\Sigma} \xrightarrow{p} \Sigma$ be independent of $\hat{\boldsymbol{\psi}}$ for every n, and $\hat{\boldsymbol{t}} \xrightarrow{p} N(0,\Sigma)$. If $||\cdot||_p$ has a finite derivative at $\boldsymbol{\psi}$ then*

$$var\left(\left\|\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\right\|_p\right) \leq M$$

for some M.

Proof. Because of the independence of the two random variables, and the convergence in distribution of $\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)$, we know that

$$\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) = \sqrt{n}\left(\hat{\boldsymbol{\psi}} + \frac{\hat{Z}}{\sqrt{n}} - \boldsymbol{\psi}\right) \xrightarrow{d} Z^* \sim N(0, \sqrt{2}\Sigma)$$

Thus, the delta method tells us that

$$\|\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\|_p \xrightarrow{d} \|Z^*\|_p \sim N\left(0, \nabla\|\boldsymbol{\psi}\|_p^\top \sqrt{2}\Sigma\nabla\|\boldsymbol{\psi}\|_p\right)$$

Because the gradient of $\|\cdot\|_p$ is bounded at $\boldsymbol{\psi}$, the variance of $\|Z^*\|_p$ will be bounded as well. □

9.4.1 Estimated Acceptance Rate

Lemma 8. *The function $\Gamma : \mathbb{R}^d \otimes \mathbb{R}^{d\times d} \to \mathbb{R}$ defined by*

$$\Gamma(t, \Sigma) = Pr_Q(\|\tilde{Z} + t\| < c_{0.8}) \text{ where } c_{\Sigma,0.8} \equiv \min_c\{c : Pr_Q(\|\tilde{Z}\| < c) \geq 0.8\} \text{ and } \tilde{Z} \sim N(0, \Sigma)$$

satisfies conditions 1 - 5.

Proof. **Continuity:** Because Γ is the integral of a function (namely the multivariate normal probability density function) that is continuous with respect to both Σ and x (see Lemma 3), it follows that Γ is also continuous.

Pr $(\Gamma(Z, \Sigma) = t) = 0$ for each t: Because Z and $\tilde{Z}$ are both normally distributed the probability mass functions of both random variables are positive over the entirety of $\mathbb{R}^d$. Considering the set of values $\{\omega : \|\omega\| = t\}$, the probability any normal random variable falls within this region is some positive value no matter what the mean of the random variable is. Therefore, $\Pr(\Gamma(Z, \Sigma) = 0) = 0$.

Convergence of $\Gamma_{\hat{\Sigma}}(\sqrt{n}\hat{\boldsymbol{\psi}})$ to zero under fixed alternatives:

$$\begin{aligned}
E[|\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma})|] &= E\left[|\Pr(\|\hat{Z} + \hat{\boldsymbol{t}}\| \leq c_\alpha)|\right] \\
&= E\left[E\left(I\left\{\|\hat{Z} + \hat{\boldsymbol{t}}\|_p \leq c_\alpha\right\}\right)\right] \\
&= E\left[I\left\{\|\hat{Z} + \sqrt{n}\hat{\boldsymbol{\psi}}\|_p \leq c_\alpha\right\}\right] = \Pr_P\left(\|\hat{Z} + \hat{\boldsymbol{t}}\|_p \leq c_\alpha\right) \\
&= \Pr_P\left(\|\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) + \sqrt{n}\boldsymbol{\psi}\|_p \leq c_\alpha\right) \\
&\leq \Pr_P\left(\|\sqrt{n}\boldsymbol{\psi}\|_p - \|\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\|_p \leq c_\alpha\right) \\
&= \Pr_P\left(\|\sqrt{n}\boldsymbol{\psi}\|_p - c_\alpha \leq \|\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\|_p\right) \\
&\leq \frac{\operatorname{var}\left(\|\hat{Z} + \sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\|_p\right)}{\|\sqrt{n}\boldsymbol{\psi}\|_p - c_\alpha}
\end{aligned}$$

Since the numerator converges to a finite value, and the denominator grows without bound, we know the fraction as a whole will converge to zero.

Central Symmetry: Because the multivariate normal distribution is centrally symmetric, so is $\Gamma(t, \Sigma)$:

$$\begin{aligned}
\Gamma(\mu, \Sigma) = \int I\left\{\|x\|_p < c\right\}\phi_\Sigma(x - \mu)dx &= \int I\left\{\|x + \mu\|_p < c\right\}\phi_\Sigma(x))dx \\
&= \int I\left\{\|x + \mu\|_p < c\right\}\phi_\Sigma(-x))dx \\
&= \int I\left\{\|x\|_p < c\right\}\phi_\Sigma(-x - \mu))dx \\
&= \int I\left\{\|x\|_p < c\right\}\phi_\Sigma(x + \mu))dx \\
&= \Gamma(-\mu, \Sigma)
\end{aligned}$$

Unimodality: First note that

$$\Gamma(\mu, \Sigma) = \int I\left\{\|x\|_p < c\right\}\phi_\Sigma(x - \mu)dx = \int I\left\{\|x + \mu\|_p < c\right\}\phi_\Sigma(x)dx = P(Z \in A_\mu) \equiv P_Z(A_\mu)$$

Where $Z \sim N(0, \Sigma)$ and $A_\mu = \{t : \|t + \mu\| < c\}$

Next consider μ_1, μ_2 such that $\Gamma(\mu_1, \Sigma) \geq k$, and $\Gamma(\mu_2, \Sigma) \geq k$. Because the multivariate normal pdf is log concave (Theorem 4.2.1 of [Tong]), we can apply theorem 1 of [Rinott] to obtain the result that:

$$k = k^t k^{1-t} \leq \Gamma(\mu_1, \Sigma)^t \Gamma(\mu_2, \Sigma)^{1-t} = P_Z(A_{\mu_1})^t P_Z(A_{\mu_2})^{1-t} \leq P_Z(tA_{\mu_1} + (1-t)A_{\mu_2}).$$

Lemma 12 shows that $tA_{\mu_1} + (1-t)A_{\mu_2} = A_{t\mu_1+(1-t)\mu_2}$ so it follows that

$$P_Z(tA_{\mu_1} + (1-t)A_{\mu_2}) = P_Z(A_{t\mu_1+(1-t)\mu_2}) = \Gamma_\Sigma(t\mu_1 + (1-t)\mu_2)$$

□

Multiplicative distance performance metric Still need to show continuity and condition 3 hold for this performance metric

Consider the other measure of performance we have used:

Lemma 9. *The function* $\Gamma : \mathbb{R}^d \otimes \mathbb{R}^{d^2} \mapsto (0, \infty)$ *defined by*

$$\Gamma(t, \Sigma) = min\{s : Pr(\|Z + st\|_p \geq c_\alpha) \geq 0.8\}$$

satisfies conditions 1-5.

Proof. **Continuity**

$\Gamma(Z, \Sigma)$ **has measure zero at zero.** Because c_α is defined in such a way that $\Pr(\|Z\|_p \geq c_\alpha) \leq 0.2 < 0.8$, and because $0 \times t = 0$ for any real-valued t, the probability that $\Gamma^*(Z, \Sigma) = 0$ is 0.

$E\left[|\Gamma(\sqrt{n}\hat{\psi}, \hat{\Sigma})|\right]$ **converges to zero for fixed alternatives:**

Consider the sequence of $s_n^* = n^{-1/4}$

$$\begin{aligned}
\Pr(\|\hat{Z} + s_n^* \hat{\boldsymbol{t}}\|_p \geq c_\alpha) &= \Pr\left(\left\|\hat{Z} + n^{-1/4}\left[\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} + \boldsymbol{\psi}\right)\right]\right\|_p \geq c_\alpha\right) \\
&= \Pr\left(\left\|\hat{Z} + n^{-1/4}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) + n^{1/4}\boldsymbol{\psi}\right\|_p \geq c_\alpha\right) \\
&\geq \Pr\left(\|n^{1/4}\boldsymbol{\psi}\|_p - \left\|\hat{Z} + n^{1/4}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\right\|_p \geq c_\alpha\right) \\
&= 1 - \Pr\left(\|n^{1/4}\boldsymbol{\psi}\|_p - c_\alpha \leq \left\|\hat{Z} + n^{1/4}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\right\|_p\right) \\
&\geq 1 - \frac{\text{var}\left(\left\|\hat{Z} + n^{1/4}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\right\|_p\right)}{\|n^{1/4}\boldsymbol{\psi}\|_p - c_\alpha}
\end{aligned}$$

Since the denominator of the second term grows without bound, we know the above quantity will tend to

1. Because $\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma})$ takes the smallest s such that $\Pr(\|\hat{Z} + s\hat{\boldsymbol{t}}\|_p \geq c_\alpha) \geq 0.8$, and the chosen sequence satisfies $\Pr(\|\hat{Z} + s_n^*\hat{\boldsymbol{t}}\|_p \geq c_\alpha) \to 1$ it follows that $\Gamma(\sqrt{n}\hat{\boldsymbol{\psi}}, \hat{\Sigma})$ must be least as small as s_n^* for large n. Thus for large n it follows that $E\left[|\Gamma(\hat{\boldsymbol{t}}, \Sigma)|\right] \leq E\left[s_n^*\right] \to 0$

Central Symmetry Lemma 8 showed that the function $f(y) = \Pr(\|Z+y\|_p \geq c)$ is centrally symmetric and unimodal for any c.

Therefore, Γ is also centrally symmetric:

$$\Gamma(-x, \Sigma) = \min\left\{s : \Pr(\|Z + -sx\|_p \geq c_\alpha) \geq 0.8\right\} = \min\left\{s : \Pr(\|Z + sx\|_p \geq c_\alpha) \geq 0.8\right\} = \Gamma(x, \Sigma)$$

Unimodality: To show unimodality, consider μ_1 and μ_2 such that $\Gamma(\mu_1, \Sigma)$ and $\Gamma(\mu_2, \Sigma) \geq k$. Let $s = \max(s_1, s_2)$. Theorem 2 and lemma 8 tell us that $\Pr(\|Z+sx\|_p \geq c_\alpha)$ is increasing in s, so if $\Gamma(\mu, \Sigma) \geq k$ then $\Pr(\|Z + k\mu\|_p \geq c_\alpha) \geq 0.8$. This means

$$\begin{aligned} \Pr(\|Z + k\mu_1\|_p \geq c_\alpha) \leq 0.8 \text{ and } \Pr(\|Z + k\mu_2\|_p \geq c_\alpha) \leq 0.8 \\ \text{or equivalently } \Pr(\|Z + k\mu_1\|_p < c_\alpha) > 0.2 \text{ and } \Pr(\|Z + k\mu_2\|_p < c_\alpha) > 0.2 \end{aligned}$$

it follows from lemma 8 that

$$\begin{aligned} \Pr(\|Z + k(t\mu_1 + (1-t)\mu_2)\|_p < c_\alpha) > 0.2 \\ \text{or quivalently } \Pr(\|Z + k(t\mu_1 + (1-t)\mu_2)\|_p \geq c_\alpha) \leq 0.8 \end{aligned}$$

Thus $\Gamma(t\mu_1 + (1-t)\mu_2, \Sigma) \geq k$ as well. □

9.5 Short, but important lemmas

Lemma 10. *The multivariate normal probability density function is unimodal for all* Σ.

Proof. Suppose that both $\phi(\mu_1)$ and $\phi(\mu_2) \geq k$. Because the multivariate normal probability density function is log concave [Tong], it follows that

$$\log\left(\phi(t\mu_1 + (1-t)\mu_2)\right) \geq t\log\left(\phi(\mu_1)\right) + (1-t)\log\left((\phi(\mu_2)\right)$$

Also, since e^x is a strictly increasing function, this would imply:

$$\exp\left(\log\left(\phi(t\mu_1 + (1-t)\mu_2)\right)\right) \geq \exp\left(t\log\left(\phi(\mu_1)\right) + (1-t)\log\left(\phi(\mu_2)\right)\right)$$
$$\phi(t\mu_1 + (1-t)\mu_2) \geq \phi(\mu_1)^t\phi(\mu_2)^{1-t} \geq k^t k^{1-t} = k$$

□

Lemma 11. *Let $f(x)$ be a centrally symmetric, unimodal function. Then $g(x) = I\{f(x) \geq c\}$ where $c \in \mathbb{R}$ is also centrally symmetric and unimodal.*

Proof. Because f is centrally symmetric and g is a function of x only through g, g is also centrally symmetric:

$$g(x) = I\{f(x) \geq c\} = I\{f(-x) \geq c\} = g(-x)$$

If $k < 0$ or $k > 1$, then $\{x : g(x) \geq k\}$ will be the empty set or all of $\mathbb{R}^d$ respectively, and both sets are convex.

Otherwise, the set $\{x : g(x) \geq k\} = \{x : I\{f(x) \geq c\} \geq k\}$. The indicator function will be greater than or equal to k whenever $f(x) \geq c$, so $\{x : g(x) \geq k\} = \{x : f(x) \geq c\}$ which is convex because f is unimodal.

Thus g is unimodal. □

Lemma 12. *If the set A is convex, and if $A_\mu = \{a + \mu : a \in A\}$, then the set $tA_{\mu_1} + (1-t)A_{\mu_2} = \{b : b = ta_{\mu_1} + (1-t)a_{\mu_2}, a_{\mu_1} \in A_{\mu_1}, a_{\mu_2} \in A_{\mu_2}\}$ is equivalent to $A_{t\mu_1+(1-t)\mu_2}$*

Proof. Let $x \in tA_{\mu_1} + (1-t)A_{\mu_2}$. Therefore

$$x = a_{\mu_1}t + a_{\mu_2}(1-t) \text{ where } a_{\mu_1} \in A_{\mu_1}, a_{\mu_2} \in A_{\mu_2}$$
$$= (a_1 + \mu_1)t + (a_2 + \mu_2)(1-t) \text{ where } a_1, a_2 \in A$$
$$= a_1t + a_2(1-t) + \mu_1t + \mu_2(1-t) \text{ where } a_1, a_2 \in A$$

Because of the convexity of A it follows that $a_1t + a_2(1-t)$ and $x \in A_{\mu_1t+\mu_2(1-t)}$. To show inclusion in the reverse direction, let $y \in A_{\mu_1t+\mu_2(1-t)}$. Then:

$$y = a + \mu_1t + \mu_2(1-t) \text{ where } a \in A$$
$$= (a + \mu_1)t + (a + \mu_2)(1-t) \text{ where } a \in A$$

so $y \in tA_{\mu_1} + (1-t)A_{\mu_2}$. □

Additional conditions for Anderson Theorem

In order to show a strict inequality

$$\int I\left\{\Gamma(x,\Sigma) \geq c_\alpha\right\}\phi(x-\underline{h})dx = \int I\left\{\Gamma(x+\underline{h},\Sigma) \geq c_\alpha\right\}\phi(x)dx < \int I\left\{\Gamma(x,\Sigma) \geq c_\alpha\right\}\phi(x)dx$$

it must be shown that for at least one u, $K_u = \{x : \phi(x) \geq u\}$ and $H_h = \{x : \Gamma(x+\underline{h},\Sigma) \geq c_\alpha\}$ that the volume of $\{K_u \cap H_h\}$ is less than $\{K_u \cap H_0\}$.

Lemma 13. *The above condition holds for all Γ satisfying conditions 1 - 5.*

Proof. Because the set H_0 is closed and bounded, let $x_0 \in \operatorname{argmin}\{\phi(x) : x \in H_0\}$. Now, consider the equivalent point inside of H_h which is the point $x_0 + \underline{h}$. Comparing these two values, compare:

$$x_0^\top \Sigma^{-1} x_0 \text{ v.s. } (x_0+\underline{h})^\top \Sigma^{-1}(x_0+\underline{h}) = x_0^\top \Sigma^{-1} x_0 + 2\underline{h}^\top \Sigma^{-1} x_0 + \underline{h}^\top \Sigma \underline{h}$$

If the value of $2\underline{h}^\top \Sigma^{-1} x_0$ is negative, instead consider $-x_0$. Because ϕ is centrally symmetric we now that $-x_0 \in \operatorname{argmin}\{\phi(x) : x \in H_0\}$ as well. Thus it is the case that $\phi(x_0) > \phi(x_0 + \underline{h}_0)$. Also, because ϕ is continuous, it is also true that for some δ all $x \in B(x_0 + \underline{h}, \delta) = \{y : |y - x_0 + \underline{h}| < \delta\}$ that $\phi(x) < \phi(x_0)$.

Because H_0 has a positive volume, we know that there exists a set of points $\{\omega_1, \ldots, \omega_{d+1}\}$ inside of H_0 which form a convex hull with positive volume (otherwise H_0 would have a volume of zero). Adding an additional point a convex hull can only increase its volume, so the convex hull H^* of the points $\{x_0 + \underline{h}, \omega_1, \ldots, \omega_{d+1}\}$ will also be a subset of H_0 and have a positive volume. Let $L = B(x_0 + \underline{h}, \delta) \cap H^*$. Because $B(x_0 + \underline{h}, \delta)$ is an open set, and both sets contain $x_0 + \underline{h}$, the set L will have a positive volume.

Because of how L is defined, we know that the volumen of $\left\{K_{\phi(x_0)} \cap H_h\right\}$ is less the volume of H_h minus the volume of L. Because the volume of H_h is the same as the volume of H_0, it follows that the volume of $\left\{K_{\phi(x_0)} \cap H_h\right\}$ is less than $\left\{K_{\phi(x_0)} \cap H_0\right\}$ which is equal to the volume of H_0 □

9.6 Consistency of norm selection

This proof would likely be difficult and require some slower than $\sqrt{n}$ covergence rates of the local alternative considered. Under fixed alternatives, all norms perform equally perfectly

9.7 Type 1 error control

Proof is so short, I am not sure if it is worth including Under the null, $Y \perp\!\!\!\perp X$. Thus the test statistic will be taken from the same distribution as all of the permutation based test statistics used to estimate the distirbution of the test statistic under the null. Therefore, as B grows, $Pr(T_n \geq F_{T^{\#}_{k,n}|X_n}(0.95)) \Rightarrow 0.05$

References

T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. 6(2):170–176. ISSN 0002-9939. doi: 10.2307/2032333. URL https://www.jstor.org/stable/2032333.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. 32(3):962–994. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053604000000265. URL https://projecteuclid.org/euclid.aos/1085408492.

Olive Jean Dunn. Estimation of the medians for dependent variables. 30(1):192–197, a. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177706374. URL https://projecteuclid.org/euclid.aoms/1177706374.

Olive Jean Dunn. Multiple comparisons among means. 56(293):52–64, b. ISSN 0162-1459. doi: 10.1080/01621459.1961.10482090. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1961.10482090.

Morris L. Eaton and Michael D. Perlman. Multivariate probability inequalities: Convolution theorems, composition theorems, and concentration inequalities. 19:104–122. URL https://zbmath.org/?q=an%3A0760.60017. MSC2010: 60E15 = Inequalities; stochastic orderings MSC2010: 62H10 = Multivariate distribution of statistics.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. 33(1):1–22. ISSN 1548-7660. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/.

Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. 75(4):800–802. ISSN 0006-3444. doi: 10.1093/biomet/75.4.800. URL https://academic.oup.com/biomet/article/75/4/800/423177.

Sture Holm. A simple sequentially rejective multiple test procedure. 6:65–70. doi: 10.2307/4615733.

Samuel D. Lendle, Joshua Schwab, Maya L. Petersen, and Mark J. van der Laan. ltmle: An r package implementing targeted minimum loss-based estimation for longitudinal data. 81(1):1–21. ISSN 1548-7660. doi: 10.18637/jss.v081.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v081i01`. Number: 1.

Ian W. McKeague and Min Qian. An adaptive resampling test for detecting the presence of significant predictors. 110(512):1422–1433. ISSN 0162-1459. doi: 10.1080/01621459.2015.1095099. URL `https://amstat.tandfonline.com/doi/abs/10.1080/01621459.2015.1095099`.

Rupert G. Jr Miller. *Simultaneous Statistical Inference.* Springer Series in Statistics. Springer-Verlag, 2 edition. ISBN 978-1-4613-8124-2. URL `https://www.springer.com/la/book/9781461381242`.

Wei Pan, Junghi Kim, Yiwei Zhang, Xiaotong Shen, and Peng Wei. A powerful and adaptive association test for rare variants. 197(4):1081–1095. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.114.165035. URL `https://www.genetics.org/content/197/4/1081`.

Iosif Pinelis. Schur2-concavity properties of gaussian measures, with applications to hypotheses testing. 124:384–397. ISSN 0047-259X. doi: 10.1016/j.jmva.2013.11.011. URL `http://www.sciencedirect.com/science/article/pii/S0047259X13002534`.

Eric Polley and Mark van der Laan. Super learner in prediction. URL `https://biostats.bepress.com/ucbbiostat/paper266`.

Yosef Rinott. On convexity of measures. 4(6):1020–1026. ISSN 0091-1798. URL `https://www.jstor.org/stable/2242963`. Publisher: Institute of Mathematical Statistics.

Burt S. Holland and Margaret DiPonzio Copenhaver. Improved bonferroni-type multiple testing procedures. 104:145–149. doi: 10.1037/0033-2909.104.1.145.

Noah Simon, Jerome Friedman, and Trevor Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. URL `http://arxiv.org/abs/1311.6529`.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. 74(2):245–266. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2011.01004.x. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4262615/`.

Y. L. Tong. *The Multivariate Normal Distribution*. Springer Science & Business Media. ISBN 978-1-4613-9655-0. Google-Books-ID: FtHgBwAAQBAJ.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press. ISBN 978-0-521-78450-4.

Gongjun Xu, Lifeng Lin, Peng Wei, and Wei Pan. An adaptive two-sample test for high-dimensional means. 103(3):609–624. ISSN 0006-3444. doi: 10.1093/biomet/asw029. URL `https://academic.oup.com/biomet/article/103/3/609/1744173`.

Yichi Zhang and Eric B. Laber. Comment. 110(512):1451–1454. ISSN 0162-1459. doi: 10.1080/01621459.2015.1106403. URL `https://amstat.tandfonline.com/doi/full/10.1080/01621459.2015.1106403`.