Abstract

It is common across scientific disciplines to be interested in predicting an outcome using some (potentially large) number of covariates. A first step to constructing accurate predictions is to decide if any covariate is associated with the outcome. Testing for the existence of an association can be done with a variety of existing methods, but frequently a trade off is made between generalizability (ability to use the test for different parameters of interest and data generating mechanisms) and power (ability to detect an association when one exists). We propose an approach testing for the existence of an association between an outcome and any number of covariates that can be carried out for most parameters of interest that performs nearly as well as tests designed for a specific parameter and data generating mechanism. We compare our method with other modern methods in a simple simulated data setting. We then study the tests performance in a setting with a complex sampling scheme, and a setting with a complex parameter of interest.

1 Introduction

One of the central goals of science is to be able to predict some outcome, often with the help of other related information. This goal is often attained using the construction of mathematical and statistical models in which many covariates are used to predict the outcome of interest. Because the accuracy of the models is dependent on the presence of an association between the outcome of interest and some number of the covariates it is important to know if any associations exist before attempting to create a model. While this first step is always important, it has become more important as it becomes more common for large numbers of covariates to be collected, often without an explicit outcome of interest in mind.

To test if any covariate is associated with the outcome, a standard approach is to test each covariates association with the outcome. 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 attain improved power compared to Bonferroni-based methods, but often rely on asymptotics to obtain these results, and don't account for the irregularity of the estimator on which their test is based. Subsequent work [McKeague and Qian, Pan et al., Xu et al.] addressed these concerns by accounting for the adaptive nature of the considered tests. While these tests obtain higher power while maintaining type one error control, they were constructed using assumptions about the data generating mechanism, and for only a single measure of association between the outcome of interest and each covariate. 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 a testing procedure is proposed that can be used for a wide variety of data generating mechanisms and parameters of interest, but also achieves comparable power to tailor made procedures. Section 2 describes the data generating mechanisms that are considered to evaluate the performance of the test and the competing test made for the given data generating mechanism. Section 3 proposes the testing procedure. Section 4 shows the performance of our testing procedure in the three considered simulation settings. Section 5 shows the application of our method to real world data. Section 6 summarizes the findings of this article and notes potential weaknesses and extensions of our procedure.

2 Working 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_j's. 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, 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:

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

Last, 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 test's performance.

2.1 Correlation Parameter

We will compare our method to a Bonferroni correct marginal testing method and the method described in [Zhang and Laber]. The settings considered will be the same as the firs setting in [McKeague and Qian]. The parameter of interest, $\psi_j(P)$ will be 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 all considered settings, $\varepsilon \sim N(0, 1)$ and is independent of all X. In the first model $Y = \varepsilon$, in the second model $Y = X_1/4 + \varepsilon$, and in the third model $Y = \sum_{k=1}^{10} \beta_k X_k + \varepsilon$ 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 ρ of $0, 0.5$, and 0.8 are considered. All combinations of model, sample size, dimension and ρ are considered, and every test's performance is measured for all considered settings.

2.2 Missing Data Example

In the second example, Y is binary, Δ is a missingness indicator, and each W_j is a covariate of interest. When $\Delta = 0$ we don't observe Y. The identifying assumption is $\Delta \perp\!\!\!\perp Y | W$, and the parameter of interest is the risk ratio under a poission working model for the probability that $Y = 1$.

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

Using the identifying assumption, the observed data parameter is:

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

The data are drawn from a binomial model:

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

There will be three settings considered which determine the values of the β's in the data generating model:

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

2.3 Marginal Structural Model

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:

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

in which our parameter of interest is:

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

The parameter of interest is β_3^*, but is worth noting that parameter is different from a usual logistic regression parameter because we are marginalizing over all w_i, not just w_j.

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

$$\text{logit}\left(Pr(Y = 1|w, a)\right) = \beta_1 a + \sum_{i=1}^{d} \beta_{i+1} w_i + \sum_{j=1}^{d} \gamma_j w_j a$$

In every setting, $\beta_1 = 0.2$, $\beta_2, \dots \beta_{d/2+1} = 2/\sqrt{d}$, and $\beta_{d/2+1} \dots \beta_{d+1} = 0$. 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 $\gamma_1 \ldots \beta_5 = 1, \gamma_6 \ldots \beta_{10} = -1$, and $\gamma_{11} \ldots \gamma_d = 0$. Data are generated from all three models with every possible combination of sample size ($n = 100$, or 200), and dimension ($d = 10, 50, 100, 150$, or 200).

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 an 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:

$$Pr_Q\{Z \notin \Theta_0(P)\} = 1 - \alpha \text{ for every } P_0 \in \mathcal{M}_0. \tag{1}$$

While there are many regions satisfying (1), 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:

$$\Theta_0(r) = \{\omega : \|\omega\|_2 \leq r\}. \tag{2}$$

A region satisfying (1) and (2) has a radius defined by:

$$r_\alpha(P) = \min\{r : Pr_Q(\|Z\|_2 \leq r) \geq 1 - \alpha\}. \tag{3}$$

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 $\boldsymbol{\psi}$. Suppose for now that $\sqrt{n}\hat{\boldsymbol{\psi}}$ 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

$$\text{reject } H_0 \text{ if } \|\sqrt{n}\hat{\boldsymbol{\psi}}\|_2 \geq r_\alpha(P_0), \tag{4}$$

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 of constructing a test of H_0, but it is easy to imagine other tests defined using different norms. The natural next question is which test will perform the best, and does the choice of norm make a difference. To explore this question, consider a simple example comparing the test described in equation (4) with a test that is identical except it uses the ℓ_∞ norm (maximum absolute value) instead of the ℓ_2 norm. Figure 1, panel (A) illustrates these two tests in $\mathbb{R}^2$. One hundred draws are taken from

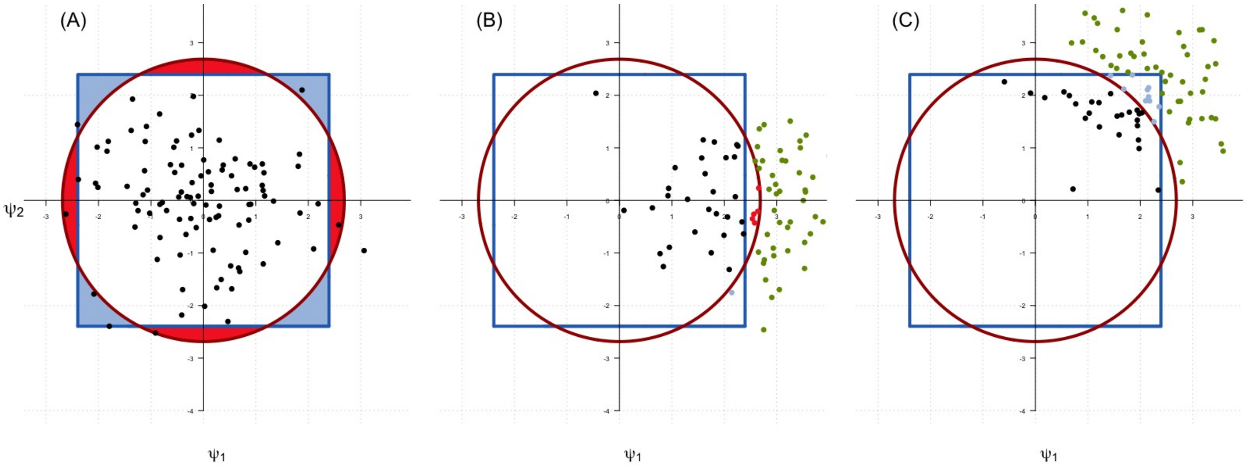

Figure 1: Plots of 100 observations from a limiting distribution of a hypothetical vector of parameter estimators in $\mathbb{R}^2$ (A) under the null, (B) under an alternative with $\psi_1 = 0, \psi_2 \neq 0$, and (C) under an alternative with $\psi_1, \psi_2 \neq 0$. The 95% quantiles for the data 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.

Y, 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 $\ell_2(Y)$ and $\ell_\infty(Y)$ respectively corresponding to $r_\alpha(P)$ from (3). The same square and circle are redrawn in panels (B) and (C) to illustrate the test's performance 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. Panels B shows draws from an alternative in which $\psi_2 \neq 0$ and $\psi_1 = 0$ and the ℓ_∞ norm performs better (achieves higher power). Panel C shows draws from an alternative in which ψ_1 and $\psi_2 \neq 0$ and the ℓ_2 norm performs better.

While both acceptance regions are created to achieve asymptotic type 1 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 for the true alternative. The procedure proposed in this section adaptively selects a norm and it will be shown that this procedure achieves greater power than a test with a fixed norm, while maintaining type 1 error control.

The test defined in (4) is a function of three objects; the chosen norm, $\sqrt{n}\hat{\boldsymbol{\psi}}$, and the limiting distribution of $\hat{t} = \sqrt{n}\hat{\boldsymbol{\psi}}$. Note that the limiting distribution of $\hat{t}$ under the null will always be a multivariate normal with mean zero with some covariance Σ. Thus potential tests can be defined using the finite dimensional Σ matrix rather than the infinite dimension distribution $Q(P)$.

In order to choose between tests (or norms) it will be important to measures the tests performance in a way that is consistent across norms. While all tests will reject as the test statistic becomes larger, if the measure of performance was the norm $\hat{t}$ then the optimal norm would always be the ℓ_1 norm. To achive this goal, we will consider other mappings from $\mathbb{R}$ into $\mathbb{R}$ that are comparable across norms. Consider a general function Γ_Σ, of a vector in $\mathbb{R}^d$ (corresponding to $\hat{\boldsymbol{t}}$). While $\Gamma_\Sigma(x)$ should get larger as x moves away from the null, Γ_Σ can be quite general. An example of a more complicated Γ_Σ is given below:

$$\Gamma_\Sigma(x) = Pr_Q(\|Y + x\| > c_{0.8}) \text{ where } c_{0.8} \equiv \min_c\{c : Pr(\|Y\| < c) \geq 0.8\} \text{ and } Y \sim N(0, \Sigma). \tag{5}$$

We can define the test using Γ_Σ as

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

Considering the Γ_Σ for which it is possible to compare across norms, we can now define a new, adaptive Γ^*_Σ which is the pointwise maximum of all of the considered Γ_Σ's. First define $\Gamma_{\Sigma,1}(x), \ldots, \Gamma_{\Sigma,p}(x)$ as collection of functions in which the functions only differ by the norm used in their definition. Next, define

$$\Gamma^*_\Sigma(x) = \max\left\{\Gamma_{\Sigma,1}(x), \Gamma_{\Sigma,2}(x), \dots, \Gamma_{\Sigma,p}(x)\right\}.$$

This could be the min if smaller values indicate larger distances away from the norm (such as p-values).

While this function is more complicated than before, distribution of $\Gamma^*_\Sigma(Y)$ can still be compared to $\Gamma^*_\Sigma(\hat{\boldsymbol{t}})$ to obtain a p-value. Also, while it may be difficult to obtain the exact distribution of $\Gamma^*_\Sigma(Y)$, the distribution is a function of Σ, so obtaining good approximations of $\Gamma^*_\Sigma(Y)$ 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 $\hat{\boldsymbol{\psi}}_1, \dots, \hat{\boldsymbol{\psi}}_d$ of $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_d$ is asymptotically linear. That is for each $j \in \{1, \dots, d\}$:

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

Where the D_j function will be referred to as an influence function. While this requirement may seem rigid, most standard measures of association are asymptotically linear, and there exists a rich and growing body of literature that describes 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 the vector of parameter estimates is asymptotically normal with mean zero, and variance covariance matrix given by $\Sigma = E_{P_0}\left[D(X)D(X)^\top\right]$:

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

Under H_0, $\sqrt{n}\hat{\boldsymbol{\psi}}$ converges to Z, and Σ can be approximated with $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} D(\boldsymbol{x}_i)D(\boldsymbol{x}_i)^\top$. 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 $\Gamma_{\hat{\Sigma}}(\hat{\boldsymbol{t}})$ will be compared to $\Gamma_{\hat{\Sigma}}(Y)$.

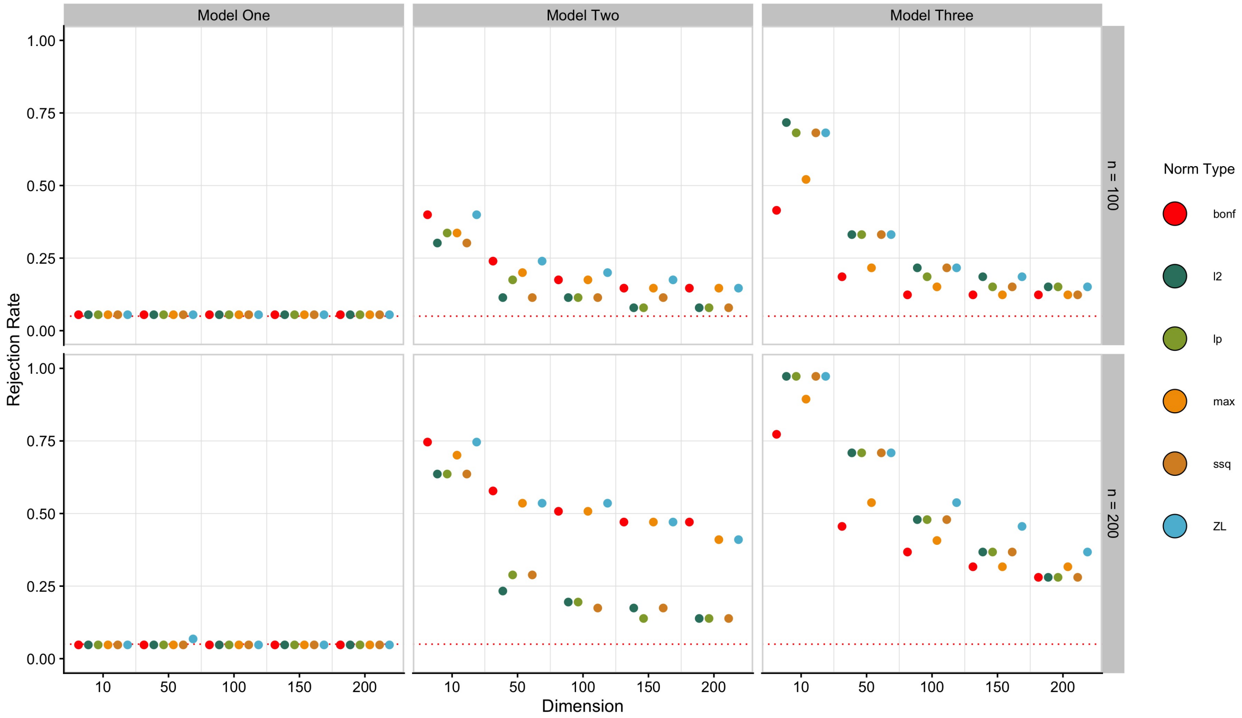

Figure 2: Display of simulations for vector of covariates in this setting will be generated from a normal distribution with mean zero and a variance covariance of Σ with Σ_{ij} equal to 0 when $i \neq j$ and equal to 1 when $i = j$. Three different models for the outcome of interest (Y) will are considered. Letting $\varepsilon \sim N(0,1)$ and be independent of X, in the first model $Y = \varepsilon$, in the second $Y = X_1/4$, and in the third $Y = \sum_{k=1}^{10} \beta_k X_k + \varepsilon$ 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.

3.3 Using a permutation test for the test statistic

While the above approach works asymptotically, there can be issues for small sample sizes. To avoid inflated type one error, a permutation based test can be used. Here, $\hat{Q}^{\#}$ is used to define Γ_{Σ}, and the test will compare $\Gamma_{\hat{Q}}(\sqrt{n}\hat{\psi})$ to $\Gamma_{\hat{Q}}(Z^{\#})$. To determine $\hat{Q}^{\#}$, the Y's from the observed data are permuted before calculating $\widehat{\Sigma}$. Draws from $Z^{\#}$ are taken by permuting all of the Y's of the observed data. There are a few more complications here that need to be figured out.

4 Simulation Study

4.1 Correlation

In the above figure, the rejection rates of six different tests are shown for a wide variety of settings. Data are generated from three different potential models. In the first model no covariates are directly associated with the outcome (H_0 holds), in the second model a single covariate is strongly directly associated with the

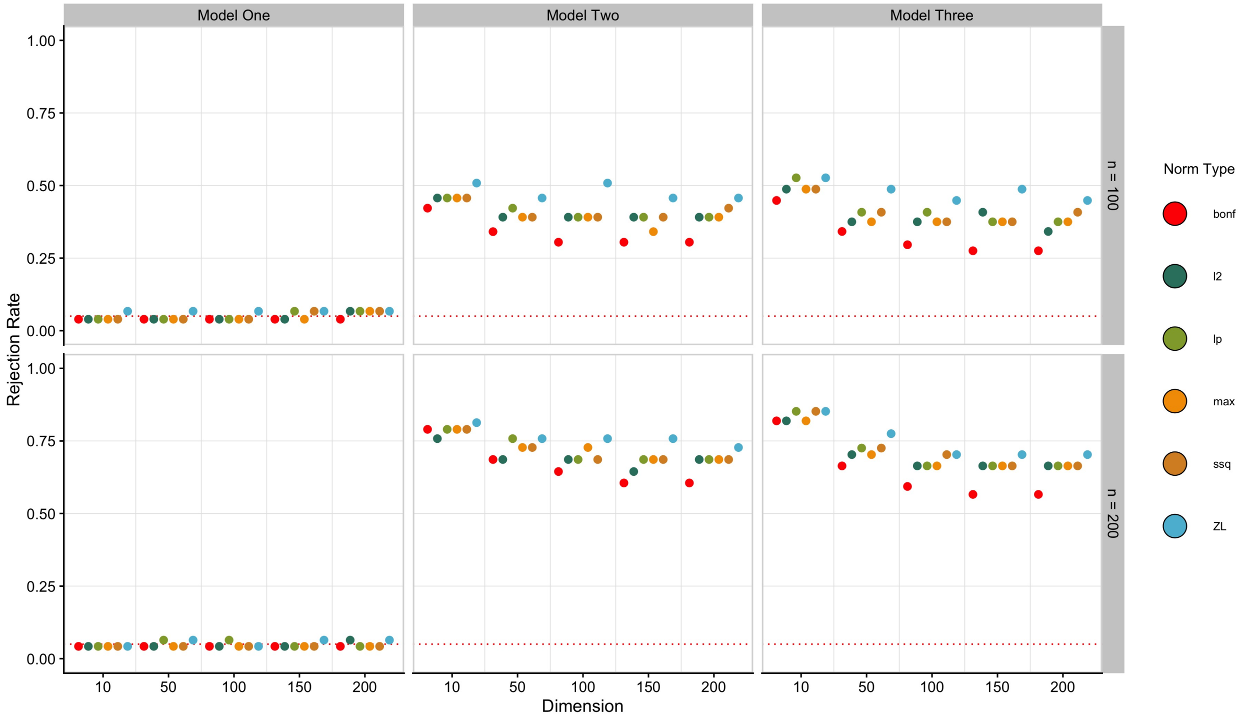

Figure 3: The same simulation settings as those used in Figure 2, but $\Sigma_{ij} = 0.5$ for $i \neq j$.

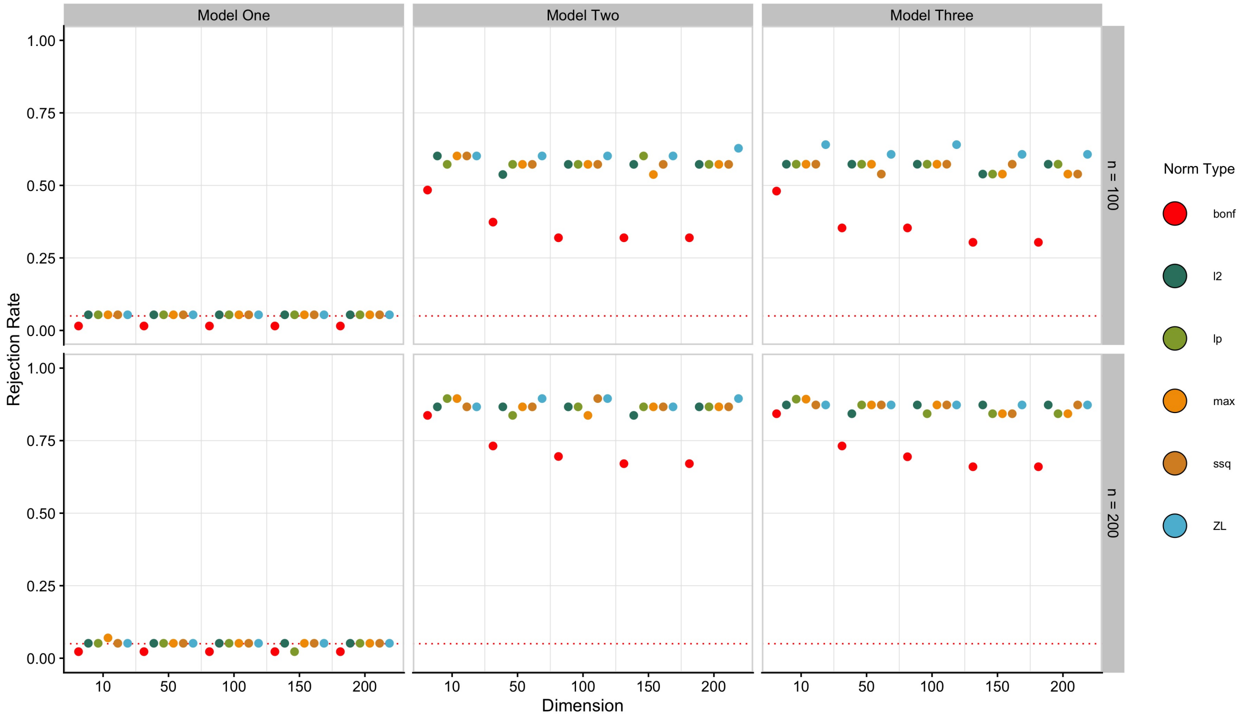

Figure 4: The same simulation settings as those used in Figure 2, but $\Sigma_{ij} = 0.8$ for $i \neq j$.

outcome (H_0 does not hold), and in the third model ten covariates are directly associated with the outcome (H_0 does not hold). In all models the covariates are generated from a mean zero multivariate normal distribution Σ, where $\Sigma_{ij} = 0.8$ for $i \neq j$ and 1 for $i = j$. Each color of dots represents a different type of test. The red dots indicate a bonferroni adjusted marginal test for each covariates estimated correlation with the outcome. The light blue dots indicate 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

Here, we would expect that the ℓ_2 norm would perform poorly in the model 2 setting while performing well in the model 3 setting, and that the max norm would perform poorly in the model three setting while performing well in the model two setting. However, we see the opposite behavior. Additionally, we find that our adaptive test performs well consistently across all alternatives.

I will eventually just chose one of these figures because performance will be very similar between the two measures. I just want to see how similar they are once I get a complete set of simulation results.

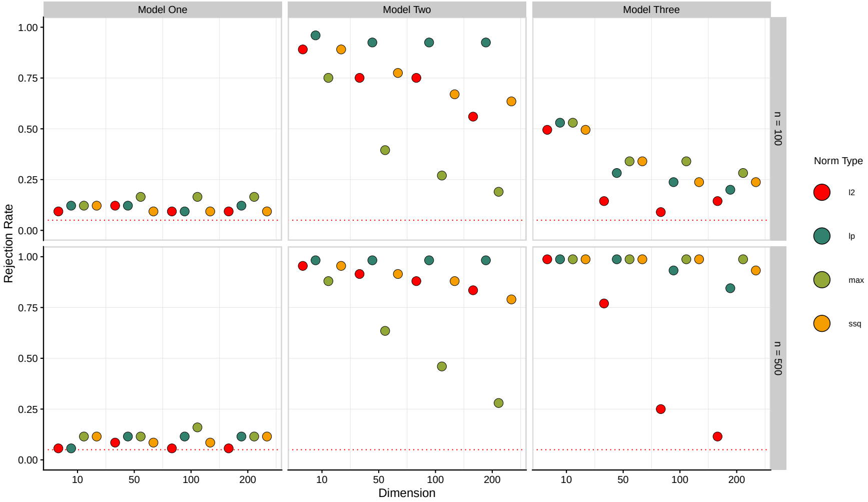

Figure 5: Rejection rates across three different data generating mechanisms. Here we are using the the multiplicative distance our test statistic is from obtaining 80% power as our measure of performance.

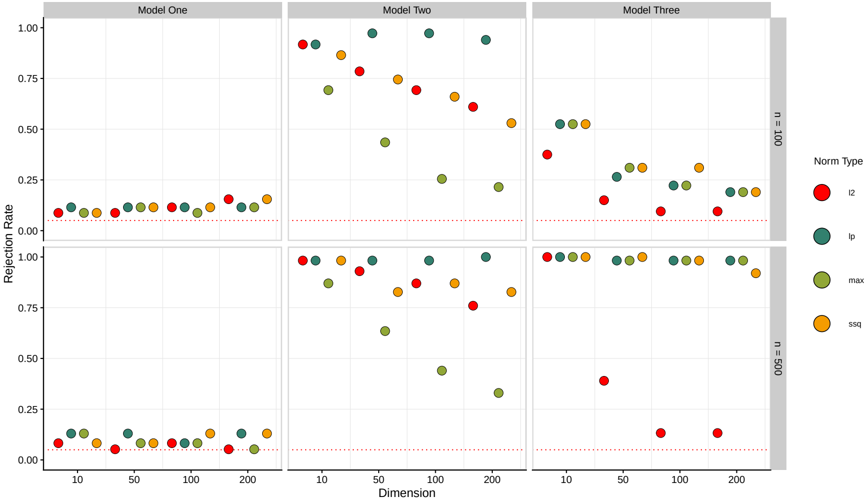

Figure 6: Rejection rates across three different data generating mechanisms. Here we are using estimated power as the measure of performance.

5 Data Application

1. Data from Peter Gilbert

Marginal Structural Model

6 Discussion

In this paper, we have discussed the difficulties of creating multivariate test of no correlated coefficients. General methods are have wide applicability, but are often underpowered. More powerful methods are frequently difficult to understand, and will only work properly in narrow settings for a single parameter of interest.

We described a test that addresses these issues, by achieving comparable performance to taylor made methods, while being applicable in a wide variety of settings. We have demonstrated the methods ability to compete with existing methods in section 4, and shown novel settings in which the method can also be performed. Data example 2 was an example of testing a relatively simple parameter (a risk ratio) in a setting with a non-standard sampling scheme. Data example 3 was an example of testing a more complex parameter (a parameter inside of a marginal structural model) for a straightforward sampling scheme.

While our procedure does rely on asymptotic results to function properly for many parameters, in certain cases, a our test can be carried out using permutations of the data to obtain better finite sample properties. This is done by repeating our estimation procedure many times after randomly permuting the outcome variable. The method is used in our first data example, and could be used for our third data example. However, in certain settings with more complicated sampling schemes, the permutation based test is more difficult to implement because many of the outcomes are not even observed.

While we have outlined some of the ways in which our method can be used, there are many other ways in which it can be extended improved.

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 Γ's by $\hat{\Sigma}$, using a multiplier bootstrap to take draws from $N(0, \Sigma)$ avoids the computation issues of having a degenerate estimate of the limiting distributions variance matrix.

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_Σ funcitons in this article, any set of reasonable Γ_Σ functions could be used. These Γ_Σ 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_Σ 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 $\sqrt{n}\hat{\boldsymbol{\psi}}$ 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 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 (say $2 \times \mathrm{var}(X)\sqrt{n}$). 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 fin influence function)

Multiple measures of association for each covariate.)

7 Conclusion

8 Appendix

8.1 Test Consistency

Theorem 1. *Assume the performance metric of choice is $\hat{r}_{\alpha,p}(a)$, and each the norms g_a considered have the following properties:*

$$\text{for } x \in \mathbb{R}^d,\ \text{ and } s, l > 0, s \cdot \max(\boldsymbol{x}) \leq g_a(\boldsymbol{x}) \leq l \cdot d \cdot \max(\boldsymbol{x}) \tag{6}$$

$$\text{for } s \leq 1, l \geq 1, g_a(s \cdot \boldsymbol{x}) \leq g_a(\boldsymbol{x}) \leq g_a(l \cdot \boldsymbol{x}) \tag{7}$$

Then for all $P \notin \mathcal{M}_0$

$$Pr\left(\frac{1}{B}\sum_{k=1}^{B} I\left\{\hat{T} \leq \hat{T}_k^{\#}\right\} < 0.05\right) \xrightarrow{p} 1 \ as \ n \to \infty$$

or Equivalently

$$Pr\left(\hat{T} \leq F^{-1}_{\hat{T}^{\#}}(0.05)\right) \xrightarrow{p} 1 \ as \ n \to \infty$$

Proof. This proof will consist of three parts. We will first show that as $n \to \infty, \hat{T}_a^{\#}$ becomes bounded away from 0 for any valid norm. Next we will show $\hat{T}_a$ converges to 0 in probability for any valid norm. Last we will show the two previous findings imply theorem 1.

Let $P_X^{\#}$ denote the distribution of the randomly permuted observations. Because $Pr(Y_i^{\#} \perp\!\!\!\perp \boldsymbol{W}_i^{\#}) \to 0$, $\boldsymbol{\Psi}(P_X^{\#}) = \mathbf{0}$. Additionally, $\sqrt{n}\left(\hat{\boldsymbol{\psi}}^{\#} - \boldsymbol{\psi}^{\#}\right) \xrightarrow{d} Z^{\#} \sim N\left(\mathbf{0}, \Sigma_{\text{perm}}\right)$. Define $\hat{T}_a^{\#} = \min_s\{s : g_a(s \cdot \sqrt{n}\hat{\boldsymbol{\psi}}^{\#}) \geq C^{\#}_{0.95,a}\}$ and $C^{\#}_{0.95,a}$ is $F^{-1}_{Z^{\#}}(0.95)$. Since we know that $\psi^{\#} = \mathbf{0}$, it follows that $\sqrt{n}\hat{\boldsymbol{\psi}}^{\#} \overset{d}{\approx} Z^{\#}$.

Now, consider:

$$\begin{aligned}
\Pr\left(\hat{T}_a^{\#} > \epsilon\right) &= \Pr\left(g_a\left(\epsilon \cdot \sqrt{n}\hat{\boldsymbol{\psi}}^{\#}\right) \leq C_{0.95,a}\right) \\
&\geq \Pr\left(\epsilon \cdot d \cdot \max\left(\sqrt{n}\hat{\boldsymbol{\psi}}^{\#}\right) \leq C_{0.95,a}\right) \\
&= \Pr\left(\max\left(\sqrt{n}\hat{\boldsymbol{\psi}}^{\#}\right) \leq C_{0.95,a}/(\epsilon \cdot d)\right)
\end{aligned}$$

Because $\max\left(\left|\sqrt{n}\hat{\boldsymbol{\psi}}^{\#}\right|\right)$ converges to a well defined, positive distribution as a result of the continuous mapping theorem, for each constant $c < 1$, we know there exists an ϵ_c such that $\Pr\left(\hat{T}_a^{\#} > \epsilon_c\right) \geq c$.

Now, shifting our focus to $\hat{T}_a$, under alternatives, $\boldsymbol{\psi} \neq \mathbf{0}$. Define $\psi_{\max} = \max(\psi_1, \ldots, \psi_d)$. Using this knowledge, and (6), note that

$$\begin{aligned}
\Pr\left(\hat{T}_a < \epsilon\right) &= \Pr\left(g_a\left(\epsilon \cdot \sqrt{n}\hat{\boldsymbol{\psi}}\right) \geq C_{0.95,a}\right) \\
&\geq \Pr\left(\epsilon \cdot \sqrt{n}\max(\hat{\psi}_1, \ldots, \hat{\psi}_d) \geq C_{0.95,a}\right) \\
&= \Pr\left(\max(\hat{\psi}_1, \ldots, \hat{\psi}_d) \geq C_{0.95,a}/\left(\epsilon \cdot \sqrt{n}\right)\right) \\
&\geq \Pr\left(\max(\hat{\psi}_1, \ldots, \hat{\psi}_d) \geq \psi_{\max}/2\right) \Pr\left(\psi_{\max}/2 \geq C_{0.95,a}/\left(\epsilon \cdot \sqrt{n}\right)\right) \qquad (8)
\end{aligned}$$

The first factor of the product in (8) will converge to 1 as $n \to \infty$ from the consistency of $\hat{\boldsymbol{\psi}}$. The second quantity will be equal to 1 for sufficiently large n. Thus $\hat{T}_a \xrightarrow{p} 0$ under any alternative.

It was shown that for each a that $\hat{T}_a \xrightarrow{p} 0$. This means that our adaptive estimator $\hat{T} \xrightarrow{p} 0$ as well. Now, let $c = 0.05/k$ and ϵ_c be small enough that $\Pr\left(\hat{T}_a^{\#} > \epsilon_c\right) \geq 1 - (0.05/k)$. The permutation version of the adaptive estimator $\hat{T}^{\#}$ has the property that

$$\Pr\left(\hat{T}^{\#} < \epsilon_c\right) \leq \Pr(\hat{T}_1^{\#} < \epsilon_c) + \cdots + \Pr(\hat{T}_k^{\#} < \epsilon_c) \leq 0.05,$$

and the theorem's conclusion follows. □

8.2 Unbiasedness at local alternatives

Consider a local alternative in which the true value of ψ is shrinking towards zero at a root n rate: $\psi = \underline{h}/\sqrt{n}$. We assume that each potential norm is convex. This assumption can be relaxed to what was described by Eaton and Perlman (1991) (Concentration inequalitites for multivariate distributions: mv normal). Under this local alternative, we will have $\sqrt{n}\hat{\psi} \xrightarrow{d} N(\underline{h}, \Sigma)$. Show that the test will reject the null with a probability greater than α for an α level test

Theorem 2. *Under local alternatives described above,*

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

Where $\hat{Z} \sim \hat{Q}$ and $Z \sim Q$. Here (unlike other parts of the paper) small values of Γ provide evidence against the null. Values of Γ can be thought of as similar to p-values.

Lemma 3. *The function:*

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

is continuous with respect to Σ and t.

This lemma will follow because Γ is the integral of a composition of bounded, continuous functions.

Proof. The multivariate normal probability density function

$$\phi(x, \mu, \Sigma) = (2\pi)^{-k} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)i\right)$$

is continuous with respect to both Σ and t.

Since the exponential function, determinate, matrix inverses, and linear operators are all differentiable, they are also all continuous. Because the multivariate normal pdf is the composition of continuous functions, it is also continuous.

It follows that if $\mu_n \to \mu$, and $\Sigma_n \to \Sigma$, then for each x, $\phi(x, \mu_n, \Sigma_n) \to \phi(x, \mu, \Sigma)$. This result and the dominated convergence theorem imply the corresponding CDF's are also continuous ($\phi(x, \hat{\mu}, \hat{\Sigma}) + \phi(x, \mu, \Sigma)$ can be used as the dominating measure): (dominating function shouldn't depend on n) To find a dominating measure, assume that Σ_n^{-1} is close enough to Σ^{-1} so the smallest eigenvalue is within ε then bound the pdf using this fact and that $(x-\mu)^\top \Sigma^{-1}(x-\mu) = (x-\mu)^\top ADA^{-1}(x-\mu) \leq \|(x-\mu)\|^2 c$ where c is the largest eigen value (can be done to provide oposite direction as well).

$$\Phi_{\hat{\Sigma},\hat{\mu}}(t) = \int_{\|x\|<t} \phi(x, \hat{\mu}, \hat{\Sigma})dx \to \int_{\|x\|<t} \phi(x, \mu, \Sigma)dx = \Phi_{\Sigma,\mu}(t).$$

Show (or find reference showing that) $\Phi^{-1}_{\Sigma_n,0}(0.8) \to \Phi^{-1}_{\Sigma,0}(0.8)$ when $\Sigma_n \to \Sigma$. E.g., implicit function theorem will do this

Because the normal cdf is continuous, we know the normal quantile function will be continuous as well. Potentially we can prove the existence of the inverse using the fact the CDF has a bounded derivative on all of $\mathbb{R}$. Let $\Sigma_n \to \Sigma$ and $h_n \to h$. Also, let Q_n be a normal distribution with mean zero and variance-covaraince Σ_n. These findings imply the following:

$$\begin{aligned}
|\Gamma_{\Sigma_n}(h) - \Gamma_{\Sigma}(h)| &= \left| Pr_{Q_n}(\|\tilde{Z} + h\| > \Phi^{-1}_{\Sigma_n,0}(0.8)) - Pr_Q(\|\tilde{Z} + h\| > \Phi^{-1}_{\Sigma,0}(0.8)) \right| \\
&= \left| \int_{\mathbb{R}^d} \phi(t, 0, \Sigma_n) I\{\|t + h\| > \Phi^{-1}_{\Sigma_n,0}(0.8)\} - \phi(t, 0, \Sigma) I\{\|t + h\| > \Phi^{-1}_{\Sigma,0}(0.8)\} dt \right| \\
&= \left| \int_{\mathbb{R}^d \setminus \left\{t: ||t|| = \Phi^{-1}_{\Sigma,0}(0.8)\right\}} \phi(t, h, \Sigma_n) I\{\|t\| > \Phi^{-1}_{\Sigma_n,0}(0.8)\} - \phi(t, h, \Sigma) I\{\|t\| > \Phi^{-1}_{\Sigma,0}(0.8)\} dt \right| \\
&\leq \int_{\mathbb{R}^d \setminus \left\{t: ||t|| = \Phi^{-1}_{\Sigma,0}(0.8)\right\}} |\phi(t, h, \Sigma_n) I\{\|t\| > \Phi^{-1}_{\Sigma_n,0}(0.8)\} - \phi(t, h, \Sigma) I\{\|t\| > \Phi^{-1}_{\Sigma,0}(0.8)\}| dt
\end{aligned}$$

The above quantity converges to zero by the dominated convergence theorem.

$$
\begin{aligned}
&= \int_{\|t\|>\Phi^{-1}_{\Sigma_n,0}(0.8),\Phi^{-1}_{\Sigma,0}(0.8)} |\phi(t,h,\Sigma_n) - \phi(t,h,\Sigma)|dt + \\
&\qquad \int_{\Phi^{-1}_{\Sigma,0}(0.8)\geq\|t\|>\Phi^{-1}_{\Sigma_n,0}(0.8)} |\phi(t,h,\Sigma_n)|dt + \int_{\Phi^{-1}_{\Sigma_n,0}(0.8)\geq\|t\|>\Phi^{-1}_{\Sigma,0}(0.8)} |\phi(t,h,\Sigma)|dt \\
&\leq \int_{\|t\|>\Phi^{-1}_{\Sigma,0}(0.8)} |\phi(t,h,\Sigma_n) - \phi(t,h,\Sigma)|dt + \\
&\qquad \int_{\Phi^{-1}_{\Sigma,0}(0.8)\geq\|t\|>\Phi^{-1}_{\Sigma_n,0}(0.8)} \phi(t,h,\Sigma_n)dt + \int_{\Phi^{-1}_{\Sigma_n,0}(0.8)\geq\|t\|>\Phi^{-1}_{\Sigma,0}(0.8)} \phi(t,h,\Sigma)dt \\
&= \int_{\|t\|>\Phi^{-1}_{\Sigma,0}(0.8)} |\phi(t,h,\Sigma_n) - \phi(t,h,\Sigma)|dt + \\
&\qquad \Phi_{\Sigma_n,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma_n,h}(\Phi^{-1}_{\Sigma_n,0}(0.8)) + \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8)) \\
&= \int_{\|t\|>\Phi^{-1}_{\Sigma,0}(0.8)} |\phi(t,h,\Sigma_n) - \phi(t,h,\Sigma)|dt + \\
&\qquad \Phi_{\Sigma_n,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma_n,h}(\Phi^{-1}_{\Sigma_n,0}(0.8)) - \Big(\Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8))\Big) + \\
&\qquad \Big(\Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8))\Big) + \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8)) \\
&= \int_{\|t\|>\Phi^{-1}_{\Sigma,0}(0.8)} |\phi(t,h,\Sigma_n) - \phi(t,h,\Sigma)|dt + \\
&\qquad \Big(\Phi_{\Sigma_n,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8))\Big) - \Big(\Phi_{\Sigma_n,h}(\Phi^{-1}_{\Sigma_n,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8))\Big) + \\
&\qquad \Big(\Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8))\Big) + \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma,0}(0.8)) - \Phi_{\Sigma,h}(\Phi^{-1}_{\Sigma_n,0}(0.8))
\end{aligned}
$$

Taking the limit as $n \to \infty$ of the quantity above, we find that the first term is zero by dominated convergence theorem. The other four terms are also zero because of the uniform continuity of $\Phi_{\Sigma,x}$ and continuity of $\Phi^{-1}_{\Sigma,x}$ □

Lemma 4. *Under local alternatives,*

$$\Gamma^*_{\hat{\Sigma}}(\sqrt{n}\hat{\boldsymbol{\psi}}) \xrightarrow{d} \Gamma^*_{\Sigma}(Z + \underline{h})$$

Proof. It was shown in Lemma 3 that for each norm $\|\cdot\|$, $\Gamma_\Sigma(x)$ is continuous with respect to Σ and x. This finding, the continuity of the max function, and because a composition of continuous functions is also continuous it follows that

$$\Gamma^*_{\Sigma} \equiv \max\{\Gamma_{1,\Sigma}, \ldots, \Gamma_{d,\Sigma}\}$$

is also continuos with respect to Σ and x. Under local alternatives, $\sqrt{n}\hat{\boldsymbol{\psi}} \xrightarrow{d} Z + \underline{h}$ and $\hat{\Sigma} \xrightarrow{p} \Sigma$. It follows from the continuous mapping theorem that $\Gamma^*_{\hat{\Sigma}}(\sqrt{n}\hat{\boldsymbol{\psi}}) \xrightarrow{d} \Gamma^*_{\Sigma}(Z + \underline{h})$ □

It has now been establish that under local alternatives the distribution of our estimate converges to $\Gamma^*_{\Sigma}(Z + \underline{h})$. Denote the CDF of $\Gamma^*_{\Sigma}(Z + t)$ this distribution by $F_{\Sigma,t}$ and the corresponding quantile function of the distribution by $F^{-1}_{\Sigma,t}$.

To prove unbiasedness at local alternatives, we show >?

$$F_{\Sigma,t}(F^{-1}_{\Sigma,0}(1-\alpha)) \geq F_{\Sigma,0}(F^{-1}_{\Sigma,0}(1-\alpha)) = \alpha$$

using results from the [Anderson] manuscript.

A result of [Anderson] is that for two centrally symmetric, unimodal functions, $f_1(x)$ and $f_2(x)$, the convolution,

$$g(\theta) = \int f_1(x) f_2(x - \theta)$$

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 decreasing function of β.

For now, we will assume that. We assume that our tests will always be define our tests in such a way that we reject the null when $\Gamma^*_{\hat{\Sigma}}(\sqrt{n}\hat{\boldsymbol{\psi}}) > c$ for some c. It is expected that $\Gamma^*_{\Sigma}(x)$ increases as x moves away from the origin. For the purposes of this proof it will be useful to define new functions

$$\Upsilon_{i,\Sigma} \equiv (\Upsilon_{i,\Sigma})^{-1} \text{ and } \Upsilon^*_{\Sigma} = \min\{\Upsilon_{1,\Sigma}, \ldots, \Upsilon_{d,\Sigma}\}$$

that decrease as x moves away from the origin.

Lemma 5. *Let $\Upsilon_1, \ldots, \Upsilon_d$ all be centrally symmetric, unimodal functions. Then $\Upsilon^* = \min\{\Upsilon_1, \ldots, \Upsilon_d\}$ is also centrally symmetric and unimodal.*

Proof. Because each Υ_i is a centrally symmetric and Υ^* is a function of x only through the Υ_i's, Υ^* is also centrally symmetric:

$$\Upsilon^*(x) = \min\{\Upsilon_1(x), \ldots, \Upsilon_d(x)\} = \min\{\Upsilon_1(-x), \ldots, \Upsilon_d(-x)\} = \Upsilon^*(-x)$$

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

Lemma 6. *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. □

Because each $\Upsilon_{\Sigma,i}$ is unimodal and centrally symmetric, lemma 5 implies $\Upsilon^*_{\Sigma,i}$ is also unimodal and centrally symmetric. It follows from lemma 6 and the previous finding that $I\left\{\Upsilon^*_Q(x) \geq c_{0.95}\right\}$ is centrally symmetric and unimodal.

$$\begin{aligned}\int I\left\{\Gamma^*_Q(x) > c_{0.95}\right\}\phi(x-\mu)dx &= \int I\left\{\Upsilon^*_Q(x) < c^{-1}_{0.95}\right\}\phi(x-\mu)dx \\ &= \int \left(1 - I\left\{\Upsilon^*_Q(x) \geq c^{-1}_{0.95}\right\}\right)\phi(x-\mu)dx \\ &= 1 - \int I\left\{\Upsilon^*_Q(x) \geq c^{-1}_{0.95}\right\}\phi(x-\mu)dx\end{aligned}$$

Since the subtracted quantity is decreasing by [Anderson], the quantity as a whole will be increasing. Thus local power is obtained.

While each set of performance measures will require a proof that they are centrally symmetric and unimodal, we will show that the estimate power performance measure is centrally symmetric and unimodal. Consider the power function.

$$\Gamma_\Sigma(\mu) = \int I\left\{\|x\|_p > c\right\}\phi_\Sigma(x-\mu)$$

Consider two values μ_1, μ_2 such that $\Gamma_\Sigma(\mu_1), \Gamma_\Sigma(\mu_2) \geq k$. Now, consider $\Gamma_\Sigma\left(t\mu_1 + (1-t)\mu_2\right)$

$$\begin{aligned}
\Gamma_\Sigma\left(t\mu_1 + (1-t)\mu_2\right) &= \int I\left\{\|x\|_p > c\right\}\phi_\Sigma(x - t\mu_1 + (1-t)\mu_2)dx \\
&= \int I\left\{\|x\|_p > c\right\}\phi_\Sigma(t(x-\mu_1) + (1-t)(x-\mu_2))dx \\
&\geq \int I\left\{\|x\|_p > c\right\}\left[t\phi_\Sigma(x-\mu_1)) + (1-t)\phi_\Sigma(x-\mu_2)\right]dx \\
&= t\int I\left\{\|x\|_p > c\right\}\phi_\Sigma(x-\mu_1))dx + (1-t)\int I\left\{\|x\|_p > c\right\}\phi_\Sigma(x-\mu_2)dx \\
&= t\Gamma_\Sigma(\mu_1) + (1-t)\Gamma_\Sigma(\mu_2) \geq k
\end{aligned}$$

The inequality on the third line comes from the fact that multivariate normal pdf is concave.

Proof Outline:

- Under local alternatives, $\sqrt{n}\hat{\boldsymbol{\psi}} \xrightarrow{d} N(\boldsymbol{c}, \Sigma)$
- With a large enough sample size, and enough MC draws, we have that for each norm:
- Think of Gamma as a funciton indexed by Σ to allow it to be similar. Also make some assumptions about how smooth Υ is with respect to this parameters. For any given value of the input (t) think about
- Right now we have given up on proving things for the permutation test, but we may try to do it again at some point. The permutation test statistic will converge in distribution to a standard normal. This paper should help: [Omelka and Pauly]
- For each norm selected, [Gupta et al.] states that the power will be non-decreasing as long as the rejection region is convex (This should be true most of our rejection regions), and the probability density is decreasing away from the mean (which is true of a normal distribution).
- Show that for n large, the norm is selected to give the best power. Thus since each norm obtains local power, the adaptive test will also obtain local power for n large.

$$\mathcal{L}\left(\Gamma_{\hat{\Sigma}}\left(\sqrt{n}\hat{\boldsymbol{\Psi}}(\boldsymbol{X})\right), \Gamma_{\Sigma}\left(\sqrt{n}\hat{\boldsymbol{\Psi}}(\boldsymbol{X})\right)\right) \to 0$$

8.3 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

8.4 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.

S. Das Gupta, M. L. Eaton, I. Olkin, M. Perlman, L. J. Savage, and M. Sobel. Inequalitites on the probability content of convex regions for elliptically contoured distributions. The Regents of the University of California. URL https://projecteuclid.org/euclid.bsmsp/1200514222.

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.

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.

M. Omelka and M. Pauly. Testing equality of correlation coefficients in two populations via permutation methods. 142(6):1396–1406. ISSN 0378-3758. doi: 10.1016/j.jspi.2011.12.018. URL http://www.sciencedirect.com/science/article/pii/S0378375811004630.

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.

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

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.