Powerful Partial Conjunction Hypothesis Testing via Conditioning

Biyonka Liang, Lu Zhang, and Lucas Janson

Department of Statistics, Harvard University

Abstract

Research questions across a diverse array of fields are formulated as a Partial Conjunction Hypothesis (PCH) test, which combines information across m base hypotheses to determine whether some subset is non-null. However, standard methods for testing a PCH can be highly conservative. In this paper, we introduce the conditional PCH (cPCH) test, a new framework for testing a single PCH that directly corrects the conservativeness of standard approaches by conditioning on certain order statistics of the base p-values. Under distributional assumptions commonly encountered in PCH testing, the cPCH test produces a p-value that is nearly uniform. Through simulations, we demonstrate that the cPCH test uniformly outperforms standard single PCH tests and maintains Type I error control even under model misspecification, and can in certain situations also be used to outperform state-of-the-art PCH multiple testing procedures. Finally, we illustrate an application of the cPCH test on a replicability analysis of four microarray studies.

Keywords: Composite null, Causal mediation analysis, Replicability analysis, Meta-analysis, Multiple hypotheses testing

1 Introduction

1.1 Motivation

Partial Conjunction Hypothesis (PCH) tests are necessary to address important statistical questions in a diverse array of fields. For example, in areas such as genomics (Liu et al., 2022; Huang, 2019; Dai et al., 2020), psychology (Baron and Kenny, 1986), and social policy evaluation (Karmakar et al., 2021), researchers are interested in understanding complex causal relationships between a cause and an effect. Methods such as causal mediation analysis and causal factor analysis formulate questions about such relationships as the conjunction of links in a causal graph, where each link in the graph represents a hypothesis relating an effect to an outcome. The primary goal of these methods is to identify true causal relationships, which requires procedures for testing whether the hypotheses representing the conjunction of certain links are individually non-null. Another major application area is replicability analysis,

where the partial conjunction of hypotheses represents the corroboration of scientific results across independent studies. In recent years, scientific replicability has become a major topic of interest in the analysis of observational studies, where subtle biases in population, data processing, and diagnostic measures can significantly influence conclusions. "Replicability crises" have been observed across a wide range of domains, including genetic epidemiology (NCI-NHGRI Working Group on Replication in Association Studies, 2007; Kraft et al., 2009; Hirschhorn and Altshuler, 2002), economics (Camerer et al., 2016), psychology (Braver et al., 2014), and medicine (Ioannidis, 2005), underscoring the need for replicability analysis to convincingly validate scientific claims.

1.2 Problem Statement

Partial Conjunction Hypothesis (PCH) testing provides a statistical framework for evaluating the partial conjunction of a set of m hypotheses. Formally, it tests whether or not at least r out of m *base hypotheses*, $H_{0,i}$, $i = 1, ..., m$, are individually non-null, i.e., letting $r^\star$ be the true number of non-null base hypotheses:

Definition 1 (Partial Conjunction Hypothesis (PCH))**.**

$$H_0^{r/m} : r^\star < r$$

where $r \in \{2, 3, ..., m\}$. $H_0^{r/m}$ can equivalently be interpreted as stating that at least $m-r+1$ of the $H_{0,i}$ are *true*, and its complement as stating that at least r of the $H_{0,i}$ are *false*. For example, in replicability analysis, where $H_{0,i}$ tests whether an effect was present in a study, the rejection of $H_0^{2/m}$ provides explicit evidence for scientific replicability (i.e., that an effect was present in two or more studies). In causal mediation and causal factor analysis, where $H_{0,i}$ tests for the presence of a link in a causal chain, the rejection of $H_0^{m/m}$ provides explicit evidence for a causal relationship between the beginning and end of the chain.

Notably, $H_0^{r/m}$ is composite, so the null space consists of r disjoint *null configurations* corresponding to $r^\star = 0, ..., r-1$. For instance, in the $r = m = 2$ case, the null configurations are:

$$H_0^{2/2} = \begin{cases} \text{Both } H_{0,i} \text{ are true } (r^\star = 0), \text{ or} \\ \text{exactly one of } H_{0,i} \text{ is false } (r^\star = 1). \end{cases}$$

In this paper, we consider the setting where the base hypotheses $H_{0,i}$ have independent *base p-values*, $p_1, \ldots, p_m$. Although we have multiple *base* null hypotheses (with corresponding *base* p-values), this paper primarily focuses on the problem of testing a single PCH (at a time), i.e., we aim to test $H_0^{r/m}$ for a *single* set of m base null hypotheses.

1.3 Background and Existing Work

1.3.1 Testing a Single PCH

Standard methods for testing a single PCH apply global null tests to the largest $m - r + 1$ base p-values. Letting $p_{(1)} \leq ... \leq p_{(m)}$ be the sorted base p-values, the resulting *PCH*

p-value, $p^{r/m}$, is of the form

$$p^{r/m}(p_1, ..., p_m) = g\left(p_{(r)}, ..., p_{(m)}\right),$$

where $g : \mathbb{R}^{m-r+1} \to \mathbb{R}$ corresponds to the combining function of the global null test that is used. These standard single PCH tests are generally referred to by the same name as the global null test that they employ, with Bonferroni's, Simes', and Fisher's global null tests being the most common (Wang et al., 2021). For instance, the combining function for Bonferron's test is $g_{\text{B}}\left(p_{(r)}, ..., p_{(m)}\right) = (m - r + 1)p_{(r)}$, which we can recognize as the standard Bonferroni correction applied to the $m - r + 1$ p-values $p_{(r)}, ..., p_{(m)}$. When $r = m$, Bonferroni's, Simes', and Fisher's tests are all equivalent and specifically when $r = m = 2$, they are collectively referred to as the *Max-P test* (Liu et al., 2022).

Intuitively, applying a global null test to the $m - r + 1$ "least promising" (i.e., largest) base p-values tests $H_0^{r/m}$ because rejecting the global null test provides evidence that strictly fewer than $m - r + 1$ base hypotheses are null (equivalently, that at least r base hypotheses are non-null) and therefore, $H_0^{r/m}$ should be rejected. Intuitively, PCH tests of this form are valid because the largest base p-values stochastically dominate the Unif(0,1), so the resulting PCH p-value will also stochastically dominate the Unif(0,1) (Benjamini and Heller, 2008). Formally, Simes' and Fisher's tests are guaranteed to be valid when the individual base p-values are independent while Bonferroni's test is valid under any dependency structure.

Despite their widespread usage in various applied and methodological studies such as Heller et al. (2007); Zuo et al. (2011); Rietveld et al. (2014); Karmakar and Small (2020); Karmakar et al. (2021), standard single PCH tests are highly conservative even when the tests for the individual base hypotheses are non-conservative (Benjamini and Heller, 2008; Zhang et al., 2016; Liu et al., 2022). For example, under the global null (where all $H_{0,i}$ are true), even when the base p-values are uniformly distributed under the null, $p_{(r)}, ..., p_{(m)}$ are highly superuniform (i.e., $\mathbb{P}\left(p_{(i)} \leq t\right) \ll t$ for all $i = r, ..., m$). Thus, the resulting PCH p-value will also be superuniform. The only null case where standard single PCH tests will produce uniform p-values is when there are exactly $r - 1$ non-null base hypotheses each having *infinite* signal strength, which we refer to as the *least favorable null* (LFN) case as it is the parameter configuration in the null that maximizes the probability of rejecting $H_0^{r/m}$ (Benjamini and Heller, 2008; Dickhaus et al., 2021). Under this setting, $p_{(r)}, ..., p_{(m)}$ are guaranteed to correspond perfectly with the true null base p-values (since the $r - 1$ non-null base p-values will all be 0). Thus, as long as the null base p-values are uniformly distributed and the global null test used is not conservative, the resulting PCH test will produce uniformly distributed PCH p-values. However, the LFN case (and in particular the infinite signal size of the non-nulls) is generally unrealistic in real-world data settings, and any other null case constitutes a situation where standard methods would be conservative to some degree.

This conservativeness is concerning primarily because it extends to the alternative space. For example, in alternative configurations where the signal strengths associated with the non-null base hypotheses are low, standard single PCH tests can have especially low power since many of the $p_{(r)}, ..., p_{(m)}$ will likely correspond to null base p-values. Thus, the low power

of standard single PCH tests in applied settings is fundamentally linked to their conservativeness under the null. However, alternative configurations with low signals are especially prevalent in applications, where effects are often subtle, such as genetic epidemiology (Sesia et al., 2021; Wang et al., 2021). Therefore, a primary challenge in PCH testing research has been to develop methods that correct the conservativeness of standard single PCH tests under the null.

1.3.2 Adjusting for Conservativeness via Multiple Testing

Several works correct the conservativeness of standard single PCH tests by sharing information across *multiple* PCH tests to infer which PCH's are most likely to be under different null configurations. Naturally, this approach can only be applied when there are multiple (ideally very many) different PCH's being tested at once. Importantly, the following methods *cannot* be applied to testing a single PCH. Broadly, we can categorize these methods into two types: empirical Bayes and filtering methods.

Empirical Bayes (EB) methods such as Heller and Yekutieli (2014); Huang (2019); Dai et al. (2020); Dreyfuss et al. (2021); Liu et al. (2022) aim to predict the proportion of PCH's belonging to each null configuration, often adapting existing methods for estimating the proportion of nulls in the multiple testing literature (Efron et al., 2001; Storey, 2002; Storey et al., 2004; Jin and Cai, 2007; Efron, 2008). These approaches usually produce asymptotically valid PCH p-values under certain regularity conditions. However, because they rely on the consistency of their estimation method for their asymptotic validity guarantees, they can experience high Type I Error inflation when the number of hypotheses is small or when violations of regularity conditions make estimation unreliable.

Filtering methods filter out unpromising hypotheses to facilitate an analysis of the remaining ones, which tend to be less conservative (Dickhaus et al., 2021; Wang et al., 2021). For example, the AdaFilter method in Wang et al. (2021) uses a data-adaptive threshold based on Bonferroni's combining function to reduce the set of PCH's to the ones closest to the LFN case. Dickhaus et al. (2021) filters and re-scales PCH p-values based on a user-provided threshold such that any multiple testing procedure like Benjamini–Hochberg (BH) can be applied to the reduced set while controlling FDR on the entire set. These methods are highly effective in settings where there is a large number of PCH's being simultaneously tested and the global null is expected to be the overwhelmingly dominant null configuration. In these settings, most of the null PCH p-values will be highly conservative, thus allowing filtering to effectively exclude unpromising hypotheses. However, outside of this particular, albeit important, setting, the performance of these methods can suffer. For instance, we show that in situations where the proportion of global nulls is small relative to other null configurations, these filtering methods can have lower power than just using standard methods for single PCH testing to compute individual PCH p-values and then applying Benjamini–Hochberg; see Section 3.3.3 for details.

Overall, since multiple testing approaches to PCH testing must correct for both the multiplicity of the hypotheses being tested and the conservativeness of the individual p-values, they tend to be tailored to certain multiple testing settings (i.e., when the global null is the

primary null configuration) and can be less powerful outside of those settings. Alternatively, a generic way to generate PCH multiple testing procedures would be to develop a powerful and non-conservative *single* PCH test (as we do in this paper), and then the p-values from such a test could be fed into any existing (non-PCH-specific) multiple testing procedure. By leveraging the vast literature on multiple testing procedures, this generic PCH multiple testing procedure has the potential to be powerfully applied in almost any setting.

1.4 Our Contributions

In this paper, we propose the conditional PCH (cPCH) test, a new approach to correcting the conservativeness of a standard (single) PCH test by conditioning on a function of the data; notably, cPCH applies directly to an individual PCH test and does not need there to be multiple PCH tests at all. In the commonly encountered situation when the underlying test statistics associated with each of the m hypotheses within a PCH test are independent and Gaussian, we show that cPCH p-values are almost exactly uniform under *any* null configuration, with small and quantifiable deviations between the realized Type I error and the nominal level. We also prove that, analogous to standard single PCH tests, the cPCH test is non-conservative under the LFN case. We demonstrate via simulation that the cPCH test is empirically more powerful than standard single PCH tests, maintains Type I error control even under model misspecification, and, in combination with different multiple testing procedures, outperforms existing PCH multiple testing methods in certain regimes while maintaining FDR control.

1.5 Outline

Section 2 provides the motivation and formal definition of the cPCH test and states our key result: When the base test statistics associated with each of the m base hypotheses are independent and Gaussian with unit-variance, the cPCH test produces p-values that are almost exactly uniform under the null with minor and quantifiable deviations between the Type I error and nominal level. Section 2.5 describes an algorithm for efficiently computing cPCH p-values. Section 3 illustrates the performance of the cPCH test on various simulated datasets for both single and PCH multiple testing and Section 4 provides a real data example. Section 5 concludes.

1.6 Reproducibility

All code, along with a tutorial for its use, is provided at `https://github.com/biyonka/cpch`.

1.7 Notation

For a distribution $P_{\boldsymbol{\theta}}$ in a parametric model $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, let $\mathbb{P}_{\boldsymbol{\theta}}$ denote a probability taken with respect to $P_{\boldsymbol{\theta}}$. We write Φ and ϕ to denote the cumulative distribution function (CDF) and probability density function (PDF) of the standard normal random variable, respectively.

2 Conditional PCH Testing

2.1 Preliminaries

Although we will ultimately argue that our method applies more broadly, we will begin by assuming that each of the base null hypotheses $H_{0,i}$ tests whether a scalar parameter $\theta_i = 0$, and that the data for such a test can be summarized into a single unit-variance Gaussian test statistic T_i with mean θ_i, for $i = 1, ..., m$:

$$H_{0,i} : \theta_i = 0 \text{ vs. } H_{1,i} : \theta_i \neq 0, \qquad T_i \sim \mathcal{N}(\theta_i, 1). \tag{1}$$

As in the problem statement in Section 1.2, we assume the test statistics are independent across i. Thus, the PCH is $H_0^{r/m} : \boldsymbol{\theta} \in \Theta_0^{r/m}$, where $\Theta_0^{r/m} = \{\boldsymbol{\theta} \in \mathbb{R}^m : \|\boldsymbol{\theta}\|_0 < r\}$ is the partial conjunction null space and $\|\cdot\|_0$ is the ℓ_0 norm. Though it may seem constraining to assume a parametric form of both the null and non-null base test statistic distributions, this distributional assumption approximates a wide array of applications, as the data for the base hypotheses are commonly summarized as asymptotically normal parameter estimators such as maximum likelihood estimators, method of moments estimators, and most causal estimators for average treatment effects (Heller et al., 2007; Rietveld et al., 2014; Zhang et al., 2016; Barfield et al., 2017; Wang et al., 2021; Liu et al., 2022). The replicability analysis of microarray studies presented in Section 4 provides a real data example that adheres to the above setting according to Wang et al. (2021).

We introduce some key notation for this setting. Recall we originally ordered the base p-values $\{p_i\}_{i=1}^m$ as $p_{(1)} \leq \cdots \leq p_{(m)}$. Since we are in a two-sided testing setting, we will analogously order the T_i by their *magnitudes*, i.e., we order $\{T_{(i)}\}_{i=1}^m$ by $|T_{(1)}| \leq ...|T_{(m)}|$ and let $\boldsymbol{T}_{(i:j)} = (T_{(i)}, ..., T_{(j)})$ for $i \leq j$. Note, however, that the indices are ordered in reverse for the $T_{(i)}$ as they are for the $p_{(i)}$, e.g., the most significant and hence *smallest* p-value $p_{(1)}$ corresponds to the *largest*-magnitude test statistic $T_{(m)}$. From now on, we also present the combining functions of global null tests as functions of the base test statistics T_i instead of the base p-values p_i. In this new formulation, a PCH test with some test-statistic-combining function f will reject $H_0^{r/m}$ when $f(\boldsymbol{T}_{(1:m-r+1)}) \geq c_\alpha$ where c_α is the rejection threshold. For instance, the Bonferroni test would have $f_{\text{B}}(\boldsymbol{T}_{(1:m-r+1)}) = |T_{(m-r+1)}|$ and $c_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2(m-r+1)}\right)$.

2.2 Motivation and Intuition

Given a test statistic $f\left(\boldsymbol{T}_{(1:m-r+1)}\right)$, the primary challenge of PCH testing is to specify a rejection threshold c_α such that the test which rejects when $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) > c_\alpha$ is valid, in that it controls Type I error at the desired level α, and powerful, in that it rejects as often as possible when the PCH is false. In particular, our design objective is to generate a test that produces uniform PCH p-values $p^{r/m}$ under *any* null configuration, i.e., $\mathbb{P}_{\boldsymbol{\theta}}(p^{r/m} \leq \alpha) = \alpha$ for all $\boldsymbol{\theta} \in \Theta_0^{r/m}$, since conservativeness in the p-value distribution at some point in the null space will imply a loss of power at nearby points in the alternative space.

A generic way to generate new PCH tests and further our understanding of existing methods in comparison is to first consider an oracle test that has the desired properties (in our case, one that produces uniform p-values under every null) and admits a plug-in version with very similar statistical behavior. For illustrative purposes, we focus on the $r = m = 2$ case where the test statistic for all standard combining functions (from Fisher's, Simes', and Bonferroni's global null tests) reduces to $f(T_{(1)}) = |T_{(1)}|$. By the symmetry of the problem (i.e., the rejection threshold is invariant to relabelling of the indices of T_1 and T_2) it is sufficient to represent all $\boldsymbol{\theta} \in \Theta_0^{2/2}$ by $\theta_{(2)}$, where, for illustrative purposes in this section, we will assume that $\theta_{(2)} \geq 0$ (recall $\theta_{(1)} = 0$ by definition of $\Theta_0^{2/2}$). A natural choice of extra information to provide to our oracle is $\theta_{(2)}$. So, we define the rejection threshold of our PCH Oracle test $c_\alpha(\theta_{(2)})$ as the value satisfying

$$\mathbb{P}_{\theta_{(2)}}\left(\left|T_{(1)}\right| > c_\alpha\left(\theta_{(2)}\right)\right) = \alpha,$$

i.e., $c_\alpha(\theta_{(2)})$ is the $1-\alpha$ quantile of $|T_{(1)}|$'s distribution when $\{\theta_1, \theta_2\} = \{0, \theta_{(2)}\}$. By definition of $c_\alpha(\theta_{(2)})$, this oracle test produces uniform PCH p-values for *every* $\boldsymbol{\theta} \in \Theta_0^{2/2}$, as desired.

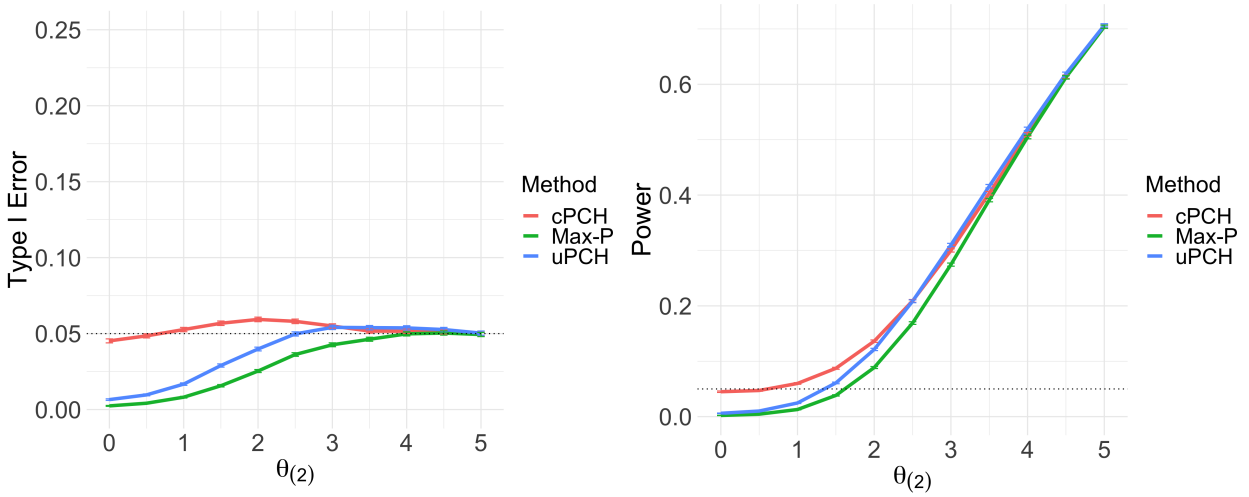
Since PCH tests do not have access to the true $\theta_{(2)}$ in practice, they must use a rejection threshold that does not require knowledge of the true $\theta_{(2)}$. One valid choice is $c_\alpha = \sup_{\theta_{(2)}} c_\alpha\left(\theta_{(2)}\right) = c_\alpha(\infty)$, which is the rejection threshold of the Max-P test: $c_\alpha(\infty) = \Phi^{-1}\left(1 - \alpha/2\right)$. Thus, we can think of the Max-P test as a *plug-in* version of the oracle test where $\hat{\theta}_{(2)} = \infty$ is a (worst-case) plug-in estimator of $\theta_{(2)}$.

At first glance, the choice of $\hat{\theta}_{(2)} = \arg\sup_{\theta_{(2)}} c_\alpha\left(\theta_{(2)}\right) = \infty$ seems to be the primary cause for the conservativeness of standard methods (recall all standard methods reduce to the Max-P test when $r = m = 2$). A promising possibility for resolving the conservativeness of the Max-P test is to choose a different estimator for $\theta_{(2)}$ that is likely to be closer to the true $\theta_{(2)}$. A natural choice would be $\hat{\theta}_{(2)} = T_{(2)}$, the maximum likelihood estimator (MLE) of $\theta_{(2)}$. This choice defines a new PCH test, which we call the *unconditional (plug-in) PCH (uPCH)* test, with rejection threshold $c_\alpha\left(T_{(2)}\right)$, the $1-\alpha$ quantile of $|T_{(1)}|$'s distribution when $\theta_{(2)} = T_{(2)}$. Though this choice of plug-in estimator does not come with any obvious Type I error guarantees, we find that the Type I error inflation of the uPCH test is remarkably small, as shown in Figure 1a. However, it is *still highly conservative* near the global null and, like the Max-P test, this conservativeness extends to the alternative space as well; see Figure 1b for details. These observations lead to our primary motivating insight for conditional PCH (cPCH) testing:

The conservativeness of existing methods is not primarily caused by the choice of the estimator for the unknown parameters, but rather the sensitivity of the test statistic's distribution to the unknown parameters.

In particular, if $|T_{(1)}|$'s distribution did not depend very much on $\theta_{(2)}$, the uPCH test, which uses a reasonable plug-in estimator for $\theta_{(2)}$, should have similar Type I error to the oracle test which uses $\theta_{(2)}$'s exact value and produces exactly uniform p-values under every null $\boldsymbol{\theta}$.

One way to make the distribution of a test statistic less sensitive to unknown parameters is by conditioning on a function of the data $h(\boldsymbol{T})$. For example, an extreme case would be



(a) Type I Error

(b) Power

Figure 1: Each point represents the proportion of Max-P, uPCH, and cPCH p-values for testing $H_0^{2/2}$ below α over 100,000 independent replicates of the data $T_i \sim \mathcal{N}(\theta_i, 1)$ where T_1 and T_2 are independent and $\{\theta_1, \theta_2\} = \{0, \theta_{(2)}\}$. Error bars represent 2 standard errors.

if the function of the data we conditioned on were sufficient for $\boldsymbol{\theta}$. Then, the conditional distribution of the test statistic would not depend on the unknown parameters at all.

Thus, we propose a new PCH test based on the distribution of the test statistic $|T_{(1)}|$ *conditional* on a function of the data. In the $r = m = 2$ case, a natural choice for conditioning is $h(\boldsymbol{T}) = T_{(2)}$, since it is all that remains after precluding conditioning on $T_{(1)}$ which uniquely determines the value of the test statistic $f(T_{(1)})$. So, we define the new conditional PCH (cPCH) Oracle test by the rejection threshold $c_\alpha(\theta_{(2)}, T_{(2)})$ satisfying

$$\mathbb{P}_{\theta_{(2)}}\left(|T_{(1)}| > c_\alpha\left(\theta_{(2)}, T_{(2)}\right) \mid T_{(2)}\right) = \alpha,$$

i.e., $c_\alpha\left(\theta_{(2)}, T_{(2)}\right)$ is the $1-\alpha$ quantile of the conditional distribution of $|T_{(1)}| \mid T_{(2)}$ when $\{\theta_1, \theta_2\} = \{0, \theta_{(2)}\}$. As with the PCH Oracle test defined above, the cPCH Oracle test produces exactly uniform p-values under every null $\boldsymbol{\theta}$ by construction. From the cPCH Oracle test, we now define a new plug-in test in the hopes that it will resolve the conservativeness of the approaches described above. We call this test the *cPCH test*, which uses the MLE $\hat{\theta}_{(2)} = T_{(2)}$ as an estimator for $\theta_{(2)}$ in the cPCH Oracle test's rejection cutoff.

In Figure 1a, we compare the Type I error of the Max-P, uPCH, and cPCH tests in the $r = m = 2$ case, and we see that the cPCH test has only small deviations between the realized Type I error and the nominal level α across the entire null space. Since the cPCH Oracle test produces uniform p-values under the null by construction, the nearly-α Type I error of the cPCH test indicates that it is matching its oracle under the null and is hence less sensitive to the estimation of $\theta_{(2)}$, thus achieving our original design objective. Importantly, in Figure 1b we see that under the alternative, the cPCH test is more powerful than the Max-P and uPCH tests in low signal settings (i.e., when θ_1 and θ_2 are both small) while all tests have similar power in high signal settings.

Additionally, like the uPCH test, the Type I error inflation of the cPCH test is remarkably small. We show in Section 2.4 that across various m and r, the maximum Type I error inflation of the cPCH test is quantifiable and small. Since it is possible to quantify the Type I error inflation of the cPCH test, one could always adjust the significance level so that the true Type I error is bounded by α. This adjustment is likely not necessary in practice, as we show that the cPCH test empirically controls the Type I error, even under model misspecification; see Section 3.2 for more details. Thus, via conditioning, we have developed a framework for PCH testing that is more robust to plug-in estimation and, as we show further in Section 3.1, is empirically more powerful than existing methods for single PCH testing.

2.3 Formal Definition

We now formally define the cPCH test for general $m \geq 2$. Analogously to the previous section, given some combining function $f : \mathbb{R}^{m-r+1} \to \mathbb{R}$, we will specify the rejection threshold of the cPCH test for $m \geq 2$ based on the quantiles of the conditional distribution of our test statistic $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid \boldsymbol{T}_{(m-r+2:m)}$ under $H_0^{r/m}$, where recall that $H_0^{r/m} : \boldsymbol{\theta} \in \Theta_0^{r/m} = \{\boldsymbol{\theta} \in \mathbb{R}^m : \|\boldsymbol{\theta}\|_0 < r\}$.We condition on $\boldsymbol{T}_{(m-r+2:m)}$ because, intuitively, we want to condition on as much information as possible to make the Type I error as insensitive as possible to the error in the estimation of $\boldsymbol{\theta}_{(m-r+2:m)}$. Hence, we define $c_\alpha(\boldsymbol{\theta}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)})$ as the $1-\alpha$ quantile of the distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid \boldsymbol{T}_{(m-r+2:m)}$ for $\boldsymbol{T} \sim \mathcal{N}(\boldsymbol{\theta}, I_m)$ where $\boldsymbol{\theta}_{(1:m-r+1)} = \boldsymbol{0}$, $\boldsymbol{\theta}_{(m-r+2:m)}$ comes from the first argument to c_α, and I_m is the $m \times m$ identity matrix. Note $c_\alpha\left(\boldsymbol{\theta}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)$ is permutation invariant to the elements of the underlying $\boldsymbol{\theta}$, i.e., the resulting quantile would be the same if $\boldsymbol{T} \sim \mathcal{N}(\sigma(\boldsymbol{\theta}), I_m)$ where $\sigma(\boldsymbol{\theta})$ is some permutation of the elements of $\boldsymbol{\theta}$. Therefore, it is sufficient to represent any $\boldsymbol{\theta} \in \Theta_0^{r/m}$ by $\boldsymbol{\theta}_{(m-r+2:m)}$.

Definition 2 (Conditional PCH (cPCH) test)**.** *The level-α cPCH test rejects $H_0^{r/m}$ when $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) > c_\alpha\left(\boldsymbol{T}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)$.*

In the definition above, $\boldsymbol{T}_{(m-r+2:m)}$ in the first argument of c_α serves as the estimator for $\boldsymbol{\theta}_{(m-r+2:m)}$, while the same quantity in the second argument denotes what is conditioned on. Analogous to our construction of the cPCH test when $r = m = 2$ in Section 2.2, $\boldsymbol{T}_{(m-r+2:m)}$ is the MLE for $\boldsymbol{\theta}_{(m-r+2:m)}$. The cPCH Oracle test has rejection threshold $c_\alpha\left(\boldsymbol{\theta}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)$ and the PCH Oracle test has rejection threshold $c_\alpha\left(\boldsymbol{\theta}_{(m-r+2:m)}\right)$, which we define as the $1-\alpha$ quantile of the distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right)$ where $\boldsymbol{T} \sim \mathcal{N}(\boldsymbol{\theta}, I_m)$ and $\boldsymbol{\theta}$ is comprised of $\boldsymbol{\theta}_{(m-r+2:m)}$ with the remaining elements $\boldsymbol{\theta}_{(1:m-r+1)} = \boldsymbol{0}$.

For all results in this paper, we focus on the combining functions of the Bonferroni, Simes, and Fisher global null tests. Note, however, that the definition of the cPCH test is general, as it allows the analyst to in principle specify any f of her choice. Our code allows the analyst to specify f and provides implementations of the cPCH test using Bonferroni's, Simes', and Fisher's combining functions as used in this paper.

2.4 Validity of the cPCH Test

2.4.1 Results under the LFN Case

We first discuss the behavior of the cPCH test under the LFN case, which we introduced in Section 1.3.1 as the setting where, informally, there are exactly $r-1$ θ_i's with $|\theta_i|$ equaling infinity. In this section, we formalize this initial description and prove a validity result for the cPCH test under the LFN case.

First, we provide some intuition for the LFN case. As discussed in Section 2.1, the data for each base hypothesis is commonly summarized as an asymptotically normal parameter estimator such as a sample mean or ordinary least squares coefficient. Under our assumed setting in Section 2.1, we expect the non-null T_i's under fixed alternatives to approach infinity in magnitude as their underlying sample size approaches infinity. For example, if each base hypothesis test reports a sample mean of n i.i.d. observations each approximated by a $\mathcal{N}(\mu_i, 1)$ distribution, then the standard z-test statistic is $T_i \sim \mathcal{N}(\theta_i, 1)$ where $\theta_i = \sqrt{n}\mu_i$. Therefore, as n approaches infinity, the non-null $|\theta_i|$'s will also approach infinity, and thus, so will their corresponding $|T_i|$'s.

When the base test statistics are independent and the null base p-values are uniformly distributed, the standard single PCH tests are exact tests under the LFN case.[1] We show that the cPCH test exhibits a similar property under the LFN case. First, we formalize our initial description of the LFN case by defining an *LFN sequence*:

Definition 3 (LFN sequence)**.** *An LFN sequence* $(\boldsymbol{\theta}^{(n)})$ *is a sequence in* $\Theta_0^{r/m}$ *such that* $\left|\boldsymbol{\theta}^{(n)}_{(j)}\right| \to \infty$ *for* $j = m-r+2, ..., m$ *as* $n \to \infty$.

Let $\boldsymbol{T}^{(n)}$ denote a test statistic vector, with the superscript (n) now allowing us to vary $\boldsymbol{T}^{(n)}$'s distribution. Let $\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \coloneqq \mathbb{1}\left\{f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) > c_\alpha\left(\boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right)\right\}$ and $\varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right) \coloneqq \mathbb{1}\left\{f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) > c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right)\right\}$ denote the decisions made by the cPCH test and PCH Oracle test, respectively.

Theorem 1 (Exactness of the cPCH test under the LFN case)**.** *Assume* $\left(\boldsymbol{\theta}^{(n)}\right)$ *is a LFN sequence and that the test statistic vector* $\boldsymbol{T}^{(n)} \sim \mathcal{N}(\boldsymbol{\theta}^{(n)}, I_m)$*. Assume* $f : \mathbb{R}^{m-r+1} \to \mathbb{R}$ *is permutation invariant, continuously differentiable, and has* $\boldsymbol{\nabla} f \neq 0$ *except on a set whose closure has measure zero. Then, for any* $\alpha \in (0, 1)$*,*

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{cPCH}\left(\boldsymbol{T}^{(n)}\right) = \varphi_\alpha^{PCHOrac}\left(\boldsymbol{T}^{(n)}\right)\right) = 1.$$

In particular, the above implies that the cPCH test's limiting Type I error under any LFN sequence is exactly α*.*

[1] This is true as stated for the Fisher's and Simes' single PCH tests. For Bonferroni's, we need the additional condition that $r = m$. Otherwise, Bonferroni's test is slightly conservative because we assume independence of the T_i while Bonferroni's test allows for arbitrary dependence.

Theorem 1 shows not only that the cPCH test achieves exactly the nominal Type I error rate in the LFN case, but that it in fact behaves *identically* to the PCH Oracle test in such a case. Note as well that Fisher, Bonferroni, and Simes combining functions satisfy the conditions on f specified in the theorem statement. The proof is provided in Appendix A.

2.4.2 Approximate Validity

As no real testing scenario falls exactly into the LFN case, the utility of a PCH test is characterized primarily by its behavior when the non-null θ_i's are finite. We find that, under the setting in Section 2.1, for a fixed m, r, and α, the cPCH test has small and quantifiable Type I error inflation. Additionally, we find that for a fixed m and r, the cPCH test produces nearly uniform p-values under the null. Thus, we call the cPCH test "approximately valid and non-conservative".

Our characterization of the cPCH test's validity for finite θ_i relies fundamentally on the fact that, for a given choice of m and r, we can estimate the cPCH p-value distribution for a given $\boldsymbol{\theta} \in \Theta_0^{r/m}$ with high accuracy via Monte Carlo sampling. Specifically, given a $\boldsymbol{\theta} \in \Theta_0^{r/m}$, we generate samples $\tilde{\boldsymbol{T}}^{(k)} \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\theta}, 1), k = 1, ..., N$ and compute the cPCH p-value for each sample to empirically estimate the distribution of the cPCH p-value under $\boldsymbol{\theta}$. Performing this sampling procedure across a sufficiently fine grid of $\Theta_0^{r/m}$ with many replicates N allows us to obtain highly accurate estimates of the p-value distribution under the null for the desired m and r. Given a fixed α, we can also use the Monte Carlo samples to obtain highly accurate estimates of the Type I error for each $\boldsymbol{\theta} \in \Theta_0^{r/m}$ by computing the proportion of the sampled cPCH p-values at $\boldsymbol{\theta} \in \Theta_0^{r/m}$ that are below α.

We pause to highlight that our approach in this section is somewhat different from a usual simulation study, which would typically explore a small (but hopefully at least somewhat representative) subset of all possible data-generating scenarios. In contrast, we provide an exhaustive characterization of the cPCH p-value distribution at effectively *every* null $\boldsymbol{\theta}$ for realistic values of m and r. Such an exhaustive search is made possible by the small dimension of the null space for most common PCH testing scenarios, such as causal mediation analysis, where $m = 2$, and replicability analysis, where m is often ≤ 4 (Bogomolov and Heller, 2013; Heller and Yekutieli, 2014; Wang et al., 2021), and hence we focus on settings with $m \leq 4$. For all analyses in this section, we fix $\alpha = 0.05$. Thus, the following results can be interpreted as essentially a computational proof that, for the configurations of m and r tested, the cPCH test produces approximately uniform p-values under the null and that, for $\alpha = 0.05$, the Type I error inflation of the cPCH test is small. We expect these results to generalize to larger m and different choices of α. Though our results assume the setting of Section 2.1 where each T_i are single, independent unit-variance Gaussians, our simulations in Section 3.2 suggest that the following results still hold under other distributional assumptions for the base test statistics.

We first quantify the closeness of the cPCH p-value distribution under the null to the Unif$[0, 1]$ distribution through various metrics of interest. To visually depict results, we focus on the $r = m = 2$ and $m = 3, r = 2$ cases since any $\boldsymbol{\theta} \in \Theta_0^{2/2} \cup \Theta_0^{2/3}$ can be represented by a single scalar $\theta_{(m)}$. Results for $r = m = 3$ are provided in Appendix C.2.3.

First, we compute the quantile-quantile plots between the empirical cPCH p-value density (estimated using the Monte Carlo sampling scheme described above) for various $\boldsymbol{\theta} \in \Theta_0^{r/m}$ and the Unif$[0, 1]$ density. As shown in Figure 2, the density of cPCH p-values closely matches that of a Unif$[0, 1]$ distribution over a fine grid of $\boldsymbol{\theta} \in \Theta_0^{r/m}$ for any choice of combining function (Bonferroni, Simes, or Fisher).

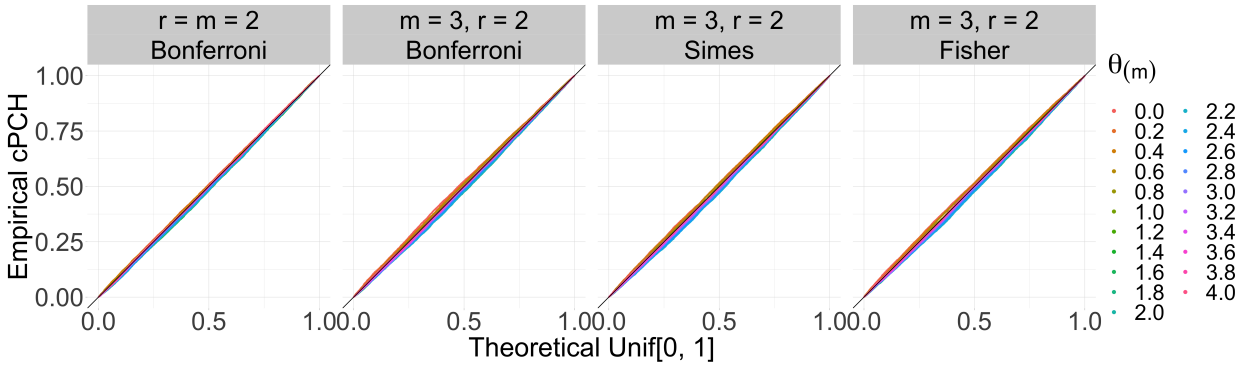


Figure 2: Quantile-quantile plot of the empirical cPCH p-value density under various $\boldsymbol{\theta} \in \Theta_0^{r/m}$ compared with the theoretical Unif[0, 1] distribution. Recall that when $r = m = 2$, Bonferroni's, Simes', and Fisher's combining functions are all equivalent. Each line represents the matched quantiles of the Unif[0, 1] density (x-coordinate) and the empirical cPCH p-value density for a given $\theta_{(m)}$ (y-coordinate), estimated using $n = 10,000$ independent replicates for the Monte Carlo sampling scheme described in Section 2.4.2.

Next, we quantify the magnitude of the deviations between the null cPCH p-value distributions and the Unif$[0, 1]$ distribution by computing the Kolmogorov–Smirnov (K–S) distances between the empirical CDF of the cPCH p-values for various $\boldsymbol{\theta} \in \Theta_0^{r/m}$ (estimated using the Monte Carlo sampling scheme described above) and the CDF of the Unif$[0, 1]$ distribution. As shown in Figure 3, we see that the K–S distances between the cPCH null p-value distributions and the Unif$[0, 1]$ distribution are small, as expected based on the results in Figure 2. The maximum K–S distance for any choice of combining function, m, and r occurs when $\theta_{(m)}$ is approximately between 1 and 3, with smaller deviations occurring under the global null, i.e., $\theta_{(m)} = 0$. Comparing Figure 3 for $r = m = 2$ with the corresponding Type I error plot for $r = m = 2$ (Figure 1a), we see that the deviation at the global null occurs because the cPCH test is slightly conservative under the global null, though far less so than its classical counterpart, the Max-P test. The deviations when $\theta_{(m)}$ is approximately between 1 and 3 occur due to the slight Type I error inflation of the cPCH test in this region. Note that the $m = 3, r = 2$ setting exhibits the same pattern of deviations as in the $r = m = 2$ setting, and we expect a similar pattern (very slight conservativeness near the global null and slight anticonservativeness when the non-null θ_i's are between 1 and 3) to generalize to other m and r.

We also provide an approach to estimating the maximum Type I error inflation of the cPCH test for any m, r, and α via stochastic gradient descent (SGD) and show that the estimated maximum Type I error inflation of the cPCH test is generally small for various $2 \leq m \leq 4$ and $2 \leq r \leq m$ when $\alpha = 0.05$. To apply SGD to the Type I error of the cPCH test, we

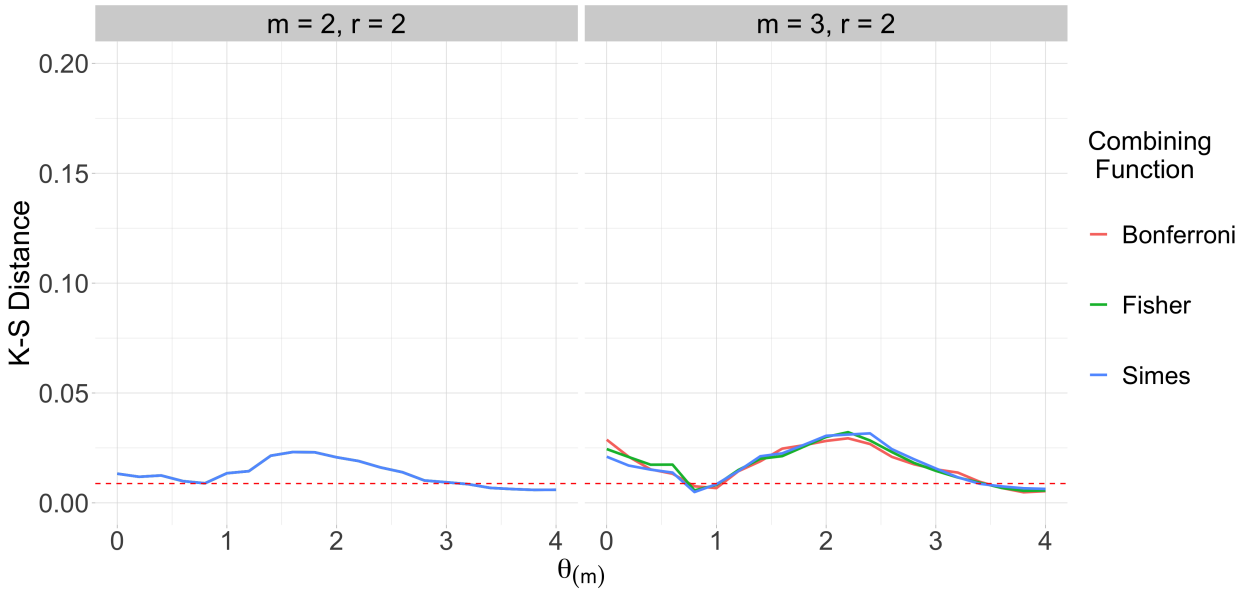


Figure 3: The K–S distances between the empirically estimated CDF of cPCH p-values and the Unif[0, 1] CDF for various $\boldsymbol{\theta} \in \Theta_0^{r/m}$. Each point represents the K–S distance between the Unif[0, 1] CDF and the empirically estimated cPCH p-value CDF for a given $\theta_{(m)}$ using $n = 10,000$ independent replicates for the Monte Carlo sampling scheme described in Section 2.4.2. To provide a baseline for comparison, the red dotted line represents the average K–S distance between the empirical CDF of the Unif[0, 1] distribution estimated over 10,000 independent replicates and the theoretical Unif[0, 1] CDF.

write the Type I error as a function of $\boldsymbol{\theta}$:

$$E(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) > c_\alpha\left(\boldsymbol{T}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)\right), \ \boldsymbol{\theta} \in \Theta_0^{r/m}. \tag{2}$$

In the above notation, we suppress the Type I error's dependence on α and the combining function f for ease of presentation. It is of interest to quantify $\max_{\boldsymbol{\theta}\in\Theta_0^{r/m}} E(\boldsymbol{\theta})$ relative to α. Though $E(\boldsymbol{\theta})$ does not admit an analytical form, we can estimate the gradient of $E(\boldsymbol{\theta})$ empirically via the Monte Carlo sampling described above, then use SGD to estimate the maximum Type I error as $-\min_{\boldsymbol{\theta}\in\Theta_0^{r/m}} -E(\boldsymbol{\theta})$. By the symmetry of the testing problem, we can conduct the SGD algorithm to search over the subset of $\Theta_0^{r/m}$ where the non-zero means are all positive.

As suggested by Figure 1a, $-E(\boldsymbol{\theta})$ is not a convex function of $\boldsymbol{\theta}$. To justify our use of SGD in this non-convex setting, we first note that when $r = m = 2$, $-E(\boldsymbol{\theta})$ appears to be smooth and quasi–convex. For continuously differentiable, quasi-convex functions, gradient descent will converge to a stationary point (Kiwiel and Murty, 1996). Since the Type I error curve appears to have a single stationary point occurring at the global minimum (where $\theta_{(2)} \approx 2$), if the SGD algorithm converges to a finite solution, that solution is likely close to the global minimum (Patel and Zhang, 2021). For $r = m = 2$, we find that across multiple, independent initializations, the SGD algorithm consistently converges to solutions where $\theta_{(2)} \approx 2$, which

is consistent with the actual location of the maximum Type I error of the cPCH test for $r = m = 2$.

As discussed with the K–S distance results, we see evidence that the behavior of the cPCH test for $r = m = 2$ generalizes to different m and r (Figure 3). Therefore, we expect that, for any m and r, the Type I error will also be quasi-convex and the maximum Type I error will occur when $\theta_{(j)} \approx 2$ for all $j = m - r + 2, ..., m$. In fact, for $m = 3, 4$, we found that the SGD algorithm consistently converges to solutions where $\theta_{(j)} \approx 2$ for all $j = m - r + 2, ..., m$, thus supporting our claim. See Appendix B.2 for further details on the SGD algorithm.

The maximum Type I error inflation estimated via SGD is overall small for various choices of $2 \leq m \leq 4$ and $2 \leq r \leq m$ at $\alpha = 0.05$; see Table 1 for details. If desired, the analyst could also use the SGD algorithm to determine how much to lower the α used by the cPCH test to achieve their desired nominal α Type I error.

		cPCH-Bonferroni	cPCH-Fisher	cPCH-Simes
m = 2	r = 2	0.060	—	—
m = 3	r = 2	0.057	0.058	0.059
	r = 3	0.053	0.053	0.053
m = 4	r = 2	0.058	0.057	0.059
	r = 3	0.057	0.058	0.060
	r = 4	0.062	0.062	0.062

Table 1: Maximum Type I error of the cPCH test computed from the SGD algorithm detailed in Section B.2. All standard errors are below 0.0015 where standard errors are calculated over many independent replicates of the cPCH p-value ($N = 10,000$) at the $\boldsymbol{\theta}^* \in \Theta_0^{r/m}$ for which the SGD algorithm terminated.

2.5 Computing cPCH p-values

Given the observed $f(\boldsymbol{T}_{(1:m-r+1)}) = f_{\text{obs}}$, the cPCH p-value can be written as:

$$\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \mid \boldsymbol{T}_{(m-r+2:m)}\right)$$

where $\hat{\boldsymbol{\theta}}$ is comprised of $m-r+1$ zeroes and the elements of $\boldsymbol{T}_{(m-r+2:m)}$. Thus, $\hat{\boldsymbol{\theta}}$ is a shorthand for the MLE of $\boldsymbol{\theta}_{(m-r+2:m)}$ as in Definition 2. Because the value of $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \,\middle|\, \boldsymbol{T}_{(m-r+2:m)}\right)$ is invariant to re-indexing of the elements of $\hat{\boldsymbol{\theta}}$, throughout this paper, we default to $\hat{\boldsymbol{\theta}} = \left(0, ..., 0, T_{(m-r+2)}, ..., T_{(m)}\right)$.

Computing this probability exactly requires us to derive the density of $\boldsymbol{T}_{(1:m-r+1)} \mid \boldsymbol{T}_{(m-r+2:m)}$ where $\boldsymbol{T}_{(1:m-r+1)}, \boldsymbol{T}_{(m-r+2:m)}$ are order statistics of independent but non-identically distributed (i.n.i.d.) random variables. Though the conditional density functions of order statistics of i.n.i.d. random variables has been studied in previous works (Bapat and Beg, 1989; Ozbey et al., 2019), calculating the conditional densities based on the techniques in such

works would be computationally prohibitive as it would involve enumerating all $m!$ possible permutations of the order statistics. Alternatively, if we can generate samples from $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid \boldsymbol{T}_{(m-r+2:m)}$, we can estimate this conditional distribution by taking many independent samples and computing the empirical distribution. Thus, even if the analytic form of the conditional density is intractable, we can compute cPCH p-values with high accuracy as long as we can efficiently sample from the conditional density.

In the following, we describe our strategy for the above sampling problem. First, we note that when $r = m = 2$, we can derive the analytic form of the conditional density $T_{(1)} \mid T_{(2)}$, thus allowing us to obtain exact cPCH p-values without sampling; see Appendix B.1.1 for details. When $m > 2$, we develop a new, efficient procedure for sampling from $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid \boldsymbol{T}_{(m-r+2:m)}$. On a high level, our approach involves conditioning on extra events about which of the $T_1, ..., T_m$ correspond to the order statistics $\boldsymbol{T}_{(1:m-r+1)}$ and $\boldsymbol{T}_{(m-r+2:m)}$. By doing so, we can express the conditional density of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid \boldsymbol{T}_{(m-r+2:m)}$ as a mixture distribution such that

- the involved mixture weights can be computed analytically;
- the mixture components can be estimated by sampling from rather simple probability distributions.

To formally describe how we condition on these extra events, we denote by S the set of all possible ways to observe some unordered set of the T_i's corresponding to the entries of $\boldsymbol{T}_{(1:m-r+1)}$ and some ordered set of the remaining T_j's corresponding to the entries of $\boldsymbol{T}_{(m-r+2:m)}$. For example, when $m = 3, r = 2$, we have

$$S = \{(\{T_1, T_2\}, T_3), (\{T_2, T_3\}, T_1), (\{T_1, T_3\}, T_2)\}$$

where the inner set corresponds to $\{T_{(1)}, T_{(2)}\}$ and the remaining term corresponds to $T_{(3)}$. Here we pause to highlight a fact: the cardinality of S is $\frac{m!}{(m-r+1)!}$, which also equals the number of mixture components. Although in general the cardinality of S grows exponentially in m, when r is small, such as when $r = 2$ for replicability analysis, the complexity of $\frac{m!}{(m-r+1)!}$ is a linear or a low-order polynomial in m, e.g., $\frac{m!}{(m-2+1)!} = O(m)$. Additionally, m is often ≤ 4 in many common PCH testing scenarios (Bogomolov and Heller, 2013; Heller and Yekutieli, 2014; Wang et al., 2021), so the number of mixture components to compute tends to be reasonably small. For each term S_ℓ from the set S, for $\ell \in \left\{1, \ldots, \frac{m!}{(m-r+1)!}\right\}$, we define the event

$$B_\ell := \left\{S_\ell = (\{T_{(1)}, ..., T_{(m-r+1)}\}, \boldsymbol{T}_{(m-r+2:m)})\right\}.$$

For example, in the $m = 3, r = 2$ case above, $B_1 = \left\{(\{T_1, T_2\}, T_3) = (\{T_{(1)}, T_{(2)}\}, T_{(3)})\right\}$. Now, we express the conditional distribution of $f(\boldsymbol{T}_{(1:m-r+1)}) \mid \boldsymbol{T}_{(m-r+2:m)}$ in the following mixture form:

$$\begin{aligned}
&\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \;\middle|\; \boldsymbol{T}_{(m-r+2:m)}\right) \\
&= \sum_{\ell=1}^{\frac{m!}{(m-r+1)!}} \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell \mid \boldsymbol{T}_{(m-r+2:m)}\right) \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \;\middle|\; B_\ell, \boldsymbol{T}_{(m-r+2:m)}\right).
\end{aligned}$$

The computational details of our strategy can be summarized in two steps:

1. We derive the analytic form of the mixture weights $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell \mid \boldsymbol{T}_{(m-r+2:m)}\right)$, which can be expressed using only evaluations of the standard normal cumulative distribution Φ and density ϕ;
2. We develop a method for sampling from the distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid B_\ell, \boldsymbol{T}_{(m-r+2:m)}$ which relies solely on sampling from truncated-Normal distributions.

Derivations and further details for the above two steps can be found in Appendix B.1. Given N independent copies $\left\{X_\ell^{(k)}\right\}_{k=1}^N$ from the conditional distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid B_\ell, \boldsymbol{T}_{(m-r+2:m)}$ for $\boldsymbol{T} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, I_m)$, we estimate $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \mid B_\ell, \boldsymbol{T}_{(m-r+2:m)}\right)$ as:

$$g\left(f_{\text{obs}}, \left\{X_\ell^{(k)}\right\}_{k=1}^N\right) := \frac{1}{N+1}\left(1 + \sum_{k=1}^N \mathbb{1}\left\{X_\ell^{(k)} \geq f_{\text{obs}}\right\}\right),$$

where the "+1" in the numerator and denominator above is standard for Monte Carlo p-values; see, e.g., Ernst (2004) for details. In the above notation, we suppress $g\left(f_{\text{obs}}, \left\{X_\ell^{(k)}\right\}_{k=1}^N\right)$'s dependence on $\hat{\boldsymbol{\theta}}$ for ease of presentation. Taking N large, we can estimate $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \mid B_\ell, \boldsymbol{T}_{(m-r+2:m)}\right)$ with high accuracy. Therefore, we compute the cPCH p-value as:

$$p_{\text{cPCH}}^{r/m}(\boldsymbol{T}) := \sum_{\ell=1}^{\frac{m!}{(m-r+1)!}} \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell \mid \boldsymbol{T}_{(m-r+2:m)}\right) g\left(f_{\text{obs}}, \left\{X_\ell^{(k)}\right\}_{k=1}^N\right).$$

Note that the smallest $p_{\text{cPCH}}^{r/m}(\boldsymbol{T})$ can be is $\frac{1}{N+1}$. In multiple testing settings where we have M cPCH p-values, many FDR controlling procedures compare p-values to thresholds on the order of α/M. Therefore, in multiple testing settings, N must be set large enough to ensure that cPCH p-values can attain values below these thresholds to make any discoveries. Thus, computing $g\left(f_{\text{obs}}, \left\{X_\ell^{(k)}\right\}_{k=1}^N\right)$ is the main computational component of calculating cPCH p-values; however, since it is possible to sample efficiently from a truncated Normal distribution, it is computationally feasible to set N very large. See Figure 13 of Appendix C.2.2 for further details on computation time.

Importantly, our sampling scheme does not rely on any specific properties of the normal distribution; the only assumption necessary is that the base test statistics are independent. Though we have generally assumed that the T_i are normally distributed, we can compute cPCH p-values assuming the base test statistics are distributed under *any* one-parameter location family, such as a t-distribution with fixed degrees of freedom. Our code contains implementations of the cPCH for both normal and t-distributed base test statistics.

3 Simulations

3.1 Single PCH Testing

In this section, we empirically evaluate the power and Type I error of the cPCH test in comparison with existing approaches (Bonferroni's, Simes', and Fisher's tests) and the cPCH Oracle test presented in Section 2.2. Recall, we define $r^\star$ as the true number of non-null hypotheses. We generate data by sampling $\boldsymbol{T}$ independently from the model:

$$T_h \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1),\ h = 1, ..., r^\star, \qquad T_l \overset{i.i.d.}{\sim} \mathcal{N}(0, 1),\ l = r^\star + 1, ..., m.$$

We generate p-values for testing $H_0^{r/m}$ using the cPCH test ($N = 10,000$), the cPCH Oracle test ($N = 10,000$), and the standard PCH tests (Fisher, Simes, and Bonferroni). As shown in Figure 4, for $m = 4$, the cPCH test has nearly identical power to the cPCH Oracle test and is uniformly more powerful than the standard single PCH tests. Figure 8 in Appendix C.2.1, the analogous plot for null values of $r^\star$, shows that the cPCH test controls Type I error for all null configurations and is far less conservative under the null than the standard single PCH tests. We perform this simulation for various m and find similar results; see Appendix C.2.2 for further details. Additionally, we find that the computation time of the cPCH test is reasonable. For $N = 10,000$, the average computation time of a single cPCH p-value is < 30 milliseconds for any $2 \leq m \leq 4$ and $2 \leq r \leq m$; for $r = m = 2$, the computation time is ≈ 1 millisecond since we are able to compute the cPCH p-value in this case without sampling (and hence, there is no N). See Figure 13 in Appendix C.2.2 for further details on computation time.

3.2 Robustness to Model Misspecification

As the cPCH test assumes a known model for the base test statistics, it is important to evaluate its robustness to model misspecification. Since we can compute cPCH p-values for $\boldsymbol{T}$'s distributed under any one-parameter location family, we can assess the performance of the cPCH test (which assumes $\boldsymbol{T}$ is normally distributed) when $\boldsymbol{T}$ is generated from a different one-parameter location family, such as the t-distribution with a fixed scale and degrees of freedom. Specifically, let $\mathrm{t}(\theta, 1, \nu)$ be the generalized t-distribution with centrality parameter θ, scale 1, and degrees-of-freedom (DOF) ν. Let Ψ be the CDF of the $\mathrm{t}(0, 1, \nu)$ distribution. We generate data by sampling $\boldsymbol{T}$ independently from the model:

$$T_h \overset{i.i.d.}{\sim} \mathrm{t}(\theta, 1, \nu),\ h = 1, ..., r^\star, \qquad T_l \overset{i.i.d.}{\sim} \mathrm{t}(0, 1, \nu),\ l = r^\star + 1, ..., m.$$

The (correctly specified) cPCH p-value ($N = 10,000$) for testing $H_0^{r/m}$ is computed from $\boldsymbol{T}$ where we use a t-distribution in the computation procedure described in Section 2.5. We then convert each base test statistic into its corresponding base p-value $p_i = 2(1 - \Psi(|T_i|), i = 1, ..., m$ and produce the test statistics $W_i = \Phi^{-1}(1 - \frac{p_i}{2})$, from which we compute the (misspecified) cPCH p-value ($N = 10,000$) using a normal distribution in the computation procedure. This simulation emulates the situation in which the analyst only views the base p-values $p_1, ..., p_m$ and incorrectly assumes that the underlying base test statistics follow a normal distribution.

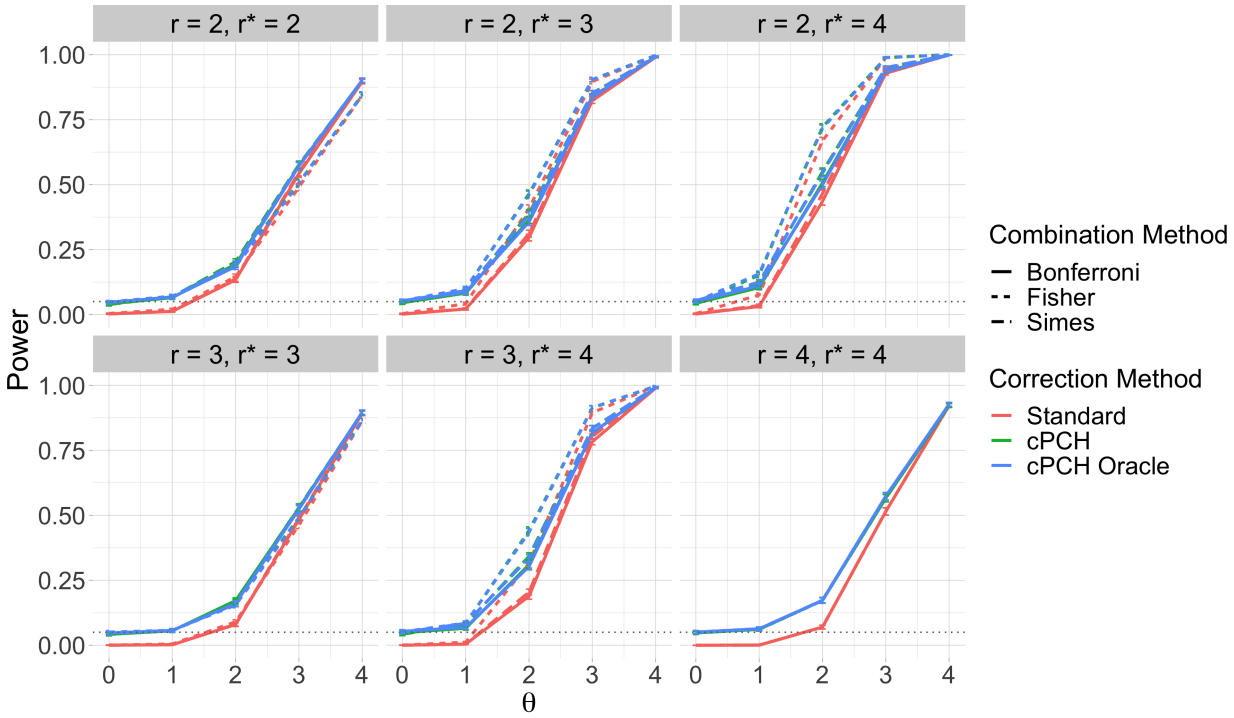


Figure 4: Power of the cPCH test, cPCH Oracle, and standard single PCH tests across all alternative cases ($r^\star \geq r$) for testing $H_0^{r/4}$ at nominal level $\alpha = 0.05$ (dotted grey line) for $m = 4$. Each point represents the proportion of cPCH p-values below α over 5000 independent replicates of the data generating procedure described in Section 3.1 for a given $(r^\star, r, \theta)$. Error bars depict ± 2 standard errors.

Figure 5 shows that the cPCH test remains powerful and maintains Type I error control for most null configurations under model misspecification, even in some settings where the base test statistics have Cauchy distribution ($\nu = 1$). Notably, the cPCH test under model misspecification maintains Type I error control for all null configurations when $\nu \geq 5$. Additionally, the cPCH test which is properly specified (i.e., the p-values are computed using the t-distribution) controls Type I error for all $\nu \geq 5$. In most applied settings, the degrees of freedom of a t-distributed base test statistic reflect the underlying sample size used to compute the base test statistic. Thus, excluding settings where the sample size is extremely small such that $\nu < 5$, we expect the cPCH test to be robust to model misspecification.

3.3 PCH Multiple Testing

Much of the existing literature on PCH testing addresses the conservativeness of standard PCH testing by sharing information across multiple PCH tests. Although the main focus of this paper is on resolving the conservativeness of standard PCH testing methods to improve the power for testing a *single* PCH, we include simulations here to show how cPCH testing can be used for PCH multiple testing and, in some cases, improves on existing PCH multiple testing procedures.

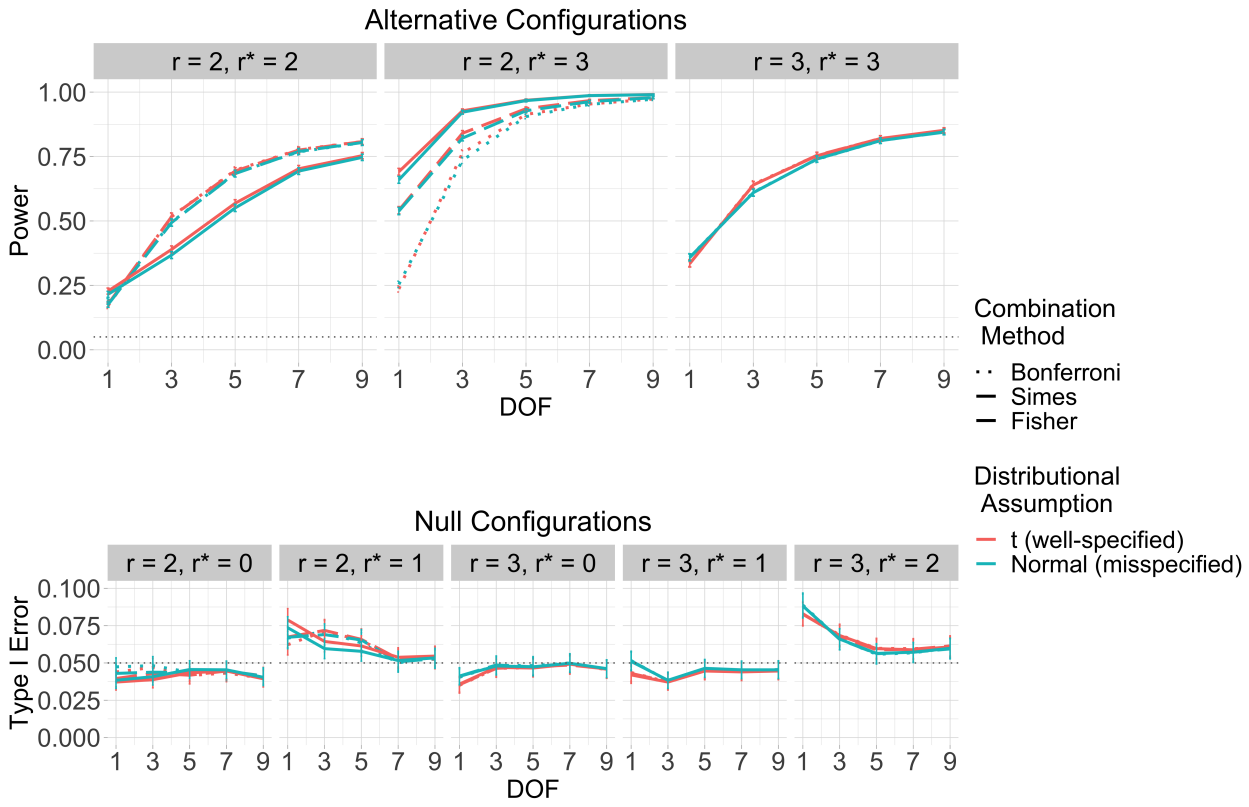


Figure 5: Power and Type I error of the properly specified (t-distributional assumption on $\boldsymbol{T}$) and the misspecified (normal distributional assumption on $\boldsymbol{T}$) cPCH tests at level $\alpha = 0.05$ (dotted grey line). Each point represents the proportion of cPCH p-values below α over 5000 independent replicates of the data generating procedure described in Section 3.2 for a given $(r^\star, r, \nu)$ with $m = 3$ and $\theta = 4$. Error bars depict ± 2 standard errors.

3.3.1 An Overview of Methods Under Comparison

In this section, we compare the performance of the cPCH test with various state-of-the-art approaches for PCH multiple testing. From now on, we assume there are M PCH's being simultaneously tested. We denote T_{ij} and p_{ij} as the ith base test statistic and p-value, respectively, for the jth PCH being tested, $i = 1, ..., m$, $j = 1, ..., M$. We focus on three FDR controlling procedures: Benjamini–Hochberg (BH) (Benjamini and Hochberg, 1995), Storey's (Storey et al., 2004), and AdaPT–GMM (Chao and Fithian, 2021), a covariate-assisted multiple testing procedure. Notably, cPCH used in combination with AdaPT–GMM provides an approach for PCH multiple testing with covariates, a setting largely unaddressed by the PCH multiple testing literature.

As discussed in Section 1.3.2, approaches designed for PCH multiple testing (in that they cannot be applied to the single PCH testing setting) can generally be grouped into two categories: empirical Bayes (EB) methods, which estimate the proportion of PCH's belonging to each null configuration, and filtering methods, which filter out unpromising PCH's and

analyze the remaining subset. Our analysis includes two state-of-the-art EB methods, the Divide-Aggregate Composite-null Test (DACT) (Liu et al., 2022) and High Dimensional Multiple Testing (HDMT) (Dai et al., 2020) and two state-of-the-art filtering methods, AdaFilter (Wang et al., 2021) and the method presented in Dickhaus et al. (2021), which we refer to by the authors' initials *"DHH"*.

Both DACT and HDMT adapt existing empirical Bayes frameworks for estimating the proportion of each null configuration (Efron et al., 2001; Storey, 2002; Storey et al., 2004). The primary difference between DACT and HDMT (and EB methods in general) lies in how the estimated proportions are used. DACT generates individual PCH p-values that are weighted sums of the estimated proportions of each null configuration, on which FDR controlling procedures like BH can be applied. HDMT estimates the FDR of the Max-P test at a given rejection threshold using these proportions, then selects the largest rejection threshold such that the estimated FDR is less than or equal to the desired level. Importantly, both HDMT and DACT are designed for causal mediation analysis, a specific case of PCH testing with $r = m = 2$, and thus, are only applicable (as implemented) when $m = 2$. However, the statistical frameworks used by EB methods for estimating the proportions of the null configurations are not limited to the $m = 2$ case and thus, these methods could reasonably be extended to larger m. As the computational cost of EB methods can grow exponentially in m (Wang et al., 2021), fixing $r = m = 2$ allows EB methods to produce accurate estimates of the true proportions within reasonable computational limits, as long as M is sufficiently large and the conditions necessary for the consistency of their estimators are met.

AdaFilter is a filtering method that infers a new (ideally less conservative) rejection threshold in a data-adaptive manner. It first reduces the set of total PCH's to the subset that would be rejected by Bonferroni PCH tests for $H_0^{(r-1)/m}$, then applies an adjusted version of the Bonferroni PCH test for $H_0^{r/m}$ to each PCH in the reduced set. Intuitively, this filtering is effective because the null PCH's that are rejected for $H_0^{(r-1)/m}$ are close to the LFN, so the Bonferroni PCH test will be less conservative on the reduced set.

DHH filters out all PCH p-values $p^{r/m}$ above some threshold τ, then applies a transformation to the smaller subset so that, under mild conditions on the p_{ij} and $p^{r/m}$, various FDR controlling procedures such as BH and Storey's procedure can be applied to the smaller subset to control FDR on the entire set. Thus, there are many variations of DHH depending on the PCH p-value and multiple testing procedure used. In our analysis, we use PCH p-values from Bonferroni's, Simes', and Fisher's tests, which are shown in Dickhaus et al. (2021) to satisfy the conditions necessary for the DHH procedure (with fixed τ) to have FDR control when using BH or Storey's procedure.

As with the cPCH test, we use the FDR controlling procedures BH, Storey's, and AdaPT–GMM (when applicable) to the PCH p-values produced from DACT and DHH. We also include the standard Bonferroni's, Simes', and Fisher's tests for single PCH testing in combination with BH, Storey's procedure, and AdaPT–GMM (when applicable) as benchmarks for comparison. For all the following simulations, we generate data such that the p_{ij} are independent. Table 2 provides a brief overview of the validity guarantees of all methods under consideration with further details in Appendix C.1.

Table 2

Method	Validity Guarantee
Empirical Bayes	
DACT	Asymptotic FDR control*, only considers $m = 2$
HDMT	Asymptotic FDR control†, only considers $m = 2$
Filtering	
AdaFilter	Asymptotic FDR control, upper bound on finite sample FDR
DHH	Finite sample FDR control
Single PCH‡	
cPCH	Approximate finite sample Type I error control⋆
Standard	Finite sample Type I error control

*: Requires that the regularity conditions of Jin and Cai (2007) hold.
†: Requires estimated null proportions (Storey, 2002; Storey et al., 2004) are consistent.
‡: With multiple testing procedures BH, Storey's Procedure, and AdaPT–GMM.
⋆: As specified in Section 2.4.

3.3.2 PCH Multiple Testing for $r = m = 2$

We first assess the FDR and power of all methods under consideration when $r = m = 2$. We begin with $r = m = 2$ because it is the simplest setting for PCH testing and the one that EB methods are designed for, allowing us to assess all the methods under consideration. We explore PCH multiple testing when $m \geq 3$ in the following subsection.

In this subsection, we consider two PCH multiple testing settings: the "classic" setting, where the methods only have access to the base test statistics T_{ij}, and the covariate-assisted setting, where the methods have access to a covariate X_j along with T_{ij}. PCH multiple testing with covariates arises in several important settings such as fMRI analysis, where fMRI scans are often taken across multiple subjects to evaluate whether certain brain regions were replicably activated across subjects (Heller et al., 2007). In such settings, it can be desirable to include covariate information like the (x, y, z) coordinates of the voxel locations, which are often highly informative of brain activation to certain stimuli. Since the cPCH test, the Max-P test, DACT, and DHH produce individual PCH p-values (though DACT and DHH rely on being in a PCH multiple testing setting to do so), they can be combined with covariate-assisted multiple testing procedures like AdaPT–GMM to leverage covariate information. It is not clear how to leverage covariate information in AdaFilter and HDMT.

We generate the data (X_j, T_{ij}), $i = 1, 2$, $j = 1, ..., 10,000$ by sampling independently from

the following model:

$$
\begin{aligned}
X_j &\sim \text{Unif}[0,1] \\
\gamma_{ij}|X_j &= \text{Bern}(\pi_1(X_j)) \text{ where } \pi_1(X_j) = \begin{cases} 1 & \text{if } X_j \geq 0.95 \\ 0.1 & \text{o/w} \end{cases} \\
T_{ij}|\gamma_{ij}, X_j &\sim \mathcal{N}(\theta\gamma_{ij}, 1)
\end{aligned}
$$

Here, X_j represents a (univariate) covariate that, when large, is highly informative that the PCH is non-null. We generate rejections using all methods under consideration: AdaFilter, DHH, HDMT, DACT, Max-P, and the cPCH test ($N = 10,000$). For the cPCH test, Max-P, DHH, and DACT, we combine the individual PCH p-values produced by each method with BH and Storey's procedure (to assess their performance without access to covariate information), and with AdaPT–GMM. All procedures are applied at nominal FDR level $q = 0.1$.

3.3.2.1 Without Covariates

We first assess the FDR and power of all methods under consideration without the use of covariate information, as shown in Figure 6. We ignore the solid lines for now, which correspond to methods that leverage covariate information via AdaPT–GMM. We defer the discussion of these methods to the following subsubsection.

In the standard no-covariate PCH multiple testing setting, DACT with BH and Storey's procedure is the most powerful but does not control FDR at the desired level. Among the methods that have FDR control, HDMT is the most powerful, outperforming both filtering approaches and the single PCH tests with BH and Storey's procedure. Through additional simulations, we find that EB methods tend to be more powerful than filtering and single PCH test-based approaches, but are more susceptible to FDR violations; see the Appendix C.2.6. These violations can occur even when M is large, which should be more favorable for EB methods to control the FDR. In particular, we find various data generating settings where HDMT does not control FDR; see Tables 6–7 in Appendix C.2.6. Additionally, we find that DACT tends to be more powerful than HDMT in settings where it does control FDR but can experience large FDR inflation in certain settings; see Table 6 in Appendix C.2.6. To provide further intuition for these results, we note that EB methods can gain power not only by resolving the conservativeness of standard single PCH tests but also potentially through learning about the distribution of the parameters. In this way, we expect EB methods to be more powerful than frequentist approaches to PCH testing (and this is reflected in our simulation), even if there was no underlying conservativeness in standard single PCH testing. However, this use of the learned parameter distribution is likely also the source of FDR violations for EB methods. Additionally, we note that the motivating application area for DACT and HDMT is causal mediation analysis, specifically for testing the effect of an exposure on a clinical outcome mediated through DNA methylations. Based on the simulations and real data analysis of Liu et al. (2022), data for this setting tend to have a very large proportion of nulls relative to alternative PCH's.

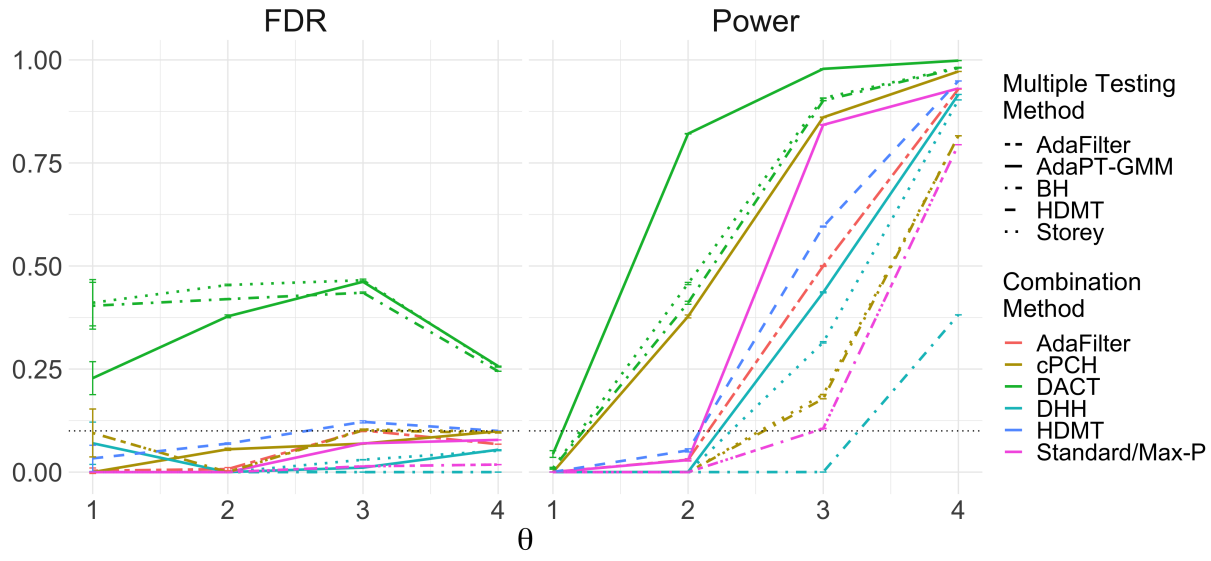


Figure 6: FDR and power of various PCH multiple testing methods at nominal FDR level $q = 0.1$ (dotted black line) for testing $H_0^{2/2}$. Each point represents 100 independent replicates of the data generating procedure described in Section 3.3.2 for a given θ, where power is estimated as the average proportion of non-null PCH's that are rejected and FDR is estimated as the average proportion of total rejections that are null. Error bars represent two standard errors.

Therefore, the FDR violations found in our simulation studies may not manifest in data that more closely emulates their desired application area.

While the filtering approaches, AdaFilter and DHH with BH and Storey's procedure, are less powerful than EB methods, they are still more powerful than the single PCH tests with BH and Storey's procedure. Generally, we find that filtering methods have relatively high power along with robust FDR control across a wide array of data generating settings, which we explore further in Section 3.3.3 and Section 4. While DHH guarantees finite-sample FDR control, we find in Section 3.3.3 that even AdaFilter, which only guaranteed FDR control at $qc(M)$ for $c(M) \approx \log(M)$, empirically controls FDR at level q across a wide range of settings. Additionally, the data generating procedure in this subsection is not particularly amenable to DHH's filtering approach, as we expect the proportion of global nulls to be relatively small, and thus, relatively few PCH p-values will be above τ. We explore a wider range of data generating settings, including those more amenable to filtering approaches in general, in Section 3.3.3.

Lastly, we note that the cPCH test with BH and Storey's procedure is slightly more powerful than its standard counterparts, but is less powerful than the filtering and EB methods. However, a primary advantage of the cPCH test is its flexibility to be combined with different multiple testing procedures depending on the problem specifics, thus allowing it to gain power over approaches that may not have been designed for the specific problem at hand. In particular, we have thus far not discussed the use of available side information in the form

the covariate X_j, and we exhibit the power gains that can be achieved through using the cPCH test when allowing it to use the covariate information in the following subsection.

3.3.2.2 With Covariates

Now including methods that leverage covariate information, we find that the cPCH test with AdaPT–GMM is more powerful than all other approaches, while still controlling FDR (Figure 6). Like with BH and Storey's procedure, DACT with AdaPT–GMM seems to be the most powerful but has large FDR inflation and thus cannot be fairly compared with the other approaches. Generally, methods that can leverage covariate information have greater power than methods that cannot. Even the highly conservative Max-P test paired with AdaPT–GMM outperforms HDMT (the best performing method excluding covariate information), AdaFilter, and the DHH–based approaches, including DHH with AdaPT–GMM. Additionally, the cPCH test with AdaPT–GMM is much more powerful than every other method in the low signal regime, which, as we discussed in Section 1.3.1, is an important regime for various application areas such as genetic epidemiology.

Interestingly, AdaPT–GMM with cPCH and AdaPT–GMM with Max-P are the two most powerful approaches that also control FDR, thus outperforming AdaPT–GMM when paired with PCH multiple testing methods like DACT and DHH. Notably, while the cPCH test with AdaPT–GMM is significantly more powerful than the cPCH test with Storey's procedure, DHH with AdaPT–GMM is only slightly more powerful than DHH with Storey's procedure. This trend likely occurs because the transformation applied to the DHH p-values causes the covariate to seem much less informative than it actually is, undercutting the benefits of AdaPT–GMM. Figure 16 in Appendix C.2.5, which shows the rejections made by AdaPT–GMM for the cPCH test and DHH, shows that AdaPT–GMM with cPCH is able to clearly detect that the covariate being above 0.95 is highly informative of the PCH being in the alternative while that trend is more muddled when using DHH.

This covariate-assisted setting also exhibits how the cPCH test can leverage learned prior information while empirically controlling FDR, which is one of the primary challenges for EB methods. Most importantly, this setting highlights what is perhaps the primary advantage of the cPCH test for PCH multiple testing: it can be powerfully combined with various multiple testing procedures, thus allowing users to leverage the vast existing literature on multiple testing to choose procedures that will allow them to gain power based on the specifics of their problem.

3.3.3 PCH Multiple Testing for $m = 4$

To approximate the real-data example in the following section, we conduct a simulation study to assess the empirical FDR and power of various PCH multiple testing approaches when $m = 4$ and $M = 2000$. As we explore a few different testing scenarios in our real data example, we choose a data generating procedure that allows us to flexibly vary the proportion of null and alternative configurations, thus emulating a wide range of possible testing scenarios for $m = 4$ and $M = 2000$. We generate the data T_{ij}, $i = 1, ..., 4$, $j =$

$1, .., 2000$ by sampling independently from the following model:

$$\begin{aligned} B_j &\sim \text{Bern}(\pi_1) \\ \gamma_{ij}|B_j &\sim \begin{cases} B_j & \text{w/p } w \\ \text{Bern}(\pi_1) & \text{w/p } 1-w \end{cases} \\ T_{ij}|\gamma_{ij}, B_j &\sim \mathcal{N}(\theta\gamma_{ij}, 1) \end{aligned}$$

π_1 tunes the expected proportion of the null and alternative PCH's and p tunes the likelihood of each null and alternative configuration. For example, when $w = 1$, we expect π_1 of the total PCH's to be alternatives, all of which are "full alternatives" (i.e., $r^\star = m$), and all remaining PCH's to be global nulls.

Discoveries are generated using AdaFilter, the cPCH ($N = 10,000$) test, the standard Bonferroni, Simes, and Fisher tests, and DHH ($\tau = 0.1$), each with BH and Storey's procedures. All procedures are applied at nominal FDR level $q = 0.1$. The EB methods under consideration focus on the $m = 2$ case and hence cannot be applied in this setting.

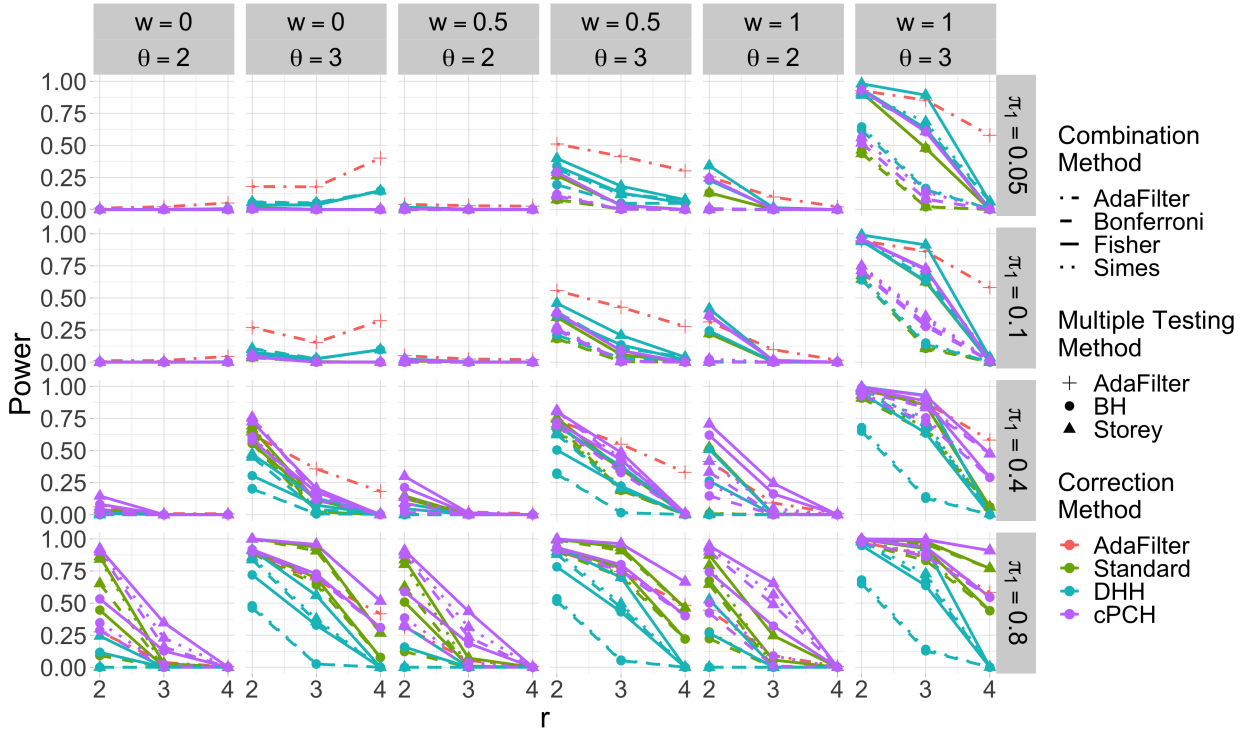


Figure 7: Power of various methods for PCH multiple testing at nominal FDR level $q = 0.1$. Each point represents the average proportion of non-null PCH's which are rejected over 1000 independent replicates of the data generating procedure described in Section 3.3.3 for a given θ. All standard errors were less than 0.015.

Appendix C.2.4 shows that all methods under consideration empirically control the FDR at the nominal level for every combination of θ, π_1, and w tested. Note, finite-sample FDR

control is guaranteed with the Max-P test (with BH and Storey's procedure) and DHH (with BH and Storey's procedure). These results, along with those of the previous section, suggest that AdaFilter and the cPCH–based approaches, neither of which guaranteed FDR control at level q in this setting, have robust FDR control across a wide array of testing settings. As shown in Figure 7, we find that the relative power of these methods is highly sensitive to the underlying data generating distribution, so no method dominates all others across all settings. However, the cPCH test is more powerful than its standard counterparts in all data generating settings. As expected, we see that the AdaFilter and DHH methods have the highest power when π_1 is small, as the construction of filtering methods allows them to be especially effective in this setting. For instance, DHH is likely to significantly reduce the multiplicity of PCH's being tested as a vast majority of the PCH p-values will be above τ.

However, in settings where π_1 is large, the cPCH test with Storey's procedure generally outperforms all other methods. As we alluded to in Section 1.3.2, even the standard single PCH tests with BH and Storey's procedure can outperform DHH and AdaFilter in this setting. There are various important applications where π_1 is expected to be large. For instance, in epigenetics, promising sets of candidate genes from pre-screening studies or prior knowledge are often re-evaluated in follow-up studies (O'Donovan et al., 2008; Benjamini et al., 2009; Rietveld et al., 2014). In this setting, researchers expect that a relatively large proportion of genes will show replicating results as they have already been identified as promising. We explore this follow-up setting through a real-data example in Section 4.

4 Differential Gene Expression Analysis

Duchenne Muscle Dystrophy (DMD) is a genetic disorder characterized by progressive muscle degeneration in young children. Understanding the genetic markers of DMD enables clinicians to link the effects of the disease with their associated genes, thus providing a pathway for targeted drug therapies. Analyzing the average power in both a single PCH testing and follow-up study setting, we demonstrate that the cPCH test has improved power for detecting replicating genetic markers that may be associated with DMD progression over existing methods in these settings.

As in Kotelnikova et al. (2012) and Wang et al. (2021), we analyze four independent DMD-related microarray datasets (GDS214, GDS563, GDS1956, and GDS3027) from the Gene Expression Omnibus (GEO) database, a public functional genomics data repository. The datasets are first pre-processed using `limma`, the standard package for processing microarray data. For each gene, `limma` outputs the test statistic from a two-sample t-test for testing whether that gene is differentially expressed in DMD patients compared to healthy subjects; the test statistics are post-processed as in Wang et al. (2021) to account for some data artifacts. Therefore, we expect the test statistics to follow an approximately normal location family with unit variance, as assumed in Condition (1).

Since the studies use various microarray platforms to measure gene expression, some genes have multiple measurements associated with different probe sets. To combine data for a single gene across the probe sets, we take an average of that gene's corresponding test

statistic for each probe set and scale appropriately so that the resulting test statistic is still approximately normally distributed with unit variance. In total, $M = 1871$ unique genes are shared among the 4 studies.

First, we can treat this data as a set of p-values and assess the empirical power, i.e., the total positive rate, calculated as the proportion of p-values below the nominal level α across the M genes, of the cPCH test compared to that of its standard counterparts for single PCH testing. As shown in Table 3, the cPCH test makes substantially more rejections than its standard counterparts at nominal level $\alpha = 0.05$.

	cPCH-Fisher	cPCH-Simes	cPCH-Bonferroni
r = 2	27.8 (+18.8%)	24.9 (+23.9%)	24.7 (+25.4%)
r = 3	18.3 (+43.0%)	17.7 (+51.3%)	16.2 (+39.7%)
r = 4	9.2 (+84.0%)	9.2 (+84.0%)	9.2 (+84.0%)

Table 3: Total positive rate of the cPCH test compared to its standard counterparts at nominal level $\alpha = 0.05$. The first value in each entry represents the proportion of cPCH p-values below α and the second value represents the percent increase in the total positive rate compared to the corresponding standard single PCH test. Treating the p-values as independent, the implied standard errors are approximately 0.9% for $r = 2, 3$ and $0.5 - 0.7\%$ for $r = 3, 4$.

Next, to emulate a follow-up study analysis, we first fix one of the studies as the primary study, then apply BH at nominal FDR level $q_0 = 0.1$ to select promising candidate genes. The remaining three studies are treated as follow-up studies to test the partial conjunction hypotheses for $r = 2, 3$ at nominal FDR level $q = 0.1$ for the candidate genes from the primary study. We repeat this for all 4 studies.

We also conduct a follow-up study analysis using two of the studies to filter candidate genes, as it is often common for the same laboratory to conduct two independent studies to verify their results, then have separate independent groups verify those results in follow-up studies (Rietveld et al., 2014). For the two-study follow-up analysis, we use the Max-P test at nominal FDR level $q_0 = 0.1$ to find the candidate genes, then test the partial conjunction hypothesis for $r = 2$ at nominal FDR level $q = 0.1$ using the remaining two studies.

Table 4 shows that cPCH using Fisher's combining function and Storey's procedure leads to the discovery of more differentially expressed genes in both the one-study-screen and two-study-screen follow-up study settings. However, in a standard PCH testing analysis, where we test $H_0^{r/4}$ using all four studies without any initial screening, AdaFilter outperforms the cPCH-based approaches for $r = 3, 4$; see Appendix D.1 for more details. Table 9 in Appendix D.2 shows 20 of the total genes discovered by cPCH using Fisher's combining function and Storey's procedure at $r = 3$ with GDS1956 set as the initial screening study. As desired, many of these genes have biological functions associated with muscle maintenance and cell growth regulation. In particular, MYH3, MYH8, MYL4, and MYL5 are known genetic markers for DMD.

		cPCH-Storey	cPCH-BH	DHH-Storey	DHH-BH	AdaFilter
One Study Screen	r = 2	254.3 (+15.4%)	179.3 (+8.47%)	106.0	124.5	168.0
	r = 3	78.5 (+49.5%)	58.5 (+29.3%)	21.0	23.8	68.8
Two Study Screen	r = 2	33.5 (+5.78%)	26.5 (+12.0%)	13.7	18.5	26.5

Table 4: The average number of rejections made for each method when screening first on one or two of the four DMD studies, then treating the remaining studies as follow-up studies using nominal FDR level $q = 0.1$ for the partial conjunction hypothesis testing. Fisher's combining function is used throughout as we found it outperforms Simes and Bonferroni in this example. Percentages represent the percent increase in rejections over the corresponding standard methods.

5 Conclusion

The cPCH test resolves the conservativeness of standard approaches by conditioning on the largest base test statistics, allowing the test statistic to be less sensitive to the estimation of unknown parameters (Section 2.2). While our implementation is tailored to the common setting where each base test statistic is a single, independent, unit-variance Gaussian, we find that the cPCH test is robust to model misspecification (Section 3.2). Additionally, our framework for computing cPCH p-values can be extended to any one-parameter location family (Section 2.5), and we expect the power and Type I error control results for Gaussian base test statistics to extend to other distributional assumptions (Section 3.2).

Through simulations and a real data example, we find that the cPCH test produces nearly uniform p-values under the null and uniformly outperforms standard single PCH testing approaches (Sections 2.4, 3.1). In certain settings, the cPCH test used in combination with different multiple testing procedures outperforms state-of-the-art approaches designed for PCH multiple testing (Section 3.3). In particular, our results on PCH multiple testing with covariates exhibit a primary advantage of the cPCH test in the PCH multiple testing setting: it can flexibly adapt to specific problems by leveraging the vast, existing literature on multiple testing procedures, many of which require or perform best when provided p-values that are uniform under the null (Section 3.3.2).

6 Acknowledgements

The authors would like to thank Nathan Cheng for his help with Lemma 4, Jingshu Wang for sharing her pre-processing code for the DMD data, James Dai and Frank Wang for providing further information on the HDMT package, and Xihong Lin for helpful discussions regarding this work. B.L. is partially supported by the National Science Foundation via the Graduate Research Fellowship Program. L.Z. and L.J. are partially supported by a grant from the National Science Foundation (Grant #DMS-2134157). B.L., L.Z., and L.J. are partially supported by a CAREER grant from the National Science Foundation (Grant

#DMS2045981).

References

Bapat, R. B. and Beg, M. I. (1989), "Order Statistics for Nonidentically Distributed Variables and Permanents", *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **51**(1), 79–93.

Barfield, R., Shen, J., Just, A. C., Vokonas, P. S., Schwartz, J., Baccarelli, A. A., VanderWeele, T. J. and Lin, X. (2017), "Testing for the Indirect Effect Under the Null for Genome-Wide Mediation Analyses", *Genetic Epidemiology* **41**(8), 824–833.

Baron, R. M. and Kenny, D. A. (1986), "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations", *Journal of Personality and Social Psychology* **51**, 1173.

Benjamini, Y. and Heller, R. (2008), "Screening for Partial Conjunction Hypotheses", *Biometrics* **64**(4), 1215–1222.

Benjamini, Y., Heller, R. and Yekutieli, D. (2009), "Selective Inference in Complex Research", *Philosophical transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences* **367**(1906), 4255–4271.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society Series B (Methodological)* **57**(1), 289–300.

Bogomolov, M. and Heller, R. (2013), "Discovering Findings That Replicate From a Primary Study of High Dimension to a Follow-Up Study", *Journal of the American Statistical Association* **108**(504), 1480–1492.

Braver, S. L., Thoemmes, F. J. and Rosenthal, R. (2014), "Continuously Cumulating Meta-Analysis and Replicability", *Perspectives on Psychological Science* **9**(3), 333–342. PMID: 26173268.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016), "Evaluating Replicability of Laboratory Experiments in Economics", *Science* **351**(6280), 1433–1436.

Chao, P. and Fithian, W. (2021), "AdaPT-GMM: Powerful and Robust Covariate-Assisted Multiple Testing". arXiv: 2106.15812.

Dai, J. Y., Stanford, J. L. and LeBlanc, M. (2020), "A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses", *Journal of the American Statistical Association* **0**(0), 1–16.

Dickhaus, T., Heller, R. and Hoang, A.-T. (2021), "Multiple Testing of Partial Conjunction Null Hypotheses, With Application to Replicability Analysis of High Dimensional Studies". arXiv: 2110.06692.

Dreyfuss, J. M., Yuchi, Y., Dong, X., Efthymiou, V., andDonald C. Simonson, H. P., Vernon, A., Halperin, F., Aryal, P., Konkar, A., Sebastian, Y., Higgs, B. W., Grimsby, J., Rondinone, C. M., Kasif, S., Kahn, B. B., Foster, K., Seeley, R., Goldfine, A., Djordjilović, V. and Patti, M. E. (2021), "High-Throughput Mediation Analysis of Human Proteome and Metabolome Identifies Mediators of Post-Variatric Surgical Diabetes Control", *Nature Communications* **21**, 6951.

Efron, B. (2008), "Microarrays, Empirical Bayes and the Two-Groups Model", *Statistical Science* **23**(1), 1 – 22.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment", *Journal of the American Statistical Association* **96**(456), 1151–1160.

Ernst, M. D. (2004), "Permutation Methods: A Basis for Exact Inference", *Statistical Science* pp. 676–685.

Folland, G. (2002), *Advanced Calculus*, Featured Titles for Advanced Calculus Series, Prentice Hall.

Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2017), "Isotonic Regression in General Dimensions". DOI: 10.48550/ARXIV.1708.09468.

Heller, R., Golland, Y., Malach, R. and Benjamini, Y. (2007), "Conjunction Group Analysis: An Alternative to Mixed/Random Effect Analysis", *NeuroImage* **37**, 1178–1185.

Heller, R. and Yekutieli, D. (2014), "Replicability Analysis for Genome-Wide Association Studies", *The Annals of Applied Statistics* **8**(1), 481 – 498.

Hirschhorn, J. N. and Altshuler, D. (2002), "Once and Again—Issues Surrounding Replication in Genetic Association Studies", *The Journal of Clinical Endocrinology and Metabolism* **87**(10), 4438–4441.

Huang, Y.-T. (2019), "Genome-wide Analyses of Sparse Mediation Effects under Composite Null Hypotheses", *Annals of Applied Statistics* **13**, 60 – 84.

Ioannidis, J. P. A. (2005), "Why Most Published Research Findings Are False", *PLoS Medicine* **2**(8), e124.

Jin, J. and Cai, T. T. (2007), "Estimating the Null and the Proportion of Nonnull Effects in Large-Scale Multiple Comparisons", *Journal of the American Statistical Association* **102**(478), 495–506.

Karmakar, B. and Small, D. S. (2020), "Assessment of the Extent of Corroboration of an Elaborate Theory of a Causal Hypothesis Using Partial Conjunctions of Evidence Factors", *The Annals of Statistics* **48**(6), 3283 – 3311.

Karmakar, B., Small, D. S. and Rosenbaum, P. R. (2021), "Reinforced Designs: Multiple Instruments Plus Control Groups as Evidence Factors in an Observational Study of the Effectiveness of Catholic Schools", *Journal of the American Statistical Association* **116**(533), 82–92.

Kelley, J. (1955), *General Topology*, Graduate texts in mathematics, D. Van Nostrand.

Kiwiel, K. and Murty, K. (1996), "Convergence of the steepest descent method for minimizing quasiconvex functions", *Journal of Optimization Theory and Applications* **89**.

Kotelnikova, E., Shkrob, M., Pyatnitskiy, M., Ferlini, A. and Daraselia, N. (2012), "Novel Approach to Meta-Analysis of Microarray Datasets Reveals Muscle Remodeling-Related Drug Targets and Biomarkers in Duchenne Muscular Dystrophy", *PLoS Computational Biology* **8**(2).

Kraft, P., Zeggini, E. and Ioannidis, J. P. A. (2009), "Replication in Genome-Wide Association Studies", *Statistical Science* **24**(4), 561 – 573.

Lee, J. (2003), *Introduction to Smooth Manifolds*, Graduate Texts in Mathematics, Springer.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A. A. and Lin, X. (2022), "Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies", *Journal of the American Statistical Association* **117**(537), 67–81.

NCI-NHGRI Working Group on Replication in Association Studies (2007), "Replicating Genotype–Phenotype Associations", *Nature* **447**, 655–660.

O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., Nikolov, I., Hamshere, M., Carroll, L., Georgieva, L., Dwyer, S., Holmans, P. and Marchini, J. L. (2008), "Identification of Loci Associated with Schizophrenia by Genome-wide Association and Follow-up", *Nature Genetics* **40**, 1053–1055.

Ozbey, F., Güngör, M. and Bulut, Y. (2019), "On Distributions of Order Statistics for Nonidentically Distributed Variables", *Appl. Math* **13**(1), 11–16.

Patel, V. and Zhang, S. (2021), "Stochastic Gradient Descent on Nonconvex Functions with General Noise Models". DOI: 10.48550/ARXIV.2104.00423.

Rietveld, C. A., Conley, D., Eriksson, N., Esko, T., Medland, S. E., Vinkhuyzen, A. A. E., Yang, J., Boardman, J. D., Chabris, C. F., Dawes, C. T., Domingue, B. W., Hinds, D. A., Johannesson, M., Kiefer, A. K., Laibson, D., Magnusson, P. K. E., Mountain, J. L., Oskarsson, S., Rostapshova, O., Teumer, A., Tung, J. Y., Visscher, P. M., Benjamin, D. J., Cesarini, D., Koellinger, P. D. and the Social Science Genetics Association Consortium (2014), "Replicability and Robustness of Genome-Wide-Association Studies for Behavioral Traits", *Psychological Science* **25**(11), 1975–1986. PMID: 25287667.

Sesia, M., Bates, S., Candès, E., Marchini, J. and Sabatti, C. (2021), "False Discovery Rate Control in Genome-Wide Association Studies With Population Structure", *Proceedings of the National Academy of Sciences* **118**(40), e2105841118.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates", *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **64**(3), 479–498.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach", *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **66**(1), 187–205.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, chapter 21.

Wang, J., Gui, L., Su, W. J., Sabatti, C. and Owen, A. B. (2021), "Detecting Multiple Replicating Signals using Adaptive Filtering Procedures".

Wedhorn, T. (2016), *Manifolds, Sheaves, and Cohomology*, Springer Studium Mathematik - Master, Springer Fachmedien Wiesbaden.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L. and Liu, L. (2016), "Estimating and Testing High-Dimensional Mediation Effects in Epigenetic Studies", *Bioinformatics* **32**(20), 3150–3154.

Zuo, L., Zhang, C. K., Wang, F., Li, C.-S. R., Zhao, H., Lu, L., Zhang, X.-Y., Lu, L., Zhang, H., Zhang, F., Krystal, J. H. and Luo, X. (2011), "A Novel, Functional and Replicable Risk Gene Region for Alcohol Dependence Identified by Genome-Wide Association Study", *PLOS ONE* **6**(11), 1–8.

A Proof of Theorem 1

In the body of this paper, we state Theorem 1 with conditions that are sufficient for the weaker conditions we use here. In particular, Assumption 3 allows for more general sequences of estimators for $\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}$ when setting the rejection threshold of the cPCH test. We show in Lemma 6 that choosing $\boldsymbol{T}^{(n)}_{(m-r+2:m)}$ as the estimator for $\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}$, as in the definition of the cPCH test in Definition 2, satisfies Assumption 3.

A.1 Preliminaries

Let $\hat{\boldsymbol{\theta}}_{(m-r+2:m)}$ be an estimator for $\boldsymbol{\theta}_{(m-r+2:m)}$. Recall that we define $c_\alpha(\boldsymbol{\theta}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)})$ to be the $1-\alpha$ quantile of the distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid \boldsymbol{T}_{(m-r+2:m)}$ where $\boldsymbol{T} \sim \mathcal{N}(\boldsymbol{\theta}, I_m)$, $\boldsymbol{\theta}_{(1:m-r+1)} = \mathbf{0}$, $\boldsymbol{\theta}_{(m-r+2:m)}$ comes from the first argument to c_α, and I_m is the $m \times m$ identity matrix. As in Section 2.1, we use order statistic notation to refer to ordering by *magnitudes*, i.e., for a vector of test statistics $\boldsymbol{T}$, we order $\{T_{(i)}\}_{i=1}^m$ by $|T_{(1)}| \leq ...|T_{(m)}|$ and let $\boldsymbol{T}_{(i:j)} = (T_{(i)}, ..., T_{(j)})$ for $i \leq j$.

Definition 4 (Generalized cPCH test)**.** *The generalized cPCH test rejects $H_0^{r/m}$ when*

$$f\left(\boldsymbol{T}_{(1:m-r+1)}\right) > c_\alpha\left(\hat{\boldsymbol{\theta}}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right).$$

Setting $\hat{\boldsymbol{\theta}}_{(m-r+2:m)} = \boldsymbol{T}_{(m-r+2:m)}$ recovers the original cPCH test as in Definition 2.

Recall from Section 2.4.1 that we define an LFN sequence $\left(\boldsymbol{\theta}^{(n)}\right)$ as a sequence in $\Theta_0^{r/m}$ such that $\left|\boldsymbol{\theta}^{(n)}_{(j)}\right| \to \infty$ for $j = m-r+2, ..., m$ as $n \to \infty$. Note, the remaining $\boldsymbol{\theta}^{(n)}_{(1)}, ..., \boldsymbol{\theta}^{(n)}_{(m-r+1)}$ are zero by definition of $\Theta_0^{r/m}$. Recall from Section 2.3 that we define the rejection threshold of the PCH Oracle test $c_\alpha\left(\boldsymbol{\theta}_{(m-r+2:m)}\right)$ as the $1-\alpha$ quantile of the distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right)$ where $\boldsymbol{T} \sim \mathcal{N}(\boldsymbol{\theta}, I_m)$ and $\boldsymbol{\theta}$ is comprised of $m-r+1$ zeroes and the elements of $\boldsymbol{\theta}_{(m-r+2:m)}$. For a test statistic vector $\boldsymbol{T}^{(n)}$, let $\varphi_\alpha^{cPCH}\left(\boldsymbol{T}^{(n)}\right) \coloneqq \mathbb{1}\left\{f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) > c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right)\right\}$ and $\varphi_\alpha^{PCHOrac}\left(\boldsymbol{T}^{(n)}\right) \coloneqq \mathbb{1}\left\{f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) > c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right)\right\}$ represent the decisions made by the generalized cPCH test and PCH Oracle test, respectively. Throughout all proofs, we will use $\xrightarrow{p}$ to denote convergence in probability and the superscript ^{c} to denote the complement of a set. In the following section, we state and prove a version of Theorem 1 for the generalized cPCH test.

A.2 Theorem 1 for the Generalized cPCH Test

Assumption 1. *Assume $\left(\boldsymbol{\theta}^{(n)}\right)$ is a LFN sequence and that $\boldsymbol{T}^{(n)} \sim \mathcal{N}(\boldsymbol{\theta}^{(n)}, I_m)$.*

Assumption 2. *Assume $f : \mathbb{R}^{m-r+1} \to \mathbb{R}$ is permutation invariant, continuously differentiable, and has $\boldsymbol{\nabla} f \neq 0$ except on a set whose closure has measure zero.*

Assumption 3. *Assume $\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}\right)$ is a sequence of estimators for the LFN sequence*

$\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right)$ *with the following property: For any* $K > 0$, $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|\hat{\boldsymbol{\theta}}^{(n)}_{(i)}\right| > K\right) = 1$ *for* $i = m-r+2, ..., m$.

Theorem 1* (Exactness of the generalized cPCH test under the LFN case)**.** *Under Assumptions 1–3, for any* $\alpha \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{cPCH}\left(\boldsymbol{T}^{(n)}\right) = \varphi_\alpha^{PCHOrac}\left(\boldsymbol{T}^{(n)}\right)\right) = 1.$$

In particular, the above implies that the generalized cPCH test's limiting Type I error under any LFN sequence is exactly α.

Proof. Without loss of generality, assume $\boldsymbol{\theta}^{(n)} = \left(0, ...0, \theta^{(n)}_{m-r+2}, ..., \theta^{(n)}_m\right)$ where $\theta^{(n)}_i \to \infty$ for $i = m-r+2, ..., m$. Let $R_f \coloneqq \text{Range}(f)$. For $x \in R_f$, let $G\left(x, \boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) \coloneqq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \le x\right)$.[2] Let $H(x) \coloneqq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{1:m-r+1}\right) \le x\right)$; note that $\boldsymbol{T}^{(n)}_{1:m-r+1} \sim \mathcal{N}(\mathbf{0}, I_{m-r+1})$ for all n, so H does not vary with n. For $a \in (0,1)$, let

$$c_a\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) = \inf\left\{x : G\left(x, \boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) \ge 1-a\right\} \text{ and } c^*_a = \inf\{x : H(x) \ge 1-a\}.$$

Lemma 1 (stated below and proved in Appendix A.3) establishes that $c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right)$, the rejection threshold of the level α PCH Oracle test, converges to c^*_α:

Lemma 1. *Under Assumptions 1-2, for* $\alpha \in (0,1)$,

$$c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) \to c^*_\alpha.$$

Lemma 2 (stated below and proved in Appendix A.3) establishes that $c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right)$, the rejection threshold of the level α generalized cPCH test, converges to c^*_α in probability:[3]

Lemma 2. *Under Assumptions 1-3, for* $\alpha \in (0,1)$,

$$c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) \xrightarrow{p} c^*_\alpha.$$

Fix $\alpha \in (0,1)$ and $\epsilon > 0$. Let $Q_n = \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c^*_\alpha\right| \le \epsilon\right)$. By

[2] By Assumption 2, G is permutation invariant with respect to $\boldsymbol{\theta}^{(n)}$, so $\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \le x\right) = \mathbb{P}_{\sigma(\boldsymbol{\theta}^{(n)})}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \le x\right)$ for any permutation of the elements of $\boldsymbol{\theta}^{(n)}$, $\sigma(\boldsymbol{\theta}^{(n)})$. Therefore, we can use $\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}$ to represent $\boldsymbol{\theta}^{(n)}$ in the input to G.

[3] Unlike $c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right)$, which is a fixed real number for each n, $c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right)$ is a random variable that is a function of $\boldsymbol{T}^{(n)}$ both through the conditioning event and $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}$.

Lemma 2, $\lim_{n\to\infty} Q_n = 1$. So,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right)\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right) \;\middle|\; |c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| \leq \epsilon\right) Q_n + \\
&\qquad \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right) \;\middle|\; |c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| > \epsilon\right)(1 - Q_n) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right) \;\middle|\; |c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| \leq \epsilon\right),
\end{aligned} \tag{3}$$

where the last line follows from Lemma 2. By Lemma 1, there exists $N_\epsilon \in \mathbb{N}$ such that for all $n \geq N_\epsilon$, $|c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| < \epsilon$. So, for $n \geq N_\epsilon$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right) \;\middle|\; |c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| \leq \epsilon\right) \\
&\leq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon) \;\middle|\; |c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| \leq \epsilon\right).
\end{aligned}$$

Since $\lim_{n\to\infty} Q_n = 1$,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon) \;\middle|\; |c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) - c_\alpha^*| \leq \epsilon\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon)\right)
\end{aligned} \tag{4}$$

where the last line follows by Lemma 3, (stated below and proved in Appendix A.4), assuming the above limit exists, which we show in Equation (5):

Lemma 3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let A_n and B_n be a sequence of events in Ω where $\lim_{n\to\infty} \mathbb{P}(B_n) = 1$ and $\lim_{n\to\infty} \mathbb{P}(A_n) = a$ for some $a \in [0, 1]$. Then, $\mathbb{P}(A_n \cap B_n) \to a$ and $\mathbb{P}(A_n \mid B_n) \to a$.*

Let $B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)$ be the event that, for $\boldsymbol{T}^{(n)} \sim \mathcal{N}(\boldsymbol{\theta}^{(n)}, I_m)$, $\left\{T^{(n)}_{(m-r+2)}, ..., T^{(n)}_{(m)}\right\} = \left\{T^{(n)}_{m-r+2}, ..., T^{(n)}_{m}\right\}$. The set $\left\{T^{(n)}_{(m-r+2)}, ..., T^{(n)}_{(m)}\right\}$ will almost surely be uniquely defined because we will almost surely have no ties among the $T_i^{(n)}$'s since $\boldsymbol{T}^{(n)} \sim \mathcal{N}(\boldsymbol{\theta}^{(n)}, I_m)$. We show in the proof of Lemma 1 that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = 1$. So,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon)\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon) \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) + \\
&\qquad \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon) \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)^c\right) \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)^c\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon) \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{1:m-r+1}\right) \in (c_\alpha^* - \epsilon, c_\alpha^* + \epsilon) \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \\
&= H(c_\alpha^* + \epsilon) - H(c_\alpha^* - \epsilon),
\end{aligned} \tag{5}$$

where the fourth line follows from the fact that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(B_n \left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = 1$ and the fifth line follows from the conditioning on $B_n \left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)$ and the permutation invariance of f in Assumption 2.

The last line follows from applying Lemma 3 using the facts that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(B_n \left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = 1$ and $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(f\left(\boldsymbol{T}^{(n)}_{1:m-r+1}\right) \in (c^*_\alpha - \epsilon, c^*_\alpha + \epsilon)\right) = H(c^*_\alpha + \epsilon) - H(c^*_\alpha - \epsilon)$. Combining Equations (3)-(5), we have that

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right)\right) \leq H(c^*_\alpha + \epsilon) - H(c^*_\alpha - \epsilon).$$

Since ϵ was arbitrary, the above is true for any $\epsilon > 0$. Additionally, H is continuous on R_f as a result of Lemma 4 (stated below and proved in Appendix A.4):

Lemma 4. *If $\boldsymbol{X} = (X_1, ..., X_l)$ is a vector of continuous random variables supported on $\mathbb{R}^l$ with density $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^l$ and $f : \mathbb{R}^l \to \mathbb{R}$ is continuously differentiable function with $\boldsymbol{\nabla} f \neq 0$ except on a set whose closure has measure zero, then the CDF of $f(\boldsymbol{X})$ is continuous and strictly increasing on the range of f.*

So, by the continuity of H, as $\epsilon \to 0$, $H(c^*_\alpha + \epsilon) - H(c^*_\alpha - \epsilon) \to 0$.

Therefore, we conclude that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) \neq \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right)\right) = 0$ and hence,

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(\varphi_\alpha^{\text{cPCH}}\left(\boldsymbol{T}^{(n)}\right) = \varphi_\alpha^{\text{PCHOrac}}\left(\boldsymbol{T}^{(n)}\right)\right) = 1.$$

□

A.3 Lemmas for Proving Theorem 1*

Lemmas 1-2 are the main results that allow us to prove Theorem 1*, with Lemmas 3-4 providing useful results that help to complete the proof of Theorem 1* and are used in the proofs of Lemmas 1-2. The proof of Lemma 2 additionally relies on Lemmas 5-6. The proofs of Lemmas 1-2 are below. Lemmas 3-6 are proved in Appendix A.4.

Lemma 1. *Under Assumptions 1-2, for $\alpha \in (0, 1)$,*

$$c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) \to c^*_\alpha.$$

Proof. As in Theorem 1*, let $\boldsymbol{\theta}^{(n)} = \left(0, ..., 0, \theta^{(n)}_{m-r+2}, ..., \theta^{(n)}_m\right)$ where $\theta^{(n)}_i \to \infty$ and define $B_n \left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)$ as the event that, for $\boldsymbol{T}^{(n)} \sim \mathcal{N}(\boldsymbol{\theta}^{(n)}, I_m)$, $\left\{T^{(n)}_{(m-r+2)}, ..., T^{(n)}_{(m)}\right\} = \left\{T^{(n)}_{m-r+2}, ..., T^{(n)}_m\right\}$. We will first show that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(B_n \left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = 1$. Let $K > 0$. Then,

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}} \left(\left|T^{(n)}_i\right| > K\right) = \lim_{n\to\infty} 1 - \Phi\left(K - \theta^{(n)}_i\right) + \Phi\left(-K - \theta^{(n)}_i\right) = 1, \tag{6}$$

where the last line follows from Assumption 1. Let $F_K^{(n)} := \left\{\min_{i=m-r+2,...,m} \left|T_i^{(n)}\right| > K\right\}$. Then,

$$\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(F_K^{(n)}\right) = \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\bigcap_{i=m-r+2}^{m}\left\{\left|T_i^{(n)}\right| > K\right\}\right) = \prod_{i=m-r+2}^{m}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T_i^{(n)}\right| > K\right) \to 1, \tag{7}$$

where the second equality follows from the fact that $T_{m-r+2}^{(n)}, ..., T_m^{(n)}$ are independent and the last result follows from Equation (6). So,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) &= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left\{T_{(m-r+2)}^{(n)}, ..., T_{(m)}^{(n)}\right\} = \left\{T_{m-r+2}^{(n)}, ..., T_m^{(n)}\right\}\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < \min_{i=m-r+2,...,m}\left|T_i^{(n)}\right|\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < \min_{i=m-r+2,...,m}\left|T_i^{(n)}\right| \,\middle|\, F_K^{(n)}\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(F_K^{(n)}\right) \\
&\quad + \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < \min_{i=m-r+2,...,m}\left|T_i^{(n)}\right| \,\middle|\, F_K^{(n)c}\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(F_K^{(n)c}\right).
\end{aligned}$$

By Equation (7),

$$\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = \lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < \min_{i=m-r+2,...,m}\left|T_i^{(n)}\right| \,\middle|\, F_K^{(n)}\right),$$

where, for any n,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < \min_{i=m-r+2,...,m}\left|T_i^{(n)}\right| \,\middle|\, F_K^{(n)}\right) \\
&\geq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < K \,\middle|\, F_K^{(n)}\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|T_j^{(n)}\right| < K\right) \\
&= (\Phi(K) - \Phi(-K))^{m-r+1}.
\end{aligned}$$

Therefore,

$$\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \geq (\Phi(K) - \Phi(-K))^{m-r+1}.$$

Since K was arbitrary, the above line holds for any $K > 0$. Note that as K gets arbitrarily large, $(\Phi(K) - \Phi(-K))^{m-r+1} \to 1$. Therefore,

$$\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = 1. \tag{8}$$

Let $x \in R_f$. Then,

$$\begin{aligned}
\lim_{n\to\infty} G\left(x, \boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) &= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \leq x\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \leq x \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) + \\
&\qquad \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \leq x \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)^c\right) \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)^c\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{(1:m-r+1)}\right) \leq x \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \\
&= \lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{1:m-r+1}\right) \leq x \;\middle|\; B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) \\
&= H(x),
\end{aligned}$$

where the fourth line follows from Equation (8), and the fifth line follows from the conditioning on $B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)$ and the permutation invariance of f in Assumption 2. The last line follows from Lemma 3 using Equation (8) and the fact that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(f\left(\boldsymbol{T}^{(n)}_{1:m-r+1}\right) \leq x\right) = H(x)$. So,

$$G\left(x, \boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) \to H(x).$$

Since x was arbitrary, the above is true for any $x \in R_f$. Additionally, H is continuous on R_f as a result of Lemma 4, where f satisfies the conditions of Lemma 4 by Assumption 2. Therefore, by Lemma 21.2 of van der Vaart (1998), for any $\alpha \in (0, 1)$,

$$c_\alpha\left(\boldsymbol{\theta}^{(n)}_{(m-r+2:m)}\right) \to c^*_\alpha.$$

□

Lemma 2. *Under Assumptions 1-3, for any* $\alpha \in (0, 1)$,

$$c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) \xrightarrow{p} c^*_\alpha.$$

Proof. Without loss of generality, let $\boldsymbol{\theta}^{(n)} = \left(0, ..., 0, \theta^{(n)}_{(m-r+2)}, ..., \theta^{(n)}_{(m)}\right)$ where $\theta^{(n)}_i \to \infty$ and let $\hat{\boldsymbol{\theta}}^{(n)} = \left(0, ..., 0, \hat{\theta}^{(n)}_{(m-r+2)}, ..., \hat{\theta}^{(n)}_{(m)}\right)$. Let

$\tilde{F}\left(x, \boldsymbol{T}^{(n)}\right) := \mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(f\left(\tilde{\boldsymbol{T}}^{(n)}_{(1:m-r+1)}\right) \leq x \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right)$ represent the probability that $f\left(\tilde{\boldsymbol{T}}^{(n)}_{(1:m-r+1)}\right) \leq x$ conditional on $\tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}$ and $\boldsymbol{T}^{(n)}$ where the subscript on the probability denotes that $\tilde{\boldsymbol{T}}^{(n)} \mid \hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}^{(n)}, I_m)$. Note, this probability is only over $\tilde{\boldsymbol{T}}^{(n)}$, which is a function of $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}$. Therefore, $\tilde{F}\left(x, \boldsymbol{T}^{(n)}\right)$ is a random variable since it is a function of $\boldsymbol{T}^{(n)}$ both through the conditioning event and through $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}$. We consider $\tilde{F}\left(x, \boldsymbol{T}^{(n)}\right)$ because $c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) = \inf\left\{x : \tilde{F}\left(x, \boldsymbol{T}^{(n)}\right) \geq 1 - \alpha\right\}$. Thus, to show our final result, we will first show that, for any $x \in R_f$, $\tilde{F}\left(x, \boldsymbol{T}^{(n)}\right) \xrightarrow{p} H(x)$.

Showing $\tilde{F}\left(x, \boldsymbol{T}^{(n)}\right) \xrightarrow{p} H(x)$ relies on Lemma 5 (stated below and proved in Appendix A.4):

Lemma 5. *Under Assumptions 1 and 3,*

$$\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right) \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right) \xrightarrow{p} 1.$$

Here, $B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right)$ represents the event that, for $\tilde{\boldsymbol{T}}^{(n)} \mid \hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}^{(n)}, I_m)$, $\left\{\tilde{T}^{(n)}_{(m-r+2)}, ..., \tilde{T}^{(n)}_{(m)}\right\} = \left\{\tilde{T}^{(n)}_{m-r+2}, ..., \tilde{T}^{(n)}_{m}\right\}$. The set $\left\{\tilde{T}^{(n)}_{(m-r+2)}, ..., \tilde{T}^{(n)}_{(m)}\right\}$ will almost surely be uniquely defined since the $\tilde{T}^{(n)}_i$'s will almost surely have no ties.

Note, $\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right) \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right)$ is also a random variable as it is a function of $\boldsymbol{T}^{(n)}$ both through the conditioning event and through $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}$. For ease of notation, let

$$\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) \coloneqq \mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(f\left(\tilde{\boldsymbol{T}}^{(n)}_{(1:m-r+1)}\right) \leq x \;\middle|\; B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right), \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right).$$

Fix $x \in R_f$. Then,

$$\begin{aligned}\tilde{F}\left(x, \boldsymbol{T}^{(n)}\right) = \tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) \mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right) \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right) + \\ \tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) \mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right)^c \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right).\end{aligned} \tag{9}$$

We will now show that $\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) \xrightarrow{p} H(x)$. By the conditioning on $B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right)$ and the permutation invariance of f in Assumption 2,

$$\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) = \mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(f\left(\tilde{\boldsymbol{T}}^{(n)}_{1:m-r+1}\right) \leq x \;\middle|\; B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right), \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right).$$

$\tilde{T}^{(n)}_1, ..., \tilde{T}^{(n)}_{m-r+1}$ are i.i.d. standard normally distributed by assumption, so,

$$\tilde{T}^{(n)}_1, ..., \tilde{T}^{(n)}_{m-r+1} \mid B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right), \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)} \overset{i.i.d.}{\sim} \text{Trunc–Norm}\left(0, 1, T^{(n)}_{(m-r+2)}\right),$$

where Trunc–Norm$(0, 1, c)$ is the standard normal distribution truncated at $-|c|$ and $|c|$ for $c \in \mathbb{R}$. Let $p_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(\boldsymbol{t} \;\middle|\; B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right), \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right)$ be the PDF of $\tilde{T}^{(n)}_1, ..., \tilde{T}^{(n)}_{m-r+1}$ conditional on $B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right), \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}$ evaluated at $\boldsymbol{t} \in \mathbb{R}^{m-r+1}$. By definition of a truncated normal random variable, we have for any $\boldsymbol{t} \in \mathbb{R}^{m-r+1}$,

$$p_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}}\left(\boldsymbol{t} \;\middle|\; B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right), \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}\right) = \prod_{i=1}^{m-r+1} \frac{\phi(t_i) \mathbb{1}_{T^{(n)}_{(m-r+2)}}(t_i)}{\Phi\left(\left|T^{(n)}_{(m-r+2)}\right|\right) - \Phi\left(-\left|T^{(n)}_{(m-r+2)}\right|\right)},$$

where $\mathbb{1}_{T^{(n)}_{(m-r+2)}}(t) \coloneqq \mathbb{1}\left\{t \in \left(-\left|T^{(n)}_{(m-r+2)}\right|, \left|T^{(n)}_{(m-r+2)}\right|\right)\right\}$ for $t \in \mathbb{R}$.

Set $\epsilon > 0$ and define $S_x = \{\boldsymbol{t} \in \mathbb{R}^{m-r+1} : f(\boldsymbol{t}) < x\}$. By Assumption 2, there exists $K_x > 0$ such that for any $\boldsymbol{t} \in S_x$, $|t_i| < K_x$ for all $i = 1, ..., m-r+1$. So, for any $K_x > 0$ satisfying this condition,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(|\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) - H(x)| > \epsilon\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(|\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) - H(x)| > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| > K_x\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(m-r+2)}\right| > K_x\right) + \\
&\qquad \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(|\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) - H(x)| > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| \leq K_x\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(m-r+2)}\right| \leq K_x\right) \quad (10)
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(|\tilde{G}\left(x, \boldsymbol{T}^{(n)}\right) - H(x)| > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| > K_x\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|\int_{S_x}\left(\prod_{i=1}^{m-r+1}\frac{\phi(t_i)\mathbb{1}_{T^{(n)}_{(m-r+2)}}(t_i)}{\Phi\left(\left|T^{(n)}_{(m-r+2)}\right|\right) - \Phi\left(-\left|T^{(n)}_{(m-r+2)}\right|\right)} - \prod_{i=1}^{m-r+1}\phi(t_i)\right)d\boldsymbol{t}\right| > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| > K_x\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|\int_{S_x}\left(\prod_{i=1}^{m-r+1}\frac{\phi(t_i)}{\Phi\left(\left|T^{(n)}_{(m-r+2)}\right|\right) - \Phi\left(-\left|T^{(n)}_{(m-r+2)}\right|\right)} - \prod_{i=1}^{m-r+1}\phi(t_i)\right)d\boldsymbol{t}\right| > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| > K_x\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\int_{S_x}\left(\prod_{i=1}^{m-r+1}\frac{\phi(t_i)}{\Phi\left(\left|T^{(n)}_{(m-r+2)}\right|\right) - \Phi\left(-\left|T^{(n)}_{(m-r+2)}\right|\right)} - \prod_{i=1}^{m-r+1}\phi(t_i)\right)d\boldsymbol{t} > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| > K_x\right) \\
&\leq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\int_{S_x}\left(\prod_{i=1}^{m-r+1}\frac{\phi(t_i)}{\Phi(K_x) - \Phi(-K_x)} - \prod_{i=1}^{m-r+1}\phi(t_i)\right)d\boldsymbol{t} > \epsilon \;\middle|\; \left|T^{(n)}_{(m-r+2)}\right| > K_x\right) \\
&= \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\int_{S_x}\left(\prod_{i=1}^{m-r+1}\frac{\phi(t_i)}{\Phi(K_x) - \Phi(-K_x)} - \prod_{i=1}^{m-r+1}\phi(t_i)\right)d\boldsymbol{t} > \epsilon\right) \\
&= \mathbb{1}\left\{\left(\frac{1}{\Phi(K_x) - \Phi(-K_x)} - 1\right)^{m-r+1}\int_{S_x}\prod_{i=1}^{m-r+1}\phi(t_i)d\boldsymbol{t} > \epsilon\right\} \\
&= \mathbb{1}\left\{\left(\frac{1}{\Phi(K_x) - \Phi(-K_x)} - 1\right)^{m-r+1}H(x) > \epsilon\right\}.
\end{aligned}$$

In the third line, we drop the indicator in the numerator because of the conditioning on $\left|T^{(n)}_{(m-r+2)}\right| > K_x$. The fourth line follows because the expression within the absolute value is non-negative almost surely. Note, there exists $K_0 > 0$ such that for all $K > K_0$,

$$\Phi(K) - \Phi(-K) > \frac{1}{1 + \left(\frac{\epsilon}{H(x)}\right)^{\frac{1}{m-r+1}}} \quad (11)$$

and therefore,

$$\left(\frac{1}{\Phi(K)-\Phi(-K)}-1\right)^{m-r+1}H(x)<\left(\frac{\epsilon}{H(x)}\right)H(x)=\epsilon.$$

For any K_x satisfying $|t_i|<K_x$ for any $\boldsymbol{t}\in S_x$, any $K>K_x$ also satisfies this condition. Therefore, we can set K_x sufficiently large such that it satisfies this condition and Equation (11). Therefore, for such a K_x,

$$\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(|\tilde{G}\left(x,\boldsymbol{T}^{(n)}\right)-H(x)|>\epsilon\;\middle|\;\left|T^{(n)}_{(m-r+2)}\right|>K_x\right)=0 \tag{12}$$

As a result of Lemma 6 (stated below and proved in Appendix A.4), for any $K>0$, $\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(m-r+2)}\right|>K\right)=1$:

Lemma 6. *Under Assumption 1, $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}=\boldsymbol{T}^{(n)}_{(m-r+2:m)}$ satisfies Assumption 3.*

Therefore, setting K_x in Equation (10) sufficiently large to satisfy Equation (11) and $|t_i|<K_x$ for any $\boldsymbol{t}\in S_x$, and applying Equation (12) and Lemma 6, we have that

$$\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|\tilde{G}\left(x,\boldsymbol{T}^{(n)}\right)-H(x)\right|>\epsilon\right)=0.$$

Since ϵ was arbitrary, we conclude that

$$\tilde{G}\left(x,\boldsymbol{T}^{(n)}\right)\xrightarrow{p}H(x).$$

Since x was arbitrary, the above holds for any $x\in R_f$. Combining the above result with Equation (9) and Lemma 5, we have that, for any $x\in R_f$,

$$\tilde{F}\left(x,\boldsymbol{T}^{(n)}\right)\xrightarrow{p}H(x). \tag{13}$$

To show the final result, we first note that H is strictly increasing and continuous on R_f as a result of Lemma 4, where f satisfies the conditions of Lemma 4 by Assumption 2. Therefore, by Lemma 4 and the fact that Equation (13) holds for all $x\in R_f$, $H(c^*_\alpha)=1-\alpha$ for any $\alpha\in(0,1)$. Set $0<\epsilon<\min(\alpha,1-\alpha)$. Then, $H(c^*_{\alpha+\epsilon})=1-\alpha-\epsilon$ and $H(c^*_{\alpha-\epsilon})=1-\alpha+\epsilon$. Set $\delta=\epsilon/2$ and let $\gamma>0$. Then, by Equation (13), there exists $N_1\in\mathbb{N}$ such that for all $n\geq N_1$,

$$\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|\tilde{F}\left(c^*_{\alpha+\epsilon},\boldsymbol{T}^{(n)}\right)-(1-\alpha-\epsilon)\right|\leq\delta\right)\geq 1-\gamma$$

and there exists $N_2\in\mathbb{N}$ such that for all $n\geq N_2$,

$$\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|\tilde{F}\left(c^*_{\alpha-\epsilon},\boldsymbol{T}^{(n)}\right)-(1-\alpha+\epsilon)\right|\leq\delta\right)\geq 1-\gamma.$$

Let $N = \max(N_1, N_2)$. Then, for $n \geq N$, $\tilde{F}\left(c^*_{\alpha+\epsilon}, \boldsymbol{T}^{(n)}\right)$ is contained in the interval $[1 - \alpha - \frac{3\epsilon}{2}, 1 - \alpha - \frac{\epsilon}{2}]$ and $\tilde{F}\left(c^*_{\alpha-\epsilon}, \boldsymbol{T}^{(n)}\right)$ is contained in the interval $[1 - \alpha + \frac{\epsilon}{2}, 1 - \alpha + \frac{3\epsilon}{2}]$ with probability $\geq 1 - \gamma$. Therefore, for any $n \geq N$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) \in (c^*_{\alpha+\epsilon}, c^*_{\alpha-\epsilon}]\right) \\
&\geq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left\{\tilde{F}\left(c^*_{\alpha+\epsilon}, \boldsymbol{T}^{(n)}\right) < 1 - \alpha\right\} \cap \left\{\tilde{F}\left(c^*_{\alpha-\epsilon}, \boldsymbol{T}^{(n)}\right) \geq 1 - \alpha\right\}\right) \\
&\geq \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\Big(\left\{\left|\tilde{F}\left(c^*_{\alpha+\epsilon}, \boldsymbol{T}^{(n)}\right) - (1 - \alpha - \epsilon)\right| \leq \delta\right\} \cap \left\{\left|\tilde{F}\left(c^*_{\alpha-\epsilon}, \boldsymbol{T}^{(n)}\right) - (1 - \alpha + \epsilon)\right| \leq \delta\right\}\Big) \\
&\geq 1 - \gamma
\end{aligned} \tag{14}$$

Since γ was arbitrary, for $0 < \epsilon < \min(\alpha, 1 - \alpha)$,

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) \in (c^*_{\alpha+\epsilon}, c^*_{\alpha-\epsilon}]\right) = 1.$$

Since H is strictly increasing on R_f, c^*_α is continuous for $\alpha \in (0, 1)$. Then, by the continuity of c^*_α and the fact that the above holds for any arbitrary $0 < \epsilon < \min(\alpha, 1 - \alpha)$, we conclude that

$$c_\alpha\left(\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)}_{(m-r+2:m)}\right) \xrightarrow{p} c^*_\alpha.$$

□

A.4 Other Helpful Results

Lemma 3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let A_n and B_n be a sequence of events in Ω where $\lim_{n\to\infty} \mathbb{P}(B_n) = 1$ and $\lim_{n\to\infty} \mathbb{P}(A_n) = a$ for some $a \in [0, 1]$. Then, $\mathbb{P}(A_n \cap B_n) \to a$ and $\mathbb{P}(A_n \mid B_n) \to a$.*

Proof.

$$\mathbb{P}(A_n \mid B_n) = \frac{\mathbb{P}(A_n \cap B_n)}{\mathbb{P}(B_n)} = \frac{\mathbb{P}(A_n) - \mathbb{P}(A_n \cap B_n^c)}{\mathbb{P}(B_n)},$$

where B_n^c represents the complement of B_n. Since $\mathbb{P}(B_n^c) \to 0$ by assumption,

$$\lim_{n\to\infty} \mathbb{P}(A_n \cap B_n^c) = \lim_{n\to\infty} \mathbb{P}(A_n \mid B_n^c)\,\mathbb{P}(B_n^c) \to 0.$$

So, we conclude that

$$\mathbb{P}(A_n) - \mathbb{P}(A_n \cap B_n^c) \to a,$$

and hence,

$$\mathbb{P}(A_n \mid B_n) \to a.$$

□

Lemma 4. *If $\boldsymbol{X} = (X_1, ..., X_l)$ is a vector of continuous random variables supported on $\mathbb{R}^l$ with density $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^l$ and $f : \mathbb{R}^l \to \mathbb{R}$ is continuously differentiable function with $\boldsymbol{\nabla} f \neq 0$ except on a set whose closure has measure zero, then the CDF of $f(\boldsymbol{X})$ is continuous and strictly increasing on the range of f.*

Proof. Let μ represent the Lebesgue measure on $\mathbb{R}^l$. Let $R_f = \text{Range}(f)$. Let $x \in R_f$ and $P_x \coloneqq \{\boldsymbol{y} : \boldsymbol{y} \in f^{-1}(x)\}$. Let H represent the CDF of $f(\boldsymbol{X})$. By the continuity of f and the fact that $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^l$, the CDF of $f(\boldsymbol{X})$ must be strictly increasing. Assume for the sake of contradiction that the CDF of $f(\boldsymbol{X})$ is not strictly increasing. Then, since f is continuous, R_f must be a connected set in $\mathbb{R}$, i.e., R_f must be an interval. Therefore, if the CDF of $f(\boldsymbol{X})$ is not strictly increasing, there exists some open interval $(a, b) \in R_f$ such that

$$\begin{aligned} 0 &= \mathbb{P}\left(f(\boldsymbol{X}) \in (a, b)\right) \\ &= \mathbb{P}\left(\boldsymbol{X} \in f^{-1}\left((a, b)\right)\right) \\ &= \int_{\boldsymbol{x} \in f^{-1}((a,b))} p_{\boldsymbol{X}}(\boldsymbol{x}) d\mu \end{aligned}$$

However, $\int_{\boldsymbol{x} \in f^{-1}((a,b))} p_{\boldsymbol{X}}(\boldsymbol{x}) d\mu > 0$ since $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^l$ and $f^{-1}\left((a, b)\right)$ is some open set in $\mathbb{R}^l$ (since for continuous functions, the pre-image of open sets are open sets). Therefore, the CDF of $f(\boldsymbol{X})$ must be strictly increasing.

To show continuity, we will show that, for any $x \in R_f$, $\mu(P_x) = 0$. If this is true, then, for any fixed $x \in R_f$,

$$\begin{aligned} 0 = \mathbb{P}(\boldsymbol{X} \in P_x) &= \lim_{n \to \infty} \mathbb{P}\left(\boldsymbol{X} \in f^{-1}\left((x - 1/n, x + 1/n]\right)\right) \\ &= \lim_{n \to \infty} \mathbb{P}\left(f(\boldsymbol{X}) \in (x - 1/n, x + 1/n]\right) \\ &= \lim_{n \to \infty} \mathbb{P}\left(f(\boldsymbol{X}) \leq x + 1/n\right) - \mathbb{P}\left(f(\boldsymbol{X}) \leq x - 1/n\right) \\ &= \lim_{n \to \infty} H(x + 1/n) - H(x - 1/n) \end{aligned}$$

Let $\epsilon > 0$. The last line above implies that there exists N sufficiently large such that for all $n \geq N$, $H(x + 1/n) - H(x - 1/n) < \epsilon$. So for any $0 < \delta < 1/n$, if $x - \delta < x' < x + \delta$, then

$$|H(x) - H(x')| \leq H(x + \delta) - H(x - \delta) < H(x + 1/n) - H(x - 1/n) < \epsilon$$

where the first inequality follows from the fact that H is a CDF and hence is nondecreasing. Since ϵ was arbitrary, H is continuous at x. Since the above holds for any $x \in R_f$, H is continuous at every $x \in R_f$.

We will now show that, for any $x \in R_f$, $\mu(P_x) = 0$. Fix an arbitrary $x \in R_f$. Let $Q = \{\boldsymbol{t} \in \mathbb{R}^l : f \text{ is not continuous at } \boldsymbol{t}\} \cup \{\boldsymbol{t} \in \mathbb{R}^l : f \text{ is not differentiable at } \boldsymbol{t}\} \cup \{\boldsymbol{t} \in \mathbb{R}^l : \boldsymbol{\nabla} f \text{ is not continuous at } \boldsymbol{t}\} \cup \{\boldsymbol{t} \in \mathbb{R}^l : \boldsymbol{\nabla} f(\boldsymbol{t}) = 0\}$. By assumption, the closure of Q, which we denote as $\overline{Q}$, has Lebesgue measure 0. Let $\boldsymbol{y} \in P_x \setminus \overline{Q}$, Since $\overline{Q}$ is closed, and $\boldsymbol{y}$ is in the complement of $\overline{Q}$, there must exist an open neighborhood of $\boldsymbol{y}$ that is in $\mathbb{R}^l \setminus \overline{Q}$. Also there must be at least one element of $\boldsymbol{\nabla} f(\boldsymbol{y})$ that is not zero, again, because $\boldsymbol{y}$ is in the complement of $\overline{Q}$. Without loss of generality, assume that the last element of $\boldsymbol{\nabla} f(\boldsymbol{y})$ is nonzero, i.e., $\boldsymbol{\nabla} f(\boldsymbol{y})_n \neq 0$. Then, by the Implicit Function Theorem (Theorem 3.1, Folland (2002)), there exists open sets $U_{\boldsymbol{y}} \subset \mathbb{R}^{l-1}$ and $V_{\boldsymbol{y}} \subset \mathbb{R}$ such that $(y_1, ..., y_{n-1}) \in U_{\boldsymbol{y}}$ and $y_n \in V_{\boldsymbol{y}}$ and there exists a unique function $g : U_{\boldsymbol{y}} \to \mathbb{R}$ such that

$$\begin{aligned} \{y_1, ..., y_{l-1}, g(y_1, ..., y_{l-1}) : (y_1, ..., y_{l-1}) \in U_{\boldsymbol{y}}\} &= \{\boldsymbol{y} \in U_{\boldsymbol{y}} \times V_{\boldsymbol{y}} \mid f(\boldsymbol{y}) = x\} \\ &= P_x \cap (U_{\boldsymbol{y}} \times V_{\boldsymbol{y}}) \end{aligned}$$

and g is continuously differentiable. In the above, $\times$ represents the Cartesian product. Let $W_{\boldsymbol{y}} = U_{\boldsymbol{y}} \times V_{\boldsymbol{y}}$. Note, $\{y_1, ..., y_{l-1}, g(y_1, ..., y_{l-1}) : (y_1, ..., y_{l-1}) \in U_{\boldsymbol{y}}\}$ is the graph of the continuous function $g : U_{\boldsymbol{y}} \to \mathbb{R}$, and so has Lebesgue measure 0 in $\mathbb{R}^l$ (Proposition 6.3, Lee (2003)). Therefore, $\forall \boldsymbol{y} \in P_x \setminus \overline{Q}$, there exists an open set $W_{\boldsymbol{y}} \subseteq \mathbb{R}^n$ such that

$$\mu\left(P_x \cap W_{\boldsymbol{y}}\right) = 0. \tag{15}$$

We can write

$$P_x \setminus \overline{Q} = \bigcup_{\boldsymbol{y} \in P_x \setminus \overline{Q}} \left(P_x \setminus \overline{Q}\right) \cap W_{\boldsymbol{y}} = \left(P_x \setminus \overline{Q}\right) \cap \left(\bigcup_{\boldsymbol{y} \in P_x \setminus \overline{Q}} W_{\boldsymbol{y}} \right).$$

Note, $\bigcup_{y \in P_x \setminus \overline{Q}} W_{\boldsymbol{y}}$ is a open cover of $P_x \setminus \overline{Q} \subseteq \mathbb{R}^l$. By the fact that $\mathbb{R}^l$ is second-countable (Example 1.2, Wedhorn (2016)), and any subspace of $\mathbb{R}^l$ is second-countable (Remark 1.3, Wedhorn (2016)), by Lindelöf's Covering Theorem (Theorem 15, Kelley (1955)), there exists a countable subcover of $P_x \setminus \overline{Q}$, i.e., for some countable set $A \in P_x \setminus \overline{Q}$, $P_x \setminus \overline{Q} \subseteq \bigcup_{\boldsymbol{y} \in A} W_{\boldsymbol{y}}$. Therefore,

$$P_x \setminus \overline{Q} \cap \left(\bigcup_{\boldsymbol{y} \in P_x \setminus \overline{Q}} W_{\boldsymbol{y}} \right) = P_x \setminus \overline{Q} \cap \left(\bigcup_{\boldsymbol{y} \in A} W_{\boldsymbol{y}} \right) = \bigcup_{\boldsymbol{y} \in A} \left(P_x \setminus \overline{Q}\right) \cap W_{\boldsymbol{y}}$$

where

$$\mu\left(\bigcup_{\boldsymbol{y} \in A} \left(P_x \setminus \overline{Q}\right) \cap W_{\boldsymbol{y}} \right) \leq \sum_{\boldsymbol{y} \in A} \mu\left(\left(P_x \setminus \overline{Q}\right) \cap W_{\boldsymbol{y}} \right) \leq \sum_{\boldsymbol{y} \in A} \mu\left(P_x \cap W_{\boldsymbol{y}}\right) = 0.$$

The last equality follows from Equation (15) and the fact that A is a countable set.

Therefore, $\mu\left(P_x \setminus \overline{Q}\right) = 0$ and

$$\mu(P_x) \leq \mu\left(\left(P_x \setminus \overline{Q}\right) \cup \overline{Q} \right) = 0$$

Since we picked x to be any arbitrary element of R_f, the above is true for all $x \in R_f$. □

Lemma 5. *Under Assumptions 1 and 3,*

$$\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}} \left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right) \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)} \right) \xrightarrow{p} 1.$$

Proof. Assume, without loss of generality, that $\hat{\boldsymbol{\theta}}^{(n)} = \left(0, ..., 0, \hat{\theta}^{(n)}_{(m-r+2)}, ..., \hat{\theta}^{(n)}_{(m)}\right)$. Let $B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right)$ be the event that $\left\{\tilde{T}^{(n)}_{(m-r+2)}, ..., \tilde{T}^{(n)}_{(m)}\right\} = \left\{\tilde{T}^{(n)}_{m-r+2}, ..., \tilde{T}^{(n)}_{m}\right\}$ for $\tilde{\boldsymbol{T}}^{(n)} \mid \hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}^{(n)}, I_m)$. Thus, $\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)} \sim \hat{\boldsymbol{\theta}}^{(n)}} \left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)}, \tilde{\boldsymbol{T}}^{(n)}\right) \;\middle|\; \tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)} = \boldsymbol{T}^{(n)}_{(m-r+2:m)}, \boldsymbol{T}^{(n)} \right)$ is a random variable as it is a function of $\boldsymbol{T}^{(n)}$ both through the conditioning event and through $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}$.

Let $p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{t}\,\middle|\,\boldsymbol{T}^{(n)}\right)$ be the PDF of the conditional distribution of $\tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)}$ given $\boldsymbol{T}^{(n)}$ evaluated at $\tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)}=\boldsymbol{t}\in\mathbb{R}^{r-1}$ and let $p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{t}\,\middle|\,B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right),\boldsymbol{T}^{(n)}\right)$ be the analogous PDF of the conditional distribution of $\tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)}$ given $\boldsymbol{T}^{(n)}$ and $B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)$, where $\tilde{\boldsymbol{T}}^{(n)}\mid\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}\sim\mathcal{N}(\hat{\boldsymbol{\theta}}^{(n)},I_m)$ and $\boldsymbol{T}^{(n)}\sim\mathcal{N}(\boldsymbol{\theta}^{(n)},I_m)$. Then,

$$\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\tilde{\boldsymbol{T}}^{(n)}_{(m-r+2:m)}=\boldsymbol{T}^{(n)}_{(m-r+2:m)},\boldsymbol{T}^{(n)}\right)=$$
$$=\frac{p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{T}^{(n)}_{(m-r+2:m)}\,\middle|\,B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right),\boldsymbol{T}^{(n)}\right)\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)}{p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{T}^{(n)}_{(m-r+2:m)}\,\middle|\,\boldsymbol{T}^{(n)}\right)},$$

where the numerator

$$\begin{aligned}&p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{T}^{(n)}_{(m-r+2:m)}\,\middle|\,B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right),\boldsymbol{T}^{(n)}\right)\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)\\&=p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{T}^{(n)}_{(m-r+2:m)}\,\middle|\,\boldsymbol{T}^{(n)}\right)-\\&\quad p_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)},\boldsymbol{T}^{(n)}\sim\boldsymbol{\theta}^{(n)}}\left(\boldsymbol{T}^{(n)}_{(m-r+2:m)}\,\middle|\,B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)^c,\boldsymbol{T}^{(n)}\right)\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)^c\,\middle|\,\boldsymbol{T}^{(n)}\right).\end{aligned}$$

We are about to show that $\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)\xrightarrow{p}1$, and therefore, the last line above (the term being subtracted) converges to 0 in probability, thus giving our final result.

For $K>0$, let $J_K^{(n)}=\left\{\bigcap_{i=m-r+2}^m\hat{\boldsymbol{\theta}}^{(n)}_{(i)}>K\right\}$. Set $\epsilon>0$. For any $K>0$,

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|1-\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)\right|<\epsilon\right)\\&=\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\right)\\&=\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)}\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(J_K^{(n)}\right)\\&\qquad+\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)c}\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(J_K^{(n)c}\right)\qquad(16)\end{aligned}$$

where the second line follows because $\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)$ is bounded between 0 and 1. First, note that for any $K>0$,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(J_K^{(n)}\right)&=\lim_{n\to\infty}1-\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\bigcup_{i=m-r+2}^m\hat{\boldsymbol{\theta}}^{(n)}_{(i)}\le K\right)\\&\ge1-\lim_{n\to\infty}\sum_{i=m-r+2}^m\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\hat{\boldsymbol{\theta}}^{(n)}_{(i)}\le K\right)\\&=1\qquad(17)\end{aligned}$$

where the last line follows by Assumption 3. We will now show that, for sufficiently large K,

$\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)}\right)=1$ for any n. For $K>0$, let $X_i^{(K)}\overset{i.i.d.}{\sim}\mathcal{N}(K,1)$, $i=1,...,r-1$ independent of $\boldsymbol{T}^{(n)}$. Then,

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)}\right)\\
&=\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=m-r+2,...,m}\left|\tilde{T}_i^{(n)}\right|\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)}\right)\\
&\geq\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_i^{(n)}\right|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)}\right)\\
&=\mathbb{P}\left(\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_i^{(n)}\right|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\right)>1-\epsilon\right)\\
&=\mathbb{1}\left\{\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\right)>1-\epsilon\right\}
\end{aligned}
\tag{18}
$$

where the third line holds because for independent random variables $Z_j\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$, $j=1,...,m-r+1$ and $Y_i\sim\mathcal{N}(\mu_i,1)$, $i=1,...,r-1$ where all $\mu_i>K$,

$$
\mathbb{P}\left(\max_{j=1,...,m-r+1}|Z_j|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\right)<\mathbb{P}\left(\max_{j=1,...,m-r+1}|Z_j|<\min_{i=1,...,r-1}|Y_i|\right).
$$

In the second to last line and onward, we drop the conditioning events and the $\boldsymbol{\theta}^{(n)}$ and $\hat{\boldsymbol{\theta}}^{(n)}$ subscripts because the $\tilde{T}_j^{(n)}$'s, which are i.i.d. standard normally distributed regardless of n, and the $X_i^{(K)}$'s do not depend on $\boldsymbol{T}^{(n)}$, and hence, do not depend on $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}$ and $\boldsymbol{\theta}^{(n)}$.

Let $L_{K-\sqrt{K}}=\left\{\min_{i=1,...,r-1}\left|X_i^{(K)}\right|>K-\sqrt{K}\right\}$. Then,

$$
\begin{aligned}
&\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\right)\\
&=\mathbb{P}\left(\max_{j=1,...,m-r+1}|\tilde{T}_j^{(n)}|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\,\middle|\,L_{K-\sqrt{K}}\right)\mathbb{P}\left(L_{K-\sqrt{K}}\right)\\
&\qquad+\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\,\middle|\,L^c_{K-\sqrt{K}}\right)\mathbb{P}\left(L^c_{K-\sqrt{K}}\right)
\end{aligned}
$$

where

$$
\mathbb{P}\left(L_{K-\sqrt{K}}\right)=\left(1-\Phi\left(-\sqrt{K}\right)+\Phi\left(-2K+\sqrt{K}\right)\right)^{r-1}.
$$

and

$$
\begin{aligned}
\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,...,r-1}\left|X_i^{(K)}\right|\,\middle|\,L_{K-\sqrt{K}}\right)&\geq\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<K-\sqrt{K}\,\middle|\,L_{K-\sqrt{K}}\right)\\
&=\mathbb{P}\left(\max_{j=1,...,m-r+1}\left|\tilde{T}_j^{(n)}\right|<K-\sqrt{K}\right)\\
&=\left(\Phi\left(K-\sqrt{K}\right)-\Phi\left(-(K-\sqrt{K})\right)\right)^{m-r+1}.
\end{aligned}
$$

As K gets arbitrarily large, $\left(1-\Phi\left(-\sqrt{K}\right)+\Phi\left(-2K+\sqrt{K}\right)\right)^{r-1}\to 1$ and $\left(\Phi\left(K-\sqrt{K}\right)-\Phi\left(-(K-\sqrt{K})\right)\right)^{m-r+1}\to 1$ and therefore,

$$\mathbb{P}\left(\max_{j=1,\ldots,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,\ldots,r-1}\left|X_i^{(K)}\right|\right)\to 1.$$

The above implies that there exists $K_\epsilon>0$ such that for any $K>K_\epsilon$, $\mathbb{P}\left(\max_{j=1,\ldots,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,\ldots,r-1}\left|X_i^{(K)}\right|\right)>1-\epsilon$. So, for $K>K_\epsilon$,

$$\mathbb{1}\left\{\mathbb{P}\left(\max_{j=1,\ldots,m-r+1}\left|\tilde{T}_j^{(n)}\right|<\min_{i=1,\ldots,r-1}\left|X_i^{(K)}\right|\right)>1-\epsilon\right\}=1$$

and by Equation (18),

$$\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)>1-\epsilon\,\middle|\,J_K^{(n)}\right)=1 \tag{19}$$

Setting $K>K_\epsilon$ in Equation (16), by Equation (19) and the fact that Equation (17) holds for any $K>0$, we have that

$$\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|1-\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)\right|<\epsilon\right)=0.$$

Since ϵ was arbitrary, we conclude that

$$\mathbb{P}_{\tilde{\boldsymbol{T}}^{(n)}\sim\hat{\boldsymbol{\theta}}^{(n)}}\left(B_n\left(\hat{\boldsymbol{\theta}}^{(n)},\tilde{\boldsymbol{T}}^{(n)}\right)\,\middle|\,\boldsymbol{T}^{(n)}\right)\xrightarrow{p}1.$$

□

Lemma 6. *Under Assumption 1, $\hat{\boldsymbol{\theta}}^{(n)}_{(m-r+2:m)}=\boldsymbol{T}^{(n)}_{(m-r+2:m)}$ satisfies Assumption 3.*

Proof. Assume without loss of generality that $\boldsymbol{\theta}^{(n)}=\left(0,\ldots 0,\theta^{(n)}_{m-r+2},\ldots,\theta^{(n)}_m\right)$ where $\theta_i^{(n)}\to\infty$. Fix $K>0$ and $i\in\{m-r+2,\ldots,m\}$. We will show that $\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(i)}\right|>K\right)=1$.

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(i)}\right|>K\right)&=\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(i)}\right|>K\,\middle|\,B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)\right)+\\
&\qquad\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(i)}\right|>K\,\middle|\,B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)^c\right)\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)^c\right)\\
&=\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(i)}\right|>K\,\middle|\,B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)\right)\\
&\geq\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T^{(n)}_{(m-r+2)}\right|>K\,\middle|\,B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)\right)\\
&=\lim_{n\to\infty}\mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\bigcap_{i=m-r+2}^{m}\left\{\left|T_i^{(n)}\right|>K\right\}\,\middle|\,B_n\left(\boldsymbol{\theta}^{(n)},\boldsymbol{T}^{(n)}\right)\right)\\
&=1,
\end{aligned}$$

where the third line follows from the fact that $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)\right) = 1$ as shown in Equation (8) of Lemma 1, the fifth line follows because of the conditioning on $B_n\left(\boldsymbol{\theta}^{(n)}, \boldsymbol{T}^{(n)}\right)$, and the last line follows from applying Lemma 3 with Equation (6) and Equation (8) of Lemma 1. Since we picked an arbitrary i and K, we can conclude that for any $K > 0$ and $i = m - r + 2, ..., m$, $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}^{(n)}}\left(\left|T_{(i)}^{(n)}\right| > K\right) = 1$. □

B Further Computational Details

B.1 Further Details on Computing cPCH p-values

Recall from Section 2.5 that the cPCH p-value can be represented as:

$$\begin{aligned}
&\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \;\middle|\; \boldsymbol{T}_{(m-r+2:m)}\right) \\
&= \sum_{\ell=1}^{\frac{m!}{(m-r+1)!}} \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell \;\middle|\; \boldsymbol{T}_{(m-r+2:m)}\right) \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \;\middle|\; B_\ell, \boldsymbol{T}_{(m-r+2:m)}\right).
\end{aligned}$$

where we let $\hat{\boldsymbol{\theta}} = (0, ..., 0, T_{(m-r+2)}, ..., T_{(m)})$. Computing this sum can be split into two parts:

1. Deriving analytic expressions of the mixture weights $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell \;\middle|\; \boldsymbol{T}_{(m-r+2:m)}\right)$;
2. Computing the mixture components by sampling from the distribution of $f\left(\boldsymbol{T}_{(1:m-r+1)}\right) \mid B_\ell, \boldsymbol{T}_{(m-r+2:m)}$.

We now provide the computational details for the above parts. Without loss of generality, we illustrate using $S_1 = (\{T_1, ..., T_{m-r+1}\}, \boldsymbol{T}_{m-r+2:m})$ and

$$B_1 = \left\{(S_1 = (\{T_{(1)}, ..., T_{(m-r+1)}\}, \boldsymbol{T}_{(m-r+2:m)})\right\}.$$

Analytical expressions of the mixture weights.

As mentioned in Section 2.5, we can express the mixture weights $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell \;\middle|\; \boldsymbol{T}_{(m-r+2:m)}\right)$ using only evaluations of the standard normal cumulative distribution Φ and density ϕ. For example, given the observed $\boldsymbol{T}_{(m-r+2:m)} = \boldsymbol{t}_{m-r+2:m}$, we have [4]

$$\begin{aligned}
&\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_1 \;\middle|\; \boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right) \\
&= \frac{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_1, \boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right)}{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(\boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right)} \\
&= \frac{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|T_1| < |t_{m-r+2}|, ..., |T_{m-r+1}| < |t_{m-r+2}|, T_{m-r+2} = t_{m-r+2}, ..., T_m = t_m\right)}{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(\boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right)} \\
&= \frac{\prod_{h=1}^{m-r+1} \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|T_h| < |t_{m-r+2}|\right) \prod_{j=m-r+2}^{m} \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_j = t_j\right)}{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(\boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right)},
\end{aligned} \tag{20}$$

[4] We employ a slight abuse of notation by writing $\mathbb{P}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m})$ to denote the PDF of $\boldsymbol{T}_{(m-r+2:m)}$ evaluated at $t_{m-r+2:m}$ and $\mathbb{P}_{\hat{\boldsymbol{\theta}}}(T_j = t_j)$ to denote the PDF of T_j evaluated at t_j, both under the model $\boldsymbol{T} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, I_m)$.

where

$$\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|T_h| < |t_{m-r+2}|\right) = \Phi(|t_{m-r+2}|) - \Phi(-|t_{m-r+2}|), h = 1, ..., m-r+1,$$
$$\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_j = t_j\right) = \phi(t_j - \hat{\theta}_j), j = m-r+2, ..., m,$$
$$\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(\boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right) = \sum_{\ell=1}^{\frac{m!}{(m-r+1)!}} \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_\ell, \boldsymbol{T}_{(m-r+2:m)} = t_{m-r+2:m}\right).$$

The summands in the last line have the same form as the numerator in Equation (20), and hence can be computed in the same way, although the means of $T_{1:m-r+1}$ above are all 0 since we calculate the above probability for B_1. The means will be different for different B_ℓ since, conditional on B_ℓ, we assume a particular subset of $T_{1:m}$ correspond to $T_{(1:m-r+1)}$.

Distributions of the mixture components.

By the conditioning on B_1, $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \mid B_1, \boldsymbol{T}_{(m-r+2:m)}\right) =$

$\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{1:m-r+1}) \geq f_{\text{obs}} \mid B_1, \boldsymbol{T}_{(m-r+2:m)}\right)$ for any f that is permutation invariant (e.g., Fisher's, Simes', or Bonferroni's). Note, we can exactly specify the distribution of $T_1, .., T_{m-r+1}$ conditional on B_1 and $\boldsymbol{T}_{(m-r+2:m)}$:

$$T_1, ..., T_{m-r+1} \mid B_1, \boldsymbol{T}_{(m-r+2:m)} \overset{i.i.d.}{\sim} \text{Trunc–Norm}(0, 1, T_{(m-r+2)}),$$

where Trunc–Norm$(\mu, 1, t)$ is the truncated normal distribution with location μ and scale 1 truncated at $|t|$ and $-|t|$. Therefore, we can utilize standard sampling procedures for truncated normal distributions to generate N independent copies $\left\{\tilde{\boldsymbol{T}}_1^{(k)}\right\}_{k=1}^N$ from the distribution of $T_1, ..., T_{m-r+1} \mid B_1, \boldsymbol{T}_{(m-r+2:m)}$, where the "1" subscript on $\tilde{\boldsymbol{T}}_1^{(k)}$ denotes the conditioning on B_1. Let $X_1^{(k)} = f\left(\tilde{\boldsymbol{T}}_1^{(k)}\right)$. Then, we estimate $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) \geq f_{\text{obs}} \mid B_1, \boldsymbol{T}_{(m-r+2:m)}\right)$ using

$$g\left(f_{\text{obs}}, \left\{X_1^{(k)}\right\}_{k=1}^N\right) = \frac{1}{N+1}\left(1 + \sum_{k=1}^N \mathbb{1}\left\{X_1^{(k)} \geq f_{\text{obs}}\right\}\right).$$

We can apply this logic to each B_ℓ to get estimates for each mixture component.

B.1.1 Calculating cPCH p-values in the $r = m = 2$ setting

In the $r = m = 2$ setting, cPCH p-values can be calculated without sampling using the following derivation. Since $r = m = 2$, $S = \{\{\{T_1\}, T_2\}, \{\{T_2\}, T_1\}\}$. As in Section 2.5, let S_ℓ be the ℓ^{th} set in S and

$$B_\ell = \{S_\ell = (\{T_{(1)}\}, T_{(2)})\}.$$

Then, given the observed $T_{(1)} = f^{\text{obs}}$,

$$\begin{aligned}\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} > f^{\text{obs}} \mid T_{(2)}\right) = \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_1 \mid T_{(2)}\right)\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} \geq f^{\text{obs}} \mid T_{(2)}, B_1\right) \\ + \mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_2 \mid T_{(2)}\right)\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} \geq f^{\text{obs}} \mid T_{(2)}, B_2\right), \qquad (21)\end{aligned}$$

where $\hat{\boldsymbol{\theta}} = (0, T_{(2)})$. The mixture weights $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(B_i \mid T_{(2)}\right)$ can be calculated as in Equation (20). For the $r = m = 2$ setting, we can derive the analytic form of the mixture components $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} \geq f^{\text{obs}} \mid T_{(2)}, B_1\right)$ as well. For example, conditional on B_1,

$$\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} \geq f^{\text{obs}} \,\middle|\, T_{(2)}, B_1\right) = \frac{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|f^{\text{obs}}| \leq |T_1| \leq |T_{(2)}| \,\middle|\, T_{(2)}\right)}{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|T_1| \leq |T_{(2)}| \,\middle|\, T_{(2)}\right)} = \frac{2\left(\Phi\left(|T_{(2)}|\right) - \Phi\left(|f^{\text{obs}}|\right)\right)}{\Phi(|T_{(2)}|) - \Phi(-|T_{(2)}|)}. \tag{22}$$

Analogously, conditional on B_2, we get

$$\begin{aligned}
&\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} \geq f^{\text{obs}} \,\middle|\, T_{(2)}, B_2\right) \\
&= \frac{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|f^{\text{obs}}| \leq |T_2| \leq |T_{(2)}| \,\middle|\, T_{(2)}\right)}{\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(|T_2| \leq |T_{(2)}| \,\middle|\, T_{(2)}\right)} \\
&= \frac{\left(\Phi\left(|T_{(2)}| - \hat{\theta}_2\right) - \Phi\left(|f^{\text{obs}}| - \hat{\theta}_2\right)\right) + \left(\Phi\left(-|f^{\text{obs}}| - \hat{\theta}_2\right) - \Phi\left(-|T_{(2)}| - \hat{\theta}_2\right)\right)}{\Phi(|T_{(2)}| - \hat{\theta}_2) - \Phi(-|T_{(2)}| - \hat{\theta}_2)}.
\end{aligned} \tag{23}$$

Plugging in Equation (22) and (23) to Equation (21) gives the final form for $\mathbb{P}_{\hat{\boldsymbol{\theta}}}\left(T_{(1)} \geq f^{\text{obs}} \,\middle|\, T_{(2)}\right)$. Note, the same strategy could be applied for $r = m$, but we do not pursue this idea here.

B.2 The SGD Algorithm for Quantifying Maximum Type I Error

In this section, we give the computation details of SGD analysis for Type I error inflation. Recall the definition of $E(\boldsymbol{\theta})$ in Equation (2),

$$\begin{aligned}
E(\boldsymbol{\theta}) &= \mathbb{P}_{\boldsymbol{\theta}}\left(f(\boldsymbol{T}_{(1:m-r+1)}) > c_\alpha\left(\boldsymbol{T}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)\right) \\
&= \int_{\mathbb{R}^m} \left(\mathbb{1}\left\{ f(\boldsymbol{T}_{(1:m-r+1)}) > c_\alpha\left(\boldsymbol{T}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)\right\} \prod_{i=1}^{m} \phi(T_i - \theta_i) \right) dT_i.
\end{aligned}$$

We simplify the notation of the cPCH test in Definition 2 and write it as a function of the data $\boldsymbol{T}$ i.e., $\varphi_\alpha^{\text{cPCH}}(\boldsymbol{T}) := \mathbb{1}\left\{ f(\boldsymbol{T}_{(1:m-r+1)}) > c_\alpha\left(\boldsymbol{T}_{(m-r+2:m)}, \boldsymbol{T}_{(m-r+2:m)}\right)\right\}$. Expressing $E(\boldsymbol{\theta})$ in terms of the integral

$$E(\boldsymbol{\theta}) = \int_{\mathbb{R}^m} \psi(\boldsymbol{t}) \prod_{i=1}^{m} \phi(t_i - \theta_i) dt_i$$

and noticing the fact that $\frac{d}{d\theta_i}\phi(t_i - \theta_i) = -(t_i - \theta_i)\phi(t_i - \theta_i)$, we have the gradient of $E(\boldsymbol{\theta})$ equals

$$\nabla E(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{T} - \boldsymbol{\theta})\psi(\boldsymbol{T})] = \mathbb{E}[\boldsymbol{Z}\psi(\boldsymbol{Z} + \boldsymbol{\theta})],$$

where $\boldsymbol{Z} = (\boldsymbol{T} - \boldsymbol{\theta}) \overset{d}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. It can be unbiasedly estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}_i \psi(\boldsymbol{Z}_i + \boldsymbol{\theta})$$

with n i.i.d. samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ and we can choose the number of repeated samples n to be sufficiently large.

Due to the symmetry of the problem, we note $\max_{\boldsymbol{\theta}\in\Theta_0^{r/m}} E(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\Theta_0^{\text{cone}}} E(\boldsymbol{\theta})$ where $\Theta_0^{\text{cone}} = \{\boldsymbol{\theta} \in \Theta_0^{r/m} : \theta_1 \geq \cdots \theta_{r-1} \geq 0 = \theta_r = \cdots = \theta_m\}$ is a convex cone and its volume is only $\frac{1}{\binom{m}{r-1}(r-1)!2^{r-1}}$ of $\Theta_0^{r/m}$. Therefore, restricting $\boldsymbol{\theta}$ in the much smaller searching space Θ_0^{cone} will speed up the convergence to a local maximum. With such a constrained parameter space of $\boldsymbol{\theta}$, we project via isotonic regression (Han et al., 2017) to ensure the move at each step is still within Θ_0^{cone}. We also make use of random initializations to help adequately search the parameter space for estimating the global maximum $\sup_{\boldsymbol{\theta}\in\Theta_0^{r/m}} E(\boldsymbol{\theta})$. The whole procedure is spelled out in Algorithm 1. For the results in Table 1, we use an exponentially decaying learning rate and terminate the algorithm after 200 batches.

Algorithm 1 SGD estimation of the cPCH test's maximum Type I error

Input: m, r, f, N, a level α, the number of repeated samples n, the maximum number of batches T, the learning rate γ.

1: draw θ_1 from $\mathcal{N}(0, \sigma^2)$ for some σ, then draw θ_k from Trunc–Norm$(0, \sigma^2, |\theta_{k-1}|)$ for $k = 2, \cdots, r-1$; set θ_k to be its absolute value when $k < r$ and 0 otherwise.
2: **repeat**
3: sample i.i.d. samples $\{\boldsymbol{Z}_i\}_{i=1}^n$ from $\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$.
4: run conditional PCH testing on data $\{\boldsymbol{Z}_i + \boldsymbol{\theta}\}_{i=1}^n$ and obtain $\{\psi(\boldsymbol{Z}_i + \boldsymbol{\theta})\}_{i=1}^n$.
5: compute the gradient estimate: $\boldsymbol{g} \leftarrow n^{-1} \sum_{i=1}^n \boldsymbol{Z}_i \psi(\boldsymbol{Z}_i + \boldsymbol{\theta})$.
6: update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \gamma \boldsymbol{g}$
7: reset $\boldsymbol{\theta}$ to be the projection of $\boldsymbol{\theta}$ on Θ_0^{cone}.
8: **until** t = T or reaching the stopping criterion

Output: estimate of the maximum Type I error: $n^{-1} \sum_{i=1}^n \psi(\boldsymbol{Z}_i + \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta_0^{r/m}$.

C Further Details on Section 3

C.1 Further Details on Methods under Comparison in Section 3.3

For the PCH multiple testing approaches under consideration (DACT, HDMT, AdaFilter, and DHH), all FDR guarantees hold under the assumption that the p_{ij} are independent, or some mild relaxation of it (e.g., AdaFilter allows weak dependence between the base p-values for their asymptotic results). Both DACT and HDMT guarantee FDR control as $M \to \infty$ under specific regularity conditions, which are primarily set to guarantee that the estimators for the proportions of each null configuration are consistent. By the linear structural model commonly used for causal mediation analysis (Baron and Kenny, 1986), both DACT and HDMT assume that the base test statistics are normally distributed and independent across m.

Under mild regularity conditions, AdaFilter guarantees FDR control as $M \to \infty$. For finite M and nominal FDR level q, AdaFilter guarantees FDR control at level $qC(M)$ where

$C(M) = \sum_{j=1}^{M} \frac{1}{j}$, when all p_{ij} are independent.

DHH guarantees FDR control for finite M when the filtering threshold τ is fixed and the PCH p-value $p^{r/m}$ used has *uniform validity*: for any $\boldsymbol{\theta} \in \Theta_0^{r/m}$ and $\alpha \in [0, \tau]$,

$$\frac{\mathbb{P}_{\boldsymbol{\theta}}\left(p^{r/m} < \alpha\right)}{\mathbb{P}_{\boldsymbol{\theta}}\left(p^{r/m} < \tau\right)} \leq \frac{\alpha}{\tau}.$$

They show that the standard Fisher, Simes, and Bonferroni PCH p-values satisfy uniform validity. Dickhaus et al. (2021) also provides an algorithm for selecting the threshold τ in a data-adaptive way, which, under certain conditions, guarantees asymptotic FDR control. We select $\tau = 0.1$ since, as shown in the empirical simulations and real data example presented in Dickhaus et al. (2021), DHH with $\tau = 0.1$ has similar power to DHH with the data-adaptive threshold in many settings.

Finally, we discuss the assumptions necessary for FDR control for BH (Benjamini and Hochberg, 1995), Storey's procedure (Storey, 2002), and AdaPT–GMM (Chao and Fithian, 2021), the multiple testing procedures used in combination with individual PCH p-values. Storey's procedure and AdaPT–GMM share the assumption that the individual PCH p-values are independent, while BH only requires positive regression dependency on a subset. AdaPT–GMM requires the additional assumption that the null p-values have a non-decreasing density. The standard PCH p-values and DHH-adjusted PCH p-values satisfy this condition, while DACT and cPCH p-values are not guaranteed to do so. However, since null cPCH p-values are *nearly* uniform, the non-decreasing density assumption is at least justified approximately for the cPCH test. Additionally, we find in our empirical simulations that FDR control is maintained when using cPCH p-values with AdaPT–GMM; see Figure 6 in Section 3.3.2.

C.2 Additional Simulation Results

C.2.1 Type I error of single PCH testing results for $m = 4$

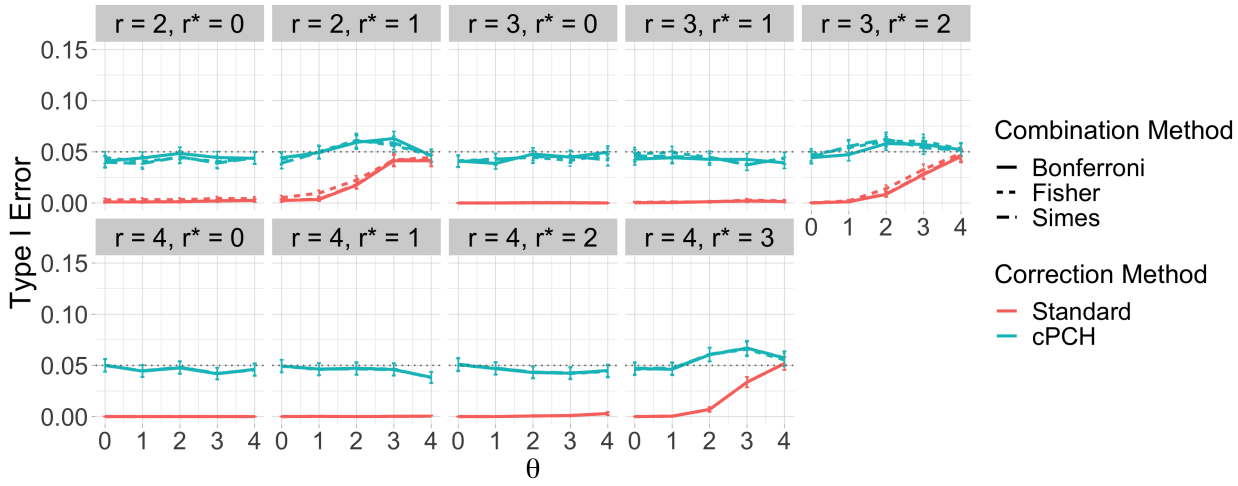


Figure 8: Type I error of the cPCH test across all null cases ($r^\star < r$) at nominal level $\alpha = 0.05$ (dotted black line) for $m = 4$. Recall that the Type I error of the cPCH Oracle test is exactly equal to the nominal level. Each point represents the proportion of cPCH p-values below α over 5000 replicates of the data generating procedure outlined in Section 3.1 for each $(r^\star, r, \theta)$ triplet. Error bars depict ± 2 standard errors.

C.2.2 Additional single PCH testing results for various m

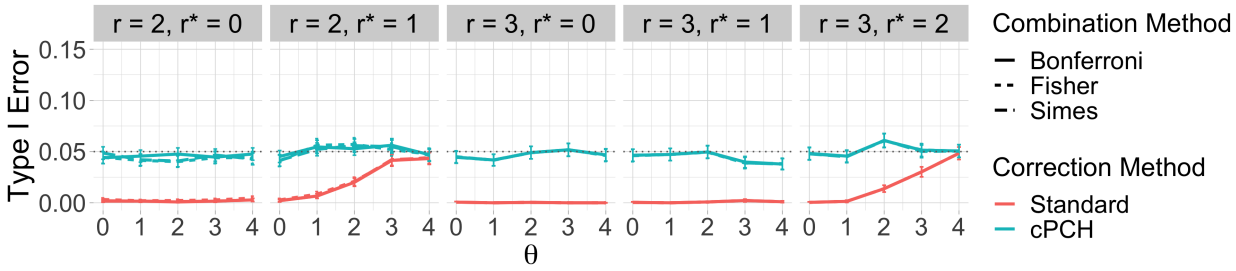


Figure 9: Type I error of the cPCH test across all null cases ($r^\star < r$) at nominal level $\alpha = 0.05$ (dotted black line) for $m = 3$. Recall that the Type I error of the cPCH Oracle test is exactly equal to the nominal level. Each point represents the proportion of cPCH p-values below α over 5000 replicates of the data generating procedure outlined in Section 3.1 for each $(r^\star, r, \theta)$ triplet. Error bars depict 2 standard errors.

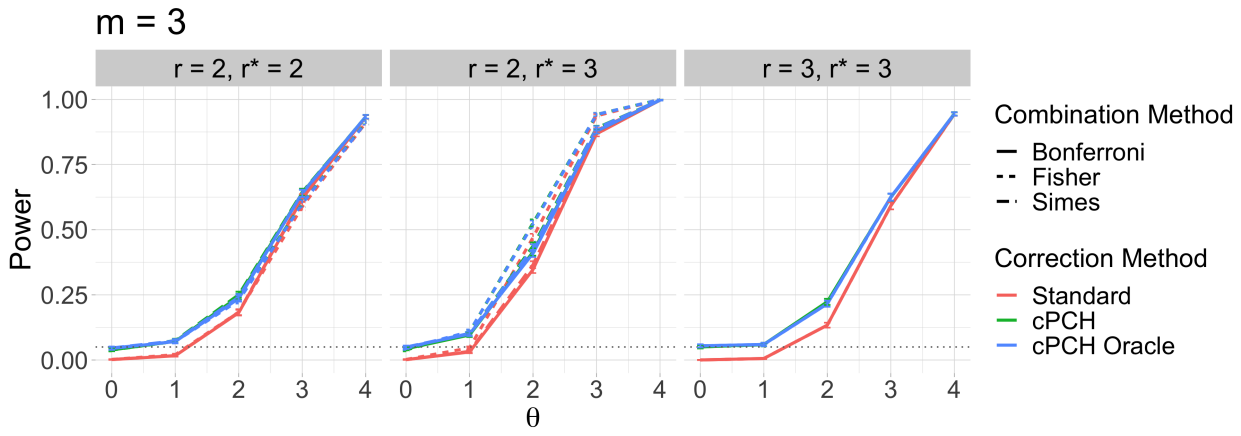


Figure 10: Power of the cPCH test across all alternative cases ($r^\star \geq r$) at nominal level $\alpha = 0.05$ (dotted black line) for $m = 3$. Each point represents the proportion of cPCH p-values below α over 5000 replicates of the data generating procedure outlined in Section 3.1 for each $(r^\star, r, \theta)$ triplet. Error bars depict 2 standard errors.

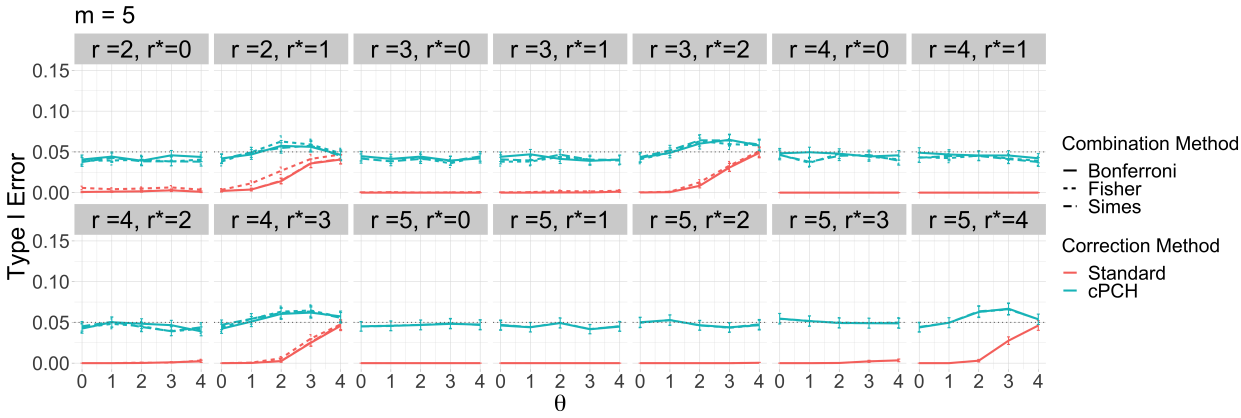


Figure 11: Type I error of the cPCH test across all null cases ($r^\star < r$) at nominal level $\alpha = 0.05$ (dotted black line) for $m = 5$. Recall that the Type I error of the cPCH Oracle test is exactly equal to the nominal level. Each point represents the proportion of cPCH p-values below α over 5000 replicates of the data generating procedure outlined in Section 3.1 for each $(r^\star, r, \theta)$ triplet. Error bars depict ± 2 standard errors.

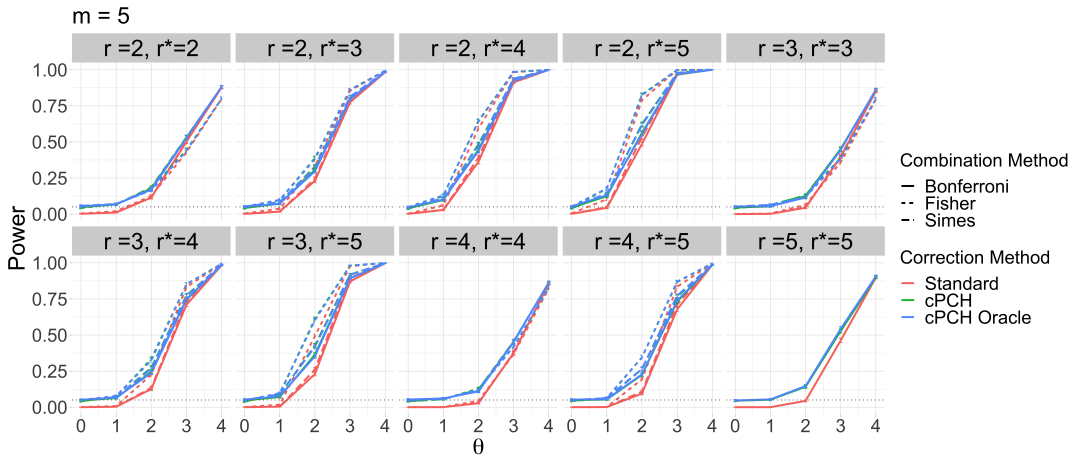


Figure 12: Power of the cPCH test across all alternative cases ($r^\star \geq r$) at nominal level $\alpha = 0.05$ (dotted black line) for $m = 5$. Each point represents the proportion of cPCH p-values below α over 5000 replicates of the data generating procedure outlined in Section 3.1 for each $(r^\star, r, \theta)$ triplet. Error bars depict ± 2 standard errors.

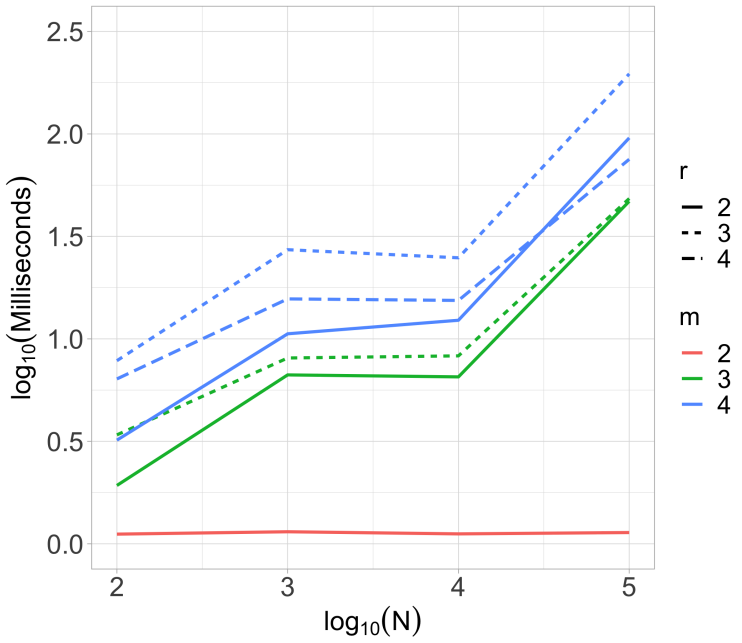


Figure 13: Each point represents the $\log_{10}$ of the average time in milliseconds to compute a single cPCH p-value (as described in Section 2.5) with Fisher's combining function over 100 replicates for each combination of m, r, and N. We use Fisher's combining function since we found that the computation times were similar across Fisher's, Simes', and Bonferroni's combining functions. The computation times for $r = m = 2$ are especially small (≈ 1 millisecond) because we are able to compute cPCH p-values in this case analytically using only evaluations of the standard normal CDF and PDF, which can be computed very efficiently; see Appendix B.1 for further details.

C.2.3 qq-plot Results for $r = m = 3$

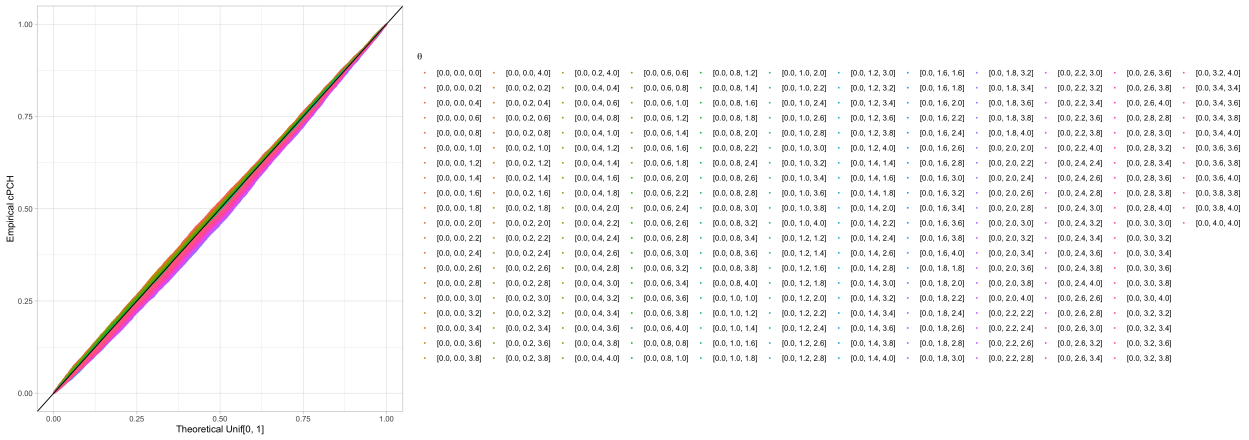


Figure 14: qq-plot of the empirical cPCH p-value density under various $\boldsymbol{\theta} \in \Theta_0^{3/3}$ compared with the theoretical Unif[0, 1] distribution. Recall that when $r = m$, Bonferroni's, Simes', and Fisher's combining functions are all equivalent. Each line represents the matched quantiles of the Unif[0, 1] density (x–coordinate) and the empirical cPCH p-value density (y–coordinate) for a given $\boldsymbol{\theta}$ estimated using $10,000$ independent replicates for the Monte Carlo sampling scheme described in Section 2.4.2.

C.2.4 Additional PCH Multiple Testing Results

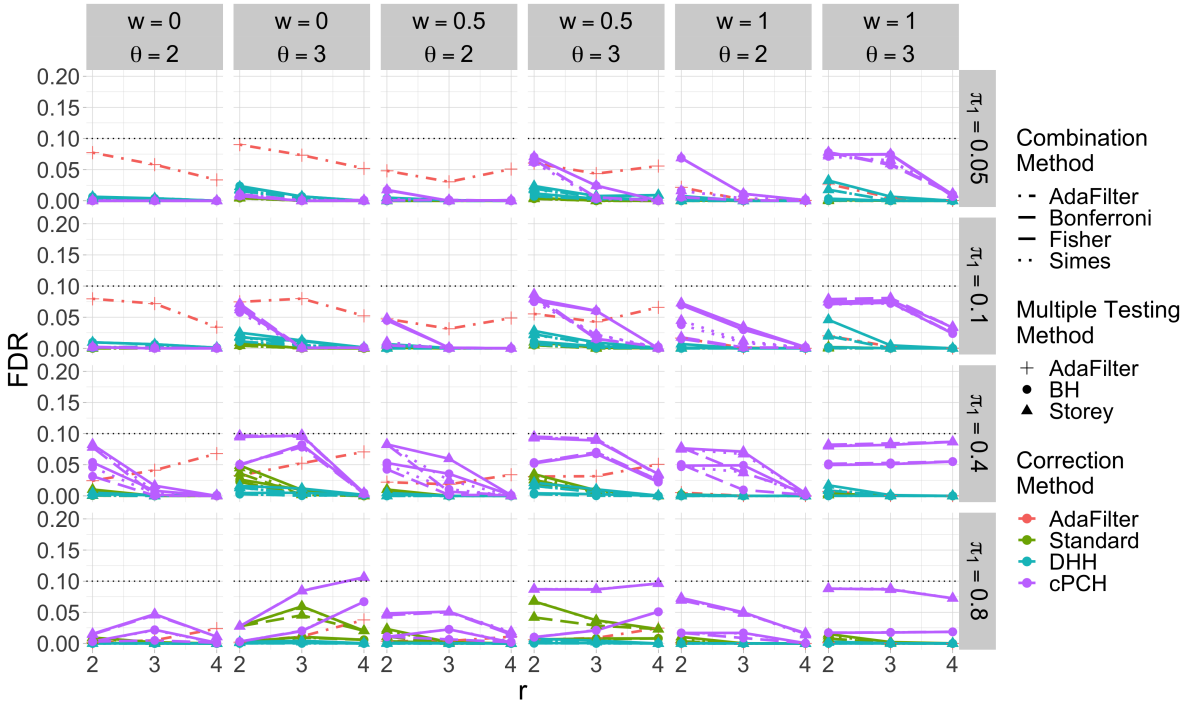
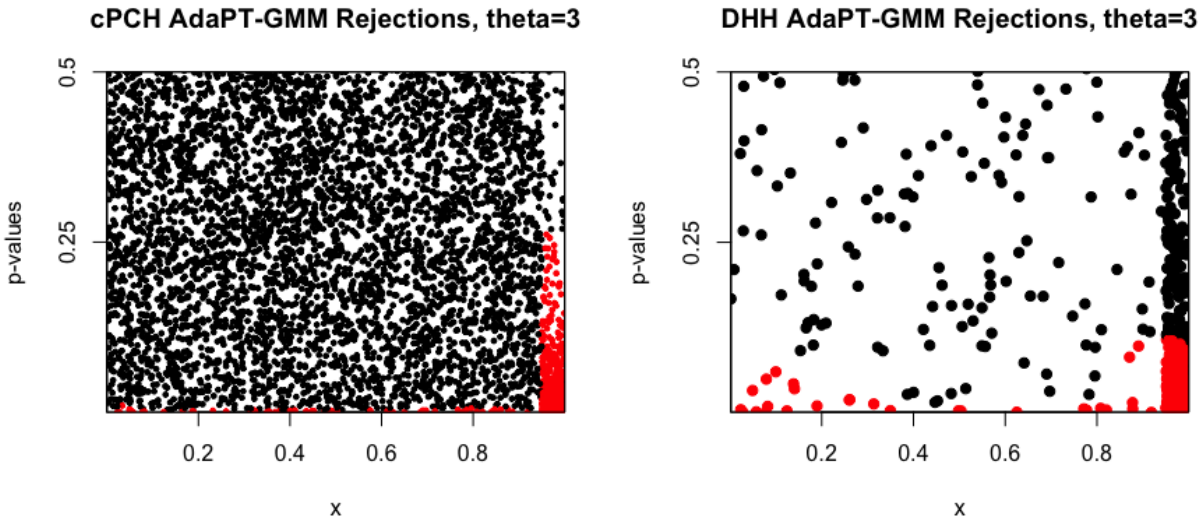


Figure 15: False Discovery Rate (FDR) results corresponding to Section 3.3.3 where all methods are implemented at nominal level $q = 0.1$ (dotted black line). Each point represents the average proportion of rejected PCH's which are null over 1000 independent replicates of the data generating procedure described in Section 3.3.3 for a given θ. Standard errors were all less than 0.008.



(a) (b)

Figure 16: Each point represents a (covariate, p-value) pair generated using the data generating scheme described in Section 3.3.2 with $\theta = 3$. The p-values rejected by AdaPT–GMM are in red. There are much fewer DHH p-values than cPCH p-values because AdaPT–GMM is applied to the DHH p-values after filtering. As indicated by the pattern of rejections, cPCH with AdaPT–GMM easily detects that the covariate being above 0.95 is highly informative of the PCH being in the alternative, which accurately reflects the true data generating procedure. DHH with AdaPT–GMM does not capture this trend as clearly.

C.2.5 AdaPT–GMM Rejections for cPCH and DHH

C.2.6 Additional Simulations Comparing cPCH with Empirical Bayes Methods

We first compare the Type I error of the cPCH test in the $r = m = 2$ setting with the Empirical Bayes methods DACT (Liu et al., 2022). We focus on the $r = m = 2$ setting since DACT is only applicable in this setting and we exclude HDMT for our Type I error analysis because HDMT does not produce individual PCH p-values.

We simulated data using the following $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta_0^{r/m}$: $(0,0), (0,1), (0,2), (0,3), (0,4)$.

For each choice of $\boldsymbol{\theta} \in \Theta_0^{r/m}$ above, we independently draw $T_1 \sim \mathcal{N}(\theta_{1j}, 1)$ and $T_2 \sim \mathcal{N}(\theta_{2j}, 1), j = 1, ..., M$ where we set $M = 100, 1000, 5000$.

The cPCH p-values are *i.i.d.* across the M draws, so any variation in the estimated Type I error for the cPCH test across M only occurs as a result of the random sampling. Since DACT controls Type I error asymptotically, we expect the estimated Type I error of DACT to vary with M not just as a result of the random sampling, but also as a result of DACT's asymptotic Type I error control guarantees.

Table 5 shows the estimated Type I error of the cPCH test, DACT, and Max-P test at nominal level $\alpha = 0.1$ over 100 independent replicates of the above data generating procedure. Note, DACT does not have Type 1 error control in the $(0, 4)$ case regardless of M.

$\boldsymbol{\theta}$		(0,0)	(0,1)	(0,2)	(0,3)	(0,4)
M = 100	cPCH	0.091	0.107	0.112	0.109	0.103
	Max-P	0.009	0.026	0.063	0.096	0.101
	DACT	0.098	0.071	0.052	0.107	0.148
M = 1000	cPCH	0.092	0.104	0.113	0.108	0.101
	Max-P	0.010	0.026	0.063	0.093	0.099
	DACT	0.098	0.059	0.041	0.096	0.177
M = 5000	cPCH	0.091	0.097	0.105	0.101	0.010
	Max-P	0.010	0.026	0.064	0.091	0.099
	DACT	0.097	0.050	0.028	0.078	0.238

Table 5: Type I Error estimated over 100 independent replicates of the data generating process described in Appendix C.2.6 using level $\alpha = 0.1$.

In a multiple testing setting, we compared the FDR and power of the cPCH test in combination with BH, DACT in combination with BH, and HDMT. For the $r = m = 2$ case there are only three possible configurations:

1. Both θ_1 and θ_2 are non-zero.
2. One of θ_1, θ_2 is zero and the other is non-zero.
3. Both θ_1 and θ_2 are zero.

with configurations 1 and 2 belonging to the null, and configuration 3 belonging to the alternative. Let $\pi_{00}, \pi_{01}, \pi_{11}$ be the true proportion of PCH's in each of the three configurations respectively across M total PCH's. As in Dai et al. (2020), we simulate the data

$(T_{1,j}, T_{2,j})_{j=1}^{M}$ where $\pi_{11}M$ of the total M pairs $(T_{1,j}, T_{2,j})$ are generated using $T_{1,j} \sim \mathcal{N}(\theta, 1)$, $T_{2,j} \sim \mathcal{N}(\theta, 1)$, $\pi_{01}M$ of the total M pairs are generated using $T_{1,j} \sim \mathcal{N}(0, 1)$, $T_{2,j} \sim \mathcal{N}(\theta, 1)$ and $\pi_{00}M$ of the total M pairs are generated using $T_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), i = 1, 2$. We set the following configurations for π_{00}, π_{01} and π_{11},

- Dense Null: $\pi_{00} = 0.6, \pi_{01} = 0.4, \pi_{11} = 0$
- Sparse Null: $\pi_{00} = 0.9, \pi_{01} = 0.1, \pi_{11} = 0$
- Complete Null: $\pi_{00} = 1, \pi_{01} = 0, \pi_{11} = 0$
- Dense Alternative: $\pi_{00} = 0.4, \pi_{01} = 0.4, \pi_{11} = 0.2$
- Sparse Alternative: $\pi_{00} = 0.88, \pi_{01} = 0.1, \pi_{11} = 0.02$

cPCH with BH, DACT with BH, and HDMT are all performed at nominal FDR level $q = 0.1$. Table 6 shows the estimated FDR and power results across the different alternative scenarios and Table 7 shows the estimated FDR results across the different null scenarios for $M = 100, 500, 5000$ and $\theta = 2$. As shown by the highlighted rows, the DACT and HDMT methods do not reliably control FDR across various settings.

		FDR	Power
Dense alternative	$\pi_{11} = 0.2$		
M = 100	cPCH-BH	0.094	0.038
	DACT	0.036	0.075
	HDMT	0.107	0.172
	MaxP-BH	0.007	0.014
M = 1000	cPCH-BH	0.080	0.008
	DACT	0.076	0.111
	HDMT	0.086	0.117
	MaxP-BH	3e-04	0.002
M = 5000	cPCH-BH	0.085	0.003
	DACT	0.142	0.220
	HDMT	0.086	0.111
	MaxP-BH	0.000	2e-04
Sparse alternative	$\pi_{11} = 0.02$		
M = 100	cPCH-BH	0.073	0.023
	DACT	0.174	0.17
	HDMT	0.159	0.196
	MaxP-BH	0.001	0.01
M = 1000	cPCH-BH	0.091	0.004
	DACT	0.304	0.200
	HDMT	0.119	0.077
	MaxP-BH	0.000	0.001
M = 5000	cPCH-BH	0.0907	0.001
	DACT	0.435	0.269
	HDMT	0.096	0.044
	MaxP-BH	0.000	1e-04

Table 6: FDR estimated over 1000 independent replicates of the data generating process described in Appendix C.2.6 for various M using level $q = 0.1$. BH was used with the cPCH and the Max-P p-values. All SEs were ≤ 0.004 for $M = 5000$, ≤ 0.009 for $M = 1000$, and ≤ 0.010 for $M = 100$. FDR values above $q + 2$SE's are highlighted.

		FDR
Dense null	$\pi_{00} = 0.6$	
M = 100	cPCH-BH	0.111
	DACT	0.252
	HDMT	0.136
	MaxP-BH	0.005
M = 1000	cPCH-BH	0.101
	DACT	0.712
	HDMT	0.099
	MaxP-BH	0.001
M = 5000	cPCH-BH	0.11
	DACT	0.982
	HDMT	0.104
	MaxP-BH	0
Sparse null	$\pi_{00} = 0.9$	
M = 100	cPCH-BH	0.097
	DACT	0.214
	HDMT	0.196
	MaxP-BH	0.001
M = 1000	cPCH-BH	0.09
	DACT	0.392
	HDMT	0.159
	MaxP-BH	0.000
M = 5000	cPCH-BH	0.104
	DACT	0.728
	HDMT	0.123
	MaxP-BH	0.000
Complete null	$\pi_{00} = 1$	
M = 100	cPCH-BH	0.083
	DACT	0.094
	HDMT	0.100
	MaxP-BH	0.000
M = 1000	cPCH-BH	0.088
	DACT	0.091
	HDMT	0.092
	MaxP-BH	0.000
M = 5000	cPCH-BH	0.098
	DACT	0.127
	HDMT	0.092
	MaxP-BH	0.000

Table 7: FDR and power estimated over 1000 independent replicates of the data generating process described in Appendix C.2.6 for various M using level $q = 0.1$. All SEs were ≤ 0.007 for $M = 5000$, ≤ 0.015 for $M = 1000$, and ≤ 0.013 for $M = 100$. FDR values above $q+2$SE's are highlighted.

D Further Details on DMD Data Analysis

D.1 Standard PCH Testing Analysis of DMD data

Single PCH Test	f	MT Procedure	$r = 2$	$r = 3$	$r = 4$
cPCH	Fisher	Storey	497	193	15
cPCH	Fisher	BH	425	167	12
cPCH	Simes	Storey	409	165	15
cPCH	Simes	BH	340	114	12
cPCH	Bonferroni	Storey	400	141	15
cPCH	Bonferroni	BH	336	109	12
Standard	Fisher	Storey	364	128	9
Standard	Fisher	BH	359	128	9
Standard	Simes	Storey	364	128	9
Standard	Simes	BH	359	128	9
Standard	Bonferroni	Storey	364	128	9
Standard	Bonferroni	BH	359	128	9
Standard	Fisher	DHH-BH	284	115	9
Standard	Fisher	DHH-Storey	382	151	21
Standard	Simes	DHH-BH	284	115	9
Standard	Simes	DHH-Storey	382	151	21
Standard	Bonferroni	DHH-BH	284	115	9
Standard	Bonferroni	DHH-Storey	382	151	21
		AdaFilter	380	217	73

Table 8: The number of rejections for each method across the $M = 1871$ unique genes shared among the $m = 4$ DMD studies described in Section 4, all at nominal FDR level $q = 0.1$. AdaFilter and DHH–Storey tend to outperforms the cPCH test when $r = 3, 4$ while the cPCH-based methods tend to outperform AdaFilter, DHH–Storey, and DHH–BH when $r = 2$.

D.2 Discovered Genes from Follow-up Analysis

Gene Symbol	cPCH p-value	Gene Function
ART3	0.00011	protein ADP-ribosylation
CFHR1	0.00075	complement activation
CHRNA1	2e-05	cation transmembrane transport/muscle cell cellular homeostasis
DAB2	1e-05	Wnt signaling pathway/apoptotic process
EEF1A1	2e-05	cellular response to epidermal growth factor stimulus/regulation of chaperone-mediated autophagy
EZR	3e-05	actin cytoskeleton reorganization/actin filament bundle assembly
HLA-DMB	0.00045	MHC class II protein complex assembly/antigen processing and presentation of exogenous peptide antigen via MHC class II
HLA-DRA	0.00062	T cell costimulation/T cell receptor signaling pathway
LAMB2	0.00034	Schwann cell development/astrocyte development
LAPTM5	5e-05	transport
MYH3	1e-05	ATP metabolic process/actin filament-based movement
MYH8	1e-05	ATP metabolic process/muscle contraction
MYL4	1e-05	cardiac muscle contraction/muscle filament sliding
MYL5	3e-05	muscle contraction/regulation of muscle contraction
S100A10	1e-05	establishment of protein localization to plasma membrane/membrane budding
S100A11	2e-05	cell-cell adhesion/negative regulation of DNA replication
S100A13	2e-05	cytokine secretion/interleukin-1 alpha secretion
S100A4	0.00041	epithelial to mesenchymal transition/positive regulation of I-kappaB kinase/NF-kappaB signaling
TMSB10	0.0001	actin filament organization

Table 9: A subset of selected genes from applying the cPCH test for $r = m = 3$ using Fisher's combining function with Storey's procedure at nominal level $q = 0.1$ on the follow-up study design described in Section 4 screening on GDS1956 then using the remaining studies (GDS214, GDS563, GDS3027) as follow-up studies. As desired, many of the genes discovered all correspond to various muscle functions. Notably, the genes MYH3, MYH8, MYL4, and MYL5 are known genetic markers for DMD.